Πανεπιστήμιο Μακεδονίας
Οικονομικών και Κοινωνικών Επιστημών
Τμήμα Εφαρμοσμένης Πληροφορικής

Πειραματικοί Σχεδιασμοί στην Ανάλυση Δεδομένων

Διδακτορική Διατριβή

Γεώργιος Χ. Μενεξές

Τριμελής Συμβουλευτική Επιτροπή:

Επιβλέπων: Καθηγητής Γιάννης Παπαδημητρίου

Μέλη: Καθηγητής Αναστάσιος Κάτος

Αν. Καθηγητής Δημήτριος Παπαναστασίου

Θεσσαλονίκη, 2006

Η παρούσα έρευνα χρηματοδοτήθηκε από το Πρόγραμμα:

«ΗΡΑΚΛΕΙΤΟΣ: ΥΠΟΤΡΟΦΙΕΣ ΕΡΕΥΝΑΣ ΜΕ ΠΡΟΤΕΡΑΙΟΤΗΤΑ ΣΤΗ ΒΑΣΙΚΗ ΕΡΕΥΝΑ»

Πανεπιστήμιο Μακεδονίας
Οικονομικών και Κοινωνικών Επιστημών
Τμήμα Εφαρμοσμένης Πληροφορικής

Πειραματικοί Σχεδιασμοί στην Ανάλυση Δεδομένων

Διδακτορική Διατριβή

Γεώργιος Χ. Μενεξές

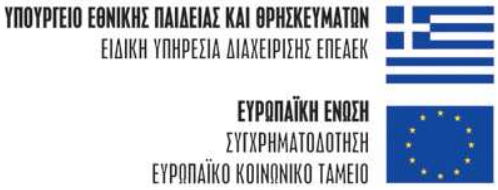
Τριμελής Συμβουλευτική Επιτροπή:

Επιβλέπων: Καθηγητής Γιάννης Παπαδημητρίου

Μέλη: Καθηγητής Αναστάσιος Κάτος

Αν. Καθηγητής Δημήτριος Παπαναστασίου

Θεσσαλονίκη 2006



Το έργο «ΗΡΑΚΛΕΙΤΟΣ: Υποτροφίες Έρευνας στο Πανεπιστήμιο Μακεδονίας» - Υποέργο «Πειραματικοί Σχεδιασμοί στην Ανάλυση Δεδομένων» υλοποιείται στα πλαίσια της Κατηγορίας Πράξεων 2.2.3.β. «Υποτροφίες Έρευνας με προτεραιότητα στη Βασική Έρευνα», Μέτρο 2.2 «Αναμόρφωση Προγραμμάτων Σπουδών – Διεύρυνση Τριτοβάθμιας Εκπαίδευσης», Ενέργεια 2.2.3 «Προγράμματα Μεταπτυχιακών Σπουδών - Έρευνα - Υποτροφίες», εκτελείται στα πλαίσια του Επιχειρησιακού Προγράμματος Εκπαίδευσης και Αρχικής Επαγγελματικής Κατάρτισης ΙΙ (ΕΠΕΑΕΚ ΙΙ) και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση [3ο Κοινοτικό Πλαίσιο Στήριξης κατά 75% Κοινοτική Συμμετοχή (ΕΚΤ) και 25% Εθνικοί Πόροι]

Στη Θεοδώρα, στο Χρήστο και στον Κωνσταντίνο.

Ευχαριστίες

Η παρούσα διατριβή είναι αποτέλεσμα ερευνητικού έργου πέντε ετών στο Τμήμα Εφαρμοσμένης Πληροφορικής του Πανεπιστημίου Μακεδονίας και χρηματοδοτήθηκε, μετά από αξιολόγηση της αντίστοιχης ερευνητικής πρότασης, με υποτροφία από το Πρόγραμμα «ΗΡΑΚΛΕΙΤΟΣ: ΥΠΟΤΡΟΦΙΕΣ ΕΡΕΥΝΑΣ ΜΕ ΠΡΟΤΕΡΑΙΟΤΗΤΑ ΣΤΗ ΒΑΣΙΚΗ ΕΡΕΥΝΑ». Η οικονομική ενίσχυση της υποτροφίας μού παρείχε τα μέσα και τη δυνατότητα: α) να διαθέσω τον απαιτούμενο χρόνο μελέτης και πειραματισμών στην επίτευξη των στόχων της διατριβής και β) να προβάλω με ανακοινώσεις, σε Πανελλήνια και Διεθνή Συνέδρια, και με δημοσιεύσεις, σε ελληνικά και διεθνή επιστημονικά περιοδικά, μέρος των αποτελεσμάτων της ερευνητικής μου προσπάθειας. Βέβαια, η ολοκλήρωση της διατριβής θα ήταν αδύνατη χωρίς τη βοήθεια αρκετών ανθρώπων.

Καταρχήν θέλω να εκφράσω την ευγνωμοσύνη μου στον επιβλέποντα Καθηγητή Γιάννη Παπαδημητρίου για την ακριβή καθοδήγηση, τη συνεχή ενθάρρυνση και τη συνεπή παρακολούθηση του ερευνητικού μου έργου. Οι σχεδόν καθημερινές συναντήσεις μας έθεσαν τις βάσεις μιας συστηματικής και προγραμματισμένης ερευνητικής εργασίας επικεντρωμένης σε στόχους. Ο Καθηγητής Γιάννης Παπαδημητρίου, εκτός από τη “μύηση” στο συναρπαστικό “κόσμο” της Ανάλυσης Δεδομένων, μου έδωσε την απαραίτητη ελευθερία και ανεξαρτησία, ώστε να μπορώ να εκφράσω τις ερευνητικές μου ανησυχίες και προβληματισμούς. Οι συχνά ακατέργαστες και αυθόρμητες ιδέες μου, κάτω από τη δημιουργική και με κατανόηση καθοδήγησή του, μετασχηματίστηκαν σε συγκεκριμένα ερευνητικά προβλήματα της Ανάλυσης Δεδομένων, πολλά από τα οποία αποτελούν θεματικές ενότητες της παρούσας μελέτης. Επίσης, τον ευχαριστώ ολόψυχα για την αμέριστη ηθική συμπαράσταση και τις ατελείωτες ώρες εποικοδομητικής συνεργασίας. Μου στάθηκε, ως δάσκαλος και ως άνθρωπος, πολύτιμος σύμμαχος. Χωρίς τη βοήθειά του η εκπόνηση της παρούσας διατριβής θα ήταν αδύνατη.

Ευχαριστώ θερμά τα μέλη της συμβουλευτικής επιτροπής, τον Καθηγητή Αναστάσιο Κάτο και τον Αν. Καθηγητή Δημήτριο Παπαναστασίου, για τη συμβολή και τις πολύτιμες υποδείξεις τους στην ολοκλήρωση και βελτίωση της διατριβής. Θέλω να

ευχαριστήσω και τα μέλη της Γενικής Συνέλευσης του Τμήματος Εφαρμοσμένης Πληροφορικής, τα οποία στις 14/11/2001 έκαναν αποδεκτή την αίτησή μου για την εκπόνηση της διδακτορικής διατριβής υπό την επίβλεψη του Καθηγητή Γιάννη Παπαδημητρίου.

Τα λόγια δεν αρκούν για να περιγράψουν τα αισθήματα μου για το φίλο και συνεργάτη Άγγελο Μάρκο, Υποψήφιο Διδάκτορα του Τμήματος Εφαρμοσμένης Πληροφορικής, για την ανεκτίμητη βοήθειά του στην ολοκλήρωση της διδακτορικής διατριβής. Θέλω να εκφράσω την εκτίμηση και τον θαυμασμό μου για την ετοιμότητα και την αποτελεσματικότητά του στην προγραμματιστική υλοποίηση κάθε "απίθανης" ιδέας και έμπνευσής μου. Τον ευχαριστώ ολόψυχα για την υπομονή και την προθυμία του.

Θέλω να πω ένα μεγάλο ευχαριστώ (για μια ακόμη φορά) στο φίλο, συνεργάτη, δάσκαλο και θείο μου, Ανδρέα Οικονόμου, Μαθηματικό και Ψυχολόγο, Καθηγητή της Α.Σ.ΠΑΙ.Τ.Ε. Θεσσαλονίκης, για τη πολύτιμη βοήθεια του όχι μόνο στη φιλολογική και δομική βελτίωση των κειμένων αλλά και για την στήριξή του στις "δύσκολες στιγμές".

Κατά τη διάρκεια των τελευταίων τεσσάρων ετών συμμετείχα με ανακοινώσεις σε δύο εξειδικευμένα Διεθνή Συνέδρια (στο International Conference on Correspondence Analysis and Related Methods 2003, CARME 2003, in Barcelona, 29 June-2 July 2003 και στο RC33 Sixth International Conference on Social Science Methodology, in Amsterdam, August 16-20, 2004). Στα συνέδρια αυτά μου δόθηκε η ευκαιρία να γνωρίσω και να ανταλλάξω απόψεις με διακεκριμένους ερευνητές της Ανάλυσης Δεδομένων. Ιδιαίτερα θέλω να ευχαριστήσω τους Καθηγητές Michael Greenacre (Universitat Pombeu Fabra, Barcelona, Spain), Ludovic Lebart (CNRS-ENST, Paris, France) και Henry Rouanet (CNRS & Université René Descartes, Paris, France) για την ενθάρρυνση και τις συμβουλές τους.

Θέλω, επίσης, να εκφράσω τις ευχαριστίες και την εκτίμησή μου στο Διδάκτορα Θεόφιλο Παπαδημητρίου, Λέκτορα του Δημοκρίτειου Πανεπιστημίου Θράκης, για τις υποδείξεις και τη βοήθειά του στην ανάπτυξη της μεθοδολογίας που παρουσιάζεται στην Ενότητα 3.3 του Κεφαλαίου 3.

Ευχαριστώ την αδελφή μου Αθανασία για τη βοήθειά της στον έλεγχο συνέπειας των βιβλιογραφικών αναφορών.

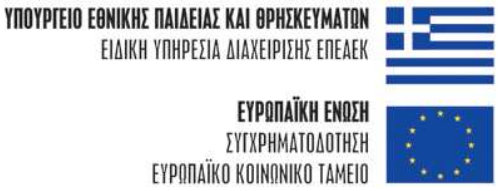
Ευχαριστώ τους γονείς μου για την πολύπλευρη συμπαράστασή τους.

Τέλος, ευχαριστώ τη σύζυγό μου Θεοδώρα που ήταν πάντα εκεί…

Ως συγγραφέας της παρούσας διατριβής, είμαι αποκλειστικά υπεύθυνος για λάθη, παραλείψεις ή ασάφειες που ενδεχομένως υπάρχουν στο κείμενο.

Γέωργιος Χ. Μενεξές

Θεσσαλονίκη, Δεκέμβριος 2006



Το έργο «ΗΡΑΚΛΕΙΤΟΣ: Υποτροφίες Έρευνας στο Πανεπιστήμιο Μακεδονίας» - Υποέργο «Πειραματικοί Σχεδιασμοί στην Ανάλυση Δεδομένων» υλοποιείται στα πλαίσια της Κατηγορίας Πράξεων 2.2.3.β. «Υποτροφίες Έρευνας με προτεραιότητα στη Βασική Έρευνα», Μέτρο 2.2 «Αναμόρφωση Προγραμμάτων Σπουδών – Διεύρυνση Τριτοβάθμιας Εκπαίδευσης», Ενέργεια 2.2.3 «Προγράμματα Μεταπτυχιακών Σπουδών - Έρευνα - Υποτροφίες», εκτελείται στα πλαίσια του Επιχειρησιακού Προγράμματος Εκπαίδευσης και Αρχικής Επαγγελματικής Κατάρτισης ΙΙ (ΕΠΕΑΕΚ ΙΙ) και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση [3ο Κοινοτικό Πλαίσιο Στήριξης κατά 75% Κοινοτική Συμμετοχή (ΕΚΤ) και 25% Εθνικοί Πόροι]

Πίνακας Περιεχομένων

Ευρετήριο Πινάκων

Ευρετήριο Διαγράμματων

Ευρετήριο Σχημάτων

Ευρετήριο Εικόνων

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

1.1 Σκοπός και Ειδικοί Στόχοι της Διατριβής

Βασικός σκοπός της παρούσας διατριβής είναι η εισαγωγή και εφαρμογή της Παραγοντικής Ανάλυσης των Αντιστοιχιών (ΠΑΑ), όπως αυτή αναδεικνύεται και εφαρμόζεται στο μεθοδολογικό πλαίσιο της Γαλλικής και Ολλανδικής Σχολής Ανάλυσης Δεδομένων, σε κατηγορικά δεδομένα, όπου υπάρχει διάκριση μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών. Η διάκριση αυτή μπορεί να είναι: α) δομική, δηλαδή να καθορίζεται από το μηχανισμό παραγωγής των δεδομένων, όπως συμβαίνει στους πειραματικούς σχεδιασμούς, ή β) εννοιολογική, δηλαδή να υπαγορεύεται θεωρητικά μέσα σε συγκεκριμένο γνωστικό ερευνητικό πεδίο, γεγονός που συχνά παρατηρείται στις δειγματοληπτικές *ex post facto* έρευνες. Το πείραμα θεωρείται εν γένει ως η πιο ενδεδειγμένη μέθοδος για την εξέταση της επίδρασης μίας ή περισσότερων μεταβλητών σε άλλες, κάτω από προκαθορισμένες και ελεγχόμενες συνθήκες. Τα αποτελέσματα είναι δυνατό να επιβεβαιώσουν σχέσεις «αιτίας – αποτελέσματος», οι οποίες, στη συνέχεια, μπορούν να οδηγήσουν και στην κατασκευή μοντέλων - υποδειγμάτων πρόβλεψης. Στο πλαίσιο της εργασίας ιδιαίτερη έμφαση δίνεται στην περίπτωση των πειραματικών σχεδιασμών και των αντίστοιχων πινάκων δεδομένων που θα δοθούν ως “είσοδος” στην ανάλυση.

Η ΠΑΑ εφαρμόζεται για τη διερεύνηση της σχέσης μεταξύ δύο ή περισσότερων κατηγορικών μεταβλητών χωρίς *a priori* υποθέσεις και ελάχιστες τεχνικές προϋποθέσεις. Βρίσκει πληθώρα εφαρμογών σε όλα σχεδόν τα ερευνητικά επιστημονικά πεδία και είναι αρκετά ευέλικτη σε ό,τι αφορά τη μορφή των πινάκων δεδομένων στους οποίους μπορεί να εφαρμοστεί. Η μέθοδος έχει συνήθως περιγραφικό και διερευνητικό χαρακτήρα και δεν συνοδεύεται από ελέγχους στατιστικής σημαντικότητας. Το φιλοσοφικό πλαίσιο, στο οποίο αναπτύχθηκε, δεν αφήνει και πολλά περιθώρια, κυρίως από επιστημολογική σκοπιά, για στατιστικούς ελέγχους υποθέσεων, όπως αυτοί εφαρμόζονται στο πλαίσιο της Επαγωγικής

Στατιστικής. Είναι προσηλωμένη στα δεδομένα καθαυτά, αφήνοντας, προσωρινά τουλάχιστον, στο περιθώριο όχι μόνο το θεωρητικό πλαίσιο που οδήγησε στη συγκρότησή τους αλλά και την ίδια τη διαδικασία παραγωγής και συλλογής τους. Το ενδιαφέρον επικεντρώνεται στις ιδιότητες και όχι στις ποσοτικές μετρήσεις που έχουν συγκεντρωθεί με βάση κάποιο προσχεδιασμένο πείραμα, όπου τα αντίστοιχα δεδομένα εξαρτώνται από γνωστές και ελεγχόμενες μεταβλητές. Τα διαθέσιμα δεδομένα αντιμετωπίζονται σαν να προέρχονται από ολόκληρο τον υπό εξέταση πληθυσμό, ανεξάρτητα με το εάν αυτά προέρχονται από ένα δείγμα του. Οι μεταβλητές αντιμετωπίζονται συμμετρικά και ισότιμα χωρίς διάκριση σε εξαρτημένες και ανεξάρτητες. Αυτό έχει ως αποτέλεσμα η μέθοδος να μην έχει βρει τη θέση της σε εμπειρικές έρευνες, όπου επιζητείται η επιβεβαίωση ή όχι σχέσεων αιτίας - αποτελέσματος, οι οποίες, κατά παράδοση, ελέγχονται με πειραματικούς σχεδιασμούς κάνοντας χρήση μεθόδων της Επαγωγικής Στατιστικής. Βέβαια, σε πολλά ερευνητικά πεδία (π.χ. Οικονομία, Οικολογία και Κοινωνιολογία) είναι σχεδόν αδύνατο να ελεγχθεί η παραγωγή κατάλληλων πειραματικών δεδομένων. Στην περίπτωση αυτή, οι σχέσεις αιτίας – αποτελέσματος είναι δυνατό να τεκμηριωθούν, μέσω συσχετιστικών δειγματοληπτικών ερευνών, μόνο θεωρητικά και όχι πειραματικά. Έτσι, σε ένα δεύτερο επίπεδο, η παρούσα μελέτη στοχεύει στο να διερευνήσει το κατά πόσο και με ποιο τρόπο μέθοδοι της Επαγωγικής Στατιστικής μπορούν να εφαρμοστούν ή/και να συνδυαστούν με την ΠΑΑ όχι μόνο σε πειραματικές αλλά και σε δειγματοληπτικές έρευνες.

Τέλος, πιστεύουμε ότι πάντα υπάρχει ενδιαφέρον από την επιστημονική κοινότητα, που είναι αποδέκτης - καταναλωτής των εκροών της μεθόδου, για βελτίωση ή καθιέρωση νέων διαδικασιών, δεικτών ή/και ελέγχων (στατιστικών ή εμπειρικών), οι οποίοι θα συμβάλουν: α) στη βελτίωση της ερμηνείας των αποτελεσμάτων, β) στην αξιολόγηση της ποιότητας και της ποσότητας της παραγόμενης πληροφορίας, γ) στην αξιολόγηση της πρακτικής ή κλινικής σημαντικότητας των ευρημάτων και δ) στην ενίσχυση της αξιοπιστίας και εγκυρότητας των συμπερασμάτων. Αυτό είναι ιδιαίτερα σημαντικό αν αναλογιστούμε ότι η ΠΑΑ αφήνει στους ερευνητές – χρήστες τη φροντίδα και την ευθύνη να ερμηνεύσουν τα αποτελέσματα και να εξάγουν συμπεράσματα. Θέτοντας ως στόχους τα παραπάνω τέσσερα σημεία, διαπραγματευόμαστε, στην εργασία, αντίστοιχα ειδικά θέματα και καταλήγουμε σε συγκεκριμένες προτάσεις. Πριν, όμως, αναφερθούμε αναλυτικά στη δομή και στη

συνεισφορά της διατριβής κρίνουμε σκόπιμο να παρουσιάσουμε πρώτα τους κυριότερους άξονες του εννοιολογικού πλαισίου της μελέτης. Θεωρούμε ότι με τον τρόπο αυτό θα καταστεί πιο αποτελεσματική η επικοινωνία με τον αναγνώστη σε ότι αφορά την οριοθέτηση του υπό εξέταση θέματος και της συνοχής των επιμέρους ζητημάτων – προβλημάτων που ανακύπτουν για την εκπλήρωση του σκοπού και των ειδικών στόχων της παρούσας διατριβής.

1.2 Πειραματικοί Σχεδιασμοί

Σύμφωνα με τον Kirk (1995), ο όρος «Πειραματικός Σχεδιασμός» (*Experimental Design*) αναφέρεται: α) σε μία διαδικασία ή μεθοδολογικό σχέδιο σύμφωνα με το οποίο οι διαθέσιμες πειραματικές μονάδες (π.χ. αντικείμενα ή υποκείμενα) θα ενταχθούν σε διάφορες πειραματικές συνθήκες (μεταχειρίσεις ή αγωγές) και β) στην κατάλληλη στατιστική ανάλυση των δεδομένων σύμφωνα με το σχέδιο αυτό. Ο Ronald Fisher, στο πρώτο τέταρτο του 20^{ου} αιώνα, έθεσε τις βάσεις των βιομετρικών πειραματικών σχεδιασμών και όχι μόνον (Pearce 1979, Preece 1990, Κίτσος 1994). Σημαντική, επίσης, ήταν μετέπειτα και η συνεισφορά του Frank Yates (Sprent, 1973), με αποτέλεσμα, το 1950, οι Πειραματικοί Σχεδιασμοί να αποτελούν ήδη ένα καλά τεκμηριωμένο πεδίο έρευνας της Στατιστικής (Chernoff, 1999). Έκτοτε η σχετική θεωρία και μεθοδολογία αποτελεί ερευνητικό αντικείμενο με συνεχή εξέλιξη και επέκταση. Χαρακτηριστικά αναφέρουμε τις περιπτώσεις των βέλτιστων πειραματικών σχεδιασμών (βλέπε Elfving 1952, Kiefer 1974 και 1959, Atkinson & Fedorov 1975, Fedorov & Khabarov 1986, Φαρμάκης 1987, Atkinson & Donev 1989, Κίτσος 1994, Atkinson 1996) και των σχεδιασμών επιφανειών απόκρισης (βλέπε Smith 1918, Box & Draper 1975, 1963 και 1959, Κίτσος 1994, Montgomery 1997, Kuehl 2000).

Πειράματα σχεδιάζονται και εκτελούνται σε όλους σχεδόν τους επιστημονικούς τομείς (Κίτσος 1994, Montgomery 1999) με σκοπό τη μελέτη της επίδρασης μιας ή περισσότερων μεταβλητών πάνω σε κάποιες άλλες, ελέγχοντας ταυτόχρονα και άλλους τοπικούς εξωγενείς, σε σχέση με το υπό εξέταση φαινόμενο, παράγοντες (Wuebben 1968, Pearce 1979). Βέβαια, οι πειραματικές διαδικασίες και μέθοδοι που εφαρμόζονται εξαρτώνται κάθε φορά από τις συνθήκες που επιβάλει το επιστημονικό πεδίο στο πλαίσιο του οποίου διενεργείται το πείραμα. Για παράδειγμα, άλλες

μέθοδοι εφαρμόζονται στη Γεωπονία και άλλες στο Βιομηχανικό Έλεγχο Ποιότητας (Hamaker, 1955).

Με βάση τα πορίσματα σχετικής βιβλιογραφίας από ποικίλες γνωστικές περιοχές (Cox 1958 και 1950, Chapin 1950, Pearce 1979, Κάτος 1986, Λελάκης 1987, Brown & Melamed 1990, Preece 1990, Polgar & Thomas 1992, Chaloner & Verdinelli 1995, Daniel 1995, Kirk 1995, Mendenhall & Sincich 1996, Cohen & Manion 1997, Κυριαζή 1998, Mertens 1998, Ταγαράς 2001, Kuehl 2000) ο πειραματικός σχεδιασμός, στο πλαίσιο της Στατιστικής και της Θεωρίας Πιθανοτήτων, θα πρέπει να υπακούει σε τρεις βασικές αρχές: α) της σύγκρισης των αγωγών, β) της τυχαίας επιλογής ή της τυχαίας ανάθεσης των πειραματικών μονάδων στις αγωγές και γ) της επανάληψης των μετρήσεων για όλους ή μερικούς από τους δυνατούς συνδυασμούς των αγωγών. Αν συνθέσουμε τις μεθοδολογικές προσεγγίσεις που προτείνουν αρκετοί ερευνητές (Cochran & Cox 1953, Cox 1958, Langley 1971, Pearce 1979, Gomez & Gomez 1984, Steel & Torrie 1986, Lipsey 1990, Mead & Curnow 1990, Kirk 1995, Zar 1996, Montgomery 1997, Lewis, Mathieu & Phan-Tan-Luu 1999, Kuehl 2000) μπορούμε να διακρίνουμε μια δομημένη πορεία στο σχεδιασμό ενός πειράματος, η οποία περιλαμβάνει γενικά τις παρακάτω συσχετιζόμενες δραστηριότητες:

1) Τη σαφή διατύπωση μιας ή περισσότερων στατιστικών υποθέσεων, οι οποίες αντιστοιχούν αμφιμονοσήμαντα σε αποδεκτές επιστημονικές ή ερευνητικές υποθέσεις που μπορούν να ελεγχθούν πειραματικά.

2) Τον καθορισμό: α) των πειραματικών συνθηκών (ανεξάρτητες μεταβλητές ή παράγοντες), β) των μετρήσεων ή αποκρίσεων που θα πρέπει να καταγραφούν (εξαρτημένες μεταβλητές) και γ) των εξωγενών συνθηκών που λειτουργούν ως πηγές ή αλλιώς ως μεταβλητές θορύβου, οι οποίες όμως μπορούν να ελεγχθούν ως ένα βαθμό. Οι μεταβλητές αυτές αποτελούν ανεπιθύμητες πηγές μεταβλητότητας που επηρεάζουν τις εξαρτημένες μεταβλητές και συνήθως δεν έχουν άμεσο ερευνητικό ενδιαφέρον. Όλες οι μεταβλητές που συμμετέχουν στο πείραμα (ανεξάρτητες, εξαρτημένες και θορύβου), σε σχέση με την κλίμακα μέτρησής τους, μπορεί να είναι ποιοτικές (κατηγορικές) ή ποσοτικές. Οι τιμές (επίπεδα) των ανεξάρτητων μεταβλητών μπορεί να είναι είτε καθορισμένες από πριν είτε να έχουν επιλεγεί με τυχαία δειγματοληψία από ένα πλήθος δυνατών τιμών.

3) Το σαφή καθορισμό του πληθυσμού, από τον οποίο θα γίνει η δειγματοληψία των πειραματικών μονάδων και τον υπολογισμό του ελάχιστου πλήθους των πειραματικών μονάδων που απαιτούνται (αριθμός επαναλήψεων και μέγεθος δείγματος).

4) Τον καθορισμό της διαδικασίας τυχαιοποίησης ή δειγματοληψίας, σύμφωνα με την οποία οι πειραματικές μονάδες θα υποβληθούν στις αγωγές. Η διαδικασία τυχαιοποίησης καθορίζει και τον τύπο (ονομασία) του πειραματικού σχεδιασμού. Ως παράδειγμα αναφέρουμε την περίπτωση του Ισορροπημένου και Πλήρως Τυχαιοποιημένου Πειραματικού Σχεδίου με έναν παράγοντα (Κάτος 1986, Κίτσος 1994). Σύμφωνα με το σχεδιασμό αυτό, N σε πλήθος διαθέσιμες πειραματικές μονάδες τυχαιοποιούνται πλήρως (π.χ. με κλήρωση) στα λ επίπεδα (στάθμες) του παράγοντα, με τον περιορισμό ότι κάθε επίπεδο θα περιλαμβάνει N / λ πειραματικές μονάδες. Βασική επιδίωξη είναι κάθε μία από τις πειραματικές μονάδες να έχει την ίδια πιθανότητα να ανατεθεί σε κάποιο από τα επίπεδα του παράγοντα.

5) Τον καθορισμό της στατιστικής ανάλυσης που θα εφαρμοστεί στα πειραματικά δεδομένα ανάλογα με τον τύπο του πειραματικού σχεδιασμού. Κατά παράδοση, και ιδιαίτερα στην περίπτωση που οι εξαρτημένες μεταβλητές είναι ποσοτικές, η στατιστική ανάλυση των πειραματικών δεδομένων πραγματοποιείται στο ευρύ μεθοδολογικό πλαίσιο της Ανάλυσης Διασποράς (Κίτσος, 1994) με προσαρμογή των δεδομένων σε κατάλληλα Γενικά Γραμμικά Μοντέλα – Υποδείγματα (βλέπε Μπόρα-Σέντα & Μωϋσιάδης 1992, Καρακώστας 1993, Kirk 1995, Stapleton 1995, Mendenhall & Sincich 1996, Kuehl 2000, Rencher 2000, Rao 2002, Kutner *et al*. 2005). Οι ερευνητικές υποθέσεις που προκύπτουν από Πειραματικούς Σχεδιασμούς ελέγχονται σχεδόν αποκλειστικά με μεθόδους της Επαγωγικής Στατιστικής, η εφαρμογή των οποίων απαιτεί την ικανοποίηση ορισμένων θεωρητικών και τεχνικών προϋποθέσεων, οι οποίες σπάνια ικανοποιούνται στην πράξη (Cohen & Cohen 1983, Μπεχράκης 1999). Όμως, οι μέθοδοι αυτοί, αν δεν εφαρμοστούν με τον ενδεδειγμένο τρόπο, εγκυμονούν κινδύνους για στοχαστικά και λογικά σφάλματα καθώς και για παρανοήσεις (Harris 2001, Huck 2000α και 2000β, Μενεξές & Οικονόμου 2002).

Ένα πείραμα διαφέρει από μία συγκριτική δειγματοληπτική έρευνα *ex post facto* στο ότι ο ερευνητής που διεξάγει το πείραμα παρεμβαίνει ενεργά, επιλέγοντας,

ελέγχοντας και επιβάλλοντας άμεσα τις αγωγές που τον ενδιαφέρουν στις πειραματικές μονάδες. Στις δειγματοληπτικές έρευνες, ο ερευνητής απλά παρατηρεί ή μετρά τα διάφορα χαρακτηριστικά και γενικά την κατάσταση των δειγματοληπτικών μονάδων, χωρίς να είναι σε θέση να αλλάξει αυτή την κατάσταση με κάποια ειδική μεταχείριση. Στην περίπτωση αυτή, οι δειγματοληπτικές έρευνες μπορούν μόνο να παρέχουν χρήσιμα στοιχεία για τη διατύπωση υποθέσεων, οι οποίες θα ελεγχθούν, στη συνέχεια, με πειραματικές μεθόδους. Η McKinlay (1975) αναφέρει ότι η έλλειψη ή η αδυναμία τυχαιοποίησης των δειγματοληπτικών μονάδων σε ομάδες σύγκρισης είναι το σημαντικότερο κριτήριο για το χαρακτηρισμό μιας εμπειρικής μελέτης ως πειραματικής ή μη.

Τα πειράματα είναι η πιο ενδεδειγμένη μέθοδος για την εξέταση της επίδρασης μιας μεταβλητής σε μία άλλη (Wuebben 1968, Mertens 1998, Bryman & Cramer 1999, De Leeuw 2005γ) στο πλαίσιο του χώρου και του χρόνου που διενεργείται ο πειραματισμός και αφού ληφθούν υπόψη τα ιδιαίτερα χαρακτηριστικά των πειραματικών μονάδων (Κυριαζή, 1998). Ο ερευνητής, επιβάλλοντας τις αγωγές και ελέγχοντας άλλες επιρροές, οι οποίες μπορούν να επηρεάσουν την εγκυρότητα των αποτελεσμάτων και των συμπερασμάτων (βλέπε Polgar & Thomas 1992, Churchill 1995, Kinnear & Taylor 1996, Cohen & Manion 1997, Mertens 1998), μπορεί να εντοπίσει την αιτία και να εκτιμήσει το μέγεθος της επίδρασης. Αντίθετα, μία δειγματοληπτική έρευνα μπορεί να δείξει ότι δύο μεταβλητές συσχετίζονται, αλλά δεν μπορεί να δείξει με ποιον τρόπο η μία μεταβλητή επηρεάζει την άλλη ή να δώσει μια πειστική μαρτυρία αιτιότητας (Cohen & Cohen, 1983).

Με το στατιστικό σχεδιασμό των πειραμάτων, η λογική του πειραματισμού και τα μαθηματικά των Πιθανοτήτων και της Στατιστικής συνδυάζονται για να δώσουν μία εν γένει αποδεκτή σχέση αιτίας και αποτελέσματος.

1.3 Η Ανάλυση Δεδομένων

Η «Ανάλυση Δεδομένων» (*Data Analysis* στα αγγλικά και *L' Analyse des Données* στα γαλλικά) αποτελεί κλάδο της ενότητας των μεθόδων της Πολυδιάστατης Στατιστικής Ανάλυσης (Καραπιστόλης 1999, Μπεχράκης 1999, Παπαδημητρίου 2006, 2002 και 1994) και περιλαμβάνει, σύμφωνα με τους Deville & Malinvaud

(1983), τρεις βασικές οικογένειες μεθόδων (βλέπε Benzécri 1992, Lebart, Morineau & Warwick 1984): α) την ΠΑΑ (διμεταβλητή και πολυμεταβλητή), β) την Ανάλυση σε Κύριες Συνιστώσες και γ) την Ταξινόμηση σε Αύξουσα Ιεραρχία. Ιδιαίτερο χαρακτηριστικό των μεθόδων αυτών είναι η συμμετρική αντιμετώπιση των μεταβλητών, όπου δεν υπάρχει διάκριση μεταξύ εξαρτημένων και ανεξάρτητων (Hair *et al*., 1995). Άλλες γνωστές και διαδεδομένες σε εφαρμογή μέθοδοι της Ανάλυσης Δεδομένων είναι η Παραγοντική Ανάλυση, η Ανάλυση Κανονικοποιημένης Συσχέτισης, η Πολυδιάστατη Κλιμάκωση και η Διακρίνουσα Ανάλυση (βλέπε Dillon & Goldstein 1984, Johnson & Wichern 1992, Hair *et al*. 1995, Sharma 1996, Tacq 1997, Johnson 1998, Stevens 2002). Βασικός σκοπός των μεθόδων είναι να αναδείξουν και να περιγράψουν λανθάνουσες δομές που ενδεχομένως εμπεριέχονται σε πολυδιάστατους πίνακες δεδομένων. Αυτό επιτυγχάνεται μέσα από διαδικασίες αλλαγής και ελάττωσης των διαστάσεων του αρχικού χώρου, στον οποίο το υπό εξέταση φαινόμενο μπορεί να περιγραφεί. Οι νέες διαστάσεις, οι οποίες δομούνται συνήθως από πολύπλοκες σχέσεις μεταξύ των διαθέσιμων μετρήσεων, ερμηνεύονται τελικά ως νέες σύνθετες μεταβλητές ή παράγοντες (Dillon & Goldstein 1984, Παπαδημητρίου 1994). Οι μέθοδοι, σε ένα πρώτο επίπεδο, δεν απαιτούν την *a priori* παραδοχή ύπαρξης κάποιας θεωρητικής κατανομής ή κάποια υπόθεση σχετικά με τις παραμέτρους του υπό εξέταση πληθυσμού ή πληθυσμών, δηλαδή την ύπαρξη κάποιου στοχαστικού μοντέλου – υποδείγματος (Benzécri, 1991). Σύμφωνα με τον Καραπιστόλη (1999),

> "*Η ανάγκη λοιπόν να μη θεωρείται εκ των προτέρων ότι ένα φαινόμενο ακολουθεί κάποιο συγκεκριμένο νόμο, οδήγησε στην εφαρμογή νέων στατιστικών μη παραμετρικών μεθόδων, κάτω από την ονομασία Ανάλυση Δεδομένων ή όπως αλλιώς μπορεί να την αποκαλέσουμε* ***Στατιστική δίχως μοντέλα***." (σ. 21).

Κάτω από αυτή τη θεώρηση, η Ανάλυση Δεδομένων φαίνεται να εκφράζει μια νέα προσέγγιση στη στατιστική συμπερασματολογία, η οποία έρχεται σε αντίθεση με την κλασική αγγλοσαξονική παράδοση του στατιστικού ελέγχου υποθέσεων (Αθανασιάδης, 1995). Μάλιστα, ο Gras (1995), φθάνοντας ίσως στην υπερβολή, μιλάει τελικά για επιστημολογική ρήξη της Ανάλυσης Δεδομένων με την Κλασική Στατιστική.

Για τους Αγγλοσάξονες ο όρος Ανάλυση Δεδομένων δηλώνει μια προσέγγιση των στατιστικών αναλύσεων με κέντρο ενδιαφέροντος και ιδιαίτερη προσήλωση στα δεδομένα καθαυτά, αφήνοντας, προσωρινά τουλάχιστον, στο περιθώριο το θεωρητικό πλαίσιο που οδήγησε στη συγκρότησή τους (Αθανασιάδης, 1995). Η θεώρηση αυτή οδήγησε στην ανάπτυξη των τεχνικών της Πολυδιάστατης Κλιμάκωσης (βλέπε Kruskal & Shepard 1974, Kruskal & Wish 1978, Hair *et al*. 1995) στις Η.Π.Α. και της μεθόδου της ΠΑΑ στη Γαλλία, κάτω από την καταλυτική επίδραση του Jean-Paul Benzécri (Deville & Malinvaud 1983, Greenacre 1993, Van Meter *et al*. 1994, Καραπιστόλης 1999, Μπεχράκης 1999, Παπαδημητρίου 2002 και 1994, Le Roux & Rouanet 2004). Για τον Benzécri η Ανάλυση Δεδομένων συνιστά Φιλοσοφία, η οποία απελευθερώνει τον ερευνητή από δεσμεύσεις, που ενδεχομένως επιβάλλουν εξωγενείς, σε σχέση με την έρευνα, παράγοντες, αφήνοντάς του τη φροντίδα και την ευθύνη να εξάγει ο ίδιος τις ερμηνείες των φαινομένων και τις συνέπειές τους.

Ο σημαντικότερος, ίσως, παράγοντας που οδήγησε στην υιοθέτηση και διάδοση των μεθόδων της Ανάλυσης Δεδομένων από ένα ευρύ φάσμα επιστημονικών πεδίων ήταν η ανάπτυξη και η γενικευμένη χρήση των Ηλεκτρονικών Υπολογιστών (Η/Υ) (Κουτσουπιάς 1999α και 1999β, Μπεχράκης 1999, Παπαδημητρίου 2002). Οι μέθοδοι απαιτούν πολύπλοκους αριθμητικούς υπολογισμούς που μόνο με τη βοήθεια Η/Υ είναι δυνατό, μέσα στα όρια της ανθρώπινης υπομονής, να επιτευχθούν, ιδιαίτερα όταν πρόκειται να αναλυθούν μεγάλα σύνολα δεδομένων. Στις μέρες μας, δημοφιλή εμπορικά στατιστικά πακέτα όπως το SAS (SAS Institute, 1999 και 1990), το BMDP (Moran & Gornbein 1988, BMDP Inc. 1992) και το SPSS (Norusis 1992α, SPSS Inc. 1998α, Meulman & Heiser 2004, Καρλής 2005), περιλαμβάνουν μεθόδους όπως η ΠΑΑ, η Ανάλυση σε Κύριες Συνιστώσες και η Ιεραρχική Ταξινόμηση. Ειδικότερα, η ΠΑΑ θεωρείται ως η πιο αποτελεσματική από τις μεθόδους της Ανάλυσης Δεδομένων για τη στατιστική επεξεργασία κατηγορικών μεταβλητών (Καραπιστόλης 1999, Κιοσέογλου 2002, Παπαδημητρίου 2004). Αποτελεί δεν το βασικό κορμό πάνω στον οποίο στηρίζονται οι μεθοδολογικές προσεγγίσεις των δύο σημαντικότερων «Σχολών» Ανάλυσης Δεδομένων, της Γαλλικής και της Ολλανδικής.

1.4 Σχολές Ανάλυσης Δεδομένων

Πριν αναφερθούμε στις Σχολές Ανάλυσης Δεδομένων, κρίνουμε σκόπιμο να παραθέσουμε μια σύντομη ιστορική ανασκόπηση σχετικά με την ανάπτυξη της ΠΑΑ. Κίνητρο για την αναδρομή αυτή δεν είναι μόνο η ικανοποίηση μιας νόμιμης ιστοριογραφικής περιέργειας, αλλά και η πεποίθηση ότι η παρακολούθηση της εσωτερικής εξέλιξης της μεθόδου συμβάλλει, ως ένα βαθμό, στην αιτιολόγηση της χρησιμότητας και της αναγκαιότητάς της τόσο σε πρακτικό όσο και σε θεωρητικό επίπεδο. Για περισσότερα στοιχεία, σχετικά με την ιστορική εξέλιξη της μεθόδου, παραπέμπουμε στους Van Rijckevorsel (1987), Greenacre (1984), Tenenhaus & Young (1985), Van Rijckevorsel και De Leeuw (1988), De Leeuw (1993), Van Meter *et al*. (1994), Nishsisato (1996 και 1980), Clausen (1998), Beh (2004) και Le Roux & Rouanet (2004).

Κατά τη διάρκεια του 20ου αιώνα η ΠΑΑ εμφανίστηκε και αναπτύχθηκε ανεξάρτητα και, σε ορισμένες περιπτώσεις, σχεδόν ταυτόχρονα σε αρκετές χώρες, όπως οι Η.Π.Α., η Μεγάλη Βρετανία, ο Καναδάς, η Γαλλία, η Ολλανδία και η Ιαπωνία (Clausen, 1998). Αυτή η παράλληλη ανάπτυξη είχε ως αποτέλεσμα να δημιουργηθούν διάφορες προσεγγίσεις και κατευθύνσεις (Σχολές), τόσο ως προς το θεωρητικό όσο και ως προς το αλγοριθμικό – υπολογιστικό υπόβαθρο της μεθόδου. Έτσι, η μέθοδος έγινε γνωστή με διαφορετικά ονόματα όπως Δυϊκή Κλιμάκωση (*Dual Scaling*), Αθροιστική Βαθμονόμηση (*Additive Scoring*), Βέλτιστη ή Άριστη Κλιμάκωση (*Optimal Scaling*), Ανάλυση Ομοιογένειας (*Homogeneity Analysis*) και άλλα (βλέπε Nishisato, 1980).

Σύμφωνα με τους Van Rijckevorsel και De Leeuw (1988) και Beh (2004), τα πρώτα μαθηματικά αποτελέσματα, με τα οποία μπορούν να συνδεθούν μεθοδολογικά στοιχεία της ΠΑΑ, παρουσιάστηκαν από τον Karl Pearson στην πρώτη δεκαετία του 20ου αιώνα. Η ΠΑΑ, ως μέθοδος ανάλυσης κατηγορικών δεδομένων, εμφανίστηκε στα μέσα της δεκαετίας του ’30 με την εργασία του Hirschfeld (1935) και με τις απαρχές της συνδέονται ονόματα και άλλων σημαντικών ερευνητών όπως των Richardson, Kuder, Horst, Fisher, Maung, Guttman και Burt (βλέπε Nishisato 1980, Greenacre 1984, Gauch 1995). Ο Παπαδημητρίου (2002) αναφέρει ότι αν και η πρώτη προσπάθεια μαθηματικής διατύπωσης της μεθόδου οφείλεται στον Η. Ο.

Hartley (ή Hirschfeld) ωστόσο σε μια πιο ολοκληρωμένη μορφή η μέθοδος παρουσιάστηκε από τον Guttman το 1941. Ο Greenacre (1984) παρατηρεί ότι ο Fisher και ο Guttman, σχεδόν ταυτόχρονα και ανεξάρτητα, παρουσίασαν σε θεωρητικό επίπεδο την ίδια ουσιαστικά μέθοδο αλλά σε διαφορετικό πλαίσιο προβληματικής και εφαρμογών: ο Fisher (1940) τη διμεταβλητή εκδοχή της μεθόδου στο χώρο της Βιομετρίας, ενώ ο Guttman (1941) την πολυμεταβλητή εκδοχή στο χώρο της Ψυχομετρίας. Σε κάθε περίπτωση, η μέθοδος δεν παρουσιάστηκε με τη σημερινή της διεθνή ονομασία, *Analyse Factoriel des Correspondances-A.F.C.* στα γαλλικά (*Correspondence Analysis–C.A.* στα αγγλικά), η οποία αποδίδεται στον Benzécri (Nishisato 1980, Van Rijckevorsel & De Leeuw 1988). Ο Benzécri, αφού συστηματοποίησε τη μαθηματική της θεμελίωση ήδη από τη δεκαετία του 1960 (Καραπιστόλης, 1999), την ανήγαγε τελικά σε ένα γενικό σύστημα στατιστικής ανάλυσης δεδομένων (Clausen, 1998).

Αξίζει να σημειωθεί ότι στις αρχές του 1950 παραλλαγές της μεθόδου γνωρίζουν σημαντική ανάπτυξη και διάδοση στην Ιαπωνία, χάρη στις εργασίες του Hayashi (βλέπε Van Rijckevorsel & De Leeuw 1988, SAS Institute 1990, Greenacre & Blasius 1994, Nishisato 1994 και 1980) και σχεδόν ταυτόχρονα στην Αγγλία από τον Burt (1950). Αργότερα, στη δεκαετία του 1970 η μέθοδος γίνεται γνωστή στον Καναδά από το Nishisato με την ονομασία Δυϊκή Κλιμάκωση (*Dual Scaling*) (Nishisato 1996, 1994, 1993, 1980 και 1978), ενώ την ίδια περίοδο οι De Leeuw, Young και Takane συστηματοποιούν υπολογιστικά και προγραμματιστικά τις μεθόδους Βέλτιστης Κλιμάκωσης (*Optimal Scaling*) (De Leeuw, Young & Takane 1976, Young, De Leeuw & Takane 1976, Takane, Young & De Leeuw 1977, Young, Takane & De Leeuw 1978), στο πλαίσιο των οποίων η ΠΑΑ μπορεί να ενταχθεί ως ειδική περίπτωση (Tenenhaus & Young, 1985). Αποτέλεσμα αυτής της προσπάθειας ήταν τελικά η δημιουργία και ανάπτυξη της Ολλανδικής Σχολής στην Ανάλυση Δεδομένων.

1.4.1 Η Γαλλική Σχολή της Ανάλυσης Δεδομένων

Στη Γαλλία οι μέθοδοι της Ανάλυσης Δεδομένων γνώρισαν σημαντική ανάπτυξη, ιδιαίτερα μετά το 1970, χάρη στο έργο και την “εμμονή” του Jean-Paul Benzécri Καθηγητή του Πανεπιστημίου Pierre et Marie Curie - Paris VI, ο οποίος

συστηματοποίησε τις μαθηματικές βάσεις των μεθόδων, τις ανέδειξε από τη λήθη και κατόρθωσε να αποκτήσουν στο χώρο της Στατιστικής θέση αντάξια της σημαντικότητας και της χρησιμότητάς τους (βλέπε Greenacre 1984, Van Meter *et al*. 1994, Clausen 1998, Καραπιστόλης 1999, Κουτσουπιάς 1999α και 1999β, Μεϊμάρης 2002, Παπαδημητρίου 2002 και 1994, Le Roux & Rouanet 2004). Ο Αθανασιάδης (1995, σ. 15) γράφει χαρακτηριστικά ότι "*Ο Benzécri είναι που θα προσδώσει στο όρο της ανάλυσης δεδομένων μια διάσταση ριζοσπαστική αν όχι διάσταση πολεμικής*." Θεμελίωσε με το έργο του μια Φιλοσοφία και μια Σχολή την οποία ασπάστηκαν και βελτίωσαν αρκετοί συνεργάτες και μαθητές του όπως οι Diday, Lebart, Escofier, Le Roux, J. Pagès, J.-P. Pagès, Morineau, Tenenhaus, Fenelon, Saporta, Greenacre, Παπαδημητρίου και άλλοι. Η δυναμική των μεθόδων ενισχύθηκε με την έκδοση των επιστημονικών περιοδικών *Cahiers d' Analyse des Données* και *Revue de Statistique Appliquée*, στα οποία δημοσιεύτηκε σημαντικός αριθμός εργασιών τόσο σε θεωρητικό όσο και σε επίπεδο εφαρμογών (Le Roux & Rouanet, 2004). Οι μέθοδοι άρχισαν να διδάσκονται σε πολλά μεταπτυχιακά τμήματα στατιστικής, ενώ σε επίπεδο εφαρμογών η ΠΑΑ, και ιδιαίτερα η πολυμεταβλητή εκδοχή της, έγινε το βασικό στατιστικό εργαλείο για την ανάλυση πολυδιάστατων κατηγορικών (ποιοτικών) δεδομένων που προέρχονταν κυρίως από τη συλλογή ερωτηματολογίων.

Η Γαλλική παράδοση έχει δώσει ιδιαίτερη έμφαση στη γεωμετρική θεώρηση της ερμηνείας των δεδομένων (Clausen 1998, Le Roux & Rouanet 2004) με βασικό σκοπό την ανάδειξη της ενδογενούς δομής που τα χαρακτηρίζει, η οποία, συνήθως, δεν είναι άμεσα αντιληπτή (Καραπιστόλης, 1999), αλλά βρίσκεται σε λανθάνουσα μορφή. Για τον Benzécri μεγαλύτερη αξία έχουν οι «παράγοντες», δηλαδή οι διαστάσεις και οι δομές που αναδεικνύονται μέσω των μεθόδων και όχι τα ίδια τα δεδομένα, τα οποία αποτελούν μόνο μια προσεγγιστική εικόνα της πραγματικότητας (Van Meter *et al*., 1994). Ως σχολή έχει να αναδείξει δύο κυρίως οικογένειες μεθόδων (Benzécri & *Collaborateurs*, 1973), την Παραγοντική Ανάλυση των Αντιστοιχιών (*Analyse Factoriel des Correspondances, A.F.C.*) και την Ταξινόμηση κατά Αύξουσα Ιεραρχία (*Classification Ascendente Hierarchique, C.A.H.*)[1]. Οι Αναστασιάδου και Παπαδημητρίου (2001α, σ. 327) υποστηρίζουν ότι οι μέθοδοι αυτές είναι οι πιο

[1] Στη Γαλία διαδεδομένες είναι επίσης η Συμβολική Ανάλυση (βλέπε Φλώρου, 1997) και η Συνεπαγωγική Στατιστική (βλέπε Gras 1995, Καραπιστόλης 1999 και 1996).

κατάλληλες για την ανάλυση ποιοτικών μεταβλητών και είναι δυνατό να αναδείξουν "*απρόβλεπτες διαστάσεις και να δημιουργήσουν νέες θεωρητικές προσεγγίσεις και προεκτάσεις*" κατά τη διερεύνηση ενός φαινομένου. Ειδικότερα, η ΠΑΑ, έφθασε να κατέχει στη Γαλλία μια τόσο ισχυρή θέση στη μεθοδολογία στατιστικής επεξεργασίας, ώστε να γίνει συνώνυμη με την έννοια της Ανάλυσης Δεδομένων (Clausen, 1998). Μέσω της ΠΑΑ είναι δυνατή η, σχεδόν, καθολική περιγραφή του υπό εξέταση φαινομένου (Αναστασιάδου & Παπαδημητρίου, 2001α), το οποίο παρουσιάζεται μέσα από ένα πίνακα κατηγορικών δεδομένων της μορφής «*αντικείμενα* × *μεταβλητές*» (βλέπε Παπαδημητρίου, 2004 και 1994). Βασικό πλεονέκτημα της μεθόδου είναι η δυνατότητα της ταυτόχρονης γραφικής απεικόνισης των μεταβλητών και των αντικειμένων, δηλαδή της οπτικής αναπαράστασης των αλληλεπιδράσεων και των σχέσεων τους σε ένα κοινό διάγραμμα (Hair *et al*. 1995, Bendixen 1995). Η μόνη απαίτηση της μεθόδου αφορά στις κλίμακες μέτρησης των μεταβλητών, οι οποίες θα πρέπει να είναι ονομαστικές (*nominal*) ή/και διάταξης (*ordinal*). Βέβαια, μπορούν να χρησιμοποιηθούν και ποσοτικές μεταβλητές, αφού πρώτα οι τιμές τους χωριστούν και ομαδοποιηθούν σε κλάσεις με βάση λογικά κριτήρια (Παπαδημητρίου, 1994). Με τον τρόπο αυτό αναδεικνύονται ιδιότητες (ποιοτικά χαρακτηριστικά) μέσα σε κάθε μεταβλητή και μέσω της ΠΑΑ οπτικοποιούνται και διαπιστώνονται οι μεταξύ τους ομοιότητες ή αντιθέσεις.

Οι μέθοδοι της Γαλλικής Σχολής, όπως αυτές αναδείχθηκαν από τον Benzécri, άρχισαν να γίνονται γνωστές στις αγγλόφωνες χώρες στα μέσα της δεκαετίας του '80 μετά τη δημοσίευση στην αγγλική γλώσσα σχετικών συγγραμμάτων, κυρίως από τους Greenacre (1984) και Lebart, Morineau & Warwick (1984). Μέχρι τότε η προσήλωση των Γάλλων ερευνητών στη συγγραφή των εργασιών τους στη δική τους γλώσσα σε συνδυασμό με τον ιδιόμορφο τρόπο της μαθηματικής παρουσίασης που υιοθέτησαν καθιστούσαν τις μεθόδους εσωτερική τους υπόθεση και "κλειστές" στο μη γαλλόφωνο αναγνωστικό κοινό (Van Rijckevorsel & De Leeuw, 1988).

Συμπερασματικά, η Γαλλική Σχολή της Ανάλυσης Δεδομένων με κύριο εκφραστή τον Benzécri αποτελεί όχι μόνο μια εναλλακτική μεθοδολογική προσέγγιση αλλά και μια νέα φιλοσοφική θεώρηση της Στατιστικής (βλέπε Van Meter *et al*. 1994, Le Roux & Rouanet 2004), μια νέα αντίληψη, ανεξάρτητη από μοντέλα, χωρίς *a priori* υποθέσεις

και χωρίς αυστηρές πιθανοθεωρητικές προϋποθέσεις, που σπάνια ικανοποιούνται στην πράξη (Gifi, 1996). Κεντρικό ρόλο στο μεθοδολογικό πλαίσιο της Γαλλικής προσέγγισης παίζει η ΠΑΑ, τόσο στη διμεταβλητή όσο και στην πολυμεταβλητή εκδοχή της, για την οποία ο Παπαδημητρίου (2004) τονίζει ότι:

> *"Η Παραγοντική Ανάλυση των Αντιστοιχιών είναι η σημαντικότερη μέθοδος που πρέπει να γνωρίζει κάποιος που επιθυμεί να αναλύσει ένα πολυμεταβλητό φαινόμενο χρησιμοποιώντας την αναλυτική διείσδυση της πολυδιάστατης στατιστικής ανάλυσης."* (σ. 3)

1.4.2 Η Ολλανδική Σχολή της Ανάλυσης Δεδομένων

Η εισαγωγή στη μεθοδολογία και τη λογική της ΠΑΑ μπορεί να γίνει με πολλούς και διαφορετικούς τρόπους (βλέπε Benzécri & *Collaborateurs* 1973, Nishisato 1980, Greenacre 1984, Tenenhaus & Young 1985, Israëls 1987, Bekker & De Leeuw 1988, Weller & Romney 1990, Andersen 1991, Benzécri 1992, Gifi 1996). Αυτός είναι, πιθανόν, ο λόγος για τον οποίο η μέθοδος "εφευρέθηκε" αρκετές φορές κατά τη διάρκεια του 20^{ου} αιώνα (Michailidis & De Leeuw, 1998). Αυτή η παράλληλη ανάπτυξη είχε ως αποτέλεσμα να δημιουργηθούν διάφορες προσεγγίσεις και κατευθύνσεις, τόσο σε σχέση με το αλγοριθμικό όσο και σε σχέση με το θεωρητικό υπόβαθρο της μεθόδου. Σε σχετικά πρόσφατες εργασίες μιας ομάδας ερευνητών του Πανεπιστημίου του Leiden (Department of Data Theory of the Faculty of Social Sciences) έγινε προσπάθεια ενοποίησης των διαφόρων παραλλαγών της μεθόδου αρχικά με την ονομασία *GIFI System* (De Leeuw 1984, Gifi 1996, Michailidis & De Leeuw 1998) και στη συνέχεια ως *Data Theory Scaling System-D.T.S.S.* (Meulman, 1999). Η ενοποίηση έγινε εφικτή με τη συστηματοποίηση και χρήση των μεθόδων της Βέλτιστης Κλιμάκωσης (Van de Geer 1993α και 1993β, Greenacre 1993α, Gifi 1996, De Leeuw 2005α). Βασικός σκοπός των μεθόδων αυτών είναι ο μετασχηματισμός ποιοτικών μεταβλητών σε ποσοτικές (Young 1981, Van Rijckevorsel & De Leeuw 1988, Greenacre 1993α). Αυτό επιτυγχάνεται με την ανάθεση βέλτιστων βαθμών (*scores*) και «βαρών» στις γραμμές (αντικείμενα) και στις στήλες (μεταβλητές) αντίστοιχα του πίνακα δεδομένων (Gifi, 1996). Οι νέες ποσοτικοποιημένες μεταβλητές μπορούν, στη συνέχεια, να χρησιμοποιηθούν σε στατιστικές διαδικασίες όπου απαιτούνται ποσοτικές μεταβλητές (Nishisato 1980, De Leeuw 2005γ, 1993 και 1988, Greenacre 1993α, Σιάρδος 1999, Meulman & Heiser 2004, Le Roux & Rouanet

2004). Η βελτιστοποίηση είναι έννοια σχετική γιατί επιτυγχάνεται με βάση τα διαθέσιμα δεδομένα που αναλύονται κάθε φορά. Δηλαδή, η κλίμακα μέτρησης μιας μεταβλητής αποκτά νόημα σε σχέση με τις υπόλοιπες. Τα κριτήρια βελτιστοποίησης είναι πολλά, όπως, για παράδειγμα, η βέλτιστη διάκριση μεταξύ των αντικειμένων (δειγματοληπτικές ή πειραματικές μονάδες), η μεγιστοποίηση της ομοιογένειας ή της εσωτερικής συνέπειας μεταξύ των μεταβλητών, η δημιουργία όσο το δυνατό ισχυρότερων γραμμικών σχέσεων μεταξύ ζευγών μεταβλητών και άλλα (Nishisato 1980, Gifi 1996, Michailidis & De Leeuw 1998, Meulman 1999, Michailidis & De Leeuw 2005). Έτσι, ο χρήστης των μεθόδων θα πρέπει, κάθε φορά, να επιλέγει το κατάλληλο κριτήριο βελτιστοποίησης καθώς και την κλίμακα ποσοτικοποίησης των μεταβλητών (ονομαστική, διάταξης ή διακριτή ποσοτική), ανάλογα με τους επιδιωκόμενους στόχους της ανάλυσης. Σε κάθε περίπτωση, μια κατάλληλη μη γραμμική «συνάρτηση απώλειας» (*loss function*) - αντικειμενική συνάρτηση - βελτιστοποιείται (Bekker & De Leeuw 1988, Gifi 1996, Michailidis & De Leeuw 2000 και 1998, SPSS Inc. 2004α και 1997) κάτω από προϋποθέσεις και δεσμεύσεις (συνθήκες) που καθορίζονται από το κριτήριο βελτιστοποίησης, τις κλίμακες μέτρησης των μεταβλητών και τους στόχους της εκάστοτε μελέτης. Υπολογιστικά, η βελτιστοποίηση επιτυγχάνεται κυρίως με την εφαρμογή του επαναληπτικού αλγόριθμου *Alternating Least Squares* (Εναλασσόμενα Ελάχιστα Τετράγωνα) (βλέπε Bekker & De Leeuw 1988, Gifi 1996, Michailidis & De Leeuw 1998, SPSS Inc. 2004α και 1997) σε αντίθεση με την αλγεβρική προσέγγιση του Benzécri. Όμως, στο πλαίσιο της Ολλανδικής προσέγγισης και η ΠΑΑ μπορεί να θεωρηθεί ως μέθοδος βέλτιστης κλιμάκωσης που έχει στόχο την καλύτερη δυνατή αναπαράσταση των δεδομένων ενός πίνακα συμπτώσεων σε ένα χώρο με λιγότερες διαστάσεις (Bendixen, 1996). Η βελτιστοποίηση έγκειται στο να βρεθούν, με κατάλληλη αλλαγή της κλίμακας μέτρησης, εκείνες οι τιμές των συντεταγμένων των προβολών των σημείων γραμμών και στηλών πάνω στους παραγοντικούς άξονες (Weller & Romney, 1990) που μεγιστοποιούν τη διακύμανση κατά τη διεύθυνση των αξόνων (Greenacre, 1993α).

Το σύστημα *GIFI* αποτελεί ένα σύνολο μεθόδων για την ανάλυση κυρίως κατηγορικών μεταβλητών και χαρακτηρίζει την Ολλανδική Σχολή της Ανάλυσης Δεδομένων (Bond & Michailidis, 1996). Σημαντικός σταθμός στην εξέλιξη των μεθόδων της Σχολής αυτής ήταν η ανάπτυξη της Ανάλυσης Ομοιογένειας

(*Homogeneity Analysis*) ή αλλιώς της Πολλαπλής Ανάλυσης των Αντιστοιχιών (*Multiple Correspondence Analysis*) (Michailidis & De Leeuw, 1998), η οποία αποτελεί, ως ένα βαθμό, την πολυμεταβλητή εκδοχή της ΠΑΑ της Γαλλικής Σχολής. Με βάση το θεωρητικό και αλγοριθμικό πλαίσιο της μεθόδου αυτής αναπτύχθηκε η Μη Γραμμική Ανάλυση σε Κύριες Συνιστώσες (*Non Linear Principal Component Analysis* ή *Principal Components Analysis for Categorical Data*) και η Μη Γραμμική Κανονικοποιημένη Συσχέτιση (*Non Linear Canonical Correlation*), οι οποίες αποτελούν επίσης σημαντικές μεθόδους του συστήματος *GIFI*. Η ερευνητική δραστηριότητα των Ολλανδών επεκτάθηκε και στην ανάπτυξη μεθόδων με τις οποίες είναι δυνατή η ταυτόχρονη χρήση δεδομένων μικτού τύπου (κατηγορικά και ποσοτικά) καθώς και στην δυνατότητα ανάλυσης πολλών ομάδων μεταβλητών (Van der Burg & De Leeuw 1988, Van der Burg, De Leeuw & Dijksterhuis 1994, Micailidis & De Leeuw 1998). Οι αντίστοιχες στατιστικές διαδικασίες περιλαμβάνονται στο υποσύστημα *Categories* του στατιστικού πακέτου SPSS (SPSS Inc. 1998α, Σιάρδος 1999 και 2000, Meulman & Heiser 2004).

Το όνομα *GIFI* αντιπροσωπεύει και την ίδια την ομάδα των Ολλανδών ερευνητών που για πρακτικούς κυρίως λόγους το χρησιμοποίησαν για κοινές δημοσιεύσεις. Μέλη της ομάδας αποτέλεσαν οι Bettonvil, Van der Burg, Van de Geer, Heiser, Meulman, Van Rijckevorsell, Stoop και άλλοι (Gifi, 1996) με πρωταγωνιστικό και καθοδηγητικό ρόλο αυτόν του Jan de Leeuw (Greenacre, 1993α). Σημαντικό μέρος του ερευνητικού έργου των Ολλανδών ερευνητών έχει δημοσιευθεί σε τεχνικές αναφορές εσωτερικής έκδοσης του Πανεπιστημίου του Leiden και στο επιστημονικό περιοδικό *Psychometrika* (βλέπε Van Rijckevorsel & De Leeuw, 1988).

Το ενδιαφέρον της Ολλανδικής Σχολής εστιάζεται όχι μόνο στη γεωμετρική ερμηνεία των δεδομένων αλλά και στην ανάδειξη των μη γραμμικών σχέσεων μεταξύ των μεταβλητών που προκύπτουν λόγω της κατηγορικής φύσης των δεδομένων (Van der Burg & De Leeuw 1983, Bekker & De Leeuw 1988, Van de Geer 1993α και 1993β, Heiser & Meulman 1994, Gifi 1996, Michailidis & De Leeuw 1998). Η προσέγγιση αυτή έρχεται σε αντίθεση με παραδοσιακές μεθόδους, όπως για παράδειγμα η Πολλαπλή Γραμμική Παλινδρόμηση, η Ανάλυση Διακύμανσης και η Ανάλυση σε Κύριες Συνιστώσες, οι οποίες προϋποθέτουν γραμμικές σχέσεις και μοντέλα -

υποδείγματα (Gifi, 1996). Είναι σημαντικό να τονιστεί ότι πολλές από τις γραμμικές μεθόδους προκύπτουν ως ειδικές περιπτώσεις των μη γραμμικών.

Συνοψίζοντας, οι μέθοδοι της Ολλανδικής Σχολής αποτελούν μια επέκταση των μεθόδων της Γαλλικής Σχολής Ανάλυσης Δεδομένων. Οι δύο σχολές έχουν αρκετές ομοιότητες αλλά και διαφορές. Έτσι, ενώ η ΠΑΑ στην πολυμεταβλητή της εκδοχή αποτελεί μία από τις σημαντικότερες μεθόδους και για τις δύο σχολές, ωστόσο το θεωρητικό και μεθοδολογικό πλαίσιο στο οποίο αναπτύχθηκε η μέθοδος σε κάθε σχολή είναι διαφορετικό. Και στις δύο σχολές, ισχυρά κίνητρα εξέλιξης και διάδοσης των προτεινόμενων μεθόδων αποτελούν η γεωμετρική ερμηνεία των δεδομένων και ο γραφικός τρόπος παρουσίασης και διάχυσης της παραγόμενης πληροφορίας. Υπάρχει όμως σημαντική διαφορά σε ό,τι αφορά την προβληματική, τη θεωρητική και την υπολογιστική προσέγγιση που ακολουθεί η κάθε Σχολή. Οι μέθοδοι της Ολλανδικής Σχολής αντιμετωπίζονται τελικά ως προβλήματα μη γραμμικής βελτιστοποίησης υπό συνθήκες. Η επίλυση τους επιτυγχάνεται με τη χρήση επαναληπτικών υπολογιστικών αλγορίθμων και όχι αλγεβρικά. Τέλος, η ιδιαίτερη προσήλωση των Ολλανδών ερευνητών στην ανάδειξη μη γραμμικών σχέσεων στα δεδομένα είχε ως αποτέλεσμα οι μέθοδοι που αναπτύχθηκαν να χαρακτηρίσουν ένα νέο πεδίο έρευνας της Στατιστικής, αυτό που οι ίδιοι ονόμασαν *Μη Γραμμική Πολυμεταβλητή Ανάλυση* (Bekker & De Leeuw 1988, Gifi 1996).

1.4.3 Η Ιταλική Σχολή της Ανάλυσης Δεδομένων

Στην Ιταλία, στις αρχές της δεκαετίας του ’90 και σε ένα γεωμετρικό πλαίσιο οι Carlo Lauro και Luigi D’Ambra παρουσίασαν νέες οικογένειες μεθόδων της Ανάλυσης Δεδομένων με τις οποίες η αντιμετώπιση των μεταβλητών δεν είναι συμμετρική, όπως συμβαίνει στις βασικές μεθόδους της Γαλλικής και ως ένα μεγάλο βαθμό της Ολλανδικής σχολής, αλλά μη συμμετρική (βλέπε Lauro & Siciliano 1989, D’ Ambra & Lauro 1992, Balbi 1998, Siciliano & Mola 1998, Kroonenberg & Lombardo 1999, Lauro & Balbi 1999, Gallo & Simonetti 2002). Πιο συγκεκριμένα, το ενδιαφέρον των Ιταλών ερευνητών στράφηκε στο διαχωρισμό των μεταβλητών σε εξαρτημένες και ανεξάρτητες και στην αξιοποίηση ενδεχόμενης διαθέσιμης εξωτερικής πληροφορίας σε σχέση με το υπό εξέταση φαινόμενο. Χαρακτηριστικά παραδείγματα των μεθόδων της Ιταλικής Σχολής αποτελούν η Μη Συμμετρική Ανάλυση των Αντιστοιχιών (*Non*

Symmetrical Correspondence Analysis-N.S.C.A.) για κατηγορικές μεταβλητές και η Ανάλυση σε Κύριες Συνιστώσες σε Υποχώρο Αναφοράς (*Principal Component Analysis onto a Reference Subspace–P.C.A.R.*) για ποσοτικές. Έκτοτε, το ενδιαφέρον τους επεκτάθηκε και στην ανάπτυξη μεθόδων με τις οποίες είναι δυνατή η ανάλυση δεδομένων μικτού τύπου (κατηγορικά και ποσοτικά) καθώς και η ανάλυση πολλών συνόλων – ομάδων μεταβλητών. Συμπερασματικά, η Ιταλική Σχολή χαρακτηρίζει έναν νέο κλάδο της Στατιστικής, τη *Μη Συμμετρική Ανάλυση Δεδομένων*. Για περισσότερες πληροφορίες σχετικά με τις μεθόδους της σχολής αυτής παραπέμπουμε στην ιστοσελίδα του Τμήματος Μαθηματικών και Στατιστικής του Πανεπιστημίου της Νάπολης «Federico II» (http://www.dms.unina.it/NSDA.html).

1.4.4 Η Ανάλυση Δεδομένων στην Ελλάδα

Στην Ελλάδα, οι μέθοδοι της Ανάλυσης Δεδομένων, όπως αυτές εφαρμόζονται στο μεθοδολογικό και φιλοσοφικό πλαίσιο κυρίως της Γαλλικής Σχολής, εμφανίστηκαν γύρω στο 1980 από έλληνες ερευνητές που είχαν αποκτήσει μεταπτυχιακό δίπλωμα ή/και διδακτορικό τίτλο κάτω από την επίβλεψη του Καθηγητή Jean Paul Benzécri ή στενών συνεργατών του. Σήμερα, πολλοί από αυτούς τους "πρωτοπόρους" είναι μέλη Δ.Ε.Π. σε διάφορα τμήματα Ελληνικών Πανεπιστημίων, ενώ ήδη από το 1985 έχει καθιερωθεί η διδασκαλία αντίστοιχων μαθημάτων, σε Προπτυχιακά και Μεταπτυχιακά Προγράμματα, διαφόρων Πανεπιστημιακών Τμημάτων. Το 2001 ιδρύθηκε το επιστημονικό σωματείο με την επωνυμία «Ελληνική Εταιρία Ανάλυσης Δεδομένων», ενώ το 2002 ξεκίνησε η έκδοση του επιστημονικού περιοδικού «Τετράδια Ανάλυσης Δεδομένων - *Data Analysis Bulletin*», στο οποίο δημοσιεύονται μετά από κρίση θεωρητικές εργασίες αλλά και εφαρμογές των μεθόδων από έλληνες και ξένους ερευνητές. Για περισσότερα στοιχεία σχετικά με την πορεία των μεθόδων στον ελληνικό χώρο παραπέμπουμε στην ιστοσελίδα της Ελληνικής Εταιρίας Ανάλυσης Δεδομένων (http://datan.uom.gr).

Ένας παράγοντας, ο οποίος συνετέλεσε στην αποδοχή των μεθόδων από την Ελληνική Πανεπιστημιακή Κοινότητα, ήταν η δυνατότητα υλοποίησης και εφαρμογής τους από εμπορικά στατιστικά λογισμικά (Παπαδημητρίου, 2002), όπως το SPSS, το SAS, το BMDP, το Statistica, το Minitab και το Systat. Όμως, η διάδοση και η εξέλιξη των μεθόδων στην Ελλάδα, ιδιαίτερα της ΠΑΑ και της Ιεραρχικής

Ταξινόμησης για κατηγορικά δεδομένα, οφείλεται κατά κύριο λόγο στην προσπάθεια, στην υπομονή και στην επιμονή του Καθηγητή Γιάννη Παπαδημητρίου (Μεϊμάρης, 2005 και 2002) του Τμήματος Εφαρμοσμένης Πληροφορικής του Πανεπιστημίου Μακεδονίας, ο οποίος διατελεί Πρόεδρος της Ελληνικής Εταιρίας Ανάλυσης Δεδομένων από το 2001 και είναι ο Διευθυντής Σύνταξης των «Τετραδίων». Είτε υπό την επίβλεψη είτε υπό την καθοδήγησή του εκπονήθηκαν διδακτορικές διατριβές, θεωρητικές και σε ποικίλα ερευνητικά πεδία εφαρμογής των μεθόδων, και αναπτύχθηκαν ειδικά λογισμικά ανάλυσης δεδομένων (βλέπε Καραπιστόλης 1996, Φλώρου 1997, Κουτσουπιάς 1999α και 1999β, Αναστασιάδου 2000, Μοσχίδης 2003, Λούκας 2004, Μπαγιάτης 2004, Δρόσος 2005, Μασούρα 2005, Μάλλιαρη 2005, Τζήμος 2006). Όλα αυτά, μεταξύ των συνεχών προσπαθειών και ενεργειών του για την προαγωγή της θεωρίας και της εφαρμογής των μεθόδων στην Ελλάδα, ώστε να αναδειχθεί η λειτουργικότητά τους και η διαλεκτική τους με άλλες μεθόδους της Στατιστικής Επιστήμης.

1.5 Η Ανάλυση Δεδομένων και η Στατιστική

1.5.1 Οι Απόψεις του Tukey

Ο John Tukey, διακεκριμένος ερευνητής στο χώρο της Μαθηματικής Στατιστικής (Cohen & Cohen, 1983), ήταν ο πρώτος που διαχώρισε την Ανάλυση Δεδομένων από τη Στατιστική και την παρουσίασε ως ανεξάρτητη επιστήμη (Gifi, 1996). Κατά την άποψή του (Tukey, 1962) η Ανάλυση Δεδομένων αποτελεί ένα γενικότερο μεθοδολογικό πλαίσιο απ’ ότι η Επαγωγική Στατιστική, ενώ υπάρχουν πεδία της Μαθηματικής Στατιστικής που είναι έξω από το πλαίσιο της Ανάλυσης Δεδομένων. Σύμφωνα με τον Tukey, η Ανάλυση Δεδομένων περιλαμβάνει: α) μεθόδους επεξεργασίας και ανάλυσης των δεδομένων, β) τεχνικές ερμηνείας των αποτελεσμάτων, γ) μεθοδολογίες σχεδιασμού των διαδικασιών συλλογής των δεδομένων και δ) όλα τα μαθηματικά και στατιστικά “εργαλεία – βοηθήματα” που μπορούν να χρησιμοποιηθούν με βασικό σκοπό την ευκολότερη ανάλυση και ερμηνεία των δεδομένων. Για τον Tukey, στην Ανάλυση Δεδομένων θα πρέπει να δοθεί ιδιαίτερα μεγάλη έμφαση στην κρίση. Τουλάχιστον τρία διαφορετικά είδη ή πηγές κρίσης είναι δυνατό να εμπλέκονται σχεδόν σε κάθε περίπτωση (Tukey, 1962)

> "*(α1) κρίση που βασίζεται σε εμπειρία από το συγκεκριμένο θεματικό πεδίο από το οποίο προέρχονται τα δεδομένα*
>
> *(α2) κρίση που βασίζεται πάνω σε μία ευρεία εμπειρία για τον τρόπο που έχουν λειτουργήσει διάφορες τεχνικές ανάλυσης δεδομένων σε σχέση με ποικίλα πεδία εφαρμογής*
>
> *(α3) κρίση που βασίζεται πάνω σε αφηρημένα αποτελέσματα για τις ιδιότητες συγκεκριμένων τεχνικών, που έχουν ληφθεί είτε μέσω μαθηματικών αποδείξεων είτε μέσω εμπειρικής δειγματοληψίας.*" (σ. 9).

Ο Tukey (1962) συνεχίζει και γράφει ότι:

> "*Η πιο σημαντική ρήση που πρέπει να λάβει υπόψη η ανάλυση δεδομένων, την οποία πολλοί στατιστικοί μοιάζουν να έχουν παραμελήσει, είναι: «Είναι πολύ προτιμότερη μία προσεγγιστική απάντηση σε ένα σωστό ερώτημα, το οποίο είναι συχνά ασαφές, παρά μία ακριβής απάντηση σε ένα λανθασμένο ερώτημα, το οποίο μπορεί εύκολα να γίνει ακριβές». Η ανάλυση δεδομένων θα πρέπει να εξελίσσεται με προσεγγιστικές απαντήσεις, στην καλύτερη περίπτωση, διότι η γνώση της σχετικά με το πραγματικό περιεχόμενο του προβλήματος θα είναι στην καλύτερη περίπτωση προσεγγιστική.*" (σσ. 13-14).

Η συνεισφορά του Tukey στην Ανάλυση Δεδομένων αφορά σε ένα πλήθος ευέλικτων γραφικών κυρίως μεθόδων και τεχνικών για τη διερευνητική και περιγραφική στατιστική ανάλυση των δεδομένων (βλέπε Tukey 1977, Hoaglin 2003). Ο Tukey θεωρεί ότι η προσέγγιση αυτή συνιστά κατά βάση μια Φιλοσοφία, η οποία επαναπροσανατολίζει τη Στατιστική στους αρχικούς της στόχους, όπου η στατιστική περιγραφή των δεδομένων παίζει το σημαντικότερο ρόλο. Δεν αντιτίθεται στον επιβεβαιωτικό χαρακτήρα της Επαγωγική Στατιστικής και διατυπώνει τελικά την άποψη ότι και οι δύο προσεγγίσεις (Επαγωγική Στατιστική και Ανάλυση Δεδομένων), μπορούν και σε ορισμένες περιπτώσεις επιβάλλεται, να χρησιμοποιούνται συμπληρωματικά (Tukey, 1980 και 1977) λαμβάνοντας υπόψη όλα τα γνωστά στοιχεία (θεωρητικά και μεθοδολογικά) που αφορούν στο συγκεκριμένο ερευνητικό πεδίο. Σε κάθε περίπτωση, η διερευνητική – περιγραφική προσέγγιση είναι αυτή που θα πρέπει να προηγείται γιατί, σύμφωνα με την άποψή του (1980), μερικές φορές είναι πιο δύσκολο αλλά και πιο ενδιαφέρον να διατυπώνεις ερωτήσεις από το να παίρνεις απαντήσεις.

1.5.2 Οι Αρχές του Benzécri

Οι αρχές της Γαλλικής Σχολής Ανάλυσης Δεδομένων, όπως εξηγούνται από τον Benzécri, διαφέρουν αρκετά από αυτές του Tukey. Ο Benzécri διατυπώνει πέντε βασικές αρχές που ορίζουν την Ανάλυση Δεδομένων (Benzécri & *Collaborateurs*, 1973):

> "*Αρχή 1: Η Στατιστική και η Θεωρία Πιθανοτήτων δεν είναι το ίδιο πράγμα. Αρκετοί συγγραφείς (οι οποίοι, και σας το λέω στα Γαλλικά, σπάνια γράφουν στη γλώσσα μας...) έχουν δομήσει μία πομπώδη επιστήμη υπό την ονομασία «Μαθηματική Στατιστική», η οποία βρίθει υποθέσεων που ποτέ δεν ικανοποιούνται στην πράξη. Δεν μπορούμε να προσδοκούμε μία λύση στα τυπολογικά μας προβλήματα από αυτούς τους συγγραφείς.* (σ. 3).
>
> *Αρχή 2: Το μοντέλο θα πρέπει να ακολουθεί τα δεδομένα, και όχι το αντίστροφο. Αυτό είναι ένα ακόμα λάθος στην εφαρμογή των μαθηματικών στις Ανθρωπιστικές Επιστήμες: η αφθονία δηλαδή των μοντέλων, που δομούνται a priori και μετά έρχονται αντιμέτωπα με τα δεδομένα μέσω του επονομαζόμενου «ελέγχου». Συχνά ο «έλεγχος» χρησιμοποιείται για να δικαιολογήσει κάποιο μοντέλο, στο οποίο ο αριθμός των παραμέτρων που πρέπει να ενταχθούν είναι μεγαλύτερος από τον αριθμό των δεδομένων σημείων. Και συχνά χρησιμοποιείται, αντιθέτως, για να απορρίψει κατηγορηματικά ως αβάσιμα, ακόμα και τα πιο επικριτικά σχόλια του πειραματιστή. Αυτό που χρειαζόμαστε όμως είναι μία αυστηρή μέθοδος για να εξάγουμε τη δομή, ξεκινώντας από τα δεδομένα.* (σ. 6).
>
> *Αρχή 3: Είναι βολικό να χειρίζεται κανείς ταυτόχρονα πληροφορίες σε όσες περισσότερες διαστάσεις μπορεί. Ως συνέπεια, το πρόβλημα της εγκυρότητας ενός «ελέγχου» - το οποίο, ομολογουμένως, είναι ορισμένες φορές δύσκολο – δεν παρουσιάζεται πλέον ως τόσο σημαντικό. Κανείς δεν γνωρίζει αν η ανισότητα 0,5 ≠ 0,7 θα πρέπει να ερμηνευτεί σε αυτές τις πρακτικές περιπτώσεις ως ένα συγκεκριμένο εμπειρικό αποτέλεσμα, ή απλώς ως τυχαίο αποτέλεσμα. Όμως το να βρει κανείς ότι σε ένα χώρο δύο διαστάσεων υπάρχουν πενήντα σημεία κατά προσέγγιση τοποθετημένα σε έναν κύκλο είναι πραγματικά μία ανακάλυψη (τουλάχιστον αν η μέθοδος υπολογισμού δεν μας παραπλανά!).* (σ. 9).
>
> *Αρχή 4: Για την ανάλυση περίπλοκων στοιχείων, και ειδικά για την ανάλυση κοινωνικών δεδομένων, χρειαζόμαστε απαραιτήτως τον ηλεκτρονικό υπολογιστή. Η αρχή αυτή είναι προφανώς ορθή ... αλλά τι άποψη θα είχαν γι' αυτό οι φοβεροί και τρομεροί πατέρες μας δεκαπέντε χρόνια πριν;* (σ. 12).

> *Αρχή 5: Η χρήση του υπολογιστή σημαίνει ότι όλες οι τεχνικές που σχεδιάστηκαν πριν την έλευση του αυτόματου υπολογισμού θα πρέπει να εγκαταλειφθούν. Και λέω τεχνικές και όχι επιστήμη: οι γεωμετρικές και αλγεβρικές αρχές των προγραμμάτων μας ήταν γνωστές στον Laplace, πριν 150 χρόνια. Όμως ο Laplace ήταν και ο συγγραφέας μίας πραγματείας πάνω στη ουράνια μηχανική, η οποία μόλις τώρα επανεκδόθηκε για να χρησιμοποιηθεί από μηχανικούς του διαστήματος ... Και να φανταστεί κανείς ότι η πραγματεία αυτή δεν ήταν αρκετή για να κατακτήσει ο Ναπολέων το φεγγάρι.*" (σ. 15).

Οι Ολλανδοί ερευνητές δεν συμφωνούν απόλυτα με τον αυστηρό και απόλυτο τρόπο που διατυπώνει ο Benzécri τις πέντε Αρχές για το τι είναι η Ανάλυση Δεδομένων (Gifi, 1996). Για παράδειγμα, δεν συμφωνούν με αυτό που ο Benzécri φαίνεται να υπονοεί στις Αρχές 3, 4 και 5 ότι δηλαδή οι μεγάλες, αδόμητες πολυμεταβλητές ομάδες δεδομένων είναι η μόνη δυνατή κατάσταση έρευνας. Εν γένει συμφωνούν με τις Αρχές 2 και 3, αλλά επισημαίνουν ότι αν οι αρχές αυτές εφαρμοστούν συστηματικά είναι δυνατό να οδηγήσουν σε ένα "τυφλό" εμπειρισμό, ο οποίος τρομοκρατεί εν γένει τους Ψυχομέτρες. Πιστεύουν ότι τουλάχιστον κάποια θεωρία – υπόδειγμα θα πρέπει να καθοδηγεί τον χρήστη των μεθόδων, για παράδειγμα στην επιλογή των σχετικών με το υπό εξέταση φαινόμενο μεταβλητών, ώστε να είναι εφικτή η ερμηνεία των αποτελεσμάτων. Για τους Ολλανδούς η Αρχή 1 του Benzécri είναι ίσως η πιο σημαντική. Τα πιο ενδιαφέροντα σημεία διατυπώνονται από τον ίδιο τον Benzécri (Benzécri & *Collaborateurs,* 1973):

> *"Τα μαθηματικά θεμέλια της στατιστικής ανάλυσης είναι περισσότερο αλγεβρικά και γεωμετρικά (και όταν εμπλέκονται πολλές διαστάσεις, οι γεωμετρικές ιδέες συγχωνεύονται με τους αλγεβρικούς υπολογισμούς) και μετά πιθανολογικά. Είναι προτιμότερο να μιλάει κανείς για κύριους άξονες κ.τ.λ. ... που καθορίζονται σε σχέση με έναν πεπερασμένο αριθμό πραγματικών δεδομένων, από το να μιλάει για τη μέση τιμή, κ.τ.λ. ... που ορίζεται σε ένα δυνητικά άπειρο Δειγματοχώρο. Όμως οι πιθανολογικές έννοιες και ιδέες μπορούν να αναδείξουν την αναγκαιότητα αλγεβρικών πράξεων και μερικές φορές μπορούν να χρησιμοποιηθούν για να αξιολογήσουν την χρησιμότητά τους."* (σ. 6).

Σύμφωνα με τους Johnson και Wichern (1992), η επιστημονική έρευνα μπορεί να θεωρηθεί ως μια επαναληπτική διαδικασία "μάθησης". Αρχικά, καθορίζονται τα αντικειμενικά στοιχεία (μετρήσεις) σχετικά με την ερμηνεία ενός κοινωνικού ή φυσικού φαινομένου, τα οποία στη συνέχεια συγκεντρώνονται και αναλύονται. Η

ανάλυση των αντίστοιχων δεδομένων μέσα από μία μόνο πειραματική ή παρατηρησιακή διαδικασία έχει ως αποτέλεσμα την περιορισμένη ερμηνεία του υπό εξέταση φαινομένου. Έτσι, κατά τη διάρκεια της επιστημονικής αναζήτησης, μέσω της επαναληπτικής διαδικασίας μάθησης, νέες μεταβλητές προστίθενται κάθε φορά ή παλιές αφαιρούνται. Για τον Benzécri η Ανάλυση Δεδομένων αποτελεί μια διαδικασία μάθησης και απόκτησης γνώσης μέσα από ποιοτικές και ποσοτικές καταγραφές πραγματικών εμπειριών, κατά τις οποίες η φροντίδα και η ευθύνη της ερμηνείας των αποτελεσμάτων και των συνεπειών τους αφήνεται κυρίως στους ερευνητές – χρήστες των μεθόδων και όχι σε πιθανολογικούς μηχανισμούς.

1.5.3 Οι Θέσεις των Ολλανδών

Στο πλαίσιο της Ολλανδικής Σχολής η Ανάλυση Δεδομένων είναι συνώνυμη με τη Στατιστική (De Leeuw, 2005γ), ενώ η Μαθηματική Στατιστική είναι ενδιαφέρουσα στο βαθμό που τα πορίσματά της έχουν πρακτικές εφαρμογές και συνέπειες (Gifi, 1996) ιδιαίτερα στη μελέτη και στον έλεγχο της «σταθερότητας» των αποτελεσμάτων που παράγονται από τις μεθόδους Ανάλυσης Δεδομένων (Markus, 1994α και 1994β). Ο όρος «σταθερότητα» αναφέρεται στο πόσο συνεπή είναι τα αποτελέσματα κάτω από διαταραχές κάποιων αρχικών συνθηκών, οι οποίες είτε αφορούν στη μεθοδολογία συλλογής των δεδομένων είτε στις ίδιες τις μεθόδους ανάλυσης, όπως για παράδειγμα στους εκάστοτε αλγόριθμους καθώς και στους αριθμητικούς υπολογισμούς. Πιο συγκεκριμένα, φαίνεται να ανοίγεται ένας μεγάλος ορίζοντας σε ό,τι αφορά τη μελέτη και τον έλεγχο της σταθερότητας (Gifi 1996, Michailidis 1996, Michailidis & De Leeuw 1998): α) των αποτελεσμάτων στη θεωρητική περίπτωση επανάληψης της έρευνας κάτω από τις ίδιες συνθήκες (*replication stability*), β) των διαφόρων δεικτών (αναφοράς, ερμηνείας και ποιότητας) με βάση τους οποίους γίνεται ή/και διευκολύνεται η παρουσίαση και η ερμηνεία των αποτελεσμάτων (*statistical stability*), γ) των αποτελεσμάτων κάτω από διαφορετικές μεθόδους συλλογής των δεδομένων ή σύνθεσης (απαλοιφή αντικειμένων ή/και μεταβλητών) του πίνακα δεδομένων που θα αναλυθεί (*stability under data selection*), δ) των αποτελεσμάτων κάτω από διαφορετικά μοντέλα - υποδείγματα διερεύνησης συσχετίσεων ή σχέσεων αιτίας-αποτελέσματος (*stability under model selection*), ε) της ακρίβειας των αριθμητικών υπολογισμών είτε λόγω σφαλμάτων στρογγυλοποίησης είτε λόγω των επαναληπτικών υπολογιστικών μεθόδων που συνήθως χρησιμοποιούνται (*numerical*

stability), στ) των αποτελεσμάτων κάτω από διαταραχές των αρχικών δεδομένων και κάτω από την ισχύ Κεντρικών Οριακών Θεωρημάτων της Θεωρίας Πιθανοτήτων (*analytical and algebraic stability*) και ζ) των αποτελεσμάτων κάτω από την επίδραση που ενδεχομένως να έχει σε αυτά η επιλογή μιας διαφορετικής μεθόδου ανάλυσης (*stability under selection of technique*). Ένα άλλο χαρακτηριστικό της Ολλανδικής Σχολής είναι η μετάθεση του ενδιαφέροντος από τις ιδιότητες και τις εφαρμογές της Πολυδιάστατης Κανονικής και της Πολυωνυμικής Κατανομής (Gifi, 1996) σε μια ενδιάμεση κατάσταση με τη χρήση μεθόδων επαναδειγματοληψίας (βλέπε Efron & Tibshirani 1993, DiCiccio & Efron 1996). Τεχνικές όπως η *Monte Carlo* και η *Bootstrap* κατέχουν σημαντική θέση κυρίως στους ελέγχους εξωτερικής και εσωτερικής εγκυρότητας των παραγόμενων αποτελεσμάτων (βλέπε Lebart *et al*. 1984, Greenacre 1993α και 1984, Markus 1994α και 1994β, Gifi 1996, Michailidis 1996, Chateau & Lebart 1996, Μπεχράκης 1999, Lebart 2006).

Έτσι, για τους Ολλανδούς η Θεωρία Πιθανοτήτων, η Στατιστική, η Υπολογιστική Στατιστική, τα Μαθηματικά, τα Υπολογιστικά Μαθηματικά και η Πληροφορική όχι μόνο ενσωματώνονται στην Ανάλυση Δεδομένων αλλά μπορούν να συνδυαστούν με αυτήν, με τρόπο ώστε να δημιουργηθούν νέοι τομείς έρευνας κυρίως σε σχέση με τη σταθερότητα των αποτελεσμάτων, η οποία εγγυάται, τελικά, την αξιοπιστία και την εγκυρότητα των συμπερασμάτων.

1.5.4 Οι Απόψεις του Παπαδημητρίου

Μια πιο λειτουργική προσέγγιση στον ορισμό της Ανάλυσης Δεδομένων και της σχέσης της με τη Στατιστική διατυπώνει ο Παπαδημητρίου (1994):

> "*Η Ανάλυση Δεδομένων αποτελεί σύνθεση μεθόδων της Περιγραφικής Στατιστικής ή καλύτερα συνάθροιση και παράθεση μιας ακολουθίας μεθόδων που ο λογικός τους δεσμός, τις περισσότερες φορές, είναι ιδιαίτερα λεπτός.*" (σ. 12).

Αυτή η σύνθεση εφαρμόζεται στους πίνακες των αρχικών δεδομένων που περιγράφουν το υπό εξέταση φαινόμενο, με σκοπό να εμφανιστούν "*στις εκροές της αποτελέσματα εύκολα ερμηνεύσιμα από τους μυημένους στις αρχές της*" (Παπαδημητρίου, 2002, σ. 6). Για τον Παπαδημητρίου (2002 και 1994) η Ανάλυση

Δεδομένων στηρίζεται στις αναλλοίωτες αρχές και μεθόδους της Γραμμικής Άλγεβρας, οι οποίες σήμερα, μέσω των Η/Υ, μπορούν πλέον να εφαρμοστούν και σε μεγάλα σύνολα δεδομένων. Παρόλο που η δυνατότητα χειρισμού των αλγεβρικών μεθόδων για πρακτική χρήση απαιτεί την κατανόησή τους σε βάθος ωστόσο αποφεύγεται ο "*εμπειρισμός*" και η "*καλή πρόθεση*" σε ό,τι αφορά την εξαγωγή συμπερασμάτων (Παπαδημητρίου, 2002, σ. 7). Με την εφαρμογή των μεθόδων της Ανάλυσης Δεδομένων είναι δυνατό να επιτευχθεί η "διαστολή" του φαινομένου που μελετάται και το αποτέλεσμα να παρουσιαστεί με τη μορφή της καλύτερης δυνατής "εικόνας-φωτογραφίας" του διασταλμένου φαινομένου και μάλιστα χωρίς καμία *a priori* υπόθεση ή προϋπόθεση.

Στην προσέγγιση αυτή είναι φανερή η αντίληψη ότι τα δεδομένα θα πρέπει να "μιλούν μόνα τους" με τρόπο ώστε ο τελικός αποδέκτης-χρήστης να εμβαθύνει στην ερμηνεία τους. Αυτό επιτυγχάνεται με τη χρήση των μεθόδων της Ανάλυσης Δεδομένων οι οποίες, μάλιστα, αναδεικνύουν – μεγεθύνουν τυχόν κρυμμένες διαστάσεις και λανθάνουσες δομές σχέσεων και τάσεων, ομοιοτήτων και διαφορών, οι οποίες είναι ενθυλακωμένες στα αρχικά δεδομένα αλλά δεν είναι εύκολο με τη χρήση απλών στατιστικών μεθόδων να ανιχνευθούν, να οπτικοποιηθούν και να ερμηνευτούν με σαφήνεια και σχετική πληρότητα.

1.5.5 Μερικά Σχόλια

Είναι γεγονός ότι συχνά δεν γνωρίζουμε τους νόμους στους οποίους ένα φαινόμενο υπακούει ή το μαθηματικό μοντέλο – υπόδειγμα (ντετερμινιστικό ή στοχαστικό), το οποίο καθορίζει τη συμπεριφορά του, είτε γιατί είναι αρκετά σύνθετο και δεν επιδέχεται λεπτομερή μαθηματική περιγραφή είτε δεν μπορεί να διατυπωθεί κάποιος νόμος που να εκφράζει μια αποδεκτή σχέση αιτίας-αποτελέσματος (Δερμάνης, 1986). Στην περίπτωση αυτή, αναδεικνύεται η χρησιμότητα της Ανάλυσης Δεδομένων, η οποία ως κλάδος της Πολυδιάστατης Στατιστικής Ανάλυσης είναι δυνατόν να περιγράψει την πολυπλοκότητα ορισμένων φαινομένων ή καταστάσεων και να αναδείξει τυχόν "άγνωστες – μη εμφανείς" πτυχές ή ιδιαιτερότητες της συμπεριφοράς και της εξέλιξής τους.

Ο χαρακτηρισμός των μεθόδων της Ανάλυσης Δεδομένων ως στατιστικές διαδικασίες ανεξάρτητες από μοντέλα, οι οποίες χρησιμοποιούνται χωρίς *a priori* υποθέσεις και προϋποθέσεις, είναι σωστός μόνο σε μια πρώτη θεώρηση και πάντα σε σύγκριση με τις διαδικασίες της Επαγωγικής Στατιστικής, οι οποίες έχουν αρκετές πιθανοθεωρητικές και τεχνικές προϋποθέσεις και εφαρμόζονται για την απόρριψη ή όχι συγκεκριμένων ερευνητικών υποθέσεων. Οι προϋποθέσεις αυτές δεν συνδέονται με τις γενικότερες επιστημονικές υποθέσεις οι οποίες προηγούνται και αποτελούν το κίνητρο για την πραγματοποίηση μιας έρευνας αλλά είναι απαιτήσεις μόνο των στατιστικών τεχνικών (Μπεχράκης, 1999). Η απόλυτη αποδοχή της παραπάνω θέσης θα μπορούσε να οδηγήσει σε μια υπέρμετρη απλοποίηση της στατιστικής σκέψης όπου το ενδιαφέρον θα επικεντρώνεται μόνο στην εξεύρεση σωστών, κατά περίπτωση, λύσεων θέτοντας στο περιθώριο τη μαθηματική αυστηρότητα και τη δυνατότητα γενίκευσης.

Η σημαντικότερη διαφορά των μεθόδων της Ανάλυσης Δεδομένων σε σχέση με αυτές της Επαγωγικής Στατιστικής είναι ότι για την εφαρμογή τους δεν απαιτείται η προσαρμογή των δεδομένων σε κάποιο στοχαστικό μοντέλο και η αντίστοιχη συμπερασματολογία δεν υπάγεται σε κάποιο μηχανισμό επαγωγικού συλλογισμού με την έννοια της στατιστικής σημαντικότητας. Η αλήθεια είναι ότι οι μέθοδοι, συνήθως, δεν συνοδεύονται από στατιστικούς ελέγχους. Αυτό όμως δεν σημαίνει ότι δεν είναι δυνατό να υπάρξουν τέτοιοι και μάλιστα με ελάχιστες τεχνικές προϋποθέσεις. Επίσης, δεν δίνεται ιδιαίτερη έμφαση στο μηχανισμό συλλογής και συγκρότησης των διαθέσιμων δεδομένων (μέθοδος δειγματοληψίας, πειραματικός σχεδιασμός, ανεξαρτησία των παρατηρήσεων) αρκεί τα δεδομένα να μπορούν να πινακοποιηθούν σε μορφή κατάλληλη για την εφαρμογή της εκάστοτε μεθόδου. Τα δεδομένα, έστω και αν προέρχονται από δείγμα, αντιμετωπίζονται σαν να αποτελούν ολόκληρο τον υπό εξέταση πληθυσμό (Greenacre, 1984) δίνοντας ένα περιγραφικό και διερευνητικό χαρακτήρα στις μεθόδους αποδυναμώνοντας οποιεσδήποτε γενικεύσεις, όπως αυτές νοούνται στο πλαίσιο της Επαγωγικής Στατιστικής.

Κάθε μέθοδος της Ανάλυσης Δεδομένων στηρίζεται σε ένα σύστημα μαθηματικών μοντέλων, τα οποία αποσκοπούν στην «προσομοίωση» του αντίστοιχου φαινομένου. Γενικά, η προσομοίωση είναι εφικτή μέσω ενός συνόλου συμβόλων και εξισώσεων που έχουν ως σκοπό να περιγράψουν τη συμπεριφορά του σε σχέση με συγκεκριμένα

χαρακτηριστικά ή ιδιότητες που έχουν επιλεγεί και θεωρούνται σημαντικές (Hillier & Lieberman, 1995). Οι συσχετισμοί μεταξύ των πραγματικών δεδομένων ή αλλιώς των πρωτοτύπων (*prototypes*), που χαρακτηρίζουν το φαινόμενο, αντικαθίστανται από ανάλογους συσχετισμούς μεταξύ μαθηματικών οντοτήτων (*entities*) (Davis & Hersh, 1980). Η βασική απαίτηση είναι το μαθηματικό μοντέλο και το φαινόμενο να είναι ισόμορφα σε όλα τα θεμελιώδη, για το πρόβλημα που εξετάζεται, θέματα (Βασιλείου, 1985). Η επιτυχία σε μια τέτοια προσπάθεια έγκειται στην επιλογή και στην κατασκευή ενός μοντέλου που να περιλαμβάνει όσο το δυνατόν περισσότερη πληροφορία σε σχέση με την πραγματικότητα, όπως αυτή αποτυπώνεται με βάση τα διαθέσιμα δεδομένα. Έτσι, αυτόματα εισάγονται αρχικές προϋποθέσεις και υποθέσεις που αφορούν στον τρόπο με τον οποίο θα μετρηθεί η ποσότητα και η ποιότητα της πληροφορίας, στην καταλληλότητα και αντιπροσωπευτικότητα των δεδομένων, στην κωδικοποίησή τους, στους θεωρούμενους συσχετισμούς μεταξύ των πρωτοτύπων και των αντίστοιχων μαθηματικών οντοτήτων, στις μαθηματικές ιδιότητες των μοντέλων, στη δυνατότητα πραγματοποίησης των διαφόρων υπολογισμών καθώς και στην ακρίβειά τους. Με άλλα λόγια, εισάγονται ορισμένες αρχικές συνθήκες - προϋποθέσεις ως προς την εγκυρότητα και αξιοπιστία των μεθόδων οι οποίες θα επιλεγούν τελικά ως οι καταλληλότερες για την ανάλυση των δεδομένων. Οι χρήστες των μεθόδων, ανάλογα με το επιστημονικό πεδίο που δραστηριοποιούνται, θα πρέπει κάθε φορά να αξιολογούν τις προϋποθέσεις εγκυρότητας και να οριοθετούν την ισχύ του αντίστοιχου μοντέλου. Οι αποφάσεις τους είναι δυνατό να επηρεάσουν την παραγόμενη πληροφορία η οποία θα πρέπει να αποκτήσει νόημα σε ένα συγκεκριμένο ερευνητικό και επιστημονικό πλαίσιο. Συνεπώς, είναι απαραίτητη η θέσπιση κάποιων λογικών κανόνων ή κριτηρίων τουλάχιστον για την επιλογή των μεταβλητών μέσω των οποίων θα περιγραφεί το υπό εξέταση φαινόμενο. Αν υπάρχει κάποια δομή στα δεδομένα οι μέθοδοι θα την αναδείξουν. Η δομή θα γίνει ενδεχομένως πιο ξεκάθαρη αν εισαχθούν στις αναλύσεις νέες μεταβλητές από το ίδιο ερευνητικό γνωστικό πεδίο. Η δομή μάλλον θα εξαφανιστεί αν προστεθούν αρκετές μεταβλητές από άλλα πεδία οι οποίες θα λειτουργήσουν συσκοτιστικά ως “θόρυβος”. Παρ’ όλα αυτά, θα πρέπει να γίνει σαφές ότι ένα μοντέλο ή σύστημα μεθόδων ανάλυσης της εμπειρίας δεν μπορεί να χρησιμοποιηθεί για να αποδείξει οτιδήποτε σχετικά με τον πραγματικό κόσμο (De Leeuw, 2005γ) αν και μπορεί, σε μερικές περιπτώσεις, να μας βοηθήσει να ανακαλύψουμε κάποια σημαντικά στοιχεία γι αυτόν. Ένα μοντέλο δεν μπορεί να χαρακτηριστεί σωστό ή λάθος παρά ως χρήσιμο ή μη χρήσιμο (Pfeiffer, 1978).

Άλλωστε, οι μέθοδοι της Ανάλυσης Δεδομένων δεν διαθέτουν κάποιο «μαγικό» μηχανισμό, ο οποίος θα λειτουργήσει με τέτοιο τρόπο ώστε πάντα ένας άμορφος και αδόμητος σωρός από δεδομένα θα μετατραπεί σε επιστημονική γνώση.

Σύμφωνα με τον Breiman (2001), ιδιαίτερη έμφαση θα πρέπει να δίνεται στην κατανόηση του εκάστοτε προβλήματος και των αντίστοιχων εμπειρικών δεδομένων πριν την προσαρμογή και τον έλεγχο οποιουδήποτε μοντέλου. Οι μέθοδοι της Ανάλυσης Δεδομένων μπορούν να συμβάλλουν προς την κατεύθυνση αυτή αποκαλύπτοντας πληροφορία σχετικά με τον συχνά άγνωστο μηχανισμό παραγωγής των δεδομένων.

1.6 Πειραματικοί Σχεδιασμοί στην Ανάλυση Δεδομένων

Πολυμεταβλητά δεδομένα εμφανίζονται στην περίπτωση όπου για κάθε πειραματική ή δειγματοληπτική μονάδα έχουν μετρηθεί περισσότερα από δύο χαρακτηριστικά ή ιδιότητες στο ίδιο ή σε διαφορετικά δείγματα ή πληθυσμούς (Dillon & Goldstein 1984, Johnson & Wichern 1992, Hair *et al*. 1995). Οι πίνακες δεδομένων που αναλύονται είναι της μορφής «*αντικείμενα* × *μεταβλητές*», όπου ο όρος «αντικείμενα» αναφέρεται σε άτομα, πράγματα, γεγονότα ή γενικά σε οποιαδήποτε οντότητα για την οποία υπάρχουν καταγραφές μετρήσεων. Οι μεταβλητές αντιστοιχούν στις διαθέσιμες μετρήσεις, οι οποίες αφορούν, για κάθε αντικείμενο, είτε σε διαφορετικές ιδιότητες είτε και στις ίδιες αλλά κάτω από διαφορετικές ή καταστάσεις στο χώρο και στο χρόνο. Από τη στιγμή που οι καταστάσεις αυτές αποτελούν διαφορετικές συνθήκες οι οποίες έχουν επιλεγεί με βάση κάποιο προκαθορισμένο σχέδιο, όπου οι διαθέσιμες μετρήσεις θεωρούνται ότι εξαρτώνται από γνωστούς και ελεγχόμενους παράγοντες, τότε η όλη διαδικασία μπορεί να ενταχθεί στο πλαίσιο των πειραματικών σχεδιασμών. Η καταλληλότητα των στατιστικών μεθόδων, που θα επιλεγούν για την επεξεργασία των αντίστοιχων δεδομένων, εξαρτάται από το πλήθος των ομάδων των μεταβλητών που θα συμμετάσχουν στην ανάλυση, από τη διάκρισή τους σε εξαρτημένες και ανεξάρτητες καθώς και από τις αντίστοιχες κλίμακες μέτρησής τους (βλέπε Tabachnick & Fidell 1989, Hair *et al*. 1995). Οι κλίμακες μέτρησης των μεταβλητών παίζουν σημαντικό ρόλο όχι μόνο γιατί καθορίζουν, κατά παράδοση, τη στατιστική μέθοδο ανάλυσης που θα χρησιμοποιηθεί αλλά και γιατί επιδρούν στην ακρίβεια και στο βαθμό βεβαιότητας που θα αποδοθούν, αντίστοιχα, στη μέτρηση και

στην ερμηνεία της μεταβλητότητας η οποία παρατηρείται στα διαθέσιμα εμπειρικά δεδομένα (Jacoby, 1999). Η διαπίστωση αυτή χρήζει ιδιαίτερης προσοχής στη στατιστική ανάλυση δεδομένων, τα οποία προέρχονται από πειραματικούς σχεδιασμούς, αφού η ακρίβεια και η αξιοπιστία της παραγόμενης πληροφορίας καθορίζει και την εγκυρότητα του συμπερασμού σχετικά με την απόδοση σχέσεων αιτίας – αποτελέσματος.

Οι κλίμακες μέτρησης των μεταβλητών μπορεί να ποικίλουν ανάλογα με το είδος της αντίστοιχης καταγραφής της μέτρησης. Έτσι, κάποιες μπορεί να είναι ποιοτικής και κάποιες άλλες ποσοτικής φύσης. Η κλίμακα (επίπεδο) μέτρησης των μεταβλητών, δηλαδή αν θα είναι ονομαστική (*nominal*), διάταξης (*ordinal*), διαστήματος (*interval*), αναλογίας (*ratio*) ή δίτιμη (*binary*), καθορίζει, στην πράξη, και την πολυμεταβλητή μέθοδο ανάλυσης που θα εφαρμοστεί σύμφωνα πάντα και με τους επιδιωκόμενους στόχους της έρευνας (βλέπε Dillon & Goldstein 1984, Johnson & Wichern 1992, Hair *et al*. 1995, Sharma 1996, Tacq 1997, Τσάντας και άλλοι 1999, Grimm & Yarnold 2000). Ωστόσο, αυτή η διάκριση των κλιμάκων μέτρησης των μεταβλητών έχει "καταδυναστεύσει" εδώ και καιρό τους ερευνητές σε διάφορα γνωστικά πεδία εισάγοντας αναγκαστικούς περιορισμούς στην επιλογή και εφαρμογή των "επιτρεπόμενων" κάθε φορά στατιστικών μεθόδων (βλέπε Velleman & Wilkinson 1993, Grimm & Yarnold 2000). Χαρακτηριστικά αναφέρουμε την περίπτωση των κλιμάκων τύπου *Likert* (βλέπε Spector 1992, Javeau 1996, Φίλιας 1996), όπου οι συμμετέχοντες στην έρευνα καλούνται να δηλώσουν το βαθμό συμφωνίας τους σε μια σειρά πολυθεματικών δηλώσεων (ερωτήσεων). Οι ερωτώμενοι απαντούν σε μια διαβαθμισμένη κλίμακα συνήθως πέντε σημείων του τύπου «Διαφωνώ Απόλυτα», «Διαφωνώ», «Ούτε Συμφωνώ, Ούτε Διαφωνώ», «Συμφωνώ» και «Συμφωνώ Απόλυτα». Στη συνέχεια, ανατίθενται στις απαντήσεις ακέραιες αριθμητικές τιμές (διαδοχικές ή μη) για τη διευκόλυνση της κωδικοποίησης και της στατιστικής ανάλυσης. Στο σημείο αυτό, αρχίζει η "διαμάχη" μεταξύ των ειδικών μεθοδολόγων σχετικά με το ποια στατιστική μέθοδος είναι κατάλληλη για την επεξεργασία δεδομένων αυτού του τύπου (βλέπε O'Brien 1979, Cohen & Cohen 1983, Srinivasan & Basu 1989, Higgs 1991, Javeau 1996, Hand 1996, Μπεχράκης 1999, Grimm & Yarnold 2000, Blasius & Thiessen 2000, Reed 2002, Πρίπορας & Μενεξές 2005). Οι αντίστοιχες μεταβλητές άλλες φορές αντιμετωπίζονται ως ποσοτικές ίσων διαστημάτων (Higgs 1991, Poon & Hung 1996) και άλλες ως ποιοτικές με

διαβαθμισμένες κατηγορίες (Κιοσέογλου, 2002). Στην πρώτη περίπτωση, οι μεταβλητές χρησιμοποιούνται χωρίς καμία επιφύλαξη σε στατιστικές διαδικασίες όπως η Ανάλυση σε Κύριες Συνιστώσες, η οποία προϋποθέτει τη χρήση ποσοτικών μεταβλητών. Στη δεύτερη, τα δεδομένα αναλύονται με μεθόδους κυρίως της Μη Παραμετρικής Στατιστικής.

Σύμφωνα με τον Kachigan (1991), η κλίμακα μέτρησης μιας μεταβλητής είναι μια έννοια θεωρητική και δεν αποτελεί παρά μια "γέφυρα" ανάμεσα στην εμπειρική παρατήρηση και στους αριθμούς. Πρόκειται απλά για ένα σχήμα αριθμητικής αντιστοίχησης των τιμών της μεταβλητής, το οποίο θα πρέπει να αποκτήσει νόημα μέσα σε ένα συγκεκριμένο ερευνητικό πεδίο, ώστε να είναι χρήσιμο τελικά. Άλλωστε, οι αριθμοί δεν γνωρίζουν τι μετρούν. Για τον Jacoby (1999) οι κλίμακες μέτρησης των μεταβλητών αποτελούν στην ουσία έλεγχο μιας θεωρίας και όχι μια *a priori* παραδοχή, αφού εκφράζουν μια αμφισβητούμενη η αβέβαιη δήλωση σχετικά με τη φύση της πραγματικότητας. Η απεικόνιση από ένα εμπειρικό σύστημα μέτρησης της πραγματικότητας σε ένα αριθμητικό δεν είναι μονοσήμαντη (Hand, 1996) αφού δεν χαρακτηρίζει με μοναδικό τρόπο τους αριθμούς ή τα σύμβολα που θα χρησιμοποιηθούν για να εκφράσουν τις αντίστοιχες καταγραφές ή μετρήσεις. Για παράδειγμα, το μήκος μπορεί να μετρηθεί σε μέτρα, σε εκατοστά, σε κλάσεις εύρους του τύπου «από … έως» ή ακόμα και να αποδοθεί λεκτικός χαρακτηρισμός για το "μεγάλο" ή το "μικρό" μήκος. Όλες οι προηγούμενες μετρήσεις είναι έγκυρες αλλά με διαφορετικό βαθμό ακρίβειας στην μέτρηση της μεταβλητότητας του μήκους. Όμως, σύμφωνα με τους στόχους της εκάστοτε μελέτης κάποια ή κάποιες απ' αυτές θα μπορούσαν να είναι περισσότερο λειτουργικές και αποκαλυπτικές για το υπό εξέταση φαινόμενο. Οι μέθοδοι της Ανάλυσης Δεδομένων που κάνουν χρήση των τεχνικών βέλτιστης κλιμάκωσης είναι δυνατό να συμβάλλουν στον έλεγχο των ιδιοτήτων των κλιμάκων μέτρησης των μεταβλητών που περιγράφουν ένα συγκεκριμένο σύνολο δεδομένων (Perreault & Young 1980, Gifi 1996, Jacoby 1999, Meulman & Heiser 2004).

Στο πλαίσιο της Ολλανδικής Σχολής Ανάλυσης Δεδομένων η κλίμακα μέτρησης μιας μεταβλητής είναι έννοια σχετική (Gifi 1996, Meulman & Heiser 2004), η οποία μπορεί να αναλυθεί σε δύο άξονες. Ο πρώτος αφορά στον τρόπο καταγραφής και κωδικοποίησης των αντίστοιχων μετρήσεων που είναι συνήθως αποτέλεσμα

συμβάσεων. Για παράδειγμα, οι τιμές της μεταβλητής «*Φύλο*» θα μπορούσαν να κωδικοποιηθούν αποδίδοντας την αριθμητική τιμή 0 στους άνδρες και την τιμή 1 στις γυναίκες (ή το αντίστροφο). Μια άλλη καταγραφή θα μπορούσε να είναι «Α» για τους άνδρες και «Γ» για τις γυναίκες ή οποιαδήποτε άλλη αριθμητική ή αλφαριθμητική ακολουθία. Στην πρώτη περίπτωση, ο υπολογισμός του μέσου όρου για τη μεταβλητή «*Φύλο*», σε ένα δείγμα μεγέθους *N*, εκφράζει το ποσοστό των γυναικών, ενώ στη δεύτερη περίπτωση απλά δεν έχει νόημα. Ο δεύτερος άξονας αφορά σε εννοιολογικές και λειτουργικές αποφάσεις, που πρέπει να ληφθούν από τους ίδιους τους ερευνητές, σε σχέση με τις κλίμακες μέτρησης των μεταβλητών ανεξάρτητα από τον πρωτογενή τρόπο καταγραφής και κωδικοποίησής τους. Οι αποφάσεις λαμβάνονται σε σχέση με τους επιδιωκόμενους στόχους αλλά κυρίως με βάση το θεωρητικό υπόβαθρο στο οποίο στηρίζεται η έρευνα και η ερμηνεία των αποτελεσμάτων (Meulman & Heiser, 2004). Χαρακτηριστικά αναφέρουμε ότι η μεταβλητή «*Φύλο*» γενικά θεωρείται ονομαστικής κλίμακας αλλά σε ένα ερευνητικό πλαίσιο όπου διερευνώνται παράμετροι μυϊκής και σωματικής δύναμης η ίδια μεταβλητή θα μπορούσε να θεωρηθεί ως διάταξης ή ακόμα και ποσοτική με την έννοια ότι η τιμή «Άνδρας» αντιστοιχεί σε μεγαλύτερη εν γένει σωματική δύναμη απ’ ότι η τιμή «Γυναίκα». Οι τεχνικές βέλτιστης κλιμάκωσης των Ολλανδών δίνουν τη δυνατότητα στο χρήστη των αντίστοιχων μεθόδων να καθορίσει την κλίμακα μέτρησης των μεταβλητών, η οποία θα διατηρηθεί και στα αποτελέσματα των αναλύσεων, εισάγοντας τους κατάλληλους περιορισμούς στον αλγόριθμο βελτιστοποίησης της αντίστοιχης συνάρτησης απώλειας. Για παράδειγμα, αν η κατηγορική μεταβλητή «*Ηλικία*» έχει καταγραφεί με 5 κλάσεις – κατηγορίες (π.χ. 20-30, 31-40, 41-50, 51-65 και «άνω των 65» ετών) και επιθυμούμε η διάταξη των κατηγοριών της να διατηρηθεί και στους παραγοντικούς άξονες που θα προκύψουν από την εφαρμογή για παράδειγμα της Κατηγορικής Ανάλυσης σε Κύριες Συνιστώσες (βλέπε Van de Geer 1993α και 1993β, Gifi 1996, De Leeuw 2005β) τότε η μεταβλητή «*Ηλικία*» θα πρέπει να δηλωθεί με κλίμακα μέτρησης διάταξης. Αν κάτι τέτοιο δεν είναι επιθυμητό ή το θεωρητικό πλαίσιο της μελέτης δεν επιτρέπει μια τέτοια θεώρηση τότε η «*Ηλικία*» θα πρέπει να δηλωθεί ως ονομαστική κλίμακα. Βέβαια, τέτοιες προσεγγίσεις έρχονται σε αντίθεση με τη φιλοσοφία του Benzécri, σύμφωνα με την οποία τα δεδομένα θα πρέπει να αναλύονται χωρίς υποθέσεις και περιορισμούς. Για παράδειγμα, στην ΠΑΑ οι κλίμακες διάταξης αντιμετωπίζονται με τον ίδιο τρόπο όπως και οι ονομαστικές κλίμακες. Η ΠΑΑ, λόγω της ιδιότητας της βέλτιστης κλιμάκωσης που τη

χαρακτηρίζει (Nishisato 1980, Greenacre 1993α και 1984, Gifi 1996), θα αναδείξει τη διάταξη των κατηγοριών αν όντως αυτή υπάρχει και δεν αποτελεί παρά μόνο μια εννοιολογική ή συμβατική παραδοχή. Σε πολλές περιπτώσεις, μεταβλητές που πρωτογενώς έχουν καταγραφεί ως ποσοτικές, στη συνέχεια, για τεχνικούς (Dillon & Goldstein, 1984) ή/και θεωρητικούς λόγους (Nishisato 1980, Meulman & Heiser 2004), οι τιμές τους κατηγοριοποιούνται σε κλάσεις, με βάση λογικά ή στατιστικά κριτήρια, με αποτέλεσμα η κλίμακα μέτρησής τους από ποσοτική να μετατραπεί σε ποιοτική και μάλιστα διάταξης (Παπαδημητρίου, 2001). Άλλωστε για τους Ολλανδούς ακόμα και οι ποσοτικές μεταβλητές μπορούν να θεωρηθούν σε ένα γενικό επίπεδο ως ονομαστικής κλίμακας με μεγάλο αριθμό κατηγοριών (Bekker & De Leeuw 1988, Gifi 1996, Michailidis & De Leeuw 1998). Η σχετικότητα με την οποία είναι συνυφασμένη η έννοια της κλίμακας μέτρησης των μεταβλητών υποστηρίζεται και από το γεγονός ότι μέθοδοι της Ανάλυσης Δεδομένων, όπως η ΠΑΑ, που είναι κατάλληλες για την ανάλυση ποιοτικών μεταβλητών, έχουν ως αποτέλεσμα την ποσοτικοποίησή τους μέσω της αντιστοίχισης αριθμητικών τιμών, με την έννοια των συντεταγμένων επί των παραγοντικών αξόνων. Όπως αναφέρθηκε στην Ενότητα 1.4.2, οι νέες ποσοτικοποιημένες μεταβλητές μπορούν να χρησιμοποιηθούν, στη συνέχεια, σε στατιστικές μεθόδους που είναι κατάλληλες για την επεξεργασία ποσοτικών δεδομένων. Η παρατήρηση αυτή είναι ιδιαίτερα σημαντική για τη στατιστική ανάλυση πειραματικών δεδομένων όπου οι εξαρτημένες μεταβλητές είναι κατηγορικές, αφού μέσω της ΠΑΑ μπορούν αρχικά να ποσοτικοποιηθούν βέλτιστα και κατόπιν να εφαρμοστούν σε αυτές μέθοδοι της Επαγωγικής Στατιστικής όπως είναι η Ανάλυση Διακύμανσης.

Η Πολυμεταβλητή Στατιστική Ανάλυση περιλαμβάνει γενικά όλες τις στατιστικές μεθόδους, οι οποίες αναλύουν ταυτόχρονα περισσότερες από δύο τυχαίες μεταβλητές (ποιοτικές ή/και ποσοτικές) με τρόπο ώστε η μελέτη της επίδρασης ή της συμπεριφοράς της κάθε μεταβλητής μεμονωμένα είτε να μην έχει νόημα είτε να μην μπορεί να ερμηνευτεί ανεξάρτητα από τις άλλες (Hair *et al*., 1995). Οι στόχοι της Πολυμεταβλητής Ανάλυσης μπορούν να αναπτυχθούν συνοπτικά σε πέντε άξονες (Harris, 2001):

1) Περιορισμός των δεδομένων, με την έννοια της μείωσης των αρχικών διαστάσεων, και απλοποίηση της δομής του πίνακα δεδομένων. Το υπό εξέταση φαινόμενο

αναπαρίσταται όσο το δυνατόν απλούστερα σε ευκλείδειους χώρους, συνήθως δύο ή τριών διαστάσεων, χωρίς να θυσιαστεί σημαντικό μέρος της πληροφορίας που περιέχεται στον αρχικό πίνακα δεδομένων, με την ελπίδα ότι μέσω αυτής της διαδικασίας θα καταστεί η ερμηνεία του φαινομένου ευκολότερη.

2) Ταξινόμηση, διάταξη και ομαδοποίηση των αντικειμένων ή/και των μεταβλητών με βάση μετρήσιμα ποιοτικά ή ποσοτικά χαρακτηριστικά.

3) Διερεύνηση των σχέσεων μεταξύ των μεταβλητών.

4) Πρόβλεψη. Οι συσχετίσεις διερευνώνται με τέτοιο τρόπο ώστε να είναι δυνατή η πρόβλεψη των τιμών μιας ή περισσότερων μεταβλητών από τις τιμές των άλλων.

5) Έλεγχοι προκαθορισμένων ερευνητικών υποθέσεων και διατύπωση νέων.

Οι αντίστοιχες στατιστικές διαδικασίες είναι δυνατό να ταξινομηθούν σε δύο βασικές κατηγορίες. Η πρώτη, περιλαμβάνει τις μεθόδους εκείνες όπου δεν γίνεται διάκριση μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών (*interdependence methods*), ενώ η δεύτερη αυτές στις οποίες υπάρχει τέτοιου είδους διάκριση (*dependence methods*) (Dillon & Goldstein 1984, Hair *et al*. 1995, Σιάρδος 1999). Χαρακτηριστικά παραδείγματα της πρώτης κατηγορίας αποτελούν οι βασικές μέθοδοι της Ανάλυσης Δεδομένων όπως η Ανάλυση σε Κύριες Συνιστώσες, η Παραγοντική Ανάλυση των Αντιστοιχιών, η Ανάλυση σε Συστάδες (Ταξινόμηση), η Παραγοντική Ανάλυση και η Πολυδιάστατη Κλιμάκωση (μετρική, μη μετρική). Η δεύτερη κατηγορία περιλαμβάνει μεθόδους όπως η Πολλαπλή Παλινδρόμηση (γραμμική, μη γραμμική), η Διακρίνουσα Ανάλυση (παραμετρική, μη παραμετρική), η Πολυμεταβλητή Ανάλυση Διακύμανσης και η Κανονικοποιημένη Συσχέτιση. Οι μέθοδοι της πρώτης κατηγορίας έχουν χαρακτήρα κυρίως διερευνητικό και στοχεύουν στην ανάδειξη και περιγραφή σύνθετων δομών που δεν είναι άμεσα αντιληπτές και μετρήσιμες. Χρησιμοποιούνται κυρίως σε έρευνες επισκόπησης όπου το ενδιαφέρον εστιάζεται απλά στη διερεύνηση των συσχετίσεων μεταξύ των μεταβλητών και όχι στην επιβεβαίωση κάποιου υποδείγματος σχέσης αιτίας – αποτελέσματος, όπως συμβαίνει στους πειραματικούς σχεδιασμούς (Tabachnick & Fidell 1989, Kachigan 1991). Οι μέθοδοι της πρώτης κατηγορίας αντιμετωπίζουν τις μεταβλητές συμμετρικά και σπανίως συνοδεύονται από ελέγχους στατιστικής σημαντικότητας αφού είναι

επικεντρωμένες στα δεδομένα. Οι διαδικασίες της δεύτερης κατηγορίας έχουν κατεύθυνση επιβεβαιωτική και χρησιμοποιούνται συνήθως για ελέγχους υποθέσεων και προβλέψεις στο πλαίσιο της Επαγωγικής Στατιστικής. Έτσι, οι μέθοδοι της δεύτερης κατηγορίας μπορούν να χρησιμοποιηθούν σε πειραματικούς σχεδιασμούς αλλά και σε έρευνες επισκόπησης, όπου υπάρχει τουλάχιστον εννοιολογική διάκριση μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών. Σε κάθε περίπτωση, η εγκυρότητα των συμπερασμάτων, που προκύπτουν από την εφαρμογή στατιστικών ελέγχων σημαντικότητας, εξαρτάται από την ικανοποίηση αρκετών τεχνικών και θεωρητικών προϋποθέσεων (Gifi 1996, De Leeuw 2005γ). Βέβαια, αν και η μη ικανοποίησή τους καθιστά οποιαδήποτε επαγωγική στατιστική συμπερασματολογία επισφαλή, ωστόσο το γεγονός αυτό δεν αποτελεί δεσμευτικό παράγοντα για τη χρήση των μεθόδων της δεύτερης κατηγορίας και για διερευνητικούς - περιγραφικούς σκοπούς (Manly, 1994).

Οι Πολυμεταβλητές Στατιστικές Αναλύσεις αναπτύχθηκαν για την ανάλυση κυρίως μη πειραματικών δεδομένων (Tabachnick & Fidell, 1989). Ωστόσο, ορισμένες από αυτές, όπως η Πολυμεταβλητή Ανάλυση Διακύμανσης, είναι αρκετά αποτελεσματικές στην ανάλυση πειραματικών σχεδιασμών που περιλαμβάνουν πολλές εξαρτημένες μεταβλητές, οι οποίες μετρούν ταυτόχρονα πολλαπλές ιδιότητες ή χαρακτηριστικά των πειραματικών μονάδων καθιστώντας τη μελέτη του υπό εξέταση φαινομένου πιο ρεαλιστική. Όμως, σχεδόν κατά αποκλειστικότητα, οι εξαρτημένες μεταβλητές είναι ποσοτικές και όχι κατηγορικές. Στο χώρο των Κοινωνικών Επιστημών η συλλογή ποσοτικών (συνεχών μετρήσεων) είναι μάλλον σπάνια (Μπεχράκης, 1999). Στις έρευνες επισκόπησης χρησιμοποιούνται ερωτηματολόγια που περιλαμβάνουν ως επί το πλείστον κλειστού τύπου ερωτήσεις με προκαθορισμένες απαντήσεις, οι οποίες οδηγούν στη συγκέντρωση κυρίως μη μετρικών (κατηγορικών) δεδομένων. Στην πράξη, η μεθοδολογική αυτή προσέγγιση έχει τα ακόλουθα πλεονεκτήματα (Dillon & Goldstein, 1984):

- Μερικές φορές είναι πιο λειτουργικό για τον ερευνητή να συλλέξει δεδομένα σε ονομαστική ή τακτική κλίμακα, δηλαδή με μικρό αριθμό διακεκριμένων τιμών, παρά σε συνεχή κλίμακα κάνοντας χρήση πραγματικών αριθμών.
- Οι ερωτώμενοι αισθάνονται μικρότερη “απειλή” όταν απαντούν κυρίως σε ευαίσθητες και προσωπικές ερωτήσεις (π.χ. εισόδημα και ηλικία) σε μια

διαβαθμισμένη κλίμακα με εύρος τιμών (π.χ. «από … έως»), απ' ότι όταν καλούνται να απαντήσουν ποσοτικά με μια ακριβή τιμή (πραγματικό αριθμό). Έτσι, θα μπορούσε να μειωθεί ο αριθμός των μη απαντημένων ερωτήσεων και να αυξηθεί ο αριθμός των συμμετεχόντων στην έρευνα.

- Ικανοποιεί την εκτίμηση του ερευνητή ότι τα υποκείμενα της έρευνας δεν έχουν την απαιτούμενη εμπειρία (γνώση, διάθεση) να κρίνουν και να απαντήσουν με ακρίβεια σε μια συνεχή αριθμητική κλίμακα.

Στα παραπάνω πλεονεκτήματα μπορούν να προστεθούν και τα ακόλουθα:

- Οι κλειστού τύπου ερωτήσεις διευκολύνουν την εισαγωγή των δεδομένων σε Η/Υ με σύγχρονες μεθόδους (π.χ. μέσω σαρωτών) ιδιαίτερα στην περίπτωση μεγάλου όγκου στοιχείων.
- Επιτυγχάνεται η πρακτική ή κλινική αξιοποίηση των στοιχείων που συγκεντρώνονται, αφού συχνά είναι δύσκολη η ποσοτική εκτίμηση ποιοτικών χαρακτηριστικών ή ιδιοτήτων. Αξίζει να σημειωθεί ότι σε ορισμένες περιπτώσεις, έχει μεγαλύτερη σημασία ή αξία η γνώση του εύρους ή της τάξης των παρατηρούμενων μεγεθών (π.χ. «μικρό, μεσαίο ή μεγάλο εισόδημα», «κέρδος ή ζημία» και «επιτυχία ή αποτυχία») και όχι η ακριβής αριθμητική τιμή τους. Με τον τρόπο αυτό, έχουμε ένα ποιοτικό πέρασμα από τη συνεχή «ακριβή μέτρηση» σε διακεκριμένες καταστάσεις στις οποίες μπορεί να βρεθεί ένα μέγεθος - μεταβλητή. Συχνά, επίσης, τίθεται και το θέμα της αξιοπιστίας των ποσοτικών μετρήσεων. *Θέση μας είναι ότι είναι προτιμότερο να έχουμε μια ακριβή ποιοτική κατάταξη που αναδεικνύει ιδιότητες μέσα σε κάθε μεταβλητή παρά μια μεροληπτική αριθμητική καταγραφή*.

Η Γαλλική Σχολή Ανάλυσης Δεδομένων έχει αναδείξει μεθόδους για την ανάλυση κυρίως πολυδιάστατων κατηγορικών δεδομένων όπου οι μεταβλητές αντιμετωπίζονται ισότιμα χωρίς διάκριση μεταξύ εξαρτημένων και ανεξάρτητων. Η συμμετρική αυτή αντιμετώπιση είχε ως αποτέλεσμα τον *a priori* αποκλεισμό των αντίστοιχων μεθόδων σε πειραματικούς σχεδιασμούς όπου εκ κατασκευής υπεισέρχονται τουλάχιστον δύο ομάδες μεταβλητών. Η ομάδα των ανεξάρτητων και η ομάδα των εξαρτημένων. Οι ανεξάρτητες μεταβλητές θεωρητικά βρίσκονται κάτω από τον άμεσο έλεγχο του πειραματιστή, ο οποίος μετά από κατάλληλη μεταβολή της

κατάστασης των ανεξάρτητων μελετά και μετρά την επίδραση των μεταβολών αυτών στη συμπεριφορά – τιμές των εξαρτημένων. Συνήθως, στους πειραματικούς σχεδιασμούς οι εξαρτημένες μεταβλητές είναι ποσοτικές. Αυτός είναι πιθανόν ένας άλλος λόγος για τον οποίο μέθοδοι όπως η ΠΑΑ, που είναι κατάλληλη για τη στατιστική επεξεργασία κατηγορικών δεδομένων, δεν βρήκε τη θέση που της αρμόζει στο χώρο των πειραματικών σχεδιασμών. Από την άλλη πλευρά, τα πειραματικά δεδομένα υποβάλλονται σε ελέγχους σημαντικότητας με σκοπό να διαπιστωθεί αν οι παρατηρούμενες μεταβολές στις τιμές των εξαρτημένων μεταβλητών είναι αποτέλεσμα της συστηματικής επίδρασης των ανεξάρτητων ή μπορούν να αποδοθούν σε άλλους τυχαίους, αστάθμητους μη ελεγχόμενους παράγοντες. Εν γένει, οι μέθοδοι της Ανάλυσης Δεδομένων έχουν διερευνητικό χαρακτήρα και αντιμετωπίζουν τα διαθέσιμα δεδομένα σαν να αποτελούν ολόκληρο τον υπό εξέταση πληθυσμό ασχέτως του γεγονότος ότι αυτά μπορεί να προέρχονται από ένα δείγμα του. Έτσι, τουλάχιστον από επιστημολογική σκοπιά, δεν φαίνεται να υπάρχουν και πολλά περιθώρια για την εφαρμογή στατιστικών ελέγχων (βλέπε Benzécri & *Collaborateurs*, 1973). Στους πειραματικούς σχεδιασμούς η στατιστική σημαντικότητα ενός ευρήματος αποτελεί ομολογουμένως έναν χρήσιμο οδηγό για τη λήψη αποφάσεων κάτω από συνθήκες σχετικής αβεβαιότητας. Όμως, η εγκυρότητα των συμπερασμάτων εξαρτάται από ένα πλήθος προϋποθέσεων που σπάνια ικανοποιούνται στην πράξη (Gifi, 1996). Οι προϋποθέσεις αυτές αφορούν στις αντίστοιχες απαιτήσεις των στατιστικών ελέγχων και δεν έχουν καμία σχέση με τα αντίστοιχα επιστημονικά ερωτήματα, τα οποία προϋπάρχουν ακόμα και αυτής της ίδιας της έρευνας (Μπεχράκης, 1999). Η λογική της επαγωγικής συμπερασματολογίας και ιδιαίτερα η απόδοση σχέσεων αιτίας – αποτελέσματος, όπως συμβαίνει στους πειραματικούς σχεδιασμούς, είναι συνάρτηση του τρόπου με τον οποίο έχουν παραχθεί τα διαθέσιμα δεδομένα και όχι της στατιστικής μεθόδου με την οποία θα αναλυθούν (Cohen & Cohen 1983, De Leeuw 2005γ). Άλλωστε, η αδυναμία πρόβλεψης με βεβαιότητα του για το τι θα συμβεί κατά την εκτέλεση ενός πειράματος, είναι αυτή που οδηγεί στην έννοια της τυχαιότητας, η οποία όμως αναφέρεται στα αποτελέσματα και όχι στο μηχανισμό εκτέλεσής του. Συμπερασματικά, μπορούμε να θεωρήσουμε τον πειραματικό σχεδιασμό απλά ως μία ακόμη διαδικασία παραγωγής δεδομένων, τα οποία, στη συνέχεια, θα πρέπει να αναλυθούν. Για τον De Leeuw (2005γ) οι σχέσεις αιτίας – αποτελέσματος και η σταθερότητα των αποτελεσμάτων των εμπειρικών ερευνών τεκμηριώνονται μέσα από

προσεχτικά σχεδιασμένα πειράματα και την επανάληψη των μελετών και όχι μέσω κάποιου μαθηματικο-στατιστικού φορμαλισμού. Ο κύριος ρόλος της Στατιστικής είναι η περιγραφή των εμπειρικών πειραματικών ή δειγματοληπτικών δεδομένων με τέτοιο τρόπο ώστε οι ερευνητές να είναι σε θέση οι ίδιοι να κάνουν τις προβλέψεις και τις γενικεύσεις τους. Όμως, η ανάγκη να κατανοήσουμε τις εν γένει πολύπλοκες σχέσεις μεταξύ ενός μεγάλου αριθμού μεταβλητών καθιστά εκ φύσεως την πολυμεταβλητή στατιστική ανάλυση ένα δύσκολο αντικείμενο (Johnson & Wichern, 1992). Η δυσκολία εντείνεται και αποκτά κρίσιμη διάσταση στην περίπτωση που το ζητούμενο είναι η απόδοση σχέσεων αιτίας - αποτελέσματος ή/και η πρόβλεψη.

Ο καθορισμός των στατιστικών μεθόδων που θα εφαρμοστούν στα διαθέσιμα πειραματικά ή δειγματοληπτικά δεδομένα απαιτεί γενικά την ένταξή τους σε μια τυπική μεθοδολογική πορεία, η οποία περιλαμβάνει έναν ορισμένο αριθμό φάσεων. Στο πλαίσιο τυποποίησης της όλης διαδικασίας, σε κάθε μία από αυτές τις φάσεις εκτελούνται συγκεκριμένες εργασίες και παράγεται αντίστοιχο υλικό τεκμηρίωσης, με βάση το οποίο θα εφαρμοστούν οι κατάλληλες στατιστικές διαδικασίες. Σε ένα πρακτικό επίπεδο, για την εκτέλεση κάθε φάσης απαιτούνται πόροι, οι οποίοι διατίθενται συνήθως με γνώμονα τους επιδιωκόμενους στόχους. Στο Παράρτημα Α παρουσιάζουμε μια δομημένη κυκλική μεθοδολογική πορεία επτά συσχετιζόμενων φάσεων στην Πολυμεταβλητή Στατιστική Ανάλυση. Η Ανάλυση Δεδομένων, ως υποσύνολο των πολυμεταβλητών στατιστικών μεθόδων, έχει να προσφέρει ορισμένα πλεονεκτήματα, τα οποία μπορούν να συμβάλλουν, σε ορισμένες περιπτώσεις, στην επιτυχημένη διεξαγωγή πολλών φάσεων ενός ερευνητικού έργου:

- Ο μη παραμετρικός χαρακτήρας των μεθόδων, οι οποίες έχουν ελάχιστες τεχνικές προϋποθέσεις και δεν απαιτούν τα διαθέσιμα δεδομένα να χαρακτηρίζονται από συγκεκριμένες στατιστικές ιδιότητες. Το γεγονός αυτό μπορεί να διευκολύνει τη διαδικασία συλλογής και επεξεργασίας των δεδομένων, περιορίζοντας έτσι τον απαιτούμενο χρόνο και το κόστος διεξαγωγής της έρευνας.
- Η δυνατότητα των μεθόδων να χειριστούν και ποιοτικά χαρακτηριστικά (μεταβλητές), τα οποία συχνά δεν διαθέτουν τις απαραίτητες στατιστικές ιδιότητες που προαπαιτούνται σε άλλες στατιστικές προσεγγίσεις.

- Η μη γραμμικότητα των σχέσεων μεταξύ των μεταβλητών. Σε αντίθεση με τις στατιστικές διαδικασίες, οι οποίες στηρίζονται αποκλειστικά σε γραμμικά υποδείγματα, οι μέθοδοι μπορούν να αναδείξουν είτε γραμμικές είτε μη γραμμικές σχέσεις ανάλογα με τα δεδομένα. Τα παραγόμενα στατιστικά αποτελέσματα (δείκτες, διαγράμματα) έχουν απτή φυσική ερμηνεία και είναι δυνατό να αναδείξουν ισχυρούς, αλλά λανθάνοντες, μηχανισμούς που οδήγησαν στη συγκρότηση των δεδομένων, στοιχείο το οποίο θεωρείται απαραίτητο για την ερμηνεία και κατανόηση του υπό εξέταση φαινομένου.
- Η αυξημένη ευελιξία στην ανάπτυξη νέων μεθόδων ή στη βελτίωση των υπαρχόντων με τη χρήση τεχνικών γραμμικής ή μη γραμμικής βελτιστοποίησης, με σκοπό την καλύτερη ερμηνεία των δεδομένων και τον περιορισμό της πολυπλοκότητάς τους.
- Η εφαρμογή τους σε ειδικές περιπτώσεις, όπου άλλες στατιστικές μέθοδοι απλά δεν μπορούν να εφαρμοστούν. Χαρακτηριστικά αναφέρουμε τις παρακάτω καταστάσεις όπου: α) το δείγμα που έχει συλλεχθεί είναι συμπτωματικό (μη πιθανότητας) και όχι τυχαίο, β) οι πειραματικοί σχεδιασμοί είναι ατελείς και δεν υπακούν στη δομημένη πορεία που παρουσιάστηκε στην Ενότητα 0 (*quasi experimental designs*) και γ) τα πειραματικά δεδομένα έχουν παραχθεί από ένα μόνο υποκείμενο (*single subject experiments*) (βλέπε Mertens, 1998). Και στις τρεις περιπτώσεις δεν μπορούν να εφαρμοστούν, για παράδειγμα, μέθοδοι της Επαγωγικής Στατιστικής.

Τα πλεονεκτήματα που αναφέρθηκαν παραπάνω περιορίζουν τους βρόγχους ανατροφοδότησης μεταξύ των φάσεων (βλέπε Σχήμα Α1 του Παραρτήματος Α) και συντομεύουν τις επαναλήψεις της κυκλικής πορείας εξοικονομώντας ταυτόχρονα και πόρους. Οι μέθοδοι της Ανάλυσης Δεδομένων αν και έχουν, σε πρακτικές εφαρμογές, υψηλή αποτελεσματικότητα, ωστόσο δεν αποτελούν πανάκεια. Δεδομένης της πληθώρας των διαθέσιμων πολυμεταβλητών στατιστικών τεχνικών, η επιλογή της πλέον αποτελεσματικής μεθοδολογίας, όποια πλεονεκτήματα κι αν έχει, μπορεί να στοιχειοθετηθεί μόνο σε συγκεκριμένα πεδία εφαρμογών και με συγκεκριμένα δεδομένα.

Στην επόμενη ενότητα θα αναφερθούμε στη δομή και τη συνεισφορά της παρούσας διατριβής έχοντας ως γνώμονα κυρίως τις Φάσεις 3, 5 και 6 της μεθοδολογικής πορείας που παραθέτουμε στην Ενότητα Α1 του Παραρτήματος Α.

1.7 Δομή και Συνεισφορά της Διατριβής

Στο Κεφάλαιο 2 επιχειρούμε μια συγκριτική παρουσίαση (αλγοριθμική και εννοιολογική) της ΠΑΑ, όπως αυτή θεμελιώνεται και εφαρμόζεται στο πλαίσιο της Γαλλικής και Ολλανδικής Σχολής Ανάλυσης Δεδομένων. Παρουσιάζουμε τις ομοιότητες και τις διαφορές τους και δείχνουμε με ποιον τρόπο οι δύο προσεγγίσεις συνδέονται. Σε κάθε περίπτωση, επισημαίνουμε τα σημεία που χρήζουν ιδιαίτερης προσοχής κατά την ερμηνεία των αποτελεσμάτων. Προβάλουμε τις ιδιότητες και τις δυνατότητες της μεθόδου ΠΑΑ και παραθέτουμε αναφορές σχετικά με τις σημαντικότερες εξελίξεις στο ερευνητικό της πλαίσιο. Θίγουμε, επίσης, θέματα που αφορούν στην εσωτερική σταθερότητα των εκροών της ΠΑΑ και αναφερόμαστε στη συμπληρωματικότητά της με άλλες στατιστικές μεθόδους (π.χ. Ταξινόμηση, Καμπύλες *Andrews* και διαγράμματα *Biplot*). Τέλος, προτείνουμε νέα εννοιολογικά και μεθοδολογικά στοιχεία στο πεδίο εφαρμογής της. Πιο συγκεκριμένα, στη διμεταβλητή εκδοχή της ΠΑΑ, προτείνουμε μέθοδο εντοπισμού των κελιών του πίνακα συμπτώσεων, τα οποία συνεισφέρουν σημαντικά στην αδράνεια των παραγοντικών αξόνων. Στην περίπτωση πολλών μεταβλητών, ορίζουμε το «Σχετικό Δείκτη Διακριτότητας» μιας μεταβλητής και παρουσιάζουμε τη χρησιμότητα του νέου αυτού δείκτη στην ερμηνεία των αποτελεσμάτων. Και οι δύο προτάσεις συμβάλουν στη διεισδυτική ερμηνεία των δομικών σχέσεων μεταξύ των μεταβλητών. Προς την ίδια κατεύθυνση συνεισφέρει και ο συνδυασμός των ιδιοτήτων της ΠΑΑ και της Ανάλυσης Ομοιογένειας, όπως αυτές αναδεικνύονται μέσα από τη συγκριτική παράθεση των μεθοδολογικών προσεγγίσεων των δύο Σχολών Ανάλυσης Δεδομένων.

Στο πρώτο μέρος του Κεφαλαίου 3 περιγράφουμε μια διαδικασία βάσει της οποίας είναι δυνατή η εφαρμογή της ΠΑΑ, μέσω του στατιστικού πακέτου SPSS, σε γενικευμένους πίνακες συμπτώσεων απολύτων συχνοτήτων (*Burt*), στο πνεύμα της Γαλλικής Σχολής. Να τονίσουμε ότι η δυνατότητα αυτή δεν είναι άμεσα διαθέσιμη στο SPSS. Δεδομένου ότι στην πολυμεταβλητή εκδοχή της η μέθοδος εφαρμόζεται σε πίνακες *Burt* και όχι στους αντίστοιχους λογικούς, οι οποίοι έχουν συνήθως μεγάλο

πλήθος γραμμών, στο δεύτερο μέρος του κεφαλαίου, προτείνουμε έναν αποτελεσματικό αλγόριθμο εφαρμογής της ΠΑΑ σε μεγάλους λογικούς πίνακες. Ο αλγόριθμος βρίσκει εφαρμογές και στη μέθοδο της Ταξινόμησης σε Αύξουσα Ιεραρχία.

Στο Κεφάλαιο 4 αποδεικνύουμε επτά νέες βασικές μαθηματικές σχέσεις που συνδέουν τις αδράνειες πινάκων συμπτώσεων, *Burt* και λογικών δύο ή περισσότερων κατηγορικών μεταβλητών. Διατυπώνουμε τις αντίστοιχες προτάσεις και συνάγουμε τρία πορίσματα. Οι σχέσεις αδράνειας εκφράζουν την ποσότητα και κυρίως την ποιότητα της πληροφορίας που παράγεται από την ΠΑΑ. Αναδεικνύουν, επίσης, τη φυσική ερμηνεία της ολικής αδράνειας, ανάλογα με τη μορφή του πίνακα που αναλύεται κάθε φορά. Στο ίδιο κεφάλαιο, ορίζουμε την έννοια της «Ενδιαφέρουσας Αδράνειας» του πίνακα *Burt* και προτείνουμε έλεγχο για τη στατιστική σημαντικότητά της. Με βάση την Ενδιαφέρουσα Αδράνεια αναπτύσσουμε διαδικασία εντοπισμού υποπίνακα του *Burt*, ο οποίος περιλαμβάνει όλες τις μεταβλητές που συμμετέχουν στην ανάλυση, και η εφαρμογή της ΠΑΑ σε αυτόν τον υποπίνακα αποδίδει την “πλησιέστερη εικόνα” του υπό εξέταση φαινομένου σε αυτήν την εικόνα που προκύπτει από την εφαρμογή της μεθόδου στον αρχικό πίνακα *Burt*. Να σημειώσουμε ότι το πρόβλημα αυτό τίθεται για πρώτη φορά στο χώρο της Ανάλυσης Δεδομένων. Στηριζόμενοι στη φυσική ερμηνεία της Ενδιαφέρουσας Αδράνειας προτείνουμε μέθοδο διόρθωσης των αδρανειών των παραγοντικών αξόνων, ώστε τα αντίστοιχα ποσοστά ερμηνείας να αντικατοπτρίζουν την πραγματική ποιότητα της λύσης της πολυμεταβλητής ΠΑΑ.

Στην εργασία διαπραγματευόμαστε, επίσης για πρώτη φορά στο χώρο της Ανάλυσης Δεδομένων, το πρόβλημα του καθορισμού του ελάχιστου απαιτούμενου μεγέθους δείγματος σε πειραματικές ή δειγματοληπτικές έρευνες, στα δεδομένα των οποίων θα εφαρμοστεί η ΠΑΑ. Το ζήτημα αντιμετωπίζεται στο Κεφάλαιο 5, όπου προτείνουμε μεθοδολογία για την *a priori* και *post hoc* Ανάλυση Ισχύος του ελέγχου ανεξαρτησίας – ομοιογένειας χ^2 στο πλαίσιο της ΠΑΑ. Με βάση την προτεινόμενη προσέγγιση, αναπτύξαμε το λογισμικό Power Analysis for AFC, το οποίο παρέχει τη δυνατότητα εκτέλεσης των σχετικών υπολογισμών και της αντίστοιχης ανάλυσης ευαισθησίας. Στο ίδιο κεφάλαιο, προτείνουμε μέθοδο καθορισμού του βέλτιστου υποχώρου, στον

οποίο το υπό εξέταση φαινόμενο προβάλλεται χωρίς στατιστικά σημαντική απώλεια πληροφορίας, κατά την εφαρμογή της διμεταβλητής εκδοχής της ΠΑΑ. Ορίζουμε, επίσης, την έννοια της «Δυναμικής Αδράνειας» (*a priori* και *post hoc*) σε απλό πίνακα συμπτώσεων δύο μεταβλητών και παρουσιάζουμε το ρόλο του νέου αυτού δείκτη στον καθορισμό του μεγέθους δείγματος και στον εμπειρικό προσδιορισμό των σημαντικών παραγοντικών αξόνων.

Το φιλοσοφικό πλαίσιο, στο οποίο αναπτύχθηκε αρχικά η ΠΑΑ, δεν επέτρεπε, κυρίως από επιστημολογική σκοπιά, τον έλεγχο της εξωτερικής εγκυρότητας των αριθμητικών και διαγραμματικών εκροών της μεθόδου μέσω στατιστικών ελέγχων υποθέσεων, όπως αυτοί εφαρμόζονται στην Επαγωγική Στατιστική. Στο Κεφάλαιο 6 δείχνουμε ότι η ΠΑΑ μπορεί να συνδυαστεί με ελέγχους σημαντικότητας και μάλιστα με ελάχιστες τεχνικές και θεωρητικές προϋποθέσεις. Για τη διμεταβλητή εκδοχή της ΠΑΑ, προτείνουμε δύο μεθόδους κατασκευής 100(1-α)% ελλείψεων εμπιστοσύνης γύρω από τα σημεία γραμμών (στηλών) με σκοπό τη στατιστική σύγκριση των αντίστοιχων προφίλ και τον έλεγχο της σταθερότητας των διαγραμματικών αποτελεσμάτων. Επίσης, κατασκευάζουμε ένα μη παραμετρικό διάστημα εμπιστοσύνης για τον έλεγχο της σημαντικότητας των παραγοντικών αξόνων, που προκύπτουν στην πολυμεταβλητή περίπτωση.

Στο Κεφάλαιο 7 δείχνουμε ότι η ΠΑΑ, όπως αυτή εφαρμόζεται στο μεθοδολογικό πλαίσιο της Γαλλικής και Ολλανδικής Σχολής Ανάλυσης Δεδομένων, μπορεί να χρησιμοποιηθεί στη στατιστική επεξεργασία κατηγορικών – ποιοτικών μεταβλητών, οι οποίες διακρίνονται σε εξαρτημένες και ανεξάρτητες. Ο διαχωρισμός μπορεί να είναι: α) δομικός, δηλαδή να καθορίζεται από το μηχανισμό παραγωγής των δεδομένων, όπως συμβαίνει στους πειραματικούς σχεδιασμούς, ή β) εννοιολογικός, δηλαδή να υπαγορεύεται θεωρητικά μέσα σε συγκεκριμένο γνωστικό ερευνητικό πεδίο. Και στις δύο Σχολές η ΠΑΑ αντιμετωπίζεται συμμετρικά, χωρίς διάκριση των μεταβλητών σε εξαρτημένες και ανεξάρτητες, γεγονός που έχει ως αποτέλεσμα η μέθοδος να μην έχει αξιοποιηθεί σε πειραματικούς σχεδιασμούς, όπως αυτοί ορίστηκαν στην Ενότητα 1.2. Με δεδομένη την προηγούμενη διαπίστωση, προτείνουμε μεθοδολογία για την ανάλυση τριών βασικών παραγοντικών πειραματικών σχεδιασμών: α) το Πλήρως Τυχαιοποιημένο Σχέδιο με Ένα Παράγοντα, β) το Πλήρως Τυχαιοποιημένο Σχέδιο με Δύο Παράγοντες και γ) το

Τυχαιοποιημένο Σχέδιο σε Πλήρη Συγκροτήματα (*blocks*) με Δύο Παράγοντες. Και στις τρεις περιπτώσεις οι μεταβλητές που εμπλέκονται είναι κατηγορικές. Οι προτεινόμενες μέθοδοι επιτρέπουν τη συμμετοχή στον πειραματισμό σε περισσότερες από μία εξαρτημένες μεταβλητές, ενώ γενικεύονται και στις περίπτωση περισσοτέρων των δύο ανεξάρτητων.

Με την παρούσα εργασία επιχειρούμε να “γεφυρώσουμε” δύο Σχολές Ανάλυσης Δεδομένων, τη Γαλλική με την Ολλανδική, και δύο φιλοσοφικές προσεγγίσεις στη στατιστική συμπερασματολογία, την Επαγωγική Στατιστική με την Ανάλυση Δεδομένων. Και για τις δύο περιπτώσεις δείχνουμε ότι το επιστημολογικό κενό, που χωρίζει τις αντίστοιχες “όχθες”, είναι δυνατό να καλυφθεί και η προσπάθεια αυτή μπορεί να συνεχιστεί με περαιτέρω έρευνα.

ΚΕΦΑΛΑΙΟ 2

Παραγοντική Ανάλυση των Αντιστοιχιών: Η Μέθοδος

2.1 Εισαγωγή

Η Παραγοντική Ανάλυση των Αντιστοιχιών (ΠΑΑ) θεωρείται ως μία περιγραφική μέθοδος για τη διερεύνηση της σχέσης μεταξύ δύο ή περισσότερων κατηγορικών μεταβλητών χωρίς *a priori* υποθέσεις και προϋποθέσεις. Μπορεί να χαρακτηριστεί ως επέκταση της Ανάλυσης σε Κύριες Συνιστώσες (Van der Heijden & De Leeuw 1989, Van de Geer 1993α και 1993β, Gifi 1996), η οποία είναι κατάλληλη για την ανάλυση μόνο ποσοτικών μεταβλητών. Πρωταρχικός σκοπός της ΠΑΑ είναι η ανάδειξη και οπτικοποίηση της ενδογενούς δομής των δεδομένων, η οποία δεν είναι άμεσα αντιληπτή, αλλά βρίσκεται σε λανθάνουσα μορφή, και μάλιστα χωρίς τη χρήση στατιστικών ελέγχων σημαντικότητας για την απόρριψη ή όχι υποθέσεων σχετικά με αυτά. Η μέθοδος χρησιμοποιείται για την ανάλυση ποιοτικών δεδομένων, τα οποία μπορούν να οργανωθούν σε απλούς και σύνθετους πίνακες συνάφειας. Αρχικά, η ΠΑΑ βρήκε εφαρμογές σε προβλήματα από το χώρο της Βιομετρίας και της Ψυχομετρίας καθώς και από το ευρύτερο ερευνητικό πεδίο των Κοινωνικών Επιστημών (βλέπε Greenacre 1984, Benzécri 1992, Blasius & Greenacre 1994, Gifi 1996, Beh 2004, Murtagh 2005). Ιδιαίτερα στη Γαλλική Σχολή οι απαρχές της μεθόδου συνδέονται με την ανάλυση περιεχομένου κειμένων και, γενικότερα, γλωσσικών δεδομένων (βλέπε Benzécri 1992, Remenyi 1992, Giegler & Klein 1994, Sapiro 2002, Κωνσταντινίδης 2002, Αθανασίου & Παπαδημητρίου 2002, Μπεχράκης 2003 και 1999, Askell-Williams & Lawson 2004, Δρόσος 2005, Μασούρα 2005, Murtagh 2005). Σήμερα, η μέθοδος εφαρμόζεται σχεδόν σε όλα τα επιστημονικά πεδία (βλέπε Ενότητα 2.4.3).

Η ΠΑΑ στη διμεταβλητή εκδοχή της εφαρμόζεται στον πίνακα συμπτώσεων απολύτων συχνοτήτων των δύο μεταβλητών. Ωστόσο, είναι αρκετά ευέλικτη, ώστε να μπορεί να εφαρμοστεί σχεδόν σε κάθε πίνακα της μορφής «*αντικείμενα* × *μεταβλητές*» με μη αρνητικά στοιχεία και μη μηδενικά αθροίσματα γραμμών και στηλών, αρκεί οι μεταβλητές να είναι ομοιογενείς (Israëls 1987, Mellinger 1987, Van der Heijden & De Leeuw 1989, Weller & Romney 1990, Benzécri 1992, Higgs 1991, Greenacre & Blasius 1994, Μαυρομάτης 1999, Murtagh 2005, Παπαδημητρίου 2006, 2004 και 1994). Η ομοιογένεια αναφέρεται στην κλίμακα μέτρησης των μεταβλητών (κοινή μονάδα μέτρησης) και στη φυσική σημασία (ερμηνεία) των αθροισμάτων των γραμμών και στηλών του πίνακα εισόδου. Στην περίπτωση πολλών μεταβλητών και στο πλαίσιο της Γαλλικής Σχολής, η μέθοδος εφαρμόζεται είτε σε Πίνακες Σχεδιασμού με λογική κωδικοποίηση 0-1 είτε στους αντίστοιχους Γενικευμένους Πίνακες Συμπτώσεων απολύτων συχνοτήτων (Lebart, Morineau & Warwick 1984, Van der Heijden & De Leeuw 1989, SAS Institute 1990, Escofier & Pagès 1998, Καραπιστόλης 1999, Μπεχράκης 1999, Παπαδημητρίου 2006, 2004 και 1994) ή αλλιώς Πίνακες *Burt* (Burt, 1950). Πρακτικά η μέθοδος μπορεί να εφαρμοστεί σε οποιονδήποτε πίνακα διπλής εισόδου με θετικά στοιχεία (Hoffman & Franke 1986, Van der Heijden & De Leeuw 1989, Higgs 1991, Groenen & Van de Velden 2004).

Στην Ολλανδική Σχολή η ΠΑΑ αντιμετωπίζεται ως πρόβλημα βελτιστοποίησης με δεσμεύσεις και υλοποιείται υπολογιστικά μέσω του επαναληπτικού αλγόριθμου *Alternating Least Squares* (Εναλλασόμενα Ελάχιστα Τετράγωνα) (βλέπε Ενότητες 1.4.2, 2.3.4 και Β3 του Παραρτήματος Β). Σε κάθε περίπτωση, κατά την εφαρμογή της, συνήθως απουσιάζουν οι *a priori* παραδοχές σχετικά με τη θεωρητική κατανομή που ακολουθούν τα δεδομένα και οι παράμετροι του υπό εξέταση πληθυσμού ή πληθυσμών. Όπως αναφέρθηκε στο Κεφάλαιο 1, οι μεταβλητές αντιμετωπίζονται συμμετρικά, χωρίς διάκριση σε εξαρτημένες και ανεξάρτητες, και εν γένει δεν λαμβάνεται υπόψη ο μηχανισμός παραγωγής των διαθέσιμων δεδομένων. Ο μόνος βασικός περιορισμός αφορά στις κλίμακες μέτρησης των μεταβλητών, οι οποίες θα πρέπει να είναι ονομαστικές ή/και διάταξης. Βέβαια, μπορούν να χρησιμοποιηθούν και ποσοτικές μεταβλητές, αφού προηγουμένως οι τιμές τους χωριστούν με βάση λογικά ή στατιστικά κριτήρια σε πεπερασμένο αριθμό κλάσεων.

Κατά τη διερεύνηση της σχέσης δύο ή περισσότερων μεταβλητών, η συνήθης διαδικασία του ελέγχου υποθέσεων και της στατιστικής σημαντικότητας αντικαθίσταται από τη γεωμετρική ερμηνεία των γραφικών αποτελεσμάτων, τα οποία παράγονται από την εφαρμογή της μεθόδου (Greenacre & Blasius 1994, Le Roux & Rouanet 2004). Αυτός ο τρόπος περιγραφής των δεδομένων αποτελεί ένα συγκεκριμένο τρόπο σκέψης, που είναι ενδεικτικός της Γαλλικής Σχολής Ανάλυσης Δεδομένων. Η φιλοσοφία της Σχολής περικλείεται στην Αρχή 2 του Benzécri «*Το μοντέλο πρέπει να προσαρμόζεται στα δεδομένα και όχι το αντίθετο*» (Benzécri & *Collaborateurs*, 1973, σ. 6). Με αυτόν το ριζοσπαστικό, για την εποχή του τρόπο, ο Benzécri θέλησε να τονίσει τη σημαντικότητα της αβίαστης αποκάλυψης της δομής των δεδομένων, σε αντίθεση με τους περιορισμούς και των υποκειμενικά, κατά την άποψή του, καθορισμένων στατιστικών υποδειγμάτων - μοντέλων.

Στο πλαίσιο της Επαγωγικής Στατιστικής η σημαντικότητα της σχέσης δύο κατηγορικών μεταβλητών ελέγχεται συνήθως με την εφαρμογή του ελέγχου χ^2 (Everitt 1979, Agresti 2002 και 1984) εφόσον, βέβαια, ικανοποιούνται ορισμένες θεωρητικές και τεχνικές προϋποθέσεις. Όμως, ο έλεγχος αυτός δεν δίνει επιπλέον πληροφορία για τη φύση της συσχέτισης μεταξύ των δύο μεταβλητών (Παπαδημητρίου, 2004 και 1994). Η εφαρμογή της ΠΑΑ είναι δυνατό να αποκαλύψει, εκτός της ύπαρξης (ή όχι) συνάφειας, και τον τρόπο με τον οποίο συνδέονται οι μεταβλητές και αλληλεπιδρούν οι κλάσεις τους. Στην πολυμεταβλητή περίπτωση, στο πνεύμα της Αγγλο-Σαξωνικής παράδοσης, η σημαντικότητα της συσχέτισης μεταξύ των μεταβλητών ελέγχεται με τη χρήση Λογαριθμογραμμικών Υποδειγμάτων (*Log-Linear Models*) (Knoke & Burke 1980, Bishop, Fienberg & Holland 1991, Andersen 1991, Ishii-Kuntz 1994, Fienberg 2000, Agresti 2002 και 1984). Στην πράξη η συγκεκριμένη μεθοδολογία μπορεί να αποτύχει λόγω των απαιτήσεων και προϋποθέσεών της, οι οποίες αφορούν (βλέπε Knoke & Burke 1980, Koehler 1986, De Leeuw 1988, Ishii-Kuntz 1994, Michailidis 1996, Gifi 1996, Verkuilen 2001, Agresti 2002):

α) Στο μέγεθος του δείγματος, το οποίο θα πρέπει να είναι αρκούντως μεγάλο.

β) Στον τρόπο συλλογής των δεδομένων. Η μέθοδος είναι έγκυρη μόνο στην περίπτωση όπου οι μονάδες του δείγματος έχουν επιλεγεί με μεθόδους της τυχαίας δειγματοληψίας και οι αντίστοιχες μετρήσεις είναι ανεξάρτητες.

γ) Στην πολυδιάστατη κοινή κατανομή των μεταβλητών, όπου η μη εμφάνιση στο δείγμα συνδυασμών κλάσεων των μεταβλητών, δηλαδή το πρόβλημα των "άδειων κελιών", δημιουργεί σοβαρό πρόβλημα στις εκτιμήσεις των παραμέτρων και γενικότερα στην εγκυρότητα των αποτελεσμάτων.

δ) Στο πλήθος των μεταβλητών. Πρακτικά η μέθοδος δεν μπορεί να χειριστεί περισσότερες από τρεις μεταβλητές με μεγάλο αριθμό κατηγοριών η κάθε μια.

Σε όλα τα παραπάνω μπορεί να προστεθεί και η δυσκολία στην ερμηνεία των παραγόμενων δεικτών και εκτιμητών (Ishii-Kuntz 1994, Michailidis 1996). Έτσι, σε ορισμένες περιπτώσεις, η ΠΑΑ αποτελεί τη μοναδική επιλογή που έχει ο ερευνητής για τη στατιστική επεξεργασία των διαθέσιμων στοιχείων του. Ωστόσο, οι δύο προσεγγίσεις έχουν κοινά σημεία και σε αρκετές περιπτώσεις έχει αναδειχθεί η συμπληρωματικότητά τους (Van der Heijden & De Leeuw 1985, Van der Heijden & Worsley 1988, Van der Heijden, De Falguerolles & De Leeuw 1989, Novak & Hoffman 1990, Van der Heijden, De Vries & Van Hooff 1990, Κιοσέογλου & Δικαίου 1993, Van der Heijden, Mooijaart & Takane 1994, Goodman 1996, 1993 και 1991, Clausen 1998, Panagiotakos & Pitsavos 2004, Καρλής 2005). Η ΠΑΑ μπορεί να θεωρηθεί ότι αναλύει τα τυποποιημένα υπόλοιπα του λογαριθμογραμμικού υποδείγματος που αντιστοιχεί στην υπόθεση της ανεξαρτησίας των μεταβλητών ή κάποιας άλλης υπόθεσης (βλέπε De Leeuw & Van der Heijden 1988, Choulakian 1988, De Falguerolles, Jmel & Whittaker 1988, Van der Heijden, De Falguerolles & De Leeuw 1989, Novak & Hoffman 1990, Andersen 1991, Van der Heijden, Mooijaart & Takane 1994). Έτσι, αρχικά μπορεί να εφαρμοστεί η ΠΑΑ ώστε να αναδειχθούν οι επικρατέστερες συσχετίσεις, οι οποίες, στη συνέχεια, είναι δυνατό να ελεγχθούν μέσω λογαριθμογραμμικών υποδειγμάτων. Αντίστροφα, η ΠΑΑ μπορεί να εφαρμοστεί μετά την προσαρμογή των δεδομένων σε κάποιο συσχετιστικό υπόδειγμα με σκοπό την οπτικοποίηση των σημαντικών σχέσεων.

Είδαμε, στο Κεφάλαιο 1, ότι η ΠΑΑ κατέχει ξεχωριστή θέση ανάμεσα στις υπόλοιπες μεθόδους της Ανάλυσης Δεδομένων. Η δυνατότητα γεωμετρικής παρουσιάσης των

δεδομένων είναι ίσως το σημαντικότερο χαρακτηριστικό της. Η μέθοδος επιτρέπει την αποκάλυψη και την οπτικοποίηση συσχετίσεων μεταξύ κατηγορικών μεταβλητών, οι οποίες δεν είναι δυνατό να εντοπιστούν με διαδοχικές συγκρίσεις των μεταβλητών ανά δύο. Επιπλέον, καταλήγει στη γραφική απεικόνιση των δεδομένων με "κομψές" και λιτές γεωμετρικές αναπαραστάσεις (νέφη σημείων, παραγοντικούς άξονες και παραγοντικά επίπεδα). Μέσω αυτών μπορεί να αναδειχθεί η φυσική ερμηνεία των πιθανών αλληλεπιδράσεων, συσχετίσεων, τάσεων, ομοιοτήτων ή αντιπαραθέσεων και κατά συνέπεια να διευκολυνθεί η κατανόηση του υπό εξέταση φαινομένου.

Αν συνθέσουμε τα συμπεράσματα που προκύπτουν από την ανασκόπηση της σχετικής βιβλιογραφίας (βλέπε Lebart, Morineau & Tabard 1977, Weller & Romney 1990, Greenacre 1993α και 1984, Heiser & Meulman 1994, Gifi 1996, Clausen 1998, Lebart, Morineau & Piron 2000, Παπαδημητρίου 2006, 2004 και 1994) διαπιστώνουμε ότι, σε γενικές γραμμές, η αποτελεσματικότητα της ΠΑΑ μπορεί να αναδειχθεί στις παρακάτω περιπτώσεις όταν:

1) Ο πίνακας δεδομένων είναι ικανοποιητικά μεγάλος. Ο λόγος είναι ότι στην αντίθετη περίπτωση δεν είναι δύσκολο να εντοπιστούν, ακόμα και με απλή παρατήρηση των πινακοποιημένων δεδομένων, ενδιαφέρουσες συσχετίσεις μεταξύ των μεταβλητών.

2) Η δομή του πίνακα δεδομένων είναι εκ των προτέρων άγνωστη στον ερευνητή – χρήστη και το ενδιαφέρον εστιάζεται στην ανακάλυψη και διερεύνηση αυτών των άγνωστων πτυχών, που ενθυλακώνει ο αρχικός πίνακας, και όχι στην επιβεβαίωση συγκεκριμένων ερευνητικών υποθέσεων.

3) Οι μεταβλητές του πίνακα με τα αρχικά δεδομένα είναι ομογενοποιημένες, έτσι ώστε να έχει νόημα ο υπολογισμός αποστάσεων (μέτρων ομοιότητας) μεταξύ των γραμμών ή των στηλών του πίνακα εισόδου.

4) Είναι επιθυμητή η ταυτόχρονη απεικόνιση των σημείων γραμμών και στηλών του πίνακα δεδομένων σε μία και μόνο γραφική παράσταση.

5) Οι προϋποθέσεις (τεχνικές και θεωρητικές) άλλων στατιστικών μεθόδων δεν ικανοποιούνται.

6) Δεν είναι επιθυμητό ή δεν έχει αξία να εξεταστεί η προσαρμογή των δεδομένων, μέσω των οποίων περιγράφεται το υπό εξέταση φαινόμενο, σε κάποιο συγκεκριμένο μαθηματικό πρότυπο. Η ΠΑΑ μπορεί να αναδείξει χωρίς περιορισμούς γραμμικές και μη γραμμικές σχέσεις μεταξύ των μεταβλητών, αν όντως αυτές υπάρχουν, σε αντίθεση με άλλες μεθόδους, όπως η Ανάλυση σε Κύριες Συνιστώσες και η Γραμμική Παλινδρόμηση, όπου η γραμμικότητα των σχέσεων μεταξύ των μεταβλητών επιβάλλεται *a priori* στον καθορισμό των αντίστοιχων μαθηματικών υποδειγμάτων.

7) Ιδιαίτερο ενδιαφέρον παρουσιάζει η μελέτη της διάταξης των κλάσεων των μεταβλητών επί των παραγοντικών αξόνων. Η ταξιθέτηση αυτή γίνεται εφικτή με την ανάθεση βέλτιστων βαθμών (*scores*) στα ποιοτικά χαρακτηριστικά που εκφράζουν οι γραμμές και οι στήλες του πίνακα συμπτώσεων. Η βέλτιστη διάταξη επιτυγχάνεται στον πρώτο παραγοντικό άξονα με βάση τις αντίστοιχες συντεταγμένες των προβολών των σημείων γραμμών ή/και στηλών.

Όπως αναφέρθηκε στην Ενότητα 1.4, η ΠΑΑ αναπτύχθηκε παράλληλα και ανεξάρτητα όχι μόνο σε πολλές χώρες αλλά και σε ποικίλα πεδία εφαρμογών. Αυτό είχε ως αποτέλεσμα να δημιουργηθούν διάφορες σχολές και κατευθύνσεις σε σχέση με το αλγοριθμικό και το θεωρητικό υπόβαθρο της μεθόδου. Είδαμε ότι οι Σχολές που ξεχωρίζουν είναι η Γαλλική και η Ολλανδική. Στα πορίσματά τους στηρίζονται οι υπολογιστικοί αλγόριθμοι των σύγχρονων εμπορικών λογισμικών στατιστικής επεξεργασίας δεδομένων που υποστηρίζουν την ΠΑΑ. Όμως, οι δύο προσεγγίσεις διαφέρουν ριζικά τόσο ως προς το υπολογιστικό όσο και ως προς το θεωρητικό τους πλαίσιο. Η "άγνοια" της μεθοδολογικής βάσης, πάνω στο οποίο τα διάφορα λογισμικά στηρίζουν τους υπολογισμούς τους, μπορεί να οδηγήσει σε εσφαλμένη συμπερασματολογία.

Στις ενότητες που ακολουθούν, επιχειρούμε μια συγκριτική παρουσίαση της ΠΑΑ, όπως αυτή θεμελιώνεται και εφαρμόζεται στο πλαίσιο της Γαλλικής και Ολλανδικής Σχολής Ανάλυσης Δεδομένων. Δείχνουμε με ποιο τρόπο οι δύο αυτές προσεγγίσεις συνδέονται και επισημαίνουμε τα σημεία που χρήζουν ιδιαίτερης προσοχής κατά την ερμηνεία των αποτελεσμάτων. Προβάλουμε τις ιδιότητες και τις δυνατότητες της μεθόδου και παραθέτουμε σύντομες αναφορές σχετικά με τις σημαντικότερες εξελίξεις στο ερευνητικό της πλαίσιο. Τέλος, εισάγουμε και προτείνουμε νέα εννοιολογικά και μεθοδολογικά στοιχεία στο πεδίο εφαρμογής της.

2.2 Η Περίπτωση Δύο Μεταβλητών: Βασικές Έννοιες και Ορισμοί

Στην ενότητα αυτή, η παρουσίαση των βασικών εννοιών συνδυάζει τις προσεγγίσεις που ακολουθούν κυρίως οι Van der Heijden και De Leeuw (1989), Weller και Romney (1990), Benzécri (1992), Greenacre (1994α, 1993α και 1984), Blasius και Greenacre (1994), Clausen (1998), Le Roux και Rouanet (2004) και Παπαδημητρίου (2006, 2004 και 1994). Να επισημάνουμε ότι μεγαλύτερη έμφαση δίνεται στη μαθηματική παρουσίαση της μεθόδου. Για μια λεπτομερή παρουσίαση της γεωμετρικής ερμηνείας της ΠΑΑ παραπέμπουμε στους Greenacre και Hastie (1987), Benzécri (1992), Greenacre (1994α, 1993α, 1993β, 1991 και 1984), Le Roux και Rouanet (2004) και Παπαδημητρίου (2006, 2004 και 1994).

Στη διμεταβλητή περίπτωση ως είσοδος στην ανάλυση δίνεται, συνήθως, ένας απλός $k \times l$ πίνακας συμπτώσεων (συνάφειας) απολύτων συχνοτήτων δύο κατηγορικών μεταβλητών X και Y, με k και l κλάσεις (ιδιότητες) αντίστοιχα (βλέπε Πίνακα 2.1). Χωρίς περιορισμό της γενικότητας, θεωρούμε ότι οι κατηγορίες της μεταβλητής X αποτελούν τις γραμμές, ενώ οι κατηγορίες της Y τις στήλες του πίνακα διπλής εισόδου. Συμβολίζουμε με:

F: τον αρχικό $k \times l$ πίνακα συμπτώσεων των απολύτων συχνοτήτων
f_{ij}: την απόλυτη συχνότητα που αντιστοιχεί στο κελί (i,j), $i = 1,\ldots, k$ και $j = 1,\ldots, l$
f_{i+} : την περιθώρια απόλυτη συχνότητα της γραμμής i, όπου:

$$f_{i+} = \sum_j f_{ij}$$

f_{+j} : την περιθώρια απόλυτη συχνότητα της στήλης j, όπου:

$$f_{+j} = \sum_i f_{ij}$$

N: το γενικό άθροισμα (σύνολο) των στοιχείων του πίνακα **F**, όπου:

$$N = \sum_i \sum_j f_{ij} \ .$$

Σε περίπτωση δειγματοληπτικής ή πειραματικής έρευνας το N εκφράζει το μέγεθος του δείγματος.

Πίνακας 2.1: Ο Πίνακας Συμπτώσεων **F** με τις Περιθώριες Κατανομές Απολύτων Συχνοτήτων

Κλάσεις της μεταβλητής X	Κλάσεις της μεταβλητής Y						
	1	2	…	j	…	l	Άθροισμα ή Περιθώρια Κατανομή της X
1	f_{11}	f_{12}	…	f_{1j}	…	f_{1l}	f_{1+}
2	f_{21}	f_{22}	…	f_{2j}	…	f_{2l}	f_{2+}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	f_{i1}	f_{i2}	…	f_{ij}	…	f_{il}	f_{i+}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	f_{k1}	f_{k2}	…	f_{kj}	…	f_{kl}	f_{k+}
Άθροισμα ή Περιθώρια Κατανομή της Y	f_{+1}	f_{+2}	…	f_{+j}	…	f_{+l}	N

2.2.1 Το Γενικό Πρόβλημα

Σύμφωνα με τον Παπαδημητρίου (2006, 2004 και 1994), το γενικό ερώτημα στο οποίο μπορεί να απαντήσει η ΠΑΑ είναι το εξής: Μπορούν να παρουσιαστούν βέλτιστα και εποπτικά τα σημεία που αντιστοιχούν στις γραμμές και στις στήλες του πίνακα **F**, με τέτοιο τρόπο ώστε να λαμβάνονται υπόψη: 1) οι προσεγγίσεις (ομοιότητες) που υπάρχουν μεταξύ των γραμμών *i*, να επιλυθεί, δηλαδή, το λεγόμενο «Πρόβλημα των Γραμμών», 2) οι προσεγγίσεις που υπάρχουν μεταξύ των στηλών *j* («Πρόβλημα των Στηλών») και 3) οι σχέσεις σύνδεσης (αλληλεπίδρασης) μεταξύ των γραμμών και των στηλών; Ειδικότερα, το Πρόβλημα των Γραμμών του πίνακα **F** συνίσταται στην προβολή των αντίστοιχων προφίλ, δηλαδή των κατανομών σχετικών συχνοτήτων των γραμμών, σε ένα πολυδιάστατο, εν γένει, χώρο και στην εύρεση ενός υποχώρου μικρότερης διάστασης από τον αρχικό, στον οποίο τα προφίλ των σημείων γραμμών θα προβάλλονται όσο το δυνατόν πλησιέστερα στα αντίστοιχα προφίλ του αρχικού χώρου. Ανάλογα, ισχύουν για τα προφίλ των στηλών.

2.2.2 Προφίλ Γραμμών και Στηλών Εφοδιασμένα με Μάζα

Σε έναν πίνακα συμπτώσεων απολύτων συχνοτήτων δύο μεταβλητών δεν είναι εφικτή η άμεση σύγκριση των γραμμών ή των στηλών του χωρίς να ληφθούν υπόψη τα αντίστοιχα αθροίσματα των γραμμών και στηλών Παπαδημητρίου (2006, 2004 και 1994). Στο πλαίσιο της ΠΑΑ, οι απόλυτες συχνότητες στα κελιά του πίνακα **F** μετασχηματίζονται σε ποσοστά των αντίστοιχων αθροισμάτων γραμμών και στηλών. Έτσι, ως «προφίλ» της *i* γραμμής ορίζεται το σύνολο που αποτελείται από τα στοιχεία:

$$\left\{ \frac{f_{i1}}{f_{i+}}, \frac{f_{i2}}{f_{i+}}, \ldots, \frac{f_{il}}{f_{i+}} \right\}, \mu\varepsilon\ f_{i+} = \sum_{j} f_{ij} \ .$$

Ανάλογα ορίζεται και το προφίλ της *j* στήλης που είναι το σύνολο με στοιχεία:

$$\left\{ \frac{f_{1j}}{f_{+j}}, \frac{f_{2j}}{f_{+j}}, \ldots, \frac{f_{kj}}{f_{+j}} \right\}, \mu\varepsilon\ f_{+j} = \sum_{i} f_{ij} \ .$$

Από τις παραπάνω σχέσεις παρατηρούμε ότι το προφίλ της *i* (*j*) γραμμής (στήλης) αντιστοιχεί στην κατανομή σχετικών συχνοτήτων της γραμμής (στήλης). Με αυτόν τον τρόπο, είναι δυνατό να κατασκευαστούν δύο διαφορετικοί πίνακες σχετικών συχνοτήτων, ο ένας με τα προφίλ των γραμμών (πίνακας **R**, βλέπε Πίνακα 2.2) και ο άλλος με τα προφίλ των στηλών (πίνακας **C**, βλέπε Πίνακα 2.3).

Το άθροισμα κάθε γραμμής του πίνακα **R** και κάθε στήλης του πίνακα **C** είναι ίσο με τη μονάδα. Στην παράδοση της Γαλλικής Σχολής Ανάλυσης Δεδομένων (Benzècri & *Collaborateurs* 1973, Benzècri 1992, Le Roux & Rouanet 2004) κάθε προφίλ γραμμής (στήλης) είναι εφοδιασμένο με «βάρος» ή, αλλιώς, «μάζα», η οποία είναι ίση με το ποσοστό του αθροίσματος της αντίστοιχης γραμμής (στήλης) ως προς το γενικό άθροισμα *N*.

Υπολογιστικά, η μάζα r_i της γραμμής *i* δίνεται από τη σχέση $r_i = \frac{f_{i+}}{N}$, ενώ η μάζα c_j της στήλης *j* από τη σχέση $c_j = \frac{f_{+j}}{N}$. Δηλαδή, οι μάζες των γραμμών και στηλών

είναι στην ουσία οι σχετικές συχνότητες των αντιστοίχων κλάσεων των δύο μεταβλητών *X* και *Y* (βλέπε Πίνακα 2.4).

Πίνακας 2.2: Ο Πίνακας **R** με τα Προφίλ των Γραμμών (στις γραμμές)

Κλάσεις της μεταβλητής *X*	Κλάσεις της μεταβλητής *Y*						
	1	2	…	*j*	…	*l*	Άθροισμα
1	$\frac{f_{11}}{f_{1+}}$	$\frac{f_{12}}{f_{1+}}$	…	$\frac{f_{1j}}{f_{1+}}$	…	$\frac{f_{1l}}{f_{1+}}$	1
2	$\frac{f_{21}}{f_{2+}}$	$\frac{f_{22}}{f_{2+}}$	…	$\frac{f_{2j}}{f_{2+}}$	…	$\frac{f_{2l}}{f_{2+}}$	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	$\frac{f_{i1}}{f_{i+}}$	$\frac{f_{i2}}{f_{i+}}$	…	$\frac{f_{ij}}{f_{i+}}$	…	$\frac{f_{il}}{f_{i+}}$	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	$\frac{f_{k1}}{f_{k+}}$	$\frac{f_{k2}}{f_{k+}}$	…	$\frac{f_{kj}}{f_{k+}}$	…	$\frac{f_{kl}}{f_{k+}}$	1
Μέσο προφίλ γραμμών ή Κέντρο βάρους γραμμών	$\frac{f_{+1}}{N}$	$\frac{f_{+2}}{N}$	…	$\frac{f_{+j}}{N}$	…	$\frac{f_{+l}}{N}$	

Πίνακας 2.3: Ο Πίνακας **C** με τα Προφίλ των Στηλών (στις στήλες)

Κλάσεις της μεταβλητής *X*	Κλάσεις της μεταβλητής *Y*						Μέσο προφίλ στηλών ή Κέντρο βάρους στηλών
	1	2	…	*j*	…	*l*	
1	$\frac{f_{11}}{f_{+1}}$	$\frac{f_{12}}{f_{+2}}$	…	$\frac{f_{1j}}{f_{+j}}$	…	$\frac{f_{1l}}{f_{+l}}$	$\frac{f_{1+}}{N}$
2	$\frac{f_{21}}{f_{+1}}$	$\frac{f_{22}}{f_{+2}}$	…	$\frac{f_{2j}}{f_{+j}}$	…	$\frac{f_{2l}}{f_{+l}}$	$\frac{f_{2+}}{N}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	$\frac{f_{i1}}{f_{+1}}$	$\frac{f_{i2}}{f_{+2}}$	…	$\frac{f_{ij}}{f_{+j}}$	…	$\frac{f_{il}}{f_{+l}}$	$\frac{f_{i+}}{N}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	$\frac{f_{k1}}{f_{+1}}$	$\frac{f_{k2}}{f_{+2}}$	…	$\frac{f_{kj}}{f_{+j}}$	…	$\frac{f_{kl}}{f_{+l}}$	$\frac{f_{k+}}{N}$
Άθροισμα	1	1	…	1	…	1	

Πίνακας 2.4: Μάζες Γραμμών και Στηλών του Πίνακα **F**

Κλάσεις της μεταβλητής X	Κλάσεις της μεταβλητής Y 1	2	…	j	…	l	Μάζες Γραμμών ή Σχετική Κατανομή της X ή Κέντρο Βάρους Στηλών
1	f_{11}	f_{12}	…	f_{1j}	…	f_{1l}	$\frac{f_{1+}}{N}=r_1$
2	f_{21}	f_{22}	…	f_{2j}	…	f_{2l}	$\frac{f_{2+}}{N}=r_2$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	f_{i1}	f_{i2}	…	f_{ij}	…	f_{il}	$\frac{f_{i+}}{N}=r_i$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	f_{k1}	f_{k2}	…	f_{kj}	…	f_{kl}	$\frac{f_{k+}}{N}=r_k$
Μάζες Στηλών ή Σχετική Κατανομή της Y ή Κέντρο Βάρους Γραμμών	$\frac{f_{+1}}{N}=c_1$	$\frac{f_{+2}}{N}=c_2$	…	$\frac{f_{+j}}{N}=c_j$	…	$\frac{f_{+l}}{N}=c_l$	Άθροισμα=1

Από στατιστική σκοπιά, οι μάζες των γραμμών μπορούν να θεωρηθούν ως οι σταθμισμένοι μέσοι όροι των αντίστοιχων γραμμών του πίνακα που περιέχει τα προφίλ των στηλών (πίνακας **C**), με σταθμίσεις τα αθροίσματα των στηλών του πίνακα **F**. Ομοίως, οι μάζες των στηλών είναι οι σταθμισμένοι μέσοι όροι των αντίστοιχων στηλών του πίνακα με στοιχεία τα προφίλ των γραμμών (πίνακας **R**), με σταθμίσεις τα αθροίσματα των γραμμών του πίνακα **F**. Για παράδειγμα, για τη μάζα της στήλης j έχουμε:

$$c_j=\frac{f_{+j}}{N}=\frac{\sum_i f_{ij}}{N}=\sum_i \frac{f_{ij}}{N}=\sum_i \frac{f_{ij}}{f_{i+}}\frac{f_{i+}}{N}=\frac{1}{N}\sum_i \frac{f_{ij}}{f_{i+}}\, f_{i+}\,.$$

Η απόδοση βάρους στα προφίλ είναι σημαντική διότι εξασφαλίζει ισότιμη συμμετοχή στο αντίστοιχο προφίλ για κάθε μία από τις *N* διαθέσιμες παρατηρήσεις. Η περιθώρια στήλη του πίνακα **C** ονομάζεται «μέσο προφίλ» στηλών ή αλλιώς «κέντρο βάρους» των στηλών. Ανάλογα, η περιθώρια γραμμή του πίνακα **R** ορίζεται ως το μέσο προφίλ των γραμμών ή διαφορετικά το κέντρο βάρους των γραμμών (βλέπε Πίνακες 2.2 και 2.3). Με βάση τους ορισμούς των εννοιών του κέντρου βάρους και της μάζας των σημείων γραμμών και στηλών, είναι φανερό ότι τα στοιχεία του κέντρου βάρους των γραμμών (στηλών) είναι στην ουσία οι μάζες των στηλών (γραμμών) (βλέπε Πίνακα 2.4).

Γεωμετρικά, τα προφίλ, για παράδειγμα, των *k* γραμμών του πίνακα **F**, μπορούν να θεωρηθούν ως διανύσματα, τα οποία ορίζουν ένα «νέφος» *k* σημείων στον πολυδιάστατο ευκλείδειο χώρο $\Re$. Ο χώρος θα έχει το πολύ *l* διαστάσεις ($\Re^l$), δηλαδή τόσες όσα είναι σε πλήθος τα στοιχεία - συντεταγμένες από τα οποία αποτελούνται τα προφίλ. Επειδή όμως τα στοιχεία κάθε προφίλ έχουν άθροισμα ίσο με τη μονάδα, τα προφίλ καταλαμβάνουν μια περιορισμένη περιοχή (υποχώρο) του αρχικού χώρου η οποία έχει *l*-1 διαστάσεις. Η περιοχή αυτή ονομάζεται «*simplex*» και περιέχει το σύνορο και το εσωτερικό του κυρτού πολυγώνου που σχηματίζεται από τα πέρατα των *l* ορθομοναδιαίων διανυσμάτων που αποτελούν τη βάση του αρχικού χώρου. Τα πέρατα αυτά ονομάζονται «κορυφές» και ορίζουν ένα «κεντροβαρικό» σύστημα συντεταγμένων ή αλλιώς τον «χώρο των προφίλ». Συνήθως, σε πρακτικές εφαρμογές, οι διαστάσεις του υποχώρου που ανήκουν τα προφίλ είναι περισσότερες από τρεις, με συνέπεια να μην είναι δυνατή η γραφική απεικόνιση και οπτικοποίησή τους. Όμως, μέσω της ΠΑΑ, οι διαστάσεις μπορούν να μειωθούν σημαντικά και τα προφίλ να απεικονιστούν “βέλτιστα” σε ένα χώρο με λιγότερες διαστάσεις από τον αρχικό. Ανάλογα ισχύουν για τα προφίλ των στηλών, τα οποία ορίζουν ένα «νέφος» *l* σημείων στον πολυδιάστατο χώρο $\Re^k$. Κατά τη γεωμετρική απεικόνιση των προφίλ των σημείων γραμμών (στηλών) του πίνακα **F**, μπορούμε να θεωρήσουμε, ότι τα προφίλ αυτά βρίσκονται γύρω από το μέσο προφίλ τους, δηλαδή το κέντρο βάρους τους, το οποίο έλκεται περισσότερο από ή βρίσκεται πλησιέστερα στα σημεία – προφίλ με τη μεγαλύτερη μάζα.

2.2.3 Ο Πίνακας των Αντιστοιχιών

Η ΠΑΑ συνήθως εφαρμόζεται στον $k \times l$ πίνακα **P** που προκύπτει από τη διαίρεση των στοιχείων του πίνακα **F** με το γενικό σύνολο N, δηλαδή ο **P** έχει στοιχεία τα:

$$p_{ij} = \frac{f_{ij}}{N}.$$

Είναι φανερό ότι: $\sum_i \sum_j p_{ij} = 1$.

Ο **P** ονομάζεται «Πίνακας των Αντιστοιχιών» και εκφράζει την κατανομή μιας ποσότητας μάζας ίση με τη μονάδα στα κελιά του αρχικού πίνακα **F**. Τα αθροίσματα των γραμμών r_i και στηλών c_j του πίνακα **P**, δηλαδή τα στοιχεία των περιθώριων κατανομών του, είναι ίσα με τις αντίστοιχες μάζες των γραμμών και στηλών. Πιο συγκεκριμένα, ισχύουν οι παρακάτω σχέσεις:

$$r_i = \sum_j p_{ij} = \sum_j \frac{f_{ij}}{N} = \frac{f_{i+}}{N}$$

και

$$c_j = \sum_i p_{ij} = \sum_i \frac{f_{ij}}{N} = \frac{f_{+j}}{N}.$$

Με βάση τις προηγούμενες σχέσεις, το προφίλ της i γραμμής (βλέπε Ενότητα 2.2.2) μπορεί να δοθεί και ως το σύνολο:

$$\left\{ \frac{p_{i1}}{r_i}, \frac{p_{i2}}{r_i}, \ldots, \frac{p_{il}}{r_i} \right\}.$$

Ανάλογα, το προφίλ της j στήλης είναι το σύνολο:

$$\left\{ \frac{p_{1j}}{c_j}, \frac{p_{2j}}{c_j}, \ldots, \frac{p_{kj}}{c_j} \right\}.$$

Αν οι X και Y θεωρηθούν ως τυχαίες μεταβλητές και το μέγεθος του δείγματος N είναι αρκούντως μεγάλο, τότε ο πίνακας **P** αποτελεί εκτίμηση της κοινής κατανομής πιθανότητας των δύο κατηγορικών μεταβλητών και τα προφίλ γραμμών και στηλών

αποτελούν εκτιμήσεις των υπό συνθήκη (δεσμευμένων) κατανομών πιθανοτήτων της Y και X αντίστοιχα.

2.2.4 Οι Αποστάσεις

Η δυνατότητα απεικόνισης των προφίλ γραμμών ή στηλών ως σημείων σε έναν πολυδιάστατο χώρο καθιστά απαραίτητο τον καθορισμό ενός μέτρου της απόστασης μεταξύ των σημείων αυτών. Στην ΠΑΑ χρησιμοποιείται το τετράγωνο της «κατά Benzécri» χ^2 απόστασης (Benzécri & *Collaborateurs* 1973, Benzécri 1992). Ειδικότερα, το τετράγωνο της απόστασης μεταξύ δύο σημείων γραμμών i και i' δίνεται από τη σχέση:

$$d^2_{\chi^2}(i,i') = \sum_{j=1}^{l} \frac{N}{f_{+j}} \left(\frac{f_{ij}}{f_{i+}} - \frac{f_{i'j}}{f_{i'+}} \right)^2 = \sum_{j=1}^{l} \frac{1}{c_j} \left(\frac{f_{ij}}{f_{i+}} - \frac{f_{i'j}}{f_{i'+}} \right)^2 = \sum_{j=1}^{l} \frac{1}{c_j} \left(\frac{p_{ij}}{r_i} - \frac{p_{i'j}}{r_{i'}} \right)^2$$

και το τετράγωνο της απόστασης μεταξύ δύο σημείων στηλών j και j' από τη σχέση:

$$d^2_{\chi^2}(j,j') = \sum_{i=1}^{k} \frac{N}{f_{i+}} \left(\frac{f_{ij}}{f_{+j}} - \frac{f_{ij'}}{f_{+j'}} \right)^2 = \sum_{i=1}^{k} \frac{1}{r_i} \left(\frac{f_{ij}}{f_{+j}} - \frac{f_{ij'}}{f_{+j'}} \right)^2 = \sum_{i=1}^{k} \frac{1}{r_i} \left(\frac{p_{ij}}{c_j} - \frac{p_{ij'}}{c_{j'}} \right)^2 .$$

Παρατηρούμε, ότι κάθε όρος του αθροίσματος τετραγώνων των διαφορών πολλαπλασιάζεται επί ένα συντελεστή στάθμισης. Πιο συγκεκριμένα, στον υπολογισμό της απόστασης μεταξύ δύο γραμμών τα τετράγωνα των διαφορών σταθμίζονται με τους αντίστροφους των μαζών των στηλών. Ανάλογα, για τον υπολογισμό της απόστασης μεταξύ δύο στηλών, τα τετράγωνα των διαφορών σταθμίζονται με τους αντίστροφους των μαζών των γραμμών. Οι σταθμίσεις αυτές έχουν ως αποτέλεσμα η απόσταση χ^2 να παρουσιάζει, έναντι της συνήθους Ευκλείδειας απόστασης, τα παρακάτω πλεονεκτήματα (Benzécri 1992, Greenacre 1993α και 1984, Καραπιστόλης 1999, Μπεχράκης 1999, Παπαδημητρίου 2006, 2004 και 1994):

1) Στον υπολογισμό της απόστασης, οι διαφορές ανάμεσα σε σημεία με χαμηλή συχνότητα (μικρή μάζα) ενισχύονται και στη συνέχεια αναδεικνύονται στη διαγραμματική αναπαράσταση, σε σχέση με τις διαφορές ανάμεσα στα σημεία με υψηλή συχνότητα (μεγάλη μάζα).

2) Η χ^2 απόσταση ικανοποιεί την «Αρχή της Ισοδυναμίας των Κατανομών» (*Principle of Distributional Equivalence*). Σύμφωνα με την αρχή αυτή, αν δύο γραμμές (στήλες) του αρχικού πίνακα **F**, οι οποίες είναι μεταξύ τους ανάλογες, αντικατασταθούν ή καλύτερα συγχωνευτούν σε μία γραμμή, που είναι το άθροισμά τους, τότε οι αποστάσεις μεταξύ των σημείων του νέφους των στηλών (γραμμών) δεν αλλοιώνονται.

3) Από την επιλογή της συγκεκριμένης απόστασης απορρέουν οι σημαντικότερες ιδιότητες της ΠΑΑ, οι οποίες σχετίζονται με τη διάσπαση της χ^2 απόστασης και τη δυνατότητα ταυτόχρονης απεικόνισης των σημείων γραμμών και στηλών σε ένα κοινό διάγραμμα.

2.2.5 Η Αδράνεια

Στη Γαλλική Σχολή Ανάλυσης Δεδομένων και ιδιαίτερα στο πλαίσιο της ΠΑΑ τα σημεία που αντιστοιχούν στα προφίλ των γραμμών (στηλών) και αποτελούν το νέφος των γραμμών (στηλών) θεωρούνται ως υλικά σημεία εφοδιασμένα με μάζα (Benzécri 1992, Greenacre 1993α και 1984, Καραπιστόλης 1999, Μπεχράκης 1999, Le Roux & Rouanet 2004, Παπαδημητρίου 2006, 2004 και 1994). Η γεωμετρική αναπαράστασή τους πραγματοποιείται μέσω της προβολής τους σε ένα κεντροβαρικό σύστημα συντεταγμένων όπου στην αρχή του προβάλλεται το κέντρο βάρους (μέσο προφίλ) του νέφους των γραμμών (στηλών) του πίνακα **F**. Ο όρος «Αδράνεια» (*Inertia*) προέρχεται από τη Μηχανική όπως και άλλες έννοιες της Στατιστικής (π.χ. οι βαθμοί ελευθερίας και οι ροπές). Είναι γνωστό ότι κάθε φυσικό αντικείμενο έχει ένα κέντρο βάρους. Αν θεωρήσουμε ότι κάθε τμήμα του αντικειμένου έχει μάζα m και απέχει απόσταση d από το κέντρο βάρους του, τότε η *αδράνεια* του αντικειμένου είναι ίση με το άθροισμα των ποσοτήτων md^2 για κάθε τμήμα του (Benzécri & *Collaborateurs* 1973, Benzécri 1992). Επομένως,

$$Αδράνεια = \sum md^2 .$$

Όπως είδαμε στα προηγούμενα, τα σημεία - προφίλ των γραμμών (στηλών) είναι εφοδιασμένα με μάζες, έχουν οριστεί μεταξύ τους αποστάσεις (χ^2) και ένα κέντρο βάρους (μέσο προφίλ). Συνεπώς, κάθε σημείο γραμμής ή στήλης συνεισφέρει στην ολική αδράνεια του αντίστοιχου νέφους σημείων, στο οποίο ανήκει, ανάλογα με τη

μάζα του και την απόστασή του από το κέντρο βάρους του. Ειδικότερα, αν με g_r και g_c συμβολίσουμε το μέσο προφίλ γραμμών και στηλών αντίστοιχα, τότε τα τετράγωνα των αποστάσεων της i γραμμής και j στήλης από τα αντίστοιχα κέντρα βάρους τους δίνονται από τις σχέσεις:

$$d^2_{\chi^2}(i, g_r) = \sum_{j=1}^{l} \frac{N}{f_{+j}} \left(\frac{f_{ij}}{f_{i+}} - \frac{f_{+j}}{N} \right)^2 = \sum_{j=1}^{l} \frac{1}{c_j} \left(\frac{f_{ij}}{f_{i+}} - c_j \right)^2,$$

και

$$d^2_{\chi^2}(j, g_c) = \sum_{i=1}^{k} \frac{N}{f_{i+}} \left(\frac{f_{ij}}{f_{+j}} - \frac{f_{i+}}{N} \right)^2 = \sum_{i=1}^{k} \frac{1}{r_i} \left(\frac{f_{ij}}{f_{+j}} - r_i \right)^2.$$

Σύμφωνα με τον ορισμό της Αδράνειας στο πλαίσιο της Μηχανικής, η αδράνεια I_r του νέφους των σημείων γραμμών θα δίνεται από τη σχέση:

$$I_r = \sum_i (\mu\acute{\alpha}\zeta\alpha \ i \ \gamma\rho\alpha\mu\mu\acute{\eta}\varsigma) \times d^2_{\chi^2}(i, g_r) = \sum_{i=1}^{k} r_i \sum_{j=1}^{l} \frac{1}{c_j} \left(\frac{f_{ij}}{f_{i+}} - c_j \right)^2 = \sum_{i=1}^{k} r_i \sum_{j=1}^{l} \frac{1}{c_j} \left(\frac{p_{ij}}{r_i} - c_j \right)^2,$$

ενώ η αδράνεια I_c του νέφους των σημείων στηλών από τη σχέση:

$$I_C = \sum_j (\mu\acute{\alpha}\zeta\alpha \ j \ \sigma\tau\acute{\eta}\lambda\eta\varsigma) \times d^2_{\chi^2}(j, g_C) = \sum_{j=1}^{l} c_j \sum_{i=1}^{k} \frac{1}{r_i} \left(\frac{f_{ij}}{f_{+j}} - r_i \right)^2 = \sum_{j=1}^{l} c_j \sum_{i=1}^{k} \frac{1}{r_i} \left(\frac{p_{ij}}{c_j} - r_i \right)^2.$$

Γεωμετρικά, η αδράνεια μπορεί να θεωρηθεί ως ένα μέτρο της διασποράς των σημείων – προφίλ στον πολυδιάστατο χώρο στον οποίο ανήκουν. Όσο μεγαλύτερη είναι η αδράνεια, τόσο μεγαλύτερη είναι και η διασπορά των σημείων στο χώρο. Από στατιστική σκοπιά, η ολική αδράνεια του νέφους των σημείων γραμμών ή στηλών μπορεί να οριστεί ως μια γενικευμένη διασπορά και, πιο συγκεκριμένα, ως ο σταθμισμένος μέσος όρος των τετραγώνων των χ^2 αποστάσεων των προφίλ γραμμών, ή ισοδύναμα των προφίλ στηλών, από το κέντρο βάρους τους. Ειδικότερα, αν συμβολίσουμε με:

r και **c** τα διανύσματα με στοιχεία τα αθροίσματα γραμμών και στηλών αντίστοιχα του πίνακα **P**, δηλαδή τα διανύσματα με στοιχεία τις μάζες των γραμμών και στηλών αντίστοιχα,

$\mathbf{D_r}$ και $\mathbf{D_c}$ τους διαγώνιους πίνακες που έχουν ως στοιχεία τα αθροίσματα γραμμών και στηλών του **P** αντίστοιχα, δηλαδή $\mathbf{D_r}=diag(\mathbf{r})$ και $\mathbf{D_c}=diag(\mathbf{c})$,

$\tilde{\mathbf{r}}_\mathbf{i}$ το διάνυσμα με το προφίλ της *i* γραμμής και με $\tilde{\mathbf{c}}_\mathbf{j}$ το διάνυσμα με το προφίλ της *j* στήλης (*i*=1,…, *k* και *j*=1,…, *l*) και

1 το διάνυσμα (κατάλληλων ανά περίπτωση διαστάσεων) με στοιχεία $[1, 1, \ldots, 1]^\mathbf{T}$,

τότε τα διανύσματα με στοιχεία τις μάζες γραμμών και στηλών μπορούν να δοθούν από τις σχέσεις:

$$\mathbf{r}=\mathbf{P1} \text{ και } \mathbf{c}=\mathbf{P}^\mathbf{T}\mathbf{1},$$

ενώ οι πίνακες **R** και **C** μπορούν να γραφούν και ως εξής:

$$\mathbf{R} = \mathbf{D_r^{-1}P} = \begin{bmatrix} \tilde{\mathbf{r}}_\mathbf{1}^\mathbf{T} \\ \tilde{\mathbf{r}}_\mathbf{2}^\mathbf{T} \\ \vdots \\ \tilde{\mathbf{r}}_\mathbf{k}^\mathbf{T} \end{bmatrix} \text{ και } \mathbf{C} = \mathbf{D_c^{-1}P^T} = \begin{bmatrix} \tilde{\mathbf{c}}_\mathbf{1}^\mathbf{T} \\ \tilde{\mathbf{c}}_\mathbf{2}^\mathbf{T} \\ \vdots \\ \tilde{\mathbf{c}}_\mathbf{l}^\mathbf{T} \end{bmatrix}.$$

Επίσης, τα μέσα προφίλ (κέντρα βάρους) των προφίλ γραμμών και στηλών δίνονται αντίστοιχα από τις σχέσεις (Greenacre, 1984):

$$\mathbf{c}=\mathbf{R}^\mathbf{T}\mathbf{r} \text{ και } \mathbf{r}=\mathbf{C}^\mathbf{T}\mathbf{c}.$$

Με βάση τους παραπάνω συμβολισμούς, η ολική αδράνεια του νέφους των σημείων γραμμών μπορεί να υπολογιστεί από τη σχέση:

$$I_r = \sum_i r_i \left(\tilde{\mathbf{r}}_\mathbf{i} - \mathbf{c}\right)^\mathbf{T} \mathbf{D_c^{-1}} \left(\tilde{\mathbf{r}}_\mathbf{i} - \mathbf{c}\right) = \sum_i r_i \sum_j \frac{\left(\frac{p_{ij}}{r_i} - c_j\right)^2}{c_j}, \qquad [2.1]$$

ενώ η ολική αδράνεια του νέφους των σημείων στηλών από τη σχέση:

$$I_c = \sum_j c_j \left(\tilde{\mathbf{c}}_\mathbf{j} - \mathbf{r}\right)^\mathbf{T} \mathbf{D_r^{-1}} \left(\tilde{\mathbf{c}}_\mathbf{j} - \mathbf{r}\right) = \sum_j c_j \sum_i \frac{\left(\frac{p_{ij}}{c_j} - r_i\right)^2}{r_i}. \qquad [2.2]$$

Παρατηρούμε ότι οι δύο προηγούμενες σχέσεις α) ικανοποιούν τον ορισμό της αδράνειας, ως άθροισμα γινομένων μαζών επί αποστάσεων από το κέντρο βάρους εις το τετράγωνο (*Αδράνεια* = $\sum md^2$) και β) επιτρέπουν να θεωρήσουμε τη χ^2 απόσταση μεταξύ των προφίλ των γραμμών (στηλών) ως μια σταθμισμένη Ευκλείδεια απόσταση με βάρη τα στοιχεία του πίνακα $\mathbf{D_c^{-1}}$ ($\mathbf{D_r^{-1}}$).

Κάτω από μια άλλη θεώρηση, η αδράνεια του νέφους των σημείων γραμμών (στηλών) είναι ίση με το σταθμισμένο άθροισμα των τετραγώνων των χ^2 αποστάσεων μεταξύ όλων των διαφορετικών $k(k-1)/2$ $\left(l(l-1)/2\right)$ ζευγαριών των προφίλ γραμμών (στηλών). Ειδικότερα, ισχύει (Greenacre, 2005 και 1994α):

$$I_r = \sum_i \sum_{i'<i} r_i r_{i'} \sum_j \frac{\left(\frac{p_{ij}}{r_i} - \frac{p_{i'j}}{r_{i'}} \right)^2}{c_j}$$

και

$$I_c = \sum_j \sum_{j'<j} c_j c_{j'} \sum_i \frac{\left(\frac{p_{ij}}{c_j} - \frac{p_{ij'}}{c_{j'}} \right)^2}{r_i}.$$

Η αδράνεια του νέφους των σημείων γραμμών είναι ίση με την αδράνεια του νέφους των σημείων στηλών. Πράγματι:

Από τη [2.1] συνεπάγεται

$$I_r = \sum_i r_i \sum_j \frac{\left(\frac{p_{ij}}{r_i} - c_j \right)^2}{c_j} = \sum_i \sum_j \frac{\left(p_{ij} - r_i c_j \right)^2}{r_i c_j}.$$

Από τη [2.2] προκύπτει

$$I_c = \sum_j c_j \sum_i \frac{\left(\frac{p_{ij}}{c_j} - r_i \right)^2}{r_i} = \sum_i \sum_j \frac{\left(p_{ij} - r_i c_j \right)^2}{r_i c_j}.$$

Συνεπώς, I_r=I_c. □[2]

Έτσι, στην περίπτωση του πίνακα **F**, μπορούμε να μιλάμε πλέον για «Ολική Αδράνεια» του **F**, η οποία υπολογιστικά δίνεται από τη σχέση:

$$Ολική\ Αδράνεια\ I = I_r = I_c = \sum_i \sum_j \frac{\left(p_{ij} - r_i c_j\right)^2}{r_i c_j} = \sum_i \sum_j \frac{\left(\frac{f_{ij}}{N} - \frac{f_{i+}}{N}\frac{f_{+j}}{N}\right)^2}{\frac{f_{i+}}{N}\frac{f_{+j}}{N}}. \qquad [2.3]$$

Αν θεωρήσουμε ότι ο πίνακας **F** παρουσιάζει την εμπειρική κοινή κατανομή δύο τυχαίων κατηγορικών μεταβλητών, τότε από τη [2.3] έχουμε:

$$I = \sum_i \sum_j \frac{\left(\frac{f_{ij}}{N} - \frac{f_{i+}}{N}\frac{f_{+j}}{N}\right)^2}{\frac{f_{i+}}{N}\frac{f_{+j}}{N}} = \sum_i \sum_j \frac{\left(\frac{1}{N}\left(f_{ij} - \frac{f_{i+}f_{+j}}{N}\right)\right)^2}{\frac{f_{i+}}{N}\frac{f_{+j}}{N}} = \frac{1}{N}\sum_i \sum_j \frac{\left(f_{ij} - \frac{f_{i+}f_{+j}}{N}\right)^2}{\frac{f_{i+}f_{+j}}{N}}.$$

Παρατηρούμε ότι το άθροισμα

$$\sum_i \sum_j \frac{\left(f_{ij} - \frac{f_{i+}f_{+j}}{N}\right)^2}{\frac{f_{i+}f_{+j}}{N}}$$

είναι της μορφής:

$$\sum_i \sum_j \frac{\left(παρατηρούμενη\ \ συχνότητα\ \ του\ \ κελιού\ \ (i,j) - αναμενόμενη\ \ συχνότητα\ \ του\ \ κελιού\ \ (i,j)\right)^2}{αναμενόμενη\ \ συχνότητα\ \ του\ \ κελιού\ \ (i,j)}$$

Επομένως, υπολογιστικά η ολική αδράνεια του πίνακα **F** συνδέεται με το γνωστό έλεγχο ανεξαρτησίας χ^2 *του Pearson* (Everitt 1979, Agresti 2002) και πιο συγκεκριμένα ισχύει:

$$I = \frac{Q}{N} = \varphi^2, \qquad [2.4]$$

---

[2] Στο εξής, το σύμβολο □ θα χρησιμοποιείται για να δηλώσει το τέλος μιας απόδειξης ή ενός παραδείγματος.

όπου Q είναι η τιμή του στατιστικού χ^2 που υπολογίζεται κατά τον έλεγχο ανεξαρτησίας (η ομοιογένειας) των δύο μεταβλητών και φ^2 ο συντελεστής συνάφειας μέσου τετραγώνου (*mean square contingency coefficient*) του *Pearson* (Bishop, Fienberg & Holland 1991, Clausen 1998).

Αν συμβολίσουμε με s_{ij} την ποσότητα $\frac{\left(p_{ij} - r_i c_j\right)}{\sqrt{r_i c_j}}$, τότε από τις σχέσεις [2.3] και [2.4] έχουμε ότι:

$$I = \frac{Q}{N} = \varphi^2 = \sum_i \sum_j s_{ij}^2 \,. \qquad [2.5]$$

Έστω, τώρα, **S** ο $k \times l$ πίνακας με στοιχεία τις ποσότητες s_{ij}. Ο **S** ονομάζεται πίνακας των «Σχετικών Τυποποιημένων Υπολοίπων». Κάθε στοιχείο του είναι η διαφορά της παρατηρούμενης σχετικής συχνότητας, που αντιστοιχεί στο κελί (i, j) του πίνακα **F**, με την αντίστοιχη αναμενόμενη (θεωρητική) σχετική συχνότητα κάτω από την ισχύ της μηδενικής υπόθεσης της ανεξαρτησίας των δύο μεταβλητών, διαιρεμένη με την τετραγωνική ρίζα της θεωρητικής αυτής συχνότητας. Από τη [2.5] είναι φανερό ότι ο πίνακας **S** σχετίζεται άμεσα με την ολική αδράνεια του **F** και γενικότερα με τη διασπορά των στοιχείων του. Αν τα στοιχεία του **S** έχουν μικρή απόλυτη τιμή, τότε η υπόθεση της ανεξαρτησίας είναι μάλλον ισχυρή και, επομένως, η αδράνεια του πίνακα αναμένεται να είναι μικρή. Στην περίπτωση της πλήρους ανεξαρτησίας των δύο μεταβλητών τα στοιχεία s_{ij} είναι ίσα με μηδέν και συνεπώς Q=I=0. Τα στοιχεία του πίνακα **S** πολλαπλασιασμένα επί την τετραγωνική ρίζα του N ονομάζονται «Τυποποιημένα Υπόλοιπα» και κάτω από την ισχύ της υπόθεσης της ανεξαρτησίας των δύο μεταβλητών έχουν μέση τιμή 0, διακύμανση μικρότερη ή ίση από τη μονάδα (Haberman 1973, Reynolds 1984, Agresti 1984) και ακολουθούν ασυμπτωτικά την Τυποποιημένη Κανονική Κατανομή (Reynolds, 1984). Κελιά του πίνακα **F** με τυποποιημένα υπόλοιπα σε απόλυτη τιμή μεγαλύτερη του 1,96≈2 συνεισφέρουν στατιστικά σημαντικά, σε επίπεδο σημαντικότητας α=0,05, στη σημαντικότητα του στατιστικού Q και σε αυτά τα κελιά οφείλεται, κυρίως, η συνάφεια ή η αλληλεπίδραση των δύο μεταβλητών (Hinkle, Wiersma & Jurs, 1988). Το πρόσημο των τυποποιημένων καταλοίπων έχει την ακόλουθη φυσική ερμηνεία:

- Αν σε κάποιο κελί το αντίστοιχο τυποποιημένο υπόλοιπο είναι σε απόλυτη τιμή μεγαλύτερο του 2 και έχει αρνητικό πρόσημο, αυτό σημαίνει ότι στο συγκεκριμένο κελί υπάρχουν στατιστικά σημαντικά λιγότερες παρατηρήσεις (f_{ij}) σε σύγκριση με αυτές που αναμένονται κάτω από την υπόθεση της ανεξαρτησίας των δύο μεταβλητών.
- Αν σε κάποιο κελί το αντίστοιχο τυποποιημένο υπόλοιπο είναι σε απόλυτη τιμή μεγαλύτερο του 2 και έχει θετικό πρόσημο, τότε στο συγκεκριμένο κελί υπάρχουν στατιστικά σημαντικά περισσότερες παρατηρήσεις (f_{ij}) σε σχέση με το αν οι δύο μεταβλητές ήταν ανεξάρτητες.

Αν κάθε στοιχείο του πίνακα **S** διαιρεθεί με την ολική αδράνεια *I* του πίνακα **F**, τότε ο νέος πίνακας που προκύπτει εκφράζει τη συνεισφορά κάθε κελιού ως ποσοστό της ολικής αδράνειας (Greenacre 1993α, Καραπιστόλης 1999).

Όπως θα διαπιστωθεί στα επόμενα, ο αλγόριθμος της ΠΑΑ στηρίζεται στην ανάλυση της βασικής δομής του πίνακα **S** με τα σχετικά τυποποιημένα υπόλοιπα.

2.2.6 Μείωση των Διαστάσεων

Λόγω της πολυδιάστατης φύσης των δεδομένων που διαχειρίζεται, η ΠΑΑ μπορεί να θεωρηθεί ως μια Πολυμεταβλητή Στατιστική Μέθοδος μείωσης των διαστάσεων του αρχικού χώρου, στον οποίο περιγράφεται το υπό εξέταση φαινόμενο (*Dimensionality Reduction Method*). Στην περίπτωση πινάκων με τρεις γραμμές (στήλες), τα αντίστοιχα προφίλ μπορούν εύκολα να απεικονιστούν σε ένα χώρο δύο διαστάσεων, δηλαδή στο επίπεδο. Όμως, σε πρακτικές εφαρμογές, οι πίνακες συμπτώσεων διαθέτουν περισσότερες από τρεις γραμμές ή στήλες, με συνέπεια η απεικόνιση των προφίλ να πρέπει να γίνει σε ένα χώρο πολλών διαστάσεων. Η αδυναμία αντίληψης και δημιουργίας νοερών εικόνων σε χώρους περισσότερων των τριών διαστάσεων, καθιστά απαραίτητη τη μείωση των διαστάσεων των δεδομένων με την ελάχιστη δυνατή απώλεια πληροφορίας.

Η βασική ιδέα είναι να αναπαρασταθούν, ή καλύτερα να προβληθούν, ταυτόχρονα τα νέφη των σημείων γραμμών και στηλών σε ένα χώρο συνήθως δύο ή τριών διαστάσεων, έτσι ώστε η νέα βέλτιστη απεικόνιση να διατηρεί όσο το δυνατό

περισσότερη από την αρχική πληροφορία (Benzécri, 1992). Λόγω του γεγονότος ότι οι θέσεις των προφίλ, σ' ένα νέο χώρο με λιγότερες διαστάσεις από τον αρχικό, απέχουν από τις πραγματικές, πρέπει να χρησιμοποιηθεί και ένα μέτρο απώλειας της πληροφορίας. Στο πλαίσιο της ΠΑΑ, χρησιμοποιείται η ολική αδράνεια ως μέτρο της πραγματικής διασποράς των σημείων γραμμών ή στηλών. Επομένως, η απώλεια πληροφορίας κατά την απεικόνιση των σημείων, για παράδειγμα, σε έναν ή δύο άξονες, μπορεί να εκφραστεί ως ποσοστό της ολικής αδράνειας. Παρόλο που με τη μείωση των διαστάσεων των δεδομένων χάνεται ένα μέρος της αρχικής πληροφορίας, ωστόσο επιτυγχάνεται η απεικόνιση των προφίλ σε ένα χώρο με λιγότερες διαστάσεις και μάλιστα με βέλτιστο τρόπο, που διαφορετικά δεν θα ήταν εφικτή. Πιο συγκεκριμένα, στόχος της ΠΑΑ είναι ο εντοπισμός ενός υποχώρου του αρχικού πολυδιάστατου χώρου, που να είναι όσο το δυνατόν πλησιέστερος στον αρχικό (Greenacre, 1984). Έστω *S* κάθε τέτοιος υποψήφιος υποχώρος. Για κάθε σημείο-προφίλ, έστω της γραμμής *i*, με μάζα r_i, υπολογίζεται η χ^2 απόσταση μεταξύ του σημείου και του υποχώρου *S*, την οποία μπορούμε να συμβολίσουμε, εν γένει, με $d_i(S)$. Η σταθμισμένη απόσταση μεταξύ του σημείου και του υποχώρου είναι ίση με $r_i\left[d_i(S)\right]^2$. Το άθροισμα των αποστάσεων όλων των σημείων γραμμών από τον υποχώρο *S* είναι $\sum_i r_i\left[d_i(S)\right]^2$. Αντικειμενικός σκοπός της ΠΑΑ είναι ο εντοπισμός του υποχώρου *S* (βέλτιστου υποχώρου) που ελαχιστοποιεί το παραπάνω κριτήριο. Το κριτήριο συνεπάγεται ότι η αδράνεια του νέφους των σημείων γραμμών στον βέλτιστο υποχώρο θα είναι η μέγιστη δυνατή (Greenacre, 1993α). Ειδικότερα, αν θεωρήσουμε στο επίπεδο ένα νέφος υλικών σημείων, το οποίο προβάλλεται σε ευθεία *ε*, η οποία διέρχεται από το κέντρο βάρους του, τότε η ολική αδράνεια του νέφους μπορεί να αναλυθεί σε δύο μέρη (Benzécri & *Collaborateurs* 1973, Benzécri 1992, Μπεχράκης 1999, Καρλής 2005, Παπαδημητρίου 2006, 2004 και 1994): α) στην αδράνεια κατά μήκος της ευθείας *ε*, η οποία ονομάζεται «ερμηνεύσιμη αδράνεια», και β) στην αδράνεια καθέτως στην *ε*, που καλείται «εγκάρσια αδράνεια». Η εγκάρσια αδράνεια εκφράζει τις σταθμισμένες αποστάσεις των σημείων από την ευθεία *ε* και αντιστοιχεί στην απώλεια αδράνειας λόγω της προβολής. Η βέλτιστη ευθεία *ε* είναι αυτή, στην οποία η κατά μήκος αδράνεια μεγιστοποιείται και ελαχιστοποιείται η εγκάρσια. Αν γενικεύσουμε την προηγούμενη διαπίστωση σε περισσότερες διαστάσεις προκύπτει το συμπέρασμα ότι ο βέλτιστος υποχώρος *S* είναι αυτός, στον

οποίο η εγκάρσια αδράνεια γίνεται ελάχιστη, που σημαίνει ότι έχουμε τη μικρότερη απώλεια αδράνειας, ή, ισοδύναμα, η ερμηνεύσιμη αδράνεια του νέφους των σημείων επί του χώρου προβολής *S* γίνεται μέγιστη. Μέσω της ΠΑΑ, αν και δεν είναι γνωστή η απόσταση και η κατεύθυνση των προφίλ από το βέλτιστο υποχώρο, επιτυγχάνεται, ωστόσο, η καλύτερη δυνατή τους αναπαράσταση με ταυτόχρονη μείωση του αριθμού των διαστάσεων. Ο λόγος της αδράνειας του υποχώρου προς την ολική αδράνεια είναι ένα μέτρο της ακρίβειας της αναπαράστασης του νέφους των σημείων στο βέλτιστο υποχώρο. Ανάλογα ισχύουν για τα προφίλ των στηλών του πίνακα **F**.

Με την ΠΑΑ προσδιορίζονται οι κύριοι άξονες της αδράνειας και για κάθε άξονα υπολογίζεται η αντίστοιχη ιδιοτιμή (*eigenvalue*), η οποία είναι ίση με την αδράνεια του νέφους προς στην κατεύθυνση του αντίστοιχου άξονα (Benzécri & *Collaborateurs* 1973, Benzécri 1992, Le Roux & Rouanet 2004). Ο πρώτος παραγοντικός άξονας είναι η ευθεία προς την κατεύθυνση της οποίας η αδράνεια του νέφους είναι μέγιστη. Ο δεύτερος παραγοντικός άξονας είναι η αμέσως επόμενη ευθεία, κάθετη στον πρώτο άξονα, για την οποία η αδράνεια του νέφους είναι επίσης μέγιστη. Το ίδιο συμβαίνει με τους υπόλοιπους άξονες. Ο βέλτιστος υποχώρος είναι αυτός που ορίζουν οι πρώτοι σε τάξη παραγοντικοί άξονες. Η «μερική» αδράνεια (*μάζα σημείου × απόσταση σημείου από την αρχή του άξονα εις το τετράγωνο*) όλων των σημείων γραμμών (ή στηλών) κατά μήκος του άξονα ισοδυναμεί με την αδράνεια του άξονα.

Η αδράνεια ενός παραγοντικού άξονα είναι ο σταθμισμένος μέσος όρος των τετραγώνων των χ^2 αποστάσεων των προβολών των σημείων γραμμών (ή στηλών) επί του άξονα, από το κέντρο βάρους τους. Με άλλα λόγια, είναι το μέτρο της διασποράς των σημείων γραμμών (ή στηλών) προς την κατεύθυνση του άξονα. Η αδράνεια ενός άξονα μπορεί να διασπαστεί στις μερικές αδράνειες κάθε σημείου επί του άξονα. Σημεία γραμμών (ή στηλών) με υψηλή συνεισφορά στην αδράνεια ενός παραγοντικού άξονα καθορίζουν σε μεγάλο βαθμό τον προσανατολισμό και την ταυτότητά του (δηλαδή τη φυσική του ερμηνεία).

Τα συνημίτονα των γωνιών που σχηματίζουν τα διανύσματα θέσης των σημείων γραμμών (ή στηλών) με τους παραγοντικούς άξονες εκφράζουν το βαθμό συσχέτισής

τους με τους αντίστοιχους άξονες. Αποτελούν δείκτες ποιότητας της απεικόνισης των σημείων στον υποχώρο που προβάλλονται (Μαυρομάτης, 1999) και εκφράζουν το πόσο κοντά στην πραγματική τους θέση βρίσκεται η απεικόνισή τους στον επιλεγμένο υποχώρο.

2.2.7 Το Δυϊκό Πρόβλημα

Το «Πρόβλημα των Γραμμών» ενός πίνακα συμπτώσεων είναι η προβολή των σημείων (προφίλ) των γραμμών σε έναν υποχώρο του αρχικού, μικρότερης διάστασης, ο οποίος βρίσκεται όσο το δυνατό πλησιέστερα στον αρχικό. Τα σημεία γραμμών προβάλλονται σ' έναν τέτοιο υποχώρο για να ερμηνευτούν οι μεταξύ τους αποστάσεις. Το ίδιο πρόβλημα πρέπει να λυθεί και για τα σημεία (προφίλ) των στηλών («Πρόβλημα των Στηλών»).

Τα δύο προβλήματα συνδέονται στενά μεταξύ τους. Όταν το πρόβλημα των γραμμών βρίσκει λύση, τότε αυτόματα βρίσκει λύση και το πρόβλημα των στηλών (Benzécri 1992, Greenacre 1993α και 1984, Μαυρομάτης 1999, Καραπιστόλης 1999, Παπαδημητρίου 2006, 2004 και 1994). Οι δύο αναλύσεις είναι ισοδύναμες διότι έχουν την ίδια ολική αδράνεια για τον ίδιο αριθμό παραγοντικών αξόνων, το ίδιο πλήθος διαστάσεων και η αδράνεια διασπάται με τον ίδιο τρόπο στους παραγοντικούς άξονες. Η δυνατότητα γραφικής ερμηνείας των αποτελεσμάτων, μέσω της σύγκρισης αποστάσεων μεταξύ σημείων, καθιστά την ΠΑΑ συγγενική μέθοδο με αυτή της Πολυδιάστατης Κλιμάκωσης (*Multidimensional Scaling*).

2.2.8 Συνεισφορές Σημείων Γραμμών και Στηλών στην Αδράνεια

Όπως αναφέρθηκε στα προηγούμενα, η ολική αδράνεια ενός πίνακα συμπτώσεων αποτελεί ένα μέτρο της συνολικής διακύμανσης των προφίλ των γραμμών ή των στηλών. Κάθε σημείο γραμμής ή στήλης συνεισφέρει με την αδράνειά του, ως ένα βαθμό, στην ολική αδράνεια. Επίσης, κάθε σημείο γραμμής ή στήλης που προβάλλεται σε άξονα *s* συνεισφέρει με τη μερική του αδράνεια στην αδράνεια του άξονα. Ο λόγος (%) της μερικής αδράνειας ενός σημείου γραμμής (στήλης), που προβάλλεται επί ενός άξονα *s*, προς την συνολική αδράνεια του άξονα συμβολίζεται με *CTR*(*s*) και επιτρέπει τον εντοπισμό των σημείων που συνεισφέρουν περισσότερο

στο να λάβει ο άξονας τη συγκεκριμένη θέση (προσανατολισμό) στο χώρο (Benzécri 1992, Παπαδημητρίου 2006, 2004 και 1994). Σε κάθε άξονα το άθροισμα των δεικτών *CTR* για τις γραμμές και τις στήλες είναι σταθερό και ίσο με τη μονάδα (Benzécri 1992, Lebart, Morineau & Piron 2000).

Ο λόγος της μερικής αδράνειας ενός σημείου γραμμής (ή στήλης), που προβάλλεται επί άξονα *s*, προς τη συνολική αδράνεια του σημείου συμβολίζεται με *COR*(*s*) και μετρά την ποιότητα αναπαράστασης του σημείου πάνω στο συγκεκριμένο άξονα. Από μια άλλη σκοπιά, ο δείκτης *COR* εκφράζει τη συμβολή του παραγοντικού άξονα *s* στην αδράνεια του αντίστοιχου σημείου (Benzécri, 1992). Γεωμετρικά, ο δείκτης ισούται με το τετράγωνο του συνημίτονου της γωνίας, που σχηματίζει ο άξονας *s* με την ευθεία που ενώνει το αντίστοιχο σημείο με την αρχή του συστήματος συντεταγμένων (κέντρο βάρους) και μπορεί να ερμηνευτεί ως το τετράγωνο του συντελεστή γραμμικής συσχέτισης του σημείου με τον άξονα. Ο δείκτης *COR* παίρνει τιμές στο διάστημα [0,1]. Όταν ο δείκτης *COR* ενός σημείου *i* πάνω στον άξονα *s* είναι κοντά στη μονάδα, τότε το αντίστοιχο σημείο αποκλίνει από την αρχή των παραγοντικών αξόνων προς την κατεύθυνση του άξονα *s* και επί αυτού του άξονα ερμηνεύεται η απόκλισή του από τη μέση κατάσταση η οποία προβάλλεται στο κέντρο βάρους του νέφους των σημείων. Αντιθέτως, όταν ο δείκτης *COR* είναι κοντά στο μηδέν, το σημείο *i* αποκλίνει από την αρχή των παραγοντικών αξόνων προς μια κατεύθυνση κάθετη στον άξονα *s* και αυτός ο άξονας δεν συμμετέχει στην ερμηνεία της απόκλισης του σημείου *i* από τη μέση κατάσταση. Σε αναλογία με την Ανάλυση σε Κύριες Συνιστώσες, για κάθε σημείο, η τετραγωνική ρίζα του *COR* με πρόσημο αυτό της αντίστοιχης προβολής της συντεταγμένης του επί του παραγοντικού άξονα *s* μπορεί να ερμηνευτεί ως το «φορτίο» (συντελεστής συσχέτισης) του σημείου στον άξονα (Blasius & Greenacre, 1994). Στην περίπτωση που επιλεγεί μια λύση, για παράδειγμα, με τρεις άξονες, τότε για κάθε σημείο το άθροισμα *QLT* των τριών αντίστοιχων δεικτών *COR* εκφράζει τη συνολική ποιότητα απεικόνισης – προβολής του στο χώρο των τριών διαστάσεων. Η λειτουργικότητα του αθροιστικού *COR* είναι ανάλογη με αυτή της Κοινής Παραγοντικής Διακύμανσης (*communality*) που χρησιμοποιείται στην Ανάλυση σε Κύριες Συνιστώσες και στην Παραγοντική Ανάλυση (βλέπε Hair *et al*., 1995). Σε κάθε γραμμή ή στήλη του πίνακα συμπτώσεων το άθροισμα των δεικτών *COR* για όλους τους $p=min\{k-1, l-1\}$ σε πλήθος άξονες, που είναι δυνατό να προκύψουν από την εφαρμογή της ΠΑΑ, είναι ίσο με τη μονάδα

(Benzécri 1992, Lebart, Morineau & Piron 2000, Παπαδημητρίου 2006, 2004 και 1994).

Και οι δύο δείκτες (*CTR* και *COR*) χρησιμοποιούνται στην επιλογή των σημαντικών σημείων που χρήζουν ερμηνείας σε κάθε άξονα (Benzécri 1992, Καραπιστόλης 1999). Με βάση τα σημαντικά σημεία επιχειρείται η απόδοση φυσικής ερμηνείας και ταυτότητας στους παραγοντικούς άξονες.

2.2.9 Μέγιστος Αριθμός Διαστάσεων

Εφόσον η περιθώρια στήλη των *l* στηλών ισούται με τα αθροίσματα των *k* γραμμών και η περιθώρια στήλη των *k* γραμμών ισούται με τα αθροίσματα των *l* στηλών του πίνακα **F** υπάρχουν, κατά μια έννοια, μόνο *l*-1 ανεξάρτητα στοιχεία σε κάθε γραμμή και *k*-1 ανεξάρτητα στοιχεία σε κάθε στήλη. Αυτό σημαίνει ότι ο μέγιστος αριθμός ιδιοτιμών και κατά συνέπεια των παραγοντικών αξόνων που μπορούν να παραχθούν από την εφαρμογή της ΠΑΑ σε έναν απλό πίνακα συμπτώσεων είναι $p=min\{k\text{-}1, l\text{-}1\}$ (βλέπε, Andersen 1991, Greenacre 1993α και 1984, Μπεχράκης 1999, Le Roux & Rouanet 2004). Αν από τα αποτελέσματα της ΠΑΑ διατηρήσουμε όλες τις δυνατές διαστάσεις, μπορούμε να ανασυστήσουμε ακριβώς τον αρχικό πίνακα συμπτώσεων.

2.2.10 Έκτοπα - Παράτυπα Σημεία (*Outliers*)

Συχνά, μετά την εφαρμογή της ΠΑΑ και την απεικόνιση των σημείων στους παραγοντικούς άξονες, εμφανίζονται σημεία που έχουν υψηλή συνεισφορά (*CTR*) στον προσανατολισμό του άξονα και ταυτόχρονα απέχουν πολύ από τα υπόλοιπα σημεία. Τα σημεία αυτά ονομάζονται «έκτοπα» ή «παράτυπα» (*outliers*) (Clausen 1998, Bendixen 2003) και γενικά δημιουργούν δυσκολίες στην ερμηνεία των παραγοντικών αξόνων. Έχουν, επίσης, αρνητική επίδραση στην «εσωτερική»[3] σταθερότητα των αποτελεσμάτων. Τα έκτοπα σημεία, τα οποία συνήθως έχουν μικρή μάζα, προβάλλονται σε απομακρυσμένη θέση σε σχέση με το κέντρο βάρους (αρχή

[3] Τα αποτελέσματα της ΠΑΑ χαρακτηρίζονται από εσωτερική σταθερότητα αν μικρές ή/και ανεπαίσθητες μεταβολές ή, καλύτερα, διαταραχές των αρχικών δεδομένων εισόδου έχουν ως αποτέλεσμα μόνο μικρές ή/και ανεπαίσθητες αλλαγές στην έξοδο της μεθόδου. Για παράδειγμα, τέτοιου είδους διαταραχές είναι δυνατό να προκαλέσουν τα παράτυπα σημεία (*outliers*), η συνένωση κλάσεων των μεταβλητών και η απομάκρυνση μεταβλητών ή αντικειμένων από την ανάλυση.

του συστήματος συντεταγμένων) και κυριαρχούν στην ερμηνεία ενός ή περισσοτέρων αξόνων. Έτσι, είναι δυνατό να αποκρύψουν τις ενδιαφέρουσες αντιθέσεις μεταξύ των σημαντικών σημείων, τα οποία εμφανίζονται στο παραγοντικό επίπεδο πιο συσπειρωμένα και πιο κοντά στην αρχή των αξόνων. Τα έκτοπα σημεία αποκλίνουν σημαντικά από την κατάσταση ανεξαρτησίας των δύο μεταβλητών και ο εντοπισμός και ο χειρισμός τους δεν είναι εν γένει απλός (Yick & Lee, 1998), ιδιαίτερα σε πολυδιάστατο χώρο (Rousseeuw & Van Zomeren 1990, Kosinski 1999, Becker & Gather 2001). Τα σημεία αυτά είτε απομακρύνονται είτε εισάγονται στην ανάλυση ως συμπληρωματικά (Καραπιστόλης 1999, Μαυρομάτης 1999, Bendixen 2003). Οι Nowak και Bar-Hen (2005) υποστηρίζουν ότι η απομάκρυνση των έκτοπων σημείων από την ανάλυση δεν είναι και η αποτελεσματικότερη πρακτική για τη βελτίωση της σταθερότητας των αποτελεσμάτων. Μεγαλύτερη επίδραση έχει η απόσταση των σημείων από το κέντρο βάρους απ’ ότι η μάζα τους. Σύμφωνα με το Rao (1995), η ιδιαίτερη βαρύτητα που δίνεται στις “σπάνιες” κατηγορίες αποτελεί μειονέκτημα της ΠΑΑ. Αντίθετα, για τον Greenacre (2006) τα σημεία με μικρή μάζα δεν έχουν κατ’ ανάγκη αρνητική επίδραση. Η μικρή τους μάζα είναι αυτή ακριβώς που αμβλύνει την επίδρασή τους. Επίσης, είναι δυνατό ορισμένες γραμμές ή/και στήλες του πίνακα συμπτώσεων ή ορισμένα κελιά του ή ακόμα και μεμονωμένες παρατηρήσεις να επηρεάζουν σημαντικά είτε τη σταθερότητα των αποτελεσμάτων (βλέπε Ενότητα 1.5.3) είτε τη βασική δομή του πίνακα που αναλύεται (ιδιοτιμές, ιδιοδιανύσματα) χωρίς τα αντίστοιχα σημεία να είναι υποχρεωτικά έκτοπα. Τα σημεία αυτά ενδεχομένως να παίζουν καθοριστικό ρόλο στη διαμόρφωση των αποτελεσμάτων (συσχετίσεις, ομοιότητες, αντιθέσεις) και να οδηγούν είτε σε συνενώσεις είτε σε απομάκρυνση κλάσεων. Μέθοδοι εντοπισμού των σημείων αυτών και εκτίμησης της επίδρασής τους στα αποτελέσματα της ΠΑΑ έχουν προταθεί από τους Escofier και Le Roux (1976), Pack και Jolliffe (1992), Krzanowski (1993), Bénasséni (1993), Nakayama (2001) και Nowak και Bar-Hen (2005). Αν και το ζήτημα παρουσιάζει θεωρητικό ενδιαφέρον, ωστόσο, στην πράξη, οι δείκτες *CTR*, *COR* και *QLT* αποτελούν επαρκή οδηγό για την ερμηνεία των αποτελεσμάτων και τον εντοπισμό των σημείων που επηρεάζουν τη διαμόρφωση των αποτελεσμάτων.

Στην περίπτωση πολλών μεταβλητών, το πρόβλημα της εμφάνισης κατηγοριών μιας μεταβλητής με χαμηλές συχνότητες μπορεί να αντιμετωπιστεί με τη σύμπτυξη των κατηγοριών αυτών με άλλες. Όμως, αυτό δεν είναι πάντα εφικτό λόγω περιορισμών

που μπορεί να εισάγονται από το θεωρητικό πλαίσιο στο οποίο θα ερμηνευτούν τα αποτελέσματα (βλέπε Παρατήρηση 2.4). Για την επίλυση του προβλήματος οι Le Roux και Chiche (2004) προτείνουν τη μέθοδο της «Ιδιάζουσας Πολλαπλής Ανάλυσης των Αντιστοιχιών» (*Specific Multiple Correspondence Analysis*), η οποία αντιμετωπίζει το ζήτημα από μια διαφορετική σκοπιά: οι κλάσεις με χαμηλή συχνότητα αντί να συμπτυχθούν ή να αφαιρεθούν από την ανάλυση απλά δεν λαμβάνονται υπόψη (αγνοούνται) στον υπολογισμό των αποστάσεων. Με τη μέθοδο αυτή μπορούν να αντιμετωπιστούν και τα προβλήματα που δημιουργούν οι ελλείπουσες τιμές καθώς και η άρνηση απάντησης σε έρευνες με ερωτηματολόγιο.

2.2.11 Συμπληρωματικά Σημεία

Ένα χρήσιμο χαρακτηριστικό της ΠΑΑ, το οποίο τη διαφοροποιεί από άλλες μεθόδους της Ανάλυσης Δεδομένων, είναι ότι κάθε επιπρόσθετη γραμμή (ή στήλη) του αρχικού πίνακα συμπτώσεων μπορεί να προβληθεί ως σημείο σε παραγοντικό άξονα ή επίπεδο, αρκεί το προφίλ του να είναι συγκρίσιμο με τα υπάρχοντα προφίλ γραμμών (ή στηλών) που συνιστούν τον άξονα ή το επίπεδο. Τα επιπλέον σημεία, ενώ παρουσιάζει ενδιαφέρον η ερμηνεία της σχετικής τους θέσης ως προς τα υπόλοιπα, δεν είναι επιθυμητό να συμμετέχουν στην κατασκευή και τον προσανατολισμό των παραγοντικών αξόνων. Τα σημεία αυτά ονομάζονται «συμπληρωματικά» σε αντίθεση με τα «ενεργά σημεία» που είναι όλα τα υπόλοιπα (Greenacre 1993α, Clausen 1998). Υπάρχουν περιπτώσεις όπου ο ερευνητής διαθέτει δεδομένα δευτερεύουσας σημασίας, τα οποία όμως είναι χρήσιμα για τη διερεύνηση των σχέσεων μεταξύ δεδομένων που είναι πρωταρχικής σημασίας. Τα συμπληρωματικά σημεία συμμετέχουν στην ανάλυση με μηδενική μάζα και απεικονίζονται στους άξονες έτσι ώστε η συνεισφορά τους στην αδράνεια του άξονα, και, συνεπώς, στην ολική αδράνεια, να είναι μηδενική (Greenacre, 1993α).

Μπορούμε να διακρίνουμε δύο γενικές περιπτώσεις στις οποίες ενδεχόμενη συμπληρωματική διαθέσιμη πληροφορία θα μπορούσε να προστεθεί σε αυτή του ενεργού συνόλου δεδομένων:

(α) Τα συμπληρωματικά στοιχεία μπορεί να προέρχονται από την ίδια έρευνα με τα ενεργά αλλά να είναι είτε δομικά είτε εννοιολογικά διαφορετικά από τα υπόλοιπα με

συνέπεια να είναι μεν επιθυμητή η ερμηνεία τους επάνω στους παραγοντικούς άξονες, χωρίς όμως να επηρεάζουν την κατασκευή τους (Benzécri 1992, Le Roux & Rouanet 2004). Ως δομικά διαφορετικά μπορούν να θεωρηθούν τα έκτοπα σημεία και οι ελλείπουσες τιμές (Μαυρομάτης, 1999), ενώ ως εννοιολογικά διαφορετικά οι μεταβλητές που χρησιμοποιούνται για την καταγραφή του δημογραφικού προφίλ των συμμετεχόντων στην έρευνα (π.χ. το φύλο, το μορφωτικό επίπεδο, το επάγγελμα και η ηλικία) (Lebart, Morineau & Warwick 1984, Le Roux & Rouanet 2004, Murtagh 2005). Στην πρώτη περίπτωση, αν μετά από μελέτη των παραγοντικών αξόνων, που προέκυψαν από την εφαρμογή της ΠΑΑ, διαπιστωθεί ότι κάποιο σημείο απέχει πολύ από τα υπόλοιπα, με αποτέλεσμα να συμμετέχει σε μεγάλο βαθμό στην κατασκευή του άξονα και να αποκρύπτει τις επικρατέστερες αντιπαραθέσεις ή να παραμορφώνει τις υπόλοιπες σχέσεις, τότε το έκτοπο σημείο μπορεί να προβληθεί ως συμπληρωματικό (Greenacre 1993α, Καραπιστόλης 1999, Μαυρομάτης 1999). Στη δεύτερη περίπτωση, τα δημογραφικά στοιχεία συμμετέχουν στις στατιστικές αναλύσεις κατά κανόνα ως ανεξάρτητες μεταβλητές. Συχνά, οι κλάσεις των μεταβλητών αυτών εισάγονται ως συμπληρωματικά σημεία, ώστε οι μεταξύ τους συσχετίσεις να μην επηρεάσουν τις συσχετίσεις των εξαρτημένων μεταβλητών (Lebart, Morineau & Warwick 1984, Greenacre 1984, Hoffman & Franke 1986, Van der Heijden & De Leeuw 1989, Benzécri 1992, Van de Geer 1993β, Gettler-Summa 1992, Thiessen, Rohlinger & Blasius 1994, Dohoo *et al*. 1996, Micheloud 1997, Γιαλαμάς & Κασιμάτη 2004, Le Roux & Rouanet 2004, Καρλής 2005, Torres & Van de Velden 2007). Επιπλέον, στην περίπτωση που είναι επιθυμητή η μελέτη της θέσης ορισμένων προφίλ σε σχέση με τη θέση ενός γενικότερου προφίλ που αποτελεί ομάδα ή υποομάδα των προηγούμενων (π.χ. το άθροισμά τους), το ομαδικό προφίλ μπορεί να εισαχθεί στην ανάλυση ως συμπληρωματικό, ώστε να μην συμμετέχει διπλά στην ανάλυση (Greenacre 1984, Micheloud 1997, Meulman & Heiser 2004, Καρλής 2005). Τέλος, αν ο ερευνητής επιθυμεί να εξετάσει τη θέση ενός προφίλ – στόχου ή "ιδανικού" προφίλ σε σχέση με τα υπάρχοντα, μπορεί να το εισάγει, και στην περίπτωση αυτή, ως συμπληρωματικό (Greenacre, 1993α).

(β) Τα συμπληρωματικά στοιχεία μπορεί να προέρχονται από διαφορετική έρευνα σε σχέση με τα ενεργά, για παράδειγμα από μια παρόμοια έρευνα που πραγματοποιήθηκε στο παρελθόν (Greenacre 1993α, Καρλής 2005). Στην περίπτωση αυτή, είναι δυνατή η μελέτη της διαχρονικής εξέλιξης του υπό εξέταση φαινομένου

με την εισαγωγή των στοιχείων του πίνακα συμπτώσεων της παλαιότερης έρευνας ως συμπληρωματικών στον πίνακα με τα ενεργά δεδομένα.

Για τα συμπληρωματικά σημεία, τις περισσότερες φορές, δεν έχει νόημα ο υπολογισμός της μάζας τους, εκτός κι αν α) η μάζα τους είναι ερμηνεύσιμη και συγκρίσιμη σε σχέση με αυτή των ενεργών, και β) πρόκειται τα συμπληρωματικά σημεία σε επόμενη ανάλυση να γίνουν ενεργά. Επιπλέον, στην περίπτωση που τα συμπληρωματικά σημεία είναι εκφρασμένα στις ίδιες μονάδες μέτρησης με τα ενεργά, τότε έχει αξία ο υπολογισμός της αδράνειάς τους, της συνεισφοράς – συμβολής (*COR*) ενός άξονα στην αδράνειά τους και η ποιότητα προβολής τους *QLT* στον υποχώρο που θα επιλεγεί (Lebart, Morineau & Warwick 1984, Benzécri 1992, Greenacre 1993α).

2.2.12 Η Βασική Δομή Πίνακα Δεδομένων

Έστω **X** ένας *n*×*m* πίνακας ποσοτικών δεδομένων της μορφής «*αντικείμενα* × *μεταβλητές*». Ο **X** μπορεί να είναι και πίνακας συμπτώσεων απολύτων ή σχετικών συχνοτήτων δύο κατηγορικών μεταβλητών. Κάθε πίνακας δεδομένων μπορεί να αναλυθεί στη βασική του δομή με τη μέθοδο της «Διάσπασης σε Χαρακτηριστικές Τιμές» (*Singular Value Decomposition – SVD*)[4] (βλέπε Golub & Van Loan 1989, Sharma 1996, Kalman 1996, Harville 1997, Meyer 2000, Strang 2001). Η πληροφορία που περικλείεται στον πίνακα **X** μπορεί να αναλυθεί σε τρεις διαφορετικούς πίνακες, που περιγράφουν τη δομή του (Weller & Romney, 1990), όπως φαίνεται στο παρακάτω σχήμα (Σχήμα 2.1). Χωρίς περιορισμό της γενικότητας, υποθέτουμε ότι *n*>*m*.

$$\underset{\mathbf{X}_{(n\times m)}}{\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}} \rightarrow \underset{\mathbf{U}_{(n\times m)}}{\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1m} \\ u_{21} & u_{22} & \cdots & u_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ u_{n1} & u_{n2} & \cdots & u_{nm} \end{bmatrix}} \underset{\mathbf{D}_{(m\times m)}}{\begin{bmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & d_{mm} \end{bmatrix}} \underset{\mathbf{V}_{(m\times m)}}{\begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1m} \\ v_{21} & v_{22} & \cdots & v_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ v_{m1} & v_{m2} & \cdots & v_{mm} \end{bmatrix}}$$

Σχήμα 2.1: Η Ανάλυση της Βασικής Δομής Πίνακα Δεδομένων

[4] Η μέθοδος είναι γνωστή και ως *Eckart-Young Decomposition*.

Ο πίνακας **U** συνοψίζει την πληροφορία που βρίσκεται στις γραμμές του πίνακα **X**. Οι γραμμές του **U** αντιστοιχούν στις γραμμές του **X**, ενώ οι στήλες του **U** αντιπροσωπεύουν τις λανθάνουσες διαστάσεις ή συνιστώσες που χαρακτηρίζουν τα αντικείμενα του **X**. Όμοια, ο πίνακας **V** συνοψίζει την πληροφορία που βρίσκεται στις στήλες του **X**. Οι γραμμές του **V** αντιστοιχούν στις στήλες του **X** και οι στήλες του **V** εκφράζουν τις λανθάνουσες διαστάσεις ή αλλιώς τους «παραγοντικούς άξονες» που χαρακτηρίζουν τις μεταβλητές (στήλες) του πίνακα **X**. Οι στήλες των πινάκων **U** και **V** αντιπροσωπεύουν τις βασικές συνιστώσες της δομής ή καλύτερα του πληροφοριακού περιεχομένου του πίνακα **X**. Ο διαγώνιος πίνακας **D** περιέχει τις χαρακτηριστικές τιμές (*singular values*) που αντιστοιχούν στις στήλες των πινάκων **U** και **V**. Το πρώτο στοιχείο d_{11} της διαγωνίου αντιστοιχεί στην πρώτη στήλη του **U** και στην πρώτη στήλη του **V**, το δεύτερο στοιχείο d_{22} στη δεύτερη στήλη των **U** και **V**, κ.ο.κ. Οι τιμές των d_{ii} μπορούν να θεωρηθούν ως «βάρη» που δηλώνουν τη σχετική σημαντικότητα κάθε συνιστώσας ή διάστασης των **U** και **V** και είναι διατεταγμένες σε φθίνουσα σειρά, δηλαδή $d_{11} \geq d_{22} \geq \cdots \geq d_{mm}$. Έτσι, οι στήλες των **U** και **V** καθώς και τα στοιχεία του πίνακα **D** είναι διατεταγμένα από το περισσότερο στο λιγότερο σημαντικό ως προς τη συνολική δομή ή το πληροφοριακό περιεχόμενο του πίνακα **X**. Αν ο **X** δεν περιέχει πλεονάζουσα πληροφορία (π.χ. γραμμικά εξαρτημένες στήλες ή γραμμές), τότε ο αριθμός στηλών των **U** και **V** καθώς και οι διαστάσεις του **D** θα είναι ίσος με τη μικρότερη διάσταση του **X** που στη συγκεκριμένη περίπτωση είναι *m*. Σε αντίθετη περίπτωση, οι ωφέλιμες ή αλλιώς χρήσιμες διαστάσεις θα είναι αριθμός μικρότερος από *m*, ίσος με τη βαθμίδα ή τάξη (*rank*) του πίνακα **X**. Αλγεβρικά, η ανάλυση ή αλλιώς διάσπαση του **X** στους τρεις πίνακες (SVD) δίνεται από τη σχέση (Golub & Van Loan, 1989):

$$\mathbf{X}_{(n\times m)} = \mathbf{U}_{(n\times m)}\mathbf{D}_{(m\times m)}\mathbf{V}^{\mathbf{T}}_{(m\times m)} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathbf{T}}_{(n\times m)}. \qquad [2.6]$$

Από τη σχέση [2.6] είναι δυνατή είτε η πλήρης είτε η μερική ανασύσταση του πίνακα **X**. Η πλήρης ανασύσταση επιτυγχάνεται αν χρησιμοποιηθούν όλες οι ωφέλιμες διαστάσεις, ενώ η μερική αν χρησιμοποιηθεί υποσύνολο των στοιχείων των πινάκων **U**, **V** και **D** (Weller & Romney, 1990). Με τον τρόπο αυτό, μπορούμε να έχουμε μονοδιάστατη, διδιάστατη κ.ο.κ., ανασύσταση ή προσέγγιση του πίνακα **X** ανάλογα με τον αριθμό στηλών των **U** και **V** και των στοιχείων του **D** που θα χρησιμοποιηθούν. Για παράδειγμα, η τριδιάστατη ανασύσταση του **X** μπορεί να

υλοποιηθεί πολλαπλασιάζοντας τις τρεις πρώτες στήλες του **U** με τα τρία πρώτα διαγώνια στοιχεία του **D** και τις τρεις στήλες του **V** (ή τις τρεις πρώτες γραμμές του $\mathbf{V^T}$).

Οι τρεις πίνακες **U**, **V** και **D**, στους οποίους αναλύεται η βασική δομή του πίνακα **X**, περιστρέφουν το νέφος των σημείων σε έναν *n*-διάστατο Ευκλείδειο χώρο (βλέπε και Εικόνα 2.1). Οι τρεις πίνακες επανατοποθετούν τα σημεία δεδομένων σε ένα νέο σύστημα συντεταγμένων, στο οποίο η νέα προβολή είναι δυνατό να αποτελεί περιστροφή, τάνυσμα ή/και ανάκλαση της αρχικής μορφής του νέφους. Οι γραμμές του πίνακα **X** (αντικείμενα) απεικονίζονται ως σημεία με συντεταγμένες τα στοιχεία των διανυσμάτων των στηλών του **U**, ενώ οι στήλες του **X** (μεταβλητές) προβάλλονται με συντεταγμένες τα στοιχεία των διανυσμάτων των στηλών του **V**. Τα διανύσματα συντεταγμένων είναι κάθετα μεταξύ τους και μάλιστα ορθοκανονικά, δηλαδή ισχύει η σχέση (Weller & Romney 1990, Strang 2001):

$$\sum_i u_{it}^2 = \sum_j v_{jt}^2 = 1 \text{ ή ισοδύναμα } \mathbf{V^T V = U^T U = I},$$

όπου **I** είναι μοναδιαίος πίνακας κατάλληλων διαστάσεων.

Έτσι, τα διανύσματα στηλών των πινάκων **U** και **V** σχηματίζουν ένα κανονικοποιημένο σύστημα συντεταγμένων, στο οποίο προβάλλεται το αρχικό νέφος των σημείων δεδομένων. Οι στήλες των **U** και **V** ονομάζονται «αριστερά» και «δεξιά» αντίστοιχα ορθογώνια χαρακτηριστικά διανύσματα (*singular vectors*) του **X** και εκφράζουν διαφορετικές και ανεξάρτητες πηγές μεταβλητότητας στα δεδομένα (Jacoby 1998, Meyer 2000). Οι διαστάσεις του χώρου προβολής επιδιώκεται να είναι μικρότερες απ' ό,τι οι διαστάσεις του αρχικού χώρου, στον οποίο το υπό εξέταση φαινόμενο περιγράφεται μέσω του πίνακα **X**. Τα στοιχεία του πίνακα **D**, δηλαδή οι χαρακτηριστικές τιμές του **X**, αποτελούν βάρη ή συντελεστές "τανύσματος", οι οποίοι από-κανονικοποιούν το νέφος των σημείων και το επαναφέρουν στην αρχική του μορφή. Για παράδειγμα (Weller & Romney 1990, Kalman 1996), αν το αρχικό νέφος σημείων έχει τη μορφή μπάλας του ράγκμπι (ελλειψοειδές), τότε στο νέο κανονικοποιημένο ορθοκανονικό σύστημα συντεταγμένων, το οποίο ορίζουν οι στήλες των πινάκων **U** και **V**, η μορφή του νέφους θα μοιάζει περισσότερο με μπάλα ποδοσφαίρου (σφαίρα). Ο πολλαπλασιασμός των στοιχείων των **U** ή **V** με τις

αντίστοιχες χαρακτηριστικές τιμές έχει ως αποτέλεσμα την επαναφορά του νέφους στην αρχική του μορφή (μπάλα του ράγκμπι). Στην ειδική περίπτωση που ο πίνακας $\mathbf{X} \in \Re^{n\times n}$ είναι τετραγωνικός μη ιδιάζων και *S* είναι η μοναδιαία σφαίρα του $\Re^n$, τότε η εικόνα της *S* μέσω του μετασχηματισμού, στον οποίο αντιστοιχεί ο **X**, είναι ένα ελλειψοειδές (βλέπε Σχήμα 2.2) που το μήκος του *j*-ιοστού ημιάξονα είναι ίσο με τη *j*-ιοστή χαρακτηριστική τιμή που προκύπτει από την SVD του **X** (Meyer, 2000). Αν περιστρέψουμε τους αρχικούς άξονες συντεταγμένων, έτσι ώστε να συμπέσουν με τους άξονες του ελλειψοειδούς, τότε μπορούμε να θεωρήσουμε ότι μέσω του πίνακα **X** επιτυγχάνεται παραμόρφωση της σφαίρας. Όμως, η παραμόρφωση δεν είναι ίδια προς την κατεύθυνση όλων των αξόνων. Η διαστολή ή συστολή των αξόνων καθορίζεται από τα αντίστοιχα στοιχεία του πίνακα **D,** ενώ ο βαθμός παραμόρφωσης της σφαίρας μετριέται μέσω του «δείκτη κατάστασης» (*condition index*) του πίνακα **X**:

$$k = \frac{max\{d_{ii}\}}{min\{d_{ii}\}} = \frac{d_{11}}{d_{nn}}.$$

Στην περίπτωση που ο πίνακας **X** είναι συμμετρικός, τότε οι συνιστώσες των γραμμών και των στηλών του θα είναι ίσες. Δηλαδή, ισχύει (Harville, 1997):

$$\mathbf{X}=\mathbf{U}\mathbf{D}\mathbf{V}^{\mathbf{T}}=\mathbf{U}\mathbf{D}\mathbf{U}^{\mathbf{T}}=\mathbf{V}\mathbf{D}\mathbf{V}^{\mathbf{T}}. \qquad [2.7]$$

Η σχέση [2.7] είναι γνωστή και ως «φασματική ανάλυση» του πίνακα **X** (Sharma 1996, Meyer 2000). Αν πολλαπλασιάσουμε είτε από δεξιά είτε από αριστερά τον **X** με τον ανάστροφό του προκύπτει ένας επίσης τετραγωνικός συμμετρικός πίνακας. Η διάσπαση των πινάκων **X**, $\mathbf{X}\mathbf{X}^{\mathbf{T}}$ και $\mathbf{X}^{\mathbf{T}}\mathbf{X}$ αποκαλύπτει την ίδια βασική δομή (Blasius & Greenacre 1994, Strang 2001):

$$\mathbf{X}=\mathbf{U}\mathbf{D}\mathbf{V}^{\mathbf{T}}$$

$$\mathbf{X}\mathbf{X}^{\mathbf{T}}=\mathbf{U}\mathbf{D}^{\mathbf{2}}\mathbf{U}^{\mathbf{T}} \qquad [2.8]$$

$$\mathbf{X}^{\mathbf{T}}\mathbf{X}=\mathbf{V}\mathbf{D}^{\mathbf{2}}\mathbf{V}^{\mathbf{T}}. \qquad [2.9]$$

Επιπλέον, ισχύει (Weller & Romney, 1990):

$$rank(\mathbf{X})=rank(\mathbf{X}^{\mathbf{T}})=rank(\mathbf{X}\mathbf{X}^{\mathbf{T}})=rank(\mathbf{X}^{\mathbf{T}}\mathbf{X}). \qquad [2.10]$$

Η ανάλυση της βασικής δομής ενός τετραγωνικού συμμετρικού πίνακα είναι ισοδύναμη με την παραγοντοποίησή του μέσω της ανάλυσης των ιδιοτιμών του (Israëls 1987, Meyer 2000). Σε αυτήν την περίπτωση, οι συνιστώσες των γραμμών και των στηλών του, δηλαδή οι στήλες των πινάκων **U** και **V**, είναι τα αντίστοιχα ιδιοανύσματα, ενώ τα στοιχεία του **D** είναι οι αντίστοιχες ιδιοτιμές. Επίσης, το άθροισμα των διαγωνίων στοιχείων του **X**, δηλαδή το ίχνος του (*trace*), είναι ίσο με το άθροισμα των διαγωνίων στοιχείων του **D**, δηλαδή με το άθροισμα των ιδιοτιμών. Δηλαδή, ισχύει:

$$trace(\mathbf{X}) = \sum_{i=1}^{r} d_i,$$

όπου *r* έιναι η βαθμίδα (τάξη) του πίνακα **X**.

Κάθε μη τετραγωνικός πίνακας μπορεί να παραγοντοποιηθεί μέσω της ανάλυσης των ιδιοτιμών των παρακάτω πινάκων (Weller & Romney 1990, Strang 2001):

$$\mathbf{X}\mathbf{X}^{\mathbf{T}} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathbf{T}} \qquad [2.11]$$

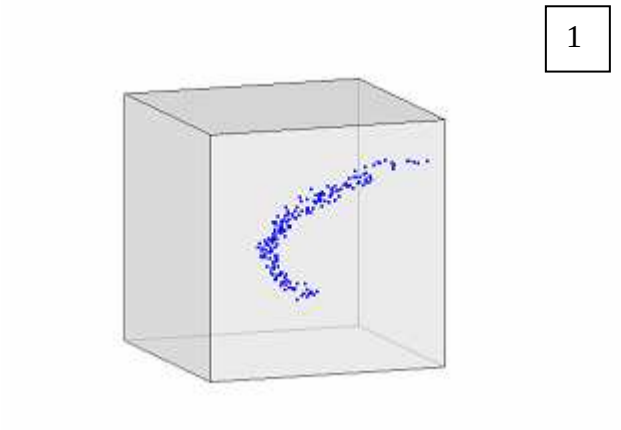
και

$$\mathbf{X}^{\mathbf{T}}\mathbf{X} = \mathbf{V}\mathbf{D}\mathbf{V}^{\mathbf{T}}. \qquad [2.12]$$

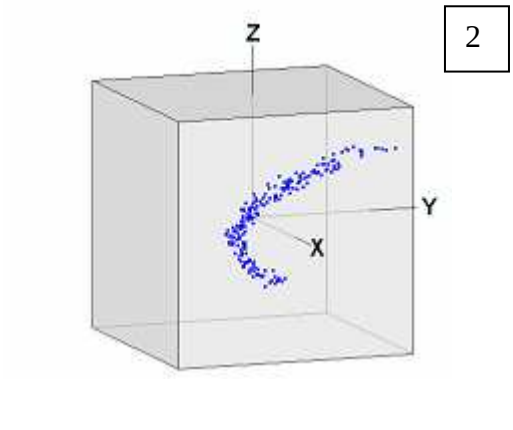
Από τις [2.11] και [2.12] συνεπάγεται ότι:

$$\mathbf{X} = \mathbf{U}\mathbf{D}^{1/2}\mathbf{V}^{\mathbf{T}}. \qquad [2.13]$$

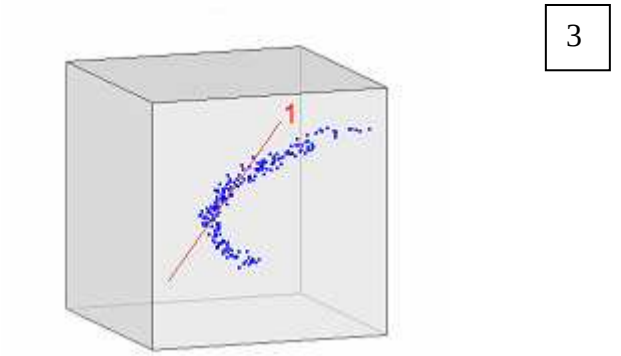
Στο πλαίσιο της ΠΑΑ ο πίνακας του οποίου η βασική δομή αναλύεται, είναι συνήθως ο πίνακας **S** με στοιχεία τα σχετικά τυποποιημένα υπόλοιπα (Greenacre, 1993α και 1984). Εκτός από την ΠΑΑ και άλλες μέθοδοι της Ανάλυσης Δεδομένων στηρίζονται στην ανάλυση της βασικής δομής του πίνακα που θα δοθεί ως είσοδος στη στατιστική ανάλυση. Χαρακτηριστικά αναφέρουμε την Ανάλυση σε Κύριες Συνιστώσες και την Κανονικοποιημένη Συσχέτιση (βλέπε Jackson 1991, Sharma 1996), όπου ως πίνακας εισόδου δίνεται είτε ο πίνακας διασπορών – συνδυασπορών ή ο πίνακας συσχετίσεων. Στην πράξη, η Διάσπαση σε Χαρακτηριστικές Τιμές ή Ιδιοτιμές δεν εφαρμόζεται απευθείας στον πίνακα δεδομένων **X** αλλά σε κάποια μετασχηματισμένη του μορφή, η οποία προκύπτει μετά από κατάλληλες για την περίπτωση κεντροποιήσεις ή τυποποιήσεις (π.χ. σε *z-scores*) των στοιχείων του (Weller & Romney, 1990).



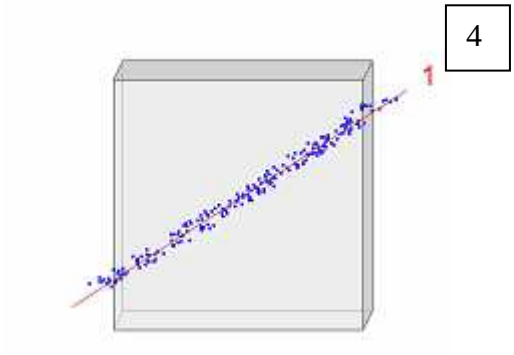
Αρχικό νέφος 100 σημείων σε σχήμα “μπανάνας”.



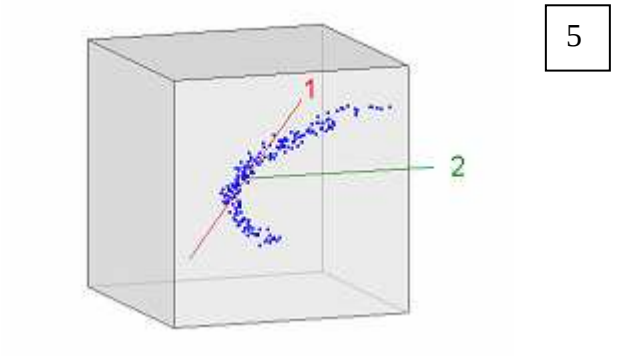
Το νέφος των σημείων σε καρτεσιανό σύστημα συντεταγμένων (*X*, *Y*, *Z*).



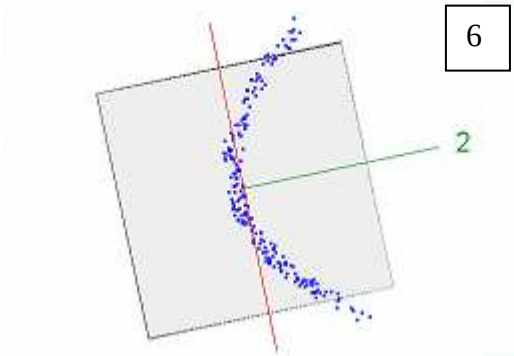
Ο πρώτος παραγοντικός άξονας (1) δείχνει την κατεύθυνση μέγιστης διακύμανσης. Αντιστοιχεί στην πρώτη χαρακτηριστική τιμή του πίνακα που αναλύεται μέσω της SVD.



Περιστροφή του διπλανού σχήματος (για καλύτερη εποπτεία).

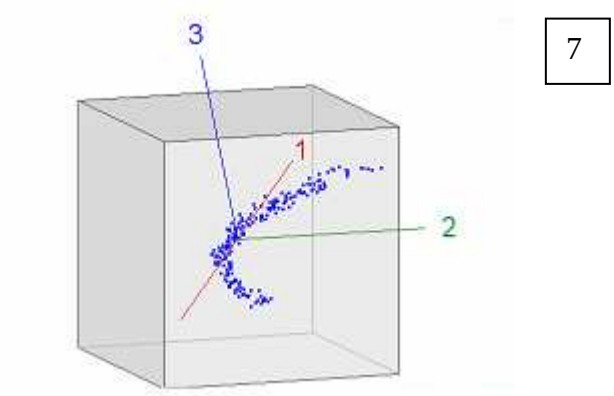


Ο δεύτερος παραγοντικός άξονας (2) είναι κάθετος στον πρώτο. Δείχνει την κατεύθυνση με την αμέσως επόμενη μεγαλύτερη διακύμανση και αντιστοιχεί στη δεύτερη χαρακτηριστική τιμή.

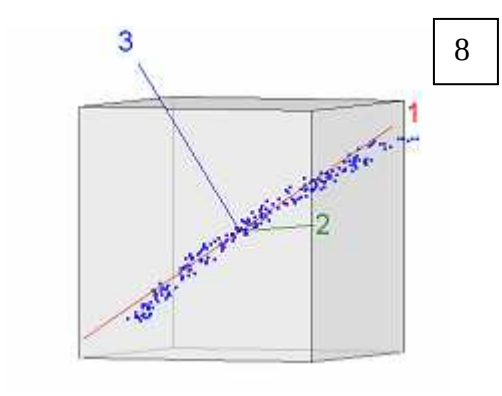


Περιστροφή του διπλανού σχήματος (για καλύτερη εποπτεία).

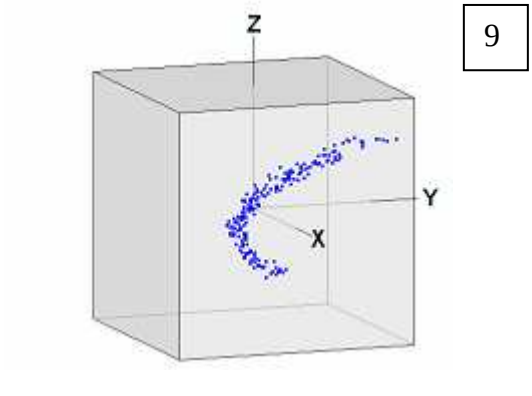
Εικόνα 2.1: Η Διαδικασία Εύρεσης των Παραγοντικών Αξόνων



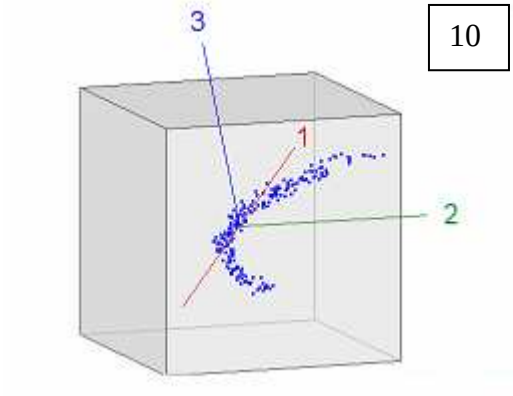
Ο τρίτος παραγοντικός άξονας (3) είναι κάθετος στο επίπεδο που ορίζουν ο πρώτος και ο δεύτερος και αντιστοιχεί στην τρίτη χαρακτηριστική τιμή.



Περιστροφή του διπλανού σχήματος (για καλύτερη εποπτεία).

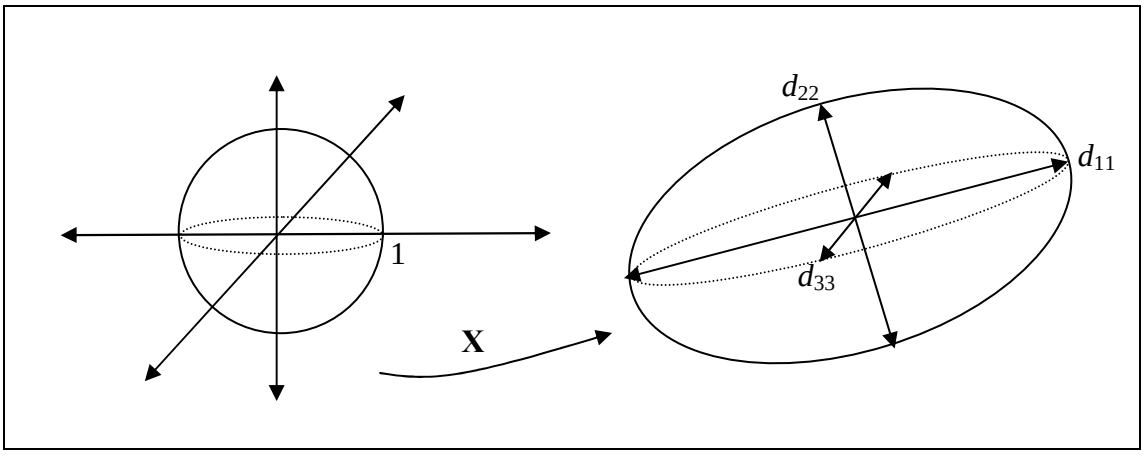


Αρχική κατάσταση.



Τελική κατάσταση. Αποτελεί στροφή του αρχικού συστήματος συντεταγμένων.

Εικόνα 2.1 (*συνέχεια*): Η Διαδικασία Εύρεσης των Παραγοντικών Αξόνων



Σχήμα 2.2: Η Μοναδιαία Σφαίρα του $\Re^3$, Μέσω της SVD του Πίνακα **X**, Μετασχηματίζεται σε Ελλειψοειδές του $\Re^3$

2.2.13 Biplots

Τα «*Biplots*» (Gabriel, 1971) είναι κυρίως ένα γραφικό εργαλείο ταυτόχρονης γραφικής απεικόνισης των γραμμών και των στηλών ενός πίνακα δεδομένων της μορφής «*αντικείμενα* × *μεταβλητές*» σε ένα κοινό χώρο, με τέτοιο τρόπο ώστε να αναδεικνύονται γραφικά και να οπτικοποιούνται οι μεταξύ τους σχέσεις. Τα *biplots* συνήθως παρουσιάζονται σε διαγράμματα δύο διαστάσεων. Το πρόθεμα "*bi*" αφορά στους δύο τύπους στοιχείων που αναπαρίστανται γραφικά (αντικείμενα και μεταβλητές) και όχι στον αριθμό των διαστάσεων. Έτσι, τα *biplots* μπορούν να γενικευτούν σε περισσότερες από δύο διαστάσεις (βλέπε Gower, 1990). Στα *biplots*, οι μεταβλητές απεικονίζονται ως διανύσματα και τα αντικείμενα ως σημεία, με τέτοιο τρόπο ώστε η τιμή – μέτρηση, για παράδειγμα, του *i* αντικειμένου στη μεταβλητή *j* να μπορεί να προσεγγιστεί (μοντελοποιηθεί) από το εσωτερικό γινόμενο των συντεταγμένων του σημείου που αναπαριστά το *i* αντικείμενο και του διανύσματος θέσης του σημείου που αντιστοιχεί στη μεταβλητή *j*. Πιο συγκεκριμένα, έστω **X** ένας $n \times m$ πίνακας ποσοτικών δεδομένων της μορφής «*αντικείμενα* × *μεταβλητές*», με $n>m$. Χωρίς περιορισμό της γενικότητας, μπορούμε να υποθέσουμε ότι οι μεταβλητές είναι σε τυποποιημένη μορφή με μέσο όρο 0 και διακύμανση ίση με 1. Σε κάθε περίπτωση, ο **X** μπορεί να παραγοντοποιηθεί στη μορφή (Gabriel, 1971):

$$\mathbf{X}=\mathbf{A}\mathbf{B}^{\mathbf{T}} .$$

Στη σχέση αυτή, ο **A** είναι ένας $n \times p$ πίνακας όπου τα στοιχεία των στηλών του εκφράζουν τις συντεταγμένες των n αντικειμένων σε ένα p-διάστατο ορθογώνιο σύστημα συντεταγμένων, με $p \le m$. Όμοια, ο πίνακας **B** είναι ένας $m \times p$ πίνακας οι γραμμές του οποίου περιέχουν τις συντεταγμένες των μεταβλητών στους ίδιους p άξονες. Οι στήλες των **A** και **B** ονομάζονται «παραγοντικοί άξονες του *biplot*» (Gabriel, 2002). Τα στοιχεία του πίνακα **X** μπορούν να ληφθούν υπολογίζοντας το εσωτερικό γινόμενο (Jacoby 1998, Gabriel 2002 και 1971):

$$x_{ij} = \sum_{r=1}^{p} \mathbf{a}_{ir}\mathbf{b}_{rj}^{\mathbf{T}} \text{ , με } i=1,\ldots, n \text{ και } j=1,\ldots, m.$$

Η παραπάνω σχέση δηλώνει ότι η συγκεκριμένη μέτρηση (τιμή) που βρίσκεται στη διασταύρωση της γραμμής *i* και της στήλης *j* του πίνακα **X** μπορεί να ανακτηθεί

συνδυάζοντας την πληροφορία για το *i* αντικείμενο, η οποία περιέχεται στην *i* γραμμή του πίνακα **A**, και την πληροφορία για τη *j* μεταβλητή που περιέχεται στη *j* στήλη του πίνακα $\mathbf{B}^{\mathbf{T}}$. Τα στοιχεία των πινάκων **A** και **B** προσδιορίζονται από την εφαρμογή της μεθόδου SVD στον πίνακα **X** (βλέπε προηγούμενη ενότητα):

$$\mathbf{X}=\mathbf{U}\mathbf{D}\mathbf{V}^{\mathbf{T}}.$$

Οι χαρακτηριστικές τιμές, δηλαδή τα στοιχεία του **D**, εκφράζουν τη συμμετοχή των αντίστοιχων χαρακτηριστικών διανυσμάτων στη συνολική διακύμανση των στοιχείων του **X**. Όπως είδαμε στην προηγούμενη ενότητα, κάθε χαρακτηριστική τιμή d_r συνδέεται με ένα συγκεκριμένο ζεύγος χαρακτηριστικών διανυσμάτων $\mathbf{u}_r$ και $\mathbf{v}_r$. Ας είναι τώρα $\mathbf{A}_{[2]}$ ο $n\times2$ πίνακας με στοιχεία τις *biplot* συντεταγμένες των αντικειμένων (γραμμών) του πίνακα **X** και $\mathbf{B}_{[2]}$ ο $m\times2$ πίνακας με στοιχεία τις *biplot* συντεταγμένες των μεταβλητών (στηλών) του **X**. Οι δύο αυτοί πίνακες μπορούν να προσδιοριστούν από τις παρακάτω σχέσεις (Lipkovich & Smith 2002, Martín-Rodríguez 2002):

$$\mathbf{A}_{[2]} = \mathbf{U}_{[2]}\mathbf{D}_{[2]}^{\boldsymbol{\alpha}}$$

και

$$\mathbf{B}_{[2]} = \mathbf{V}_{[2]}\mathbf{D}_{[2]}^{\mathbf{1-}\boldsymbol{\alpha}} .$$

Στις προηγούμενες σχέσεις, $\mathbf{U}_{[2]}$ και $\mathbf{V}_{[2]}$ είναι οι δύο πρώτες στήλες των **U** και **V** αντίστοιχα και $\mathbf{D}_{[2]}$ ο διαγώνιος πίνακας που σχηματίζεται από τις δύο πρώτες χαρακτηριστικές τιμές. Η τιμή του συντελεστή *α*, ο οποίος εμφανίζεται στους εκθέτες, μπορεί να μεταβάλλεται από 0 έως 1 (Jacoby 1998, Gower 2004). Η τιμή του *α* εκφράζει το βαθμό διάχυσης της μεταβλητότητας στο νέφος των δεδομένων (νέφος αντικειμένων ή μεταβλητών). Συνήθως, η τιμή του *α* τίθεται ίση με 0,5. Αυτό έχει ως αποτέλεσμα η διακύμανση να μοιράζεται εξίσου στα αντικείμενα και στις μεταβλητές. Από γεωμετρική σκοπιά, η τιμή α=0,5 επιφέρει μια αλλαγή στη κλίμακα μέτρησης των συντεταγμένων, τέτοια ώστε τα αντικείμενα και οι μεταβλητές να καταλαμβάνουν την ίδια περιοχή στο χώρο σχεδίασης. Κάθε γραμμή (αντικείμενο) του $\mathbf{A}_{[2]}$ προβάλλεται ως σημείο σε ένα διδιάστατο σύστημα συντεταγμένων, ενώ οι γραμμές (μεταβλητές) του $\mathbf{B}_{[2]}$ ως διανύσματα θέσης στον ίδιο χώρο (Gabriel 1971, Greenacre 1993β, Jacoby 1998, Meulman & Heiser 2004). Η γραφική αναπαράσταση των γραμμών του πίνακα $\mathbf{B}_{[2]}$ μπορεί να έχει τη μορφή σημείων με διαφορετική

κατάδειξη (*marker*) από αυτή των αντικειμένων ή βελών, τα οποία συνδέουν την αρχή του συστήματος συντεταγμένων με τα αντίστοιχα σημεία, ή ευθειών, οι οποίες διέρχονται από την αρχή και τα αντίστοιχα σημεία των μεταβλητών (Udina, 2005). Έτσι, τα αντικείμενα και οι μεταβλητές εύκολα διακρίνονται οπτικά κατά την ταυτόχρονη γραφική αναπαράστασή τους στο *biplot*. Η γεωμετρική ερμηνεία των *biplots* στηρίζεται στις παρακάτω ιδιότητες (Jacoby 1998, Gabriel 2002 και 1971):

α) Το συνημίτονο της γωνίας που σχηματίζουν τα διανύσματα θέσης δύο μεταβλητών εκφράζει το βαθμό γραμμικής συσχέτισης μεταξύ των μεταβλητών. Αν η γωνία είναι οξεία, τότε οι αντίστοιχες μεταβλητές συσχετίζονται θετικά. Αν είναι αμβλεία, οι μεταβλητές συσχετίζονται αρνητικά. Τέλος, στην περίπτωση που η γωνία είναι ορθή, τότε οι μεταβλητές είναι γραμμικά ανεξάρτητες.

β) Οι αποστάσεις μεταξύ των αντικειμένων καθορίζουν το βαθμό ομοιότητάς τους ως προς τις τιμές τους στις αντίστοιχες μεταβλητές που δομούν και χαρακτηρίζουν τον κάθε παραγοντικό άξονα. Για παράδειγμα, δύο αντικείμενα με παρόμοιες συντεταγμένες σε όλους τους άξονες, δηλαδή με παρόμοιο προφίλ, θα βρίσκονται κοντά το ένα με το άλλο στο χώρο προβολής, ενώ δύο σημεία με αρκετά διαφορετικό προφίλ θα απέχουν σημαντικά.

γ) Η σύνδεση ενός αντικειμένου *i* με μια μεταβλητή *j*, επιτυγχάνεται με την ορθογώνια προβολή του σημείου *i* στην ευθεία που είναι συγγραμμική με το διάνυσμα θέσης της μεταβλητής *j*. Αν η τιμή x_{ij} του *i* αντικειμένου στη μεταβλητή *j* είναι θετική, τότε η προβολή του σημείου *i* θα βρίσκεται επάνω στο διάνυσμα θέσης της *j*. Αν η τιμή x_{ij} είναι αρνητική, τότε η προβολή του σημείου *i* θα βρίσκεται στην προέκταση του διανύσματος θέσης της *j* και, μάλιστα, προς την αντίθετη κατεύθυνση από αυτή που βρίσκεται το σημείο που αντιστοιχεί στη μεταβλητή *j*. Όσο πιο μακριά από την αρχή των αξόνων βρίσκεται η ορθογώνια προβολή ενός σημείου, τόσο πιο ακραία (υψηλή) είναι η τιμή του στην αντίστοιχη μεταβλητή-διάνυσμα. Η ιδιότητα αυτή του *biplot* εξασφαλίζει ότι τα διανύσματα (μεταβλητές) θα είναι προσανατολισμένα προς τα αντικείμενα για τα οποία έχουν τις μεγαλύτερες τιμές.

Όπως και στην περίπτωση ανάλυσης της βασικής δομής ενός πίνακα δεδομένων (βλέπε προηγούμενη ενότητα) έτσι και στην περίπτωση των *biplots* μπορούμε να

επιτύχουμε μερική ή ολική ανασύσταση των στοιχείων του πίνακα εισόδου **X** ανάλογα με το πλήθος των χαρακτηριστικών διανυσμάτων που θα χρησιμοποιηθούν στην εκτίμηση των στοιχείων των πινάκων **A** και **B**. Η μέθοδος SVD αποτελεί ένα μετασχηματισμό μεγιστοποίησης της διακύμανσης των στοιχείων του πίνακα **X**. Αυτό σημαίνει, για παράδειγμα, ότι οι *s* πρώτες χαρακτηριστικές τιμές και τα αντίστοιχα χαρακτηριστικά διανύσματα μπορούν να χρησιμοποιηθούν για την όσο το δυνατό πιο ακριβή αναπαραγωγή της αρχικής διακύμανσης που περικλείεται στον **X** (Jacoby, 1998). Με αυτήν την έννοια, τα σημεία και τα διανύσματα που προβάλλονται στα *biplots* αντιπροσωπεύουν την "καλύτερη δυνατή" διδιάστατη προσέγγιση των αρχικών δεδομένων. Η έκφραση "καλύτερη δυνατή" θα πρέπει να ερμηνευτεί στο πλαίσιο των σταθμισμένων ελαχίστων τετραγώνων (Gabriel 1971, Osmond 1985, Andersen 1991). Με άλλα λόγια, τα στοιχεία του πίνακα $\mathbf{A}_{[2]}\mathbf{B}_{[2]}^{\mathbf{T}}$ έχουν τη μέγιστη δυνατή γραμμική συσχέτιση με τα αντίστοιχα στοιχεία του **X**. Συνεπώς, το αντίστοιχο διδιάστατο *biplot* θα περιέχει όσο το δυνατό περισσότερη πληροφορία, με την έννοια της διακύμανσης, από αυτή που περιέχεται στο *m*-διάστατο χώρο στον οποίο το υπό εξέταση φαινόμενο περιγράφεται μέσω του πίνακα **X**. Δηλαδή, το *biplot* αποτελεί μια βέλτιστη σύνοψη δύο διαστάσεων των αρχικών δεδομένων. Η ερμηνευτική αποτελεσματικότητα του *biplot*, με την έννοια της ανασύστασης της αρχικής ολικής διακύμανσης του πίνακα **X**, μπορεί να μετρηθεί μέσω δεικτών καλής προσαρμογής ή ποιότητας (βλέπε Heo & Gabriel 2001, Gabriel 2002, Gower 2004). Συνήθως, χρησιμοποιούνται δείκτες που εκφράζουν το ποσοστό της ολικής διακύμανσης που περιέχεται στο *biplot*. Για παράδειγμα, ο λόγος

$$\frac{d_1^2 + d_2^2}{\sum_{r=1}^{p} d_r^2}$$

εκφράζει το ποσοστό της ολικής διακύμανσης που ερμηνεύεται ή αιτιολογείται από τους δύο πρώτους άξονες του *biplot* (Gabriel 1971, Andersen 1991, Jacoby 1998, Martín-Rodríguez 2002, Udina 2005). Βέβαια, η αποτελεσματικότητα του *biplot* στην ανάδειξη της βασικής δομής ενός πίνακα δεδομένων εξαρτάται από το κατά πόσο είναι εφικτό η δομή αυτή να αναπαρασταθεί σε ένα χώρο με λιγότερες διαστάσεις από τον αρχικό (Jacoby, 1998). Όταν η λύση είναι πολυδιάστατη ενδεχομένως στο *biplot* να μην περιέχεται ικανοποιητικό ποσοστό της ολικής πληροφορίας του αρχικού

πίνακα δεδομένων. Το παραπάνω πρόβλημα μπορεί να αντιμετωπιστεί με κατάλληλη προεργασία στα αρχικά δεδομένα, η οποία μπορεί να περιλαμβάνει διαδικασίες επιλογής υποσυνόλων των αρχικών δεδομένων (αντικειμένων ή/και μεταβλητών). Για παράδειγμα, μπορούν να αποκλειστούν από την ανάλυση μεταβλητές που δεν έχουν σημαντική συσχέτιση με όλες τις υπόλοιπες (Hair *et al*. 1995, Sharma 1996).

Αν και έχουν προταθεί αρκετές παραλλαγές, βελτιώσεις και γενικεύσεις των *biplots* (βλέπε Osmond 1985, Gower & Harding 1988, Greenacre 1993α και 1993β, Carlier & Kroonenberg 1996, Jacoby 1998, Gabriel, Galindo & Vicente-Villardón 1998, Gower, Meulman & Arnold 1999, Aitchison & Greenacre 2002, Martín-Rodríguez 2002, Gower 2003, 1993, 1992 και 1990), ωστόσο ο βασικός σκοπός τους είναι η οπτικοποίηση πολυμεταβλητών δεδομένων, με ταυτόχρονη γραφική αναπαράσταση αντικειμένων και μεταβλητών στον ίδιο χώρο σε συνδυασμό με στατιστικές διαδικασίες μείωσης των διαστάσεων του χώρου των αρχικών δεδομένων. Έτσι, τα *biplots* μπορούν να συνδυαστούν με αρκετές από τις μεθόδους της Ανάλυσης Δεδομένων όπως είναι, για παράδειγμα, η Ανάλυση σε Κύριες Συνιστώσες, η Πολυδιάστατη Κλιμάκωση και η Κανονικοποιημένη Συσχέτιση (Lipkovich & Smith 2002, Udina 2005). Η εφαρμογή των *biplots* στην οπτικοποίηση των αποτελεσμάτων της ΠΑΑ αποτελεί και τη σημαντικότερη διαφοροποίηση της Γαλλικής μεθοδολογικής προσέγγισης από αυτήν της Ολλανδικής.

2.2.14 Ο Αλγόριθμος της Παραγοντικής Ανάλυσης των Αντιστοιχιών

Ο βασικός αλγόριθμος της ΠΑΑ, όπως εφαρμόζεται υπολογιστικά στα στατιστικά πακέτα (SAS Institute 1990, SPSS Inc. 1997, SPSS Inc. 2004α), μπορεί να συνοψιστεί σε τέσσερα βήματα.

Πρώτο Βήμα

Στο πρώτο βήμα δημιουργείται ο βοηθητικός πίνακας **S** με στοιχεία τα σχετικά τυποποιημένα υπόλοιπα (βλέπε Ενότητα 2.2.5). Το γενικό στοιχείο του **S** δίνεται από την παρακάτω σχέση (Benzécri 1992, Greenacre 1993α, Παπαδημητρίου 2006, 2004 και 1994):

$$s_{ij} = \frac{\left(p_{ij} - r_i c_j\right)}{\sqrt{r_i c_j}},\ i = 1,\ldots, k \text{ και } j = 1,\ldots, l.$$

Με μορφή πινάκων ο **S** γράφεται:

$$\mathbf{S} = \mathbf{D_r^{-1/2}}\left(\mathbf{P - rc^T}\right)\mathbf{D_c^{-1/2}}.$$

Δεύτερο Βήμα

Το δεύτερο βήμα περιλαμβάνει την ανάλυση της βασικής δομής του πίνακα **S** με τη μέθοδο SVD (Israëls 1987, SAS Institute 1990, Weller & Romney 1990, Andersen 1991, Greenacre 1993α και 1984, Blasius & Greenacre 1994, SPSS Inc. 1997, βλέπε και Ενότητα 2.2.12):

$\mathbf{S}=\mathbf{UDV^T}$, με τους περιορισμούς: $\mathbf{U^TU}=\mathbf{I}$, $\mathbf{V^TV}=\mathbf{I}$ και **D** διαγώνιος.

Ειδικότερα, ο **U** είναι ο πίνακας του οποίου οι στήλες είναι τα αριστερά ορθοκανονικά χαρακτηριστικά διανύσματα του **S**, **D** είναι ο διαγώνιος πίνακας με τις μη αρνητικές χαρακτηριστικές τιμές $d_1, d_2, \ldots$, σε φθίνουσα σειρά κατά μήκος της κύριας διαγωνίου και $\mathbf{V^T}$ είναι ο ανάστροφος του πίνακα **V** του οποίου οι στήλες είναι τα δεξιά ορθοκανονικά χαρακτηριστικά διανύσματα του **S**. Οι χαρακτηριστικές τιμές d_i αποκαλούνται συχνά και «συντελεστές κανονικοποιημένης συσχέτισης» (*canonical correlations*) (Van de Geer 1993β, Nishisato 1994 και 1980, Gifi 1996). Τα τετράγωνα των χαρακτηριστικών τιμών (d_i^2) εκφράζουν τις αδράνειες των αντίστοιχων παραγοντικών αξόνων.

Ένας άλλος τρόπος διαγωνοποίησης του πίνακα **S** είναι ο παρακάτω (Greenacre 1993α, Μαυρομάτης 1999, Παπαδημητρίου 2006, 2004 και 1994):

Κατασκευάζεται ο τετραγωνικός και συμμετρικός πίνακα $\mathbf{S^TS}$ (ή ο $\mathbf{SS^T}$) (βλέπε και Ενότητα 2.2.12), ο οποίος ονομάζεται πίνακας «διακύμανσης-συνδυακύμανσης», και υπολογίζονται οι ιδιοτιμές $\lambda_1, \lambda_2, \ldots$, και τα αντίστοιχα ιδιοδιανύσματα $\mathbf{v}_1, \mathbf{v}_2,\ldots,$. Οι ιδιοτιμές είναι ίσες με τις αδράνειες των αντίστοιχων αξόνων από την ανάλυση του **S** και ικανοποιούν τις σχέσεις (Benzécri 1992, Καραπιστόλης 1999, Van de Velden & Neudecker 2000):

i) $0 \le \lambda_s \le 1$

ii) $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_p$

Οι χαρακτηριστικές τιμές της προηγούμενης προσέγγισης είναι ίσες με τις τετραγωνικές ρίζες των ιδιοτιμών $d_1=\sqrt{\lambda_1}$, $d_2=\sqrt{\lambda_2}$,…,. Τα δεξιά χαρακτηριστικά διανύσματα είναι οι στήλες (διανύσματα) $\mathbf{v}_1$, $\mathbf{v}_2$,…, του πίνακα **V** και τα αριστερά χαρακτηριστικά διανύσματα προκύπτουν από τον παρακάτω γραμμικό μετασχηματισμό του πίνακα **V** (Greenacre, 1993α):

$$\mathbf{U}=\mathbf{SVD}^{-1}.$$

Στην περίπτωση αυτή, το γενικό στοιχείο του πίνακα **U** δίνεται από τη σχέση:

$$u_{is}=\frac{1}{d_s}\sum_j s_{ij}v_{js} = \frac{1}{\sqrt{\lambda_s}}\sum_j s_{ij}v_{js} ,$$

όπου v_{js} είναι το γενικό στοιχείο του πίνακα **V** και s δείκτης για τη δήλωση των παραγοντικών αξόνων με s=1,…, p, όπου $p=min\{k\text{-}1,l\text{-}1\}$ είναι ο μέγιστος αριθμός διαστάσεων του υποχώρου στον οποίο μπορεί να προβληθεί το υπό εξέταση φαινόμενο χωρίς απώλεια πληροφορίας.

Στην περίπτωση του πίνακα $\mathbf{S^TS}$ ισχύει (Greenacre 1993α και 1984, Blasius & Greenacre 1994):

$$trace(\mathbf{S^TS}) = \sum_i \sum_j s_{ij}^2 = \sum_{s=1}^{p} \lambda_s .$$

Από τη σχέση [2.5] έχουμε ότι $I = \sum_i \sum_j s_{ij}^2$.

Συνεπώς, η ολική αδράνεια I του πίνακα συμπτώσεων **F** δίνεται και από τη σχέση:

$$I = trace(\mathbf{S^TS}) = \sum_{s=1}^{p} \lambda_s .$$

Από την παραπάνω σχέση φαίνεται ότι η ανάλυση της βασικής δομής του πίνακα $\mathbf{S^TS}$ ή του πίνακα **S** έχει ως αποτέλεσμα τη διάσπαση της ολικής αδράνειας του πίνακα **F** σε p μέρη. Αν θεωρήσουμε την ολική αδράνεια του νέφους των σημείων γραμμών ή

στηλών ως ένα μέτρο γενικευμένης διασποράς – απόκλισης από το κέντρο βάρους του νέφους, δηλαδή την κατάσταση ανεξαρτησίας των δύο κατηγορικών μεταβλητών, τότε διαισθητικά μπορούμε να πούμε ότι η ολική αδράνεια διασπάται και διαχέεται σε διαφορετική έκταση (εύρος) κατά μήκος των παραγοντικών αξόνων.

Παρατήρηση 2.1

Εκτός από τους πίνακες $\mathbf{S}$ και $\mathbf{S^TS}$ μπορούν να χρησιμοποιηθούν ως είσοδος στον αλγόριθμο της ΠΑΑ και οι πίνακες $\mathbf{P}$ και $\mathbf{P-rc^T}$ (Greenacre 1984, SAS Institute 1990, Jackson 1991) καθώς και ο πίνακας $\mathbf{D_r^{-1/2}PD_c^{-1/2}}$ (Weller & Romney 1990, Gifi 1996, Le Roux & Rouanet 2004). Μια άλλη μεθοδολογική προσέγγιση είναι η διαγωνοποίηση των πινάκων $\mathbf{P^TD_r^{-1}PD_c^{-1}}$ και $\mathbf{PD_r^{-1}P^TD_c^{-1}}$ (Lebart, Morineau & Warwick 1984, Lebart, Morineau & Piron 2000) ή των πινάκων $\mathbf{D_r^{1/2}\left(D_r^{-1}P-1c^T\right)D_c^{-1/2}}$ και $\mathbf{D_c^{1/2}\left(D_c^{-1}P^T-1r^T\right)D_r^{-1/2}}$ (Greenacre 1984, Blasius & Greenacre 1994) για την επίλυση του προβλήματος των γραμμών και στηλών αντίστοιχα.

Αν υποθέσουμε, χωρίς περιορισμό της γενικότητας, ότι $k>l$, τότε κατά την εφαρμογή της SVD στον πίνακα $\mathbf{S}$, ο πίνακας $\mathbf{D}$ είναι διαστάσεων $l\times l$ (βλέπε Ενότητα 2.2.12). Η l-οστή σε τάξη χαρακτηριστική τιμή είναι πάντα ίση με μηδέν (τετριμμένη χαρακτηριστική τιμή) και απομακρύνεται από την ανάλυση μαζί με τα αντίστοιχα αριστερά και δεξιά χαρακτηριστικά διανύσματα (Greenacre 1993α, Μαυρομάτης 1999). Η εφαρμογή της «Γενικευμένης Διάσπασης σε Χαρακτηριστικές Τιμές[5]» (*Generalized SVD-GSVD*) (βλέπε Greenacre 1984, SAS Institute 1990, Gifi 1996) στον πίνακα $\mathbf{P-rc^T}$, αφού απομακρυνθούν η τελευταία σε τάξη χαρακτηριστική τιμή, η οποία είναι ίση με μηδέν, και τα αντίστοιχα αριστερά και δεξιά χαρακτηριστικά διανύσματα, αναδεικνύει την ίδια βασική δομή με αυτή που προκύπτει από την εφαρμογή της GSVD στους πίνακες $\mathbf{P}$ και $\mathbf{D_r^{-1/2}PD_c^{-1/2}}$, αφού απομακρυνθούν η πρώτη σε τάξη χαρακτηριστική τιμή, η οποία είναι ίση με τη μονάδα, και τα αντίστοιχα

---

[5] Η GSVD, για παράδειγμα του πίνακα $\mathbf{P}$, ορίζεται ως εξής: GSVD($\mathbf{P}$)=$\mathbf{UDV^T}$, με $\mathbf{U^TD_r^{-1}U = V^TD_c^{-1}V = I}$.

αριστερά και δεξιά χαρακτηριστικά διανύσματα (SAS Institute 1990, Weller & Romney 1990, Jackson 1991). Η χαρακτηριστική τιμή d_0=1 και συνεπώς η αντίστοιχη ιδιοτιμή $\lambda_0 = d_0^2 = 1$ αντιστοιχεί στην κατάσταση ανεξαρτησίας των μεταβλητών *X* και *Y* (Weller & Romney 1990, Greenacre 1993β και 1984) και συνδέεται με τον πρώτο παραγοντικό άξονα (τετριμμένος άξονας ή άξονας μηδέν), ο οποίος δεν παρουσιάζει, προς το παρόν, ενδιαφέρον. Η ανάλυση της βασικής δομής των πινάκων **P**, $\mathbf{D_r^{-1/2}PD_c^{-1/2}}$ και $\mathbf{P - rc^T}$ είναι εννοιολογικά προτιμότερη, όταν ο αρχικός πίνακας **F** δεν περιγράφει την κοινή κατανομή συχνοτήτων δύο τυχαίων κατηγορικών μεταβλητών, αλλά είναι εν γένει ένας ομοιογενής πίνακας της μορφής «*αντικείμενα* × *μεταβλητές*» με μη αρνητικά στοιχεία και μη μηδενικά αθροίσματα γραμμών και στηλών. Στην περίπτωση αυτή, η ανάλυση του πίνακα **S** των σχετικών τυποποιημένων υπολοίπων κάτω από την υπόθεση της ανεξαρτησίας των γραμμών και στηλών του πίνακα στερείται εννοιολογικού περιεχομένου.

Τρίτο Βήμα

Στο βήμα αυτό, τα αριστερά και τα δεξιά χαρακτηριστικά διανύσματα του πίνακα **S** σταθμίζονται έτσι ώστε οι συντεταγμένες της προβολής των γραμμών και των στηλών πάνω στον άξονα, έστω, *s*, να τυποποιηθούν ως προς τις μάζες των γραμμών και των στηλών αντίστοιχα του πίνακα αντιστοιχιών **P.** Με τον τρόπο αυτό υπολογίζονται οι «τυποποιημένες συντεταγμένες» (*standardized coordinates*) (SPSS Inc. 1997, SAS Institute 1990, Greenacre 1993α, Blasius & Greenacre 1994, Bendixen 2003) των προβολών των σημείων γραμμών και στηλών πάνω στους παραγοντικούς άξονες. Πιο συγκεκριμένα, οι τυποποιημένες συντεταγμένες των γραμμών και στηλών δίνονται από τις παρακάτω σχέσεις:

$$r_{is} = \frac{u_{is}}{\sqrt{\frac{f_{i+}}{N}}} = \frac{u_{is}}{\sqrt{r_i}}, \qquad [2.14]$$

και

$$c_{js} = \frac{v_{js}}{\sqrt{\frac{f_{+j}}{N}}} = \frac{v_{js}}{\sqrt{c_j}}. \qquad [2.15]$$

Μέχρι και το τρίτο βήμα ο αλγόριθμος της Παραγοντικής Ανάλυσης των Αντιστοιχιών εφαρμόζεται με τον ίδιο τρόπο τόσο στο σύστημα *GIFI* όσο και στη Γαλλική Σχολή. Η διαφοροποίηση ξεκινά από το τέταρτο βήμα.

Τέταρτο Βήμα

Στο πλαίσιο της Γαλλικής Σχολής, οι τυποποιημένες συντεταγμένες κανονικοποιούνται ως προς την απόσταση χ^2 και με αυτό τον τρόπο υπολογίζονται οι «κύριες συντεταγμένες» (*principal coordinates*) των προβολών των σημείων (προφίλ) γραμμών και στηλών πάνω στους παραγοντικούς άξονες (Benzécri 1992, SAS Institute 1990, Greenacre 1993α και 1984, Blasius & Greenacre 1994, Bendixen 2003, SPSS Inc. 2004α και 1997). Πιο συγκεκριμένα για τον παραγοντικό άξονα, έστω, *s*, ισχύουν οι παρακάτω σχέσεις:

$$\varphi_{is} = d_s r_{is} = \sqrt{\lambda_s}\ r_{is} \ \text{με} \ \sum_i r_i \varphi_{is} = 0, \qquad [2.16]$$

και

$$\gamma_{js} = d_s c_{js} = \sqrt{\lambda_s}\ c_{js} \ \text{με} \ \sum_j c_j \gamma_{js} = 0. \qquad [2.17]$$

Οι ισότητες $\sum_i r_i \varphi_{is} = 0$ και $\sum_j c_j \gamma_{js} = 0$ προκύπτουν άμεσα από το γεγονός ότι οι κύριες συντεταγμένες των γραμμών (στηλών) επί των παραγοντικών αξόνων δηλώνουν τη θέση των αντίστοιχων προφίλ ως προς το κέντρο βάρους τους, δηλαδή το σταθμισμένο μέσο όρο τους, ο οποίος συμπίπτει με την αρχή του νέου συστήματος συντεταγμένων που ορίζουν οι παραγοντικοί άξονες (Greenacre, 1984). Είναι φανερό ότι επιπλεόν ισχύει $\sum_i r_i r_{is} = 0$ και $\sum_j c_j c_{js} = 0$.

Μπορεί να δειχθεί ότι (βλέπε Greenacre 1984, Benzécri 1992, Lebart, Morineau & Piron 2000, Le Roux & Rouanet 2004, Murtagh 2005):

$$\varphi_{is} = \frac{1}{d_s} \sum_j \frac{p_{ij}}{r_i} \gamma_{js} = \frac{1}{\sqrt{\lambda_s}} \sum_j \frac{p_{ij}}{r_i} \gamma_{js}, \qquad [2.18]$$

και

$$\gamma_{js} = \frac{1}{d_s}\sum_i \frac{p_{ij}}{c_j}\varphi_{is} = \frac{1}{\sqrt{\lambda_s}}\sum_i \frac{p_{ij}}{c_j}\varphi_{is}. \qquad [2.19]$$

Οι σχέσεις [2.18] και [2.19] ονομάζονται «σχέσεις μετάβασης» (Παπαδημητρίου, 2006, 2004 και 1994) ή αλλιώς «εξισώσεις μεταφοράς» (Μεϊμάρης, 2002) και παίζουν σημαντικό ρόλο τόσο στους αριθμητικούς υπολογισμούς όσο και στην ερμηνεία των παραγοντικών αξόνων και επιπέδων (Murtagh, 2005). Οι σχέσεις αυτές δηλώνουν ότι μπορούμε να μεταβούμε εύκολα από το χώρο γραμμών στο χώρο στηλών και αντίστροφα. Έτσι, από τη στιγμή που επιλυθεί το Πρόβλημα των Γραμμών επιλύεται αυτόματα και το δυϊκό του, δηλαδή το Πρόβλημα των Στηλών. Η σχέση [2.18] εκφράζει ότι κατά προσέγγιση ενός συντελεστή $\left(\sqrt{\lambda_s}\right)^{-1}$ κάθε σημείο γραμμής προβάλλεται στο κέντρο βάρους (σταθμισμένο μέσο όρο) των σημείων στηλών. Από τη σχέση [2.19] προκύπτει ότι κατά προσέγγιση ενός συντελεστή $\left(\sqrt{\lambda_s}\right)^{-1}$ κάθε σημείο στήλης προβάλλεται στο κέντρο βάρους των σημείων γραμμών (Benzécri 1992, Μπεχράκης 1999, Le Roux & Rouanet 2004, Murtagh 2005). Οι «βαρυκεντρικές» αυτές σχέσεις επιτρέπουν την ταυτόχρονη προβολή και ερμηνεία των σημείων γραμμών και στηλών σε ένα κοινό παραγοντικό διάγραμμα. Μέσω της ΠΑΑ οι τιμές των συντεταγμένων φ_{is} και γ_{js} υπολογίζονται με τέτοιο τρόπο ώστε τα υποδείγματα [2.18] και [2.19] να αντιστοιχούν σε ευθείες παλινδρόμησης. Μπορεί να δειχθεί (Hirschfeld 1935, Fisher 1941, Guttman 1941, Nishisato 1980, Greenacre 1984, Gifi 1996, Van Rijckevorsel 1987) ότι επί του πρώτου παραγοντικού άξονα μεγιστοποιείται η (κανονικοποιημένη) συσχέτιση μεταξύ των συντεταγμένων των γραμμών και στηλών του πίνακα συμπτώσεων και ο συντελεστής συσχέτισής τους είναι ίσος με τη χαρακτηριστική τιμή $d_1 = \sqrt{\lambda_1}$ του άξονα. Στο πλαίσιο της Ολλανδικής Σχολής οι σχέσεις μετάβασης εκφράζουν την «Αρχή των Αντιστρόφων Μέσων» (*Principal of Reciprocal Averaging*) (Van de Geer, 1993β). Στην αρχή αυτή στηρίζεται ο ομώνυμος υπολογιστικός επαναληπτικός αλγόριθμος (Hill 1974 και 1973, Nishisato 1980, Greenacre 1984, Gifi 1996) που χρησιμοποιείται σε πολλές παραλλαγές της ΠΑΑ όπως είναι η Δυική Κλιμάκωση (*Dual Scaling*) (Nishisato 1980).

Η σχέση [2.18] αν λάβουμε υπόψη τη [2.17] γράφεται:

$$\varphi_{is} = \frac{1}{d_s}\sum_j \frac{p_{ij}}{r_i}\gamma_{js} = \frac{1}{\sqrt{\lambda_s}}\sum_j \frac{p_{ij}}{r_i}\gamma_{js} = \frac{1}{\sqrt{\lambda_s}}\sum_j \frac{p_{ij}}{r_i}\sqrt{\lambda_s}\ c_{js} = \sum_j \frac{p_{ij}}{r_i}c_{js}. \qquad [2.20]$$

Η σχέση [2.19] αν λάβουμε υπόψη τη [2.16] γράφεται:

$$\gamma_{js} = \frac{1}{d_s}\sum_i \frac{p_{ij}}{c_j}\varphi_{is} = \frac{1}{\sqrt{\lambda_s}}\sum_i \frac{p_{ij}}{c_j}\varphi_{is} = \frac{1}{\sqrt{\lambda_s}}\sum_i \frac{p_{ij}}{c_j}\sqrt{\lambda_s}\ r_{is} = \sum_i \frac{p_{ij}}{c_j}r_{is}. \qquad [2.21]$$

Η σχέση [2.20] εκφράζει ότι η κύρια συντεταγμένη της γραμμής *i* επί του άξονα *s* είναι ο σταθμισμένος μέσος όρος των τυποποιημένων συντεταγμένων των στηλών επί του *s*. Οι όροι p_{ij}/r_i δεν είναι παρά τα στοιχεία του διανύσματος του προφίλ της *i* γραμμής. Σύμφωνα με τη σχέση [2.21], η κύρια συντεταγμένη της στήλης *j* επί του άξονα *s* είναι ο σταθμισμένος μέσος όρος των τυποποιημένων συντεταγμένων των γραμμών πάνω στον ίδιο άξονα. Ανάλογα, οι όροι p_{ij}/c_j είναι στοιχεία του προφίλ της *j* στήλης. Όπως θα διαπιστωθεί στα επόμενα (βλέπε Ενότητα 2.2.14.2), οι σχέσεις [2.20] και [2.21] είναι σημαντικές για την ερμηνεία των γραφικών αποτελεσμάτων που παράγονται κατά την εφαρμογή της ΠΑΑ.

Στο σημείο αυτό, θα πρέπει να τονιστεί ότι η λύση του προβλήματος της ΠΑΑ μέσω των σχέσεων [2.16] και [2.17] οδηγεί σε μια ποιοτική μετάβαση από τη χ^2 απόσταση στην ευκλείδεια (Andersen 1991, Jackson 1991, Benzécri 1992, Greenacre 1993α και 1984, Blasius & Greenacre 1994, Gifi 1996, Clausen 1998, Murtagh 2005, Le Roux & Rouanet 2004, Meulman & Heiser 2004). Στο βέλτιστο υποχώρο προβολής οι χ^2 αποστάσεις ($d^2_{\chi^2}$) των προφίλ των σημείων γραμμών και στηλών προσεγγίζονται από ευκλείδειες (d^2_e). Συνεπώς, στον *p*-διάστατο υποχώρο προβολής με *p=min*{*k*-1, *l*-1} οι ευκλείδειες αποστάσεις μεταξύ των σημείων είναι ίσες με τις αντίστοιχες χ^2 στον αρχικό χώρο. Έτσι, με βάση τις σχέσεις [2.16], [2.17] και όσα αναφέρθηκαν στις Ενότητες 2.1.2.4, 2.1.2.5 και 2.1.2.8, έχουμε τα παρακάτω αποτελέσματα (βλέπε Andersen 1991, Benzécri 1992, Greenacre 1993α και 1984, Blasius & Greenacre 1994, Le Roux & Rouanet 2004, Murtagh 2005):

για την απόσταση της *i* γραμμής από το κέντρο βάρους:

$$d^2_{\chi^2}\left(i, g_r\right) = \sum_{s=1}^{p} \varphi^2_{is} = d^2_e(i, g_r),$$

για την απόσταση της *j* στήλης από το κέντρο βάρους:

$$d^2_{\chi^2}\left(j, g_c\right) = \sum_{s=1}^{p} \gamma^2_{js} = d^2_e(j, g_c),$$

για την απόσταση μεταξύ δύο σημείων (προφίλ) γραμμών:

$$d^2_{\chi^2}\left(i, i'\right) = \sum_{s=1}^{p} \left(\varphi_{is} - \varphi_{i's}\right)^2 = d^2_{e^2}\left(i, i'\right),$$

για την απόσταση μεταξύ δύο σημείων (προφίλ) στηλών:

$$d^2_{\chi^2}\left(j, j'\right) = \sum_{s=1}^{p} \left(\gamma_{is} - \gamma_{i's}\right)^2 = d^2_{e^2}\left(j, j'\right),$$

για την ολική αδράνεια της *i* γραμμής:

$$I_i = \left(\mu\acute{\alpha}\zeta\alpha \ i \ \gamma\rho\alpha\mu\mu\acute{\eta}\varsigma\right) \times d^2_{\chi^2}\left(i, g_r\right) = \sum_{s=1}^{p} r_i \varphi^2_{is},$$

για την ολική αδράνεια της *j* στήλης:

$$I_j = \left(\mu\acute{\alpha}\zeta\alpha \ j \ \sigma\tau\acute{\eta}\lambda\eta\varsigma\right) \times d^2_{\chi^2}\left(j, g_C\right) = \sum_{s=1}^{p} c_j \gamma^2_{js},$$

για τη μερική αδράνεια της *i* γραμμής επί του άξονα *s*:

$$I_{i(s)} = (\mu\acute{\alpha}\zeta\alpha \times \alpha\pi\acute{o}\sigma\tau\alpha\sigma\eta \ \alpha\pi\acute{o} \ \tau\eta\nu \ \alpha\rho\chi\acute{\eta} \ \tau o\upsilon \ s \ \varepsilon\iota\varsigma \ \tau o \ \tau\varepsilon\tau\rho\acute{\alpha}\gamma\omega\nu o) = r_i \varphi^2_{is},$$

για τη μερική αδράνεια της *j* στήλης επί του άξονα *s*:

$$I_{j(s)} = (\mu\acute{\alpha}\zeta\alpha \times \alpha\pi\acute{o}\sigma\tau\alpha\sigma\eta \ \alpha\pi\acute{o} \ \tau\eta\nu \ \alpha\rho\chi\acute{\eta} \ \tau o\upsilon \ s \ \varepsilon\iota\varsigma \ \tau o \ \tau\varepsilon\tau\rho\acute{\alpha}\gamma\omega\nu o) = c_j \gamma^2_{is},$$

για την αδράνεια του άξονα *s*:

$$\lambda_s = \sum_{i=1}^{k} r_i \varphi^2_{is} = \sum_{j=1}^{l} c_j \gamma^2_{js},$$

για την ολική αδράνεια του νέφους των σημείων γραμμών (στηλών):

$$I = I_c = I_r = \sum_{i=1}^{k}\sum_{s=1}^{p} r_i \varphi_{is}^2 = \sum_{j=1}^{l}\sum_{s=1}^{p} c_j \gamma_{js}^2 .$$

Τέλος, υπολογίζονται τα παρακάτω βασικά αριθμητικά αποτελέσματα (Greenacre 1993α και 1984, Blasius & Greenacre 1994, Καραπιστόλης 1996, SAS Institute 1990, SPSS Inc. 1997, Murtagh 2005, Παπαδημητρίου 2006, 2004 και 1994):

α) Η αδράνεια του κάθε άξονα:

$$d_s^2 = \lambda_s \text{ , } s=1,\ldots, p \text{ με } p=min\{k\text{-}1,\ l\text{-}1\}.$$

Ισχύει, επίσης,

$$\lambda_s = \sum_i r_i \varphi_{is}^2 = \sum_j c_j \gamma_{js}^2 .$$

β) Η ολική αδράνεια I του νέφους των σημείων (γραμμών, στηλών):

$$I = \sum_{s=1}^{p} \lambda_s = \frac{Q}{N} = trace(\mathbf{S^T S}) = \sum_i \sum_j s_{ij}^2 .$$

γ) Το ποσοστό % της ολικής αδράνειας που ερμηνεύει ο κάθε άξονας:

$$\frac{\lambda_s}{I} \times 100 .$$

δ) Το αθροιστικό % της ολικής αδράνειας που ερμηνεύουν οι κ πρώτοι παραγοντικοί άξονες:

$$\frac{\sum_{s=1}^{\kappa} \lambda_s}{I} \times 100 \text{ με } \kappa=1,\ldots, p.$$

δ) Οι ολικές αδράνειες των γραμμών και των στηλών:

Ολική αδράνεια της i γραμμής, $I_i = \sum_{j=1}^{l} s_{ij}^2 = \sum_{s=1}^{p} r_i \varphi_{is}^2 .$

Ολική αδράνεια της j στήλης, $I_j = \sum_{i=1}^{k} s_{ij}^2 = \sum_{s=1}^{p} c_j \gamma_{js}^2 .$

ε) Οι μερικές αδράνειες των γραμμών και στηλών επί των παραγοντικών αξόνων:

Μερική αδράνεια της γραμμής *i* επί του άξονα *s*, $I_{i(s)} = r_i \varphi_{is}^2$.

Μερική αδράνεια της στήλης *j* επί του άξονα *s*, $I_{j(s)} = c_j \gamma_{is}^2$.

στ) Οι συνεισφορές (*CTR*) των σημείων γραμμών και στηλών στην αδράνεια των αξόνων (βλέπε Ενότητα 2.2.8):

Συνεισφορά της γραμμής *i* στον άξονα *s*,

$$CTR(i,s) = \frac{I_{i(s)}}{\lambda_s} = r_i \frac{\varphi_{is}^2}{\lambda_s} \text{ με } \sum_{i=1}^{k} CTR(i,s) = 1.$$

Συνεισφορά της στήλης *j* στον άξονα *s*,

$$CTR(j,s) = \frac{I_{j(s)}}{\lambda_s} = c_j \frac{\gamma_{js}^2}{\lambda_s} \text{ με } \sum_{j=1}^{l} CTR(j,s) = 1.$$

ζ) Οι συνεισφορές των κύριων αξόνων (*COR*) στην αδράνεια των σημείων γραμμών και στηλών (βλέπε Ενότητα 2.2.8):

Συνεισφορά του άξονα *s* στη γραμμή *i*:

$$COR(s,i) = \frac{I_{i(s)}}{I_i} = r_i \frac{\varphi_{is}^2}{I_i} = \frac{r_i \varphi_{is}^2}{\sum_{t=1}^{p} r_i \varphi_{it}^2} = \frac{r_i \varphi_{is}^2}{r_i \sum_{t=1}^{p} \varphi_{it}^2} = \frac{\varphi_{is}^2}{\sum_{t=1}^{p} \varphi_{it}^2} = cos^2\,\theta, \qquad [2.22]$$

όπου θ είναι η γωνία που σχηματίζει ο άξονας *s* με το διάνυσμα που ενώνει το σημείο γραμμής *i* με την αρχή του συστήματος συντεταγμένων (κέντρο βάρους). Επιπλέον, ισχύει: $\forall i, \sum_{s=1}^{p} COR(s,i) = 1.$

Συνεισφορά του άξονα *s* στη στήλη *j*:

$$COR(s,j) = \frac{I_{j(s)}}{I_j} = c_j \frac{\gamma_{js}^2}{I_j} = \frac{c_j \gamma_{js}^2}{\sum_{t=1}^{p} c_j \gamma_{jt}^2} = \frac{c_j \gamma_{js}^2}{c_j \sum_{t=1}^{p} \gamma_{jt}^2} = \frac{\gamma_{js}^2}{\sum_{t=1}^{p} \gamma_{jt}^2} = cos^2\,\phi, \qquad [2.23]$$

όπου ϕ είναι η γωνία που σχηματίζει ο άξονας s με το διάνυσμα που ενώνει το σημείο στήλης j με το κέντρο βάρους. Επιπλέον ισχύει: $\forall j, \sum_{s=1}^{p} COR(s, j) = 1.$

η) Η ποιότητα απεικόνισης των σημείων γραμμών και στηλών σε μια λύση με κ διαστάσεις, δηλαδή πάνω στους κ πρώτους κύριους άξονες (βλέπε Ενότητα 2.2.8):

$$\text{Ποιότητα της γραμμής } i,\ QLT(i) = \sum_{s=1}^{\kappa} COR(s, i),$$

$$\text{Ποιότητα της στήλης } j,\ QLT(j) = \sum_{s=1}^{\kappa} COR(s, j).$$

Στην περίπτωση συμπληρωματικών σημείων (γραμμών, στηλών) τα προφίλ και οι μάζες τους υπολογίζονται όπως και στα ενεργά (βλέπε Ενότητα 2.2.2.). Οι κύριες συντεταγμένες της προβολής μιας συμπληρωματικής γραμμής i^{sup} πάνω στον άξονα, έστω, s, υπολογίζονται απ' τη σχέση:

$$\varphi_{is}^{sup} = \sum_{j} \frac{f_{ij}^{sup}}{f_{i+}^{sup}} c_{js},$$

όπου c_{js} είναι οι τυποποιημένες συντεταγμένες των ενεργών στηλών επί του άξονα s.

Παρόμοια, οι κύριες συντεταγμένες της προβολής μιας συμπληρωματικής στήλης j^{sup} πάνω στον άξονα s, δίνονται από τη σχέση:

$$\gamma_{js}^{sup} = \sum_{i} \frac{f_{ij}^{sup}}{f_{+j}^{sup}} r_{is},$$

όπου r_{is} είναι οι τυποποιημένες συντεταγμένες των ενεργών γραμμών επί του άξονα s.

Η συμβολή ενός άξονα, έστω s, στην αδράνεια ενός συμπληρωματικού σημείου γραμμής υπολογίζεται από τη σχέση:

$$COR(s, sup\ i) = \frac{r_i^{sup}\ (\varphi_{is}^{sup})^2}{I_i^{sup}},$$

όπου $r_i^{sup} = \sum_j p_{ij}^{sup}$, η μάζα της συμπληρωματικής *i* γραμμής.

Αντίστοιχα, ο δείκτης *COR* για μια συμπληρωματική στήλη δίνεται από τη σχέση:

$$COR(s, sup\ j) = \frac{c_j^{sup}\ (\gamma_{js}^{sup})^2}{I_j^{sup}},$$

όπου $c_j^{sup} = \sum_i p_{ij}^{sup}$, η μάζα της συμπληρωματικής *j* στήλης.

Επιπλέον, στον υποχώρο που ορίζουν έστω οι *κ* πρώτοι παραγοντικοί άξονες, οι επιμέρους δείκτες *COR* μπορούν να προστεθούν για να ορίσουν το μέτρο ποιότητας *QLT* της προβολής κάθε σημείου συμπληρωματικής γραμμής στον υποχώρο αυτό:

$$QLT(i\ sup) = \frac{\sum_{s=1}^{\kappa} r_i^{sup} (\varphi_{is}^{sup})^2}{I_i^{sup}}.$$

Παρόμοια για ένα συμπληρωματικό σημείο στήλης έχουμε:

$$QLT(j\ sup) = \frac{\sum_{s=1}^{\kappa} c_j^{sup} (\gamma_{js}^{sup})^2}{I_j^{sup}}.$$

Στο πλαίσιο της Γαλλικής Σχολής Ανάλυσης Δεδομένων εφαρμόζεται μία μόνο μέθοδος υπολογισμού των παραγοντικών συντεταγμένων. Η βασική απαίτηση είναι οι χ^2 αποστάσεις μεταξύ των προφίλ γραμμών (στηλών) του πίνακα **F** να προσεγγίζονται από ευκλείδειες. Στην Ολλανδική Σχολή, εκτός από τη χ^2, ανάλογα με τον πίνακα δεδομένων που θα δοθεί ως είσοδος στην ανάλυση, μπορεί να χρησιμοποιηθεί και η ευκλείδεια απόσταση, μετά από κατάλληλη κεντροποίηση (*centering*) και στάθμιση των δεδομένων (βλέπε SPSS Inc., 2004α, 1998 και 1997). Επίσης, στους χρήστες της ΠΑΑ δίνεται η δυνατότητα να επιλέξουν ανάμεσα σε τέσσερις βασικές μεθόδους κανονικοποίησης των συντεταγμένων των προβολών των σημείων γραμμών και στηλών επί των παραγοντικών αξόνων.

2.2.14.1 Κανονικοποίηση των Συντεταγμένων των Προβολών των Σημείων πάνω στους Παραγοντικούς Άξονες

Στο σύστημα *GIFI* μια σημαντική επιλογή του χρήστη της μεθόδου είναι η μέθοδος κανονικοποίησης (*normalization*) των συντεταγμένων των προβολών των σημείων γραμμών και στηλών, που θα εφαρμοστεί (Gifi 1996, Meulman & Heiser 2004). Ο όρος κανονικοποίηση αναφέρεται στη μέθοδο "διάχυσης" ή ανακατανομής της αδράνειας πάνω στους παραγοντικούς άξονες. Κάποια από τα αποτελέσματα της ΠΑΑ δεν επηρεάζονται από τη μέθοδο κανονικοποίησης (όπως για παράδειγμα οι ιδιοτιμές των παραγοντικών αξόνων, η ολική αδράνεια και η αδράνεια που ερμηνεύει κάθε άξονας). Αυτά που επηρεάζονται κυρίως είναι οι συντεταγμένες των σημείων γραμμών και στηλών πάνω στους παραγοντικούς άξονες καθώς και οι διασπορές τους. Το αποτέλεσμα είναι ότι τα διάφορα παραγοντικά επίπεδα, ανάλογα με τη μέθοδο κανονικοποίησης, επιδέχονται διαφορετική ερμηνεία.

Για παράδειγμα, η διαδικασία της ΠΑΑ μέσω του SPSS διαθέτει τέσσερις βασικούς τρόπους κανονικοποίησης (SPSS Inc. 2004α, Meulman & Heiser 2004):

1. Την Κύρια Κανονικοποίηση κατά Γραμμές (*Row Principal Normalization-RPN*).
2. Την Κύρια Κανονικοποίηση κατά Στήλες (*Column Principal Normalization-CPN*).
3. Τη Συμμετρική Κανονικοποίηση των συντεταγμένων των προβολών των γραμμών και των στηλών (*Symmetrical ή Canonical Normalization-SN ή CN*).
4. Την Κύρια Κανονικοποίηση (*Principal Normalization-PN*), που, όπως θα δείξουμε στη συνέχεια, είναι η ίδια με τη μέθοδο υπολογισμού των κύριων συντεταγμένων φ_{is} και γ_{js} των προβολών των σημείων γραμμών και στηλών πάνω στους παραγοντικούς άξονες, σύμφωνα με την παράδοση της Γαλλικής Σχολής.

Διαισθητικά μπορούμε να πούμε ότι στην *RPN* η αδράνεια ανακατανέμεται μόνο πάνω στις συντεταγμένες των γραμμών, στην *CPN* μόνο πάνω στις συντεταγμένες των στηλών και στην *SN* συμμετρικά και ταυτόχρονα πάνω στις συντεταγμένες των γραμμών και των στηλών. Στην *PN* η αδράνεια ανακατανέμεται δύο φορές. Μία φορά πάνω στις συντεταγμένες των προβολών των σημείων γραμμών και μία φορά πάνω στις συντεταγμένες των στηλών.

Ειδικότερα, ανάλογα με τη μέθοδο κανονικοποίησης που θα χρησιμοποιηθεί, οι τυποποιημένες συντεταγμένες των προβολών των σημείων γραμμών και στηλών πάνω στους παραγοντικούς άξονες τροποποιούνται κατάλληλα, γεγονός που σημαίνει ότι επιβάλλεται μια αλλαγή κλίμακας στις συντεταγμένες (Greenacre, 1993α). Με την κανονικοποίηση επηρεάζεται η θέση των σημείων πάνω στους άξονες και, επομένως, διαφοροποιείται και η ερμηνεία των παραγοντικών επιπέδων. Στο πλαίσιο της Ολλανδικής Σχολής, το κριτήριο κανονικοποίησης καθορίζεται από την επιβολή περιορισμών ως προς την επιθυμητή τιμή του σταθμισμένου αθροίσματος τετραγώνων (αδράνειας, διασποράς) των συντεταγμένων των προβολών των σημείων γραμμών και στηλών επί των παραγοντικών αξόνων. Οι γενικές σχέσεις υπολογισμού του σταθμισμένου αθροίσματος τετραγώνων που προκύπτει από την αλλαγή κλίμακας των συντεταγμένων είναι οι εξής (Gifi 1996, SPSS Inc. 2004α):

για τις νέες κανονικοποιημένες συντεταγμένες r'_{is} των προβολών των γραμμών:

$$r'_{is} = r_{is} d_s^{\alpha} = r_{is} \left(\sqrt{\lambda_s}\right)^{\alpha}$$

και

$$\sum_i f_{i+} r'^2_{is} = N d_s{}^{2\alpha} = N \left(\sqrt{\lambda_s}\right)^{2\alpha} \Rightarrow \sum_i r_i r'^2_{is} = \left(\sqrt{\lambda_s}\right)^{2\alpha}, \qquad [2.24]$$

για τις νέες κανονικοποιημένες συντεταγμένες c'_{js} των προβολών των στηλών:

$$c'_{is} = c_{js} d_s^{\beta} = c_{js} \left(\sqrt{\lambda_s}\right)^{\beta}$$

και

$$\sum_j f_{+j} c'^2_{js} = N d_s{}^{2\beta} = N \left(\sqrt{\lambda_s}\right)^{2\beta} \Rightarrow \sum_j c_j c'^2_{js} = \left(\sqrt{\lambda_s}\right)^{2\beta}, \qquad [2.25]$$

όπου α=(1+q)/2, β=(1-q)/2 και η παράμετρος q μπορεί να επιλεγεί ελεύθερα στο διάστημα [-1,+1] σύμφωνα με την παρακάτω προτυποποίηση:

q = 0 για *SN*
q = 1 για *RPN*
q = -1 για *CPN*

Από τις σχέσεις [2.24] και [2.25] είναι φανερό ότι η διασπορά των συντεταγμένων των σημείων γραμμών και στηλών διαμορφώνεται ανάλογα με τη μέθοδο κανονικοποίησης που θα εφαρμοστεί.

Κατά τη *RPN* ισχύει ότι α=1 και β=0. Συνεπώς:

Από τη [2.24] έχουμε:

$$\sum_i f_{i+} r_{is}'^2 = N\lambda_s \Rightarrow \sum_i r_i r_{is}'^2 = \lambda_s . \qquad [2.26]$$

Από τη [2.25] προκύπτει:

$$\sum_j f_{+j} c_{js}'^2 = N \Rightarrow \sum_j c_j c_{js}'^2 = 1. \qquad [2.27]$$

Κατά την *CPN* ισχύει ότι α=0 και β=1. Επομένως:

Από τη [2.24] συνεπάγεται:

$$\sum_i f_{i+} r_{is}'^2 = N \Rightarrow \sum_i r_i r_{is}'^2 = 1. \qquad [2.28]$$

Από τη [2.25] προκύπτει:

$$\sum_j f_{+j} c_{js}'^2 = N\lambda_s \Rightarrow \sum_j c_j c_{js}'^2 = \lambda_s . \qquad [2.29]$$

Κατά τη *SN* έχουμε ότι $\alpha=\beta$=1/2. Άρα:

Από τη [2.24] συνεπάγεται: $\sum_i f_{i+} r_{is}'^2 = N\sqrt{\lambda_s} \Rightarrow \sum_i r_i r_{is}'^2 = \sqrt{\lambda_s}$.

Από τη [2.25] έχουμε: $\sum_j f_{+j} c_{js}'^2 = N\sqrt{\lambda_s} \Rightarrow \sum_j c_j c_{js}'^2 = \sqrt{\lambda_s}$.

Επίσης, για τη *SN* και για $\alpha=\beta$=1/2 έχουμε:

$$r_{is}' = r_{is}\lambda_s^{1/4} \qquad [2.30]$$

και

$$c_{is}' = c_{js}\lambda_s^{1/4}. \qquad [2.31]$$

Για την *PN* τα α και β τίθενται ίσα με 1 και το σταθμισμένο άθροισμα τετραγώνων για τα δύο σύνολα συντεταγμένων είναι:

Για τις γραμμές:

$$\sum_i f_{i+} r'^2_{is} = N d_s^2 = N\lambda_s . \qquad [2.32]$$

Για τις στήλες:

$$\sum_j f_{+j} c'^2_{js} = N d_s^2 = N\lambda_s . \qquad [2.33]$$

Από τη [2.32] προκύπτει:

$$\frac{1}{N}\sum_i f_{i+} r'^2_{is} = \sum_i r_i r'^2_{is} = \lambda_s . \qquad [2.32.1]$$

Από τη [2.33] έχουμε:

$$\frac{1}{N}\sum_j f_{+j} c'^2_{js} = \sum_j c_j c'^2_{js} = \lambda_s . \qquad [2.33.1]$$

Γενικά, αν συμβολίσουμε με **X** και **Y** τους πίνακες με στοιχεία τις συντεταγμένες των προβολών των σημείων γραμμών και στηλών αντίστοιχα, τότε με τη μορφή πινάκων οι τέσσερις μέθοδοι κανονικοποίησης δίνονται από τις παρακάτω σχέσεις (SAS Institute 1990, Jackson, 1991):

Για τη *RPN*:
$$\mathbf{X} = \mathbf{D_r^{-1/2}UD}, \quad \mathbf{Y} = \mathbf{D_c^{-1/2}V} \qquad [2.34]$$

Για την *CPN*:
$$\mathbf{X} = \mathbf{D_r^{-1/2}U}, \quad \mathbf{Y} = \mathbf{D_c^{-1/2}VD} \qquad [2.35]$$

Για την *PN*:
$$\mathbf{X} = \mathbf{D_r^{-1/2}UD}, \quad \mathbf{Y} = \mathbf{D_c^{-1/2}VD} \qquad [2.36]$$

Για τη *SN*:
$$\mathbf{X} = \mathbf{D_r^{-1/2}UD^{1/2}}, \quad \mathbf{Y} = \mathbf{D_c^{-1/2}VD^{1/2}} \qquad [2.37]$$

Με βάση όσα αναφέρθηκαν στα προηγούμενα, προκύπτει μια εύλογη απορία σχετικά με το ποια από τις μεθόδους κανονικοποίησης είναι κοινή στις δύο Σχολές Ανάλυσης

Δεδομένων, αφού σε κάθε Σχολή η κανονικοποίηση που εφαρμόζεται καθορίζεται από διαφορετικά κριτήρια. Για το λόγο αυτό, θα δείξουμε ότι:

Η μέθοδος κανονικοποίησης *PN* του συστήματος *GIFI* ταυτίζεται με τη μέθοδο που χρησιμοποιείται στη Γαλλική Σχολή για τον υπολογισμό των κύριων συντεταγμένων φ_{is} και γ_{js}.

Απόδειξη:

Από τη σχέση [2.16] έχουμε:

$$\varphi_{is} = d_s r_{is} = \sqrt{\lambda_s}\ r_{is}\,.$$

Από τη σχέση [2.14] προκύπτει:

$$r_{is} = \frac{u_{is}}{\sqrt{\frac{f_{i+}}{N}}} = \frac{u_{is}}{\sqrt{r_i}}\,.$$

Συνεπώς,

$$\varphi_{is} = \sqrt{\lambda_s}\,\frac{u_{is}}{\sqrt{r_i}} \Rightarrow \varphi_{is}\sqrt{r_i} = \sqrt{\lambda_s}\ u_{is} \Rightarrow\ \varphi_{is}^2 r_i = \lambda_s u_{is}^2\ \Rightarrow$$

$$\Rightarrow \sum_i \varphi_{is}^2 r_i = \lambda_s \sum_i u_{is}^2 \Rightarrow \sum_i \varphi_{is}^2 \frac{f_{i+}}{N} = \lambda_s .1 \Rightarrow$$

$$\Rightarrow \sum_i \varphi_{is}^2 f_{i+} = N\lambda_s, \qquad [2.38]$$

δηλαδή, ίδια με τη σχέση [2.32], αν λάβουμε υπόψη ότι τα διανύσματα $\mathbf{u_1}$, $\mathbf{u_2}$,…, είναι ορθοκανονικά και άρα $\sum_i u_{is}^2 = 1$. Με ανάλογο τρόπο μπορεί να δειχθεί ότι:

$$\sum_j \gamma_{js}^2 f_{+j} = N\lambda_s \qquad [2.39]$$

ίδια, δηλαδή, με τη σχέση [2.33]. □

Παρατήρηση 2.2

Στην Ολλανδική Σχολή, το γενικό πρόβλημα, στο οποίο καλείται να δώσει λύση η ΠΑΑ, μπορεί να διατυπωθεί και ως πρόβλημα βελτιστοποίησης (βλέπε Nishisato 1980, Van de Geer 1993α και 1993β, Gifi 1996, Greenacre 1998, SPSS Inc. 2004α):

Να βρεθούν τα σημεία $\{ r'_{is} \}$ που αντιστοιχούν στις γραμμές και τα σημεία $\{ c'_{js} \}$ που αντιστοιχούν στις στήλες του πίνακα συμπτώσεων $\mathbf{F}_{k\times l}$ τέτοια ώστε η συνάρτηση:

$$\sigma(\{ r'_{is} \}, \{ c'_{js} \})=\sum_i \sum_j f_{ij} \sum_s \left(r'_{is} - c'_{js} \right)^2 ,$$

να γίνεται ελάχιστη, κάτω από τους περιορισμούς:

$$\sum_i f_{i+} r'_{is} = 0 \text{ και } \sum_i f_{i+} r'_{is} r'_{it} = \delta^{st} \quad (1)$$

και

$$\sum_j f_{+j} c'_{js} = 0 \text{ και } \sum_j f_{+j} c'_{js} c'_{jt} = \delta^{st} \quad (2),$$

όπου δ^{st} είναι το δέλτα του *Kronecker,* s και *t* δείκτες για τη δήλωση των παραγοντικών αξόνων και i=1,…,k και j=1,…l.

Από τη στιγμή που το πρόβλημα επιλυθεί οι τυποποιημένες συντεταγμένες των σημείων γραμμών και στηλών ορίζουν τους παραγοντικούς άξονες. Επειδή ο προσανατολισμός του συστήματος συντεταγμένων και οι μονάδες μέτρησης επί των αξόνων είναι αυθαίρετες (Greenacre 1984, Michailidis 1996) μια συνηθισμένη επιλογή είναι ο σταθμισμένος μέσος όρος των συντεταγμένων των προβολών των σημείων γραμμών και στηλών επί των παραγοντικών αξόνων να είναι ίσος με μηδέν (Greenacre 1993α και 1984, Nishisato 1994 και 1980, Παπαδημητρίου 2006, 2004 και 1994). Έτσι, δεν θα προκύψει η τετριμμένη λύση όπου τα ζητούμενα σημεία θα είναι ίσα με μηδέν. Αν υποθέσουμε ότι ο πίνακας $\mathbf{F}_{k\times l}$ περιγράφει την κοινή κατανομή δύο τυχαίων κατηγορικών μεταβλητών και R'_s και C'_s είναι οι τυχαίες μεταβλητές, που αντιστοιχούν στις τυποποιημένες συντεταγμένες των γραμμών και στηλών αντίστοιχα του **F** επί του άξονα *s*, τότε από τους περιορισμούς (1) και (2) συνεπάγεται ότι:

$E(R'_s)=E(C'_s)=0$,

$Var(R'_s)=1$ και $Var(C'_s)=1$

και

$Covar(R'_s, C'_s)= 0$, για $s=1,\ldots, p$.

Με βάση τις παραπάνω σχέσεις διαπιστώνουμε ότι οι άξονες που προκύπτουν χρησιμοποιώντας τις τυποποιημένες συντεταγμένες των σημείων, αν θεωρηθούν ως νέες σύνθετες ή/και λανθάνουσες μεταβλητές, έχουν μέση τιμή μηδέν, διασπορά ίση με τη μονάδα και είναι γραμμικά ανεξάρτητοι. Αν υποθέσουμε ότι οι τυποποιημένες συντεταγμένες ακολουθούν Κανονική Κατανομή, τότε μπορούν να ερμηνευτούν ως τιμές από την Τυποποιημένη Κανονική Κατανομή (*z-scores*). Αν συμβολίσουμε με Φ_s και Γ_s τις τυχαίες μεταβλητές, που αντιστοιχούν στις κύριες συντεταγμένες των γραμμών και στηλών αντίστοιχα του **F** επί του άξονα *s*, και λάβουμε υπόψη τις σχέσεις [2.16], [2.17], [2.32.1] και [2.33.1] προκύπτει:

$E(\Phi_s)=E(\Gamma_s)=0$,

$Var(\Phi_s)=Var(\Gamma_s)= \lambda_s$

και

$Covar(\Phi_s, \Gamma_s)= 0$, για $s=1,\ldots, p$.

Παρατηρούμε ότι οι παράγοντες που προκύπτουν με Κύρια Κανονικοποίηση έχουν διακύμανση ίση με την αδράνεια του αντίστοιχου άξονα και μέση τιμή ίση με 0.

2.2.14.2 Συνέπειες της Κανονικοποίησης

Πρώτη Συνέπεια: Διαφορετικά Αριθμητικά Αποτελέσματα

Α) Όπως είδαμε στα προηγούμενα, ανάλογα με τη μέθοδο κανονικοποίησης υπολογίζονται και διαφορετικά οι συντεταγμένες των προβολών των σημείων γραμμών και στηλών πάνω στους παραγοντικούς άξονες.

Β) Στην περίπτωση συμπληρωματικών σημείων οι συντεταγμένες των προβολών τους επί των παραγοντικών αξόνων δίνονται από τις παρακάτω σχέσεις μετάβασης (SPSS Inc., 2004α):

Για συμπληρωματική γραμμή, $r'^{\sup}_{is} = \sum_j \frac{f_{ij}}{f_{i+}} c'_{js} d_s^{2\alpha-2}$.

Για συμπληρωματική στήλη, $c'^{\sup}_{js} = \sum_i \frac{f_{ij}}{f_{+j}} r'_{is} d_s^{2\beta-2}$.

Από τις σχέσεις [2.22] και [2.23] μπορούν να υπολογιστούν οι δείκτες *COR* για τα συμπληρωματικά στοιχεία γραμμών και στηλών αντίστοιχα. Η ποιότητα της προβολής των συμπληρωματικών σημείων μπορεί να βελτιστοποιηθεί με τη μέθοδο που προτείνουν οι (Graffelman & Aluja-Banet, 2003). Η βέλτιστη ποιότητα προβολής, με την έννοια των ελαχίστων τετραγώνων, επιτυγχάνεται αν ληφθεί υπόψη ο συντελεστής γραμμικής συσχέτισης του συμπληρωματικού σημείου με τους παραγοντικούς άξονες κατάλληλα κανονικοποιημένους.

Γ) Με την προϋπόθεση ότι τα δεδομένα που πρόκειται να αναλυθούν προέρχονται από τυχαίο δείγμα από κάποιο άγνωστο πληθυσμό, τότε οι συχνότητες στα κελιά του πίνακα **F** ακολουθούν Πολυωνυμική Κατανομή (SAS Institute 1990, Andersen 1991, Gifi 1996, SPSS Inc. 2004α). Με βάση τη διαπίστωση αυτή, στο πλαίσιο της Ολλανδικής Σχολής, είναι δυνατό, μέσω της μεθόδου «*Δέλτα*» (*Delta Method*) (βλέπε Israëls 1987, Gifi 1996, Rao 2002, SPSS Inc. 2004α, Ενότητα ΣΤ1 του Παραρτήματος ΣΤ), να υπολογιστούν εκτιμητές των διασπορών και συνδιασπορών των χαρακτηριστικών τιμών και των συντεταγμένων των προβολών των σημείων γραμμών και στηλών σε κάθε άξονα. Οι αριθμητικές τιμές των εκτιμήσεων εξαρτώνται από τη μέθοδο κανονικοποίησης που θα επιλεγεί. Ο ρόλος των παραπάνω εκτιμήσεων στην ερμηνεία των αποτελεσμάτων είναι ο εξής (Gifi 1996, Meulman & Heiser 2004):

Αν η τυπική απόκλιση της δειγματικής συντεταγμένης ενός σημείου (γραμμής, στήλης) είναι μεγάλη, τότε υπάρχει αβεβαιότητα για τη θέση (προβολή) του αντίστοιχου σημείου του πληθυσμού. Αντίθετα, αν η τυπική απόκλιση είναι μικρή, τότε με αρκετή σιγουριά αναμένεται το αντίστοιχο σημείο του πληθυσμού να βρίσκεται πολύ κοντά στο δειγματικό. Μικρές τιμές της τυπικής απόκλισης των χαρακτηριστικών τιμών υποδηλώνουν σταθερότητα στη βασική δομή του πίνακα που αναλύεται. Αυτό σημαίνει ότι η ανάλυση θα έδινε τα ίδια αποτελέσματα αν εφαρμοζόταν σε ένα άλλο δείγμα από τον ίδιο πληθυσμό, λίγο διαφορετικό από το

διαθέσιμο. Αν οι χαρακτηριστικές τιμές θεωρηθούν ως τυχαίες μεταβλητές, τότε στην περίπτωση που η μεταξύ τους συσχέτιση είναι υψηλή η θέση των σημαντικών σημείων πάνω στους άξονες παρουσιάζει αστάθεια. Έτσι, ένα σημείο, ενώ αρχικά είναι σημαντικό για τον πρώτο άξονα, μπορεί, στη συνέχεια, κάτω από μικρές διαταραχές των αρχικών δεδομένων, να χαρακτηρίζει τον δεύτερο.

Δεύτερη Συνέπεια: Διαφορετικός Τρόπος Ανασύστασης του Αρχικού Πίνακα Συμπτώσεων

Η ανασύσταση του αρχικού πίνακα αντιστοιχιών **P** επιτυγχάνεται μέσω των παρακάτω σχέσεων (Andersen 1991, Greenacre 1993α, Blasius & Greenacre 1994):

$$p_{ij} = r_i c_j \left(1 + \sum_{s=1}^{p} \sqrt{\lambda_s}\ r_{is} c_{js} \right), \qquad [2.40]$$

$$p_{ij} = r_i c_j \left(1 + \sum_{s=1}^{p} \varphi_{is} c_{js} \right), \qquad [2.41]$$

$$p_{ij} = r_i c_j \left(1 + \sum_{s=1}^{p} \frac{\varphi_{is} \gamma_{js}}{\sqrt{\lambda_s}} \right), \qquad [2.42]$$

$$p_{ij} = r_i c_j \left(1 + \sum_{s=1}^{p} r_{is} \gamma_{js} \right). \qquad [2.43]$$

Αν τα δεύτερα μέλη των παραπάνω σχέσεων πολλαπλασιαστούν επί N, τότε επιτυγχάνεται η ανασύσταση των στοιχείων του αρχικού πίνακα συμπτώσεων **F**.

Με βάση όσα αναφέρθηκαν στην Ενότητα 2.2.14.1 σχετικά με τις μεθόδους κανονικοποίησης των συντεταγμένων των προβολών των σημείων γραμμών και στηλών επί των παραγοντικών αξόνων, είναι φανερό ότι η σχέση [2.41] αντιστοιχεί στην Κύρια Κανονικοποίηση κατά Γραμμές (*RPN*), η [2.43] στην Κύρια Κανονικοποίηση κατά Στήλες (*CPN*) και η [2.42] στην Κύρια Κανονικοποίηση (*PN*). Η σχέση [2.40] μπορεί να γραφεί και ως εξής:

$$p_{ij} = r_i c_j \left(1 + \sum_{s=1}^{p} \sqrt{\lambda_s}\ r_{is} c_{js} \right) = r_i c_j \left(1 + \sum_{s=1}^{p} \lambda_s^{\frac{1}{4}} \lambda_s^{\frac{1}{4}} r_{is} c_{js} \right) = r_i c_j \left(1 + \sum_{s=1}^{p} \lambda_s^{\frac{1}{4}} r_{is} \lambda_s^{\frac{1}{4}} c_{js} \right).$$

Επομένως, η σχέση [2.40], λαμβάνοντας υπόψη και τις σχέσεις [2.30] και [2.31], αντιστοιχεί στη Συμμετρική Κανονικοποίηση (*SN*).

Η ανασύσταση είναι ακριβής μόνο όταν επιλεγεί λύση με όλους τους δυνατούς $p=min\{k\text{-}1, l\text{-}1\}$ παραγοντικούς άξονες (Andersen 1991, Blasius & Greenacre 1994). Σε περίπτωση που επιλεγεί μια λύση με κ παραγοντικούς άξονες, όπου $\kappa < p$, τότε η ανασύσταση είναι προσεγγιστική και συνεπώς:

$$p_{ij} \approx r_i c_j \left(1+\sum_{\kappa} \sqrt{\lambda_\kappa}\, r_{i\kappa} c_{j\kappa}\right), \qquad [2.44]$$

$$p_{ij} \approx r_i c_j \left(1+\sum_{\kappa} \varphi_{i\kappa} c_{j\kappa}\right), \qquad [2.45]$$

$$p_{ij} \approx r_i c_j \left(1+\sum_{\kappa} \frac{\varphi_{i\kappa}\gamma_{j\kappa}}{\sqrt{\lambda_\kappa}}\right), \qquad [2.46]$$

$$p_{ij} \approx r_i c_j \left(1+\sum_{\kappa} r_{i\kappa}\gamma_{j\kappa}\right). \qquad [2.47]$$

Παρατηρούμε ότι στην περίπτωση που $\kappa < p$ υπάρχει απώλεια της αρχικής πληροφορίας ιδιαίτερα όταν κ=2 και p αρκετά μεγαλύτερο από το κ. Έτσι, στο παραγοντικό επίπεδο 1×2 το υπό εξέταση φαινόμενο δεν θα προβάλλεται ικανοποιητικά. Σύμφωνα με τον Κουτσουπιά (1999), μέρος της χαμένης πληροφορίας μπορεί να ανακτηθεί αν χρησιμοποιηθούν και άλλα παραγοντικά επίπεδα, όπως το 2×3 και το 1×3. Ενημερωτικά σημειώνουμε ότι οι Groenen & Van de Velden (2004) μελέτησαν τη δυνατότητα εξεύρεσης ενός ή περισσοτέρων πινάκων συμπτώσεων στους οποίους θα μπορούσε να αντιστοιχεί μια μερική λύση ($\kappa < p$) της ΠΑΑ. Το πρόβλημα αυτό παρουσιάζει μόνο θεωρητικό ενδιαφέρον και αποτελεί το αντικείμενο της «Αντίστροφης Ανάλυσης των Αντιστοιχιών» (*Inverse Correspondence Analysis*).

Αν η σχέση [2.40] θεωρηθεί ως ένα σύστημα εξισώσεων με αγνώστους τα λ_s, r_{is} και c_{js}, τότε οι αντίστοιχες τιμές που προκύπτουν από την ΠΑΑ μπορούν να θεωρηθούν ως εκτιμητές σταθμισμένων ελαχίστων τετραγώνων των άγνωστων παραμέτρων (Greenacre 1988α, Andersen 1991). Έτσι, στην περίπτωση που τα δεδομένα προέρχονται από τυχαία δειγματοληψία και τα λ_s, r_{is} και c_{js} θεωρηθούν ως τυχαίες

μεταβλητές, τότε οι αντίστοιχες τιμές που υπολογίζονται μέσω της ΠΑΑ αποτελούν εκτιμήσεις των αντίστοιχων αναμενόμενων τιμών. Ανάλογα συμπεράσματα ισχύουν για τις σχέσεις [2.41], [2.42] και [2.43]. Τέλος, από τη σχέση [2.44] έχουμε (Greenacre, 1993α):

$$p_{ij} = r_i c_j \left(1 + \sum_{\kappa} \sqrt{\lambda_\kappa}\ r_{i\kappa} c_{j\kappa} \right) + e_{ij}, \qquad [2.48]$$

όπου η διαφορά (σφάλμα) e_{ij} χρησιμοποιείται για την εξισορρόπηση της ανασύστασης των αρχικών δεδομένων, με την έννοια της πληροφορίας που χάνεται και δεν ερμηνεύεται από τη λύση με κ άξονες. Παρόμοια ισχύουν για τις υπόλοιπες τρεις σχέσεις [2.45], [2.46] και [2.47].

Πληροφοριακά αναφέρουμε ότι στις σχέσεις ανασύστασης [2.40] και [2.48] στηρίζονται ποικίλα συσχετιστικά υποδείγματα – μοντέλα, που έχουν προταθεί κατά καιρούς, για την ανάλυση της σχέσης μεταξύ δύο ή περισσότερων κατηγορικών μεταβλητών (βλέπε Gilula 1986 και 1984, Gilula & Haberman 1988 και 1986, Choulakian 1988, Gilula & Krieger 1989, Gilula & Ritov 1990, Faust & Wasserman 1993, Van der Heijden, Mooijaart & Takane 1994, Haberman 1995 και 1981, Goodman 1996, 1993 και 1991, Mirkin 2001). Τα αριθμητικά αποτελέσματα, εστιάζονται μόνο στις εκτιμήσεις μεγίστης πιθανοφάνειας των παραμέτρων λ_κ, $r_{i\kappa}$ και $c_{j\kappa}$, και είναι συγκρίσιμα με τα αντίστοιχα που υπολογίζονται από την ΠΑΑ. Η σύνδεση της μεθόδου με συσχετιστικά και λογαριθμογραμμικά υποδείγματα αποσκοπεί στη “μοντελοποίηση” και την ένταξή της στο μεθοδολογικό πλαίσιο της Επαγωγικής Στατιστικής. Κάτω από την ισχύ συγκεκριμένων υποθέσεων και προϋποθέσεων οι εκτιμήσεις των εκτιμώμενων παραμέτρων των υποδειγμάτων συνοδεύονται από ελέγχους στατιστικής σημαντικότητας.

Τρίτη Συνέπεια: Διαφορετική Ερμηνεία των Παραγοντικών Διαγραμμάτων

Ας συμβολίσουμε με $\mathbf{\Phi}$ ($\mathbf{\Phi} = \mathbf{D}_\mathbf{r}^{-1/2}\mathbf{UD}$) και $\mathbf{\Gamma}$ ($\mathbf{\Gamma} = \mathbf{D}_\mathbf{c}^{-1/2}\mathbf{VD}$) τους πίνακες που περιέχουν, αντίστοιχα, τις κύριες συντεταγμένες φ_{is} και γ_{js} και με $\mathbf{R}^*$ ($\mathbf{R}^* = \mathbf{D}_\mathbf{r}^{-1/2}\mathbf{U}$) και $\mathbf{C}^*$ ($\mathbf{C}^* = \mathbf{D}_\mathbf{c}^{-1/2}\mathbf{V}$) τους πίνακες που περιέχουν, αντίστοιχα, τις τυποποιημένες

συντεταγμένες r_{is} και c_{js}. Οι στήλες (διανύσματα) των πινάκων $\mathbf{R}^*$ και $\mathbf{C}^*$ ορίζουν τους «κύριους άξονες», ενώ οι στήλες των **Φ** και **Γ** τους «παράγοντες» (De Leeuw & Van der Heijden 1988, Benzécri 1992, Παπαδημητρίου 2006, 2004 και 1994). Οι στήλες του πίνακα **Φ** αντιστοιχούν στα προφίλ των γραμμών, ενώ οι στήλες του πίνακα $\mathbf{C}^*$ ορίζουν το *simplex*, στις κορυφές του οποίου προβάλλονται οι τυποποιημένες συντεταγμένες των στηλών του **F**. Η ταυτόχρονη απεικόνιση των δύο συνόλων συντεταγμένων **Φ** και $\mathbf{C}^*$ στο ίδιο παραγοντικό επίπεδο αποτελεί το μη συμμετρικό διάγραμμα των γραμμών (Micheloud 1997, Greenacre 1993α και 1993β, Gabriel 1971 και 2002, Meulman & Heiser 2004) και επιλύει το «Πρόβλημα των Γραμμών». Ανάλογα, οι στήλες του πίνακα **Γ** αντιστοιχούν στα προφίλ των στηλών και οι αντίστοιχες στήλες του πίνακα $\mathbf{R}^*$ ορίζουν το *simplex* των γραμμών. Η απεικόνιση των δύο συνόλων συντεταγμένων **Γ** και $\mathbf{R}^*$ στο ίδιο παραγοντικό επίπεδο αποτελεί το μη συμμετρικό διάγραμμα των στηλών και το «Πρόβλημα των Στηλών» βρίσκει τη λύση του. Σύμφωνα με τους Greenacre (1993α), Andersen (1991), Micheloud (1997) και Bendixen (2003), μια συνηθισμένη επιλογή είναι η σχεδίαση ενός συμμετρικού διαγράμματος όπου απεικονίζονται ταυτόχρονα στο ίδιο παραγοντικό επίπεδο οι κύριες συντεταγμένες φ_{is} και γ_{js}, οι οποίες όμως δεν αντιστοιχούν στον ίδιο χώρο (Lebart, Morineau & Tabard 1977, Carroll, Green & Schaffer 1989, 1987 και 1986, Greenacre 1993α και 1993β, 1989 και 1984, Gifi 1996, Nishisato 1998, 1995, 1994 και 1980, Escofier & Pagès 1998, Nishisato & Clavel 2003), και έτσι είναι δύσκολο να ερμηνεύσει κανείς τις σχέσεις μεταξύ των σημείων γραμμών και στηλών με βάση τις αποστάσεις τους, όπως αυτές προβάλλονται στο παραγοντικό επίπεδο. Τα διαγράμματα αυτά αποτελούν παράδοση στη Γαλλική Σχολή (Novak & Hoffman 1990, Benzécri 1992, Καραπιστόλης 1999, Lebart, Morineau & Piron 2000, Le Roux & Rouanet 2004, Murtagh 2005, Παπαδημητρίου 2006, 2004 και 1994) και για το λόγο αυτό ονομάζονται «*French plots*» (Bendixen, 2003) ή «*Benzécri plots*» (Gabriel, 2002).

Στα *French plots* η βασική ερμηνεία των αποτελεσμάτων είναι η ακόλουθη (Benzécri 1992, Παπαδημητρίου 2006, 2004 και 1994):

- Σημεία γραμμών που ταυτίζονται ή γειτνιάζουν έχουν αντίστοιχα το ίδιο ή παρόμοιο προφίλ. Το ίδιο ισχύει για τα σημεία στηλών.

- Σημεία γραμμών (στηλών) που προβάλλονται κοντά στην αρχή των αξόνων (κέντρο βάρους) δεν διαφέρουν σημαντικά από τη μέση κατάσταση (μέσο προφίλ), δηλαδή από την περιθώρια σχετική κατανομή των στηλών (γραμμών). Στην περίπτωση που ο πίνακας **F** περιγράφει την κοινή κατανομή απολύτων συχνοτήτων δύο τυχαίων κατηγορικών μεταβλητών, τότε τα σημεία που προβάλλονται κοντά στο κέντρο βάρους δεν συνεισφέρουν στη συσχέτιση (συνάφεια) ή στην εξάρτηση των δύο μεταβλητών.
- Σημεία γραμμών ή στηλών που προβάλλονται, επί του παραγοντικού επιπέδου, σε απομακρυσμένες θέσεις σε σχέση με τα υπόλοιπα έχουν σημαντικά διαφορετικό προφίλ από το αντίστοιχο μέσο προφίλ και, συνεπώς, παρουσιάζουν ιδιαίτερο ενδιαφέρον.

Στα *French plots* η ευκλείδεια απόσταση μεταξύ ενός σημείου γραμμής και ενός σημείου στήλης δεν ορίζεται και συνεπώς δεν έχει νόημα η ερμηνεία της μεταξύ τους απόστασης (Lebart, Morineau & Warwick 1984, SAS Institute 1990, Novak & Hoffman 1990, Andersen 1991, Greenacre 1993α, Micheloud 1997, Clausen 1998, Bendixen 2003, Meulman & Heiser 2004). Στο πλαίσιο της Ολλανδικής Σχολής, η κατασκευή και ερμηνεία των παραγοντικών επιπέδων της ΠΑΑ στηρίζεται στα *biplots* (Meulman & Heiser, 2004). Η κατασκευή των *biplots* είναι στενά συνδεδεμένη με τις κανονικοποιήσεις *SN*, *RPN* και *CPN* που παρουσιάστηκαν στην προηγούμενη ενότητα. Πιο συγκεκριμένα για τη Συμμετρική Κανονικοποίηση (*SN*) έχουμε τα παρακάτω αποτελέσματα:

Από τη σχέση [2.40] προκύπτει ότι:

$$\left(\frac{p_{ij}}{r_i} - c_j\right) = \sum_{s=1}^{p} \varphi_{is}\left(c_j c_{js}\right). \qquad [2.49]$$

Η σχέση [2.49] δηλώνει ότι αν θέλουμε να οπτικοποιήσουμε τις διαφορές μεταξύ των προφίλ των γραμμών από το μέσο προφίλ τους, τότε η απεικόνιση των σημείων (προφίλ) των γραμμών με κύριες συντεταγμένες (φ_{is}) και των σημείων (κορυφών) των στηλών με σταθμισμένες τις τυποποιημένες συντεταγμένες $\left(c_j c_{js}\right)$ αποτελεί *biplot* (Greenacre 1993α και 1993β, Gabriel 2002 και 1971), αφού το εσωτερικό γινόμενο των αντίστοιχων διανυσμάτων μπορεί να ανασυστήσει την αρχική

πληροφορία (βλέπε Ενότητα 2.2.13). Επομένως, τα παραγοντικά διαγράμματα που παράγονται με Συμμετρική Κανονικοποίηση (*SN*) είναι *biplots*. Το ίδιο ισχύει για τα διαγράμματα που παράγονται εφαρμόζοντας την Κύρια Κανονικοποίηση κατά Γραμμές (*RPN*) και την Κύρια Κανονικοποίηση κατά Στήλες (*CPN*). Αυτό είναι φανερό αν γράψουμε τις σχέσεις [2.41] και [2.43] ως εξής:

Από τη [2.41] συνεπάγεται ότι:

$$\frac{\left(\frac{p_{ij}}{r_i} - c_j\right)}{c_j} = \sum_{s=1}^{p} \varphi_{is} c_{js} \ . \qquad [2.50]$$

Από τη [2.43] έχουμε:

$$\frac{\left(\frac{p_{ij}}{c_j} - r_i\right)}{r_i} = \sum_{s=1}^{p} r_{is} \gamma_{js} \ . \qquad [2.51]$$

Από τη σχέση [2.50] ([2.51]) προκύπτει ότι αν θέλουμε να προβάλλουμε τις σχετικές διαφορές μεταξύ των προφίλ των γραμμών (στηλών) από το μέσο προφίλ τους, τότε η απεικόνιση των σημείων (προφίλ) γραμμών (στηλών) με κύριες συντεταγμένες φ_{is} (γ_{js}) και των σημείων (κορυφές) στηλών (γραμμών) με τυποποιημένες c_{js} (r_{is}) αποτελεί *biplot* (βλέπε Greenacre 1993α και 1993β, Gabriel 2002 και 1971). Άλλωστε, και στις τρεις περιπτώσεις ο πίνακας **P** μπορεί να ανασυσταθεί από το εσωτερικό γινόμενο των αντίστοιχων συντεταγμένων. Από τη σχέση [2.42] διαπιστώνουμε ότι η ανασύσταση των στοιχείων του πίνακα **P** δεν είναι εφικτή μόνο από το εσωτερικό γινόμενο των κύριων συντεταγμένων χωρίς τη χρήση του συντελεστή $\left(\sqrt{\lambda_s}\right)^{-1}$, ο οποίος "τανύζει" τις ευκλείδειες αποστάσεις, ώστε να προσεγγίζουν τις χ^2. Από τις σχέσεις μετάβασης [2.18] και [2.19] προκύπτει ότι χωρίς το συντελεστή $\left(\sqrt{\lambda_s}\right)^{-1}$ τα σημεία των γραμμών (στηλών), σε κάθε άξονα, θα προβαλλόταν στο κέντρο βάρους του νέφους των στηλών (γραμμών) γεγονός που θα καθιστούσε αδύνατη τη ταυτόχρονη απεικόνιση των γραμμών και στηλών (Μπεχράκης, 1999). Επομένως, η Κύρια Κανονικοποίηση (*PN*) δεν οδηγεί στην κατασκευή *biplot* (Greenacre & Hastie 1987, Jackson 1991, Greenacre 1993α, Gifi

1996, Gabriel 2002, Meulman & Heiser 2004). Ο Gabriel (2002) έδειξε ότι παρόλο που το εσωτερικό γινόμενο δεν ορίζεται μεταξύ των κύριων συντεταγμένων ωστόσο η οπτικοποίηση της σχέσης μεταξύ, για παράδειγμα, ενός σημείου γραμμής και ενός σημείου στήλης δεν επηρεάζεται σημαντικά επί των *French plots*.

Για να δείξουμε εμπειρικά τις διαφορές μεταξύ των τεσσάρων μεθόδων κανονικοποίησης, σε ότι αφορά την ερμηνεία των παραγοντικών επιπέδων, θα αναλύσουμε τα δεδομένα του Πίνακα 2.5 με το στατιστικό πακέτο SPSS, έκδοση 13.

Πίνακας 2.5: Κατανομή του Είδους των Διακοπών των Φοιτητών κατά Επάγγελμα του Πατέρα. Πίνακας Συμπτώσεων Απολύτων Συχνοτήτων με Σύνολα Γραμμών και Στηλών

	Είδος Διακοπών Φοιτητή								
Επάγγελμα Πατέρα	δ1	δ2	δ3	δ4	δ5	δ6	δ7	δ8	*Σύνολο*
ε1	160	28	0	321	36	141	45	65	796
ε2	35	34	1	178	8	0	4	0	260
ε3	700	354	229	959	185	292	119	140	2.978
ε4	961	471	633	1.580	305	360	162	148	4.620
ε5	572	537	279	1.689	206	748	155	112	4.298
ε6	441	404	166	1.079	178	434	178	92	2.972
ε7	783	1.114	387	4.052	497	1.464	525	387	9.208
ε8	65	43	21	294	79	57	18	6	583
ε9	77	60	189	839	53	124	28	53	1.423
ε10	741	332	327	1.789	311	236	102	102	3.940
Σύνολο	4.535	3.377	2.232	12.780	1.858	3.856	1.336	1.105	31.079

Πηγή: Παπαδημητρίου (1994, σ. 206).

Η *RPN* (βλέπε Διάγραμμα 2.1) έχει ως αποτέλεσμα η ευκλείδεια απόσταση ανάμεσα σε ένα σημείο (προφίλ) γραμμής και της αρχής των αξόνων να προσεγγίζει την απόσταση χ^2 ανάμεσα στο προφίλ της γραμμής και στο μέσο προφίλ (κέντρο βάρους) των γραμμών. Η ευκλείδεια απόσταση μεταξύ των σημείων γραμμών προσεγγίζει την απόσταση χ^2 ανάμεσα στα προφίλ γραμμών. Οι συντεταγμένες των γραμμών είναι οι σταθμισμένοι μέσοι όροι των συντεταγμένων των στηλών (βλέπε σχέση [2.20]). Οι συντεταγμένες των στηλών τυποποιούνται ώστε το σταθμισμένο άθροισμα των τετραγώνων των αποστάσεων τους από το κέντρο βάρος να είναι ίσο με τη μονάδα (βλέπε σχέση [2.27]). Όπως αναφέρθηκε στα προηγούμενα, στη *RPN* για την απεικόνιση των γραμμών χρησιμοποιούνται οι κύριες συντεταγμένες, ενώ για τις στήλες χρησιμοποιούνται οι τυποποιημένες. Έτσι, τα σημεία γραμμών προβάλλονται στο χώρο των στηλών. Αυτή η μέθοδος κανονικοποίησης μεγιστοποιεί τις αποστάσεις μεταξύ των σημείων στηλών και είναι κατάλληλη όταν ενδιαφερόμαστε να

εξετάσουμε με ποιο τρόπο τα προφίλ των γραμμών διαφοροποιούνται μεταξύ τους. Μπορούμε, δηλαδή, να ερμηνεύσουμε μόνο τις αποστάσεις μεταξύ των σημείων γραμμών. Στο παραγοντικό επίπεδο τα σημεία των στηλών είναι πιο "απλωμένα" από τα σημεία των γραμμών τα οποία εμφανίζονται πιο συσπειρωμένα γύρω από την αρχή των αξόνων. Ο Greenacre (1993α και 1993β) έχει προτείνει μεθόδους διόρθωσης των *biplot* διαγραμμάτων της ΠΑΑ, ώστε να αμβλυνθούν οι διαφορές στις κλίμακες μέτρησης μεταξύ των κύριων και τυποποιημένων συντεταγμένων και να βελτιωθεί η ευκρίνεια των διαγραμμάτων. Η διόρθωση έγκειται στο να σταθμιστούν οι τυποποιημένες συντεταγμένες των στηλών ή των γραμμών με βάρη που είναι συναρτήσεις των μαζών τους. Πιο συγκεκριμένα, η καλύτερη προσέγγιση είναι να πολλαπλασιατούν με την τετραγωνική ρίζα των μαζών τους (Greenacre, 2006). Κατά τη *RPN*, οποιαδήποτε ερμηνεία των αποστάσεων μεταξύ των σημείων στηλών ή/και μεταξύ σημείων γραμμών και σημείων στηλών δεν έχει νόημα (Greenacre & Hastie 1987, SAS Institute 1990, Greenacre 1993α, Gabriel 2002 και 1971, Bendixen 2003, Meulman & Heiser 2004).

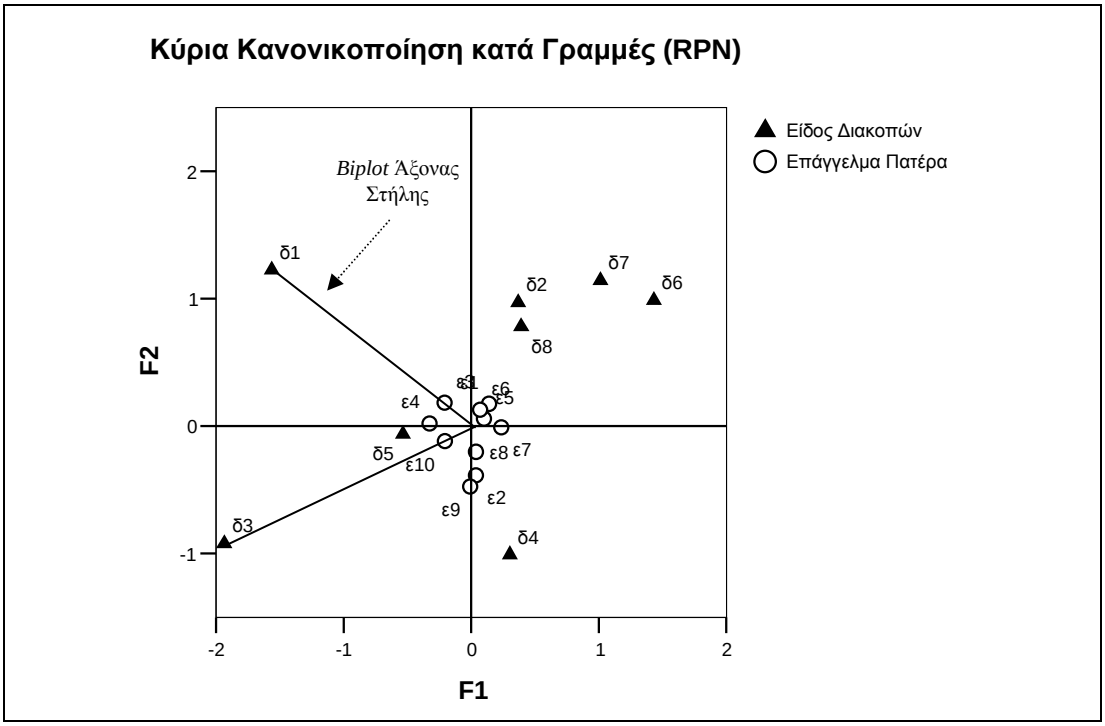
Αν και τα σημεία των στηλών, λόγω της *RPN*, είναι πιο απλωμένα στο παραγοντικό επίπεδο μπορούμε να διαπιστώσουμε τις σχέσεις μεταξύ των δύο μεταβλητών (μεταβλητή γραμμής και μεταβλητή στήλης) με τον παρακάτω τρόπο (Greenacre & Hastie 1987, Greenacre 1993α και 1993β, Meulman & Heiser 2004, βλέπε και Ενότητα 2.2.13):

Σχεδιάζουμε την ευθεία που ενώνει την αρχή των αξόνων με ένα σημείο στήλης. Η ευθεία αυτή ονομάζεται «*biplot* άξονας» (Greenacre, 1993β). Στη συνέχεια, παίρνουμε τις προβολές των σημείων γραμμών πάνω σε αυτή την ευθεία[6]. Οι αποστάσεις των προβολών από το σημείο στήλης αποτελούν ένδειξη για το πώς τα σημεία γραμμών σχετίζονται (συνδέονται) με το συγκεκριμένο σημείο στήλης. Ειδικότερα, τα σημεία γραμμών που οι προβολές τους είναι πλησιέστερα στο σημείο στήλης παρουσιάζουν υψηλότερες τιμές τυποποιημένων υπολοίπων σε σχέση με τα σημεία γραμμών που οι προβολές τους είναι πιο απομακρυσμένες. Η διαδικασία

---

[6] Η ορθογώνια προβολή ενός διανύσματος **u** του $\Re^n$ επί ένός άλλου διανύσματος **v** δίνεται από τη σχέση: $P_{\mathbf{v}}\mathbf{u} = \frac{< \mathbf{u}, \mathbf{v} >}{< \mathbf{u}, \mathbf{u} >}\mathbf{v}$.

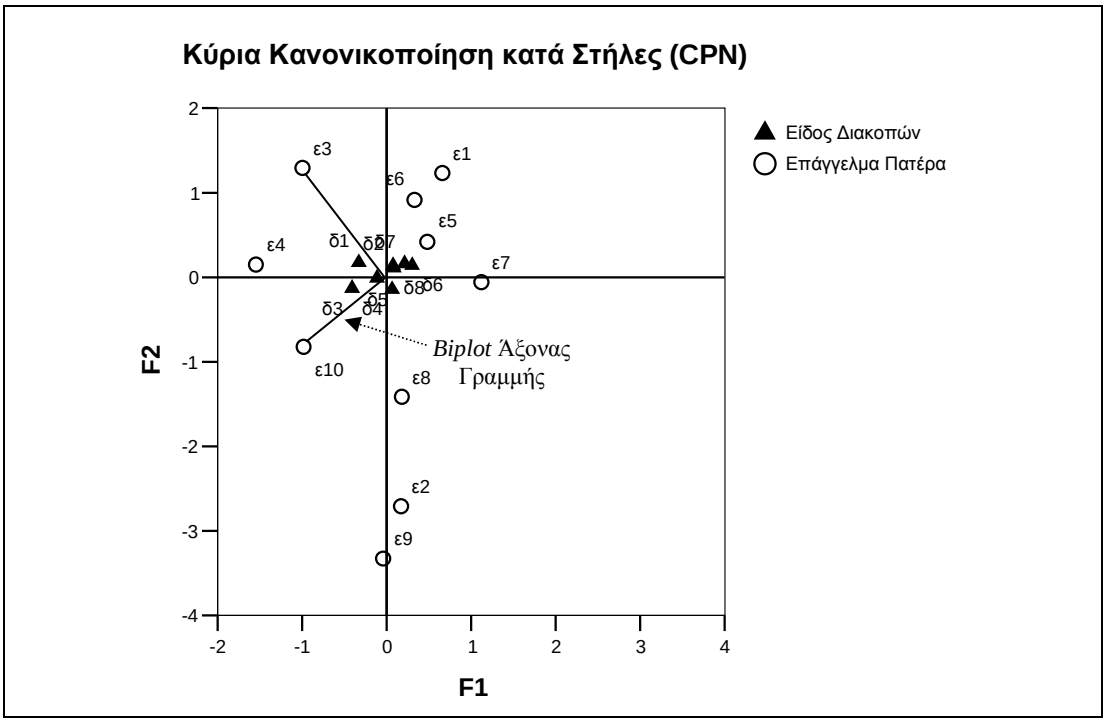
επαναλαμβάνεται και για τα υπόλοιπα σημεία στηλών. Αυτός ο τρόπος ερμηνείας, που είναι χαρακτηριστικός για τα *biplots*, είναι ο ίδιος και για τις μεθόδους κανονικοποίησης *RPN*, *CPN* και *SN*. Δεν έχει όμως εφαρμογή στη *PN*, αφού το αντίστοιχο παραγοντικό επίπεδο δεν είναι *biplot*. Μέσω των *biplots* τα παραγοντικά επίπεδα της ΠΑΑ εμπλουτίζονται με επιπλέον άξονες αναφοράς, τους *biplot* άξονες, επί των οποίων μπορεί να μελετηθεί η διάταξη των προβαλόμενων σημείων.



Διάγραμμα 2.1: Παραγοντικό Επίπεδο 1×2 με Κύρια Κανονικοποίση κατά Γραμμές (*RPN*)

Κατά την κανονικοποίηση *CPN* (βλέπε Διάγραμμα 2.2) η ευκλείδεια απόσταση ανάμεσα στα σημεία στηλών του παραγοντικού διαγράμματος προσεγγίζει την απόσταση χ^2 ανάμεσα στα προφίλ των στηλών του πίνακα συμπτώσεων. Η μέθοδος αυτή είναι κατάλληλη αν ενδιαφερόμαστε να εξετάσουμε το πώς διαφοροποιούνται ή ομοιάζουν οι κατηγορίες (κλάσεις) της μεταβλητής των στηλών. Στην περίπτωση αυτή, οι συντεταγμένες των στηλών είναι οι σταθμισμένοι μέσοι των συντεταγμένων των γραμμών (βλέπε σχέση [2.21]). Οι συντεταγμένες των γραμμών τυποποιούνται ώστε το σταθμισμένο άθροισμα των τετραγώνων των αποστάσεων τους από το κέντρο βάρος να είναι ίσο με τη μονάδα (βλέπε σχέση [2.28]). Έτσι, στο παραγοντικό επίπεδο τα σημεία των γραμμών είναι πιο απλωμένα από τα σημεία των στηλών, τα οποία εμφανίζονται συγκεντρωμένα στην αρχή των αξόνων. Οποιαδήποτε ερμηνεία των αποστάσεων μεταξύ των σημείων γραμμών ή/και μεταξύ σημείων γραμμών και

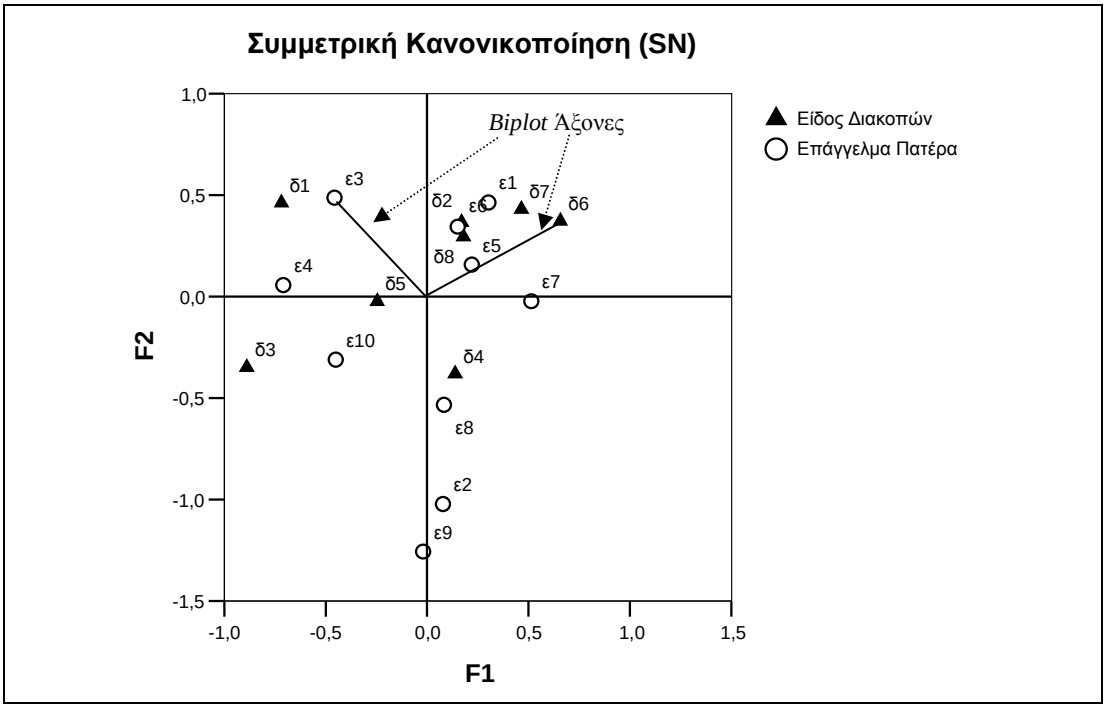
σημείων στηλών δεν έχει νόημα (Greenacre & Hastie 1987, SAS Institute 1990, Greenacre 1993α, Bendixen 2003, Meulman & Heiser 2004).



Διάγραμμα 2.2: Παραγοντικό Επίπεδο 1×2 με Κύρια Κανονικοποίηση κατά Στήλες (*CPN*)

Η μέθοδος *SN* (βλέπε Διάγραμμα 2.3) δίνει τη δυνατότητα να χειριστούμε τις γραμμές και τις στήλες συμμετρικά διαχέοντας την αδράνεια πάνω στις γραμμές και τις στήλες ταυτόχρονα. Η αδράνεια μοιράζεται εξίσου πάνω στις συντεταγμένες των γραμμών και των στηλών. Για κάθε άξονα, οι συντεταγμένες των γραμμών είναι οι σταθμισμένοι μέσοι των σκορ των στηλών διαιρεμένα με την τετραγωνική ρίζα της αντίστοιχης χαρακτηριστικής τιμής του άξονα και οι συντεταγμένες των στηλών είναι οι σταθμισμένοι μέσοι των συντεταγμένων των γραμμών διαιρεμένα με την τετραγωνική ρίζα της αντίστοιχης χαρακτηριστικής τιμής του άξονα (βλέπε σχέσεις [2.37]). Να σημειωθεί ότι στην περίπτωση αυτή ούτε οι αποστάσεις μεταξύ των σημείων γραμμών ούτε οι αποστάσεις μεταξύ των σημείων στηλών είναι προσεγγίσεις της απόστασης χ^2 (Meulman & Heiser, 2004) και δεν ισχύουν πλέον οι βαρυκεντρικές σχέσεις μεταξύ των συντεταγμένων των σημείων γραμμών και στηλών. Έτσι, δεν έχει νόημα ούτε η ερμηνεία των αποστάσεων μεταξύ των σημείων γραμμών ούτε η ερμηνεία των αποστάσεων μεταξύ των σημείων στηλών. Έχει νόημα μόνο η *biplot* ερμηνεία των αποστάσεων μεταξύ σημείων γραμμών και στηλών. Η μέθοδος αυτή είναι χρήσιμη όταν θέλουμε να εξετάσουμε τις διαφορές ή τις

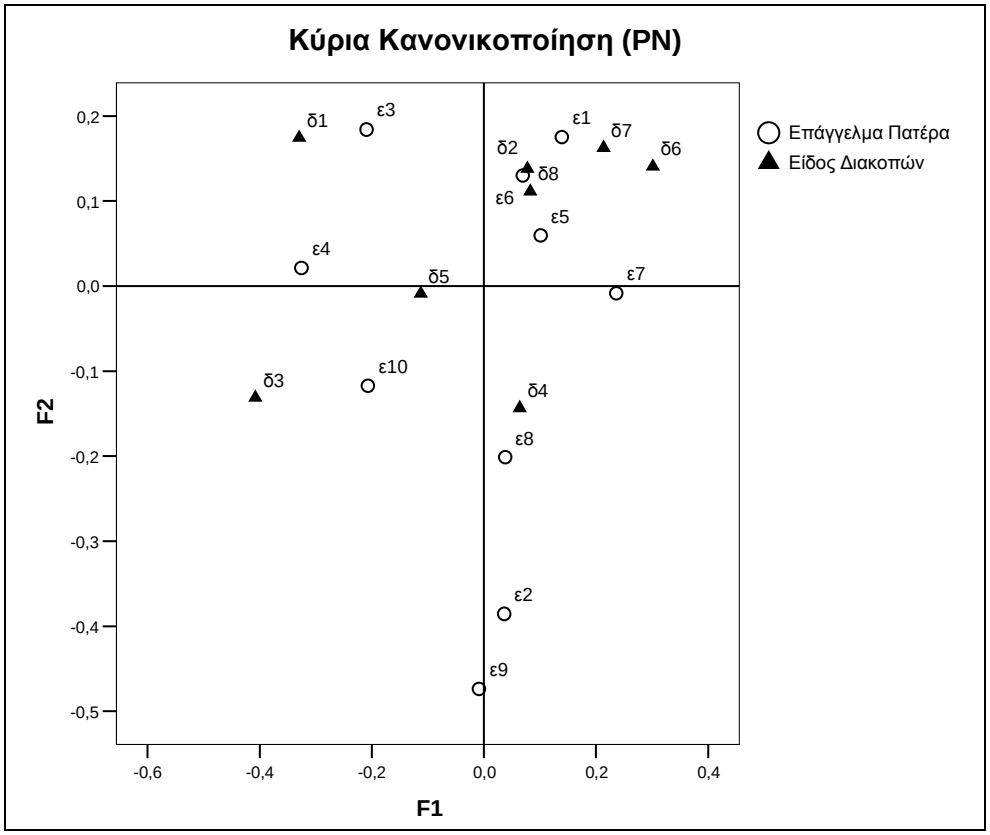
ομοιότητες μεταξύ των δύο μεταβλητών (μεταβλητή γραμμής, μεταβλητή στήλης). Στο σύστημα *GIFI*, αυτή είναι και η προτεινόμενη μέθοδος για την ταυτόχρονη απεικόνιση των σημείων γραμμών και σημείων στηλών στα παραγοντικά επίπεδα (Jackson 1991, Meulman & Heiser 2004). Μάλιστα, στο SPSS αποτελεί την προκαθορισμένη (*default*) επιλογή κανονικοποίησης. Η περαιτέρω διερεύνηση των σχέσεων, στηρίζεται στη *biplot* ερμηνεία του παραγοντικού επιπέδου. Μπορούμε, για παράδειγμα, να φέρουμε την ευθεία που ενώνει ένα σημείο γραμμής με την αρχή των αξόνων, στη συνέχεια να φέρουμε τις προβολές των σημείων στηλών πάνω σε αυτή την ευθεία και, τέλος, να ερμηνεύσουμε τις αποστάσεις των προβολών από το σημείο στήλης όπως και στην περίπτωση της *RPN*.



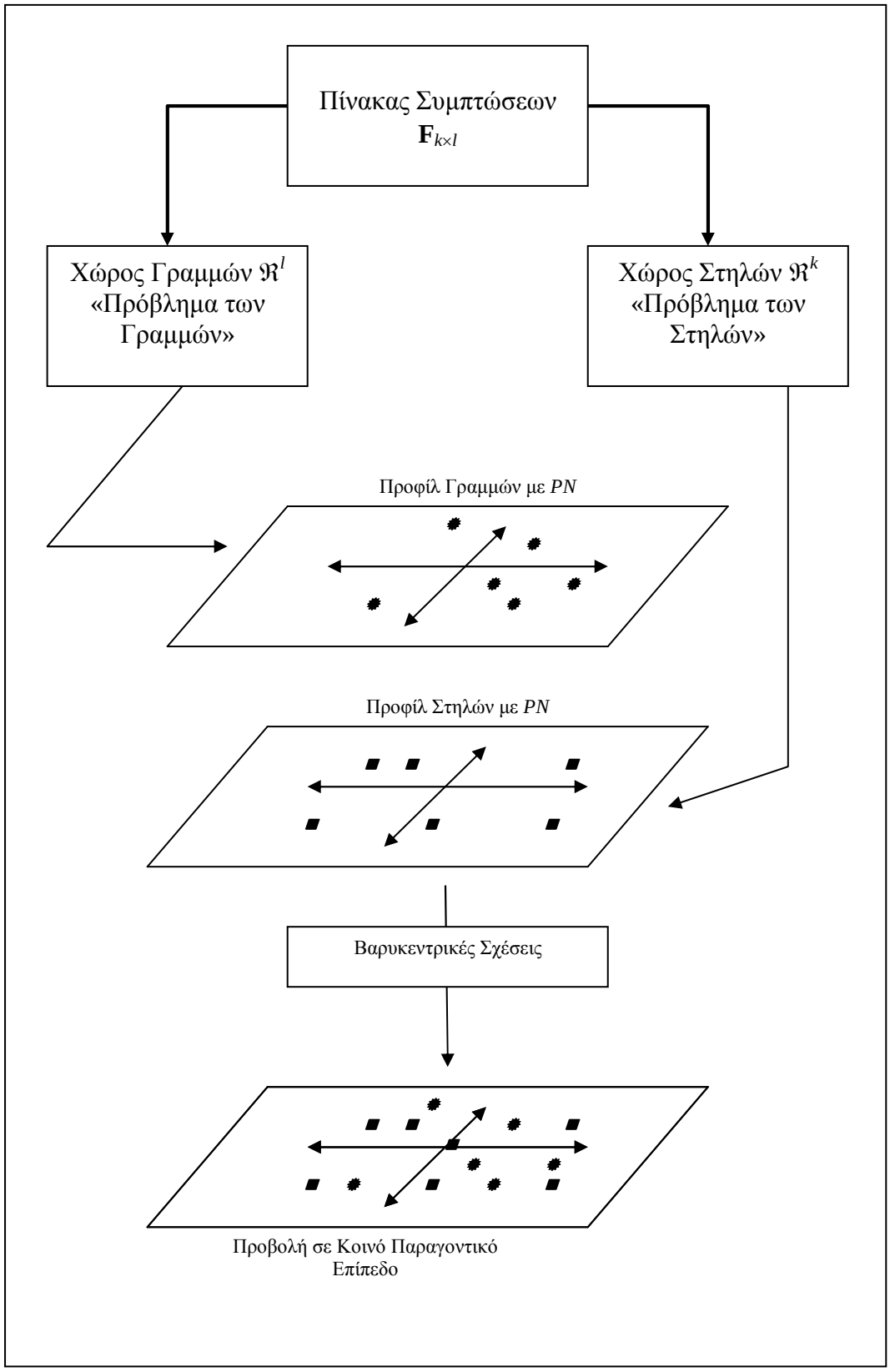
Διάγραμμα 2.3: Παραγοντικό Επίπεδο 1×2 με Συμμετρική Κανονικοποίση (*SN*)

Στην κανονικοποίηση *PN* (βλέπε Διάγραμμα 2.4) η αδράνεια διαχέεται δύο φορές. Μία φορά πάνω στις συντεταγμένες των γραμμών και μία φορά πάνω στις συντεταγμένες των στηλών. Όπως είδαμε στα προηγούμενα, επί του παραγοντικού επιπέδου, η ευκλείδεια απόσταση τόσο ανάμεσα στα σημεία των γραμμών όσο και ανάμεσα στα σημεία των στηλών προσεγγίζει την απόσταση χ^2 ανάμεσα στα αντίστοιχα σημεία (προφίλ) γραμμών και στηλών του πίνακα συμπτώσεων **F**. Η μέθοδος αυτή είναι πιο χρήσιμη όταν θέλουμε να εξετάσουμε τις αποστάσεις μεταξύ των γραμμών και τις αποστάσεις μεταξύ των στηλών ξεχωριστά και όχι όταν θέλουμε να εξετάσουμε το πώς τα σημεία των γραμμών και τα σημεία των στηλών σχετίζονται

το ένα με το άλλο. Παραγοντικά επίπεδα με αυτή τη μέθοδο κανονικοποίησης δεν είναι κατάλληλα (SAS Institute 1990, Andersen 1991, Greenacre 1993α και 1993β, Hair *et al.* 1995, Micheloud 1997, Clausen 1998, Bendixen 2003, Meulman & Heiser 2004) και μπορεί να οδηγήσουν σε λανθασμένη ερμηνεία. Για το λόγο αυτό, η επιλογή δημιουργίας παραγοντικών επιπέδων (*French plots*) δεν είναι διαθέσιμη στη διαδικασία *Correspondence Analysis* του SPSS όταν επιλεγεί ως μέθοδος κανονικοποίησης η *PN* (SPSS Inc. 1998, Meulman & Heiser 2004). Κατά την *PN* στο αντίστοιχο παραγοντικό επίπεδο είναι σαν να έχουν τοποθετηθεί δύο διαγράμματα - χάρτες με την ίδια κλίμακα και προσανατολισμό ο ένας πάνω στον άλλο (βλέπε Σχήμα 2.3). Για παράδειγμα, ο χάρτης του νομού *Θ* πάνω στο χάρτη του νομού *Σ*. Οι αποστάσεις μεταξύ των πόλεων του νομού *Θ* έχουν νόημα, το ίδιο και οι αποστάσεις μεταξύ των πόλεων του νομού *Σ*. Οι αποστάσεις όμως μεταξύ πόλεων του νομού *Θ* και πόλεων του νομού *Σ* δεν έχουν κανένα νόημα.



Διάγραμμα 2.4: Παραγοντικό Επίπεδο 1×2 με Κύρια Κανονικοποίση (*PN*)



Σχήμα 2.3: Η Κύρια Κανονικοποίηση (*PN*) της ΠΑΑ στο Πλαίσιο της Γαλλικής Σχολής (προσαρμογή από τους Lebart, Morineau & Piron, 2000, σ. 86)

Παρατήρηση 2.3

Κατά παράδοση, μετά τον υπολογισμό και την παρουσίαση των αριθμητικών αποτελεσμάτων της ΠΑΑ ακολουθεί η κατασκευή και η παρουσίαση ενός ή περισσοτέρων παραγοντικών επιπέδων (Andersen, 1991). Σκοπός της ανάλυσης είναι να αναδειχθεί η φυσική ερμηνεία των σημείων γραμμών ή/και των σημείων στηλών του αρχικού πίνακα συμπτώσεων μέσα από τις σχέσεις και τις αλληλεπιδράσεις τους, όπως αυτές απορρέουν από τα ίδια τα δεδομένα και σχηματοποιούνται γραφικά στα παραγοντικά επίπεδα (Κουτσουπιάς, 1999). Όμως, η προσπάθεια διατύπωσης γενικών κανόνων για την ερμηνεία των νεφών των σημείων δεν είναι απλή υπόθεση και, πολλές φορές, είναι σε βάρος της ιδιαιτερότητας των δεδομένων που αναλύονται και απεικονίζονται γραφικά κάθε φορά. Πάνω σε αυτό ο Παπαδημητρίου (1994, σ.166) σημειώνει σχετικά: "*Η ερμηνεία των παραγοντικών επιπέδων και η εξαγωγή συμπερασμάτων από αυτά, χρειάζεται μια κάποια εξοικείωση αλλά και εμπειρία*". Στην ελληνική βιβλιογραφία, γενικές κατευθύνσεις, σχετικά με την ερμηνεία των παραγοντικών αξόνων και επιπέδων, συναντάμε στους Αθανασιάδη (1995), Μπεχράκη (1999), Κουτσουπιά (1999), Κουτσουπιά & Παπαδημητρίου (1999), Μαυρομάτη (1999), Καραπιστόλη (1999 και 2002), Καρλή (2005) και Παπαδημητρίου (2006, 2004 και 1994).

Συχνά η ανάγκη για εύκολη και γρήγορη ερμηνεία των παραγοντικών αξόνων και των αντίστοιχων παραγοντικών επιπέδων είναι πιθανό να οδηγήσει τον ερευνητή στο να στηρίξει τις αποφάσεις του στα παραγοντικά επίπεδα και όχι στα αριθμητικά αποτελέσματα της ανάλυσης. Δεν είναι λίγες οι περιπτώσεις όπου σε δημοσιευμένες επιστημονικές εργασίες παρουσιάζονται μόνο τα παραγοντικά επίπεδα (Micheloud, 1997) και όλη η συζήτηση και η ερμηνεία των αποτελεσμάτων της Παραγοντικής Ανάλυσης των Αντιστοιχιών βασίζεται πάνω σε αυτά.

Είδαμε ότι το να στηρίζει κανείς την ερμηνεία των αποτελεσμάτων και τις αποφάσεις του για μελλοντικές ενέργειες μόνο στην εικόνα που αποδίδουν τα παραγοντικά επίπεδα εγκυμονεί σοβαρό κίνδυνο για λανθασμένη συμπερασματολογία. Ο κίνδυνος αυτός είναι άμεσος αν δεν ληφθεί υπόψη η Σχολή Ανάλυσης Δεδομένων, στο πλαίσιο της οποίας εφαρμόζονται οι υπολογιστικοί αλγόριθμοι που χρησιμοποιούνται από τα διάφορα στατιστικά πακέτα. Για παράδειγμα, συγκρίναμε 16 λογισμικά στατιστικής επεξεργασίας, που περιλαμβάνουν στις διαδικασίες τους την ΠΑΑ, ως προς τις

μεθόδους κανονικοποίησης και τις δυνατότητες κατασκευής παραγοντικών επιπέδων που προσφέρουν. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 2.6.

Πίνακας 2.6: Σύγκριση Λογισμικών

Λογισμικό	Κανονικοποίηση *PN*	Άλλες Μέθοδοι Κανονικοποίησης	French Plot
SPSS ver. 13	ΝΑΙ	ΝΑΙ	**ΟΧΙ**
SAS ver. 9	ΝΑΙ	ΝΑΙ	ΝΑΙ
STATISTICA ver. 6	ΝΑΙ	ΝΑΙ	ΝΑΙ
SYSTAT ver.10	ΝΑΙ	ΟΧΙ	ΝΑΙ
JMP IN ver.5.1	ΝΑΙ	ΟΧΙ	ΝΑΙ
MINITAB ver. 14	ΝΑΙ	ΝΑΙ	ΝΑΙ
SPAD ver. 5	ΝΑΙ	ΟΧΙ	ΝΑΙ
XLSTAT ver. 7.5.3.	ΝΑΙ	ΝΑΙ	ΝΑΙ
STATBOXPRO ver. 5	ΝΑΙ	ΟΧΙ	ΝΑΙ
WINVISTA ver. 6.4	ΝΑΙ	ΝΑΙ	ΝΑΙ
STATA ver. 9.1	ΝΑΙ	ΝΑΙ	ΝΑΙ
BIPLOTS	ΝΑΙ	ΝΑΙ	ΝΑΙ
BMDP ver. 7	ΝΑΙ	ΝΑΙ	ΝΑΙ
MAD ver. 4 (*)	ΝΑΙ	ΟΧΙ	ΝΑΙ
S-PRO (*)	ΝΑΙ	ΟΧΙ	ΝΑΙ
ΠΡΑΞΙΤΕΛΗΣ (*)	ΝΑΙ	ΟΧΙ	ΝΑΙ

(*) Πρόκειται για λογισμικά τα οποία έχουν κατασκευαστεί από Έλληνες ερευνητές. Για περισσότερες πληροφορίες σχετικά με τις δυνατότητες των λογισμικών αυτών παραπέμπουμε στον Καραπιστόλη (1999, 2000 και 2002) για το MAD, στον Κουτσουπιά (1999, 2002 και 2005) για το S-PRO και στον Καράκο (2003) για το ΠΡΑΞΙΤΕΛΗΣ.

Όλα τα λογισμικά που συμμετείχαν στη σύγκριση προσφέρουν τη δυνατότητα Κύριας Κανονικοποίησης (*PN*) των Γραμμών και Στηλών και συνεπώς είναι συμβατά με τη Γαλλική Σχολή Ανάλυσης Δεδομένων. Μάλιστα, ορισμένα από αυτά, όπως το SPAD και το MAD, δεν παρέχουν άλλες δυνατότητες κανονικοποίησης. Είναι χαρακτηριστικό ότι το SPSS δεν δίνει τη δυνατότητα κατασκευής *french plot* στην περίπτωση που επιλεγεί ως μέθοδος κανονικοποίησης η *PN*. Το SAS διαθέτει και μια επιπλέον μέθοδο κανονικοποίησης που πρότειναν οι Carroll, Green & Schaffer (1986 και 1987). Η συγκεκριμένη κανονικοποίηση αποτελεί αντικείμενο αντιπαράθεσης γιατί, σύμφωνα με τους εμπνευστές της, είναι δυνατή, επί των παραγοντικών επιπέδων, η άμεση ερμηνεία των ευκλείδειων αποστάσεων μεταξύ των σημείων γραμμών και στηλών. Ο Greenacre (1989) υποστήριξε ότι η κανονικοποίηση αυτή δεν έχει καμία λογική βάση και δεν μπορεί να ερμηνευτεί γεωμετρικά. Οι Carroll, Green & Schaffer (1989) αντίκρουσαν τα επιχειρήματα του Greenacre. Σύμφωνα με τη μέθοδο αυτή οι κανονικοποιημένες συντεταγμένες των σημείων γραμμών (**X**) και στηλών (**Y)** δίνονται από τις σχέσεις:

$$\mathbf{X} = \mathbf{D}_r^{-1/2}\mathbf{U}\left[\mathbf{D} + \mathbf{I}\right]^{1/2}$$

και

$$\mathbf{Y} = \mathbf{D}_c^{-1/2}\mathbf{V}\left[\mathbf{D} + \mathbf{I}\right]^{1/2}.$$

Πρακτικά, η παραπάνω κανονικοποίηση είναι ισοδύναμη με το να εφαρμοστεί η ΠΑΑ στο Γενικευμένο Πίνακα Συμπτώσεων (πίνακας *Burt*) των δύο μεταβλητών. Πάντως, το ζήτημα της ερμηνείας των αποστάσεων μεταξύ των σημείων γραμμών και στηλών αποτελεί ακόμα θέμα προς διερεύνηση (Nishisato & Clavel, 2003).

Το SPSS παρέχει τη δυνατότητα προσαρμογής του βαθμού διάχυσης της αδράνειας πάνω στους άξονες των *biplot* (SPSS Inc. 2004α, Meulman & Heiser 2004). Διαθέτει την επιλογή **Custom**, όπου ο χρήστης μπορεί να εισάγει τιμές στο διάστημα [-1, +1]. Η τιμή –1 αντιστοιχεί στη *CPN*, η τιμή +1 στη *RPN* και η τιμή 0 στη *SN*. Άλλες τιμές εντός του διαστήματος [-1, +1] διαχέουν την αδράνεια των συντεταγμένων των προβολών των σημείων γραμμών και στηλών πάνω στους παραγοντικούς άξονες ασύμμετρα και σε διαφορετικό βαθμό. Αυτή η μέθοδος έχει ως αποτέλεσμα τα σημεία του ενός συνόλου (π.χ. γραμμής) να προβάλλονται πιο απλωμένα στα παραγοντικά επίπεδα και τα σημεία του άλλου συνόλου (π.χ. στήλης) πιο συσπειρωμένα.

2.2.14.3 Σημαντικοί Άξονες-Σημαντικά Σημεία

Α) Σημαντικοί Άξονες

Η επιλογή των σημαντικών παραγοντικών αξόνων, δηλαδή αυτών που θα πρέπει να μελετηθούν με λεπτομέρεια, στηρίζεται κυρίως σε εμπειρικές τεχνικές ανάλογες με αυτές που χρησιμοποιούνται στις Παραγοντικές Αναλύσεις (βλέπε Hair *et al*. 1995, Sharma 1996). Αποτελεί κλασικό πρόβλημα στο χώρο της Πολυδιάστατης Στατιστικής Ανάλυσης και επιδέχεται πολλών και διαφορετικών προσεγγίσεων (Κιοσέογλου, 2003) με αντιφατικά αποτελέσματα. Οι προσεγγίσεις αυτές περιλαμβάνουν εμπειρικά και στατιστικά κριτήρια. Τα εμπειρικά κριτήρια εμπεριέχουν μεγάλο βαθμό υποκειμενικότητας, ενώ τα στατιστικά στηρίζονται είτε στις ασυμπτωτικές κατανομές πιθανότητας των ιδιοτιμών είτε σε μεθόδους επαναδειγματοληψίας (*Bootstrap*) (Karlis, Saporta & Spinakis, 2003). Στην ΠΑΑ ο

προσδιορισμός των παραγοντικών αξόνων που χρήζουν ερμηνείας επιχειρείται συνήθως με τις παρακάτω ευρετικές μεθόδους που εφαρμόζονται και στην Ανάλυση σε Κύριες Συνιστώσες:

- Μετά από εξέταση του διαγράμματος των ιδιοτιμών (*scree plot*). Στο διάγραμμα αυτό ο οριζόντιος άξονας *x* βαθμονομείται με την τάξη των ιδιοτιμών (πρώτη, δεύτερη κ.ο.κ.), ενώ ο κατακόρυφος άξονας *y* με τις τιμές των ιδιοτιμών. Τα ζεύγη των σημείων (*x*, *y*) απεικονίζονται στο επίπεδο και ενώνονται με συνεχή γραμμή. Σύμφωνα με το κριτήριο αυτό επιλέγονται οι άξονες για τους οποίους εμφανίζεται απότομη μεταβολή στις τιμές των ιδιοτιμών[7] με συνέπεια η γραμμή να εμφανίζει τη μορφή "αγκώνα" (*elbow*). Ο εμπειρικός αυτός έλεγχος είναι γνωστός ως κριτήριο του *Cattell* (Harman 1976, Rummel 1979, Κιοσέογλου 2003).
- Με επιλογή των *κ* πρώτων αξόνων που εξηγούν αθροιστικά πάνω από κάποιο συγκεκριμένο ποσοστό της ολικής αδράνειας που κρίνεται ικανοποιητικό (συνήθως 60%).
- Με επιλογή των αξόνων με ιδιοτιμή μεγαλύτερη από 1/*p*, με $p=min\{k\text{-}1,\ l\text{-}1\}$ (Lebart, Morineau & Tabard 1977, Bendixen 2003). Το κριτήριο αυτό είναι αντίστοιχο του κριτηρίου των *Kaiser–Guttman* (βλέπε Harman 1976, Rummel 1979) που εφαρμόζεται στην Ανάλυση σε Κύριες Συνιστώσες και βασίζεται στο παρακάτω σκεπτικό: Κάτω από την υπόθεση της τυχαίας διακύμανσης των δύο κατηγορικών μεταβλητών χωρίς την ύπαρξη συστηματικής σχέσης μεταξύ τους κάθε άξονας, κατά μέσο όρο, θα ερμηνεύει το $(1/p)\times 100$ της ολικής αδράνειας. Επομένως, παραγοντικοί άξονες με ιδιοτιμή μεγαλύτερη από (1/*p*) συνεισφέρουν άνω του μέσου όρου και μάλλον θα πρέπει να διατηρηθούν προς ερμηνεία.

Εν γένει, το πρόβλημα του καθορισμού των σημαντικών αξόνων στην ΠΑΑ δεν έχει μελετηθεί διεξοδικά. Αντίθετα, στην Ανάλυση σε Κύριες Συνιστώσες αποτελεί βασικό θέμα έρευνας (Harman 1976, Rummel 1979, Eastment & Krzanowski 1982, Krzanowski 1987α, Habbard & Allen 1987, Jackson 1991, Huang & Tseng 1992, Jackson 1993, Qian, Gabor & Gupta 1994, Ferré 1995, Hair *et al*. 1995, Sharma 1996, Valle, Li & Qin 1999, Keeling 2000, Meloun *et al*. 2000, Karlis, Saporta & Spinakis 2003, Peres-Neto, Jackson & Somers 2005, Καρλής 2005). Το βασικό συμπέρασμα

[7] Στην ίδια λογική στηρίζεται και η ερμηνεία του «Ιστογράμματος» των ιδιοτιμών (Benzécri, 1992).

που προκύπτει είναι ότι μάλλον θα πρέπει να χρησιμοποιείται συνδυασμός κριτηρίων για την επιλογή των αξόνων, οι οποίοι χρήζουν ερμηνείας και δεν αποτελούν "θόρυβο", ο οποίος πιθανώς να οφείλεται σε τυχαίες διακυμάνσεις, στην ποιότητα των δεδομένων ή/και στην ύπαρξη πολυσυγραμμικότητας μεταξύ των μεταβλητών. Στο πλαίσιο της ΠΑΑ, ο Κιοσέογλου (2003) πρότεινε και σύγκρινε τρεις μεθόδους προσδιορισμού των ωφέλιμων αξόνων, οι οποίες χρησιμοποιούνται στην Ανάλυση σε Κύριες Συνιστώσες και στηρίζονται στην εφαρμογή κατάλληλων υποδειγμάτων γραμμικής παλινδρόμησης επί των σημείων που προβάλλονται στο διάγραμμα των ιδιοτιμών. Πάντως, αξιοσημείωτο είναι το γεγονός ότι πιο απλές τεχνικές, όπως είναι η μέθοδος της «*σπασμένης ράβδου* (*broken stick*)» (βλέπε Καρλής, 2005), οι οποίες χρησιμοποιούνται στις Παραγοντικές Αναλύσεις, δεν έχουν χρησιμοποιηθεί στην ΠΑΑ. Στο Κεφάλαιο 5 προτείνουμε ένα ακόμη εμπειρικό κριτήριο για την επιλογή των σημαντικών αξόνων που βασίζεται στην έννοια της «Δυναμικής Αδράνειας» την οποία ορίζουμε στην Ενότητα 5.10, ενώ στην Ενότητα 5.13.4 παρουσιάζουμε μια εφαρμογή της μεθόδου της «*σπασμένης ράβδου*» στο πλάσιο της ΠΑΑ. Αξίζει, επίσης, να τονιστεί ότι ο Lebart (1976) πρότεινε μέθοδο επιλογής των στατιστικά σημαντικών αξόνων η οποία στηρίζεται στο ότι οι ιδιοτιμές που προκύπτουν από την εφαρμογή της ΠΑΑ σε πίνακα συμπτώσεων δύο τυχαίων κατηγορικών μεταβλητών ακολουθούν την ίδια κατανομή με τις ιδιοτιμές ενός αντίστοιχου πίνακα που ακολουθεί την Κατανομή Wishart με παραμέτρους (l-1) και (k-1) (βλέπε Lebart, Morineau & Tabard 1977, Lebart, Morineau & Warwick 1984, Lebart, Morineau & Piron 2000). Στο Κεφάλαιο 6 (Ενότητα 6.2.1.3) δίνουμε περισσότερα στοιχεία σχετικά με τη στατιστική σημαντικότητα των παραγοντικών αξόνων.

Β) Σημαντικά Σημεία

Κατά παράδοση, ως σημαντικά σημεία επιλέγονται αυτά που σε κάθε άξονα απεικονίζονται καλύτερα και ταυτόχρονα συνεισφέρουν περισσότερο στο να λάβει ο άξονας τη συγκεκριμένη θέση (Lebart, Morineau & Warwick 1984, Benzécri 1992, Le Roux & Rouanet 2004). Για το σκοπό αυτό χρησιμοποιούνται οι δείκτες *COR* και *CTR*.

Όπως αναφέρθηκε στην Ενότητα 2.2.8 ο δείκτης *CTR*(i, s) μετρά τη συνεισφορά του σημείου i, ώστε να λάβει ο παραγοντικός άξονας s τη συγκεκριμένη θέση στο χώρο.

Όσο πιο υψηλή είναι η τιμή του τόσο υψηλότερη είναι και η συνεισφορά του σημείου στην κατασκευή του άξονα. Ο δείκτης *CTR* χρησιμοποιείται για την ερμηνεία των παραγόντων (Benzécri 1992, Greenacre 1993β, Μπεχράκης 1999, Μαυρομάτης 1999, Murtagh 2005). Συνήθως επιλέγονται τα σημεία γραμμών (στηλών) με $CTR(i,s) \geq 1/k$ ($CTR(j,s) \geq 1/l$), όπου k (l) είναι το πλήθος των γραμμών (στηλών) του πίνακα συμπτώσεων **F** (Bendixen, 2003). Ο Benzécri (1992) παρατηρεί ότι τα επιλεγμένα σημαντικά σημεία θα πρέπει να αιτιολογούν τουλάχιστον τα 2/3 της αδράνειας του αντίστοιχου άξονα, ενώ σημεία για τα οποία ο δείκτης *CTR* είναι μεγαλύτερος από 0,25 θα πρέπει να αντιμετωπίζονται με προσοχή γιατί είναι πιθανό να πρόκειται για έκτοπα σημεία, τα οποία ίσως θα πρέπει να εξαιρεθούν από την ανάλυση και να εισαχθούν στη συνέχεια ως συμπληρωματικά.

Είδαμε, επίσης, ότι ο δείκτης *COR*(*s*, *i*) μετρά τη συμβολή του παραγοντικού άξονα *s* στην ερμηνεία της απόστασης του σημείου *i* από το κέντρο βάρους (αρχή ων αξόνων) και μπορεί να ερμηνευτεί ως συντελεστής συσχέτισης του σημείου με τον άξονα. Σύμφωνα με τον Greenacre (1993β), ο δείκτης *COR* βοηθά στην ερμηνεία των προφίλ των σημείων πάνω στους παραγοντικούς άξονες. Το *COR* ενός σημείου γραμμής (ή στήλης) θεωρείται υψηλό όταν είναι μεγαλύτερο από 0,20 (Καραπιστόλης 1999, Bendixen 2003). Ο δείκτης *QLT* μετρά την ποιότητα της απεικόνισης των σημείων στον υποχώρο προβολής (συνήθως 2 ή 3 διαστάσεων), με συνέπεια, όσο μεγαλύτερη είναι η τιμή του τόσο καλύτερη να είναι η απεικόνιση των αντίστοιχων σημείων στο υποχώρο. Για τους Le Roux & Rouanet (2004) ένας δείκτης ποιότητας της αναπαράστασης μεγαλύτερος από 0,50 θεωρείται, εν γένει, ικανοποιητικός.

Τα προηγούμενα κριτήρια αποτελούν εμπειρικές προσεγγίσεις για τον εντοπισμό των σημαντικών σημείων. Στο Κεφάλαιο 6 (Ενότητα 6.3) προτείνουμε μέθοδο εντοπισμού των στατιστικά σημαντικών γραμμών ή/και στηλών, η οποία στηρίζεται στην κατασκευή 100(1-α)% ελλείψεων εμπιστοσύνης γύρω από τα αντίστοιχα σημεία του πίνακα συμπτώσεων.

Ο Καραπιστόλης (1999) θεωρεί σημαντικά, για την ερμηνεία των παραγοντικών αξόνων, τα σημεία που έχουν ικανοποιητικές τιμές και στους δύο δείκτες. Οι Παπαδημητρίου και Φλώρου (1999) πρότειναν ένα νέο δείκτη (*PF*), ο οποίος

λαμβάνει υπόψη ταυτόχρονα το *CTR* και το *COR*. Όσο η τιμή του δείκτη προσεγγίζει, για κάποιο σημείο, την τιμή 0 τόσο πιο σημαντικό είναι το σημείο για την ερμηνεία και δημιουργία του αντίστοιχου άξονα. Οι Benzécri (1992) και Gettler-Summa (1992) υποστηρίζουν ότι η ερμηνεία των πρώτων σε τάξη παραγοντικών αξόνων θα πρέπει να δίνεται με βάση τις ομοιότητες ή/και τις αντιπαραθέσεις των σημείων που παρουσιάζουν υψηλούς δείκτες *CTR*, ενώ ο δείκτης *COR* είναι προτιμότερο να χρησιμοποιείται για την ερμηνεία των αξόνων που ερμηνεύουν χαμηλά ποσοστά της ολικής αδράνειας.

Για τον Benzécri μεγαλύτερη αξία έχουν οι παράγοντες, δηλαδή οι διαστάσεις και οι δομές που αναδεικνύονται μέσω της ΠΑΑ και όχι τα ίδια τα δεδομένα, τα οποία αποτελούν μόνο μια προσεγγιστική εικόνα της πραγματικότητας (Meter *et al*., 1994). Η ερμηνεία ενός παραγοντικού άξονα έγκειται στο να διαπιστωθούν οι ομοιότητες μεταξύ των σημείων που βρίσκονται από τη μια και από την άλλη πλευρά (κατεύθυνση) του άξονα και στη συνέχεια να διατυπωθούν με σαφήνεια οι αντιθέσεις ανάμεσα στις δύο πλευρές (Benzécri, 1992). Οι Le Roux & Rouanet (2004) διακρίνουν τους παραγοντικούς άξονες σε «γενικούς» (*general*) και «ειδικούς» (*specific*). Οι γενικοί άξονες αναδεικνύουν τις αντιθέσεις όλων σχεδόν των σημείων, ενώ οι ειδικοί χαρακτηρίζονται από συγκεκριμένα σημεία ή υποσύνολα σημείων. Οι ίδιοι ερευνητές (1998 και 2004) προτείνουν τον υπολογισμό του δείκτη *CTI* (*Intra Contribution*) για την ερμηνεία των αντιθέσεων μεταξύ σημείων επί των παραγοντικών αξόνων. Ο δείκτης αυτός εκφράζει τη συνεισφορά της αντίθεσης, με την έννοια της απόκλισης, μεταξύ δύο σημείων στην αδράνεια (διακύμανση) του άξονα. Ο υπολογισμός του δείκτη *CTI* μπορεί να επεκταθεί και για την ερμηνεία των αντιθέσεων μεταξύ ομάδων σημείων.

Πάντως, η προσπάθεια απόδοσης απλής φυσικής ερμηνείας και λεκτικής περιγραφής στους παράγοντες, αν τους θεωρήσουμε ως νέες σύνθετες και λανθάνουσες μεταβλητές, απαιτεί, αφενός, καλή γνώση του ερευνητικού και θεωρητικού πεδίου στο πλαίσιο του οποίου θα ερμηνευτούν τα αποτελέσματα (Kachigan 1991, Hair *et al*. 1995) και, αφετέρου, αρκετή εμπειρία (Παπαδημητρίου, 1994). Η όλη προσπάθεια, από ένα σημείο και μετά, ξεφεύγει από το χώρο της Στατιστικής και καθίσταται μάλλον “τέχνη” (*art*) (Manly, 1994).

2.2.14.4 Πρόταση Μεθόδου Καθορισμού των Σημαντικών Κελιών του Πίνακα Συμπτώσεων

Με βάση τη μεθοδολογία που παρουσίασε χωρίς απόδειξη ο Van de Geer (1993β), προτείνουμε και στηρίζουμε με απόδειξη μια τροποποιημένη εκδοχή της, σύμφωνα με την οποία οι τυποποιημένες συντεταγμένες r_{is} και c_{js}, των σημείων γραμμών και στηλών αντίστοιχα, μπορούν να χρησιμοποιηθούν για τον εντοπισμό των κελιών του πίνακα συμπτώσεων που συνεισφέρουν στην αδράνεια των αξόνων. Η βασική ιδέα της εκδοχής, που προτείνουμε, είναι να διασπάσουμε την ολική αδράνεια του πίνακα συμπτώσεων σε συνεισφορές των κελιών του πίνακα σε κάθε άξονα. Πιο συγκεκριμένα:

Έστω **F** ο πίνακας απολύτων συχνοτήτων δύο κατηγορικών μεταβλητών *X* και *Y*, με *k* και *l* κλάσεις αντίστοιχα, και γενικό σύνολο *N*. Συμβολίζουμε με **P** τον πίνακα αντιστοιχιών και με **r** και **c** τα διανύσματα με στοιχεία τις μάζες των γραμμών και στηλών αντίστοιχα. Αν $\mathbf{S}=N(\mathbf{P}-\mathbf{r}\mathbf{c}^{\mathrm{T}})$ είναι ο πίνακας με στοιχεία τις διαφορές μεταξύ των παρατηρούμενων και των αναμενόμενων συχνοτήτων του πίνακα $\mathbf{F}_{k\times l}$, τότε, όπως θα δείξουμε στη συνέχεια, ο **S** μπορεί να διασπαστεί σε *p* πίνακες $\mathbf{\Omega}^{(s)}_{k\times l}$ (s=1,…,p, p=min{k-1, l-1}), έτσι ώστε το άθροισμα των στοιχείων ω^s_{ij} κάθε πίνακα $\mathbf{\Omega}^{(s)}$ να είναι ίσο με την αδράνεια του άξονα *s*. Δηλαδή:

$$\sum_{i=1}^{k}\sum_{j}^{l}\omega_{ij}^{s}=\lambda_{s}\,.$$

Απόδειξη:

Δημιουργούμε αρχικά τους πίνακες $\mathbf{B}^{(s)}$ (s=1,…, p) με στοιχεία:

$$b_{ij}^{s}=Nr_{i}c_{j}\sqrt{\lambda_{s}}\;r_{is}c_{js},\ i=1,\ldots,k \text{ και } j=1,\ldots,l, \qquad [2.52]$$

όπου r_i είναι η μάζα της γραμμής *i* και c_j η μάζα της στήλης *j*.

Το γενικό στοιχείο b^s_{ij} προκύπτει από τη σχέση [2.40] ως εξής:

$$p_{ij} = r_i c_j \left(1 + \sum_{s=1}^{p} \sqrt{\lambda_s}\ r_{is} c_{js} \right) \Rightarrow N\left(p_{ij} - r_i c_j \right) = N r_i c_j \sum_{s=1}^{p} \sqrt{\lambda_s}\ r_{is} c_{js},$$

από όπου:

$$\sum_{s=1}^{p} N r_i c_j \sqrt{\lambda_s}\ r_{is} c_{js} = s_{ij}\ . \qquad [2.53]$$

Παρατηρούμε ότι τα στοιχεία s_{ij}, του πίνακα **S,** τα οποία εκφράζουν τις αποκλίσεις από την κατάσταση ανεξαρτησίας μπορούν να δοθούν και από τη σχέση [2.53]. Αν θέσουμε $b_{ij}^s = N r_i c_j \sqrt{\lambda_s}\ r_{is} c_{js}$, τότε $s_{ij} = \sum_{s=1}^{p} b_{ij}^s$ και, συνεπώς, ο πίνακας **S** μπορεί να γραφεί ως άθροισμα των πινάκων $\mathbf{B}_{k\times l}^{(s)}$ που έχουν ως στοιχεία τα b_{ij}^s. Δηλαδή:

$$\mathbf{S} = \sum_{s=1}^{p} \mathbf{B}^{(s)}\ .$$

Επομένως, έχουμε πετύχει μια διάσπαση, ανά παραγοντικό άξονα, των αποκλίσεων από την κατάσταση ανεξαρτησίας. Υψηλές τιμές των στοιχείων b_{ij}^s δηλώνουν σύνδεση ή αλληλεπίδραση της γραμμής *i* και της στήλης *j* του πίνακα **F**, η οποία αναδεικνύεται επί του άξονα *s* (Van de Geer, 1993β). Το πρόσημό τους μπορεί να ερμηνευτεί με τον ίδιο τρόπο όπως και στην περίπτωση των τυποποιημένων υπολοίπων του πίνακα **F** (βλέπε Ενότητα 2.2.5).

Στη συνέχεια, κατασκευάζουμε τους πίνακες $\mathbf{W}^{(s)}$ με στοιχεία:

$$w_{ij}^s = \frac{\left(b_{ij}^s \right)^2}{N r_i c_j}\ . \qquad [2.54]$$

Για τα στοιχεία w_{ij}^s ισχύει:

$$\sum_i \sum_j w_{ij}^s = \sum_i \sum_j \frac{\left(b_{ij}^s \right)^2}{N r_i c_j} = \sum_i \sum_j \frac{\left(N r_i c_j \sqrt{\lambda_s}\ r_{is} c_{js} \right)^2}{N r_i c_j} =$$

$$= \sum_i \sum_j N r_i c_j \lambda_s r_{is}^2 c_{js}^2 = N \lambda_s \sum_i \sum_j r_i c_j r_{is}^2 c_{js}^2 =$$

$$= N\lambda \sum_i r_i r_{is}^2 \sum_j c_j c_{js}^2 .$$

Αν λάβουμε υπόψη τις σχέσεις $\sum_j c_j c_{js}'^2 = 1$ (σχέση [2.27]) και $\sum_i r_i r_{is}'^2 = 1$ (σχέση [2.28]), τότε:

$$\sum_i \sum_j w_{ij}^s = N\lambda_s . \qquad [2.55]$$

Από τη [2.55] προκύπτει ότι:

$$\sum_s \sum_i \sum_j w_{ij}^s = \sum_s N\lambda_s = N\sum_s \lambda_s = NI = Q, \qquad [2.56]$$

όπου I η ολική αδράνεια του **F** και Q το στατιστικό χ^2 που υπολογίζεται κάτω από την υπόθεση της ανεξαρτησίας των μεταβλητών X και Y (βλέπε Ενότητες 2.1.2.5 και 2.1.2.14.2).

Από τη [2.56] μπορούμε να συμπεράνουμε ότι τα στοιχεία των πινάκων $\mathbf{W}^{(s)}$ δηλώνουν τις συνεισφορές των κελιών στο στατιστικό χ^2 ανά άξονα. Στη συνέχεια εκφράζουμε τις συνεισφορές αυτές σε σχέση με τις αδράνειες των αξόνων.

Από τη [2.55], προφανώς, ισχύει:

$$\frac{1}{N}\sum_i \sum_j w_{ij}^s = \lambda_s .$$

Αν θέσουμε τώρα $\omega_{ij}^s = (1/N) w_{ij}^s$, τότε οι p σε πλήθος πίνακες $\mathbf{\Omega}^{(s)}=(1/N)\mathbf{W}^{(s)}$ είναι διαστάσεων $k\times l$ και το άθροισμα των στοιχείων τους είναι ίσο με τη αδράνεια του αντίστοιχου άξονα. □

Αν το στοιχείο ω_{ij}^s έχει υψηλή τιμή σε σχέση με τα υπόλοιπα, τότε το κελί (i, j) συνεισφέρει σημαντικά στην αδράνεια του άξονα s. Έτσι, η μελέτη των στοιχείων των πινάκων $\mathbf{\Omega}^{(s)}$ μπορεί να αναδείξει τα σημεία γραμμών και στηλών που συνεισφέρουν ταυτόχρονα στην αδράνεια του κάθε άξονα και, κατ' επέκταση, στη συσχέτιση των δύο μεταβλητών. Για να διευκολυνθούμε περισσότερο στην ερμηνεία

των αποτελεσμάτων μπορούμε να εκφράσουμε τα στοιχεία των πινάκων $\mathbf{\Omega}^{(s)}$ με σχετικές τιμές, ως ποσοστά (%) των αδρανειών των αντίστοιχων αξόνων. Με την προτεινόμενη μεθοδολογία εντοπίζονται τα σημαντικά κελιά ανά άξονα, δηλαδή τα σημεία γραμμών και στηλών που αλληλεπιδρούν και συνεισφέρουν από κοινού στην αδράνεια του αντίστοιχου άξονα. Μάλιστα, η κοινή συνεισφορά τους εκφράζεται ως ποσοστό της αδράνειας του αντίστοιχου άξονα. Η επιλογή μεμονωμένων σημείων γραμμών ή στηλών του πίνακα **F** με βάση το δείκτη *CTR* (ή/και *COR*) εγκυμονεί τον κίνδυνο κάποια από αυτά να μη ληφθούν υπόψη κατά την ερμηνεία των αποτελεσμάτων λόγω χαμηλών τιμών των αντίστοιχων δεικτών.

Στην Ενότητα Β1 του Παραρτήματος Β δίνουμε ένα αριθμητικό παράδειγμα εφαρμογής της προτεινόμενης μεθόδου.

2.2.14.5 Μερικές Επισημάνσεις

Στην ενότητα αυτή παραθέτουμε πρόσθετα πληροφοριακά στοιχεία σχετικά με την ερμηνεία των αποτελεσμάτων που παράγονται από την ΠΑΑ.

Α) Ερμηνεία των γραφικών αποτελεσμάτων της ΠΑΑ

Αν συνθέσουμε τα ευρήματα των προηγούμενων ενοτήτων μπορούμε να διακρίνουμε τρεις τουλάχιστον γενικές μεθοδολογικής προσεγγίσεις για την ερμηνεία των γραφικών αποτελεσμάτων της ΠΑΑ:

Πρώτη: Δημιουργούνται δύο τύποι συμμετρικών παραγοντικών διαγραμμάτων. Ο ένας τύπος κατασκευάζεται με βάση τη Συμμετρική Κανονικοποίηση *SN* και ο άλλος τύπος με την Κύρια Κανονικοποίηση *PN*. Με τον πρώτο τύπο, που όπως είδαμε είναι *biplot*, είναι δυνατή η ερμηνεία των συνδέσεων – σχέσεων μεταξύ των σημείων γραμμών και στηλών, ενώ με το δεύτερο (*french plot*) είναι δυνατή η ερμηνεία των αποστάσεων είτε μόνο μεταξύ των σημείων γραμμών είτε μόνο μεταξύ των σημείων στηλών. Ο διαχωρισμός αυτός είναι περισσότερο τυπικός, αφού αν παρατηρήσουμε τα αντίστοιχα διαγράμματα (βλέπε Διαγράμματα 2.3 και 2.4) θα διαπιστώσουμε ότι οι σχετικές θέσεις των σημείων επί του παραγοντικού επιπέδου είναι ίδιες. Αλλάζουν μόνο οι αριθμητικές τιμές των συντεταγμένων των σημείων γραμμών και στηλών.

Δεύτερη: Κατασκευάζονται δύο μη συμμετρικά παραγοντικά διαγράμματα. Το πρώτο δημιουργείται με βάση την Κύρια Κανονικοποίηση κατά Γραμμές *RPN*, το οποίο διατηρεί τις αποστάσεις χ^2 μεταξύ των γραμμών. Τα προφίλ των γραμμών προβάλλονται στο χώρο (*simplex*) των στηλών και μπορούν να ερμηνευτούν οι αποστάσεις μεταξύ των σημείων γραμμών πάνω στο παραγοντικό επίπεδο. Οι συντεταγμένες των προβολών τους, μέσω της ΠΑΑ, έχουν κανονικοποιηθεί έτσι ώστε οι μεταξύ τους ευκλείδειες αποστάσεις να προσεγγίζουν τις αρχικές χ^2 αποστάσεις. Το δεύτερο διάγραμμα δημιουργείται με βάση την Κύρια Κανονικοποίηση κατά Στήλες *CPN*, στο οποίο διατηρούνται οι αποστάσεις χ^2 μεταξύ των στηλών. Τα προφίλ των στηλών προβάλλονται στο χώρο (*simplex*) των γραμμών και μπορούν να ερμηνευτούν οι αποστάσεις μεταξύ των σημείων στηλών. Οι δύο τύποι διαγραμμάτων είναι *biplot* και επομένως, σύμφωνα με όσα αναφέρθηκαν στις Ενότητες 2.2.13 και 2.2.14.2, είναι δυνατή η ερμηνεία και των σχέσεων σύνδεσης μεταξύ των σημείων γραμμών και στηλών. Η φυσική ερμηνεία και η ταυτότητα των παραγοντικών αξόνων μπορεί να δοθεί με βάση είτε τα σημεία γραμμών είτε τα σημεία στηλών. Στη συνέχεια, μπορεί να δημιουργηθεί ένα διάγραμμα που να απεικονίζει μόνο τα σημεία στηλών (αντίστοιχα γραμμών) πάνω στους άξονες που ήδη έχουμε ερμηνεύσει ως προς τα σημεία γραμμών (αντίστοιχα στηλών) (Greenacre 1993α, Clausen 1998, Μαυρομάτης 1999, Bendixen 2003). Η μέθοδος αυτή ονομάζεται «αναλυτική κατά παράγοντες ερμηνεία» (*dimensional* ή *factor-analytic interpretation*) (Greenacre, 1993α, Bacher 1995) και μπορεί να εφαρμοστεί και στην πρώτη προσέγγιση.

Τρίτη: Σύμφωνα με το Micheloud (1997), ένας τρόπος για να ερμηνεύσουμε το συμμετρικό παραγοντικό διάγραμμα που προκύπτει από την Κύρια Κανονικοποίηση *PN*, δηλαδή στο πλαίσιο της Γαλλικής Σχολής, είναι, για παράδειγμανα, να εξετάσουμε τη γωνία θ που σχηματίζουν το ευθύγραμμο τμήμα που ενώνει ένα σημείο γραμμής με την αρχή των αξόνων και το ευθύγραμμο τμήμα που ενώνει ένα σημείο στήλης με την αρχή των αξόνων. Αν η γωνία θ είναι μικρότερη από 90^o, τότε τα δύο χαρακτηριστικά, που αντιπροσωπεύουν τα δύο σημεία, είναι μάλλον θετικά συσχετισμένα μεταξύ τους (έλκονται). Αν η γωνία είναι μεγαλύτερη από 90^o, τότε τα δύο χαρακτηριστικά είναι μάλλον αρνητικά συσχετισμένα (απωθούνται). Τέλος, αν η γωνία είναι ορθή, τότε τα δύο χαρακτηριστικά μάλλον δεν αλληλεπιδρούν. Αυτός ο

τρόπος ερμηνείας μπορεί να αιτιολογηθεί μέσω τη σχέσης [2.46] (Andersen, 1991) ως εξής:

Για το παραγοντικό επίπεδο 1×2, από τη [2.46] έχουμε:

$$p_{ij} \approx r_i c_j \left(1+\sum_{s=1}^{2}\frac{\varphi_{is}\gamma_{js}}{\sqrt{\lambda_s}}\right) \Rightarrow \frac{p_{ij}}{r_i c_j} - 1 \approx \sum_{s=1}^{2}\frac{\varphi_{is}\gamma_{js}}{\sqrt{\lambda_s}}.$$

Η παραπάνω σχέση εκφράζει ότι η διαφορά μεταξύ του λόγου $\frac{p_{ij}}{r_i c_j}$ και του αριθμού 1, που είναι η αναμενόμενη τιμή του λόγου κάτω από την υπόθεση της ανεξαρτησίας μεταξύ των δύο μεταβλητών, είναι ανάλογη με το συνημίτονο της γωνίας που σχηματίζουν το διάνυσμα θέσης της γραμμής *i* και της στήλης *j*. Συνεπώς, όσο μικρότερη είναι η γωνία τόσο μεγαλύτερη είναι η διαφορά μεταξύ της παρατηρούμενης σχετικής συχνότητας p_{ij} και της αναμενόμενης $r_i c_j$ κάτω από την υπόθεση της ανεξαρτησίας των δύο μεταβλητών. Αν το συνημίτονο είναι ίσο με μηδέν, τότε η παρατηρούμενη σχετική συχνότητα p_{ij} είναι ίση με την αναμενόμενη $r_i c_j$. Αυτό σημαίνει ότι τα αντίστοιχα διανύσματα θέσης είναι κάθετα μεταξύ τους με αποτέλεσμα το κελί (*i*, *j*) να μη συνεισφέρει στην εξάρτηση ή τη συσχέτιση των δύο μεταβλητών. Μάλιστα, ο Καραπιστόλης (1999) ορίζει το λόγο $d_{ij} = \frac{p_{ij}}{r_i c_j}$ ως «δείκτη έλξης-άπωσης» μιας γραμμής *i* και μιας στήλης *j*. Η σχετική θέση των σημείων *i* και *j* μπορεί να χαρακτηριστεί ως εξής (Snelders & Stokmans 1994, Clausen 1998, Καραπιστόλης 1999):

- Αν $\frac{p_{ij}}{r_i c_j} > 1$, τότε υπάρχει «έλξη» μεταξύ των σημείων *i* και *j* και, αφού η παρατηρούμενη συχνότητα είναι μεγαλύτερη από την αναμενόμενη, τότε $cos(\theta) > 0$. Αν η ποιότητα προβολής στο παραγοντικό επίπεδο είναι υψηλή, τότε τα σημεία που αντιστοιχούν στη γραμμή *i* και στη στήλη *j* θα βρίσκονται σχετικά κοντά το ένα με το άλλο.

- Αν $\frac{p_{ij}}{r_i c_j}$<1, τότε υπάρχει «άπωση» μεταξύ των σημείων *i* και *j* και, επειδή η παρατηρούμενη συχνότητα είναι μικρότερη από την αναμενόμενη, το $cos(\theta)$<0. Αντίστοιχα, αν η ποιότητα προβολής στο παραγοντικό επίπεδο είναι υψηλή, τότε τα σημεία που αντιστοιχούν στη γραμμή *i* και στη στήλη *j* θα βρίσκονται σε απομακρυσμένες θέσεις.

- Αν $\frac{p_{ij}}{r_i c_j}$=1, τότε η παρατηρούμενη συχνότητα είναι ίση με την αναμενόμενη ($cos(\theta)$=0) και το αντίστοιχο κελί δεν συνεισφέρει στην αλληλεπίδραση της μεταβλητής γραμμής και της μεταβλητής στήλης.

Η παραπάνω προσέγγιση μπορεί να συνδυαστεί και με τη μεθοδολογία που προτείναμε στην Ενότητα 2.2.14.4, ώστε να εντοπιστούν και τα ζεύγη της μορφής (γραμμή *i*, στήλη *j*) που συνεισφέρουν και είναι σημαντικά σε κάθε παραγοντικό άξονα.

Β) Συμπληρωματικότητα της ΠΑΑ με Μεθόδους Ταξινόμησης

Συχνά η ΠΑΑ συνδυάζεται με μεθόδους Ταξινόμησης (Lebart, Marineau & Warwick 1984, Benzécri 1992, Lebart & Mirkin 1993, Greenacre 1993α και 1988β, Lebart 1994, Bendixen 1995, Κιοσέογλου 1997, Μπεχράκης 1999, Καραπιστόλης 1999, Lebart, Marineau & Piron 2000, Γναρδέλλης & Κουλιεράκης 2002, Αθανασιάδης 2002 και 1995, Le Roux & Rouanet 2004, Καρλής 2005), οι οποίες καθιστούν δυνατή την ομαδοποίηση σημείων με όμοια χαρακτηριστικά ή ιδιότητες (βλέπε Aldenderfer & Blashfield 1984, Everitt 1993, Hair *et al*. 1995, Sharma 1996). Σε πολλές περιπτώσεις αναδεικνύεται η συμπληρωματικότητα των δύο μεθόδων στην ερμηνεία των αποτελεσμάτων (Lebart & Mirkin 1993, Lebart 1994, Bacher 1995, Lebart, Morineau & Piron 2000, Παπαδημητρίου & Αθανασίου 2003, Καρλής 2005). Τα αποτελέσματα της ΠΑΑ και της Ταξινόμησης (ομάδες σημείων) είναι δυνατό να παρουσιαστούν ταυτόχρονα επί των παραγοντικών επιπέδων είτε με την ενσωμάτωση του δενδρογράμματος στα διαγράμματα της ΠΑΑ (βλέπε Escofier & Pagès 1998, Lebart, Morineau & Piron 2000, Δρόσος 2005) είτε με τη σχεδίαση χωρίων επί των παραγοντικών επιπέδων που να περικλείουν τα σημεία που ανήκουν στην ίδια ομάδα (βλέπε De Lagarde 1995, Μασούρα & Παπαδημητρίου 2002, Μπαγιάτης &

Παπαδημητρίου 2002, Μάλλιαρη 2004, Παπαδημητρίου & Μάρκος 2004, Τζήμος & Παπαδημητρίου 2004, Αναστασιάδου & Κοσμά 2004 και Διάγραμμα Β2.1 της Ενότητας Β2 του Παραρτήματος Β). Μια άλλη συνηθισμένη πρακτική είναι η προβολή των κέντρων των ομάδων ως συμπληρωματικά σημεία στα παραγοντικά επίπεδα (Benzécri 1992, Μπεχράκης 1999, Καραπιστόλης 1999). Η συνδυασμένη εφαρμογή της ΠΑΑ με μεθόδους Ταξινόμησης είναι μάλλον επιβεβλημένη ιδιαίτερα στην περίπτωση κατά την οποία, επί του παραγοντικού επιπέδου 1×2, η ποιότητα προβολής του υπό εξέταση φαινομένου δεν είναι ικανοποιητική. Άλλωστε, δύο σημεία τα οποία προβάλλονται σε κοντινές θέσεις στο επίπεδο μπορεί να απέχουν σημαντικά σε χώρο τριών ή περισσοτέρων διαστάσεων. Από μια άλλη σκοπιά, οι μέθοδοι Ταξινόμησης, αν και συμβάλλουν στην ανάδειξη ομοιογενών ομάδων σημείων αντικειμένων (συνήθως των γραμμών του πίνακα συμπτώσεων που θα δοθεί ως είσοδος στην ΠΑΑ), ωστόσο δεν προσφέρονται για τη διάταξη των σημείων αυτών ή των ομάδων τους επί των παραγοντικών αξόνων (Παπαδημητρίου & Αθανασίου, 2003). Η δυνατότητα αυτή παρέχεται μόνο μέσω της ΠΑΑ.

Τα αποτελέσματα της ΠΑΑ είναι δυνατό να συνδυαστούν και με μεθόδους οπτικοποίησης πολυδιάστατων δεδομένων, όπως είναι για παράδειγμα οι καμπύλες *Andrews* (Andrews, 1972). Οι μέθοδοι αυτές μπορούν να χρησιμοποιηθούν για τη γραφική σύγκριση και ομαδοποίηση των προφίλ γραμμών ή στηλών του πίνακα συμπτώσεων. Στην Ενότητα Β2 του Παραρτήματος Β παρουσιάζουμε ένα παράδειγμα συνδυασμού της ΠΑΑ με τις καμπύλες *Andrews* και την Ταξινόμηση σε Αύξουσα Ιεραρχία.

Γ) Η Τετριμμένη Ιδιοτιμή $\lambda_0=1$

Από τις σχέσεις [2.40], [2.41], [2.42] και [2.43] διαπιστώνουμε ότι για την ανασύσταση του αρχικού πίνακα συμπτώσεων **F** είναι απαραίτητη και η χρήση της τετριμμένης ιδιοτιμής $\lambda_0 = 1$, η οποία αντιστοιχεί στην κατάσταση ανεξαρτησίας των δύο μεταβλητών (βλέπε Παρατήρηση 2.1). Μπορούμε να θεωρήσουμε ότι η πληροφορία που περιέχεται στον πίνακα **F** είναι δυνατό να αναλυθεί σε δύο τμήματα: α) σε ένα που περιέχει την πληροφορία, η οποία οφείλεται στις αποκλίσεις από την κατάσταση ανεξαρτησίας και μετριέται μέσω της ολικής αδράνειας του **F**, και β) στο τμήμα που περιέχει την πληροφορία που αντιστοιχεί στην κατάσταση ανεξαρτησίας

και η οποία, μέσω της ΠΑΑ, τυποποιείται στην τιμή 1. Με βάση το συλλογισμό αυτό, οι Weller και Romney (1990) προτείνουν να λαμβάνεται υπόψη και η ιδιοτιμή 1 για τον έλεγχο της ποιότητας της λύσης της ΠΑΑ Για παράδειγμα, ας υποθέσουμε ότι κατά την εφαρμογή της ΠΑΑ σε έναν 3×3 πίνακα συμπτώσεων δύο μεταβλητών οι δύο παραγοντικοί άξονες έχουν ιδιοτιμές λ_1=0,02 και $\lambda_2 = 0{,}0001$ αντίστοιχα. Η ολική αδράνεια I του **F** είναι ίση με $\lambda_1 + \lambda_2 \Rightarrow I$=0,02+0,0001=0,0201. Ο πρώτος άξονας ερμηνεύει το $(0{,}02/0{,}0201)\times 100 = 99{,}5\%$ της ολικής αδράνειας, ενώ ο δεύτερος το $(0{,}0001/0{,}0201)\times 100 = 0{,}5\%$. Το ολικό πληροφοριακό περιεχόμενο του **F** είναι ίσο με 1+0,0201=1,0201. Παρόλο που ο πρώτος άξονας ερμηνεύει σχεδόν όλη την πληροφορία που αντιστοιχεί στις αποκλίσεις από την κατάσταση ανεξαρτησίας ωστόσο το $\{1/(1+0{,}0201)\}\times 100$ =98,3% της ολικής πληροφορίας του πίνακα **F** οφείλεται στην ανεξαρτησία των δύο μεταβλητών.

Αξίζει να αναφερθεί ότι έχουν προταθεί αρκετοί δείκτες καθορισμού της ποσότητας πληροφορίας που ερμηνεύουν οι άξονες και τα παραγοντικά επίπεδα της ΠΑΑ (βλέπε Nishisato 1994, Gabriel 2002). Όμως, οι δείκτες αυτοί εξαρτώνται είτε από τον τύπο και τις διαστάσεις του πίνακα που θα δοθεί ως είσοδος στην ανάλυση είτε από τη μέθοδο κανονικοποίησης των συντεταγμένων των προβολών των σημείων γραμμών και στηλών στους παραγοντικούς άξονες. Πάντως, σε πρακτικό επίπεδο, ο λόγος της αδράνειας ενός άξονα προς την ολική αδράνεια εκφράζει την αναλογία της μεταβλητότητας των σημείων γύρω από το κέντρο βάρους, η οποία εξηγείται από τον αντίστοιχο άξονα και ως τέτοια θα πρέπει να ερμηνεύεται (Μπεχράκης, 1999).

Δ) Χρήση Άλλων Αποστάσεων και Γενικεύσεις της ΠΑΑ

Κατά καιρούς έχουν προταθεί ποικίλες "γενικεύσεις" της μεθόδου με τη χρήση άλλων αποστάσεων πέραν της χ^2 (βλέπε Novak & Hoffman 1990, Rao 1995, Nishisato 1995, Cuadras & Cuadras 2006). Ενδεικτικά αναφέρουμε τις αποστάσεις *Hamming* και *Hellinger*. Στην Ιταλική Σχολή, η Μη Συμμετρική Ανάλυση των Αντιστοιχιών (Lauro & Balbi 1999, Kroonenberg & Lombardo 1999) στηρίζεται στην ανάλυση του συντελεστή συνάφειας τ των Goodman & Kruskal (1954) και όχι στο συντελεστή συνάφειας μέσου τετραγώνου φ^2 του Pearson, που όπως είδαμε στην Ενότητα 2.2.5

είναι ίσος με την ολική αδράνεια του πίνακα συμπτώσεων. *Κατά την άποψή μας η ΠΑΑ είναι άρρηκτα συνδεδεμένη με την απόσταση* χ^2. *Η χρήση οποιασδήποτε άλλης μετρικής την καθιστά διαφορετική μέθοδο και όχι απλά μια γενίκευση.*

2.3 Η Περίπτωση Πολλών Μεταβλητών

Στην πολυμεταβλητή περίπτωση θεωρούμε ότι ο αρχικός πίνακας δεδομένων, έστω **D**, έχει διαστάσεις $N \times q$ και είναι της μορφής «*αντικείμενα* × *μεταβλητές*». Για παράδειγμα, τα στοιχεία του **D** θα μπορούσαν να αντιστοιχούν στις απαντήσεις N ατόμων σε q ερωτήσεις ενός ερωτηματολογίου. Στο πλαίσιο της Γαλλικής Σχολής Ανάλυσης Δεδομένων η πολυμεταβλητή εκδοχή της ΠΑΑ (*Analyse des Correspondances Multiples* στα Γαλλικά και *Multiple Correspondence Analysis* στα Αγγλικά) εφαρμόζεται αλγοριθμικά και εννοιολογικά ακριβώς με τον ίδιο τρόπο όπως και στην περίπτωση των δύο μεταβλητών (Lebart, Morineau & Tabard 1977, Lebart, Morineau & Warwick 1984, Israëls 1987, Benzécri 1992, Escofier & Pagès 1998, Καραπιστόλης 1999, Lebart, Morineau & Piron 2000, Le Roux & Rouanet 2004, Murtagh 2005, Παπαδημητρίου 2006, 2004 και 1994). Εκείνο που αλλάζει είναι ο πίνακας που θα δοθεί ως είσοδος στην ανάλυση. Πιο συγκεκριμένα, όταν έχουμε q κατηγορικές μεταβλητές η μέθοδος μπορεί να εφαρμοστεί είτε στον «Πίνακα Λογικής Περιγραφής» είτε στον αντίστοιχο «Γενικευμένο Πίνακα Συμπτώσεων» των q μεταβλητών (Israëls 1987, Benzécri 1992, Παπαδημητρίου 2006, 2004 και 1994). Συνήθως εφαρμόζεται στο δεύτερο πίνακα για υπολογιστική ευκολία. Στην Ολλανδική Σχολή η μέθοδος είναι γνωστή ως «Ανάλυση Ομοιογένειας με Εναλλασσόμενα Ελάχιστα Τετράγωνα» (*Homogeneity Analysis by Alternating Least Squares-HOMALS*) (Greenacre 1993α, Van de Geer 1993α και 1993β, Heiser & Meulman 1994, Gifi 1996, Michailidis 1996, SPSS Inc. 1998, Michailidis & De Leeuw 1998) και αποτελεί, όπως είδαμε στο Κεφάλαιο 1, τη βασική μέθοδο πάνω στην οποία αναπτύχθηκε το σύστημα *GIFI*. Αντιμετωπίζεται ως πρόβλημα βελτιστοποίησης με δεσμεύσεις και υλοποιείται υπολογιστικά μέσω του επαναληπτικού αλγόριθμου *Alternating Least Squares* (Gifi 1996, De Leeuw 1993, Michailidis 1996, Michailidis & De Leeuw 1998). Τα αποτελέσματα των δύο Σχολών, αν και είναι συγκρίσιμα, διαφέρουν ριζικά, όχι μόνο στην προβληματική αλλά και στο θεωρητικό και υπολογιστικό υπόβαθρο της μεθόδου. Στη Γαλλική Σχολή βασικός σκοπός της ΠΑΑ είναι η διερεύνηση, η ανάδειξη και η οπτικοποίηση των σχέσεων μεταξύ των κατηγορικών μεταβλητών που συμμετέχουν στην ανάλυση. Στο πλαίσιο της Ολλανδικής Σχολής πρωταρχικός στόχος είναι η ανάθεση αριθμητικών βαρών (*weights*) στις κλάσεις (κατηγορίες) των μεταβλητών και βαθμών (*scores*) στα

αντικείμενα, ώστε να μεγιστοποιείται η ομοιογένεια (Van Rijckevorsel 1987, Greenacre 1993α, Van de Geer 1993α και 1993β, Gifi 1996, Michailidis 1996, SPSS Inc. 1998, Michailidis & De Leeuw 2000 και 1998, Meulman & Heiser 2004), με την έννοια της εσωτερικής συνέπειας (Nishisato 1980, Spector 1992, Traub 1994, Strub 2000), μεταξύ των μεταβλητών. Η ομοιογένεια έγκειται στην ταυτόχρονη ανάθεση αριθμητικών τιμών στις κατηγορίες (ιδιότητες) των μεταβλητών και στα αντικείμενα, έτσι ώστε η τιμή κάθε αντικειμένου να είναι όσο το δυνατό πιο "όμοια" (αντιπροσωπευτική, συνεπής) με τις τιμές των ιδιοτήτων που το χαρακτηρίζουν. Η ομοιότητα καθορίζεται ως ελαχιστοποίηση ενός αθροίσματος διαφορών εις το τετράγωνο με συνέπεια το αρχικό πρόβλημα να καθίσταται τελικά πρόβλημα ελαχίστων τετραγώνων (βλέπε Ενότητα 2.3.4). Η μέθοδος εντάσσεται στις τεχνικές βέλτιστης κλιμάκωσης που χαρακτηρίζουν την Ολλανδική Σχολή και μπορεί να θεωρηθεί ως ένας μετασχηματισμός ποιοτικών δεδομένων σε ποσοτικά (Young, 1981). Η Ανάλυση Ομοιογένειας, ως τεχνική βέλτιστης κλιμάκωσης, δίνει συγκρίσιμα αποτελέσματα και με την πολυμεταβλητή εκδοχή της «Δυϊκής Κλιμάκωσης» (*Dual Scaling*) (βλέπε Nishisato, 1994 και 1980).

Στη συνέχεια, θα παρουσιάσουμε τους βασικούς πίνακες που μπορούν να δοθούν ως είσοδος στην ΠΑΑ

2.3.1 Λογικοί Πίνακες 0-1

Έστω **D** ο $N \times q$ πίνακας δεδομένων της μορφής «*αντικείμενα* × *μεταβλητές*». Συμβολίζουμε με $X_1, X_2, \ldots, X_q$, τις q σε πλήθος κατηγορικές μεταβλητές με $j_1, j_2, \ldots, j_q$ κλάσεις (ιδιότητες) αντίστοιχα. Ας είναι j ο συνολικός αριθμός των ιδιοτήτων των q μεταβλητών, δηλαδή $j = \sum_{i=1}^{q} j_i$. Για κάθε μεταβλητή X_i (i=1,…, q) κατασκευάζουμε τον $N \times j_i$ πίνακα $\mathbf{Z_i}$, ο οποίος έχει ως στοιχεία μόνο τους αριθμούς 0 και 1 και οι στήλες του αντιστοιχούν στις κατηγορίες της μεταβλητής X_i. Για κάθε αντικείμενο n (n=1,…, N), που χαρακτηρίζεται από την ιδιότητα δ_v^i (v=1,…, j_i), η αντίστοιχη στήλη του πίνακα $\mathbf{Z_i}$ παίρνει την τιμή 1, ενώ οι υπόλοιπες την τιμή 0. Οι πίνακες $\mathbf{Z_i}$ συχνά αποκαλούνται «Λογικοί Πίνακες 0-1» ή «Πίνακες Σχεδιασμού». Στη συνέχεια, τοποθετούμε τους q πίνακες $\mathbf{Z_i}$ τον ένα δίπλα στον άλλο, οπότε

παίρνουμε τον $N \times q$ πίνακα $\mathbf{Z_{0\text{-}1}}$ που ονομάζεται «Πίνακας Πλήρους Διαζευκτικής Περιγραφής» με λογική κωδικοποίηση 0-1 (Lebart, Marineau & Warwick 1984, Παπαδημητρίου 2006 και 2004). Δηλαδή,

$$\mathbf{Z_{0\text{-}1}} = \left[\mathbf{Z_1} \,|\, \mathbf{Z_2} \,|\, \cdots \,|\, \mathbf{Z_q} \right].$$

Ο πίνακας $\mathbf{Z_{0\text{-}1}}$ έχει το ίδιο πληροφοριακό περιεχόμενο με τον αρχικό πίνακα δεδομένων **D**. Η διαφορά έγκειται στον αριθμό των στηλών και στην λογική κωδικοποίηση των δεδομένων όπου χρησιμοποιούνται μόνο οι αριθμοί 0 και 1. Στη Γαλλική Σχολή ο πίνακας $\mathbf{Z_{0\text{-}1}}$ αντιμετωπίζεται ως πίνακας διπλής εισόδου της μορφής «*αντικείμενα* × *ιδιότητες*» και μπορεί να δοθεί ως είσοδος στην ΠΑΑ. Στην περίπτωση αυτή, ο αλγόριθμος της ΠΑΑ που περιγράψαμε στην Ενότητα 2.2.14 (διμεταβλητή περίπτωση) εφαρμόζεται ακριβώς με τον ίδιο τρόπο αν στη θέση του **F** αντικαταστήσουμε τον πίνακα $\mathbf{Z_{0\text{-}1}}$.

Παράδειγμα Κατασκευής του Πίνακα $\mathbf{Z_{0\text{-}1}}$.

Ας υποθέσουμε ότι έχουμε στη διάθεσή μας στοιχεία για 5 αντικείμενα, για τα οποία έχουμε καταγράψει ποιοτικές μετρήσεις σε 3 μεταβλητές X_1, X_2 και X_3. Έστω ότι η μεταβλητή X_1 έχει 3 κατηγορίες και οι μεταβλητές X_2 και X_3 από δύο η κάθε μια. Ο πίνακας δεδομένων **D** θα μπορούσε να περιέχει τα παρακάτω στοιχεία:

$$\mathbf{D} = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 2 & 2 \\ 3 & 2 & 1 \\ 2 & 1 & 2 \end{bmatrix}.$$

Θα πρέπει να τονιστεί ότι οι αριθμοί που εμφανίζονται ως στοιχεία του **D** αντιστοιχούν στις διαφορετικές κατηγορίες των τριών μεταβλητών και δεν έχουν την έννοια της μέτρησης.

Σε αυτή την περίπτωση προφανώς N=5, q=3 και j=3+2+2=7, ενώ οι πίνακες $\mathbf{Z_1}$, $\mathbf{Z_2}$, $\mathbf{Z_3}$ και $\mathbf{Z_{0\text{-}1}}$ θα είναι:

$$\mathbf{Z_1}=\begin{bmatrix}0&1&0\\1&0&0\\1&0&0\\0&0&1\\0&1&0\end{bmatrix},\ \mathbf{Z_2}=\begin{bmatrix}1&0\\0&1\\0&1\\1&0\\1&0\end{bmatrix},\ \mathbf{Z_3}=\begin{bmatrix}1&0\\1&0\\0&1\\1&0\\0&1\end{bmatrix},\ \mathbf{Z_{0\text{-}1}}=\begin{bmatrix}0&1&0&1&0&1&0\\1&0&0&0&1&1&0\\1&0&0&0&1&0&1\\0&0&1&0&1&1&0\\0&1&0&1&0&0&1\end{bmatrix}.$$

$$\underbrace{}_{\mathbf{Z_1}}\quad\underbrace{}_{\mathbf{Z_2}}\quad\underbrace{}_{\mathbf{Z_3}}$$

2.3.1.1 Γενικές Ιδιότητες του Πίνακα $\mathbf{Z_{0\text{-}1}}$

1) Το άθροισμα κάθε γραμμής του είναι σταθερό και ίσο με *q*, όσο δηλαδή και το πλήθος των μεταβλητών. Αυτό έχει ως αποτέλεσμα τα αντικείμενα να συμμετέχουν στην ανάλυση με το ίδιο βάρος. Στο συγκεκριμένο παράδειγμα, το άθροισμα κάθε γραμμής είναι ίσο με 3.

2) Το άθροισμα κάθε στήλης του είναι ίσο με τη συχνότητα της αντίστοιχης κλάσης. Στο παράδειγμα, η συχνότητα της κλάσης 2 της μεταβλητής X_1 είναι ίση με 2. Με 2 είναι ίσο και το άθροισμα της πρώτης στήλης του πίνακα $\mathbf{Z_{0\text{-}1}}$.

3) Το γενικό άθροισμα (σύνολο) των στοιχείων του είναι ίσο με *Nq*. Στο παράδειγμα, το γενικό άθροισμα είναι ίσο με 5×3=15.

4) Μέσω του πίνακα $\mathbf{Z_i}$ κάθε κατηγορική μεταβλητή X_i διασπάται σε j_i νέες. Οι νέες μεταβλητές αντιστοιχούν στις ιδιότητες (κλάσεις, κατηγορίες) της X_i και αντιμετωπίζονται ως ομάδα. Μπορεί να δειχθεί ότι στην περίπτωση δύο μεταβλητών X_1 και X_2 τα αποτελέσματα της ΠΑΑ είναι συγκρίσιμα με αυτά που παράγονται κατά την εφαρμογή της μεθόδου της Κανονικοποιημένης Συσχέτισης στις δύο ομάδες μεταβλητών που ορίζονται από τις ιδιότητες j_1 και j_2 των X_1 και X_2 αντίστοιχα (βλέπε Greenacre 1984, Lebart, Morineau & Warwick 1984, Tenenhaus & Young 1985, Gower 1990, Andersen 1991, Van de Geer 1993β, De Leeuw, Wang & Michailidis 1999).

2.3.2 Γενικευμένοι Πίνακες Συμπτώσεων

Στην πολυμεταβλητή περίπτωση η ΠΑΑ εφαρμόζεται συνήθως στο Γενικευμένο Πίνακα Συμπτώσεων απολύτων συχνοτήτων των q μεταβλητών. Ο πίνακας αυτός είναι γνωστός και ως πίνακας *Burt* (Burt, 1950). Υπολογιστικά ο πίνακας *Burt* **B** δίνεται από τη σχέση (Burt 1950, Israëls 1987, SAS Institute 1990, Greenacre 1994β, 1993α, 1991 και1990):

$$\mathbf{B} = \mathbf{Z}_{\mathbf{0\text{-}1}}^{\mathbf{T}}\mathbf{Z}_{\mathbf{0\text{-}1}}\,.$$

Ο πίνακας **B** είναι διαστάσεων $j \times j$ και περιέχει της διασταυρώσεις των j ιδιοτήτων (κλάσεων) των q μεταβλητών μεταξύ τους. Αποτελεί στην ουσία έναν block πίνακα που σχηματίζεται από q^2 σε πλήθος υποπίνακες. Η γενική μορφή του πίνακα **B** δίνεται στον Πίνακα 2.7.

Στον Πίνακα 2.7, με $f_{v+}^{X_i}$ συμβολίζουμε τη συχνότητα της κλάσης v της X_i μεταβλητής (v=1,…, j_i, i=1,…, q) και με $f_{vu}^{X_iX_m} = f_{uv}^{X_mX_i}$ τη συχνότητα που αντιστοιχεί στο κελί (v, u) (v=1,…, j_i και u=1,…, j_m) του πίνακα που διασταυρώνει τις j_i κλάσεις της μεταβλητής X_i με τις j_m κλάσεις της μεταβλητής X_m (i, m=1,…, q). Αν συμβολίσουμε με $\mathbf{F_{im}}$ τον πίνακα που διασταυρώνει τη μεταβλητή i (στις γραμμές) με τη μεταβλητή m (στις στήλες), τότε ο πίνακας **B** μπορεί να γραφεί και ως εξής:

$$\mathbf{B} = \begin{bmatrix} \mathbf{F_{11}} & \mathbf{F_{12}} & \mathbf{F_{13}} & \cdots & \mathbf{F_{1q}} \\ \mathbf{F_{12}^T} & \mathbf{F_{22}} & \mathbf{F_{23}} & \cdots & \mathbf{F_{2q}} \\ \mathbf{F_{13}^T} & \mathbf{F_{23}^T} & \mathbf{F_{33}} & \cdots & \mathbf{F_{3q}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{F_{1q}^T} & \mathbf{F_{2q}^T} & \mathbf{F_{3q}^T} & \cdots & \mathbf{F_{qq}} \end{bmatrix}.$$

Πίνακας 2.7: Πίνακας *Burt* των q Μεταβλητών

		Ιδιότητες της X_1				Ιδιότητες της X_2					Ιδιότητες της X_q				
		1	2	…	j_1	1	2	…	j_2	…	1	2		j_q	*Σύνολο*
Ιδιότητες της X_1	1	$f_{1+}^{X_1}$	0	…	0	$f_{11}^{X_1X_2}$	$f_{12}^{X_1X_2}$	…	$f_{1j_2}^{X_1X_2}$	…	$f_{11}^{X_1X_q}$	$f_{12}^{X_1X_q}$	…	$f_{1j_q}^{X_1X_q}$	$q\, f_{1+}^{X_1}$
	2	0	$f_{2+}^{X_1}$	…	0	$f_{21}^{X_1X_2}$	$f_{22}^{X_1X_2}$	…	$f_{2j_2}^{X_1X_2}$	…	$f_{21}^{X_1X_q}$	$f_{22}^{X_1X_q}$	…	$f_{2j_q}^{X_1X_q}$	$q\, f_{2+}^{X_1}$
	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋮	⋮	…	⋮	⋮	⋮	⋮	⋮
	j_1	0	0	0	$f_{j_1+}^{X_1}$	$f_{j_11}^{X_1X_2}$	$f_{j_12}^{X_1X_2}$	…	$f_{j_1j_2}^{X_1X_2}$	…	$f_{j_11}^{X_1X_q}$	$f_{j_12}^{X_1X_q}$	…	$f_{j_1j_q}^{X_1X_q}$	$q\, f_{j_1+}^{X_1}$
	1	$f_{11}^{X_2X_1}$	$f_{12}^{X_2X_1}$	…	$f_{1j_1}^{X_2X_1}$	$f_{1+}^{X_2}$	0	…	0	…	$f_{11}^{X_2X_q}$	$f_{12}^{X_2X_q}$	…	$f_{1j_q}^{X_2X_q}$	$q\, f_{1+}^{X_2}$
	2	$f_{21}^{X_2X_1}$	$f_{22}^{X_2X_1}$	…	$f_{2j_1}^{X_2X_1}$	0	$f_{2+}^{X_2}$	…	0	…	$f_{21}^{X_2X_q}$	$f_{22}^{X_2X_q}$	…	$f_{2j_q}^{X_2X_q}$	$q\, f_{2+}^{X_2}$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	…	⋮	⋮	⋮	⋮	⋮
	j_2	$f_{j_21}^{X_2X_1}$	$f_{j_22}^{X_2X_1}$	…	$f_{j_2j_1}^{X_2X_1}$	0	0	0	$f_{j_2+}^{X_2}$	…	$f_{j_21}^{X_2X_q}$	$f_{j_22}^{X_2X_q}$	…	$f_{j_2j_q}^{X_2X_q}$	$q\, f_{j_2+}^{X_2}$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋮	⋮
	1	$f_{11}^{X_qX_1}$	$f_{12}^{X_qX_1}$	…	$f_{1j_1}^{X_qX_1}$	$f_{11}^{X_qX_2}$	$f_{12}^{X_qX_2}$	…	$f_{1j_2}^{X_qX_2}$	…	$f_{1+}^{X_q}$	0	…	0	$q\, f_{1+}^{X_q}$
	2	$f_{21}^{X_qX_1}$	$f_{22}^{X_qX_1}$	…	$f_{2j_1}^{X_qX_1}$	$f_{21}^{X_qX_2}$	$f_{22}^{X_qX_2}$	…	$f_{2j_2}^{X_qX_2}$	…	0	$f_{2+}^{X_q}$	…	0	$q\, f_{2+}^{X_q}$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	…	⋮	⋮	⋱	⋮	⋮
	j_q	$f_{j_q1}^{X_qX_1}$	$f_{j_q2}^{X_qX_1}$	…	$f_{j_qj_1}^{X_qX_1}$	$f_{j_q1}^{X_qX_2}$	$f_{j_q2}^{X_qX_2}$	…	$f_{j_qj_2}^{X_qX_2}$	…	0	0	0	$f_{j_q+}^{X_q}$	$q\, f_{j_q+}^{X_q}$
	Σύνολο	$q\, f_{1+}^{X_1}$	$q\, f_{2+}^{X_1}$	…	$q\, f_{j_1+}^{X_1}$	$q\, f_{1+}^{X_2}$	$q\, f_{2+}^{X_2}$	…	$q\, f_{j_2+}^{X_2}$	…	$q\, f_{1+}^{X_q}$	$q\, f_{2+}^{X_q}$	…	$q\, f_{j_q+}^{X_q}$	Nq^2

2.3.2.1 Γενικές Ιδιότητες του Πίνακα *Burt* (**B**)

1) Ο πίνακας **B** είναι τετράγωνος και συμμετρικός ως προς την κύρια διαγώνιό του.

2) Στους υποπίνακες $\mathbf{F_{ii}}$ που σχηματίζονται από τις ενδοδιασταυρώσεις των ιδιοτήτων της ίδιας μεταβλητής X_i, υπάρχουν μη μηδενικά στοιχεία μόνο στην κύρια διαγώνιο. Προφανώς, το άθροισμα των στοιχείων κάθε διαγώνιου πίνακα $\mathbf{F_{ii}}$ είναι ίσο με το πλήθος των αντικειμένων N. Επίσης, σε κάθε υποπίνακα $\mathbf{F_{im}}$ για το άθροισμα της στήλης u (u=1,…, j_m) ισχύει (Παπαδημητρίου, 2006, 2004, 2001 και 1990):

$$\sum_{v=1}^{j_i} f_{vu}^{X_i X_m} = f_{u+}^{X_m}, \qquad [2.57]$$

και, λόγω συμμετρίας, για το άθροισμα της γραμμής v (v=1,…, j_i)

$$\sum_{u=1}^{j_m} f_{vu}^{X_i X_m} = f_{v+}^{X_i}. \qquad [2.58]$$

3) Το άθροισμα κάθε γραμμής (στήλης) του πίνακα **B** είναι ίσο με q φορές τη συχνότητα της κλάσης στην οποία αντιστοιχεί η γραμμή (στήλη) του **B**.

4) Το γενικό σύνολο του πίνακα **B** είναι ίσο με Nq^2.

Αν το ενδιαφέρον της μελέτης εστιάζεται μόνο στη διερεύνηση της σχέσης μεταξύ των μεταβλητών και όχι στα αντικείμενα της μελέτης, τότε είναι προτιμότερο η ΠΑΑ να εφαρμοστεί στον πίνακα **B** που έχει μικρότερες διαστάσεις από τον $\mathbf{Z_{0\text{-}1}}$ και η ενδεχόμενη ύπαρξη ελλειπουσών τιμών δεν δημιουργεί ιδιαίτερο πρόβλημα. Στο Κεφάλαιο 3 (Ενότητα 3.3), προτείνουμε έναν αποτελεσματικό αλγόριθμο για την ανάλυση μεγάλων πινάκων $\mathbf{Z_{0\text{-}1}}$.

2.3.3 Ο Αλγόριθμος της Πολυμεταβλητής ΠΑΑ

Όπως προαναφέρθηκε, στο πλαίσιο της Γαλλικής Σχολής ο αλγόριθμος της πολυμεταβλητής εκδοχής της ΠΑΑ δεν διαφοροποιείται από αυτόν της διμεταβλητής περίπτωσης. Στη συνέχεια, λόγω της ιδιαιτερότητας του πίνακα $\mathbf{Z_{0\text{-}1}}$ θα περιγράψουμε συνοπτικά την αλγοριθμική διαδικασία της μεθόδου. Αν εφαρμόσουμε τους βασικούς

ορισμούς και βήματα, όπως και στην περίπτωση των δύο μεταβλητών, έχουμε (βλέπε Israëls 1987, Greenacre 1991 και Ενότητα 2.2.14):

1. $\mathbf{P} = (\frac{1}{Nq})\mathbf{Z_{0\text{-}1}}$, είναι ο $N \times j$ Πίνακας των Αντιστοιχιών, όπου Nq είναι το γενικό άθροισμα του $\mathbf{Z_{0\text{-}1}}$.

2. Το προφίλ μιας γραμμής (στήλης) του πίνακα $\mathbf{Z_{0\text{-}1}}$ (ή ισοδύναμα του $\mathbf{P}$) εκφράζεται σε σχετικές τιμές ως προς το άθροισμα (σύνολο) της αντίστοιχης γραμμής (στήλης).

3. $\mathbf{r} = \mathbf{P1}$, είναι το $N \times 1$ διάνυσμα με στοιχεία τις μάζες των γραμμών, ταυτίζεται με το μέσο προφίλ των στηλών και τα στοιχεία του είναι όλα ίσα με $\frac{q}{Nq} = \frac{1}{N}$.

4. $\mathbf{c} = \mathbf{P^T 1}$, είναι το $j \times 1$ διάνυσμα με στοιχεία τις μάζες των στηλών, ταυτίζεται με το μέσο προφίλ των γραμμών και τα στοιχεία του είναι ίσα με τα αντίστοιχα στοιχεία της περιθώριας γραμμής του πίνακα $\mathbf{Z_{0\text{-}1}}$ διαιρεμένα δια Nq.

5. $\mathbf{D_r}$ είναι ο $N \times N$ διαγώνιος πίνακας με στοιχεία επί της διαγωνίου τις μάζες των γραμμών και $\mathbf{D_c}$ είναι ο $j \times j$ διαγώνιος πίνακας με στοιχεία τις μάζες των στηλών.

6. Η απόσταση χ^2 μεταξύ δύο προφίλ γραμμών-αντικειμένων (στηλών-ιδιοτήτων) είναι η σταθμισμένη Ευκλείδεια απόσταση με βάρη που ορίζονται από τους πίνακες $\mathbf{D_c^{-1}}$ ($\mathbf{D_r^{-1}}$) (βλέπε σχέσεις [2.1] και [2.2]). Οι αποστάσεις των αντικειμένων και των κατηγοριών από τα κέντρα βάρους τους ορίζονται όπως και στη διμεταβλητή περίπτωση (βλέπε Ενότητα 2.2.14). Ειδικότερα, η απόσταση της κλάσης u^{X_i} (u^{X_i}=1,…, j_i) της μεταβλητής X_i από το κέντρο βάρους του αντίστοιχου νέφους δίνεται από τη σχέση (Lebart, Morineau & Warwick 1984, Greenacre 1984, Μπεχράκης 1999, Le Roux & Rouanet 2004):

$$d^2_{\chi^2}\left(u^{X_i}, g\right) = \sum_{n=1}^{N} \frac{N}{1}\left(\frac{z_{nu}}{f_{u+}^{X_i}} - \frac{1}{N}\right)^2 = N\sum_{n=1}^{N}\left(\frac{z_{nu}}{f_{u+}^{X_i}} - \frac{1}{N}\right)^2 =$$

$$= N\sum_{n=1}^{N}\left(\frac{\left(z_{nu}\right)^2}{\left(f_{u+}^{X_i}\right)^2} - 2\frac{z_{nu}}{f_{u+}^{X_i}}\frac{1}{N} + \frac{1}{N^2}\right) =$$

$$= \frac{N}{\left(f_{u+}^{X_i}\right)^2}\sum_{n=1}^{N}\left(z_{nu}\right)^2 - \frac{2N}{f_{u+}^{X_i}N}\sum_{n=1}^{N} z_{nu} + N\sum_{n=1}^{N}\frac{1}{N^2} =$$

$$= \frac{N}{\left(f_{u+}^{X_i}\right)^2}\left(f_{u+}^{X_i}\right) - \frac{2N}{f_{u+}^{X_i}N}\left(f_{u+}^{X_i}\right) + \frac{N^2}{N^2} =$$

$$= \frac{N}{f_{u+}^{X_i}} - 2 + 1 = \frac{N}{f_{u+}^{X_i}} - 1,$$

όπου z_{nu} είναι στοιχείο του πίνακα $\mathbf{Z_{0\text{-}1}}$ και $f_{u+}^{X_i}$ η συχνότητα της κλάσης u^{X_i}. Στα ενδιάμεσα βήματα χρησιμοποιήσαμε το γεγονός ότι τα στοιχεία z_{nu} είναι 0 ή 1 και ότι $\sum_{n=1}^{N} z_{nu} = \sum_{n=1}^{N}\left(z_{nu}\right)^2 = f_{u+}^{X_i}$.

Συνεπώς, η αδράνεια της κλάσης u^{X_i} θα είναι ίση με:

$$I_{u^{X_i}} = \frac{f_{u+}^{X_i}}{Nq}\left(\frac{N}{f_{u+}^{X_i}} - 1\right) = \frac{1}{q} - \frac{f_{u+}^{X_i}}{Nq} = \frac{1}{q}\left(1 - \frac{f_{u+}^{X_i}}{N}\right).$$

7. Η ανάλυση συνίσταται στην εφαρμογή της Διάσπασης σε Χαρακτηριστικές Τιμές SVD (συνήθως) στον πίνακα:

$$\mathbf{S} = \mathbf{D_r^{-1/2}}\left(\mathbf{P} - \mathbf{rc^T}\right)\mathbf{D_c^{-1/2}} = \left(\frac{1}{N^{1/2}q}\right)\left(\mathbf{Z_{0\text{-}1}} - q\mathbf{1c^T}\right)\mathbf{D_c^{-1/2}} \qquad [2.59]$$

Έχουμε, δηλαδή:

$$\mathbf{S}=\mathbf{UDV^T}.$$

Επειδή στην πράξη το πλήθος των αντικειμένων *N* είναι πολύ μεγαλύτερο από το συνολικό πλήθος των κλάσεων *j* των *q* μεταβλητών είναι προτιμότερο οι υπολογισμοί να βασιστούν στην ανάλυση του πίνακα $\mathbf{S}^{\mathbf{T}}\mathbf{S}$, ο οποίος αντιστοιχεί στον πίνακα που διαγωνοποιείται κατά την εφαρμογή της ΠΑΑ στον πίνακα *Burt* (Greenacre, 1984). Βέβαια, μπορούν να χρησιμοποιηθούν και οι πίνακες που παρουσιάστηκαν στην Παρατήρηση 2.1.

8. Τέλος, υπολογίζονται τα βασικά αριθμητικά αποτελέσματα όπως και στην περίπτωση δύο μεταβλητών (οι δείκτες *CTR*, *COR* και *QLT*, οι ιδιοτιμές, τα ποσοστά και τα αθροιστικά ποσοστά ερμηνείας των αξόνων). Θα πρέπει να τονιστεί ότι στην πολυμεταβλητή περίπτωση η αθροιστική συνεισφορά *CTR* των κατηγοριών μιας μεταβλητής επί ενός άξονα *s* αποκτά ιδιαίτερη σημασία, αφού μπορεί να θεωρηθεί ως δείκτης σημαντικότητας της μεταβλητής στην κατασκευή και στον προσανατολισμό του *s*.

2.3.3.1 Ολική Αδράνεια και Συνεισφορές Ιδιοτήτων και Μεταβλητών

Οι αδράνειες λ_s των παραγοντικών αξόνων βρίσκονται στη διαγώνιο του πίνακα $\mathbf{D}^2$, ο μέγιστος αριθμός μη τετριμμένων διαστάσεων που μπορούν να προκύψουν από τη λύση της ΠΑΑ είναι $p=j-q$ (με δεδομένο ότι στην πράξη $N>>j$), ενώ η συνολική αδράνεια I_{0-1} του πίνακα $\mathbf{Z_{0\text{-}1}}$ είναι ίση με (Lebart, Marineau & Warwick 1984, Greenacre 1991, Benzécri 1992, Μπεχράκης 1999):

$$I_{0-1} = \frac{j}{q} - 1 = \frac{(j-q)}{q} = trace\left(\mathbf{D}^2\right) = \sum_{s=1}^{p} \lambda_s \ . \qquad [2.60]$$

Η συνεισφορά I_{X_i} της μεταβλητής X_i στην ολική αδράνεια του πίνακα $\mathbf{Z_{0\text{-}1}}$ δίνεται από τη σχέση (Lebart, Marineau & Warwick 1984, Greenacre 1984, Benzécri 1992, Μπεχράκης 1999, Le Roux & Rouanet 2004):

$$I_{X_i} = \frac{1}{q}\left(j_i - 1\right), \qquad [2.61]$$

όπου j_i είναι το πλήθος των κλάσεων (ιδιοτήτων) της μεταβλητής X_i.

Από την παραπάνω σχέση είναι φανερό ότι όσο περισσότερες κλάσεις (κατηγορίες) έχει μια μεταβλητή τόσο μεγαλύτερη είναι και η αδράνειά της. Αυτό έχει ως αποτέλεσμα κάποιες ερωτήσεις να συμμετέχουν στην ανάλυση με μεγαλύτερο "βάρος" απ' ότι κάποιες άλλες, ανάλογα με το πλήθος των κατηγοριών τους. Κάτι τέτοιο είναι πιθανό να αποκρύψει την ιδιαιτερότητα και τη σημασία των μεταβλητών με λίγες κατηγορίες. Αν αυτό δεν είναι επιθυμητό, τότε θα πρέπει είτε κατά το σχεδιασμό της έρευνας είτε κατά το στάδιο της κωδικοποίησης των δεδομένων να ληφθεί μέριμνα ώστε οι μεταβλητές (π.χ. ερωτήσεις) να έχουν τον ίδιο ή σχεδόν τον ίδιο αριθμό κλάσεων (Μπεχράκης, 1999). Οι Μοσχίδης, Παπαδημητρίου και Χατζηπαντελής (2005) προτείνουν μια τροποποιημένη εκδοχή της απόστασης χ^2, σύμφωνα με την οποία οι αδράνειες των μεταβλητών με διαφορετικό πλήθος κατηγοριών εξισορροπούνται και συμβάλλουν με την ίδια βαρύτητα στην ολική αδράνεια του πίνακα $\mathbf{Z_{0\text{-}1}}$. Ο Greenacre (1994β) υποστηρίζει ότι η εξισορρόπηση της συμβολής των μεταβλητών στην ολική αδράνεια μπορεί να επιτευχθεί με τροποποίηση του πίνακα *Burt* (**B**), σύμφωνα με την οποία τα στοιχεία κάθε υποπίνακα $\mathbf{F_{ij}}$ του **B** θα πρέπει να διαιρεθούν με τον αριθμό $min\{ j_i - 1,\ j_j - 1 \}$, όπου j_i είναι το πλήθος κλάσεων της μεταβλητής X_i και j_j το πλήθος κλάσεων της μεταβλητής X_j, με $i,\ j = 1, \ldots, q$. Στην περίπτωση αυτή, η αδράνεια κάθε υποπίνακα $\mathbf{F_{ij}}$ ($i \neq j$) καθίσταται ίση με το τετράγωνο του συντελεστή συνάφειας *V* του *Cramer* των μεταβλητών X_i και X_j (Blasius, 1994).

Όπως είδαμε και παραπάνω, η συνεισφορά της κλάσης u^{X_i} ($u^{X_i} = 1, \ldots, j_i$) της μεταβλητής X_i στην ολική αδράνεια του πίνακα $\mathbf{Z_{0\text{-}1}}$ υπολογίζεται από τη σχέση:

$$I_{u^{X_i}} = \frac{1}{q}\left(1 - \frac{f_{u+}^{X_i}}{N}\right), \qquad [2.62]$$

όπου $\frac{f_{u+}^{X_i}}{N}$ είναι η σχετική συχνότητα της κλάσης u^{X_i}.

Αν αθροίσουμε τις συνεισφορές των κλάσεων της μεταβλητής X_i, τότε προκύπτει η συνεισφορά I_{X_i} της αντίστοιχης μεταβλητής (βλέπε σχέση [2.61]):

$$\sum_{u=1}^{j_i} I_{u^{X_i}} = \sum_{u=1}^{j_i} \frac{1}{q}\left(1-\frac{f_{u+}^{X_i}}{N}\right) = \frac{1}{q}\sum_{u=1}^{j_i}\left(1-\frac{f_{u+}^{X_i}}{N}\right) = \frac{1}{q}\left(\sum_{u=1}^{j_i}1-\sum_{u=1}^{j_i}\frac{f_{u+}^{X_i}}{N}\right) = \frac{1}{q}\left(j_i-1\right) = I_{X_i}.$$

Το άθροισμα των συνεισφορών των q μεταβλητών δίνει την ολική αδράνεια του πίνακα $\mathbf{Z_{0\text{-}1}}$ (βλέπε σχέση [2.60]):

$$\sum_{i=1}^{q}\sum_{u=1}^{j_i} I_{u^{X_i}} = \sum_{i=1}^{q}\frac{1}{q}\left(j_i-1\right) = \frac{1}{q}\sum_{i=1}^{q}\left(j_i-1\right) = \frac{1}{q}\left(\sum_{i=1}^{q}j_i-\sum_{i=1}^{q}1\right) = \frac{1}{q}\left(j-q\right) = \left(\frac{j}{q}-1\right) = I_{0-1}.$$

Από τη σχέση [2.62] προκύπτει ότι η αδράνεια μιας κλάσης είναι τόσο μεγαλύτερη όσο μικρότερη είναι η σχετική της συχνότητα στο σύνολο των N αντικειμένων. Με άλλα λόγια, όσο πιο σπάνια είναι η κατηγορία μιας μεταβλητής τόσο μεγαλύτερη είναι και η αδράνειά της. Σύμφωνα με τον Μπεχράκη (1999), θα πρέπει πριν την εφαρμογή της ΠΑΑ να γίνεται έλεγχος των κατανομών σχετικών συχνοτήτων των μεταβλητών, ώστε να διαπιστωθεί αν υπάρχουν κατηγορίες μεταβλητών με πολύ χαμηλές συχνότητες (μικρές μάζες). Είναι δυνατό αυτές οι σπάνιες κατηγορίες να συμβάλουν σημαντικά στην κατασκευή των παραγοντικών αξόνων με αποτέλεσμα να αλλοιώνεται η εικόνα του υπό εξέταση φαινομένου (βλέπε Ενότητες 2.1.2.10 και 2.1.2.11). Μια πρακτική λύση είναι η συγχώνευση (ομαδοποίηση) των κλάσεων αυτών με άλλες της ίδιας μεταβλητής, αν αυτό, βέβαια, είναι εννοιολογικά επιτρεπτό στο πλαίσιο της εκάστοτε μελέτης (βλέπε Παρατήρηση 2.4 στο τέλος της τρέχουσας ενότητας). Αλλιώς, θα μπορούσαν οι κλάσεις αυτές να εισαχθούν στην ανάλυση ως συμπληρωματικά σημεία. Γενικά, στην περίπτωση πολλών μεταβλητών η συνένωση κλάσεων, οι οποίες έχουν στους παραγοντικούς άξονες τις ίδιες ή σχεδόν τις ίδιες συντεταγμένες, δεν έχει σημαντική επίδραση στις χαρακτηριστικές τιμές και κατά συνέπεια στις αδράνειες των αξόνων με συνέπεια να μη μεταβάλλεται η ποιότητα της λύσης της ΠΑΑ (Gifi 1996, Michailidis 1996, Michailidis & De Leeuw 1998). Το ίδιο ισχύει και στη διμεταβλητή περίπτωση (Andersen, 1991). Για τις σπάνιες κατηγορίες, η συγχώνευσή τους με άλλες όχι μόνο βελτιώνει γενικά την ποιότητα και σταθερότητα των αποτελεσμάτων, αλλά και μερικές φορές είναι επιβεβλημένη ιδιαίτερα στην περίπτωση που η απόλυτη συχνότητά τους είναι μικρότερη του 8 (Markus, 1994α). Επομένως, κατά την εφαρμογή της ΠΑΑ είναι σημαντικό να αποφεύγονται κατηγορίες με πολύ χαμηλές συχνότητες (Lebart, Morineau & Warwick 1984, Van de Geer 1993α και 1993β). Ασήμαντη επίδραση στις ιδιοτιμές φαίνεται να έχει και η απομάκρυνση μιας μεταβλητής που έχει αμελητέα συσχέτιση

με τους άξονες (Gifi 1996, Michailidis 1996, Michailidis & De Leeuw 1998). Η ΠΑΑ είναι μια αρκετά εύρωστη μέθοδος, με την έννοια ότι η προσθήκη μιας νέας κατηγορίας στη διμεταβλητή περίπτωση (Καρλής, 2005) ή μιας νέας μεταβλητής στην πολυμεταβλητή εκδοχή της (Blasius, 1994) δεν επηρεάζει σημαντικά τη σταθερότητα των δομών μεταξύ των μεταβλητών. Άλλωστε, η «Αρχή της Ισοδυναμίας των Κατανομών» (βλέπε Ενότητα 2.2.4) εξασφαλίζει ότι η προσθήκη ή η απομάκρυνση, από την ανάλυση, σημείων με ίδια ή παρόμοια προφίλ έχει αμελητέα επίδραση στα αποτελέσματα. Για αναλυτικά αποτελέσματα σχετικά με τη σταθερότητα των δομών ενός νέφους σημείων στον Ευκλείδειο χώρο παραπέμπουμε στους Le Roux και Rouanet (2004). Στις μέρες μας, η μεγάλη υπολογιστική ισχύς των Η/Υ επιτρέπει, μέσω δοκιμών, τον εμπειρικό έλεγχο της επίδρασης στις ιδιοτιμές, στις σχέσεις και στις δομές, που αναδεικνύονται από την εφαρμογή της ΠΑΑ, στις περιπτώσεις συνένωσης ή/και απομάκρυνσης αντικειμένων, κατηγοριών και μεταβλητών. Αν, για παράδειγμα, υπάρχει προβληματισμός σχετικά με το εάν δύο κλάσεις θα πρέπει να συνενωθούν δεν έχουμε παρά να εκτελέσουμε την ανάλυση δύο φορές. Μία πριν και μία μετά τη συνένωση των κατηγοριών και να συγκρίνουμε τα αποτελέσματα.

2.3.3.2 Κανονικοποίηση των Συντεταγμένων Αντικειμένων και Ιδιοτήτων

Σε αντιστοιχία με τη διμεταβλητή περίπτωση, οι κύριες και τυποποιημένες συντεταγμένες των γραμμών (αντικειμένων) και στηλών (ιδιοτήτων) δίνονται από τις παρακάτω σχέσεις (βλέπε Ενότητα 2.2.14.1 και τις σχέσεις [2.34] και [2.35]):

Κύριες Συντεταγμένες Γραμμών	$\mathbf{\Phi} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}$	[2.63]
Κύριες Συντεταγμένες Στηλών	$\mathbf{\Gamma} = \mathbf{D}_c^{-1/2}\mathbf{V}\mathbf{D}$	[2.64]
Τυποποιημένες Συντεταγμένες Γραμμών	$\mathbf{R}^* = \mathbf{D}_r^{-1/2}\mathbf{U}$	[2.65]
Τυποποιημένες Συντεταγμένες Στηλών	$\mathbf{C}^* = \mathbf{D}_c^{-1/2}\mathbf{V}$	[2.66]

Αφού προσδιοριστούν οι συντεταγμένες των προβολών των σημείων γραμμών και στηλών επί των παραγοντικών αξόνων μπορούν να υπολογιστούν και τα υπόλοιπα

αριθμητικά αποτελέσματα που προκύπτουν από την εφαρμογή της ΠΑΑ όπως και στη διμεταβλητή περίπτωση (βλέπε Ενότητα 2.2.14).

2.3.3.3 Συμπληρωματικά Σημεία

Γενικά, ισχύουν όσα έχουν αναφερθεί στην περίπτωση των δύο μεταβλητών (βλέπε Ενότητες 2.2.11, 2.2.14 και 2.2.14.2). Εκείνο που θα πρέπει να επισημάνουμε είναι ότι στην πολυμεταβλητή εκδοχή της ΠΑΑ δεν έχει ιδιαίτερη αξία η προβολή μεμονωμένων κλάσεων μιας μεταβλητής ως συμπληρωματικών σημείων, αλλά η προβολή της ίδιας της μεταβλητής ως συμπληρωματικής. Η δυνατότητα αυτή είναι αρκετά χρήσιμη, ιδιαίτερα στην περίπτωση που οι μεταβλητές μπορούν να χωριστούν εννοιολογικά σε δύο ομάδες (Benzécri, 1992): α) την ομάδα των βασικών μεταβλητών που μετρούν το υπό εξέταση φαινόμενο και β) την ομάδα των πληροφοριακών ή «οργανισμικών» (βλέπε Kirk, 1995) μεταβλητών (π.χ. δημογραφικά στοιχεία), οι οποίες χαρακτηρίζουν τις δειγματοληπτικές ή πειραματικές μονάδες της έρευνας. Συχνά, οι πληροφοριακές μεταβλητές εισάγονται στην ανάλυση ως συμπληρωματικές για να μη “διαστρεβλωθούν” ή αποκρυφτούν οι σχέσεις και αλληλεπιδράσεις των βασικών μεταβλητών (βλέπε Ενότητα 2.2.11). Για πιο εξειδικευμένες εφαρμογές των συμπληρωματικών σημείων παραπέμπουμε στον Greenacre (2005).

2.3.3.4 Ελλείπουσες Τιμές

Ο χειρισμός των ενδεχόμενων ελλειπουσών τιμών των μεταβλητών στηρίζεται συνήθως στη δημιουργία μιας νέας κατηγορίας για κάθε μεταβλητή. Η νέα αυτή κατηγορία που αντιστοιχεί στις ελλείπουσες τιμές μπορεί να συμμετέχει στην ανάλυση είτε ως ενεργή είτε ως συμπληρωματική (Van Rijckevorsel 1987, SAS Institute 1990, Van de Geer 1993α και 1993β). Συχνά, οι ελλείπουσες τιμές αντικαθίστανται με την επικρατούσα τιμή της αντίστοιχης κατηγορικής μεταβλητής ή με μια άλλη τιμή, η οποία προκύπτει από την εφαρμογή κάποιας μεθόδου συμπλήρωσης (*imputation*) ελλειπουσών τιμών (Rubin 1987, SPSS Inc. 2004α και 1997). Στην πράξη, αν το μέγεθος του δείγματος το επιτρέπει, τα αντικείμενα με ελλείπουσες τιμές απομακρύνονται από την ανάλυση. Για πιο εξειδικευμένες μεθόδους χειρισμού των ελλειπουσών τιμών, στο πλαίσιο της ΠΑΑ, παραπέμπουμε

στους Young 1981, Greenacre 1984, Van Rijckevorsel 1987, Van der Heijden και Escofier 1988, De Leeuw και Van der Heijden 1988, Van der Heijden, De Vries και Van Hooff 1990, Van de Geer 1993α και 1993β, Nishisato 1994 και 1980, Gifi 1996, Van der Heijden, Teunissen και Van Orlé 1997, Le Roux και Chiche 2004. Δύο εναλλακτικοί τρόποι αντιμετώπισης των ελλειπουσών τιμών είναι μέσω: α) της Ιδιάζουσας Πολλαπλής Ανάλυσης των Αντιστοιχιών, που πρότειναν οι Le Roux και Chiche (2004) (βλέπε Ενότητα 2.2.10 και Le Roux & Rouanet, 2004) και β) της μεθόδου που παρουσιάσαν οι Greenacre και Pardo (2005), σύμφωνα με την οποία η ΠΑΑ μπορεί να τροποποιηθεί, ώστε η ανάλυση να πραγματοποιηθεί σε υποσύνολα των κατηγοριών των μεταβλητών, χωρίς όμως να διαταραχθεί η δομή του αρχικού και πλήρους πίνακα δεδομένων. Οι ελλείπουσες τιμές δεν δημιουργούν ιδιαίτερο υπολογιστικό πρόβλημα, όταν η ΠΑΑ εφαρμόζεται στον πίνακα *Burt*, αφού αυτές δεν λαμβάνονται υπόψη κατά το σχηματισμό των υποπινάκων $\mathbf{F}_{im}$.

2.3.3.5 Ισοδυναμία Λογικού Πίνακα 0-1 με Πίνακα Burt

Στη Γαλλική Σχολή, είτε η ΠΑΑ εφαρμοστεί στον πίνακα $\mathbf{Z}_{0\text{-}1}$ είτε στον **B**, η μέθοδος κανονικοποίησης που χρησιμοποιείται είναι η κύρια (*PN*). Στην περίπτωση του πίνακα $\mathbf{Z}_{0\text{-}1}$, επί των παραγοντικών επιπέδων προβάλλονται μόνο οι κλάσεις των μεταβλητών, δηλαδή οι ιδιότητες των αντικειμένων. Τα αντικείμενα δεν προβάλλονται στα παραγοντικά επίπεδα εκτός και αν είναι "επώνυμα" και σχετικά λίγα σε πλήθος. Όταν η ανάλυση εφαρμοστεί στον πίνακα **B**, τότε δεν υπάρχει πληροφορία για τα αντικείμενα, αφού ο **B** περιλαμβάνει τις ενδοδιασταυρώσεις μεταξύ των ιδιοτήτων των μεταβλητών, και, συνεπώς, επί των παραγοντικών επιπέδων προβάλλονται και πάλι οι ιδιότητες των αντικειμένων με κύρια κανονικοποίηση. Αν και ο πίνακας **B** περιέχει λιγότερη πληροφορία από τον $\mathbf{Z}_{0\text{-}1}$ (Van de Geer, 1993α), ωστόσο οι δύο προσεγγίσεις είναι ισοδύναμες και η εικόνα του υπό εξέταση φαινομένου επί των παραγοντικών διαγραμμάτων είναι η ίδια (Lebart, Morineau & Tabard 1977, Lebart, Morineau & Warwick 1984, Israëls 1987, Van der Heijden & De Leeuw 1989, Greenacre 1994β, 1993α, 1991, 1990 και 1984, Gifi 1996, Escofier & Pagès 1998, Lebart, Morineau & Piron 2000, Μάρκος & Παπαδημητρίου 2003). Αυτό οφείλεται στο γεγονός ότι η βασική δομή των στηλών του πίνακα **S** είναι ίδια με αυτή των στηλών (γραμμών) του πίνακα $\mathbf{S}^{\mathbf{T}}\mathbf{S}$ ($\mathbf{S}\mathbf{S}^{\mathbf{T}}$). Οι σχέσεις που συνδέουν τις δύο αναλύσεις είναι οι εξής (βλέπε Ενότητα 2.2.12 και σχέσεις [2.8] και [2.9]:

$$\mathbf{S} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathbf{T}} \qquad [2.67]$$

$$\mathbf{S}^{\mathbf{T}}\mathbf{S} = \mathbf{V}\mathbf{D}^{\mathbf{2}}\mathbf{V}^{\mathbf{T}} \qquad [2.68]$$

$$\mathbf{S}\mathbf{S}^{\mathbf{T}} = \mathbf{U}\mathbf{D}^{\mathbf{2}}\mathbf{U}^{\mathbf{T}} \qquad [2.69]$$

Από τις παραπάνω σχέσεις διαπιστώνουμε ότι τα δεξιά (αριστερά) χαρακτηριστικά διανύσματα του πίνακα **S** ταυτίζονται με τα ιδιοανύσματα του $\mathbf{S}^{\mathbf{T}}\mathbf{S}$ ($\mathbf{S}\mathbf{S}^{\mathbf{T}}$) και τα τετράγωνα των χαρακτηριστικών τιμών του **S** είναι ίσα με τις αντίστοιχες ιδιοτιμές του $\mathbf{S}^{\mathbf{T}}\mathbf{S}$ ($\mathbf{S}\mathbf{S}^{\mathbf{T}}$). Επειδή, σε κάθε περίπτωση, τα τετράγωνα των χαρακτηριστικών τιμών είναι ίσα με τις αδράνειες των αντίστοιχων παραγοντικών αξόνων, έχουμε ότι:

Αν λ_s είναι η αδράνεια του άξονα *s* από την ανάλυση του πίνακα **Z**, τότε $\lambda_s^2 = \lambda_{Bs}$ θα είναι η αδράνεια του ίδιου άξονα από την ανάλυση του **B**. Επομένως, η ολική αδράνεια I_B του πίνακα **B** θα δίνεται από τη σχέση:

$$I_B = \sum_{s=1}^{p} \lambda_s^2 = \sum_{s=1}^{p} \lambda_{Bs}, \quad \text{με} \quad p = j - q.$$

Κατά την εφαρμογή της ΠΑΑ στους πίνακες **Z** και **B**, οι τυποποιημένες συντεταγμένες των σημείων (στηλών για τον **Z** και γραμμών ή στηλών για τον **B**) στους υποχώρους προβολής είναι ίσες (Greenacre 1984, βλέπε και Κεφάλαιο 3, Ενότητα 3.3.1). Επομένως, το μόνο που αλλάζει είναι η κλίμακα μέτρησης του συστήματος συντεταγμένων. Στην περίπτωση του πίνακα **B**, οι τυποποιημένες συντεταγμένες κανονικοποιούνται μέσω των στοιχείων του πίνακα $\mathbf{D}^{\mathbf{2}}$ και όχι του **D**. Συνεπώς, οι σχετικές θέσεις των σημείων δεν μεταβάλλονται. Το μόνο μειονέκτημα κατά την ανάλυση του πίνακα *Burt* είναι ότι χάνεται πληροφορία για τα αντικείμενα και ο αρχικός πίνακας δεδομένων δεν μπορεί να ανασυσταθεί. Έτσι, η ΠΑΑ δεν μπορεί να συνδυαστεί άμεσα με μεθόδους όπως η Ιεραρχική Ταξινόμηση για την ομαδοποίηση των αντικειμένων και την περαιτέρω εμβάθυνση στην ερμηνεία των αποτελεσμάτων. Στο Κεφάλαιο 3 (Ενότητα 3.3.1) προτείνουμε μεθοδολογία, σύμφωνα με την οποία είναι δυνατό να έχουμε πληροφορία για τα αντικείμενα, έστω κι αν η μέθοδος εφαρμοστεί στον πίνακα *Burt*.

Ο πίνακας **B** χαρακτηρίζεται από δύο σημαντικές στατιστικές ιδιότητες:

1) Περιέχει όλη την πληροφορία “πρώτης τάξης”, στην οποία θα βασιστεί η ανάλυση της εσωτερικής δομής των μεταβλητών, και έχει τον ίδιο ρόλο με αυτόν του πίνακα “διασπορών – συνδυασπορών” στην Ανάλυση σε Κύριες Συνιστώσες, στην Παραγοντική Ανάλυση και στην Γραμμική Παλινδρόμηση (Greenacre 1994β, Verkuilen 2001). Με άλλα λόγια, η εφαρμογή της ΠΑΑ στον πίνακα *Burt* αναλύει τις αλληλεπιδράσεις (συσχετίσεις) των μεταβλητών ανά ζεύγη, αγνοώντας τις αλληλεπιδράσεις των μεταβλητών σε τριάδες, τετράδες κ.ο.κ. Το ίδιο ισχύει και σε άλλες πολυμεταβλητές αναλύσεις όπως είναι η Παραγοντική Ανάλυση και η Ανάλυση σε Κύριες Συνιστώσες, στις οποίες λαμβάνονται υπόψη μόνο οι ροπές πρώτης και δεύτερης τάξης. Αυτός ο περιορισμός δεν μπορεί να θεωρηθεί σημαντικό μειονέκτημα της μεθόδου (McDonald 1981, Van de Geer 1993α, Gifi 1996, Clausen 1998, Verkuilen 2001), γιατί, σε πρακτικό επίπεδο, οι αλληλεπιδράσεις δεύτερης και ανώτερης τάξης είναι πάρα πολύ δύσκολο, έως αδύνατο, να ερμηνευτούν κλινικά. Επίσης, σπάνια αναδεικνύονται ενδιαφέρουσες δομές των αλληλεπιδράσεων αυτών, όταν οι αλληλεπιδράσεις κατά ζεύγη (πρώτης τάξης) δεν είναι σημαντικές. Τέλος, απαιτούνται τεράστια, σε μέγεθος, δείγματα, ώστε τα αποτελέσματα και η ερμηνεία των αλληλεπιδράσεων να κινούνται σε αποδεκτά όρια αξιοπιστίας.

Μια ενδιαφέρουσα παρατήρηση προκύπτει από το γεγονός ότι, παρόλο που ο πίνακας $\mathbf{Z_{0\text{-}1}}$ περιέχει όλη την πληροφορία τού υπό εξέταση φαινομένου, δηλαδή την κοινή κατανομή όλων των j ιδιοτήτων των q μεταβλητών, ωστόσο, λόγω της ισοδυναμίας του με τον πίνακα **B**, η εφαρμογή της ΠΑΑ στον $\mathbf{Z_{0\text{-}1}}$ ισοδυναμεί, τελικά, με την ανάλυση διμεταβλητών και όχι πολυμεταβλητών συσχετίσεων. Στο Κεφάλαιο 4 αποδεικνύουμε νέες αναλυτικές σχέσεις που συνδέουν τις αδράνειες απλών πίνακων συμπτώσεων, γενικευμένων και λογικών δύο ή περισσότερων μεταβλητών.

2) Η εφαρμογή της ΠΑΑ στον πίνακα *Burt* αναλύει ταυτόχρονα τις αποκλίσεις από την κατάσταση ανεξαρτησίας των μεταβλητών ανά δύο (Van der Heijden & De Leeuw 1989, Verkuilen 2001). Αυτό, όπως θα δούμε στο Κεφάλαιο 4 (Ενότητα 4.8.2) επιτρέπει τη σύνδεση της μεθόδου με το στατιστικό έλεγχο χ^2 (στην περίπτωση που τα διαθέσιμα δεδομένα έχουν συγκεντρωθεί με απλή τυχαία δειγματοληψία).

2.3.3.6 Σημαντικοί Άξονες –Σημαντικά Σημεία-Σημαντικές Μεταβλητές

Όπως στη διμεταβλητή περίπτωση έτσι και στην πολυμεταβλητή εκδοχή της ΠΑΑ χρησιμοποιούνται, συνήθως, εμπειρικά κριτήρια για τον προσδιορισμό των σημαντικών σημείων και αξόνων (Saporta & Tambrea, 1993). Πιο συγκεκριμένα, σημαντικά θεωρούνται τα σημεία με $CTR \geq 1/j$ και $COR \geq 0{,}20$ (Καραπιστόλης 1999, Bendixen 2003). Σύμφωνα με τους Le Roux και Rouanet (2004), μια μεταβλητή θεωρείται σημαντική για κάποιον άξονα αν η αθροιστική συνεισφορά της είναι μεγαλύτερη από $1/q$, πάνω, δηλαδή, από τη μέση αναμενόμενη αδράνεια $\left(I/p = (j-q)/q(j-q)\right)$ των αξόνων κάτω από την υπόθεση της τυχαίας διακύμανσης των δεδομένων. Οι ίδιοι ερευνητές προτείνουν να επιλέγονται μόνο οι μεταβλητές που ερμηνεύουν αθροιστικά τουλάχιστον το 75% της αδράνειας του άξονα. Όταν η ανάλυση εφαρμόζεται στον πίνακα $\mathbf{Z_{0\text{-}1}}$, σημαντικοί θεωρούνται οι άξονες με αδράνεια μεγαλύτερη από $1/q$ (Van Rijckevorsel 1987, Greenacre 1993α και 1984, Nishisato 1994 και 1980, Le Roux & Rouanet 2004, βλέπε και Ενότητα 2.3.4.2, Παρατήρηση Η). Αν η ΠΑΑ εφαρμοστεί στον πίνακα *Burt*, τότε λόγω της σχέσεων που συνδέουν τις δύο αναλύσεις (βλέπε σχέσεις [2.67], [2.68] και [2.69]), σημαντικοί θεωρούνται οι παραγοντικοί άξονες με αδράνειες μεγαλύτερες από $1/q^2$ (βλέπε και Greenacre, 1984). Γενικά, το πρόβλημα της στατιστικής σημαντικότητας των παραγοντικών αξόνων αντιμετωπίζεται σχεδόν αποκλειστικά με μεθόδους επαναδειγματοληψίας (Ringrose 1992, Van de Geer 1993β, Markus 1994α και 1994β, Gifi 1996, Michailidis 1996, Bond & Michailidis 1997 και 1996, Michailidis & De Leeuw 1998). Στο Κεφάλαιο 6 (Ενότητα 6.4) προτείνουμε ένα "Μη Παραμετρικό" 100(1-α)% διάστημα εμπιστοσύνης για τον εντοπισμό των στατιστικά σημαντικών αξόνων στην πολυμεταβλητή ΠΑΑ.

Παρατήρηση 2.4

Στο πλαίσιο της φιλοσοφίας των μεθόδων της Ανάλυσης Δεδομένων, μια *a priori* προσέγγιση (έλεγχος, κριτήριο), με ελάχιστες τεχνικές και θεωρητικές προϋποθέσεις, για τη σύμπτυξη κλάσεων μεταβλητών, έτσι ώστε να διατηρείται η ομοιογένεια και η δομή του πίνακα συμπτώσεων, είναι μάλλον δύσκολο να βρεθεί. Αυτό οφείλεται στο γεγονός ότι η Ανάλυση Δεδομένων δεν είναι μια "τυφλή" διαδικασία, ανεξάρτητη

από τους στόχους του εκάστοτε ερευνητή και από το γνωστικό-επιστημονικό πεδίο του φαινομένου που μελετά. Για παράδειγμα, σε ένα πίνακα δεδομένων, όπου οι γραμμές του αντιστοιχούν στους νομούς της Ελλάδας ένα κριτήριο μπορεί να υποδείξει ότι οι νομοί Χαλκιδικής και Αττικής μπορούν να συγχωνευθούν. Όμως, αυτό μπορεί να μην έχει "κλινική" σημασία για την ερμηνεία των αποτελεσμάτων ή να μην είναι επιτρεπτό με βάση τους στόχους της συγκεκριμένης μελέτης (Le Roux & Chiche, 2004).

Αντίθετα, στην Επαγωγική Στατιστική υπάρχουν αρκετές *post hoc* προσεγγίσεις για τον καθορισμό των κλάσεων γραμμών ή στηλών που μπορούν συγχωνευτούν (Goodman 1968 και 1981, Gabriel 1966, Hirotsu 1983, Gilula 1986 και 1985, Gilula & Krieger 1989 και 1983, Ocerin, Mohedano & Segador 1999). Οι προτεινόμενες μεθοδολογίες στηρίζονται κυρίως στα αποτελέσματα που προκύπτουν από την εφαρμογή και σύγκριση συσχετιστικών υποδειγμάτων ή σε στατιστικούς ελέγχους ομοιογένειας. Οι μέθοδοι απαιτούν τυχαία δείγματα και, εφόσον στηρίζονται σε ελέγχους σημαντικότητας, έχουν αρκετές τεχνικές και θεωρητικές προϋποθέσεις. Το ενδιαφέρον είναι ότι η προσπάθεια για την εξεύρεση κριτηρίων για τη συγχώνευση κατηγοριών μεταβλητών φαίνεται να οδηγεί τελικά στην εφαρμογή της ΠΑΑ. Ειδικότερα, σύμφωνα με τους Gilula (1986) και Andersen (1991) μια ικανή και αναγκαία συνθήκη για την ομαδοποίηση δύο κλάσεων είναι οι συντεταγμένες των προβολών των αντίστοιχων σημείων επί των παραγοντικών αξόνων να είναι ίσες για όλους τους άξονες. Επομένως, σε ένα πρώτο επίπεδο, μπορεί να εφαρμοστεί η ΠΑΑ ώστε να υποδείξει σημεία γραμμών ή στηλών με ίδιο ή παρόμοιο προφίλ και, στη συνέχεια, να ελεγχθεί η ομοιογένεια τους με *post hoc* ελέγχους σημαντικότητας.

2.3.4 Η Ανάλυση Ομοιογένειας

Όπως αναφέρθηκε στο Κεφάλαιο1, οι μέθοδοι του συστήματος *GIFI,* της Ολλανδικής Σχολής Ανάλυσης Δεδομένων, βασίζονται στη διαδικασία της «Ανάλυσης Ομοιογένειας (ΑΟ) με Εναλλασσόμενα Ελάχιστα Τετράγωνα» (*Homogeneity Analysis by Alternating Least Squares-HOMALS*), η οποία είναι αντίστοιχη με την πολυμεταβλητή εκδοχή της ΠΑΑ στη Γαλλική Σχολή. Η Ανάλυση Ομοιογένειας αποτελεί μια "εκμοντερνισμένη" εκδοχή της μεθόδου κλιμάκωσης που πρότεινε ο Guttman (1941). Το γενικό πρόβλημα, στο οποίο καλείται να δώσει λύση η ΑΟ,

μπορεί να διατυπωθεί ως εξής (Israëls 1987, Van Buuren & De Leeuw 1992, Van de Geer 1993β, Gifi 1996, Michailidis 1996, Michailidis & De Leeuw 2000 και 1998, SPSS Inc. 2004α και 1997):

Δοθέντος ενός συνόλου δεδομένων N αντικειμένων και q κατηγορικών μεταβλητών με j_i κατηγορίες η κάθε μία (i=1,…, q και $\sum_{i=1}^{q} j_i = j$), να βρεθούν και να ανατεθούν αριθμητικές τιμές ή αλλιώς βαθμοί (*scores*) στα αντικείμενα και στις κατηγορίες των μεταβλητών, ώστε να ελαχιστοποιείται η παρακάτω συνάρτηση απώλειας (*loss function*):

$$\sigma\left(\mathbf{X}; \mathbf{Y_1}, \mathbf{Y_2}, \cdots \mathbf{Y_q}\right) = q^{-1}\sum_{i=1}^{q} SSQ\left(\mathbf{X} - \mathbf{Z_i}\mathbf{Y_i}\right) =$$

$$= q^{-1}\sum_{i=1}^{q} trace\left(\left(\mathbf{X} - \mathbf{Z_i}\mathbf{Y_i}\right)^{\mathbf{T}}\left(\mathbf{X} - \mathbf{Z_i}\mathbf{Y_i}\right)\right), \qquad [2.70]$$

όπου ο συμβολισμός *SSQ*(**H**) δηλώνει το άθροισμα τετραγώνων των στοιχείων του πίνακα **H**, **X** είναι ο $N\times s$ πίνακας με τις βέλτιστες αριθμητικές τιμές των αντικειμένων, $\mathbf{Y_i}$ είναι ο $j_i\times s$ πίνακας με τις βέλτιστες αριθμητικές τιμές των κατηγοριών της μεταβλητής i και $\mathbf{Z_i}$ είναι ο $N\times j_i$ λογικός πίνακας 0-1 που αντιστοιχεί στη μεταβλητή i.

Η τιμή του s καθορίζεται από το χρήστη της μεθόδου. Σε κάθε περίπτωση η μέγιστη τιμή του s είναι $p = j - q$. Για να αποφευχθεί η τετριμμένη λύση $\mathbf{X}=\mathbf{0}$ και $\mathbf{Y_i}=\mathbf{0}$ $\forall$ i, επιβάλλονται οι παρακάτω περιορισμοί κανονικοποίησης:

i) $\mathbf{X}^{\mathbf{T}}\mathbf{X}=N\mathbf{I}$
ii) $\mathbf{1}^{\mathbf{T}}\mathbf{X}=\mathbf{0}$,

όπου **I** είναι ο $s\times s$ μοναδιαίος πίνακας, **0** είναι πίνακας με τα στοιχεία του όλα ίσα με 0 και **1** είναι το $N\times 1$ διάνυσμα στήλης με στοιχεία ίσα με 1.

Ο περιορισμός ii) εξασφαλίζει ότι ο μέσος όρος των βαθμών των αντικειμένων θα είναι ίσος με μηδέν και, συνεπώς, κατά τη γραφική απεικόνισή τους θα είναι κεντροποιημένα γύρω από την αρχή των αξόνων. Ο περιορισμός i) τυποποιεί τις τιμές

των αντικειμένων ώστε το άθροισμα των τετραγώνων τους να είναι ίσο με *N* και, στην περίπτωση που s≥2, εξασφαλίζει ότι οι στήλες του πίνακα **X** θα είναι μεταξύ τους ορθογώνιες (γραμμικά ανεξάρτητες). Η έννοια της «πλήρους ομοιογένειας» ορίζεται σε σχέση με την εξεύρεση και ανάθεση βέλτιστων τιμών στα αντικείμενα και στις κατηγορίες των μεταβλητών, ώστε να παρουσιάζουν τέλεια εσωτερική συνέπεια (Israëls 1987, Van Rijckevorsel 1987, Van de Geer 1993β, Michailidis 1996, Michailidis & De Leeuw 2000 και 1998), δηλαδή, όταν ισχύει:

$$\mathbf{X} = \mathbf{Z}_1\mathbf{Y}_1 = \ldots = \mathbf{Z}_i\mathbf{Y}_i \, .$$

Η φυσική ερμηνεία της παραπάνω σχέσης είναι ότι όλες οι μεταβλητές μετρούν το ίδιο χαρακτηριστικό ή ιδιότητα των αντικειμένων. Ας υποθέσουμε ότι έχουν υπολογιστεί οι βέλτιστες τιμές των κατηγοριών και ότι οι νέες ποσοτικοποιημένες μεταβλητές έχουν μέσο όρο ίσο με 0. Τότε η πληροφορία που περιέχεται στον αρχικό πίνακα δεδομένων μπορεί να συνοψιστεί λαμβάνοντας για κάθε αντικείμενο μόνο τους μέσους όρους των ιδιοτήτων που τα χαρακτηρίζουν. Ο βαθμός ομοιογένειας μπορεί να εκφραστεί ως ο λόγος δύο διακυμάνσεων: της διακύμανσης των μέσων όρων των αντικειμένων προς την ολική διακύμανση των ποσοτικοποιημένων μεταβλητών (Van de Geer, 1993β). Συχνά ο λόγος αυτός αναφέρεται και ως «Λόγος Συσχέτισης» (*Correlation Ratio-CR*) (Guttman 1941, Nishisato 1980, Greenacre 1984) και συνδέει την ΑΟ με έννοιες που χρησιμοποιούνται στην Ανάλυση Διακύμανσης. Αντίστροφα, ας υποθέσουμε ότι έχουν υπολογιστεί οι βέλτιστοι βαθμοί για τα αντικείμενα. Τότε το ζητούμενο είναι να ανατεθούν τιμές στις κατηγορίες των αρχικών μεταβλητών, έτσι ώστε οι νέες ποσοτικοποιημένες μεταβλητές να αποκλίνουν, με την έννοια των ελαχίστων τετραγώνων, όσο το δυνατό λιγότερο από το διάνυσμα των βαθμών των αντικειμένων. Βέβαια, στην περίπτωση που δύο αντικείμενα χαρακτηρίζονται από τις ίδιες ιδιότητες, τότε οι βαθμοί τους θα πρέπει να είναι ίσοι. Μέσω της ΑΟ επιτυγχάνεται η εύρεση ενός ολικού και αρκετών τοπικών "μεγίστων" για το Λόγο Συσχέτισης *CR*. Οι διαφορετικές λύσεις που αντιστοιχούν στα "μέγιστα" χαρακτηρίζουν τους παραγοντικούς άξονες στους οποίους αναλύεται η βασική δομή του πίνακα δεδομένων.

Επειδή στην πράξη είναι μάλλον δύσκολο να επιτευχθεί η πλήρης ομοιογένεια (Guttman 1944, Gifi 1996), είναι προτιμότερο η λύση να αναζητηθεί μέσω της

ελαχιστοποίησης των αποκλίσεων από την τέλεια εσωτερική συνέπεια. Τις αποκλίσεις αυτές δηλώνει η συνάρτηση απώλειας μέσω της σχέσης:

$$\sigma\left(\mathbf{X};\mathbf{Y_1},\mathbf{Y_2},\cdots\mathbf{Y_q}\right)=q^{-1}\sum_{i=1}^{q}SSQ\left(\mathbf{X}-\mathbf{Z_i}\mathbf{Y_i}\right).$$

Σε αντιστοιχία με την ΠΑΑ της Γαλλικής Σχολής, ο πίνακας **X** περιέχει τις παραγοντικές συντεταγμένες των αντικειμένων, ενώ οι $\mathbf{Y_i}$ τις συντεταγμένες των κατηγοριών (κλάσεων, ιδιοτήτων) των μεταβλητών στους *s* άξονες.

Η σχέση [2.70] αποτελεί τον πυρήνα του συστήματος *GIFI*. Άλλες μέθοδοι της Ολλανδικής Σχολής προκύπτουν είτε από παραλλαγές της [2.70] είτε από την επιβολή διαφορετικών περιορισμών στις τιμές των αντικειμένων και κατηγοριών (βλέπε De Leeuw, Young & Takane 1976, Young, De Leeuw & Takane 1976, Takane, Young & De Leeuw 1977, Young, Takane & De Leeuw 1978, Perreault & Young 1980, Young 1981, Van Rijckevorsel 1987, Israëls 1987, Van Rijckevorsel & De Leeuw 1988, Bekker & De Leeuw 1988, Van Buuren & De Leeuw 1992, Van der Burg & De Leeuw 1994, Gifi 1996, Michailidis 1996, Michailidis & De Leeuw 2000, 1998 και 1997, Adachi 2002, Hwang & Takane 2002, SPSS Inc. 2004α και 1997). Η ελαχιστοποίηση της [2.70], κάτω από τους περιορισμούς i) και ii), επιτυγχάνεται με την εφαρμογή του επαναληπτικού αλγόριθμου *Alternating Least Squares-ALS*. Να τονιστεί ότι ο χρήστης της μεθόδου θα πρέπει να προκαθορίσει την τιμή του *s*, δηλαδή το πλήθος των παραγοντικών αξόνων που επιθυμεί ως λύση. Τα βασικά βήματα του αλγόριθμου ALS παρουσιάζονται στην Ενότητα Β3 του Παραρτήματος Β.

Με βάση τον αλγόριθμο ALS υλοποιούνται οι σημαντικότερες από τις μεθόδους των Ολλανδών, οι οποίες περιλαμβάνονται στο υποσύστημα *Categories* του στατιστικού πακέτου SPSS (SPSS Inc. 2004α και 1997, Meulman & Heiser 2004 και 2001). Το SPSS είναι το μόνο εμπορικό λογισμικό το οποίο μπορεί να θεωρηθεί ότι αντιπροσωπεύει, σε μεγάλο βαθμό, τις μεθόδους της Ολλανδικής Σχολής. Το συγκεκριμένο λογισμικό μέχρι και την έκδοση 12 περιλάμβανε τη διαδικασία *HOMALS*. Από την έκδοση 13 και μετά, η *HOMALS* αντικαταστάθηκε από τη διαδικασία *Multiple Correspondence Analysis*, η οποία είναι πιο "κοντά" στην ΠΑΑ της Γαλλικής Σχολής. Η ΑΟ έχει υλοποιηθεί και στο λογισμικό WinVista (Young

1992, Young, Faldowski & McFarlane 1993, Young, Valero-Mora & Ledesma-Mouripo 2000) καθώς και στη συναρτησιακή γλώσσα Lisp-Stat (Bond & Michailidis, 1997 και 1996). Στο πλαίσιο του συστήματος *GIFI*, δεν έχει εννοιολογικό περιεχόμενο η αδράνεια όπως ορίζεται στη Γαλλική Σχολή, ενώ τα αριθμητικά αποτελέσματα της μεθόδου είναι αρκετά "φτωχά" και δεν προσφέρουν δείκτες για τη διευκόλυνση της ερμηνείας. Για παράδειγμα, αφού ο χρήστης καθορίσει των αριθμό *s* των παραγοντικών αξόνων, τότε για κάθε άξονα υπολογίζεται το τετράγωνο της αντίστοιχης χαρακτηριστικής τιμής, ενώ για τις κατηγορίες των μεταβλητών υπολογίζονται οι συντεταγμένες τους στους *s* άξονες. Για τους Ολλανδούς, δείκτες όπως οι *CTR*, *COR* και *QLT* δεν έχουν φυσική ερμηνεία και για αυτό δεν υπολογίζονται. Επίσης, δεν υπολογίζεται η ολική αδράνεια και, συνεπώς, ούτε τα ποσοστά και τα αντίστοιχα αθροιστικά ποσοστά ερμηνείας των αξόνων. Ακόμη, στη διαδικασία *HOMALS* του SPSS δεν υπάρχει η δυνατότητα εισαγωγής συμπληρωματικών σημείων. Τέλος, δεν υπάρχει καμία σύνδεση με την απόσταση χ^2 ούτε και με τις έννοιες «μάζα» και «κέντρο βάρους», όπως αυτές αποκτούν νόημα και περιεχόμενο στο πνεύμα της Γαλλικής Σχολής. Το μόνο αριθμητικό αποτέλεσμα που υπολογίζεται στην AO, το οποίο συντελεί στην ερμηνεία των αποτελεσμάτων, είναι ο «Δείκτης Διακριτότητας» η_{is}^2 (*Discrimination Measure*) για κάθε μεταβλητή *i* σε κάθε άξονα *s*. Ο δείκτης παίρνει τιμές στο διάστημα [0, 1] και δίνεται από την παρακάτω σχέση (Van Rijckevorsel 1987, Van de Geer 1993β, Gifi 1996, Michailidis 1996, Michailidis & De Leeuw 1998):

$$\eta_{is}^2 = \frac{\mathbf{Y}_{\mathbf{i}(s)}^{\mathbf{T}} \mathbf{D}_{\mathbf{i}} \mathbf{Y}_{\mathbf{i}(s)}}{N}, \qquad [2.71]$$

όπου $\mathbf{Y}_{\mathbf{i}(s)}$ είναι ο πίνακας με στοιχεία τις αριθμητικές τιμές των κατηγοριών (κλάσεων) της μεταβλητής *i* επί του άξονα *s* και $\mathbf{D}_{\mathbf{i}}$ είναι ο διαγώνιος πίνακας με στοιχεία τις συχνότητες των κατηγοριών της μεταβλητής *i*.

Ο δείκτης η_{is}^2 εκφράζει τη σταθμισμένη μέση απόσταση (εις το τετράγωνο) των τιμών των κατηγοριών της μεταβλητής *i* στον άξονα *s* από την αρχή του συστήματος συντεταγμένων, δηλαδή τη διακύμανση της ποσοτικοποιημένης μεταβλητής *i*. Συνεπώς, μεταβλητές στις οποίες αντιστοιχεί υψηλή τιμή του δείκτη η^2

παρουσιάζουν μεγαλύτερη διασπορά επί του αντίστοιχου άξονα και άρα οι κατηγορίες της μεταβλητής διακρίνονται ή διαφοροποιούνται εντονότερα. Μπορεί να δειχθεί (Van Rijckevorsel 1987, Gifi 1996) ότι, στην περίπτωση που δεν υπάρχουν ελλείπουσες τιμές, ο δείκτης διακριτότητας της μεταβλητής *i* είναι ίσος με το τετράγωνο του συντελεστή γραμμικής συσχέτισης του *Pearson* μεταξύ τις βέλτιστα ποσοτικοποιημένης μεταβλητής $\mathbf{D_i Y_{i(}}_{s)}$ και της αντίστοιχης στήλης του πίνακα $\mathbf{X}_{(s)}$, δηλαδή των βέλτιστα ποσοτικοποιημένων αντικειμένων στον άξονα *s*. Κάτω από αυτή τη θεώρηση, η συνάρτηση απώλειας της σχέσης [2.70] μπορεί να γραφεί και ως εξής (Michailidis 1996, Michailidis & De Leeuw 1998):

$$N\left(p-\frac{1}{q}\sum_{i=1}^{q}\sum_{s=1}^{p}\eta_{is}^{2} \right)=N\left(p-\sum_{s=1}^{p}\lambda_{s} \right), \qquad [2.72]$$

όπου η ποσότητα $\lambda_s=\frac{1}{q}\sum_{i=1}^{q}\eta_{is}^{2}$ εκφράζει το μέσο όρο των δεικτών διακριτότητας των *q* μεταβλητών επί του *s*.

Εμπειρικά μπορεί να δειχθεί ότι η ποσότητα λ_s αντιστοιχεί στην αδράνεια του άξονα *s*, η οποία προκύπτει από την ανάλυση της βασικής δομής του πίνακα **S**[8] της σχέσης [2.59] [9]. Παρατηρούμε ότι στο πλαίσιο της ΑΟ η αδράνεια ενός άξονα συνδέεται με τους αντίστοιχους δείκτες διακριτότητας των μεταβλητών. Συνεπώς, μπορούμε να θεωρήσουμε ότι επί του πρώτου άξονα μεγιστοποιείται το τετράγωνο του μέσου συντελεστή συσχέτισης (*average squared correlation*) μεταξύ των βαθμών των αντικειμένων και των ποσοτικοποιημένων πλέον μεταβλητών (βλέπε Greenacre, 1993α). Σύμφωνα με το Michailidis (1996), η σημαντικότητα μιας μεταβλητής στην κατασκευή ενός παραγοντικού άξονα μπορεί να μετρηθεί μέσω της ποσότητας:

$$I_{mp}(i,s)=\eta_{is}^{2}-\lambda_{s}\ (s=1,\ldots,p,\ i=1,\ldots,q).$$

Από πρακτική σκοπιά, ο παραπάνω δείκτης μπορεί να χρησιμοποιηθεί για τον εντοπισμό των μεταβλητών που δεν είναι σημαντικές για την ερμηνεία των παραγοντικών αξόνων. Μάλιστα, αν διαπιστωθεί ότι κάποιες μεταβλητές έχουν

---

[8] Όπως αυτή ορίστηκε στην Ενότητα 2.2.12.

[9] Η διαπίστωση αυτή ελέγχεται μόνο με εμπειρικές συγκρίσεις των αποτελεσμάτων που παράγονται από τις δύο μεθόδους, αφού η ΑΟ μέσω του αλγόριθμου ALS δεν επιδέχεται αλγεβρικού χειρισμού.

χαμηλούς δείκτες I_{mp} στους πρώτους σε τάξη παραγοντικούς άξονες, τότε αυτές μπορούν να απομακρυνθούν από την ανάλυση, χωρίς σημαντική επίδραση στην ποιότητα των αποτελεσμάτων (Michailidis 1996, Michailidis & De Leeuw 1998).

Επίσης, μπορεί να δειχθεί (Van de Geer 1993β, Gifi 1996) ότι, στην περίπτωση που δεν υπάρχουν ελλείπουσες τιμές, για κάθε μεταβλητή το άθροισμα των δεικτών διακριτότητας σε όλους του δυνατούς $p=j-q$ άξονες είναι ίσο με το πλήθος των κατηγοριών της μεταβλητής πλην 1. Ισχύει, δηλαδή:

$$\sum_{s=1}^{p} \eta_{is}^2 = j_i - 1, \ \forall\ i=1,\ldots, q. \qquad [2.73]$$

Από τη σχέση [2.73] προκύπτει ότι το άθροισμα των δεικτών διακριτότητας κάθε μεταβλητής σε όλους τους δυνατούς p άξονες είναι σταθερό. Επομένως, είναι δυνατό κάποιες μεταβλητές με ασήμαντη συνεισφορά στους πρώτους άξονες να είναι σημαντικές σε επόμενους. Για το λόγο αυτό προτείνουμε να εκφράζονται οι δείκτες διακριτότητας και ως ποσοστά του αθροίσματός τους στους p άξονες. Πιο συγκεκριμένα, ορίζουμε ως «Σχετικό Δείκτη Διακριτότητας» η_{is}^{*2} της μεταβλητής i επί του άξονα s το λόγο:

$$\eta_{is}^{*2} = \frac{\eta_{is}^2}{\sum_{s=1}^{p} \eta_{is}^2} = \frac{\eta_{is}^2}{j_i - 1}.$$

Μέσω του Σχετικού Δείκτη Διακριτότητας είναι δυνατό να εντοπιστούν οι μεταβλητές που εκφράζουν τις “κύριες επιδράσεις” στους πρώτους παραγοντικούς άξονες, και οι μεταβλητές με “τοπική επίδραση”, οι οποίες είναι σημαντικές σε επόμενους άξονες, μικρής ερμηνευτικής ικανότητας. Για παράδειγμα, έστω ότι έχουμε 6 μεταβλητές Y_i με 3, 4, 3, 2, 3 και 2 κατηγορίες αντίστοιχα. Στην περίπτωση αυτή, το συνολικό πλήθος των κατηγοριών j είναι ίσο με 17, ενώ ο μέγιστος αριθμός παραγοντικών αξόνων p ισούται με 11. Ας υποθέσουμε, επίσης, ότι κατά την εφαρμογή της ΑΟ (ή της ΠΑΑ) προέκυψαν οι δείκτες διακριτότητας που δίνονται στον Πίνακα 2.8. Ο Πίνακας 2.9 παρουσιάζει τους αντίστοιχους σχετικούς δείκτες διακριτότητας. Από τον Πίνακα 2.8, παρατηρούμε ότι οι μεταβλητές Y_3 και Y_2 φαίνεται να έχουν την υψηλότερη συσχέτιση με τον άξονα F1 (η^2=0,562 και

η^2=0,530 αντίστοιχα). Την αμέσως υψηλότερη συσχέτιση δείχνει να έχει η Y_6 (η^2=0,331). Μάλιστα, οι δείκτες διακριτότητας των Y_3 και Y_2 στον F1 είναι περίπου της ίδιας τάξης μεγέθους. Αν όμως λάβουμε υπόψη τους σχετικούς δείκτες διακριτότητας (Πίνακας 2.9), τότε διαπιστώνουμε ότι η Y_3 έχει αρκετά μεγαλύτερη συνεισφορά στον άξονα F1 (η^{*2}=28,08) απ' ότι η Y_2 (η^{*2}=17,67). Μάλιστα, η μεταβλητή Y_6 αναδεικνύεται ως η μεταβλητή με την υψηλότερη συνεισφορά στον F1 (η^{*2}=33,09). Οι μεταβλητές Y_6 και Y_3 φαίνεται να ασκούν τις κύριες επιδράσεις στον πρώτο άξονα, αφού επί αυτού έχουν τους υψηλότερους σχετικούς δείκτες με μεγάλη διαφορά από τους υπόλοιπους. Η μεταβλητή Y_2 εμφανίζει τοπική επίδραση στους άξονες F1, F2, F4 και F7. Τοπική επίδραση φαίνεται να έχει και η μεταβλητή Y_1 στους F2, F8 και F3. Η μεταβλητή Y_4 χαρακτηρίζει κυρίως τον F3 (η^{*2}=52,74), ενώ φαίνεται να ασκεί τοπική επίδραση στον άξονα F6 (η^{*2}=19,91).□

Οι τετραγωνικές ρίζες των δεικτών διακριτότητας μπορούν να θεωρηθούν ως "φορτία" (*loadings*) ή συνεισφορές των μεταβλητών στους παραγοντικούς άξονες και έχουν στην ερμηνεία των αποτελεσμάτων την ίδια λειτουργικότητα με τους ανάλογους δείκτες (φορτία) στην Ανάλυση σε Κύριες Συνιστώσες (Van de Geer 1993β, Gifi 1996). Σύμφωνα με τους Hair *et al*. (1995), φορτία με απόλυτη τιμή μεγαλύτερη ή ίση από 0,30 έχουν γενικά πρακτική ή κλινική σημαντικότητα. Επομένως, ένας δείκτης διακριτότητας, μεγαλύτερος ή ίσος από 0,09, αντιστοιχεί σε φορτίο μεγαλύτερο ή ίσο με 0,30 και μπορεί να θεωρηθεί σημαντικός για τις περισσότερες πρακτικές εφαρμογές.

Ο δείκτης η_{is}^2, ως συνάρτηση των αποστάσεων των προβολών των κατηγοριών της μεταβλητής *i* στον άξονα *s* από την αρχή του συστήματος συντεταγμένων, είναι ανάλογος της αθροιστικής συνεισφοράς *CTR* της μεταβλητής. Πιο συγκεκριμένα, ισχύει (Greenacre, 1991):

$$\eta_{is}^2 = q\lambda_s \sum_{u_i=1}^{j_i} CTR(u_i, s), \quad (\forall\, i = 1,\ldots q, \forall\, s = 1,\ldots p),$$

όπου j_i είναι οι κλάσεις της μεταβλητής *i* και λ_s η αδράνεια του άξονα *s*.

Πίνακας 2.8: Μέτρα Διακριτότητας των Μεταβλητών στους Παραγοντικούς Άξονες

	Παραγοντικοί Άξονες (p=11)											
Μεταβλητές	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	*Άθροισμα*
Y_1	0,084	0,427	0,330	0,130	0,123	0,249	0,163	0,393	0,064	0,021	0,017	2
Y_2	0,530	0,513	0,160	0,360	0,415	0,222	0,344	0,009	0,096	0,246	0,107	3
Y_3	0,562	0,278	0,172	0,200	0,141	0,063	0,119	0,105	0,213	0,040	0,107	2
Y_4	0,020	0,071	0,527	0,005	0,004	0,199	0,040	0,000	0,065	0,032	0,036	1
Y_5	0,230	0,293	0,120	0,380	0,338	0,141	0,170	0,145	0,057	0,031	0,095	2
Y_6	0,331	0,112	0,014	0,174	0,046	0,024	0,011	0,040	0,101	0,147	0,000	1

Πίνακας 2.9: Σχετικά Μέτρα Διακριτότητας των Μεταβλητών στους Παραγοντικούς Άξονες

	Παραγοντικοί Άξονες (p=11)											
Μεταβλητές	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	*Άθροισμα*
Y_1	4,18	21,37	16,49	6,49	6,16	12,44	8,16	19,63	3,18	1,03	0,86	100
Y_2	17,67	17,09	5,32	11,99	13,82	7,38	11,45	0,29	3,21	8,21	3,55	100
Y_3	28,08	13,92	8,58	10,00	7,04	3,14	5,95	5,27	10,67	2,00	5,36	100
Y_4	2,01	7,11	52,74	0,51	0,41	19,91	3,98	0,04	6,47	3,20	3,61	100
Y_5	11,49	14,66	5,99	19,02	16,91	7,06	8,51	7,24	2,84	1,55	4,73	100
Y_6	33,09	11,20	1,36	17,36	4,61	2,44	1,09	4,02	10,14	14,65	0,03	100

2.3.4.1 Βασικές Ιδιότητες της Ανάλυσης Ομοιογένειας

Αν συνθέσουμε τα πορίσματα της σχετικής βιβλιογραφίας (Israëls 1987, De Leeuw & Van Rijckevorsel 1988, Hoffman & De Leeuw 1992, Van de Geer 1993α και 1993β, Greenacre 1993α και 1991, Heiser & Meulman 1994, Nishisato 1994 και 1980, Gifi 1996, Michailidis 1996, Michailidis & De Leeuw 2000 και 1998, Meulman & Heiser 2004 και 2001) διαπιστώνουμε ότι η ΑΟ έχει τις παρακάτω ιδιότητες:

1. Τα αντικείμενα και οι κατηγορίες (κλάσεις, ιδιότητες) των μεταβλητών μπορούν να προβληθούν σε κοινό χώρο.

2. Ένα σημείο που αντιστοιχεί σε κατηγορία μεταβλητής προβάλλεται στο κεντροειδές των αντικειμένων που χαρακτηρίζονται από την αντίστοιχη ιδιότητα. Ισχύει και το αντίστροφο, δηλαδή ένα σημείο που αντιστοιχεί σε κάποιο αντικείμενο προβάλλεται στο κεντροειδές των ιδιοτήτων που το χαρακτηρίζουν.

3. Αντικείμενα, τα οποία χαρακτηρίζονται από τις ίδιες ιδιότητες, έχουν δηλαδή το ίδιο προφίλ, λαμβάνουν την ίδια ποσοτικοποίηση επί των παραγοντικών αξόνων. Γενικά, η απόσταση μεταξύ δύο αντικειμένων συνδέεται με την ομοιότητα των προφίλ τους.

4. Μια μεταβλητή χαρακτηρίζει έναν άξονα, στο βαθμό που οι κατηγορίες της διακρίνονται σημαντικά επί αυτού. Δηλαδή, στο βαθμό που οι κατηγορίες προβάλλονται σε σχετικά απομακρυσμένες θέσεις επί του άξονα.

5. Αν μια ιδιότητα χαρακτηρίζει μόνο ένα αντικείμενο, τότε τα αντίστοιχα σημεία θα ταυτίζονται στο χώρο προβολής.

6. Τα αντικείμενα με σημαντικά διαφορετικό προφίλ από τα υπόλοιπα προβάλλονται σε απομακρυσμένη θέση ως προς την αρχή των αξόνων, ενώ αντικείμενα με προφίλ παρόμοιο με το μέσο προφίλ προβάλλονται κοντά στην αρχή.

7. Το σταθμισμένο άθροισμα των βαθμών των ποσοτικοποιημένων μεταβλητών σε κάθε άξονα είναι ίσο με μηδέν.

8. Οι λύσεις της ΑΟ είναι ιεραρχικά διακλαδιζόμενες. Αυτό σημαίνει ότι αν ο χρήστης ζητήσει μια λύση με s_1 άξονες και στη συνέχεια μια λύση με s_2 άξονες όπου $s_1 < s_2$, τότε οι πρώτοι s_1 άξονες της δεύτερης λύσης (με s_2 άξονες) ταυτίζονται με τους άξονες της πρώτης. Ο μέγιστος αριθμός παραγοντικών αξόνων που μπορούν να προκύψουν για οποιοδήποτε σύνολο δεδομένων είναι:

$$p = min\left\{N-1,\ j-max\left\{q^*, 1\right\}\right\},$$

όπου q^* είναι ο αριθμός των μεταβλητών χωρίς ελλείπουσες τιμές.

9. Οι λύσεις είναι διατεταγμένες. Αυτό σημαίνει ότι ο πρώτος άξονας έχει την απόλυτα μεγαλύτερη χαρακτηριστική τιμή. Ο δεύτερος έχει την αμέσως μεγαλύτερη χαρακτηριστική τιμή, κάτω από τον περιορισμό ότι θα πρέπει να είναι ορθογώνιος με τον πρώτο κ.ο.κ.

10. Οι βαθμοί των αντικειμένων σε διαδοχικούς άξονες είναι ασυσχέτιστοι ανά δύο, ενώ οι ποσοτικοποιήσεις των κατηγοριών των μεταβλητών δεν είναι. Μάλιστα, η μορφή της συσχέτισής τους είναι μάλλον απρόβλεπτη.

11. Οι λύσεις είναι αμετάβλητες από περιστροφή των αντικειμένων και των κατηγοριών στο χώρο των p διαστάσεων. Με άλλα λόγια, οι αποστάσεις μεταξύ των σημείων δεν μεταβάλλονται αν αλλάξει ο προσανατολισμός του αρχικού συστήματος συντεταγμένων.

Οι Ιδιότητες 1, 2, 3, 5, 6 και 7 είναι συνέπεια του αλγόριθμου ALS (Βήματα 2 και 3, βλέπε Ενότητα Β3 του Παραρτήματος Β), των περιορισμών i) και ii) και των κανονικοποιήσεων και κεντροποιήσεων που πραγματοποιούνται κατά τις επαναλήψεις του αλγόριθμου. Η Ιδιότητα 4 είναι αποτέλεσμα της σχέσης $\lambda_s = \frac{1}{q}\sum_{i=1}^{q}\eta_{is}^2$ που συνδέει τις αδράνειες των αξόνων με τους δείκτες διακριτότητας. Οι Ιδιότητες 8, 9 και 10 στηρίζονται στην ισοδυναμία των αποτελεσμάτων του αλγόριθμου ALS με αυτά που παράγονται από την εφαρμογή της SVD σε κατάλληλους πίνακες εισόδου, τους οποίους θα παρουσιάσουμε στην επόμενη ενότητα. Στην επόμενη ενότητα, επίσης, θα αιτιολογήσουμε και την ισχύ της Ιδιότητας 11.

Στις παραπάνω ιδιότητες στηρίζεται και η ερμηνεία των γραφικών αποτελεσμάτων που παράγονται από τη μέθοδο της Ανάλυσης Ομοιογένειας.

2.3.4.2 Παρατηρήσεις

A) Καταρχήν θα πρέπει να τονιστεί ότι τα αποτελέσματα της λύσης της ΑΟ ως προς τις κατηγορίες των μεταβλητών είναι συγκρίσιμα με αυτά που προκύπτουν από την εφαρμογή της ΠΑΑ στον πίνακα $\mathbf{Z}_{\mathbf{0\text{-}1}}$ (Israëls 1987, Van Buuren & De Leeuw 1992, Greenacre 1993α, Gifi 1996, Michailidis 1996, Michailidis & De Leeuw 2000 και 1998). Εμπειρικά μπορούμε να δείξουμε ότι στην ΑΟ οι συντεταγμένες των προβολών των κατηγοριών των μεταβλητών επί των παραγοντικών αξόνων είναι οι κύριες, όπως αυτές υπολογίζονται από την εφαρμογή της ΠΑΑ στον πίνακα $\mathbf{Z}_{\mathbf{0\text{-}1}}$. Από τους περιορισμούς i) και ii) έχουμε αντίστοιχα:

$$\mathbf{X}^{\mathbf{T}}\mathbf{X}=N\mathbf{I} \Rightarrow \frac{1}{N}\mathbf{X}^{\mathbf{T}}\mathbf{X}=\mathbf{I} \qquad [2.74]$$

και

$$\mathbf{1}^{\mathbf{T}}\mathbf{X}=\mathbf{0} \Rightarrow \frac{\mathbf{1}^{\mathbf{T}}\mathbf{X}}{N}=\mathbf{0}. \qquad [2.75]$$

Οι σχέσεις [2.74] και [2.75] δηλώνουν, αντίστοιχα, ότι η διασπορά των βέλτιστων τιμών των αντικειμένων σε κάθε άξονα θα είναι ίση με 1 και ο μέσος όρος τους ίσος με 0. Αν λάβουμε υπόψη και την Ιδιότητα 2 καταλήγουμε στο συμπέρασμα ότι για την κοινή προβολή των αντικειμένων και των κατηγοριών σε ένα κοινό χώρο θα πρέπει τα αντικείμενα να προβάλλονται με τυποποιημένες συντεταγμένες, ενώ οι κατηγορίες των μεταβλητών με κύριες. Συνεπώς, αν η προβολή γίνει στο επίπεδο, το αντίστοιχο διάγραμμα θα είναι *biplot* και η κανονικοποίηση των συντεταγμένων των προβολών των αντικειμένων και των κατηγοριών θα αντιστοιχεί σε Κύρια Κανονικοποίηση κατά Στήλες (*CPN*) (βλέπε Ενότητες 2.1.2.14.5 και 2.1.2.14.6). Δηλαδή, τα αποτελέσματα της ΑΟ είναι ίδια με αυτά που προκύπτουν από την εφαρμογή της ΠΑΑ στον πίνακα $\mathbf{Z}_{\mathbf{0\text{-}1}}$ με *CPN*. Κατ' αναλογία, με τη διμεταβλητή περίπτωση οι κύριες συντεταγμένες των αντικειμένων και των κατηγοριών των μεταβλητών επί ενός άξονα *s* θα έχουν μέσο όρο ίσο με 0 και διακύμανση ίση με την

αδράνεια του *s* (Israëls, 1987). Η ισοδυναμία της ΑΟ με την ΠΑΑ αρκεί για να αιτιολογήσει και τις περισσότερες από τις ιδιότητες που αναφέρθηκαν στην προηγούμενη ενότητα.

Β) Οι σχέσεις [Β3.1] και [Β3.2] (βλέπε Ενότητα Β3 του Παραρτήματος Β) είναι αντίστοιχες με τις βαρυκεντρικές σχέσεις της Γαλλικής Σχολής. Αποτελούν συνέπεια των Βημάτων 2 και 3 του αλγόριθμου ALS και αιτιολογούν την προβολή των αντικειμένων και των κλάσεων των μεταβλητών σε ένα κοινό χώρο. Στο πλαίσιο της Ολλανδικής Σχολής οι σχέσεις αυτές εκφράζουν την «Αρχή των Αντιστρόφων Μέσων» (*Principal of Reciprocal Averaging*) (Israëls 1987, Van Rijckevorsel 1987, Van de Geer, 1993β). Στην αρχή αυτή στηρίζεται και ο ομώνυμος υπολογιστικός επαναληπτικός αλγόριθμος (Hill 1974, Greenacre 1984, Bekker & De Leeuw 1988, Nishisato 1994 και 1980), τον οποίο συναντάμε ήδη στις εργασίες του Fisher (1940) και του Guttman (1941).

Γ) Σύμφωνα με την Ιδιότητα 11, ο κοινός χώρος προβολής των αντικειμένων και των κατηγοριών δεν ορίζεται μονοσήμαντα (Michailidis 1996, Michailidis & De Leeuw 2000 και 1998, SPSS Inc. 2004α). Ας υποθέσουμε ότι επιλέγουμε μια άλλη βάση για το χώρο στηλών του πίνακα **X** έτσι ώστε:

$$\mathbf{X}^{*} = \mathbf{XR}\,,$$

όπου **R** είναι $p \times N$ πίνακας περιστροφής με $\mathbf{R}^{\mathbf{T}}\mathbf{R} = \mathbf{R}\mathbf{R}^{\mathbf{T}} = \mathbf{I}$.

Από τη σχέση [Β3.1] (βλέπε Ενότητα Β3 του Παραρτήματος Β) έχουμε:

$$\mathbf{Y}_{\mathbf{i}}^{*} = \mathbf{D}_{\mathbf{i}}^{-1}\mathbf{Z}_{\mathbf{i}}^{\mathbf{T}}\mathbf{X}^{*} = \mathbf{D}_{\mathbf{i}}^{-1}\mathbf{Z}_{\mathbf{i}}^{\mathbf{T}}\mathbf{XR} = \hat{\mathbf{Y}}_{\mathbf{i}}\mathbf{R}\,. \qquad [2.76]$$

Από τη σχέση [2.76] είναι φανερό ότι αν ο χώρος στηλών του πίνακα **X** περιστραφεί μέσω του **R**, τότε θα περιστραφούν και οι τιμές των κατηγοριών. Επομένως, η λύση της ΑΟ δεν είναι μοναδική. Όπως και στη διμεταβλητή περίπτωση της ΠΑΑ, ο προσανατολισμός του συστήματος συντεταγμένων του υποχώρου προβολής είναι αυθαίρετος (Greenacre, 1984). Πληροφοριακά αναφέρουμε ότι οι Carroll και Green (1988) πρότειναν μέθοδο για την ανάλυση μιας κανονικοποιημένης εκδοχής του πίνακα *Burt* όπου ο προσανατολισμός του συστήματος συντεταγμένων ορίζεται μονοσήμαντα. Ο Van de Velden (2000) μελέτησε τη δυνατότητα ορθογώνιας

περιστροφής μεγίστης διακύμανσης (*varimax*) των παραγοντικών αξόνων με σκοπό την ανάδειξη απλούστερων, στην ερμηνεία, δομών. Η συγκεκριμένη μέθοδος περιστροφής εφαρμόζεται στην Ανάλυση σε Κύριες Συνιστώσες και την Παραγοντική Ανάλυση (Hair *et al*. 1995, Sharma 1996). Σύμφωνα με τον Greenacre (2006), στη διμεταβλητή εκδοχή της ΠΑΑ η περιστροφή δεν αιτιολογείται θεωρητικά, διότι τα προφίλ γραμμών και στηλών δεν είναι μεταβλητές, όπως για παράδειγμα στην Ανάλυση σε Κύριες Συνιστώσες, αλλά σημεία εφοδιασμένα με μάζα, τα οποία προβάλλονται όχι σε έναν υποχώρο ενός "απεριόριστου" διανυσματικού χώρου, αλλά σε έναν υποχώρο του περιορισμένου *simplex* στο οποίο ανήκουν τα προφίλ. Στην περίπτωση πολλών μεταβλητών, όπου το συνολικό πλήθος *j* των κατηγοριών των μεταβλητών είναι εν γένει μεγάλο, η περιστροφή του συστήματος συντεταγμένων είναι δυνατό να διευκολύνει σημαντικά μόνο τη διαγραμματική ερμηνεία των αποτελεσμάτων, ιδιαίτερα όταν υπάρχουν πολλές ελλείπουσες τιμές ή/και κλάσεις μεταβλητών χωρίς σημαντικό ενδιαφέρον[10]. Επίσης, οι προτεινόμενοι δείκτες βοήθειας στην ερμηνεία (συνεισφορές σημείων και αξόνων), αν και από μαθηματικής πλευράς είναι ικανοποιητικοί, ωστόσο δεν έχουν ισχυρή θεωρητική βάση (Van de Velden, 2000). Γενικά, ο Greenacre (2006) υποστηρίζει ότι η περιστροφή του συστήματος συντεταγμένων έχει λειτουργικότητα μόνο σε σπάνιες περιπτώσεις.

Δ) Επίσης, οι ποσοτικοποιήσεις μεταβλητές δεν είναι γραμμικά ανεξάρτητες στο χώρο των *p* διαστάσεων. Αυτό οφείλεται στο ότι μια μεταβλητή με *k* κλάσεις δεν μπορεί να έχει περισσότερες από *k*-1 ανεξάρτητες ποσοτικοποιήσεις (Van de Geer, 1993α και 1993β). Έτσι, σε μια λύση της ΑΟ με περισσότερους από *k*-1 παραγοντικούς άξονες θα υπάρχει γραμμική εξάρτηση μεταξύ των ποσοτικοποίησεων της μεταβλητής.

Ε) Μπορεί να δειχθεί ότι η λύση της ΑΟ είναι δυνατό να προκύψει και από την εφαρμογή της Διάσπασης σε Χαρακτηριστικές Τιμές SVD στους παρακάτω πίνακες (Michailidis 1996, Michailidis & De Leeuw 1998, De Leeuw, Wang & Michailidis 1999):

---

[10] Για παράδειγμα οι απαντήσεις του τύπου «Δεν γνωρίζω», «Δεν απαντώ» και «Δεν έχω γνώμη», οι οποίες εμφανίζονται συχνά σε έρευνες με ερωτηματολόγιο.

$$\mathbf{P}^{*} = \frac{1}{q}\sum_{i=1}^{q}\mathbf{P}_i\,, \qquad [2.77]$$

όπου $\mathbf{P_i} = \mathbf{Z_i}\mathbf{D_i^{-1}}\mathbf{Z_i^T}$,

και

$$\mathbf{P}^{**} = q^{-1/2}\,\ell\left(\mathbf{Z_{0\text{-}1}}\right)\mathbf{D}^{-1/2}, \qquad [2.78]$$

όπου $\ell(*) = \mathbf{I} - \frac{\mathbf{1}\mathbf{1}^{\mathbf{T}}}{\mathbf{1}^{\mathbf{T}}\mathbf{1}}$ είναι τελεστής κεντροποίησης που αφαιρεί από τα στοιχεία στηλών του πίνακα τους μέσους όρους των αντίστοιχων στηλών και $\mathbf{D} = \oplus_{i=1}^{q}\mathbf{D_i}$ είναι το ευθύ άθροισμα των πινάκων $\mathbf{D_i}$ δηλαδή:

$$\mathbf{D} = \begin{bmatrix} \mathbf{D_1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D_2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{D_q} \end{bmatrix}.$$

Οι πίνακες που ορίζονται από τις σχέσεις [2.77] και [2.78] θα μπορούσαν να αποτελέσουν εναλλακτικούς πίνακες εισόδου στον αλγόριθμο της ΠΑΑ, όπως αυτή εφαρμόζεται στο πλαίσιο της Γαλλικής Σχολής. Η ισοδυναμία των αποτελεσμάτων που προκύπτουν από την εφαρμογή του αλγόριθμου ALS και της SVD στους παραπάνω πίνακες αιτιολογεί την ισχύ των Ιδιοτήτων 8, 9 και 10.

ΣΤ) Για τους Ολλανδούς, ο πίνακας $\mathbf{Z_{0\text{-}1}}$ και ο αντίστοιχος πίνακας *Burt* **B** δεν αποτελούν καλά ορισμένους πίνακες συχνοτήτων διπλής εισόδου, ώστε η εφαρμογή της πολυμεταβλητής ΠΑΑ να θεωρηθεί ως μια απλή γενίκευση της διμεταβλητής εκδοχής (De Leeuw & Van Rijckevorsel 1988, De Leeuw 1993, Saporta & Tambrea 1993, Gifi 1996). Δυσκολίες φαίνεται να υπάρχουν και στην αποδοχή της καταλληλότητας της χ^2 απόστασης στην περίπτωση των πινάκων $\mathbf{Z_{0\text{-}1}}$ και **B**. Τις ίδιες ενστάσεις φαίνεται να έχει και ο Greenacre (2005, 1994β, 1993γ, 1993α, 1991, 1990, 1989 και 1988α) ιδιαίτερα στην περίπτωση του πίνακα *Burt*. Αντίθετα, στη Γαλλική Σχολή και οι δύο πίνακες είναι καλά ορισμένοι και τεκμηριώνεται η καταλληλότητα της χρήσης της απόστασης χ^2 (Benzécri & *Collaborateurs* 1973,

Lebart, Morineau & Warwick 1984, Israëls 1987, SAS Institute 1990, Benzécri 1992, Escofier & Pagès 1998, Καραπιστόλης 1999, Μπεχράκης 1999, Lebart, Morineau & Piron 2000, Le Roux & Rouanet 2004, Παπαδημητρίου 2006, 2004 και 1994). Στο Κεφάλαιο 3 (Ενότητες 3.2 και Γ1 του Παραρτήματος Γ) προτείνουμε μέθοδο σύμφωνα με την οποία μπορούμε να προγραμματίσουμε το SPSS, ώστε η διαδικασία *Correspondence Analysis* να δέχεται ως είσοδο τον πίνακα *Burt* και τα παραγόμενα αποτελέσματα να είναι σύμφωνα με την παράδοση της Γαλλικής Σχολής Ανάλυσης Δεδομένων.

Ζ) Στα προηγούμενα είδαμε ότι κατά την εφαρμογή του αλγόριθμου ALS θα πρέπει ο χρήστης να προκαθορίσει τον αριθμό των διαστάσεων (αξόνων) που επιθυμεί. Αυτό είναι κοινό χαρακτηριστικό σχεδόν όλων των μεθόδων βέλτιστης κλιμάκωσης της Ολλανδικής Σχολής. Αν και από υπολογιστική σκοπιά οι μέθοδοι παρουσιάζουν σχετική αποτελεσματικότητα, περιορίζοντας το χρόνο επεξεργασίας και τις απαιτήσεις σε μνήμη του Η/Υ, ωστόσο η γεωμετρική πολυδιάστατη θεώρηση των δεδομένων δεν θεωρείται δεδομένη *a priori*, όπως συμβαίνει στη Γαλλική Σχολή (Van Rijckevorsel, 1987). Συχνά το ζητούμενο είναι μόνο η βέλτιστη ποσοτικοποίηση των κατηγορικών μεταβλητών ή/και η βέλτιστη ταξιθέτηση των αντικειμένων, που επιτυγχάνονται στον πρώτο παραγοντικό άξονα της ΑΟ (ή της ΠΑΑ). Αυτό έχει ως αποτέλεσμα η έννοια της ομοιογένειας να συνδέεται με το ερώτημα του κατά πόσο το υπό εξέταση φαινόμενο είναι μονοδιάστατο, δηλαδή αν οι μεταβλητές που συμμετέχουν στην ανάλυση αποτελούν μία ενιαία κλίμακα, η οποία μετρά τελικά το "ίδιο πράγμα" (π.χ. χαρακτηριστικό, ιδιότητα, λανθάνουσα δομή, παράγοντα). Αν έχουμε στη διάθεσή μας μια κλίμακα με ποσοτικές μεταβλητές, τότε αυτή θεωρείται ομοιογενής ή αξιόπιστη (εσωτερικά συνεπής) όταν όλες οι μεταβλητές που την αποτελούν είναι γραμμικά συσχετισμένες (Carmines & Zeller 1979, Spector 1992, Traub 1994, Hair *et al*., 1995). Στην περίπτωση που η κλίμακα αποτελείται από μικτού τύπου μεταβλητές (ονομαστικές, διάταξης και ποσοτικές), τότε αυτή είναι "εν δυνάμει ομοιογενής" αν όλες οι μεταβλητές που την αποτελούν μπορούν να μετασχηματιστούν ή να ποσοτικοποιηθούν, έτσι ώστε η νέα κλίμακα που θα προκύψει να είναι ομοιογενής (Michailidis, 1996). Η ομοιογένεια μιας κλίμακας με κεντροποιημένες μεταβλητές, δηλαδή με μέσο όρο ίσο με μηδέν, συνδέεται με το άθροισμα τετραγώνων (μεταβλητότητα) των βαθμών μεταξύ των αντικειμένων (*between objects*) και το άθροισμα τετραγώνων των βαθμών ανάμεσα στα

αντικείμενα (*within objects*). Πλήρης ή τέλεια ομοιογένεια επιτυγχάνεται όταν το άθροισμα τετραγώνων των βαθμών μέσα στα αντικείμενα είναι ίσο με μηδέν (Traub 1994, Κ. Μπαγιάτης 1997). Έτσι, ο βαθμός της ομοιογένειας της κλίμακας μπορεί να μετρηθεί μέσω του λόγου του αθροίσματος τετραγώνων μεταξύ των αντικειμένων προς το ολικό άθροισμα τετραγώνων (Nishisato, 1980). Πρωταρχικός στόχος της ΑΟ είναι η ανάθεση βέλτιστων αριθμητικών τιμών στα αντικείμενα και στις κατηγορίες των ποιοτικών μεταβλητών, έτσι ώστε να μεγιστοποιείται η ομοιογένεια. Το ολικό μέγιστο της ομοιογένειας επιτυγχάνεται στον πρώτο παραγοντικό άξονα, η ιδιοτιμή του οποίου εκφράζει και το βαθμό ή την έντασή της (Tenenhaus & Young 1985, De Leeuw 1993 και 1988, Gower 1990, Greenacre 1993α και 1984, Nishisato 1994 και 1980, Gifi 1996, Michailidis 1996, Michailidis & De Leeuw 1998). Οι επόμενοι άξονες αποτελούν τοπικά μέγιστα. Η εξέταση των αλληλεπιδράσεων μεταξύ ων κατηγοριών των μεταβλητών, η ερμηνεία των παραγοντικών αξόνων και επιπέδων έρχονται σε δεύτερο επίπεδο (Greenacre, 1991).

Η) Θα πρέπει να τονιστεί ότι μια ενδεχόμενη υψηλή τιμή της πρώτης χαρακτηριστικής τιμής δεν συνεπάγεται υποχρεωτικά ότι οι μεταβλητές μετρούν την ίδια λανθάνουσα δομή. Στην περίπτωση αυτή, εκείνο που είναι δεδομένο είναι ότι οι συσχετίσεις των μεταβλητών ανά δύο είναι υψηλές. Βέβαια, υψηλή συσχέτιση παρουσιάζεται και όταν οι μεταβλητές οδηγούν στον ίδιο διαμελισμό των αντικειμένων, με άλλα λόγια, όταν οι μεταβλητές τμηματοποιούν τα αντικείμενα σε ομοιογενείς ομάδες με τις ίδιες ή σχεδόν ίδιες ιδιότητες (Meulman & Heiser, 2004). Έτσι, σύμφωνα με το Nishisato (1994 και 1980), είναι συχνά χρήσιμο τα αποτελέσματα της ΑΟ να συνοδεύονται και από μια εκτίμηση του δείκτη α_s του *Cronbach* για κάθε παραγοντικό άξονα (βλέπε Nunnally 1978, Carmines & Zeller 1979, Spector 1992, Norusis 1992α, Traub 1994, Hair *et al*. 1995, Κ. Μπαγιάτης 1997, Strub 2000). Ο δείκτης α_s είναι συνάρτηση της αδράνειας λ_s και του πλήθους q των μεταβλητών. Στην ΑΟ υπολογίζεται από τη σχέση (Nishisato 1994 και 1980, SPSS Inc. 2004α):

$$\alpha_s = 1 - \frac{(1-\lambda_s)}{(q-1)\lambda_s}. \qquad [2.79]$$

Ο δείκτης α_s αποτελεί μέτρο της εσωτερικής συνέπειας ή αξιοπιστίας του άξονα *s* και εκφράζει το βαθμό της συσχέτισης των βαθμών των αντικειμένων επί του *s* με τις αντίστοιχες ποσοτικοποιημένες μεταβλητές. Σύμφωνα με το Nishisato (1994 και 1980), επί του πρώτου παραγοντικού άξονα μεγιστοποιείται η τιμή του δείκτη α_s (βλέπε και Heiser & Meulman, 1994). Η μέγιστη δυνατή τιμή του δείκτη είναι 1, ενώ μπορεί να πάρει και αρνητικές τιμές. Κατά σύμβαση, ικανοποιητικοί θεωρούνται δείκτες εσωτερικής συνέπειας της τάξης του 0,70 και πάνω (Nunnaly, 1978) ή ακόμα και πάνω από 0,60 (Malhotra 1996, Σιάρδος 1999). Στην περίπτωση της ΑΟ, ο δείκτης α_s έχει την παρακάτω ιδιότητα:

Ακόμη και στην περίπτωση που οι μεταβλητές έχουν σχετικά χαμηλή συσχέτιση μεταξύ τους, το κοινό τους χαρακτηριστικό, αν όντως αυτό υπάρχει και με δεδομένο ότι το πλήθος των μεταβλητών που το μοιράζονται είναι σχετικά μεγάλο, θα "κυριαρχήσει", δηλαδή θα επηρεάσει σημαντικά τη διαμόρφωση των βαθμών των αντικειμένων, οι οποίοι, όπως είδαμε, δεν είναι παρά οι μέσοι όροι των βέλτιστα ποσοτικοποιημένων ιδιοτήτων που τα χαρακτηρίζουν.

Από τη σχέση [2.79] είναι φανερό ότι για να είναι ο δείκτης θετικός α_s θα πρέπει η αδράνεια λ_s να είναι μεγαλύτερη από $1/q$. Αρνητικές τιμές του α_s θεωρούνται απαράδεκτες (Nishisato 1980, Spector 1992) και, συνεπώς, η ποσότητα $1/q$ μπορεί να χρησιμοποιηθεί ως εμπειρική οριακή τιμή για την επιλογή των σημαντικών αξόνων. Η τιμή $1/q$ εκφράζει τη μέση αναμενόμενη αδράνεια ανά άξονα, κάτω από την υπόθεση της τυχαίας μεταβλητότητας των δεδομένων. Παραγοντικοί άξονες με αδράνεια μικρότερη ή ίση από $1/q$ θεωρούνται, εν γένει, "θόρυβος". Για τον Greenacre (2003)[11] το υπό εξέταση φαινόμενο σταματά στους άξονες με αδράνεια μεγαλύτερη από $1/q$, κατά την ανάλυση του πίνακα **Z**, και στους άξονες με αδράνεια μεγαλύτερη από $1/q^2$ για την περίπτωση του πίνακα *Burt*.

Θ) Λόγω της ισοδυναμίας της ΑΟ με την ΠΑΑ, για την ερμηνεία των γραφικών αποτελεσμάτων μπορούμε να συνδυάσουμε τις ιδιότητες και των δύο μεθόδων. Αξίζει

[11] Μετά από προσωπική επικοινωνία στο International Conference on Correspondence Analysis and Related Methods 2003 (CARME 2003) in Barcelona, Spain, 29 June-2 July, 2003.

να παρατηρήσουμε ότι στο λογισμικό SyStat, στα παραγοντικά επίπεδα σχεδιάζονται και τα διανύσματα θέσης των προβαλλόμενων σημείων. Έτσι, είναι δυνατό να αναδειχθούν ομάδες ιδιοτήτων ή αντικειμένων που προβάλλονται στο ίδιο τεταρτημόριο του παραγοντικού επιπέδου και τα αντίστοιχα διανύσματα θέσης έχουν περίπου την ίδια διεύθυνση (SAS Institute, 1990). Με τον τρόπο αυτό, η κατά παράγοντες ερμηνεία των αποτελεσμάτων μπορεί να συνδυαστεί με την κατά ομάδες ερμηνεία, η οποία είναι ανάλογη με αυτή που χρησιμοποιείται στην Ανάλυση Συστάδων (Bacher, 1995). Γενικά, η ΠΑΑ, όπως αυτή εφαρμόζεται στη Γαλλική Σχολή, υπερτερεί έναντι της ΑΟ τόσο ως προς την παρεχόμενη βοήθεια στην ερμηνεία των αποτελεσμάτων όσο και ως προς την ευελιξία που έχει στην απευθείας ανάλυση πολλών τύπων πινάκων εισόδου.

2.4 Άλλοι Πίνακες Εισόδου στην Παραγοντική Ανάλυση των Αντιστοιχιών

Στην πράξη, εκτός από τους πίνακες **F, Z_{0-1}** και **B,** η μέθοδος της ΠΑΑ μπορεί να εφαρμοστεί απευθείας, χωρίς προσαρμογές, και σε άλλους πίνακες εισόδου που ικανοποιούν τις προϋποθέσεις ομοιογένειας που αναφέρθηκαν στην Εισαγωγή του τρέχοντος Κεφαλαίου. Στην ενότητα αυτή παραθέτουμε τους σημαντικότερους τύπους πινάκων συμπτώσεων, επί των οποίων μπορεί να εφαρμοστεί η μέθοδος, ενώ στην Ενότητα Β4 του Παραρτήματος Β παρουσιάζουμε ορισμένους ειδικούς πίνακες που μπορούν να δοθούν ως είσοδος στην ΠΑΑ.

2.4.1 Πίνακες Τύπου «Στοίβας», Πίνακες Τύπου «Φέτας» και Υποπίνακες του Πίνακα *Burt*

Οι πίνακες τύπου «στοίβας» αποτελούνται από πίνακες συμπτώσεων με κοινή μεταβλητή στήλης, οι οποίοι τοποθετούνται ο ένας πάνω στον άλλο, ενώ οι πίνακες «φέτα», κατασκευάζονται από πίνακες συμπτώσεων με κοινή μεταβλητή γραμμής που τοποθετούνται ο ένας δίπλα στον άλλο (βλέπε Σχήμα 2.4). Εφαρμογές της ΠΑΑ σε πίνακες τύπου «στοίβας» και «φέτας» συναντάμε στους Cazes (1980), Benzécri (1983), Moussa και Ouda (1988), Van der Heijden και De Leeuw (1989), Weller και Romney (1990), Higgs (1991), Παπαδημητρίου και Φλώρου (1992), Van de Geer

(1993β), Greenacre (1994β και 1993α), Aluja-Banet και Nonell-Torrent (1993), Blasius (1994), Martens (1994), De Lagarde (1995), Clausen (1998), Romney, Moore και Brazill (1998) και Lebart, Morineau και Piron (2000). Η μέθοδος μπορεί να εφαρμοστεί και σε υποπίνακες του πίνακα *Burt* (βλέπε Σχήμα 2.3) (Benzécri, 1992). Στην περίπτωση αυτή, η μέθοδος χαρακτηρίζεται ως «Σύνθετη» (*Composite*) ΠΑΑ (Israëls, 1987). Μάλιστα, αν έχουμε Q σε πλήθος κατηγορικές μεταβλητές X_i (i=1,…, Q), οι οποίες μπορούν να διαμεριστούν σε δύο υποσύνολα με Q_1 και Q_2 μεταβλητές αντίστοιχα, τέτοια ώστε οι μεταβλητές μέσα σε κάθε υποσύνολο να είναι ανά δύο ανεξάρτητες ή ασυσχέτιστες, τότε μπορεί να δειχθεί (Greenacre 1984, Lebart, Morineau & Warwick 1984) ότι η εφαρμογή της ΠΑΑ στις Q μεταβλητές είναι ισοδύναμη με την εφαρμογή της μεθόδου στον πίνακα συμπτώσεων **F** που δημιουργείται, όταν οι κλάσεις των Q_1 μεταβλητών τοποθετηθούν στις γραμμές και οι κλάσεις των Q_2 μεταβλητών τοποθετηθούν στις στήλες του πίνακα **F** (ή αντίστροφα). Στο Κεφάλαιο 4 (Ενότητα 4.8.1) προτείνουμε μέθοδο εντοπισμού υποπίνακα του *Burt*, που να περιλαμβάνει όλες τις μεταβλητές που συμμετέχουν στην ανάλυση και η εφαρμογή της ΠΑΑ σε αυτόν να αποδίδει την πλησιέστερη “εικόνα” του φαινομένου σε αυτή που προκύπτει από την εφαρμογή της μεθόδου στον αρχικό πίνακα *Burt*.

$$\begin{bmatrix} \mathbf{F}_{15} \\ \mathbf{F}_{25} \\ \mathbf{F}_{35} \\ \mathbf{F}_{45} \end{bmatrix}$$

Πίνακας Στοίβα

$$\begin{bmatrix} \mathbf{F}_{12} & \mathbf{F}_{13} & \mathbf{F}_{14} & \mathbf{F}_{15} \end{bmatrix}$$

Πίνακας Φέτα

$$\begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} & \mathbf{F}_{13} & \mathbf{F}_{14} & \mathbf{F}_{15} \\ \mathbf{F}_{21} & \mathbf{F}_{22} & \mathbf{F}_{23} & \mathbf{F}_{24} & \mathbf{F}_{25} \\ \mathbf{F}_{31} & \mathbf{F}_{32} & \mathbf{F}_{33} & \mathbf{F}_{34} & \mathbf{F}_{35} \\ \mathbf{F}_{41} & \mathbf{F}_{42} & \mathbf{F}_{43} & \mathbf{F}_{44} & \mathbf{F}_{45} \\ \mathbf{F}_{51} & \mathbf{F}_{52} & \mathbf{F}_{53} & \mathbf{F}_{54} & \mathbf{F}_{55} \end{bmatrix} \longrightarrow \begin{bmatrix} \mathbf{F}_{14} & \mathbf{F}_{15} \\ \mathbf{F}_{24} & \mathbf{F}_{25} \\ \mathbf{F}_{34} & \mathbf{F}_{35} \end{bmatrix}$$

Πίνακας *Burt* Υποπίνακας του *Burt*

όπου $\mathbf{F}_{ij}$ είναι ο πίνακας συμπτώσεων απολύτων συχνοτήτων της μεταβλητής i με τη μεταβλητή j (i, j=1,…, 5).

Σχήμα 2.4: Πίνακας «Στοίβα», Πίνακας «Φέτα» και Υποπίνακας του *Burt*.

Να τονίσουμε ότι κατά την εφαρμογή της ΠΑΑ στους τρεις πίνακες, που προαναφέρθηκαν, αναλύονται μόνο οι συσχετίσεις των μεταβλητών για τις οποίες ο αντίστοιχος απλός πίνακας συμπτώσεων συμμετέχει στην κατασκευή τους. Έτσι, στην ανάλυση του πίνακα «φέτα», του Σχήματος 2.4, θα ληφθούν υπόψη μόνο οι συσχετίσεις μεταξύ των μεταβλητών (X_1, X_2), (X_1, X_3), (X_1, X_4) και (X_1, X_5). Ανάλογα ισχύουν για τους άλλους δύο πίνακες.

Στην περίπτωση που δύο ή περισσότερες μεταβλητές δεν παρουσιάζουν σημαντική συσχέτιση μεταξύ τους ή η συσχέτισή τους δεν παρουσιάζει ενδιαφέρον, τότε είναι χρήσιμο από τις μεταβλητές αυτές να δημιουργήσουμε μια νέα μεταβλητή «αλληλεπίδρασης», οι τιμές της οποίας θα είναι οι συνδυασμοί των τιμών των αρχικών μεταβλητών (Israëls 1987, Van der Heijden & De Leeuw 1989, Andersen 1991, Van de Geer 1993β, Greenacre 1993α, Michailidis 1996, Clausen 1998). Πιο συγκεκριμένα, αν έχουμε δύο μεταβλητές *X* και *Y* με 2 και 3 κλάσεις αντίστοιχα, τότε μπορούμε να δημιουργήσουμε μια νέα μεταβλητή *Z* που θα παίρνει 2×3=6 διακεκριμένες τιμές, όσοι είναι δηλαδή και οι συνδυασμοί των κλάσεων των δύο αρχικών μεταβλητών *X* και *Y*. Για τη διευκόλυνση των υπολογισμών είναι χρήσιμο να εισάγουμε έναν τελεστή αλληλεπίδρασης, τον οποίο συμβολίζουμε με *Int*(*), ο οποίος δίνεται από την παρακάτω σχέση:

$$Int(X_1, X_2, \ldots, X_q) = 10^{q-1} v_{X_1} + 10^{q-2} v_{X_2} + \ldots + 10^0 v_{X_q},$$

όπου v_{X_1} είναι κωδικοποιημένη τιμή[12] της μεταβλητής X_i με i=1,…, q.

Παράδειγμα Δημιουργίας Μεταβλητής Αλληλεπίδρασης

Έστω ότι η μεταβλητή *Φύλο* παίρνει δύο τιμές {*Άνδρας*=1, *Γυναίκα*=2} και η μεταβλητή *Ηλικία* τρεις τιμές {20-30 *ετών*=1, 31-50 *ετών*=2, 51 *και άνω*=3. Τότε ο τελεστής *Int* (*Φύλο*, *Ηλικία*), για q=2, δίνει στη νέα μεταβλητή *Z* το σύνολο τιμών {11, 12, 13, 21, 22, 23} όπου, για παράδειγμα, η τιμή "12" δεν εκφράζει μέτρηση, δηλαδή τον αριθμό 12, αλλά το συνδυασμό ιδιοτήτων "*Άνδρας, Ηλικίας 31-50 ετών*".

[12] Θεωρούμε ότι οι κλάσεις της μεταβλητής έχουν κωδικοποιηθεί με τους διαδοχικούς θετικούς αριθμούς 1, 2, …, q.

Οι νέες μεταβλητές αλληλεπίδρασης μπορούν να χρησιμοποιηθούν για την κατασκευή πινάκων των τριών τύπων που παρουσιάστηκαν ανωτέρω.

2.4.2 Γενικευμένοι Λογικοί Πίνακες

Σύμφωνα με τους Le Roux και Rouanet (2004), η βαρυκεντρική και ασαφής κωδικοποίηση των μεταβλητών επεκτείνει τους λογικούς πίνακες $\mathbf{Z}_{0\text{-}1}$ σε «γενικευμένους» πίνακες σχεδιασμού της μορφής «*αντικείμενα* × *μεταβλητές*». Στους πίνακες αυτούς, για τη κωδικοποίηση των μεταβλητών, δεν χρησιμοποιούνται πλέον οι αριθμοί 0 και 1, αλλά ποσοστά, τα οποία αντιπροσωπεύουν ένα είδος σχετικής κατανομής των ιδιοτήτων (κατηγοριών) των μεταβλητών για κάθε αντικείμενο (SAS Institute, 1990). Τα ποσοστά αυτά, για κάθε αντικείμενο και για κάθε μεταβλητή έχουν άθορισμα ίσο με τη μονάδα, με αποτέλεσμα το βάρος κάθε αντικειμένου να εξακολουθεί να είναι ίσο με το πλήθος των μεταβλητών *q*, όπως, δηλαδή, και στους λογικούς πίνακες $\mathbf{Z}_{0\text{-}1}$. Έτσι, στην τιμή κάθε ιδιότητας μιας μεταβλητής ανατίθεται θετική μάζα ή, κατά μία έννοια, ένα μέτρο πιθανότητας[13], η οποία κατανέμεται σε δύο ή περισσότερους πόλους (Greenacre, 1984). Αν υποθέσουμε ότι οι τιμές 0 και 1, που χρησιμοποιούνται για την κατασκευή των συνήθων λογικών πινάκων, εκφράζουν την πιθανότητα ένα αντικείμενο να έχει μια συγκεκριμένη ιδιότητα ή να ανήκει σε μια συγκεκριμένη κλάση, τότε μπορούμε να θεωρήσουμε ότι η λογική κωδικοποίηση 0-1 αποτελεί μια “τετριμμένη” βαρυκεντρική διάσπαση. Κάτω από αυτή τη θεώρηση, οι λογικοί πίνακες αποτελούν ειδικές περιπτώσεις των γενικευμένων. Στη συνέχεια, δίνουμε μερικά παραδείγματα βαρυκεντρικής και ασαφούς κωδικοποίησης.

Παραδείγματα

Ο Πίνακας 2.10 παρουσιάζει στοιχεία για το φύλο και την ηλικία πέντε ατόμων με βαρυκεντρική κωδικοποίηση για τις ελλείπουσες τιμές, κάτω από την υπόθεση ότι η κατανομή των ανδρών και των γυναικών καθώς και των τριών ηλικιακών ομάδων είναι ομοιόμορφη για το συγκεκριμένο δείγμα ή πληθυσμό.

[13] Σχετικά με τη σχέση μάζας και πιθανότητας παραπέμπουμε στον Menexes (1998).

Αν με βάση προηγούμενη γνώση ή εμπειρία, για το συγκεκριμένο πληθυσμό, η κατανομή των ανδρών και των γυναικών είναι 35% και 65% αντίστοιχα, τότε η βαρυκεντρική κωδικοποίηση των ελλειπουσών τιμών θα μπορούσε να είναι 0,35 για τη στήλη των ανδρών και 0,65 για τη στήλη των γυναικών. Στην περίπτωση αυτή, οι ελλείπουσες τιμές αντικαθίστανται με ένα ζεύγος συμπληρωματικών πιθανοτήτων, όπου το πρώτο στοιχείο του ζεύγους εκφράζει την πιθανότητα κάποιο άτομο, επιλεγμένο τυχαία, να είναι άνδρας, και το δεύτερο την πιθανότητα να είναι γυναίκα.

Πίνακας 2.10: Γενικευμένος Λογικός Πίνακας-Βαρυκεντρική Κωδικοποίηση Ελλειπουσών Τιμών

	Δεδομένα		Γενικευμένη Λογική Κωδικοποίηση					
Άτομα	Φύλο	Ηλικία	Άνδρας	Γυναίκα	20-30 ετών	31-40 ετών	41-60 ετών	*Σύνολο*
Α	Άνδρας	20-30 ετών	1	0	1	0	0	2
Β	Γυναίκα	31-40 ετών	0	1	0	1	0	2
Γ	Γυναίκα	41-60 ετών	0	1	0	0	1	2
Δ	Άνδρας	*	1	0	0,333	0,333	0,333	2
Ε	*	31-40 ετών	0,5	0,5	0	1	0	2

* Ελλείπουσα Τιμή

Αν με βάση προηγούμενη γνώση ή εμπειρία, για το συγκεκριμένο πληθυσμό, η κατανομή των ανδρών και των γυναικών είναι 35% και 65% αντίστοιχα, τότε η βαρυκεντρική κωδικοποίηση των ελλειπουσών τιμών θα μπορούσε να είναι 0,35 για τη στήλη των ανδρών και 0,65 για τη στήλη των γυναικών. Στην περίπτωση αυτή, οι ελλείπουσες τιμές αντικαθίστανται με ένα ζεύγος συμπληρωματικών πιθανοτήτων, όπου το πρώτο στοιχείο του ζεύγους εκφράζει την πιθανότητα κάποιο άτομο, επιλεγμένο τυχαία, να είναι άνδρας, και το δεύτερο την πιθανότητα να είναι γυναίκα. Αυτή η προσέγγιση, αποτελεί για τον Verkuilen (2001) την «πιθανολογική κωδικοποίηση» (*probabilistic coding*) που εκφράζει το βαθμό στον οποίο ένα αντικείμενο είναι δυνατό να ανήκει σε ένα ορισμένο σύνολο ιδιοτήτων. Άλλωστε, όπως είδαμε και στη διμεταβλητή εκδοχή της ΠΑΑ (βλέπε Ενότητα 2.2.3), τα προφίλ των γραμμών (στηλών) εκφράζουν κατανομές δεσμευμένων πιθανοτήτων.

Ο Πίνακας 2.11 παρουσιάζει τη βαρυκεντρική διάσπαση σε ποσοστά που πρότεινε ο Μοσχίδης (2003β, σ. 342), για τις απαντήσεις πέντε ατόμων σε μια ερώτηση με πέντε διαβαθμισμένες κατηγορίες. Στο παράδειγμα αυτό, οι πέντε διαβαθμίσεις θα μπορούσαν να αντιστοιχούν σε κλάσεις μιας ποσοτικής μεταβλητής που

κατηγοριοποιήθηκε με μια από τις μεθόδους που αναφέρθηκαν στην προηγούμενη ενότητα.

Πίνακας 2.11: Γενικευμένος Λογικός Πίνακας-Βαρυκεντρική Κωδικοποίηση

		Λογική Κωδικοποίηση					Γενικευμένη Λογική Κωδικοποίηση					
Άτομα	Ερώτηση*	ΣΑ	Σ	Ο	Δ	ΔΑ	ΣΑ	Σ	Ο	Δ	ΔΑ	*Σύνολο*
Α	*ΣΑ*	1	0	0	0	0	0,83	0,11	0,04	0,014	0,006	1
Β	*Σ*	0	1	0	0	0	0,17	0,66	0,11	0,04	0,02	1
Γ	*Ο*	0	0	1	0	0	0,06	0,11	0,66	0,11	0,06	1
Δ	*Δ*	0	0	0	1	0	0,02	0,04	0,11	0,66	0,17	1
Ε	*ΑΔ*	0	0	0	0	1	0,006	0,014	0,04	0,11	0,83	1

* ΣΑ: Συμφωνώ Απόλυτα, Σ: Συμφωνώ, Ο: Ουδέτερη Άποψη, Δ: Διαφωνώ και ΔΑ: Διαφωνώ Απόλυτα

Ο Πίνακας 2.12 παρουσιάζει τα αρχικά δεδομένα για εννέα τύπους αυτοκινήτων. Για κάθε τύπο έχει μετρηθεί η τιμή του σε $ και ο κυβισμός της μηχανής του (Van Rijckevorsel, 1987, p. 148). Το εύρος των τιμών των δύο ποσοτικών μεταβλητών (τιμή και κυβισμός) έχει μοιραστεί σε μια σειρά διαδοχικών διαστημάτων με αυθαίρετα όρια, ως εξής:

Τιμή: 4000-5000, 5000-6000, 6000-7000, 7000-8000 και 8000-9000
Κυβισμός: 50-100, 100-150, 150-200, 200-250, 250-300 και 300-350

Επομένως, για τη λογική κωδικοποίηση 0-1 της "Τιμής" θα πρέπει να χρησιμοποιηθούν πέντε διαστήματα (κλάσεις) του τύπου «*από-έως*» και έξι διαστήματα για τον "Κυβισμό".

Πίνακας 2.12: Τιμή και Κυβισμός για Εννέα Τύπους Αυτοκινήτων

Τύπος Αυτοκινήτου	Τιμή	Κυβισμός
Α	5700	134
Β	4350	92
Γ	6100	97
Δ	6850	152
Ε	8200	258
Ζ	7200	173
Η	7200	232
Θ	6700	225
Ι	7550	318

Στον Πίνακα 2.13 έχουμε τη συνηθισμένη κωδικοποίηση 0-1, ενώ στον Πίνακα 2.14 μία πρόταση ασαφούς κωδικοποίησης των δύο ποσοτικών μεταβλητών (Van Rijckevorsel, 1987, σ. 149).

Πίνακας 2.13: Λογική Κωδικοποίηση των Κατηγοριοποιημένων Ποσοτικών Μεταβλητών

Τύπος	Λογική Κωδικοποίηση											*Σύνολο*
	Διαστήματα της "Τιμής"					Διαστήματα του "Κυβισμού"						
Α	0	1	0	0	0	0	1	0	0	0	0	2
Β	1	0	0	0	0	1	0	0	0	0	0	2
Γ	0	0	1	0	0	1	0	0	0	0	0	2
Δ	0	0	1	0	0	0	0	1	0	0	0	2
Ε	0	0	0	0	1	0	0	0	0	1	0	2
Ζ	0	0	0	1	0	0	0	1	0	0	0	2
Η	0	0	0	1	0	0	0	0	1	0	0	2
Θ	0	0	1	0	0	0	0	0	1	0	0	2
Ι	0	0	0	1	0	0	0	0	0	0	1	2

Πίνακας 2.14: Ασαφής Κωδικοποίηση των Κατηγοριοποιημένων Ποσοτικών Μεταβλητών

	Γενικευμένος Λογικός Πίνακας													
	Τιμή						Κυβισμός							
Τύπος	4000	5000	6000	7000	8000	9000	50	100	150	200	250	300	350	*Σύνολο*
Α	0	0,3	0,7	0	0	0	0	0,32	0,68	0	0	0	0	2
Β	0,65	0,35	0	0	0	0	0,16	0,84	0	0	0	0	0	2
Γ	0	0	0,9	0,1	0	0	0,06	0,94	0	0	0	0	0	2
Δ	0	0	0,15	0,85	0	0	0	0	0,95	0,04	0	0	0	2
Ε	0	0	0	0	0,8	0,2	0	0	0	0	0,84	0,16	0	2
Ζ	0	0	0	0,8	0,2	0	0	0	0,54	0,46	0	0	0	2
Η	0	0	0	0,8	0,2	0	0	0	0	0,36	0,64	0	0	2
Θ	0	0	0,3	0,7	0	0	0	0	0	0,5	0,5	0	0	2
Ι	0	0	0	0,45	0,55	0	0	0	0	0	0	0,64	0,36	2

Παρατηρούμε ότι στην ασαφή κωδικοποίηση οι στήλες του γενικευμένου λογικού πίνακα δεν αντιστοιχούν σε διαστήματα αλλά στα όρια των διαστημάτων. Κάτω από μια γενική θεώρηση, μπορούμε να ισχυριστούμε ότι για κάθε τύπο αυτοκινήτου η τιμή κάθε στήλης εκφράζει ένα μέτρο πιθανότητας για την αντίστοιχη οριακή τιμή της μεταβλητής. Αν και η ΠΑΑ θα μπορούσε να εφαρμοστεί σε πίνακες όπως ο 2.14 χωρίς υπολογιστικό πρόβλημα, ωστόσο, στο συγκεκριμένο παράδειγμα, προκύπτουν δυσκολίες σε σχέση με την ερμηνεία των αποτελεσμάτων, αφού οι στήλες του ασαφούς πίνακα δεν αντιστοιχούν κατ' ανάγκη σε παρατηρούμενες ιδιότητες των αυτοκινήτων. Συνεπώς, είναι ανάγκη η κατηγοριοποίηση μιας ποσοτικής μεταβλητής

σε κλάσεις να πραγματοποιείται λαμβάνοντας υπόψη τόσο στατιστικά όσο και θεωρητικά κριτήρια, ώστε τα διαστήματα που θα προκύψουν να έχουν φυσική ερμηνεία και κλινική ή πρακτική σημαντικότητα, στο πλαίσιο του επιστημονικού χώρου που διεξάγεται ή έρευνα. Πάντως, από τη στιγμή που επιλεγεί (ή/και συμφωνηθεί) η κατάλληλη κατηγοριοποίηση, οι κλάσεις των μεταβλητών αναδεικνύουν πλέον συγκεκριμένες ιδιότητες των αντικειμένων και ως προς αυτές θα πρέπει να ερμηνευτούν τα αποτελέσματα. Διαφορετική κατηγοριοποίηση αποδίδει, εν γένει, και διαφορετικές ιδιότητες.

2.4.3 Επεκτάσεις και Πεδία Εφαρμογής της Παραγοντικής Ανάλυσης των Αντιστοιχιών

Αξιοσημείωτο είναι ότι σε όλες τις περιπτώσεις πινάκων, που παρουσιάστηκαν στην προηγούμενη ενότητα (βλέπε και Ενότητα Β4 του Παραρτήματος Β), η ΠΑΑ εφαρμόζεται αλγοριθμικά και εννοιολογικά με τον ίδιο ακριβώς τρόπο. Οι πίνακες αντιμετωπίζονται, τελικά, ως απλοί πίνακες συμπτώσεων δύο μεταβλητών, με τη γενική μορφή «*αντικείμενα* × *ιδιότητες*». Κάτω από αυτή την ενιαία θεώρηση των πινάκων εισόδου σε συνδυασμό με τη διεισδυτική προσέγγιση στην ανάλυση των δεδομένων, η ΠΑΑ έχει αναδειχθεί σε ένα γενικό σύστημα ή, καλύτερα, "πλατφόρμα" στατιστικής επεξεργασίας που βρίσκει εφαρμογές σε όλα σχεδόν τα ερευνητικά πεδία (βλέπε Greenacre 1984, Benzécri 1992, Blasius & Greenacre 1994, Gifi 1996, Beh 2004). Η μέθοδος εφαρμόζεται είτε όπως έχει είτε με παραλλαγές και τροποποιήσεις, οι οποίες επιβάλλονται από περιορισμούς ή ιδιαιτερότητες του επιστημονικού πεδίου, στο πλαίσιο του οποίου θα ερμηνευτούν τα αποτελέσματα της έρευνας. Αξίζει να αναφερθούμε στο χώρο της Οικολογίας των Φυτών, όπου η ΠΑΑ χρησιμοποιείται κυρίως ως μέθοδος ταξιθέτησης (*ordination*) σταθμών (περιοχών) με βάση την εμφάνιση (ποικιλότητα, πυκνότητα και υπεροχή) των παρατηρούμενων φυτικών ειδών (Hill 1974, Kent & Coker 1996, Ter Braak 2002 και 1985). Στο συγκεκριμένο επιστημονικό χώρο έχουν αναπτυχθεί δύο σημαντικές παραλλαγές της μεθόδου. Η πρώτη, που ονομάζεται «*Detrended Correspondence Analysis*» (Hill & Gauch, 1980), αφορά στη διόρθωση των συντεταγμένων των προβαλλόμενων σημείων επί των παραγοντικών επιπέδων, ώστε να απαλειφθεί η επίδραση του φαινομένου «*Guttman*» στην ερμηνεία των αποτελεσμάτων. Ως φαινόμενο *Guttman* χαρακτηρίζεται η παραβολοειδής διάταξη των προβαλλόμενων σημείων στο

παραγοντικό επίπεδο 1×2 (Greenacre 1984, Van Rijckevorsel 1987, Weller & Romney 1990, Manly 1994). Εμφανίζεται συχνά στις μεθόδους ταξιθέτησης και θεωρείται, εν γένει, μειονέκτημα της ΠΑΑ. Νεότερες προσεγγίσεις στην αντιμετώπιση του προβλήματος έχουν δείξει ότι από το φαινόμενο *Guttman* είναι δυνατό να εξαχθεί χρήσιμη πληροφορία (βλέπε Camiz, 2005) και ότι δεν θα πρέπει υποχρεωτικά να θεωρείται μειονέκτημα της μεθόδου ή ένδειξη ακαταλληλότητας των δεδομένων που αναλύονται (Gifi 1996, Verkuilen 2001). Η φυσική του ερμηνεία είναι ότι ο πρώτος παραγοντικός άξονας συνδέεται με σχέση δευτέρου βαθμού με τον δεύτερο, με σχέση τρίτου βαθμού με τον τρίτο κ.ο.κ. (Baccini, Caussinus & De Falguerolles 1993, Verkuilen 2001). Έτσι, στο παραγοντικό επίπεδο 1×2 οι ποσοτικοποιήσεις των σημείων επί του δεύτερου άξονα αποτελούν συναρτήσεις δευτέρου βαθμού των ποσοτικοποιήσεων του πρώτου. Το πρακτικό συμπέρασμα είναι ότι ο πίνακας δεδομένων που αναλύεται έχει στην ουσία μονοδιάστατη δομή και, επομένως, ο πρώτος παραγοντικός άξονας εκφράζει μια κυρίαρχη και σχετικά ομοιογενή σύνθετη μεταβλητή, ως προς την οποία θα πρέπει να ερμηνευτούν τα αποτελέσματα (Gifi 1996, De Leeuw, Wang & Michailidis 1999). Από μια άλλη σκοπιά, η εμφάνιση του φαινομένου *Guttman* θα μπορούσε να θεωρηθεί ως δείκτης ποιότητας της λύσης της ΠΑΑ, ιδιαίτερα στην περίπτωση που το ζητούμενο είναι ο έλεγχος ή η επιβεβαίωση του μονοδιάστατου χαρακτήρα του υπό εξέταση φαινομένου (Guttman 1941, Gifi 1996, Verkuilen 2001). Άλλες μέθοδοι αποκατάστασης του φαινομένου *Guttman* έχουν προταθεί από τους Van Rijckevorsel (1987) και Camiz (2005). Η δεύτερη παραλλαγή της μεθόδου ονομάζεται «Κανονικοποιημένη ή Κανονική Ανάλυση των Αντιστοιχιών» (*Canonical Correspondence Analysis*) (Ter Braak, 1986) και αναπτύχθηκε ώστε κατά την εφαρμογή της ΠΑΑ να λαμβάνεται υπόψη και επιπλέον διαθέσιμη εξωτερική πληροφορία σχετικά με κλιματολογικές, εδαφολογικές και άλλες περιβαλλοντικές παραμέτρους, οι οποίες μπορεί να είναι ποσοτικές ή ποιοτικές (Kent & Coker 1996, Ter Braak 2002). Χαρακτηριστικό της μεθόδου είναι ότι κατά την υλοποίηση της ΠΑΑ εφαρμόζεται ο αλγόριθμος *Reciprocal Averaging*, στον οποίο εισάγονται περιορισμοί, ώστε οι παραγοντικοί άξονες να αποτελούν γραμμικούς συνδυασμούς των εξωτερικών περιβαλλοντικών μεταβλητών.

Τόσο οι πίνακες εισόδου όσο και τα συστήματα κωδικοποίησης στα οποία αναφερθήκαμε δεν είναι τα μοναδικά αλλά τα πιο συχνά χρησιμοποιούμενα. Μεγάλη ποικιλία τεχνικών πινακοποίησης και κωδικοποίησης των δεδομένων συναντάμε στα ερευνητικά πεδία του Μάρκετινγκ και της Οικονομίας. Ενδεικτικά αναφέρουμε ότι ειδική κωδικοποίηση απαιτείται όταν: α) οι προς ανάλυση μεταβλητές αντιστοιχούν σε ερωτήσεις πολλαπλής απάντησης, κατάταξης ή προτίμησης (Kaciak & Louviere 1990, Nishisato 1996, 1994, 1993, 1980 και 1978, Van der Heijden, Teunissen & Van Orlé 1997, Van de Velden 2000, Torres & Greenacre 2002, Torres & Van de Velden 2007), β) η μέθοδος πρόκειται να εφαρμοστεί σε πίνακες με δεδομένα κυριαρχίας-υπεροχής (Dzhafarov 1999, Greenacre & Torres 1999) ή ζευγαρωτών συγκρίσεων (Greenacre, 2003) και γ) πρόκειται να αναλυθούν ειδικές κατηγορίες πινάκων (συμμετρικών και μη) που περιγράφουν μεταβολές μιας κατάστασης ή φαινομένου, για παράδειγμα, μετακινήσεις (εργασιακές, μεταναστευτικές) ατόμων και πληθυσμών και οικονομικά αποτελέσματα εισροών-εκροών (Van der Heijden, De Vries & Van Hooff 1990, Blasius 1994, Gower & Greenacre 1996, Greenacre & Clavel 1998, Greenacre 2000, Τζήμος & Παπαδημητρίου 2004). Θα πρέπει να παρατηρήσουμε ότι στις περισσότερες από τις παραπάνω περιπτώσεις η ΠΑΑ δεν εφαρμόζεται άμεσα, αλλά μετά από τροποποιήσεις στον αλγόριθμο και στην ερμηνεία των αποτελεσμάτων της. Προσαρμογές, επίσης, απαιτούνται για την ανάλυση κοινωνιομετρικών πινάκων διπλής εισόδου (βλέπε Noma & Smith 1985, Greenacre 2000, Roberts 2000 και 1996) και για τη σύγκριση δύο ή περισσότερων πινάκων συμπτώσεων με ζευγαρωτές μετρήσεις (βλέπε Greenacre, 2003).

Στην περίπτωση που για τη συλλογή των δεδομένων έχει εφαρμοστεί κάποιο ιεραρχικό δειγματοληπτικό σχέδιο πολλών επιπέδων, έχουν προταθεί παραλλαγές της ΠΑΑ, οι οποίες λαμβάνουν υπόψη την ενδεχόμενη επίδραση του δειγματοληπτικού σχήματος στα αποτελέσματα. Ειδικότερα, η Escofier (1987) πρότεινε την «Υπό Συνθήκη ή Δεσμευμένη Παραγοντική Ανάλυση των Αντιστοιχιών» (*Analyse des Correspondances Multiples Conditionelles*), κατά την οποία οι μεταβλητές που αναλύονται ελέγχονται ως προς τα επίπεδα μιας τρίτης μεταβλητής, οι τιμές της οποίας μπορεί να δηλώνουν στρώματα ή συστάδες του υπό εξέταση πληθυσμού. Κατά την προσέγγιση αυτή, πραγματοποιούνται δύο αναλύσεις, η «Ενδοταξική» (*Intra-Classes*) ΠΑΑ και η «Διαταξική» (*Inter-Classes*) ΠΑΑ. Η μέθοδος εφαρμόζεται στην περίπτωση που είναι επιθυμητό να απαλειφθεί η επίδραση της

διαταξικής αδράνειας μεταξύ των ομάδων (στρώματα, συστάδες) του πληθυσμού ΠΑΑ από τις σχέσεις των υπολοίπων μεταβλητών. Αποτελεί μια ειδική περίπτωση της «Εσωτερικής Παραγοντικής Ανάλυσης των Αντιστοιχιών» (*Internal Correspondence Analysis*) (βλέπε Lobry & Chessel 2003, Bècue-Bertaut & Pagès 2004), που αναπτύχθηκε από τους Benzécri (1983) και Escofier και Drouet (1983). Ο Michalidis (1996) παρουσίασε μεθοδολογία σύμφωνα με την οποία η Ανάλυση Ομοιογένειας και άλλες μέθοδοι της Ολλανδικής Σχολής μπορούν να χρησιμοποιηθούν σε περιπτώσεις όπου τα δεδομένα έχουν συγκεντρωθεί με δειγματοληψία κατά συστάδες πολλών σταδίων. Η προσέγγιση του Michailidis είναι περισσότερο γνωστή ως «Ανάλυση Ομοιογένειας Πολλαπλών Επιπέδων» (*Multi-Level Homogeneity Analysis*). Χρησιμοποιείται για την ανάλυση και σύγκριση δεδομένων, τα οποία έχουν συγκεντρωθεί ιεραρχικά (κατά στάδια). Σκοπός της μεθόδου είναι να ληφθούν υπόψη τόσο τα ιδιαίτερα χαρακτηριστικά των ομάδων αντικειμένων, που συγκροτούνται σε κάθε στάδιο-επίπεδο, όσο και η επίδραση τους στη διαμόρφωση των σχέσεων των μεταβλητών που εξετάζονται (βλέπε Michailidis & De Leeuw, 2000 και 1997). Στον Israëls (1987) συναντάμε τη «Μερική Παραγοντική Ανάλυση των Αντιστοιχιών» (*Partial Correspondence Analysis*), όπου η συνάφεια δύο μεταβλητών ελέγχεται, αφού αφαιρεθεί ή επίδραση μιας τρίτης. Οι Bècue-Bertaut και Pagès (2004) πρότειναν την «Πολλαπλή Παραγοντική Ανάλυση για Πίνακες Συνάφειας» (*Multiple Factor Analysis for Contingency Tables-MFACT*) με σκοπό τη σύγκριση δύο ή περισσοτέρων πινάκων συμπτώσεων, που προέρχονται από διαφορετικά δείγματα ή πληθυσμούς, με κοινές γραμμές (στήλες) αλλά διαφορετικές στήλες (γραμμές). Γενικά, τροποποιήσεις της ΠΑΑ για την ανάλυση δεδομένων που είναι οργανωμένα σε πίνακες τριπλής εισόδου[14] (ή μπορούν να θεωρηθούν ως τέτοιοι) έχουν προταθεί από πολλούς ερευνητές. Ενδεικτικά αναφέρουμε τις εργασίες των Cazes (1980), Benzécri (1983, 1982 και 1979), Escofier (1983), Escofier και Drouet (1983), Deville και Saporta (1983), De Leeuw και Van der Heijden (1988), Van der Heijden και De Leeuw (1989 και 1985), Novak και Hoffman (1990), Παπαδημητρίου (1991), Burtschy και Papadimitriou (1991), Παπαδημητρίου και Φλώρου (1992), Aluja-Banet και Nonell-Torrent (1993), Martens (1994), Carlier και Kroonenberg (1998 και 1996), Escofier και Pagès (1998),

[14] Συχνά η τρίτη διάσταση καθορίζεται από μεταβλητή οι τιμές της οποίας εκφράζουν διαφορετικές χρονικές περιόδους.

Μπεχράκη (1999), Greenacre (2003), Lobry και Chessel (2003) και Bècue-Bertaut και Pagès (2004). Σε κάθε περίπτωση, στόχος των παραπάνω γενικεύσεων είναι η ανάδειξη και η σύγκριση των εσωτερικών δομών των επιμέρους πινάκων διπλής εισόδου που συνθέτουν τον αρχικό τριδιάστατο πίνακα. Να επισημάνουμε ότι σε ορισμένες περιπτώσεις, όπως για παράδειγμα της Escofier (1983), η απόσταση που χρησιμοποιείται δεν μπορεί να ερμηνευτεί ως χ^2.

Θα πρέπει να παρατηρήσουμε ότι εφαρμογές και παραλλαγές της ΠΑΑ για την ανάλυση πινάκων με ιδιαίτερη δομή ως προς την κατασκευή ή/και το πληροφοριακό τους περιεχόμενο συναντάμε, πλέον, και σε ερευνητικούς τομείς, όπου ο ρόλος των Πολυδιάστατων Στατιστικών Μεθόδων ήταν τουλάχιστον κατά το παρελθόν λιγότερο εμφανής. Στις μέρες μας, η μέθοδος χρησιμοποιείται και σε νέα ερευνητικά πεδία στα οποία έδωσε ώθηση κυρίως η ανάπτυξη της Τηλεπληροφορικής. Η ΠΑΑ έχει "εισαχθεί" και εφαρμόζεται στο χώρο της Μηχανικής Μάθησης (Merz, 1997), της Αναγνώρισης Προτύπων (Queiros, Gelsema & Timmers, 1983), της Βιο-Ιατρικής (Perrière & Thioulouse, 2003), της Ιατρικής (Greenacre 1992, Kaminska *et al*. 1999), της Επιδημιολογίας (Guinot *et al*. 2001, Grassi *et al*. 2003), της Κοινωνικής Ιατρικής (Kakai *et al*., 2003), της Βιο-Πληροφορικής (Busold *et al*., 2005), της Μηχανικής (Bouilland & Loslever, 1998), των Πολυκριτηρίων Αποφάσεων (Cheung, 1994 και 1991, Λούκας & Παπαδημητρίου 2005), της Γεωλογίας (Teil, 1975), της Ωκεανογραφίας (Malmgren *et al*., 1978), της Γεωχημείας (Hongjin, Yongzheng & Xisheng, 1995), της Βιο-Χημείας (Hans, Ojasoo & Doré, 2000), της Μικροβιολογίας (Sieber, Petrini & Greenacre, 1998), της Γενετικής (Lobry & Chessel, 2003), της Χημείας (Mellinger, 1987), της Τεχνολογίας Τροφίμων (McEwan & Schlich, 1992), της Αρχαιολογίας (Bolviken *et al*. 1982, Ringrose 1992), της Βιβλιοθηκονομίας (Μάλλιαρη, 2004), της Μουσικής (Purwins *et al*., 2003), της Διαχείρισης Ποιότητας (Ngai & Cheng, 1997), της Διαχείρισης Πόρων (Kishino *et al*., 1998), της Χρηματοοικονομικής Διαχείρισης (Καραπιστόλης, 1999 και 1996), της Οικονομετρίας (Deville & Saporta, 1983), της Εργονομίας (Miyake, Loslever & Hancock, 2001), της Κοινωνιολογίας (De Nooy, 2003), της Εκπαίδευσης (Αναστασιάδου 2000, Αναστασιάδου & Καράκος 2005, Αθανασιάδης 2005) και του Αθλητισμού (Μαυρομάτης, 1999).

Ένα γενικό συμπέρασμα που προκύπτει από τη ανασκόπηση της παραπάνω ενδεικτικής βιβλιογραφίας είναι ότι η ΠΑΑ είναι αρκετά ευέλικτη και μπορεί να προσαρμοστεί στις ιδιαιτερότητες του εκάστοτε ερευνητικού προβλήματος και στους περιορισμούς που θέτει το θεωρητικό πλαίσιο, εντός του οποίου θα ερμηνευτούν τα αποτελέσματα. Στην πράξη, η μόνη αυστηρή απαίτηση για την εφαρμογή της μεθόδου είναι η ύπαρξη ενός πίνακα διπλής εισόδου με μη αρνητικά στοιχεία. Βέβαια, ιδιαίτερη προσοχή απαιτείται στην κατάστρωση του πίνακα, που θα δοθεί ως είσοδος στην ανάλυση, ώστε να μπορεί να αναδειχθεί η φυσική ερμηνεία των δομών (σχέσεων, ομοιοτήτων, τάσεων και αντιθέσεων) που παρουσιάζουν ενδιαφέρον σε κάθε ερευνητική περίπτωση. Στο Κεφάλαιο 7 προτείνουμε μεθοδολογία για την εφαρμογή της ΠΑΑ σε πίνακες σχεδιασμού που προκύπτουν από πειραματικές διατάξεις.

2.5 Ιδιότητες Βέλτιστης Κλιμάκωσης της Παραγοντικής Ανάλυσης των Αντιστοιχιών

Καταρχήν θα πρέπει να επισημάνουμε ότι η ΠΑΑ ως μέθοδος βέλτιστης κλιμάκωσης, που ποσοτικοποιεί ποιοτικά δεδομένα (Young 1981, Greenacre 1993α, De Leeuw 1993, Meulman 1999), είναι στενά συνδεδεμένη και με άλλες στατιστικές μεθόδους, όπως είναι η Κανονικοποιημένη Συσχέτιση (Greenacre 1984, Lebart, Morineau & Warwick 1984, Tenenhaus & Young 1985, Gower 1990, Andersen 1991, Van de Geer 1993β, De Leeuw, Wang & Michailidis 1999), η Διακρίνουσα Ανάλυση (Nishisato 1980, Greenacre 1984, Lebart, Morineau & Warwick 1984, Van der Heijden, Mooijaart & Takane 1994), η Ανάλυση σε Κύριες Συνιστώσες (Tenenhaus & Young 1985, Greenacre 1991 και 1984, Van de Geer 1993α και 1993β, Heiser & Meulman 1994, Gifi 1996, Van de Velden 2000), η Ανάλυση Διακύμανσης (Tenenhaus & Young 1985, Nishisato 1994 και 1980), η Γραμμική Παλινδρόμηση (Nishisato 1980, Greenacre 1984, Van Rijckevorsel 1987, Buja 1990, De Leeuw 1993 και 1988, Heiser & Meulman 1994, Michailidis 1996, Michailidis & De Leeuw 1998, De Leeuw, Wang & Michailidis 1999), η Ανάλυση Λανθανουσών Κλάσεων (*Latent Class Analysis*) (Israëls 1987, Andersen 1991, Gifi 1996, Van der Heijden, Gilula & Van der Ark 1999) και η Πολυδιάστατη Κλιμάκωση (Heiser 1987, Hoffman & De Leeuw 1992, Gifi 1996, Romney, Moore & Brazill 1998). Σε κάθε περίπτωση, παρόλο που

τίθενται διαφορετικά κριτήρια βελτιστοποίησης, τα οποία ικανοποιούνται με διαφορετικούς υπολογιστικούς αλγόριθμους εν γένει, ωστόσο τα παραγόμενα βασικά αριθμητικά αποτελέσματα ταυτίζονται (με προσέγγιση συντελεστών κανονικοποίησης). Όπως είδαμε στα προηγούμενα, μέσω της ΠΑΑ για κάθε αντικείμενο σε κάθε παραγοντικό άξονα υπολογίζεται μία αριθμητική τιμή (βαθμός ή σκορ). Οι βαθμοί αυτοί δεν είναι παρά οι συντεταγμένες των προβολών των αντικειμένων πάνω στους αντίστοιχους άξονες. Ανάλογα, στις κατηγορίες (ιδιότητες ή κλάσεις) των κατηγορικών μεταβλητών ανατίθενται αριθμητικά βάρη, με τέτοιο τρόπο ώστε οι κατηγορίες να προβάλλονται στο κέντρο βάρους των αντικειμένων που χαρακτηρίζονται από τις αντίστοιχες ιδιότητες. Η ποσοτικοποίηση αντικειμένων και μεταβλητών έχει τις παρακάτω συνέπειες:

- Μεγιστοποιείται η διάκριση μεταξύ των αντικειμένων ως προς τη λανθάνουσα δομή που εκφράζει ο κάθε άξονας. Ειδικότερα, στον πρώτο παραγοντικό άξονα η διακύμανση (αδράνεια) των βαθμών των αντικειμένων, και συνεπώς η διάκρισή τους, είναι η μέγιστη δυνατή.
- Στην περίπτωση πολλών μεταβλητών, επί του πρώτου παραγοντικού άξονα μεγιστοποιείται το τετράγωνο της μέσης συσχέτισης μεταξύ των βαθμών των αντικειμένων και των ποσοτικοποιημένων πλέον κατηγορικών μεταβλητών. Στη διμεταβλητή περίπτωση μεγιστοποιείται η (κανονικοποιημένη) συσχέτιση των δύο μεταβλητών. Οι προβολές των αντικειμένων και των κλάσεων των μεταβλητών στον πρώτο άξονα αποτελούν θέσεις βέλτιστης ταξιθέτησης που αναδεικνύουν τη διάταξη των αντίστοιχων σημείων.
- Μεγιστοποιείται η ομοιογένεια των μεταβλητών και η αξιοπιστία (εσωτερική συνέπεια) των παραγοντικών αξόνων.
- Τα σύνολα των βαθμών των αντικειμένων σε κάθε άξονα είναι ανά δύο γραμμικά ασυσχέτιστα. Με άλλα λόγια, οι παραγοντικοί άξονες, ως νέες σύνθετες ποσοτικές μεταβλητές, εκφράζουν γραμμικά ανεξάρτητες δομές ή διαστάσεις του υπό εξέταση φαινομένου. Οι νέες μεταβλητές ή, αλλιώς, οι “παράγοντες” μπορούν, στη συνέχεια, να χρησιμοποιηθούν σε περαιτέρω στατιστικές αναλύσεις. Βέβαια, η κλίμακα μέτρησης των βέλτιστων βαθμών των αντικειμένων είναι αυθαίρετη. Από τη στιγμή που δοθεί φυσική ερμηνεία στους άξονες, οι αντίστοιχοι παραγοντικοί βαθμοί, για πρακτικούς λόγους, μπορούν να

αναχθούν σε μια νέα κλίμακα μέτρησης με συγκεκριμένο εύρος και τυποποίηση. Για παράδειγμα, ο παρακάτω μετασχηματισμός:

$$z_{ij} = \frac{x_{ij} - min(x_{ij})}{max(x_{ij}) - min(x_{ij})} \times 100,$$

όπου x_{ij} είναι ο βαθμός του *i* αντικειμένου στον *j* άξονα, μετασχηματίζει και τυποποιεί τις παραγοντικές συντεταγμένες των αντικειμένων στην κλίμακα από 0 έως 100. Το 0 αντιστοιχεί σε αντικείμενο με τις συγκριτικά χαμηλότερες τιμές στη δομή (ιδιότητα, χαρακτηριστικό) που εκφράζει ο αντίστοιχος παράγοντας, ενώ το 100 σε αντικείμενο με τις συγκριτικά υψηλότερες. Έτσι, αν τα αντικείμενα είναι "επώνυμα", οι νέοι βαθμοί z_{ij} μπορούν να χρησιμοποιηθούν είτε για την κατάταξη (*ranking*) είτε για την ανάπτυξη τυπολογιών των αντικειμένων.

Να υπενθυμίσουμε ότι κατά τον υπολογισμό των βέλτιστων τιμών:

- Έχουν ληφθεί υπόψη οι συσχετίσεις ανά δύο όλων των κατηγορικών μεταβλητών που συμμετέχουν στην ανάλυση.
- Η χ^2 απόσταση εξασφαλίζει ότι για κάθε ιδιότητα λαμβάνεται υπόψη, κατά την ανάλυση, η κατανομή της σε σχέση με την ιδιαιτερότητά της (σπανιότητα) ως προς τις υπόλοιπες ιδιότητες.

Πληροφοριακά αναφέρουμε ότι έχουν γίνει προσπάθειες ενοποίησης των μεθόδων βέλτιστης κλιμάκωσης με μεθόδους που στηρίζονται στα Γενικά Γραμμικά Υποδείγματα, όπως είναι η Γραμμική Παλινδρόμηση, η Κανονικοποιημένη Συσχέτιση και η Ανάλυση Διακύμανσης (βλέπε Perreault & Young 1980, Sands & Young 1980, Young 1981). Οι προτεινόμενες μέθοδοι στηρίζονται στον αλγόριθμο ALS με εναλλασσόμενες φάσεις και έχουν ως σκοπό την επέκταση των δυνατοτήτων που προσφέρουν τα γραμμικά μοντέλα σε καταστάσεις, όπου οι διαθέσιμες μεταβλητές είναι μικτού τύπου, δηλαδή κατηγορικές και ποσοτικές (συνεχείς ή ασυνεχείς). Οι προς εκτίμηση παράμετροι μπορούν να χωριστούν σε δύο ομάδες. Στη μία ομάδα ανήκουν οι παράμετροι των γραμμικών υποδειγμάτων και στην άλλη οι παράμετροι βέλτιστης κλιμάκωσης των δεδομένων. Σε κάθε επανάληψη του

αλγόριθμου, οι υπό συνθήκη[15] εκτιμητές ελαχίστων τετραγώνων των παραμέτρων της μιας ομάδας εναλλάσονται με αυτούς της άλλης. Οι νέοι εκτιμητές αντικαθιστούν τους παλιούς και η διαδικασία συνεχίζεται εωσότου τα κριτήρια βελτιστοποίησης της αντίστοιχης συνάρτησης απώλειας ικανοποιηθούν. Όμως, οι μέθοδοι αυτοί παρουσιάζουν δύο βασικά μειονεκτήματα: α) τα αποτελέσματα παρουσιάζουν αστάθεια λόγω του υπερβολικά μεγάλου αριθμού παραμέτρων που εκτιμώνται κατά την υλοποίηση του αλγόριθμου και β) οι περιορισμοί που επιβάλλονται από τις τεχνικές και θεωρητικές προϋποθέσεις καταλληλότητας εφαρμογής των γραμμικών υποδειγμάτων σε συνδυασμό με αυτούς των τεχνικών βέλτιστης κλιμάκωσης είναι δυνατό να οδηγήσουν σε τοπικά και όχι ολικά βέλτιστες λύσεις. Για το λόγο αυτό θα πρέπει τα αποτελέσματα να ελέγχονται τόσο ως προς την εγκυρότητά τους όσο και ως προς τη διαμόρφωση των αρχικών τυχαίων τιμών που ανατίθενται στις δύο ομάδες παραμέτρων κατά την εκκίνηση του αλγόριθμου.

2.6 Σχόλια και Συμπεράσματα Κεφαλαίου

Η ΠΑΑ θεωρείται κυρίως ως μία περιγραφική μέθοδος για τη διερεύνηση της σχέσης μεταξύ δύο ή περισσότερων κατηγορικών μεταβλητών, χωρίς *a priori* υποθέσεις και προϋποθέσεις. Μπορεί να χαρακτηριστεί ως επέκταση της Ανάλυσης σε Κύριες Συνιστώσες για την ανάλυση κατηγορικών δεδομένων. Πρωταρχικός σκοπός της μεθόδου είναι η ανάδειξη και οπτικοποίηση της ενδογενούς δομής του πίνακα που θα δοθεί ως είσοδος στην ανάλυση σε χώρους προβολής με μικρότερες διαστάσεις από τον αρχικό, στον οποίο μπορεί να αναλυθεί το υπό εξέταση φαινόμενο. Μέσω της ΠΑΑ το πληροφοριακό περιεχόμενο του πίνακα εισόδου, όπως αυτό μετριέται μέσω της αδράνειας, διασπάται σε παραγοντικούς άξονες και “μεγεθύνεται”. Έτσι, αναδεικνύονται ομοιότητες, αντιπαραθέσεις και αλληλεπιδράσεις μεταξύ των προβαλλόμενων σημείων, οι οποίες δεν είναι άμεσα αντιληπτές αλλά βρίσκονται σε λανθάνουσα μορφή. Μάλιστα, αυτό επιτυγχάνεται χωρίς τη χρήση στατιστικών ελέγχων σημαντικότητας για την απόρριψη ή όχι υποθέσεων σχετικά με αυτά. Η ΠΑΑ χρησιμοποιείται για την ανάλυση ποιοτικών ή/και κατάλληλα κατηγοριοποιημένων (κωδικοποιημένων) ποσοτικών δεδομένων, τα οποία μπορούν να οργανωθούν σε απλούς και σύνθετους πίνακες συνάφειας. Είναι αρκετά ευέλικτη,

[15] Διατηρώντας σταθερές τις παραμέτρους της άλλης ομάδας.

ώστε να μπορεί να εφαρμοστεί σχεδόν σε κάθε πίνακα της μορφής «*αντικείμενα* × *μεταβλητές*», με μη αρνητικά στοιχεία και μη μηδενικά αθροίσματα γραμμών και στηλών, αρκεί οι μεταβλητές να είναι ομοιογενείς. Στην πράξη, μπορεί να εφαρμοστεί σε οποιονδήποτε πίνακα διπλής εισόδου με θετικά στοιχεία. Η δυνατότητα αυτή σε συνδυασμό με τις ιδιότητες των αποτελεσμάτων της την καθιστούν ένα γενικό σύστημα ανάλυσης κατηγορικών δεδομένων. Ο πολυδιάστατος και πολυμεταβλητός χαρακτήρας της μεθόδου επιτρέπει την ανάδειξη σχέσεων και δομών που δεν θα μπορούσαν να εντοπιστούν με απλή εξέταση των μεταβλητών κατά ζεύγη. Σήμερα, η μέθοδος χρησιμοποιείται σχεδόν σε όλα τα ερευνητικά επιστημονικά πεδία είτε όπως έχει είτε με παραλλαγές και τροποποιήσεις στο αλγοριθμικό της μέρος.

Η παράλληλη και ανεξάρτητη εξέλιξη της ΠΑΑ σε ερευνητικά πεδία ποικίλων επιστημονικών περιοχών είχε ως αποτέλεσμα να δημιουργηθούν διάφορες σχολές και κατευθύνσεις τόσο σε σχέση με το αλγοριθμικό όσο και σε σχέση με το θεωρητικό υπόβαθρο της μεθόδου. Οι δύο Σχολές Ανάλυσης Δεδομένων που ξεχωρίζουν είναι η Γαλλική και η Ολλανδική (Σύστημα *GIFI*). Στα πορίσματά τους στηρίζονται οι υπολογιστικοί αλγόριθμοι των σύγχρονων εμπορικών λογισμικών στατιστικής επεξεργασίας δεδομένων που διαθέτουν την ΠΑΑ. Στη Γαλλική Σχολή δίνεται μεγαλύτερη έμφαση στη γεωμετρική ερμηνεία των δεδομένων μέσω της απόστασης χ^2 κατά Benzécri και το υπολογιστικό της μέρος είναι κατά βάση αλγεβρικό. Η πολυμεταβλητή εκδοχή της ΠΑΑ εφαρμόζεται αλγοριθμικά και εννοιολογικά ακριβώς με τον ίδιο τρόπο όπως και στην περίπτωση των δύο μεταβλητών. Στην Ολλανδική, η μέθοδος, ιδιαίτερα στην πολυμεταβλητή εκδοχή της, αντιμετωπίζεται ως πρόβλημα βέλτιστης κλιμάκωσης ή, κατά μία άλλη έννοια, ως μέθοδος ποσοτικοποίησης κατηγορικών (ποιοτικών) δεδομένων. Στην περίπτωση αυτή, στόχος είναι η εξεύρεση, μέσω του επαναληπτικού αλγόριθμου ALS, βέλτιστων βαθμών για τις γραμμές και τις στήλες του πίνακα που αναλύεται, ώστε να ελαχιστοποιείται μια συγκεκριμένη συνάρτηση απώλειας κάτω από περιορισμούς. Οι βέλτιστες τιμές μπορούν να χρησιμοποιηθούν, στη συνέχεια, σε περαιτέρω στατιστικές αναλύσεις. Η προσέγγιση των Ολλανδών οδήγησε στην ανάπτυξη της μεθόδου της Ανάλυσης Ομοιογένειας, η οποία αποτελεί τον πυρήνα του συστήματος *GIFI*. Η μέθοδος παράγει ισοδύναμα αποτελέσματα με αυτά που προκύπτουν από την εφαρμογή της ΠΑΑ στο λογικό πίνακα q κατηγορικών μεταβλητών. Όμως, οι δύο προσεγγίσεις

διαφέρουν ριζικά τόσο ως προς το υπολογιστικό όσο και ως προς το θεωρητικό πλαίσιο και την προβληματική που οδήγησε στην ανάπτυξή τους. Η "άγνοια" της μεθοδολογικής βάσης, πάνω στο οποίο τα διάφορα λογισμικά στηρίζουν τους υπολογισμούς τους, μπορεί να οδηγήσει σε εσφαλμένη συμπερασματολογία. Ειδικότερα, στο πνεύμα της Γαλλικής Σχολής, η μέθοδος κανονικοποίησης των τυποποιημένων συντεταγμένων των προβολών των σημείων γραμμών και στηλών, του πίνακα συμπτώσεων, πάνω στους παραγοντικούς άξονες είναι η Κύρια Κανονικοποίηση, ενώ η Ολλανδική Σχολή προσφέρει και άλλες μεθόδους, οι οποίες οδηγούν στην κατασκευή *biplot* διαγραμμάτων. Η μέθοδος κανονικοποίησης που εφαρμόζεται σε κάθε Σχολή καθορίζεται από διαφορετικά κριτήρια. Στη Γαλλική Σχολή η Κύρια Κανονικοποίηση έχει ως αποτέλεσμα οι αποστάσεις μεταξύ των σημείων γραμμών (στηλών) να προσεγγίζουν τις αποστάσεις χ^2. Στην Ολλανδική, το κριτήριο κανονικοποίησης καθορίζεται από την επιβολή περιορισμών ως προς την επιθυμητή τιμή του σταθμισμένου αθροίσματος τετραγώνων (αδράνειας, διασποράς) των συντεταγμένων των προβολών των σημείων γραμμών και στηλών επί των παραγοντικών αξόνων. Σε κάθε περίπτωση, η ερμηνεία των αποτελεσμάτων, αριθμητικών και διαγραμματικών είναι διαφορετική. Ιδιαίτερη προσοχή χρειάζεται στην ερμηνεία των *French plots* όπου στα παραγοντικά επίπεδα η ευκλείδεια απόσταση μεταξύ ενός σημείου γραμμής και ενός σημείου στήλης δεν ορίζεται. Έτσι, κατά τη γεωμετρική ερμηνεία των αποτελεσμάτων, το γεγονός ότι ένα σημείο γραμμής βρίσκεται κοντά με ένα σημείο στήλης δεν θα πρέπει να θεωρηθεί ως απόδειξη υψηλής συσχέτισης μεταξύ των δύο αντίστοιχων κατηγοριών. Αυτό είναι ιδιαίτερα σημαντικό αν αναλογιστούμε ότι στις δημοσιευμένες εργασίες αυτό που συνήθως παρουσιάζεται, από τα αποτελέσματα της ΠΑΑ, είναι ένα ή περισσότερα παραγοντικά διαγράμματα και η όλη συζήτηση στηρίζεται πάνω σε αυτά.

Στο κεφάλαιο αυτό, επιχειρήσαμε μια συγκριτική παρουσίαση της ΠΑΑ, όπως αυτή θεμελιώνεται και εφαρμόζεται στο πλαίσιο της Γαλλικής και Ολλανδικής Σχολής Ανάλυσης Δεδομένων. Δείξαμε με ποιο τρόπο οι δύο αυτές προσεγγίσεις συνδέονται και επισημάναμε, δίνοντας ταυτόχρονα και τις σχετικές κατευθύνσεις (υποδείξεις), τα σημεία που χρήζουν ιδιαίτερης προσοχής κατά την ερμηνεία των αποτελεσμάτων (αριθμητικών και γραφικών). Προβάλαμε τις ιδιότητες και τις δυνατότητες της μεθόδου και παραθέσαμε αναφορές σχετικά με τις σημαντικότερες εξελίξεις στο

ερευνητικό της πλαίσιο. Επίσης, εισαγάγαμε νέα εννοιολογικά και μεθοδολογικά στοιχεία στο πεδίο εφαρμογής της. Πιο συγκεκριμένα, στην Ενότητα 2.2.14.4 προτείναμε μέθοδο εντοπισμού, σε κάθε άξονα, των σημαντικών κελιών σε πίνακα συμπτώσεων δύο μεταβλητών, ενώ στην πολυμεταβλητή περίπτωση ορίσαμε το Σχετικό Δείκτη Διακριτότητας μιας μεταβλητής (Ενότητα 2.3.4). Και οι δύο προτάσεις συνεισφέρουν στη διεισδυτική ερμηνεία των παραγόμενων αποτελεσμάτων. Προς την ίδια κατεύθυνση συμβάλει και ο συνδυασμός των ιδιοτήτων της ΠΑΑ και της Ανάλυσης Ομοιογένειας, όπως αυτές αναδεικνύονται από τη συγκριτική παράθεση των μεθοδολογικών προσεγγίσεων των δύο Σχολών Ανάλυσης Δεδομένων.

ΚΕΦΑΛΑΙΟ 3

Ειδικά Υπολογιστικά Θέματα: Δύο Μεθοδολογικές Προτάσεις

3.1 Εισαγωγή

Όπως αναφέρθηκε στο Κεφάλαιο 1 (Ενότητα 1.3), ο σημαντικότερος, ίσως, παράγοντας, που οδήγησε στην υιοθέτηση και διάδοση των μεθόδων της Ανάλυσης Δεδομένων, ήταν η ανάπτυξη και η ευρεία χρήση των Η/Υ. Οι μέθοδοι απαιτούν πολύπλοκους αριθμητικούς υπολογισμούς που μόνο με τη βοήθεια Η/Υ είναι δυνατό, μέσα στα όρια της ανθρώπινης υπομονής, να επιτευχθούν, ιδιαίτερα όταν πρόκειται να αναλυθούν μεγάλα σύνολα δεδομένων. Έτσι, σήμερα, διαδεδομένα εμπορικά (αμερικανικά) στατιστικά πακέτα, όπως το SPSS και το SAS, διαθέτουν τη στατιστική διαδικασία της ΠΑΑ, γεγονός το οποίο οδήγησε στην εφαρμογή της μεθόδου σε ερευνητικές μελέτες από όλα σχεδόν τα επιστημονικά πεδία. Ειδικότερα, το SPSS θεωρείται ένα από τα πιο δημοφιλή και φιλικά στο χρήστη προγράμματα στατιστικής επεξεργασίας δεδομένων και αποτελεί λογισμικό αναφοράς. Διαθέτει το υποσύστημα *Categories*, το οποίο υλοποιεί τις σημαντικότερες μεθόδους βέλτιστης κλιμάκωσης του συστήματος *GIFI* της Ολλανδικής Σχολής (Gifi 1996, SPSS Inc. 2004α, 1998α και 1997, Meulman & Heiser 2004 και 2001). Περιλαμβάνει, επίσης, και τη στατιστική διαδικασία *Correspondence Analysis* (Ανάλυση των Αντιστοιχιών), η οποία μπορεί να χρησιμοποιηθεί για τη διερεύνηση της σχέσης μεταξύ δύο μόνον κατηγορικών μεταβλητών. Στο SPSS, μέχρι και την έκδοση 12, η πολυμεταβλητή εκδοχή της ΠΑΑ είναι διαθέσιμη, ως ένα βαθμό, μέσω της διαδικασίας *HOMALS*, δηλαδή της Ανάλυσης Ομοιογένειας με τη χρήση του αλγόριθμου ALS (βλέπε Ενότητα 2.3.4). Όμως, τα παραγόμενα αποτελέσματα δεν έχουν την ίδια αμεσότητα στην ερμηνεία με τα αντίστοιχα αποτελέσματα, τα οποία προκύπτουν από την εφαρμογή της ΠΑΑ στο πλαίσιο της Γαλλικής Σχολής (Μενεξές, 2001). Κατά την εφαρμογή της ΑΟ, τα μόνα αριθμητικά αποτελέσματα που υπολογίζονται είναι οι

ιδιοτιμές των παραγοντικών αξόνων, τα μέτρα διακριτότητας των μεταβλητών και οι κύριες συντεταγμένες των κλάσεων των μεταβλητών. Ο χρήστης θα πρέπει να προκαθορίσει τον αριθμό των επιθυμητών λύσεων (αξόνων) της ΑΟ, ενώ για τα αντικείμενα υπολογίζονται μόνο οι τυποποιημένες συντεταγμένες των προβολών τους πάνω στους παραγοντικούς άξονες. Με εξαίρεση τα μέτρα διακριτότητας, δεν παρέχεται καμία άλλη βοήθεια στην ερμηνεία των αποτελεσμάτων. Από την έκδοση 13 και μετά, του SPSS, η διαδικασία *HOMALS* αντικαταστάθηκε από τη *Multiple Correspondence Analysis* (Πολλαπλή Ανάλυση των Αντιστοιχιών), η οποία είναι "συμβατή" με το πνεύμα της Γαλλικής Σχολής (βλέπε Meulman & Heiser 2004, SPSS Inc. 2004α). Η διαδικασία παράγει τα ίδια βασικά αποτελέσματα με αυτά που υπολογίζονται κατά την εφαρμογή της ΠΑΑ σε λογικό πίνακα 0-1, μετά από επιλογή των κατάλληλων κανονικοποιήσεων των συντεταγμένων των σημείων (αντικειμένων και ιδιοτήτων) (βλέπε Ενότητα 2.3.3.2). Και στην περίπτωση αυτή, η μέθοδος υλοποιείται μέσω του επαναληπτικού αλγόριθμου ALS και βασικός στόχος εξακολουθεί να είναι η ποσοτικοποίηση ποιοτικών μεταβλητών και η ανάθεση βέλτιστων βαθμών στα αντικείμενα (Meulman & Heiser, 2004).

Στο SPSS η ΠΑΑ δεν μπορεί να εφαρμοστεί άμεσα σε γενικευμένους πίνακες συμπτώσεων *Burt*, δυνατότητα η οποία, στην περίπτωση μεγάλων συνόλων δεδομένων, θα βελτίωνε σημαντικά τους χρόνους επεξεργασίας και τη διαχείριση μνήμης του Η/Υ. Στη Γαλλική Σχολή, η ΠΑΑ εφαρμόζεται σχεδόν αποκλειστικά σε πίνακες *Burt*, με αποτέλεσμα να μην είναι εφικτός ο άμεσος υπολογισμός των συντεταγμένων των αντικειμένων επί των παραγοντικών αξόνων, ο οποίος, όμως, είναι εφικτός αν η ανάλυση εφαρμοστεί στον αντίστοιχο λογικό πίνακα 0-1. Η δυνατότητα αυτή είναι ιδιαίτερα χρήσιμη, διότι οι παραγοντικές συντεταγμένες (βαθμοί) των αντικειμένων μπορούν να χρησιμοποιηθούν στη συνέχεια σε άλλες αναλύσεις, όπως είναι η Ταξινόμηση. Από την άλλη πλευρά, η εφαρμογή της μεθόδου σε μεγάλους λογικούς πίνακες είναι προβληματική. Είτε η επεξεργασία καθυστερεί σημαντικά είτε δεν είναι εφικτή. Στις ενότητες που ακολουθούν: α) περιγράφουμε μια διαδικασία, με την οποία μπορούμε να χρησιμοποιήσουμε τις δυνατότητες του SPSS για την εφαρμογή της ΠΑΑ σε πίνακες *Burt* και β) προτείνουμε έναν αποτελεσματικό αλγόριθμο εφαρμογής της ΠΑΑ σε μεγάλους λογικούς πίνακες στο πνεύμα της Γαλλικής παράδοσης.

3.2 Ανάλυση Πινάκων *Burt* Μέσω του SPSS

Η υλοποίηση της προτεινόμενης διαδικασίας απαιτεί την επίλυση δύο βασικών προβλημάτων: α) τη δημιουργία του πίνακα *Burt* και β) την κωδικοποίησή του, ώστε να τεθεί σε μορφή αναγνωρίσιμη και επεξεργάσιμη από το SPSS. Από τη στιγμή που ο πίνακας *Burt* εισαχθεί στο λογισμικό, τότε μέσω της διαδικασίας *Correspondence Analysis* είναι δυνατό να αναλυθεί ως απλός πίνακας συμπτώσεων δύο μεταβλητών επιλέγοντας μία από τις τρεις μεθόδους κανονικοποίησης *RPN*, *CPN* ή *PN* (βλέπε Ενότητες 2.2.14.1, 2.3.3.2 και 2.3.3.5).

3.2.1 Σύντομη Περιγραφή της Διαδικασίας

Για την εφαρμογή της ΠΑΑ σε πίνακες *Burt* μέσω του SPSS ακολουθούμε την παρακάτω πορεία:

- Καταχωρούμε τα αρχικά δεδομένα (*αντικείμενα* × *μεταβλητές*) σε αρχείο τύπου *.sav στο SPSS.
- Στο ίδιο αρχείο δεδομένων δημιουργούμε νέες μεταβλητές με λογική κωδικοποίηση, ώστε να κατασκευαστεί ο πίνακας 0-1.
- Δημιουργούμε τον πίνακα *Burt* με “τέχνασμα” που στηρίζεται στις δυνατότητες παραγωγής αναφορών (υποσύστημα *Tables*) του SPSS.
- Εισάγουμε με κατάλληλη κωδικοποίηση τον πίνακα *Burt* σε ένα νέο αρχείο δεδομένων και αποδίδουμε βάρη στις γραμμές του νέου πίνακα.
- Αναλύουμε τον πίνακα *Burt* με τη στατιστική διαδικασία *Correspondence Analysis* εφαρμόζοντας μία από τις τρεις μεθόδους κανονικοποίησης *RPN*, *CPN* ή *PN*.
- Σε ένα νέο αρχείο δεδομένων αποθηκεύουμε τις συντεταγμένες των προβολών των σημείων γραμμών (ή στηλών) πάνω στους παραγοντικούς άξονες για περεταίρω αναλύσεις.

Στην Ενότητα Γ1 του Παραρτήματος Γ παρουσιάζουμε την αναλυτική περιγραφή της διαδικασίας κάνοντας χρήση αριθμητικού παραδείγματος.

3.2.2 Συμπεράσματα και Σχόλια

Η διαδικασία που προτείνουμε αποτελείται από πολλά επιμέρους βήματα και απαιτεί καλή γνώση του προγράμματος SPSS. Με δεδομένο ότι το SPSS είναι ένα δημοφιλές και διαδεδομένο πακέτο, η συγκεκριμένη διαδικασία αποτελεί μια λύση για τους ερευνητές, οι οποίοι δεν έχουν πρόσβαση σε ειδικό λογισμικό εφαρμογής της πολλαπλής εκδοχής της ΠΑΑ στο πρότυπο της Γαλλικής Σχολής, όπως είναι για παράδειγμα το SPAD.

Το Βήμα 3 (βλέπε Ενότητα Γ1 του Παραρτήματος Γ) είναι γενικότερα χρήσιμο για την εισαγωγή και κωδικοποίηση στο SPSS δευτερογενών (π.χ. δημοσιευμένων ή ήδη επεξεργασμένων) πινάκων συμπτώσεων απολύτων συχνοτήτων, αφού, στην περίπτωση αυτή, δεν είναι διαθέσιμα τα αναλυτικά πρωτογενή δεδομένα. Ο τρόπος κωδικοποίησης που προτείνουμε μπορεί να χρησιμοποιηθεί και για την εισαγωγή πινάκων τύπου “στοίβας”, “φέτας” ή άλλης μορφής υποπινάκων του *Burt*. Γενικά, μπορούμε να εισάγουμε όλους τους τύπους πινάκων που παρουσιάστηκαν στις Ενότητες 2.4 και Β3 (βλέπε Παράρτημα Β). Ειδικότερα, αν παραλείψουμε το Βήμα 2 (βλέπε Ενότητα Γ1), μπορούμε να εισάγουμε με την “κατακόρυφη” μορφή του Πίνακα Γ1.7 της Ενότητας Γ1 και τον ίδιο το λογικό πίνακα 0-1. Αν εισάγουμε τον πίνακα *Burt* που αντιστοιχεί σε δύο μόνο μεταβλητές, τότε η εφαρμογή της ΠΑΑ δίνει τα ίδια αποτελέσματα με τη μέθοδο κανονικοποίησης των Carroll, Green & Schaffer (1986 και 1987) (βλέπε Παρατήρηση 2.3 της Ενότητας 2.2.14.2).

Με την εφαρμογή της ΠΑΑ απευθείας στον πίνακα *Burt*, οι ελλείπουσες τιμές δεν δημιουργούν υπολογιστικό πρόβλημα, αφού αυτές μπορούν να μην ληφθούν υπόψη κατά την κατασκευή των επιμέρους απλών πινάκων συμπτώσεων που απαρτίζουν τον *Burt*.

Κατά την εφαρμογή της διαδικασίας *Correspondence Analysis* είναι δυνατό να υπολογιστούν, μέσω της μεθόδου «*Δέλτα*» (βλέπε Ενότητα 2.2.14.2 και Ενότητα ΣΤ1 του Παραρτήματος ΣΤ), εκτιμητές των διασπορών και συνδυασπορών τόσο των χαρακτηριστικών τιμών των παραγοντικών αξόνων όσο και των προβαλλόμενων σημείων (SPSS Inc., 2004α και 1997). Η δυνατότητα αυτή, όπως θα δούμε στο Κεφάλαιο 6 (Ενότητα 6.3), είναι σημαντική στην κατασκευή διαστημάτων ή

περιοχών εμπιστοσύνης για τον έλεγχο της σταθερότητας και εγκυρότητας των παραγόμενων αποτελεσμάτων.

Η προτεινόμενη διαδικασία έχει ελεγχθεί για τις εκδόσεις 10, 11.5 και 13 του SPSS και επαληθεύτηκε με την ανάλυση συνόλων δεδομένων από τη βιβλιογραφία.

3.3 Ένας Αποτελεσματικός Αλγόριθμος Εφαρμογής της ΠΑΑ σε Μεγάλα Σύνολα Δεδομένων

Όπως αναφέρθηκε στην προηγούμενη ενότητα, στην περίπτωση πολλών μεταβλητών και στο πλαίσιο της Γαλλικής Σχολής, η ΠΑΑ εφαρμόζεται συνήθως στον πίνακα *Burt*, με αποτέλεσμα να μην είναι εφικτός ο άμεσος υπολογισμός των συντεταγμένων των αντικειμένων, ο οποίος, όμως, είναι εφικτός αν η ανάλυση πραγματοποιηθεί στον αντίστοιχο λογικό πίνακα 0-1. Στα επόμενα επιχειρούμε μια σύνδεση των δύο προσεγγίσεων. Η βασική ιδέα της προτεινόμενης μεθοδολογίας στηρίζεται στο γεγονός ότι, παρόλο που κατά την ανάλυση του *Burt* "χάνεται" η πληροφορία για τα αντικείμενα και ο αρχικός πίνακας δεδομένων δεν μπορεί να ανασυσταθεί, ωστόσο είναι δυνατός ο υπολογισμός των παραγοντικών συντεταγμένων των αντικειμένων, ώστε τα αποτελέσματα να ταυτίζονται με αυτά που προκύπτουν από την ανάλυση του πίνακα 0-1.

3.3.1 Πρόταση Μεθόδου Υπολογισμού των Τυποποιημένων και Κύριων Συντεταγμένων των Αντικειμένων από τον Πίνακα *Burt*

Έστω ότι έχουμε N αντικείμενα που χαρακτηρίζονται από q κατηγορικές μεταβλητές με l_k κλάσεις (ιδιότητες) η κάθε μια (k=1,…, q). Ας είναι $j = \sum_{k=1}^{q} l_k$ το πλήθος των κατηγοριών των q μεταβλητών. Έστω $\mathbf{\Delta}$ ο $N \times q$ πίνακας δεδομένων της μορφής «*αντικείμενα* × *μεταβλητές*». Από τον $\mathbf{\Delta}$ κατασκευάζουμε τον αντίστοιχο $N \times j$ λογικό πίνακα $\mathbf{Z}$ ($\mathbf{Z_{0\text{-}1}}$). Από τη σχέση $\mathbf{B}=\mathbf{Z}^{\mathbf{T}}\mathbf{Z}$ (βλέπε Ενότητα 2.3.2) προκύπτει ο συμμετρικός $j \times j$ γενικευμένος πίνακας συμπτώσεων απολύτων συχνοτήτων $\mathbf{B}$, δηλαδή ο πίνακας *Burt*.

Όπως αναφέρθηκε στο Κεφάλαιο 2 (Ενότητα 2.3.4.2), στην Ολλανδική Σχολή η διαδικασία της Ανάλυσης Ομοιογένειας (ΑΟ) παράγει αποτελέσματα συγκρίσιμα με αυτά της ΠΑΑ της Γαλλικής Σχολής. Κατά την υλοποίηση της ΑΟ βελτιστοποιείται με χρήση του επαναληπτικού αλγορίθμου ALS η συνάρτηση απώλειας [2.70] κάτω από τους περιορισμούς i) και ii) (βλέπε Ενότητα 2.3.4). Οι παραγοντικές συντεταγμένες των N αντικειμένων, που υπολογίζονται στην ΑΟ, είναι οι τυποποιημένες, οι οποίες προκύπτουν αν η ΠΑΑ εφαρμοστεί απευθείας στον **Z** (βλέπε Ενότητα 2.3.4.2, Παρατήρηση Α). Όμως, η εφαρμογή της ΠΑΑ στον **Z** είτε δεν είναι πάντα εφικτή υπολογιστικά είτε οι υπολογισμοί καθυστερούν σημαντικά. Για παράδειγμα, έστω ότι έχουμε 20 κατηγορικές μεταβλητές με 82 συνολικά κατηγορίες σε ένα δείγμα 12.480 αντικειμένων. Στην περίπτωση αυτή ο πίνακας *Burt* είναι διαστάσεων 82×82, ενώ ο αντίστοιχος πίνακας **Z** είναι διαστάσεων 12.480×82. Είδαμε στο Κεφάλαιο 2, ότι ο αλγόριθμος της ΠΑΑ στηρίζεται στη μέθοδο SVD κατάλληλα κεντροποιημένων πινάκων ίδιων διαστάσεων με τους αρχικούς (*Burt* ή **Z**). Ενδεικτικά, αναφέρουμε ότι στο λογισμικό MATLAB, έκδοση 6.5 R13, η εφαρμογή της SVD σε πίνακα διαστάσεων 12.480×82 είναι περίπου 130 φορές πιο αργή από την εφαρμογή της SVD στον αντίστοιχο πίνακα *Burt* (82×82)[16]. Έτσι, από τη μια πλευρά, η εφαρμογή της ΠΑΑ στον *Burt* έχει υπολογιστικό πλεονέκτημα αλλά, από την άλλη, έχει ως αποτέλεσμα να μην είναι διαθέσιμη η πληροφορία για τα 12.480 αντικείμενα, πληροφορία η οποία θα μπορούσε να χρησιμοποιηθεί σε περαιτέρω αναλύσεις, όπως η είναι η Ανάλυση Συστάδων (Ταξινόμηση).

Ας είναι $\mathbf{S_Z}$ και $\mathbf{S_B}$ οι πίνακες που διαγωνοποιούνται κατά την εφαρμογή της ΠΑΑ στους πίνακες **Z** και **B** αντίστοιχα (βλέπε Ενότητα 2.3.3). Από την εφαρμογή της SVD προκύπτουν οι σχέσεις (βλέπε Ενότητα 2.3.3.5 και σχέσεις [2.67] και [2.68]):

$$\mathbf{S_Z = UDV^T} \qquad [3.1]$$

και

$$\mathbf{S_B = S_Z^T S_Z = VD^2V^T} \qquad [3.2]$$

με τους περιορισμούς $\mathbf{U^TU=I}$, $\mathbf{V^TV=I}$, όπου **I** είναι ο μοναδιαίος πίνακας, **D** και $\mathbf{D^2}$ διαγώνιοι πίνακες διαστάσεων, το πολύ, $(j\text{-}q) \times (j\text{-}q)$.

[16] Η δοκιμή πραγματοποιήθηκαν σε Η/Υ με επεξεργαστή Pentium(R) 4, CPU 2.8 GHz και 1GB RAM.

Στη σχέση [3.1] ο **U** είναι ο πίνακας με στήλες τα αριστερά ορθοκανονικά χαρακτηριστικά διανύσματα (χ.δ.) του $\mathbf{S_Z}$, **D** είναι ο διαγώνιος πίνακας με τις χαρακτηριστικές τιμές (χ.τ.) $d_1, d_2, \ldots, d_p$ ($p=j-q$, βλέπε Ενότητα 2.3.3.1) σε φθίνουσα σειρά κατά μήκος της κύριας διαγωνίου και $\mathbf{V^T}$ είναι ο ανάστροφος του **V**, του οποίου οι στήλες είναι τα δεξιά ορθοκανονικά (χ.δ.) του $\mathbf{S_Z}$. Στη σχέση [3.2] ο $\mathbf{D^2}$ είναι ο διαγώνιος πίνακας με στοιχεία τις ιδιοτιμές $\lambda_1 = d_1^2$, $\lambda_2 = d_2^2, \ldots, \lambda_p = d_p^2$ του $\mathbf{S_B}$. Οι ιδιοτιμές λ_s ($s=1,\ldots, p$) είναι ίσες με τις αδράνειες των αντίστοιχων παραγοντικών αξόνων από την ανάλυση του $\mathbf{S_Z}$. Τα ιδιοδιανύσματα του $\mathbf{S_B}$ είναι οι στήλες $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$ του **V**, δηλαδή είναι ίσα με τα δεξιά χ.δ. του $\mathbf{S_Z}$. Τα αριστερά χ.δ. του $\mathbf{S_Z}$ μπορούν να υπολογιστούν από το γραμμικό μετασχηματισμό του **V**: $\mathbf{U}=\mathbf{S_Z V D^{-1}}$. Κατά την εφαρμογή της ΠΑΑ (βλέπε Τρίτο Βήμα της Ενότητας 2.2.14), οι τυποποιημένες συντεταγμένες (τ.σ.) των στηλών των πινάκων **Z** και **B**, επί του άξονα *s*, δίνονται από τις σχέσεις [3.3] και [3.4] αντίστοιχα:

$$c_{ms} = \frac{v_{ms}}{\sqrt{c_m}}, \qquad [3.3]$$

όπου $s=1,\ldots, p$ και ο δείκτης m ($m=1,\ldots, j$) δηλώνει τη στήλη m του **Z,** η οποία έχει μάζα ίση με c_m,

και

$$c_{ms} = \frac{v_{ms}}{\sqrt{c'_m}}, \qquad [3.4]$$

όπου $s=1,\ldots, p$ και ο δείκτης m ($m=1,\ldots, j$) αντιστοιχεί στη στήλη m του **B** η οποία έχει μάζα c'_m.

Η ποσότητα c_m στη σχέση [3.3] είναι ίση με το σύνολο f_{+m} της στήλης m διαιρεμένο με το γενικό σύνολο Nq του πίνακα **Z** (βλέπε Ενότητα 2.3.1.1). Αντίστοιχα, η μάζα c'_m της σχέσης [3.4] είναι ίση με το σύνολο qf_{+m} της στήλης m διαιρεμένο με το γενικό σύνολο Nq^2 του πίνακα **B** (βλέπε Ενότητα 2.3.2.1). Συνεπώς έχουμε:

$$c_m = \frac{f_{+m}}{Nq}$$

και

$$c'_m = \frac{qf_{+m}}{Nq^2}.$$

Από τις παραπάνω σχέσεις είναι φανερό ότι:

$$c_m = c'_m. \qquad [3.5]$$

Έτσι, αν $\mathbf{c_Z}(s)$ είναι το διάνυσμα των τ.σ. των j κατηγοριών (ιδιοτήτων) των q μεταβλητών πάνω στον άξονα s από την ανάλυση του **Z** και $\mathbf{c_B}(s)$ το αντίστοιχο διάνυσμα των τ.σ. των κατηγοριών των μεταβλητών από την ανάλυση του **B**, τότε από τις σχέσεις [3.3], [3.4] και [3.5] συνεπάγεται ότι:

$$\mathbf{c_Z}(s) = \mathbf{c_B}(s), \qquad [3.6]$$

δηλαδή, οι τ.σ. των προβολών των κατηγοριών των q μεταβλητών επί των παραγοντικών αξόνων είναι ίσες για τους πίνακες **Z** και **B**.

Τώρα, αν $\mathbf{k_Z}(s)$ είναι το διάνυσμα των κύριων συντεταγμένων (κ.σ.) των j κατηγοριών των q μεταβλητών στον άξονα s από την ανάλυση του **Z** και $\mathbf{k_B}(s)$ το αντίστοιχο διάνυσμα των κ.σ. των κατηγοριών από την ανάλυση του **B**, τότε έχουμε (βλέπε Τέταρτο Βήμα της Ενότητας 2.2.14 και σχέσεις [2.16] και [2.17]):

$$\mathbf{k_B}(s) = d_s^2 \times \mathbf{c_B}(s) = \lambda_s \times \mathbf{c_B}(s), \qquad [3.7]$$

και

$$\mathbf{k_Z}(s) = d_s \times \mathbf{c_Z}(s) = \sqrt{\lambda_s} \times \mathbf{c_Z}(s), \qquad [3.8]$$

όπου λ_s και d_s είναι οι χ.τ. των πινάκων $\mathbf{S_B}$ και $\mathbf{S_Z}$ που αντιστοιχούν στον άξονα s.

Λόγω της σχέσης [3.6] η [3.8] μπορεί να γραφεί και ως εξής:

$$\mathbf{k_Z}(s) = \sqrt{\lambda_s} \times \mathbf{c_B}(s) \qquad [3.9]$$

Από την [3.9] προκύπτει ότι οι τ.σ. των κατηγοριών των μεταβλητών από την ανάλυση του **B** θα πρέπει να πολλαπλασιαστούν επί $\sqrt{\lambda_s} = d_s$ για να δώσουν τις κ.σ., όπως αυτές προκύπτουν από την ανάλυση του **Z**.

Χρησιμοποιώντας τους συμβολισμούς της Ενότητας 2.1.14.2, ας είναι φ_{is} και r_{is} οι κύριες και τυποποιημένες συντεταγμένες αντίστοιχα του αντικειμένου i (i=1,…, N) στον άξονα s (s=1,…, (j-q)) και γ_{ts} η κ.σ. της ιδιότητας t (t=1,…, j) επί του ίδιου άξονα. Από τις σχέσεις μετάβασης έχουμε τα εξής:

Από τη σχέση [2.16]: $\varphi_{is} = d_s r_{is}$.

Από τη σχέση [2.18]: $\varphi_{is} = \frac{1}{d_s}\sum_{t=1}^{j}\frac{p_{it}}{r_i}\gamma_{ts}$.

Από τις [2.16] και [2.17] συνεπάγεται ότι:

$$d_s r_{is} = \frac{1}{d_s}\sum_{t=1}^{j}\frac{p_{it}}{r_i}\gamma_{ts} \Rightarrow r_{is} = \frac{1}{d_s^2}\sum_{t=1}^{j}\frac{p_{it}}{r_i}\gamma_{ts}. \qquad [3.10]$$

Η σχέση [3.10] στην περίπτωση του πίνακα **Z** γράφεται (βλέπε Ενότητα 2.3.3):

$$r_{is} = \frac{1}{d_s^2}\sum_{t=1}^{j}\frac{p_{it}}{r_i}\gamma_{ts} = \frac{1}{\lambda_s}\sum_{t=1}^{j}\frac{\frac{z_{it}}{Nq}}{\frac{q}{Nq}}\gamma_{ts} = \frac{1}{\lambda_s}\frac{1}{q}\sum_{t=1}^{j} z_{it}\gamma_{ts}. \qquad [3.11]$$

όπου z_{it} είναι στοιχείο του πίνακα **Z** με τιμές 0 ή 1, λ_s και d_s οι χ.τ. του άξονα s από την ανάλυση των πινάκων $\mathbf{S_B}$ και $\mathbf{S_Z}$ αντίστοιχα.

Από την [3.11] διαπιστώνουμε ότι οι τ.σ. των αντικειμένων σε κάθε άξονα μπορούν να υπολογιστούν ως ο μέσος όρος ανά μεταβλητή των νέων κανονικοποιημένων συντεταγμένων των κατηγοριών, οι οποίες προέκυψαν από την ανάλυση του πίνακα *Burt* και τη σχέση [3.9], διαιρεμένος με τη χ.τ. του αντίστοιχου άξονα ($\lambda_s = d_s^2$). Επομένως, λόγω της [2.16] οι κ.σ. των αντικειμένων σε κάθε άξονα μπορούν να υπολογιστούν πολλαπλασιάζοντας τις τ.σ. επί την τετραγωνική ρίζα της χ.τ. του αντίστοιχου άξονα ($d_s = \sqrt{\lambda_s}$). Πιο συγκεκριμένα, αν το αντικείμενο i (i=1,…, N) χαρακτηρίζεται από τις ιδιότητες l_{gk} (k=1,…, q και g=1,…, l_k) και αν συμβολίσουμε τώρα με $y_{gk}(s)$ τις κ.σ. των προβολών των ιδιοτήτων πάνω στον άξονα s, τότε η τυποποιημένη r_{is} και η κύρια φ_{is} συντεταγμένη του αντικειμένου i στον άξονα s δίνονται από τις παρακάτω σχέσεις αντίστοιχα:

$$r_{is} = \frac{\sum_{k=1}^{q} y_{gk}(s)}{q\lambda_s} = \frac{\sum_{k=1}^{q} y_{gk}(s)}{qd_s^2}, \qquad [3.12]$$

και

$$\varphi_{is} = r_{is}\sqrt{\lambda_s} = r_{is}d_s . \qquad [3.13]$$

Οι κ.σ. των *N* αντικειμένων πάνω στον άξονα *s*, οι οποίες υπολογίζονται από τη σχέση [3.11], έχουν μέση τιμή 0 και διακύμανση ίση με την χ.τ. λ_s, δηλαδή την αδράνεια του άξονα *s* από την ανάλυση του πίνακα **Z** (βλέπε Παρατήρηση 2.2 της Ενότητας 2.2.14.1 και Παρατήρηση Α της Ενότητας 2.3.4.2).

Συνοπτικά και σε πρακτικό επίπεδο η παραπάνω διαδικασία περιλαμβάνει τα ακόλουθα βήματα:

1) Εφαρμογή της ΠΑΑ στον πίνακα *Burt*.
2) Υπολογισμός και κανονικοποίηση των τ.σ. των γραμμών (ή στηλών) του *Burt* μέσω της σχέσης [3.9] για κάθε παραγοντικό άξονα *s*.
3) Αντικατάσταση στον πίνακα δεδομένων **Δ** των κωδικοποιημένων τιμών των ιδιοτήτων των μεταβλητών με τις νέες κανονικοποιημένες συντεταγμένες, που υπολογίστηκαν στο προηγούμενο βήμα, για κάθε άξονα *s*.
4) Υπολογισμός των τ.σ. των αντικειμένων μέσω της σχέσης [3.12] για κάθε *s*.
5) Υπολογισμός των κ.σ. των αντικειμένων μέσω της σχέσης [3.13] για κάθε *s*.

Τα βήματα 3), 4) και 5) επαναλαμβάνονται το πολύ *p* φορές, όσες είναι δηλαδή και οι διαστάσεις του χώρου, στον οποίο μπορεί να αναλυθεί η βασική δομή των πινάκων $\mathbf{S_B}$ και $\mathbf{S_Z}$ χωρίς απώλεια πληροφορίας. Αν επιθυμούμε λύση με τους πρώτους $t < p$ παραγοντικούς άξονες, τότε τα βήματα 3), 4) και 5) επαναλαμβάνονται μόνο *t* φορές με αποτέλεσμα να εξοικονομείται επιπλέον υπολογιστικό έργο.

Στην Ενότητα Γ2 του Παραρτήματος Γ δίνουμε ένα αριθμητικό παράδειγμα εφαρμογής του προτεινόμενου αλγόριθμου. Στην ενότητα που ακολουθεί παρουσιάζουμε δύο εφαρμογές που αναδεικνύουν την αποτελεσματικότητα του

προτεινόμενου αλγόριθμου. Η πρώτη αφορά στην Ταξινόμηση αντικειμένων και η δεύτερη στην ανάλυση μεγάλων πινάκων δεδομένων.

3.3.2 Αποτελεσματικότητα του Αλγόριθμου

3.3.2.1 Εφαρμογή στην Ταξινόμηση Αντικειμένων

Οι παράγοντες που επηρεάζουν τα αποτελέσματα (δομή και ιεραρχία), τα οποία προκύπτουν από την εφαρμογή μιας μεθόδου Ιεραρχικής Ταξινόμησης στα αντικείμενα (γραμμές) ως προς τις μεταβλητές (στήλες) ενός πίνακα δεδομένων της μορφής «*αντικείμενα* × *μεταβλητές*», μπορούν να συνοψιστούν στους εξής: 1) οι μονάδες μέτρησης των μεταβλητών, 2) η διακύμανση των μεταβλητών, 3) η ύπαρξη πολυ-συγγραμμικότητας μεταξύ των μεταβλητών, 4) η ύπαρξη έκτοπων τιμών (*outliers*), 5) η συμμετοχή αντικειμενικά "άσχετων" μεταβλητών με το υπό εξέταση φαινόμενο (μεταβλητές θορύβου), και 6) η αλληλεπίδραση της μεθόδου ταξινόμησης με το είδος των δεδομένων. Για περισσότερες πληροφορίες σχετικά με τον τρόπο και την κατεύθυνση που οι πρώτοι πέντε παράγοντες μπορούν να επηρεάσουν τα αποτελέσματα μιας ταξινόμησης παραπέμπουμε στους Williams (1971), Aldenderfer και Blashfield (1984), Kaufman και Rousseeuw (1990), Everitt (1993), Hair *et al*. (1995), Johnson (1998) και Hair και Black (2000). Κρίνουμε ότι ιδιαίτερο ενδιαφέρον παρουσιάζει ο έκτος παράγοντας. Η μέθοδος ταξινόμησης αναφέρεται στην απόσταση μεταξύ των αντικειμένων και στο κριτήριο σχηματισμού των συστάδων, ενώ το είδος των δεδομένων καθορίζεται τόσο από την κλίμακα μέτρησης των μεταβλητών (δυαδική, ονομαστική, διάταξης, διαστήματος και αναλογίας) όσο και από την ακρίβεια μέτρησης των τιμών των μεταβλητών. Η αλληλεπίδραση της μεθόδου ταξινόμησης με το είδος των δεδομένων είναι αρκετά σύνθετη. Ενδεικτικά αναφέρουμε τις παρακάτω περιπτώσεις:

- Για το ίδιο σύνολο δεδομένων και για την ίδια απόσταση η εφαρμογή διαφορετικών κριτηρίων σχηματισμού των συστάδων οδηγεί, εν γένει, σε διαφορετικά αποτελέσματα. Το ίδιο ισχύει και στην περίπτωση που για το ίδιο κριτήριο χρησιμοποιηθούν δύο διαφορετικές αποστάσεις.

- Στην περίπτωση κατηγορικών μεταβλητών χρησιμοποιούνται, εν γένει, διαφορετικές αποστάσεις και κριτήρια απ' ότι στην περίπτωση των ποσοτικών μεταβλητών.

Η παραπάνω αλληλεπίδραση γίνεται πιο πολύπλοκη αν συνδυαστεί με το επιστημονικό αντικείμενο, στο πλαίσιο του οποίου θα ερμηνευθούν τα αποτελέσματα της ταξινόμησης. Έτσι, για παράδειγμα, για το ίδιο σύνολο (ή είδος) δεδομένων η εφαρμογή μιας μεθόδου ταξινόμησης *A* μπορεί να έχει καλύτερη φυσική ερμηνεία από μια άλλη μέθοδο *B* όταν τα αποτελέσματα θα ερμηνευθούν από έναν κοινωνιολόγο, ενώ μπορεί να συμβεί το αντίθετο, δηλαδή η μέθοδος *B* να κριθεί ως καταλληλότερη, στην περίπτωση που τα αποτελέσματα ερμηνευθούν από έναν οικονομολόγο. Σε κάθε περίπτωση, η απόφαση σχετικά με την επιλογή της μεθόδου ταξινόμησης φαίνεται να εξαρτάται σε μεγάλο βαθμό από την ιδιαιτερότητα των δεδομένων και το σκοπό της έρευνας. Στο πλαίσιο αυτό, μια άλλη σημαντική απόφαση που πρέπει να ληφθεί είναι σχετικά με την τυποποίηση ή, αντίθετα, την απόδοση "βάρους" στις μεταβλητές κατά την εφαρμογή της ταξινόμησης.

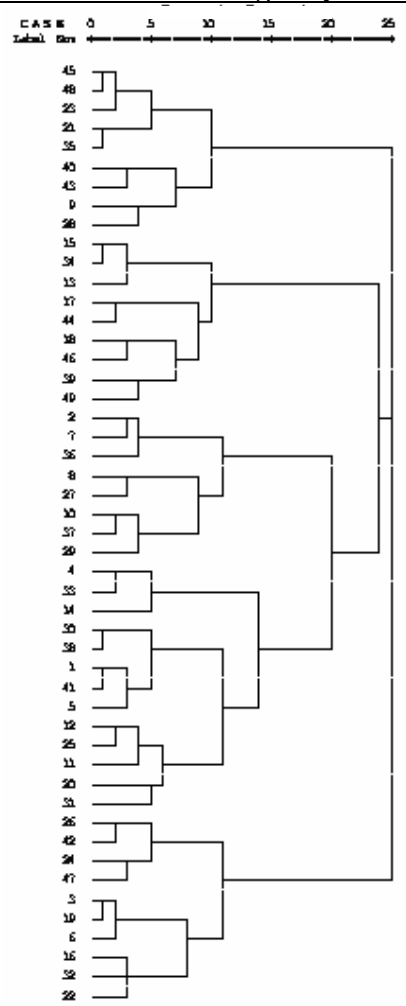
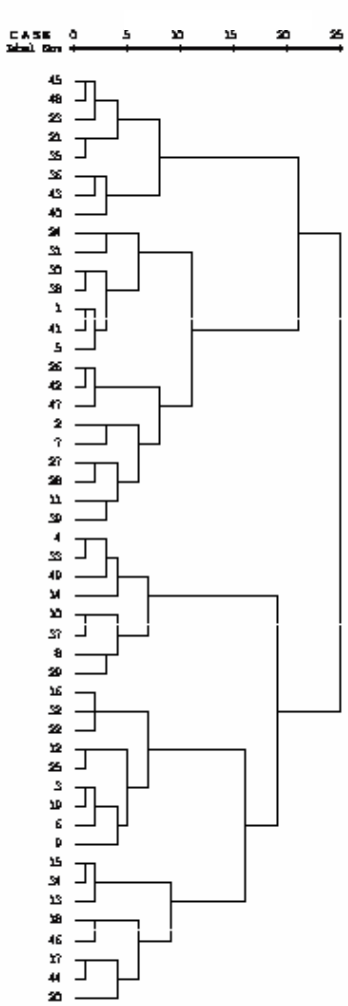
Η τυποποίηση των μεταβλητών (π.χ. σε *z-scores*) αποτελεί ένα συνηθισμένο προπαρασκευαστικό στάδιο μετασχηματισμού των δεδομένων πριν την εφαρμογή μιας μεθόδου ταξινόμησης (Manly, 1994). Με την τυποποίηση των μεταβλητών επιτυγχάνονται δύο στόχοι: α) η απαλοιφή της επίδρασης των μονάδων μέτρησης των μεταβλητών στα αποτελέσματα και β) η απόδοση ίδιου βάρους στις μεταβλητές που συμμετέχουν στην ταξινόμηση (Williams, 1971). Όμως, σε ορισμένες περιπτώσεις, η τυποποίηση των μεταβλητών οδηγεί σε λανθασμένες τυπολογίες και ταξινομήσεις (Aldenderfer & Blashfield 1984, Kaufman & Rousseeuw 1990, Everitt 1993, Manly 1994, Hair *et al*. 1995, Johnson 1998, Hair & Black 2000). Δυστυχώς, δεν υπάρχει ένας γενικός κανόνας τυποποίησης, ο οποίος να οδηγεί στη βέλτιστη λύση όλων των προβλημάτων. Το ζήτημα αντιμετωπίζεται τοπικά ανάλογα με τους στόχους της εκάστοτε μελέτης και την αλληλεπίδραση της μεθόδου ταξινόμησης με το είδος των δεδομένων (βλέπε Williams 1971, Stoddard 1979, Schaffer & Green 1996, Cao *et al*. 1999).

Στον αντίποδα του προηγούμενου προβληματισμού βρίσκεται η έννοια της "σημαντικότητας" των μεταβλητών. Ο όρος σημαντικότητα χρησιμοποιείται για να

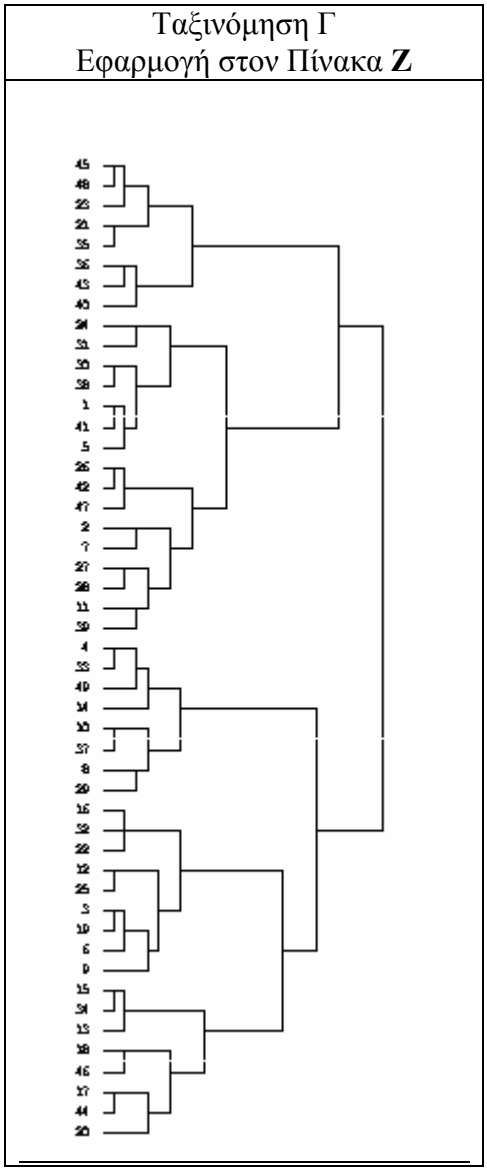
δηλώσει τη βαρύτητα που πρέπει να έχουν κάποιες μεταβλητές, ώστε να επηρεάσουν τα αποτελέσματα της ταξινόμησης (Morrison 1967, Williams 1971, Makarenkov & Legendre 2001). Σε πολλές περιπτώσεις τα βάρη των μεταβλητών καθορίζονται υποκειμενικά, από τους ερευνητές, και εισάγονται ως συντελεστές βαρύτητας στη μαθηματική έκφραση της απόστασης μεταξύ των αντικειμένων (Morrison, 1967). Ο προβληματισμός σχετικά με τη σημαντικότητα των μεταβλητών γίνεται πιο έντονος στην περίπτωση κατά την οποία ως μεταβλητές στην ταξινόμηση χρησιμοποιηθούν οι συντεταγμένες των προβολών των αντικειμένων σε παραγοντικούς άξονες, οι οποίοι προκύπτουν από την εφαρμογή της Ανάλυσης σε Κύριες Συνιστώσες (*Principal Component Analysis – PCA*) ή της ΠΑΑ στις αρχικές μεταβλητές. Οι δύο αυτές διαδικασίες μπορούν να παράγουν δύο τύπους συντεταγμένων: τις τυποποιημένες και τις κύριες (Greenacre 1993α, Gifi 1996, Johnson 1998). Οι τ.σ. των αντικειμένων, εκ κατασκευής, έχουν μέση τιμή 0 και διασπορά ίση με 1, ενώ οι κ.σ. έχουν, επίσης, μέση τιμή 0 αλλά διασπορά ίση με την αδράνεια του αντίστοιχου παραγοντικού άξονα. Το ερώτημα που προκύπτει είναι το κατά πόσο είναι σκόπιμο η ταξινόμηση των αντικειμένων να πραγματοποιείται με τις τ.σ. στους παραγοντικούς άξονες, με δεδομένη τη διαφορετική βαρύτητα, τουλάχιστον στην ερμηνεία, που έχει ο κάθε άξονας με βάση το ποσοστό της ολικής διακύμανσης (αδράνειας) που ερμηνεύει. Η χρήση των τ.σ. σε ρόλο μεταβλητών έχει ως αποτέλεσμα να συμμετέχουν στην ταξινόμηση με το ίδιο βάρος, γεγονός το οποίο έρχεται σε αντίθεση με τη διαφορετική σημαντικότητα-βαρύτητα που έχει ο κάθε άξονας στην ερμηνεία των αποτελεσμάτων. Χαρακτηριστικά αναφέρουμε ότι στο στατιστικό πακέτο SPSS οι συντεταγμένες των αντικειμένων επί των παραγοντικών αξόνων, οι οποίες υπολογίζονται από τις διαδικασίες *PCA* και Ανάλυση Ομοιογένειας (*HOMALS*), είναι οι τυποποιημένες (βλέπε Ενότητα 2.3.4). Έτσι, αν δεν υπάρχει κάποιο *a priori* ή *post-hoc* σύστημα (θεωρητικό, μεθοδολογικό) ανάθεσης βάρους στους παραγοντικούς άξονες, μια φυσική επιλογή θα μπορούσε να είναι η κανονικοποίηση των συντεταγμένων των αντικειμένων με τέτοιο τρόπο, ώστε η συμμετοχή των παραγοντικών αξόνων στην ταξινόμηση να γίνεται με βάρος ίσο με την αδράνειά τους. Με άλλα λόγια, αντί των τυποποιημένων να χρησιμοποιηθούν οι κύριες. Με τον αλγόριθμο, που προτείναμε στα προηγούμενα, οι παράγοντες (συντεταγμένες των αντικειμένων), οι οποίοι προκύπτουν από την εφαρμογή της ΠΑΑ, μπορούν να εισαχθούν, ως νέες σύνθετες μεταβλητές, σε ανάλυση Ταξινόμησης των αντικειμένων με βάρος ίσο με την αδράνεια τους.

Είδαμε στο Κεφάλαιο 2 ότι η εφαρμογή της ΠΑΑ σε πίνακα συμπτώσεων έχει ως αποτέλεσμα την προβολή των αντικειμένων (γραμμών) και των αντίστοιχων ιδιοτήτων τους (στηλών) σε ένα κοινό χώρο, συνήθως μικρότερης διάστασης απ' ότι ο χώρος των αρχικών δεδομένων, ώστε οι κατά Benzécri χ^2 αποστάσεις μεταξύ των αντικειμένων ή των ιδιοτήτων να προσεγγίζονται από ευκλείδειες. Όταν πρόκειται να αναλυθούν q κατηγορικές μεταβλητές που έχουν j σε πλήθος κλάσεις-ιδιότητες, τότε ο μέγιστος αριθμός παραγοντικών αξόνων, στους οποίους μπορεί να αναλυθεί η αδράνεια του πίνακα συμπτώσεων, είναι ίσος με $p=j-q$. Σε περίπτωση που χρησιμοποιηθούν όλες οι δυνατές διαστάσεις (άξονες) της λύσης που προκύπτει από την ΠΑΑ, τότε το τετράγωνο της απόστασης χ^2 μεταξύ δύο αντικειμένων είναι ίσο με το τετράγωνο της ευκλείδειας απόστασής τους στο χώρο των p διαστάσεων. Η ισότητα αυτή είναι εφικτή μόνο όταν χρησιμοποιηθούν οι κ.σ. των αντικειμένων στους p παραγοντικούς άξονες (Greenacre, 1993α και 1984). Έτσι, οδηγούμαστε στην ανάλυση πινάκων διαστάσεων $N\times p$ με διάσταση στηλών αρκετά μικρότερη απ' ότι του λογικού πίνακα **Z**, αφού $p<j$. Για να διαπιστώσουμε εμπειρικά τα προηγούμενα, πραγματοποιήσαμε στα δεδομένα του Πίνακα Γ1.1 (βλέπε Ενότητα Γ1 του Παραρτήματος Γ) δύο ταξινομήσεις των αντικειμένων με το SPSS: μία με βάση τις τυποποιημένες και μία άλλη με τις κύριες συντεταγμένες τους επί των παραγοντικών αξόνων, οι οποίοι προέκυψαν από την εφαρμογή της ΠΑΑ σύμφωνα με τη μεθοδολογία της Ενότητας 3.3.1. Στις αναλύσεις χρησιμοποιήθηκε ως απόσταση το τετράγωνο της ευκλείδειας και οι συστάδες σχηματίστηκαν σύμφωνα με το κριτήριο του *Ward* (Ward, 1963). Συγκρίναμε, στη συνέχεια, τα αποτελέσματα των δύο αναλύσεων με τα αποτελέσματα της ταξινόμησης στον αντίστοιχο πίνακα **Z**. Στην περίπτωση αυτή χρησιμοποιήσαμε ως απόσταση τη χ^2 και το κριτήριο του *Ward* (Lebart, Morineau & Warwick 1984, Benzécri 1992, Lebart 1994, Καραπιστόλης 1999, Lebart, Marineau & Piron 2000, Παπαδημητρίου 2006, 2004 και 1994). Η ανάλυση έγινε με το λογισμικό MAD (Καραπιστόλης, 2002). Τα αντίστοιχα δενδρογράμματα παρουσιάζονται στα Διαγράμματα 3.1 και 3.2. Ένα άτομο εξαιρέθηκε από την ανάλυση διότι θεωρήθηκε ως έκτοπο (*outlier*). Τελικά, ο πίνακας δεδομένων **Δ** που αναλύθηκε ήταν διαστάσεων 49×6. Για το συγκεκριμένο παράδειγμα έχουμε: N=49, q=6, j=16, p=10, ο πίνακας **B** (*Burt*) είναι διαστάσεων 16×16, ενώ ο **Z** είναι πίνακας 49×16. Η ταξινόμηση των αντικειμένων ως προς τις κύριες και τυποποιημένες συντεταγμένες τους έγινε σε πίνακες διαστάσεων 49×10. Η

σύγκριση των δενδρογραμμάτων οδηγεί στο συμπέρασμα ότι η εφαρμογή της ταξινόμησης, με απόσταση το τετράγωνο της ευκλείδειας, στις κ.σ. των αντικειμένων (Ταξινόμηση Β), δίνει όμοια αποτελέσματα (δομή και ιεραρχία) με την εφαρμογή της ταξινόμησης στον πίνακα **Z** με απόσταση τη χ^2 (Ταξινόμηση Γ).

Ταξινόμηση Α με Τυποποιημένες Συντεταγμένες	Ταξινόμηση Β με Κύριες Συντεταγμένες
	

Διάγραμμα 3.1: Αποτελέσματα Ταξινομήσεων: Δενδρογράμματα



Διάγραμμα 3.2: Δενδρόγραμμα Ταξινόμησης Αντικειμένων στο Λογικό Πίνακα **Z**

3.3.2.2 Εφαρμογή σε Μεγάλα Σύνολα Δεδομένων

Για να ελέγξουμε την αποτελεσματικότητα του προτεινόμενου αλγόριθμου σε μεγάλα σύνολα δεδομένων πραγματοποιήσαμε μια σειρά πειραμάτων με δεδομένα τα οποία προέκυψαν από προσομοίωση. Κατασκευάσαμε 10.000 τεχνητά σύνολα δεδομένων για πέντε ομάδες πινάκων της μορφής «*αντικείμενα* × *μεταβλητές*». Για κάθε ομάδα δημιουργήσαμε 2.000 πίνακες δεδομένων με διαφορετικό αριθμό γραμμών (αντικειμένων). Πιο συγκεκριμένα, η πρώτη ομάδα περιλάμβανε πίνακες με 100.000 γραμμές, η δεύτερη με 200.000, η τρίτη με 300.000, η τέταρτη με 400.000 και η πέμπτη με 500.000. Σε κάθε περίπτωση, διατηρήσαμε σταθερό τον αριθμό των

μεταβλητών, ίσο με 15, και το συνολικό πλήθος των κατηγοριών τους, ίσο με 120. Οι κατηγορίες των μεταβλητών κυμαίνονταν από 2 έως 6 με περιθώριες απόλυτες συχνότητες από την Ομοιόμορφη Κατανομή. Σε κάθε σύνολο δεδομένων εφαρμόσαμε τον προτεινόμενο αλγόριθμο και τον "παραδοσιακό" που περιγράψαμε στην Ενότητα 2.2.14. Η αποτελεσματικότητα των δύο προσεγγίσεων αξιολογήθηκε με βάση το χρόνο εκτέλεσης, σε *sec*, στον Η/Υ. Η χρονομέτρηση άρχιζε από το στάδιο κατασκευής των πινάκων **Z** και τελείωνε με τον υπολογισμό των κύριων συντεταγμένων των αντικειμένων στους p άξονες (p=120-15=105). Τα πειράματα πραγματοποιήθηκαν σε Η/Υ με επεξεργαστή Pentium(R) 4, CPU 2.8 GHz και 1GB RAM. Τα τεχνητά δεδομένα και οι αναλύσεις των πειραμάτων υλοποιήθηκαν με τη συναρτησιακή γλώσσα προγραμματισμού MATLAB, έκδοση 6.5 R13. Η αποτελεσματικότητα των δύο αλγόριθμων μετρήθηκε με το δείκτη:

$$E = \frac{(\overline{t}_1 - \overline{t}_2)}{\overline{t}_1} \times 100 ,$$

όπου $\overline{t}_1$ και $\overline{t}_2$ είναι οι μέσοι χρόνοι εκτέλεσης (για 2.000 μετρήσεις) του παραδοσιακού και του προτεινόμενου αλγόριθμου αντίστοιχα.

Ο δείκτης E εκφράζει τη διαφορά των μέσων χρόνων εκτέλεσης των δύο αλγόριθμων ως ποσοστό (%) του μέσου χρόνου εκτέλεσης του παραδοσιακού.

Για κάθε μία από τις πέντε ομάδες συνόλων δεδομένων υπολογίστηκαν οι μέσοι χρόνοι εκτέλεσης των δύο αλγόριθμων και οι αντίστοιχοι δείκτες αποτελεσματικότητας E. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 3.1.

Πίνακας 3.1: Αποτελεσματικότητα των Δύο Αλγόριθμων

	Ομάδες Πινάκων Δεδομένων- Διαστάσεις Λογικού Πίνακα				
	100.000 ×120	200.000 ×120	300.000 ×120	400.000 ×120	500.000 ×120
$\overline{t}_1$ (*sec*)	9,59	22,13	32,13	47,65	*
$\overline{t}_2$ (*sec*)	6,00	13,79	19,15	27,32	36,21
E (%)	37,35	37,66	40,39	42,66	-

*Δεν ήταν εφικτός ο υπολογισμός του πίνακα $\mathbf{S}_\mathbf{Z}$.

Καταρχήν παρατηρούμε ότι για την πέμπτη ομάδα δεδομένων δεν ήταν εφικτός ο υπολογισμός του πίνακα $\mathbf{S_Z}$ με αποτέλεσμα να μην μπορεί να εφαρμοστεί ο παραδοσιακός αλγόριθμος. Αυτό αποτελεί και τη βασικότερη ένδειξη της αποτελεσματικότητας της προτεινόμενης μεθοδολογίας. Επομένως, η εφαρμογή της ΠΑΑ απευθείας σε λογικούς πίνακες της τάξης των 500.000 αντικειμένων, τα οποία χαρακτηρίζονται από ένα πλήθος 120 ιδιοτήτων, δεν ήταν εφικτό να υλοποιηθεί στο πλαίσιο των συνθηκών και των μέσων του πειραματισμού. Αντίθετα, δεν παρατηρήθηκε τέτοιο πρόβλημα εφαρμόζοντας τον προτεινόμενο αλγόριθμο. Επομένως, οι παραγοντικές συντεταγμένες των αντικειμένων μπορούν να υπολογιστούν και, στη συνέχεια, να χρησιμοποιηθούν σε άλλες αναλύσεις, όπως είναι η Ταξινόμηση (βλέπε προηγούμενη ενότητα). Κατά την εφαρμογή του παραδοσιακού αλγόριθμου οι πίνακες $\mathbf{S_Z}$, που αναλύονται μέσω της SVD, έχουν μεταβλητό πλήθος γραμμών, ενώ οι πίνακες $\mathbf{S_B}$, οι οποίοι χρησιμοποιούνται στον προτεινόμενο, είναι πάντα 120×120. Για να γίνει περισσότερο αντιληπτή η διαφορά των δύο προσεγγίσεων δίνουμε το ακόλουθο ακραίο παράδειγμα: Έστω ότι σε μια επιδημιολογική μελέτη έχουν συγκεντρωθεί στοιχεία για 50.000 άτομα σε σχέση με την εμφάνιση ή όχι 5 κλινικών σημείων. Στην περίπτωση αυτή ο πίνακας $\mathbf{S_Z}$ είναι διαστάσεων 50.000×10, ενώ ο $\mathbf{S_B}$ είναι πίνακας 10×10.

Από τον Πίνακα 3.1 διαπιστώνουμε την υπεροχή του προτεινόμενου αλγόριθμου και στις υπόλοιπες τέσσερις ομάδες δεδομένων με αριθμό γραμμών έως 400.000. Μάλιστα παρατηρούμε αυξητική τάση του δείκτη *E* καθώς το πλήθος γραμμών των πινάκων γίνεται όλο και μεγαλύτερο. Ο προτεινόμενος αλγόριθμος φαίνεται να είναι περισσότερο αποτελεσματικός όταν εφαρμόζεται σε μεγαλύτερα σύνολα δεδομένων.

Αξίζει να παρατηρήσουμε ότι πειραματιστήκαμε και με το SPSS στο οποίο εφαρμόσαμε τις διαδικασίες *HOMALS*, στην έκδοση 11.5 του πακέτου, και *Multiple Correspondence Analysis*, στην έκδοση 13. Κατασκευάσαμε τυχαία σύνολα δεδομένων με 60.000 αντικείμενα και 12 μεταβλητές. Οι αντίστοιχες ιδιότητες ήταν συνολικά 60. Αν και δεν μπορούμε να συγκρίνουμε άμεσα την αποτελεσματικότητα του αλγόριθμου ALS με αυτή της προτεινόμενης μεθοδολογίας, ενδεικτικά αναφέρουμε ότι η διαδικασία *HOMALS* δεν ήταν δυνατό να υλοποιηθεί, ενώ η

Multiple Correspondence Analysis ολοκληρώθηκε σε 64 δευτερόλεπτα κατά μέσο όρο.

3.3.3 Συμπεράσματα και Σχόλια

Στην περίπτωση πολλών μεταβλητών, η ΠΑΑ εφαρμόζεται συνήθως στο γενικευμένο πίνακα συμπτώσεων *Burt*, με αποτέλεσμα να μην είναι εφικτός ο άμεσος υπολογισμός των συντεταγμένων των αντικειμένων επί των παραγοντικών αξόνων, ο οποίος, όμως, είναι εφικτός, αν η ανάλυση εφαρμοστεί στον αντίστοιχο λογικό πίνακα 0-1. Η εφαρμογή της ΠΑΑ στον πίνακα *Burt* **B** απαιτεί μικρότερο υπολογιστικό έργο (λόγω μικρότερων διαστάσεων του πίνακα), αλλά δεν μπορεί να οδηγήσει σε άμεση ανασύσταση του αρχικού πίνακα δεδομένων **Δ**. Αντίθετα, η εφαρμογή της ΠΑΑ στον **Z** απαιτεί μεγαλύτερο υπολογιστικό έργο, αλλά μπορεί να γίνει η ανασύσταση του **Δ**. Συνεπώς, με τη μεθοδολογία που προτείνουμε εξισορροπούνται οι δύο περιπτώσεις.

Τα αποτελέσματα των πειραματισμών επιβεβαιώνουν την αποτελεσματικότητα του προτεινόμενου αλγόριθμου στην ανάλυση μεγάλων πινάκων δεδομένων με πιθανές εφαρμογές, για παράδειγμα, στις τεχνικές «Εξόρυξης Δεδομένων» (*Data Mining*) (βλέπε Βαζιργιάννης & Χαλκίδη 2003, Dunham 2004) και στις επιδημιολογικές έρευνες μεγάλης κλίμακας (βλέπε Lindelöf *et al*. 1991, Kroke *et al*. 2001, Magnussen 2003, Manfredi 2004).

Η ταξινόμηση των αντικειμένων, ως προς τις κύριες συντεταγμένες τους, οδηγεί στην επεξεργασία πινάκων μικρότερων διαστάσεων απ' ότι στην περίπτωση της ανάλυσης του **Z**, με ταυτόσημα αποτελέσματα. Με την προτεινόμενη μεθοδολογία, οι παραγοντικοί άξονες, ως νέες σύνθετες μεταβλητές, κανονικοποιούνται με τρόπο, ώστε να συμμετέχουν στην ταξινόμηση με βάρος ίσο με την αδράνειά τους. Βέβαια, αυτό είναι αληθινό μόνο στην περίπτωση που οι αναλύσεις και τα κριτήρια που αναφέρθηκαν (ΠΑΑ, Ιεραρχική Ταξινόμηση, χ^2 απόσταση, ευκλείδεια απόσταση και κριτήριο του *Ward*) κριθούν ως κατάλληλα για την ανάλυση ενός μεγάλου συνόλου κατηγορικών δεδομένων. Τέλος, να σημειώσουμε ότι το πλήθος των υπολογισμών μπορεί να μειωθεί περισσότερο στην περίπτωση που ο χρήστης εφαρμόσει την ανάλυση Ταξινόμησης όχι στους p παράγοντες αλλά στους t πρώτους (με $t<p$).

ΚΕΦΑΛΑΙΟ 4

Σχέσεις Αδράνειας σε Πίνακες Συμπτώσεων, Γενικευμένους και Λογικούς Δύο ή Περισσότερων Μεταβλητών

4.1 Εισαγωγή

Στην περίπτωση δύο μεταβλητών η ΠΑΑ μπορεί να εφαρμοστεί σε τρεις τουλάχιστον πίνακες συμπτώσεων: α) στον απλό πίνακα συμπτώσεων **F**, β) στον αντίστοιχο πίνακα σχεδιασμού **Z** ή, αλλιώς, πίνακα λογικής περιγραφής 0-1, και γ) στον αντίστοιχο γενικευμένο πίνακα συμπτώσεων **B** (πίνακας *Burt*). Όπως είδαμε στο Κεφάλαιο 3, στην πολυμεταβλητή περίπτωση η ΠΑΑ εφαρμόζεται συνήθως στον πίνακα **B**. Τόσο στη διμεταβλητή όσο και στην πολυμεταβλητή εκδοχή της ΠΑΑ, η εικόνα επί των παραγοντικών επιπέδων του φαινομένου που εξετάζεται είναι η ίδια, ανεξάρτητα από τον πίνακα δεδομένων, στον οποίο θα εφαρμοστεί η ανάλυση (Lebart, Morineau & Tabard 1977, Greenacre 1984, Israëls 1987, Gifi 1996, Lebart, Morineau & Piron 2000, Μάρκος & Παπαδημητρίου 2003). Η ισοδυναμία των πινάκων **Z** και **B** έχει ήδη εξηγηθεί στην Ενότητα 2.3.3.5. Όμως, η ολική αδράνεια και η αδράνεια που ερμηνεύει κάθε παραγοντικός άξονας είναι διαφορετικές, ανάλογα με τον πίνακα, στον οποίο θα εφαρμοστεί η ανάλυση. Αυτό έχει ως αποτέλεσμα το ποσοστό της ολικής αδράνειας που ερμηνεύουν, για παράδειγμα, οι δύο πρώτοι παραγοντικοί άξονες να αποτελεί άλλοτε “φτωχότερη” και άλλοτε “πλουσιότερη” ένδειξη προσαρμογής των δεδομένων και της πληροφορίας που αναλύεται κάθε φορά (Greenacre 2005 και 1993α, Μάρκος & Παπαδημητρίου 2003). Θα πρέπει να τονιστεί ότι, κατά παράδοση, σε στατιστικές διαδικασίες, όπως είναι η Γραμμική Παλινδρόμηση και η Ανάλυση σε Κύριες Συνιστώσες, επιβάλλονται, μάλλον αυθαίρετα, όρια στο ποσοστό διακύμανσης (συνήθως ≥60%), το οποίο θα

“πρέπει” να ερμηνεύεται κάθε φορά, ώστε τα αποτελέσματα των αναλύσεων να είναι κατάλληλα για την εξαγωγή συμπερασμάτων. Έτσι, στην περίπτωση που κατά την εφαρμογή της ΠΑΑ οι δύο πρώτοι παραγοντικοί άξονες ερμηνεύουν, για παράδειγμα, το 27% της ολικής αδράνειας, η τιμή αυτή θα μπορούσε να θεωρηθεί ως δείκτης κακής προσαρμογής των δεδομένων και ποιότητας της πληροφορίας. Ένας τέτοιος ισχυρισμός μάλλον θα ήταν βιαστικός, αν προηγουμένως δεν έχει ληφθεί υπόψη το μέγεθος του υπό ανάλυση πίνακα δεδομένων (Μάρκος & Παπαδημητρίου, 2003). Ενδεικτικά αναφέρουμε ότι για τα δεδομένα του συνόλου Α (βλέπε αρχείο data_files.xls, στο Παράρτημα CDA του CD που συνοδεύει τη διατριβή), η εφαρμογή της ΠΑΑ στους πίνακες **F**, **Z** και **B**, οι οποίοι περιγράφουν, για 138 φοιτητές, τη σχέση των μεταβλητών «είδος διακοπών φοιτητή» και «επάγγελμα του πατέρα» με 6 και 7 κατηγορίες αντίστοιχα, έδωσε τα παρακάτω αποτελέσματα:

Οι δύο πρώτοι άξονες ερμηνεύουν:

- Το 30,8% της ολικής αδράνειας, από την ανάλυση του 138×13 πίνακα **Z** (p=11).
- Το 43,9% της ολικής αδράνειας, στην περίπτωση του 13×13 πίνακα *Burt* (p=11).
- Το 93,2% της ολικής αδράνειας, κατά την επεξεργασία του 6×7 πίνακα **F** (p=5).

Και στις τρεις περιπτώσεις, οι σχετικές θέσεις των σημείων επί των παραγοντικών επιπέδων είναι οι ίδιες (Μάρκος & Παπαδημητρίου, 2003). Όμως, τα ποσοστά ερμηνείας των παραγοντικών αξόνων είναι διαφορετικά καθώς και η “εντύπωση” που δημιουργούν για την ποιότητα της λύσης της ΠΑΑ. Γενικά, τα ποσοστά ερμηνείας εξαρτώνται από το μέγεθος του πίνακα που αναλύεται και φυσικά από τις δομές που ενθυλακώνουν τα ίδια τα δεδομένα. Στο συγκεκριμένο παράδειγμα, οι τρεις πίνακες περιγράφουν το ίδιο φαινόμενο. Επομένως, η διαφοροποίηση οφείλεται μόνο στο μέγεθος των πινάκων που αναλύονται και συνεκδοχικά στο μέγιστο αριθμό p των παραγοντικών αξόνων, οι οποίοι μπορούν να προκύψουν από την εφαρμογή της ΠΑΑ σε κάθε περίπτωση.

Αν θεωρήσουμε την ολική αδράνεια ως ένα μέτρο της πληροφορίας που περιέχεται στον πίνακα ή ως ένα δείκτη μεγέθους του αποτελέσματος (*effect size*), τότε είναι χρήσιμο να εξετάσουμε τις σχέσεις που συνδέουν τις ολικές αδράνειες των πινάκων, στους οποίους μπορεί να εφαρμοστεί η ΠΑΑ. Οι σχέσεις αυτές αναδεικνύουν κυρίως

την ποιότητα της πληροφορίας που παράγεται από την ΠΑΑ, όπως, για παράδειγμα, τη φυσική ερμηνεία της ολικής αδράνειας ανάλογα με τον πίνακα που αναλύεται κάθε φορά και την ανάγκη για κάποιου είδους διόρθωση ή τροποποίηση των βασικών αριθμητικών αποτελεσμάτων της μεθόδου. Επίσης, μπορούν να χρησιμοποιηθούν για την εξεύρεση πρακτικών κριτηρίων σχετικά με την επιλογή (υπό)πινάκων που θα συμπεριληφθούν στην ανάλυση. Έτσι, στις επόμενες ενότητες:

α) Αρχικά, εξετάζουμε τις μαθηματικές σχέσεις που συνδέουν τις ολικές αδράνειες των τριών πινάκων δεδομένων στην περίπτωση δύο μεταβλητών (Ενότητες 4.2 έως 4.4) και, στη συνέχεια, τις αντίστοιχες γενικεύσεις των σχέσεων στην περίπτωση περισσότερων μεταβλητών (Ενότητα 4.6). Το ενδιαφέρον, στην προσέγγισή μας, είναι ότι οι αποδείξεις των σχέσεων δεν στηρίζονται στην εφαρμογή της συνηθισμένης στο χώρο άλγεβρας πινάκων, αλλά κυρίως στις εξής επισημάνσεις:

1. Κατά το σχεδιασμό μιας μελέτης, η πειραματική ή δειγματοληπτική μονάδα άλλες φορές ταυτίζεται με τη μονάδα παρατήρησης και άλλες φορές όχι (Kirk, 1995). Για παράδειγμα: Σε μια μελέτη εκπαιδευτικής έρευνας η πειραματική ή δειγματοληπτική μονάδα μπορεί να είναι η τάξη, αλλά μονάδα παρατήρησης οι μαθητές της τάξης.

2. Στην πράξη, εκτός από τους πίνακες **F**, **Z** και **B**, η ΠΑΑ εφαρμόζεται και σε άλλους τύπους πινάκων δεδομένων που πληρούν τις προϋποθέσεις εφαρμογής της (βλέπε Ενότητα 2.4). Χαρακτηριστικά αναφέρουμε τους πίνακες τύπου “στοίβας” και “φέτας», όπου πίνακες συμπτώσεων τοποθετούνται είτε ο ένας κάτω από τον άλλο είτε ο ένας δίπλα στον άλλο (βλέπε Ενότητα 2.4.1). Σε κάθε περίπτωση, ο αλγόριθμος της ΠΑΑ εφαρμόζεται με τον ίδιο τρόπο (βλέπε Ενότητα 2.4). Οι πίνακες αντιμετωπίζονται, τελικά, ως απλοί πίνακες συμπτώσεων δύο μεταβλητών, οι οποίοι έχουν τη γενική μορφή «*αντικείμενα* × *ιδιότητες*» (Benzécri & *Collaborateurs* 1973, Israëls 1987, Weller & Romney 1990, Andersen 1991, Benzécri 1992, Greenacre 1993α και 1984, Gifi 1996, Clausen 1998, Escofier & Pagès 1998, Lebart, Marineau & Piron 2000). Έτσι, η ολική αδράνεια I του πίνακα δεδομένων, ο οποίος αναλύεται κάθε φορά, δίνεται υπολογιστικά από τη σχέση:

$$I = \frac{Q}{N},$$

όπου Q είναι το στατιστικό χ^2 που αντιστοιχεί στον πίνακα που αναλύεται και N είναι το πλήθος των μονάδων παρατήρησης.

Όπως θα φανεί στη συνέχεια, ο ρόλος και το πλήθος των μονάδων παρατήρησης μεταβάλλονται κατάλληλα ανάλογα με τον τύπο του πίνακα δεδομένων.

β) Στην Ενότητα 4.7 εισάγουμε την έννοια της «Ενδιαφέρουσας Αδράνειας», κατά την εφαρμογή της πολυμεταβλητής εκδοχής της ΠΑΑ στον πίνακα *Burt*, βάσει της οποίας προτείνουμε μέθοδο εντοπισμού υποπίνακα του *Burt*, ο οποίος περιλαμβάνει όλες τις μεταβλητές που συμμετέχουν στην ανάλυση και η εφαρμογή της ΠΑΑ σε αυτόν αποδίδει την "πλησιέστερη εικόνα" του φαινομένου σε αυτή που προκύπτει από την εφαρμογή της μεθόδου στον αρχικό πίνακα *Burt* (Ενότητα 4.8.1).

γ) Προτείνουμε έλεγχο της στατιστικής σημαντικότητας της ενδιαφέρουσας αδράνειας (Ενότητα 4.8.2).

δ) Με αφετηρία τη φυσική ερμηνεία της ενδιαφέρουσας αδράνειας, προτείνουμε στην Ενότητα 4.8.3 μέθοδο διόρθωσης των αδρανειών των παραγοντικών αξόνων, ώστε τα αντίστοιχα ποσοστά ερμηνείας να αντικατοπτρίζουν την πραγματική ποιότητα της λύσης της ΠΑΑ.

4.2 Περίπτωση 1η: Απλός Πίνακας Συμπτώσεων Δύο Μεταβλητών

Έστω $\mathbf{F}_{k\times l}$ ο απλός πίνακας συμπτώσεων απολύτων συχνοτήτων δύο κατηγορικών μεταβλητών X και Y. Στην περίπτωση αυτή θεωρούμε ότι ο αντίστοιχος σχεδιασμός της μελέτης περιλαμβάνει ως πειραματική ή δειγματοληπτική μονάδα το αντικείμενο και μάλιστα η πειραματική μονάδα και η μονάδα παρατήρησης ταυτίζονται.

Πίνακας 4.1: Ο Απλός Πίνακας Συμπτώσεων **F** με Περιθώρια Στήλη και Γραμμή

		Μεταβλητή Y Κλάσεις ή ιδιότητες της Y				
	Μεταβλητή X	1	2	…	l	Άθροισμα
Κλάσεις ή ιδιότητες της X	1	f_{11}	f_{12}	…	f_{1l}	f_{1+}
	2	f_{21}	f_{22}	…	f_{2l}	f_{2+}
	⋮	⋮	⋮	⋮	⋮	⋮
	k	f_{k1}	f_{k2}	...	f_{kl}	f_{k+}
	Άθροισμα	f_{+1}	f_{+2}	…	f_{+l}	N

Γνωρίζουμε ότι η ολική αδράνεια I_F του πίνακα **F** συνδέεται με τις χαρακτηριστικές τιμές και ιδιοτιμές των πινάκων, οι οποίοι διαγωνοποιούνται κατά την εφαρμογή της ΠΑΑ, μέσω των παρακάτω σχέσεων (βλέπε Δεύτερο Βήμα της Ενότητας 2.2.14):

$$I_F = \sum_{s=1}^{p} d_s^2 \text{ ή } I_F = \sum_{s=1}^{p} \lambda_s,$$

όπου d_i είναι οι χαρακτηριστικές τιμές του πίνακα **S**, λ_i είναι οι ιδιοτιμές του πίνακα $\mathbf{S^T S}$, $d_s^2 = \lambda_s$ και $p=min\{k\text{-}1, l\text{-}1\}$.

Υπολογιστικά η I_F δίνεται και από την παρακάτω σχέση (βλέπε Ενότητα 2.2.5):

$$I_F = \frac{Q}{N}, \qquad [4.1]$$

όπου N είναι το πλήθος των μονάδων παρατήρησης και Q το στατιστικό χ^2 που υπολογίζεται, με βάση τους συμβολισμούς του Πίνακα 4.1, από τη σχέση:

$$Q = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(f_{ij} - \frac{f_{i+} f_{+j}}{N} \right)^2}{\frac{f_{i+} f_{+j}}{N}}. \qquad [4.2]$$

Παρατήρηση 4.1

Από τη σχέση [4.1] φαίνεται ότι η ολική αδράνεια του πίνακα **F** εκφράζει τη μέση απόκλιση ανά μονάδα παρατήρησης από την κατάσταση ανεξαρτησίας των δύο μεταβλητών X και Y, όπως αυτή μετριέται μέσω του στατιστικού χ^2.

4.3 Περίπτωση 2^{η}: Λογικός Πίνακας 0-1 Δύο Μεταβλητών

Έστω $\mathbf{Z}_{N\times(k+l)}$ ο αντίστοιχος λογικός πίνακας του **F** με στοιχεία 0 ή 1. Το σύνολο κάθε γραμμής του **Z** είναι ίσο με 2 και η περιθώρια γραμμή αποτελεί την ένωση της περιθώριας γραμμής και της περιθώριας στήλης του πίνακα **F** (βλέπε Πίνακα 4.2). Στην περίπτωση αυτή, μπορούμε να θεωρήσουμε ότι ο πίνακας **Z** δεν είναι παρά ένας "αραιός" αλλά απλός πίνακας συμπτώσεων δύο μεταβλητών. Η μεταβλητή που ορίζεται από το σύνολο των γραμμών αφορά τα αντικείμενα ή υποκείμενα της μελέτης και θεωρούμε ότι έχει *N* κλάσεις, ενώ η μεταβλητή που ορίζεται από το σύνολο των στηλών αφορά στο προφίλ των γραμμών, ως προς τις εξεταζόμενες δύο αρχικές μεταβλητές *X* και *Y*, και θεωρούμε ότι έχει *k+l* κλάσεις ή ιδιότητες. Υποθέτουμε ότι η μεταβλητή των στηλών αποτελεί μια νέα μεταβλητή, έστω *E*, η οποία δημιουργείται από την ένωση των μεταβλητών *X* και *Y*. Στην περίπτωση αυτή, επιβάλουμε μια εννοιολογική αλλαγή στον πειραματικό σχεδιασμό της μελέτης. Θεωρούμε ότι το σχέδιο της μελέτης περιλαμβάνει ως πειραματική ή δειγματοληπτική μονάδα το αντικείμενο ή το υποκείμενο, ενώ ως μονάδα παρατήρησης την απόκρισή του (*response*) στη νέα μεταβλητή *E*. Με άλλα λόγια, για κάθε αντικείμενο ή υποκείμενο οι μονάδες παρατήρησης αντιστοιχούν στις ιδιότητες, από τις οποίες αυτό χαρακτηρίζεται. Ως ιδιότητες λαμβάνονται οι κατηγορίες της μεταβλητής *E*. Κάτω από την προϋπόθεση ότι δεν υπάρχουν ελλείπουσες τιμές, το σύνολο των μονάδων παρατήρησης είναι ίσο με 2*N*. Στο πλαίσιο αυτό και σύμφωνα με την Παρατήρηση 4.1, μπορούμε να υποθέσουμε ότι η ολική αδράνεια $I_{0\text{-}1}$ του πίνακα **Z** εκφράζει τη μέση απόκλιση ανά μονάδα παρατήρησης από την κατάσταση ανεξαρτησίας μεταξύ της μεταβλητής των γραμμών και της μεταβλητής των στηλών, όπως αυτή μετριέται μέσω του στατιστικού χ^2. Πιο συγκεκριμένα, βάσει της επισήμανσης (2) ισχύει:

$$I_{0-1} = \frac{Q_{0-1}}{2N}, \qquad [4.3]$$

όπου Q_{0-1} είναι το στατιστικό χ^2 που αντιστοιχεί στο λογικό πίνακα **Z**.

Πίνακας 4.2: Ο Λογικός Πίνακας **Z** με Περιθώρια Στήλη και Γραμμή

	Μεταβλητή X				Μεταβλητή Y				
	Κλάσεις ή ιδιότητες της X				Κλάσεις ή ιδιότητες της Y				
	1	2	…	k	1	2	…	l	
	Κλάσεις ή ιδιότητες της νέας μεταβλητής E								
Αντικείμενα ή Υποκείμενα	1	2	...	k	k+1	k+2	...	$k+l$	Άθροισμα
1	b_{11}	b_{12}	...	b_{1k}	$b_{1(k+1)}$	$b_{1(k+2)}$	...	$b_{1(k+l)}$	2
2	b_{21}	b_{22}	...	b_{2k}	$b_{2(k+1)}$	$b_{2(k+2)}$	...	$b_{2(k+l)}$	2
3	b_{31}	b_{32}	...	b_{3k}	$b_{3(k+1)}$	$b_{3(k+2)}$	...	$b_{3(k+l)}$	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
N	b_{N1}	b_{N2}	...	b_{Nk}	$b_{N(k+1)}$	$b_{N(k+2)}$	...	$b_{N(k+l)}$	2
Άθροισμα	t_1	t_2	...	t_k	t_{k+1}	t_{k+2}	...	t_{k+l}	$2N$
Άθροισμα	f_{1+}	f_{2+}	...	f_{k+}	f_{+1}	f_{+2}	...	f_{+l}	$2N$

Στη συνέχεια, χρησιμοποιώντας τους συμβολισμούς του Πίνακα 4.2, αποδεικνύουμε την παρακάτω πρόταση:

<u>Πρόταση 1</u>

Η ολική αδράνεια $I_{0\text{-}1}$ του λογικού πίνακα **Z**, στην περίπτωση δύο κατηγορικών μεταβλητών X και Y, δίνεται από την παρακάτω σχέση:

$$I_{0-1} = \frac{Q_{0-1}}{2N} = \frac{k+l}{2} - 1,$$

όπου Q_{0-1} είναι το στατιστικό χ^2 που αντιστοιχεί στον πίνακα **Z**, $2N$ είναι το πλήθος των μονάδων παρατήρησης, k είναι οι κλάσεις (ιδιότητες) της μεταβλητής X και l είναι οι κλάσεις (ιδιότητες) της μεταβλητής Y.

Προϋπόθεση: Δεν υπάρχουν ελλείπουσες τιμές και όλες οι γραμμές και οι στήλες του πίνακα **Z** είναι ενεργές για την εφαρμογή της ΠΑΑ.

Απόδειξη:

Υπολογίζουμε την ποσότητα Q_{0-1}, η οποία αντιστοιχεί στο λογικό πίνακα, με το συνήθη τρόπο (βλέπε Ενότητα 2.2.5):

$$Q_{0-1} = \sum\sum \frac{(O-E)^2}{E}, \qquad [4.4]$$

όπου O είναι η παρατηρούμενη και E η αναμενόμενη συχνότητα ενός κελιού.

Έχουμε:

$$Q_{0-1} = \sum_{i=1}^{N}\sum_{j=1}^{k+l} \frac{\left(b_{ij} - \frac{2t_j}{2N} \right)^2}{\frac{2t_j}{2N}} = \sum_{i=1}^{N}\sum_{j=1}^{k+l} \frac{\left[b_{ij}^2 - 2b_{ij}\frac{t_j}{N} + \left(\frac{t_j}{N} \right)^2 \right]}{\frac{t_j}{N}}.$$

Επειδή $\sum_{i=1}^{N}\sum_{j=1}^{k+l} b_{ij} = 2N$ και $\sum_{j=1}^{k+l} t_j = 2N$, προκύπτει ότι:

$$Q_{0-1} = \sum_{i=1}^{N}\sum_{j=1}^{k+l} \frac{b_{ij}^2}{\frac{t_j}{N}} - 2\sum_{i=1}^{N}\sum_{j=1}^{k+l} b_{ij} + \sum_{i=1}^{N}\sum_{j=1}^{k+l} \frac{t_j}{N} = N\sum_{i=1}^{N}\sum_{j=1}^{k+l} \frac{b_{ij}^2}{t_j} - 2\cdot 2N + \frac{N}{N}\cdot 2N.$$

Επιπλέον, επειδή $b_{ij}^2 = b_{ij}$ προκύπτει ότι:

$$Q_{0-1} = N\sum_{i=1}^{N}\sum_{j=1}^{k+l} \frac{b_{ij}^2}{t_j} - 4N + 2N = N\left(\sum_{i=1}^{N}\sum_{j=1}^{k+l} \frac{b_{ij}}{t_j} - 2 \right). \qquad [4.5]$$

Για το άθροισμα $\sum_{i=1}^{N}\sum_{j=1}^{k+l} \frac{b_{ij}}{t_j}$ ισχύει το εξής: από τους $N(k+l)$ όρους οι $2N$ είναι διάφοροι του μηδενός και οι υπόλοιποι είναι ίσοι με μηδέν. Οι $2N$ όροι που είναι διάφοροι του μηδενός είναι της μορφής $\frac{1}{t_j}$. Επομένως:

$$\sum_{i=1}^{N}\sum_{j=1}^{k+l} \frac{b_{ij}}{t_j} = \sum_{i=1}^{N}\sum_{j=1}^{k+l} \frac{1}{t_j} = \underbrace{t_1\frac{1}{t_1} + t_2\frac{1}{t_2} + \cdots + t_j\frac{1}{t_j}}_{k+l \text{ όροι}} = \underbrace{1+1+\cdots+1}_{k+l \text{ όροι}} = k+l. \qquad [4.6]$$

Η δεύτερη ισότητα στην παραπάνω σχέση προκύπτει από το γεγονός ότι ο όρος $\frac{1}{t_1}$ εμφανίζεται t_1 φορές, ο όρος $\frac{1}{t_2}$ εμφανίζεται t_2 φορές,… και ο όρος $\frac{1}{t_j}$ εμφανίζεται t_j φορές, με $\sum t_j = 2N$ για j=1, …, k+l.

Έτσι, από τις σχέσεις [4.5] και [4.6] έχουμε τελικά ότι:

$$Q_{0-1} = N\left(k+l-2\right).$$

Με βάση τη σχέση [4.3], η ολική αδράνεια του λογικού πίνακα είναι ίση με:

$$I_{0-1} = \frac{Q_{0-1}}{2N} = \frac{N\left(k+l-2\right)}{2N} = \frac{1}{2}\left(k+l-2\right) = \frac{k+l}{2} - 1.$$

Συνεπώς,

$$I_{0-1} = \frac{k+l}{2} - 1. \square \qquad [4.7]$$

Στο αποτέλεσμα αυτό έχουν καταλήξει και άλλοι ερευνητές με διαφορετική όμως μεθοδολογία (Lebart, Morineau & Tabard 1977, Greenacre 1984, Israëls 1987, Gifi 1996, Lebart, Morineau & Piron 2000). Στην προτεινόμενη προσέγγιση, είναι η εννοιολογική αλλαγή στον πειραματικό σχεδιασμό που επέτρεψε να καταλήξουμε στο ίδιο αποτέλεσμα. Η σχέση [4.7] αποτελεί ειδική περίπτωση της σχέσης [2.60] (βλέπε Ενότητα 2.3.3.1) για j=k+l και q=2.

Παρατήρηση 4.2

Από τη σχέση [4.7] διαπιστώνουμε ότι η ολική αδράνεια του πίνακα **Z** εξαρτάται μόνο από το συνολικό αριθμό κλάσεων ή ιδιοτήτων των δύο μεταβλητών και, φυσικά, από το πλήθος των μεταβλητών (ο αριθμός 2 που υπάρχει στον παρονομαστή του κλάσματος). Δεν φαίνεται να λαμβάνεται υπόψη η συνάφεια (συσχέτιση, αλληλεπίδραση) μεταξύ των δύο μεταβλητών ούτε και το πλήθος N των αντικειμένων της μελέτης. Τελικά, η φυσική ερμηνεία της ολικής αδράνειας, όπως αυτή υπολογίζεται από τη σχέση [4.7], είναι δύσκολο να δοθεί.

4.3.1 Δύο Ειδικοί Πίνακες

Θεωρούμε τον πίνακα **X**, στον οποίο οι κλάσεις ή ιδιότητες της μεταβλητής X διασταυρώνονται μεταξύ τους:

Πίνακας 4.3.1: Ο Πίνακας **X** με Περιθώρια Στήλη και Γραμμή

		Μεταβλητή X Κλάσεις ή ιδιότητες				
	Μεταβλητή X	1	2	…	k	Άθροισμα
Κλάσεις ή ιδιότητες	1	f_{11}	0	0	0	f_{1+}
	2	0	f_{22}	0	0	f_{2+}
	⋮	⋮	⋮	⋮	⋮	⋮
	k	0	0	0	f_{kk}	f_{k+}
	Άθροισμα	f_{+1}	f_{+2}	...	f_{+k}	N

Παρατηρούμε ότι ο $k \times k$ πίνακας **X** είναι διαγώνιος και ισχύει:

$$f_{ij} = f_{ii} = f_{jj} = f_{i+} = f_{+j}\text{, για } i=j \qquad [4.8]$$

Στη συνέχεια, χρησιμοποιώντας τους συμβολισμούς του Πίνακα 4.3.1, αποδεικνύουμε την παρακάτω πρόταση:

Πρόταση 2

Η ολική αδράνεια I_{XX} του πίνακα **X** δίνεται από την παρακάτω σχέση:

$$I_{XX} = \frac{Q}{N} = k - 1,$$

όπου Q είναι το στατιστικό χ^2 που αντιστοιχεί στον πίνακα **X**, N είναι το πλήθος των μονάδων παρατήρησης και k οι κλάσεις (ιδιότητες) της μεταβλητής X.

Απόδειξη:

Με συλλογισμό ανάλογο με αυτό που χρησιμοποιήσαμε για την απόδειξη της Πρότασης 1, υπολογίζουμε την ποσότητα χ^2 (Q) που αντιστοιχεί στον πίνακα **X**. Έχουμε:

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{\left(f_{ij} - \dfrac{f_{i+}\, f_{j+}}{N} \right)^2}{\dfrac{f_{i+}\, f_{j+}}{N}}.$$

Στο άθροισμα αυτό, από τους k^2 όρους οι k^2–k είναι ίσοι με 0 και οι υπόλοιποι k όροι διάφοροι του 0 και λόγω της σχέσης [4.8] το άθροισμα μπορεί να γραφεί ως εξής:

$$Q = \sum_{i=1}^{k} \frac{\left(f_{ii} - \dfrac{f_{ii}\, f_{ii}}{N} \right)^2}{\dfrac{f_{ii}\, f_{ii}}{N}} = \sum_{i=1}^{k} \frac{\left(f_{ii}^2 - 2\dfrac{f_{ii}^3}{N} + \dfrac{f_{ii}^2}{N^2} \right)}{\dfrac{f_{ii}^2}{N}} = \sum_{i=1}^{k} \left(N - 2\, f_{ii} + \frac{f_{ii}^2}{N} \right) =$$

$$= \sum_{i=1}^{k} N - 2\sum_{i=1}^{k} f_{ii} + \sum_{i=1}^{k} \frac{f_{ii}^2}{N} = Nk - 2N + N = N\left(k-1\right).$$

Επομένως, η ολική αδράνεια I_{XX} του πίνακα **X** δίνεται από τη σχέση:

$$I_{XX} = \frac{Q}{N} = \frac{N\left(k-1\right)}{N} = k-1. \square \qquad [4.9]$$

Ανάλογα εργαζόμενοι για τη μεταβλητή Y, μπορούμε να δείξουμε ότι η ολική αδράνεια του αντίστοιχου $l \times l$ διαγώνιου πίνακα **Y** (Πίνακας 4.3.2) είναι ίση με:

$$I_{YY} = l-1. \qquad [4.10]$$

Πίνακας 4.3.2: Ο Πίνακας **Y** με Περιθώρια Στήλη και Γραμμή

	Μεταβλητή Y Κλάσεις ή ιδιότητες					
	Μεταβλητή Y	1	2	…	l	Άθροισμα
Κλάσεις ή ιδιότητες	1	f_{11}	0	0	0	f_{1+}
	2	0	f_{22}	0	0	f_{2+}
	⋮	⋮	⋮	⋮	⋮	⋮
	l	0	0	0	f_{ll}	f_{l+}
	Άθροισμα	f_{+1}	f_{+2}	...	f_{+l}	N

Παρατήρηση 4.3

Από τις σχέσεις [4.7], [4.9] και [4.10] έχουμε:

$$I_{0-1} = \frac{k+l}{2} - 1 = \frac{k+l-2}{2} = \frac{k-1+l-1}{2} = \frac{I_{XX} + I_{YY}}{2}.$$

Συνεπώς,

$$I_{0-1} = \frac{I_{XX} + I_{YY}}{2}. \qquad [4.11]$$

Από τη σχέση [4.11] προκύπτει μια νέα ερμηνεία της ολικής αδράνειας του πίνακα **Z** στην περίπτωση δύο κατηγορικών μεταβλητών *X* και *Y*: η $I_{0\text{-}1}$ εκφράζει τη μέση αδράνεια ανά μεταβλητή των αδρανειών των ειδικών πινάκων **X** και **Y**. Ο πίνακας **X** αντιστοιχεί στην περιθώρια κατανομή απολύτων συχνοτήτων της μεταβλητής *X* και η αδράνειά του είναι ίση με τη μέγιστη αδράνεια που μπορεί να έχει ένας *k*×*k* πίνακας συνάφειας (βλέπε Κεφάλαιο 5, Ενότητα 5.10). Ανάλογα ισχύουν για τον πίνακα **Y**.

Παρατήρηση 4.4

Γνωρίζουμε ότι στην περίπτωση του λογικού πίνακα δύο κατηγορικών μεταβλητών η αδράνεια μιας μεταβλητής *U* δίνεται από τη σχέση (βλέπε σχέση [2.61] της Ενότητας 2.3.3.1):

$$I_U = \frac{1}{2}(m-1), \qquad [4.12]$$

όπου *m* είναι ο αριθμός των κλάσεων ή ιδιοτήτων της μεταβλητής *U*.

Έτσι, στην περίπτωσή που εξετάζουμε, η ολική αδράνεια $I_{0\text{-}1}$ του πίνακα **Z** λόγω της [4.12] μπορεί να γραφεί και ως εξής:

$$I_{0-1} = \frac{k-1+l-1}{2} = \frac{k-1}{2} + \frac{l-1}{2} = I_X + I_Y, \qquad [4.13]$$

όπου I_X και I_Y είναι οι αδράνειες των μεταβλητών *X* και *Y* αντίστοιχα στην περίπτωση του λογικού πίνακα. Συνεπώς, μια άλλη προσέγγιση στη φυσική ερμηνεία της ολικής αδράνειας $I_{0\text{-}1}$ του πίνακα **Z** είναι ότι εκφράζει το άθροισμα των αδρανειών των δύο μεταβλητών.

4.4 Περίπτωση 3η: Γενικευμένος Πίνακας Συμπτώσεων Δύο Μεταβλητών

Έστω $\mathbf{B}_{(k+l)\times(k+l)}$ ο γενικευμένος πίνακας συμπτώσεων απολύτων συχνοτήτων (πίνακας *Burt*) των δύο κατηγορικών μεταβλητών *X* και *Y* με γενικό στοιχείο b_{ij}, $i=1,\ldots, k+l$, $j=1,\ldots, k+l$. Χρησιμοποιώντας τους συμβολισμούς του Πίνακα 4.4 και θεωρώντας τον πίνακα **B** ως ένα απλό πίνακα συμπτώσεων, μπορούμε να εφαρμόσουμε τη μεθοδολογία, την οποία χρησιμοποιήσαμε για την απόδειξη της Πρότασης 1, για να δείξουμε την παρακάτω πρόταση:

Πρόταση 3

Η ολική αδράνεια I_B του γενικευμένου πίνακα συμπτώσεων **B**, στην περίπτωση δύο κατηγορικών μεταβλητών *X* και *Y*, δίνεται από την παρακάτω σχέση:

$$I_B = \frac{Q_B}{4N} = \frac{I_{0-1} + I_F}{2},$$

όπου Q_B είναι το στατιστικό χ^2 που αντιστοιχεί στον πίνακα **B**, $4N$ είναι το πλήθος των μονάδων παρατήρησης, $I_{0\text{-}1}$ είναι η ολική αδράνεια του αντίστοιχου λογικού πίνακα **Z** και I_F η ολική αδράνεια του απλού πίνακα συμπτώσεων **F** απολύτων συχνοτήτων των δύο μεταβλητών *X* και *Y*.

Προϋπόθεση: Δεν υπάρχουν ελλείπουσες τιμές και όλες οι γραμμές και οι στήλες του πίνακα **B** είναι ενεργές για την εφαρμογή της ΠΑΑ.

Σημείωση: Για τη διευκόλυνση των αλγεβρικών πράξεων στην απόδειξη, στον Πίνακα 4.4 αντί του γενικού συμβολισμού b_{ij} για τη δήλωση των στοιχείων του πίνακα *Burt*, θα χρησιμοποιήσουμε τους συμβολισμούς του απλού πίνακα συμπτώσεων των δύο μεταβλητών όπως αυτοί συνδέονται με τους αντίστοιχους του *Burt*.

Πίνακας 4.4: Ο Γενικευμένος Πίνακας Συμπτώσεων **B** με Περιθώρια Γραμμή και Στήλη

	Μεταβλητή X Κλάσεις ή ιδιότητες της X						Μεταβλητή Y Κλάσεις ή ιδιότητες της Y							
Κλάσεις ή ιδιότητες	1	2	…	i	…	k	1	2	…	j	…	l	Άθροισμα	Άθροισμα
1	f_{1+}	0	…	0	…	0	f_{11}	f_{12}	…	f_{ij}	…	f_{1l}	t_1	$2\,f_{1+}$
2	0	f_{2+}	…	0	…	0	f_{21}	f_{22}	…	f_{2j}	…	f_{2l}	t_2	$2\,f_{2+}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	0	0	0	f_{i+}	…	0	f_{i1}	f_{i2}	…	f_{ij}	…	f_{il}	t_i	$2\,f_{i+}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	0	0	…	0	…	f_{k+}	f_{k1}	f_{k2}	...	f_{kj}	...	f_{kl}	t_k	$2\,f_{k+}$
1	f_{11}	f_{21}	…	f_{i1}	...	f_{k1}	f_{+1}	0	...	0	...	0	t_{k+1}	$2\,f_{+1}$
2	f_{12}	f_{22}	…	f_{i2}	…	f_{k2}	0	f_{+2}	…	0	…	0	t_{k+2}	$2\,f_{+2}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
j	f_{1j}	f_{2j}	…	f_{ij}	…	f_{kj}	0	0	…	f_{+j}	…	0	t_{k+j}	$2\,f_{+j}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
l	f_{1l}	f_{2l}	…	f_{il}	…	f_{kl}	0	0	…	0	…	f_{+l}	t_{k+l}	$2\,f_{+l}$
Άθροισμα	t_1	t_2	...	...	...	t_k	t_{k+1}	t_{k+2}	...	t_{k+j}	...	t_{k+l}	$4N$	$4N$
Άθροισμα	$2\,f_{1+}$	$2\,f_{2+}$	...	...	...	$2\,f_{k+}$	$2\,f_{+1}$	$2\,f_{+2}$	...	$2\,f_{+j}$	...	$2\,f_{+l}$	$4N$	

Απόδειξη:

Καταρχήν θα πρέπει να παρατηρήσουμε ότι στην περίπτωση του πίνακα **B** το πλήθος των μονάδων παρατήρησης είναι ίσο με $4N$. Στη συνέχεια, υπολογίζουμε την ποσότητα Q_B που αντιστοιχεί στον πίνακα **B** με τη βοήθεια της σχέσης [4.4]. Έχουμε:

$$Q_B = \sum_{i=1}^{k+l}\sum_{j=1}^{k+l} \frac{\left(b_{ij} - \frac{t_i t_j}{4N}\right)^2}{\frac{t_i t_j}{4N}} = \sum_{i=1}^{k+l}\sum_{j=1}^{k+l} \frac{\left[b_{ij}^2 - 2b_{ij}\frac{t_i t_j}{4N} + \left(\frac{t_i t_j}{4N}\right)^2\right]}{\frac{t_i t_j}{4N}} =$$

$$= \sum_{i=1}^{k+l}\sum_{j=1}^{k+l} \frac{b_{ij}^2}{\frac{t_i t_j}{4N}} - 2\sum_{i=1}^{k+l}\sum_{j=1}^{k+l} b_{ij} + \sum_{i=1}^{k+l}\sum_{j=1}^{k+l} \frac{t_i t_j}{4N} = \sum_{i=1}^{k+l}\sum_{j=1}^{k+l} \frac{b_{ij}^2}{\frac{t_i t_j}{4N}} - 2\cdot 4N + 4N =$$

$$= \sum_{i=1}^{k+l}\sum_{j=1}^{k+l} \frac{b_{ij}^2}{\frac{t_i t_j}{4N}} - 4N.$$

Στα ενδιάμεσα βήματα των υπολογισμών χρησιμοποιήσαμε τις σχέσεις:

$$\sum_{i=1}^{k+l}\sum_{j=1}^{k+l} b_{ij} = 4N \quad \text{και} \quad \sum_{i=1}^{k+l}\sum_{j=1}^{k+l} \frac{t_i t_j}{4N} = 4N.$$

Ο πίνακας **B** περιέχει $(k+l)^2$ κελιά από τα οποία τα $2kl+k+l$ δεν περιέχουν μηδενικά. Έτσι το άθροισμα:

$$\sum_{i=1}^{k+l}\sum_{j=1}^{k+l} \frac{b_{ij}^2}{\frac{t_i t_j}{4N}}, \qquad [4.14]$$

αποτελείται από $2kl+k+l$ μη μηδενικούς όρους από τους οποίους οι $k+l$ αφορούν τα στοιχεία της κύριας διαγωνίου του πίνακα **B**. Το μέρος του αθροίσματος [4.14] που αντιστοιχεί στα $k+l$ στοιχεία της κύριας διαγωνίου είναι της μορφής:

$$\frac{4Nf_{1+}^2}{2f_{1+}2f_{1+}}+\frac{4Nf_{2+}^2}{2f_{2+}2f_{2+}}+\cdots+\frac{4Nf_{k+}^2}{2f_{k+}2f_{k+}}+\frac{4Nf_{+1}^2}{2f_{+1}2f_{+1}}+\frac{4Nf_{+2}^2}{2f_{+2}2f_{+2}}+\cdots+\frac{4Nf_{+l}^2}{2f_{+l}2f_{+l}}=$$

$$=\frac{4Nf_{1+}^2}{4f_{1+}^2}+\cdots+\frac{4Nf_{+l}^2}{4f_{+l}^2}=\underbrace{N+N+\cdots+N}_{k\ \text{όροι}}+\underbrace{N+N+\cdots+N}_{l\ \text{όροι}}=N\left(k+l\right). \qquad [4.15]$$

Το μέρος του αθροίσματος [4.14] που αντιστοιχεί στα *kl* μη μηδενικά στοιχεία, τα οποία στην ουσία αποτελούν τα στοιχεία του αρχικού απλού πίνακα συμπτώσεων **F**, είναι της μορφής:

$$\frac{4Nf_{11}^2}{2f_{1+}2f_{+1}}+\frac{4Nf_{12}^2}{2f_{1+}2f_{+2}}+\cdots+\frac{4Nf_{1l}^2}{2f_{1+}2f_{+l}}+\cdots+\frac{4Nf_{k1}^2}{2f_{k+}2f_{+1}}+\frac{4Nf_{k2}^2}{2f_{k+}2f_{+2}}+\cdots+\frac{4Nf_{kl}^2}{2f_{k+}2f_{+l}}=$$

$$=\frac{f_{11}^2}{\frac{f_{1+}f_{+1}}{N}}+\cdots+\frac{f_{kl}^2}{\frac{f_{k+}f_{+l}}{N}}=\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{f_{ij}^2}{\frac{f_{i+}f_{+j}}{N}}=Q_F+N, \qquad [4.16]$$

όπου Q_F είναι το στατιστικό χ^2 που αντιστοιχεί στον πίνακα **F**. Στην τελευταία ισότητα της παραπάνω σχέσης κάναμε χρήση της σχέσης:

$$Q_F=\left(\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{f_{ij}^2}{\frac{f_{i+}f_{+j}}{N}}\right)-N\Rightarrow Q_F+N=\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{f_{ij}^2}{\frac{f_{i+}f_{+j}}{N}}. \qquad [4.17]$$

Τέλος, το μέρος του αθροίσματος [4.14] που αντιστοιχεί στα τελευταία *kl* μη μηδενικά στοιχεία, τα οποία αποτελούν τα στοιχεία του πίνακα $\mathbf{F^T}$, είναι ίσο και αυτό με:

$$Q_F+N \qquad [4.18]$$

Από τις σχέσεις [4.15], [4.16] και [4.18] έχουμε τελικά:

$$Q_B=\sum_{i=1}^{k+l}\sum_{j=1}^{k+l}\frac{b_{ij}^2}{\frac{t_it_j}{4N}}-4N=\left[N\left(k+l\right)+2\left(Q_F+N\right)\right]-4N=N\left(k+l-2\right)+2Q_F \qquad [4.19]$$

Βάσει της επισήμανσης (2) η ολική αδράνεια του πίνακα **B** δίνεται από την παρακάτω σχέση:

$$I_B = \frac{Q_B}{4N}. \qquad [4.20]$$

Λόγω, όμως, της [4.19], η [4.20] γράφεται:

$$I_B = \frac{N\left(k+l-2\right)+2Q_F}{4N} = \frac{k+l-2}{4} + \frac{Q_F}{2N}. \qquad [4.21]$$

Η σχέση [4.21] λόγω της [4.1] και της [4.7] γράφεται:

$$I_B = \frac{k+l-2}{4} + \frac{1}{2}I_F = \frac{1}{2}\left(\frac{k+l}{2} - 1 + I_F\right) = \frac{1}{2}\left(I_{0-1} + I_F\right). \qquad [4.22]$$

Τελικά, η σχέση που συνδέει την ολική αδράνεια του γενικευμένου πίνακα συμπτώσεων **B** με τις ολικές αδράνειες των πινάκων **F** και **Z** δίνεται από την παρακάτω σχέση:

$$I_B = \frac{I_{0-1} + I_F}{2}. \square \qquad [4.23]$$

Παρατήρηση 4.5

Από την [4.23] προκύπτει μια νέα προσέγγιση στον υπολογισμό και στην ερμηνεία της ολικής αδράνεια του πίνακα **B** δύο κατηγορικών μεταβλητών X και Y: η I_B εκφράζει τη μέση αδράνεια των πινάκων **F** και **Z**. Ο πίνακας **B** είναι ένας block πίνακας που αποτελείται από 4 υποπίνακες τους **X**, **Y**, **F** και $\mathbf{F}^{\mathbf{T}}$. Η σχέση [4.23] λόγω της σχέσης [4.11] μπορεί να γραφεί και ως εξής:

$$I_B = \frac{\frac{I_{XX} + I_{YY}}{2} + \frac{2I_F}{2}}{2} \Rightarrow \; I_B = \frac{I_{XX} + I_{yy} + I_F + I_{F^T}}{4}, \qquad [4.24]$$

όπου I_{XX} και I_{YY} είναι οι ολικές αδράνειες των πινάκων **X** και **Y** αντίστοιχα και $I_F = I_{F^T}$ οι ολικές αδράνειες των πινάκων **F** και $\mathbf{F}^{\mathbf{T}}$. Με τον τρόπο αυτό, επιτυγχάνουμε μια νέα προσέγγιση στον υπολογισμό και στην ερμηνεία της ολικής αδράνειας του πίνακα **B** δύο κατηγορικών μεταβλητών X και Y: η I_B εκφράζει τη μέση αδράνεια των 4 υποπινάκων που απαρτίζουν τον πίνακα **B**.

4.5 Σχέσεις Αδρανειών των Αξόνων στη Διμεταβλητή Περίπτωση

Στην ενότητα αυτή θα αναφερθούμε συνοπτικά, μέσω του Πίνακα 4.5, στις σχέσεις, οι οποίες συνδέουν τις αδράνειες των αντίστοιχων πινάκων που διαγωνοποιούνται κατά την ΠΑΑ, δηλαδή τις σχέσεις που συνδέουν τις αδράνειες των παραγοντικών αξόνων στις τρεις περιπτώσεις. Για περισσότερες πληροφορίες παραπέμπουμε στους Greenacre (1984), Israëls (1987), Gifi (1996) και Lebart, Morineau και Piron (2000).

Από την τελευταία στήλη του Πίνακα 4.5, η οποία παρουσιάζει τις σχέσεις που συνδέουν τις αδράνειες των παραγοντικών αξόνων των τριών πινάκων, μπορούμε να αιτιολογήσουμε, σε ένα πρώτο επίπεδο, γιατί η εικόνα επί των παραγοντικών επιπέδων του φαινομένου που εξετάζεται είναι η ίδια και στις τρεις περιπτώσεις. Φαίνεται, λοιπόν, ότι η ανάλυση των πινάκων **Z** και **B** έχει ως αποτέλεσμα μια αλλαγή στην κλίμακα μέτρησης, διαφορετική κάθε φορά, των συντεταγμένων των προβολών των σημείων γραμμών και στηλών επί των παραγοντικών αξόνων σε σχέση με τις συντεταγμένες των σημείων, όπως αυτές υπολογίζονται αρχικά από την ανάλυση του πίνακα **F**.

Πίνακας 4.5: Σχέσεις Αδρανειών

Τύπος Πίνακα	Διαστάσεις Πίνακα	Μέγιστη Διάσταση Λύσης της ΠΑΑ	Αδράνεια Άξονα
Πίνακας Συμπτώσεων **F**	$k\times l$	$p=min\{k\text{-}1, l\text{-}1\}$	λ_F
Λογικός Πίνακας 0-1 **Z**	$N\times(k+l)$	$k+l$-2 με $N>>k+l$-2	$\lambda=\frac{1+\sqrt{\lambda_F}}{2}$ *
Γενικευμένος Πίνακας Συμπτώσεων (*Burt*) $\mathbf{B}=\mathbf{Z}^T\mathbf{Z}$	$(k+l)\times(k+l)$	$k+l$-2	λ^2 *

* Στη διμεταβλητή περίπτωση, οι σχέσεις ισχύουν για τους πρώτους p άξονες. Στην πολυμεταβλητή ισχύουν για όλους τους j-q άξονες, όπου j είναι το συνολικό πλήθος των κατηγοριών των q μεταβλητών (βλέπε Ενότητα 2.3.3.5).

Οι χαρακτηριστικές τιμές των πινάκων, που διαγωνοποιούνται στις τρεις περιπτώσεις, παίρνουν τιμές στο διάστημα [0,1] (βλέπε Δεύτερο Βήμα της Ενότητας 2.2.14) και είναι εύκολο να δειχθεί ότι γενικά ισχύει $\lambda\geq\lambda^2\geq\lambda_F$. Μάλιστα, αν εξαιρέσουμε τις τετριμμένες χαρακτηριστικές τιμές 0 και 1 (βλέπε Παρατήρηση 2.1), τότε $\lambda>\lambda^2>\lambda_F$.

Χωρίς περιορισμό της γενικότητας, μπορούμε να υποθέσουμε ότι $p=min\{k\text{-}1, l\text{-}1\}=k\text{-}1$ οπότε $k+l\text{-}2=p+l\text{-}1$, δηλαδή $p<k+l\text{-}2$. Στην περίπτωση αυτή, οι ολικές αδράνειες των τριών πινάκων δίνονται από τις παρακάτω σχέσεις (βλέπε Δεύτερο Βήμα της Ενότητας 2.2.14):

$$I_F = \sum_{i=1}^{s} \lambda_{Fi} \ , \ I_{0-1} = \sum_{i=1}^{s+l-1} \lambda_i \ , \ I_B = \sum_{i=1}^{s+l-1} \lambda_i^2 .$$

Από τις παραπάνω σχέσεις αν λάβουμε υπόψη ότι, εν γένει, $\lambda>\lambda^2>\lambda_F$ είναι φανερό ότι:

$$I_{0-1} > I_B > I_F \ .$$

Επίσης, από τις σχέσεις της τελευταίας στήλης του Πίνακα 4.5 μπορεί να δειχθεί ότι $0,5<\lambda<1$ και $0,25<\lambda^2<1$. Συνεπώς, ένα κριτήριο για την αξιολόγηση της ποιότητας της πληροφορίας είναι το εξής: στην περίπτωση του **Z** θα πρέπει για την ερμηνεία των αποτελεσμάτων να επιλέξουμε τους παραγοντικούς άξονες, για τους οποίους οι αντίστοιχες αδράνειες είναι μεγαλύτερες από $0,5=1/q$ (για $q=2$), ενώ στην περίπτωση του **B** τους άξονες για τους οποίους οι αδράνειές τους είναι μεγαλύτερες από $0,25=1/q^2$ ($q=2$). Η διαπίστωση αυτή συμφωνεί με τα εμπειρικά κριτήρια για τη σημαντικότητα των αξόνων που παρουσιάσαμε στις Ενότητες 2.3.3.6 και 2.3.4.2 (βλέπε Παρατήρηση Η).

Παρατήρηση 4.6

Σχετικά με την ποσότητα της πληροφορίας, με την έννοια της αδράνειας, η οποία αναλύεται μέσω της ΠΑΑ, έχουμε να παρατηρήσουμε ότι ο πίνακας **Z** περιέχει περισσότερη πληροφορία από τον **B** και αυτός με τη σειρά του περισσότερη από τον **F**. Όμως, η αδράνεια σε κάθε περίπτωση διαχέεται σε διαφορετικό αριθμό παραγοντικών αξόνων με αποτέλεσμα τα ποσοστά της συνολικής αδράνειας που ερμηνεύουν οι 2 ή 3 πρώτοι παραγοντικοί άξονες να δίνουν μερικές φορές μια απαισιόδοξη εικόνα σχετικά με την ερμηνεία του φαινομένου που εξετάζεται.

4.6 Γενικεύσεις των Προτάσεων 1 και 3

Στην ενότητα αυτή μελετάμε τις γενικεύσεις των σχέσεων, οι οποίες δείχθηκαν στα προηγούμενα, στην περίπτωση πολλών μεταβλητών.

4.6.1 Γενίκευση της Πρότασης 1

Πρόταση 4

Η ολική αδράνεια $I_{0\text{-}1}$ του λογικού πίνακα **Z**, στην περίπτωση q κατηγορικών μεταβλητών X_i, με i=1, …, q, δίνεται από την παρακάτω σχέση:

$$I_{0-1} = \frac{Q_{0-1}}{qN} = \frac{j-q}{q} = \frac{j}{q} - 1, \qquad [4.25]$$

όπου Q_{0-1} είναι το στατιστικό χ^2 που αντιστοιχεί στον πίνακα **Z**, qN είναι το πλήθος των μονάδων παρατήρησης, j είναι το συνολικό πλήθος των κλάσεων (ιδιοτήτων) των q μεταβλητών X_i, δηλαδή $j = \sum_{i=1}^{q} j_i$, όπου j_i είναι οι κλάσεις της μεταβλητής X_i.

Προϋπόθεση: Δεν υπάρχουν ελλείπουσες τιμές και όλες οι γραμμές και οι στήλες του πίνακα **Z** είναι ενεργές για την εφαρμογή της ΠΑΑ.

Σημείωση: Το πλήθος των πειραματικών ή δειγματοληπτικών μονάδων της μελέτης εξακολουθεί να είναι N αλλά το πλήθος των μονάδων παρατήρησης είναι qN. Επίσης, η ολική αδράνεια του πίνακα **Z** υπολογίζεται και ως εξής (βλέπε Ενότητα 2.3.3.1):

$$I_{0-1} = \sum_{s=1}^{p} \lambda_s ,$$

όπου λ_s είναι η αδράνεια του άξονα s και p=j-q.

Απόδειξη (α):

Η σχέση [4.25] είναι ήδη γνωστή από τη βιβλιογραφία (Lebart, Morineau & Tabard 1977, Greenacre 1984, Israëls 1987, Gifi 1996, Lebart, Morineau & Piron 2000, βλέπε και Ενότητα 2.3.3.1). Εναλλακτικά, μπορούμε να εφαρμόσουμε το συλλογισμό και τη

μεθοδολογία που χρησιμοποιήσαμε για την απόδειξη της Πρότασης 1. Μπορούμε, δηλαδή, να θεωρήσουμε το $N \times j$ λογικό πίνακα **Z** ως ένα "αραιό" αλλά απλό πίνακα συμπτώσεων δύο μεταβλητών, όπου η μεταβλητή που ορίζεται από το σύνολο των γραμμών αφορά στα αντικείμενα ή υποκείμενα της μελέτης και έχει N κλάσεις, ενώ η μεταβλητή που ορίζεται από το σύνολο των στηλών αφορά στο προφίλ των γραμμών, ως προς τις εξεταζόμενες μεταβλητές X_i, και έχει j κλάσεις ή ιδιότητες. Υποθέτουμε και πάλι ότι η μεταβλητή των στηλών αποτελεί μια νέα μεταβλητή, η οποία δημιουργείται από την ένωση των μεταβλητών X_i. Κάτω από την προϋπόθεση ότι δεν υπάρχουν ελλείπουσες τιμές, το σύνολο των μονάδων παρατήρησης είναι ίσο με qN. Στο πλαίσιο αυτό και σύμφωνα με την Παρατήρηση 4.1, μπορούμε πάλι να υποθέσουμε ότι η ολική αδράνεια $I_{0\text{-}1}$ του πίνακα **Z** εκφράζει τη μέση απόκλιση ανά μονάδα παρατήρησης από την κατάσταση ανεξαρτησίας μεταξύ της μεταβλητής των γραμμών και της μεταβλητής των στηλών, όπως αυτή μετριέται μέσω του στατιστικού χ^2. Πιο συγκεκριμένα, ισχύει:

$$I_{0-1} = \frac{Q_{0-1}}{qN},$$

όπου Q_{0-1} είναι το στατιστικό χ^2 που αντιστοιχεί στο λογικό πίνακα **Z**. Εργαζόμενοι στο δεύτερο μέλος της σχέσης καταλήγουμε στη σχέση [4.25].

Μια άλλη προσέγγιση για την απόδειξη της σχέσης [4.25] είναι η εξής:

Απόδειξη (β):

Από την Παρατήρηση 4.4 έχουμε ότι στη διμεταβλητή περίπτωση η ολική αδράνεια του αντίστοιχου πίνακα **Z** είναι ίση με το άθροισμα των αδρανειών των δύο μεταβλητών. Γνωρίζουμε, επίσης, (βλέπε Ενότητα 2.3.3.1) ότι στην περίπτωση των q μεταβλητών η αδράνεια I_{X_i} κάθε μεταβλητής X_i, όπως αυτή υπολογίζεται από το λογικό πίνακα **Z**, είναι ίση με:

$$I_{X_i} = \frac{1}{q}(j_i - 1),$$

όπου j_i είναι οι κλάσεις ή ιδιότητες της μεταβλητής X_i.

Γενικεύοντας, είναι εύκολο να δειχθεί ότι η ολική αδράνεια του πίνακα **Z** είναι ίση με το άθροισμα των αδρανειών των q μεταβλητών, δηλαδή:

$$I_{0-1} = \sum_{i=1}^{q} I_{X_i}. \qquad [4.26]$$

Πράγματι, εργαζόμενοι στο δεύτερο μέλος της σχέσης [4.26] έχουμε:

$$\sum_{i=1}^{q} I_{X_i} = \sum_{i}^{q} \frac{1}{q}(j_i - 1) = \frac{1}{q}(j - q) = \frac{j}{q} - 1 = I_{0-1}, \text{ αφού } \sum_{i=1}^{q} j_i = j \,. \square$$

Τέλος, με συλλογισμό ανάλογο με αυτόν της Παρατήρησης 4.3, μπορούμε εύκολα να γενικεύσουμε τη σχέση [4.11] στην περίπτωση των q μεταβλητών:

$$I_{0-1} = \frac{j}{q} - 1 = \frac{\sum_{i=1}^{q} j_i}{q} - 1 = \frac{\sum_{i=1}^{q} j_i - q}{q} = \frac{\sum_{i=1}^{q} j_i - \sum_{i=1}^{q} 1}{q} = \frac{\sum_{i=1}^{q} (j_i - 1)}{q} \Rightarrow$$

$$I_{0-1} = \frac{\sum_{i=1}^{q} I_{X_i X_i}}{q}, \qquad [4.27]$$

όπου $I_{X_i X_i}$ είναι η αδράνεια του πίνακα **Xi** (**i**=1, …, q) που διασταυρώνει τις κλάσεις ή ιδιότητες της μεταβλητής X_i μεταξύ τους.□

Παρατήρηση 4.7

Οι σχέσεις [4.26] και [4.27] οδηγούν στη διαπίστωση ότι η ολική αδράνεια του πίνακα **Z** εκφράζει είτε το άθροισμα των αδρανειών των q μεταβλητών είτε τη μέση αδράνεια, ανά μεταβλητή, των πινάκων **Xi** (**i**=1, …, q). Η διαπίστωση αυτή οδηγεί σε μια νέα προσέγγιση στην ερμηνεία της ολικής αδράνειας του πίνακα **Z**.

4.6.2 Γενίκευση της Πρότασης 3

Πρόταση 5

Η ολική αδράνεια I_B του γενικευμένου πίνακα συμπτώσεων **B**, q κατηγορικών μεταβλητών X_i, δίνεται από την παρακάτω σχέση:

$$I_B = \frac{Q_B}{Nq^2} = \frac{I_{0-1}}{q} + \frac{2}{q}\left(\frac{\sum_{h<w} I_{hw}}{q} \right), \text{ με } h, w=1,\ldots, q \qquad [4.28]$$

όπου Q_B είναι το στατιστικό χ^2 που αντιστοιχεί στον πίνακα **B**, Nq^2 είναι το πλήθος των μονάδων παρατήρησης, $I_{0\text{-}1}$ είναι η ολική αδράνεια του αντίστοιχου λογικού πίνακα **Z** και $\sum_{h<w} I_{hw}$ το άθροισμα των αδρανειών των $q(q-1)/2$ διαφορετικών απλών πινάκων συμπτώσεων των q μεταβλητών ανά δύο. Η ποσότητα I_{hw} εκφράζει την αδράνεια του υποπίνακα που σχηματίζεται από τη διασταύρωση της μεταβλητής X_h με τη X_w.

Προϋπόθεση: Δεν υπάρχουν ελλείπουσες τιμές και όλες οι γραμμές και οι στήλες του πίνακα **B** είναι ενεργές για την εφαρμογή της ΠΑΑ.

Σημείωση: Το πλήθος των πειραματικών ή δειγματοληπτικών μονάδων της μελέτης εξακολουθεί να είναι N αλλά το πλήθος των μονάδων παρατήρησης είναι Nq^2. Επίσης, η ολική αδράνεια του πίνακα **B** υπολογίζεται και ως εξής (βλέπε Ενότητα 2.3.3.5):

$$I_B = \sum_{s=1}^{p} \lambda_{Bs},$$

όπου λ_{Bs} είναι η αδράνεια του άξονα s και $p=j-q$.

Απόδειξη:

Μπορούμε, και στην περίπτωση αυτή, να θεωρήσουμε τον πίνακα **B** ως ένα απλό πίνακα συμπτώσεων και να υπολογίσουμε το αντίστοιχο στατιστικό χ^2, όπως ακριβώς εργαστήκαμε για την απόδειξη της Πρότασης 3.

Θα πρέπει όμως να επισημάνουμε τα παρακάτω:

1. Ο πίνακας **B** είναι διαστάσεων $j \times j$, με $j = \sum_{i=1}^{q} j_i$, όπου j_i είναι οι κλάσεις ή ιδιότητες της μεταβλητής X_i.

2. Ο **B** είναι ένας block πίνακας που αποτελείται από q^2 υποπίνακες από τους οποίους: οι q σε πλήθος είναι διαγώνιοι, στους οποίους οι κλάσεις ή ιδιότητες της κάθε μεταβλητής X_i διασταυρώνονται μεταξύ τους (όπως ο ειδικός πίνακας **X** που παρουσιάστηκε στα προηγούμενα), οι $q(q-1)/2$ υποπίνακες αντιστοιχούν στους πίνακες συμπτώσεων των q μεταβλητών ανά δύο και οι υπόλοιποι $q(q-1)/2$ υποπίνακες αντιστοιχούν στους ανάστροφους απλούς πίνακες συμπτώσεων των q μεταβλητών ανά δύο (βλέπε Ενότητα 2.3.2).
3. Τα στοιχεία της περιθώριας γραμμής των συνόλων στηλών ταυτίζονται με τα στοιχεία της περιθώριας στήλης των συνόλων γραμμών (βλέπε Ενότητα 2.3.2.1).
4. Το συνολικό πλήθος αποκρίσεων (μονάδων παρατήρησης) είναι ίσο με Nq^2.
5. Αν j_i είναι σε πλήθος οι κλάσεις της μεταβλητής X_i και **di**, με **i**=1, …, q, είναι ο $1\times j_i$ πίνακας γραμμή με στοιχεία τις απόλυτες συχνότητες των j_i κλάσεων ή ιδιοτήτων της μεταβλητής X_i, τότε η περιθώρια γραμμή (αντίστοιχα στήλη) των στηλών (αντίστοιχα γραμμών) του πίνακα **B** αποτελεί ένα block πίνακα γραμμή (αντίστοιχα στήλη) με στοιχεία τους **di** πίνακες όπου, όμως, τα στοιχεία τους είναι πολλαπλασιασμένα επί q (βλέπε Πίνακα 2.7 της Ενότητας 2.3.2 και Ενότητα 2.3.2.1).

Αρχικά, υπολογίζουμε την ποσότητα Q_B που αντιστοιχεί στον πίνακα **B** με τη βοήθεια της σχέσης [4.4]. Αν b_{pr} είναι το γενικό στοιχείο του **B** με p, r=1, …, j ($j=\sum_{i=1}^{q} j_i$) και t_l (l=1, …, j) το γενικό στοιχείο της περιθώριας γραμμής (αντίστοιχα στήλης) της l στήλης (αντίστοιχα γραμμής) του πίνακα **B**, τότε έχουμε:

$$Q_B=\sum_{p=1}^{j}\sum_{r=1}^{j}\frac{\left(b_{pr}-\frac{t_p t_r}{Nq^2}\right)^2}{\frac{t_p t_r}{Nq^2}}=\sum_{p=1}^{j}\sum_{r=1}^{j}\frac{\left[b_{pr}^2-2b_{pr}\frac{t_p t_r}{Nq^2}+\left(\frac{t_p t_r}{Nq^2}\right)^2\right]}{\frac{t_p t_r}{Nq^2}}=$$

$$=\sum_{p=1}^{j}\sum_{r=1}^{j}\frac{b_{pr}^2}{\frac{t_p t_r}{Nq^2}}-2\sum_{p=1}^{j}\sum_{r=1}^{j}b_{pr}+\sum_{p=1}^{j}\sum_{r=1}^{j}\frac{t_p t_r}{Nq^2}=\sum_{p=1}^{j}\sum_{r=1}^{j}\frac{b_{pr}^2}{\frac{t_p t_r}{Nq^2}}-2Nq^2+Nq^2=$$

$$= \sum_{p=1}^{j}\sum_{r=1}^{j} \frac{b_{pr}^2}{\frac{t_p t_r}{Nq^2}} - Nq^2 .$$

Στα ενδιάμεσα βήματα των υπολογισμών χρησιμοποιήσαμε τις σχέσεις:

$$\sum_{p=1}^{j}\sum_{r=1}^{j} b_{pr} = Nq^2 \quad \text{και} \quad \sum_{p=1}^{j}\sum_{r=1}^{j} \frac{t_p t_r}{Nq^2} = Nq^2 .$$

Τώρα, με συλλογισμό ανάλογο με αυτό που χρησιμοποιήσαμε στην απόδειξη της Πρότασης 3 για τον επιμερισμό των όρων του αθροίσματος

$$\sum_{p=1}^{j}\sum_{r=1}^{j} \frac{b_{pr}^2}{\frac{t_p t_r}{Nq^2}} .$$

και λαμβάνοντας υπόψη τις παραπάνω πέντε επισημάνσεις, καταλήγουμε τελικά στο ότι:

$$Q_B = \left[N\left(\sum_{i=1}^{q} j_i - q \right) + 2\sum_{h<w} Q_{hw} \right] = N\left(j - q \right) + 2\sum_{h<w} Q_{hw}, \text{ με } h, w = 1, \ldots, q . \qquad [4.29]$$

όπου Q_{hw} είναι το στατιστικό χ^2 που αντιστοιχεί στον απλό πίνακα συμπτώσεων των μεταβλητών X_h και X_w και $\sum_{h<w} Q_{hw}$ είναι το άθροισμα των στατιστικών χ^2 που αντιστοιχούν στους $q(q-1)/2$ διαφορετικούς απλούς πίνακες συμπτώσεων των q μεταβλητών ανά δύο. Επομένως, η ποσότητα $\frac{\sum_{h<w} Q_{hw}}{N}$ εκφράζει το άθροισμα των αδρανειών των παραπάνω πινάκων.

Βάσει της επισήμανσης (2) η ολική αδράνεια του πίνακα **B** δίνεται από τη σχέση:

$$I_B = \frac{Q_B}{Nq^2} . \qquad [4.30]$$

Λόγω, όμως, της [4.29], η [4.30] γράφεται:

$$I_B = \frac{N(j-q)+2\sum_{h<w} Q_{hw}}{Nq^2} = \frac{j-q}{q^2} + \frac{2}{N}\cdot\frac{\sum_{h<w} Q_{hw}}{q^2} = \frac{j-q}{q^2} + \frac{2}{q^2}\cdot\frac{\sum_{h<w} Q_{hw}}{N} =$$

$$= \frac{j-q}{q^2} + \frac{2}{q^2}\cdot\sum_{h<w} I_{hw}. \qquad [4.31]$$

Η [4.31] μπορεί να γραφεί και ως εξής:

$$I_B = \frac{j-q}{q^2} + \frac{2}{q}\cdot\frac{\sum_{h<w} I_{hw}}{q} = \frac{\frac{j}{q}-1}{q} + \frac{2}{q}\cdot\frac{\sum_{h<w} I_{hw}}{q} \Rightarrow$$

$$I_B = \frac{I_{0-1}}{q} + \frac{2}{q}\left(\frac{\sum_{h<w} I_{hw}}{q}\right).\square \qquad [4.32]$$

Λόγω της Πρότασης 4, $(j/q)-1$ είναι η συνολική αδράνεια $I_{0\text{-}1}$ του αντίστοιχου λογικού πίνακα **Z** των q μεταβλητών X_i. Από τη σχέση [4.32], η οποία συνδέει για πρώτη φορά τη I_B με τη $I_{0\text{-}1}$ και τις αδράνειες των απλών πινάκων συμπτώσεων που απαρτίζουν τον πίνακα **B**, παρατηρούμε ότι η ολική αδράνεια του πίνακα **B** εκφράζει ένα μέσο μέγεθος αποτελέσματος ανά μεταβλητή. Ένα μάλλον πολύπλοκο μέσο μέγεθος αποτελέσματος. Όμως, από την Παρατήρηση 4.5 και τη σχέση [4.24] διαπιστώσαμε ότι μια άλλη προσέγγιση στην ερμηνεία της ολικής αδράνειας του πίνακα **B** δύο κατηγορικών μεταβλητών X και Y είναι ότι εκφράζει τη μέση αδράνεια των 4 υποπινάκων που απαρτίζουν τον πίνακα **B**. Γενικεύοντας την παρατήρηση αυτή θα πρέπει η ολική αδράνεια του πίνακα **B,** στην περίπτωση q μεταβλητών, να δίνεται ως ο μέσος όρος των αδρανειών των q^2 σε πλήθος υποπινάκων που απαρτίζουν τον πίνακα **B**. Πράγματι:

$$\frac{\sum_{q} I_{diag} + 2\sum_{h<w} I_{hw}}{q^2} = \frac{\sum_{i=1}^{q}(j_i - 1) + 2\sum_{h<w} I_{hw}}{q^2} = \frac{j-q+2\sum_{h<w} I_{hw}}{q^2} = \frac{j-q}{q^2} + \frac{2}{q}\left(\frac{\sum_{h<w} I_{hw}}{q}\right) =$$

$$= \frac{\frac{j-q}{q}}{q} + \frac{2}{q}\left(\frac{\sum_{h<w} I_{hw}}{q}\right) = \frac{\frac{j}{q}-1}{q} + \frac{2}{q}\left(\frac{\sum_{h<w} I_{hw}}{q}\right) = \frac{I_{0-1}}{q} + \frac{2}{q}\left(\frac{\sum_{h<w} I_{hw}}{q}\right) =$$

$$= I_B,$$

όπου το άθροισμα $\sum_q I_{diag}$ εκφράζει την αδράνεια των q διαγώνιων υποπινάκων του πίνακα **B**.

Συνεπώς,

$$I_B = \frac{\sum_q I_{diag} + 2\sum_{h<w} I_{hw}}{q^2}. \square \qquad [4.33]$$

Άμεσες συνέπειες της Πρότασης 5 και της σχέσης [4.33] είναι τα παρακάτω Πορίσματα:

Πόρισμα 1

Η ολική αδράνεια ενός πίνακα "φέτα", ο οποίος αποτελείται από *t* πίνακες συμπτώσεων τοποθετημένους τον ένα δίπλα στον άλλο (βλέπε Ενότητα 2.4.1), είναι ίση με τη μέση αδράνεια των *t* υποπινάκων.

Πόρισμα 2

Η ολική αδράνεια ενός πίνακα "στοίβα", ο οποίος αποτελείται από *t* πίνακες συμπτώσεων τοποθετημένους τον ένα κάτω από τον άλλο (βλέπε Ενότητα 2.4.1), είναι ίση με τη μέση αδράνεια των *t* υποπινάκων.

Πόρισμα 3

Η ολική αδράνεια ενός υποπίνακα του *Burt*, ο οποίος αποτελείται από *t* υποπίνακες συμπτώσεων (βλέπε Ενότητα 2.4.1), είναι ίση με τη μέση αδράνεια των *t* υποπινάκων.

Παρατήρηση 4.8

Από τη σχέση [4.33] προκύπτει μια νέα προσέγγιση στον υπολογισμό και στην ερμηνεία της ολικής αδράνεια του πίνακα **B**: η ολική αδράνεια του πίνακα **B,** στην περίπτωση q μεταβλητών, εκφράζει τη μέση αδράνεια των αδρανειών των q^2, σε πλήθος, υποπινάκων που τον απαρτίζουν. Από τις γενικές σχέσεις που απορρέουν από την Πρόταση 5 (σχέσεις [4.28], [4.32] και [4.33]) διαπιστώνουμε ότι οι αδράνειες των απλών πινάκων συμπτώσεων των μεταβλητών ανά δύο συμμετέχουν δύο φορές στον υπολογισμό της ολικής αδράνειας του πίνακα **B**. Έτσι, δημιουργείται η εντύπωση ότι η ΠΑΑ αναλύει τελικά ένα διπλό μέγεθος του αποτελέσματος, με την έννοια της αδράνειας, δημιουργώντας την αίσθηση της "υπερπληροφορίας". Τέλος, η ύπαρξη των διαγώνιων υποπινάκων του **B** δημιουργεί πρόβλημα σε ό,τι αφορά τουλάχιστον τη γεωμετρική ερμηνεία των αποτελεσμάτων (Israëls 1987, Gifi 1996, Greenacre 2005, 1994β, 1993γ, 1993α, 1991, 1990, 1989 και 1988α). Φαίνεται να αποτελεί περισσότερο "θόρυβο" παρά πληροφορία. Για το λόγο αυτό έχουν προταθεί διάφορες διορθώσεις τόσο σε σχέση με τα αριθμητικά αποτελέσματα, τα οποία παράγονται από την ΠΑΑ, όσο και σε σχέση με τη μορφή του πίνακα **B** που θα αναλυθεί (Greenacre, 2005, 1994β και 1993γ). Ενδεικτικά αναφέρουμε τις παρακάτω διορθώσεις και τροποποιήσεις:

α) Διορθώσεις των Αδρανειών

Ο Benzécri (1979) προτείνει οι αδράνειες που προκύπτουν από την ανάλυση του πίνακα **B** να διορθωθούν σύμφωνα με την παρακάτω σχέση:

$$\lambda_{Bs}^{adj} = \left(\frac{q}{q-1}\right)^2 \times \left(\sqrt{\lambda_{Bs}} - \frac{1}{q}\right)^2, \text{ για } \sqrt{\lambda_{Bs}} > \frac{1}{q}.$$

όπου q είναι ο αριθμός των μεταβλητών και λ_{Bs} η αδράνεια του άξονα s που προκύπτει από την ανάλυση του πίνακα *Burt*. Στη συνέχεια, οι διορθωμένες αδράνειες εκφράζονται ως ποσοστό του αθροίσματος των διορθωμένων αδρανειών των αξόνων για τους οποίους η αντίστοιχη ιδιοτιμή $\sqrt{\lambda_{Bs}}$ είναι μεγαλύτερη από $1/q$.

Η διόρθωση του Greenacre συνίσταται στον ίδιο μετασχηματισμό με αυτόν του Benzécri με τη διαφορά οι διορθωμένες αδράνειες εκφράζονται ως ποσοστό της ποσότητας (Greenacre 2005, 1994β και 1993α):

$$I_m = \frac{q}{q-1}\left(I_B - \frac{j-q}{q^2} \right),$$

όπου I_B είναι η ολική αδράνεια του πίνακα *Burt* και *j* το συνολικό πλήθος των κατηγοριών των *q* μεταβλητών.

Οι δύο αυτές διορθώσεις υπάρχουν ως επιλογές στην διαδικασία της ΠΑΑ στο στατιστικό πακέτο SAS (SAS Institute, 1999).

β) Τροποποιήσεις του πίνακα *Burt*

Κατά τον Israëls (1987) οι διαγώνιοι υποπίνακες του πίνακα **B** μπορούν να αντικατασταθούν με τους αντίστοιχους πίνακες που προκύπτουν κάτω από την υπόθεση της ανεξαρτησίας των γραμμών και των στηλών τους, δηλαδή με πίνακες όπου τα στοιχεία τους είναι οι αντίστοιχες αναμενόμενες συχνότητες. Με τον τρόπο αυτό, οι αδράνειες των διαγώνιων υποπινάκων είναι ίσες με μηδέν, οπότε δεν συνεισφέρουν στην ολική αδράνεια του πίνακα **B**. Ο Greenacre (1993γ και 1984) προτείνει τα στοιχεία των διαγώνιων πινάκων να αντικατασταθούν με μηδενικά ή να θεωρηθούν ως ελλείπουσες τιμές, ώστε και πάλι οι αδράνειές τους να μη συμμετέχουν στον υπολογισμό της ολικής αδράνειας του πίνακα **B**. Τέλος, η πιο ενδιαφέρουσα εκδοχή φαίνεται να είναι η εφαρμογή της *Joint Correspondence Analysis* που προτείνει ο Greenacre (2005, 1998, 1994β, 1993α και 1988), κατά την οποία αναλύονται μόνο οι $q(q-1)/2$ σε πλήθος υποπίνακες συμπτώσεων των *q* μεταβλητών ανά δύο (βλέπε επίσης Boik 1996, Tateneni & Browne 2000 και Van de Velden 2000). Να παρατηρήσουμε ότι: α) η *Joint Correspondence Analysis* δεν περιλαμβάνεται σε κανένα από τα εμπορικά στατιστικά πακέτα που εξετάσαμε στην Ενότητα 2.2.14.2, β) οι λύσεις της δεν είναι ιεραρχικά διακλαδιζόμενες και γ) δεν ισχύουν πλέον οι ιδιότητες βέλτιστης κλιμάκωσης της ΠΑΑ. Σύμφωνα με τον Greenacre (2005, 1993γ και 1993α), η διόρθωση που προτείνει έχει ως αποτέλεσμα τα παραγόμενα αποτελέσματα να είναι συγκρίσιμα με αυτά της *Joint Correspondence*

Analysis. Εκείνο που έχουμε να επισημάνουμε, επίσης, είναι ότι η διόρθωση του Benzécri αποσκοπεί μόνο στη βελτίωση της ποιότητας της λύσης της ΠΑΑ, με την έννοια της αύξησης του ποσοστού της αδράνειας, το οποίο ερμηνεύουν οι παραγοντικοί άξονες, ώστε να μη δημιουργείται η αίσθηση της φτωχής προσαρμογής των δεδομένων. Η διόρθωση του Greenacre έχει διπλό στόχο: α) να βελτιώσει το ποσοστό ερμηνείας της αδράνειας και β) να διορθώσει το πρόβλημα που δημιουργούν, κατά τη γεωμετρική ερμηνεία των αποστάσεων χ^2 μεταξύ των σημείων επί των παραγοντικών επιπέδων, οι διαγώνιοι πίνακες και οι ανάστροφοι των απλών πινάκων συμπτώσεων που συγκροτούν τον πίνακα *Burt*. Οι δύο διορθώσεις έχουν ως αποτέλεσμα να τροποποιούνται τόσο οι συντεταγμένες των προβολών των σημείων στους παραγοντικούς άξονες όσο και τα άλλα αριθμητικά αποτελέσματα της ΠΑΑ, όπως για παράδειγμα ο δείκτης *CTR*, στον υπολογισμό των οποίων λαμβάνεται υπόψη η αδράνεια των αξόνων. Και στις δύο περιπτώσεις οι σχετικές θέσεις των σημείων δεν μεταβάλλονται. Στην Ενότητα 4.8.3 παραθέτουμε περισσότερα στοιχεία για τις διορθώσεις *του* Greenacre και του Benzécri και προτείνουμε μία εναλλακτική μέθοδο διόρθωσης των αδρανειών του πίνακα *Burt*, η οποία βασίζεται στην έννοια της «Ενδιαφέρουσας Αδράνειας» την οποία ορίζουμε στην επόμενη ενότητα.

4.7 Ενδιαφέρουσα Αδράνεια του Πίνακα *Burt*

Σύμφωνα με τη Γαλλική παράδοση, στην πολυμεταβλητή περίπτωση, η ΠΑΑ εφαρμόζεται στον πίνακα **B** (*Burt*). Οι q^2 σε πλήθος υποπίνακες, οι οποίοι απαρτίζουν τον πίνακα **B**, αναλύονται σε q πίνακες που αντιστοιχούν στους διαγώνιους υποπίνακες και σε $2\times q\left(q-1\right)/2$ υποπίνακες που αντιστοιχούν στους απλούς πίνακες συμπτώσεων των q μεταβλητών μεταξύ τους. Από τη σχέση [4.33] προκύπτει ότι η ολική αδράνεια του πίνακα *Burt* εκφράζει ένα μέσο μέγεθος του αποτελέσματος, αφού είναι ίση με τη μέση αδράνεια των q^2 υποπινάκων από τους οποίους δομείται. Φαίνεται, λοιπόν, ότι το μέρος της ολικής αδράνειας του πίνακα *Burt*, που εκφράζει όλες τις σχέσεις μεταξύ των μεταβλητών ανά δύο και παρουσιάζει ενδιαφέρον να μελετηθεί, είναι η μέση αδράνεια των $q\left(q-1\right)/2$ σε πλήθος διαφορετικών απλών πινάκων συμπτώσεων των q μεταβλητών ανά δύο (Israëls 1987, Boik 1996, Gifi 1996, Tateneni & Browne 2000, Greenacre 2005, 1994β, 1993γ, 1993α, 1991, 1990,

1989 και 1988α). Έτσι, ορίζουμε ως «Ενδιαφέρουσα Αδράνεια» $I_{\varepsilon \mathrm{B}}$ του πίνακα *Burt* (**B**) την ποσότητα:

$$I_{\varepsilon \mathrm{B}} = \frac{\sum_{h<w} I_{hw}}{\frac{q(q-1)}{2}}, \quad h, w = 1, \ldots, q \,. \qquad [4.34]$$

όπου $\sum_{h<w} I_{hw}$ είναι το άθροισμα των αδρανειών των $q(q-1)/2$ διαφορετικών απλών πινάκων συμπτώσεων των q μεταβλητών ανά δύο. Η ποσότητα I_{hw} εκφράζει την αδράνεια του υποπίνακα που σχηματίζεται από τη διασταύρωση της μεταβλητής X_h με τη X_w.

Η φυσική ερμηνεία της ενδιαφέρουσας αδράνειας είναι ότι εκφράζει τη μέση αδράνεια των $q(q-1)/2$ σε πλήθος διαφορετικών απλών πινάκων συμπτώσεων των q μεταβλητών ανά δύο, οι οποίοι συμμετέχουν στην κατασκευή του πίνακα *Burt*. Στην περίπτωση που οι q μεταβλητές είναι ανά δύο ασυσχέτιστες, τότε η αδράνεια κάθε απλού υποπίνακα συμπτώσεων I_{hw} είναι ίση με μηδέν με αποτέλεσμα και η τιμή της ενδιαφέρουσας αδράνειας να είναι ίση με μηδέν.

Με βάση τα παραπάνω, διαπιστώνουμε ότι η ΠΑΑ αναλύει, τελικά, ένα “πακέτο” διμεταβλητών σχέσεων ή, αλλιώς, τις αλληλεπιδράσεις πρώτης τάξης όλων των μεταβλητών ανά δύο (βλέπε και Ενότητα 2.3.3.5). Αν, μάλιστα, θεωρήσουμε ότι η ολική αδράνεια εκφράζει ένα μέτρο της πληροφορίας που περιέχει ο αντίστοιχος πίνακας ή ένα συνολικό μέγεθος αποτελέσματος, τότε μπορούμε να ισχυριστούμε ότι κάθε φορά αναλύεται ένα μέσο μέγεθος του αποτελέσματος, το οποίο προκύπτει από διμεταβλητές και όχι από πολυμεταβλητές σχέσεις (βλέπε σχέση [4.32]). Στο σημείο αυτό θα πρέπει να τονίσουμε ότι στην περίπτωση που οι μεταβλητές είναι ανά δύο ασυσχέτιστες ή σχεδόν ασυσχέτιστες, τότε η ποσότητα

$$\frac{2}{q}\left(\frac{\sum_{h<w} I_{hw}}{q}\right),$$

της σχέσης [4.32], θα είναι ίση με μηδέν ή θα τείνει στο μηδέν. Έτσι, δεν συμβάλει στην ολική αδράνεια και δεν παρέχει πρόσθετη πληροφορία. Στην περίπτωση αυτή (ασυσχέτιστες μεταβλητές), η ολική αδράνεια του πίνακα **B** είναι ίση με

$$I_B = \frac{I_{0-1}}{q} = \frac{\frac{j}{q} - 1}{q} = \frac{j-q}{q^2}$$

ή πολύ κοντά στην ποσότητα αυτή. Με άλλα λόγια, φαίνεται ο πίνακας **B** να έχει αδράνεια διάφορη του μηδενός και να περιέχει πληροφορία, γεγονός καθόλου ρεαλιστικό. Αυτό μπορεί να οδηγήσει σε εσφαλμένα συμπεράσματα ειδικά από "αφελείς" χρήστες της μεθόδου, οι οποίοι εμπιστεύονται πλήρως τα αποτελέσματα που παράγονται από τα στατιστικά πακέτα.

Συμπερασματικά, επιβάλλεται ένας προ-έλεγχος των διμεταβλητών συσχετίσεων των μεταβλητών που θα αναλυθούν με την ΠΑΑ. Στην περίπτωση που τα δεδομένα προέρχονται από απλή τυχαία δειγματοληψία και το μέγεθος του δείγματος είναι αρκούντος μεγάλο μπορεί να εφαρμοστεί ο στατιστικός έλεγχος ανεξαρτησίας χ^2 για τις $q(q-1)/2$, σε πλήθος, συσχετίσεις των μεταβλητών ανά δύο. Μια ανάλογη πρακτική εφαρμόζεται για την επιλογή μεταβλητών στην Ανάλυση σε Κύριες Συνιστώσες. Μεταβλητές που δεν έχουν σημαντική συσχέτιση με τις υπόλοιπες απομακρύνονται από την ανάλυση (Hair *et al*. 1995, Sharma 1996, Coakes & Steed 1999). Μάλιστα, ορισμένοι ερευνητές θεωρούν απόλυτα επιβεβλημένο τον έλεγχο των διμεταβλητών συσχετίσεων στις Παραγοντικές Αναλύσεις (Cohen & Cohen 1983, Fouladi & Steiger 1999). Αν δεν ανιχνευθούν σημαντικές συσχετίσεις, τότε δεν υπάρχει λόγος εφαρμογής των μεθόδων αυτών. Στο μεθοδολογικό πλαίσιο της ΠΑΑ, μια εναλλακτική προσέγγιση θα μπορούσε να είναι ο έλεγχος αρχικά της προσαρμογής του λογαριθμογραμμικού υποδείγματος, που αντιστοιχεί στην υπόθεση της πλήρους ανεξαρτησίας των q μεταβλητών, και, στη συνέχεια, αφού η προηγούμενη υπόθεση απορριφθεί, ο έλεγχος της καλής προσαρμογή του υποδείγματος, στο οποίο συμμετέχουν μόνο οι κύριες επιδράσεις των μεταβλητών και οι αλληλεπιδράσεις πρώτης τάξης. Για περισσότερες πληροφορίες σχετικά με τα λογαριθμογραμμικά υποδείγματα και τις εφαρμογές τους παραπέμπουμε στους Knoke και Burke (1980), Bishop, Fienberg και Holland (1991), Andersen (1991), Ishii-Kuntz (1994), Fienberg (2000 και 1991) και Agresti (2002 και 1984). Στην Ενότητα 4.8.2

προτείνουμε μέθοδο για τον έλεγχο της στατιστικής σημαντικότητας της ενδιαφέρουσας αδράνειας του πίνακα *Burt*, η οποία, όπως είδαμε στα προηγούμενα, είναι συνάρτηση των διμεταβλητών συσχετίσεων των μεταβλητών που συμμετέχουν στην ανάλυση.

4.8 Προτάσεις Εφαρμογών

Στην παρούσα ενότητα προτείνουμε ορισμένες μεθοδολογικές και θεωρητικές εφαρμογές των ευρημάτων που συζητήθηκαν στις προηγούμενες ενότητες.

4.8.1 Μέθοδος Επιλογής Υποπίνακα με την Πλησιέστερη Απεικόνιση μέσω της ΠΑΑ σε αυτήν του Πίνακα *Burt*

Ένα από τα σημαντικότερα αριθμητικά αποτελέσματα της ΠΑΑ, τα οποία λαμβάνονται υπόψη για την εξαγωγή συμπερασμάτων, είναι το ποσοστό της αδράνειας που ερμηνεύει ο κάθε παραγοντικός άξονας. Στην περίπτωση μεγάλου αριθμού μεταβλητών, ο συνολικός αριθμός των κλάσεών τους μπορεί να αυξηθεί σε τέτοιο βαθμό, ώστε: α) τα ποσοστά ερμηνείας της ολικής αδράνειας, ιδιαίτερα στους δύο πρώτους άξονες, να θεωρηθούν χαμηλά και η απεικόνιση, επί των παραγοντικών αξόνων, φτωχή και β) η ερμηνεία του υπό εξέταση φαινομένου να καθίσταται πλέον δυσχερής. Είναι γνωστό ότι όσο ελαττώνονται οι διαστάσεις του υπό ανάλυση πίνακα τόσο λεπτομερέστερη είναι η απεικόνιση επί του παραγοντικού επιπέδου (Μάρκος & Παπαδημητρίου, 2003). Επομένως, το πρόβλημα που τίθεται είναι το κατά πόσο είναι δυνατό να επιλεγεί υποπίνακας του πίνακα *Burt*, με μικρότερες διαστάσεις, ο οποίος να περιλαμβάνει όλες τις υπό εξέταση μεταβλητές και η εφαρμογή της ΠΑΑ σε αυτόν να αποδίδει την “πλησιέστερη εικόνα” του φαινομένου σε αυτή που προκύπτει από την εφαρμογή της μεθόδου στον αρχικό πίνακα *Burt*. Με την έκφραση “πλησιέστερη εικόνα” εννοούμε ότι οι σχετικές θέσεις και οι ομάδες σημείων, επί του παραγοντικού επιπέδου 1×2, θα πρέπει να είναι σχεδόν ίδιες για τις δύο αναλύσεις. Αξίζει να τονιστεί ότι το πρόβλημα αυτό τίθεται για πρώτη φορά στο χώρο της Ανάλυσης Δεδομένων.

Για την πληρέστερη κατανόηση του προβλήματος δίνουμε ένα παράδειγμα εντοπισμού των διαφορετικών υποπινάκων ενός πίνακα *Burt*. Έστω **B** ο γενικευμένος πίνακας συμπτώσεων τριών μεταβλητών, X_1, X_2 και X_3 με 7, 6 και 2 κατηγορίες αντίστοιχα. Οι διαφορετικοί υποπίνακες του **B** που περιλαμβάνουν και τις τρεις μεταβλητές είναι οι εξής:

- Ο 7×8 πίνακας "φέτα" $\mathbf{F_{1(2\cup 3)}}$ που διασταυρώνει τη μεταβλητή X_1 με τις X_2 και X_3.
- Ο 6×9 πίνακας "φέτα" $\mathbf{F_{2(1\cup 3)}}$ που διασταυρώνει τη μεταβλητή X_2 με τις X_1 και X_3.
- Ο 2×13 πίνακας "φέτα" $\mathbf{F_{3(1\cup 2)}}$ που διασταυρώνει τη μεταβλητή X_3 με τις X_1 και X_2.

Στη θέση των παραπάνω πινάκων θα μπορούσαν να χρησιμοποιηθούν και οι ανάστροφοί τους, δηλαδή οι αντίστοιχοι πίνακες "στοίβες". Οι απλοί πίνακες συμπτώσεων, όπως, για παράδειγμα ο $\mathbf{F_{12}}$, που διασταυρώνει τις μεταβλητές X_1 και X_2, αν και είναι υποπίνακες του **B** ωστόσο δεν παρουσιάζουν ενδιαφέρον για την περίπτωση στην οποία αναφερόμαστε, γιατί περιλαμβάνουν μόνο δύο από τις τρεις μεταβλητές. Θα πρέπει να επισημάνουμε ότι σε έναν πίνακα "φέτα", για παράδειγμα στον $\mathbf{F_{1(2\cup 3)}}$, κατά την εφαρμογή της ΠΑΑ λαμβάνεται υπόψη μόνο η συσχέτιση (συνάφεια) της μεταβλητής X_1 με τη X_2 και της X_1 με τη X_3. Δεν λαμβάνεται υπόψη η συσχέτιση της X_2 με τη X_3. Αν η συσχέτιση των X_2 και X_3 είναι αμελητέα, τότε δεν έχει ιδιαίτερη αξία η ερμηνεία της σχέσης των δύο αυτών μεταβλητών επί των παραγοντικών επιπέδων. Ανάλογα συμπεράσματα ισχύουν και για τους πίνακες τύπου "στοίβας" και τους υποπίνακες του *Burt*. Σε κάθε περίπτωση, μέσω της ΠΑΑ αναλύονται μόνο οι συσχετίσεις μεταξύ των μεταβλητών για τις οποίες υπάρχει πίνακας διασταύρωσης των κατηγοριών τους.

Το προτεινόμενο κριτήριο επιλογής του "καλύτερου" υποπίνακα του *Burt* βασίζεται στην έννοια της ενδιαφέρουσας αδράνειας, η οποία, όπως είδαμε στην Ενότητα 4.7, ορίζεται ως η μέση αδράνεια των $q(q-1)/2$, σε πλήθος, διαφορετικών απλών πινάκων συμπτώσεων που σχηματίζουν q μεταβλητές ανά δύο (σχέση [4.34]). Αν θεωρήσουμε όλους τους δυνατούς υποπίνακες ως block πίνακες, που αποτελούνται

από απλούς πίνακες συμπτώσεων, τότε ως καλύτερος υποπίνακας επιλέγεται αυτός που η αδράνειά του μεγιστοποιεί το λόγο:

$$g = \frac{I_{F^*}}{I_{\varepsilon B}}, \qquad [4.35]$$

όπου I_{F^*} είναι η αδράνεια του εξεταζόμενου υποπίνακα $\mathbf{F}^*$ και $I_{\varepsilon B}$ η ενδιαφέρουσα αδράνεια του πίνακα *Burt*.

Η φυσική ερμηνεία του λόγου *g* είναι ότι εκφράζει την αναλογία του πληροφοριακού περιεχομένου του υποπίνακα $\mathbf{F}^*$, με την έννοια της αδράνειας, ως προς την ενδιαφέρουσα αδράνεια του πίνακα *Burt*. Ο λόγος μπορεί να πάρει και τιμές μεγαλύτερες από 1.

Οι αδράνειες των υποπινάκων $\mathbf{F}^*$ μπορούν να υπολογιστούν μέσω των Πορισμάτων 1, 2 και 3 (βλέπε Ενότητα 3.6.2). Με βάση την Πρόταση 5 και τα Πορίσματα 1, 2 και 3 η πληροφορία που ενθυλακώνει ο υποπίνακας $\mathbf{F}^*$ αποτελεί, κατά μία έννοια, μέρος του πληροφοριακού περιεχομένου του *Burt*, αφού σε κάθε περίπτωση τόσο οι αδράνειες των υποπινάκων όσο και η ενδιαφέρουσα αδράνεια του *Burt* εκφράζονται ως μέσοι όροι των αδρανειών των απλών πινάκων συμπτώσεων από τους οποίους συντίθενται.

Η προτεινόμενη μεθοδολογία επιλογής υποπίνακα από τον πίνακα *Burt*, περιλαμβάνει τα παρακάτω βήματα:

Βήμα 1: Υπολογίζεται η ενδιαφέρουσα αδράνεια του πίνακα *Burt* μέσω της σχέσης [4.34].

Βήμα 2: Δημιουργούνται όλοι οι δυνατοί υποπίνακες $\mathbf{F}^*$, που περιλαμβάνουν συνδυασμούς όλων των μεταβλητών, και υπολογίζεται η αδράνειά τους ως η μέση αδράνεια των επιμέρους πινάκων από τους οποίους δομούνται (βλέπε Πορίσματα 1, 2 και 3). Αν ο αριθμός *q* των μεταβλητών είναι άρτιος, τότε το πλήθος των προς εξέταση διαφορετικών υποπινάκων, που περιλαμβάνουν συνδυασμούς όλων των μεταβλητών, είναι (Μάρκος, Μενεξές & Παπαδημητρίου, 2005):

$$\binom{q}{1}+\binom{q}{2}+\ldots+\binom{q}{\frac{q}{2}}\Big/2$$

Αν το q είναι περιττός, τότε το πλήθος των προς εξέταση πινάκων είναι:

$$\binom{q}{1}+\binom{q}{2}+\ldots+\binom{q}{\frac{q-1}{2}}$$

Βήμα 3: Για κάθε υποπίνακα υπολογίζεται ο λόγος g από τη σχέση [4.35] και τελικά, επιλέγεται εκείνος ο υποπίνακας που η αδράνειά του μεγιστοποιεί την ποσότητα g.

Η προτεινόμενη μέθοδος είναι διαθέσιμη στο λογισμικό CHIC Analysis (Μάρκος, 2006).

Για να ελέγξουμε την αποτελεσματικότητα της προτεινόμενης μεθοδολογίας εφαρμόσαμε την παραπάνω διαδικασία σε διαφορετικά σύνολα δεδομένων με γνωστή *a priori* δομή. Σε όλες τις περιπτώσεις, που εξετάσαμε, οι απεικονίσεις επί των παραγοντικών επιπέδων ήταν “κοντά” σε αυτές που προκύπτουν από την ανάλυση των αντίστοιχων πινάκων *Burt*. Ενδεικτικά παρουσιάζουμε τα αποτελέσματα για τέσσερα σύνολα δεδομένων.

Α) Σύνολο δεδομένων *A* που προέκυψε από πραγματική έρευνα.

Το σύνολο δεδομένων *A* (βλέπε αρχείο data_files.xls στο Παράρτημα CDA του CD που συνοδεύει τη διατριβή) προέκυψε από έρευνα που πραγματοποιήθηκε σε 138 φοιτητές και περιλαμβάνει τρεις μεταβλητές, το «είδος διακοπών του φοιτητή» με 6 κλάσεις, το «επάγγελμα του πατέρα» με 7 κλάσεις και το «φύλο του φοιτητή» με 2 κλάσεις (βλέπε Μάρκος & Παπαδημητρίου, 2003). Από την εφαρμογή της προτεινόμενης μεθόδου στον 15×15 πίνακα *Burt* προέκυψαν τρεις διαφορετικοί πίνακες “φέτες”, δηλαδή υποπίνακες του *Burt*, όπου μία από τις μεταβλητές εμφανίζεται στις γραμμές του πίνακα και οι υπόλοιπες δύο στις στήλες:

α) Ο υποπίνακας $\mathbf{A}_{2\times13}$ με το «φύλο» στις γραμμές και το «επάγγελμα» μαζί με τις «διακοπές» στις στήλες, β) ο υποπίνακας $\mathbf{B}_{7\times8}$ με τις «διακοπές» στις γραμμές και τις

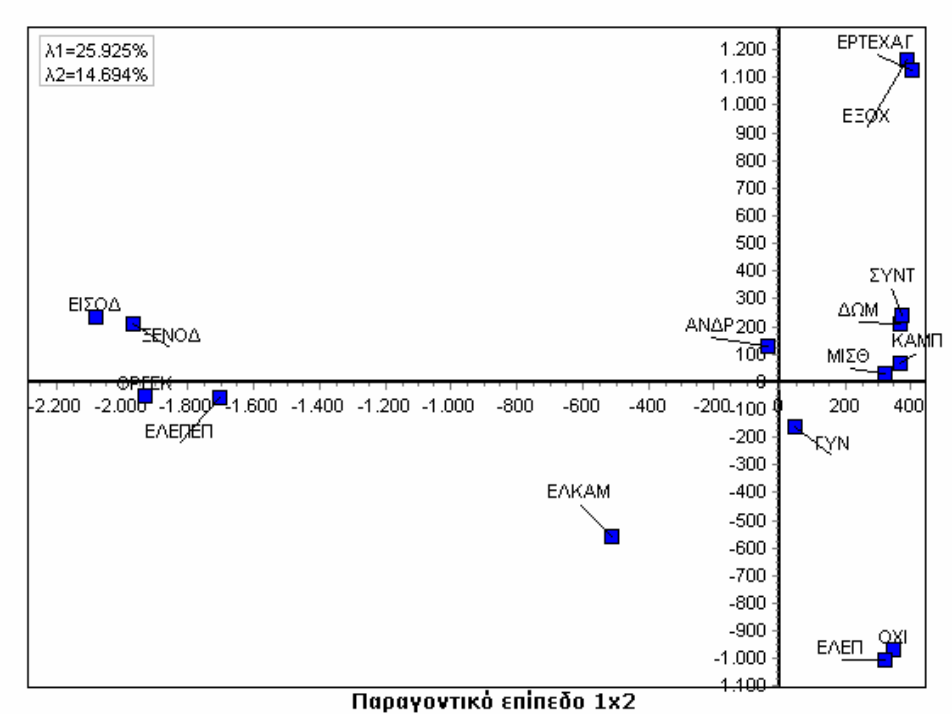
άλλες δύο μεταβλητές στις στήλες και γ) ο υποπίνακας $\mathbf{\Gamma}_{6\times9}$ με το «επάγγελμα» στις γραμμές και τις άλλες στις στήλες. Για κάθε υποπίνακα υπολογίστηκε η αδράνειά του καθώς και ο λόγος *g* (βλέπε Σχήμα 4.1). Οι αδράνειες των υποπινάκων $\mathbf{A}_{2\times13}$, $\mathbf{B}_{7\times8}$ και $\mathbf{\Gamma}_{6\times9}$ υπολογίστηκαν ως οι μέσοι όροι των αδρανειών των δύο απλών πινάκων συμπτώσεων, από τους οποίους συγκροτούνται (βλέπε Πόρισμα 1). Με βάση τους λόγους *g*, το πληροφοριακό περιεχόμενο (αδράνεια) του πρώτου υποπίνακα είναι ίσο με το 10,5% της ενδιαφέρουσας αδράνειας του πίνακα *Burt*, του δεύτερου με το 144% και του τρίτου με το 146%. Επομένως, σύμφωνα με το προτεινόμενο κριτήριο ως "καλύτερος" υποπίνακας επιλέγεται ο $\mathbf{\Gamma}_{6\times9}$ με g_{Γ} =146%.

Για τον έλεγχο της αποτελεσματικότητας της μεθόδου εφαρμόσαμε την ΠΑΑ τόσο στον πίνακα *Burt* όσο και στον επιλεγμένο υποπίνακα-φέτα $\mathbf{\Gamma}_{6\times9}$. Πήραμε τα αποτελέσματα που παρουσιάζονται στον Πίνακα 4.6, στα Διαγράμματα 4.1 και 4.2 και στο Σχήμα 4.1.

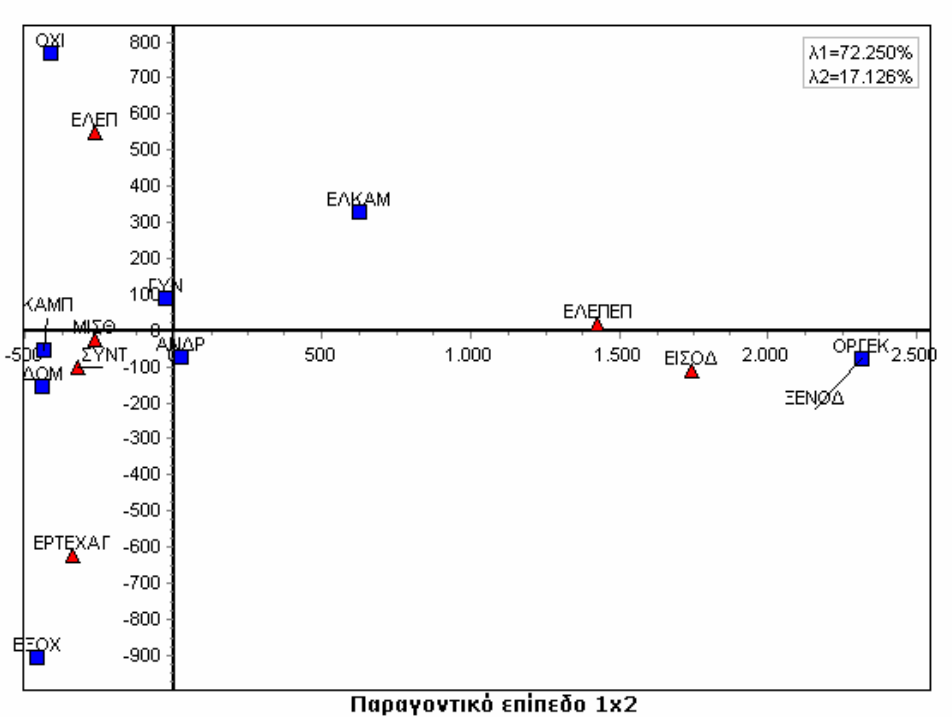
Πίνακας 4.6: Αποτελέσματα Εφαρμογής της ΠΑΑ στους Πίνακες *Burt* και **Γ**

Πίνακας *Burt* (15×15)				Υποπίνακας **Γ** (6×9)			
Άξονας	Αδράνεια	%	Αθρ. %	Άξονας	Αδράνεια	%	Αθρ. %
1	0,417	25,925	25,925	1	0,439	72,250	72,250
2	0,237	14,694	40,619	2	0,104	17,126	89,376
3	0,194	12,049	52,668	3	0,039	6,488	95,864
4	0,158	9,783	62,450	4	0,022	3,646	99,510
5	0,139	8,647	71,097	5	0,003	0,490	100,000

Από τον Πίνακα 4.6 παρατηρούμε ότι στην περίπτωση της εφαρμογής της ΠΑΑ στον πίνακα *Burt* οι δύο πρώτοι άξονες ερμηνεύουν το 40,6%, ενώ στην περίπτωση του επιλεγμένου υποπίνακα $\mathbf{\Gamma}_{6\times9}$ το 89,4% της ολικής αδράνειας. Αν εξετάσουμε την απεικόνιση των σημείων επί των παραγοντικών επίπεδων (βλέπε Διαγράμματα 4.1 και 4.2), που σχηματίζουν ο πρώτος με το δεύτερο άξονα, θα διαπιστώσουμε ότι οι σχετικές θέσεις των σημείων είναι σχεδόν ίδιες και για τις δύο αναλύσεις.



Διάγραμμα 4.1: Παραγοντικό Επίπεδο 1×2 από την Ανάλυση του *Burt* (Σύνολο Δεδομένων *A*)



Διάγραμμα 4.2: Παραγοντικό Επίπεδο 1×2 από την Ανάλυση του Υποπίνακα **Γ** (Σύνολο Δεδομένων *A*)

$I_{0-1} = \frac{13}{3} - 1 = 3,333$, από τη σχέση [2.60]

$I_{\varepsilon B} = \frac{0,057 + 0,03 + 1,160}{3} = \frac{1,247}{3} = 0,416$, από τη σχέση [4.34]

$I_B = \frac{3,333}{3} + \frac{2}{3}\left(\frac{1,247}{3}\right) = 1,388$, από τη σχέση [4.28]

Υποπίνακας $\mathbf{A_{2\times13}}$

	Επάγγελμα (7 κλάσεις)	Διακοπές (6 κλάσεις)
Φύλο (2 κλάσεις)	0,057 (p>0,05) *	0,03 (p>0,05)
	χ^2(6)=7,866, Cramer's V=0,239	χ^2(5)=4,14, Cramer's V=0,173

$g_{\mathbf{A}} = \frac{I_{\mathbf{A}}}{I_{\varepsilon B}} = \frac{\frac{0,057 + 0,03}{2}}{0,416} = \frac{0,0435}{0,416} = 0,105$ ή $10,5\%$,

όπου $I_{\mathbf{A}}$ η αδράνεια του $\mathbf{A_{2\times13}}$

Υποπίνακας $\mathbf{B_{7\times8}}$

	Φύλο (2 κλάσεις)	Επάγγελμα (7 κλάσεις)
Διακοπές (6 κλάσεις)	0,03	1,160 (p<0,05)
		χ^2(30)=160,08, Cramer's V=0,482

$g_{\mathbf{B}} = \frac{I_{\mathbf{B}}}{I_{\varepsilon B}} = \frac{\frac{0,03 + 1,160}{2}}{0,416} = \frac{0,60}{0,416} = 1,44$ ή 144%,

όπου $I_{\mathbf{B}}$ η αδράνεια του $\mathbf{B_{7\times8}}$

Υποπίνακας $\mathbf{\Gamma_{6\times9}}$

	Διακοπές (6 κλάσεις)	Φύλο (2 κλάσεις)
Επάγγελμα (7 κλάσεις)	1,160 (p<0,05)	0,057 (p>0,05)

$g_{\Gamma} = \frac{I_{\Gamma}}{I_{\varepsilon B}} = \frac{\frac{1,160 + 0,057}{2}}{0,416} = \frac{0,61}{0,416} = 1,46$ ή 146%,

όπου I_{Γ} η αδράνεια του $\mathbf{\Gamma_{6\times9}}$

*Στα κελιά του υποπίνακα εμφανίζεται η αδράνεια του αντίστοιχου απλού πίνακα συμπτώσεων και μέσα σε παρένθεση η παρατηρούμενη στάθμη σημαντικότητας (p-value) του ελέγχου ανεξαρτησίας χ^2 των δύο μεταβλητών.

Σχήμα 4.1: Αποτελέσματα της Εφαρμογής της Προτεινόμενης Μεθοδολογίας για το Σύνολο Δεδομένων *A*

Από το Σχήμα 4.1 παρατηρούμε ότι η ολική αδράνεια του πίνακα *Burt* είναι ίση με 1,388, ενώ η αδράνεια του αντίστοιχου λογικού πίνακα είναι ίση 3,333. Είναι φανερό ότι ο λογικός πίνακας περιέχει περισσότερη πληροφορία, με την έννοια της αδράνειας, απ' ότι ο αντίστοιχος πίνακας *Burt* (βλέπε Παρατήρηση 4.6). Δεδομένου ότι ο μέγιστος αριθμός αξόνων και στις δύο περιπτώσεις είναι ίδιος (p=13-3=10), είναι αναμενόμενο ότι η διάχυση της αδράνειας στους παραγοντικούς άξονες θα οδηγήσει σε διαφορετικά ποσοστά ερμηνείας στις δύο περιπτώσεις. Ειδικότερα, κατά την εφαρμογή της ΠΑΑ στο λογικό πίνακα, που αντιστοιχεί στα δεδομένα του συνόλου *A*, οι δύο πρώτοι άξονες ερμηνεύουν το 28,3% της ολικής αδράνειας I_{0-1} (16,15% ο πρώτος και 12,15% ο δεύτερος), ενώ στην ανάλυση του *Burt* οι δύο πρώτοι άξονες αιτιολογούν το 40,6% της αδράνειας I_B (25,9% ο πρώτος και 14,7% ο δεύτερος). Όταν όμως η ΠΑΑ εφαρμοστεί στον "καλύτερο" υποπίνακα $\mathbf{\Gamma}_{\mathbf{6\times9}}$, τότε επί του παραγοντικού επιπέδου 1×2 ερμηνεύεται το 89,35% της ολικής αδράνειας του πίνακα (72,25% στον πρώτο και 17,10% στον δεύτερο άξονα). Συνεπώς, έχουμε μια σημαντική αύξηση της ποιότητας της πληροφορίας καθώς το μέγεθος του πίνακα που αναλύεται μικραίνει. Όμως, και στις τρεις περιπτώσεις το υπό εξέταση φαινόμενο είναι στην ουσία το ίδιο. Το μόνο που αλλάζει είναι ο πίνακας περιγραφής του.

Από το Σχήμα 4.1 φαίνεται, επίσης, ότι η μεταβλητή «φύλλο» έχει μικρή συσχέτιση με τις μεταβλητές «επάγγελμα» και «διακοπές» συγκρινόμενη με τη συσχέτιση που έχουν αυτές οι δύο μεταβλητές μεταξύ τους. Επίσης, οι απλοί πίνακες συμπτώσεων, στους οποίους συμμετέχει το «φύλο», παρουσιάζουν αμελητέα αδράνεια σε σχέση με την αδράνεια του πίνακα που διασταυρώνει τις άλλες δύο μεταβλητές. Αυτό έχει ως αποτέλεσμα, όλη σχεδόν η σημαντική πληροφορία να ενθυλακώνεται τελικά στον απλό πίνακα συμπτώσεων των μεταβλητών «επάγγελμα» και «διακοπές». Η μεταβλητή «φύλο» πρακτικά δεν επηρεάζει τα αποτελέσματα και θα μπορούσε να απομακρυνθεί από την ανάλυση. Μάλιστα, αν θεωρήσουμε ότι τα δεδομένα προέρχονται από τυχαία δειγματοληψία, διαπιστώνουμε και πάλι ότι η μόνη στατιστικά σημαντική συσχέτιση, σε επίπεδο σημαντικότητας α=0,05, είναι αυτή μεταξύ των μεταβλητών «επάγγελμα» και «διακοπές» (p<0,05). Το φύλο δεν παρουσιάζει στατιστικά σημαντική συσχέτιση με καμία από τις δύο άλλες μεταβλητές (p>0,05). Εφόσον οι απλοί πίνακες συμπτώσεων που περιλαμβάνουν το «φύλο» δεν παρέχουν επιπλέον πληροφορία, για το συγκεκριμένο παράδειγμα, οι υποπίνακες $\mathbf{\Gamma}_{\mathbf{6\times9}}$

και $\mathbf{B}_{7\times8}$ είναι ισοδύναμοι. Αποδίδουν και οι δύο σχεδόν την ίδια εικόνα με αυτή που προκύπτει από την εφαρμογή της ΠΑΑ στον αρχικό πίνακα *Burt*. Άλλωστε, και οι λόγοι $g_{\mathbf{B}}$ και g_{Γ} δεν διαφέρουν σημαντικά.

Β) Σύνολο δεδομένων *B* με γνωστό *a priori* "καλύτερο" υποπίνακα.

Ομοίως, εφαρμόσαμε την προτεινόμενη μέθοδο στον 15×15 πίνακα *Burt*, που προέκυψε από το σύνολο δεδομένων *B* (βλέπε αρχείο data_files.xls στο Παράρτημα CDA του CD που συνοδεύει τη διατριβή), το οποίο περιλαμβάνει 360 αντικείμενα και 4 μεταβλητές με 15 συνολικά κλάσεις. Το σύνολο *B* κατασκευάστηκε έτσι ώστε ο "καλύτερος" υποπίνακας να είναι γνωστός από πριν. Τα δεδομένα δημιουργήθηκαν με τρόπο ώστε μόνο ορισμένοι απλοί πίνακες συμπτώσεων του *Burt* να έχουν σημαντικές αδράνειες. Έτσι, προέκυψε ως "καλύτερος" ο 9×6 υποπίνακας, έστω $\mathbf{K_2}$, με τις μεταβλητές X_1 και X_4 στις γραμμές και τις X_2 και X_3 στις στήλες. Ο αντίστοιχος λόγος g βρέθηκε ίσος με 146%. Επομένως, κατά την ερμηνεία των αποτελεσμάτων θα πρέπει η προσοχή του αναλυτή να στραφεί στις συσχετίσεις της μεταβλητής X_1 με τις X_2 και X_3 και της X_4 με τις X_2 και X_3 και όχι στη συνάφεια των X_2 και X_3.

Τα αποτελέσματα της ΠΑΑ για τους πίνακες *Burt* και $\mathbf{K_2}$ δίνονται στον Πίνακα 4.7 και στα Διαγράμματα 4.3 και 4.4.

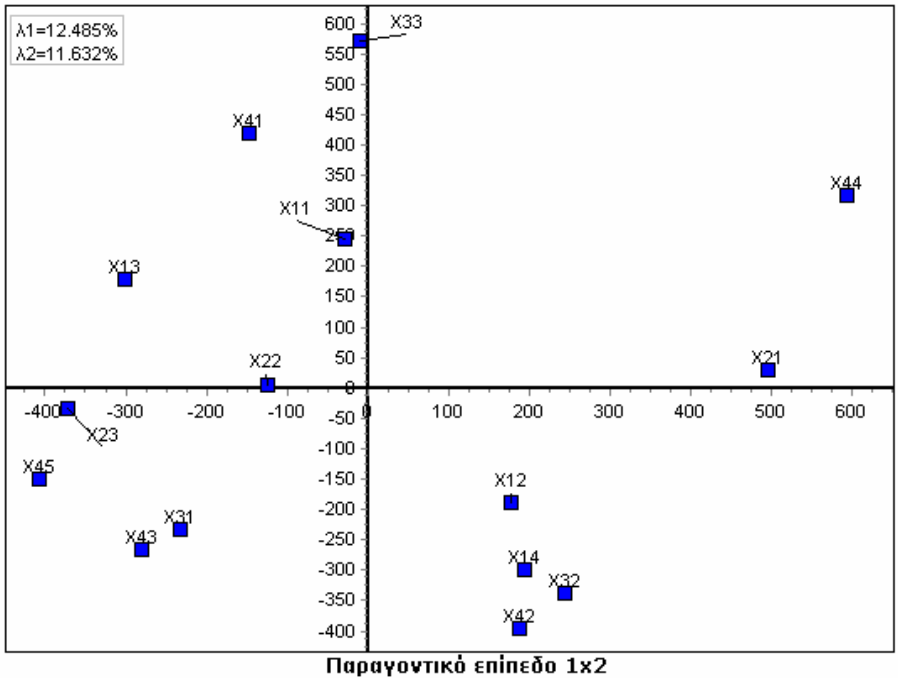
Πίνακας 4.7: Αποτελέσματα Εφαρμογής της ΠΑΑ στους Πίνακες *Burt* και $\mathbf{K_2}$

Πίνακας *Burt* (15×15)				Υποπίνακας $\mathbf{K_2}$ (9×6)			
Άξονας	Αδράνεια	%	Αθρ. %	Άξονας	Αδράνεια	%	Αθρ. %
1	0,087	12,485	12,485	1	0,008	50,778	50,778
2	0,081	11,632	24,117	2	0,005	32,632	83,411
3	0,076	10,932	35,050	3	0,002	14,776	98,186
4	0,067	9,702	44,751	4	0,000	1,814	100,000
5	0,065	9,418	54,169	5	0,000	0,000	100,000
6	0,061	8,827	62,996				
7	0,060	8,575	71,571				
8	0,058	8,318	79,890				
9	0,052	7,448	87,338				
10	0,046	6,558	93,896				

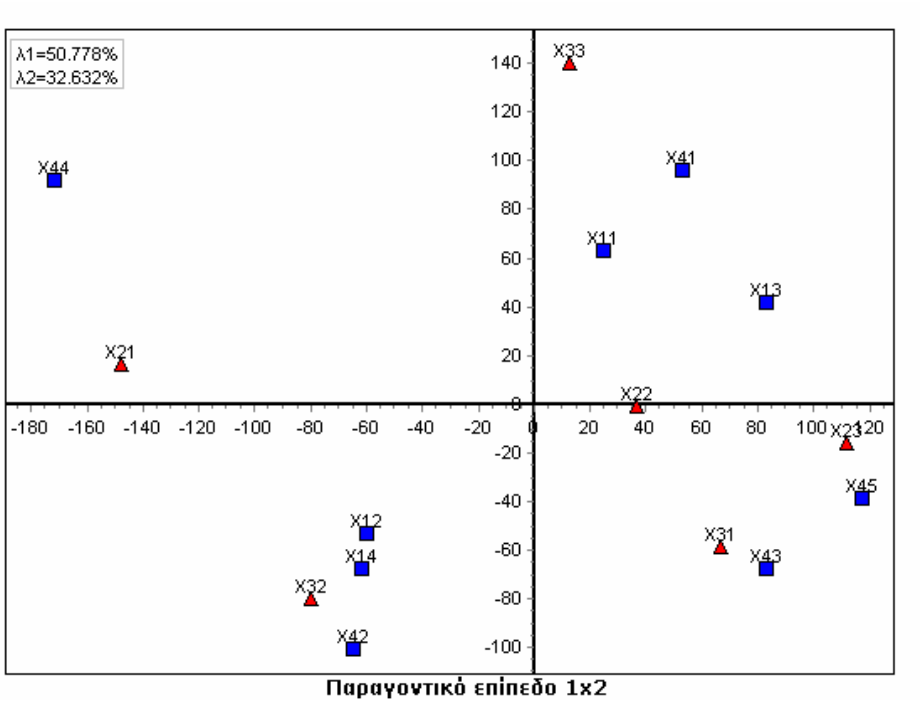
Παρατηρούμε ότι η εφαρμογή της ΠΑΑ στον πίνακα *Burt* δίνει αρκετά χαμηλά ποσοστά ερμηνείας της ολικής αδράνειας σε όλους τους παραγοντικούς άξονες, με τις αντίστοιχες αδράνειες να μη διαφέρουν σημαντικά (βλέπε Πίνακα 4.7). Αυτό καθιστά σχεδόν απαγορευτική την εφαρμογή της μεθόδου στον παραπάνω πίνακα, με την

έννοια ότι κάθε άξονας εκφράζει μια διαφορετική “τοπική” και όχι καθολική λανθάνουσα δομή, η οποία μπορεί να ερμηνευτεί από συγκεκριμένες, λίγες σε πλήθος, μεταβλητές, οι οποίες δεν συσχετίζονται σημαντικά με τις υπόλοιπες.

Στην περίπτωση, όμως, της ανάλυσης του επιλεγμένου υποπίνακα, οι δύο πρώτοι άξονες ερμηνεύουν το 83,4% της ολικής. Γεγονός αναμενόμενο καθώς η πληροφορία συγκεντρώνεται σε μικρότερο αριθμό παραγοντικών αξόνων. Από τα Διαγράμματα 4.3 και 4.4. διαπιστώνουμε και πάλι ότι οι σχετικές θέσεις των σημείων είναι σχεδόν ίδιες και στις δύο αναλύσεις.



Διάγραμμα 4.3: Παραγοντικό Επίπεδο 1×2 από την Ανάλυση του *Burt* (Σύνολο Δεδομένων *B*)



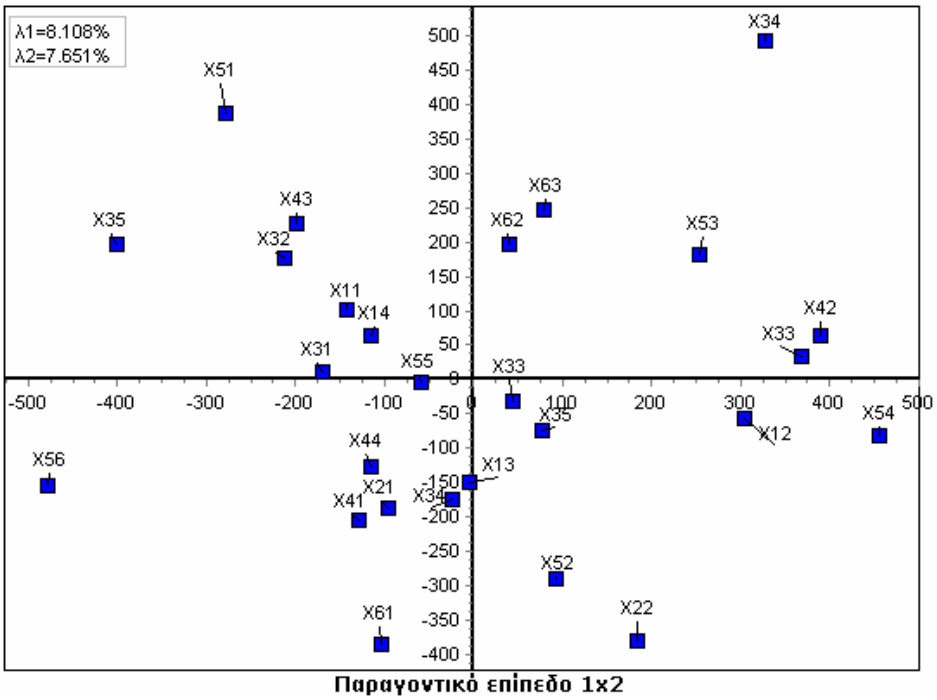
Διάγραμμα 4.4: Παραγοντικό Επίπεδο 1×2 από την Ανάλυση του Υποπίνακα **K₂** (Σύνολο Δεδομένων *B*)

Γ) Σύνολο δεδομένων *Γ* που κατασκευάστηκε με γεννήτρια τυχαίων αριθμών

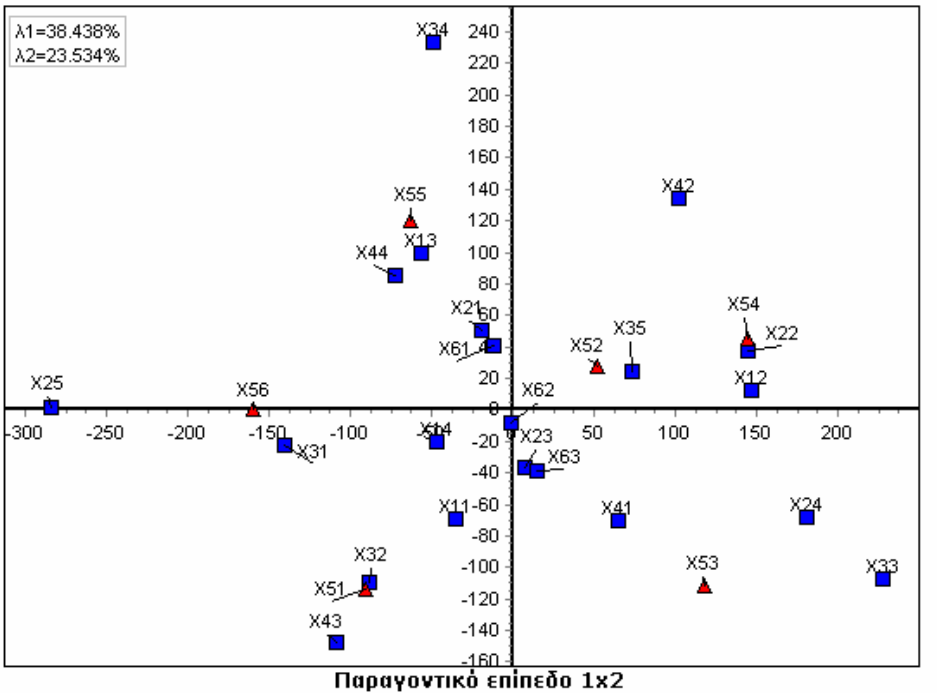
Το τρίτο σύνολο δεδομένων *Γ* (βλέπε αρχείο data_files.xls στο Παράρτημα CDA του CD που συνοδεύει τη διατριβή), με 500 αντικείμενα και 6 μεταβλητές με 27 συνολικά κλάσεις, προέκυψε από γεννήτρια ψευδοτυχαίων αριθμών, μέσω του λογισμικού Excel (Blattner, 2001). Η εφαρμογή της μεθόδου στον αντίστοιχο 27×27 πίνακα *Burt* ανέδειξε ως "καλύτερο" τον 21×6 υποπίνακα "φέτα", έστω **K₃**, με τη μεταβλητή X_5 στις στήλες και τις υπόλοιπες στις γραμμές (g=138%).

Πίνακας 4.8: Αποτελέσματα Εφαρμογής της ΠΑΑ στους Πίνακες *Burt* και **K₃**

Πίνακας *Burt* (27×27)				Υποπίνακας **K₃** (21×6)			
Άξονας	Αδράνεια	%	Αθρ. %	Άξονας	Αδράνεια	%	Αθρ. %
1	0,049	8,108	8,108	1	0,012	38,438	38,438
2	0,046	7,651	15,759	2	0,007	23,534	61,971
3	0,043	7,231	22,990	3	0,006	19,739	81,710
4	0,039	6,460	29,450	4	0,003	9,608	91,318
5	0,037	6,078	35,528	5	0,003	8,682	100,000



Διάγραμμα 4.5: Παραγοντικό Επίπεδο 1×2 από την Ανάλυση του *Burt* (Σύνολο Δεδομένων *Γ*)



Διάγραμμα 4.6: Παραγοντικό Επίπεδο 1×2 από την Ανάλυση του Υποπίνακα $\mathbf{K_3}$ (Σύνολο Δεδομένων *Γ*)

Η εφαρμογή της ΠΑΑ στον πίνακα *Burt* έδωσε αρκετά χαμηλό ποσοστό ερμηνεύσιμης αδράνειας στους δύο πρώτους παραγοντικούς άξονες ίσο με 15,8%

(βλέπε Πίνακα 4.8). Όταν όμως η μέθοδος εφαρμόστηκε στον επιλεγμένο υποπίνακα το ποσοστό αυξήθηκε σε 62% για τους δύο πρώτους άξονες. Η εικόνα του φαινομένου στα παραγοντικά επίπεδα των δύο αναλύσεων είναι σε γενικές γραμμές η ίδια, ενώ παρατηρήθηκε εναλλαγή των σημαντικότερων σημείων μεταξύ 2ου και 3ου άξονα. Η εναλλαγή μπορεί να αιτιολογηθεί λόγω της αστάθειας των παραγοντικών αξόνων που οφείλεται στη μικρή διαφορά των αδρανειών τους. Συμπερασματικά, καλό είναι, στη συγκεκριμένη περίπτωση, η ερμηνεία του φαινομένου να μην περιοριστεί μόνο στους δύο πρώτους παραγοντικούς άξονες.

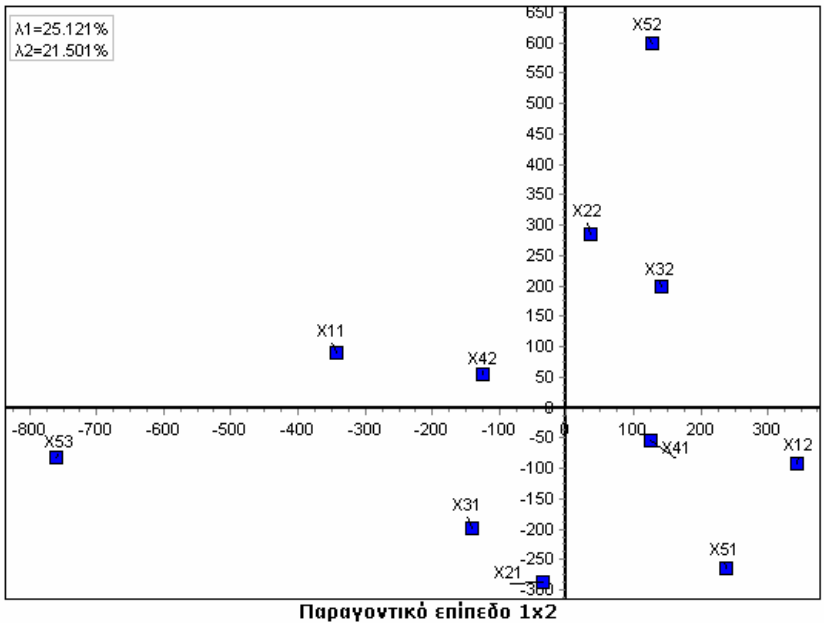
Δ) Σύνολο δεδομένων *Δ* όπου μόνο μία απ' τις μεταβλητές συσχετίζεται με όλες τις υπόλοιπες, οι οποίες είναι μεταξύ τους ασυσχέτιστες

Το σύνολο δεδομένων *Δ* (βλέπε αρχείο data_files.xls στο Παράρτημα CDA του CD που συνοδεύει τη διατριβή) κατασκευάστηκε έτσι ώστε οι 4 από τις συνολικά 5 μεταβλητές να είναι μεταξύ τους τελείως ασυσχέτιστες. Οι αδράνειες των αντίστοιχων πινάκων συμπτώσεων ήταν ίσες με μηδέν. Το σύνολο περιλαμβάνει 1600 αντικείμενα και 5 μεταβλητές με 11 συνολικά κλάσεις. Ο καλύτερος υποπίνακας, έστω $\mathbf{K_4}$, είναι τύπου "στοίβας" διαστάσεων 8×3 με τη μεταβλητή X_5 στις στήλες και τις υπόλοιπες στις γραμμές (g=249%). Κατά την εφαρμογή της μεθόδου στον πίνακα *Burt*, το ποσοστό ερμηνείας της ολικής αδράνειας για τους δύο πρώτους άξονες είναι 46,6%, ενώ φθάνει το 100% στην περίπτωση του επιλεγμένου υποπίνακα (βλέπε Πίνακα 4.9).

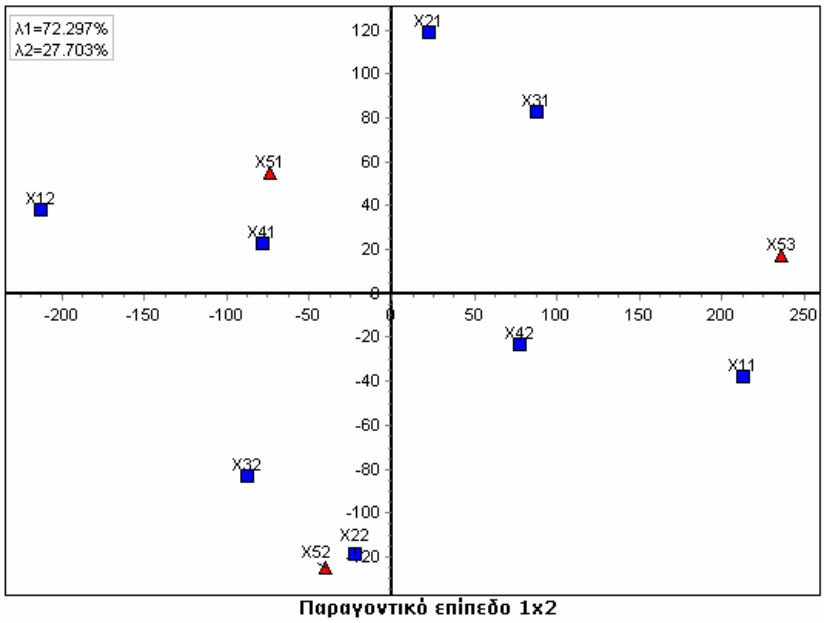
Πίνακας 4.9: Αποτελέσματα Εφαρμογής της ΠΑΑ στους Πίνακες *Burt* και $\mathbf{K_4}$

Πίνακας *Burt* (11×11)				Υποπίνακας $\mathbf{K_4}$ (8×3)			
Άξονας	Αδράνεια	%	Αθρ. %	Άξονας	Αδράνεια	%	Αθρ. %
1	0,062	25,121	25,121	1	0,015	72,297	72,297
2	0,053	21,501	46,622	2	0,006	27,703	100,000
3	0,040	16,220	62,842				
4	0,040	16,220	79,061				
5	0,029	11,681	90,743				

Η εικόνα του φαινομένου και για τις δύο αναλύσεις είναι σχεδόν η ίδια (βλέπε Διαγράμματα 4.7 και 4.8). Το σημαντικότερο πλεονέκτημα που προκύπτει είναι ότι στο παραγοντικό επίπεδο της εφαρμογής της μεθόδου στον "καλύτερο" υποπίνακα, αποφεύγεται η ερμηνεία σχέσεων μεταξύ μεταβλητών με ελάχιστη ή καθόλου συσχέτιση.



Διάγραμμα 4.7: Παραγοντικό Επίπεδο 1×2 από την Ανάλυση του *Burt* (Σύνολο Δεδομένων *Δ*)



Διάγραμμα 4.8: Παραγοντικό Επίπεδο 1×2 από την Ανάλυση του Υποπίνακα $\mathbf{K_4}$ (Σύνολο Δεδομένων *Δ*)

4.8.1.1 Παρατηρήσεις

Η προτεινόμενη μέθοδος παρέχει σημαντική βοήθεια κατά την ερμηνεία των αποτελεσμάτων της ΠΑΑ. Συγκεκριμένα, αποφεύγεται η ερμηνεία σχέσεων μεταξύ μεταβλητών με ελάχιστη ή καθόλου συνάφεια και επιτυγχάνεται συγκέντρωση ουσιαστικότερης πληροφορίας σε μικρότερο αριθμό αξόνων. Άλλωστε, ο Greenacre (1984) έδειξε ότι αν *Q* κατηγορικές μεταβλητές μπορούν να διαμεριστούν σε δύο

υποσύνολα με *Q*1 και *Q*2 μεταβλητές αντίστοιχα, τέτοια ώστε οι μεταβλητές μέσα σε κάθε υποσύνολο να είναι ανά δύο ανεξάρτητες ή ασυσχέτιστες, τότε η εφαρμογή της ΠΑΑ στις *Q* μεταβλητές είναι ισοδύναμη με την εφαρμογή της στον πίνακα συμπτώσεων που δημιουργείται όταν οι κλάσεις των *Q*1 μεταβλητών τοποθετηθούν στις γραμμές και οι κλάσεις των *Q*2 μεταβλητών τοποθετηθούν στις στήλες του πίνακα (ή αντίστροφα) (βλέπε επίσης Lebart, Morineau & Warwick 1984, Παπαδημητρίου 2006).

Από την εφαρμογή της μεθόδου και σε άλλα σύνολα δεδομένων προέκυψαν αρκετά εμπειρικά ευρήματα. Αναφέρουμε τα σημαντικότερα:

Όταν οι αδράνειες μεταξύ 2ου και 3ου άξονα δεν διαφέρουν σημαντικά, παρατηρήθηκε εναλλαγή των σημαντικότερων σημείων των δύο αξόνων. Πιο συγκεκριμένα, σημεία τα οποία αρχικά, κατά την ανάλυση του *Burt*, ήταν σημαντικά στο δεύτερο άξονα, εμφανίζονται να είναι σημαντικά στον τρίτο (και το αντίστροφο) όταν η ΠΑΑ εφαρμόζεται στον επιλεγμένο υποπίνακα. Επομένως, η ερμηνεία της ΠΑΑ στον "καλύτερο" υποπίνακα δεν θα πρέπει να περιορίζεται μόνο στους δύο πρώτους άξονες, ιδιαίτερα όταν το αθροιστικό ποσοστό ερμηνείας της ολικής αδράνειας είναι χαμηλό και οι αδράνειες των πρώτων σε τάξη αξόνων δεν παρουσιάζουν απότομη μεταβολή σε σχέση με τις αδράνειες των υπόλοιπων. Γενικά, διαπιστώθηκε ότι σημεία, τα οποία με βάση τους δείκτες *CTR* και *COR* είναι σημαντικά κατά την ανάλυση του πίνακα *Burt*, εξακολουθούν να είναι σημαντικά και κατά την ανάλυση του "καλύτερου" υποπίνακα. Λόγω του γεγονότος ότι κατά την ανάλυση του "καλύτερου" υποπίνακα η πληροφορία συγκεντρώνεται σε μικρότερο αριθμό αξόνων, διαπιστώθηκε, σε αρκετές περιπτώσεις, η απεικόνιση επί του παραγοντικού επιπέδου να είναι λεπτομερέστερη απ' ότι κατά την ανάλυση του πίνακα *Burt*. Σημεία, τα οποία στην περίπτωση του *Burt* αναδεικνύονται σε άξονες μικρής ερμηνευτικής ικανότητας, εμφανίζονται πλέον ως σημαντικά στους δύο πρώτους παραγοντικούς άξονες του "καλύτερου" υποπίνακα.

Επειδή σε κάθε περίπτωση υπάρχει ένας υποπίνακας που να μεγιστοποιεί το λόγο *g*, από τα εμπειρικά ευρήματα των πειραματισμών μας φάνηκε ότι αν υπάρχει υποπίνακας με την πλησιέστερη απεικόνιση, μέσω της ΠΑΑ, στον αντίστοιχο πίνακα *Burt*, αυτός θα πρέπει να αναζητηθεί στον υποπίνακα με το μεγαλύτερο λόγο *g*.

Βέβαια, περαιτέρω δοκιμές θα μπορούσαν να ενισχύσουν τον προηγούμενο ισχυρισμό και γενικότερα την αποτελεσματικότητα της προτεινόμενης μεθοδολογίας.

4.8.2 Στατιστική Σημαντικότητα της Ενδιαφέρουσας Αδράνειας

Στην περίπτωση που τα διαθέσιμα δεδομένα αποτελούν τυχαίο δείγμα, τότε από τη σχέση [4.34] μέσω της [4.31] έχουμε ότι:

$$I_{\varepsilon B} = \frac{\sum_{h<w} I_{hw}}{\frac{q(q-1)}{2}} = \frac{\frac{\sum_{h<w} Q_{hw}}{N}}{\frac{q(q-1)}{2}} = \frac{\sum_{h<w} Q_{hw}}{\frac{q(q-1)}{2} N}, \quad h, w = 1, \ldots, q$$

όπου $\sum_{h<w} Q_{hw}$ είναι το άθροισμα των στατιστικών χ^2 που αντιστοιχούν στους $q(q-1)/2$ διαφορετικούς απλούς πίνακες συμπτώσεων των q μεταβλητών ανά δύο.

Κάτω από την υπόθεση της ανεξαρτησίας των q μεταβλητών ανά δύο η ποσότητα $\sum_{h<w} Q_{hw}$ ακολουθεί ασυμπτωτικά την Κατανομή χ^2 με $\frac{1}{2}\left[(j-q)^2 - \sum_{i=1}^{q} (k_i - 1)^2 \right]$ βαθμούς ελευθερίας (β.ε.) (Bekker & De Leeuw 1988, Van der Heijden & De Leeuw 1989, Gifi 1996), όπου j είναι το συνολικό πλήθος των κλάσεων των q μεταβλητών και k_i ο αριθμός των κατηγοριών της μεταβλητής i (i=1,…, q). Επομένως, η στατιστική σημαντικότητα της ενδιαφέρουσας αδράνειας $I_{\varepsilon B}$ μπορεί να ελεγχθεί συγκρίνοντας το στατιστικό

$$Q^* = \frac{q(q-1)}{2} N I_{\varepsilon B}$$

με την κρίσιμη τιμή της Κατανομής χ^2 με $\frac{1}{2}\left[(j-q)^2 - \sum_{i=1}^{q} (k_i - 1)^2 \right]$ β.ε., σε επίπεδο σημαντικότητας α. Μια ισοδύναμη προσέγγιση είναι η σύγκριση της παρατηρούμενης στάθμης σημαντικότητας (p-value) (βλέπε Κεφάλαιο 5, Ενότητα 5.5) του στατιστικού

$\sum_{h<w} Q_{hw}$ με την προκαθορισμένη τιμή του *α*. Προφανώς, στην περίπτωση που οι *q* μεταβλητές είναι ανά δύο ασυσχέτιστες, τότε $I_{\varepsilon B} = 0$ (ή πολύ κοντά στο μηδέν).

Με βάση τα προηγούμενα, οι στατιστικές υποθέσεις που μπορούν να ελεγχθούν είναι οι εξής:

Μηδενική Υπόθεση	H_0: $I_{\varepsilon B} = 0$
Εναλλακτική Υπόθεση	H_1: $I_{\varepsilon B} \neq 0$, σε επίπεδο σημαντικότητας *α*.

Προϋποθέσεις για την εγκυρότητα του προτεινόμενου ελέγχου στατιστικής σημαντικότητας της ενδιαφέρουσας αδράνειας του πίνακα *Burt* είναι οι εξής: α) τα δεδομένα να έχουν συγκεντρωθεί με απλή τυχαία δειγματοληψία, β) το μέγεθος του δείγματος να είναι αρκούντος μεγάλο, ώστε να ισχύει η ασυμπτωτική προσέγγιση της ποσότητας $\sum_{h<w} Q_{hw}$ από την Κατανομή χ^2 και γ) οι αναμενόμενες συχνότητες των κελιών σε κάθε έναν από τους $q(q-1)/2$ απλούς πίνακες συμπτώσεων, που συγκροτούν τον *Burt*, να είναι εν γένει ≥5, ώστε τα στατιστικά Q_{hw} να ακολουθούν ασυμπτωτικά την Κατανομή χ^2 (Κάτος 1986, Hinkle, Wiersma & Jurs 1988, Fienberg 1991).

<u>Παράδειγμα Υπολογισμών</u>

Για τα δεδομένα του συνόλου *A*, που χρησιμοποιήσαμε στην προηγούμενη ενότητα, έχουμε:

N=138, *q*=3, *j*=6+7+2=15, $I_{\varepsilon B}$=0,416 και έστω ότι προκαθορίζουμε το επίπεδο σημαντικότητας του ελέγχου σε *α*=0,05.

Έχουμε, επίσης,

$$\sum_{i=1}^{3} (k_i - 1)^2 = (6-1)^2 + (7-1)^2 + (2-1)^2 = 62$$

και

$$\text{β.ε.} = \frac{1}{2}\left[(j-q)^2 - \sum_{i=1}^{q} (k_i - 1)^2 \right] = \frac{1}{2}\left[(15-3)^2 - 62 \right] = 41 .$$

Συνεπώς, η ποσότητα Q^* θα είναι ίση με:

$$Q^* = \frac{3(3-1)}{3} 138 \times 0,416 = 172,224 > 56,94 = \chi^2_{critical;a=0,05}(41) \text{ και } p=0,000 < \alpha=0,05.$$

Επειδή η παρατηρούμενη τιμή του στατιστικού Q^* είναι μεγαλύτερη από την κρίσιμη τιμή της Κατανομής χ^2 με 41 β.ε., σε επίπεδο σημαντικότητας α=0,05, απορρίπτεται η Μηδενική Υπόθεση και, συνεπώς, η ενδιαφέρουσα αδράνεια του αντίστοιχου πίνακα *Burt* είναι στατιστικά σημαντική σε α=0,05, εφόσον βέβαια ικανοποιούνται οι προϋποθέσεις α), β) και γ) του ελέγχου. Το προηγούμενο αποτέλεσμα οδηγεί στο συμπέρασμα ότι το πληροφοριακό περιεχόμενο του πίνακα *Burt*, όπως αυτό μετριέται μέσω της $I_{\varepsilon B}$, διαφέρει στατιστικά σημαντικά από το μηδέν, δηλαδή δεν είναι αποτέλεσμα τυχαίων διακυμάνσεων ή επιδράσεων, και, επομένως, έχει αξία να προχωρήσουμε σε περαιτέρω ανάλυση των συσχετίσεων των μεταβλητών μέσω της ΠΑΑ.

4.8.3 Πρόταση Μεθόδου Διόρθωσης των Αδρανειών του Πίνακα *Burt*

Στην Ενότητα 4.6.3 (βλέπε Παρατήρηση 4.8) αναφερθήκαμε σε δύο μεθόδους διόρθωσης των αδρανειών των παραγοντικών αξόνων, οι οποίοι προκύπτουν από την εφαρμογή της πολυμεταβλητής εκδοχής της ΠΑΑ επί του γενικευμένου πίνακα συμπτώσεων απολύτων συχνοτήτων q μεταβλητών (πίνακας *Burt*). Η πρώτη διόρθωση (ΔB) οφείλεται στο Benzécri (1979), ενώ η δεύτερη (ΔG) στον Greenacre (2005, 1994β, 1993α και 1984). Βασικός στόχος και των δύο μεθόδων διόρθωσης είναι η βελτίωση της ποιότητας της λύσης της ΠΑΑ, με την έννοια της αύξησης του ποσοστού της αδράνειας που ερμηνεύουν οι παραγοντικοί άξονες, ώστε να μη δημιουργείται η αίσθηση φτωχής προσαρμογής των δεδομένων και χαμηλής ποιότητας της λύσης της ΠΑΑ. Επιπλέον κίνητρο για τη πρόταση του Greenacre είναι η διόρθωση του προβλήματος που δημιουργούν οι διαγώνιοι πίνακες και οι ανάστροφοι των απλών πινάκων συμπτώσεων που συγκροτούν τον πίνακα *Burt*, κατά

τη γεωμετρική ερμηνεία των αποστάσεων χ^2 μεταξύ των προβαλλόμενων σημείων επί των παραγοντικών επιπέδων (Greenacre 2005, 1994β, 1993γ, 1993α, 1991, 1990, 1989 και 1988α). Οι διορθώσεις *ΔB* και *ΔG* έχουν ως αποτέλεσμα να τροποποιούνται τόσο οι συντεταγμένες των προβολών των σημείων στους παραγοντικούς άξονες όσο και τα άλλα αριθμητικά αποτελέσματα της ΠΑΑ, όπως για παράδειγμα οι δείκτες *CTR*, στον υπολογισμό των οποίων συμμετέχει η αδράνεια των αξόνων. Θα πρέπει να τονιστεί ότι και στις δύο περιπτώσεις οι σχετικές θέσεις των σημείων, επί των παραγοντικών επιπέδων, δεν μεταβάλλονται. Απλά, επιβάλλεται μια αλλαγή στην κλίμακα μέτρησης του συστήματος συντεταγμένων του υποχώρου προβολής (Greenacre, 2005 και 1994β).

Σύμφωνα με τη διόρθωση *ΔB* του Benzécri, οι αδράνειες των παραγοντικών αξόνων που προκύπτουν από την ανάλυση του πίνακα *Burt* θα πρέπει να διορθωθούν σύμφωνα με τη σχέση:

$$\lambda_{Bs}^{adj} = \left(\frac{q}{q-1} \right)^2 \times \left(\sqrt{\lambda_{Bs}} - \frac{1}{q} \right)^2 , \text{για } \sqrt{\lambda_{Bs}} > \frac{1}{q} \qquad (\Delta B)$$

όπου q είναι ο αριθμός των μεταβλητών, λ_{Bs} η αδράνεια του άξονα s που προκύπτει από την ανάλυση του πίνακα *Burt* και $\sqrt{\lambda_{Bs}}$ η αδράνεια του s από την ανάλυση του λογικού πίνακα **Z** (βλέπε Ενότητα 2.3.3.5).

Στη συνέχεια, οι διορθωμένες αδράνειες εκφράζονται ως ποσοστό του αθροίσματος των διορθωμένων αδρανειών των αξόνων, για τους οποίους η αντίστοιχη ιδιοτιμή $\sqrt{\lambda_{Bs}}$ είναι μεγαλύτερη από $1/q$ (βλέπε επίσης Rovan 1994, SAS Institute 1999, Guinot *et al*. 2001, Greenacre 2005 και 1984). Με τον τρόπο αυτό, τα ποσοστά % ερμηνείας των παραγοντικών αξόνων έχουν άθροισμα 100%.

Η διόρθωση *ΔG* του Greenacre συνίσταται στον ίδιο μετασχηματισμό με αυτόν του Benzécri με τη διαφορά οι διορθωμένες αδράνειες εκφράζονται ως ποσοστό της ποσότητας (Greenacre 2005, 1994β και 1993α):

$$I_m = \frac{q}{q-1}\left(I_B - \frac{j-q}{q^2} \right), \qquad [4.36]$$

όπου I_B είναι η ολική αδράνεια του πίνακα *Burt* και j το συνολικό πλήθος των κατηγοριών των q μεταβλητών. Με τη διόρθωση του Greenacre τα ποσοστά % ερμηνείας των παραγοντικών αξόνων έχουν άθροισμα μικρότερο από 100%.

Η προσέγγιση του Greenacre στηρίζεται στο εύρημα ότι οι διορθωμένες αδράνειες που προκύπτουν από την εφαρμογή της ΠΑΑ στον πίνακα *Burt* δύο κατηγορικών μεταβλητών, είναι ίσες με τις αδράνειες που προκύπτουν από την εφαρμογή της μεθόδου στον αντίστοιχο απλό πίνακα συμπτώσεων (Greenacre, 2005 και 1993α).

Θα δείξουμε ότι η ποσότητα I_m είναι ίση με την ενδιαφέρουσα αδράνεια $I_{\varepsilon B}$ του πίνακα *Burt*, την οποία ορίσαμε στην Ενότητα 4.7.

Απόδειξη:

Από τη σχέση [4.34] έχουμε:

$$I_{\varepsilon B} = \frac{\sum_{h<w} I_{hw}}{\frac{q(q-1)}{2}} = \frac{2\sum_{h<w} I_{hw}}{q(q-1)} \Rightarrow$$

$$\Rightarrow \frac{2\sum_{h<w} I_{hw}}{q} = (q-1) I_{\varepsilon B} \qquad [4.37]$$

Από τη σχέση [4.28] μέσω της [4.37] και της [2.60] προκύπτει:

$$I_B = \frac{I_{0-1}}{q} + \frac{2}{q}\left(\frac{\sum_{h<w} I_{hw}}{q} \right) = \frac{\frac{j-q}{q}}{q} + \frac{(q-1) I_{\varepsilon B}}{q} = \frac{j-q}{q^2} + \frac{(q-1) I_{\varepsilon B}}{q} \Rightarrow$$

$$\Rightarrow I_{\varepsilon B} = \frac{q}{q-1}\left(I_B - \frac{j-q}{q^2} \right) = I_m . \square \qquad [4.38]$$

Από την παραπάνω σχέση διαπιστώνουμε ότι η ποσότητα I_m εκφράζει τελικά τη μέση αδράνεια των μη διαγώνιων υποπινάκων του *Burt*, δηλαδή την ενδιαφέρουσα αδράνειά του.

Στις Ενότητες 2.3.3.5, 4.6.2 και 4.7 διαπιστώσαμε ότι η ΠΑΑ αναλύει διμεταβλητές σχέσεις και όχι πολυμεταβλητές. Η διαπίστωση αυτή είναι αληθινή είτε η ανάλυση εφαρμοστεί στον πίνακα *Burt* είτε στο λογικό πίνακα **Z**, λόγω της ισοδυναμίας των δύο αναλύσεων (βλέπε Ενότητες 2.3.3.5 και 3.3). Είδαμε, επίσης, ότι το τμήμα της ολικής αδράνειας του πίνακα *Burt*, το οποίο έχει ενδιαφέρον να αναλυθεί και περιέχει, στην ουσία, όλη την απαραίτητη πληροφορία που αφορά στις συσχετίσεις των μεταβλητών ανά δύο, είναι η ενδιαφέρουσα αδράνεια. Επομένως, αν συνδέσουμε τις αρχικές αδράνειες των παραγοντικών αξόνων (ή συναρτήσεις τους) με την ενδιαφέρουσα αδράνεια και επιτύχουμε μια διαμέρισή της ανά άξονα, θα είναι δυνατή η αναγωγή της «μερικής» ενδιαφέρουσας αδράνειας του κάθε άξονα ως ποσοστό είτε της (ολικής) ενδιαφέρουσας αδράνειας του πίνακα *Burt*, όπως συμβαίνει στη διόρθωση *ΔG*, είτε ως προς το διορθωμένο μέρος της που αντιστοιχεί στους σημαντικούς άξονες, όπως στην περίπτωση *ΔB*. Με βάση τον προηγούμενο συλλογισμό προτείνουμε, στη συνέχεια, μέθοδο διόρθωσης των αδρανειών των παραγοντικών αξόνων που προκύπτουν από την εφαρμογή της ΠΑΑ στον πίνακα *Burt q* κατηγορικών μεταβλητών.

Μπορεί να δειχθεί ότι (Gifi, 1996):

$$\sum_{h<w} Q_{hw} = \frac{1}{2} N \sum_{s=1}^{p} \left(q\sqrt{\lambda_{Bs}} - 1 \right)^2, \qquad [4.39]$$

όπου Q_{hw} είναι το στατιστικό χ^2 που αντιστοιχεί στον απλό πίνακα συμπτώσεων των μεταβλητών X_h και X_w (h, w=1,…, q), p=j-q και λ_{Bs} η αδράνεια του άξονα s από την ανάλυση του πίνακα *Burt*.

Πράγματι:

$$\frac{1}{2}N\sum_{s=1}^{p}\left(q\sqrt{\lambda_{Bs}}-1\right)^2=\frac{1}{2}N\sum_{s=1}^{p}\left(q^2\lambda_{Bs}-2q\sqrt{\lambda_{Bs}}+1\right)=\frac{1}{2}Nq^2\sum_{s=1}^{p}\lambda_{Bs}-\frac{1}{2}N2q\sum_{s=1}^{p}\sqrt{\lambda_{Bs}}+\frac{1}{2}Np=$$

$$=\frac{1}{2}Nq^2\left(\frac{j-q}{q^2}+\frac{2\sum_{h<w}I_{hw}}{q^2}\right)-\frac{1}{2}N2q\frac{j-q}{q}+\frac{1}{2}Np=$$

$$=\frac{1}{2}Nq^2\left(\frac{p}{q^2}+\frac{2\frac{\sum_{h<w}Q_{hw}}{N}}{q^2}\right)-\frac{1}{2}N2q\frac{p}{q}+\frac{1}{2}Np=$$

$$=\frac{1}{2}Np+\sum_{h<w}Q_{hw}-\frac{1}{2}Np=\sum_{h<w}Q_{hw}.\quad \square$$

Στην παραπάνω απόδειξη λάβαμε υπόψη:

α) ότι το άθροισμα $\sum_{s=1}^{p}\lambda_{Bs}=I_B$, το οποίο μέσω της σχέσης [4.32] και της σχέσης $I_{0-1}=\frac{j-q}{q}$, γράφεται $\sum_{s=1}^{p}\lambda_{Bs}=\frac{j-q}{q^2}+\frac{2\sum_{h<w}I_{hw}}{q^2}$

β) ότι το άθροισμα $\sum_{s=1}^{p}\sqrt{\lambda_{Bs}}=I_{0-1}=\frac{j-q}{q}=\frac{p}{q}$, και

γ) ότι $\sum_{h<w}I_{hw}=\frac{\sum_{h<w}Q_{hw}}{N}$ (βλέπε απόδειξη της Πρότασης 5).

Από την [4.39] κάνοντας χρήση και της [4.34] συνεπάγεται ότι:

$$\frac{\sum_{h<w}Q_{hw}}{N}=\frac{1}{2}\sum_{s=1}^{p}\left(q\sqrt{\lambda_{Bs}}-1\right)^2\Rightarrow\frac{\frac{\sum_{h<w}Q_{hw}}{N}}{\frac{q(q-1)}{2}}=\frac{\frac{1}{2}\sum_{s=1}^{p}\left(q\sqrt{\lambda_{Bs}}-1\right)^2}{\frac{q(q-1)}{2}}\Rightarrow$$

$$\Rightarrow I_{\varepsilon B} = \frac{\sum_{s=1}^{p}\left(q\sqrt{\lambda_{Bs}} - 1\right)^2}{q(q-1)}. \qquad [4.40]$$

Παρατηρούμε ότι μέσω της σχέσης [4.40] έχουμε πετύχει τη σύνδεση της ενδιαφέρουσας αδράνειας με τις ιδιοτιμές $\sqrt{\lambda_{Bs}}$ των παραγοντικών αξόνων που προκύπτουν από την ανάλυση του πίνακα *Burt* ή, ισοδύναμα, με τις αδράνειες των αντίστοιχων αξόνων από την ανάλυση του λογικού πίνακα **Z**. Συνεπώς, η ποσότητα

$$\frac{\left(q\sqrt{\lambda_{Bs}} - 1\right)^2}{q(q-1)}$$

εκφράζει για κάθε s=1,…, p τη «μερική» ενδιαφέρουσα αδράνεια που αντιστοιχεί σε κάθε παραγοντικό άξονα.

Προτείνουμε, λοιπόν, οι αδράνειες των παραγοντικών αξόνων του πίνακα *Burt* να διορθωθούν σύμφωνα με τη σχέση:

$$\lambda_{Bs}^{*adj} = \frac{\left(q\sqrt{\lambda_{Bs}} - 1\right)^2}{q(q-1)}. \qquad [4.41]$$

Το ερώτημα που τίθεται τώρα είναι αν για τον υπολογισμό των ποσοστών ερμηνείας των παραγοντικών αξόνων θα πρέπει να ακολουθήσουμε την προσέγγιση του Benzécri ή του Greenacre. Στην Ενότητα 2.3.4.2 (Παρατήρηση Η) είδαμε ότι οι παραγοντικοί άξονες με αδράνεια μικρότερη ή ίση από $1/q$, στην περίπτωση του πίνακα **Z**, ή μικρότερη ή ίση από $1/q^2$, στην περίπτωση του πίνακα *Burt*, θεωρούνται, εν γένει, "θόρυβος" και έχουν ως αποτέλεσμα οι συντελεστές εσωτερικής συνέπειας των αξόνων να παίρνουν αρνητικές τιμές. Επίσης, εμπειρικά ευρήματα (βλέπε Παράδειγμα Εφαρμογής του Παραρτήματος Δ) ενισχύουν τον προηγούμενο ισχυρισμό. Συνεπώς, η πρότασή μας είναι οι διορθωμένες αδράνειες των αξόνων, μέσω της σχέσης [4.41], να εκφραστούν ως ποσοστό του αθροίσματος των διορθωμένων αδρανειών των αξόνων, για τους οποίους η αρχική, μη διορθωμένη, αδράνειά τους είναι μεγαλύτερη από $1/q^2$. Έτσι, η προτεινόμενη μέθοδος διόρθωσης των αδρανειών του αξόνων που προκύπτουν από την ανάλυση του πίνακα *Burt* παίρνει τελικά τη μορφή:

$$\lambda_{Bs}^{*adj} = \frac{\left(q\sqrt{\lambda_{Bs}} - 1\right)^2}{q(q-1)}, \text{ για } \lambda_{Bs} > \frac{1}{q^2}, \qquad [4.42]$$

Ποσοστό ερμηνείας του άξονα $s = \dfrac{\lambda_{Bs}^{*adj}}{\sum_{t=1}^{p^*} \lambda_{Bt}^{*adj}}$, για p^* τέτοιο ώστε $\lambda_{Bs} > \dfrac{1}{q^2}$, [4.43]

όπου s=1,…, p^* και λ_{Bs} είναι η αρχική αδράνεια του άξονα s από την ανάλυση του *Burt*.

Με την προτεινόμενη μέθοδο διόρθωσης τα ποσοστά % ερμηνείας των παραγοντικών αξόνων έχουν άθροισμα 100%.

Παρατήρηση 4.9

Από τη σχέση διόρθωσης των αδρανειών που προτείνουν οι Benzécri και Greenacre, έχουμε:

$$\lambda_{Bs}^{adj} = \left(\frac{q}{q-1}\right)^2 \left(\sqrt{\lambda_{Bs}} - \frac{1}{q}\right)^2 \Rightarrow \lambda_{Bs}^{adj} = \left(\frac{q}{q-1}\right)^2 \left(\lambda_{Bs} - 2\frac{\sqrt{\lambda_{Bs}}}{q} + \frac{1}{q^2}\right) =$$

$$= \left(\frac{1}{(q-1)^2}\right)\left(q^2\lambda_{Bs} - 2q\sqrt{\lambda_{Bs}} + 1\right) = \frac{\left(q\sqrt{\lambda_{Bs}} - 1\right)^2}{(q-1)^2}.$$

Παρατηρούμε ότι ο αριθμητής του κλάσματος της παραπάνω σχέσης είναι ίσος με τον αριθμητή του κλάσματος της σχέσης [4.42]. Διαφέρουν, όμως, οι αντίστοιχοι παρονομαστές. Επομένως, οι διορθώσεις του Benzécri και του Greenacre δεν επιμερίζουν την ενδιαφέρουσα αδράνεια του πίνακα *Burt* σε κάθε άξονα. Δηλαδή

$$\sum_{s=1}^{p} \lambda_{Bs}^{adj} = \sum_{s=1}^{p} \frac{\left(q\sqrt{\lambda_{Bs}} - 1\right)^2}{(q-1)^2} \neq I_{\varepsilon B} = \sum_{s=1}^{p} \lambda_{Bs}^{*adj} = \sum_{s=1}^{p} \frac{\left(q\sqrt{\lambda_{Bs}} - 1\right)^2}{q(q-1)}.$$

Το προηγούμενο συμπέρασμα οδηγεί σε ένα “παράδοξο” σχετικά με τη μέθοδο διόρθωσης του Greenacre. Παρόλο που η αναγωγή των διορθωμένων αδρανειών γίνεται ως προς την ενδιαφέρουσα αδράνεια του πίνακα *Burt* (βλέπε σχέση [4.38]), ωστόσο οι διορθωμένες αδράνειες δεν αποτελούν διαμελισμό αυτής. Η προσέγγιση αυτή δημιουργεί όχι μόνο εννοιολογικό αλλά και μαθηματικό πρόβλημα. Συνεπώς, τα

ποσοστά ερμηνείας των παραγοντικών αξόνων με τη μέθοδο *ΔG* δεν είναι καλά ορισμένα.

Προφανώς, ισχύει η σχέση:

$$\lambda_{Bs}^{*adj} = \frac{q-1}{q}\lambda_{Bs}^{adj}, \qquad [4.44]$$

η οποία δηλώνει ότι οι διορθώσεις *ΔB* και *ΔG* είναι ανάλογες με την προτεινόμενη. Αποτέλεσμα της σχέσης [4.44] είναι ότι τα διορθωμένα ποσοστά ερμηνείας των αξόνων, τα οποία προκύπτουν από τη μέθοδο του Benzécri, είναι ίσα με αυτά που υπολογίζονται με την προτεινόμενη διόρθωση. Πράγματι, από τη σχέση [4.43] λόγω της [4.44] έχουμε:

$$\text{Ποσοστό ερμηνείας του άξονα } s = \frac{\lambda_{Bs}^{*adj}}{\sum_{t=1}^{p^*}\lambda_{Bt}^{*adj}} = \frac{\frac{q-1}{q}\lambda_{Bs}^{adj}}{\frac{q-1}{q}\sum_{t=1}^{p^*}\lambda_{Bt}^{adj}} = \frac{\lambda_{Bs}^{adj}}{\sum_{t=1}^{p^*}\lambda_{Bt}^{adj}}, \qquad [4.45]$$

για p^* τέτοιο ώστε $\lambda_{Bs} > 1/q^2$.

Βέβαια, στην περίπτωση που το πλήθος q των μεταβλητών είναι μεγάλο, τότε $q(q-1) \approx (q-1)^2$ και συνεπώς $\lambda_{Bs}^{adj} \approx \lambda_{Bs}^{*adj}$. Συμπερασματικά, με την προτεινόμενη μέθοδο τροποποιούμε τις μαθηματικές σχέσεις διόρθωσης των μεθόδων *ΔB* και *ΔG*, ώστε να είναι εννοιολογικά συμβατές με τη φυσική ερμηνεία της ενδιαφέρουσας αδράνειας. Από πρακτική σκοπιά ή προτεινόμενη μέθοδος είναι υπολογιστικά ισοδύναμη με την προσέγγιση του *ΔB* του Benzécri.

Στο Παράρτημα Δ δίνουμε ένα αριθμητικό παράδειγμα εφαρμογής της προτεινόμενης μεθόδου διόρθωσης των αδρανειών του πίνακα *Burt*.

4.9 Σχόλια και Συμπεράσματα Κεφαλαίου

Καταρχήν, θα πρέπει να παρατηρήσουμε ότι η γενίκευση της ΠΑΑ, όπως αυτή εφαρμόζεται στο πλαίσιο της Γαλλικής Σχολής, σε περισσότερες από δύο μεταβλητές δεν είναι άμεση και, για ορισμένους ερευνητές, δεν είναι και καλά ορισμένη (βλέπε

Παρατήρηση ΣΤ της Ενότητας 2.3.4.2). Είδαμε, για παράδειγμα, ότι η ολική αδράνεια του λογικού πίνακα **Z** εξαρτάται μόνο από τον αριθμό των μεταβλητών και από το συνολικό πλήθος των κατηγοριών τους. Φαίνεται να μη λαμβάνεται υπόψη ούτε το πλήθος των αντικειμένων ούτε τη συνάφεια - συσχέτιση μεταξύ των μεταβλητών που συμμετέχουν στην ανάλυση. Επομένως, η ολική αδράνεια στην περίπτωση του πίνακα **Z** δεν έχει την ίδια φυσική ερμηνεία με αυτή της διμεταβλητής περίπτωσης. Κατά την ανάλυση του πίνακα *Burt*, αν και συμμετέχουν όλες οι συσχετίσεις των μεταβλητών ανά δύο στον υπολογισμό της ολικής αδράνειας, ωστόσο οι διαγώνιοι πίνακες και το γεγονός ότι οι διμεταβλητές συσχετίσεις λαμβάνονται υπόψη δύο φορές δημιουργούν προβλήματα στη γεωμετρική ερμηνεία των αποτελεσμάτων και στη φυσική ερμηνεία της αδράνειας (De Leeuw & Van Rijckevorsel 1988, De Leeuw 1993, Saporta & Tambrea 1993, Gifi 1996, Greenacre 2005, 1994β, 1993γ, 1993α, 1991, 1990, 1989 και 1988α). Οι προηγούμενοι προβληματισμοί μεταφέρονται αυτόματα και στη φυσική ερμηνεία και καταλληλότητα της απόστασης χ^2 για τη σύγκριση των προφίλ γραμμών και στηλών των πινάκων **Z** και **B**.

Η διμεταβλητή εκδοχή της ΠΑΑ αναδεικνύει τη συσχέτιση μεταξύ δύο μεταβλητών ή δύο συνόλων μεταβλητών, όπως στην περίπτωση των πινάκων “φέτας”, “στοίβας” και των υποπινάκων του *Burt*. Αντίθετα, στην πολυμεταβλητή περίπτωση το ενδιαφέρον εστιάζεται στην ανάλυση των ενδοσυσχετίσεων που ενθυλακώνει ένα μόνο σύνολο μεταβλητών. Σύμφωνα με τον Greenacre (2005), αυτή η διαφοροποίηση θα πρέπει να διέπει τόσο την ανάλυση όσο και την ερμηνεία των αποτελεσμάτων. Κατά προτίμηση οι μεταβλητές θα πρέπει να έχουν την ίδια κλίμακα μέτρησης, το ίδιο πλήθος κατηγοριών και να διαπραγματεύονται ένα συγκεκριμένο θέμα (Greenacre, 2005), όπως συμβαίνει στις έρευνες διερεύνησης στάσεων όπου, κατά κανόνα, οι κλάσεις των μεταβλητών – ερωτήσεων αντιστοιχούν σε διαβαθμισμένες κατηγορίες τύπου *Likert*. Με άλλα λόγια, ένα καλύτερο πλαίσιο ερμηνείας των αποτελεσμάτων της ΠΑΑ είναι αυτό της Ανάλυσης Ομοιογένειας (βλέπε Ενότητα 2.3.4). Έτσι, από μια άλλη οπτική, είναι προτιμότερο να θεωρήσουμε την πολυμεταβλητή εκδοχή της ΠΑΑ ως μια επέκταση της Ανάλυσης σε Κύριες Συνιστώσες για κατηγορικές μεταβλητές, όπου αναλύονται όλες οι διμεταβλητές συσχετίσεις μεταξύ των μεταβλητών (Van de Geer 1993α και 1993β, Greenacre 1994β, Gifi 1996). Με την προσέγγιση αυτή, η μέθοδος καθίσταται καθαρά μια

τεχνική μείωσης των διαστάσεων του χώρου των αρχικών δεδομένων. Επίσης, η ΠΑΑ, ως μέθοδος κλιμάκωσης που ποσοτικοποιεί ποιοτικά δεδομένα, ικανοποιεί μια πληθώρα κριτηρίων βελτιστοποίησης (βλέπε Ενότητα 2.5). Υπό αυτό το πρίσμα, αν τα κριτήρια αυτά συνάδουν με τους στόχους της εκάστοτε μελέτης, τότε η μέθοδος μπορεί να εφαρμοστεί όπως έχει στον λογικό πίνακα **Z** ή, ισοδύναμα, στον πίνακα *Burt* μέσω της μεθοδολογίας που προτείναμε στην Ενότητα 3.3. Σε περιπτώσεις όπου δεν έχει ιδιαίτερο ενδιαφέρον η μελέτη των αντικειμένων του πίνακα δεδομένων, η ΠΑΑ μπορεί να εφαρμοστεί στον αντίστοιχο πίνακα *Burt*, με τη διόρθωση των αδρανειών των παραγοντικών αξόνων που προτείναμε στην Ενότητα 4.8.3. Γενικά, αν αποδεσμεύσουμε τη διμεταβλητή εκδοχή της ΠΑΑ από την ανάλυση των αποκλίσεων από την κατάσταση ανεξαρτησίας των μεταβλητών, η οποία αποκτά νόημα μόνο όταν πρόκειται να αναλυθούν δεδομένα που έχουν συγκεντρωθεί με μεθόδους της τυχαίας δειγματοληψίας, τότε η χρήση της απόστασης χ^2 ως μιας σταθμισμένης Ευκλείδειας (βλέπε Ενότητα 2.2.5) και η βαρυκεντρική ερμηνεία των αποτελεσμάτων (βλέπε Ενότητες 2.3.3 και 2.3.4 και Τέταρτο Βήμα της Ενότητας 2.2.14) εγγυώνται τη γενίκευση της μεθόδου σε οποιοδήποτε πίνακα συμπτώσεων που έχει τη γενική μορφή «*αντικείμενα* × *ιδιότητες*» και ικανοποιεί τις προϋποθέσεις εφαρμογής της ΠΑΑ που αναφέρθηκαν στην Ενότητα 2.1.

Η μεθοδολογία υπολογισμού της ολικής αδράνειας των τριών πινάκων **F**, **Z** και **B** στηρίχθηκε, αρχικά, στην παρατήρηση ότι ο πίνακας που αναλύεται κάθε φορά μπορεί να θεωρηθεί ως ένας απλός πίνακας συμπτώσεων δύο μεταβλητών, επιβάλλοντας έτσι μια εννοιολογική αλλαγή στο σχεδιασμό της μελέτης, ανάλογα με τον πίνακα εισόδου στην ΠΑΑ. Οι έννοιες της δειγματοληπτικής ή πειραματικής μονάδας και της μονάδας παρατήρησης τροποποιούνται κατάλληλα, ανάλογα με τον πίνακα που αναλύεται. Στη συνέχεια, ο υπολογισμός της ολικής αδράνειας I πραγματοποιήθηκε με την εφαρμογή της γνωστής σχέσης:

$$I = \frac{Q}{N},$$

όπου Q είναι το στατιστικό χ^2 που αντιστοιχεί στον πίνακα που αναλύεται και N το πλήθος των αντίστοιχων μονάδων παρατήρησης.

Με τον τρόπο αυτό επιτυγχάνουμε μια ενιαία μέθοδο υπολογισμού της ολικής αδράνειας ενός πίνακα συμπτώσεων της μορφής «*αντικείμενα* × *ιδιότητες*». Στην πολυμεταβλητή περίπτωση, η σχέση [4.33], που συνδέει την αδράνεια του πίνακα **B** με τις αδράνειες των υποπινάκων που τον απαρτίζουν, είναι ιδιαίτερα χρήσιμη, αφού η ολική αδράνεια του πίνακα **B** μπορεί να υπολογιστεί *a priori* και όχι με την εφαρμογή της SVD στον πίνακα που διαγωνοποιείται κατά την εφαρμογή της ΠΑΑ. Όλες οι νέες σχέσεις που αποδείχθηκαν ([4.11], [4.23], [4.24], [4.27], [4.28], [4.32] και [4.33]) καθώς και τα Πορίσματα 1, 2 και 3 μπορούν να επαληθευτούν εμπειρικά μέσω των στατιστικών πακέτων που περιλαμβάνουν τη διαδικασία της ΠΑΑ. Οι μαθηματικοί τύποι υπολογισμού της ολικής αδράνειας καθώς και οι σχέσεις που συνδέουν τις ολικές αδράνειες των πινάκων **F**, **Z** και **B**, στην περίπτωση των δύο μεταβλητών, και των πινάκων **Z** και **B**, στην περίπτωση *q* μεταβλητών, αναδεικνύουν ποικίλες και ορισμένες νέες φυσικές ερμηνείες της ολικής αδράνειας του πίνακα που αναλύεται κάθε φορά. Στην Ενότητα 2.2.5 είδαμε ότι η ολική αδράνεια του πίνακα **F** εκφράζει μια γενικευμένη διασπορά και, πιο συγκεκριμένα, το σταθμισμένο μέσο όρο των τετραγώνων των χ^2 αποστάσεων των προφίλ γραμμών (ή ισοδύναμα των προφίλ των στηλών) από το κέντρο βάρους τους. Από τη σχέση [4.1] διαπιστώνουμε ότι η ολική αδράνεια του πίνακα **F** εκφράζει, επίσης, τη μέση απόκλιση ανά μονάδα παρατήρησης από την κατάσταση ανεξαρτησίας μεταξύ των δύο μεταβλητών *X* και *Y*, όπως αυτή μετριέται μέσω του στατιστικού χ^2. Αυτή η ερμηνεία είναι έγκυρη μόνο όταν αναλύονται δύο τυχαίες μεταβλητές.

Στην περίπτωση του πίνακα **Z** για δύο μεταβλητές φαίνεται να έχουμε τρεις τουλάχιστον επιλογές σχετικά με τη φυσική ερμηνεία της ολικής αδράνειας:

1) Μπορεί να θεωρηθεί ως η μέση απόκλιση ανά μονάδα παρατήρησης, από την κατάσταση ανεξαρτησίας της μεταβλητής των γραμμών και της μεταβλητής των στηλών, όπως αυτή μετριέται μέσω του στατιστικού χ^2. Στην περίπτωση αυτή, η έννοια της ανεξαρτησίας είναι προτιμότερο να αντικατασταθεί με την έννοια της ομοιογένειας των γραμμών (Greenacre, 2005). Δηλαδή, του κατά πόσο τα αντικείμενα της μελέτης έχουν το ίδιο προφίλ ως προς τη νέα μεταβλητή των στηλών *E* που σχηματίζεται από την ένωση των μεταβλητών *X* και *Y*.

2) Μπορεί να θεωρηθεί ως η μέση αδράνεια ανά μεταβλητή, των αδρανειών των πινάκων **X** και **Y**, όπου **X** και **Y** είναι οι διαγώνιοι πίνακες, οι οποίοι διασταυρώνουν τις κλάσεις ή ιδιότητες των μεταβλητών *X* και *Y* μεταξύ τους.

3) Μπορεί να θεωρηθεί ως το άθροισμα των αδρανειών των δύο μεταβλητών.

Οι παραπάνω τρεις εκδοχές-επιλογές μπορούν να γενικευτούν για τον πίνακα **Z** και στην περίπτωση *q* μεταβλητών.

Η ολική αδράνεια του πίνακα **B** (*Burt*), στην περίπτωση των δύο μεταβλητών, εκφράζει είτε τη μέση αδράνεια ανά μεταβλητή των πινάκων **Z** και **F** είτε τη μέση αδράνεια των 4 υποπινάκων **X**, **Y**, **F** και $\mathbf{F}^{\mathbf{T}}$, από τους οποίους αποτελείται. Η τελευταία αυτή εκδοχή μπορεί να γενικευτεί και για περισσότερες μεταβλητές. Από τις σχέσεις [4.32] και [4.33] φαίνεται ότι στον υπολογισμό της ολικής αδράνειας του **B** συμμετέχουν δύο μέρη. Το πρώτο αφορά την αδράνεια των διαγώνιων πινάκων και το δεύτερο την αδράνεια των απλών πινάκων συμπτώσεων των μεταβλητών ανά δύο (και μάλιστα δύο φορές). Το πρώτο δίνει την αίσθηση του θορύβου, ενώ το δεύτερο την αίσθηση της πλεονάζουσας πληροφορίας. Για το λόγο αυτό έχουν προταθεί διορθώσεις τόσο σε σχέση με κάποια αριθμητικά αποτελέσματα (π.χ. αδράνειες) που παράγονται από την ΠΑΑ όσο και σε σχέση με τη μορφή του πίνακα **B** που θα αναλυθεί.

Η εφαρμογή της ΠΑΑ είτε στον πίνακα **Z** είτε στον **B**, λόγω της σχέσης που συνδέει τους δύο πίνακες $\mathbf{B}=\mathbf{Z}^{\mathbf{T}}\mathbf{Z}$ (βλέπε Ενότητα 2.3.3.5), έχει ως αποτέλεσμα η εικόνα, επί των παραγοντικών επιπέδων, του φαινομένου που εξετάζεται να είναι η ίδια και στις δύο περιπτώσεις. Διατηρεί, μάλιστα, τις ιδιότητες της βέλτιστης ή άριστης κλιμάκωσης (βλέπε Ενότητα 2.5) των συντεταγμένων των προβολών των σημείων πάνω στους παραγοντικούς άξονες (Greenacre 2005 και 1993α, Gifi 1996).

Η φυσική ερμηνεία της ενδιαφέρουσας αδράνειας που ορίσαμε στην Ενότητα 4.7 αποδείχθηκε χρήσιμη στην ανάπτυξη των μεθόδων: α) εντοπισμού υποπίνακα του πίνακα *Burt*, ο οποίος περιλαμβάνει όλες τις μεταβλητές που συμμετέχουν στην ανάλυση και η εφαρμογή της ΠΑΑ σε αυτόν αποδίδει την “πλησιέστερη εικόνα” του φαινομένου σε αυτήν την εικόνα που προκύπτει από την εφαρμογή της μεθόδου στον αρχικό πίνακα *Burt* και β) διόρθωσης των αδρανειών των παραγοντικών αξόνων,

ώστε τα αντίστοιχα ποσοστά ερμηνείας να μη δίνουν πλέον την εντύπωση φτωχής προσαρμογής των δεδομένων και χαμηλής ποιότητας της λύσης της ΠΑΑ. Επιπλέον, ο έλεγχος της στατιστικής σημαντικότητας της ενδιαφέρουσας αδράνειας αποτελεί έναν ακόμη συνδετικό κρίκο της Ανάλυσης Δεδομένων με την Επαγωγική Στατιστική. Και οι τρεις εφαρμογές που προτείναμε στην Ενότητα 4.8 έχουν αλγοριθμοποιηθεί και υλοποιούνται στο λογισμικό CHIC Analysis (βλέπε Μάρκος, 2006).

ΚΕΦΑΛΑΙΟ 5

Ανάλυση Ισχύος και Καθορισμός Μεγέθους Δείγματος στην Παραγοντική Ανάλυση των Αντιστοιχιών

5.1 Εισαγωγή

Η ΠΑΑ θεωρείται ως μία περιγραφική μέθοδος διερεύνησης της σχέσης δύο ή περισσότερων κατηγορικών μεταβλητών (βλέπε Κεφάλαιο 1). Τα αποτελέσματα παρέχουν πληροφορία, η οποία είναι παρόμοια στη φύση της με αυτή που παράγεται από τις τεχνικές της Παραγοντικής Ανάλυσης (*Factor Analysis*) και επιτρέπουν τη διερεύνηση των δομών των σχέσεων μεταξύ των μεταβλητών που συμμετέχουν στην ανάλυση. Χαρακτηριστικό της μεθόδου είναι ότι τα δεδομένα αντιμετωπίζονται σαν να αποτελούν ολόκληρο τον υπό εξέταση πληθυσμό, ανεξάρτητα με το εάν αυτά προέρχονται από ένα δείγμα του. Όμως, στην περίπτωση δύο μεταβλητών, όπου η συλλογή των αντίστοιχων δεδομένων έχει γίνει με απλή τυχαία δειγματοληψία, η στατιστική σημαντικότητα της ολικής αδράνειας I_F του πίνακα συμπτώσεων μπορεί να ελεγχθεί μέσω της χ^2 Κατανομής (Lebart, Morineau & Tabard 1977, Lebart, Morineau & Piron 2000). Έτσι, αρκετοί συγγραφείς (Weller & Romney 1990, Greenacre 1993α, Van de Geer 1993β, Blasius 1994, Micheloud 1997, Clausen 1998) συνδυάζουν την εφαρμογή της ΠΑΑ με τον έλεγχο ανεξαρτησίας χ^2.

Κατά παράδοση, ο στατιστικός έλεγχος υποθέσεων στην επιστημονική έρευνα έχει να επιδείξει μια σαφή προτίμηση στη χρήση της στατιστικής σημαντικότητας ως κριτηρίου απόρριψης ή όχι της μηδενικής υπόθεσης H_0 (Τσάντας και άλλοι 1999, Huck 2000α και 2000β), με αποτέλεσμα να δοθεί μεγαλύτερη έμφαση στον έλεγχο και στη διαχείριση του Σφάλματος Τύπου Ι. Όμως, τα τελευταία χρόνια, και ιδιαίτερα μετά τις εργασίες του Jacob Cohen (1962, 1965 και 1988) σχετικά με την Ανάλυση Ισχύος των στατιστικών ελέγχων (*Statistical Power Analysis*) στις Επιστήμες της

Συμπεριφοράς, η προσοχή των ερευνητών αρχίζει να στρέφεται και στον έλεγχο του Σφάλματος Τύπου ΙΙ και στην αναγκαιότητα ανάλυσης της ισχύος των στατιστικών ελέγχων (Cohen 1988, Murphy & Myors 1998). Η αναγκαιότητα αυτή ξεπέρασε το χώρο των Επιστημών της Συμπεριφοράς και εμφανίζεται πλέον σε αρκετά επιστημονικά πεδία (βλέπε Cascio & Zedeck 1983, Κάτος 1986, Cohen & Nee 1987, Muller *et al*. 1992, Hubbard & Armstrong 1992, Verma & Goodale 1995, Aguinis 1995, Buhl-Mortensen 1996, Thomas & Juanes 1996, Heidelbaugh & Nelson 1996, Meyer & Mark 1996, Mone, Mueller & Mauland 1996, Miller *et al*. 1997, MacCallum & Hong 1997, Sheppard 1999, Lee & Zelen 2000, Evans & Viengkham 2001, Foster 2001, Di Stefano 2001, Nutahara *et al*. 2001, Pan 2001, Desmond & Glover 2002, Mumby 2002, Χατζηνικολάου 2002). Για μια εισαγωγική παρουσίαση της λογικής και των εννοιών που εμπλέκονται στην Ανάλυση Ισχύος των στατιστικών ελέγχων παραπέμπουμε στους Cohen (1988), Mohr (1990), Nemec (1991), Hallahan και Rosenthal (1996), Murphy και Myors (1998), Sheppard (1999), Kramer και Rosenthal (1999), Lewis (2000), Huck (2000α), Lenth (2001), Hoenig και Heisey (2001), Χατζηνικολάου (2002), Μενεξέ (2002), Stevens (2002) και Berger (2003). Οι Goldstein (1989), Thomas και Krebs (1997), Iwane, Palensky και Plante (1997) και Chernick και Liu (2002) επιχειρούν μια συγκριτική παρουσίαση λογισμικών που πραγματοποιούν Ανάλυση Ισχύος, ενώ ο Μενεξές (2002) παραθέτει διευθύνσεις ιστοσελίδων με πληροφοριακό υλικό σχετικά με ειδικά λογισμικά. Μάλιστα, οι Gatti και Harwell (1998) προτείνουν τη χρήση λογισμικών για Ανάλυση Ισχύος έναντι των παραδοσιακών Διαγραμμάτων Ισχύος (*Power Charts*), τα οποία δημιουργούν δυσκολίες στην ανάγνωση των αριθμητικών τιμών και εσφαλμένες εκτιμήσεις λόγω της συχνά απαιτούμενης γραμμικής παρεμβολής.

Στο πλαίσιο της Ανάλυσης Ισχύος των στατιστικών ελέγχων, που προτείνει ο Cohen (1988), είναι δυνατός και ο *a priori* υπολογισμός του ελάχιστου απαιτούμενου μεγέθους δείγματος N, ώστε ο στατιστικός έλεγχος ανεξαρτησίας ή ομοιογένειας χ^2 δύο κατηγορικών μεταβλητών, σε επίπεδο σημαντικότητας a, με ισχύ γ να ανιχνεύσει ένα προκαθορισμένο μέγεθος αποτελέσματος (*Effect Size-ES*) ως στατιστικά σημαντικό. Το θέμα της Ανάλυσης Ισχύος του στατιστικού ελέγχου χ^2 έχει απασχολήσει και άλλους ερευνητές (Meng & Chapman 1966, Nathan 1972, Guenther 1977, Lachin 1977). Οι μεθοδολογίες που προτείνουν φαίνεται να λύνουν τοπικά το

πρόβλημα είτε για συγκεκριμένα δειγματοληπτικά ή πειραματικά σχέδια είτε για συγκεκριμένες εναλλακτικές υποθέσεις και όχι μέσα σε ένα γενικό μεθοδολογικό πλαίσιο, όπως η πρόταση του Cohen, η οποία, επιπλέον, όπως θα δειχθεί στη συνέχεια, επιτρέπει τη σύνδεση της Ανάλυσης Ισχύος με την έννοια της ολικής αδράνειας του πίνακα συμπτώσεων απολύτων συχνοτήτων δύο κατηγορικών μεταβλητών. Σε περίπτωση πειραματικής ή δειγματοληπτικής έρευνας σημαντική μέριμνα θα πρέπει να λαμβάνεται σχετικά με τον υπολογισμό του ελάχιστου απαιτούμενου μεγέθους δείγματος (Hansen, Hurwitz & Madow 1953, Cochran 1977, Cohen & Cohen 1983, Kraemer & Thiemann 1987, Cohen 1988, Hinkle, Wiersma & Jurs 1988, Lipsey 1990, Hair *et al*. 1995, Kirk 1995, Φίλιας 1996, Kirkwood 1996, Περσίδης 1997, Lindley 1997, Nicewander & Price 1997, Shih & Zhao 1997, Adcock 1997, Murphy & Mayors 1998, Hoyle 1999, Lenth 2001, Lewis 2000, Horn & Vollandt 2000, Φαρμάκης 2003 και 1994, Δαφέρμος 2005). Το μέγεθος του δείγματος αποτελεί τον κρισιμότερο παράγοντα που επηρεάζει: α) τη σημαντικότητα των αποτελεσμάτων και τη δυνατότητα γενίκευσής τους, β) την απορρόφηση και την κατανομή των διαθέσιμων χρονικών και οικονομικών πόρων και γ) την αποδοχή της αποτελεσματικότητας και της χρηστικής αξίας της εκάστοτε έρευνας σε σχέση με τους δεοντολογικούς και ηθικούς κανόνες, οι οποίοι διέπουν τα διάφορα επιστημονικά ερευνητικά πεδία.

Στο Κεφάλαιο αυτό, θεωρούμε την ολική αδράνεια ως έναν ολικό δείκτη μεγέθους του αποτελέσματος, βάσει του οποίου μπορεί να υπολογιστεί η παρατηρούμενη ισχύς (*observed power*) του στατιστικού ελέγχου χ^2. Εισάγουμε την έννοια της «Δυναμικής Αδράνειας» ενός πίνακα συμπτώσεων δύο κατηγορικών μεταβλητών και προτείνουμε μεθοδολογία για την *a priori* και *post-hoc* Ανάλυση Ισχύος του στατιστικού ελέγχου χ^2 χρησιμοποιώντας την ολική ή/και τη δυναμική αδράνεια ως δείκτες μεγέθους του αποτελέσματος. Με την προτεινόμενη μεθοδολογία είναι δυνατή η εκτίμηση του ελάχιστου απαιτούμενου μεγέθους δείγματος σε δειγματοληπτική ή πειραματική έρευνα, όπου στα δεδομένα θα εφαρμοστεί η ΠΑΑ. Θα πρέπει να τονιστεί ότι αν και το πρόβλημα της εκτίμησης του απαιτούμενου μεγέθους δείγματος είναι αντικείμενο έρευνας στο χώρο της Παραγοντικής Ανάλυσης και της Ανάλυσης σε Κύριες Συνιστώσες (βλέπε Guilford 1954, Comrey & Lee 1973, Everitt 1975, Cattell 1978, Barrett & Kline 1981, Gorsuch 1983, Arrindell & Van der Ende 1985, Browne &

Cudeck 1993, Hair *et al*. 1995, MacCallum *et al*. 1999, MacCallum *et al*. 2001), ωστόσο αυτό τίθεται για πρώτη φορά στο πλαίσιο της ΠΑΑ.

Ειδικότερα, στην Ενότητα 5.1 αναφερόμαστε στον έλεγχο στατιστικής σημαντικότητας της ολικής αδράνειας, ενώ στις Ενότητες 5.2 έως 5.5 παρουσιάζουμε τους βασικούς συμβολισμούς και έννοιες που σχετίζονται με τη διαδικασία των στατιστικών ελέγχων υποθέσεων (για παράδειγμα, Σφάλμα Τύπου Ι, Σφάλμα Τύπου ΙΙ, Σφάλμα Τύπου ΙΙ½, Ισχύς και Παρατηρούμενη Στάθμη Σημαντικότητας). Στη συνέχεια, επιχειρούμε μια σύνοψη των κύριων επιχειρημάτων, πάνω στα οποία στηρίζεται η κριτική κατά των στατιστικών ελέγχων (Ενότητα 5.6). Στην Ενότητα 5.7 κάνουμε μια εισαγωγή στην *a priori* και *post hoc* Ανάλυση Ισχύος. Κατόπιν, προτείνουμε μεθοδολογία Ανάλυσης Ισχύος του στατιστικού ελέγχου ανεξαρτησίας (ή ομοιογένειας) χ^2, στην περίπτωση δύο κατηγορικών μεταβλητών, βάσει της σχέσης της ολικής αδράνειας του αντίστοιχου πίνακα συμπτώσεων με το μέγεθος αποτελέσματος, όπως αυτό ορίζεται από τον Cohen για τον έλεγχο ανεξαρτησίας ή ομοιογένειας χ^2 (Ενότητα 5.8). Στην Ενότητα 5.9 προτείνουμε τρόπους προκαθορισμού του μεγέθους αποτελέσματος και παρουσιάζουμε τις σχέσεις που συνδέουν την ολική αδράνεια με άλλους δείκτες συνάφειας. Ακολούθως, εισάγουμε τις έννοιες της «*a priori*» και «*post hoc*» Δυναμικής Αδράνειας ενός πίνακα συμπτώσεων δύο κατηγορικών μεταβλητών (Ενότητα 5.10). Στην Ενότητα 5.11 επιχειρούμε γενίκευση της μεθόδου υπολογισμού του μεγέθους του δείγματος στην περίπτωση πολλών μεταβλητών. Στην επόμενη ενότητα (5.12), παρουσιάζουμε μεθοδολογία Ανάλυσης Ισχύος του ελέγχου χ^2 καλής προσαρμογής για τον προσδιορισμό του αριθμού των σημαντικών παραγοντικών αξόνων που θα πρέπει να διατηρηθούν μετά την εφαρμογή της ΠΑΑ σε πίνακα συμπτώσεων δύο μεταβλητών. Για τον ίδιο σκοπό, προτείνουμε και ένα εμπειρικό κριτήριο βασισμένο στη φυσική ερμηνεία της Δυναμικής Αδράνειας. Τέλος, στην Ενότητα 5.12, δίνουμε διάφορα αριθμητικά παραδείγματα. Στο παράδειγμα της Ενότητας 5.13.4 παραθέτουμε εφαρμογή του κριτηρίου της «*σπασμένης ράβδου*» για τον εντοπισμό των σημαντικών αξόνων στο πλαίσιο της ΠΑΑ.

Με βάση την προτεινόμενη μεθοδολογία, αναπτύξαμε[17] το λογισμικό Power Analysis for AFC, το οποίο υλοποιεί *a priori* και *post hoc* Ανάλυση Ισχύος στο πλαίσιο της ΠΑΑ (βλέπε Παράρτημα CDB στο CD που συνοδεύει τη διατριβή). Σύντομη περιγραφή των δυνατοτήτων και του τρόπου χρήσης του λογισμικού παραθέτουμε στην Ενότητα Ε2 του Παραρτήματος Ε.

5.2 Στατιστική Σημαντικότητα της Ολικής Αδράνειας

Έστω X και Y δύο κατηγορικές μεταβλητές με k και l κλάσεις αντίστοιχα. Συμβολίζουμε με (βλέπε Ενότητα 2.2):

F τον $k \times l$ πίνακα συμπτώσεων απολύτων συχνοτήτων με στοιχεία f_{ij} (i=1,…,k και j=1,…,l), ο οποίος εκφράζει την από κοινού εμπειρική κατανομή των μεταβλητών X και Y

f_{i+} : την περιθώρια απόλυτη συχνότητα της γραμμής i, i = 1,…,k του πίνακα **F**

f_{+j} : την περιθώρια απόλυτη συχνότητα της στήλης j, j = 1,…, l του πίνακα **F**

N: το γενικό σύνολο του πίνακα **F**, το οποίο αντιστοιχεί στο μέγεθος του δείγματος, όπου:

$$\sum_i f_{i+} = \sum_j f_{+j} = N$$

P: τον $k \times l$ πίνακα των αντιστοιχιών, του οποίου τα στοιχεία είναι τα στοιχεία του πίνακα **F** διαιρεμένα με το N, δηλαδή τα στοιχεία του πίνακα **P** δίνονται από τη σχέση:

$$p_{ij} = \frac{f_{ij}}{N},\ i = 1,\ldots,k \text{ και } j = 1,\ldots,\ l$$

r_i: τη μάζα της γραμμής i του πίνακα **P**, όπου:

---

[17] Σε συνεργασία με το συνάδελφο κ. Άγγελο Μάρκο, Υποψήφιο Διδάκτορα του Τμήματος Εφαρμοσμένης Πληροφορικής του Πανεπιστημίου Μακεδονίας.

$$r_i = \sum_j p_{ij} = \sum_j \frac{f_{ij}}{N}, = \frac{f_{i+}}{N} \;\; i = 1,\ldots,k \text{ και } j = 1,\ldots, l$$

c_j: τη μάζα της στήλης j του πίνακα **P**, όπου:

$$c_j = \sum_i p_{ij} = \sum_i \frac{f_{ij}}{N} = \frac{f_{+j}}{N}, \;\; i = 1,\ldots,k \text{ και } j = 1,\ldots, l$$

Στην Ενότητα 2.2.5 είδαμε ότι η ολική αδράνεια I_F του πίνακα **F** εκφράζει μια γενικευμένη διασπορά και, πιο συγκεκριμένα, το σταθμισμένο μέσο όρο των τετραγώνων των χ^2 αποστάσεων των προφίλ γραμμών (ή ισοδύναμα των προφίλ των στηλών) από το κέντρο βάρους τους. Υπολογιστικά, όμως, η I_F δίνεται και από τις παρακάτω σχέσεις (βλέπε Ενότητα 2.2.5):

$$I_F = \sum_i \sum_j \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} \qquad [5.1]$$

και

$$I_F = \frac{Q}{N}. \qquad [5.2]$$

Από τη σχέση [5.2] συνεπάγεται ότι:

$$Q = NI_F. \qquad [5.3]$$

Στη σχέση [5.2] η ποσότητα Q είναι το στατιστικό χ^2 που αντιστοιχεί στον πίνακα **F** και υπολογίζεται από την παρακάτω σχέση:

$$Q = \sum_i \sum_j \frac{\left(f_{ij} - \frac{f_{i+} f_{+j}}{N} \right)^2}{\frac{f_{i+} f_{+j}}{N}} = N \sum_i \sum_j \frac{(p_{ij} - r_i c_j)^2}{r_i c_j}. \qquad [5.4]$$

Στην περίπτωση που τα δεδομένα έχουν συγκεντρωθεί με τη μέθοδο της απλής τυχαίας δειγματοληψίας, από τη σχέση [5.3] και με την παραδοχή ότι ισχύουν οι προϋποθέσεις εφαρμογής του στατιστικού ελέγχου ανεξαρτησίας χ^2 (βλέπε Haberman 1988, Agresti 1990), η στατιστική σημαντικότητα της ποσότητας NI_F μπορεί να ελεγχθεί μέσω της χ^2 Κατανομής με (k-1)(l-1) βαθμούς ελευθερίας σε επίπεδο

σημαντικότητας α (Lebart, Morineau & Tabard 1977, Lebart, Morineau & Warwick 1984, Lebart, Morineau & Piron 2000).

5.3 Σφάλμα Τύπου Ι και Σφάλμα Τύπου ΙΙ

Σε ένα στατιστικό έλεγχο η απόφαση σχετικά με την απόρριψη της H_0 μπορεί να είναι σωστή ή λανθασμένη. Λανθασμένη απόφαση λαμβάνεται όταν:

α) απορριφθεί η H_0, ενώ στην πραγματικότητα είναι αληθής. Λέμε τότε ότι έχει διαπραχθεί «Σφάλμα Τύπου Ι» ή «σφάλμα πρώτου είδους». Η πιθανότητα να διαπραχθεί Σφάλμα Τύπου Ι συμβολίζεται με α και είναι η δεσμευμένη πιθανότητα:

$$\alpha = P(\text{απόρριψης της } H_0 \,/\, H_0 \text{ αληθής}). \qquad [5.5]$$

β) δεν απορριφθεί η H_0, ενώ στην πραγματικότητα είναι ψευδής. Λέμε τότε ότι έχει διαπραχθεί «Σφάλμα Τύπου ΙΙ» ή «σφάλμα δεύτερου είδους». Η πιθανότητα να διαπραχθεί Σφάλμα Τύπου ΙΙ συμβολίζεται με β και είναι η δεσμευμένη πιθανότητα:

$$\beta = P(\text{μη απόρριψης της } H_0 \,/\, H_0 \text{ λανθασμένη}). \qquad [5.6]$$

Όταν ελέγχεται η H_0 επιλέγεται ως α μία τιμή που εκφράζει τη μέγιστη πιθανότητα αποδοχής διάπραξης Σφάλματος Τύπου Ι. Η πιθανότητα αυτή ονομάζεται «επίπεδο σημαντικότητας» (ε.σ.) και είναι απαραίτητο να καθορίζεται από τον ερευνητή πριν από τη δειγματοληψία ή την εκτέλεση ενός πειράματος, ώστε τα αποτελέσματα των στατιστικών αναλύσεων να μην επηρεάσουν την τιμή της (Cohen 1988, Hinkle, Wiersma & Jurs 1988, Kachigan 1991, Pagano & Gauvreau 2000). Έτσι, η τιμή του α δεν θα πρέπει να καθορίζεται μετά από προκαταρκτικές αναλύσεις των δεδομένων, ούτε θα πρέπει να τροποποιείται έτσι ώστε να εξυπηρετεί την απόρριψη ή μη συγκεκριμένων μηδενικών υποθέσεων. Επίσης, το ε.σ. α εκφράζει την πιθανότητα να διαπραχθεί Σφάλμα Τύπου Ι μόνο όταν: α) οι μετρήσεις είναι έγκυρες και αξιόπιστες και β) ισχύουν οι προϋποθέσεις εφαρμογής του αντίστοιχου στατιστικού ελέγχου.

Στην πράξη χρησιμοποιούνται παραδοσιακά οι συμβατικές (αυθαίρετες) τιμές α = 0,10 ή α = 0,05 ή α= 0,01 (Hinkle, Wiersma & Jurs 1988, Kirk 1995, Hair *et al*. 1995, Hopkins 1997, Huck 2000α). Αν για παράδειγμα σε κάποιον έλεγχο καθοριστεί ως ε.σ. α = 0,05 ή 5% και απορριφθεί η H_0, τότε θεωρητικά σε 100 όμοιες περιπτώσεις ή

σε 100 επαναλήψεις του πειράματος μόνο σε 5 αναμένεται να ληφθεί λάθος απόφαση, δηλαδή να απορριφθεί η H_0, ενώ στην πραγματικότητα είναι σωστή. Έτσι, φαίνεται ότι το ε.σ. εκφράζει ένα ρυθμό σφάλματος που συνδέεται κυρίως με τη στατιστική διαδικασία και όχι με την τιμή του στατιστικού (π.χ. t, F και χ^2) του ελέγχου (Lohninger, 1999).

Η πιθανότητα να μην απορριφθεί μία όντως αληθή H_0 καθορίζεται από το ε.σ. α και δίνεται από τη σχέση:

$$1 - \alpha = P(\text{μη απόρριψης της } H_0 \,/\, H_0 \text{ αληθής}). \qquad [5.7]$$

Η πιθανότητα γ να απορριφθεί μία όντως ψευδής H_0 καθορίζεται από την πιθανότητα β και ονομάζεται «ισχύς» (*power*) του στατιστικού ελέγχου:

$$\gamma = 1 - \beta = P(\text{απόρριψης της } H_0 \,/\, H_0 \text{ λανθασμένη}). \qquad [5.8]$$

Οι σχέσεις [5.7] και [5.8] εκφράζουν την πιθανότητα να έχει ληφθεί σωστή απόφαση σε ένα στατιστικό έλεγχο.

Άρα, για να μπορέσει κανείς, με βάση τα δεδομένα, να καταλήξει σε σχετικά ασφαλή και αξιόπιστα συμπεράσματα θα πρέπει ο στατιστικός έλεγχος να ελαχιστοποιεί τα α και β. Όμως, κάθε προσπάθεια μείωσης του ενός κινδύνου αυξάνει τον άλλο (Κολυβά-Μαχαίρα & Μπόρα-Σέντα 1996, Τσάντας και άλλοι 1999). Σε πρακτικό επίπεδο, επιχειρείται η μείωση του σπουδαιότερου από τους δύο κινδύνους. Ένας τρόπος να μειωθούν και οι δύο κίνδυνοι ταυτόχρονα είναι να αυξηθεί το μέγεθος του δείγματος (Zar 1996, Pagano & Gauvreau 2000, Χάλκος 2000), το οποίο δεν είναι πάντοτε εφικτό λόγω φυσικών, τεχνικών, οικονομικών, χρονικών και δεοντολογικών περιορισμών.

Όμως, ποιο από τα δύο σφάλματα είναι σημαντικότερο; Η απάντηση είναι σχετική και εξαρτάται από πολλούς παράγοντες, όπως από το γενικό σκοπό και τους ειδικούς στόχους της έρευνας, από το θεωρητικό πλαίσιο, τις γνώσεις του ερευνητή ή από σκοπιμότητες (Μενεξές & Οικονόμου, 2001). Σε κάθε περίπτωση, στην απόφαση απόρριψης ή όχι μίας υπόθεσης πρέπει να υπολογίζονται και να λαμβάνονται υπόψη τόσο το α όσο και το β (Κολυβά-Μαχαίρα & Μπόρα-Σέντα, 1996).

Οι στατιστικοί έλεγχοι περιλαμβάνουν αρκετές συμβάσεις σε ότι αφορά τον προκαθορισμό του α και του β. Για παράδειγμα, πολλοί ερευνητές θέτουν το $\alpha \leq 0,05$ και το $\beta \leq 0,20$ (Kirk, 1995). Αυτό σημαίνει ότι θεωρούν ως σοβαρότερο τον κίνδυνο να διαπράξουν Σφάλμα Τύπου Ι απ' ότι να διαπράξουν Σφάλμα Τύπου ΙΙ. Αν υπολογίσουμε το λόγο:

P(να διαπραχθεί Σφάλμα Τύπου ΙΙ) / P(να διαπραχθεί Σφάλμα Τύπου Ι),

για α=0,05 και β=0,20, έχουμε, 0,20 / 0,05 = 4. Στην περίπτωση αυτή, το Σφάλμα Τύπου Ι θεωρείται τέσσερις φορές πιο σοβαρό, πιο κρίσιμο, απ' ότι το Σφάλμα Τύπου ΙΙ. Αν το β = 0,20, τότε η ισχύς είναι γ = 0,80. Πολλοί ερευνητές θέτουν ως ελάχιστη αποδεκτή ισχύ ενός στατιστικού ελέγχου την τιμή 0,80 και αν ο στατιστικός έλεγχος έχει ισχύ μικρότερη δεν εκτελούν ή ξανασχεδιάζουν την έρευνα (Kirk, 1995).

5.4 Σφάλμα Τύπου ΙΙ ½

Ιδιαίτερη προσοχή χρειάζεται για την αποφυγή του λανθασμένου συμπεράσματος ότι η "αδυναμία" του ελέγχου να αποκαλύψει ένα στατιστικά σημαντικό αποτέλεσμα, όπως μια διαφορά, επίδραση ή συσχέτιση, αποτελεί απόδειξη ότι η διαφορά, ή η επίδραση, ή η συσχέτιση δεν υπάρχει στους αντίστοιχους πληθυσμούς. Ο εσφαλμένος αυτός συμπερασμός συχνά ονομάζεται «Σφάλμα Τύπου ΙΙ ½» (Kritzer, 1996) ή, σύμφωνα με το Hopkins (1997) «Σφάλμα Τύπου 0». Το σφάλμα αυτό είναι λογικό και αφορά την περίπτωση κατά την οποία, σε έναν υποθετικοπαραγωγικό συλλογισμό, θεωρείται ότι ισχύει η πρόταση του συμπεράσματος (Dometrius 1992, Kargopoulos & Raftopoulos 1998). Θα πρέπει να τονιστεί ότι η μη απόρριψη της Μηδενικής Υπόθεσης δεν σημαίνει την αποδοχή της (Kirkwood, 1996). Για ένα τέτοιο συμπέρασμα θα πρέπει να υπολογιστεί και η πιθανότητα να έχει διαπραχθεί Σφάλμα Τύπου ΙΙ. Ένα μη στατιστικά σημαντικό αποτέλεσμα σημαίνει απλά ότι τα διαθέσιμα δειγματικά ή πειραματικά δεδομένα δεν είναι αρκετά "ισχυρά" για να υποστηρίξουν την εναλλακτική υπόθεση ή, ισοδύναμα, ότι η H_0 δεν είναι αληθής. Με άλλα λόγια, τα δεδομένα δεν παρέχουν αρκετή μαρτυρία για την απόρριψη της μηδενικής υπόθεσης και συνεπώς η H_0 "παραμένει".

5.5 Παρατηρούμενη Στάθμη Σημαντικότητας (*p*-value)

Στα στατιστικά πακέτα υπολογίζεται η «παρατηρούμενη στάθμη σημαντικότητας» (π.σ.σ.) (*p-value* ή *probability level*) των στατιστικών ελέγχων. Η π.σ.σ. εκφράζει την πιθανότητα να παρατηρηθεί μια τιμή του αντίστοιχου στατιστικού του ελέγχου μεγαλύτερη ή ίση από αυτήν που έδωσε το δείγμα με δεδομένο ότι η H_0 είναι αληθής (James *et al*. 1997, Sackrowitz & Samuel-Cahn 1999), δηλαδή:

$$p = P(Z \geq |z'| \,/\, H_0 \text{ είναι αληθής}), \qquad [5.9]$$

όπου Z είναι η τυχαία μεταβλητή που αντιστοιχεί στο στατιστικό του ελέγχου και z' η τιμή του στατιστικού για το συγκεκριμένο δείγμα (π.χ. t, F και χ^2).

Η τιμή της π.σ.σ. στηρίζεται στα δεδομένα και αποτελεί τη βάση πάνω στην οποία θα στηριχθεί η απόφαση σχετικά με το αν θα απορριφθεί η H_0 ή όχι. Αν η π.σ.σ. ενός ελέγχου είναι μικρότερη ή το πολύ ίση με το ε.σ. α που έχει προκαθοριστεί, τότε απορρίπτεται η H_0 σε ε.σ. α (Dometrius 1992, Kirk 1995, Kinnear & Gray 1999, Pagano & Gauvreau 2000). Αν η π.σ.σ. είναι μεγαλύτερη από το ε.σ. α που προκαθορίστηκε, τότε δεν απορρίπτεται η H_0. Η π.σ.σ. εκφράζει την πιθανότητα ένα στατιστικό αποτέλεσμα, τόσο μεγάλο ή μεγαλύτερο από το παρατηρούμενο, θα μπορούσε να συμβεί στην "τύχη" αν η H_0 είναι αληθής (Bryman & Cramer, 1999). Η τιμή της π.σ.σ. εκφράζει το χαμηλότερο ε.σ. στο οποίο μπορεί να απορριφθεί η H_0 (Τσάντας και άλλοι 1999, Χάλκος 2000). Να τονιστεί ότι σε κάθε περίπτωση το τι ισχύει στην πραγματικότητα σχετικά με την H_0 είναι άγνωστο.

5.6 Κριτική στους Ελέγχους Σημαντικότητας της H_0

Παρόλο που στους στατιστικούς ελέγχους υποθέσεων η Θεωρία Πιθανοτήτων και η Στατιστική συνδυάζονται ώστε να αποτελέσουν έναν πολύτιμο οδηγό για τη λήψη αποφάσεων κάτω από συνθήκες αβεβαιότητας, ωστόσο σχετικά με τη διαδικασία ελέγχου σημαντικότητας της H_0 (*Null-Hypothesis Significance-Test Procedure-NHSTP*) έχει ασκηθεί κριτική ήδη από τη δεκαετία του '60, η οποία και επαναλαμβάνεται περιοδικά από αρκετούς ερευνητές (Yates1951, Kish 1959, Rozeboom 1960, Bakan 1966, Pratt 1976, Cox 1977, Carver 1978, Parkhurst 1985,

Guttman 1985, Oakes 1986, Cohen 1988, Chatfield 1991, Loftus 1991, Yoccuz 1991, Habbard & Armstrong 1992, Schmidt 1996, Murphy & Myors 1998, Nix & Barnett 1998, Brandstätter 1999, Huck 2000β, Haller & Krauss 2002). Μπορούμε να συνοψίσουμε τα βασικά επιχειρήματα κατά της NHSTP ως εξής:

- Η στατιστική σημαντικότητα ενός αποτελέσματος μπορεί να είναι αποτέλεσμα της κατάλληλης επιλογής του μεγέθους του δείγματος και του επιπέδου σημαντικότητας α.
- Η H_0 δεν μπορεί ποτέ να είναι αληθινή.
- Από τη στατιστική σημαντικότητα δεν μπορούμε να εξάγουμε συμπεράσματα για την αντίστροφη πιθανότητα της υπόθεσης, δηλαδή την πιθανότητα η H_0 να είναι αληθής με βάση τα δεδομένα.
- Η στατιστική σημαντικότητα δεν δίνει πληροφορίες σχετικά με τις τιμές των παραμέτρων του ή των πληθυσμών.
- Ο έλεγχος του σφάλματος Τύπου ΙΙ και η Ανάλυση Ισχύος των ελέγχων παραμελούνται αδικαιολόγητα.
- Από τη στατιστική σημαντικότητα δεν μπορεί να εξαχθεί κάποιο συμπέρασμα σχετικά με την πρακτική ή κλινική σημαντικότητα ενός αποτελέσματος.
- Η δυαδική λογική της NHSTP (η H_0 απορρίπτεται ή όχι) δεν συμβαδίζει με το γεγονός ότι η γνώση αποχτιέται βήμα προς βήμα.
- Οι προϋποθέσεις (τεχνικές και θεωρητικές) εφαρμογής των στατιστικών ελέγχων σπάνια ικανοποιούνται στην πράξη.
- Η διαδικασία εγκυμονεί κινδύνους για στοχαστικά και λογικά σφάλματα καθώς και για παρανοήσεις.

Όμως, στις περισσότερες περιπτώσεις, η κριτική που έχει ασκηθεί στηρίζεται σε λόγους που δεν αφορούν στην ίδια τη στατιστική διαδικασία, αλλά κυρίως στο γεγονός ότι οι λανθασμένες αντιλήψεις των ερευνητών και ο στοχαστικός αναλφαβητισμός είναι οι παράγοντες που οδηγούν σε εσφαλμένη χρήση και ερμηνεία των αποτελεσμάτων των στατιστικών ελέγχων σημαντικότητας της H_0 (Huck 2000α και 2000β, Μενεξές & Οικονόμου 2001, Harris 2001).

5.7 Ανάλυση Ισχύος

Η Ανάλυση Ισχύος (ΑΙ) πραγματοποιείται, συνήθως, κατά τη διάρκεια σχεδιασμού της έρευνας ή του πειράματος, δηλαδή πριν ακόμα από τη συλλογή των δεδομένων (*a priori*), και χρησιμοποιείται για την εκτίμηση της πιθανότητας να απορριφθεί μια όντως λανθασμένη H_0. Με άλλα λόγια, με την ΑΙ επιχειρείται μια προσέγγιση του βαθμού εμπιστοσύνης που θα αποδοθεί στην "ικανότητα" ή τη "δύναμη" του ελέγχου να αναδείξει όντως στατιστικά σημαντικά αποτελέσματα (Χατζηνικολάου, 2002). Η ισχύς ενός στατιστικού ελέγχου γ εξαρτάται κυρίως από τρεις παράγοντες (Cohen 1988, Murphy & Myors 1998):

α) Το επίπεδο σημαντικότητας α
β) Το μέγεθος του δείγματος N, και
γ) Το μέγεθος ή την "ένταση" του αποτελέσματος (*Effect Size-ES*).

Το μέγεθος του αποτελέσματος μπορεί να οριστεί γενικά ως η έκταση ή η ένταση του υπό εξέταση φαινομένου (Cohen & Cohen, 1983). Αποτελεί ένα μέτρο του βαθμού στον οποίο ένα φαινόμενο πραγματοποιείται (Cohen, 1965). Από στατιστική σκοπιά, το *ES* μπορεί να θεωρηθεί ως ο βαθμός απόκλισης του παρατηρούμενου αποτελέσματος, όπως αυτό δηλώνεται στην εναλλακτική υπόθεση H_1, από το αποτέλεσμα που δηλώνεται κάτω από την ορθότητα της H_0 (Kramer & Rosental, 1999). Η λειτουργικότητα του *ES*, μέσα στο πλαίσιο της ΑΙ, έγκειται στο γεγονός ότι είναι συνήθως καθαρός αριθμός, απαλλαγμένος από μονάδες μέτρησης, και, κυρίως, ανεξάρτητος από το μέγεθος του δείγματος (Cohen, 1988). Σε κάθε στατιστικό έλεγχο αντιστοιχεί και διαφορετικό *ES* και μπορεί να μετρηθεί με δύο τρόπους (Cohen 1988, Kramer & Rosental 1999, Murphy & Myors 1998, Χατζηνικολάου 2002, Μενεξές 2002): α) ως διαφορά μέσων όρων ή ποσοστών, τυποποιημένη ή μη (για παράδειγμα, οι διαφορές *Cohen's d*, *Hedges's g* και *Glass's delta*), ή β) ως συντελεστής συσχέτισης - συνάφειας ή συνάρτησής του (για παράδειγμα, οι συντελεστές r, R^2, η^2, ω^2 και φ). Στην πράξη, το *ES* προκαθορίζεται από τον ερευνητή, που σχεδιάζει τη μελέτη, κάτω από "ενδιαφέρουσες" εναλλακτικές υποθέσεις (Κάτος, 1986, Sackrowitz & Samuel-Cahn 1999).

Οι τρεις παραπάνω παράγοντες μαζί με την ισχύ αποτελούν ένα κλειστό σύστημα[18], με την έννοια ότι εάν τρία από τα στοιχεία του συστήματος είναι γνωστά και σταθερά, τότε το τέταρτο μπορεί να καθοριστεί πλήρως (Cohen & Cohen 1983, Cohen 1988, Nemec 1991, Murphy & Myors 1998). Ειδικότερα, για δοσμένα (σταθερά) *N* και *ES* η ισχύς του ελέγχου μεγαλώνει καθώς αυξάνεται το *α*, για δοσμένα *α* και *ES* η ισχύς αυξάνεται καθώς μεγαλώνει το *N* και για δοσμένα *α* και *N* η ισχύς μεγαλώνει καθώς αυξάνεται το *ES* (βλέπε Διαγράμματα Ε2.5, Ε2.1 και Ε2.4 αντίστοιχα, στην Παρατήρηση Ε2.1 της Ενότητας Ε2 στο Παράρτημα Ε). Σκοπός της ΑΙ είναι να εξισορροπήσει κατάλληλα τους τέσσερις παράγοντες του συστήματος λαμβάνοντας υπόψη τόσο τους θεωρητικούς όσο και τους πρακτικούς στόχους της έρευνας σε συνδυασμό με τους πόρους (οικονομικούς, χρονικούς και τεχνολογικούς) που έχει στη διάθεσή του ο ερευνητής. Η εξισορρόπηση αυτή δεν θα πρέπει να αντιτίθεται στους ηθικούς-δεοντολογικούς περιορισμούς της έρευνας.

Σε πρακτικό επίπεδο η ΑΙ, μεταξύ άλλων, μπορεί να απαντήσει στα δύο παρακάτω βασικά ερωτήματα:

α) Ποιο είναι το ελάχιστο μέγεθος δείγματος *N* ώστε σε επίπεδο σημαντικότητας *α* και για επίπεδο ισχύος *γ* ο στατιστικός έλεγχος που θα εφαρμοστεί να διαγνώσει ως στατιστικά σημαντικό ένα *ES* ίσο με *d*; Στην περίπτωση αυτή, το *d*, π.χ. 0,20, αποτελεί μια εκτίμηση του μικρότερου *ES* που έχει πρακτική ή κλινική σημαντικότητα για τον ερευνητή και έχει αξία να ανιχνευθεί ως στατιστικά σημαντικό.

β) Δοθέντος του μεγέθους δείγματος *N*, του επιπέδου σημαντικότητας *α* και του παρατηρούμενου *ES*, ποια είναι ισχύς *γ* του στατιστικού ελέγχου;

Η απάντηση στο ερώτημα α) αποτελεί την *a priori* προσέγγιση στην ανάλυση ισχύος, ενώ η απάντηση στο ερώτημα β) την *post-hoc*.

---

[18] Ο Μενεξές (2002) το ονομάζει «Σύστημα *α*-*γ*-*N*-*ES*».

5.8 Ανάλυση Ισχύος στην Παραγοντική Ανάλυση των Αντιστοιχιών

Έστω **F** ο πίνακας συμπτώσεων απολύτων συχνοτήτων δύο κατηγορικών μεταβλητών X και Y με k και l κλάσεις αντίστοιχα. Έστω, επίσης, ότι ισχύουν οι προϋποθέσεις εφαρμογής του στατιστικού ελέγχου ανεξαρτησίας χ^2. Τότε ο στατιστικός έλεγχος που πραγματοποιείται είναι ο παρακάτω:

Μηδενική Υπόθεση	H_0: Οι X και Y είναι ανεξάρτητες
Εναλλακτική Υπόθεση	H_1: όχι η H_0,
	σε επίπεδο σημαντικότητας α.

Η H_0 απορρίπτεται αν $Q > \chi^2_{(k-1)(l-1);\alpha}$, όπου Q είναι το στατιστικό χ^2 που αντιστοιχεί στον πίνακα **F** και $\chi^2_{(k-1)(l-1);\alpha}$ η κρίσιμη τιμή της χ^2 Κατανομής, σε επίπεδο σημαντικότητας α , με (k-1)(l-1) βαθμούς ελευθερίας (β.ε.).

Αν η H_0 είναι αληθής, τότε το Q ακολουθεί ασυμπτωτικά την Κατανομή χ^2 με (k-1)(l-1) β.ε. Αν, όμως, η H_1 είναι αληθής, τότε το Q ακολουθεί οριακά τη Μη Κεντρική (*non-central*) χ^2 Κατανομή με (k-1)(l-1) β.ε και παράμετρο μη κεντρικότητας ή εκκεντρότητας λ (*non-centrality parameter*) (Cochran 1952, Chapman & Nam 1968, Lachin 1977, Guenther 1977, Χατζηνικολάου 2002). Για περισσότερες πληροφορίες σχετικά με τη Μη Κεντρική χ^2 Κατανομή παραπέμπουμε στους Patnaik (1949), Sankaran (1963), Guenther (1964), Johnson και Pearson (1969), Han (1975), Lal Saxena και Alam (1982), Bishop, Fienberg και Holland (1991), Kent και Hainsworth (1995), López-Blázquez (2000) και Benton και Krishnamoorthy (2003).

Γενικά, για την παράμετρο μη κεντρικότητας λ ισχύει (Lachin, 1977):

$$\lambda = Nf(\boldsymbol{\theta}^0, \boldsymbol{\theta}^a), \qquad [5.10]$$

όπου N είναι το μέγεθος δείγματος και f η συνάρτηση των διανυσμάτων των παραμέτρων $\boldsymbol{\theta}^{\mathbf{0}}$ και $\boldsymbol{\theta}^{\mathbf{a}}$ που εμπλέκονται στο στατιστικό έλεγχο χ^2 κάτω από την ισχύ των H_0 και H_1 αντίστοιχα.

Από μια άλλη σκοπιά, η f μπορεί να θεωρηθεί ως ο βαθμός απόκλισης του παρατηρούμενου αποτελέσματος από την κατάσταση που δηλώνεται μέσω της H_0 και, συνεπώς, αποτελεί συνάρτηση του αντίστοιχου *ES* του στατιστικού ελέγχου. Από τη σχέση [5.10] προκύπτει:

$$N=\lambda / f(\boldsymbol{\theta}^{\mathbf{0}}, \boldsymbol{\theta}^{\mathbf{a}}). \qquad [5.11]$$

Επομένως, αν εκτιμηθεί η παράμετρος λ και το αντίστοιχο *ES*, τότε από τη σχέση [5.11] μπορεί να υπολογιστεί το ελάχιστο μέγεθος δείγματος N που απαιτείται ώστε σε επίπεδο σημαντικότητας α και για επίπεδο ισχύος γ ο στατιστικός έλεγχος χ^2 να διαγνώσει ως στατιστικά σημαντικό το αντίστοιχο *ES*.

5.8.1 *Post-hoc* Ανάλυση Ισχύος

Λαμβάνοντας υπόψη τη σχέση [5.6] και όσα αναφέρθηκαν στην προηγούμενη ενότητα, το παρατηρούμενο Σφάλμα Τύπου ΙΙ β_{obs} υπολογίζεται ως εξής:

$$\beta_{obs} = P\left(Q < \chi^2_{(k-1)(l-1);a} / H_0 \text{ λανθασμένη}\right) = P\left(\chi^2_{nc(k-1)(l-1);\alpha}(\lambda) < \chi^2_{(k-1)(l-1);a}\right), \qquad [5.12]$$

όπου $\chi^2_{nc(k-1)(l-1)}(\lambda)$ είναι η τιμή της μη κεντρικής χ^2 Κατανομής με παράμετρο λ και $(k\text{-}1)(l\text{-}1)$ β.ε.

Λόγω της σχέσης [5.8] η παρατηρούμενη ισχύς γ_{obs} του ελέγχου χ^2 δίνεται από την παρακάτω σχέση:

$$\gamma_{obs} = 1-\beta_{obs} = P\left(\chi^2_{nc(k-1)(l-1);\alpha}(\lambda) \geq \chi^2_{(k-1)(l-1);a}\right). \qquad [5.13]$$

Για τον υπολογισμό της γ_{obs} θα πρέπει να εκτιμηθεί η παράμετρος λ. Από τον Cohen (1988) έχουμε ότι:

$$\lambda=Nw^2, \qquad [5.14]$$

όπου *N* είναι το μέγεθος του δείγματος και *w* μια εκτίμηση του *ES*, όπως ορίζεται από τον Cohen για τον έλεγχο ανεξαρτησίας χ^2 δύο μεταβλητών.

Το *ES w* εκτιμάται γενικά από την παρακάτω σχέση (Cohen, 1988):

$$w = \sqrt{\sum_{i}^{kl} \frac{(p_{1i} - p_{0i})^2}{p_{0i}}}, \qquad [5.15]$$

όπου p_{1i} και p_{0i} είναι οι σχετικές συχνότητες του κελιού *i* του πίνακα συμπτώσεων κάτω από την ισχύ των H_1 και H_0 αντίστοιχα. Στην συγκεκριμένη περίπτωση του πίνακα **F**, η σχέση [5.15] μέσω της σχέσης [5.1] γράφεται και ως εξής:

$$w = \sqrt{\sum_{i}\sum_{j} \frac{(p_{ij} - r_i c_j)^2}{r_i c_j}} = \sqrt{I_F}\ . \qquad [5.16]$$

Ο δείκτης *w* κυμαίνεται από 0 δηλώνοντας την ανεξαρτησία των δύο μεταβλητών έως τη μέγιστη τιμή $\sqrt{p}$, όπου *p*=*min*(*k*-1, *l*-1) (Cohen, 1988). Η οριακή μέγιστη τιμή του *w* δηλώνει απόλυτη συνάφεια μεταξύ των δύο μεταβλητών. Επομένως, η μέγιστη δυνατή τιμή της αδράνειας $I_F = w^2$ ενός πίνακα συμπτώσεων δύο μεταβλητών είναι ίση με *p*. Στο ίδιο συμπέρασμα καταλήγουμε αν θεωρήσουμε τη σχέση [2.4] (βλέπε Ενότητα 2.2.5), από την οποία έχουμε ότι $I_F = \varphi^2$, όπου φ^2 είναι ο συντελεστής συνάφειας μέσου τετραγώνου του *Pearson*. Σύμφωνα με τους Kalantari *et al*., (1993) η μέγιστη δυνατή τιμή του φ^2 για ένα πίνακα συμπτώσεων *k*×*l* είναι *min*(*k*-1, *l*-1), δηλαδή ίση με *p* (βλέπε επίσης Leik & Gove 1971, Acock & Stavig 1979).

Από τη σχέση [5.16] συνεπάγεται ότι:

$$w^2 = I_F\ . \qquad [5.17]$$

Επίσης, από τη σχέση [5.2] είναι φανερό ότι:

$$w^2 = \frac{Q}{N}\ . \qquad [5.18]$$

Μέσω των σχέσεων [5.3] και [5.17], η σχέση [5.14] γράφεται:

$$\lambda = Nw^2 = NI_F = Q. \qquad [5.19]$$

Συνεπώς, η παρατηρούμενη ισχύς του ελέγχου ανεξαρτησίας χ^2 δίνεται από την παρακάτω σχέση:

$$\gamma_{obs} = P\left(\chi^2_{nc(k-1)(l-1);\alpha}(NI_F) \geq \chi^2_{(k-1)(l-1);\alpha}\right) = P\left(\chi^2_{nc(k-1)(l-1);\alpha}(Q) \geq \chi^2_{(k-1)(l-1);\alpha}\right). \qquad [5.20]$$

Στην πράξη, ο αριθμητικός υπολογισμός της [5.20] μπορεί να γίνει με τη βοήθεια πινάκων της μη κεντρικής χ^2 Κατανομής (Johnson & Pearson 1969, Haynam, Govindarajulu & Leone 1970) ή τη χρήση λογισμικών, όπως το SAS (βλέπε Castelloe 2000, Bergerud & Sit 2001, Castelloe & O'Brien 2001) και το SPSS (βλέπε Χατζηνικολάου, 2002), τα οποία διαθέτουν ειδικές συναρτήσεις για τους σχετικούς υπολογισμούς.

5.8.2 *A priori* Ανάλυση Ισχύος

Από τις σχέσεις [5.11] και [5.19] είναι δυνατός ο *a priori* υπολογισμός του ελάχιστου απαιτούμενου μεγέθους δείγματος όταν γνωρίζουμε μια εκτίμηση της παραμέτρου λ και το μεγέθους αποτελέσματος *w*. Στην περίπτωση αυτή, το μέγεθος του δείγματος δίνεται από την παρακάτω σχέση:

$$N = \frac{\lambda}{w^2} = \frac{\lambda}{I_F}. \qquad [5.21]$$

Για τη Μη Κεντρική χ^2 Κατανομή, τιμές της παραμέτρου λ (*a, β, u)* που αντιστοιχούν σε ισχύ γ=1-β και σε επίπεδο σημαντικότητας α, με *u* βαθμούς ελευθερίας, υπάρχουν σε πίνακες (Haynam Govindarajulu & Leone 1970, Pearson & Hartley 1972) ή υπολογίζονται με τη βοήθεια λογισμικών. Το πρόβλημα είναι να προκαθοριστεί μια εκτίμηση του *w* ή της I_F που να έχει κλινική ή πρακτική σημαντικότητα στο πλαίσιο της έρευνας που πρόκειται να πραγματοποιηθεί.

Παρατήρηση 5.1

Τα λογισμικά GPower[19] και Sample Power[20] διαθέτουν διαδικασίες για την AI του ελέγχου χ^2. Όμως, και στα δύο δεν υπάρχει σύνδεση με την ΠΑΑ, ενώ στο δεύτερο υπάρχει και σημαντικός περιορισμός ως προς τον αριθμό των γραμμών και στηλών των πινάκων συμπτώσεων που μπορεί χειριστεί (μέχρι 8 στην έκδοση 1.2). Οι ίδιοι περιορισμοί ισχύουν και για τη χρήση των δημοσιευμένων πινάκων (βλέπε Cohen, 1988) με τιμές για το απαιτούμενο μέγεθος δείγματος και την ισχύ του ελέγχου χ^2.

5.9 Προκαθορισμός του *ES*

Ο προκαθορισμός του *w* ή της I_F μπορεί να επιτευχθεί είτε από πιλοτικές έρευνες είτε από μετα-αναλύσεις προηγούμενων συγκρίσιμων ερευνών στο ίδιο ερευνητικό αντικείμενο (βλέπε Wolf, 1986). Μια εναλλακτική προσέγγιση είναι ο καθορισμός του *ES* μετά από δοκιμές σε πίνακες συμπτώσεων με τεχνητά δεδομένα, τέτοια ώστε να εκφράζουν κλινικά σημαντικά δομές ή διαφορές για το υπό εξέταση φαινόμενο. Με τον τρόπο αυτό, δημιουργείται μια σειρά από "ενδιαφέρουσες" εναλλακτικές υποθέσεις. Επίσης, μπορούν να χρησιμοποιηθούν οι συμβάσεις κατά Cohen (1988) σχετικά με το τι μπορεί να θεωρηθεί ως "μικρό", "μεσαίο" και "μεγάλο" μέγεθος του αποτελέσματος στο πλαίσιο του στατιστικού ελέγχου ανεξαρτησίας χ^2 (βλέπε Πίνακα 5.1).

Πίνακας 5.1: Συμβάσεις κατά Cohen και Αντιστοιχία Μεταξύ *w* και I_F

Μικρό *ES*	Μέτριο *ES*	Μεγάλο *ES*
w=0,10	*w*=0,30	*w*=0,50
I_F =0,01	I_F =0,09	I_F =0,25

Πιο ενδιαφέρουσα φαίνεται να είναι η σύνδεση του *w* ή της I_F με δείκτες συνάφειας που βασίζονται στο στατιστικό *Q*. Δύο είναι οι βασικοί λόγοι που οδηγούν στην

[19] Το λογισμικό είναι freeware και είναι διαθέσιμο στη διεύθυνση: http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/

[20] Το λογισμικό είναι εμπορικό και διατίθεται από την SPSS Inc. (για περισσότερες πληροφορίες βλέπε http://www.spss.com/samplepower/).

παραπάνω πρόταση: α) Οι δείκτες αυτοί μπορούν να υπολογιστούν σχετικά εύκολα από δημοσιευμένα αποτελέσματα αντίστοιχων ερευνών και β) εκφράζουν την ένταση ή το βαθμό της συνάφειας μεταξύ των μεταβλητών σε μια κλίμακα από 0 έως μια μέγιστη τιμή ≤1. Δύο από τους πιο συχνά χρησιμοποιούμενους δείκτες συνάφειας για πίνακες συμπτώσεων δύο μεταβλητών είναι ο δείκτης *Contigency Coefficient C* και ο δείκτης *Cramer's V*.

5.9.1 Σχέση των *w* και I_F με το δείκτη *Contingency Coefficient C*

Ο δείκτης *C* δίνεται από τη σχέση (Goodman & Kruskal 1954, Reynolds 1984):

$$C = \sqrt{\frac{Q}{Q+N}}. \qquad [5.22]$$

Η σχέση [5.22] μέσω των σχέσεων [5.18] και [5.17] γράφεται και ως εξής:

$$C = \sqrt{\frac{\frac{Q}{N}}{\frac{Q}{N}+\frac{N}{N}}} = \sqrt{\frac{w^2}{w^2+1}} = \sqrt{\frac{I_F}{I_F+1}}. \qquad [5.23]$$

Από τη σχέση [5.23] προκύπτει:

$$w = \sqrt{\frac{C^2}{1-C^2}} = \sqrt{I_F}\ . \qquad [5.24]$$

Συνεπώς,

$$I_F = \frac{C^2}{1-C^2}. \qquad [5.25]$$

Η μέγιστη τιμή του δείκτη *C* εξαρτάται από τις διαστάσεις του πίνακα συμπτώσεων με αποτέλεσμα να μην είναι εφικτή η άμεση σύγκριση δεικτών *C* που προέρχονται από πίνακες διαφορετικών διαστάσεων (Cohen, 1988).

5.9.2 Σχέση των *w* και I_F με το δείκτη *Cramer's V*

Ο δείκτης *Cramer's V* δίνεται από τη σχέση (Goodman & Kruskal 1954, Reynolds 1984):

$$V = \sqrt{\frac{Q}{Np}}, \qquad [5.26]$$

όπου $p=min(k\text{-}1, l\text{-}1)$.

Η σχέση [5.26] λόγω των σχέσεων [5.18] και [5.17] μπορεί να γραφεί και ως εξής:

$$V = \frac{w}{\sqrt{p}} = \sqrt{\frac{I_F}{p}}. \qquad [5.27]$$

Από τη σχέση [5.27] έχουμε ότι:

$$w = V\sqrt{p}. \qquad [5.28]$$

Συνεπώς,

$$\sqrt{I_F} = V\sqrt{p} \Rightarrow I_F = V^2 p. \qquad [5.29]$$

5.10 Δυναμική Αδράνεια του Πίνακα Συμπτώσεων Δύο Μεταβλητών

Είδαμε στο Κεφάλαιο 2 ότι στην ΠΑΑ κεντρικό ρόλο παίζει η έννοια της αδράνειας. Από μια σκοπιά, η ολική αδράνεια ενός πίνακα συμπτώσεων δύο μεταβλητών εκφράζει μια γενικευμένη διασπορά. Από μια άλλη, μπορεί να θεωρηθεί ως ένας δείκτης της περιεχόμενης στον πίνακα πληροφορίας. Στα προηγούμενα δείχθηκε ότι η ολική αδράνεια, μέσω της σχέσης που τη συνδέει με το *w*, μπορεί να θεωρηθεί και ως ένας ολικός δείκτης μεγέθους του αποτελέσματος *ES*, ο οποίος εκφράζει το βαθμό ή την ένταση της συνάφειας μεταξύ δύο κατηγορικών μεταβλητών. Στην Ενότητα 5.8.1 παρατηρήσαμε ότι η μέγιστη δυνατή τιμή, έστω $I_{\max}$, της ολικής αδράνειας, με μόνο περιορισμό τον αριθμό γραμμών και στηλών του πίνακα συμπτώσεων, είναι ίση με *p*

όπου $p=min(k\text{-}1,\ l\text{-}1)$. Από τη σχέση [5.2] συνεπάγεται ότι η μέγιστη δυνατή τιμή $Q_{\max}$ της ποσότητας Q είναι ίση με Np.

Αν I_F είναι η (παρατηρούμενη) αδράνεια του πίνακα συμπτώσεων **F** και $I_{\max}=p$ η μέγιστη δυνατή αδράνεια του **F** με μόνο περιορισμό τον αριθμό γραμμών και στηλών του πίνακα για δοσμένο μέγεθος δείγματος N, τότε ορίζουμε ως «Δυναμική Αδράνεια» I_D του πίνακα **F** το λόγο:

$$I_D = \frac{I_F}{I_{\max}}. \qquad [5.30]$$

Η δυναμική αδράνεια εκφράζει την παρατηρούμενη αδράνεια ως ποσοστό της μέγιστης δυνατής $I_{\max}$ και μπορεί να πάρει τιμές στο διάστημα [0, 1]. Η σχέση [5.30] , μέσω των σχέσεων [5.2] και [5.26] μπορεί να γραφεί και ως εξής:

$$I_D = \frac{Q}{Np} = V^2. \qquad [5.31]$$

Συνεπώς, η δυναμική αδράνεια του πίνακα συμπτώσεων είναι ίση με το τετράγωνο του συντελεστή συνάφειας *Cramer's V*. Η ποσότητα $I_D\times 100$ εκφράζει τη I_F ως ποσοστό % της $I_{\max}$. Η στατιστική σημαντικότητα της I_D μπορεί να ελεγχθεί με τη σύγκριση της ποσότητας $NpI_D = Q$ με την κρίσιμη τιμή της Κατανομής χ^2 με $(k-1)(l-1)$ β.ε., σε ε.σ. α (βλέπε Srikantan, 1970). Επίσης, από τη σχέση [5.31], αν λάβουμε υπόψη ότι $I_F = Q/N$ και ότι p είναι ο μέγιστος αριθμός παραγοντικών αξόνων που μπορούν να προκύψουν από την εφαρμογή της ΠΑΑ στον πίνακα **F**, προκύπτει ότι:

$$I_D = \frac{I_F}{p}. \qquad [5.32]$$

Η σχέση [5.32] δηλώνει ότι η I_D είναι ίση με τη μέση αδράνεια των p αξόνων.

Για τους ερευνητές που είναι εξοικειωμένοι με την έννοια της αδράνειας η δυναμική αδράνεια αποτελεί μια εναλλακτική προσέγγιση στον προκαθορισμό του μεγέθους

του αποτελέσματος. Στη συνέχεια, δίνουμε τις σχέσεις που συνδέουν τη I_D με το w και τη I_F.

Από τις σχέσεις [5.28] και [5.31] συνεπάγεται ότι:

$$w = \sqrt{pI_D} \ . \qquad [5.33]$$

Από τη σχέση [5.31] προκύπτει ότι:

$$I_F = pI_D \, . \qquad [5.34]$$

Η δυναμική αδράνεια, όπως ορίστηκε στα προηγούμενα, στηρίζεται σε έναν *a priori*, με γενική ισχύ, υπολογισμό της μέγιστης αδράνειας $I_{\max}$ που μπορεί να έχει ένας $k \times l$ πίνακας συμπτώσεων δύο κατηγορικών μεταβλητών. Η εκτίμηση αυτή είναι ανεξάρτητη από τις περιθώριες κατανομές συχνοτήτων των δύο μεταβλητών και λαμβάνει υπόψη μόνο τον αριθμό γραμμών και στηλών του πίνακα συμπτώσεων. Με άλλα λόγια, η τιμή της μέγιστης αδράνειας $I_{\max}$, με την έννοια του μέγιστου πληροφοριακού περιεχόμενου του πίνακα συμπτώσεων, είναι δεδομένη πριν ακόμα την πραγματοποίηση του υπό εξέταση φαινομένου, εφόσον τα k και l είναι προκαθορισμένα. Για το λόγο αυτό μπορούμε να την ονομάσουμε και «*a priori* Δυναμική Αδράνεια». Το ερώτημα που τίθεται τώρα είναι το εξής:

Ποια είναι η μέγιστη αδράνεια I_D^{ph} που μπορεί να έχει ο πίνακας συμπτώσεων **F** μετά την πραγματοποίηση του υπό εξέταση φαινομένου (*post hoc*), δοθέντων, δηλαδή, των κατανομών συχνοτήτων των δύο μεταβλητών ή ισοδύναμα της περιθώριας γραμμής και στήλης του **F**.

Το παραπάνω ερώτημα μπορεί να αποδοθεί ως πρόβλημα μη γραμμικής βελτιστοποίησης με δεσμεύσεις. Ειδικότερα, ζητάμε τη μεγιστοποίηση της αντικειμενικής συνάρτησης $f(\mathbf{F}) = I_F(\mathbf{F})$, δηλαδή:

$$\text{maximize}(\, f(\mathbf{F})) = \text{maximize}\,(I_F(\mathbf{F}) = \frac{Q}{N}) = \text{maximize}\left(\sum_i \sum_j \frac{\left(f_{ij} - \frac{f_{i+} f_{+j}}{N} \right)^2}{f_{i+} f_{+j}} \right),$$

με $\mathbf{F}=[\, f_{ij}\,]$, με $i = 1, \ldots k$ και $l = 1, \ldots l$, κάτω από τους περιορισμούς:

1) $\sum_{j=1}^{l} f_{1j} = f_{1+} = a_1$, $\sum_{j=1}^{l} f_{2j} = f_{2+} = a_2$, …, $\sum_{j=1}^{l} f_{kj} = f_{k+} = a_k$,

2) $\sum_{i=1}^{k} f_{i1} = f_{+1} = b_1$, $\sum_{i=1}^{k} f_{i2} = f_{+2} = b_2$, …, $\sum_{i=1}^{k} f_{il} = f_{+l} = b_l$,

3) $\sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij} = N$

και

4) $f_{ij} \geq 0, \quad \forall i, j$,

όπου τα a_i και b_j είναι οι παρατηρούμενες απόλυτες συχνότητες των κλάσεων της μεταβλητής γραμμών και στηλών αντίστοιχα, δηλαδή οι τιμές της περιθώριας γραμμής και στήλης.

Επειδή το μέγεθος του δείγματος *N*, κάτω από την ισχύ των δεσμεύσεων 1) και 2) για την περιθώρια γραμμή και στήλη του πίνακα συμπτώσεων, είναι καθορισμένο, το αρχικό πρόβλημα είναι ισοδύναμο με το:

$$\text{maximize}(Q) = \text{maximize}\left(\sum_i \sum_j \frac{\left(f_{ij} - \frac{f_{i+} f_{+j}}{N} \right)^2}{\frac{f_{i+} f_{+j}}{N}} \right),$$

με τους περιορισμούς:

1*) $\sum_{j=1}^{l} f_{1j} = f_{1+} = a_1$, $\sum_{j=1}^{l} f_{2j} = f_{2+} = a_2$, …, $\sum_{j=1}^{l} f_{kj} = f_{k+} = a_k$,

2*) $\sum_{i=1}^{k} f_{i1} = f_{+1} = b_1$, $\sum_{i=1}^{k} f_{i2} = f_{+2} = b_2$, …, $\sum_{i=1}^{k} f_{il} = f_{+l} = b_l$

και

3*) $f_{ij} \geq 0, \quad \forall i, j$.

Η παραπάνω κατάστρωση αντιστοιχεί σε τυπικό πρόβλημα βέλτιστης μεταφοράς φορτίου με *k* αφετηρίες, *l* προορισμούς, a_i εφόδια (προϊόντα, προμήθειες, εμπορεύματα, προσφορά) και b_j απαιτήσεις (ζήτηση) και έχει βέλτιστη λύση (βλέπε Kalantari *et al*. 1993, Tofallis 1999). Πρόκειται για πρόβλημα μη γραμμικής βελτιστοποίησης[21] και υπολογιστικά μπορεί να αντιμετωπιστεί με τη βοήθεια ειδικού λογισμικού, όπως το επιπρόσθετο *Solver* (Επίλυση) του EXCEL (βλέπε Walsh & Diamond 1995, Fylstra *et al*. 1998, Brown 2001, Τσάντας & Βασιλείου 2000, Ασημακόπουλος & Αραμπατζής 2002, Stokes & Plummer 2004, Buttrey 2005). Από τη στιγμή που υπολογιστεί η μέγιστη δυνατή τιμή, έστω $Q_F^{\max}$, του *Q* και, συνεπώς, η μέγιστη αδράνεια $I_F^{\max} = Q_F^{\max} / N$ του παρατηρούμενου πλέον πίνακα συμπτώσεων **F**, τότε μπορούμε να υπολογίσουμε και την *post-hoc* δυναμική αδράνεια I_D^{ph} από τη σχέση:

$$I_D^{ph} = \frac{I_F}{I_F^{\max}} . \qquad [5.35]$$

Όπως αναφέρθηκε στα προηγούμενα, η *a priori* μέγιστη δυνατή τιμή $Q_{\max}$ της ποσότητας *Q*, με μόνο περιορισμό τον αριθμό κλάσεων των μεταβλητών, είναι ίση με *Np*. Αν $Q_F^{\max}$ είναι η *post hoc* μέγιστη δυνατή τιμή του *Q* με επιπλέον περιορισμούς (1* και 2*) ως προς την περιθώρια γραμμή και στήλη, τότε:

$$Q_{\max} = Np \geq Q_F^{\max} \text{ που συνεπάγεται ότι } I_{\max} \geq I_F^{\max} \text{ και συνεπώς } I_D \leq I_D^{ph} .$$

Επιπλέον, η βέλτιστη τιμή $Q_F^{\max}$ ικανοποιεί τις ανισότητες (Kalantari *et al*., 1993):

$$\frac{N}{N-1}(k-1)(l-1) \leq Q_F^{\max} \leq Nmin(k-1, l-1) = Np .$$

Η παραπάνω σχέση μπορεί να χρησιμοποιηθεί για τη μερική επαλήθευση της βέλτιστης λύσης του προβλήματος.

---

[21] Η αντικειμενική συνάρτηση είναι δευτέρου βαθμού, διαχωρίσιμη και αυστηρά κυρτή (Kalantari *et al*., 1993).

Η *post-hoc* δυναμική αδράνεια I_D^{ph} αποτελεί μία ακόμη προσέγγιση στον καθορισμό του μεγέθους του αποτελέσματος *ES*, ειδικά στην περίπτωση που υπάρχουν διαθέσιμοι πίνακες συμπτώσεων από προηγούμενες έρευνες για το ίδιο υπό εξέταση φαινόμενο. Όταν το ε.σ. α και η ισχύς γ είναι σταθερά, τότε το μέγεθος του δείγματος είναι φθίνουσα συνάρτηση του *ES* (Cohen & Cohen 1983, Cohen 1988, Murphy & Myors 1998). Επομένως, στην περίπτωση που χρησιμοποιηθεί η I_D^{ph} ως εκτίμηση του *ES* για τον υπολογισμό του απαιτούμενου μεγέθους του δείγματος, αναμένουμε, εν γένει, μικρότερες τιμές για το *N*.

Στην Ενότητα Ε1 του Παραρτήματος Ε παραθέτουμε ένα παράδειγμα υπολογισμού της *post-hoc* δυναμικής αδράνειας ενός πίνακα συμπτώσεων δύο κατηγορικών μεταβλητών.

Παρατήρηση 5.2.

Σε πρακτικές εφαρμογές, για τον *a priori* καθορισμό του μεγέθους του δείγματος, είναι χρήσιμο να πραγματοποιείται σχετικός έλεγχος ευαισθησίας, δίνοντας όρια για το *ES*, και ανάλυση κόστους - ωφέλειας για την εξισορρόπηση των διαθέσιμων πόρων σε σχέση με τους στόχους της έρευνας. Στο λογισμικό Power Analysis for AFC έχουμε ενσωματώσει τη δυνατότητα ανάλυσης ευαισθησίας των τεσσάρων παραμέτρων (α, γ, *N* και *ES*) που εμπλέκονται στην ΑΙ του στατιστικού ελέγχου χ^2.

5.11 Καθορισμός του Μεγέθους Δείγματος στην Περίπτωση Πολλών Μεταβλητών

Στην περίπτωση *q* κατηγορικών μεταβλητών η ΠΑΑ μπορεί να εφαρμοστεί είτε στο λογικό πίνακα **Z** είτε στον αντίστοιχο γενικευμένο πίνακα συμπτώσεων **B** (*Burt*). Στις Ενότητες 2.3.3.5, 4.6, 4.7 και 4.9 είδαμε ότι, σε κάθε περίπτωση, η ΠΑΑ αναλύει, τελικά, ένα "πακέτο" διμεταβλητών σχέσεων ή, αλλιώς, τις αλληλεπιδράσεις όλων των μεταβλητών ανά δύο. Αν θεωρήσουμε ότι η ολική αδράνεια εκφράζει ένα μέτρο της πληροφορίας που περιέχει ο αντίστοιχος πίνακας ή ένα συνολικό μέγεθος αποτελέσματος, τότε μπορούμε να ισχυριστούμε ότι κάθε φορά αναλύεται ένας μέσος δείκτης μεγέθους αποτελέσματος, ο οποίος προκύπτει από διμεταβλητές και όχι από

πολυμεταβλητές σχέσεις. Διαπιστώσαμε, επίσης, ότι κατά την ανάλυση του πίνακα **B**, το μέρος της ολικής αδράνειας που εκφράζει τη συνάφεια όλων των μεταβλητών ανά δύο και έχει ενδιαφέρον να αναλυθεί είναι η ενδιαφέρουσα αδράνεια, η οποία εκφράζει τη μέση αδράνεια των $q(q-1)/2$, σε πλήθος, διαφορετικών απλών πινάκων συμπτώσεων που συγκροτούν τον **B**. Μάλιστα, όταν το δείγμα προέρχεται από απλή τυχαία δειγματοληψία επιβάλλεται ένας προ-έλεγχος των διμεταβλητών συσχετίσεων των μεταβλητών που θα αναλυθούν με την ΠΑΑ (βλέπε Ενότητα 4.7). Την πιο απλή προσέγγιση αποτελεί η εφαρμογή του ελέγχου ανεξαρτησίας χ^2 για τις $q(q-1)/2$, σε πλήθος, συσχετίσεις των μεταβλητών ανά δύο. Αν το πλήθος q των μεταβλητών είναι μεγάλο, τότε είναι επιβεβλημένη και κάποιου είδους διόρθωση, όπως για παράδειγμα η διόρθωση κατά *Bonferroni* (Brown & Melamed 1990, Girden 1992, Τσάντας και άλλοι 1999), του ε.σ. *a* των ελέγχων.

Με βάση τα παραπάνω, προτείνουμε το ελάχιστο απαιτούμενο μέγεθος δείγματος να υπολογιστεί για κάθε ζεύγος μεταβλητών χωριστά και, στη συνέχεια, να επιλεγεί η μεγαλύτερη από τις $q(q-1)/2$ σε πλήθος τιμές, που θα προκύψουν, ώστε να ικανοποιούνται οι απαιτήσεις σε ό,τι αφορά το μέγεθος δείγματος, όλων των ελέγχων. Βέβαια, η διαδικασία αυτή απαιτεί τον προκαθορισμό ενός *ES* για κάθε διμεταβλητή συσχέτιση.

5.12 Ανάλυση Ισχύος του Ελέγχου χ^2 Καλής Προσαρμογής και Σημαντικότητα του Υποχώρου Προβολής

Κατά την εφαρμογή της ΠΑΑ στον πίνακα συμπτώσεων $\mathbf{F}_{k\times l}$ δύο μεταβλητών η ανασύσταση των αρχικών συχνοτήτων του πίνακα επιτυγχάνεται από τη σχέση μετάβασης [2.42] (βλέπε Ενότητα 2.2.14.2):

$$p_{ij} = r_i c_j \left(1 + \sum_{s=1}^{p} \frac{\varphi_{is}\gamma_{js}}{\sqrt{\lambda_s}}\right), \text{ με } i=1,\ldots,k \text{ και } j=1,\ldots,l,$$

όπου λ_s είναι η ιδιοτιμή (αδράνεια) του παραγοντικού άξονα s και $\{\varphi_{is}, \gamma_{js}\}$ είναι οι κύριες συντεταγμένες των γραμμών και των στηλών αντίστοιχα του πίνακα **F** επί του παραγοντικού άξονα *s*, με $s=0,\ldots,p$ και $p=mim(k\text{-}1, l\text{-}1)$.

Αν πολλαπλασιάσουμε και τα δύο μέλη της παραπάνω σχέσης επί N, τότε προκύπτει:

$$f_{ij} = \frac{f_{i+} f_{+j}}{N}\left(1 + \sum_{s=1}^{p} \frac{\varphi_{is}\gamma_{js}}{\sqrt{\lambda_s}}\right). \qquad [5.36]$$

Ο στατιστικός έλεγχος καλής προσαρμογής χ^2 (Cohen, 1988) μπορεί να εφαρμοστεί για τη σύγκριση των παρατηρούμενων συχνοτήτων f_{ij} του πίνακα **F**, από τυχαίο δείγμα μεγέθους N, με τις αναμενόμενες συχνότητες μ_{ij} κάτω από την ισχύ της μηδενικής υπόθεσης H_s, η οποία δηλώνει ότι μόνο οι πρώτοι s ιδιοτιμές είναι διάφορες του μηδενός (Saporta & Tambrea 1993, Baccini, Caussinus & De Falguerolles 1993). Οι εκτιμητές σταθμισμένων ελαχίστων τετραγώνων των αναμενόμενων συχνοτήτων μ_{ij} είναι οι τιμές που προκύπτουν από τη σχέση μετάβασης [5.36] όταν στο άθροισμα χρησιμοποιούνται μόνο οι s πρώτοι όροι (Andersen, 1991). Ο έλεγχος καλής προσαρμογής μπορεί να πραγματοποιηθεί με το στατιστικό:

$$Q_s = \sum_i \sum_j \frac{\left(\mu_{ij} - f_{ij}\right)^2}{\mu_{ij}}. \qquad [5.37]$$

Στην περίπτωση που s=0, δηλαδή όταν οι δύο μεταβλητές είναι ανεξάρτητες, το στατιστικό Q_0 μπορεί να συγκριθεί, εφόσον ικανοποιούνται οι προϋποθέσεις εφαρμογής του ελέγχου, με την κρίσιμη τιμή της χ^2 Κατανομής με (k-1)(l-1) βαθμούς ελευθερίας, σε ε.σ. α. Αν p=1, τότε το στατιστικό Q_1 ακολουθεί ασυμπτωτικά τη χ^2 Κατανομή με (k-2)(l-2) βαθμούς ελευθερίας. Στη γενική περίπτωση, κάτω από την ισχύ της H_s το στατιστικό Q_s μπορεί να συγκριθεί με την κρίσιμη τιμή της χ^2 Κατανομής με (k-s-1)(l-s-1) βαθμούς ελευθερίας (Andersen 1991, Saporta & Tambrea 1993, Rao 1995). Ο στατιστικός έλεγχος χ^2 μπορεί να εφαρμοστεί για να ελέγξουμε την υπόθεση ότι οι s πρώτοι παραγοντικοί άξονες είναι αρκετοί και στατιστικά σημαντικοί για την ανασύσταση του πίνακα **F**. Κάτω από μια άλλη οπτική, με τη διαδικασία αυτή μπορούμε να εντοπίσουμε τον ελάχιστο στατιστικά σημαντικό υποχώρο, στον οποίο μπορεί να προβληθεί το υπό εξέταση φαινόμενο χωρίς ουσιαστική απώλεια πληροφορίας. Έτσι, στην πράξη, πραγματοποιούμε μια σειρά ελέγχων, ξεκινώντας με s=0 μέχρι που η μηδενική υπόθεση H_s να γίνει αποδεκτή, σε επίπεδο σημαντικότητας α. Συνεπώς, στην περίπτωση αυτή, το ενδιαφέρον εστιάζεται

στον υπολογισμό του β_{obs}, ώστε να αποφευχθεί η διάπραξη Σφάλματος Τύπου II και Τύπου II½. Η μέθοδος υπολογισμού της παρατηρούμενης ισχύος γ_{obs} του στατιστικού ελέγχου καλής προσαρμογής χ^2 μπορεί να εφαρμοστεί με τον ίδιο τρόπο όπως και στην περίπτωση του ελέγχου ανεξαρτησίας που παρουσιάσαμε στην Ενότητα 5.7.1 (βλέπε Cohen 1988, Χατζηνικολάου 2002).

Θα πρέπει να τονιστεί ότι ο έλεγχος της στατιστικής σημαντικότητας του υποχώρου προβολής είναι ανεξάρτητος από τη στατιστική σημαντικότητα της ολικής αδράνειας του πίνακα συμπτώσεων **F** (Lebart 1976, Lebart, Morineau & Tabard 1977, Lebart, Morineau & Warwick 1984, Lebart, Morineau & Piron 2000). Αυτό σημαίνει ότι η ολική αδράνεια του **F** μπορεί να μην είναι στατιστικά σημαντική αλλά η ανασύσταση του **F**, για παράδειγμα, από το παραγοντικό επίπεδο 1×2 να μην οδηγεί σε απώλεια στατιστικά σημαντικής πληροφορίας για τον υπό εξέταση πίνακα.

Παρατήρηση 5.3

Το στατιστικό Q_s έχει το μειονέκτημα ότι στην περίπτωση που οι παρατηρούμενες συχνότητες f_{ij} είναι μικρές, τότε οι εκτιμώμενες συχνότητες μ_{ij} μπορούν να πάρουν και αρνητικές τιμές και, επομένως, δεν μπορεί να εφαρμοστεί ο στατιστικός έλεγχος. Για να ξεπεραστεί το παραπάνω πρόβλημα ο Malinvaud (1987) προτείνει τη χρήση της ποσότητας $\frac{f_{i+} f_{+j}}{N}$ στον παρονομαστή του κλάσματος της σχέσης [5.37]. Αυτό οδηγεί σε ένα τροποποιημένο στατιστικό Q_s^*, το οποίο δίνεται από τη σχέση:

$$Q_s^* = \sum_i \sum_j \frac{\left(\mu_{ij} - f_{ij}\right)^2}{\frac{f_{i+} f_{+j}}{N}} = N\left(\lambda_{s+1} + \lambda_{s+2} + \ldots + \lambda_p\right). \qquad [5.38]$$

Το στατιστικό Q_s^* είναι συνάρτηση των υπόλοιπων p-s, σε πλήθος, αδρανειών και ακολουθεί και αυτό ασυμπτωτικά την χ^2 Κατανομή με (k-s-1)(l-s-1) βαθμούς ελευθερίας. Στην ουσία, μέσω του στατιστικού Q_s^*, ελέγχεται η στατιστική σημαντικότητα της πληροφορίας που χάνεται και δεν ερμηνεύεται από τους πρώτους s άξονες. Κάτω από μια σχετικά "χαλαρή" θεώρηση ο έλεγχος καλής προσαρμογής αντιστοιχεί στον έλεγχο σημαντικότητας του ποσοστού της ολικής αδράνειας που

ερμηνεύει ο υποχώρος προβολής. Στατιστικούς ελέγχους για τη σημαντικότητα του κάθε άξονα ξεχωριστά παρουσιάζουμε στην Ενότητα 6.2.13 του Κεφαλαίου 6.

Αν το διαθέσιμο δείγμα δεν μπορεί να θεωρηθεί τυχαίο ή αν δεν ικανοποιούνται οι προϋποθέσεις εφαρμογής του ελέγχου χ^2, τότε ένα ακόμη εμπειρικό κριτήριο για τον εντοπισμό των σημαντικών αξόνων μπορεί να στηριχθεί στη φυσική ερμηνεία της δυναμικής αδράνειας του πίνακα **F**, όπως αυτή ορίζεται μέσω της σχέσης [5.32]. Εφόσον η I_D εκφράζει τη μέση αδράνεια των p αξόνων ως σημαντικοί μπορούν να θεωρηθούν οι άξονες που η αδράνειά τους είναι μεγαλύτερη από το μέσο όρο, δηλαδή μεγαλύτερη από τη δυναμική αδράνεια.

5.13 Παραδείγματα Εφαρμογών

Οι απαραίτητοι αριθμητικοί υπολογισμοί πραγματοποιήθηκαν στο EXCEL με τη βοήθεια του επιπρόσθετου *π-face*[22] το οποίο είναι διαθέσιμο στη διεύθυνση: ftp://ftp.stat.uiowa.edu/pub/rlenth/PiFace/.

Προτιμήθηκε το λογισμικό EXCEL, ως πλατφόρμα για τους υπολογισμούς, για δύο κυρίως λόγους: α) είναι διαδεδομένο, και β) παρέχει τη δυνατότητα δημιουργίας σεναρίων και *What if Analysis*.

Το *π-face* διαθέτει εκτός άλλων και τις συναρτήσεις *Chi2Power* και *Chi2PowerNC*. Η συνάρτηση *Chi2Power* δέχεται ως ορίσματα το επιθυμητό επίπεδο σημαντικότητας α του ελέγχου χ^2, την παράμετρο μη κεντρικότητας λ και τους αντίστοιχους βαθμούς ελευθερίας. Η συνάρτηση επιστρέφει την παρατηρούμενη ισχύ γ_{obs} του ελέγχου και, επομένως, μπορεί να χρησιμοποιηθεί για την *post-hoc* προσέγγιση της AI. Η συνάρτηση *Chi2PowerNC* δέχεται ως ορίσματα το επιθυμητό επίπεδο σημαντικότητας του ελέγχου χ^2, την επιθυμητή ισχύ γ του ελέγχου και τους αντίστοιχους βαθμούς ελευθερίας. Επιστρέφει την παράμετρο μη κεντρικότητας λ και, συνεπώς, μπορεί να χρησιμοποιηθεί στην *a priori* προσέγγιση της AI.

---

[22] Το *PiFace* αναπτύχθηκε από τον Russell V. Lenth (Associate Professor, Department of Statistics and Actuarial Science, The University of Iowa). Περιλαμβάνεται στο Παράρτημα CDB του CD που συνοδεύει τη διατριβή.

5.13.1 *Post hoc* Ανάλυση Ισχύος

Έστω ότι έχουμε δύο κατηγορικές μεταβλητές *X* και *Y* με τρεις κλάσεις η κάθε μία. Το μέγεθος του δείγματος είναι *N*=80. Ας υποθέσουμε ότι ο στατιστικός έλεγχος χ^2 έδωσε *Q*=14,32. Η τιμή του *Q* για 4 β.ε. είναι στατιστικά σημαντική σε ε.σ. α=0,05 (p<0,05). Το πρόβλημα που τίθεται είναι να υπολογιστεί η παρατηρούμενη ισχύς γ_{obs} του ελέγχου χ^2 που αντιστοιχεί σε ε.σ. α=0,05. Η αδράνεια *I* του αντίστοιχου πίνακα συμπτώσεων των δύο μεταβλητών μπορεί να υπολογιστεί από τη σχέση [5.2] και είναι *I* =0,179. Από τη σχέση [5.16] προκύπτει ότι η αδράνεια *I* αντιστοιχεί σε μέγεθος αποτελέσματος *w*=0,423 (τείνει προς μεγάλο *ES* σύμφωνα με τις συμβάσεις κατά Cohen, βλέπε Πίνακα 5.1).

Από τη σχέση [5.19] μπορεί να εκτιμηθεί η παράμετρος μη κεντρικότητας λ η οποία είναι λ=14,32. Κάνοντας χρήση της σχέσης [5.20] και με τη βοήθεια της συνάρτησης *Chi2Power* η παρατηρούμενη ισχύς του ελέγχου εκτιμάται σε γ_{obs}=0,875. Συνεπώς, η πιθανότητα ο έλεγχος χ^2 να ανιχνεύσει ένα *ES* ίσο με το παρατηρούμενο ως στατιστικά σημαντικό, σε ε.σ. α=0,05, είναι περίπου 87,5%.

Η ίδια διαδικασία μπορεί να χρησιμοποιηθεί και για την *post hoc* AI του ελέγχου σημαντικότητας της ενδιαφέρουσας αδράνειας (βλέπε Ενότητες 4.7 και 4.8.2). Στην περίπτωση αυτή, κάτω από την υπόθεση της αμοιβαίας ανεξαρτησίας των *q* μεταβλητών που συμμετέχουν στην ανάλυση, η παράμετρος μη κεντρικότητας του ελέγχου είναι ίση με $\lambda = \sum_{h<w} Q_{hw} = \sum_{h<w} \lambda_{hw} = Q^*$, για $h, w = 1, \ldots, q$, όπου λ_{hw} είναι η παράμετρος μη κεντρικότητας που αντιστοιχεί στο στατιστικό Q_{hw} (Rencher 2000, Rao 2002).

5.13.2 *A priori* Ανάλυση Ισχύος

Έστω ότι κατά το στάδιο σχεδιασμού μιας έρευνας το ενδιαφέρον εστιάζεται στον έλεγχο της συνάφειας δύο κατηγορικών μεταβλητών *X* και *Y* με τρεις και τέσσερις κλάσεις αντίστοιχα. Από προηγούμενη εμπειρία είναι γνωστό ότι *ES* που αντιστοιχεί σε αδράνεια *I* τουλάχιστον της τάξης του 0,04 έχει κλινική ή πρακτική

σημαντικότητα σύμφωνα με του στόχους και το θεωρητικό πλαίσιο της έρευνας. Το πρόβλημα που τίθεται είναι να εκτιμηθεί το ελάχιστο απαιτούμενο μέγεθος δείγματος N ώστε ο στατιστικός έλεγχος χ^2 (για 6 β.ε.), σε ε.σ. α=0,10 και με ισχύ γ=0,99 να ανιχνεύσει το προκαθορισμένο *ES* ως στατιστικά σημαντικό.

Από τις σχέσεις [5.16], [5.27] για p=2 και [5.31] προκύπτει ότι η αδράνεια I=0,04 αντιστοιχεί σε w=0,20, σε δείκτη *Cramer's V*=0,14 και σε δυναμική αδράνεια I_D =0,019. Με τη βοήθεια της συνάρτησης *Chi2PowerNC* υπολογίζεται η παράμετρος μη κεντρικότητας λ=24,65. Επομένως, από τη σχέση [5.21] το ζητούμενο μέγεθος δείγματος υπολογίζεται σε N=617 δειγματοληπτικές μονάδες. Αν το επιθυμητό επίπεδο ισχύος του ελέγχου μειωθεί σε 0,95, τότε το εκτιμώμενο μέγεθος δείγματος είναι N=447. Στην περίπτωση που τεθεί γ=0,90, το αντίστοιχο μέγεθος δείγματος εκτιμάται σε N=367.

Στο παράδειγμα αυτό, ο προκαθορισμός του κλινικά σημαντικού *ES* θα μπορούσε από την αρχή να στηριχθεί στο δείκτη *V* ή στη δυναμική αδράνεια.

Θα πρέπει να επισημάνουμε ότι το μέγεθος του δείγματος, το οποίο θα επιδιωχθεί να συγκεντρωθεί τελικά, εξαρτάται και από τους διαθέσιμους πόρους (για παράδειγμα, οικονομικούς και χρονικούς). Επομένως, κατά το στάδιο σχεδιασμού της έρευνας θα πρέπει να ληφθούν υπόψη και τυχόν άλλοι, εξωγενείς σε σχέση με το αντικείμενο της έρευνας, περιορισμοί.

5.13.3 Καθορισμός του Μεγέθους Δείγματος στην Περίπτωση Τριών Μεταβλητών

Έστω ότι κατά το στάδιο σχεδιασμού μιας έρευνας το ενδιαφέρον εστιάζεται στον έλεγχο της συνάφειας ανά δύο, τριών κατηγορικών μεταβλητών *X*, *Y* και *Z* με τρεις, τέσσερις και πέντε κατηγορίες αντίστοιχα. Από προηγούμενη εμπειρία είναι γνωστό ότι για το ζεύγος μεταβλητών (*X*, *Y*) το *ES* που αντιστοιχεί σε δυναμική αδράνεια τουλάχιστον της τάξης του 0,005 έχει κλινική ή πρακτική σημαντικότητα, ενώ για τα ζεύγη (*X*, *Z*) και (*Y*, *Z*) τα αντίστοιχα *ES* είναι της τάξης του 0,045 και 0,085. Το πρόβλημα που τίθεται είναι να εκτιμηθεί το ελάχιστο απαιτούμενο μέγεθος δείγματος

για κάθε έναν από τους τρεις ελέγχους χ^2, σε ε.σ. α=0,05, με ισχύ γ=0,80. Τα δεδομένα και τα αποτελέσματα του προβλήματος δίνονται στον Πίνακα 5.2.

Πίνακας 5.2: Μέγεθος Δείγματος για Κάθε Ζεύγος Συσχετίσεων

Ζεύγη Συσχετίσεων	p	Κλινικά Σημαντικό *ES*: Δυναμική Αδράνεια I_D	Μέγεθος Δείγματος N για α=0,05 και γ=0,80
(X, Y)	2	0,005	1503
(X, Z)	2	0,045	152
(Y, Z)	3	0,085	68

Εφαρμόζοντας τη μεθοδολογία της *a priori* AI μπορεί να υπολογιστεί το μέγεθος δείγματος για κάθε ζεύγος μεταβλητών. Από την τελευταία στήλη του Πίνακα 5.2 διαπιστώνεται ότι με μέγεθος δείγματος N=1503 ικανοποιούνται οι απαιτήσεις και των τριών ελέγχων χ^2. Λόγω του γεγονότος ότι θα πραγματοποιηθούν τρεις στατιστικοί έλεγχοι χ^2 το επίπεδο σημαντικότητας μπορεί να διορθωθεί κατά *Bonferroni*, ώστε το Αθροιστικό Σφάλμα Τύπου I να είναι ίσο με 0,05 (βλέπε Ενότητα Ε3.2 του Παραρτήματος Ε). Στην περίπτωση αυτή, το ε.σ. α για κάθε έλεγχο μπορεί να προκαθοριστεί στο $0,05/3 \approx 0,0167$ και, στη συνέχεια, να γίνουν οι σχετικοί υπολογισμοί. Βέβαια, μόνο για τρεις ελέγχους η παραπάνω διόρθωση είναι αρκετά "αυστηρή", με την έννοια ότι το απαιτούμενο μέγεθος δείγματος θα αυξηθεί σημαντικά, ενώ ταυτόχρονα θα μειωθεί η ισχύς των αντίστοιχων στατιστικών ελέγχων. Απλά, παρουσιάζουμε τη διόρθωση για τις ανάγκες του παραδείγματος.

5.13.4 Ανάλυση Ισχύος του Ελέγχου χ^2 Καλής Προσαρμογής και Σημαντικότητα του Υποχώρου Προβολής

Ο Πίνακας 5.3 παρουσιάζει την κοινή κατανομή απολύτων συχνοτήτων δύο κατηγορικών τυχαίων μεταβλητών X και Y με k=7 και l=6 κατηγορίες αντίστοιχα. Επομένως, ο μέγιστος αριθμός παραγοντικών αξόνων που μπορούν να προκύψουν από την εφαρμογή της ΠΑΑ στον αντίστοιχο πίνακα συμπτώσεων **F** είναι p=5. Καταρχήν, να παρατηρήσουμε ότι ο στατιστικός έλεγχος ανεξαρτησίας χ^2 έδειξε ότι υπάρχει στατιστικά σημαντική συσχέτιση - συνάφεια μεταξύ των δύο μεταβλητών, σε ε.σ. α=0,05 (Q=160,061, β.ε.=30, p=0,000 και *Cramer's V*=0,482). Λόγω την υψηλής τιμής του δείκτη V ή ένταση της σχέσης φαίνεται να είναι ισχυρή. Από την εφαρμογή της ΠΑΑ προέκυψαν τα βασικά αποτελέσματα που δίνονται στον Πίνακα 5.4.

Από τον Πίνακα 5.4 διαπιστώνουμε ότι ο πρώτος παραγοντικός άξονας ερμηνεύει το 75,5% της ολικής αδράνειας, ενώ ο δεύτερος το 17,6%. Συνεπώς, επί του παραγοντικού επιπέδου 1×2 αιτιολογείται το 93,3%, σχεδόν το σύνολο, της ολικής πληροφορίας του πίνακα συμπτώσεων. Λόγω της υψηλής ερμηνευτικής ικανότητας του πρώτου άξονα, με μεγάλη διαφορά από το δεύτερο, το υπό εξέταση φαινόμενο θα μπορούσε να θεωρηθεί ως μονοδιάστατο. Στο ίδιο συμπέρασμα καταλήγουμε αν χρησιμοποιήσουμε και το κριτήριο που προτείναμε στην Ενότητα 5.12 (Παρατήρηση 5.3), σύμφωνα με το οποίο ως σημαντικοί μπορούν να θεωρηθούν οι άξονες για τους οποίους η αδράνειά τους είναι μεγαλύτερη από τη δυναμική αδράνεια του πίνακα συμπτώσεων. Για το συγκεκριμένο παράδειγμα, η δυναμική αδράνεια είναι ίση με $I_D = V^2 = (0{,}482)^2 \approx 0{,}232$ (βλέπε σχέση [5.31]). Παρατηρούμε ότι μόνο για τον πρώτο άξονα ικανοποιείται το κριτήριο αυτό. Η αδράνεια του δεύτερου άξονα είναι ελάχιστα μικρότερη από την οριακή τιμή του κριτηρίου. Στην Ενότητα 2.2.14.3 αναφέρθηκε ότι ένα εμπειρικό κριτήριο που χρησιμοποιείται συχνά στην Ανάλυση σε Κύριες Συνιστώσες, για την επιλογή των σημαντικών συνιστωσών, είναι αυτό της «*σπασμένης ράβδου* (*broken stick*)». Να τονίσουμε ότι η μέθοδος αυτή δεν έχει εφαρμοστεί στο πλαίσιο της ΠΑΑ. Το κριτήριο βασίζεται στη διαπίστωση ότι αν έχουμε μία ράβδο με μήκος 1 (ή 100%) και τη σπάσουμε τυχαία σε *p* κομμάτια, τότε το μεγαλύτερο σε μήκος από αυτά, έστω το κομμάτι *s*, θα έχει (μέσο) αναμενόμενο μήκος L_s ίσο με (Barton & David, 1956):

$$L_s = \frac{1}{p}\sum_{h=1}^{p}\left(\frac{1}{h}\right).$$

Το κριτήριο συνίσταται στην επιλογή των αξόνων για τους οποίους ισχύει (Καρλής, 2005):

$$\frac{\lambda_s}{\sum_{h=1}^{p}\lambda_h} > L_s,$$

όπου *p* είναι το πλήθος των μεταβλητών που συμμετέχουν στην ανάλυση (μέγιστος αριθμός συνιστωσών), λ_s είναι η ιδιοτιμή της συνιστώσας *s* και $\sum_{h=1}^{p}\lambda_h = p$ είναι η συνολική διασπορά.

Η φυσική ερμηνεία της παραπάνω σχέσης είναι ότι επιλέγονται ως σημαντικές οι συνιστώσες, για τις οποίες το αντίστοιχο ποσοστό ερμηνείας της ολικής διασποράς είναι μεγαλύτερο από την τιμή που αναμένεται κάτω από την υπόθεση της τυχαίας μεταβλητότητας των δεδομένων, χωρίς δηλαδή την ύπαρξη συστηματικών συσχετίσεων μεταξύ των μεταβλητών. Για περισσότερες πληροφορίες σχετικά με το κριτήριο παραπέμπουμε στους Jackson (1993), Ferré (1995) και Καρλή (2005).

Πίνακας 5.3: Πίνακας Συμπτώσεων των Μεταβλητών *X* και *Y* με Περιθώρια Γραμμή και Στήλη

Μεταβλητή *X*	Μεταβλητή *Y*						
	Σ1	Σ2	Σ3	Σ4	Σ5	Σ6	Σύνολο
Γ1	0	8	0	0	4	0	12
Γ2	0	4	0	0	2	0	6
Γ3	16	0	10	6	0	6	38
Γ4	14	0	10	4	0	6	34
Γ5	4	4	0	0	0	2	10
Γ6	8	0	4	0	0	12	24
Γ7	6	0	2	6	0	0	14
Σύνολο	48	16	26	16	6	26	138

Πίνακας 5.4: Αποτελέσματα της ΠΑΑ (Αδράνειες και Ποσοστά Ερμηνείας Αξόνων, Τιμές Ελέγχου για το Κριτήριο της «*Σπασμένης Ράβδου*»)

Άξονας	Αδράνεια	%	Αθροιστικό %	L_s *
1	0,878	0,757	0,757	0,457
2	0,204	0,176	0,933	0,257
3	0,049	0,042	0,975	0,157
4	0,029	0,025	1,000	0,090
5	0,000	0,000	1,000	0,040
Σύνολο	1,160	1,000		

* Τιμή σύγκρισης για τη σημαντικότητα των αξόνων με τη μέθοδο της «σπασμένης ράβδου».

Για να εφαρμόσουμε το κριτήριο της «σπασμένης ράβδου» στο πλαίσιο της ΠΑΑ θα πρέπει απλά να παρατηρήσουμε ότι στην περίπτωση ανάλυσης ενός $k \times l$ πίνακα συμπτώσεων **F** δύο μεταβλητών ισχύουν τα εξής: α) ο μέγιστος αριθμός παραγοντικών αξόνων είναι ίσος με $p=min\{k\text{-}1, l\text{-}1\}$, β) λ_s είναι η αδράνεια του

άξονα *s*, γ) το άθροισμα $\sum_{h=1}^{p} \lambda_h = I$ είναι ίσο με την ολική αδράνεια του **F** και δ) ο λόγος

$$\frac{\lambda_s}{\sum_{h=1}^{p} \lambda_h}$$

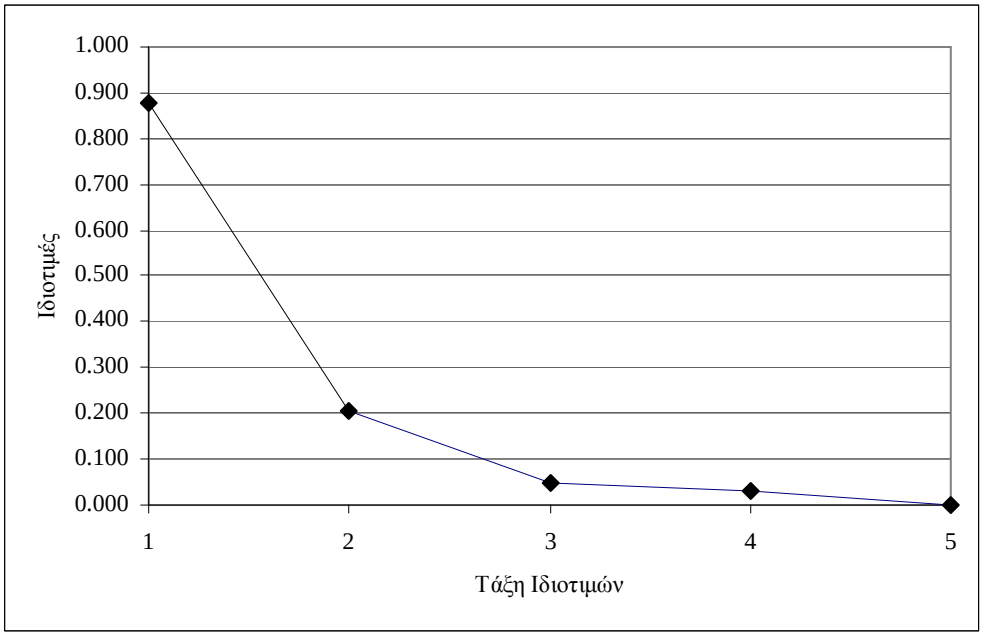
εκφράζει το ποσοστό της ολικής αδράνειας που ερμηνεύει ο άξονας *s*.

Στον Πίνακα 5.4 έχουμε υπολογίσει την ποσότητα L_s για κάθε άξονα. Παρατηρούμε ότι σύμφωνα με το κριτήριο της «σπασμένης ράβδου» μόνο ο πρώτος άξονας μπορεί να θεωρηθεί, και πάλι, σημαντικός, αφού το αντίστοιχο ποσοστό ερμηνείας της ολικής αδράνειας 0,757 είναι μεγαλύτερο από την τιμή L_1 =0,457.

Στην Ενότητα 2.2.14.3 είδαμε, επίσης, ότι ένα συνηθισμένο εμπειρικό κριτήριο, το οποίο εφαρμόζεται στο πλαίσιο της ΠΑΑ, για την επιλογή των σημαντικών αξόνων είναι αυτό βάσει του οποίου επιλέγονται οι άξονες που η αδράνειά τους είναι μεγαλύτερη από $1/p$. Για τα δεδομένα του παραδείγματος έχουμε ότι $1/p = 1/5 = 0,20$. Με μεγάλη αυστηρότητα θα επιλέγαμε τους δύο πρώτους άξονες, αλλά με κάποια αμφιβολία διότι η αδράνεια του δεύτερου άξονα είναι σχεδόν ασήμαντα μεγαλύτερη από 0,20. Η προηγούμενη διαπίστωση ενισχύεται και από τη μελέτη του Διαγράμματος των Ιδιοτιμών - *Scree Plot* (Διάγραμμα 5.1) από όπου φαίνεται το φαινόμενο να είναι μάλλον διδιάστατο.

Το ερώτημα που τίθεται είναι αν μόνο ο πρώτος άξονας αρκεί για την ανασύσταση, μέσω της σχέσης [5.35], του αρχικού πίνακα συμπτώσεων. Με άλλα λόγια, θέλουμε να ελέγξουμε αν ο πίνακας συμπτώσεων, ο οποίος θα προκύψει από την ανασύσταση, διαφέρει στατιστικά σημαντικά, έστω σε ε.σ. α=0,05, από τον αρχικό πίνακα **F**. Το γενικότερο ερώτημα, στο οποίο πρέπει να δοθεί απάντηση, είναι το πόσες είναι οι ελάχιστες διαστάσεις του βέλτιστου υποχώρου προβολής που απαιτούνται ώστε ο πίνακας συμπτώσεων, που θα προκύψει από την ανασύσταση, να μη διαφέρει στατιστικά σημαντικά, σε ε.σ. α, από τον **F**.

Για να απαντήσουμε στα παραπάνω ερωτήματα θα εφαρμόσουμε τη μεθοδολογία που παρουσιάσαμε στην Ενότητα 5.12. Ειδικότερα, θα πρέπει να ελέγξουμε διαδοχικά τις υποθέσεις $H_{s,}$ με s=0,…, 5, οι οποίες δηλώνουν ότι οι s πρώτοι άξονες είναι στατιστικά σημαντικοί για την ανασύσταση του αρχικού πίνακα συμπτώσεων. Για λόγους υπολογιστικής ευκολίας και την αποφυγή του προβλήματος των ενδεχόμενων αρνητικών εκτιμήσεων των συχνοτήτων θα χρησιμοποιήσουμε τη σχέση [5.38].



Διάγραμμα 5.1: Διάγραμμα των Ιδιοτιμών (*Scree Plot*)

Υπόθεση $H_{s=0}$: Αντιστοιχεί στην υπόθεση ανεξαρτησίας των δύο μεταβλητών η οποία έχει ήδη απορριφθεί από τον αρχικό έλεγχο χ^2. Πράγματι, από τη [5.38] για s=0 έχουμε:

$$Q_0^* = N\left(\lambda_1 + \lambda_2 + \ldots + \lambda_5\right) \Rightarrow Q_0^* = 138(0{,}878+0{,}204+\ldots+0{,}000)=138\times 1{,}160=160{,}061.$$

Δηλαδή $Q_0^*=Q$.

Η τιμή του στατιστικού Q_0^* για (k-s-1)(l-s-1)=(7-0-1)(6-0-1)=6×5=30 β.ε. είναι στατιστικά σημαντική σε ε.σ. α=0,05. Η παρατηρούμενη ισχύς του ελέγχου (*post-hoc* AI, βλέπε Ενότητα 5.13.1) υπολογίστηκε σε γ_{obs}=1. Επομένως, η απόφαση για την απόρριψη της υπόθεσης $H_{s=0}$ λαμβάνεται με βεβαιότητα. Η απόρριψη της $H_{s=0}$ είναι

ισοδύναμη με το αυτονόητο συμπέρασμα ότι και οι πέντε άξονες είναι στατιστικά σημαντικοί για την ανασύσταση του αρχικού πίνακα **F**.

Υπόθεση $H_{s=1}$: Αντιστοιχεί στην υπόθεση ότι ο πρώτος άξονας είναι στατιστικά σημαντικός και αρκεί για την ανασύσταση του αρχικού πίνακα **F**. Εργαζόμενοι όπως και στην περίπτωση της $H_{s=0}$, έχουμε:

$Q_1^* = N\left(\lambda_2 + \lambda_3 \ldots + \lambda_5\right) \Rightarrow Q_1^* = 138(0{,}204+0{,}049+\ldots+0{,}000)=138\times 0{,}282=38{,}853.$

Η τιμή του στατιστικού Q_1^* για (k-s-1)(l-s-1)=(7-1-1)(6-1-1)=5×4=20 β.ε. είναι στατιστικά σημαντική σε ε.σ. α=0,05 (p=0,007 και η κρίσιμη τιμή της Κατανομής $\chi^2_{0,05}(20) = 31{,}410<38{,}853$). Επίσης, γ_{obs}= 0,988 και η απόφαση για την απόρριψη της $H_{s=1}$ λαμβάνεται σχεδόν με βεβαιότητα. Επομένως, μόνο ο πρώτος άξονας δεν αρκεί για την ανασύσταση του αρχικού πίνακα συμπτώσεων (βλέπε επίσης Πίνακα 5.5 και Διάγραμμα 5.2). Το υπόλοιπο της πληροφορίας, με την έννοια της αδράνειας, που ερμηνεύουν οι επόμενοι τέσσερις άξονες είναι σημαντικό και δεν θα πρέπει να αγνοηθεί.

Υπόθεση $H_{s=2}$: Αντιστοιχεί στην υπόθεση ότι οι δύο πρώτοι άξονες είναι στατιστικά σημαντικοί και αρκούν για την ανασύσταση του αρχικού πίνακα **F**.

$Q_2^* = N\left(\lambda_3 + \lambda_4 + \lambda_5\right) \Rightarrow Q_2^* = 138(0{,}049+0{,}029+0{,}000)=138\times 0{,}078=10{,}743.$

Η τιμή του στατιστικού Q_2^* για (k-s-1)(l-s-1)=(7-2-1)(6-2-1)=4×3=12 β.ε. δεν είναι στατιστικά σημαντική σε ε.σ. α=0,05 (p=0,0,551 και η κρίσιμη τιμή της Κατανομής $\chi^2_{0,05}(12) = 21{,}026>10{,}743$). Συνεπώς, οι δύο πρώτοι άξονες αρκούν για την ανασύσταση του αρχικού πίνακα συμπτώσεων. Ο πίνακας που προκύπτει από την ανασύσταση δεν διαφέρει στατιστικά σημαντικά, σε ε.σ. α=0,05, από τον αρχικό πίνακα **F** (βλέπε επίσης Πίνακα 5.6 και Διάγραμμα 5.3). Πρακτικά, η διαδικασία σταματά εδώ. Αν συνεχίσουμε τους ελέγχους για τους επόμενους άξονες τα αντίστοιχα στατιστικά προφανώς και δεν θα είναι σημαντικά. Ενδεικτικά δίνουμε τις τιμές των στατιστικών Q_3^*=3,950 (p=0,683>0,05) και Q_4^*=0,007 (p=0,997>0,05). Έτσι, κάτω από μια διαφορετική θεώρηση, μπορούμε να ισχυριστούμε ότι ο

υποχώρος προβολής που ορίζουν οι δύο πρώτοι παραγοντικοί άξονες είναι ο ελάχιστος στατιστικά σημαντικός υποχώρος, στον οποίο μπορεί να προβληθεί το υπό εξέταση φαινόμενο χωρίς ουσιαστική απώλεια πληροφορίας.

Κατά τον έλεγχο της υπόθεσης $H_{s=2}$ η Μηδενική Υπόθεση του ελέγχου σημαντικότητας χ^2 μέσω του στατιστικού Q_2^* δηλώνει ότι οι δύο πρώτοι άξονες είναι σημαντικοί για την ανασύσταση του πίνακα **F**. Εμείς, με βάση τα αποτελέσματα του ελέγχου δεν την έχουμε απορρίψει και, κατά μία έννοια, την "αποδεχόμαστε, αφού, μάλιστα, δεν συνεχίζουμε τη διαδικασία για τους επόμενους άξονες με s>2. Επομένως, επιβάλλεται να εκτιμήσουμε την πιθανότητα να έχουμε διαπράξει Σφάλμα Τύπου ΙΙ στην απόφασή μας να "αποδεχθούμε" την $H_{s=2}$ ή ισοδύναμα να εκτιμήσουμε την παρατηρούμενη ισχύ γ_{obs} του ελέγχου χ^2. Αν η ισχύς γ_{obs} είναι σε αποδεκτά επίπεδα, τότε μπορούμε με μεγαλύτερη σιγουριά να ισχυριστούμε ότι τελικά οι δύο πρώτοι άξονες είναι σημαντικοί, σε ε.σ. α=0,05, για την ανασύσταση του υπό εξέταση φαινομένου. Θα εφαρμόσουμε τη μεθοδολογία που προτείναμε στην Ενότητα 5.13.1 για την *post hoc* ΑΙ του ελέγχου χ^2.

Από τη σχέση [5.19] μπορούμε να εκτιμήσουμε την παράμετρο μη κεντρικότητας λ, η οποία, στην περίπτωση που εξετάζουμε, είναι ίση με $Q_2^*=\lambda$=10,743. Κάνοντας χρήση της σχέσης [5.20] και με τη βοήθεια της συνάρτησης *Chi2Power* η παρατηρούμενη ισχύς του ελέγχου εκτιμάται σε γ_{obs}=0,543. Από τη σχέση [5.13] προκύπτει ότι β_{obs}=0,457. Επομένως, η πιθανότητα να έχουμε διαπράξει σφάλμα Τύπου ΙΙ είναι περίπου 45,7%, τιμή που εκφράζει ότι η απόφασή μας με βάση τα διαθέσιμα δεδομένα να "αποδεχθούμε" την υπόθεση ότι οι δύο πρώτοι άξονες είναι σημαντικοί σε ε.σ. α=0,05, είναι αρκετά επισφαλής για το δοσμένο μέγεθος δείγματος N=138. Αυτό που τελικά μπορούμε να ισχυριστούμε είναι ότι δεν έχουμε αρκετή δειγματική μαρτυρία ώστε να απορρίψουμε την υπόθεση $H_{s=2}$ και, συνεπώς, η $H_{s=2}$ "παραμένει". Αν το ε.σ. α είχε προκαθοριστεί στην τιμή 0,10, τότε η γ_{obs} θα ήταν ίση με 0,668 ή 66,8% και η πιθανότητα β_{obs} ίση με 0,332 ή 33,2%. Επομένως, η απόφασή μας για την διατήρηση των δύο αξόνων θα "ενέπνεε" μεγαλύτερη σιγουριά. Επίσης, αν το μέγεθος δείγματος ήταν διπλάσιο (δηλαδή N=276), τότε γ_{obs}=89,5% και

β_{obs}=10,5% (σε ε.σ. α=0,05). Στην περίπτωση αυτή θα ισχυριζόμασταν με μεγαλύτερη βεβαιότητα ότι η $H_{s=2}$ είναι αληθής.

Παρατήρηση 5.4

Οι Πίνακες 5.5 και 5.6 παρουσιάζουν την ανασύσταση του πίνακα **F** με βάση τον πρώτο και τους δύο πρώτους άξονες αντίστοιχα. Η ανασύσταση πραγματοποιήθηκε μέσω της σχέσης [5.37] και οι τιμές στα κελιά εμφανίζονται με στρογγυλοποίηση στον πλησιέστερο ακέραιο. Από την απλή οπτική εξέταση και σύγκριση των αρχικών τιμών των κελιών του πίνακα **F** με αυτές που προέκυψαν από την ανασύσταση δεν είναι εύκολο να βγάλουμε συμπεράσματα σχετικά με την ποιότητα της ανασύστασης στις δύο περιπτώσεις. Για το λόγο αυτό κατασκευάσαμε και τα διαγράμματα διασποράς (Διαγράμματα 5.2 και 5.3), όπου στον οριζόντιο άξονα *x* προβάλλονται οι τιμές των συχνοτήτων των κελιών του αρχικού πίνακα **F**, ενώ στον άξονα *y* οι αντίστοιχες συχνότητες, που προέκυψαν από την ανασύσταση. Σύμφωνα με τον Greenacre (1993α), ένα εμπειρικό κριτήριο ελέγχου καλής προσαρμογής στηρίζεται στο συλλογισμό ότι στην περίπτωση τέλειας ή πλήρους προσαρμογής τα προβαλλόμενα σημεία θα πρέπει να είναι διατεταγμένα κατά μήκος της ευθείας, η οποία διχοτομεί τη γωνία των αξόνων *x* και *y* στο επίπεδο προβολής (βλέπε διακεκομμένη γραμμή). Έτσι, όσο πλησιέστερα στη διχοτόμο βρίσκονται τα σημεία τόσο καλύτερη είναι η ποιότητα της ανασύστασης. Επίσης, η ύπαρξη πολλών απομακρυσμένων σημείων αποτελεί ένδειξη κακής προσαρμογής. Για περαιτέρω διευκόλυνση των συγκρίσεων σε κάθε διάγραμμα σχεδιάστηκε η ευθεία ελαχίστων τετραγώνων (βλέπε συνεχόμενη γραμμή) και υπολογίσαμε τους αντίστοιχους συντελεστές γραμμικής συσχέτισης *r* του *Pearson*, τους συντελεστές γραμμικού προσδιορισμού R^2 και τους συντελεστές συσχέτισης τ_b του *Kendal*.

Από την εξέταση και σύγκριση των δύο διαγραμμάτων είναι φανερό ότι η προσαρμογή βελτιώνεται σημαντικά αν πραγματοποιηθεί από τους δύο άξονες και όχι μόνο από τον πρώτο. Στο Διάγραμμα 5.2 παρατηρούμε μεγαλύτερες αποκλίσεις των σημείων τόσο από την ευθεία ελαχίστων τετραγώνων όσο και από τη διχοτόμο. Επίσης, τα σημεία που αντιστοιχούν στις συχνότητες των κελιών 74 (Γ7, Σ4) και 66 (Γ6, Σ6) δείχνουν αρκετά απομακρυσμένα και από τις δύο ευθείες.

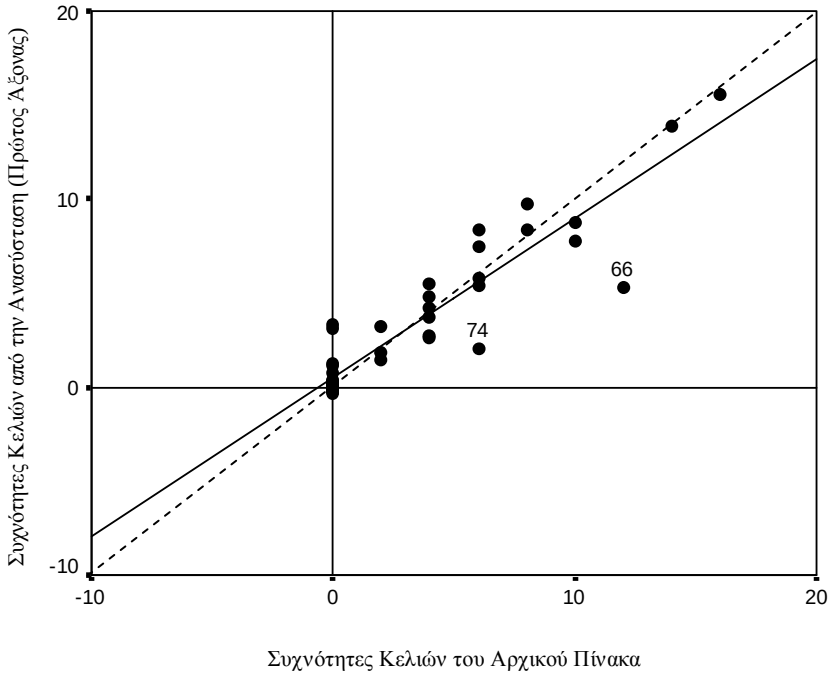
Πίνακας 5.5: Ανασύσταση του Πίνακα Συμπτώσεων των Μεταβλητών *X* και *Y* με βάση τον Πρώτο Άξονα

Μεταβλητή *X*	Μεταβλητή *Y*						
	Σ1	Σ2	Σ3	Σ4	Σ5	Σ6	Σύνολο
Γ1	0	8	0	0	4	0	12
Γ2	0	4	0	0	2	0	6
Γ3	16	0	9	5	0	8	38
Γ4	14	0	8	5	0	8	34
Γ5	3	3	1	1	1	1	10
Γ6	10	0	5	3	0	5	24
Γ7	6	0	3	2	0	3	14
Σύνολο	48	16	26	16	6	26	138

Πίνακας 5.6: Ανασύσταση του Πίνακα Συμπτώσεων των Μεταβλητών *X* και *Y* με βάση τους Δύο Πρώτους Άξονες

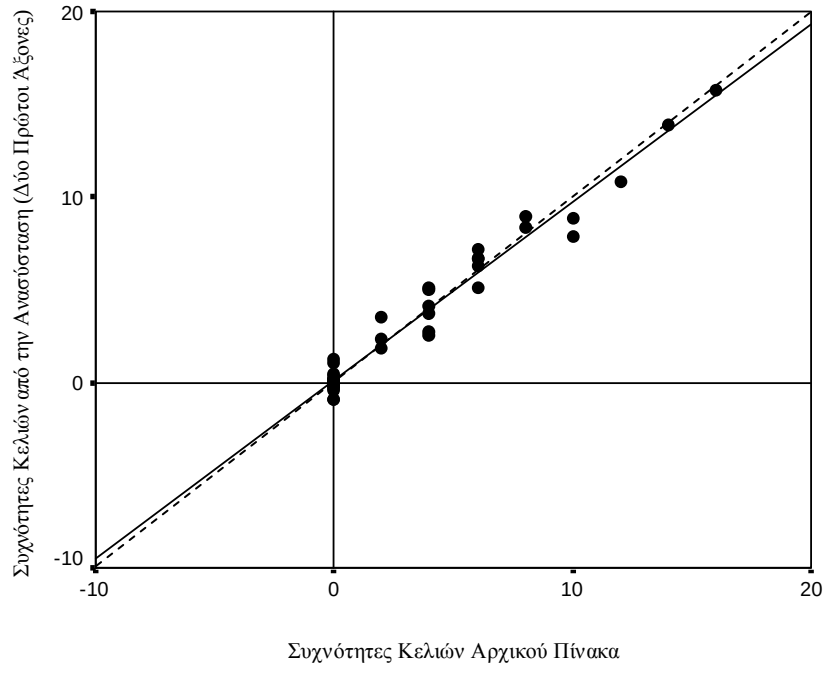
Μεταβλητή *X*	Μεταβλητή *Y*						
	Σ1	Σ2	Σ3	Σ4	Σ5	Σ6	Σύνολο
Γ1	0	8	0	0	4	0	12
Γ2	0	4	0	0	2	0	6
Γ3	16	0	9	7	0	7	38
Γ4	14	0	8	5	0	7	34
Γ5	2	3	1	0	1	2	10
Γ6	9	0	5	-1	0	11	24
Γ7	6	0	4	5	0	-1	14
Σύνολο	48	16	26	16	6	26	138

Η ποιότητα της ανασύστασης βελτιώνεται σημαντικά στην δεύτερη περίπτωση (Διάγραμμα 5.3). Η ευθεία παλινδρόμησης συγκλίνει με τη διχοτόμο, οι αποκλίσεις των σημείων είναι μικρότερες και για τις δύο ευθείες και οι συχνότητες των δύο κελιών, (Γ7, Σ4) και (Γ6, Σ6), προσαρμόζονται πλέον ικανοποιητικά. Βέβαια, και οι τιμές των δεικτών r, R^2 και τ_b δηλώνουν καλύτερη προσαρμογή, και μάλιστα με διαφορά, στην περίπτωση ανασύστασης με χρήση των δύο πρώτων αξόνων, αν και οι αντίστοιχοι δείκτες στην πρώτη περίπτωση έχουν και αυτοί υψηλές τιμές.



r=0,922, R^2=0,851 και τ_b =0,754

Διάγραμμα 5.2: Διάγραμμα Διασποράς των Αρχικών Συχνοτήτων και των Συχνοτήτων μετά την Ανασύσταση (Πρώτος Άξονας)



r=0,983, R^2=0,966 και τ_b =0,836

Διάγραμμα 5.3: Διάγραμμα Διασποράς των Αρχικών Συχνοτήτων και των Συχνοτήτων μετά την Ανασύσταση (Δύο Πρώτοι Άξονες)

Ένα βασικό συμπέρασμα που προκύπτει από την ενότητα αυτή είναι ότι τελικά απαιτείται μεγάλη προσοχή στην απόφαση σχετικά με την επιλογή των σημαντικών διαστάσεων του υποχώρου προβολής. Θα πρέπει να λαμβάνονται υπόψη αρκετά κριτήρια τα οποία είναι τις περισσότερες φορές αντιφατικά. Μάλλον, στην απόφαση θα πρέπει να επικουρήσουν οι ιδιαιτερότητες του ερευνητικού πεδίου καθώς και οι στόχοι της εκάστοτε μελέτης.

5.14 Σχόλια και Συμπεράσματα Κεφαλαίου

Στο Κεφάλαιο αυτό, εισαγάγαμε την έννοια της Δυναμικής Αδράνειας (*a priori* και *post hoc*) και προτείναμε μεθοδολογία με την οποία μπορεί να εκτιμηθεί τόσο η παρατηρούμενη ισχύς του στατιστικού ελέγχου χ^2 (*post hoc*) όσο και το ελάχιστο απαιτούμενο μέγεθος δείγματος (*a priori*) σε δειγματοληπτική ή πειραματική έρευνα, όπου στα δεδομένα θα εφαρμοστεί η ΠΑΑ. Ο προκαθορισμός του μεγέθους του δείγματος είναι, ίσως, το πιο κρίσημο στάδιο κατά το σχεδιασμό μιας έρευνας. Απαιτεί τη συνεργασία των ερευνητών (συνήθως μη-στατιστικών) με το στατιστικό αναλυτή για τον καθορισμό του μεγέθους του αποτελέσματος (*ES*), το οποίο έχει κλινική ή πρακτική σημαντικότητα και επομένως έχει νόημα να ανιχνευφθεί ως στατιστικά σημαντικό. Ένα αποτέλεσμα μπορεί να είναι στατιστικά σημαντικό, αλλά η κλινική του αξία να είναι ασήμαντη ή τελείως αναμενόμενη. Άλλωστε, όταν το μέγεθος του δείγματος είναι μεγάλο, αναμένουμε και στατιστικά σημαντικά ευρήματα. Να τονίσουμε ότι το πρόβλημα του καθορισμού του μεγέθους δείγματος τέθηκε για πρώτη φορά στο πλαίσιο της ΠΑΑ, αν και αποτελεί ερευνητικό αντικείμενο άλλων στατιστικών μεθόδων, όπως η Παραγοντική Ανάλυση και η Ανάλυση σε Κύριες Συνιστώσες. Για την ανάπτυξη της προτεινόμενης μεθοδολογίας θεωρήσαμε την ολική και τη δυναμική αδράνεια του πίνακα συμπτώσεων δύο μεταβλητών ως εναλλακτικούς δείκτες μεγέθους του αποτελέσματος και στηριχθήκαμε στο πλαίσιο Ανάλυσης Ισχύος του στατιστικού ελέγχου χ^2 που προτάθηκε από τον Cohen.

Αν και η ΠΑΑ θεωρείται εν γένει περιγραφική μέθοδος, ωστόσο κατά τη διερεύνηση της σχέσης μεταξύ δύο κατηγορικών μεταβλητών, όπου τα δεδομένα έχουν συγκεντρωθεί με τη μέθοδο της απλής τυχαίας δειγματοληψίας, είναι δυνατός ο

συνδυασμός της ΠΑΑ με τους στατιστικούς ελέγχους ανεξαρτησίας και καλής προσαρμογής χ^2. Ο έλεγχος ανεξαρτησίας μπορεί να εφαρμοστεί πριν από την ΠΑΑ για να διαπιστώσουμε αν η συνάφεια – συσχέτιση μεταξύ των μεταβλητών είναι στατιστικά σημαντική. Ο έλεγχος καλής προσαρμογής μπορεί να χρησιμοποιηθεί μετά την εφαρμογή της ΠΑΑ για τον καθορισμό του μικρότερου σε διαστάσεις στατιστικά σημαντικού υποχώρου προβολής, στον οποίο η παρατηρούμενη αλληλεπίδραση των μεταβλητών προβάλλεται χωρίς σημαντική απώλεια πληροφορίας. Οι δύο διαδικασίες είναι ανεξάρτητες. Ο έλεγχος καλής προσαρμογής μπορεί να εφαρμοστεί ασχέτως του αποτελέσματος που θα προκύψει από τον αρχικό έλεγχο ανεξαρτησίας. Κάτω από μια άλλη οπτική, μπορούμε να θεωρήσουμε ότι ο έλεγχος καλής προσαρμογής αντιστοιχεί στον έλεγχο σημαντικότητας του ποσοστού της ολικής αδράνειας που ερμηνεύει ο υποχώρος προβολής.

Στην περίπτωση που το διαθέσιμο δείγμα δεν μπορεί να θεωρηθεί τυχαίο ή αν δεν ικανοποιούνται οι προϋποθέσεις εφαρμογής του ελέγχου καλής προσαρμογής χ^2, τότε μπορούμε να χρησιμοποιήσουμε ένα νέο εμπειρικό κριτήριο για τον εντοπισμό των σημαντικών αξόνων. Το κριτήριο στηρίζεται στη φυσική ερμηνεία της *a priori* δυναμικής αδράνειας του πίνακα συμπτώσεων, η οποία εκφράζει τη μέση αδράνεια των αξόνων. Σύμφωνα με το προτεινόμενο κριτήριο, σημαντικοί θεωρούνται οι παραγοντικοί άξονες που η αντίστοιχη αδράνειά τους είναι μεγαλύτερη από τη δυναμική αδράνεια. Είναι σημαντικό, η απόφασή μας σχετικά με την επιλογή των αξόνων που χρήζουν ερμηνείας να λαμβάνεται μετά από εξέταση και άλλων κριτηρίων, όπως για παράδειγμα αυτό της «σπασμένης ράβδου», το οποίο προσαρμόσαμε στο μεθοδολογικό πλαίσιο της ΠΑΑ. Γενικά, απαιτείται μεγάλη προσοχή στην επιλογή των σημαντικών διαστάσεων του υποχώρου προβολής. Λόγω του γεγονότος ότι τα διαθέσιμα κριτήρια είναι τις περισσότερες φορές αντιφατικά θα ήταν ενδεχομένως χρήσιμο να λαμβάνονται υπόψη και οι ιδιαιτερότητες του ερευνητικού πεδίου καθώς και οι στόχοι της εκάστοτε μελέτης.

Τα στοχαστικά και λογικά σφάλματα, τα οποία είναι δυνατό να διαπραχθούν στους ελέγχους σημαντικότητας, σε συνδυασμό με τα επιχειρήματα αυτών που ασκούν κριτική στη διαδικασία ελέγχου της Μηδενικής Υπόθεσης θέτουν τις βάσεις για έναν περαιτέρω προβληματισμό σχετικά με την εγκυρότητα της γνώσης, η οποία

παράγεται από τους στατιστικούς ελέγχους υποθέσεων. Για παράδειγμα, ένα αποτέλεσμα μη στατιστικά σημαντικό είναι πιο ενδιαφέρον από ένα αντίστοιχο στατιστικά σημαντικό, ιδιαίτερα στην περίπτωση που το μέγεθος του δείγματος είναι αρκετά μεγάλο (άλλωστε τα μη στατιστικά σημαντικά αποτελέσματα σπάνια δημοσιεύονται). Θα συμφωνήσουμε με τον Gras (1995) ότι

> «*Χρειάζεται, κατά τη διερεύνηση της εγκυρότητας υποθέσεων, να βρεθεί μια σωστή ισορροπία ανάμεσα στην απλοϊκή χρήση των στατιστικών μεθόδων, την άρνηση επένδυσης σε αυτό το πεδίο και την στατιστικομανία που οδηγεί σε μια πληθώρα ανεξερεύνητων αποτελεσμάτων, που συνοδεύονται από μια ψευδαίσθηση διαφάνειας*». (σ. 98).

Κλείνοντας, θα πρέπει να επισημάνουμε τον εξής κίνδυνο: ένας έμπειρος στατιστικός αναλυτής, ο οποίος κάνει αντιδεοντολογική χρήση της Στατιστικής, μπορεί να σχεδιάσει μια έρευνα με τρόπο ώστε να εξισορροπήσει κατάλληλα τα Σφάλματα Τύπου Ι και Τύπου ΙΙ, με αποτέλεσμα, μετά από κατάλληλη επιλογή των α, β, του μεγέθους του αποτελέσματος και του μεγέθους του δείγματος, να προσανατολίσει τα συμπεράσματα προς ορισμένες “επιθυμητές” κατευθύνσεις (βλέπε Παρατήρηση Ε2.1 της Ενότητας Ε2 στο Παράρτημα Ε).

ΚΕΦΑΛΑΙΟ 6

Εξωτερική Εγκυρότητα των Αποτελεσμάτων της Παραγοντικής Ανάλυσης των Αντιστοιχιών: Έλεγχοι Στατιστικής Σημαντικότητας και Περιοχές Εμπιστοσύνης

6.1 Εισαγωγή

Όπως διαπιστώθηκε στο Κεφάλαιο 2, η ΠΑΑ αποτελεί ένα γενικό σύστημα ανάλυσης κατηγορικών δεδομένων και χρησιμοποιείται, πλέον, σχεδόν σε όλα τα ερευνητικά επιστημονικά πεδία. Παρέχει στις εκροές της χρήσιμες οπτικοποιήσεις των προτύπων και των σχέσεων μεταξύ των κατηγορικών μεταβλητών, οι οποίες συμμετέχουν στην ανάλυση. Ωστόσο, τα παραγόμενα αποτελέσματα (αριθμητικά και διαγραμματικά) συχνά είναι δύσκολο να αξιολογηθούν (Lebart, 2005). Στην περίπτωση αυτή, "υποβόσκει" το ερώτημα αν οι παρατηρούμενες δομές είναι πραγματικές ή αποτέλεσμα τυχαίων επιδράσεων. Στις μέρες μας, στην εποχή των Ηλεκτρονικών Υπολογιστών, δεν αρκεί να βασιζόμαστε μόνο σε εμπειρικά κριτήρια ερμηνείας των αποτελεσμάτων, τα οποία είχαν ευρεία εφαρμογή τις πρώτες δεκαετίες της Ανάλυσης Δεδομένων. Στο φιλοσοφικό πλαίσιο της Γαλλικής Σχολής και κάτω από την καταλυτική επίδραση του Benzécri (βλέπε Ενότητα 1.5.2), η ΠΑΑ αναδείχθηκε ως μέθοδος ανεξάρτητη από μοντέλα – υποδείγματα και η εφαρμογή της δεν συνδέεται με *a priori* υποθέσεις και προϋποθέσεις (βλέπε Ενότητες 1.3, 1.4 και 1.4.1). Άλλωστε, για τον Benzécri η Ανάλυση Δεδομένων αποτελεί μια διαδικασία "μάθησης" και απόκτησης γνώσης μέσα από ποιοτικές και ποσοτικές καταγραφές πραγματικών εμπειριών, κατά τις οποίες η φροντίδα και η ευθύνη της ερμηνείας των αποτελεσμάτων και των συνεπειών τους αφήνεται κυρίως στους ερευνητές – χρήστες

των μεθόδων και όχι σε πιθανολογικούς μηχανισμούς (βλέπε Ενότητα 1.5.2). Όμως, όπως τονίστηκε στην Ενότητα 1.5.5, η απόλυτη αποδοχή της παραπάνω θέσης θα μπορούσε να οδηγήσει σε μια υπέρμετρη απλοποίηση της στατιστικής σκέψης, όπου το ενδιαφέρον θα επικεντρώνεται μόνο στην εξεύρεση σωστών, κατά περίπτωση, λύσεων θέτοντας στο περιθώριο τη μαθηματική αυστηρότητα και τη δυνατότητα γενίκευσης των αποτελεσμάτων. Η γενίκευση μπορεί να επιτευχθεί με την εφαρμογή μεθόδων της Επαγωγικής Στατιστικής. Βέβαια, οι μέθοδοι αυτές απαιτούν την ικανοποίηση αρκετών πιθανοθεωρητικών και τεχνικών προϋποθέσεων (Gifi, 1996), που δεν συνδέονται με τις γενικότερες επιστημονικές υποθέσεις, οι οποίες προηγούνται και αποτελούν το κίνητρο για την πραγματοποίηση μιας έρευνας, αλλά είναι απαιτήσεις μόνο των στατιστικών τεχνικών (Μπεχράκης, 1999). Μια εναλλακτική προσέγγιση αποτελεί η χρήση τεχνικών επαναδειγματοληψίας, όπως η *Bootstrap*, όπου οι προϋποθέσεις είναι λιγότερο αυστηρές (Greenacre 1993α και 1984, Gifi 1996, Michailidis 1996, Michailidis & De Leeuw 1998, Lebart 2005), αλλά δημιουργούν αφορμές για περαιτέρω προβληματισμούς (βλέπε Ενότητα 6.3.1).

Στην Ενότητα 1.5.5 αναφέρθηκε ότι η σημαντικότερη διαφορά των μεθόδων της Ανάλυσης Δεδομένων σε σχέση με αυτές της Επαγωγικής Στατιστικής είναι ότι για την εφαρμογή τους δεν απαιτείται η προσαρμογή των δεδομένων σε κάποιο στοχαστικό μοντέλο και η αντίστοιχη συμπερασματολογία δεν υπάγεται σε κάποιο μηχανισμό επαγωγικού συλλογισμού, με την έννοια της στατιστικής σημαντικότητας. Η αλήθεια είναι ότι οι μέθοδοι και ιδιαίτερα η ΠΑΑ, συνήθως, δεν συνοδεύονται από στατιστικούς ελέγχους. Αυτό όμως δεν σημαίνει ότι δεν είναι δυνατό να υπάρξουν τέτοιοι και μάλιστα με ελάχιστες τεχνικές προϋποθέσεις. Επίσης, δεν δίνεται ιδιαίτερη έμφαση στο μηχανισμό συλλογής και συγκρότησης των διαθέσιμων δεδομένων (μέθοδος δειγματοληψίας, πειραματικός σχεδιασμός, ανεξαρτησία των παρατηρήσεων) αρκεί τα δεδομένα να μπορούν να πινακοποιηθούν σε μορφή κατάλληλη για την εφαρμογή της μεθόδου. Τα δεδομένα, έστω και αν προέρχονται από δείγμα, αντιμετωπίζονται σαν να αποτελούν ολόκληρο τον υπό εξέταση πληθυσμό (Greenacre, 1984), δίνοντας ένα περιγραφικό και διερευνητικό χαρακτήρα στην ΠΑΑ αποδυναμώνοντας οποιεσδήποτε γενικεύσεις, όπως αυτές νοούνται στην Επαγωγική Στατιστική.

Η δυνατότητα γενίκευσης των δομών και σχέσεων, οι οποίες οπτικοποιούνται και αναδεικνύονται μέσω της ΠΑΑ, συνδέεται κυρίως με τον έλεγχο της «εξωτερικής» εγκυρότητας των αποτελεσμάτων που παράγονται από την εφαρμογή της μεθόδου. Στο χώρο της Πολυδιάστατης Στατιστικής Ανάλυσης, ο όρος εξωτερική εγκυρότητα αναφέρεται στην περίπτωση που τα δεδομένα έχουν συγκεντρωθεί από τον υπό εξέταση πληθυσμό με μεθόδους της τυχαίας δειγματοληψίας και εστιάζεται στον έλεγχο τού κατά πόσο συνεπή (σταθερά) είναι τα παραγόμενα αποτελέσματα, στη θεωρητική περίπτωση που η ανάλυση επαναληφθεί σε άλλα τυχαία δείγματα από τον ίδιο πληθυσμό (Greenacre 1993α και 1984, Markus 1994α, Chateau & Lebart 1996, Gifi 1996, Lebart 2005). Στο συγκεκριμένο χώρο, το ζήτημα αντιμετωπίζεται σχεδόν αποκλειστικά με την εφαρμογή μεθόδων επαναδειγματοληψίας. Η τυχαία δειγματοληψία εξασφαλίζει, ως ένα βαθμό, την απαίτηση της αντιπροσωπευτικότητας, ώστε το δείγμα να θεωρείται μικρογραφία του υπό εξέταση πληθυσμού. Σε ένα πιο αυστηρό πλαίσιο, η εξωτερική εγκυρότητα συνδέεται με τη στατιστική σημαντικότητα των αποτελεσμάτων (Greenacre 1984, Gifi 1996, Michailidis & De Leeuw 1998, Lebart 2005) και την απόδοση ενός βαθμού εμπιστοσύνης στα συμπεράσματα που θα προκύψουν.

Στο κεφάλαιο αυτό, αρχικά, επιχειρούμε μια σύντομη ανασκόπηση των στατιστικών ελέγχων σημαντικότητας, που είναι δυνατό να εφαρμοστούν ή/και να συνδυαστούν με την ΠΑΑ, κατά την επεξεργασία δεδομένων, τα οποία έχουν συγκεντρωθεί με μεθόδους της τυχαίας δειγματοληψίας (Ενότητα 6.2). Οι έλεγχοι αφορούν στην εξωτερική εγκυρότητα των βασικών αριθμητικών αποτελεσμάτων, τα οποία παράγονται στις εκροές της μεθόδου, τόσο στη διμεταβλητή όσο και στην πολυμεταβλητή εκδοχή της. Στην Ενότητα 6.3, προτείνουμε, για τη διμεταβλητή περίπτωση, δύο μεθόδους κατασκευής 100(1-α)% ελλείψεων εμπιστοσύνης επί των παραγοντικών επιπέδων με κέντρα τις προβολές των γραμμών (στηλών) του αντίστοιχου πίνακα συμπτώσεων. Στην πρώτη προσέγγιση, οι συντεταγμένες των σημείων θεωρούνται ως τυχαίες μεταβλητές, ενώ στη δεύτερη ως μέσοι όροι βέλτιστα ποσοτικοποιημένων (μετασχηματισμένων) βαθμών των πειραματικών ή δειγματοληπτικών μονάδων (βλέπε Ενότητα 2.5), οι οποίες συμμετέχουν στην ανάλυση. Η βασική ιδέα και στις δύο προσεγγίσεις είναι η εφαρμογή μιας «τοπικής» Ανάλυσης σε Κύριες Συνιστώσες γύρω από κάθε σημείο γραμμής (στήλης). Ο προσανατολισμός των αξόνων των ελλείψεων και τα μήκη τους καθορίζονται με

μέθοδο, η οποία συνδυάζει, ανάλογα με την προσέγγιση, ιδιότητες της Διδιάστατης Κανονικής Κατανομής και το επίπεδο σημαντικότητας *α*. Τέλος, στην Ενότητα 6.4 προτείνουμε μέθοδο κατασκευής ενός "μη παραμετρικού" διαστήματος εμπιστοσύνης για τον έλεγχο σημαντικότητας των παραγοντικών αξόνων που αναδεικνύονται στην πολυμεταβλητή εκδοχή της ΠΑΑ.

6.2 Έλεγχοι Στατιστικής Σημαντικότητας στην ΠΑΑ

Στην ενότητα αυτή παρουσιάζουμε συνοπτικά τους στατιστικούς ελέγχους σημαντικότητας που έχουν προταθεί κατά καιρούς και είναι δυνατό να εφαρμοστούν ή/και να συνδυαστούν με την ΠΑΑ, στη διμεταβλητή και στην πολυμεταβλητή εκδοχή της.

6.2.1 Η Περίπτωση Δύο Μεταβλητών

6.2.1.1 Στατιστική Σημαντικότητα της Ολικής Αδράνειας του Πίνακα Συμπτώσεων

Είδαμε στην Ενότητα 5.2 ότι η στατιστική σημαντικότητα της ολικής αδράνειας I_F του πίνακα συμπτώσεων απολύτων συχνοτήτων **F**, δύο κατηγορικών τυχαίων μεταβλητών *X* και *Y* με *k* και *l* κλάσεις αντίστοιχα, μπορεί να ελεγχθεί μέσω της Κατανομής χ^2 συγκρίνοντας την ποσότητα $Q = NI_F$ με την κρίσιμη τιμή της Κατανομής χ^2 με $(k-1)(l-1)$ βαθμούς ελευθερίας (β.ε.), σε επίπεδο σημαντικότητας (ε.σ.) *α*. Η ποσότητα *Q* αντιστοιχεί στο στατιστικό χ^2 που υπολογίζεται κάτω από την υπόθεση ανεξαρτησίας των δύο μεταβλητών και *N* είναι το μέγεθος του δείγματος.

6.2.1.2 Στατιστική Σημαντικότητα των Τυποποιημένων Υπολοίπων του Πίνακα Συμπτώσεων

Στην Ενότητα 2.2.5 παρατηρήσαμε ότι τα τυποποιημένα υπόλοιπα που αντιστοιχούν στα κελιά του πίνακα **F** κάτω από την ισχύ της υπόθεσης της ανεξαρτησίας των δύο μεταβλητών έχουν μέση τιμή μηδέν, διακύμανση μικρότερη ή ίση από τη μονάδα και ακολουθούν ασυμπτωτικά την Τυποποιημένη Κανονική Κατανομή (Agresti, 2002).

Κελιά με τυποποιημένα υπόλοιπα σε απόλυτη τιμή μεγαλύτερη του 1,96≈2 συνεισφέρουν στατιστικά σημαντικά, σε ε.σ. α=0,05, στη σημαντικότητα του στατιστικού Q και σε αυτά τα κελιά οφείλεται, κυρίως, η συνάφεια ή η αλληλεπίδραση των δύο μεταβλητών. Η τιμή 1,96 αντιστοιχεί στην κρίσιμη τιμή της Τυποποιημένης Κανονικής Κατανομής για $\alpha/2$=0,025. Αν θέσουμε το α=0,01, τότε η τιμή σύγκρισης των τυποποιημένων υπολοίπων είναι 2,58, ενώ αυτή μειώνεται σε 1,64 για α=0,10. Τα τυποποιημένα υπόλοιπα διαιρεμένα με μια εκτίμηση της τυπικής τους απόκλισης ονομάζονται «Διορθωμένα Τυποποιημένα Υπόλοιπα» (Haberman 1973, Everitt 1979, Agresti 2002). Έχουν μέση τιμή μηδέν και διακύμανση ίση με τη μονάδα και η ασυμπτωτική τους συμπεριφορά προσεγγίζει καλύτερα την Τυποποιημένη Κανονική Κατανομή. Υπολογιστικά τα διορθωμένα τυποποιημένα υπόλοιπα δίνονται από τη σχέση:

$$\frac{f_{ij}-\frac{f_{i+}f_{+j}}{N}}{\sqrt{\frac{f_{i+}f_{+j}}{N}\left(1-\frac{f_{i+}}{N}\right)\left(1-\frac{f_{+j}}{N}\right)}},$$

όπου f_{i+} και f_{+j} είναι οι περιθώριες συχνότητες της γραμμής i και της στήλης j αντίστοιχα του πίνακα συμπτώσεων **F** με γενικό στοιχείο f_{ij} $\left(i=1,\ldots k,\ j=1,\ldots,l\right)$.

Ο έλεγχος της στατιστικής σημαντικότητας των τυποποιημένων υπολοίπων εγείρει το πρόβλημα των πολλαπλών ελέγχων και την εξάπλωση του Αθροιστικού Σφάλματος Τύπου Ι (βλέπε Ενότητα Ε3.2 του Παραρτήματος Ε). Κατά συνέπεια, οι $k\times l$ σε πλήθος έλεγχοι θα πρέπει να συνοδεύονται και από κατάλληλη διόρθωση του ε.σ. α.

6.2.1.3 Στατιστική Σημαντικότητα των Παραγοντικών Αξόνων

Η στατιστική σημαντικότητα των παραγοντικών αξόνων ανάγεται στον έλεγχο σημαντικότητας των αντίστοιχων ιδιοτιμών (αδρανειών) ή/και χαρακτηριστικών τιμών.

Έστω λ_s η ιδιοτιμή (αδράνεια) που αντιστοιχεί στον παραγοντικό άξονα s (s=1,…, p), με $p=min\{k\text{-}1, l\text{-}1\}$. Από την Ενότητα 2.2.14 (Δεύτερο Βήμα) έχουμε ότι:

$$I_F = \sum_s \lambda_s \text{ , με } \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \text{ .}$$

Λόγω της σχέσης $I_F = \frac{Q}{N}$, ισχύει:

$$Q = N\sum_s \lambda_s = N\lambda_1 + N\lambda_2 + \cdots + N\lambda_p \text{ .}$$

Ο Lebart (1976) έδειξε ότι κάτω από την υπόθεση της ανεξαρτησίας των δύο μεταβλητών η ποσότητα $N\lambda_s$ έχει ασυμπτωτικά την ίδια κατανομή με αυτήν της *s*-ιοστής ιδιοτιμής μιας πολυδιάστατης τυχαίας μεταβλητής που ακολουθεί την Κεντρική Κατανομή *Wishart* με παραμέτρους (*l*-1) και (*k*-1) (βλέπε και Lebart, Morineau & Tabard 1977, Haberman 1981, Hirotsu 1983, Lebart, Morineau & Warwick 1984, Lebart, Morineau & Piron 2000, Lebart 2005). Ειδικότερα, αν $l \leq k$, τότε η ποσότητα $N\lambda_s$ θα πρέπει να συγκριθεί με την κρίσιμη τιμή της $W_{(l-1)}$ με (*k*-1) β.ε., σε ε.σ. *α*. Η προσέγγιση αυτή στηρίζεται: α) στο γεγονός ότι η Πολυωνυμική Κατανομή που ορίζεται από τα *k*×*l* κελιά του πίνακα συμπτώσεων μπορεί να προσεγγιστεί ασυμπτωτικά από την πολυδιάστατη Κανονική Κατανομή και β) στο ότι μια κανονικοποιημένη μορφή του πίνακα $\mathbf{S^T S}$, ο οποίος δίνεται ως είσοδος στην SVD κατά την εφαρμογή της ΠΑΑ, είναι πίνακας *Wishart*. Αν και η πρόταση του Lebart αποτελεί την καλύτερη προσέγγιση (Greenacre, 1984), στην πράξη αποδεικνύεται δύσχρηστη (Lebart, 2005). Θα πρέπει να καταφύγει κανείς σε ειδικούς πίνακες ή διαγράμματα για τον αντίστοιχο στατιστικό έλεγχο (βλέπε Hanumara & Thompson 1968, Pearson & Hartley 1972, Clemm, Krishnaiah & Waikar 1973, Lebart 1976, Lebart, Morineau & Tabard 1977, Lebart, Morineau & Warwick 1984, Lebart, Morineau & Piron 2000) με περιορισμένο, όμως, εύρος τιμών για τις παραμέτρους (*l*-1) και (*k*-1). Για περισσότερες πληροφορίες σχετικά με την Κατανομή *Wishart* παραπέμπουμε στους Aitken (1949), Mathew και Nordström (1997), Rao (2002), Mardia, Kent και Bibby (2003), Καρλή (2005) και Nadarajah και Kotz (2005).

Ο Williams (1952), στο μεθοδολογικό πλαίσιο της Κανονικοποιημένης Συσχέτισης για πίνακες συνάφειας δύο κατηγορικών μεταβλητών, έδειξε ότι η στατιστική σημαντικότητα της ποσότητας $N\lambda_s$ μπορεί να ελεγχθεί μέσω της χ^2 Κατανομής με (*k*-*s*)(*l*-*s*)-(*k*-*s*-1)(*l*-*s*-1) βαθμούς ελευθερίας σε ε.σ. *α*. Την προσέγγιση αυτή

ακολουθεί και ο Van de Geer (1993β) για τη στατιστική σημαντικότητα των παραγοντικών αξόνων της ΠΑΑ. Αντίθετα, ο Greenacre (1993α) υποστηρίζει ότι όλες οι ποσότητες $N\lambda_s$ θα πρέπει να συγκρίνονται με την κρίσιμη τιμή της Κατανομής χ^2 με $(k-1)(l-1)$ β.ε., σε ε.σ. *α*. Όσες από αυτές υπερβαίνουν την κρίσιμη τιμή, τότε οι αντίστοιχοι άξονες μπορούν να θεωρηθούν ως στατιστικά σημαντικοί σε ε.σ. *α*. Στο πλαίσιο της Δυϊκής Κλιμάκωσης (*Dual Scaling*) (Nishisato, 1980), η οποία παράγει συγκρίσιμα αποτελέσματα με αυτά της ΠΑΑ, συναντάμε την εκδοχή ότι το στατιστικό:

$$-\left[N-1-\frac{1}{2}(k+l-1)\right]\ln(1-\lambda_s)\,, \qquad [6.1]$$

ακολουθεί ασυμπτωτικά τη χ^2 Κατανομή με *k*+*l*-1-2*s* βαθμούς ελευθερίας (Nishisato, 1980). Οι άξονες για τους οποίους η αντίστοιχη τιμή του στατιστικού [6.1] είναι μεγαλύτερη από την κρίσιμη τιμή της Κατανομής χ^2 μπορούν να δηλωθούν ως στατιστικά σημαντικοί. Έτσι, μέσω του Πίνακα 6.1 μπορεί να ελεγχθεί η σημαντικότητα των *p* αξόνων. Για την προσέγγιση του Nishisato ασκήθηκε κριτική από τον Greenacre (1984) με βασικό επιχείρημα το ότι η προτεινόμενη μέθοδος στηρίζεται στον έλεγχο χ^2 του *Bartlett* για τη σημαντικότητα των συντελεστών Κανονικοποιημένης Συσχέτισης μεταξύ δύο συνόλων μεταβλητών, η οποία προϋποθέτει ότι οι μεταβλητές τουλάχιστον του ενός συνόλου θα πρέπει να ακολουθούν την Πολυδιάστατη Κανονική Κατανομή. Σύμφωνα με τον Greenacre (1984), η προϋπόθεση αυτή είναι δύσκολο, εν γένει, να γίνει δεκτή στην περίπτωση της ΠΑΑ. Επιχειρήματα, με βάση τα οποία οι επιφυλάξεις του Greenacre μπορούν να αντικρουστούν, συναντάμε στους Pearson (1906), Fisher (1941), Williams (1952), Nishisato (1980) και De Leeuw (1993).

Σύμφωνα με τον Lancaster (1963α και 1963β) οι προσεγγίσεις των ποσοτήτων $N\lambda_s$ μέσω της Κατανομής χ^2 δεν είναι εν γένει ικανοποιητικές (βλέπε και Lebart, Morineau & Warwick 1984, Greenacre 1984) και μπορούν να οδηγήσουν σε "αισιόδοξα" στατιστικά σημαντικά αποτελέσματα για τους άξονες με υψηλές αδράνειες (ιδιοτιμές) και σε "απαισιόδοξα", δηλαδή μη στατιστικά σημαντικά, για τους άξονες μικρής ερμηνευτικής ικανότητας (χαμηλές αδράνειες). Και αυτό διότι δεν

συμβαίνει το ίδιο με τις ποσότητες $N\lambda_s$ (Lancaster, 1963α και 1969), παρόλο που το στατιστικό Q ακολουθεί ασυμπτωτικά την Κατανομή χ^2. Μάλιστα, η μέση (αναμενόμενη) τιμή της πρώτης ιδιοτιμής είναι πάντα μεγαλύτερη από την κρίσιμη τιμή της αντίστοιχης Κατανομής χ^2 (Lancaster, 1957). Επομένως, στην πράξη η κατάσταση μπορεί να βελτιωθεί καθιστώντας τους ελέγχους, με βάση το χ^2, περισσότερο αυστηρούς διορθώνοντας το ε.σ. α προς τα κάτω, τουλάχιστον για τους δύο πρώτους άξονες. Για παράδειγμα, να προκαθοριστεί το α σε 0,01 αντί σε 0,05.

Πίνακας 6.1: χ^2 Ανάλυση των p Ιδιοτιμών του Πίνακα **F**

Άξονας	χ^2	β.ε.
1	$-c\ln(1-\lambda_1)$	k+l-3
2	$-c\ln(1-\lambda_2)$	k+l-5
3	$-c\ln(1-\lambda_3)$	k+l-7
⋮	⋮	⋮
p	$-c\ln(1-\lambda_p)$	k+l-(2p+1)
Σύνολο	$-c\sum_{s=1}^{p}\ln(1-\lambda_s)$	(k-1)(l-1)

Με $c=\left[N-1-\frac{1}{2}(k+l-1)\right]$.

Ο O'Neill (1978α, 1978β, 1980 και 1981) σε μια σειρά άρθρων μελέτησε την ασυμπωτική κατανομή των συντελεστών Κανονικοποιημένης Συσχέτισης για πίνακες συνάφειας δύο κατηγορικών μεταβλητών (βλέπε Williams 1952, Srinkantan 1970, Haberman 1981, Greenacre 1984, Lebart, Morineau & Warwick 1984, Tenenhaus & Young 1985, Gower 1990, Andersen 1991, Van de Geer 1993β, De Leeuw, Wang & Michailidis 1999). Οι συντελεστές αυτοί είναι ίσοι με τις χαρακτηριστικές τιμές που αντιστοιχούν στους παραγοντικούς άξονες οι οποίοι προκύπτουν από την ΠΑΑ. Το βασικό συμπέρασμα είναι ότι χαρακτηριστικές τιμές ακολουθούν ασυμπτωτικά Κανονική Κατανομή. Μάλιστα, ο O'Neill κάνοντας χρήση της μεθόδου *Δέλτα* (βλέπε Ενότητα ΣΤ1 του Παραρτήματος ΣΤ) δίνει και τις μαθηματικές σχέσεις για την εκτίμηση των αντίστοιχων διακυμάνσεων των κατανομών. Οι προσεγγίσεις που προτείνει είναι αρκετά πολύπλοκες και στηρίζονται σε *a priori* παραδοχές σχετικά με τις πραγματικές (θεωρητικές) τιμές των χαρακτηριστικών τιμών και των

συντεταγμένων των προβολών των σημείων γραμμών και στηλών του πίνακα συμπτώσεων. Προς την ίδια κατεύθυνση κινείται και η εργασία των Eaton & Tyler (1994). Για μια σύνοψη των αποτελεσμάτων του O' Neill παραπέμπουμε στον Greenacre (1984).

Στην Ενότητα 2.2.14.2 αναφέρθηκε ότι μέσω της μεθόδου *Δέλτα* μπορούν να υπολογιστούν για κάθε άξονα εκτιμητές των διασπορών των χαρακτηριστικών τιμών και των συντεταγμένων των προβολών των σημείων γραμμών και στηλών του πίνακα συμπτώσεων. Με βάση τις εκτιμήσεις αυτές είναι δυνατό να κατασκευαστούν ξεχωριστά 100(1-α)% διαστήματα εμπιστοσύνης για τις χαρακτηριστικές τιμές και τις παραγοντικές συντεταγμένες των σημείων γραμμών ή/και στηλών. Τα διαστήματα αυτά είναι της μορφής (Gifi, 1996):

$$\left(x - z_{a/2}s_x,\, x + z_{a/2}s_x\right),$$

όπου $z_{\alpha/2}$ είναι η κρίσιμη τιμή της Τυποποιημένης Κανονικής Κατανομής σε επίπεδο σημαντικότητας $\alpha/2$ και s_x η εκτίμηση της τυπικής απόκλισης της αντίστοιχης παραμέτρου x (χαρακτηριστικής τιμής ή παραγοντικής συντεταγμένης). Διαστήματα εμπιστοσύνης τα οποία δεν περιέχουν το 0 δηλώνουν στατιστική σημαντικότητα, σε επίπεδο σημαντικότητας α, του αντίστοιχου μεγέθους x που εξετάζεται.

Συνοψίζοντας, οι προσεγγίσεις που βασίζονται στη Κατανομή χ^2 θα πρέπει να αντιμετωπίζονται με επιφύλαξη, εκτός και αν η παρατηρούμενη στάθμη σημαντικότητας (p-value) των αντίστοιχων ελέγχων είναι αρκετά μικρή (π.χ. p<0,001). Στις περιπτώσεις που η σημαντικότητα των ελέγχων χ^2 είναι οριακή (0,01<p<0,05), τότε θα πρέπει να χρησιμοποιούνται περισσότερο ως οδηγός και όχι ως απόδειξη της στατιστικής σημαντικότητας των αξόνων (Greenacre, 1993α). Μάλλον θα πρέπει να συνδυάζονται και με εμπειρικά κριτήρια. Πιο ασφαλής είναι η μέθοδος ελέγχου της στατιστικής σημαντικότητας του υποχώρου προβολής, την οποία παρουσιάσαμε στην Ενότητα 5.12. Η μόνη απαίτηση της συγκεκριμένης μεθόδου είναι το τυχαίο δείγμα να είναι αρκούντος μεγάλο. Βέβαια, και στην περίπτωση των ελέγχων σημαντικότητας των παραγοντικών αξόνων ιδιαίτερη μέριμνα θα πρέπει να ληφθεί για την διόρθωση του ε.σ. α για την αποφυγή της

εξάπλωσης του Αθροιστικού Σφάλματος Τύπου Ι. Περισσότερα στοιχεία και προσεγγίσεις σχετικά με τη σημαντικότητα των παραγοντικών αξόνων παραθέτουν οι Nishisato (1980), Greenacre (1984) και Lebart, Marineau και Warwick (1984).

6.2.1.4 Στατιστική Σημαντικότητα του Ποσοστού της Ολικής Αδράνειας που Ερμηνεύουν οι Παραγοντικοί Άξονες

Μπορεί να δειχθεί ότι ποσοστό της ολικής αδράνειας που ερμηνεύει ο κάθε παραγοντικός άξονας είναι ανεξάρτητο από την ολική αδράνεια του πίνακα **F** (Lebart, 1976). Αυτό πρακτικά σημαίνει ότι η ολική αδράνεια του **F** μπορεί να μην είναι στατιστικά σημαντική αλλά η ανασύστασή του, για παράδειγμα, από το παραγοντικό επίπεδο 1×2 να μην οδηγεί σε απώλεια στατιστικά σημαντικής πληροφορίας (βλέπε Ενότητα 5.12). Η στατιστική σημαντικότητα του ποσοστού της ολικής αδράνειας, το οποίο ερμηνεύουν οι παραγοντικοί άξονες, μπορεί να ελεγχθεί με τη χρήση διαγραμμάτων που έχουν κατασκευαστεί με μεθόδους επαναδειγματοληψίας (Lebart 1976, Lebart, Morineau & Tabard 1977, Lebart, Morineau & Warwick 1984, Lebart, Morineau & Piron 2000).

6.2.1.5 Στατιστική Σημαντικότητας του Υποχώρου Προβολής

Στις Ενότητες 5.12 και 5.13.4 παρουσιάσαμε μεθοδολογία για τον εντοπισμό του ελάχιστου στατιστικά σημαντικού υποχώρου προβολής.

Μια εναλλακτική προσέγγιση αποτελεί η πρόταση του Nishisato (1980) σύμφωνα με την οποία η ποσότητα:

$$-\left[N-1-\frac{1}{2}(k+l-1)\right]\sum_{s=t}^{p}\ln(1-\lambda_s), \qquad [6.2]$$

ακολουθεί την Κατανομή χ^2 με $(k\text{-}t)(l\text{-}t)$ β.ε. και ελέγχει τη στατιστική σημαντικότητα της αδράνειας[23] που δεν ερμηνεύεται από τους πρώτους t-1 άξονες. Σύμφωνα με την προσέγγιση αυτή, αν η υπόλοιπη αδράνεια των p-t αξόνων είναι

[23] Μεταφέροντας τη σχέση [6.2] από το πλαίσιο της Δυϊκής Κλιμάκωσης σε αυτό της ΠΑΑ.

στατιστικά σημαντική, τότε ο υποχώρος προβολής με *t* διαστάσεις δεν επαρκεί για την ανασύσταση του πίνακα συμπτώσεων χωρίς απώλεια στατιστικά σημαντικής πληροφορίας. Μέσω του Πίνακα 6.2 μπορεί να ελεγχθεί η σημαντικότητα του αντίστοιχου υποχώρου προβολής.

Πίνακας 6.2: Στατιστική Σημαντικότητα του Υποχώρου Προβολής (Πρόταση Nishisato)

Διαστάσεις Υποχώρου	Ιδιοτιμές	χ^2	β.ε.
1	$\lambda_2 .\cdots, \lambda_s, \cdots, \lambda_p$	$-c\sum_{z=2}^{p}\ln(1-\lambda_z)$	$(k\text{-}2)(l\text{-}2)$
2	$\lambda_3 .\cdots, \lambda_s, \cdots, \lambda_p$	$-c\sum_{z=3}^{p}\ln(1-\lambda_z)$	$(k\text{-}3)(l\text{-}3)$
⋮	⋮	⋮	⋮
s-1	$\lambda_s, \cdots, \lambda_p$	$-c\sum_{z=s}^{p}\ln(1-\lambda_z)$	$(k\text{-}s)(l\text{-}s)$

Με $c = \left[N - 1 - \frac{1}{2}(k + l - 1) \right]$.

6.2.1.6 Στατιστική Σημαντικότητα των Προφίλ Γραμμών ή/και Στηλών

Στην Ανάλυση Δεδομένων, το πρόβλημα του ελέγχου της στατιστικής σημαντικότητας των προφίλ γραμμών και στηλών του πίνακα συμπτώσεων αντιμετωπίζεται σχεδόν αποκλειστικά με μεθόδους επαναδειγματοληψίας με σκοπό την κατασκευή περιοχών εμπιστοσύνης γύρω από τα σημεία που προβάλλονται στα παραγοντικά επίπεδα. Για περισσότερες πληροφορίες παραπέμπουμε στην Ενότητα 6.3 όπου προτείνουμε και δύο προσεγγίσεις για την κατασκευή 100(1-α)% ελλείψεων εμπιστοσύνης γύρω από τα προβαλλόμενα σημεία γραμμών ή/και στηλών του πίνακα συμπτώσεων **F**.

Στην Ενότητα 2.2.14.2 παρατηρήσαμε ότι στις σχέσεις ανασύστασης [2.40] και [2.48] στηρίζονται ποικίλα συσχετιστικά υποδείγματα – μοντέλα, που έχουν προταθεί στο πλαίσιο της Επαγωγικής Στατιστικής, για την ανάλυση της σχέσης μεταξύ δύο ή περισσότερων κατηγορικών μεταβλητών (βλέπε Gilula 1986 και 1984, Gilula & Haberman 1988 και 1986, Choulakian 1988, Gilula & Krieger 1989, Gilula & Ritov 1990, Faust & Wasserman 1993, Van der Heijden, Mooijaart & Takane 1994, Haberman 1995 και 1981, Goodman 1996, 1993 και 1991, Mirkin 2001, Aït-Sidi-

Allal, Baccini & Mondot 2004). Τα αριθμητικά αποτελέσματα, εστιάζονται μόνο στις εκτιμήσεις μεγίστης πιθανοφάνειας των παραμέτρων λ_κ, $r_{i\kappa}$ και $c_{j\kappa}$, και είναι συγκρίσιμα με τα αντίστοιχα που υπολογίζονται από την ΠΑΑ. Κάτω από την ισχύ συγκεκριμένων υποθέσεων και προϋποθέσεων οι εκτιμήσεις των εκτιμώμενων παραμέτρων των υποδειγμάτων συνοδεύονται και από ελέγχους στατιστικής σημαντικότητας. Στην Ενότητα 6.2.1.3 είδαμε, επίσης, ότι μπορούμε να κατασκευάσουμε και ξεχωριστά 100(1-α)% διαστήματα εμπιστοσύνης για τις παραγοντικές συντεταγμένες των προφίλ γραμμών και στηλών.

6.2.1.7 Ανάλυση Ισχύος (*a priori* & *post hoc*) του Ελέγχου χ^2

Στο Κεφάλαιο 5 προτείναμε σχετική μεθοδολογία. Η *post hoc* ανάλυση ισχύος μπορεί να εφαρμοστεί και στους ελέγχους σημαντικότητας των παραγοντικών αξόνων που στηρίζονται στην Κατανομή χ^2, όπως για παράδειγμα στην πρόταση του Nishisato.

6.2.2 Η Περίπτωση Πολλών Μεταβλητών

Είδαμε στην Ενότητα 2.3 ότι στην περίπτωση πολλών μεταβλητών η ΠΑΑ μπορεί να εφαρμοστεί είτε στον λογικό πίνακα 0-1 είτε στον αντίστοιχο πίνακα *Burt*. Συνήθως εφαρμόζεται στον *Burt*, αλλά οι δύο αναλύσεις είναι ισοδύναμες (βλέπε Ενότητα 2.3.3.5). Στην Ενότητα 4.9 παρατηρήσαμε ότι στην περίπτωση του λογικού πίνακα η ολική αδράνεια δεν έχει την ίδια φυσική ερμηνεία όπως στην περίπτωση δύο μεταβλητών. Επίσης, η ολική αδράνεια του πίνακα *Burt* περιέχει πλεονάζουσα πληροφορία (βλέπε Ενότητα 4.6.2, Παρατήρηση 4.8) και για αυτό προτείναμε την ενδιαφέρουσα αδράνεια ως εναλλακτικό μέτρο του πληροφοριακού περιεχομένου του πίνακα *Burt* (βλέπε Ενότητα 4.7).

6.2.2.1 Στατιστική Σημαντικότητα της Ενδιαφέρουσας Αδράνειας

Προτείναμε σχετική μεθοδολογία στην Ενότητα 4.8.2.

6.2.2.2 Στατιστική Σημαντικότητα των Παραγοντικών Αξόνων

Γενικά, το πρόβλημα της εξωτερικής εγκυρότητας – σταθερότητας των παραγοντικών αξόνων αντιμετωπίζεται με μεθόδους επαναδειγματοληψίας (Ringrose 1992, Van de

Geer 1993β, Markus 1994α και 1994β, Gifi 1996, Michailidis 1996, Bond & Michailidis 1997 και 1996, Michailidis & De Leeuw 1998). Οι μέθοδοι προσφέρονται για την εκτίμηση ορίων εμπιστοσύνης, εντός των οποίων αναμένεται να βρίσκεται η πραγματική πληθυσμιακή τιμή των χαρακτηριστικών τιμών ή ιδιοτιμών των παραγοντικών αξόνων.

Μια εναλλακτική προσέγγιση, αυτή τη φορά στο πλαίσιο της Επαγωγικής Στατιστικής, προτείνει ο Nishisato (1980) σύμφωνα με την οποία το στατιστικό:

$$-[Nq\ -1-\frac{1}{2}(N+m-q-1)]\ln(1-\lambda_s), \qquad [6.3]$$

όπου m είναι το πλήθος των κλάσεων των q μεταβλητών, ακολουθεί ασυμπτωτικά τη χ^2 Κατανομή με $N+m-q-1-2s$ βαθμούς ελευθερίας. Μεταφέροντας το πρόβλημα από τη Δυϊκή Κλιμάκωση στην ΠΑΑ η στατιστική σημαντικότητα των παραγοντικών αξόνων μπορεί να ελεγχθεί μέσω της σχέσης [6.3]. Στην Ενότητα 6.4 προτείνουμε μέθοδο για την κατασκευή ενός μη παραμετρικού διαστήματος εμπιστοσύνης για τον έλεγχο της στατιστικής σημαντικότητας των παραγοντικών αξόνων στην πολυμεταβλητή εκδοχή της ΠΑΑ.

6.2.2.3 Στατιστική Σημαντικότητα των Προφίλ των Ιδιοτήτων των Μεταβλητών

Και στην πολυμεταβλητή περίπτωση το ζήτημα αντιμετωπίζεται σχεδόν αποκλειστικά με μεθόδους επαναδειγματοληψίας. Σκοπός είναι η κατασκευή περιοχών εμπιστοσύνης γύρω από τα σημεία που αντιστοιχούν στις ιδιότητες των μεταβλητών και προβάλλονται στα παραγοντικά επίπεδα (βλέπε Greenacre 1984, Markus 1994α και 1994β, Gifi 1996, Chateau & Lebart 1996, Michailidis 1996, Michailidis & De Leeuw 1998, Lebart 2005).

6.2.2.4 Στατιστική Σημαντικότητα των Συμπληρωματικών Στοιχείων

Στην περίπτωση αυτή ελέγχεται η σημαντικότητα της συσχέτισης των συμπληρωματικών στοιχείων με τους παραγοντικούς άξονες (Lebart, 1005). Πιο συγκεκριμένα:

Έστω η συμπληρωματική κλάση *j* με συχνότητα n_j. Η μηδενική υπόθεση δηλώνει ότι οι n_j παρατηρήσεις επιλέγονται τυχαία (χωρίς επανατοποθέτηση) από τις *N* συνολικά παρατηρήσεις ενός $N \times m$ λογικού πίνακα **Z** και, συνεπώς, δεν υπάρχει συστηματική σύνδεση της συμπληρωματικής κλάσης με τον άξονα, έστω *s*. Κάτω από την ισχύ της μηδενικής υπόθεσης, μπορεί να δειχθεί (Lebart, Morineau & Tabard 1977, Lebart, Morineau & Warwick 1984, Lebart, Morineau & Piron 2000) ότι επί του άξονα *s* η κύρια συντεταγμένη g_{js} της συμπληρωματικής κλάσης *j* είναι τυχαία μεταβλητή με μέση τιμή ίση με μηδέν και διακύμανση $Var(j)$, η οποία δίνεται από τη σχέση (Lebart, Morineau & Warwick, 1984):

$$Var(j) = \frac{N - n_j}{n_j (N - 1)}.$$

Η τυχαία μεταβλητή:

$$t_s(j) = \frac{g_{js}}{\sqrt{Var(j)}} \qquad [6.4]$$

ονομάζεται «*T*-Τιμή» (*Test-Value*), έχει μέση τιμή ίση με μηδέν, διακύμανση ίση με τη μονάδα και ακολουθεί ασυμπτωτικά την Τυποποιημένη Κανονική Κατανομή. Επομένως, μια *T*-τιμή ≥2 υποδεικνύει σημαντική θέση του σημείου *j* στον άξονα *s* (με δίπλευρο έλεγχο, σε επίπεδο σημαντικότητας α=0,05). Η σημαντικότητα της θέσης του σημείου εκφράζει ότι η συσχέτιση – σύνδεση του συμπληρωματικού σημείου με τον αντίστοιχο άξονα δεν μπορεί να αποδοθεί σε τυχαίες επιδράσεις. Οι *T*-τιμές μπορούν να χρησιμοποιηθούν και για την κατασκευή διαστημάτων εμπιστοσύνης για τις θέσεις των αντίστοιχων συμπληρωματικών σημείων επί των παραγοντικών αξόνων. Αν υποθέσουμε ότι οι συντεταγμένες των προβολών των συμπληρωματικών σημείων ακολουθούν την Κανονική Κατανομή, τότε αναμένουμε το 95% περίπου των προβολών να κυμαίνονται στο διάστημα:

$$\left(-2\sqrt{\frac{N - n_j}{n_j (N - 1)}}, +2\sqrt{\frac{N - n_j}{n_j (N - 1)}} \right),$$

όπου n_j είναι η συχνότητα της συμπληρωματικής κλάσης *j*.

Ακόμα και στην περίπτωση που η υπόθεση της Κανονικής Κατανομής των συντεταγμένων των προβολών δεν ισχύει, τότε, αν το μέγεθος του δείγματος είναι αρκούντος μεγάλο, χάρη στο Κεντρικό Οριακό Θεώρημα τα παραπάνω διαστήματα εμπιστοσύνης αποτελούν ένα χρήσιμο πλαίσιο αναφοράς για την επιλογή των σημαντικών συμπληρωματικών σημείων. Η διαπίστωση αυτή είναι ιδιαίτερα χρήσιμη στην περίπτωση δειγματοληπτικών ερευνών επισκόπησης όπου συνήθως ελέγχονται οι συσχετίσεις πολλών συμπληρωματικών μεταβλητών με τους παραγοντικούς άξονες (Lebart, 2005). Οι *T*-τιμές σε συνδυασμό με τους αντίστοιχους δείκτες *COR* μπορούν να χρησιμοποιηθούν για την ιεράρχηση του βαθμού της συσχέτισης των συμπληρωματικών σημείων με τους άξονες. Αν το πλήθος των συμπληρωματικών σημείων είναι μεγάλο, τότε και πάλι θα πρέπει να αντιμετωπίσουμε το πρόβλημα των πολλαπλών ελέγχων και την εξάπλωση του Αθροιστικού Σφάλματος Τύπου Ι.

6.2.2.5 Συνδυασμός της ΠΑΑ με Λογαριθμογραμμικά Υποδείγματα

Στην Ενότητα 4.7 διαπιστώσαμε ότι η ΠΑΑ μπορεί να συνδυαστεί αρχικά με τον έλεγχο της προσαρμογής του λογαριθμογραμμικού υποδείγματος, το οποίο αντιστοιχεί στην υπόθεση της πλήρους ανεξαρτησίας των q μεταβλητών και, στη συνέχεια, αφού η προηγούμενη υπόθεση απορριφθεί, με τον έλεγχο της καλής προσαρμογή του υποδείγματος, στο οποίο συμμετέχουν μόνο οι κύριες επιδράσεις των μεταβλητών και οι αλληλεπιδράσεις πρώτης τάξης. Άλλες εφαρμογές συνδυασμένης και συμπληρωματικής χρήσης της ΠΑΑ με λογαριθμογραμμικά υποδείγματα συναντάμε στους Van der Heijden και De Leeuw (1985), Van der Heijden και Worsley (1988), Van der Heijden, De Falguerolles και De Leeuw (1989), Novak και Hoffman (1990), Van der Heijden, De Vries και Van Hooff (1990), Κιοσέογλου και Δικαίου (1993), Van der Heijden, Mooijaart και Takane (1994), Goodman (1996, 1993 και 1991), Clausen (1998) και Panagiotakos και Pitsavos (2004).

6.3 Δύο Προτάσεις - Προσεγγίσεις στην Κατασκευή Ελλείψεων Εμπιστοσύνης στα Παραγοντικά Επίπεδα της ΠΑΑ

6.3.1 Εισαγωγή

Στην ενότητα αυτή παρουσιάζουμε και προτείνουμε δύο μεθοδολογικές προσεγγίσεις κατασκευής 100(1-α)% ελλείψεων εμπιστοσύνης με κέντρα τις προβολές των σημείων γραμμών ή/και στηλών επί των παραγοντικών επιπέδων, τα οποία παράγονται κατά την εφαρμογή της ΠΑΑ σε πίνακα συμπτώσεων απολύτων συχνοτήτων δύο κατηγορικών μεταβλητών. Το ζήτημα αυτό συνδέεται κυρίως με τον έλεγχο της εξωτερικής εγκυρότητας των γραφικών αποτελεσμάτων που παράγονται από την ΠΑΑ. Στο χώρο της Πολυδιάστατης Στατιστικής Ανάλυσης, ο όρος εξωτερική εγκυρότητα αναφέρεται στην περίπτωση που τα δεδομένα έχουν συγκεντρωθεί με απλή τυχαία δειγματοληψία από τον υπό εξέταση πληθυσμό και εστιάζεται στον έλεγχο του κατά πόσο συνεπή (σταθερά) είναι τα παραγόμενα αποτελέσματα, στη θεωρητική περίπτωση που η ανάλυση επαναληφθεί σε άλλα τυχαία δείγματα από τον ίδιο πληθυσμό (Greenacre 1993α και 1984, Markus 1994α, Gifi 1996, Michailidis & De Leeuw 1998). Με την έννοια αυτή, η εξωτερική εγκυρότητα των αποτελεσμάτων συνδέεται με τη στατιστική σταθερότητα - σημαντικότητα των δομών και σχέσεων, οι οποίες παρατηρούνται και ερμηνεύονται στα παραγοντικά επίπεδα που προκύπτουν από την εφαρμογή της ΠΑΑ. Γενικά, η απεικόνιση των σημείων σε ένα παραγοντικό επίπεδο μπορεί να χαρακτηριστεί ως εξωτερικά σταθερή αν ο προσανατολισμός των παραγοντικών αξόνων δεν μεταβάλλεται σημαντικά κατά την εφαρμογή της ΠΑΑ σε διαφορετικά ανεξάρτητα τυχαία δείγματα από τον ίδιο πληθυσμό (Greenacre, 1984). Όμως, στο χώρο των Κοινωνικών Επιστημών, η συλλογή πολλών και ανεξάρτητων δειγμάτων για τον ίδιο σκοπό και στον ίδιο χρόνο, είναι μάλλον ανέφικτη και, πολλές φορές, μη ρεαλιστική. Έτσι, για την αντιμετώπιση του προβλήματος της επαναληψιμότητας των αποτελεσμάτων θα πρέπει να καταφύγει κανείς είτε σε πορίσματα και μεθόδους της Επαγωγικής Στατιστικής είτε σε τεχνικές προσομοίωσης που βασίζονται στην

επαναδειγματοληψία (Greenacre 1984, Gifi 1996, Chateau & Lebart 1996, Michailidis & De Leeuw 1998, Lebart 2005).

Οι δύο προτεινόμενες προσεγγίσεις στηρίζονται στη θεώρηση ότι η εξωτερική εγκυρότητα των γραφικών αποτελεσμάτων μπορεί να ελεγχθεί ως προς τη σταθερότητα της θέσης της προβολής των σημείων γραμμών – στηλών επί των παραγοντικών επιπέδων. Οι προτάσεις μας βασίζονται στο γεγονός ότι στην περίπτωση δύο κατηγορικών μεταβλητών, όπου η συλλογή των δεδομένων έχει γίνει με τυχαία δειγματοληψία, οι συντεταγμένες των προβολών των σημείων γραμμών και στηλών επί των παραγοντικών επιπέδων, τα οποία παράγονται από την εφαρμογή της ΠΑΑ, αποτελούν εκτιμήσεις των πραγματικών συντεταγμένων, αυτών, δηλαδή, που θα προέκυπταν από την εφαρμογή της ΠΑΑ στα δεδομένα ολόκληρου του υπό εξέταση πληθυσμού. Έτσι, το πρόβλημα που τίθεται είναι ο προσδιορισμός περιοχών εμπιστοσύνης 100(1-α)% για τις πραγματικές θέσεις των σημείων (γραμμών, στηλών) επί των παραγοντικών επιπέδων της ΠΑΑ. Στο χώρο της Πολυδιάστατης Ανάλυσης Δεδομένων και, ιδιαίτερα, στο πλαίσιο των μεθόδων βέλτιστης κλιμάκωσης, όπως είναι η ΠΑΑ, η Μη Γραμμική Ανάλυση σε Κύριες Συνιστώσες και η Μη Γραμμική Κανονικοποιημένη Συσχέτιση, το ζήτημα αντιμετωπίζεται σχεδόν αποκλειστικά με τεχνικές *Bootstrap* (Heiser & Meulman 1983, Weinberg, Carroll & Cohen 1984, Van der Burg & De Leeuw 1988, Takane & Shibayama 1991, Greenacre 1993α και 1984, Gifi 1996, Michailidis & De Leeuw 1998, Lebart 2005). Για περισσότερες πληροφορίες σχετικά με την μέθοδο *Bootstrap* παραπέμπουμε στους Efron και Tibshirani (1993), Mooney και Duval (1993) και DiCiccio και Efron (1996). Όμως, στην περίπτωση της ΠΑΑ, η εφαρμογή των μεθόδων αυτών δημιουργεί τεχνικούς και θεωρητικούς προβληματισμούς σχετικά με την επιλογή της κατάλληλης μεθοδολογίας για την υλοποίηση και αξιοποίησή τους. Σημαντικές αποφάσεις θα πρέπει να ληφθούν σχετικά: α) με τη μορφή του πίνακα δεδομένων από τον οποίο θα γίνει η δειγματοληψία με επανατοποθέτηση, β) το πλήθος των δειγμάτων, γ) τη μέθοδο διόρθωσης (ή όχι) μεροληψίας των εκτιμητών, δ) την προβολή ή όχι των δειγμάτων σε ένα χώρο αναφοράς, ε) την αντιμετώπιση του προβλήματος της ενδεχόμενης μη εμφάνισης ορισμένων κατηγοριών των μεταβλητών στα επαναλαμβανόμενα δείγματα και στ) τον καθορισμό και έλεγχο της πιθανότητας κάλυψης (*coverage probability*) των περιοχών εμπιστοσύνης. Λαμβάνοντας υπόψη και τους προβληματισμούς που προαναφέρθηκαν, η μέθοδος υλοποίησης των τεχνικών επαναδειγματοληψίας

διαφοροποιείται ανάλογα με το αν το ενδιαφέρον επικεντρώνεται στην εκτίμηση παραμέτρων ή στον έλεγχο της «εσωτερικής» σταθερότητας των δομών και σχέσεων των μεταβλητών, οι οποίες αναδεικνύονται από τα αποτελέσματα της ΠΑΑ. Η εσωτερική σταθερότητα τεκμηριώνεται όταν μικρές ή/και ανεπαίσθητες “διαταραχές” των αρχικών δεδομένων δεν έχουν σημαντική επίδραση στα παρατηρούμενα αποτελέσματα (βλέπε Ενότητα 6.5). Χαρακτηριστικά αναφέρουμε ότι κατά την εφαρμογή της *Bootstrap* για την εκτίμηση παραμέτρων απαιτούνται τουλάχιστον 1000 τυχαία δείγματα (Markus, 1994α), ενώ για τον έλεγχο της εσωτερικής εγκυρότητας 30 είναι αρκετά (Lebart, 2005).

Για μια πληρέστερη παρουσίαση της *Bootstrap* στο πλαίσιο της ΠΑΑ παραπέμπουμε στους Greenacre (1993α και 1984), Markus (1994α και 1994β), Gifi (1996), Michailidis (1996), Michailidis και De Leeuw (1998) και Lebart (2005). Εφαρμογές της τεχνικής *Bootstrap* για την εκτίμηση παραμέτρων κατηγορικών, εν γένει, δεδομένων παραθέτουν οι Langeheine, Pannekoek και Van de Pol (1996) και Jhun και Jeong (2000), ενώ οι Milian και Whittaker (1995) προτείνουν μεθοδολογία εφαρμογής της μεθόδου σε αναλύσεις όπου εμπλέκεται η SVD. Σε άλλες μορφές εγκυρότητας, πέραν της εξωτερικής και της εσωτερικής, έχουμε αναφερθεί στην Ενότητα 1.5.3.

Στην παρούσα ενότητα προτείνουμε δύο προσεγγίσεις στην κατασκευή ελλείψεων εμπιστοσύνης γύρω από τις προβολές των γραμμών (στηλών) επί των παραγοντικών επιπέδων. Στην πρώτη προσέγγιση, οι συντεταγμένες των σημείων θεωρούνται ως τυχαίες μεταβλητές, ενώ στη δεύτερη ως μέσοι όροι βέλτιστα ποσοτικοποιημένων (μετασχηματισμένων) βαθμών των πειραματικών ή δειγματοληπτικών μονάδων (βλέπε Ενότητα 2.5). Η βασική ιδέα και στις δύο προσεγγίσεις είναι η εφαρμογή μιας «τοπικής» Ανάλυσης σε Κύριες Συνιστώσες γύρω από κάθε σημείο γραμμής (στήλης). Στη συνέχεια, ο προσανατολισμός των αξόνων των ελλείψεων και τα μήκη τους καθορίζονται με μέθοδο, η οποία συνδυάζει, ανάλογα με την προσέγγιση, ορισμένες ιδιότητες της Διδιάστατης Κανονικής Κατανομής και το επίπεδο σημαντικότητας α.

6.3.2 Το Γενικό Πρόβλημα

Το γενικό πρόβλημα που προτείνουμε να επιλυθεί αρχικά μπορεί να διατυπωθεί ως εξής:

Δοθείσης της κατ' εκτίμηση θέσης ενός σημείου σε ένα ορθογώνιο σύστημα συντεταγμένων του $\Re^2$, να προσδιοριστεί περιοχή στην οποία αναμένεται, με προκαθορισμένη πιθανότητα 1-α, να βρίσκεται η πραγματική του θέση θεωρώντας τις συντεταγμένες του ως τυχαίες μεταβλητές.

Η "μεταφορά" και η επίλυση του προβλήματος στο πλαίσιο της ΠΑΑ παρουσιάζει ορισμένες ιδιαιτερότητες όπως: α) ο προσανατολισμός του συστήματος συντεταγμένων είναι αυθαίρετος, β) υπάρχει διαφορετικός βαθμός "αβεβαιότητας", με την έννοια της αδράνειας, κατά μήκος των παραγοντικών αξόνων και γ) οι δειγματικές συντεταγμένες των σημείων μπορούν να θεωρηθούν είτε ως εκτιμήσεις των πραγματικών είτε ως μέσοι όροι βέλτιστα ποσοτικοποιημένων βαθμών των πειραματικών ή δειγματοληπτικών μονάδων, οι οποίες συμμετέχουν στην έρευνα. Οι παραπάνω ιδιαιτερότητες γεννούν ερωτήματα σχετικά με το:

- Τι σημαίνει "πραγματική θέση" των σημείων;
- Τι μορφή θα έχει η περιοχή εμπιστοσύνης;
- Ποια είναι η κατανομή των συντεταγμένων των σημείων;
- Πώς θα εκτιμηθούν οι παράμετροι της κατανομής τους;
- Ποια είναι η πρακτική χρησιμότητα της λύσης του προβλήματος και πώς μπορεί να συνδυαστεί με άλλες στατιστικές μεθόδους για την περαιτέρω αξιοποίηση των αποτελεσμάτων;

Στα επόμενα, αρχικά επιλύουμε το γενικό πρόβλημα και κατόπιν προτείνουμε δύο μεθοδολογικές προσεγγίσεις για τη μεταφορά και επίλυση του προβλήματος στο πλαίσιο της ΠΑΑ.

6.3.3 Επίλυση του Γενικού Προβλήματος

Έστω σημείο Α(u,v), όπου u και v είναι οι διαθέσιμες αμερόληπτες εκτιμήσεις των συντεταγμένων του σε ένα ορθογώνιο σύστημα συντεταγμένων (U×V) του $\Re^2$. Αν

θεωρήσουμε τα *u* και *v* ως τυχαίες μεταβλητές, τότε ο πίνακας διασπορών – συνδυασπορών **Σ**:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_u^2 & \sigma_{uv}^2 \\ \sigma_{uv}^2 & \sigma_v^2 \end{bmatrix}, \text{ με } \sigma_u^2 \neq \sigma_v^2 \text{ εν γένει, } \sigma_u^2\sigma_v^2 \neq 0 \text{ και } \sigma_u^2, \sigma_v^2 < \infty,$$

περιγράφει την αβεβαιότητα της κατ' εκτίμηση θέσης του σημείου Α κατά μήκος των δύο αξόνων του συστήματος συντεταγμένων. Αν οι εκτιμήσεις *u* και *v* ακολουθούν τη $N_2(\mathbf{\mu}, \mathbf{\Sigma})$, όπου $\mathbf{\mu}=[\mu_u, \mu_v]^{\mathbf{T\alpha}}$ το διάνυσμα των αντίστοιχων μέσων τιμών, τότε μπορούμε να θεωρήσουμε ότι οι αναμενόμενες τιμές μ_u και μ_v είναι οι άγνωστες πραγματικές συντεταγμένες της θέσης του σημείου Α επί του επιπέδου. Έτσι, είναι δυνατό να κατασκευαστούν ξεχωριστά 100(1-*α*)% διαστήματα εμπιστοσύνης για τις αναμενόμενες τιμές μ_u και μ_v μέσω των παρακάτω σχέσεων:

$$P(u\text{-}z_{\alpha/2}\sigma_u \leq \mu_u \leq u+z_{\alpha/2}\sigma_u) = 1\text{-}\alpha$$

και

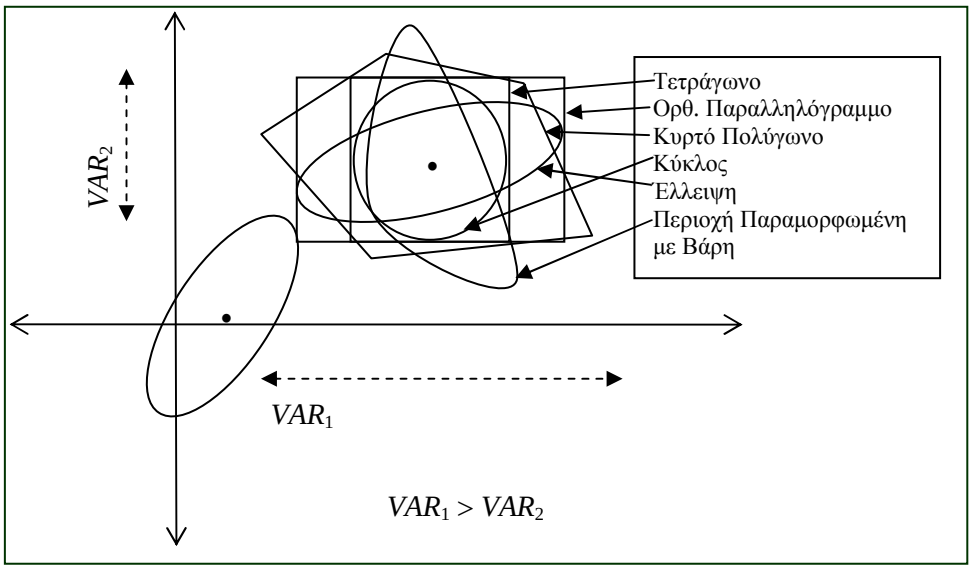
$$P(v\text{-}z_{\alpha/2}\sigma_v \leq \mu_v \leq v+z_{\alpha/2}\sigma_v) = 1\text{-}\alpha,$$

όπου $z_{\alpha/2}$ είναι η κρίσιμη τιμή της Τυποποιημένης Κανονικής Κατανομής σε επίπεδο σημαντικότητας *α*/2.

Όμως, τα παραπάνω διαστήματα δεν ισχύουν ταυτόχρονα (Δερμάνης 1986, Srivastava 2002), δηλαδή:

$$P(u\text{-}z_{\alpha/2}\sigma_u \leq \mu_u \leq u+z_{\alpha/2}\sigma_u \text{ και } v\text{-}z_{\alpha/2}\sigma_v \leq \mu_v \leq v+z_{\alpha/2}\sigma_v) \neq 1\text{-}\alpha.$$

Από πρακτική σκοπιά, θα ήταν χρήσιμο να αναζητήσουμε μια «περιοχή εμπιστοσύνης» στο επίπεδο ανεξάρτητα από το εν γένει αυθαίρετο σύστημα αναφοράς, τέτοια ώστε η πραγματική θέση του Α να βρίσκεται μέσα σε αυτή με προκαθορισμένη πιθανότητα *P*=1-*α* (Knight 2000, Srivastava 2002). Σε ό,τι αφορά τη μορφή της περιοχής έχουμε να κάνουμε ορισμένες επισημάνσεις, οι οποίες αφορούν τόσο στην περίπτωση του γενικού προβλήματος όσο και στην περίπτωση της ΠΑΑ (βλέπε και Σχήμα 6.1).



Σχήμα 6.1: Περιοχές Εμπιστοσύνης 100(1-α)%

Η περιοχή θα μπορούσε να είναι ένα ορθογώνιο παραλληλόγραμμο. Όμως, αυτό δεν θα οδηγούσε σε μονοσήμαντη λύση, εφόσον με κατάλληλη αλλαγή των διαστάσεων του αρχικού θα μπορούσε να προκύψει νέο ορθογώνιο παραλληλόγραμμο με το ίδιο εμβαδόν. Η περιοχή θα μπορούσε να είναι ένα τετράγωνο. Σε αυτή την περίπτωση δεν λαμβάνεται υπόψη ο διαφορετικός βαθμός αβεβαιότητας – διακύμανσης της θέσης του σημείου κατά μήκος των αξόνων. Η περιοχή θα μπορούσε να είναι ένας κύκλος. Μάλιστα, οι Lebart, Morineau και Warwick (1984) προτείνουν μεθοδολογία κατασκευής κύκλων εμπιστοσύνης γύρω από τα ενεργά ή/και συμπληρωματικά σημεία γραμμών (στηλών) επί των παραγοντικών επιπέδων που παράγονται από την ΠΑΑ (βλέπε και Beh, 2001). Όμως, και πάλι δεν λαμβάνεται υπόψη ο διαφορετικός βαθμός αβεβαιότητας της θέσης των σημείων κατά μήκος των αξόνων. Η περιοχή θα μπορούσε να είναι ένα κυρτό πολύγωνο. Τέτοιες περιοχές προκύπτουν, στο πλαίσιο της ΠΑΑ, από την εφαρμογή τεχνικών *Bootstrap* και, μάλιστα, μετά από μια διαδικασία “απολέπισης” (*peeling*), δηλαδή απομάκρυνσης των ακραίων τιμών και εξομάλυνσης της πολυγωνικής γραμμής (Green & Silverman 1979, Green 1981, Greenacre 1993α και 1984). Ο Snee (1974) προτείνει την κατασκευή περιοχών εμπιστοσύνης “παραμορφωμένων” ανάλογα με το βάρος των σημείων, οι οποίες έχουν εφαρμογή κυρίως σε πίνακες συμπτώσεων με τρεις γραμμές (στήλες). Τέλος, η ζητούμενη περιοχή θα μπορούσε να είναι μια έλλειψη, τέτοια ώστε να λαμβάνεται υπόψη τόσο ο διαφορετικός βαθμός αβεβαιότητας της θέσης του σημείου κατά μήκος των αξόνων όσο και η πιθανή συσχέτιση των εκτιμητριών των συντεταγμένων. Έτσι,

αν $\mathbf{x}=[u,v]^{\mathbf{T}}$ και $\mathbf{x}\sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, τότε η ποσότητα Q, που ορίζεται από την παρακάτω σχέση [6.5], ακολουθεί την κατανομή χ^2 με 2 βαθμούς ελευθερίας (John 1968, Penny 1996, Becker & Gather 2001):

$$Q=(\mathbf{x}-\boldsymbol{\mu})^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \sim \chi_2^2 \ . \qquad [6.5]$$

Η ποσότητα Q δεν είναι παρά η απόσταση *Mahalanobis* του διανύσματος **x** από το διάνυσμα **μ** (Jackson 1991, De Maesschalck, Jouan-Rimbaud & Massart 2000, Srivastava 2002, Mardia, Kent & Bibby 2003). Από την [6.5] συνεπάγεται ότι:

$$P((\mathbf{x}-\boldsymbol{\mu})^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \leq \chi_2^2) = 1-\alpha \ . \qquad [6.6]$$

Συνεπώς, η ζητούμενη περιοχή εμπιστοσύνης είναι το εσωτερικό και το σύνορο της καμπύλης με εξίσωση:

$$c\text{: } (\mathbf{x}-\boldsymbol{\mu})^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \chi_{2;\alpha}^2 \ , \qquad [6.7]$$

όπου $\chi_{2;\alpha}^2$ είναι η κρίσιμη τιμή της κατανομής χ^2 με 2 βαθμούς ελευθερίας, σε επίπεδο σημαντικότητας α.

Η καμπύλη c της σχέσης [6.7], όπως θα δείξουμε στη συνέχεια, είναι εξίσωση έλλειψης με κέντρο το σημείο (μ_u, μ_v). Η βασική ιδέα της προτεινόμενης μεθοδολογίας κατασκευής ελλείψεων εμπιστοσύνης στηρίζεται στην εφαρμογή μιας «τοπικής» Ανάλυσης σε Κύριες Συνιστώσες στο σημείο Α (Mardia, Kent & Bibby, 2003). Για το λόγο αυτό, αρχικά, αλλάζουμε το σύστημα αναφοράς μετατοπίζοντας την αρχή του στο σημείο (μ_u, μ_v) και, στη συνέχεια, το στρέφουμε κατά γωνία φ (θετική φορά) σύμφωνα με τη σχέση (Δερμάνης1986, Mardia, Kent & Bibby 2003):

$$\mathbf{x}^*=\mathbf{R}(\mathbf{x}-\boldsymbol{\mu}), \qquad [6.8]$$

όπου $\mathbf{x}^*$ είναι το διάνυσμα συντεταγμένων του σημείου Α στο νέο σύστημα αναφοράς και **R** ορθογώνιος πίνακας περιστροφής. Πιο αναλυτικά έχουμε (Sharma 1996, Lipschutz & Lipson 2003):

$$\mathbf{x}^*=\begin{bmatrix} u^* \\ v^* \end{bmatrix}=\begin{bmatrix} \sigma\upsilon\nu\varphi & \eta\mu\varphi \\ -\eta\mu\varphi & \sigma\upsilon\nu\varphi \end{bmatrix}\begin{bmatrix} u-\mu_u \\ v-\mu_v \end{bmatrix}.$$

Ο πίνακας $\boldsymbol{\Sigma}$ στο νέο σύστημα αναφοράς δίνεται από τη σχέση (Rao, 2002):

$$\mathbf{\Xi}=\mathbf{R\Sigma R^T}. \qquad [6.9]$$

Η γωνία φ μπορεί να επιλεγεί με τρόπο ώστε ο ορθογώνιος πίνακας **R** να διαγωνιοποιεί τον **Σ** (Δερμάνης 1998 και 1986, Lipschutz & Lipson 2003, Mardia, Kent & Bibby 2003) με αποτέλεσμα ο **Ξ** να μπορεί να γραφεί:

$$\mathbf{\Xi}=\begin{bmatrix}\sigma_{u^*}^2 & 0\\ 0 & \sigma_{v^*}^2\end{bmatrix},$$

όπου οι διασπορές $\sigma_{u^*}^2$ και $\sigma_{v^*}^2$ είναι ταυτόχρονα και ιδιοτιμές του **Σ**.

Από την [6.9] συνεπάγεται ότι:

$$\mathbf{\Sigma=R^T\Xi R}\Rightarrow\mathbf{\Sigma^{-1}=R^T\Xi^{-1}R}. \qquad [6.10]$$

Από τις σχέσεις [6.10], [6.8] και [6.7] έχουμε:

$$\mathbf{(x-\mu)^T\Sigma^{-1}(x-\mu)=x^{*T}\Xi^{-1}x^*}=$$

$$=\begin{bmatrix}u^*,v^*\end{bmatrix}\begin{bmatrix}\sigma_{u^*}^2 & 0\\ 0 & \sigma_{v^*}^2\end{bmatrix}^{-1}\begin{bmatrix}u^*\\ v^*\end{bmatrix}=\frac{u^{*2}}{\sigma_{u^*}^2}+\frac{v^{*2}}{\sigma_{v^*}^2}=\chi_{2;\alpha}^2. \qquad [6.11]$$

Αν θέσουμε $p_\alpha=\sqrt{\chi_{2;\alpha}^2}$, τότε η [6.11] γράφεται:

$$\frac{u^{*2}}{p_\alpha^2\sigma_{u^*}^2}+\frac{v^{*2}}{p_\alpha^2\sigma_{v^*}^2}=1 \qquad [6.12]$$

Η [6.12] είναι εξίσωση έλλειψης με κέντρο την αρχή του νέου συστήματος αναφοράς, με μήκη ημιαξόνων $p_\alpha\sigma_{u^*}$ και $p_\alpha\sigma_{v^*}$ αντίστοιχα, των οποίων οι διευθύνσεις είναι ίδιες με αυτές των αξόνων των *u** και *v**. Οι $\sigma_{u^*}^2$, $\sigma_{v^*}^2$ και η γωνία φ μπορούν να υπολογιστούν από τις παρακάτω σχέσεις (βλέπε Δερμάνης, 1998 και 1986)[24]:

$$\sigma_{u^*}^2=\frac{\sigma_u^2+\sigma_v^2}{2}+\sqrt{\left(\frac{\sigma_u^2-\sigma_v^2}{2}\right)^2+\sigma_{uv}^2}, \qquad [6.13]$$

[24] Οι σχέσεις [6.13] και [6.14] προκύπτουν από την επίλυση της δευτεροβάθμιας χαρακτηριστικής εξίσωσης του πίνακα **Σ,** $|\mathbf{\Sigma}-\lambda\mathbf{I}|=0$. Η [6.15] προκύπτει από τη σχέση $\mathbf{\Sigma U}=\lambda\mathbf{U}$ αν θέσουμε $\mathbf{U}=[\sigma\upsilon\nu\varphi, \eta\mu\varphi]$.

$$\sigma_{v^*}^2 = \frac{\sigma_u^2 + \sigma_v^2}{2} - \sqrt{\left(\frac{\sigma_u^2 - \sigma_v^2}{2}\right)^2 + \sigma_{uv}^2}\ , \qquad [6.14]$$

και

$$\varepsilon\phi 2\varphi = \frac{2\sigma_{uv}}{\sigma_u^2 - \sigma_v^2}\ . \qquad [6.15]$$

Σε πρακτικές εφαρμογές, προτείνουμε η έλλειψη εμπιστοσύνης να κατασκευάζεται με κέντρο το σημείο που ορίζεται από τις διαθέσιμες εκτιμήσεις *u* και *v* (βλέπε Σχήμα 6.2). Τα μήκη των ημιαξόνων $m_1 = p_\alpha \sigma_{u^*}$ και $m_2 = p_\alpha \sigma_{v^*}$ καθώς και η γωνία φ μπορούν να υπολογιστούν μέσω των σχέσεων [6.13], [6.14] και [6.15] με αντικατάσταση των δειγματικών εκτιμήσεων των παραμέτρων, αφού οι πραγματικές τους τιμές είναι συνήθως άγνωστες. Στην περίπτωση αυτή, η ποσότητα:

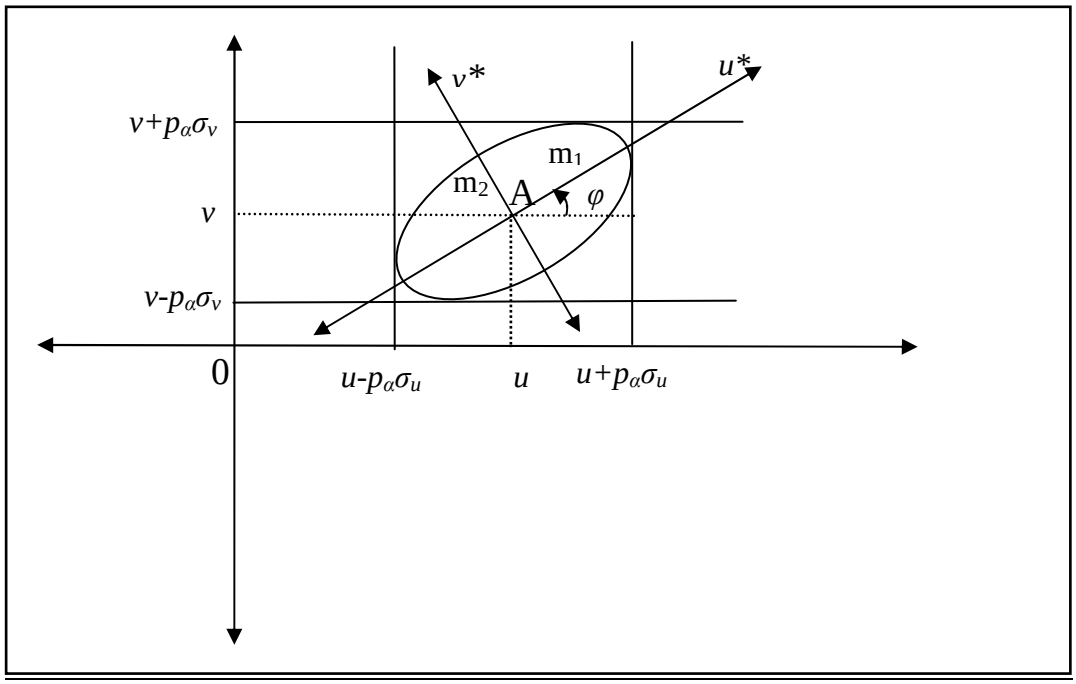
$$Q' = (\mathbf{x}-\boldsymbol{\mu})^{\mathbf{T}}\mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \sim \chi_2^2\ ,$$

όπου **S** είναι ο δειγματικός πίνακας διασπορών - συνδυασπορών, ακολουθεί ασυμπτωτικά τη χ^2 Κατανομή με 2 βαθμούς ελευθερίας (Rousseeuw & Van Zomeren 1990, Atkinson 1994, Becker & Gather 2001) και η αντίστοιχη περιοχή αποτελεί ασυμπτωτική έλλειψη εμπιστοσύνης. Είναι φανερό ότι η εγκυρότητα της μεθόδου βασίζεται τελικά στο πόσο "καλές" εκτιμήσεις θα έχουμε στη διάθεσή μας για τα στοιχεία του πίνακα **Σ**.

Η επίλυση του Γενικού Προβλήματος στηρίζεται κυρίως στις γεωμετρικές ιδιότητες της Διδιάστατης Κανονικής Κατανομής για την κατασκευή ελλείψεων σταθερής πυκνότητας πιθανότητας (*contours*) (Chew 1966, Ζαχαροπούλου 1994, Mardia, Kent & Bibby 2003), τον καθορισμό «περιοχών ανοχής»[25] (*tolerance regions*) (Wilks 1942, Wald & Wolfowitz 1946, Proschan 1953, Burrows 1963, Chew 1966, John 1968) και τον εντοπισμό πολυδιάστατων παράτυπων σημείων (*multivariate outliers*) (Rousseeuw & Van Zomeren 1990, Atkinson 1994, Rocke & Woodruff 1996, Penny 1996, Kosinski 1999, Becker & Gather 2001). Ανάλογα προβλήματα εμφανίζονται στο χώρο των Γεωδαιτικών Επιστημών για τον προσδιορισμό περιοχής εμπιστοσύνης,

[25] Πρόκειται για περιοχές όπου με προκαθορισμένη πιθανότητα β αναμένεται να βρίσκεται ένα συγκεκριμένο ποσοστό γ των παρατηρήσεων του αντίστοιχου πληθυσμού.

στην οποία βρίσκεται η πραγματική θέση ενός σημείου επί ενός συστήματος γεωγραφικών συντεταγμένων (βλέπε Δερμάνης, 1986).



Σχήμα 6.2: Κατασκευή Έλλειψης Εμπιστοσύνης 100(1-*α*)%

6.3.4 "Μεταφορά" και Επίλυση του Προβλήματος στην ΠΑΑ

6.3.4.1 Πρώτη Προσέγγιση

Η προσέγγιση αυτή στηρίζεται στο ότι οι δειγματικές συντεταγμένες των προβολών των σημείων γραμμών (στηλών) του πίνακα συμπτώσεων, επί των παραγοντικών επιπέδων της ΠΑΑ, μπορούν να θεωρηθούν ως εκτιμήσεις των πραγματικών, οι οποίες αντιμετωπίζονται ως τυχαίες μεταβλητές.

Το πρόβλημα που πρέπει να επιλυθεί στην περίπτωση της ΠΑΑ είναι η εκτίμηση των στοιχείων του πίνακα **Σ**, δηλαδή ο καθορισμός των στοιχείων του πίνακα **S** των δειγματικών διασπορών - συνδυασπορών για κάθε σημείο γραμμής (στήλης) του πίνακα συμπτώσεων απολύτων συχνοτήτων **F** των δύο μεταβλητών, στον οποίο θα εφαρμοστεί η ΠΑΑ. Στη σχετική βιβλιογραφία αναφέρονται δύο βασικές μέθοδοι εκτίμησης: α) μέθοδοι που στηρίζονται στις τεχνικές *Bootstrap* και *Jackknife* (Heiser & Meulman 1983, Weinberg, Carroll & Cohen 1984, Van der Burg & De Leeuw 1988, Markus 1994α, Takane & Shibayama 1991, Gifi 1996, Michailidis & De Leeuw 1998, Lebart 2005) και β) η μέθοδος *Δέλτα* (Israëls 1987, Stark 1990, De Leeuw

1993, Markus 1994α, Gifi 1996, Agresti 2002, Rao 2002). Για την πρώτη προσέγγιση προτείνουμε τη μέθοδο *Δέλτα* διότι: α) τα τυπικά σφάλματα των εκτιμητών της μεθόδου είναι "κοντά" στα τυπικά σφάλματα που προκύπτουν από την εφαρμογή των μεθόδων *Bootstrap* (τουλάχιστον για 500 επαναλήψεις), β) η μέθοδος είναι διαθέσιμη, μέσω προγραμματισμού, στο υποσύστημα *Categories* του στατιστικού πακέτου SPSS (Meulman & Heiser, 2004 και 2001), και 3) αποφεύγονται οι σχετικοί προβληματισμοί που αναφέρθηκαν στην Ενότητα 6.3.1. Σύντομη περιγραφή της μεθόδου *Δέλτα* παραθέτουμε στην Ενότητα ΣΤ1 του Παραρτήματος ΣΤ.

Στο σημείο αυτό λαμβάνουμε υπόψη ένα βασικό συμπέρασμα που προκύπτει από το συνδυασμό αποτελεσμάτων σχετικών μελετών (Nishisato 1980, Andersen 1991, De Leeuw 1993, Markus 1994α, Aït-Sidi-Allal, Baccini & Mondot 2004), το οποίο συμπυκνώνεται στην παρακάτω πρόταση:

Οι παραγοντικές συντεταγμένες των σημείων γραμμών και στηλών, όπως αυτές υπολογίζονται από την ΠΑΑ, στο πλαίσιο της Γαλλικής Σχολής Ανάλυσης Δεδομένων, είναι αμερόληπτοι εκτιμητές σταθμισμένων ελαχίστων τετραγώνων. Τα αντίστοιχα διανύσματα των συντεταγμένων ακολουθούν ασυμπτωτικά Πολυδιάστατη Κανονική Κατανομή.

Στη συνέχεια, αφού εκτιμηθούν τα στοιχεία του πίνακα **S** για κάθε σημείο γραμμής (στήλης), λαμβάνουμε υπόψη την προηγούμενη πρόταση και εφαρμόζουμε την διαδικασία κατασκευής των ελλείψεων εμπιστοσύνης, όπως στο γενικό πρόβλημα. Η κατασκευή "συντηρητικών" ελλείψεων μπορεί να επιτευχθεί με τη χρήση ορίων *Bonferroni* (Krzanowski & Radley 1989, Penny 1996). Πιο συγκεκριμένα, το επίπεδο εμπιστοσύνης α διορθώνεται σε α/r για τα σημεία γραμμών και σε α/c για τα σημεία στηλών, όπου r και c είναι το πλήθος γραμμών και στηλών αντίστοιχα του πίνακα **F**. Τα αποτελέσματα της ΠΑΑ μπορούν να "ενισχυθούν" με ταυτόχρονες πολλαπλές συγκρίσεις (ελέγχους) των προφίλ, ανά δύο, των γραμμών (ή/και στηλών) με τη μέθοδο του Gabriel (1966) (βλέπε Ενότητα ΣΤ3 του Παραρτήματος ΣΤ) ή του Hirotsou (1983).

6.3.4.2 Δεύτερη Προσέγγιση

Κατά την αλγοριθμική υλοποίηση της ΠΑΑ, επί των παραγοντικών επιπέδων, υπολογίζονται δύο ειδών συντεταγμένες των προβολών των σημείων γραμμών (στηλών) του πίνακα συμπτώσεων **F**: οι τυποποιημένες και οι κύριες (κανονικοποιημένες) (βλέπε Τρίτο και Τέταρτο Βήμα της Ενότητας 2.2.14). Λόγω των βαρυκεντρικών σχέσεων, οι οποίες συνδέουν τις συντεταγμένες των προβολών επί των παραγοντικών αξόνων των σημείων γραμμών και στηλών του πίνακα συμπτώσεων (Greenacre, 1993α και 1984, βλέπε και Τέταρτο Βήμα της Ενότητας 2.2.14), οι κύριες συντεταγμένες των προβολών των σημείων γραμμών (στηλών) αποτελούν σταθμισμένους μέσους όρους των τυποποιημένων συντεταγμένων των στηλών (γραμμών) του **F**. Στη διαπίστωση αυτή στηρίζεται η δεύτερη προτεινόμενη μέθοδος κατασκευής ελλείψεων εμπιστοσύνης γύρω από τις προβολές των σημείων γραμμών (στηλών) επί των παραγοντικών επιπέδων.

Έστω Α το σημείο που αντιστοιχεί στην i γραμμή του $k \times l$ πίνακα συμπτώσεων **F** και n_i το βάρος της στήλης i, δηλαδή το πλήθος των δειγματοληπτικών ή πειραματικών μονάδων που ανήκουν στην κλάση i της μεταβλητής γραμμών του **F**. Οι n_i μονάδες κατανέμονται στις l στήλες με βάρη f_{ij} με j=1,…, l και ισχύει $\sum_{j=1}^{l} f_{ij} = n_i$. Αν u_{is} είναι η κύρια συντεταγμένη του σημείου Α επί του παραγοντικού άξονα s και c_{1s}, c_{2s},…, c_{ls} οι τυποποιημένες συντεταγμένες των σημείων στηλών επί του s, τότε, σύμφωνα με σχέση [2.20] για $f_{i+} = n_i$, ισχύει:

$$u_{is} = \frac{\sum_{j=1}^{l} f_{ij} c_{js}}{n_i}. \qquad [6.16]$$

Έτσι, σε κάθε γραμμή του **F** που προβάλλεται με κύριες συντεταγμένες επί ενός παραγοντικού άξονα, αντιστοιχεί μια δεσμευμένη εμπειρική κατανομή n_i δειγματοληπτικών ή πειραματικών μονάδων με τιμές τις αντίστοιχες τυποποιημένες συντεταγμένες των στηλών του **F**. Συνεπώς, αν σε κάθε μία από τις n_i δειγματοληπτικές μονάδες της γραμμής i αντιστοιχίσουμε την τυποποιημένη συντεταγμένη της στήλης στην οποία ανήκει, δημιουργούμε μια εμπειρική κατανομή (για κάθε άξονα), όπου ο μέσος όρος της είναι η κύρια συντεταγμένη της γραμμής i.

Στην περίπτωση αυτή, το πρόβλημα κατασκευής περιοχής εμπιστοσύνης μπορεί να διατυπωθεί ως εξής:

Με βάση ένα τυχαίο δείγμα $\mathbf{x_1}, \mathbf{x_2},\ldots, \mathbf{x_n}$ *n παρατηρήσεων από την κατανομή* $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *με* $\boldsymbol{\mu}=[\mu_u, \mu_v]^{\mathbf{T}}$, *να κατασκευαστεί περιοχή εμπιστοσύνης, η οποία με πιθανότητα* (1-α)% *να περιέχει το σημείο* (μ_u, μ_v).

Γνωρίζουμε ότι αν $\mathbf{x} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$[26], τότε (Mardia, Kent & Bibby, 2003):

i. το διάνυσμα των αντίστοιχων δειγματικών μέσων όρων $\mathbf{m} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$
ii. και η ποσότητα $Q^*=(\mathbf{m}-\boldsymbol{\mu})^{\mathbf{T}} (\boldsymbol{\Sigma}/n)^{-1}(\mathbf{m}-\boldsymbol{\mu}) \sim \chi_2^2$.

Αν ο πίνακας $\boldsymbol{\Sigma}$ αντικατασταθεί με τον δειγματικό $\mathbf{S}$, τότε η ποσότητα Q^* ακολουθεί ασυμπτωτικά την Κατανομή *Hotteling's* $T^2(2, n\text{-}2)$ και λόγω της παρακάτω σχέσης [6.17] (Chew 1966, Jackson 1991, Srivastava 2002, Mardia, Kent & Bibby 2003, Nadarajah & & Kotz 2005):

$$T^2(2, n\text{-}2) = \frac{2(n-1)}{n-2} F(2, n-2), \qquad [6.17]$$

η έλλειψη με εξίσωση:

$$c^*=(\mathbf{m}-\boldsymbol{\mu})^{\mathbf{T}} (\mathbf{S}/n)^{-1}(\mathbf{m}-\boldsymbol{\mu}) = \frac{2(n-1)}{n-2} F(2, n-2; \alpha), \qquad [6.18]$$

αποτελεί την 100(1-α)% περιοχή εμπιστοσύνης για το σημείο (μ_u, μ_v).

Στη σχέση [6.17], $T^2(2, n\text{-}2)$ είναι η κατανομή T^2 του *Hotteling* με 2 και n-2 βαθμούς ελευθερίας και F η κρίσιμη τιμή της F κατανομής με 2 και n-2 βαθμούς ελευθερίας, σε επίπεδο σημαντικότητας α. Θα πρέπει να επισημάνουμε ότι το στατιστικό T^2 είναι ασυμπτωτικά μη παραμετρικό. Σύμφωνα με τους Oja και Randles (2004) αν $\mathbf{x_1}, \mathbf{x_2}, \ldots \mathbf{x_n}$ είναι ένα τυχαίο δείγμα από οποιαδήποτε p-Διάστατη Κατανομή με μέση

[26] Στην περίπτωση που οι περιθώριες συχνότητες των γραμμών (στηλών) του $\mathbf{F}$ είναι ≥30, τότε ισχύει το Κεντρικό Οριακό Θεώρημα (Johnson & Wichern 1992, Mardia, Kent & Bibby 2003) και η παραδοχή ότι $\mathbf{x} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ δεν είναι απαραίτητη.

τιμή **0** και πεπερασμένες ροπές δεύτερης τάξης, τότε η ποσότητα T^2, ως τυχαία μεταβλητή, τείνει ασυμπτωτικά προς την Κατανομή χ^2 με *p* βαθμούς ελευθερίας:

$$T^2 \to \chi^2(p).$$

Επομένως, όταν το μέγεθος δείγματος είναι μεγάλο, τότε η προϋπόθεση της Πολυδιάστατης Κανονικής Κατανομής δεν είναι και τόσο αυστηρή ή απαραίτητη. Δεν έχουμε παρά να αντικαταστήσουμε το δεξιό σκέλος της σχέσης [6.18] με την ποσότητα $\chi_a^2(p)$.

Μεταφέροντας το πρόβλημα στην περίπτωση της ΠΑΑ έχουμε: έστω Α(*u*,*v*) το σημείο που αντιστοιχεί στην *i* γραμμή του πίνακα **F**, η οποία έχει βάρος *n*, (*u*, *v*) οι κύριες συντεταγμένες του σημείου στους δύο παραγοντικούς άξονες και $\mathbf{x_1}$, $\mathbf{x_2}$,…, $\mathbf{x_n}$ τα διανύσματα του $\Re^2$ των τιμών των τυποποιημένων συντεταγμένων των στηλών, στις οποίες ανήκουν οι *n* μονάδες της *i* γραμμής. Λόγω του ότι οι κύριες συντεταγμένες *u* και *v* είναι οι μέσοι όροι των τυποποιημένων συντεταγμένων των στηλών προκύπτει ότι $\mathbf{m}=[u,v]^{\mathbf{T}}$, ενώ ο πίνακας **S** μπορεί να εκτιμηθεί από τις εμπειρικές κατανομές των *n* δειγματοληπτικών ή πειραματικών μονάδων στους δύο άξονες. Στη συνέχεια, εφαρμόζουμε την ίδια μεθοδολογία όπως και στην διαδικασία επίλυσης του γενικού προβλήματος, με μόνη διαφορά στους συντελεστές εμπιστοσύνης στην εξίσωση της έλλειψης, οι οποίοι υπολογίζονται λαμβάνοντας υπόψη και τη σχέση [6.18]. Τελικά, η έλλειψη της σχέσης [6.12] γράφεται ως εξής:

$$c^* = \frac{u^{*2}}{p_\alpha^2 \sigma_{u^*}^2} + \frac{v^{*2}}{p_\alpha^2 \sigma_{v^*}^2} = 1, \qquad [6.19]$$

με $p_\alpha = \sqrt{\frac{2(n-1)}{n(n-2)} F(2, n-2; \alpha)}$.

Θα πρέπει να παρατηρήσουμε ότι στην προσέγγιση αυτή υποθέτουμε ότι καμία από τις γραμμές (στήλες) του πίνακα **F** δεν είναι «αυστηρά» *unimodal*, δηλαδή να διασταυρώνεται αποκλειστικά με μία μόνο στήλη (γραμμή) του πίνακα **F**. Σε αντίθετη περίπτωση, η διακύμανση του αντίστοιχου προφίλ θα είναι ίση με μηδέν και δεν ορίζονται τα κλάσματα της σχέσης [6.19].

6.3.5 Παράδειγμα Εφαρμογής

Στον Πίνακα 6.3 παρουσιάζεται ο πίνακας συμπτώσεων απολύτων συχνοτήτων, τα σύνολα γραμμών και στηλών (περιθώριες κατανομές) καθώς και τα προφίλ (%) των γραμμών, για δύο κατηγορικές μεταβλητές *X* και *Y*, με 5 και 6 κλάσεις αντίστοιχα, σε ένα τυχαίο δείγμα 437 δειγματοληπτικών μονάδων. Χωρίς περιορισμό της γενικότητας, θεωρούμε ότι το ενδιαφέρον της μελέτης εστιάζεται στη σύγκριση μόνο των προφίλ των 5 κλάσεων της μεταβλητής *X* και στην αιτιολόγηση των ομοιοτήτων ή/και διαφορών, που ενδεχομένως θα προκύψουν, σε σχέση με τη μεταβλητή *Y*. Η εφαρμογή της ΠΑΑ, μέσω του SPSS έκδοση 13, στα δεδομένα του Πίνακα 6.3, ανέδειξε δύο παραγοντικούς άξονες που ερμηνεύουν το 88,8% της ολικής αδράνειας. Για λόγους οικονομίας, παραλείπουμε το σχολιασμό και την ερμηνεία των αναλυτικών αποτελεσμάτων που παράγονται από την ΠΑΑ και παρουσιάζουμε μόνο την εφαρμογή των προτεινόμενων μεθοδολογιών για την κατασκευή 95% ελλείψεων εμπιστοσύνης γύρω από τα σημεία που αντιστοιχούν στις 5 κλάσεις της *X* στο παραγοντικό επίπεδο 1×2. Στο Διάγραμμα 6.1 απεικονίζονται οι 95% ελλείψεις γύρω από τα σημεία που αντιστοιχούν στις 5 κλάσεις της μεταβλητής *X* (Tr_1, Tr_2, Tr_3, Tr_4 και Tr_5). Η σχεδίαση των ελλείψεων εμπιστοσύνης έγινε μέσω προγράμματος που αναπτύχθηκε στο λογισμικό Matlab έκδοση 7.0.

Σύμφωνα με την πρώτη προσέγγιση, η εκτίμηση των στοιχείων του πίνακα **S** πραγματοποιήθηκε με τη μέθοδο *Δέλτα* και τα αποτελέσματα παρουσιάζονται στον Πίνακα 6.4. Στον Πίνακα 6.5 παρουσιάζονται τα αποτελέσματα της δεύτερης προσέγγισης, όπου η εκτίμηση των στοιχείων του **S** πραγματοποιήθηκε μέσω της εμπειρικής κατανομής των 437 αντικειμένων ως προς τις ποσοτικοποιημένες τιμές των κλάσεων των στηλών του πίνακα συμπτώσεων. Τα αναλυτικά αριθμητικά αποτελέσματα και για τις δύο προσεγγίσεις, μαζί με σχετικά παραδείγματα υπολογισμών παρατίθενται στην Ενότητα ΣΤ2 του Παραρτήματος ΣΤ.

Στο Διάγραμμα 6.1 απεικονίζονται οι 95% ελλείψεις εμπιστοσύνης γύρω από τα σημεία γραμμών του πίνακα **F** κατασκευασμένες και με τις δύο προτεινόμενες μεθοδολογίες. Με έντονο περίγραμμα είναι σχεδιασμένες οι ελλείψεις εμπιστοσύνης με βάση την πρώτη προσέγγιση, ενώ με πιο αχνό οι ελλείψεις σύμφωνα με τη δεύτερη.

Πίνακας 6.3: Πίνακας Συμπτώσεων των *X* και *Y*

Μεταβλητή *X*		Μεταβλητή *Y* C1	C2	C3	C4	C5	C6	Σύνολα
Tr_1	Συχνότητες	18	17	16	3	4	16	74
	%	24,3%	23,0%	21,6%	4,1%	5,4%	21,6%	100,0%
Tr_2	Συχνότητες	12	17	20	3	7	3	62
	%	19,4%	27,4%	32,3%	4,8%	11,3%	4,8%	100,0%
Tr_3	Συχνότητες	5	16	43	10	36	6	116
	%	4,3%	13,8%	37,1%	8,6%	31,0%	5,2%	100,0%
Tr_4	Συχνότητες	1	7	11	10	6	4	39
	%	2,6%	17,9%	28,2%	25,6%	15,4%	10,3%	100,0%
Tr_5	Συχνότητες	7	9	29	11	90	0	146
	%	4,8%	6,2%	19,9%	7,5%	61,6%	0,0%	100,0%
Σύνολα	Συχνότητες	43	66	119	37	143	29	437
	%	9,8%	15,1%	27,2%	8,5%	32,7%	6,6%	100,0%

Στη συνέχεια παραθέτουμε τους βασικούς κανόνες ερμηνείας της πληροφορίας που παρουσιάζεται στο Διάγραμμα 6.1.

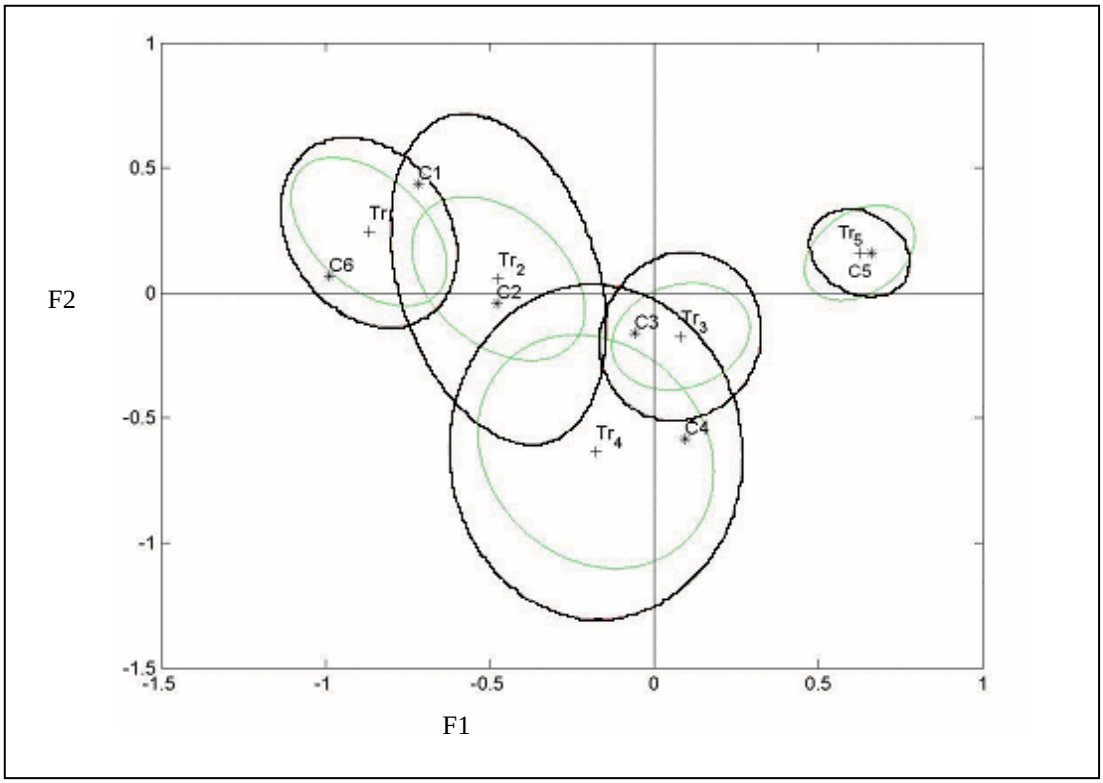
Τα σημεία γραμμών που η αντίστοιχη έλλειψη δεν περιέχει την αρχή των αξόνων είναι σημαντικά, δηλαδή διαφέρουν από το κέντρο βάρους (περιθώρια σχετική κατανομή) του νέφους των σημείων γραμμών, και συνεισφέρουν στη συσχέτιση (εξάρτηση) των δύο μεταβλητών, όπως αυτή ερμηνεύεται από το παραγοντικό επίπεδο 1×2. Από το διάγραμμα παρατηρούμε ότι τα σημεία Tr_1, Tr_2, Tr_4 και Tr_5 είναι σημαντικά, ενώ το Tr_3 δεν είναι. Για το συγκεκριμένο παράδειγμα, η προηγούμενη διαπίστωση ισχύει και για τις δύο προτεινόμενες προσεγγίσεις.

Πίνακας 6.4: Εκτίμηση Διασπορών – Συνδυασπορών με τη Μέθοδο *Δέλτα* (Πρώτη Προσέγγιση)

	Παραγοντικοί Άξονες *F1*	*F2*		Παραγοντικοί Άξονες *F1*	*F2*	
	Var(*u*)	*Var*(*v*)	*Covar*(*u*, *v*)	*SD*(*u*)	*SD*(*v*)	*Cor*(*u*, *v*)
Tr_1	0,012	0,024	-0,004	0,110	0,156	-0,242
Tr_2	0,018	0,073	-0,011	0,133	0,270	-0,296
Tr_3	0,010	0,019	0,001	0,099	0,136	0,105
Tr_4	0,033	0,075	-0,001	0,181	0,274	-0,017
Tr|_5	0,004	0,005	-0,001	0,063	0,069	-0,144

Πίνακας 6.5: Εκτίμηση Διασπορών – Συνδυασπορών μέσω της Εμπειρικής Κατανομής (Δεύτερη Προσέγγιση)

	Παραγοντικοί Άξονες *F1* *F2*			Παραγοντικοί Άξονες *F1* *F2*		
	Var(*u*)	*Var*(*v*)	*Covar*(*u*, *v*)	*SD*(*u*)	*SD*(*v*)	*Cor*(*u*, *v*)
Tr_1	0,659	1,031	-0,382	0,812	1,015	-0,463
Tr_2	0,670	1,039	-0,315	0,819	1,019	-0,378
Tr_3	0,822	0,852	0,140	0,907	0,923	0,167
Tr_4	0,751	1,276	-0,165	0,867	1,130	-0,169
Tr_5	0,665	0,857	0,257	0,815	0,926	0,340



Διάγραμμα 6.1: 95% Ελλείψεις Εμπιστοσύνης Γύρω από τα Σημεία Γραμμών του Πίνακα **F** στο Παραγοντικό Επίπεδο 1×2

Τα σημεία γραμμών που οι αντίστοιχες ελλείψεις δεν τέμνονται έχουν διαφορετικό προφίλ ιδιαίτερα στην περίπτωση που οι αντίστοιχοι παραγοντικοί άξονες ερμηνεύουν υψηλό ποσοστό της ολικής αδράνειας. Για παράδειγμα, τα προφίλ των γραμμών Tr_1, Tr_4 και Tr_5 διαφέρουν σημαντικά μεταξύ τους, ενώ τα προφίλ των γραμμών Tr_1 και Tr_2 καθώς και των γραμμών Tr_2 και Tr_4 δεν διαφέρουν σημαντικά. Η παρατήρηση αυτή ισχύει και για τις δύο προσεγγίσεις κατασκευής ελλείψεων εμπιστοσύνης. Οι συγκρίσεις των προφίλ μπορούν να "ενισχυθούν" περαιτέρω με τη μέθοδο Πολλαπλών Συγκρίσεων του *Gabriel* (βλέπε Ενότητα ΣΤ3

του Παραρτήματος ΣΤ), τα αποτελέσματα της οποίας προσαρμόστηκαν σε μορφή συνοπτικής παρουσίασης (βλέπε Πίνακα 6.6) ανάλογης με αυτή που χρησιμοποιείται στο SPSS για τις πολλαπλές συγκρίσεις μέσων όρων με μεθόδους όπως του *Tukey* και του *Duncan*. Οι πολλαπλοί έλεγχοι είναι μάλλον επιβεβλημένοι στην περίπτωση που οι αντίστοιχοι παραγοντικοί άξονες δεν ερμηνεύουν σημαντικό ποσοστό της ολικής αδράνειας.

Τα σημεία γραμμών που οι αντίστοιχες ελλείψεις έχουν σχετικά μικρή επιφάνεια έχουν και πιο σταθερή απεικόνιση, με την έννοια της εξωτερικής εγκυρότητας - σταθερότητας. Από το Διάγραμμα 6.1 φαίνεται ότι η θέση του σημείου Tr_4 παρουσιάζει τη μεγαλύτερη αστάθεια, με αποτέλεσμα η ερμηνεία του επί του πρώτου παραγοντικού άξονα να είναι προβληματική, αφού η θέση του θα μπορούσε να αποδοθεί είτε στα δεξιά είτε στα αριστερά του άξονα και με τις δύο προσεγγίσεις. Το ίδιο ισχύει για την ερμηνεία του σημείου Tr_2 ως προς τη θέση του στο δεύτερο άξονα (πάνω ή κάτω).

Αν και στο συγκεκριμένο παράδειγμα υπάρχει συμφωνία σε πολλά σημεία ως προς την ερμηνεία και αξιολόγηση των αποτελεσμάτων, ωστόσο τα αποτελέσματα των δύο προσεγγίσεων κατασκευής ελλείψεων εμπιστοσύνης δεν είναι ταυτόσημα. Για παράδειγμα, με την πρώτη προσέγγιση οι ελλείψεις που αντιστοιχούν στα σημεία Tr_2 και Tr_3 φαίνεται, έστω και οριακά, ότι έχουν κοινά σημεία, ενώ με τη δεύτερη προσέγγιση οι δύο ελλείψεις είναι αρκετά απομακρυσμένες και συνεπώς διακεκριμένες. Από το Διάγραμμα 6.1 φαίνεται, επίσης, ότι για το συγκεκριμένο σύνολο δεδομένων, οι ελλείψεις που κατασκευάστηκαν με την πρώτη προσέγγιση είναι πιο “συντηρητικές”, καλύπτουν δηλαδή μεγαλύτερη επιφάνεια επί του παραγοντικού επιπέδου, σε σχέση με τις ελλείψεις της δεύτερης προσέγγισης. Εξαίρεση αποτελεί το σημείο Tr_5, για το οποίο οι δύο ελλείψεις φαίνεται να έχουν το ίδιο εμβαδόν. Επίσης, ο προσανατολισμός των αξόνων των ελλείψεων, για όλα τα σημεία, δεν είναι ίδιος και για τις δύο προτεινόμενες μεθόδους, όπως συμβαίνει, για παράδειγμα, με το σημείο Tr_5. Τέλος, παρατηρώντας το διάγραμμα και τα αποτελέσματα του Πίνακα 6.6 φαίνεται ότι για το συγκεκριμένο παράδειγμα υπάρχει σχεδόν απόλυτη συμφωνία των αποτελεσμάτων της πρώτης προσέγγισης με αυτά που προκύπτουν από τις πολλαπλές συγκρίσεις των προφίλ γραμμών με τη μέθοδο του

Gabriel, η οποία, όμως, λόγω του *post hoc* χαρακτήρα της είναι συντηρητική εκ κατασκευής.

Πίνακας 6.6: Ταυτόχρονες Πολλαπλές Συγκρίσεις των Προφίλ των Γραμμών (Προσαρμοσμένη Μέθοδος του *Gabriel*)

Ζεύγος Αγωγών	*Likelihood Ratio-LR*	Ομοιογενής Ομάδα Γραμμών 1	Ομοιογενής Ομάδα Γραμμών 2	Ομοιογενής Ομάδα Γραμμών 3	Αποτελέσματα Συγκρίσεων**
Tr_1 *vs* Tr_2	11,187	Tr_1			Tr_1 **c**
Tr_1 *vs* Tr_3	49,435*	Tr_2	Tr_2		Tr_2 **bc**
Tr_1 *vs* Tr_4	24,800	Tr_4	Tr_4		Tr_3 **bc**
Tr_1 *vs* Tr_5	111,587*		Tr_3		Tr_4 **b**
Tr_2 *vs* Tr_3	21,351			Tr_5	Tr_5 **a**
Tr_2 *vs* Tr_4	16,831				
Tr_2 *vs* Tr_5	63,789*				
Tr_3 *vs* Tr_4	10,968				
Tr_3 *vs* Tr_5	33,893*				
Tr_4 *vs* Tr_5	41,596*				

(*) Στατιστικά σημαντική διαφορά σε επίπεδο σημαντικότητας α=0,05. Η κρίσιμη τιμή της $\chi^2_{20;0,05}$=31,140. (**) Τα προφίλ των γραμμών που ακολουθούνται από κοινό γράμμα δεν διαφέρουν στατιστικά σημαντικά, σε α=0,05, σύμφωνα με το στατιστικό έλεγχο του *Gabriel*.

6.4 Πρόταση Κατασκευής Μη Παραμετρικού Διαστήματος Εμπιστοσύνης για τον Έλεγχο της Στατιστικής Σημαντικότητας των Αξόνων στην Πολυμεταβλητή ΠΑΑ

Η προτεινόμενη μέθοδος στηρίζεται στην εύρεση των μαθηματικών σχέσεων για τον υπολογισμό της μέσης τιμής μ και της διακύμανσης σ^2 των ιδιοτιμών των παραγοντικών αξόνων που προκύπτουν από την εφαρμογή της ΠΑΑ στον πίνακα *Burt q* κατηγορικών μεταβλητών, οι οποίες είναι ανά δύο ανεξάρτητες. Έστω ότι *q* κατηγορικές μεταβλητές X_i $(i=1,\ldots,q)$, με *m* συνολικά κατηγορίες (κλάσεις), χαρακτηρίζουν *N* αντικείμενα. Ας είναι λ_{Bs} η αδράνεια του άξονα *s* που προκύπτει από την εφαρμογή της ΠΑΑ στον αντίστοιχο $m\times m$ πίνακα *Burt* **B** και $\sqrt{\lambda_{Bs}}$ η αντίστοιχη ιδιοτιμή. Αν οι *q* μεταβλητές είναι ανά δύο ανεξάρτητες, τότε η

ενδιαφέρουσα αδράνεια $I_{\varepsilon B}$ του πίνακα **B** είναι ίση με μηδέν (βλέπε Ενότητα 4.7). Από τη σχέση [4.40] της Ενότητας 4.8.3 έχουμε ότι:

$$I_{\varepsilon B} = \frac{\sum_{s=1}^{p}\left(q\sqrt{\lambda_{Bs}}-1\right)^2}{q(q-1)}, \text{ με } p=m-q.$$

Κάτω από την υπόθεση της ανεξαρτησίας των μεταβλητών ανά δύο ισχύει:

$$I_{\varepsilon B}=0 \Rightarrow \frac{\sum_{s=1}^{p}\left(q\sqrt{\lambda_{Bs}}-1\right)^2}{q(q-1)}=0 \Rightarrow \sum_{s=1}^{p}\left(q\sqrt{\lambda_{Bs}}-1\right)^2=0 \Rightarrow \left(q\sqrt{\lambda_{Bs}}-1\right)^2=0, \forall s \Rightarrow$$

$$\Rightarrow \left(q\sqrt{\lambda_{Bs}}-1\right)=0, \forall s \Rightarrow \sqrt{\lambda_{Bs}}=\frac{1}{q}, \forall s.$$

Συνεπώς, αν θεωρήσουμε την ιδιοτιμή $\sqrt{\lambda_{Bs}}$ ως τυχαία μεταβλητή, τότε η μέση τιμή της είναι:

$$E\left(\sqrt{\lambda_{Bs}}\right)=\mu=\frac{1}{q}.$$

Στη συνέχεια, ορίζουμε ένα μέτρο μεταβλητότητας V^2 της $\sqrt{\lambda_{Bs}}$ γύρω από τη μέση τιμή της ως εξής:

$$V^2=\frac{1}{p}\sum_{s=1}^{p}\left(\sqrt{\lambda_{Bs}}-\frac{1}{q}\right)^2=\frac{1}{p}\sum_{s=1}^{p}\left(\lambda_{Bs}-2\frac{\sqrt{\lambda_{Bs}}}{q}+\frac{1}{q^2}\right)=\frac{1}{p}\sum_{s=1}^{p}\left(\lambda_{Bs}-\frac{2}{q^2}+\frac{1}{q^2}\right)=$$

$$=\frac{1}{p}\sum_{s=1}^{p}\left(\lambda_{Bs}-\frac{1}{q^2}\right)=\frac{1}{p}\sum_{s=1}^{p}\lambda_{Bs}-\frac{1}{p}p\frac{1}{q^2}=\frac{1}{p}\sum_{s=1}^{p}\lambda_{Bs}-\frac{1}{q^2}.$$

Στις παραπάνω πράξεις λάβαμε υπόψη ότι $\forall s, \sqrt{\lambda_{Bs}}=\frac{1}{q}$.

Συνεπώς,

$$V^2=\frac{1}{p}\sum_{s=1}^{p}\lambda_{Bs}-\frac{1}{q^2}. \qquad [6.20]$$

Γνωρίζουμε ότι (βλέπε Ενότητα 2.3.3.5) η ολική αδράνεια I_B του πίνακα *Burt* δίνεται από τη σχέση:

$$I_B = \sum_{s=1}^{p} \lambda_{Bs} ,$$

η οποία λόγω της [4.28] γράφεται:

$$I_B = \sum_{s=1}^{p} \lambda_{Bs} = \frac{I_{0-1}}{q} + \frac{2}{q}\left(\frac{\sum_{h<w} I_{hw}}{q} \right), \text{ με } h, w=1,\ldots, q,$$

όπου $I_{0\text{-}1}$ είναι η ολική αδράνεια του αντίστοιχου λογικού πίνακα **Z** και $\sum_{h<w} I_{hw}$ το άθροισμα των αδρανειών των $q(q-1)/2$ διαφορετικών απλών πινάκων συμπτώσεων των q μεταβλητών ανά δύο. Η ποσότητα I_{hw} εκφράζει την αδράνεια του υποπίνακα που σχηματίζεται από τη διασταύρωση της μεταβλητής X_h με τη X_w και, επομένως, (βλέπε απόδειξη Πρότασης 5, Ενότητα 4.6.2):

$$\sum_{h<w} I_{hw} = \sum_{h<w} \frac{Q_{hw}}{N}, \qquad [6.21]$$

όπου Q_{hw} είναι το στατιστικό χ^2, το οποίο που αντιστοιχεί στον απλό πίνακα συμπτώσεων των μεταβλητών X_h και X_w, και $\sum_{h<w} Q_{hw}$ είναι το άθροισμα των στατιστικών χ^2 που αντιστοιχούν στους $q(q-1)/2$ διαφορετικούς απλούς πίνακες συμπτώσεων των q μεταβλητών ανά δύο.

Γνωρίζουμε, επίσης, ότι (βλέπε σχέση [4.25] για $j=m$):

$$I_{0-1} = \frac{m}{q} - 1 = \frac{m-q}{q} = \frac{p}{q}. \qquad [6.22]$$

Αν λάβουμε υπόψη τις σχέσεις [6.21] και [6.22] η I_B γράφεται:

$$I_B = \sum_{s=1}^{p} \lambda_{Bs} = \frac{p}{q^2} + \frac{2}{q}\left(\frac{\sum_{h<w} Q_{hw}}{Nq} \right) = \frac{p}{q^2} + \frac{1}{Nq^2} \sum_{h \neq w} Q_{hw} \ . \qquad [6.23]$$

Στην τελευταία ισότητα λάβαμε υπόψη ότι:

$$2\sum_{h<w} Q_{hw} = \sum_{h \neq w} Q_{hw} \ .$$

Αν οι q μεταβλητές είναι ανά δύο ανεξάρτητες, τότε το στατιστικό Q_{hw} ακολουθεί ασυμπτωτικά την Κατανομή χ^2 με $(k_h - 1)(k_w - 1)$ β.ε., όπου k_h και k_w είναι οι κλάσεις των μεταβλητών X_h και X_w αντίστοιχα. Οι ποσότητες Q_{hw}, ως τυχαίες μεταβλητές, κάτω από την υπόθεση της ανεξαρτησίας, έχουν μέση τιμή (Rencher 2000, Rao 2002):

$$E\left(Q_{hw}\right) = (k_h - 1)(k_w - 1) \qquad [6.24]$$

και εν γένει ισχύει:

$$E\left(\sum Q_{hw} \right) = \sum (k_h - 1)(k_w - 1) \, . \qquad [6.25]$$

Επομένως, από την [6.23] μέσω της [6.25] προκύπτει ότι:

$$E\left(I_B\right) = E\left(\sum_{s=1}^{p} \lambda_{Bs} \right) = E\left(\frac{p}{q^2} + \frac{1}{Nq^2} \sum_{h \neq w} Q_{hw} \right) = \frac{p}{q^2} + \frac{1}{Nq^2} E\left(\sum_{h \neq w} Q_{hw} \right) \Rightarrow$$

$$\Rightarrow E\left(I_B\right) = E\left(\sum_{s=1}^{p} \lambda_{Bs} \right) = \frac{p}{q^2} + \frac{1}{Nq^2} \sum_{h \neq w} (k_h - 1)(k_w - 1) \, . \qquad [6.26]$$

Στη συνέχεια, από τη σχέση [6.20] κάνοντας χρήση της [6.26] έχουμε ότι:

$$E\left(V^2\right)=E\left(\frac{1}{p}\sum_{s=1}^{p}\lambda_{Bs}-\frac{1}{q^2}\right)=\frac{1}{p}E\left(\sum_{s=1}^{p}\lambda_{Bs}\right)-\frac{1}{q^2}=$$

$$=\frac{1}{p}\left(\frac{p}{q^2}+\frac{1}{Nq^2}\sum_{h\neq w}\left(k_h-1\right)\left(k_w-1\right)\right)-\frac{1}{q^2}=$$

$$=\frac{1}{q^2}+\frac{1}{Npq^2}\sum_{h\neq w}\left(k_h-1\right)\left(k_w-1\right)-\frac{1}{q^2}=$$

$$=\frac{1}{Npq^2}\sum_{h\neq w}\left(k_h-1\right)\left(k_w-1\right).$$

Η μέση τιμή $E\left(V^2\right)$ εκφράζει τη διακύμανση σ^2 της τυχαίας μεταβλητής $\sqrt{\lambda_{Bs}}$ κάτω από την υπόθεση της ανεξαρτησίας των q μεταβλητών ανά δύο. Έτσι, μπορούμε να γράψουμε:

$$E\left(V^2\right)=Var\left(\sqrt{\lambda_{Bs}}\right)=\sigma^2=\frac{1}{Npq^2}\sum_{h\neq w}\left(k_h-1\right)\left(k_w-1\right). \qquad [6.27]$$

Μέχρι τώρα έχουμε δείξει ότι:

$$E\left(\sqrt{\lambda_{Bs}}\right)=\mu=\frac{1}{q}\quad \text{και}\quad Var\left(\sqrt{\lambda_{Bs}}\right)=\sigma^2=\frac{1}{Npq^2}\sum_{h\neq w}\left(k_h-1\right)\left(k_w-1\right). \qquad [6.28]$$

Συνεπώς, αν οι q μεταβλητές είναι ανά δύο ανεξάρτητες, τότε από την ανισότητα C*hebychev*[27] (βλέπε Hogg & Craig 1978, Κουνιάς & Μωϋσιάδης 1985, Bishop, Fienberg & Holland 1991) αναμένουμε τουλάχιστον το 75% των ιδιοτιμών $\sqrt{\lambda_{Bs}}$ να περιέχονται στο διάστημα:

$$\mu\pm 2\sigma,$$

δηλαδή στο:

$$\frac{1}{q}\pm 2\sqrt{\frac{1}{Npq^2}\sum_{h\neq w}\left(k_h-1\right)\left(k_w-1\right)}. \qquad [6.29]$$

---

[27] Για κάθε τυχαία μεταβλητή X ισχύει: $P(|X-\mu|\geq\lambda\sigma)\leq 1/\lambda^2$.

Έτσι, τα όρια της [6.29] ορίζουν ένα 75% διάστημα εμπιστοσύνης. Αν ορίσουμε το διάστημα:

$$\mu \pm 3\sigma ,$$

τότε αναμένουμε το 89% των ιδιοτιμών να περιέχονται σε αυτό.

Αν το μέγεθος του δείγματος είναι αρκούντος μεγάλο, τότε με βάση το Κεντρικό Οριακό Θεώρημα (Knight 2000, Agresti 2002) αναμένουμε περίπου το 95% των ιδιοτιμών να περιέχονται στο διάστημα [6.29]. Στην περίπτωση αυτή, μπορούμε να μιλάμε για ένα ασυμπτωτικό 95% διάστημα εμπιστοσύνης.

Το προτεινόμενο διάστημα εμπιστοσύνης μπορεί να χαρακτηριστεί ως “Μη Παραμετρικό” με την έννοια ότι για την κατασκευή του στηριχθήκαμε μόνο στις βασικές ιδιότητες της μέσης τιμής και της διακύμανσης των τυχαίων μεταβλητών.

Επομένως, κατά την ανάλυση του πίνακα *Burt*, μπορούμε να θεωρήσουμε ως στατιστικά σημαντικούς, σε επίπεδο σημαντικότητας α=0,05, τους άξονες για τους οποίους η αντίστοιχη ιδιοτιμή είναι μεγαλύτερη από το άνω όριο του προτεινόμενου διαστήματος.

Παρατηρούμε, όμως, ότι η ιδιοτιμή $\sqrt{\lambda_{Bs}}$ είναι ίση με την αδράνεια του άξονα *s* που προκύπτει από την ανάλυση του αντίστοιχου λογικού πίνακα **Z** (βλέπε Ενότητα 2.3.3.5). Άρα, στην περίπτωση που η ΠΑΑ εφαρμοστεί στον **Z**, το ίδιο κριτήριο μπορεί να χρησιμοποιηθεί αυτή τη φορά για τον έλεγχο της στατιστικής σημαντικότητας των αδρανειών των παραγοντικών αξόνων.

Στην Ενότητα 2.3.3.6 είδαμε ότι όταν η ανάλυση εφαρμόζεται στον πίνακα **Z** σημαντικοί θεωρούνται οι παραγοντικοί άξονες με αδράνεια μεγαλύτερη από $1/q$, ενώ στην περίπτωση του πίνακα *Burt* σημαντικοί θεωρούνται οι άξονες με αδράνειες μεγαλύτερες από $1/q^2$. Βέβαια, τα δύο αυτά κριτήρια είναι εμπειρικά. Το προτεινόμενο στατιστικό κριτήριο αποτελεί μια διόρθωση των δύο προηγούμενων.

6.4.1 Παράδειγμα Εφαρμογής

Ας υποθέσουμε ότι q=7, m=25, p=18, N=300, $\sum(k_h-1)(k_w-1)$=302 και ότι η εφαρμογή της ΠΑΑ στον αντίστοιχο λογικό πίνακα **Z** έδωσε τα αποτελέσματα που δίνονται στον Πίνακα 6.7.

Επειδή το μέγεθος του δείγματος είναι μεγάλο τα όρια της [6.29] ορίζουν ένα 95% ασυμπτωτικό διάστημα εμπιστοσύνης για τις αδράνειες των παραγοντικών αξόνων. Από τα δεδομένα του παραδείγματος έχουμε:

$$\mu = \frac{1}{q} \Rightarrow \mu = \frac{1}{7}$$

και

$$\sigma = \sqrt{\frac{1}{Npq^2}\sum_{h\neq w}(k_h-1)(k_w-1)} \Rightarrow \sigma = \sqrt{\frac{1}{300\times 18\times (7)^2}\times 302} = 0,034\,.$$

Επομένως, το κάτω και το άνω όριο του διαστήματος είναι αντίστοιχα:

$$\frac{1}{7} - 2\times 0,034 = 0,075$$

και

$$\frac{1}{7} + 2\times 0,034 = 0,210\,.$$

Συνεπώς, σύμφωνα με το προτεινόμενο κριτήριο ως στατιστικά σημαντικοί μπορούν να θεωρηθούν μόνο οι πέντε πρώτοι άξονες για τους οποίους η αντίστοιχη αδράνεια είναι μεγαλύτερη από 0,210.

Αν εφαρμόσουμε το εμπειρικό κριτήριο βάσει του οποίου σημαντικοί θεωρούνται οι παραγοντικοί άξονες με αδράνεια μεγαλύτερη από $1/q = 1/7 = 0,143$, τότε πρέπει να διατηρήσουμε τους οκτώ πρώτους άξονες για ερμηνεία.

Πίνακας 6.7: Αδράνειες Αξόνων

Άξονας	Αδράνεια
1	0,325
2	0,293
3	0,290
4	0,287
5	0,269
6	0,156
7	0,155
8	0,151
9	0,139
10	0,134
11	0,129
12	0,125
13	0,119
14	0,000
15	0,000
16	0,000
17	0,000
18	0,000

Για συγκριτικούς λόγους παραθέτουμε και τα αποτελέσματα της μεθόδου που προτείνει ο Nishisato (βλέπε Πίνακα 6.8), την οποία παρουσιάσαμε στην Ενότητα 6.2.2.2. Εκτός από την κρίσιμη τιμή της Κατανομής χ^2 σε ε.σ. α=0,05 υπολογίσαμε την αντίστοιχη τιμή στο διορθωμένο κατά *Bonferroni* ε.σ. $\alpha = \frac{0,05}{18} \approx 0,003$ καθώς και την *post hoc* ισχύ του ελέγχου σε ε.σ. α=0,003. Ο υπολογισμός της ισχύος έγινε με το λογισμικό Power Analysis for AFC μέσω του module **Post Hoc: Goodness of Fit** (βλέπε Ενότητα Ε2 του Παραρτήματος Ε).

Από την εξέταση του Πίνακα 6.8, διαπιστώνουμε ότι και με την προσέγγιση του Nishsisato μόνο οι πέντε πρώτοι άξονες είναι στατιστικά σημαντικοί, ακόμη και σε επίπεδο σημαντικότητας α=0,003. Στο ίδιο συμπέρασμα καταλήξαμε και προηγουμένως, εφαρμόζοντας το προτεινόμενο κριτήριο, αλλά με πολύ λιγότερους αριθμητικούς υπολογισμούς.

Πίνακας 6.8: Αποτελέσματα Ελέγχων Σημαντικότητας των Αξόνων: Μέθοδος Nishisato

Άξονας	Αδράνεια Άξονα	Στατιστικό Ελέγχου	β.ε.	Κρίσιμη Τιμή χ^2 σε α=0,05	Κρίσιμη Τιμή χ^2 σε α=0,003	*p*	*Post hoc Ισχύς*
1	0,325	762,699*	315	357,391	389,064	0,0000	1
2	0,293	672,819*	313	355,260	386,843	0,0000	1
3	0,290	664,602*	311	353,128	384,621	0,0000	1
4	0,287	656,420*	309	350,995	382,398	0,0000	1
5	0,269	608,040*	307	348,863	380,175	0,0000	1
6	0,156	329,114	305	346,730	377,951	0,1639	1
7	0,155	326,816	303	344,596	375,726	0,1660	1
8	0,151	317,652	301	342,462	373,500	0,2440	1
9	0,139	290,417	299	340,328	371,274	0,6282	1
10	0,134	279,180	297	338,193	369,047	0,7639	1
11	0,129	268,009	295	336,058	366,819	0,8685	1
12	0,125	259,118	293	333,922	364,590	0,9236	1
13	0,119	245,857	291	331,786	362,361	0,9744	1
14	0,000	0,000	289	329,649	360,130	1,0000	
15	0,000	0,000	287	327,512	357,899	1,0000	
16	0,000	0,000	285	325,374	355,667	1,0000	
17	0,000	0,000	283	323,236	353,435	1,0000	
18	0,000	0,000	281	321,097	351,201	1,0000	

* Στατιστικά σημαντικό σε ε.σ. α=0,003.

6.5 Σχόλια και Συμπεράσματα Κεφαλαίου

Το βασικό συμπέρασμα, το οποίο προκύπτει από όσα αναφέρθηκαν στο κεφάλαιο αυτό, είναι ότι στην περίπτωση που τα διαθέσιμα δεδομένα έχουν συγκεντρωθεί με τυχαία δειγματοληψία, η ΠΑΑ μπορεί να συνδυαστεί με μεθόδους της Επαγωγικής Στατιστικής. Στις μεθοδολογικές προσεγγίσεις που παρουσιάσαμε, η μόνη βασική απαίτηση είναι το μέγεθος του δείγματος να είναι αρκούντως μεγάλο, κάτι που συνήθως ισχύει στις περιπτώσεις όπου εφαρμόζεται η ΠΑΑ. Φαίνεται ότι το επιστημολογικό "χάσμα", που κατά τον Gras (1995), χωρίζει την Ανάλυση Δεδομένων με την Επαγωγική Στατιστική τελικά δεν είναι και τόσο μεγάλο.

Στο πλαίσιο της ΠΑΑ, οι έλεγχοι στατιστικής σημαντικότητας συνδέονται με την εξωτερική εγκυρότητα των αριθμητικών και διαγραμματικών εκροών της. Στο χώρο της Ανάλυσης Δεδομένων, και ιδιαίτερα στην Ολλανδική Σχολή, το πρόβλημα αντιμετωπίζεται κυρίως με τη χρήση τεχνικών επαναδειγματοληψίας, όπως είναι η *Bootstrap* (Van de Geer 1993β, Markus 1994α και 1994β, Gifi 1996, Michailidis & De Leeuw 1998). Η εξωτερική εγκυρότητα αναφέρεται στην περίπτωση κατά την

οποία τα διαθέσιμα δεδομένα έχουν συγκεντρωθεί με μεθόδους της τυχαίας δειγματοληψίας, από τον υπό εξέταση πληθυσμό, και εστιάζεται στον έλεγχο του κατά πόσο συνεπή (σταθερά) είναι τα παραγόμενα αποτελέσματα, στη θεωρητική περίπτωση που η ανάλυση επαναληφθεί σε άλλα τυχαία δείγματα από τον ίδιο πληθυσμό. Κατά τον έλεγχο της εξωτερικής εγκυρότητας, βασικός στόχος είναι η εκτίμηση παραμέτρων και η κατασκευή διαστημάτων εμπιστοσύνης που επιτρέπουν τη γενίκευση των συμπερασμάτων στον πληθυσμό, από τον οποίο προέρχεται το δείγμα. Οι μικρές τιμές των τυπικών σφαλμάτων στην εκτίμηση των παραμέτρων, οι μικρές επιφάνειες των περιοχών εμπιστοσύνης και η στατιστική σημαντικότητα των δεικτών εξασφαλίζουν την εξωτερική εγκυρότητα των αποτελεσμάτων. Πάντως, σε αυτές τις προσεγγίσεις θα πρέπει να λαμβάνονται υπόψη και μερικά από τα αναπόφευκτα προβλήματα των «πολλαπλών συγκρίσεων» (βλέπε Ενότητα Ε3.2 του Παραρτήματος Ε). Να σημειώσουμε ότι, γενικά, ο έλεγχος της εξωτερικής σταθερότητας δεν περιλαμβάνει μόνο στατιστικούς ελέγχους ή την κατασκευή διαστημάτων εμπιστοσύνης, αλλά και τη χρήση διαθέσιμης εξωτερικής πληροφορίας με τη μορφή “μετα-δεδομένων” (τα οποία συνήθως εισάγονται στην ΠΑΑ ως συμπληρωματικά στοιχεία)[28] (Lebart, 2005) καθώς και ελέγχους που περιλαμβάνουν διαδικασίες διασταυρούμενης εγκυρότητας (*cross validation*)[29] (Aldenderfer & Blashfield 1984, Hair *et al*. 1995, Jackson 1991, Lebart 2005).

Όταν τα δεδομένα δεν αποτελούν τυχαίο δείγμα ή, εν γένει, δεν μπορούν να εφαρμοστούν έλεγχοι και διαδικασίες της Επαγωγικής Στατιστικής, τότε προκύπτει το πρόβλημα της «εσωτερικής» εγκυρότητας ή σταθερότητας των δομών και των σχέσεων που αναδεικνύονται μέσω της ΠΑΑ (Greenacre 1984, Gifi 1996, Michailidis & De Leeuw 1998, Lebart 2005). Στην περίπτωση αυτή, το ενδιαφέρον εστιάζεται μόνο στα διαθέσιμα δεδομένα και στο κατά πόσο τα αποτελέσματα αποτελούν μια καλή σύνοψη της πληροφορίας που ενθυλακώνει ο αρχικός πίνακας δεδομένων. Τα παραγόμενα αποτελέσματα χαρακτηρίζονται από εσωτερική σταθερότητα αν μικρές ή/και ανεπαίσθητες μεταβολές ή, καλύτερα, διαταραχές των αρχικών δεδομένων

---

[28] Στην περίπτωση αυτή το ενδιαφέρον εστιάζεται στον έλεγχο των δομών και των σχέσεων, όπως αυτές διαμορφώνονται λαμβάνοντας υπόψη δεδομένα, τα οποία δεν συμμετείχαν στην αρχική ανάλυση. Ο έλεγχος συνίσταται στην εξέταση της σταθερότητας ή όχι των αρχικών συμπερασμάτων.

[29] Σε γενικές γραμμές η μέθοδος εφαρμόζεται ως εξής: Το αρχικό δείγμα χωρίζεται τυχαία σε δύο μέρη. Η στατιστική μέθοδος εφαρμόζεται και στα δύο υποσύνολα και τα αποτελέσματα συγκρίνονται ως προς τη σύγκλισή τους.

εισόδου έχουν ως αποτέλεσμα μόνο μικρές ή/και ανεπαίσθητες αλλαγές στην έξοδο της μεθόδου. Για παράδειγμα, τέτοιου είδους διαταραχές είναι δυνατό να προκαλέσουν τα παράτυπα σημεία (*outliers*), η συνένωση κλάσεων των μεταβλητών και η απομάκρυνση μεταβλητών ή αντικειμένων από την ανάλυση. Στις συνέπειές τους έχουμε αναφερθεί στο Κεφάλαιο 2. Για τον έλεγχο της εσωτερικής εγκυρότητας η εφαρμογή των μεθόδων *Bootstrap* και *Jackknife* αποτελούν μάλλον "μονόδρομο". Ενδεικτικά αναφέρουμε δύο βασικές τεχνικές που εφαρμόζονται στο πλαίσιο της ΠΑΑ (Greenacre 2006 και 1984, Lebart 2005): α) Η «μερική» *Bootstrap*, στην οποία τα επαναληπτικά δείγματα εισάγονται ως συμπληρωματικά στοιχεία στην ανάλυση των αρχικών δεδομένων και β) Η «ολική» *Bootstrap*, σύμφωνα με την οποία κάθε επανάληψη οδηγεί σε μια ξεχωριστή Απλή ή Πολλαπλή Ανάλυση των Αντιστοιχιών. Όμως, όπως παρατηρήσαμε στην Ενότητα 6.3.1, οι μέθοδοι επαναδειγματοληψίας, όπως αυτές εφαρμόζονται στην ΠΑΑ για τον έλεγχο της εξωτερικής εγκυρότητας, εγείρουν αρκετές επιφυλάξεις σχετικά με τη διαδικασία υλοποίησής τους και τις αποφάσεις που θα πρέπει να ληφθούν από τους χρήστες (βλέπε και Lebart, 2005). Οι συγκεκριμένες μέθοδοι μάλλον εισάγουν περισσότερα προβλήματα απ' ότι επιλύουν. Για το Lebart[30] είναι προτιμότερο η *Bootstrap* να χρησιμοποιείται μόνο για τον έλεγχο της εσωτερικής εγκυρότητας.

Οι προτεινόμενες μεθοδολογίες κατασκευής ελλείψεων εμπιστοσύνης γύρω από τα σημεία γραμμών ή/και στηλών, επί των παραγοντικών επιπέδων που παράγονται από την ΠΑΑ, παρέχουν όχι μόνο στατιστικό, αλλά και οπτικό έλεγχο της εξωτερικής εγκυρότητας - σταθερότητας και της σημαντικότητας των προβολών των αντίστοιχων σημείων επί των παραγοντικών επιπέδων και αξόνων. Οι μέθοδοι μπορούν να επεκταθούν και στις τρεις διαστάσεις με την κατασκευή ελλειψοειδών εμπιστοσύνης. Συνήθως, όμως, τα γραφικά αποτελέσματα της ΠΑΑ παρουσιάζονται σε δύο διαστάσεις. Είναι σημαντικό να τονίσουμε ότι η απόφαση των αναλυτών τόσο σε σχέση με τον αριθμό των διαστάσεων του χώρου (2 ή 3) όσο και σε σχέση με το σε ποιους υποχώρους θα προβληθούν τα σημεία (π.χ. παραγοντικό επίπεδο 1×2 ή/και 1×3, κ.λπ.) θα πρέπει να ληφθεί *a priori*, ώστε η σημαντικότητα της θέσης των σημείων και οι συγκρίσεις των αντίστοιχων προφίλ, μέσω των ελλείψεων

[30] Μετά από προσωπική επικοινωνία στο 3ο Πανελλήνιο Συνέδριο Ανάλυσης Δεδομένων, με Διεθνή Συμμετοχή, Σιθωνία, Χαλκιδικής, 15-18/09, 2005.

εμπιστοσύνης, να εντάσσονται στο μεθοδολογικό πλαίσιο της Επαγωγικής Στατιστικής. Αντίθετα, η μέθοδος του *Gabriel* μπορεί να εφαρμοστεί και εκ των υστέρων, δηλαδή μετά την παρατήρηση των αποτελεσμάτων της ΠΑΑ και των ελλείψεων. Οι μέθοδοι μπορούν να εφαρμοστούν και σε πειραματική μελέτη, όπου υπάρχει διάκριση μεταξύ εξαρτημένης και ανεξάρτητης μεταβλητής. Στο παράδειγμα που παραθέσαμε στην Ενότητα 6.3.5, η μεταβλητή *X* θα μπορούσε να θεωρηθεί ως ένας ποιοτικός ή ποσοτικός παράγοντας με 5 επίπεδα (αγωγές), ενώ η *Y* ως μεταβλητή απόκρισης με 6 διακεκριμένες κατηγορικές τιμές.

Το ερώτημα που τίθεται είναι ποια από τις δύο προσεγγίσεις είναι προτιμότερη. Η απάντηση στο ερώτημα δεν είναι απλή. Στην πρώτη προσέγγιση, οι προβολές των σημείων γραμμών (στηλών), επί των παραγοντικών επιπέδων, θεωρούνται ως τυχαίες μεταβλητές και μέσω της μεθόδου *Δέλτα* υπολογίζονται για κάθε σημείο τα στοιχεία του πίνακα **S** των δειγματικών διασπορών – συνδυασπορών. Η μέθοδος *Δέλτα* στηρίζεται στην υπόθεση ότι οι συχνότητες στα κελιά του πίνακα συμπτώσεων ακολουθούν Πολυωνυμική Κατανομή, η οποία μπορεί να προσεγγιστεί από την Κανονική, και η ιδιότητα αυτή χρησιμοποιείται, στη συνέχεια, στον υπολογισμό των στοιχείων του πίνακα **S**. Στην περίπτωση αυτή, κάθε σημείο γραμμής (στήλης), ανεξάρτητα από τον υποχώρο που προβάλλεται (παραγοντικό επίπεδο 1×2, 2×3, κ.λπ.), είναι διαρκώς συνδεδεμένο με την αρχική του κατανομή. Στη δεύτερη προσέγγιση, και ειδικά για το παραγοντικό επίπεδο 1×2, το οποίο απεικονίζει στις δύο διαστάσεις τη βέλτιστη δυνατή θέση των σημείων γραμμών (στηλών), η θεώρηση είναι διαφορετική και στηρίζεται στις ιδιότητες βέλτιστης κλιμάκωσης της ΠΑΑ. Στην προσέγγιση αυτή, θεωρούμε ότι τα σημεία έχουν τοποθετηθεί – προβληθεί βέλτιστα στο παραγοντικό επίπεδο, αφού η ΠΑΑ αξιοποιεί όλη την πληροφορία για την κατανομή, τις δομές και τις σχέσεις που ενθυλακώνει ο αρχικός πίνακας συμπτώσεων. Στη συνέχεια, τα σημεία αποδεσμεύονται από την αρχική τους κατανομή και τα στοιχεία του πίνακα **S** υπολογίζονται με βάση την εμπειρική κατανομή των τυποποιημένων συντεταγμένων των προβολών των δειγματοληπτικών ή πειραματικών μονάδων επί των παραγοντικών αξόνων. Η απόφαση σχετικά με το ποια, τελικά, μέθοδος θα πρέπει να ακολουθηθεί είναι μάλλον θέμα επιστημολογικής προσέγγισης στη λογική και στη φιλοσοφία της ΠΑΑ. Πειραματισμοί με τυχαία σύνολα δεδομένων και εφαρμογή μεθόδων προσομοίωσης *Monte Carlo* θα

μπορούσαν να οδηγήσουν στην επιλογή μιας εκ των δύο προσεγγίσεων με βάση τη σύγκριση της “τοπικής” εγκυρότητάς τους.

Πάντως, οι δύο μέθοδοι κατασκευής ελλείψεων εμπιστοσύνης (Ενότητα 6.3) μαζί με το διάστημα εμπιστοσύνης για τον έλεγχο της στατιστικής σημαντικότητας των παραγοντικών αξόνων (Ενότητα 6.4) αποτελούν εναλλακτικές προτάσεις απέναντι στη χρήση της τεχνικής *Bootstrap*. Να σημειώσουμε ότι οι ελλείψεις εμπιστοσύνης και η μέθοδος του *Gabriel* μπορούν να χρησιμοποιηθούν και για τον έλεγχο της ομοιογένειας των προφίλ των ομάδων σημείων, οι οποίες προκύπτουν από την εφαρμογή μεθόδων Ταξινόμησης στις γραμμές (στήλες) του πίνακα συμπτώσεων δύο μεταβλητών.

Αρκετές από τις μεθόδους ελέγχου της στατιστικής σημαντικότητας των αποτελεσμάτων που παράγονται από την εφαρμογή της ΠΑΑ υλοποιούνται μέσω του λογισμικού CHIC Analysis (βλέπε Μάρκος, 2006).

ΚΕΦΑΛΑΙΟ 7

Πρόταση Μεθόδου Εφαρμογής της Παραγοντικής Ανάλυσης των Αντιστοιχιών σε Πειραματικούς Σχεδιασμούς

7.1 Εισαγωγή

Στην Ενότητα 1.6 τονίστηκε ότι η ΠΑΑ δεν βρήκε τη θέση που της αρμόζει στην ανάλυση δεδομένων, τα οποία προέρχονται από πειραματικούς σχεδιασμούς, όπως αυτοί ορίστηκαν στην Ενότητα 1.2. Αυτό οφείλεται στους παρακάτω λόγους:

- Στην ΠΑΑ οι μεταβλητές αντιμετωπίζονται συμμετρικά, χωρίς διάκριση σε εξαρτημένες και ανεξάρτητες (βλέπε Ενότητα 1.3). Το χαρακτηριστικό αυτό καθιστά *a priori* απαγορευτική την εφαρμογή της μεθόδου σε πειραματικές διατάξεις, όπου η λογική του πειραματισμού στηρίζεται ακριβώς στο διαχωρισμό αυτό (βλέπε Ενότητες 1.2 και 1.6). Στα πειράματα εκ κατασκευής υπεισέρχονται τουλάχιστον δύο ομάδες μεταβλητών. Η ομάδα των ανεξάρτητων και η ομάδα των εξαρτημένων. Οι ανεξάρτητες μεταβλητές θεωρητικά βρίσκονται κάτω από τον άμεσο έλεγχο του πειραματιστή, ο οποίος μετά από κατάλληλη μεταβολή της κατάστασης των ανεξάρτητων μελετά και μετρά την επίδραση των μεταβολών αυτών στις τιμές των εξαρτημένων. Κατά την πειραματική διαδικασία ελέγχονται ταυτόχρονα και άλλοι εξωγενείς παράγοντες, οι οποίοι αποτελούν πηγές ανεπιθύμητης μεταβλητότητας. Με τον πειραματισμό είναι δυνατό να τεκμηριωθούν σχέσεις αιτίας – αποτελέσματος, ενώ αυτό δεν είναι εφικτό με τις δειγματοληπτικές έρευνες επισκόπησης (βλέπε Ενότητα 1.2), στις οποίες εφαρμόζεται συχνά η ΠΑΑ για τη διερεύνηση μόνο της συσχέτισης μεταξύ των εξεταζόμενων μεταβλητών.

- Το Φιλοσοφικό πλαίσιο στο οποίο αναπτύχθηκε η ΠΑΑ (ιδιαίτερα στη Γαλλία) καθιστούσε απαγορευτική την εφαρμογή ελέγχων στατιστικής σημαντικότητας στα παραγόμενα αποτελέσματα (βλέπε Ενότητες 1.4.1 και 1.5.2). Αντίθετα, στα πειράματα η στατιστική σημαντικότητα ενός ευρήματος αποτελεί χρήσιμο οδηγό για τη λήψη αποφάσεων κάτω από συνθήκες σχετικής αβεβαιότητας (βλέπε Ενότητες 1.2 και 1.6). Οι έλεγχοι εφαρμόζονται με σκοπό να διαπιστωθεί αν οι παρατηρούμενες μεταβολές στις τιμές των εξαρτημένων μεταβλητών είναι αποτέλεσμα της συστηματικής επίδρασης των ανεξάρτητων ή μπορούν να αποδοθούν μόνο σε τυχαίους, αστάθμητους και μη ελεγχόμενους παράγοντες.

- Η ΠΑΑ αναπτύχθηκε για την ανάλυση κατηγορικών μεταβλητών. Κατά κανόνα, στους πειραματικούς σχεδιασμούς τουλάχιστον η μία από τις δύο εμπλεκόμενες ομάδες μεταβλητών (εξαρτημένες, ανεξάρτητες) αποτελείται από ποσοτικές μεταβλητές (βλέπε Ενότητα 1.6). Η στατιστική επεξεργασία των αντίστοιχων αποτελεσμάτων μπορεί να γίνει με μια πληθώρα μεθόδων, οι οποίες έχουν αναπτυχθεί στο χώρο της Επαγωγικής Στατιστικής. Ενδεικτικά αναφέρουμε την Ανάλυση Διακύμανσης (μονομεταβλητή, πολυμεταβλητή), τη Διακρίνουσα Ανάλυση, την Ανάλυση Παλινδρόμησης (απλή, πολλαπλή) και τη Λογιστική Παλινδρόμηση (βλέπε Hair *et al*. 1995, Sharma 1996).

- Η ΠΑΑ, παρόλο που ως μέθοδος ανάλυσης κατηγορικών δεδομένων εμφανίστηκε γύρω στα μέσα της δεκαετίας του ’30 (βλέπε Ενότητα 1.4), άργησε ωστόσο να γίνει γνωστή στην αγγλόφωνη επιστημονική κοινότητα, η οποία είχε ήδη “επενδύσει” στην ανάπτυξη άλλων μεθόδων για τη στατιστική επεξεργασία δεδομένων από πειραματικούς σχεδιασμούς (βλέπε Ενότητα 1.2). Η ΠΑΑ, όπως αυτή αναδείχθηκε από τον Benzécri, άρχισε να γίνεται γνωστή στις αγγλόφωνες χώρες στα μέσα της δεκαετίας του ’80 μετά τη δημοσίευση στην αγγλική γλώσσα σχετικών συγγραμμάτων κυρίως από τους Greenacre (1984) και Lebart, Morineau και Warwick (1984). Μέχρι τότε η προσήλωση των Γάλλων ερευνητών στη συγγραφή των εργασιών τους στη δική τους γλώσσα, σε συνδυασμό με τον ιδιόμορφο τρόπο της μαθηματικής παρουσίασης που υιοθέτησαν, καθιστούσαν τις μεθόδους εσωτερική τους υπόθεση και “κλειστές” στο μη γαλλόφωνο αναγνωστικό κοινό.

Στο χώρο της Ανάλυσης Δεδομένων οι μόνες συστηματικές προσπάθειες για την εφαρμογή παραλλαγών της ΠΑΑ σε σύνολα δεδομένων, στα οποία υπάρχει διάκριση μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών, είναι αυτές των Ιταλών ερευνητών (βλέπε Ενότητα 1.4.3) και του Nishisato (1994, 1990 και 1980). Στην Ιταλική Σχολή έχει αναπτυχθεί η Μη Συμμετρική Ανάλυση των Αντιστοιχιών (Lauro & Balbi 1999, Kroonenberg & Lombardo 1999), στην οποία όμως δεν χρησιμοποιείται η απόσταση χ^2 (βλέπε Επισήμανση Δ της Ενότητας 2.2.14.5). Η μέθοδος στηρίζεται στη διάσπαση του συντελεστή συνάφειας τ των Goodman και Kruskal (1954) και όχι στο συντελεστή συνάφειας μέσου τετραγώνου φ^2 του *Pearson*, ο οποίος, όπως δείχθηκε στην Ενότητα 2.2.5, είναι ίσος με την ολική αδράνεια του πίνακα συμπτώσεων που αναλύεται. Ο Nishisato στο πλαίσιο της Δυικής Κλιμάκωσης (*Dual Scaling*), η οποία παράγει συγκρίσιμα αποτελέσματα με αυτά της ΠΑΑ (Greenacre 1984, SAS Institute 1990, Gifi 1996), έχει προτείνει μεθοδολογία ανάλυσης κατηγορικών δεδομένων που προέρχονται από πειραματικές διατάξεις και τα αποτελέσματα συνοδεύονται, μάλιστα, από ελέγχους στατιστικής σημαντικότητας. Η Δυική Κλιμάκωση είναι καθαρά μια τεχνική βελτιστοποίησης, η οποία χαρακτηρίζεται από τις ιδιότητες που παρουσιάστηκαν στην Ενότητα 2.5. Η θεωρητική της τεκμηρίωση βασίζεται σε έννοιες της Ανάλυσης Διασποράς και δεν υπάρχει καμία σύνδεση ούτε με την απόσταση χ^2 ούτε με την ΠΑΑ, όπως αυτή εφαρμόζεται και ερμηνεύεται στη Γαλλική Σχολή. Η μέθοδος υλοποιείται μέσω επαναληπτικού αλγόριθμου και σκοπός της είναι η ανάθεση βαθμών στα αντικείμενα και βαρών στις κατηγορίες των μεταβλητών, ώστε να μεγιστοποιούνται, κατά περίπτωση, η κανονικοποιημένη συσχέτιση, ο δείκτης εσωτερικής συνέπειας α του *Cronbach* και η ομοιογένεια των μεταβλητών (βλέπε Ενότητα 2.3.4). Οι εκροές της είναι κυρίως αριθμητικές και η ερμηνεία των αποτελεσμάτων στηρίζεται στις σχετικές θέσεις των (τυποποιημένων) συντεταγμένων των αντικειμένων και των μεταβλητών κυρίως επί του πρώτου άξονα. Σε γενικές γραμμές, η Δυική Κλιμάκωση προσεγγίζει την Ολλανδική Σχολή και διέπεται από το πνεύμα της Ιαπωνικής παράδοσης σε ό,τι αφορά την ποσοτικοποίηση ποιοτικών δεδομένων (Van Rijckevorse & De Leeuw 1988, SAS Institute 1990, Greenacre & Blasius 1994, Nishisato 1994, 1990 και 1980). Θα πρέπει να σημειωθεί ότι ο διαχωρισμός των μεταβλητών σε εξαρτημένες και ανεξάρτητες δεν εμφανίζεται μόνο στους πειραματικούς σχεδιασμούς, αλλά και σε έρευνες επισκόπησης, όπου η διάκριση είναι

τουλάχιστον εννοιολογική. Σύμφωνα με τους Le Roux και Rouanet (2004), στις περισσότερες έρευνες η αναζήτηση δομικών σχέσεων μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών αποτελεί τον κανόνα και όχι την εξαίρεση. Οι ίδιοι ερευνητές προτείνουν τον όρο «*Structured Data Analysis*» (Δομημένη Ανάλυση Δεδομένων) για να διαφοροποιήσουν την επιβεβαιωτική από τη διερευνητική προσέγγιση στη στατιστική επεξεργασία των δεδομένων (βλέπε Ενότητα 1.6).

Εκτός από την Ιταλική, και η Ολλανδική Σχολή έχει να προτείνει γενικές μεθόδους, στις οποίες υπάρχει διάκριση μεταξύ εξαρτημένων και ανεξάρτητων κατηγορικών μεταβλητών. Ενδεικτικά αναφέρουμε την Κατηγορική Παλινδρόμηση και τη Μη Γραμμική Κανονικοποιημένη Συσχέτιση (Gifi 1996, SPSS Inc. 2004α). Στην Ενότητα 2.5 αναφέρθηκε ότι κατά καιρούς έχουν γίνει προσπάθειες ενοποίησης των μεθόδων βέλτιστης κλιμάκωσης με μεθόδους που στηρίζονται στα Γενικά Γραμμικά Υποδείγματα, όπως είναι η Γραμμική Παλινδρόμηση, η Κανονικοποιημένη Συσχέτιση και η Ανάλυση Διακύμανσης (Perreault & Young 1980, Sands & Young 1980, Young 1981). Στο πλαίσιο της Επαγωγικής Στατιστικής[31], μοναδικές, ίσως, εξαιρέσεις αποτελούν: α) η Πολυωνυμική Λογιστική Παλινδρόμηση (*Multinomial Logistic Regression*), στην οποία τόσο η εξαρτημένη όσο και οι ανεξάρτητες μεταβλητές μπορούν να είναι κατηγορικές (Fienberg 1991, Zelen 1991, Liao 1994, SAS Institute 1999 και 1990, Agresti 2002, SPSS Inc. 2004β, Kutner *et al*. 2005) και β) ορισμένες επεκτάσεις της Ανάλυσης Διακύμανσης σε κατηγορικά δεδομένα (Light & Margolin 1971, Margolin & Light 1974, Landis & Koch 1977, Beitler & Landis 1985). Οι διαδικασίες που προαναφέρθηκαν μπορούν να εφαρμοστούν τόσο σε πειραματικά όσο και σε δειγματοληπτικά δεδομένα.

Στην Ενότητα 2.4.3 έγινε αναφορά σε εργασίες στις οποίες έχουν προταθεί παραλλαγές της ΠΑΑ, οι οποίες λαμβάνουν υπόψη την ενδεχόμενη επίδραση του δειγματοληπτικού σχεδίου στα αποτελέσματα. Κάτω από συνθήκες, τα πορίσματα των εργασιών αυτών θα μπορούσαν να αξιοποιηθούν στην ανάλυση πειραματικών σχεδιασμών, αν στα αντίστοιχα δειγματοληπτικά σχήματα αντιστοιχηθούν γνωστοί πειραματικοί σχεδιασμοί. Οι προτεινόμενες μέθοδοι αποτελούν μεμονωμένες

[31] Για αναλυτική παρουσίαση των μεθόδων στατιστικής ανάλυσης κατηγορικών δεδομένων παραπέμπουμε στους Fienberg (1991), Bishop, Fienberg & Holland (1991) και Agresti (2002 και 1984).

προσπάθειες, που έχουν ως στόχο την επίλυση συγκεκριμένων προβλημάτων, και στις περισσότερες περιπτώσεις απαιτούν τροποποιήσεις της ΠΑΑ τόσο στο αλγοριθμικό όσο και στο ερμηνευτικό της μέρος.

Στο κεφάλαιο αυτό προτείνουμε μεθοδολογία για την εφαρμογή της ΠΑΑ, όπως αυτή εφαρμόζεται στη Γαλλική και Ολλανδική Σχολή Ανάλυσης Δεδομένων, σε τρεις βασικούς πειραματικούς σχεδιασμούς: α) το Πλήρως Τυχαιοποιημένο Σχέδιο με Ένα Παράγοντα, β) το Πλήρως Τυχαιοποιημένο Σχέδιο με Δύο Παράγοντες και γ) το Τυχαιοποιημένο Σχέδιο σε Πλήρη Συγκροτήματα (*blocks*) με Δύο Παράγοντες. Η μέθοδος συνδυάζει εφαρμογές και ιδιότητες των Πινάκων Σχεδιασμού (*Design Matrices*) και των Πινάκων Προβολής «καπέλο» (*Hat Matrices*) καθώς και την Αρχή της Ισοδυναμίας των Σχετικών Κατανομών (βλέπε Ενότητα 2.2.3). Στην περίπτωση του Τυχαιοποιημένου Σχεδίου σε Πλήρη Συγκροτήματα, η απαλοιφή της επίδρασης στη διαμόρφωση των αποτελεσμάτων, την οποία ενδεχομένως επιφέρει η ομαδοποίηση των πειραματικών μονάδων σε συγκροτήματα, επιτυγχάνεται με την εφαρμογή της Προκρούστιας Προβολής – Περιστροφής (Golub & Van Loan 1989, Mardia, Kent & Bibby 2003). Οι πίνακες σχεδιασμού και προβολής χρησιμοποιούνται ευρέως στη στατιστική επεξεργασία δειγματοληπτικών και πειραματικών δεδομένων μέσω των Γενικών Γραμμικών Μοντέλων (βλέπε Μπόρα-Σέντα & Μωϋσιάδης 1992, Καρακώστας 1993, Ζαχαροπούλου 1994, Kirk 1995, Stapleton 1995, Neter *et al*. 1996, Mendenhall & Sincich 1996, Kleinbaum *et al*. 1998, SAS Institute 1999 και 1990, Kuehl 2000, Rencher 2000, Rao 2002, SPSS Inc. 2004α και 1997, Srivastava 2002, Kutner *et al*. 2005). Σε κάθε πειραματικό σχεδιασμό αντιστοιχεί, εν γένει, και ένας διαφορετικός πίνακας σχεδιασμού. Στο γενικό πλαίσιο της προτεινόμενης προσέγγισης, αρχικά γίνεται διάκριση των μεταβλητών σε εξαρτημένες και ανεξάρτητες. Στη συνέχεια, οι πίνακες σχεδιασμού των εξαρτημένων μεταβλητών προβάλλονται (ορθογώνια) στο χώρο σχεδιασμού των ανεξάρτητων. Τέλος, η ΠΑΑ εφαρμόζεται είτε στον πίνακα προβολής, που περιγράφει το υπό εξέταση φαινόμενο στο χώρο των ανεξάρτητων μεταβλητών, είτε στον αντίστοιχο συμπτυγμένο πίνακα, ο οποίος προκύπτει από την εφαρμογή της Αρχής της Ισοδυναμίας των Σχετικών Κατανομών. Στην Ενότητα Z2 του Παραρτήματος Z δίνουμε ένα παράδειγμα εφαρμογής της προτεινόμενης μεθοδολογίας.

Στην ενότητα που ακολουθεί, δείχνουμε τον τρόπο με τον οποίο η ΠΑΑ μπορεί να συνδεθεί με Πίνακες Σχεδιασμού και Πίνακες Προβολής. Η σύνδεση αυτή αποτελεί τη βάση της προτεινόμενης μεθοδολογίας για την εφαρμογή της ΠΑΑ σε πειραματικούς σχεδιασμούς ή δειγματοληπτικές έρευνες, στις οποίες υπάρχει διάκριση μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών. Χωρίς περιορισμό της γενικότητας, θα παρουσιάσουμε την περίπτωση δύο κατηγορικών μεταβλητών, όπου η μία θεωρείται εξαρτημένη και ή άλλη ανεξάρτητη.

7.2 Σύνδεση της ΠΑΑ με Πίνακες Σχεδιασμού και Πίνακες Προβολής

Έστω X και Y δύο κατηγορικές μεταβλητές με k και l κλάσεις αντίστοιχα. Συμβολίζουμε με $f_i^X \left(i=1,\ldots,k\right)$ και $f_j^Y \left(j=1,\ldots,l\right)$ τα στοιχεία της περιθώριας κατανομής απολύτων συχνοτήτων των X και Y αντίστοιχα. Υποθέτουμε ότι τα διαθέσιμα δεδομένα, για N σε πλήθος αντικείμενα, είναι συγκεντρωμένα στον $N\times 2$ πίνακα **D** και ότι δεν υπάρχουν ελλείπουσες τιμές. Οι στήλες του **D** περιέχουν τις τιμές των δύο μεταβλητών για τα N αντικείμενα. Ας είναι $\mathbf{C_X}$ και $\mathbf{Z_Y}$ οι λογικοί πίνακες των μεταβλητών X και Y αντίστοιχα. Ο $\mathbf{C_X}$ είναι διαστάσεων $N\times k$, ενώ ο $\mathbf{Z_Y}$ είναι πίνακας $N\times l$. Στην Ενότητα 2.3.1 αναφέραμε ότι οι λογικοί πίνακες ονομάζονται και πίνακες σχεδιασμού (βλέπε και Mardia, Kent & Bibby, 2003). Πίνακες με λογική κωδικοποίηση 0-1 (*dummy* ή *indicator coding*) εμφανίζονται συχνά στις μαθηματικές αναπαραστάσεις των πινάκων σχεδιασμού, οι οποίοι χρησιμοποιούνται στα Γενικά Γραμμικά Μοντέλα[32] για την εκτίμηση των συντελεστών των κύριων επιδράσεων (*main effects*) και των αλληλεπιδράσεων (*interaction effects*) των ανεξάρτητων μεταβλητών (Raktoe & Federer 1973 και 1970, Elswick *et al*. 1991, Μπόρα-Σέντα & Μωϋσιάδης 1992, Καρακώστας 1993, Ζαχαροπούλου 1994, Kirk 1995, Mendenhall & Sincich 1996, Neter *et al*. 1996, Montgomery 1997, Kleinbaum *et al*. 1998, SAS Institute 1999 και 1990, Rencher 2000, Kuehl 2000, Mardia, Kent & Bibby 2003,

[32] Στους αριθμητικούς υπολογισμούς, οι πίνακες σχεδιασμού με λογική κωδικοποίηση 0-1 έχουν μία στήλη λιγότερη απ' ότι είναι οι κατηγορίες της αντίστοιχης μεταβλητής. Έτσι, αν η μεταβλητή έχει 4 κατηγορίες, τότε 3 στήλες είναι αρκετές για τη λογική περιγραφή των κατηγοριών της, αφού, για παράδειγμα, η τέταρτη κατηγορία μπορεί να αναπαρασταθεί με την τριάδα (0, 0, 0), δηλαδή τους κωδικούς που δηλώνουν ότι η συγκεκριμένη κατηγορία δεν είναι η 1 ούτε ή 2 ούτε και η 3.

Kutner *et al*. 2005). Η κωδικοποίηση αυτή εφαρμόζεται συνήθως όταν οι ανεξάρτητες μεταβλητές είναι κατηγορικές[33]. Πίνακες σχεδιασμού με κωδικοποίηση 0-1 χρησιμοποιούνται για την αριθμητική αναπαράσταση κατηγορικών μεταβλητών στην Ανάλυση Κανονικοποιημένης Συσχέτισης (Greenacre 1984, Mardia, Kent & Bibby 2003), στο Γενικό Σύστημα Συσχετιστικής Ανάλυσης (Cohen & Cohen, 1983) και στο μεθοδολογικό πλαίσιο των Λογαριθμογραμμικών Υποδειγμάτων (Lipsitz, Laird & Harrington 1990, Von Eye & Spiel 1996).

Σύμφωνα με τα παραπάνω, οι λογικοί πίνακες $\mathbf{C_X}$ και $\mathbf{Z_Y}$ αντιστοιχούν στους πίνακες σχεδιασμού των μεταβλητών *X* και *Y*. Χωρίς περιορισμό της γενικότητας, μπορούμε να θεωρήσουμε τη μεταβλητή *X* ως ανεξάρτητη και την *Y* ως εξαρτημένη. Η βασική ιδέα της προτεινόμενης μεθοδολογίας στηρίζεται στην ορθογώνια προβολή των στοιχείων του λογικού πίνακα $\mathbf{Z_Y}$, της εξαρτημένης μεταβλητής, στο χώρο που ορίζεται από τις στήλες του πίνακα σχεδιασμού $\mathbf{C_X}$, της ανεξάρτητης.

Η τεχνική της προβολής των εξαρτημένων μεταβλητών στο χώρο των ανεξάρτητων χρησιμοποιείται και στους σχεδιασμούς επιφανειών απόκρισης (*response surface designs*) (Montgomery 1997, Kuehl 2000) για τον εντοπισμό της περιοχής του πειραματισμού στην οποία επιτυγχάνεται η βελτιστοποίηση (μεγιστοποίηση ή ελαχιστοποίηση) της εξαρτημένης μεταβλητής. Έχει εφαρμοστεί σε τροποποιήσεις της Ανάλυσης σε Κύριες Συνιστώσες, για να ληφθεί υπόψη διαθέσιμη εξωτερική πληροφορία[34] (Takane & Shibayama, 1991), και της ΠΑΑ, με σκοπό την επιβολή γραμμικών περιορισμών στις ποσοτικοποιήσεις των κατηγοριών των μεταβλητών (Takane, Yanai & Mayekawa 1991, Bφckenholt & Takane 1994). Επίσης, οι Lebart, Morineau και Warwick (1984) προτείνουν οι κλάσεις των εξαρτημένων μεταβλητών να προβάλλονται ως συμπληρωματικά σημεία στα παραγοντικά επίπεδα, τα οποία προκύπτουν από την εφαρμογή της ΠΑΑ στις ανεξάρτητες. Με αυτόν τον τρόπο τα διαγραμματικά αποτελέσματα της ΠΑΑ καθίστανται «προβλεπτικοί χάρτες» (*predictive maps*) και η όλη διαδικασία μπορεί να χαρακτηριστεί ως μια διερευνητική "οπτική" παλινδρόμηση, η οποία αποσκοπεί στην ερμηνεία της επίδρασης των

---

[33] Για περισσότερες πληροφορίες σχετικά με τους τρόπους κωδικοποίησης κατηγορικών δεδομένων παραπέμπουμε στους Cohen και Cohen (1983), Kirk (1995) και SPSS Inc. (2004β).
[34] Συνήθως δημογραφικά στοιχεία.

ανεξάρτητων μεταβλητών στις εξαρτημένες. Όμως, η μέθοδος αυτή δεν μπορεί να εφαρμοστεί σε ισορροπημένους πειραματικούς σχεδιασμούς[35] γιατί οι ανεξάρτητες μεταβλητές είναι ασυσχέτιστες με αποτέλεσμα να μην είναι δυνατό να εφαρμοστεί η ΠΑΑ. Η αδράνεια των αντίστοιχων πινάκων συμπτώσεων είναι ίση με μηδέν. Την αντίστροφη διαδικασία, δηλαδή την προβολή των ανεξάρτητων μεταβλητών στο χώρο των εξαρτημένων, συναντάμε σε πολλές περιπτώσεις δειγματοληπτικών ερευνών επισκόπησης. Κατά κανόνα, οι δημογραφικές μεταβλητές θεωρούνται ως ανεξάρτητες μεταβλητές. Συχνά, κατά την εφαρμογή της ΠΑΑ, οι κατηγορίες των μεταβλητών αυτών εισάγονται ως συμπληρωματικά σημεία, ώστε οι μεταξύ τους συσχετίσεις να μην επηρεάσουν τις συσχετίσεις των εξαρτημένων μεταβλητών (βλέπε Ενότητα 2.2.10). Στην προσέγγιση αυτή, το ενδιαφέρον εστιάζεται κατά πρώτο λόγο στη διερεύνηση των δομών των σχέσεων των εξαρτημένων και, κατά δεύτερο, στη σχέση των δομών αυτών με τις ανεξάρτητες μεταβλητές. Στο πλαίσιο της Δυικής Κλιμάκωσης, ο Nishisato (1994, 1990 και 1980) προτείνει τρεις τρόπους με τους οποίους μπορεί να μελετηθεί η σχέση μεταξύ ενός συνόλου εξαρτημένων και ενός συνόλου ανεξάρτητων κατηγορικών μεταβλητών. Και οι τρεις προσεγγίσεις στηρίζονται στην ορθογώνια προβολή των πινάκων σχεδιασμού με κωδικοποίηση 0-1 των εξαρτημένων μεταβλητών στο χώρο που ορίζουν οι ανεξάρτητες. Αυτό επιτυγχάνεται με τη χρήση κατάλληλων πινάκων προβολής, ανάλογα με την προσέγγιση. Στην περίπτωση πειραματικών σχεδιασμών, οι λογικοί πίνακες των εξαρτημένων μεταβλητών προβάλλονται σε χώρο που ορίζουν οι στήλες του αντίστοιχου πίνακα σχεδιασμού. Ο χώρος προβολής επιλέγεται ανάλογα με τις επιδράσεις (*effects*) που είναι επιθυμητό να μεγιστοποιηθούν. Για παράδειγμα, αν το πειραματικό σχέδιο είναι πλήρως τυχαιοποιημένο με δύο παράγοντες (βλέπε Κάτος, 1986) και ο πειραματιστής επιθυμεί τη μεγιστοποίηση της αλληλεπίδρασης των δύο παραγόντων, τότε η προβολή των εξαρτημένων μεταβλητών γίνεται στο χώρο που ορίζεται από τις στήλες του πίνακα σχεδιασμού που αντιστοιχεί στην αλληλεπίδραση των δύο παραγόντων. Στη συνέχεια, η μέθοδος της Δυικής Κλιμάκωσης εφαρμόζεται στον πίνακα που περιγράφει τις εξαρτημένες μεταβλητές στο χώρο προβολής. Στην περίπτωση που εξετάζουμε, ένας τρόπος για να προβάλουμε τον πίνακα $\mathbf{Z}_{\mathbf{Y}}$ στον $\mathbf{C}_{\mathbf{X}}$

[35] Είναι οι σχεδιασμοί όπου κάθε αγωγή ή συνδυασμός αγωγών περιλαμβάνει τον ίδιο αριθμό πειραματικών μονάδων.

είναι μέσω της παρακάτω σχέσης (Nishisato 1980, Harville 1997, Rencher 2000, Kuehl 2000, Srivastava 2002):

$$\mathbf{P}_{\text{proj}} = \mathbf{C}_{\mathbf{X}} \left(\mathbf{C}_{\mathbf{X}}^{\mathbf{T}} \mathbf{C}_{\mathbf{X}} \right)^{-1} \mathbf{C}_{\mathbf{X}}^{\mathbf{T}} \mathbf{Z}_{\mathbf{Y}} . \qquad [7.1]$$

Ο πίνακας $\mathbf{P}_{\text{proj}}$ αποτελεί την προβολή του $\mathbf{Z}_{\mathbf{Y}}$ στο χώρο στηλών του $\mathbf{C}_{\mathbf{X}}$, είναι διαστάσεων $N \times l$ και, όπως θα δείξουμε στη συνέχεια, περιγράφει τα προφίλ των N δειγματοληπτικών ή πειραματικών μονάδων ως προς τη μεταβλητή Y στο χώρο του $\mathbf{C}_{\mathbf{X}}$. Ο πίνακας $\mathbf{C}_{\mathbf{X}} \left(\mathbf{C}_{\mathbf{X}}^{\mathbf{T}} \mathbf{C}_{\mathbf{X}} \right)^{-1} \mathbf{C}_{\mathbf{X}}^{\mathbf{T}}$ είναι αντίστοιχος με τον "πίνακα καπέλο" (*hat matrix*) $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}$, ο οποίος χρησιμοποιείται στη Γραμμική Παλινδρόμηση για την εκτίμηση των τιμών της εξαρτημένης μεταβλητής, μετά την προσαρμογή των δεδομένων στο γραμμικό υπόδειγμα $\mathbf{Y}=\mathbf{Xb}+\mathbf{e}$, και τον εντοπισμό ασυνήθιστων σημείων (παράτυπα, μόχλευσης) του σχεδιασμού (Fox 1991, Ζαχαροπούλου 1994, Mendenhall & Sincich 1996, Rencher 2000, Kuehl 2000, Kutner *et al*. 2005). Οι εκτιμώμενες τιμές $\hat{y}_i$ της εξαρτημένης μεταβλητής Y δεν είναι παρά οι ορθογώνιες προβολές των παρατηρούμενων y_i στο χώρο που ορίζεται από τον πίνακα σχεδιασμού $\mathbf{X}$, δηλαδή $\hat{\mathbf{Y}} = \mathbf{HY}$. Με άλλα λόγια, ο πίνακας $\mathbf{H}$ μετασχηματίζει το διάνυσμα παρατηρήσεων στο διάνυσμα προβλέψεων. Για το λόγο αυτό, συχνά αποκαλείται και «πίνακας πρόβλεψης» (*prediction matrix*).

Στη συνέχεια, θα εξετάσουμε τη μορφή των πινάκων που εμφανίζονται στο δεξιό σκέλος της [7.1]. Χωρίς περιορισμό της γενικότητας, υποθέτουμε ότι: α) η διάταξη των τιμών (κλάσεων) των μεταβλητών X και Y είναι $1, 2, \ldots, k$ και $1, 2, \ldots, l$ αντίστοιχα και β) τα στοιχεία του πίνακα $\mathbf{D}$ έχουν ταξινομηθεί σε αύξουσα σειρά ως προς τις τιμές της μεταβλητής X. Αυτό έχει ως αποτέλεσμα την αντιμετάθεση των στοιχείων του πίνακα $\mathbf{C}_{\mathbf{X}}$ με τρόπο ώστε στην πρώτη στήλη του να εμφανίζονται συνεχόμενα οι f_1^X σε πλήθος μονάδες που αντιστοιχούν στην πρώτη κλάση της μεταβλητής X, στη δεύτερη στήλη οι f_2^X μονάδες της δεύτερης κλάσης της X κ.ο.κ. Η ταξινομημένη μορφή του πίνακα $\mathbf{C}_{\mathbf{X}}$ είναι η εξής:

$$\mathbf{C_X} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Το άθροισμα των στοιχείων της στήλης *i* του $\mathbf{C_X}$ είναι ίσο με f_i^X .

Έστω **F** ο $k \times l$ πίνακας συμπτώσεων απολύτων συχνοτήτων των μεταβλητών *X* και *Y* (Πίνακας 7.1). Είναι γνωστό ότι $\mathbf{F} = \mathbf{C_X^T Z_Y}$ (Greenacre 1984, Lebart, Morineau & Warwick 1984, Gifi 1996).

Πίνακας 7.1: Πίνακας Συμπτώσεων Απολύτων Συχνοτήτων των *X* και *Y*

	Κλάσεις της *Y*				
Κλάσεις της *X*	1	2	…	*l*	Σύνολα
1	f_{11}	f_{12}	…	f_{1l}	f_1^X
2	f_{21}	f_{22}	…	f_{2l}	f_2^X
⋮	⋮	⋮	⋮	⋮	⋮
k	f_{k1}	f_{k2}	…	f_{kl}	f_k^X
Σύνολα	f_1^Y	f_2^Y	…	f_l^Y	*N*

Ο πίνακας $\mathbf{C_X^T C_X}$ είναι διαγώνιος, διαστάσεων $k \times k$ με διαγώνια στοιχεία ίσα με τις συχνότητες των κλάσεων της *X* (περιθώρια κατανομή της *X*). Ειδικότερα:

$$\mathbf{C_X^T C_X} = \begin{bmatrix} 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \Rightarrow$$

$$\Rightarrow \mathbf{C_X^T C_X} = \begin{bmatrix} f_1^X & 0 & \cdots & 0 \\ 0 & f_2^X & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & f_k^X \end{bmatrix}.$$

Η μορφή του πίνακα $\mathbf{C_X^T C_X}$ προκύπτει ως εξής:

Αν c_{ni} είναι το γενικό στοιχείο του $\mathbf{C_X}$, με $n = 1,\ldots,N$ και $i = 1,\ldots,k$, τότε:

$$c_{ni} = \begin{cases} 1 & \text{αν το αντικείμενο } n \text{ ανήκει στην κλάση } i \\ 0 & \text{αν το αντικείμενο } n \text{ δεν ανήκει στην κλάση } i \end{cases}.$$

Επίσης ισχύει (βλέπε Ενότητα 2.3.1):

$$\sum_{n=1}^{N} c_{ni} = f_i^X,\ \forall\, i = 1,\ldots,k\,.$$

Αφού τα στοιχεία του $\mathbf{C_X}$ είναι 0 ή 1 προκύπτει ότι:

$$\sum_{n=1}^{N} c_{ni} = \sum_{n=1}^{N} c_{ni}^2 = f_i^X,\ \forall\, i = 1,\ldots,k\,. \qquad [7.2]$$

Εφόσον οι στήλες του πίνακα $\mathbf{C_X}$ είναι συμπληρωματικές, με την έννοια ότι αν στη γραμμή i η μία στήλη έχει την τιμή 1 οι άλλες θα έχουν υποχρεωτικά 0, προκύπτει ότι

το εσωτερικό γινόμενο των διανυσμάτων $\mathbf{c_j}$ και $\mathbf{c_s}$ των στηλών *i* και *s* αντίστοιχα, με *i*≠*s*, είναι ίσο με 0. Συνεπώς, ισχύει:

$$\mathbf{c_i^T c_s} = \sum_{n=1}^{N} c_{in} c_{ns} = 0 \text{, για } i, s = 1, \ldots, k \text{ και } i \neq s. \qquad [7.3]$$

Επομένως, λόγω της [7.3] έχουμε ότι:

$$\mathbf{C_X^T C_X} = \begin{bmatrix} \sum_{n=1}^{N} c_{1n}c_{n1} & \sum_{n=1}^{N} c_{1n}c_{n2} & \cdots & \sum_{n=1}^{N} c_{1n}c_{nk} \\ \sum_{n=1}^{N} c_{2n}c_{n1} & \sum_{n=1}^{N} c_{2n}c_{n2} & \cdots & \sum_{n=1}^{N} c_{2n}c_{nk} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{n=1}^{N} c_{kn}c_{n1} & \sum_{n=1}^{N} c_{kn}c_{n2} & \cdots & \sum_{n=1}^{N} c_{kn}c_{nk} \end{bmatrix} = \begin{bmatrix} \sum_{n=1}^{N} c_{n1}^2 & 0 & \cdots & 0 \\ 0 & \sum_{n=1}^{N} c_{n2}^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sum_{n=1}^{N} c_{nk}^2 \end{bmatrix}.$$

Τώρα, αν λάβουμε υπόψη τη σχέση [7.2] έχουμε τελικά:

$$\mathbf{C_X^T C_X} = \begin{bmatrix} f_1^X & 0 & \cdots & 0 \\ 0 & f_2^X & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & f_k^X \end{bmatrix}.$$

Ο πίνακας $\left(\mathbf{C_X^T C_X}\right)^{-1}$ είναι διαγώνιος, διαστάσεων *k*×*k*, με διαγώνια στοιχεία ίσα με τους αντίστροφους των στοιχείων του $\mathbf{C_X^T C_X}$. Ειδικότερα:

$$\left(\mathbf{C_X^T C_X}\right)^{-1} = \begin{bmatrix} \left(f_1^X\right)^{-1} & 0 & \cdots & 0 \\ 0 & \left(f_2^X\right)^{-1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \left(f_k^X\right)^{-1} \end{bmatrix}.$$

Να παρατηρήσουμε ότι ο $\left(\mathbf{C_X^T C_X}\right)^{-1}$ ορίζεται πάντα, αφού $\forall\, i = 1, \ldots, k,\ f_i^X > 0$.

Ο *N*×*k* πίνακας $\mathbf{C_X}\left(\mathbf{C_X^T C_X}\right)^{-1}$ έχει τη μορφή:

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \left(f_1^{\,X}\right)^{-1} & 0 & \cdots & 0 \\ 0 & \left(f_2^{\,X}\right)^{-1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \left(f_k^{\,X}\right)^{-1} \end{bmatrix} = \begin{bmatrix} \left(f_1^{\,X}\right)^{-1} & 0 & \cdots & 0 \\ \left(f_1^{\,X}\right)^{-1} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \left(f_1^{\,X}\right)^{-1} & 0 & \cdots & 0 \\ 0 & \left(f_2^{\,X}\right)^{-1} & \cdots & 0 \\ 0 & \left(f_2^{\,X}\right)^{-1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \left(f_2^{\,X}\right)^{-1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \left(f_k^{\,X}\right)^{-1} \\ 0 & 0 & \cdots & \left(f_k^{\,X}\right)^{-1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \left(f_k^{\,X}\right)^{-1} \end{bmatrix}.$$

Στον παραπάνω πίνακα το στοιχείο $\left(f_i^{\,X}\right)^{-1}$ εμφανίζεται $f_i^{\,X}$ φορές, με $i = 1,\ldots,k$.

Ο πίνακας προβολής $\mathbf{C_X}\left(\mathbf{C_X^T C_X}\right)^{-1}\mathbf{C_X^T}$ είναι διαστάσεων $N \times N$ και είναι της μορφής:

$$\begin{bmatrix} \left(f_1^{\,X}\right)^{-1} & 0 & \cdots & 0 \\ \left(f_1^{\,X}\right)^{-1} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \left(f_1^{\,X}\right)^{-1} & 0 & \cdots & 0 \\ 0 & \left(f_2^{\,X}\right)^{-1} & \cdots & 0 \\ 0 & \left(f_2^{\,X}\right)^{-1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \left(f_2^{\,X}\right)^{-1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \left(f_k^{\,X}\right)^{-1} \\ 0 & 0 & \cdots & \left(f_k^{\,X}\right)^{-1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \left(f_k^{\,X}\right)^{-1} \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 1 & 1 & \cdots & 1 \end{bmatrix} =$$

$$= \begin{bmatrix} \left(f_1^X\right)^{-1} & \cdots & \left(f_1^X\right)^{-1} & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \left(f_1^X\right)^{-1} & \cdots & \left(f_1^X\right)^{-1} & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \left(f_2^X\right)^{-1} & \cdots & \left(f_2^X\right)^{-1} & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \left(f_2^X\right)^{-1} & \cdots & \left(f_2^X\right)^{-1} & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & \left(f_k^X\right)^{-1} & \cdots & \left(f_k^X\right)^{-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & \left(f_k^X\right)^{-1} & \cdots & \left(f_k^X\right)^{-1} \end{bmatrix},$$

όπου το πρώτο διαγώνιο block τιμών έχει διαστάσεις $f_1^X \times f_1^X$ και περιλαμβάνει $\left(f_1^X\right)^2$ στοιχεία, το δεύτερο είναι διαστάσεων $f_2^X \times f_2^X$ με $\left(f_2^X\right)^2$ στοιχεία κ.ο.κ. Το τελευταίο block είναι $f_k^X \times f_k^X$ και περιλαμβάνει $\left(f_k^X\right)^2$ στοιχεία.

Επειδή η μορφή του πίνακα $\mathbf{Z_Y}$ δεν είναι συγκεκριμένη, εισάγουμε τον παρακάτω συμβολισμό:

$$\mathbf{Z_Y} = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ \vdots & p(1)_{f_2^Y} & \vdots & \vdots \\ p(1)_{f_1^Y} & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & p(1)_{f_l^Y} \\ \vdots & \vdots & \vdots & \vdots \\ p(0)_{N-f_1^Y} & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & p(0)_{N-f_l^Y} \\ \vdots & p(0)_{N-f_2^Y} & \vdots & \vdots \end{bmatrix},$$

όπου το σύμβολο $p(1)_{f_j^Y}$, με $j = 1,\ldots,l$, δηλώνει ότι στη στήλη j του πίνακα $\mathbf{Z_Y}$ σε κάποιες θέσεις υπάρχουν f_j^Y, σε πλήθος, 1, ενώ το σύμβολο $p(0)_{N-f_j^Y}$ ότι σε κάποιες

θέσεις υπάρχουν $N - f_j^Y$, σε πλήθος, 0. Με βάση τους παραπάνω συμβολισμούς ο πίνακας $\mathbf{P}_{\text{proj}}$ γράφεται:

$$\mathbf{P}_{\text{proj}}=\begin{bmatrix} \left(f_1^X\right)^{-1} & \cdots & \left(f_1^X\right)^{-1} & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \left(f_1^X\right)^{-1} & \cdots & \left(f_1^X\right)^{-1} & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \left(f_2^X\right)^{-1} & \cdots & \left(f_2^X\right)^{-1} & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \left(f_2^X\right)^{-1} & \cdots & \left(f_2^X\right)^{-1} & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & \left(f_k^X\right)^{-1} & \cdots & \left(f_k^X\right)^{-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & \left(f_k^X\right)^{-1} & \cdots & \left(f_k^X\right)^{-1} \end{bmatrix} \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ \vdots & p(1)_{f_2^Y} & \vdots & \vdots \\ p(1)_{f_1^Y} & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & p(1)_{f_l^Y} \\ \vdots & \vdots & \vdots & \vdots \\ p(0)_{N-f_1^Y} & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & p(0)_{N-f_l^Y} \\ \vdots & p(0)_{N-f_2^Y} & \vdots & \vdots \end{bmatrix}$$

Αν συμβολίσουμε με $\mathbf{H}$ τον πίνακα $\mathbf{C_X}\left(\mathbf{C_X^T C_X}\right)^{-1}\mathbf{C_X^T}$, τότε $\mathbf{P}_{\text{proj}}=\mathbf{H}\,\mathbf{Z_Y}$. Το στοιχείο p_{11} του $\mathbf{P}_{\text{proj}}$ προκύπτει από το εσωτερικό γινόμενο του διανύσματος της πρώτης γραμμής του $\mathbf{H}$ και του διανύσματος της πρώτης στήλης του $\mathbf{Z_Y}$. Δηλαδή $p_{11}=\sum_{n=1}^{N} h_{1n}z_{n1}$, όπου $h_{\mu n}$ και z_{nj} με $(\mu, n=1,\ldots,N,)$ και $(j=1,\ldots,l)$, είναι τα γενικά στοιχεία των πινάκων $\mathbf{H}$ και $\mathbf{Z_Y}$ αντίστοιχα. Στο άθροισμα αυτό, υπάρχουν f_{11} σε πλήθος όροι που είναι διάφοροι του μηδενός και ίσοι με $\left(f_1^X\right)^{-1}$, αφού f_{11} σε πλήθος αντικείμενα της κλάσης 1 της μεταβλητής X ανήκουν στην κλάση 1 της Y (βλέπε Πίνακα 7.1). Συνεπώς,

$$p_{11}=\sum_{n=1}^{N} h_{1n}z_{n1}=\underbrace{\frac{1}{f_1^X}+\frac{1}{f_1^X}+\ldots\frac{1}{f_1^X}}_{f_{11}\ \text{όροι}}=\frac{f_{11}}{f_1^X}.$$

Όμοια, για το στοιχείο $p_{12}=\sum_{n=1}^{N} h_{1n}z_{n2}$ έχουμε ότι το άθροισμα περιλαμβάνει f_{12} όρους ίσους με $\left(f_1^X\right)^{-1}$, αφού f_{12} σε πλήθος αντικείμενα ανήκουν ταυτόχρονα στην κλάση 1 της X και στην κλάση 2 της Y (βλέπε Πίνακα 7.1). Επομένως,

$$p_{12} = \sum_{n=1}^{N} h_{1n} z_{n2} = \underbrace{\frac{1}{f_1^X} + \frac{1}{f_1^X} + \ldots \frac{1}{f_1^X}}_{f_{12}\ \text{όροι}} = \frac{f_{12}}{f_1^X}.$$

Ανάλογα εργαζόμενοι μπορούμε να δείξουμε ότι:

$$p_{1l} = \frac{f_{1l}}{f_1^X}.$$

Συνεχίζοντας με την παραπάνω διαδικασία προκύπτει η τελική μορφή του πίνακα $\mathbf{P}_{\text{proj}}$ που παρουσιάζεται στον Πίνακα 7.2. Διαπιστώνουμε ότι ο $\mathbf{P}_{\text{proj}}$ αποτελείται από k block υποπινάκων. Το πρώτο block είναι διαστάσεων $f_1^X \times l$, το δεύτερο $f_2^X \times l$ κ.ο.κ. Το τελευταίο block έχει διαστάσεις $f_k^X \times l$. Λόγω της ταξινόμησης του πίνακα $\mathbf{C_X}$ οι πρώτες f_1^X γραμμές του $\mathbf{P}_{\text{proj}}$ αντιστοιχούν στα αντικείμενα που ανήκουν στην κλάση 1 της μεταβλητής X, οι επόμενες f_2^X γραμμές στα αντικείμενα που ανήκουν στην κλάση 2 της X κ.ο.κ. Ο πολλαπλασιασμός της γραμμής n του πίνακα $\mathbf{C_X}\left(\mathbf{C_X^T C_X}\right)^{-1}\mathbf{C_X^T}$ με τη στήλη s του πίνακα $\mathbf{Z_Y}$ έχει ως αποτέλεσμα το γενικό στοιχείο p_{ns} του πίνακα $\mathbf{P}_{\text{proj}}$ να δίνεται από τη σχέση:

$$p_{ns} = \frac{f_s^Y}{f_i^X} \text{ για } s = 1, \ldots, l,$$

όπου $i = 1, \ldots, k$ είναι η κλάση της μεταβλητής X στην οποία ανήκει το αντικείμενο n.

Όπως προαναφέρθηκε, για να καταλήξουμε στην παραπάνω σχέση λάβαμε υπόψη ότι τα στοιχεία του $\mathbf{Z_Y}$ είναι 0 ή 1 και ότι στη στήλη s (του $\mathbf{Z_Y}$) υπάρχουν συνολικά f_s^Y σε πλήθος μονάδες (1). Η περιθώρια γραμμή του $\mathbf{P}_{\text{proj}}$ αντιστοιχεί στην περιθώρια κατανομή απολύτων συχνοτήτων της Y, ενώ το άθροισμα των στοιχείων του είναι ίσο με N.

Παρατηρούμε ότι οι γραμμές του πίνακα $\mathbf{P}_{\text{proj}}$ εκφράζουν τα προφίλ των N δειγματοληπτικών ή πειραματικών μονάδων (αντικειμένων) ως προς τις κλάσεις της εξαρτημένης μεταβλητής Y ανάλογα με το σε ποια κλάση της ανεξάρτητης μεταβλητής X αυτά ανήκουν (βλέπε Πίνακα 7.2). Με άλλα λόγια, ο $\mathbf{P}_{\text{proj}}$ περιγράφει

τα προφίλ των N αντικειμένων ως προς τη μεταβλητή Y στο χώρο που ορίζουν οι στήλες του πίνακα $\mathbf{C_X}$. Ο $\mathbf{P}_{\text{proj}}$ είναι ένας γενικευμένος λογικός πίνακας (βλέπε Ενότητα 2.4.2). Τα στοιχεία κάθε γραμμής του είναι ποσοστά, δηλαδή αριθμοί στο διάστημα [0, 1] και έχουν άθροισμα ίσο με τη μονάδα. Επίσης, τα σύνολα των στηλών του αντιστοιχούν στην περιθώρια κατανομή απολύτων συχνοτήτων των κλάσεων της μεταβλητής Y. Επομένως, αν τα διαθέσιμα δεδομένα έχουν συγκεντρωθεί με τυχαία δειγματοληψία, οι τιμές που αντιστοιχούν σε κάθε διάνυσμα γραμμής μπορούν να θεωρηθούν ως δεσμευμένες πιθανότητες. Πιο συγκεκριμένα, η τιμή p_{ns} εκφράζει την πιθανότητα το αντικείμενο n να ανήκει στην κλάση s της εξαρτημένης μεταβλητής Y δοθέντος ότι ανήκει στην κλάση i της ανεξάρτητης μεταβλητής X. Στην Ενότητα 2.4 είδαμε ότι η ΠΑΑ μπορεί να εφαρμοστεί απευθείας σε πίνακες όπως ο $\mathbf{P}_{\text{proj}}$.

Πίνακας 7.2: Τελική Μορφή του Πίνακα $\mathbf{P}_{\text{proj}}$

	Κλάσεις της Y				
Κλάσεις της X	1	2	…	l	Σύνολα
1	$\frac{f_{11}}{f_1^X}$	$\frac{f_{12}}{f_1^X}$	…	$\frac{f_{1l}}{f_1^X}$	1
⋮	⋮	⋮	⋮	⋮	
1	$\frac{f_{11}}{f_1^X}$	$\frac{f_{12}}{f_1^X}$	…	$\frac{f_{1l}}{f_1^X}$	1
2	$\frac{f_{21}}{f_2^X}$	$\frac{f_{22}}{f_2^X}$	…	$\frac{f_{2l}}{f_2^X}$	1
⋮	⋮	⋮	⋮	⋮	
2	$\frac{f_{21}}{f_2^X}$	$\frac{f_{22}}{f_2^X}$		$\frac{f_{2l}}{f_2^X}$	1
⋮	⋮	⋮	⋮	⋮	⋮
k	$\frac{f_{k1}}{f_k^X}$	$\frac{f_{k2}}{f_k^X}$	…	$\frac{f_{kl}}{f_k^X}$	1
⋮	⋮	⋮	⋮	⋮	
k	$\frac{f_{k1}}{f_k^X}$	$\frac{f_{k2}}{f_k^X}$	…	$\frac{f_{kl}}{f_k^X}$	1
Σύνολα	f_1^Y	f_2^Y	…	f_l^Y	N

Αν συνενώσουμε τις γραμμές (αντικείμενα) του $\mathbf{P}_{\text{proj}}$, οι οποίες έχουν ίδιο προφίλ, τότε προκύπτει ο συμπτυγμένος πίνακας $\mathbf{P_C}$, ο οποίος, όπως εύκολα μπορεί να διαπιστωθεί, ταυτίζεται με τον αρχικό πίνακα συμπτώσεων **F** (βλέπε Πίνακα 7.3). Επομένως, με βάση την Αρχή της Ισοδυναμίας των Σχετικών Κατανομών, η εφαρμογή της ΠΑΑ στον $\mathbf{P}_{\text{proj}}$ παράγει ισοδύναμα αποτελέσματα με αυτά που προκύπτουν από την εφαρμογή της μεθόδου στον $\mathbf{P_C}$ (Benzécri, 1992). Το πλεονέκτημα είναι ότι κατά την ανάλυση του πίνακα $\mathbf{P}_{\text{proj}}$ έχουμε πληροφορία για τις *N* δειγματοληπτικές ή πειραματικές μονάδες, η οποία μπορεί να χρησιμοποιηθεί σε περεταίρω αναλύσεις. Η πληροφορία αυτή δεν είναι άμεσα διαθέσιμη κατά την ανάλυση του **F** ή ισοδύναμα του $\mathbf{P_C}$. Για την εφαρμογή της ΠΑΑ υποθέτουμε ότι δεν υπάρχουν ελλείπουσες τιμές και ότι όλα τα σημεία είναι ενεργά. Στην Ενότητα Z1 του Παραρτήματος Z δίνουμε ένα αριθμητικό παράδειγμα υπολογισμού των πινάκων $\mathbf{P}_{\text{proj}}$ και $\mathbf{P_C}$.

Πίνακας 7.3: Ο Συμπτυγμένος Πίνακας $\mathbf{P_C}$

	Στήλες (*Y*)				
Γραμμές (*X*)	1	2	…	*l*	Σύνολα
1	$f_1^X \frac{f_{11}}{f_1^X} = f_{11}$	$f_1^X \frac{f_{12}}{f_1^X} = f_{12}$	…	$f_1^X \frac{f_{1l}}{f_1^X} = f_{1l}$	f_1^X
2	$f_2^X \frac{f_{21}}{f_2^X} = f_{21}$	$f_2^X \frac{f_{22}}{f_2^X} = f_{22}$	…	$f_2^X \frac{f_{2l}}{f_2^X} = f_{2l}$	f_2^X
⋮	⋮	⋮	⋮	⋮	⋮
k	$f_k^X \frac{f_{k1}}{f_k^X} = f_{k1}$	$f_k^X \frac{f_{k2}}{f_k^X} = f_{k2}$	…	$f_k^X \frac{f_{kl}}{f_k^X} = f_{kl}$	f_k^X
Σύνολα	f_1^Y	f_2^Y	…	f_l^Y	*N*

Το ερώτημα που τίθεται τώρα είναι σχετικά με την επιλογή του είδους της κανονικοποίησης των παραγοντικών συντεταγμένων. Με βάση όσα αναφέρθηκαν στις Ενότητες 2.2.14.1 και 2.2.14.2 προτείνουμε τα παρακάτω:

Αν η ΠΑΑ εφαρμοστεί στον πίνακα **F** (ή ισοδύναμα στον $\mathbf{P_C}$), για την κατασκευή των παραγοντικών επιπέδων και τη διαγραμματική ερμηνεία των αποτελεσμάτων προτείνουμε την Κύρια Κανονικοποίηση (*Principal Normalization - PN*) των παραγοντικών συντεταγμένων, στο πνεύμα της Γαλλικής Σχολής. Για τη χρήση των συντεταγμένων, ως βέλτιστα ποσοτικοποιημένων τιμών, σε περαιτέρω αναλύσεις

προτείνουμε την Κύρια Κανονικοποίηση κατά Γραμμές (*Row Principal Normalization - RPN*), σύμφωνα με την οποία υπολογίζονται οι κύριες συντεταγμένες για τα σημεία που αντιστοιχούν στις κλάσεις της ανεξάρτητης μεταβλητής (*X*) και οι τυποποιημένες για τις κλάσεις της εξαρτημένης μεταβλητής (*Y*). Με τη *RPN* οι κλάσεις της ανεξάρτητης μεταβλητής προβάλλονται στο κέντρο βάρους των κλάσεων της εξαρτημένης από τις οποίες χαρακτηρίζονται. Την ίδια μέθοδο κανονικοποίησης προτείνουμε και στην περίπτωση που η ΠΑΑ εφαρμοστεί στον πίνακα $\mathbf{P}_{\text{proj}}$. Για τα *N* αντικείμενα υπολογίζονται οι κύριες, ενώ για τις κλάσεις της μεταβλητής *Y* οι τυποποιημένες συντεταγμένες. Έτσι, και στις δύο περιπτώσεις μπορούμε να έχουμε και *biplot* ερμηνεία των αντίστοιχων παραγοντικών διαγραμμάτων (βλέπε Ενότητες 2.2.14.1 και 2.2.14.2). Κατά την ανάλυση του $\mathbf{P}_{\text{proj}}$, με *RPN*, οι *N* δειγματοληπτικές ή πειραματικές μονάδες προβάλλονται, επί των παραγοντικών επιπέδων, στο κέντρο βάρους των κλάσεων της μεταβλητής *Y* στις οποίες ανήκουν (βλέπε Τέταρτο Βήμα της Ενότητας 2.2.14). Αν τα αντικείμενα ταυτοποιηθούν επί των παραγοντικών επιπέδων ως προς τις κλάσεις της ανεξάρτητης μεταβλητής *X*, τότε η εικόνα του υπό εξέταση φαινομένου ταυτίζεται με αυτή που προκύπτει από την εφαρμογή της ΠΑΑ στον αρχικό πίνακα συμπτώσεων **F**.

Στις ενότητες που ακολουθούν παρουσιάζουμε εφαρμογές της προτεινόμενης μεθοδολογίας σε τρεις βασικούς πειραματικούς σχεδιασμούς, στους οποίους οι εμπλεκόμενες μεταβλητές είναι κατηγορικές. Σε κάθε περίπτωση υποθέτουμε ότι δεν υπάρχουν ελλείπουσες τιμές και ότι όλα τα σημεία είναι ενεργά για την εφαρμογή της ΠΑΑ. Οι βασικές αρχές των πειραμάτων, στο πλαίσιο της Στατιστικής και της Θεωρίας Πιθανοτήτων, αναφέρθηκαν στην Ενότητα 1.2. Για περισσότερες πληροφορίες σχετικά με τη μεθοδολογία σχεδιασμού και την ανάλυση πειραματικών δεδομένων παραπέμπουμε στους Cochran και Cox (1953), Cox (1958), Gomez και Gomez (1984), Κάτο (1986), Steel και Torrie (1986), Brown και Melamed (1990), Μπόρα-Σέντα και Μωϋσιάδη (1992), Girden (1992), Καρακώστα (1993), Κίτσο (1994), Kirk (1995), Daniel (1995), Stapleton (1995), Zar (1996), Mendenhall και Sincich (1996), Montgomery (1997), Kuehl (2000), Rencher (2000), Rao (2002) και Kutner *et al*. (2005).

Παρατηρήσεις

Α) Είναι πιθανό ο πίνακας $\mathbf{P}_{proj}$ να περιλαμβάνει μεγάλο πλήθος γραμμών – αντικειμένων, όπως για παράδειγμα συμβαίνει σε έρευνες δημοσκόπησης. Στην περίπτωση αυτή, η ΠΑΑ μπορεί να εφαρμοστεί στον **F** και, στη συνέχεια, οι βέλτιστες τιμές, που θα προκύψουν, να αντικατασταθούν στον αρχικό πίνακα δεδομένων **D**, ώστε να χρησιμοποιηθούν σε περαιτέρω αναλύσεις.

Β) Η προτεινόμενη διαδικασία μπορεί να συνδυαστεί με τους ελέγχους σημαντικότητας που παρουσιάσαμε στην Ενότητα 6.2.1 καθώς και με τις μεθόδους κατασκευής ελλείψεων εμπιστοσύνης, τις οποίες προτείναμε στην Ενότητα 6.3. Εφόσον το ενδιαφέρον της μελέτης επικεντρώνεται στη σύγκριση των προφίλ των γραμμών του πίνακα **F**, δηλαδή των κλάσεων της ανεξάρτητης μεταβλητής, οι ελλείψεις θα πρέπει να σχεδιαστούν γύρω από τα σημεία γραμμών. Εναλλακτικά, μπορεί να εφαρμοστεί και η μέθοδος πολλαπλών συγκρίσεων του *Gabriel* (βλέπε Ενότητες 6.3.5 και ΣΤ3 του Παραρτήματος ΣΤ). Αν τα διαθέσιμα δεδομένα δεν έχουν συγκεντρωθεί με μεθόδους της τυχαίας δειγματοληψίας ή ο πειραματισμός είναι ατελής και δεν υπακούει στη δομημένη πορεία που παρουσιάστηκε στην Ενότητα 1.2, τότε η ΠΑΑ, ενδεχομένως σε συνδυασμό με άλλες μεθόδους της Ανάλυσης Δεδομένων, όπως είναι η Ταξινόμηση (βλέπε Επισήμανση Β της Ενότητας 2.2.14.5), είναι ίσως η μοναδική μέθοδος για τη διερεύνηση είτε της επίδρασης της ανεξάρτητης μεταβλητής στην εξαρτημένη είτε, γενικότερα, της σχέσης μεταξύ των δύο κατηγορικών μεταβλητών. Βέβαια, στην περίπτωση αυτή, δεν μπορούν να αποδοθούν σχέσεις αιτίας αποτελέσματος.

Γ) Στη συγκεκριμένη περίπτωση και σε αντιστοιχία με την Ανάλυση Διασποράς με ένα παράγοντα, σκοπός της μελέτης είναι η σύγκριση των προφίλ των κλάσεων της ανεξάρτητης μεταβλητής *X*. Στις Ενότητες 2.2.14.1 και 2.2.14.2 είδαμε ότι όταν το ερευνητικό ενδιαφέρον εστιάζεται στη σύγκριση των προφίλ της μεταβλητής γραμμών, τότε είναι προτιμότερη η εφαρμογή της *RPN* (βλέπε και Greenacre, 2006).

7.3 Πλήρως Τυχαιοποιημένο Σχέδιο[36] με Ένα Παράγοντα και Μία Εξαρτημένη Μεταβλητή

Στο σχεδιασμό αυτό εμπλέκονται μία ανεξάρτητη μεταβλητή και μία εξαρτημένη που παίρνουν k και l διακεκριμένες τιμές αντίστοιχα. Σκοπός του πειράματος είναι η μελέτη της επίδρασης της ανεξάρτητης μεταβλητής στην εξαρτημένη. Η ανεξάρτητη μεταβλητή ονομάζεται συνήθως «παράγοντας» και οι k τιμές της «επίπεδα», «στάθμες», «μεταχειρίσεις» ή «αγωγές». Οι τιμές της εξαρτημένης μεταβλητής αποτελούν τις «αποκρίσεις» των πειραματικών μονάδων στο υπό εξέταση μέγεθος, όπως αυτό μετριέται μέσω της εξαρτημένης μεταβλητής. Τα επίπεδα του παράγοντα είτε προκαθορίζονται από τον πειραματιστή είτε προκύπτουν από τυχαία επιλογή από ένα σύνολο δυνατών τιμών. Η πειραματική διάταξη του συγκεκριμένου σχεδιασμού απαιτεί την τυχαιοποίηση (π.χ. με κλήρωση) των διαθέσιμων N σε πλήθος πειραματικών μονάδων στα k επίπεδα του παράγοντα. Στην περίπτωση «ισορροπημένου» σχεδίου η μόνη απαίτηση είναι κάθε αγωγή να περιλαμβάνει N/k πειραματικές μονάδες. Σε μη ισορροπημένα σχέδια το πλήθος των πειραματικών μονάδων μπορεί να διαφέρει από αγωγή σε αγωγή. Στους κλασικούς πειραματικούς σχεδιασμούς η ανεξάρτητη μεταβλητή είναι συνήθως κατηγορική και η εξαρτημένη ποσοτική. Στο πλαίσιο της Επαγωγικής Στατιστικής, η στατιστική επεξεργασία των διαθέσιμων δεδομένων περιλαμβάνει διαδικασίες, όπως είναι η Ανάλυση Διακύμανσης (Παραμετρική, Μη Παραμετρική) και οι συγκρίσεις μέσων όρων. Αν η ανεξάρτητη μεταβλητή είναι ποσοτική, τότε τα στοιχεία αναλύονται συνήθως μέσω της Ανάλυσης Παλινδρόμησης (γραμμική, μη γραμμική).

Στην περίπτωση που εξετάζουμε, και οι δύο μεταβλητές είναι κατηγορικής φύσης. Η ανάλυση του συγκεκριμένου πειραματικού σχεδιασμού αποτελεί άμεση εφαρμογή της βασικής μεθοδολογίας που παρουσιάσαμε στην προηγούμενη ενότητα.

---

[36] Ο αγγλικός όρος είναι *Complete Randomized Design* (*CRD*).

7.4 Πλήρως Τυχαιοποιημένο Σχέδιο με Ένα Παράγοντα και Δύο ή Περισσότερες Εξαρτημένες Μεταβλητές

Ο σχεδιασμός αυτός είναι ίδιος με τον προηγούμενο με τη διαφορά ότι συμμετέχουν περισσότερες εξαρτημένες κατηγορικές μεταβλητές. Έστω X η ανεξάρτητη μεταβλητή με k κατηγορίες και Y_i $(i=1,\ldots,q)$ οι q σε πλήθος κατηγορικές μεταβλητές με l_i κλάσεις η κάθε μία. Συμβολίζουμε με $j=\sum_{i=1}^{q} l_i$ το συνολικό αριθμό κλάσεων των q μεταβλητών. Ας είναι **D** ο $N\times(q+1)$ αρχικός πίνακας δεδομένων, όπου N είναι το συνολικό πλήθος των πειραματικών μονάδων. Κατασκευάζουμε τον πίνακα σχεδιασμού $\mathbf{C_X}$ της ανεξάρτητης μεταβλητής και τον $N\times j$ λογικό πίνακα $\mathbf{Z_Y}$, ο οποίος αντιστοιχεί στις q μεταβλητές. Και στην περίπτωση αυτή μπορούμε να προβάλουμε τον πίνακα $\mathbf{Z_Y}$ στο χώρο σχεδιασμού της ανεξάρτητης μεταβλητής μέσω του πίνακα:

$$\mathbf{P}_{\text{proj}}=\mathbf{C_X}(\mathbf{C_X^T C_X})^{-1}\mathbf{C_X^T}\,\mathbf{Z_Y}.$$

Ο πίνακας $\mathbf{P}_{\text{proj}}$ είναι τώρα διαστάσεων $N\times j$ και περιγράφει τα προφίλ των N πειραματικών μονάδων, ως προς τις μεταβλητές Y_i, στο χώρο που ορίζουν οι στήλες του πίνακα σχεδιασμού $\mathbf{C_X}$. Η ΠΑΑ μπορεί να εφαρμοστεί είτε στον $\mathbf{P}_{\text{proj}}$ είτε στο συμπτυγμένο πίνακα $\mathbf{P_C}$, ο οποίος προκύπτει από την συνένωση των αντικειμένων με όμοιο προφίλ (Αρχής της Ισοδυναμίας των Σχετικών Κατανομών). Ακολουθώντας τη μεθοδολογία της Ενότητας 7.2 καταλήγουμε στην τελική μορφή του $\mathbf{P}_{\text{proj}}$, η οποία δίνεται στον Πίνακα 7.4. Ο πίνακας αποτελείται από $k\times q$ block υποπινάκων. Το πρώτο block που αντιστοιχεί στις κλάσεις της μεταβλητής Y_i και αποτελείται από l_i στήλες και f_1^X γραμμές, το δεύτερο από l_i στήλες και f_2^X γραμμές κ.ο.κ. Το τελευταίο, k block, έχει διαστάσεις $f_k^X\times l_i$. Το στοιχείο $f_{ts}^{XY_i}$ του $\mathbf{P}_{\text{proj}}$ εκφράζει την απόλυτη συχνότητα που αντιστοιχεί στο κελί (t, s) του πίνακα συμπτώσεων

$$\mathbf{F_{XY_i}}=\begin{bmatrix} f_{11} & \cdots & f_{1l_i} \\ \vdots & \vdots & \vdots \\ f_{k1} & \cdots & f_{kl_i} \end{bmatrix},$$

ο οποίος διασταυρώνει τις κλάσεις της ανεξάρτητης μεταβλητής *X*, στις γραμμές, με τις κλάσεις της εξαρτημένης Y_i, στις στήλες. Το άθροισμα κάθε γραμμής του $\mathbf{P}_{\text{proj}}$ είναι ίσο με *q*, ενώ το γενικό άθροισμα των στοιχείων του είναι ίσο με *Nq*. Τα πρώτα l_1 κελιά της περιθώριας γραμμής του $\mathbf{P}_{\text{proj}}$ αντιστοιχούν στην περιθώρια κατανομή της μεταβλητής Y_1, τα επόμενα l_2 στην περιθώρια κατανομή της Y_2 κ.ο.κ. Παρατηρούμε ότι οι γραμμές του πίνακα $\mathbf{P}_{\text{proj}}$ εκφράζουν τα προφίλ των αντικειμένων για κάθε εξαρτημένη μεταβλητή Y_i ανάλογα με την κλάση της ανεξάρτητης στην οποία αυτά ανήκουν. Αν συνενώσουμε τα αντικείμενα του $\mathbf{P}_{\text{proj}}$ που έχουν το ίδιο προφίλ προκύπτει ο συμπτυγμένος $k \times j$ πίνακας $\mathbf{P_C}$:

$$\mathbf{P_C} = \begin{bmatrix} f_{11}^{XY_1} & \cdots & f_{1l_1}^{XY_1} & f_{11}^{XY_2} & \cdots & f_{1l_2}^{XY_2} & \cdots & f_{11}^{XY_q} & \cdots & f_{1l_q}^{XY_q} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{k1}^{XY_1} & \cdots & f_{kl_1}^{XY_1} & f_{k1}^{XY_2} & \cdots & f_{kl_2}^{XY_2} & \cdots & f_{k1}^{XY_q} & \cdots & f_{kl_q}^{XY_q} \end{bmatrix},$$

ο οποίος δεν είναι παρά ο πίνακας "φέτα" που διασταυρώνει την ανεξάρτητη μεταβλητή με τις *q* εξαρτημένες. Το άθροισμα της γραμμής *t* του $\mathbf{P_C}$ είναι ίσο με qf_t^X, όπου f_t^X είναι η συχνότητα της κλάσης *t* της X $(t = 1, \ldots, k)$, ενώ το γενικό άθροισμα των στοιχείων του είναι ίσο με *Nq*. Τα πρώτα l_1 κελιά της περιθώριας γραμμής του $\mathbf{P_C}$ αντιστοιχούν στην περιθώρια κατανομή της μεταβλητής Y_1, τα επόμενα l_2 στην περιθώρια κατανομή της Y_2 κ.ο.κ. Τα τελευταία l_p κελιά αντιστοιχούν στην περιθώρια κατανομή της Y_p.

Παρατηρήσεις

Α) Κατά την εφαρμογή της ΠΑΑ στον $\mathbf{P_C}$ λαμβάνονται υπόψη μόνο οι συσχετίσεις της ανεξάρτητης μεταβλητής με τις εξαρτημένες και όχι οι ενδοσυσχετίσεις μεταξύ των *q* εξαρτημένων μεταβλητών (βλέπε Ενότητα 2.4.1). Επομένως, μπορούμε να εξετάσουμε τη στατιστική σημαντικότητα της συσχέτισης της ανεξάρτητης με κάθε μία από τις εξαρτημένες μέσω του ελέγχου χ^2 μετά από κατάλληλη διόρθωση του επιπέδου σημαντικότητας *α*, ώστε να διατηρηθεί το Αθροιστικό ή καλύτερα το Πειραματικό Σφάλμα Τύπου Ι σταθερό (βλέπε Ενότητα Ε3.2 του Παραρτήματος Ε).

Πίνακας 7.4: Τελική Μορφή του Πίνακα $\mathbf{P}_{\text{proj}}$

	Εξαρτημένες Μεταβλητές										
	Κλάσεις της Y_1			Κλάσεις της Y_2				Κλάσεις της Y_q			
Κλάσεις της X	1	$\cdots$	l_1	1	$\cdots$	l_2	$\cdots$	1	$\cdots$	l_q	Σύνολα
1	$f_{11}^{XY_1}/f_1^X$	$\cdots$	$f_{1l_1}^{XY_1}/f_1^X$	$f_{11}^{XY_2}/f_1^X$	$\cdots$	$f_{1l_2}^{XY_2}/f_1^X$	$\cdots$	$f_{11}^{XY_q}/f_1^X$	$\cdots$	$f_{1l_q}^{XY_q}/f_1^X$	q
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	$f_{11}^{XY_1}/f_1^X$	$\cdots$	$f_{1l_1}^{XY_1}/f_1^X$	$f_{11}^{XY_2}/f_1^X$	$\cdots$	$f_{1l_2}^{XY_2}/f_1^X$	$\cdots$	$f_{11}^{XY_q}/f_1^X$	$\cdots$	$f_{1l_q}^{XY_q}/f_1^X$	q
2	$f_{21}^{XY_1}/f_2^X$	$\cdots$	$f_{2l_1}^{XY_1}/f_2^X$	$f_{21}^{XY_2}/f_2^X$	$\cdots$	$f_{2l_2}^{XY_2}/f_2^X$	$\cdots$	$f_{21}^{XY_q}/f_2^X$	$\cdots$	$f_{2l_q}^{XY_q}/f_2^X$	q
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2	$f_{21}^{XY_1}/f_2^X$	$\cdots$	$f_{2l_1}^{XY_1}/f_2^X$	$f_{21}^{XY_2}/f_2^X$	$\cdots$	$f_{2l_2}^{XY_2}/f_2^X$	$\cdots$	$f_{21}^{XY_q}/f_2^X$	$\cdots$	$f_{2l_q}^{XY_q}/f_2^X$	q
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$f_{k1}^{XY_1}/f_k^X$	$\cdots$	$f_{kl_1}^{XY_1}/f_k^X$	$f_{k1}^{XY_2}/f_k^X$	$\cdots$	$f_{kl_2}^{XY_2}/f_k^X$	$\cdots$	$f_{k1}^{XY_q}/f_k^X$	$\cdots$	$f_{kl_q}^{XY_q}/f_k^X$	q
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$f_{k1}^{XY_1}/f_k^X$	$\cdots$	$f_{kl_1}^{XY_1}/f_k^X$	$f_{k1}^{XY_2}/f_k^X$	$\cdots$	$f_{kl_2}^{XY_2}/f_k^X$	$\cdots$	$f_{k1}^{XY_q}/f_k^X$	$\cdots$	$f_{kl_q}^{XY_q}/f_k^X$	q
Σύνολα	$f_1^{Y_1}$	$\cdots$	$f_{l_1}^{Y_1}$	$f_1^{Y_2}$	$\cdots$	$f_{l_2}^{Y_2}$	$\cdots$	$f_1^{Y_q}$	$\cdots$	$f_{l_q}^{Y_q}$	Nq

Οι έλεγχοι μπορούν να συνδυαστούν με τον υπολογισμό δεικτών συνάφειας PRE (*Proportional Reduction of Error-Αναλογικής Μείωσης του Σφάλματος Πρόβλεψης*), όπως είναι ο συντελεστής τ των Goodman και Kruskal (1954), για να διαπιστωθεί η προβλεπτική ικανότητα της ανεξάρτητης μεταβλητής (βλέπε Everitt 1979, Reynolds 1984, Hinkle, Wiersma & Jurs 1988, Dometrius 1992, Τσάντας και άλλοι 1999).

Β) Η ολική αδράνεια του πίνακα $\mathbf{P_C}$ είναι ίση με τη μέση αδράνεια των q σε πλήθος απλών υποπινάκων συμπτώσεων, από τους οποίους δομείται (βλέπε Ενότητα 4.6.2, Πόρισμα 1). Κάτω από την υπόθεση της ανεξαρτησίας της μεταβλητής X με τις εξαρτημένες Y_i η στατιστική σημαντικότητα της ολικής αδράνειας του $\mathbf{P_C}$ μπορεί να ελεγχθεί μέσω της Κατανομής χ^2. Ειδικότερα, αν I είναι η ολική αδράνεια του πίνακα, τότε (βλέπε Ενότητα 4.6.2, Πρόταση 5 και Πόρισμα 3):

$$I = \frac{\sum_{i=1}^{q} I_{XY_i}}{q} = \frac{1}{qN} \sum_{i=1}^{q} Q_{XY_i} ,$$

όπου I_{XY_i} και Q_{XY_i} είναι, αντίστοιχα, η αδράνεια και το στατιστικό χ^2 του απλού πίνακα συμπτώσεων $\mathbf{F}_{XY_i}$, ο οποίος διασταυρώνει τις μεταβλητές X και Y_i.

Κάτω από την υπόθεση της ανεξαρτησίας της X με τις Y_i το στατιστικό $\sum_{i=1}^{q} Q_{XY_i}$ ακολουθεί ασυμπτωτικά την Κατανομή χ^2 με $\sum_{i=1}^{q} d_i$ βαθμούς ελευθερίας, όπου d_i είναι οι βαθμοί ελευθερίας που αντιστοιχούν στο στατιστικό Q_{XY_i} (Hogg & Craig 1978, Rencher 2000, Rao 2002). Συνεπώς, για τον έλεγχο της στατιστικής σημαντικότητας της ολικής αδράνειας I του $\mathbf{P_C}$ η ποσότητα qNI μπορεί να συγκριθεί με την κρίσιμη τιμή της Κατανομής χ^2 με $\sum_{i=1}^{q} d_i$ βαθμούς ελευθερίας σε επίπεδο σημαντικότητας α.

Γ) Στην ανάλυση του $\mathbf{P_C}$ έχει ενδιαφέρον να εξετάσουμε τη φυσική ερμηνεία της κατά Benzécri χ^2 απόστασης μεταξύ των προφίλ δύο κλάσεων της ανεξάρτητης

μεταβλητής. Θεωρούμε τις κλάσεις *v* και *w* της μεταβλητής X $(v, w = 1, \ldots, k)$. Το τετράγωνο της χ^2 απόστασής τους δίνεται από τη σχέση (βλέπε Ενότητα 2.2.3):

$$d^2_{\chi^2}(v,w) = \frac{Nq}{f_1^{Y_1}}\left(\frac{f_{v1}^{XY_1}}{qf_v^X} - \frac{f_{w1}^{XY_1}}{qf_w^X}\right)^2 + \ldots + \frac{Nq}{f_{l_1}^{Y_1}}\left(\frac{f_{vl_1}^{XY_1}}{qf_v^X} - \frac{f_{wl_1}^{XY_1}}{qf_w^X}\right)^2 + \cdots +$$

$$+ \frac{Nq}{f_1^{Y_2}}\left(\frac{f_{v1}^{XY_2}}{qf_v^X} - \frac{f_{w1}^{XY_2}}{qf_w^X}\right)^2 + \ldots + \frac{Nq}{f_{l_2}^{Y_2}}\left(\frac{f_{vl_2}^{XY_2}}{qf_v^X} - \frac{f_{wl_2}^{XY_2}}{qf_w^X}\right)^2 + \cdots +$$

$$+ \frac{Nq}{f_1^{Y_q}}\left(\frac{f_{v1}^{XY_q}}{qf_v^X} - \frac{f_{w1}^{XY_q}}{qf_w^X}\right)^2 + \ldots + \frac{Nq}{f_{l_q}^{Y_q}}\left(\frac{f_{vl_q}^{XY_q}}{qf_v^X} - \frac{f_{wl_q}^{XY_q}}{qf_w^X}\right)^2 =$$

$$= \frac{Nq}{f_1^{Y_1}q^2}\left(\frac{f_{v1}^{XY_1}}{f_v^X} - \frac{f_{w1}^{XY_1}}{f_w^X}\right)^2 + \ldots + \frac{Nq}{f_{l_1}^{Y_1}q^2}\left(\frac{f_{vl_1}^{XY_1}}{f_v^X} - \frac{f_{wl_1}^{XY_1}}{f_w^X}\right)^2 + \cdots +$$

$$+ \frac{Nq}{f_1^{Y_2}q^2}\left(\frac{f_{v1}^{XY_2}}{f_v^X} - \frac{f_{w1}^{XY_2}}{f_w^X}\right)^2 + \ldots + \frac{Nq}{f_{l_2}^{Y_2}q^2}\left(\frac{f_{vl_2}^{XY_2}}{f_v^X} - \frac{f_{wl_2}^{XY_2}}{f_w^X}\right)^2 + \cdots +$$

$$+ \frac{Nq}{f_1^{Y_q}q^2}\left(\frac{f_{v1}^{XY_q}}{f_v^X} - \frac{f_{w1}^{XY_q}}{f_w^X}\right)^2 + \ldots + \frac{Nq}{f_{l_q}^{Y_q}q^2}\left(\frac{f_{vl_q}^{XY_q}}{f_v^X} - \frac{f_{wl_q}^{XY_q}}{f_w^X}\right)^2 =$$

$$= \frac{1}{q}\left[\frac{N}{f_1^{Y_1}}\left(\frac{f_{v1}^{XY_1}}{f_v^X} - \frac{f_{w1}^{XY_1}}{f_w^X}\right)^2 + \ldots + \frac{N}{f_{l_1}^{Y_1}}\left(\frac{f_{vl_1}^{XY_1}}{f_v^X} - \frac{f_{wl_1}^{XY_1}}{f_w^X}\right)^2\right] + \cdots +$$

$$+ \frac{1}{q}\left[\frac{N}{f_1^{Y_2}}\left(\frac{f_{v1}^{XY_2}}{f_v^X} - \frac{f_{w1}^{XY_2}}{f_w^X}\right)^2 + \ldots + \frac{N}{f_{l_2}^{Y_2}}\left(\frac{f_{vl_2}^{XY_2}}{f_v^X} - \frac{f_{wl_2}^{XY_2}}{f_w^X}\right)^2\right] + \cdots +$$

$$+ \frac{1}{q}\left[\frac{N}{f_1^{Y_q}}\left(\frac{f_{v1}^{XY_q}}{f_v^X} - \frac{f_{w1}^{XY_q}}{f_w^X}\right)^2 + \ldots + \frac{N}{f_{l_q}^{Y_q}}\left(\frac{f_{vl_q}^{XY_q}}{f_v^X} - \frac{f_{wl_q}^{XY_q}}{f_w^X}\right)^2\right] =$$

$$= \frac{1}{q}\left(d^2_{\chi^2}(v,w)^{XY_1} + d^2_{\chi^2}(v,w)^{XY_2} + \ldots + d^2_{\chi^2}(v,w)^{XY_q}\right),$$

όπου:

$d^2_{\chi^2}(v,w)^{XY_i}$: είναι το τετράγωνο της απόστασης χ^2 των προφίλ των κλάσεων *v* και *w* στον απλό πίνακα συμπτώσεων που διασταυρώνει τη μεταβλητή X με την Y_i $(i=1,\ldots,q)$

Nq: είναι το γενικό άθροισμα των στοιχείων του $\mathbf{P_C}$

$f^{Y_i}_{s_i}$: είναι η συχνότητα της κλάσης *s* της εξαρτημένης μεταβλητής Y_i $(s=1,\ldots l_i, i=1,\ldots,q)$

και

qf^X_v και qf^X_w: είναι τα σύνολα των γραμμών του $\mathbf{P_C}$ που αντιστοιχούν στις κλάσεις *v* και *w* της μεταβλητής *X*.

Από τη σχέση:

$$d^2_{\chi^2}(v,w)=\frac{1}{q}\left(d^2_{\chi^2}(v,w)^{XY_1}+d^2_{\chi^2}(v,w)^{XY_2}+\ldots+d^2_{\chi^2}(v,w)^{XY_q}\right),$$

προκύπτει το συμπέρασμα ότι στον πίνακα "φέτα" $\mathbf{P_C}$ το τετράγωνο της χ^2 απόστασης μεταξύ των προφίλ δύο κλάσεων της ανεξάρτητης μεταβλητής γραμμών είναι ίσο με το μέσο όρο των τετραγώνων των αποστάσεων χ^2 των αντίστοιχων προφίλ όπως αυτά υπολογίζονται ξεχωριστά σε καθέναν από τους *q* σε πλήθος απλούς πίνακες συμπτώσεων που διασταυρώνουν τη μεταβλητή X με τις Y_i. Με την έννοια αυτή, η απόσταση $d^2_{\chi^2}(v,w)$ θεωρείται ως μία γενικευμένη απόσταση *Manhattan* μεταξύ των προφίλ των κλάσεων *v* και *w* (Everitt, 1993). Με ανάλογο τρόπο μπορούμε να δείξουμε ότι η απόσταση $d^2_{\chi^2}(v,g)$ της κλάσης *v* από το κέντρο βάρους *g* των γραμμών του πίνακα $\mathbf{P_C}$ δίνεται από τη σχέση:

$$d^2_{\chi^2}(v,g)=\frac{1}{q}\left(d^2_{\chi^2}(v,g_1)^{XY_1}+d^2_{\chi^2}(v,g_2)^{XY_2}+\ldots+d^2_{\chi^2}(v,g_{\mathrm{q}})^{XY_q}\right),$$

όπου $d^2_{\chi^2}\left(v, g_i\right)^{XY_i}$ είναι η απόσταση του προφίλ της κλάσης *v* από το κέντρο βάρους g_i των γραμμών του απλού πίνακα συμπτώσεων, ο οποίος διασταυρώνει τη μεταβλητή X με τη Y_i $\left(i = 1, \ldots q\right)$.

Διαπιστώνουμε ότι η απόσταση $d^2_{\chi^2}\left(v, g\right)$ εκφράζει και πάλι το μέσο όρο των αποστάσεων του προφίλ της κλάσης *v* από τα κέντρα βάρους των γραμμών σε καθέναν από τους *q* απλούς πίνακες συμπτώσεων, από τους οποίους συντίθεται ο $\mathbf{P_C}$. Το αποτέλεσμα αυτό είναι αναμενόμενο και κάτω από την ισχύ του Πορίσματος 3 της Ενότητας 4.6.2. Και στην περίπτωση αυτή, η απόσταση $d^2_{\chi^2}\left(v, g\right)$ αποτελεί γενικευμένη απόσταση *Manhattan*. Φυσικά, τα παραπάνω συμπεράσματα σχετικά με τις αποστάσεις $d^2_{\chi^2}\left(v, w\right)$ και $d^2_{\chi^2}\left(v, g\right)$ μπορούν να γενικευτούν σε οποιονδήποτε πίνακα τύπου "φέτας".

Δ) Οι διαθέσιμες πειραματικές μονάδες θεωρούνται ομοιογενείς ως προς τα φυσιολογικά, βιολογικά ή άλλα χαρακτηριστικά ή ιδιότητές τους (Κάτος, 1986). Θεωρούμε, δηλαδή, ότι η ύπαρξη συστηματικής μεταβλητότητας στις τιμές των εξαρτημένων μεταβλητών είναι αποτέλεσμα κυρίως της επίδρασης των ανεξάρτητων. Βέβαια, στη διαμόρφωση των τιμών των εξαρτημένων μεταβλητών επιδρούν και τυχαίοι μη ελεγχόμενοι παράγοντες, οι οποίοι έχουν ως αποτέλεσμα την εμφάνιση πειραματικών ή δειγματοληπτικών σφαλμάτων, συστηματικών ή μη, το μέγεθος των οποίων μπορεί να εκτιμηθεί και να απομονωθεί μέσω του πειραματισμού (Cox, 1961). Η "προσδοκία" του πειραματιστή είναι η παρατηρούμενη διακύμανση των τιμών των μεταβλητών απόκρισης να μπορεί να αποδοθεί κυρίως στη συστηματική επίδραση των ανεξάρτητων παραγόντων.

Ε) Η προτεινόμενη μεθοδολογία μπορεί να εφαρμοστεί και στην περίπτωση που ο σχεδιασμός δεν είναι ισορροπημένος.

Ε) Σε έρευνες επισκόπησης οι κλάσεις της ανεξάρτητης μεταβλητής είναι δυνατό να αντιστοιχούν σε στρώματα ή ομάδες του υπό εξέταση πληθυσμού που καθορίζονται από το δειγματοληπτικό σχέδιο συλλογής των δεδομένων όπως, για παράδειγμα,

συμβαίνει στην Στρωματοποιημένη Τυχαία Δειγματοληψία. Μια άλλη εκδοχή είναι οι τιμές της ανεξάρτητης μεταβλητής να αντιστοιχούν σε διαφορετικές χρονικές περιόδους συγκέντρωσης των δεδομένων.

7.5 Τυχαιοποιημένο Σχέδιο με Δύο Παράγοντες και Μία ή Περισσότερες Εξαρτημένες Μεταβλητές

Ο σχεδιασμός αυτός περιλαμβάνει δύο ανεξάρτητες μεταβλητές (παράγοντες) και μία ή περισσότερες εξαρτημένες. Σκοπός του πειραματισμού είναι να εξεταστούν οι «κύριες επιδράσεις» (*main effects*) των παραγόντων καθώς και η αλληλεπίδρασή τους (*interaction*) πάνω στις εξαρτημένες μεταβλητές. Αλληλεπίδραση μεταξύ των δύο παραγόντων διαπιστώνεται όταν η μεταβολή των επιπέδων του ενός έχει ως αποτέλεσμα διαφορετική συμπεριφορά των τιμών της εξαρτημένης μεταβλητής στα επίπεδα του άλλου. Στην πράξη, στο συγκεκριμένο σχεδιασμό το ερευνητικό ενδιαφέρον εστιάζεται στη μελέτη της συνδυασμένης δράσης των δύο παραγόντων πάνω στη μεταβλητή απόκρισης, δηλαδή στην αλληλεπίδρασή τους (Cox 1958, Gomez & Gomez 1984, Hinkle, Wiersma & Jurs 1988, Jaccard, Turrisi & Wan 1990, Kirk 1995, Hair *et al*. 1995, Montgomery 1997, Jaccard 1998, Kinnear & Gray 1999 και 1995, Schabenberger, Gregoire & Kong 2000). Αν ανιχνευθεί σημαντική αλληλεπίδραση, τότε οι κύριες επιδράσεις, δηλαδή η μεμονωμένη και ανεξάρτητη δράση των δύο παραγόντων, σπάνια αποτελούν αντικείμενο περεταίρω διερεύνησης (Κάτος 1986, Girden 1992, Καρακώστας 1993, Mendenhall & Sincich 1996, Μαυρομάτης 1999, Huck 2000β). Σύμφωνα με τους Anderson και Whitcomb (2000) η αλληλεπίδραση των παραγόντων είναι αυτή που μπορεί να αποκαλύψει τις βαθύτερες αιτίες παραγωγής των πειραματικών δεδομένων.

Αν υποθέσουμε ότι οι δύο παράγοντες έχουν k και m επίπεδα αντίστοιχα, τότε οι N διαθέσιμες πειραματικές μονάδες τυχαιοποιούνται στους $k \times m$ συνδυασμούς των επιπέδων των δύο παραγόντων με την απαίτηση κάθε συνδυασμός να περιλαμβάνει N/km πειραματικές μονάδες. Μεθοδολογικά, η διαδικασία ανάθεσης των πειραματικών μονάδων στις $k \times m$ αγωγές είναι η ίδια όπως και στον προηγούμενο σχεδιασμό. Απλά, στην περίπτωση που εξετάζουμε, μπορούμε να θεωρήσουμε ότι τελικά έχουμε έναν μόνο παράγοντα με $k \times m$ στάθμες. Ο παραπάνω σχεδιασμός

ονομάζεται και «Πλήρες $k\times m$ Παραγοντικό Πείραμα» (*Complete* $k\times m$ *Factorial Experiment*). Ας συμβολίσουμε με A και B τους δύο παράγοντες. Ο συγκεκριμένος πειραματισμός περιλαμβάνει τρεις πίνακες σχεδιασμού (Kirk 1995, Mendenhall & Sincich 1996, Neter *et al*. 1996, Montgomery 1997, Kleinbaum *et al*. 1998, SAS Institute 1999 και 1990, Rencher 2000, Kuehl 2000, Kutner *et al*. 2005). Τον $N\times k$ πίνακα σχεδιασμού $\mathbf{C_A}$ που αντιστοιχεί στην κύρια επίδραση του παράγοντα Α, τον $N\times m$ πίνακα σχεδιασμού $\mathbf{C_B}$ της κύριας επίδρασης του παράγοντα B και τον $N\times km$ πίνακα $\mathbf{C_{A\times B}}$ που αντιστοιχεί στην αλληλεπίδραση $A\times B$ των δύο παραγόντων. Για παράδειγμα, αν $k=m=2$ και $N=4$, τότε οι τρεις πίνακες είναι οι παρακάτω:

$$\mathbf{C_A}=\begin{bmatrix}1&0\\1&0\\0&1\\0&1\end{bmatrix},\ \mathbf{C_B}=\begin{bmatrix}1&0\\0&1\\1&0\\0&1\end{bmatrix}\ \text{και}\ \mathbf{C_{A\times B}}=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}.$$

Οι δύο στήλες του πίνακα $\mathbf{C_A}$ αντιστοιχούν στα επίπεδα $\alpha 1$ και $\alpha 2$ του παράγοντα A, οι στήλες του $\mathbf{C_B}$ στα επίπεδα $\beta 1$ και $\beta 2$ του B και οι στήλες του $\mathbf{C_{A\times B}}$ στους τέσσερις συνδυασμούς (αγωγές), $\alpha 1\beta 1$, $\alpha 1\beta 2$, $\alpha 2\beta 1$ και $\alpha 2\beta 2$, των επιπέδων των δύο παραγόντων. Ο Nishisato (1994, 1990 και 1980) προτείνει η προβολή του λογικού πίνακα $\mathbf{Z_Y}$, των q εξαρτημένων μεταβλητών, στο χώρο των ανεξάρτητων να προκύπτει ανάλογα με το ποια επίδραση είναι επιθυμητό να μεγιστοποιηθεί μέσω της Δυικής Κλιμάκωσης. Έτσι, αν το ενδιαφέρον της μελέτης εστιάζεται στη μεγιστοποίηση της κύριας επίδρασης του παράγοντα A, τότε μπορεί να χρησιμοποιηθεί ο πίνακας:

$$\mathbf{P_A}=\mathbf{C_A}\left(\mathbf{C_A^T C_A}\right)^{-1}\mathbf{C_A^T Z_Y},$$

του οποίου η συμπτυγμένη μορφή[37] είναι ο $k\times j$ πίνακας "φέτα" $\mathbf{P_{C_A}}$ που διασταυρώνει τις κλάσεις (επίπεδα) του παράγοντα A με τις κατηγορίες των εξαρτημένων μεταβλητών (j είναι το συνολικό πλήθος των κλάσεων των q εξαρτημένων μεταβλητών). Ειδικότερα,

---

[37] Μετά από την εφαρμογή της μεθοδολογίας που παρουσιάσαμε στην Ενότητα 7.2.

$$\mathbf{P}_{\mathbf{C_A}} = \begin{bmatrix} f_{11}^{AY_1} & \cdots & f_{1l_1}^{AY_1} & f_{11}^{AY_2} & \cdots & f_{1l_2}^{AY_2} & \cdots & f_{11}^{AY_q} & \cdots & f_{1l_q}^{AY_q} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{k1}^{AY_1} & \cdots & f_{kl_1}^{AY_1} & f_{k1}^{AY_2} & \cdots & f_{kl_2}^{AY_2} & \cdots & f_{k1}^{AY_q} & \cdots & f_{kl_q}^{AY_q} \end{bmatrix}.$$

Στην περίπτωση που είναι επιθυμητή η μεγιστοποίηση της κύριας επίδρασης του παράγοντα *B*, τότε η προβολή του $\mathbf{Z_Y}$ στο χώρο στηλών του $\mathbf{C_B}$ είναι δυνατό να επιτευχθεί μέσω του πίνακα:

$$\mathbf{P_B} = \mathbf{C_B}\left(\mathbf{C_B^T C_B}\right)^{-1}\mathbf{C_B^T Z_Y},$$

στον οποίο αντιστοιχεί ο $m \times j$ πίνακας "φέτα" $\mathbf{P}_{\mathbf{C_B}}$, ο οποίος διασταυρώνει τα επίπεδα του παράγοντα *B* με τις κλάσεις των εξαρτημένων μεταβλητών. Πιο συγκεκριμένα, ο $\mathbf{P}_{\mathbf{C_B}}$ είναι της μορφής:

$$\mathbf{P}_{\mathbf{C_B}} = \begin{bmatrix} f_{11}^{BY_1} & \cdots & f_{1l_1}^{BY_1} & f_{11}^{BY_2} & \cdots & f_{1l_2}^{BY_2} & \cdots & f_{11}^{BY_q} & \cdots & f_{1l_q}^{BY_q} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{m1}^{BY_1} & \cdots & f_{ml_1}^{BY_1} & f_{m1}^{BY_2} & \cdots & f_{ml_2}^{BY_2} & \cdots & f_{m1}^{BY_q} & \cdots & f_{ml_q}^{BY_q} \end{bmatrix}$$

Τέλος, για τη μεγιστοποίηση της αλληλεπίδρασης των δύο παραγόντων η Δυική Κλιμάκωση μπορεί να εφαρμοστεί στον πίνακα:

$$\mathbf{P_{A\times B}} = \mathbf{C_{A\times B}}\left(\mathbf{C_{A\times B}^T C_{A\times B}}\right)^{-1}\mathbf{C_{A\times B}^T Z_Y},$$

του οποίου η συμπτυγμένη του εκδοχή είναι ο $km \times j$ πίνακας "φέτα" $\mathbf{P}_{\mathbf{C_{A\times B}}}$ που διασταυρώνει τους $k \times m$ συνδυασμούς επιπέδων των δύο παραγόντων με τις κλάσεις των εξαρτημένων μεταβλητών. Η γενική μορφή του είναι η παρακάτω:

$$\mathbf{P}_{\mathbf{C_{A\times B}}} = \begin{bmatrix} f_{11}^{(A\times B)Y_1} & \cdots & f_{1l_1}^{(A\times B)Y_1} & f_{11}^{(A\times B)Y_2} & \cdots & f_{1l_2}^{(A\times B)Y_2} & \cdots & f_{11}^{(A\times B)Y_q} & \cdots & f_{1l_q}^{(A\times B)Y_q} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{(km)1}^{(A\times B)Y_1} & \cdots & f_{(km)l_1}^{(A\times B)Y_1} & f_{(km)1}^{(A\times B)Y_2} & \cdots & f_{(km)l_2}^{(A\times B)Y_2} & \cdots & f_{(km)1}^{(A\times B)Y_q} & \cdots & f_{(km)l_q}^{(A\times B)Y_q} \end{bmatrix}.$$

Οι δύο πρώτες από τις παραπάνω περιπτώσεις δεν έχουν πρακτικό ενδιαφέρον, αφού ο χώρος σχεδιασμού της αλληλεπίδρασης $A \times B$ των δύο παραγόντων περιέχει τους χώρους σχεδιασμού των δύο κύριων επιδράσεων (SAS Institute, 1999 και 1990). Η

πρότασή μας για την ανάλυση του πειραματικού σχεδίου, που εξετάζουμε, είναι να εφαρμόσουμε την ΠΑΑ στον πίνακα

$$\mathbf{P}_{\mathbf{A\times B}} = \mathbf{C}_{\mathbf{A\times B}}\left(\mathbf{C}_{\mathbf{A\times B}}^{\mathbf{T}}\mathbf{C}_{\mathbf{A\times B}}\right)^{-1}\mathbf{C}_{\mathbf{A\times B}}^{\mathbf{T}}\mathbf{Z}_{\mathbf{Y}}$$

ή στην συμπτυγμένη του μορφή $\mathbf{P}_{\mathbf{C}_{\mathbf{A\times B}}}$, και, στη συνέχεια, να προβάλουμε, επί των παραγοντικών διαγραμμάτων, τα σημεία των κλάσεων των δύο παραγόντων ως συμπληρωματικά. Με τον τρόπο αυτό στις διαγραμματικές εκροές της ΠΑΑ θα οπτικοποιείται όχι μόνο η αλληλεπίδραση των δύο παραγόντων αλλά και οι κύριες επιδράσείς τους, σε αντιστοιχία με την Ανάλυση Διασποράς με δύο παράγοντες. Πρακτικά, αυτό σημαίνει ότι θα πρέπει αρχικά να δημιουργήσουμε μια νέα μεταβλητή με $k\times m$ κλάσεις, η οποία θα δηλώνει την αλληλεπίδραση των δύο ανεξάρτητων μεταβλητών - παραγόντων (βλέπε Ενότητα 2.4.1), στη συνέχεια να θεωρήσουμε ότι στον πειραματισμό υπεισέρχεται μόνο μία ανεξάρτητη μεταβλητή, η μεταβλητή αλληλεπίδρασης, και τέλος να υπολογίσουμε τους πίνακες $\mathbf{P}_{\mathbf{A\times B}}$ και $\mathbf{P}_{\mathbf{C}_{\mathbf{A\times B}}}$ επαυξημένους κατακόρυφα με τους πίνακες “φέτα” $\mathbf{P}_{\mathbf{C}_{\mathbf{A}}}$ και $\mathbf{P}_{\mathbf{C}_{\mathbf{B}}}$. Έτσι, δημιουργούμε τους πίνακες τύπου “στοίβας”:

$$\begin{bmatrix}\mathbf{P}_{\mathbf{A\times B}}\\ \mathbf{P}_{\mathbf{C}_{\mathbf{A}}}\\ \mathbf{P}_{\mathbf{C}_{\mathbf{B}}}\end{bmatrix} \text{ και } \begin{bmatrix}\mathbf{P}_{\mathbf{C}_{\mathbf{A\times B}}}\\ \mathbf{P}_{\mathbf{C}_{\mathbf{A}}}\\ \mathbf{P}_{\mathbf{C}_{\mathbf{B}}}\end{bmatrix},$$

επί των οποίων μπορούμε να εφαρμόσουμε την ΠΑΑ θέτοντας τις γραμμές των πινάκων $\mathbf{P}_{\mathbf{C}_{\mathbf{A}}}$ και $\mathbf{P}_{\mathbf{C}_{\mathbf{B}}}$ ως συμπληρωματικές. Με τον τρόπο αυτό, επί των παραγοντικών επιπέδων, προβάλλονται και τα κέντρα βάρους (μέσα προφίλ) των κλάσεων των δύο παραγόντων (Παπαδημητρίου, 1991), τα οποία αντιστοιχούν στις κύριες επιδράσεις τους. Διαπιστώνουμε ότι η ανάλυση του πειραματικού σχεδιασμού που εξετάζουμε ανάγεται στη μεθοδολογία της προηγούμενης ενότητας όπου εμπλέκεται μία μόνο ανεξάρτητη μεταβλητή (βλέπε και Παρατηρήσεις Α, Β, Γ, Δ και Ε της Ενότητας 7.4). Η προτεινόμενη μεθοδολογία μπορεί να επεκταθεί και σε ισορροπημένους παραγοντικούς σχεδιασμούς που περιλαμβάνουν περισσότερες από δύο ανεξάρτητες μεταβλητές. Στην περίπτωση αυτή, τα στοιχεία του λογικού πίνακα $\mathbf{Z}_{\mathbf{Y}}$ των εξαρτημένων μεταβλητών προβάλλονται στο χώρο στηλών του πίνακα σχεδιασμού που αντιστοιχεί στην αλληλεπίδραση μεγαλύτερης τάξης. Για

παράδειγμα, αν στο πείραμα εμπλέκονται τρεις παράγοντες *Α*, *Β* και *Γ* με *κ*, *λ* και *μ* στάθμες αντίστοιχα, τότε η προβολή του $\mathbf{Z}_{\Upsilon}$ θα πρέπει να γίνει στον *Ν*×*κλμ* πίνακα σχεδιασμού της αλληλεπίδρασης δεύτερης τάξης *Α*×*Β*×*Γ*, ο οποίος μπορεί να θεωρηθεί ως πίνακας σχεδιασμού μίας μόνο ανεξάρτητης μεταβλητής με *κ*×*λ*×*μ*, σε πλήθος, επίπεδα. Τα επίπεδα αυτά εκφράζουν τους συνδυασμούς των κλάσεων των τριών ανεξάρτητων μεταβλητών *Α*, *Β* και *Γ*.

Θα πρέπει να σημειώσουμε ότι, ο Nishisato (1994, 1990 και 1980) εκτός από τους πίνακες προβολής "καπέλο" χρησιμοποιεί και άλλους δύο τύπους πινάκων προβολής, των οποίων οι αντίστοιχοι πίνακες $\mathbf{P}_{\text{proj}}$ δεν έχουν την ίδια φυσική ερμηνεία με αυτή που προκύπτει από την χρήση των πινάκων τύπου "καπέλο". Πιο συγκεκριμένα, οι γραμμές τους δεν εκφράζουν τα προφίλ των αντικειμένων ως προς τις εξαρτημένες μεταβλητές, με τη συνήθη έννοια, και η συμπτυγμένη τους εκδοχή δεν είναι πίνακες τύπου "φέτας". Επίσης, οδηγούν σε διαφορετικά μεταξύ τους αποτελέσματα και στη χρήση γενικευμένων αντιστρόφων πινάκων κατά τους υπολογισμούς.

Παρατήρηση

Η μέθοδος που παρουσιάσαμε μπορεί να εφαρμοστεί και όταν το πείραμα δεν είναι ισορροπημένο ή στην περίπτωση έρευνας επισκόπησης όπου υπάρχει εννοιολογική διάκριση μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών.

7.6 Τυχαιοποιημένο Σχέδιο σε Πλήρη Συγκροτήματα[38] (*blocks*) με Δύο Παράγοντες και Μία ή Περισσότερες Εξαρτημένες Μεταβλητές

Στους σχεδιασμούς, τους οποίους παρουσιάσαμε στις προηγούμενες ενότητες, οι διαθέσιμες πειραματικές μονάδες θεωρούνται ομοιογενείς ως προς τα φυσιολογικά, βιολογικά ή άλλα χαρακτηριστικά τους. Έτσι, οι συστηματικές μεταβολές των τιμών των εξαρτημένων μεταβλητών μπορούν να αποδοθούν στην επίδραση κυρίως των ανεξάρτητων και όχι στη μεταβλητότητα που εισάγει η ενδεχόμενη ανομοιογένεια

[38] Ο αγγλικός όρος είναι *Randomized Complete Block Design* (*RCBD*).

των πειραματικών μονάδων. Στις περιπτώσεις κατά τις οποίες η ανομοιογένεια είναι σημαντική ή έχουν εντοπιστεί παράγοντες που επιδρούν ως “θόρυβος” στο υλικό πειραματισμού, η χρήση των πλήρως τυχαιοποιημένων παραγοντικών σχεδιασμών είναι δυνατό να οδηγήσει σε λανθασμένα συμπεράσματα (Cochran & Cox 1953, Cox 1961 και 1958, Gomez & Gomez 1984, Κάτο 1986, Steel & Torrie 1986, Brown & Melamed 1990, Καρακώστας 1993, Kirk 1995, Daniel 1995, Zar 1996, Mendenhall & Sincich 1996, Montgomery 1997, Kuehl 2000 και Kutner *et al*. 2005). Η επίδραση των ανεξάρτητων μεταβλητών μπορεί να είναι συσκοτισμένη (*confounded*) με τη συστηματική δράση εξωγενών παραγόντων, η μελέτη των οποίων δεν εντάσσεται στους στόχους του πειραματισμού. Όμως, είναι επιθυμητός ο έλεγχος και η απομόνωση της επίδρασής τους στη διακύμανση των εξαρτημένων μεταβλητών, ώστε να μελετηθούν με μεγαλύτερη ακρίβεια οι επιδράσεις των παραγόντων, οι οποίοι αποτελούν τις σημαντικές ανεξάρτητες μεταβλητές του πειράματος. Το Τυχαιοποιημένο σε Πλήρη Συγκροτήματα Σχέδιο αποτελεί μία λύση για τον *a priori* έλεγχο της “ανεπιθύμητης” μεταβλητότητας (Brown & Melamed, 1990). Σύμφωνα με το σχεδιασμό αυτό, οι διαθέσιμες πειραματικές μονάδες αρχικά χωρίζονται σε όσο το δυνατόν πιο ομοιογενείς ομάδες οι οποίες ονομάζονται «συγκροτήματα» (*blocks*). Στη συνέχεια, μέσα σε κάθε συγκρότημα οι πειραματικές μονάδες τυχαιοποιούνται στις πειραματικές αγωγές με βάση το τυχαιοποιημένο σχέδιο με έναν ή περισσότερους παράγοντες. Αν κάθε συγκρότημα περιλαμβάνει τον ίδιο αριθμό πειραματικών μονάδων, τότε ο πειραματισμός ονομάζεται «πλήρης» ή «ισορροπημένος». Στην ουσία, ο σχεδιασμός σε συγκροτήματα αποτελείται από μια σειρά (επαναλήψεις) ανεξάρτητων πλήρως τυχαιοποιημένων παραγοντικών σχεδιασμών, όπως αυτοί που παρουσιάσαμε στις προηγούμενες ενότητες. Τα κριτήρια ομαδοποίησης σε συγκροτήματα καθορίζονται από χαρακτηριστικά ή ιδιότητες (Kirk 1995, Kutner *et al*. 2005):

α) Των πειραματικών μονάδων. Για παράδειγμα, αν πρόκειται για άτομα, παράγοντες ομαδοποίησης συνήθως αποτελούν το φύλο, η ηλικία, το εισόδημα, η ευφυΐα, το μορφωτικό επίπεδο, η επαγγελματική εμπειρία και οι συνήθειές τους. Στην περίπτωση γεωγραφικών περιοχών, ο σχεδιασμός σε συγκροτήματα μπορεί να είναι τέτοιος ώστε να λαμβάνεται υπόψη το μέγεθος του αντίστοιχου πληθυσμού και το μέσο εισόδημα των κατοίκων της περιοχής.

β) Της πειραματικής διάταξης. Η ανάγκη για συγκρότηση ομάδων ομοιογενών πειραματικών μονάδων μπορεί να οφείλεται στην εμπλοκή περισσοτέρων του ενός παρατηρητών ή πειραματιστών, στο διαφορετικό χρόνο και τόπο διεξαγωγής των επαναλήψεων του πειράματος, στη χρήση πολλαπλών οργάνων μέτρησης και στην ύπαρξη διαφορετικών μηχανισμών παραγωγής των πειραματικών δεδομένων.

Σε κάθε περίπτωση, βασική επιδίωξη είναι οι πειραματικές μονάδες μέσα σε κάθε συγκρότημα να είναι όσο το δυνατόν πιο ομοιογενείς, ενώ ταυτόχρονα να παρουσιάζουν τη μέγιστη δυνατή ανομοιογένεια μεταξύ των συγκροτημάτων. Με το σχεδιασμό αυτό αναμένεται μείωση του πειραματικού σφάλματος και ακριβέστερη εκτίμηση των επιδράσεων των αγωγών. Όμως, το πειραματικό σχέδιο που εξετάζουμε έχει ορισμένες ιδιαιτερότητες. Οι παρατηρήσεις – μετρήσεις επί των πειραματικών μονάδων μέσα σε κάθε συγκρότημα δεν μπορούν να θεωρηθούν τελείως ανεξάρτητες. Για παράδειγμα, στην Εκπαιδευτική Έρευνα, οι μαθητές στο ίδιο σχολείο, λόγω γεωγραφικής τοποθεσίας, μάλλον έχουν μεγαλύτερη ομοιότητα μεταξύ τους ως προς το Κοινωνικο-Οικονομικό Επίπεδο (ΚΟΕ) των γονέων τους. Όμως, ορισμένα σχολεία προσελκύουν μαθητές από υψηλότερα κοινωνικά στρώματα σε σύγκριση με κάποια άλλα, στα οποία το ΚΟΕ των οικογενειών των μαθητών είναι χαμηλό. Η κοινωνική διαστρωμάτωση είναι πιθανό να επηρεάζει τις εξαρτημένες μεταβλητές. Αυτό έχει ως αποτέλεσμα η μέση συσχέτιση (*intra class correlation*) των εξεταζόμενων μεταβλητών στους μαθητές του ιδίου σχολείου να είναι μεγαλύτερη από την αντίστοιχη μέση συσχέτιση μετρημένη σε μαθητές από διαφορετικά σχολεία (Hox, 1995). Αν στο προηγούμενο παράδειγμα οι μαθητές αποτελούν τις μονάδες ενός πειραματικού σχεδιασμού, τότε το σχολείο τους μπορεί να θεωρηθεί ως παράγοντας *block*. Σε μια άλλη περίπτωση, στα γεωργικά πειράματα που εγκαθίστανται σε αγρούς, είναι δυνατό εξωγενείς παράγοντες, όπως η κλίση, τα θρεπτικά συστατικά και η υγρασία του εδάφους, να επηρεάσουν την παραγωγή και την ποιότητα της φυτικής παραγωγής. Αν οι παράγοντες αυτοί δεν ληφθούν υπόψη κατά το σχεδιασμό της μελέτης, τότε η δράση τους θα είναι συσκοτισμένη με αυτή των πειραματικών αγωγών.

Οι κλασικοί στατιστικοί έλεγχοι υποθέσεων βασίζονται στην υπόθεση της ανεξαρτησίας των παρατηρήσεων. Αν αυτό δεν ισχύει, τότε τα αποτελέσματα των ελέγχων είναι αναξιόπιστα, ενώ το μέγεθος και η κατεύθυνση της μεροληψίας είναι

απρόβλεπτη. Περιπτώσεις, στις οποίες η προϋπόθεση της ανεξαρτησίας των μετρήσεων δεν ικανοποιείται, παρουσιάζονται και στην Κατά Συστάδες Δειγματοληψία Πολλών Επιπέδων, η οποία εφαρμόζεται κυρίως στις έρευνες επισκόπησης (βλέπε Thomas, DiPrete & Forristal 1994, Hox 1995, Goldstein 1999 και 1991, Guo & Zhao 2000). Συχνά, οι δειγματοληπτικές μονάδες δεν επιλέγονται με τυχαία διαδικασία αλλά συμπεριλαμβάνονται στο δείγμα συμπτωματικά γιατί απλά συνέβη να βρίσκονται στις επιλεγμένες συστάδες.

Με βάση τις παραπάνω επισημάνσεις, κατά την εφαρμογή της ΠΑΑ σε κατηγορικά δεδομένα που έχουν παραχθεί από Τυχαιοποιημένο σε Πλήρη Συγκροτήματα Σχέδιο, θα πρέπει να ληφθεί υπόψη η ομαδοποίηση των πειραματικών μονάδων και να απομονωθεί η επίδρασή της στη μεταβλητότητα των εξαρτημένων μεταβλητών. Έστω, λοιπόν, ότι οι N διαθέσιμες πειραματικές μονάδες έχουν ομαδοποιηθεί σε c συγκροτήματα με τον περιορισμό κάθε ένα από αυτά να περιλαμβάνει N/c μονάδες. Χωρίς περιορισμό της γενικότητας, θεωρούμε ότι στο σχεδιασμό εμπλέκονται δύο παράγοντες A και B, με k και m στάθμες αντίστοιχα, και q σε πλήθος εξαρτημένες μεταβλητές Y_i με l_i κλάσεις η κάθε μία ($\sum_{i=1}^{q} l_i = j$). Μέσα σε κάθε συγκρότημα οι $k \times m$ πειραματικές αγωγές τυχαιοποιούνται πλήρως, σύμφωνα με το σχέδιο που παρουσιάσαμε στην προηγούμενη ενότητα. Πρακτικά αυτό σημαίνει ότι το τυχαιοποιημένο σχέδιο με δύο παράγοντες επαναλαμβάνεται c φορές, με διαφορετική τυχαιοποίηση των πειραματικών μονάδων στις αγωγές σε κάθε επανάληψη. Χρησιμοποιώντας τους συμβολισμούς της προηγούμενης ενότητας, μια πρώτη απόπειρα για την ανάλυση των δεδομένων βασίζεται στην εφαρμογή της ΠΑΑ στον παρακάτω $N \times j$ πίνακα τύπου "στοίβας":

$$\mathbf{P}_{\text{Int}} = \begin{bmatrix} \mathbf{P}^{(1)}_{\mathbf{A\times B}} \\ \mathbf{P}^{(2)}_{\mathbf{A\times B}} \\ \vdots \\ \mathbf{P}^{(c)}_{\mathbf{A\times B}} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_{\mathbf{A\times B}} \left(\mathbf{C}^{\mathbf{T}}_{\mathbf{A\times B}} \mathbf{C}_{\mathbf{A\times B}} \right)^{-1} \mathbf{C}^{\mathbf{T}}_{\mathbf{A\times B}} \mathbf{Z}_{\mathbf{Y}(\text{block}=1)} \\ \mathbf{C}_{\mathbf{A\times B}} \left(\mathbf{C}^{\mathbf{T}}_{\mathbf{A\times B}} \mathbf{C}_{\mathbf{A\times B}} \right)^{-1} \mathbf{C}^{\mathbf{T}}_{\mathbf{A\times B}} \mathbf{Z}_{\mathbf{Y}(\text{block}=2)} \\ \vdots \\ \mathbf{C}_{\mathbf{A\times B}} \left(\mathbf{C}^{\mathbf{T}}_{\mathbf{A\times B}} \mathbf{C}_{\mathbf{A\times B}} \right)^{-1} \mathbf{C}^{\mathbf{T}}_{\mathbf{A\times B}} \mathbf{Z}_{\mathbf{Y}(\text{block}=c)} \end{bmatrix},$$

όπου $\mathbf{P}^{(w)}_{\mathbf{A\times B}}$ είναι ο $(N/c)\times j$ πίνακας που περιγράφει στο συγκρότημα w (με $w = 1, \ldots, c$) τα προφίλ των πειραματικών μονάδων ως προς τις εξαρτημένες

μεταβλητές στο χώρο σχεδιασμού της αλληλεπίδρασης $A\times B$ των δύο παραγόντων και $\mathbf{Z}_{\mathbf{Y}(\mathbf{block}=w)}$ είναι ο λογικός πίνακας των εξαρτημένων μεταβλητών $Y_i \ \left(i=1,\ldots q\right)$ στο συγκρότημα *w*.

Οι γραμμές του πίνακα $\mathbf{P}_{\text{Int}}$ αντιστοιχούν στις πειραματικές μονάδες, όπως αυτές κατανέμονται στους συνδυασμούς των επιπέδων των παραγόντων *A* και *B* με τα συγκροτήματα σχεδιασμού. Αν συμβολίσουμε με *C* τη μεταβλητή οι τιμές της οποίας δηλώνουν το συγκρότημα πειραματισμού, τότε μέσω του $\mathbf{P}_{\text{Int}}$ τα προφίλ των πειραματικών μονάδων περιγράφονται στο χώρο σχεδιασμού της αλληλεπίδρασης δεύτερης τάξης $A\times B\times C$ των τριών μεταβλητών. Ειδικότερα, η μορφή του $kmc\times j$ συμπτυγμένου πίνακα που αντιστοιχεί στον $\mathbf{P}_{\text{Int}}$

$$\begin{bmatrix} f_{11}^{(A\times B\times C)Y_1} & \cdots & f_{1l_1}^{(A\times B\times C)Y_1} & f_{11}^{(A\times B\times C)Y_2} & \cdots & f_{1l_2}^{(A\times B\times C)Y_2} & \cdots & f_{11}^{(A\times B\times C)Y_q} & \cdots & f_{1l_q}^{(A\times B\times C)Y_q} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{(kmc)1}^{(A\times B\times C)Y_1} & \cdots & f_{(kmc)l_1}^{(A\times B\times C)Y_1} & f_{(kmc)1}^{(A\times B\times C)Y_2} & \cdots & f_{(kmc)l_2}^{(A\times B\times C)Y_2} & \cdots & f_{(kmc)1}^{(A\times B\times C)Y_q} & \cdots & f_{(kmc)l_q}^{(A\times B\times C)Y_q} \end{bmatrix}$$

αιτιολογεί το προηγούμενο συμπέρασμα.

Η προσέγγιση αυτή είναι ισοδύναμη με την προσθήκη στο σχεδιασμό μιας τρίτης ανεξάρτητης μεταβλητής *C* με *c* επίπεδα και την επιβολή του περιορισμού οι κλάσεις των εξαρτημένων μεταβλητών να λάβουν την ίδια ποσοτικοποίηση[39] για όλα τα συγκροτήματα (βλέπε Van Buuren & De Leeuw 1992, Michailidis 1996, Michailidis & De Leeuw 2000). Όμως, βασικός σκοπός του συγκεκριμένου πειράματος είναι να απομονωθεί η επίδραση της ομαδοποίησης των πειραματικών μονάδων και όχι να συνδυαστεί με την δράση των ανεξάρτητων μεταβλητών *A* και *B*. Η επίδραση των συγκροτημάτων στις εξαρτημένες μεταβλητές, αν όντως υπάρχει, θα πρέπει να είναι ανεξάρτητη από αυτή των παραγόντων. Από την ανάλυση του πίνακα $\mathbf{P}_{\text{Int}}$ το πιο πιθανό είναι να αναδειχθούν οι διαφορές μεταξύ των συγκροτημάτων (Michailidis, 1996), αποτέλεσμα που δεν παρουσιάζει ενδιαφέρον. Συνεπώς, ο πίνακας $\mathbf{P}_{\text{Int}}$ (ή η συμπτυγμένη του εκδοχή) δεν προσφέρεται για την εφαρμογή της ΠΑΑ στα δεδομένα του πειραματικού σχεδιασμού που εξετάζουμε. Η λύση για την εξεύρεση του πίνακα,

[39] Δηλαδή, οι αντίστοιχες προβολές επί των παραγοντικών αξόνων να συμπίπτουν.

ο οποίος θα δοθεί ως είσοδος στην ΠΑΑ, στηρίζεται στην ίδια την κατάστρωση του συγκεκριμένου σχεδίου, η οποία συνίσταται στην εκτέλεση *c* ανεξάρτητων τυχαιοποιημένων παραγοντικών πειραμάτων (Kutner *et al*., 2005). Η διαπίστωση αυτή οδηγεί στο συμπέρασμα ότι η ΠΑΑ θα πρέπει να εφαρμοστεί ξεχωριστά στους πίνακες $\mathbf{P}_{\mathbf{A\times B}}^{(w)}$ μέσα σε κάθε συγκρότημα. Ένας γενικός τρόπος για να εφαρμόσουμε ταυτόχρονα την ΠΑΑ σε *c* διαφορετικούς πίνακες συμπτώσεων $\mathbf{F}_s$ $(s=1,\ldots,c)$, της μορφής «*αντικείμενα*×*ιδιότητες*», είναι να κατασκευάσουμε αρχικά τον $c\times c$ block διαγώνιο πίνακα[40]:

$$\begin{bmatrix} \mathbf{F}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{F}_c \end{bmatrix},$$

όπου **0** είναι πίνακες με στοιχεία ίσα με 0 και, στη συνέχεια, να εφαρμόσουμε την ΠΑΑ επί αυτού (Michailidis 1996, Michailidis & De Leeuw 2000).

Αν στηριχθούμε στην παραπάνω μέθοδο, τότε ο πίνακας που μπορεί να δοθεί ως είσοδος στην ΠΑΑ για την ανάλυση του τυχαιοποιημένου πειραματικού σχεδίου σε συγκροτήματα έχει τη μορφή:

$$\mathbf{P}_{\text{proj}} = \begin{bmatrix} \mathbf{P}_{\mathbf{A\times B}}^{(1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{\mathbf{A\times B}}^{(2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{P}_{\mathbf{A\times B}}^{(c)} \end{bmatrix} =$$

$$= \begin{bmatrix} \mathbf{C}_{\mathbf{A\times B}}\left(\mathbf{C}_{\mathbf{A\times B}}^{\mathbf{T}}\mathbf{C}_{\mathbf{A\times B}}\right)^{-1}\mathbf{C}_{\mathbf{A\times B}}^{\mathbf{T}}\mathbf{Z}_{\mathbf{Y}(\text{block}=1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{\mathbf{A\times B}}\left(\mathbf{C}_{\mathbf{A\times B}}^{\mathbf{T}}\mathbf{C}_{\mathbf{A\times B}}\right)^{-1}\mathbf{C}_{\mathbf{A\times B}}^{\mathbf{T}}\mathbf{Z}_{\mathbf{Y}(\text{block}=2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{C}_{\mathbf{A\times B}}\left(\mathbf{C}_{\mathbf{A\times B}}^{\mathbf{T}}\mathbf{C}_{\mathbf{A\times B}}\right)^{-1}\mathbf{C}_{\mathbf{A\times B}}^{\mathbf{T}}\mathbf{Z}_{\mathbf{Y}(\text{block}=c)} \end{bmatrix}$$

[40] Διαγώνιοι block πίνακες ανάλογης μορφής χρησιμοποιούνται και για την ταυτόχρονη ανάλυση δύο ή περισσοτέρων υποδειγμάτων γραμμικής παλινδρόμησης (Mardia, Kent & Bibby, 2003).

Η εφαρμογή της ΠΑΑ στον πίνακα $\mathbf{P}_{\text{proj}}$ ή στον αντίστοιχο συμπτυγμένο παρουσιάζει στην πράξη δύο δυσκολίες:

1) Από την ανάλυση θα προκύψουν *c*-1 χαρακτηριστικές τιμές ίσες με 1 (Greenacre, 1984).

2) Δεν είναι εύκολο να διακρίνουμε σε ποιους παραγοντικούς άξονες αντιστοιχούν οι *c* αναλύσεις.

Ο ασφαλέστερος τρόπος είναι να εφαρμόσουμε ανεξάρτητα την ΠΑΑ σε κάθε έναν από τους πίνακες $\mathbf{P}^{(w)}_{\text{A×B}}$ ή στους αντίστοιχους συμπτυγμένους πίνακες "φέτα". Το γενικό πρόβλημα που προκύπτει είναι ότι τα γραφικά και αριθμητικά αποτελέσματα από τις *c* αναλύσεις δεν είναι άμεσα συγκρίσιμα, αφού ο προσανατολισμός του συστήματος συντεταγμένων, επί των παραγοντικών επιπέδων, είναι αυθαίρετος και εν γένει διαφορετικός σε κάθε ανάλυση (βλέπε Παρατήρηση Γ της Ενότητας 2.3.4.2). Συνεπώς, δεν μπορούμε να συγκρίνουμε άμεσα τις δομές των σχέσεων που αναδεικνύονται σε κάθε ένα από τους πίνακες $\mathbf{P}^{(w)}_{\text{A×B}}$. Ειδικότερα, οι σχετικές θέσεις των προφίλ των πειραματικών αγωγών μπορεί να είναι παρόμοιες σε κάποιες από τις *c* αναλύσεις, αλλά αυτό να μην είναι άμεσα αντιληπτό. Λύση στο προηγούμενο πρόβλημα μπορεί να δώσει η «Προκρούστια Ορθογώνια Προβολή ή Περιστροφή - *Procrustes Orthogonal Projection or Rotation*». Σύμφωνα με τη μέθοδο αυτή, δοθέντων δύο *μ*×*ν* πινάκων **A** και **B**, εκ των οποίων ο ένας, έστω ο **B**, θεωρείται σταθερός, προσδιορίζεται *ν*×*ν* πίνακας μετασχηματισμού **T**, τέτοιος ώστε ο πίνακας **AT** να προσεγγίζει βέλτιστα, με την έννοια των ελαχίστων τετραγώνων, τον **B** (Schönemann 1966, Golub & Van Loan 1989, Mardia, Kent & Bibby 2003). Χωρίς περιορισμό της γενικότητας, θεωρούμε ότι οι στήλες των πινάκων **A** και **B** έχουν μέσο όρο ίσο με μηδέν. Σε αντιστοιχία με τον πειραματικό σχεδιασμό που εξετάζουμε, το ζητούμενο είναι οι χώροι των λύσεων, οι οποίοι προκύπτουν από την εφαρμογή της ΠΑΑ στους *c* πίνακες $\mathbf{P}^{(w)}_{\text{A×B}}$, να προβληθούν βέλτιστα μετά από ορθογώνια περιστροφή των αντίστοιχων παραγοντικών αξόνων σε έναν κοινό χώρο "στόχο" – αναφοράς (*target or reference space*). Λόγω του γεγονότος ότι οι πίνακες $\mathbf{P}^{(w)}_{\text{A×B}}$ έχουν τον ίδιο αριθμό γραμμών και στηλών, ο μέγιστος αριθμός παραγοντικών

αξόνων *p*, που μπορούν να προκύψουν από την εφαρμογή της ΠΑΑ, είναι ίδιος σε κάθε μία από τις *c* αναλύσεις. Άρα, οι αντίστοιχοι χώροι έχουν την ίδια μέγιστη διάσταση. Επίσης, οι στήλες των πινάκων $\mathbf{P}_{\mathbf{A\times B}}^{(w)}$ έχουν μέσο όρο ίσο με 0 και εκφράζουν τις αποκλίσεις των αντίστοιχων σημείων από το κέντρο βάρους τους (βλέπε Ενότητα 2.2.14, Παρατήρηση 2.2 της Ενότητας 2.2.14.1 και Παρατήρηση Α της Ενότητας 2.3.4.2). Επομένως, τα συστήματα συντεταγμένων των χώρων που ορίζονται από τις στήλες των $\mathbf{P}_{\mathbf{A\times B}}^{(w)}$ έχουν κοινή αρχή. Συνεπώς, οποιαδήποτε από τις *c* λύσεις της ΠΑΑ μπορεί να χρησιμοποιηθεί ως χώρος αναφοράς, επί του οποίου θα προβληθούν οι υπόλοιποι *c*-1 χώροι. Πιο συγκεκριμένα, έστω $\mathbf{X}_t$ και $\mathbf{X}_w$ οι πίνακες με στοιχεία τις παραγοντικές συντεταγμένες των πειραματικών μονάδων (ή των κλάσεων των εξαρτημένων και ανεξάρτητων μεταβλητών) στο συγκρότημα αναφοράς *t* και στο συγκρότημα *w* αντίστοιχα $\left(t, w = 1,\ldots, c \text{ με } t \neq w\right)$. Οι $\mathbf{X}_t$ και $\mathbf{X}_w$ προκύπτουν από την εφαρμογή της ΠΑΑ αντίστοιχα στους πίνακες $\mathbf{P}_{\mathbf{A\times B}}^{(t)}$ και $\mathbf{P}_{\mathbf{A\times B}}^{(w)}$. Το πρόβλημα είναι να προσδιοριστεί πίνακας μετασχηματισμού **R**, τέτοιος ώστε ο πίνακας $\mathbf{X}_w\mathbf{R}$, δηλαδή ο πίνακας με στοιχεία τις παραγοντικές συντεταγμένες των πειραματικών μονάδων μετά την περιστροφή μέσω του **R**, να είναι όσο το δυνατόν πλησιέστερα στον πίνακα $\mathbf{X}_t$. Σύμφωνα με την Προκρούστια μέθοδο πρέπει να ελαχιστοποιηθεί η ποσότητα:

$$D = trace\left[\left(\mathbf{X}_t - \mathbf{X}_w\mathbf{R}\right)^{\mathbf{T}}\left(\mathbf{X}_t - \mathbf{X}_w\mathbf{R}\right)\right],$$

όπου **R** είναι ο ζητούμενος ορθογώνιος $p \times p$ πίνακας περιστροφής με:

$$\mathbf{R}^{\mathbf{T}}\mathbf{R} = \mathbf{R}\mathbf{R}^{\mathbf{T}} = \mathbf{I}_p.$$

Πρόκειται για ένα κλασικό πρόβλημα ελαχίστων τετραγώνων και η λύση του επιτυγχάνεται με την παρακάτω διαδικασία (Schönemann 1966, Sibson 1978, Rao 1980, Krzanowski 1987β, Mardia, Kent & Bibby 2003, Andrade *et al*. 2004)[41]:

---

[41] Η διαδικασία στηρίζεται στο ότι το κλασικό πρόβλημα ελαχίστων τετραγώνων $\min\|\mathbf{Ax}\text{-}\mathbf{b}\|^2$ έχει δύο ισοδύναμες λύσεις (βλέπε Kalman, 1996): $\mathbf{x} = \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathbf{T}}\mathbf{b}$ και $\mathbf{x} = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^{\mathbf{T}}\mathbf{b}$.

Αρχικά εφαρμόζεται ή μέθοδος SVD (βλέπε Ενότητα 2.2.11) στον πίνακα $\mathbf{X}_w^{\mathbf{T}}\mathbf{X}_t$:

$$\mathbf{X}_w^{\mathbf{T}}\mathbf{X}_t = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathbf{T}}.$$

Στη συνέχεια, ο πίνακας περιστροφής **R** υπολογίζεται από τη σχέση:

$$\mathbf{R} = \mathbf{U}\mathbf{V}^{\mathbf{T}}.$$

Μέσω της εφαρμογής της SVD στον πίνακα $\mathbf{X}_w^{\mathbf{T}}\mathbf{X}_t$ είναι δυνατό να επιτευχθεί περιστροφή, τάνυσμα ή/και ανάκλαση των στοιχείων του (Weller & Romney, 1990). Ο πίνακας μετασχηματισμού $\mathbf{R} = \mathbf{U}\mathbf{V}^{\mathbf{T}}$ περιστρέφει ή/και ανακλά τα στοιχεία του πίνακα $\mathbf{X}_w$ στο χώρο στηλών του $\mathbf{X}_t$ έτσι ώστε να ελαχιστοποιείται η ποσότητα *D*. Αν η προσαρμογή είναι τέλεια, τότε τα στοιχεία του πίνακα $\mathbf{X}_t - \mathbf{X}_w\mathbf{R}$, ο οποίος μετρά την απώλεια προσαρμογής μετά την περιστροφή, θα είναι ίσα με μηδέν με αποτέλεσμα και *D*=0. Η ποσότητα *D* θεωρείται ως απόσταση ή δείκτης καλής προσαρμογής και η φυσική του ερμηνεία είναι ότι εκφράζει το βαθμό ομοιότητας των προτύπων και των δομικών σχέσεων που αναδεικνύονται από την εφαρμογή της ΠΑΑ στους πίνακες $\mathbf{P}^{(t)}_{\text{A×B}}$ και $\mathbf{P}^{(w)}_{\text{A×B}}$. Δύο τυποποιημένες εκδοχές της απόστασης *D* αποτελούν οι δείκτες $D^{*} = D/N$ και $D^{**} = D/trace\left(\mathbf{X}_t\mathbf{X}_t^{\mathbf{T}}\right)$ (Sibson, 1978).

Κατά την εφαρμογή της ΠΑΑ τα στοιχεία κάθε στήλης των πινάκων $\mathbf{X}_t$ και $\mathbf{X}_w$ είναι κεντροποιημένα εκ κατασκευής, ως προς τους μέσους όρους των στηλών και, επομένως, τα συστήματα συντεταγμένων των αντίστοιχων χώρων, οι οποίοι ορίζονται από τις στήλες τους, έχουν κοινή αρχή. Έτσι, δεν απαιτείται μεταφορά του συστήματος συντεταγμένων του χώρου που ορίζουν οι στήλες του πίνακα $\mathbf{X}_w$. Στο πλαίσιο της Γαλλικής Σχολής, οι παραγοντικές συντεταγμένες που συγκροτούν τα στοιχεία στηλών των πινάκων $\mathbf{X}_t$ και $\mathbf{X}_w$ υπολογίζονται με Κύρια Κανονικοποίηση - *PN* (βλέπε Τέταρτο Βήμα της Ενότητας 2.2.14). Αυτό έχει ως αποτέλεσμα το άθροισμα τετραγώνων των συντεταγμένων σε κάθε άξονα, δηλαδή οι διακυμάνσεις των στηλών των $\mathbf{X}_t$ και $\mathbf{X}_w$, να είναι ίσες με τις αδράνειες των αντίστοιχων παραγοντικών αξόνων (βλέπε Ενότητα 2.2.14.1 και Παρατήρηση 2.2). Επομένως, το νέφος σημείων του $\mathbf{X}_w$ κατά την προβολή του στο χώρο αναφοράς $\mathbf{X}_t$ θα εμφανίζει διαφορετική διασπορά απ' ότι το νέφος των στοιχείων του $\mathbf{X}_t$, αφού κάθε

ορθογώνιος μετασχηματισμός διατηρεί τις αποστάσεις (Peay 1988, Groenen & Frances 2000). Για να είναι συγκρίσιμα τα αποτελέσματα των δύο λύσεων θα πρέπει να ληφθεί υπόψη και η διαφορετική μεταβλητότητα των αντίστοιχων νεφών (Sibson 1978, Mardia, Kent & Bibby 2003). Ένας απλός τρόπος για να αντιμετωπιστεί το ζήτημα είναι τα στοιχεία στηλών των πινάκων $\mathbf{X}_t$ και $\mathbf{X}_w$, πριν την Προκρούστια Προβολή, να μετασχηματιστούν έτσι ώστε η διακύμανσή τους να είναι ίση με 1 (Sibson 1978, Peay 1988, Krzanowski 1993), δηλαδή να χρησιμοποιηθούν οι τυποποιημένες συντεταγμένες αντί για τις κύριες. Εναλλακτικά, στη γενική περίπτωση όπου τα στοιχεία των πινάκων $\mathbf{X}_t$ και $\mathbf{X}_w$ έχουν διαφορετικές κλίμακες μέτρησης, η συμβατότητα των δύο λύσεων μπορεί να επιτευχθεί με την επιβολή του περιορισμού τα στοιχεία του $\mathbf{X}_w$, μετά την περιστροφή, να ικανοποιούν τη σχέση:

$$z\mathbf{X}_w\mathbf{R}\text{, με } z>0. \qquad [7.4]$$

Μπορεί να δειχθεί (Sibson 1978, Mardia, Kent & Bibby 2003) ότι η ελαχιστοποίηση της ποσότητας *D* επιτυγχάνεται θέτοντας:

$$z = \frac{\left[trace\left(\mathbf{D}\right)\right]}{\left[trace\left(\mathbf{X}_w\mathbf{X}_w^{\mathbf{T}}\right)\right]}, \qquad [7.5]$$

όπου **D** είναι ο διαγώνιος πίνακας με στοιχεία τις χαρακτηριστικές τιμές που προκύπτουν από την εφαρμογή της SVD στον πίνακα $\mathbf{X}_w^{\mathbf{T}}\mathbf{X}_t$. Με τον τρόπο αυτό, η κλίμακα μέτρησης των στοιχείων του $\mathbf{X}_w$ μετασχηματίζεται σε αυτή των στοιχείων του $\mathbf{X}_t$. Θα πρέπει να παρατηρήσουμε ότι οι μετασχηματισμοί $z\mathbf{X}_w\mathbf{R}$ και $\mathbf{X}_w\mathbf{R}$ δεν είναι συμμετρικοί ως προς τις κλίμακες μέτρησης των στοιχείων των πινάκων $\mathbf{X}_t$ και $\mathbf{X}_w$ (Rummel 1970, Gower 1975, Sibson 1978, Mardia, Kent & Bibby 2003). Τα στοιχεία του $\mathbf{X}_t$ δεν μπορούν να προκύψουν από τον αντίστροφο μετασχηματισμό των στοιχείων του $\mathbf{X}_w$. Η συμμετρία μπορεί να επιτευχθεί με τον καθορισμό μιας κοινής κλίμακας μέτρησης, τέτοιας ώστε τα αθροίσματα τετραγώνων των στοιχείων των $\mathbf{X}_t$ και $\mathbf{X}_w$ να είναι ίσα (Mardia, Kent & Bibby, 2003). Δηλαδή να ισχύει:

$$trace\left(\mathbf{X}_w\mathbf{X}_w^{\mathbf{T}}\right) = trace\left(\mathbf{X}_t\mathbf{X}_t^{\mathbf{T}}\right).$$

Προφανώς, η χρήση των τυποποιημένων παραγοντικών συντεταγμένων ικανοποιεί την παραπάνω απαίτηση, αφού κατά την εφαρμογή της ΠΑΑ ισχύει (Gifi, 1996):

$$trace\left(\mathbf{X}_w\mathbf{X}_w^{\mathrm{T}}\right) = trace\left(\mathbf{X}_t\mathbf{X}_t^{\mathrm{T}}\right) = pN\ ,$$

όπου N είναι το πλήθος γραμμών των πινάκων $\mathbf{X}_t$ και $\mathbf{X}_w$.

Στην πράξη, αφού επιλεγεί ο πίνακας αναφοράς $\mathbf{X}_t$, οι υπόλοιποι $c-1$ πίνακες προβάλλονται, με την προτεινόμενη μεθοδολογία, στο χώρο που ορίζουν οι στήλες του $\mathbf{X}_t$. Στο πειραματικό σχέδιο που εξετάζουμε το ενδιαφέρον εστιάζεται στη σύγκριση των προφίλ των αγωγών μεταξύ των συγκροτημάτων, αφού αφαιρεθεί η (ενδεχόμενη) επίδρασή τους στη διαμόρφωση των αποτελεσμάτων. Η οπτικοποίηση της όλης διαδικασίας, επί του παραγοντικού επιπέδου, επιτυγχάνεται για p=2. Η διαγραμματική ερμηνεία των αποτελεσμάτων στηρίζεται στη σύγκριση των σχετικών θέσεων των ομόλογων σημείων προφίλ ή ομάδων τους επί του κοινού παραγοντικού επιπέδου. Τα αριθμητικά αποτελέσματα της Προκρούστιας προβολής μπορούν να εμπλουτιστούν με τον υπολογισμό των παρακάτω δεικτών (βλέπε Rummel, 1970):

α) Των συντελεστών γραμμικής συσχέτισης του *Pearson* μεταξύ των παραγοντικών αξόνων που προκύπτουν από την εφαρμογή της ΠΑΑ στους πίνακες $\mathbf{P}^{(t)}_{\mathrm{A\times B}}$ και $\mathbf{P}^{(w)}_{\mathrm{A\times B}}$ μετά την περιστροφή. Για τον υπολογισμό τους χρησιμοποιούνται οι στήλες των πινάκων $\mathbf{X}_t$ και $\mathbf{X}_w\mathbf{R}$. Ιδιαίτερο ενδιαφέρον για την ποιότητα της προσαρμογής παρουσιάζουν οι συντελεστές συσχέτισης των ομόλογων παραγοντικών αξόνων στα συγκροτήματα *t* (αναφοράς) και *w* μετά την περιστροφή. Οι τιμές τους εκφράζουν το βαθμό σύμπτωσης (*congruence*) των προτύπων ή δομών που αναδεικνύουν οι αντίστοιχοι άξονες.

β) Των συντελεστών συσχέτισης (γραμμικής του *Pearson* και τάξεων του *Spearman*) μεταξύ των ομόλογων αγωγών στα συγκροτήματα *t* και *w* μετά την περιστροφή. Για τον υπολογισμό τους χρησιμοποιούνται οι τυποποιημένες συντεταγμένες των αγωγών στους p άξονες. Οι συντεταγμένες βρίσκονται στις γραμμές των πινάκων $\mathbf{X}_t$ και $\mathbf{X}_w\mathbf{R}$. Οι τιμές των συντελεστών αυτών εκφράζουν το βαθμό ομοιότητας των προφίλ των ομόλογων αγωγών στα δύο συγκροτήματα. Χαμηλές τιμές των δεικτών

υποδεικνύουν τις αγωγές που δέχθηκαν τη μεγαλύτερη επίδραση από την ομαδοποίηση (*blocking*) των πειραματικών μονάδων. Βέβαια, το ίδιο μπορεί να διαπιστωθεί και από την εξέταση των στοιχείων του πίνακα $\mathbf{X}_t - \mathbf{X}_w\mathbf{R}$, ο οποίος περιέχει τις αποκλίσεις προσαρμογής ανά άξονα. Ο μέσος όρος (*ΜΣ*) των συντελεστών συσχέτισης μεταξύ των ομόλογων αγωγών εκφράζει ένα συνολικό δείκτη καλής προσαρμογής της Προκρούστιας προβολής.

γ) Των συντελεστών συσχέτισης (γραμμικής του *Pearson* και τάξεων του *Spearman*) μεταξύ των ομόλογων αξόνων, οι οποίοι προκύπτουν από την εφαρμογή της ΠΑΑ στο συγκρότημα *w*, πριν και μετά την περιστροφή – προβολή. Για τον υπολογισμό τους χρησιμοποιούνται οι στήλες των πινάκων $\mathbf{X}_w$ και $\mathbf{X}_w\mathbf{R}$. Οι χαμηλές τιμές και το αρνητικό πρόσημο των συντελεστών δηλώνουν μεταβολή της διάταξης των αγωγών και της δομής των αξόνων μετά την περιστροφή.

Οι δείκτες *CTR*, αν και είναι δυνατό να υπολογιστούν, ωστόσο δεν έχουν φυσική ερμηνεία και θεωρητική τεκμηρίωση (βλέπε Van de Velden, 2000), αφού σε κάθε προβολή η αδράνεια του αντίστοιχου πίνακα είτε ανακατανέμεται είτε τυποποιείται (βλέπε Greenacre, 2006).

Με την Προκρούστια προβολή – περιστροφή οι πειραματικές μονάδες εντός των συγκροτημάτων προβάλλονται σε κοινό χώρο και οι σχετικές τους θέσεις είναι συγκρίσιμες. Φυσικά, το ίδιο ισχύει και για τις κατηγορίες των μεταβλητών, ανεξάρτητων και εξαρτημένων. Την ίδια μέθοδο περιστροφής προτείνουν οι Michailidis (1996) και Michailidis και De Leeuw (2000) για τη σύγκριση των αποτελεσμάτων που προκύπτουν από την εφαρμογή της Ανάλυσης Ομοιογένειας σε *c* συστάδες δεδομένων, τα οποία έχουν συγκεντρωθεί με το δειγματοληπτικό σχήμα των Πολλαπλών Επιπέδων (*Multilevel*). Σε αντιστοιχία με την περίπτωση του σχεδιασμού που εξετάζουμε οι *c* συστάδες μπορούν να θεωρηθούν ως συγκροτήματα. Επίσης, η μέθοδος χρησιμοποιείται για τη σύγκριση των υποχώρων που προκύπτουν από την εφαρμογή της Παραγοντικής Ανάλυσης (Harman 1967, Gruvaeus 1970, Raykov & Little 1999) και της Πολυδιάστατης Κλιμάκωσης (Sibson 1978, Peay 1988) σε δύο ή περισσότερα ανεξάρτητα σύνολα δεδομένων. Βρίσκει εφαρμογή στους ελέγχους εξωτερικής και εσωτερικής σταθερότητας των αποτελεσμάτων της

ΠΑΑ όταν αυτοί πραγματοποιούνται με μεθόδους επαναδειγματοληψίας, όπως είναι η *Bootstrap* (Markus 1994α και 1994β, Greenacre 2006), καθώς και σε μεθοδολογίες για την επιλογή μεταβλητών στην Ανάλυση σε Κύριες Συνιστώσες (Krzanowski 1987β, Andrade *et al*. 2004). Γενικεύσεις και επιπλέον εφαρμογές της Προκρούστιας Προβολής συναντάμε στους Schönemann (1966), Harman (1967), Rummel (1970), Gower (1975), Ten Berge (1977), Sibson (1978), Broken (1983), Ten Berge και Knol (1984), Goodal (1991), Dijksterhuis και Gower (1991/2), Arnold και Collins (1993), Raykov και Little (1999), Balbi και Esposito (2000), Lebart, Morineau και Piron (2000), Trendafilov και Lippert (2002), Andrade *et al*. (2004), Héberger και Andrade (2004), Pagès (2004), Dijksterhuis, Martens και Martens (2005) και Gardner, Gower και Le Roux (2006).

Συμπερασματικά, με την προτεινόμενη μεθοδολογία ανάλυσης του παραγοντικού σε πλήρη συγκροτήματα σχεδιασμού επιτυγχάνονται τα εξής:

1) απομονώνεται η συνολική επίδραση της ομαδοποίησης των πειραματικών μονάδων από την δράση των παραγόντων ενδιαφέροντος,

2) δίνεται η δυνατότητα ταυτόχρονης ανάλυσης της επίδρασης των παραγόντων στις εξαρτημένες μεταβλητές μέσα σε κάθε συγκρότημα,

3) τα συγκροτήματα ομογενοποιούνται, είναι συγκρίσιμα και επιπλέον αναδεικνύονται διαγραμματικά οι μεταξύ τους διαφοροποιήσεις, ως προς τις σχέσεις των προφίλ των πειραματικών αγωγών, που ενθυλακώνουν οι πίνακες που αναλύονται. Ως μέτρο ομοιότητας μεταξύ του συγκροτήματος αναφοράς και των προβαλλόμενων επί αυτού συγκροτημάτων μπορεί να χρησιμοποιηθεί η Προκρούστια απόσταση *D*, οι τυποποιημένες της εκδοχές D^{*} και D^{**} καθώς και ο μέσος όρος των συντελεστών συσχέτισης (*ΜΣ*) μεταξύ των ομόλογων αγωγών στο συγκρότημα *t* και στο *w* μετά την περιστροφή.

Στην Ενότητα Z2 του Παραρτήματος Z δίνουμε ένα παράδειγμα εφαρμογής της προτεινόμενης μεθοδολογίας σε πειραματικά δεδομένα, τα οποία συγκεντρώθηκαν από τυχαιοποιημένο σχέδιο σε δύο πλήρη συγκροτήματα με δύο παράγοντες και δύο εξαρτημένες κατηγορικές μεταβλητές.

Παρατηρήσεις

Α) Η προτεινόμενη μέθοδος ανάλυσης δεδομένων, τα οποία προέρχονται από τυχαιοποιημένο σχέδιο σε πλήρη συγκροτήματα, μπορεί να επεκταθεί και στην περίπτωση τριών ή περισσότερων ανεξάρτητων μεταβλητών με τον ίδιο τρόπο που παρουσιάσαμε στον παραγοντικό πειραματισμό της προηγούμενης ενότητας. Βέβαια, η εισαγωγή επιπλέον ανεξάρτητων μεταβλητών έχει ως συνέπεια την αύξηση του πλήθους των πειραματικών μονάδων και της πολυπλοκότητας της ερμηνείας των αποτελεσμάτων. Στην πράξη, η αλληλεπίδραση δεύτερης τάξης $A\times B\times \Gamma$ τριών παραγόντων είναι πολύ δύσκολο να αιτιολογηθεί. Στο ευρύτερο πλαίσιο των Επιστημών της Συμπεριφοράς η πολυπλοκότητα που εισάγει η αλληλεπίδραση $A\times B\times \Gamma$ στην ερμηνεία των παρατηρούμενων δομών και σχέσεων είναι συχνά μη ρεαλιστική και σπάνια αποτελεί το κέντρο του ερευνητικού ενδιαφέροντος (Cohen & Cohen, 1983).

Β) Η εφαρμογή της ΠΑΑ στους πίνακες $\mathbf{P}^{(w)}_{\mathbf{A\times B}}$ (ή στους αντίστοιχους συμπτυγμένους) μπορεί να εμπλουτιστεί με την προσθήκη επιπλέον πινάκων συμπτώσεων, οι οποίοι παρουσιάζουν ενδιαφέρον στο πλαίσιο της εκάστοτε μελέτης (βλέπε Παπαδημητρίου, 1991). Τα στοιχεία τους μπορούν να εισαχθούν στις αναλύσεις ως συμπληρωματικά (βλέπε Ενότητα 7.5).

Γ) Οι παραγοντικές συντεταγμένες των πειραματικών μονάδων και των κλάσεων των μεταβλητών, ανεξάρτητων και εξαρτημένων, υπολογίζονται όχι ως προς το κέντρο βάρους όλου του νέφους τους αλλά ως προς το κέντρο βάρους του αντίστοιχου νέφους εντός του κάθε συγκροτήματος. Αυτό έχει ως συνέπεια οι κλάσεις των εξαρτημένων μεταβλητών να λαμβάνουν διαφορετική ποσοτικοποίηση μέσα σε κάθε συγκρότημα και όχι την ίδια, όπως συμβαίνει κατά την ανάλυση του πίνακα $\mathbf{P}_{\text{Int}}$.

Δ) Με την Προκρούστια προβολή οι πίνακες $\mathbf{X}_w$ καθίστανται όσο το δυνατόν πιο "όμοιοι" με τον πίνακα αναφοράς $\mathbf{X}_t$ αλλά και μεταξύ τους (Michailidis 1996, Michailidis & De Leeuw 2000). Εξομαλύνεται η μεταβλητότητα που εισάγει η ομαδοποίηση των πειραματικών μονάδων σε συγκροτήματα και απαλείφεται η επίδρασή της από τις μεταβλητές ενδιαφέροντος (Rummel, 1970).

Ε) Η συμπτυγμένη μορφή του πίνακα $\mathbf{P}_{\text{proj}}$ είναι η παρακάτω:

$$\mathbf{P}_{\mathbf{C}} = \begin{bmatrix} \mathbf{P}_{\mathbf{C}_{\mathbf{A}\times\mathbf{B}}}^{(1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{\mathbf{C}_{\mathbf{A}\times\mathbf{B}}}^{(2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{P}_{\mathbf{C}_{\mathbf{A}\times\mathbf{B}}}^{(c)} \end{bmatrix},$$

όπου $\mathbf{P}_{\mathbf{C}_{\mathbf{A}\times\mathbf{B}}}^{(w)}$ είναι ο $km\times j$ πίνακας "φέτα" που διασταυρώνει, στο συγκρότημα w ($w=1,\ldots,c$), τη μεταβλητή αλληλεπίδρασης, η οποία αντιστοιχεί στους $k\times m$ συνδυασμούς των επιπέδων των παραγόντων A και B, με τις j κλάσεις των εξαρτημένων μεταβλητών. Η αναλυτική μορφή των πινάκων $\mathbf{P}_{\mathbf{C}_{\mathbf{A}\times\mathbf{B}}}^{(w)}$ είναι η εξής:

$$\mathbf{P}_{\mathbf{C}_{\mathbf{A}\times\mathbf{B}}}^{(w)} =$$

$$= \begin{bmatrix} f_{11}^{(A\times B)Y_1} & \cdots & f_{1l_1}^{(A\times B)Y_1} & f_{11}^{(A\times B)Y_2} & \cdots & f_{1l_2}^{(A\times B)Y_2} & \cdots & f_{11}^{(A\times B)Y_q} & \cdots & f_{1l_q}^{(A\times B)Y_q} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{(km)1}^{(A\times B)Y_1} & \cdots & f_{(km)l_1}^{(A\times B)Y_1} & f_{(km)1}^{(A\times B)Y_2} & \cdots & f_{(km)l_2}^{(A\times B)Y_2} & \cdots & f_{(km)1}^{(A\times B)Y_q} & \cdots & f_{(km)l_q}^{(A\times B)Y_q} \end{bmatrix}_{(\text{block}=w)}.$$

ΣΤ) Στην περίπτωση που η προτεινόμενη μέθοδος εφαρμοστεί στη συμπτυγμένη εκδοχή του πίνακα $\mathbf{P}_{\text{proj}}$, τότε ως πίνακας αναφοράς μπορεί να επιλεγεί ο $\mathbf{X}_t$, ο οποίος προκύπτει από την εφαρμογή της ΠΑΑ στον $km\times j$ πίνακα "φέτα":

$$\mathbf{P}_{\mathbf{A}\times\mathbf{B}} = \mathbf{C}_{\mathbf{A}\times\mathbf{B}}\left(\mathbf{C}_{\mathbf{A}\times\mathbf{B}}^{\mathbf{T}}\mathbf{C}_{\mathbf{A}\times\mathbf{B}}\right)^{-1}\mathbf{C}_{\mathbf{A}\times\mathbf{B}}^{\mathbf{T}}\mathbf{Z}_{\mathbf{Y}},$$

που διασταυρώνει τη μεταβλητή αλληλεπίδρασης, η οποία αντιστοιχεί στους $k\times m$ συνδυασμούς των επιπέδων των παραγόντων A και B, με τις j κλάσεις των εξαρτημένων μεταβλητών στο σύνολο των παρατηρήσεων του πειραματισμού. Στην παραπάνω σχέση, $\mathbf{Z}_{\mathbf{Y}}$ είναι ο $N\times j$ λογικός πίνακας 0-1 των εξαρτημένων μεταβλητών που προκύπτει αν δεν ληφθεί υπόψη η ομαδοποίηση των πειραματικών μονάδων σε c συγκροτήματα. Ο πίνακας $\mathbf{P}_{\mathbf{A}\times\mathbf{B}}$ αντιστοιχεί στον πίνακα άθροισμα των c συμπτυγμένων πινάκων $\mathbf{P}_{\mathbf{C}_{\mathbf{A}\times\mathbf{B}}}^{(w)}$ και εκφράζει την περιθώρια ή, αλλιώς, τη μέση κατανομή των $k\times m$ πειραματικών αγωγών (βλέπε Παπαδημητρίου, 1991). Αν οι c

σε πλήθος πίνακες $\mathbf{X}_w$ προβληθούν με την Προκρούστια μέθοδο στο χώρο που ορίζουν οι στήλες του πίνακα αναφοράς $\mathbf{X}_t$, τότε επί των παραγοντικών επιπέδων είναι δυνατό να μελετήσουμε όχι μόνο τη συμπεριφορά των κέντρων βάρους των πειραματικών αγωγών αλλά και τις ομοιότητες ή αντιθέσεις των προφίλ των αγωγών στα *c* συγκροτήματα ως προς το κέντρο βάρους τους.

Z) Η μέθοδος που παρουσιάσαμε δεν είναι εφικτό να υλοποιηθεί άμεσα αν οι πίνακες $\mathbf{P}^{(w)}_{\mathrm{A\times B}}$ δεν έχουν το ίδιο πλήθος γραμμών[42], δηλαδή όταν τα συγκροτήματα δεν περιέχουν τον ίδιο αριθμό πειραματικών μονάδων. Εκ κατασκευής όλοι οι πίνακες $\mathbf{P}^{(w)}_{\mathrm{A\times B}}$ έχουν *j* στήλες. Όμως, μπορεί να εφαρμοστεί στους συμπτυγμένους πίνακες $\mathbf{P}^{(w)}_{\mathbf{C}_{\mathrm{A\times B}}}$ οι οποίοι έχουν το ίδιο πλήθος γραμμών και στηλών. Έτσι, αν δεν είναι επιθυμητό να χρησιμοποιήσουμε τις παραγοντικές συντεταγμένες των αντικειμένων σε άλλες στατιστικές αναλύσεις, η προτεινόμενη μέθοδος μπορεί να εφαρμοστεί στις συμπτυγμένες μορφές των πινάκων $\mathbf{P}^{(w)}_{\mathrm{A\times B}}$. Σε αντίθετη περίπτωση, μπορούμε να εφαρμόσουμε τη μεθοδολογία που προτείναμε στην Ενότητα 3.3, σύμφωνα με την οποία οι κύριες συντεταγμένες των πειραματικών μονάδων υπολογίζονται ως οι μέσοι όροι των τυποποιημένων συντεταγμένων των ιδιοτήτων (κλάσεων των μεταβλητών) από τις οποίες χαρακτηρίζονται. Εναλλακτικά, αν για παράδειγμα ο πίνακας $\mathbf{X}_w$ έχει λιγότερες γραμμές από τον $\mathbf{X}_t$, τότε οι υπολειπόμενες γραμμές του $\mathbf{X}_w$ μπορούν να συμπληρωθούν με μηδενικά, ώστε τελικά οι δύο πίνακες να έχουν τον ίδιο αριθμό γραμμών (Mardia, Kent & Bibby, 2003). Με τους παραπάνω τρόπους είναι δυνατό να αναλυθούν μη ισορροπημένοι πειραματικοί σχεδιασμοί σε συγκροτήματα καθώς και δεδομένα που έχουν συγκεντρωθεί με την Κατά Συστάδες Τυχαία Δειγματοληψία σε έρευνες επισκόπησης, στις οποίες υπάρχει εννοιολογική διάκριση μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών. Σε κάθε συστάδα - συγκρότημα το πλήθος των δειγματοληπτικών μονάδων είναι, εν γένει, διαφορετικό.

H) Στη γενική περίπτωση όπου δεν υπάρχει διαχωρισμός σε εξαρτημένες και ανεξάρτητες μεταβλητές και τα διαθέσιμα αντικείμενα μπορούν να ομαδοποιηθούν σε

---

[42] Πιο συγκεκριμένα, δεν μπορεί να εφαρμοστεί άμεσα η Προκρούστια Προβολή, αφού οι πίνακες δεν έχουν το ίδιο μέγεθος.

c συγκροτήματα, τότε η προτεινόμενη μέθοδος μπορεί να εφαρμοστεί στους υποπίνακες του πίνακα:

$$\mathbf{B}^{*} = \begin{bmatrix} \mathbf{B}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{B}_c \end{bmatrix},$$

όπου $\mathbf{B}_w$ είναι ο πίνακας *Burt* που αντιστοιχεί στο συγκρότημα *w* ($w = 1, \ldots, c$).

Πιο συγκεκριμένα, αν $\mathbf{X}_t$ είναι ο πίνακας αναφοράς, οι στήλες του οποίου περιέχουν τις παραγοντικές συντεταγμένες των *j* κλάσεων των *q* μεταβλητών, που υπολογίζονται από την εφαρμογή της ΠΑΑ στον πίνακα *Burt* του συγκροτήματος *t*, τότε οι χώροι που ορίζονται από τις στήλες των υπόλοιπων *c*-1 σε πλήθος πινάκων $\mathbf{X}_w$ προβάλλονται με την Προκρούστια μέθοδο στο χώρο στηλών του $\mathbf{X}_t$. Βέβαια, εκτός των πινάκων $\mathbf{B}_w$ μπορούν να χρησιμοποιηθούν και άλλοι πίνακες εισόδου που πληρούν τις προϋποθέσεις εφαρμογής της ΠΑΑ (βλέπε Ενότητα 2.4) αρκεί να έχουν το ίδιο πλήθος γραμμών και στηλών. Ιδιαίτερο ενδιαφέρον παρουσιάζουν οι πίνακες, στους οποίους αναφερθήκαμε στην Ενότητα Β4.1 του Παραρτήματος Β. Αν υπάρχουν *c* σε πλήθος συγκεντρωτικοί πίνακες $\mathbf{K}_w$, τότε η ανάλυση μπορεί να εφαρμοστεί στους υποπίνακες του πίνακα:

$$\mathbf{K}^{*} = \begin{bmatrix} \mathbf{K}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{K}_c \end{bmatrix},$$

όπου $\mathbf{K}_w$ είναι ο συγκεντρωτικός πίνακας που αντιστοιχεί στο συγκρότημα *w* ($w = 1, \ldots, c$). Μέσω του παραπάνω πίνακα, οι κοινές διακεκριμένες τιμές των μεταβλητών αποκτούν την ίδια βέλτιστη ποσοτικοποίηση για όλες τις μεταβλητές μέσα σε κάθε συγκρότημα αλλά διαφορετική μεταξύ των συγκροτημάτων.

Θ) Μια άλλη ενδιαφέρουσα κατάσταση προκύπτει όταν έχουμε να αναλύσουμε *c* απλούς πίνακες συμπτώσεων απολύτων συχνοτήτων δύο κατηγορικών μεταβλητών,

όπου ο διαχωρισμός των πινάκων σε συγκροτήματα δεν έγινε με τη θέληση του ερευνητή αλλά επιβλήθηκε από περιορισμούς του δειγματοληπτικού σχεδιασμού ή της κατανομής (χωρικής, χρονικής) των δειγματοληπτικών μονάδων. Αν το ενδιαφέρον της μελέτης εστιάζεται στη διερεύνηση της σχέσης μεταξύ των δύο μεταβλητών με όσο το δυνατό μεγαλύτερη ακρίβεια, τότε είναι σκόπιμο να ελεγχθεί η ομοιογένεια των επιμέρους *c* πινάκων. Αν διαπιστωθεί ότι οι ομόλογες γραμμές ή στήλες των πινάκων έχουν παρόμοια προφίλ σε κάθε συγκρότημα, τότε η ΠΑΑ μπορεί να εφαρμοστεί στον απλό πίνακα συμπτώσεων τα στοιχεία του οποίου είναι το άθροισμα των ομόλογων στοιχείων των *c* πινάκων. Σύμφωνα με τον Everitt (1979), ο πιο απλός τρόπος για να ελεγχθεί αν η ακρίβεια και η ευαισθησία στη διερεύνηση της σχέσης των δύο μεταβλητών αυξάνεται με τη συνένωση των πρωτογενών δεδομένων είναι ο εξής: εφαρμόζουμε τον έλεγχο ανεξαρτησίας χ^2 στους *c* επιμέρους πίνακες και στον συγκεντρωτικό. Αν τα αποτελέσματα υποδείξουν ότι στον πίνακα άθροισμα η ένταση ή/και η στατιστική σημαντικότητα της σχέσης μεταξύ των δύο μεταβλητών είναι μεγαλύτερη απ' ότι κατά την ανάλυση των επιμέρους πινάκων, τότε είναι προτιμότερο να διατηρήσουμε το συγκεντρωτικό πίνακα. Σε αντίθετη περίπτωση, θα πρέπει να πραγματοποιηθούν *c* ξεχωριστές αναλύσεις. Μια άλλη προσέγγιση για την ανάλυση του συγκεντρωτικού πίνακα βασίζεται στον υπολογισμό, αρχικά, των αναμενομένων συχνοτήτων σε κάθε έναν από τους *c* πίνακες και, στη συνέχεια, στη σύγκριση της παρατηρούμενης κατανομής των στοιχείων του συγκεντρωτικού πίνακα με την κατανομή που προκύπτει μετά την άθροιση των αντίστοιχων αναμενόμενων συχνοτήτων των ομόλογων κελιών των *c* πινάκων. Η σύγκριση μπορεί να επιτευχθεί μέσω του ελέγχου καλής προσαρμογής χ^2, σε επίπεδο σημαντικότητας *α*, με $(k-1)(l-1)$ βαθμούς ελευθερίας, όπου *k* και *l* είναι οι κλάσεις των δύο μεταβλητών (Everitt, 1979). Με τη μέθοδο αυτή ελέγχεται η συνάφεια των δύο μεταβλητών στο σύνολο των διαθέσιμων παρατηρήσεων. Επομένως, στην περίπτωση αυτή, η ΠΑΑ μπορεί να εφαρμοστεί στον πίνακα άθροισμα, με τη διαφορά ότι οι αναμενόμενες συχνότητες, που απαιτούνται για τον υπολογισμό των στοιχείων του πίνακα **S** (βλέπε Ενότητα 2.2.4 και Πρώτο Βήμα της Ενότητας 2.2.14), θα είναι ίσες με το άθροισμα των αντίστοιχων αναμενόμενων συχνοτήτων, οι οποίες εκτιμώνται κάτω από την υπόθεση της ανεξαρτησίας των δύο μεταβλητών στους *c* επιμέρους πίνακες.

7.7 Σχόλια και Συμπεράσματα Κεφαλαίου

Στο κεφάλαιο αυτό δείξαμε ότι η ΠΑΑ μπορεί να εφαρμοστεί σε δεδομένα που προέρχονται από τρεις βασικούς πειραματικούς σχεδιασμούς, στους οποίους τόσο οι ανεξάρτητες όσο και οι εξαρτημένες μεταβλητές είναι κατηγορικές. Η προτεινόμενη μεθοδολογία ακολουθεί την πορεία ανάλυσης των κλασικών πειραματικών σχεδιασμών, η οποία αποσκοπεί κυρίως στη μελέτη της κοινής επίδρασης των ανεξάρτητων μεταβλητών – παραγόντων πάνω στις εξαρτημένες. Συνδυάζει εφαρμογές και ιδιότητες των Πινάκων Σχεδιασμού και των Πινάκων Προβολής «καπέλο» καθώς και την Αρχή της Ισοδυναμίας των Σχετικών Κατανομών. Στο γενικό πλαίσιο της προσέγγισης που παρουσιάσαμε, γίνεται, αρχικά, διάκριση των μεταβλητών σε εξαρτημένες και ανεξάρτητες. Στη συνέχεια, οι πίνακες σχεδιασμού των εξαρτημένων μεταβλητών προβάλλονται (ορθογώνια) στο χώρο που ορίζουν οι στήλες του πίνακα σχεδιασμού της αλληλεπίδρασης των ανεξάρτητων. Τέλος, η ΠΑΑ εφαρμόζεται είτε στον πίνακα προβολής, ο οποίος περιγράφει το υπό εξέταση φαινόμενο στο χώρο που ορίζεται από την αλληλεπίδραση των παραγόντων, είτε στον αντίστοιχο συμπτυγμένο πίνακα, ο οποίος προκύπτει από την εφαρμογή της Αρχής της Ισοδυναμίας των Σχετικών Κατανομών. Στην περίπτωση του Τυχαιοποιημένου Σχεδίου σε Πλήρη Συγκροτήματα, η απαλοιφή της επίδρασης στη διαμόρφωση των αποτελεσμάτων, την οποία ενδεχομένως επιφέρει η ομαδοποίηση των πειραματικών μονάδων σε συγκροτήματα (*blocks*), επιτυγχάνεται με την εφαρμογή της Προκρούστιας προβολής – περιστροφής. Με τη μέθοδο αυτή τα συγκροτήματα ομογενοποιούνται, είναι συγκρίσιμα και, επιπλέον, αναδεικνύονται διαγραμματικά οι μεταξύ τους διαφοροποιήσεις ως προς τις σχέσεις των προφίλ των πειραματικών αγωγών που ενθυλακώνουν οι πίνακες που αναλύονται. Να τονίσουμε ότι ο όρος «συγκρότημα» αναφέρεται όχι μόνο σε μια χωρική ομαδοποίηση των πειραματικών μονάδων, αλλά και σε οποιαδήποτε τμηματοποίηση που αποσκοπεί στη μεγιστοποίηση της ομοιογένειας του υλικού πειραματισμού. Οι μέθοδοι που προτείναμε μπορούν να εφαρμοστούν και σε δειγματοληπτικές έρευνες, στις οποίες υπάρχει εννοιολογικός διαχωρισμός των μεταβλητών σε εξαρτημένες και ανεξάρτητες.

Πρέπει να επισημάνουμε ότι σε όλα τα πειραματικά σχέδια που εξετάσαμε δεν λαμβάνεται υπόψη, κατά την ανάλυση των αντίστοιχων δεδομένων, η συσχέτιση – συνάφεια μεταξύ των εξαρτημένων μεταβλητών. Αν οι σχεδιασμοί είναι ισορροπημένοι, τότε οι συσχετίσεις μεταξύ των ανεξάρτητων μεταβλητών είναι ίσες με μηδέν. Όμως, δεν ισχύει το ίδιο σε μη ισορροπημένους σχεδιασμούς. Αν θεωρήσουμε ένα πείραμα, απλά, ως μία διαδικασία παραγωγής δεδομένων, όπου η εμπλεκόμενη αβεβαιότητα αφορά στα αποτελέσματα και όχι στο μηχανισμό εκτέλεσής του (βλέπε Ενότητα 1.6), τότε η ΠΑΑ μπορεί να εφαρμοστεί στο σύνολο των μεταβλητών, οι οποίες συμμετέχουν στον πειραματισμό, χωρίς διάκριση σε εξαρτημένες και ανεξάρτητες. Κάτω από αυτήν τη θεώρηση, το πειραματικό σχέδιο δεν παίζει σημαντικό ρόλο στην ανάλυση των αντίστοιχων δεδομένων. Η λογική της επαγωγικής συμπερασματολογίας, και ιδιαίτερα η απόδοση σχέσεων αιτίας – αποτελέσματος, είναι συνάρτηση του τρόπου με τον οποίο έχουν παραχθεί τα διαθέσιμα δεδομένα και όχι της στατιστικής μεθόδου με την οποία θα αναλυθούν (Cohen & Cohen 1983, De Leeuw 2005γ). Οι ερευνητές – χρήστες της μεθόδου έχουν την τελική ευθύνη για την ερμηνεία και τη γενίκευση των αποτελεσμάτων. Επομένως, στην περίπτωση που τα διαθέσιμα δεδομένα προέρχονται από πειραματική μελέτη και το ενδιαφέρον εστιάζεται στη σύγκριση των διαφόρων αγωγών, λαμβάνοντας ταυτόχρονα υπόψη τις συσχετίσεις ανά δύο όλων των εμπλεκόμενων μεταβλητών, ανεξάρτητων και εξαρτημένων, μπορούμε να εφαρμόσουμε την ΠΑΑ κατά το συνήθη τρόπο και να ερμηνεύσουμε τα αποτελέσματα σύμφωνα με τους περιορισμούς που θέτει το συγκεκριμένο πειραματικό σχέδιο, μέσω του οποίου παρήχθησαν τα προς ανάλυση δεδομένα.

Σε αντίθεση με την προηγούμενη θεώρηση, μπορούμε να αναλύσουμε τα πειραματικά δεδομένα στηριζόμενοι στις ιδιότητες βέλτιστης κλιμάκωσης της ΠΑΑ (βλέπε Ενότητα 2.5). Αρχικά διαχωρίζουμε τις εμπλεκόμενες μεταβλητές σε εξαρτημένες και ανεξάρτητες. Στη συνέχεια, εφαρμόζουμε την ΠΑΑ μόνο στις εξαρτημένες με σκοπό τη βέλτιστη ποσοτικοποίησή τους και την απόδοση φυσικής ερμηνείας στους παραγοντικούς άξονες, που θα αναδειχθούν. Οι άξονες αποτελούν νέες, σύνθετες, ποσοτικές πλέον μεταβλητές, οι οποίες μπορούν να πάρουν τη θέση των εξαρτημένων. Η στατιστική ανάλυση ολοκληρώνεται με την εφαρμογή κλασικών μεθόδων της Επαγωγικής Στατιστικής, για την ανάλυση πειραματικών δεδομένων (π.χ. Ανάλυση Διακύμανσης), λαμβάνοντας υπόψη, στο στάδιο αυτό, το

συγκεκριμένο πειραματικό σχέδιο, από το οποίο παρήχθησαν τα διαθέσιμα δεδομένα. Σύμφωνα με τον De Leeuw (1993) τα τυπικά σφάλματα των ποσοτικοποιημένων κατηγοριών των εξαρτημένων μεταβλητών ή των παραγοντικών βαθμών των αντικειμένων του πίνακα που αναλύεται είναι ασυμπτωτικά ακριβή και μπορούν να υπολογιστούν από τις αντίστοιχες εμπειρικές κατανομές χωρίς να είναι απαραίτητη η εφαρμογή της μεθόδου *Δέλτα*. Οι παραγοντικοί άξονες της ΠΑΑ θεωρούνται ποσοτικές μεταβλητές, εν δυνάμει συνεχείς, οι οποίες ακολουθούν ασυμπτωτικά την Κανονική Κατανομή. Μπορούν να συμμετάσχουν στις μετέπειτα αναλύσεις σαν να ήταν οι τιμές τους γνωστές από την αρχή και όχι παράγωγες, δηλαδή βασισμένες στα ίδια τα δεδομένα (βλέπε Pearson 1906, De Leeuw 1993). Στο πλαίσιο αυτό, η ΠΑΑ θεωρείται ως μέθοδος μετασχηματισμού των αρχικών δεδομένων, ώστε να αποκτήσουν επιθυμητές βέλτιστες ιδιότητες. Άλλωστε, διαδικασίες μετασχηματισμού των αρχικών μετρήσεων για να επιτευχθεί ομοιογένεια της διασποράς και καλύτερη προσαρμογή στην Κανονική Κατανομή εφαρμόζονται συχνά στην Ανάλυση Διακύμανσης (Kirk 1995, Μαυρομάτης 1999). Αν η παραδοχή της Κανονικής Κατανομής δεν επαληθεύεται εμπειρικά ή δημιουργεί θεωρητικούς προβληματισμούς, σχετικά με τη μεροληψία των συμπερασμάτων, τα πειραματικά δεδομένα μπορούν να αναλυθούν με μεθόδους της Μη Παραμετρικής Στατιστικής (Thomas & Kiwanga, 1993).

Οι προηγούμενες θεωρήσεις, αν και εισάγουν προβληματισμό, ως προς τη μέθοδο που θα ακολουθηθεί για την ανάλυση πειραματικών δεδομένων μέσω της ΠΑΑ, ωστόσο αναδεικνύουν την ευελιξία καθώς και το εύρος των δυνατοτήτων και ιδιοτήτων της μεθόδου, όπως αυτή έχει διαμορφωθεί και θεμελιωθεί κάτω από την επίδραση φιλοσοφικών κυρίως προσεγγίσεων. Η ΠΑΑ αποτελεί ένα γενικό σύστημα ανάλυσης κατηγορικών δεδομένων και η απόφαση σχετικά με το ποια, τελικά, μέθοδο θα επιλέξει ο ερευνητής είναι μάλλον θέμα επιστημολογικής αντιμετώπισης (ίσως και προτίμησης) του θεωρητικού ή/και μεθοδολογικού της πλαισίου.

ΚΕΦΑΛΑΙΟ 8

Γενικά Συμπεράσματα και Προτάσεις για Περαιτέρω Έρευνα

Στην παρούσα διατριβή δείξαμε ότι η Παραγοντική Ανάλυση των Αντιστοιχιών (ΠΑΑ), όπως αυτή αναδεικνύεται στο μεθοδολογικό πλαίσιο της Γαλλικής και Ολλανδικής Σχολής Ανάλυσης Δεδομένων, μπορεί να εφαρμοστεί στη στατιστική επεξεργασία κατηγορικών – ποιοτικών μεταβλητών, οι οποίες διακρίνονται σε εξαρτημένες και ανεξάρτητες. Η διάκριση μπορεί να είναι: α) δομική, δηλαδή να καθορίζεται από το μηχανισμό παραγωγής των δεδομένων, όπως συμβαίνει στους πειραματικούς σχεδιασμούς, ή β) εννοιολογική, δηλαδή να υπαγορεύεται θεωρητικά μέσα σε συγκεκριμένο γνωστικό ερευνητικό πεδίο. Ο εννοιολογικός διαχωρισμός παρατηρείται συχνά στις δειγματοληπτικές *ex post facto* έρευνες. Στο Κεφάλαιο 7 προτείναμε μεθόδους, στο πρότυπο της Ανάλυσης Διακύμανσης, για την ανάλυση τριών βασικών παραγοντικών πειραματικών σχεδιασμών: α) το Πλήρως Τυχαιοποιημένο Σχέδιο με Ένα Παράγοντα, β) το Πλήρως Τυχαιοποιημένο Σχέδιο με Δύο Παράγοντες και γ) το Τυχαιοποιημένο Σχέδιο σε Πλήρη Συγκροτήματα (*blocks*) με Δύο Παράγοντες. Οι προτεινόμενες μεθοδολογικές προσεγγίσεις επιτρέπουν τη συμμετοχή στον πειραματισμό δύο ή περισσότερων εξαρτημένων μεταβλητών, ενώ μπορούν να γενικευτούν και στην περίπτωση περισσοτέρων των δύο ανεξάρτητων. Οι μέθοδοι μπορούν να εφαρμοστούν και σε αντίστοιχα δειγματοληπτικά σχήματα. Με δεδομένο ότι η ΠΑΑ, λόγω της συμμετρικής αντιμετώπισης των μεταβλητών, δεν έχει αξιοποιηθεί στην ανάλυση πειραματικών σχεδιασμών, όπως αυτοί ορίστηκαν στην Ενότητα 1.2, και ιδιαίτερα σε αυτούς όπου οι εξαρτημένες μεταβλητές είναι κατηγορικής φύσης, ενδιαφέρον έχει η επέκταση εφαρμογής της μεθόδου και σε άλλες πειραματικές διατάξεις. Ενδεικτικά αναφέρουμε τα ιεραρχικά διακλαδιζόμενα πειράματα (*nested experimental designs*), τους σχεδιασμούς σε λατινικά και ελληνολατινικά τετράγωνα και τα μη πλήρη ή ατελή πειραματικά σχέδια. Αξίζει, επίσης, να διερευνηθεί η εφαρμογή της ΠΑΑ σε πειραματικούς σχεδιασμούς στους οποίους εμπλέκονται, εκτός από κατηγορικές, και ποσοτικές ανεξάρτητες μεταβλητές.

Ανοιχτό, τέλος, είναι και το ζήτημα του ελέγχου της στατιστικής σημαντικότητας των αποκλίσεων μεταξύ των σημείων του χώρου αναφοράς και των ομόλογων σημείων των προβαλλόμενων επί αυτού χώρων μέσω της Προκρούστιας μεθόδου.

Το φιλοσοφικό πλαίσιο, στο οποίο αναπτύχθηκε αρχικά η ΠΑΑ, δεν επέτρεπε, κυρίως από επιστημολογική σκοπιά, τον έλεγχο της εξωτερικής εγκυρότητας των αριθμητικών και διαγραμματικών εκροών της μεθόδου μέσω στατιστικών ελέγχων υποθέσεων, όπως αυτοί εφαρμόζονται στην Επαγωγική Στατιστική. Στα Κεφάλαια 5 και 6 δείξαμε ότι η ΠΑΑ μπορεί να συνδυαστεί με ελέγχους σημαντικότητας και μάλιστα με ελάχιστες τεχνικές και θεωρητικές προϋποθέσεις. Ειδικότερα, στο Κεφάλαιο 6, για τη διμεταβλητή εκδοχή της ΠΑΑ, προτείναμε δύο μεθόδους κατασκευής 100(1-α)% ελλείψεων εμπιστοσύνης γύρω από τα σημεία γραμμών (στηλών) με σκοπό τη στατιστική σύγκριση των αντίστοιχων προφίλ και τον έλεγχο της σταθερότητας των διαγραμματικών αποτελεσμάτων. Κατασκευάσαμε, επίσης, ένα μη παραμετρικό διάστημα εμπιστοσύνης για τον έλεγχο της σημαντικότητας των παραγοντικών αξόνων, που προκύπτουν στην πολυμεταβλητή περίπτωση. Ανοιχτά θέματα προς διερεύνηση αποτελούν οι έλεγχοι στατιστικής σημαντικότητας των αριθμητικών αποτελεσμάτων της ΠΑΑ, στην περίπτωση ανάλυσης ειδικών πινάκων της μορφής «*αντικείμενα* × *ιδιότητες*», όπως είναι για παράδειγμα οι πίνακες τύπου "φέτας".

Στην εργασία τέθηκε για πρώτη φορά στο χώρο της Ανάλυσης Δεδομένων το πρόβλημα του καθορισμού του ελάχιστου απαιτούμενου μεγέθους δείγματος σε πειραματικές ή δειγματοληπτικές έρευνες, στα δεδομένα των οποίων θα εφαρμοστεί η ΠΑΑ. Το ζήτημα αντιμετωπίστηκε στο Κεφάλαιο 5, όπου προτείναμε μεθοδολογία για την *a priori* και *post hoc* Ανάλυση Ισχύος του ελέγχου ανεξαρτησίας – ομοιογένειας χ^2 στο πλαίσιο της ΠΑΑ. Με βάση την προτεινόμενη προσέγγιση αναπτύξαμε το λογισμικό Power Analysis for AFC για την εκτέλεση των σχετικών υπολογισμών. Στο ίδιο κεφάλαιο, παρουσιάσαμε μέθοδο καθορισμού του βέλτιστου υποχώρου, στον οποίο το υπό εξέταση φαινόμενο προβάλλεται χωρίς στατιστικά σημαντική απώλεια πληροφορίας, κατά την εφαρμογή της διμεταβλητής εκδοχής της ΠΑΑ. Ορίσαμε, επίσης, την έννοια της Δυναμικής Αδράνειας (*a priori* και *post hoc*) σε απλό πίνακα συμπτώσεων δύο μεταβλητών και παρουσιάσαμε το ρόλο του νέου

αυτού δείκτη στον καθορισμό του μεγέθους δείγματος και στον εμπειρικό προσδιορισμό των σημαντικών αξόνων. Ιδιαίτερο ενδιαφέρον παρουσιάζει ο καθορισμός νέων δεικτών μέτρησης του μεγέθους του αποτελέσματος (*effect size*), οι οποίοι θα μπορούσαν να χρησιμοποιηθούν στον υπολογισμό του ελάχιστου απαιτούμενου μεγέθους δείγματος στην πολυμεταβλητή εκδοχή της ΠΑΑ.

Στο Κεφάλαιο 4 δείξαμε επτά νέες βασικές μαθηματικές σχέσεις που συνδέουν τις αδράνειες πινάκων συμπτώσεων, γενικευμένων (*Burt*) και λογικών δύο ή περισσότερων κατηγορικών μεταβλητών. Διατυπώσαμε τις αντίστοιχες προτάσεις και καταλήξαμε και σε τρία πορίσματα. Οι σχέσεις αδράνειας αναδεικνύουν κυρίως την ποιότητα της πληροφορίας που παράγεται από την ΠΑΑ και τη φυσική ερμηνεία της ολικής αδράνειας, ανάλογα με τη μορφή του πίνακα που αναλύεται κάθε φορά. Στο ίδιο κεφάλαιο, ορίσαμε την έννοια της Ενδιαφέρουσας Αδράνειας του πίνακα *Burt* και προτείναμε έλεγχο για τη στατιστική σημαντικότητά της. Με βάση το νέο αυτό δείκτη αναπτύξαμε διαδικασία εντοπισμού υποπίνακα του *Burt*, ο οποίος περιλαμβάνει όλες τις μεταβλητές που συμμετέχουν στην ανάλυση, και η εφαρμογή της ΠΑΑ σε αυτόν αποδίδει την "πλησιέστερη εικόνα" του υπό εξέταση φαινομένου σε αυτή που προκύπτει από την εφαρμογή της μεθόδου στον αρχικό πίνακα *Burt*. Το πρόβλημα αυτό τέθηκε, επίσης, για πρώτη φορά στο χώρο της Ανάλυσης Δεδομένων και αντιμετωπίστηκε στην παρούσα μελέτη. Θεωρούμε ότι η εγκυρότητα της πρότασης, αν και επαληθεύτηκε εμπειρικά, αποτελεί αντικείμενο περαιτέρω πειραματισμών και ελέγχων. Στηριζόμενοι στη φυσική ερμηνεία της ενδιαφέρουσας αδράνειας, προτείναμε μέθοδο διόρθωσης των αδρανειών των παραγοντικών αξόνων, ώστε τα αντίστοιχα ποσοστά ερμηνείας να αντικατοπτρίζουν την πραγματική ποιότητα της λύσης της πολυμεταβλητής ΠΑΑ.

Στο Κεφάλαιο 2 επιχειρήσαμε μια συγκριτική παρουσίαση της ΠΑΑ όπως αυτή θεμελιώνεται και εφαρμόζεται στο πλαίσιο της Γαλλικής και Ολλανδικής Σχολής Ανάλυσης Δεδομένων. Δείξαμε με ποιον τρόπο οι δύο αυτές προσεγγίσεις συνδέονται και επισημάναμε τα σημεία που χρήζουν ιδιαίτερης προσοχής κατά την ερμηνεία των αποτελεσμάτων. Προβάλαμε τις ιδιότητες και τις δυνατότητες της μεθόδου και παραθέσαμε αναφορές σχετικά με τις σημαντικότερες εξελίξεις στο ερευνητικό της πλαίσιο. Θίξαμε, επίσης, θέματα που αφορούν στην εσωτερική σταθερότητα των εκροών της ΠΑΑ και αναφερθήκαμε στη συμπληρωματικότητά της με άλλες

στατιστικές μεθόδους (π.χ. Ταξινόμηση, Καμπύλες *Andrews* και διαγράμματα *Biplot*). Τέλος, εισαγάγαμε νέα εννοιολογικά και μεθοδολογικά στοιχεία στο πεδίο εφαρμογής της. Πιο συγκεκριμένα, στη διμεταβλητή εκδοχή της ΠΑΑ, προτείναμε μέθοδο εντοπισμού των κελιών του πίνακα συμπτώσεων, τα οποία συνεισφέρουν σημαντικά στην αδράνεια των παραγοντικών αξόνων. Στην περίπτωση πολλών μεταβλητών, ορίσαμε το Σχετικό Δείκτη Διακριτότητας μιας μεταβλητής και παρουσιάσαμε τη χρησιμότητα του νέου αυτού δείκτη στην ερμηνεία των αποτελεσμάτων. Και οι δύο προτάσεις συμβάλουν στη διεισδυτική ερμηνεία των δομικών σχέσεων μεταξύ των μεταβλητών. Προς την ίδια κατεύθυνση συνεισφέρει και ο συνδυασμός των ιδιοτήτων της ΠΑΑ και της Ανάλυσης Ομοιογένειας, όπως αυτές αναδεικνύονται από τη συγκριτική παράθεση των μεθοδολογικών προσεγγίσεων των δύο Σχολών Ανάλυσης Δεδομένων. Ανοιχτά προβλήματα προς διερεύνηση αποτελούν οι έλεγχοι σημαντικότητας των βασικών αριθμητικών αποτελεσμάτων της Ανάλυσης Ομοιογένειας, όπως είναι οι δείκτες Διακριτότητας των μεταβλητών και οι συντελεστές Εσωτερικής Συνέπειας – Αξιοπιστίας (*Cronbach's a*) των παραγοντικών αξόνων.

Στο Κεφάλαιο 3 περιγράψαμε μια διαδικασία με την οποία μπορούμε να χρησιμοποιήσουμε τις δυνατότητες του SPSS για να εφαρμόσουμε την ΠΑΑ σε πίνακες *Burt*, στο πνεύμα της Γαλλικής Σχολής, και προτείναμε έναν αποτελεσματικό αλγόριθμο εφαρμογής της μεθόδου σε μεγάλους λογικούς πίνακες. Ο αλγόριθμος βρίσκει εφαρμογές και στη μέθοδο της Ταξινόμησης σε Αύξουσα Ιεραρχία.

Θεωρούμε ότι πάντα θα υπάρχει ενδιαφέρον για βελτίωση ή καθιέρωση νέων διαδικασιών, δεικτών ή/και ελέγχων (στατιστικών ή εμπειρικών), οι οποίοι θα συμβάλουν: α) στη βελτίωση της ερμηνείας των αποτελεσμάτων, β) στην αξιολόγηση της ποιότητας και της ποσότητας της παραγόμενης πληροφορίας, γ) στην αξιολόγηση της πρακτικής ή κλινικής σημαντικότητας των ευρημάτων, δ) στον έλεγχο της εσωτερικής και εξωτερικής σταθερότητας των αριθμητικών και γραφικών αποτελεσμάτων και ε) στην ενίσχυση της αξιοπιστίας και εγκυρότητας των συμπερασμάτων. Έτσι, αν και στη μελέτη διαπραγματευτήκαμε σχετικά ζητήματα και προτείναμε συγκεκριμένες μεθόδους, διαδικασίες και δείκτες, σε όλα σχεδόν τα κεφάλαια, ωστόσο πιστεύουμε ότι τα θέματα αυτά θα είναι συνεχώς ανοιχτά για περαιτέρω έρευνα.

Στο Κεφάλαιο 1 αναπτύξαμε τους κυριότερους άξονες του εννοιολογικού πλαισίου της μελέτης, ώστε να καταστεί πιο αποτελεσματική η επικοινωνία με τον αναγνώστη, σε ότι αφορά την οριοθέτηση του υπό εξέταση θέματος και της συνοχής των επιμέρους ζητημάτων – προβλημάτων, τα οποία αντιμετωπίστηκαν για την εκπλήρωση του σκοπού και των ειδικών στόχων της διατριβής.

Η ΠΑΑ αποτελεί ένα γενικό σύστημα ανάλυσης κατηγορικών δεδομένων. Είναι αρκετά ευέλικτη και προσαρμόζεται στις ιδιαιτερότητες του εκάστοτε ερευνητικού προβλήματος καθώς και στους περιορισμούς που θέτει το θεωρητικό πλαίσιο, εντός του οποίου θα ερμηνευτούν τα αποτελέσματα. Στην πράξη, η μόνη αυστηρή απαίτηση για την εφαρμογή της μεθόδου είναι η ύπαρξη ενός πίνακα διπλής εισόδου της μορφής «*αντικείμενα* × *ιδιότητες*» με μη αρνητικά στοιχεία. Βέβαια, ιδιαίτερη προσοχή απαιτείται στην κατάστρωση του πίνακα, που θα δοθεί ως είσοδος στην ανάλυση, ώστε να μπορεί να αναδειχθεί η φυσική ερμηνεία των δομών (συσχετίσεων, ομοιοτήτων, τάσεων και αντιθέσεων), που παρουσιάζουν ενδιαφέρον σε κάθε ερευνητική περίπτωση. Ο πολυδιάστατος και πολυμεταβλητός χαρακτήρας της μεθόδου επιτρέπει την ανάδειξη κυρίως μη γραμμικών σχέσεων, οι οποίες δεν θα μπορούσαν να εντοπιστούν με απλή εξέταση των μεταβλητών κατά ζεύγη. Σήμερα, η μέθοδος χρησιμοποιείται σχεδόν σε όλα τα ερευνητικά επιστημονικά πεδία, είτε όπως έχει είτε με παραλλαγές και τροποποιήσεις στο αλγοριθμικό της μέρος.

Στα πορίσματα της Γαλλικής και Ολλανδικής Σχολής στηρίζονται οι υπολογιστικοί αλγόριθμοι των σύγχρονων εμπορικών λογισμικών στατιστικής επεξεργασίας δεδομένων, τα οποία διαθέτουν την ΠΑΑ. Στη Γαλλική Σχολή δίνεται μεγαλύτερη έμφαση στη γεωμετρική ερμηνεία των δεδομένων μέσω της απόστασης χ^2 και το υπολογιστικό της μέρος είναι κατά βάση αλγεβρικό. Η πολυμεταβλητή εκδοχή της ΠΑΑ εφαρμόζεται αλγοριθμικά και εννοιολογικά ακριβώς με τον ίδιο τρόπο, όπως και στην περίπτωση των δύο μεταβλητών. Στην Ολλανδική, η μέθοδος, ιδιαίτερα στην πολυμεταβλητή εκδοχή της, αντιμετωπίζεται ως πρόβλημα βέλτιστης κλιμάκωσης ή, κατά μία άλλη θεώρηση, ως μέθοδος ποσοτικοποίησης κατηγορικών (ποιοτικών) δεδομένων. Στην περίπτωση αυτή, στόχος είναι η εξεύρεση, μέσω του επαναληπτικού αλγόριθμου ALS, βέλτιστων βαθμών για τις γραμμές και τις στήλες του πίνακα που αναλύεται, ώστε να ελαχιστοποιείται μια συγκεκριμένη συνάρτηση απώλειας κάτω από περιορισμούς. Οι βέλτιστες τιμές μπορούν να χρησιμοποιηθούν,

στη συνέχεια, σε περαιτέρω στατιστικές αναλύσεις. Η προσέγγιση των Ολλανδών οδήγησε στην ανάπτυξη της μεθόδου της Ανάλυσης Ομοιογένειας, η οποία αποτελεί τον πυρήνα του συστήματος *GIFI* για τη μη γραμμική ανάλυση κατηγορικών δεδομένων. Η Ανάλυση Ομοιογένειας παράγει ισοδύναμα αποτελέσματα με αυτά που προκύπτουν από την εφαρμογή της ΠΑΑ στο λογικό πίνακα q κατηγορικών μεταβλητών. Όμως, οι δύο προσεγγίσεις – μέθοδοι διαφέρουν ριζικά τόσο ως προς το υπολογιστικό όσο και ως προς το θεωρητικό πλαίσιο και την προβληματική που οδήγησε στην ανάπτυξή τους.

Με την παρούσα εργασία επιχειρήσαμε να "γεφυρώσουμε" το επιστημολογικό κενό που χωρίζει δύο Σχολές Ανάλυσης Δεδομένων, τη Γαλλική με την Ολλανδική, και δύο φιλοσοφικές προσεγγίσεις στη στατιστική συμπερασματολογία, την Επαγωγική Στατιστική με την Ανάλυση Δεδομένων. Σε μελλοντικές έρευνες θα συνεχίσουμε το έργο αυτό, σε μια φιλόδοξη προσπάθεια να προτείνουμε μια ενιαία θεώρηση της ΠΑΑ, η οποία θα την καταστήσει ακόμη ισχυρότερο εργαλείο για την ανάλυση και ερμηνεία κατηγορικών δεδομένων.

Βιβλιογραφία

Ξενόγλωσση

Acock, A. & Stavig, G. (1979). A Measure for Association for Nonparametric Statistics. *Social Forces*, **57**(4), 1381-1386.

Adachi, K. (2002). Optimal Quantification of a Longitudinal Indicator Matrix: Homogeneity and Smoothness Analysis. *Journal of Classification*, **19**, 215-248.

Adcock, C. J. (1997). Sample Size Determination: A Review. *The Statistician*, **46**(2), 261-283.

Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. New York: John Willey & Sons, Inc.

Agresti, A. (2002). *Categorical Data Analysis*. New Jersey: John Willey & Sons, Inc.

Aguinis, H. (1995). Statistical Power Problems with Moderated Multiple Regression in Management Research. *Journal of Management*, **21**(6), 1141-1158.

Aitchison, J. & Greenacre, M. (2002). Biplots of Compositional Data. *Applied Statistics*, **51**(4), 375-392.

Aitken, A. C. (1949). On the Wishart Distribution in Statistics. *Biometrika*, **36**(1/2), 59-62.

Aït-Sidi-Allal, M., Baccini, A. & A. Mondot (2004). A New Algorithm for Estimating the Parameters and their Asymptotic Covariance in Correlation and Association Models. *Computations Statistics & Data Analysis*, **45**, 389-421.

Aldenderfer, M. & Blashfield, R. (1984). *Cluster Analysis*. Beverly Hills: Sage Publications.

Aluja-Banet, T. & Nonell-Torrent, R. (1993). Multiple Correspondence Analysis on Panel Data. In C. M. Cuadras and C. R. Rao (Eds), *Multivariate Analysis: Future Directions 2*, (pp. 233-244). Amsterdam: North-Holland.

Andersen, E. (1991). *The Statistical Analysis of Categorical Data*. Berlin-Heidelberg: Springer-Verlag.

Anderson, M. & Whitcomp, P. (2000). *DOE Simplified: Practical Tools for Effective Experimentation*. Portland, Oregon: Productivity, Inc.

Andrade, J. M., Gómez – Carracedo, M., Krzanowski, W. & Kubista, M. (2004). Procrustes Rotation in Analytical Chemistry, a Tutorial. *Chemometrics and Intelligent Laboratory Systems*, **72**, 123-132.

Andrews, D. F. (1972). Plots of High Dimensional Data. *Biometrics*, **28**, 125-136.

Arnold, G. M. & Collins, A. J. (1993). Interpretation of Transformed Axes in Multivariate Analysis. *Applied Statistics*, **42**(2), 381-400.

Arrindell, W. A. & Van der Ende, J. (1985). An Empirical Test of the Utility of the Observations-to-Variables Ratio in Factor and Components Analysis. *Applied Psychological Measurement*, **9**, 165-178.

Askell-Williams, H. & Lawson, M. (2004). A Correspondence Analysis of Child-Care Students' and Medical Students' Knowledge about Teaching and Learning. *International Education Journal*, **5**(2), 176-204.

Atkinson, A. C. & Donev, A. N. (1989). The Construction of Exact *D*-Optimum Experimental Designs With Application to Blocking Response Surface Designs. *Biometrika*, **76**(3), 515-526.

Atkinson, A. C. & Fedorov, V. V. (1975). Optimal Design: Experiments for Discrimination between Several Models. *Biometrika*, **62**(2), 289-303.

Atkinson, A. C. (1996). The Usefulness of Optimum Experimental Designs. *Journal of the Royal Statistical Society, Series B (Methodological)*, **58**(1), 59-76.

Atkinson, A. C. (1994). Fast Very Robust Methods for the Detection on Multiple Outliers. *Journal of the American Statistical Association*, **89**(428), 1329-1339.

Baccini, A., Caussinus, H. & De Falguerolles, A. (1993). Analysing Dependence in Large Contingency Tables: Dimensionality and Patterns in Scatter-Plots. In C. M. Cuadras and C. R. Rao (Eds), *Multivariate Analysis: Future Directions 2*, (pp. 245-263). Amsterdam: North-Holland.

Bacher, J. (1995). Goodness-of-fit Measures for Multiple Correspondence Analysis. *Quality & Quantity*, **29**, 1-16.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, **66**, 423-437.

Balbi, S. & Esposito, V. (2000). Rotated Canonical Analysis Onto a Reference Subspace. *Computational Statistics & Data Analysis*, **32**, 395-410.

Balbi, S. (1998). Graphical Displays in Nonsymmetric Correspondence Analysis. In J. Blasius and M. Greenacre (Eds), *Visualization of Categorical Data* (pp. 297-309). San Diego: Academic Press.

Barrett, P. T. & Kline, P. (1981). The Observation to Variable Ratio in Factor Analysis. *Personality Study in Group Behavior*, **1**, 23-33.

Barton, D. E. & David, F. N. (1956). Some Notes on Ordered Random Intervals. *Journal of the Royal Statistical Society. Series B (Methodological)*, **18**(1), 79-94.

Becker, C. & Gather, U. (2001). The Largest Nonidentifiable Outlier: A Comparison of Multivariate Simultaneous Outlier Identification Rules. *Computational Statistics & Data Analysis*, **36**, 119-127.

Bécue-Bertaut, M. & Pagès, J. (2004). A Principal Axes Method for Comparing Contingency Tables: MFACT. *Computational Statistics & Data Analysis*, **45**, 481-503.

Beh, E. (1997). Simple Correspondence Analysis of Ordinal Cross-Classifications Using Orthogonal Polynomials. *Biometrical Journal*, **39**, 589-613.

Beh, E. (1998). A Comparative Study of Scores for Correspondence Analysis With Ordered Categories. *Biometrical Journal*, **40**(4), 413-429.

Beh, E. (2001). Confidence Circles for Correspondence Analysis Using Orthogonal Polynomials. *Journal of Applied Mathematics and Decision Sciences*, **5**(1), 35-45.

Beh, E. (2004). Simple Correspondence Analysis: A Bibliographic Review. *International Statistical Review*, **72**(2), 257–284.

Beitler, P. & Landis, R. (1985). A Mixed-Effects Model for Categorical Data. *Biometrics*, **41**(4), 991-1000.

Bekker, P. & De Leeuw, J. (1988). Relations Between Variants of Non-Linear Principal Components Analysis. In J. Van Rijckevorsel and J. De Leeuw (Eds), *Component and Correspondence Analysis. Dimension Reduction by Functional Approximation*, (pp. 1-31). Chichester: John Willey & Sons Ltd.

Bénasséni, J. (1993). Perturbational Aspects in Correspondence Analysis. *Computational Statistics & Data Analysis*, **15**, 393-410.

Bendixen, M. (1995). Compositional Perceptual Mapping Using Chi-Squared Trees Analysis and Correspondence Analysis. *Journal of Marketing Management*, **11**, 571-581.

Bendixen, M. (2003). A Practical Guide to the Use of Correspondence Analysis in Marketing Research. *Marketing Bulletin, Techical Note 2*, 14. Διαθέσιμο στην ιστοσελίδα: http://marketing-bulletin.massey.ac.nz.

Benton, D. & Krishnamoorthy, L. (2003). Computing Discrete Mixtures of Continuous Distributions: Noncentral Chisquare, Noncentral *t* and the Distribution of the Square of the Sample Multiple Correlation Coefficient. *Computational Statistics & Data Analysis*, **43**, 249-267.

Benzécri J.-P. & Collaborateurs (1973). *L' Analyse des Données. Vol. 1: Taxinomie. Vol. 2: Analyse des Correspondances*. Paris: Dunod.

Benzécri, J.-P. (1982). Sur la Généralisation du Tableau de Burt et son Analyse per Bandes. *Les Cahiers de l' Analyse des Données*, **7**(1), 33-43.

Benzécri, J.-P. (1983). Analyse de l'Inertie Intraclasse par l'Analyse d'un Tableau de Contingence. *Les Cahiers de l' Analyse des Données*, **8**(3), 351-358.

Benzécri, J.-P. (1991). Measures, Models, and Graphical Displays in the Analysis of Cross-Classified Data: A Comment. *Journal of the American Statistical Association*, **86**(416), 1112-115.

Benzécri, J.-P. (1992). *Correspondence Analysis Handbook*. New York: Marcel Dekker, Inc.

Benzécri, J-P. (1979). Sur le Calcul des taux d'inertie dans l'analyse d'un questionnaire. Addendum et erratum á [BIN.MULT.]. *Cahiers de l'Analyse des Données*, **4**, 377–378.

Berdie, D., Anderson, J. & Niebuhr, M. (1986). *Questionnaires: Design and Use*. London: The Scarecrow Press, Inc.

Berger, J. (2003). Could Fisher, Jeffreys and Neyman Have Agreed on Testing? *Statistical Science*, **18**(1), 1-32.

Bergerud, W. & Sit, V. (2001). *What is Power*?...*Why Should we Use it*? Paper Presented at the 2001 Power Analysis Workshop, Ministry of Forests Research Program, British Columbia.

Besse, P. (1988). Spline Functions and Optimal Metric in Linear Principal Components Analysis. In J. Van Rijckevorsel and J. De Leeuw (Eds), *Component and Correspondence Analysis. Dimension Reduction by Functional Approximation*, (pp. 81-101). Chichester: John Willey & Sons Ltd.

Bishop, Y., Fienberg, S. & Holland, P. (1991). *Discrete Multivariate Analysis: Theory and Practice*. Massachusetts: The MIT Press.

Blasius, J. & Greenacre, M. (1994). Computation of Correspondence Analysis. In M. Greenacre and J. Blasius (Eds), *Correspondence Analysis in the Social Sciences. Recent Developments and Applications*, (pp. 53-78). London: Academic Press.

Blasius, J. & Thiessen, V. (2000). Methodological Artifacts in Measures of Political Efficacy and Trust: A Multiple Correspondence Analysis. *Political Analysis*, **9**(1), 1-21.

Blasius, J. & Thiessen, V. (2001). The Use of Neutral Responses in Survey Questions: An Application of Multiple Correspondence Analysis. *Journal of Official Statistics*, **17**(3), 351-367.

Blasius, J. (1994). Correspondence Analysis in Social Science Research. In M. Greenacre and J. Blasius (Eds), *Correspondence Analysis in the Social Sciences*, (pp. 23-52). London: Academic Press.

Blattner, P. (2001). *Οι Συναρτήσεις του Microsoft Excel 2000 στην Πράξη*. Αθήνα: Εκδόσεις Κλειδάριθμος.

BMDP Statistical Software Inc. (1992). *BMDP Statistical Software Manual Release 7*, vols. 1 and 2. Los Angeles.

Böckenholt, U. & Takane, Y. (1994). Linear Constraints in Correspondence Analysis. In M. Greenacre and J. Blasius (Eds), *Correspondence Analysis in the Social Sciences. Recent Developments and Applications*, (pp. 112-127). London: Academic Press.

Boik, R. (1996). An Efficient Algorithm for Joint Correspondence Analysis. *Psychometrika*, **61**(2), 255-269.

Bolviken *et al*. (1982). Correspondence Analysis: An Alternative to Principal Components. *World Archaeology*, **14**(1), *Quantitative Methods*, 41-60.

Bond, J. & Michailidis, G. (1996). Homogeneity Analysis in Lisp-Stat. *Journal of Statistical Software*, 1(2). Διαθέσιμο στην ιστοσελίδα: http://www.jstatsoft.org/v01/i02/PAPER/paper.html.

Bond, J. & Michailidis, G. (1997). Interactive Correspondence Analysis in a Dynamic Object-Oriented Environment. *Journal of Statistical Software*, **2**(8). Διαθέσιμο στην ιστοσελίδα: http://www.jstatsoft.org/v02/i08/jss.pdf

Borkowf, C. (2000). On Multidimentional Contingency Tables with Categories Defined by the Empirical Quantiles of the Marginal Data. *Journal of Statistical Planning and Inference*, **91**, 33-51.

Bouilland, S. & Loslever, P. (1998). Multiple Correspondence Analysis of Biomechanical Signals Characterized Through Fuzzy Histograms. *Journal of Biomechanics*, **31**, 663-666.

Box, G. E. P. & Draper, N. (1959). A Basis for the Selection of a Response Surface Design. *Journal of the American Statistical Association*, **54**(287), 622-654.

Box, G. E. P. & Draper, N. (1963). The Choise of a Second Order Rotatable Design. *Biometrika*, **50**(3/4), 335-352.

Box, G. E. P. & Draper, N. (1975). Robust Designs. *Biometrika*, **62**(2), 347-352.

Brandstätter, E. (1999). Confidence Intervals as an Alternative to Significance Testing. *Methods of Psychological Research Online*, **4**(2). Διαθέσιμο στην ιστοσελίδα: http://www.ipn.uni-kiel.de/mpr/.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, **16**(3), 199-231.

Brokken, F. (1983). Orthogonal Procrustes Rotation Maximizing Congruence. *Psychometrika*, **48**(3), 343-352.

Brown, A. (2001). A Step-by-Step Guide to Non-Linear Regression Analysis of Experimental Data Using a Microsoft Excel Spreadsheet. *Computer Methods and Programs in Biomedicine*, **65**, 191-200.

Brown, S. & Melamed, L. (1990). *Experimental Design and Analysis*. Newbury Park: Sage Publications.

Browne, M. W. & Cudeck, R. (1993). Alternative Ways of Assessing Model Fit. In K. A. Bollen & J. S. Long (Eds), *Testing Structural Equation Models*, (pp. 136-162). Newbury Park: Sage Publications.

Bryant, F. (2000). Assessing the Validity of Measurement. In L. Grimm and P. Yarnold (Eds), *Reading and Understanding More Multivariate Statistics*, (pp. 99-146). Washington: American Psychological Association.

Bryman, A. & Cramer, D. (1999). *Quantitative Data Analysis with SPSS Release 8 for Windows: A Guide to Social Scientists*. London: Routledge.

Buhl-Mortensen, L. (1996). Type-II Statistical Errors in Environmental Science and the Precautionary Principle. *Marine Pollution Bulletin*, **32**(7), 528-531.

Buja, A. (1990). Remarks on Functional Canonical Variates, Alternating Least Squares and Ace. *The Annals of Statistics*, **18**(3), 1032-1069.

Burrows, G. L. (1963). Statistical Tolerance Limits-What Are They? *Applied Statistics*, **12**(2), 133-144.

Burt, C. (1950). The Factorial Analysis of Qualitative Data. *British Journal of Psychology, Statistical Section*, **3**, 166-185.

Burtschy, B. & Papadimitriou, I. (1991). La Matrice de Leontief de la Grèce: Analyse Diachronique de 1958 à 1977. *Les Cahiers de l' Analyse des Données*, Vol. **XVI**, No 4, 403-418. Paris: Dunod.

Busold, C. *et al*. (2005). Integration of GO Annotations in Correspondence Analysis: Facilitating the Interpretation of Microarray Data. *Bioinformatics*, **21**(10), 2424-2429.

Buttrey, S. (2005). Calling the lp_solve Linear Program Software from R, S-Plus and Excel. *Journal of Statistical Software*, **14**(4). Διαθέσιμο στην ιστοσελίδα: http://www.jstatsoft.org/.

Camiz, S. (2005). The Guttman Effect: Its Intepretation and a New Redressing Method. *Τετράδια Ανάλυσης Δεδομένων*, **5**/05, 7-34.

Cao, Y., Williams, D. & Williams, N. (1999). Data Transformations and Standardization in the Multivariate Analysis of River Water Quality. *Ecological Applications*, **9**(2), 669-677.

Carlier, A. & Kroonenberg, P. (1996). Biplots and Decompositions in Two-Way and Three-Way Correspondence Analysis. *Psychometrika*, **61**(2), 355-373.

Carlier, A. & Kroonenberg, P. (1998). Three-Way Correspondence Analysis, the Case of the French Cantons. In J. Blasius and M. Greenacre (Eds), *Visualization of Categorical Data*, (pp. 253-276). San Diego: Academic Press.

Carmines, E. & Zeller, R. (1979). *Reliability and Validity Assessment*. Newbury Park: Sage Publications.

Carroll, D. & Green P. (1988). An INDSCAL-Based Approach to Multiple Correspondence Analysis. *Journal of Marketing Research*, **XXV**, 193-203.

Carroll, D., Green, P. & Schaffer, C. (1986). Interpoint Distance Comparisons in Correspondence Analysis. *Journal of Marketing Research*, **XXIII**, 271-280.

Carroll, D., Green, P. & Schaffer, C. (1987). Comparing Interpoint Distances in Correspondence Analysis: A Clarification. *Journal of Marketing Research*, **XXIV**, 445-450.

Carroll, D., Green, P. & Schaffer, C. (1989). Reply to Greenacre's Commentary on the Carrol-Green-Schaffer Scaling of Two-Way Correspondence Analysis Solutions. *Journal of Marketing Research*, **XXVI**, 366-368.

Carver, P. (1978). The Case Against Statistical Testing. *Harvard Educational Review*, **48**, 378-399.

Cascio, W. & Zedeck, S. (1983). Open a New Window in Rational Research Planning: Adjust Alpha to Maximize Statistical Power. *Personnel Psychology*, **36**, 517-526.

Castelloe, J. & O'Brien, R. (2001). *Power and Sample Size Determination for Linear Models*. Paper 240-26, SUGI Proceedings, SAS Institute Inc., Cary, NC.

Castelloe, J. (2000). *Sample Size Computations and Power Analysis with the SAS® System*. Paper 265-25, SUGI Proceedings, SAS Institute Inc., Cary, NC.

Cattell, R. B. (1978). *The Scientific Use of Factor Analysis in Behavioral and Life Sciences*. New York: Plenum.

Cazes, P. (1980). Analyse de Certains Tableaux Rectangulaires Décomposés en Blocks. *Les Cahiers de l' Analyse des Données*, **5**(4), 387-403.

Chaloner, K. & Verdinelli, I. (1995). Bayesian Experimental Design: A Review. *Statistical Science*, **10**(3), 273-304.

Chapin, F. S. (1950). Experimental Design in Sociology: Limitations and Abuses. *Social Forces*, **29**(1), 25-28.

Chapman, D. & Nam, J. (1968). Asymptotic Power of Chi-Square Tests for Linear Trends in Proportions. *Biometrics*, **24**(2), 315-327.

Chateau, F. & Lebart, L. (1996). Assessing Sample Variability in the Visualization Techniques Related to Principal Components Analysis: Bootstrap and Alternative Simulation Methods. *Proceedings: Computational Statistics COMPSTAT*, 205-210.

Chatfield, C. (1991). Avoiding Statistical Pitfalls. *Statistical Science*, **6**, 240-268.

Chernick, M. & Liu, C. (2002). The Saw-Toothed Behavior of Power Versus Sample Size and Software Solutions: Single Binomial Proportion Using Exact Methods. *The American Statistician*, **56**(2), 149-155.

Chernoff, H. (1999).Gustav Elfving's Impact on Experimental Design. *Statistical Science*, **14**(2), 201-205.

Cheung, Y.-L. (1991). Correspondence Analysis as an Aid to Multicriteria Decision Making. *OMEGA*, **19**(2/3), 149-155.

Cheung, Y.-L. (1994). Categorical Criteria Values: Correspondence Analysis. *OMEGA*, **22**(4), 371-380.

Chew, V. (1966). Confidence, Prediction, and Tolerance Regions for the Multivariate Normal Distribution. *Journal of the American Statistical Association*, **61**(315), 605-617.

Choulakian, V. (1988). Exploratory Analysis of Contingency Tables by Loglinear Formulation and Generalizations of Correspondence Analysis. *Psychometrika*, **53**(2), 235-250.

Churchill, G. (1995). *Marketing Research: Methodological Foundations*. Fort Worth: The Dryden Press Harcourt Brace College Publisher.

Clausen, S.-E. (1998). *Applied Correspondence Analysis: An Introduction*. Thousand Oakes: Sage Publications.

Clemm, D. S., Krishnaiah, P. R. & Waikar, V. B. (1973). Tables for the Extreme Roots of the Wishart Matrix. *J. Statist. Comput. Simul.*, 2, 65-92.

Coakes, S. & Steed, L. (1999). *SPSS: Analysis Without Anguish*. Singapore: John Willey & Sons.

Cochran, W. & Cox, G. (1953). *Experimental Designs*. New York: John Willey & Sons, Inc.

Cochran, W. (1952). The chi^2 Test of Goodness of Fit. *The Annals of Mathematical Statistics*, **23**(3), 315-345.

Cochran, W. (1977). *Sampling Techniques*. New York: John Wiley & Sons.

Cohen, J. & Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.* New Jersey: Lawrence Erlbaum Associates Inc.

Cohen, J. & Nee, J. (1987). A Comparison of Two Noncentral *F* Approximations, with Applications to Power Analysis in Set Correlation. *Multivariate Behavioral Research*, **22**, 483-490.

Cohen, J. (1962). The Statistical Power of Abnormal-Social Psychological Research: A Review. *Journal of Abnormal and Social Psychology*, **65**, 145-153.

Cohen, J. (1965). Some Statistical Issues in Psychological Research. In B. B. Wolman (Ed.), *Handbook of Clinical Psychology*, (pp. 95-121). New York: McGraw-Hill.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New Jersey: Lawrence Erlbaum Associates, Inc.

Cohen, L. & Manion L. (1997). *Μεθοδολογία Εκπαιδευτικής Έρευνας*. Αθήνα: Εκδόσεις μεταίχμιο – Εκδόσεις Έκφραση.

Comrey, A. L. & Lee, H. B. (1973). *A First Course in Factor Analysis*. New York: Academic Press.

Corston, R. (1992). *Research Methods and Statistics in the Social Sciences*. Birtley: Casdec Ltd.

Cox, D. R. (1957). Note on Grouping. *Journal of the American Statistical Association*, **52**(280), 543-547.

Cox, D. R. (1958). *Planning of Experiments*. New York: John Willey & Sons, Inc.

Cox, D. R. (1961). Design of Experiments: The Control of Error. *Journal of the Royal Statistical Society. Series A (General)*, **124**(1), 44-48.

Cox, D. R. (1977). The Role of Significance Tests. *Scandanavian Journal of Statistics*, **4**, 49-70.

Cox. G. (1950). Discussion on Experimental Design. *Biometrics*, **6**(3), 317-319.

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

Cuadras, C. (2002). Correspondence Analysis and Diagonal Expansions in Terms of Distribution Functions. *Journal of Statistical Planning and Inference*, **103**, 137-150.

D'Ambra, L. & Lauro, C. (1992). Non Symmetrical Exploratory Data Analysis. *Statistica Applicata*, **4**(4), 511-529.

Daniel, W. (1995). *Biostatistics: A Foundation for Analysis in the Health Sciences*. Singapore: John Willey & Sons, Inc.

Davis, P. & Hersh, R. (1980). *The Mathematical Experience*. London: Penguin Books.

De Falguerolles, A., Jmel, S. & Whittaker, J. (1988). Correspondence Analysis and Association Models Constrained by a Conditional Independence Graph. *Psychometrika*, **60**(2), 161-180.

De Lagarde, J. (1995). *Initiation à l'Analyse des Données*. Paris: Dunod.

De Leeuw, J. & Van der Heijden, P. (1988). Correspondence Analysis of Incomplete Contingency Tables. *Psychometrika*, **53**(2), 223-233.

De Leeuw, J. & Van Rijckvorsel, J. (1988). Beyond Homogeneity Analysis. In J. Van Rijckevorsel and J. De Leeuw (Eds), *Component and Correspondence Analysis. Dimension Reduction by Functional Approximation*, (pp. 55-80). Chichester: John Willey & Sons Ltd.

De Leeuw, J. (1984). The GIFI-System of Non-Linear Multivariate Analysis. In E. Diday, M. Jambu, L. Lebart and R. Tomassone (Eds), *Data Analysis and Informatics III*, (pp.415-424). Amsterdam: North Holland.

De Leeuw, J. (1988). Multivariate Analysis With Linearizable Regressions. *Psychometrika*, **53**(4), 437-454.

De Leeuw, J. (1993). Some Generalizations of Correspondence Analysis. In C. M. Cuadras and C. R. Rao (Eds), *Multivariate Analysis: Future Directions 2*, (pp. 359-375). Amsterdam: North-Holland.

De Leeuw, J. (2005α). *Multivariate Analysis With Optimal Scaling*. Department of Statistics, UCLA. Department of Statistics Papers. Paper 2005103002. Διαθέσιμο στην ιστοσελίδα: http://repositories.cdlib.org/uclastat/papers/2005103002

De Leeuw, J. (2005β). *Nonlinear Principal Component Analysis*. Department of Statistics, UCLA. Department of Statistics Papers. Paper 2005103001. Διαθέσιμο στην ιστοσελίδα: http://repositories.cdlib.org/uclastat/papers/2005103001

De Leeuw, J. (2005γ). *Models and Techniques*. Department of Statistics, UCLA. Department of Statistics Papers. Paper 2005102703. Διαθέσιμο στην ιστοσελίδα: http://repositories.cdlib.org/uclastat/papers/2005102703

De Leeuw, J., Wang, D. & Michailidis, G. (1999). Correspondence Analysis Techniques. In S. Ghosh (Ed.), *Multivariate Analysis, Design of Experiments, and Survey Sampling*, (pp. 523-545). Basel: Marcel Dekker, Inc.

De Leeuw, J., Young, F. & Takane, Y. (1976). Additive Structure in Qualitative Data: An Alternating Least Squares Methods with Optimal Scaling Features. *Psychometrika*, **41**(4), 471-503.

De Maesschalck, R., Jouan-Rimbaud, D. & Massart, D. L. (2000). The Mahalanobis Distance. *Chemometrics and Intelligent Laboratory Systems*, **50**, 1-18.

De Nooy, W. (2003). Fields and Networks: Correspondence Analysis and Social Network Analysis in the Framework of Field Theory. *Poetics*, **31**, 305-327.

Desmond , J. & Glover, G. (2002). Estimating Sample Size in Functional MRI (fMRI) Neuroimaging Studies: Statistical Power Analysis. *Journal of Neuroscience Methods*, **118**, 115-128.

Deville J.-C. & Malinvaud, E. (1983). Data Analysis and Official Socio-Economic Statistics. *Journal of the Royal Statistical Society, Series A (General)*, **146**(4), 335-361.

Deville, J.-C. & Saporta, G. (1983). Correspondence Analysis, with an Extension Towards Nominal Time Series. *Journal of Econometrics*, **22**, 169-189.

Di Stefano, J. (2001). Power Analysis and Sustainable Forest Management. *Forest Ecology and Management*, **154**, 141-153.

DiCiccio, T. & Efron, B. (1996). Bootstrap Confidence Intervals. *Statistical Science*, **11**(3), 189-228.

Didow, N., Perreault, W. & Williamson, N. (1983). A Cross-Sectional Optimal Scaling Analysis of the Index of Consumer Sentiment. *Journal of Consumer Research*, **10**, 339-347.

Dijksterhuis, G. B. & Gower, J. (1991/2). The Interpretation of Generalized Procrustes Analysis and Allied Methods. *Food Quality and Preferences*, **3**, 67-87.

Dijksterhuis, G., Martens, H. & Martens, M. (2005). Combined Procrustes Analysis and PLSR for Internal and External Mapping of Data from Multiple Sources. *Computational Statistics & Data Analysis*, **48**, 47-62.

Dillon, W. & Goldstein, M. (1984). *Multivariate Analysis: Methods and Applications*. New York: John Willey & Sons, Inc.

DiPrete, T. & Forristal, J. (1994). Multilevel Models: Methods and Substance. *Annual Review of Sociology*, **20**, 331-357.

Dohoo, I. R. et al. (1996). An Overview of Techniques for Dealing with Large Number of Independent Variables in Epidimiologic Studies. *Preventine Veterinary Medicine*, **29**, 221-239.

Dometrius, N. (1992). *Social Statistics Using SPSS*. New York: Harper Collins Publishers, Inc.

Dunham, M. (2004). *Data Mining: Εισαγωγικά και Προηγμένα Θέματα Εξόρυξης Γνώσης από Δεδομένα*. Αθήνα: Εκδόσεις Νέων Τεχνολογιών.

Dzhafarov, E. (1999). Double Skew-Dual Scaling: A Conjoint Scaling of Two Sets of Objects Related by a Dominance Matrix. *Journal of Mathematical Psychology*, **43**, 483-515.

Eastment, H. T. & Krzanowski, W. J. (1982). Cross-Validation Choise of the Number of Components From Principal Components Analysis. *Technometrics*, **24**(1), 73-77.

Eaton, M. & Tyler, D. (1994). The Asymptotic Distribution of Singular Values With Applications to Canonical Correlations and Correspondence Analysis. *Journal of Multivariate Analysis*, **50**, 238-264.

Edsall, R. (2003). The Parallel Coordinate Plot in Action: Design and Use for Geographic Visualization. *Computational Statistics & Data Analysis*, **43**, 605-619.

Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Elfving, G. (1952). Optimum Allocation in Linear Regression Theory. *The Annals of Mathematical Statistics*, **23**(2), 255-262.

Elswick, Jr.; R. K., Gennings, C., Chinchilli, V. & Dawson, K. (1991). A Simple Approach for Finding Estimable Functions in Linear Models. *The American Statistician*, **45**(1), 51-53.

Embrechts, P. & Herzberg, A. M. (1991). Variations of Andrews' Plots. *International Statistical Review*, **59**(2), 175-194.

Escofier, B. & Drouet, D. (1983). Analyse des Différences Entre Plusieurs Tableaux de Fréquence. *Les Cahiers de l' Analyse des Données*, **8**(4), 491-499.

Escofier, B. & Le Roux, B. (1976). Influence d'un Elément sur les Facteurs en Analyse Des Correspondances. *Les Cahiers de l' Analyse des Données*, **1**(3), 297-318.

Escofier, B. & Pagès, J. (1994). Multiple Factor Analysis (AFMULT package). *Computational Statistics & Data Analysis*, **18**, 121-140.

Escofier, B. & Pagès, J. (1998). *Analyses Factorielles Simples et Multiples: Objectifs, Mèthodes et Interprètation*. Paris: Dunod.

Escofier, B. (1979). Traitement Simultané de Variables Qualitatives et Quantitative en Analyse Factorielle. *Les Cahiers de l' Analyse des Données*, **4**, 137-146.

Escofier, B. (1983). Analyse de la Difference Entre Deux Measures sur le Produit de Deux Memes Ensembles. *Cahiers de l'Analyse des Données*, **3**, 325-329.

Evans, T. & Viengkham, O. (2001). Inventory Time-Cost and Statistical Power: A Case Study of a Lao Rattan. *Forest Ecology and Management*, **150**, 313-322.

Everitt, B. & Nicholls, P. (1975). Visual Techniques for Representing Multivariate Data. *The Statistician*, **24**(1), 37-49.

Everitt, B. (1975). Multivariate Analysis: The Need for Data, and Other Problems. *British Journal of Psychiatry*, **126**, 237-240.

Everitt, B. (1979). *The Analysis of Contingency Tables*. London: Chapman and Hall.

Everitt, B. (1993). *Cluster Analysis*. London: Edward Arnold A division of Hodder & Stoughton.

Faust, K. & Wasserman, S. (1993). Correlation and Association Models for Studying Measurements on Ordinal Relations. *Sociological Methodology*, **23**, 177-215.

Fedorov, V. & Khabarov, V. (1986). Duality of Optimal Designs for Model Discrimination and Parameter Estimation. *Biometrika*, **73**(1), 183-190.

Ferré, L. (1995). Selection of Components in Principal Components Analysis: A Comparison of Methods. *Computational Statistics & Data Analysis*, **19**, 669-682.

Fienberg, S. (1991). *The Analysis of Cross-Classified Categorical Data*. Massachusetts: The MIT Press.

Fienberg, S. (2000). Contingency Tables and Log-Linear Models: Basic Results and New Developments. *Journal of the American Statistical Association*, **95**(450), 643-647.

Fisher, R. A. (1940). The Precision of Discriminant Functions. *Annals of Eugenics*, **10**, 422-429.

Foster, J., (2001). Statistical Power in Forest Monitoring. *Forest Ecology and Management*, **151**, 211-222.

Fouladi, R. & Steiger, J. (1999). Tests of an Identity Correlation Structure. In R. Hoyle (Ed.), *Statistical Strategies for Small Sample Research*, (pp. 167-194). Thousand Oakes: Sage Publications, Inc.

Fox, J. (1991). *Regression Diagnostics*. Newbury Park: Sage Publications.

Fox, M. (1993). A Multi-Dimensional Exploration of the Decision Process Using Correspondence Analysis. *Marketing Bulletin*, **4**, 30-42.

Franke, G. (1985). Evaluating Measures Through Data Quantification: Applying Dual Scaling to an Advertising Copytest. *Journal of Business Research*, **13**, 61-69.

Friendly, M. (1994). Mosaic Displays for Multi-Way Contingency Tables. *Journal of the American Statistical Association*, **89**(425), 190-200.

Friendly, M. (1998). Extending Mosaic Displays: Marginal, Partial, and Conditional Views of Categorical Data. Paper presented at the Workshop on "*Data Visualization in Statistics*", July 6-10, 1998 at Drew University.

Fylstra, D., Lasdon, L., Watson, J. & Waren, A. (1998). Design and Use of the Microsoft Excel Solver. *Interfaces*, **28**(5), 29-55.

Gabriel, R. (1966). Simultaneous Test Procedures for Multiple Comparisons on Categorical Data. *Journal of American Statistical Association*, **61**(316), 1081-1096.

Gabriel, R. (1971). The Biplot-Graphic Display of Matrices with Application to Principal Component Analysis. *Biometrika*, **58**, 453-467.

Gabriel, R. (2002). Goodness of Fit of Biplots and Correspondence Analysis. *Biometrika*, **89**, 423-436.

Gabriel, R., Galindo, P. & Vicente-Villardón, J. L. (1998). Use of Biplots to Diagnose Independence Models in Three-Way Contingency Tables. In J. Blasius and M. Greenacre (Eds), *Visualization of Categorical Data*, (pp. 391-404). San Diego: Academic Press.

Gallo, M. & Simonetti, B. (2002). Alternative Interpretations to the Non Symmetrical Correspondence Analysis. *Caribb. J. Math. Comput. Sci.*, **12**, 18-22.

Gardner, S. Gower, J. & Le Roux, N. J. (2006). A Synthesis of Canonical Variate Analysis, Generalized Canonical Correlation and Procrustes Analysis. *Computational Statistics & Data Analysis*, **50**, 107-134.

Gatti, G. & Harwell, M. (1998). Advantages of Computer Programs Over Power Charts for the Estimation of Power. *Journal of Statistics Education*, **6**(3). Διαθέσιμο στην ιστοσελίδα: http://www.amstat.org/publications/jse/v6n3/gatti.html

Gauch, H. G. Jr. (1995). *Multivariate Analysis in Community Ecology*. Cambridge: Cambridge University Press.

Gettler-Summa, M. (1992). Factorial Axis Interpretation by Symbolic Objects. In Lice Ceremade (Ed.), *TRAITEMENT DES CONNAISSANCES "SYMBOLIQUES-NUMÉRIQUES"*, (pp. 53-64). Paris: Université Paris IX-Dauphine.

Giegler, H. & Klein, H. (1994). Correspondence Analysis of Textual Data from Personal Advertisements. In M. Greenacre & J. Blasius (Eds), *Correspondence Analysis in the Social Sciences. Recent Developments and Applications*, (pp. 283-301). London: Academic Press.

Gifi, A., (1996). *Non-Linear Multivariate Analysis*. Chichester: John Willey & Sons Ltd.

Gilula, Z. & Haberman, S. (1986). Canonical Analysis of Contingency Tables by Maximum Likelihood. *Journal of the American Statistical Association*, **81**(395), 780-788.

Gilula, Z. & Haberman, S. (1988). The Analysis of Multivariate Contingency Tables by Restricted Canonical and Restricted Association Models. *Journal of the American Statistical Association*, **83**(403), 760-771.

Gilula, Z. & Krieger, A. M. (1983). The Decomposability and Monotonicity of Pearson's Chi-Square for Collapsed Contingency Tables. *Journal of American Statistical Association*, **78**(381), 176-180.

Gilula, Z. & Krieger, A. M. (1989). Collapsed Two-Way Contingency Tables and the Chi-Square Reduction Principle. *Journal of the Royal Statistical Society. Series B* (*Methodological*), **51**(3), 425-433.

Gilula, Z. & Ritov, Y. (1990). Inferential Ordinal Correspondence Analysis: Motivation, Derivation and Limitations. *International Statistical Review*, **58**(2), 99-108.

Gilula, Z. (1984). On Some Similarities Betwwen Canonical Correlation Models and Latent Class Models for Two-Way Contingency Tables. *Biometrika*, **71**(3), 523-529.

Gilula, Z. (1985). On the Analysis of Heterogeneity Among Populations. *Journal of the Royal Statistical Society. Series B* (*Methodological*), **47**(1), 76-83.

Gilula, Z. (1986). Grouping and Asoociation in Contingency Tables: An exploratory Canonical Correlation Approach. *Journal of American Statistical Association*, **81**(395), 773-779.

Girden, E. (1992). *ANOVA: Repeated Measures*. Newbury Park: Sage Publications.

Goldstein, H. (1991). Multilevel Modelling of Survey Data. *The Statistician*, **40**(2), Special Issue: Survey Design, Methodology and Analysis, 235-244.

Goldstein, H. (1999). *Multilevel Statistical Models*. Διαθέσιμο στην ιστοσελίδα: www.ioe.ac.uk/multilevel/.

Goldstein, R. (1989). Power and Sample Size via MS/PC-DOS Computers. *The American Statistician*, **43**(4), 253-260.

Golub, G. & Van Loan, C. (1989). *Matrix Computations*, 2nd edition. Baltimore and London: The Johns Hopkins University Press.

Gomez, K. & Gomez, A. (1984). *Statistical Procedures for Agricultural Research*. Singapore: John Willey & Sons, Inc.

Goodal, C. (1991). Procrustes Methods in the Statistical Analysis of Shape. *Journal of the Royal Statistical Society. Series B (Methodological)*, **53**(2), 285-339.

Goodman, L. & Kruskal, W. (1954). Measures of Association for Cross Classifications. *Journal of the American Statistical Association*, **49**(268), 732-764.

Goodman, L. (1968). The Analysis of Cross-Classified Data: Independence, Quasi-Independence and Interactions in Contingency Tables With or Without Missing Entries. *Journal of American Statistical Association*, **63**(324), 1091-1131.

Goodman, L. (1981). Criteria for Determining Whether Certain Categories in a Cross-Classification Table Shoud Be Combined with Special Reference to Occupational Categories in Occupational Mobility Tables. *Amer. J. Sociol.*, **87**(3), 612-650.

Goodman, L. (1991). Measures, Models, and Graphical Displays in the Analysis of Cross-Classified Data: Rejoinder. *Journal of the American Statistical Association*, **86**(416), 1124-1138.

Goodman, L. (1993). Correspondence Analysis, Association Analysis, and Generalized Nonindependence Analysis of Contingency Tables: Saturated and Unsaturated Models, and Appropriate Graphical Displays. In C. M. Cuadras & C. R. Rao (Eds), *Multivariate Analysis: Future Directions 2*, (pp. 265-294). Amsterdam: North-Holland.

Goodman, L. (1996). A Single General Method for the Analysis of Cross-Classified Data: Reconciliation and Synthesis of Some Methods of Pearson, Yule, and Fisher, and Also Some Methods of Correspondence Analysis and Association Analysis. *Journal of the American Statistical Association*, **91**(433), 408-428.

Gorsuch, R. L. (1983). *Factor Analysis*. Hillsdale, NJ: Erlbaum.

Gower, J. & Greenacre M. (1996). Unfolding a Symmetric Matrix. *Journal of Classification*, **13**, 81-105.

Gower, J. & Harding, S. A. (1988). Nonlinear Biplots. *Biometrika*, **75**(3), 445-455.

Gower, J. (1975). Generalized Procrustes Analysis. *Psychometrika*, **40**(1), 33-51.

Gower, J. (1990). Fisher's Optimal Scores and Multiple Correspondence Analysis. *Biometrics*, **46**(4), 947-961.

Gower, J. (1990). Three-Dimensional Biplots. *Biometrika*, **77**(4), 773-885.

Gower, J. (1992). Generalized Biplots. *Biometrika*, **79**(3), 475-493.

Gower, J. (1993). Recent Advances in Biplot Methodology. In C. M. Cuadras and C. R. Rao (Eds), *Multivariate Analysis: Future Directions 2*, (pp. 295-325). Amsterdam: North-Holland.

Gower, J. (2003). Unified Biplot Geometry. *Development in Applied Statistics*, 3-22.

Gower, J. (2004). The Geometry of Biplot Scaling. *Biometrika*, **91**, 705-714.

Gower, J., Meulman, J. & Arnold, G. M. (1999). Nonmetric Linear Biplots. *Journal of Classification*, **16**, 181-196.

Graffelman, J. & Aluja-Banet, T. (2003). Optimal Representation od Supplementary Variables in Biplots from Principal Components Analysis and Correspondence Analysis. *Biometrical Journal*, **45**(4), 491-509.

Gras, R. (1995). Ανάλυση ενός Ερωτηματολογίου με τη Συνεπαγωγική Μέθοδο. Στο Α. Γαγάτσης (Ed.), *Διδακτική και Ιστορία των Μαθηματικών*, (σ.σ. 97-109). Θεσσαλονίκη: ERASMUS ICP-94-G-2011/11.

Grassi, M., Rezzani, C., Biino, G. & Marinoni, A. (2003). Asthma-Like Syndroms Assessment Trough ECRHS Screening Questionnaire Scoring. *Journal of Clinical Epidemiology*, **56**, 238-247.

Green, P. J. & Silverman, B. W. (1979). Constructing the Convex Hull of a Set of Points in the Plane. *The Computer Journal*, **22**(3), 262-266.

Green, P.J. (1981). Peeling Bivariate Data. In V. Barnett (Ed.), *Interpreting Multivariate Data*, (pp. 3-20). Chichester: John Willey & Sons.

Greenacre, M & Pardo, R. (2005). Multiple Correspondence Analysis of a Subset of Response Categories. *Economics Working Papers 881*, Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain.

Greenacre, M. & Blasius, J. (Eds) (1994). *Correspondence Analysis in the Social Sciences: Recent Developments and Applications*. London: Academic Press.

Greenacre, M. & Clavel, J. G. (1998). *Analysis of a Pair of Transition Matrices*. *Economics Working Paper 298*, Department of Economics and Business, Universitat Pombeu Fabra, Barcelona, Spain.

Greenacre, M. & Hastie, T. (1987). The Geometric Interpretation of Correspondence Analysis. *Journal of the American Statistical Association*, **82/83**(98), 437-447.

Greenacre, M. & Torres, A. (1999). *Dual Scaling of Dominance Data and its Relationship to Correspondence Analysis*. *Economics Working Paper 430*, Department of Economics and Business, Universitat Pombeu Fabra, Barcelona, Spain.

Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.

Greenacre, M. (1988α). Correspondence Analysis of Multivariate Categorical Data by Weighted Least-Squares. *Biometrika*, **75**(3), 457-467.

Greenacre, M. (1988β). Clustering the Rows and Columns of a Contingency Table. *Journal of Classification*, **5**, 39-51.

Greenacre, M. (1989). The Carrol-Green-Schaffer Scaling in Correspondence Analysis: A Theoritical and Empirical Appraisal. *Journal of Marketing Research*, **XXVI**, 358-365.

Greenacre, M. (1990). Some Limitations of Multiple Correspondence Analysis. *Computational Statistics Quarterly*, **3**, 249-256.

Greenacre, M. (1991). Interpreting Multiple Correspondence Analysis. *Applied Stochastic Models and Data Analysis*, **7**, 195-210.

Greenacre, M. (1992). Correspondence Analysis in Medical Research. *Statistical Methods in Medical Research*, **1**, 97-117.

Greenacre, M. (1993α). *Correspondence Analysis in Practice*. London: Academic Press.

Greenacre, M. (1993β). Biplots in Correspondence Analysis. *Journal of Applied Statistics*, **26**, 251-269.

Greenacre, M. (1993γ). Multivariate Generalisations of Correspondence Analysis. In C. M. Cuadras and C. R. Rao (Eds), *Multivariate Analysis: Future Directions 2*, (pp. 327-340). Amsterdam: North-Holland.

Greenacre, M. (1994α). Correspondence Analysis and its Interpretation. In M. Greenacre and J. Blasius (Eds), *Correspondence Analysis in the Social Sciences. Recent Developments and Applications*, (pp. 3-22). London: Academic Press.

Greenacre, M. (1994β). Multiple and Joint Correspondence Analysis. In M. Greenacre and J. Blasius (Eds), *Correspondence Analysis in the Social Sciences. Recent Developments and Applications*, (pp. 141-161). London: Academic Press.

Greenacre, M. (1998). Diagnostics for Joint Displays in Correspondence Analysis. In J. Blasius and M. Greenacre (Eds), *Visualization of Categorical Data*, (pp. 221-238). San Diego: Academic Press.

Greenacre, M. (2000). Correspondence Analysis of Square Asymetric Matrices. *Appl. Statist.*, **49**(3), 297-310.

Greenacre, M. (2003). Singular Value Decomposition of Matched Matrices. *Journal of Applied Statistics*, **30**(10), 1101-1113.

Greenacre, M. (2005). From Correspondence Analysis to Multiple and Joint Correspondence Analysis. *Economics Working Papers 883*, Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain.

Greenacre, M. (2006). Tying Up the Loose Ends in Simple Correspondence Analysis. *Economics Working Papers 940*, Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain.

Grimm, L. & Yarnold, P. (2000). Introduction to Multivariate Analysis. In L. Grimm and P. Yarnold (Eds), *Reading and Understanding More Multivariate Statistics*, (pp. 3-21). Washington, DC: American Psychological Association.

Groenen, P. & Franses, P. (2000). Visualizing Time-Varying Correlations Across Stock Market. *Journal of Empirical Finance*, **7**, 155-172.

Groenen, P. & Van de Velden, M. (2004). Inverse Correspondence Analysis. *Linear Algebra and its Applications*, **388**, 221-238.

Gruvaeus, G. (1970). A General Approach to Procrustes Pattern Rotation. *Psychometrika*, **35**(4), 493-505.

Guenther, W. (1964). Another Derivation of the Non-Central Chi-Square Distribution. *Journal of the American Statistical Association*, **59**(307), 957-960.

Guenther, W. (1977). Power and Sample Size for Approximate Chi-Square Tests. *The American Statistician*, **31**(2), 83-85.

Guilford, J. P. (1954). *Psychometric Methods*. New York: McGraw-Hill.

Guinot, C. *et al*. (2001). Use of Multiple Correspondence Analysis and Cluster Analysis to Study Dietary Behaviour: Food Consumption Questionnaire in the SU.VI.MAX. Cohort. *European Journal of Epidemiology*, **17**, 505-516.

Guo, G. & Zhao, H. (2000). Multilevel Modelling for Binary Data. *Annual Review of Sociology*, **26**, 441-462.

Guttman, L. (1941). The Quantification of a Class of Attributes: A Theory and Method of Scale Construction. In P. Horst *et al*. (Eds), *The Prediction of Personal Adjustment*, (pp. 321-348). New York: Social Science Research Council.

Guttman, L. (1944). A Basis for Scaling Qualitative Data. *American Sociological Review*, **9**(2), 139-150.

Guttman, L. (1985). The Illogic of Statistical Inference for Cumulative Science. *Applied Stochastic Models and Data Analysis*, **1**(1), 3-9.

Habbard, R. & Allen, S. (1987). An Empirical Comparison of Alternative Methods for Principal Component Extraction. *J. Bus. Res*., **15**, 173-190.

Haberman, S. (1973). The Analysis of Residuals in Cross-Classified Tables. *Biometrics*, **29**(1), 205-220.

Haberman, S. (1981). Tests for Independence in Two-Way Contingency Tables Based on Canonical Correlation and on Linear-By-Linear Interaction. The *Annals of Statistics*, **9**(6), 1178-1186.

Haberman, S. (1988). A Warning on the Use of Chi-Squared Statistics With Frequency Tables With Small Expected Cell Counts. *Journal of the American Statistical Association*, **83**(402), 555-560.

Haberman, S. (1995). Computation of Maximum Likelihood Estimates in Association Models. *Journal of the American Statistical Association*, **90**(432), 1438-1446.

Hair, J. & Black, W. (2000). Cluster Analysis. In L. Grimm and P. Yarnold (Eds), *Reading and Understanding More Multivariate Statistics*, (pp. 147-205). Washington, DC: American Psychological Association.

Hair, J., Anderson, R., Tatham, R. & Black, W. (1995). *Multivariate Data Analysis With Readings*. New Jersey: Prentice-Hall International, Inc.

Hallahan, M. & Rosenthal, R. (1996). Statistical Power: Concepts, Procedures, and Applications. *Behav. Res. Ther*., **34**(5/6), 489-499.

Haller, H. & Krauss, S. (2002). Misinterpretations of Significance: A Problem Students Share with their Teachers. *Methods of Psychological Research Online*, **7**(1). Διαθέσιμο στην ιστοσελίδα: http://www.mpr-online.de.

Hamaker, H. (1955). Experimental Design in Industry. *Biometrics*, **11**(3), 257-286.

Hamdam, M. A. (1968). Optimum Choise of Classes for Contingency Tables. *Journal of the American Statistical Association*, **63**(321), 291-297.

Han, C-P. (1975). Some Relationships Betwwen Noncentral Chi-Squared and Normal Distributions. *Biometrika*, **62**(1), 213-214.

Hand, D. (1996). Statistics and the Theory of Measurement. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **159**(3), 445-492.

Hans, D., Ojasoo, T. & Doré, J.-C. (2000). Deaths from Breast Cancer: Tackling Multidimensionality and Non-Linearity by Correspondence Analysis. *Journal of Steroid Biochemistry & Molecular Biology*, **74**, 195-202.

Hansen, M., Hurwitz, W. & Madow, W. (1953). *Sample Survey Methods and Theory. Volume I, Methods and Applications*. New York: John Wiley & Sons, Inc.

Hanumara, C & Thompson, W. A. (1968). Percentage Points of the Extreme Roots of a Wishart Matrix. *Biometrika*, **55**(3), 505-512.

Harman, H. H. (1967). *Modern Factor Analysis*. Chicago: University of Chicago Press.

Harris, R. (2001). *A Primer of Multivariate Statistics*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Harville, D. (1997). *Matrix Algebra from a Statistician's Perspective*. New York: Springer – Verlag, Inc.

Haynam, G. E., Govindarajulu, Z. & Leone, F. C. (1970). Tables of the Cumulative Non-central Chi-Square Distribution. In H. L. Harter and D. B. Owen (Eds), *Selected Tables in Mathematical Statistics, Vol. I.* Chicago: Markham Publishing Co.

Héberger, K. & Andrade, J. (2004). Procrustes Rotation and Pair-Wise Correaltion: A Parametric and a Non Parametric Method for Variable Selection. *Croatica Chemica Acta*, **77**(1-2), 117-125.

Hedrick, T., Bickman, L. & Rog, D. (1993). *Applied Research Design: A Practical Guide*. Newbury Park: Sage Publications.

Heidelbaugh, S. & Nelson, W. (1996). A Power Analysis of Methods for Assessment of Change in Seagrass Cover. *Aquatic Botany*, **53**, 227-233.

Heiser, W. & Meulman, J. (1983). Constrained Multidimensional Scaling, Including Confirmation. *Applied Psychological Measurement*, **7**(4), 381-404.

Heiser, W. & Meulman, J. (1994). Homogeneity Analysis: Exploring the Distribution of Variables and their Non-Linear Relationships. In M. Greenacre and J. Blasius (Eds), *Correspondence Analysis in the Social Sciences: Recent Developments and Applications*, (pp. 179-209). London: Academic Press.

Heiser, W. (1987). Correspondence Analysis with Least Absolute Residuals. *Computational Statistics & Data Analysis*, **5**, 337-356.

Henry, G. (1990). *Practical Sampling*. Newbury Park: Sage Publications.

Heo, M. & Gabriel, R. (2001). The Fit of Graphical Displays to Patterns of Expectations. *Computational Statistics & Data Analysis*, **36**, 47-67.

Higgs, N. T. (1991). Practical and Innovative Uses of Correspondence Analysis. *The Statistician*, **40**(2), 183-194.

Hill, M. O. & Gauch, H. G. (1980). Detrended Correspondence Analysis, an Improved Ordination Technique. *Vegetatio*, **42**, 47-58.

Hill, M. O. (1973). Reciprocal Averaging: An Eigenvector Method for Ordination. *The Journal of Ecology*, **61**(1), 237-249.

Hill, M. O. (1974). Correspondence Analysis: A Neglected Multivariate Method. *Applied Statistics*, **23**(3), 340-354.

Hillier, F. & Lieberman, G. (1995). *Introduction to Operations Research*. Singapore: McGraw-Hill, Inc.

Hinkle, D., Wiersma, W. & Jurs, S. (1988). *Applied Statistics for the Behavioral Sciences*. Boston: Houghton Mifflin Company.

Hirotsu, C. (1983). Defining the Pattern of Association in Two-Way Contingency Tables. *Biometrika*, **70**(3), 579-589.

Hirschfeld, H. O. (1935). A Connection Between Correlation and Contingency. *Cambridge Philosophical Society Proceedings*, **31**, 520-524.

Hoaglin, D. (2003). John W. Tukey and Data Analysis. *Statistical Science*, **18**(3), 311-318.

Hoenig, J. & Heisey, D. (2001). The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. *The American Statistician*, **55**(1), 19-24.

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

Hoffman, D. & Franke, G. (1986). Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research. *Journal of Marketing Research*, **XXIII**, 213-227.

Hofmann, H. (2003). Constructing and Reading Mosaic Plots. *Computational Statistics & Data Analysis*, **43**, 565-580.

Hogg, R. & Craig, A. (1978). *Introduction to Mathematical Statistics*. New York: Macmillan Publishing Co., Inc.

Hongjin, J., Yongzheng, Z. & Xisheng, W. (1995). Correspondence Cluster Analysis and its Application in Exploration Geochemistry. *Journal of Geochemical Exploration*, **55**, 137-144.

Hopkins, W. (1997). *A New View of Statistics*. Διαθέσιμο στην ιστοσελίδα: http://www.sportsci.org/resource/stats/index.html.

Horn, M. & Vollandt, R. (2000). A Survey of Sample Size Formulas for Pairwise and Many-One Multiple Comparisons in the Parametric, Nonparametric and Binomial Case. *Biometrical Journal*, **42**(1), 27-44.

Hotelling, H. (1936). Relations Between Two Sets of Variates. *Biometrika*, **28**(3/4), 321-377.

Hox, J. J. (1995). *Applied Multilevel Analysis*. Amsterdam: TT-Puplikaties.

Hoyle, R. (Ed.) (1999). *Statistical Strategies for Small Sample Research*, (pp. 167-194). Thousand Oakes: Sage Publications, Inc.

Huang, D.-Y. & Tseng, S.-T. (1992). A Decision Procedure for Determining the Number of Components in Principal Component Analysis. *Journal of Statistical Planning and Inference*, **30**, 63-71.

Hubbard, R. & Armstrong, S. (1992). Are Null Results Becoming an Endangered Species in Marketing? *Marketing Letters*, **3**(2), 127-136.

Hubert, L. & Arabie, P. (1992). Correspondence Analysis and Optimal Structurals Representations. *Psychometrika*, **56**(1), 119-140.

Huck, S. (2000α). Misconceptions. In *RSR: Reading Statistics & Research-Student Help*, Chapter 9. Διαθέσιμο στην ιστοσελίδα: http://www.readingstats.com.

Huck, S. (2000β). *Reading Statistics and Research*. New York: Addison Wesley Longman, Inc.

Hung, J., O'Neill, R., Bauer, P. & Kohne, K. (1997). The Behavior of the P-Value When the Alternative Hypothesis is True. *Biometrics*, **53**(1), 11-22.

Hwang, H. & Takane, Y. (2002). Generalized Constrained Multiple Correspondence Analysis. *Psychometrika*, **67**(2), 211-224.

Inselberg, A. & Dimsdale, B. (1994). Multidimensional Lines I: Representation. *SIAM Journal of Applied Mathematics*, **54**(2), 559-577.

Ishii-Kuntz, M. (1994). *Ordinal Log-Linear Models*. Thousand Oakes: Sage Publications.

Israëls, A. (1987). *Eigenvalue techniques for Qualitative Data*. Leiden: DSWO Press.

Iwane, M., Palensky, J. & Plante, K. (1997). A User's Review of Commercial Sample Size Software for Design of Biomedical Studies Using Survival Data. *Controlled Clinical Trials*, **18**, 65-83.

Jaccard, J. (1998). *Interaction Effects in Factorial Analysis of Variance*. Thousand Oakes: Sage Publications.

Jaccard, J., Turrisi, R. & Wan, C. (1990). *Interaction Effects in Multiple Regression*. Newbury Park: Sage Publications.

Jackson, D. (1993). Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches. *Ecology*, **74**(8), 2204-2214.

Jackson, E. (1991). *A User's Guide to Principal Component's*. New York: John Willey & Sons, Inc.

Jacoby, W. (1998). *Statistical Graphics for Visualizing Multivariate Data*. Thousands Oakes: Sage Publications.

Jacoby, W. (1999). Levels of Measurements and Political Research: An Optimistic View. *American Journal of Political Science*, **43**(1), 271-301.

Javeau, C., (1996). *Η Έρευνα με Ερωτηματολόγιο*. Αθήνα: Τυπωθήτω, ΓΙΩΡΓΟΣ ΔΑΡΔΑΝΟΣ.

Jhun, M. & Jeong, H.-C. (2000). Applications of Bootstrap Methods for Categorical Data Analysis. *Computational Statistics & Data Analysis*, **35**, 83-91.

John, S. (1968). A Central Tolerance Region for the Multivariate Normal Distribution. *Journal of the Royal Statistical Society. Series B (Methodological)*, **30**(3), 599-601.

Johnson, D. (1998). *Applied Multivariate Methods for Data Analysis*. Pacific Grove: Brook/Cole Publishing Company.

Johnson, N. L. & Pearson, E. S. (1969). Tables of Percentage Points of Non-Central χ^2. *Biometrika*, **56**(2), 255-272.

Johnson, R. & Wichern, D. (1992). *Applied Multivariate Statistical Analysis*. New Jersey: Prentice – Hall, Inc.

Kachigan, S. K. (1991). *Multivariate Statistical Analysis: A Conceptual Introduction*. New York: Radius Press.

Kaciak, E. & Louviere, J. (1990). Multiple Correspondence Analysis of Multiple Choise Experiment Data. *Journal of Marketing Research*, **XXVII**, 455-465.

Kakai, H. *et al*. (2003). Ethnic Differences in Choices of Health Information by Cancer Patients Using Complementary and Alternative Medicine: An Exploratory Study with Correspondence Analysis. *Social Science & Medicine*, **56**, 851-862.

Kalantari, B., Lari, I., Rizzi, A. & Simeone, B. (1993). Sharp Bounds for the Maximum of the Chi-Square Index in a Class of Contingency Tables with Given Marginals. *Computational Statistics & Data Analysis*, **16**, 19-34.

Kale, B. K. (1964). A Note on the Loss of Information Due to Grouping of Observations. *Biometrika*, **51**(3/4), 495-497.

Kalman, D. (1996). A Singularly Valuable Decomposition: The SVD of a Matrix. *The College Mathematics Journal*, **27**(1), 2-23.

Kaniska, A. *et al*. (1999). Delineation of Cryptogenetic Lennox-Gastaut Syndrome and Myoclonic Astatic Epilepsy Using Multiple Correspondence Analysis. *Epilepsy Research*, **36**, 15-29.

Kargopoulos P. & Raftopoulos T. (1998). *The Science of Logic & The Art of Thinking*. Thessaloniki: Vanias Publishing House.

Karlis, D., Saporta, G. & Spinakis, A. (2003). A Simple Rule for the Selection of Principal Components. *Communication in Statistics, Theory and Methods*, **32**(3), 643-666.

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

Keeling, K. (2000). A Regression Equation for Determining the Dimensionality of Data. *Multivariate Behavioral Research*, **25**(4), 457-468.

Kent, J. & Hainsworth, T. (1995). Confidence Intervals for the Noncentral Chi-Squared Distribution. *Journal of Statistical Planning and Inference*, **46**, 147-159.

Kent, M. & Coker, P. (1996). *Vegetation Description and Analysis: A Practical Approach*. Chichester: John Willey & Sons Ltd.

Kent, R. (1993). *Marketing Research in Action*. London: Routledge.

Khattree, R. & Naik, D. N. (2002). Andrews Plots for Multivariate Data: Some New Suggestions and Applications. *Journal of Statistical Planning and Inference*, **100**, 411-425.

Kiefer, J. (1959). Optimum Experimental Designs. *Journal of the Royal Statistical Society, Series B (Methodological)*, **21**(2), 272-319).

Kiefer, J. (1974). General Equivalence Theory for Optimum Designs (Approximate Theory). *The Annals of Statistics*, **2**(5), 849-879.

Kim, J. O. & Mueller, C. (1978). *Factor Analysis Statistical Methods and Practical Issues*. Beverly Hills: Sage.Publications.

Kimball, A. W. (1957). Errors of the Third Kind in Statistical Consulting. *Journal of the American Statistical Association*, **52**(278), 133-142.

Kinnear, P. & Gray, C. (1995). *SPSS for Windows Made Simple*. Hove, East Sussex: Erlbaum (UK) Taylor & Francis.

Kinnear, T. & Taylor, J. (1996). *Marketing Research: An Applied Approach*. New York: McGraw-Hill, Inc.

Kirk, R. (1995). *Experimental Design: Procedures for the Behavioral Sciences*. Pacific Grove: Brooks/Cole Publishing Company.

Kirkwood, B. (1996). *Essentials of Medical Statistics*. London: Blackwell Science, Ltd.

Kish, L., (1959). Some Statistical Problems in Research Design. *American Sociological Review*, **24**, 328-338.

Kishino, H., Hanyu, K., Yamashita, H. & Hayashi, C. (1998). Correspondence Analysis of Paper Recycling Society: Consumers and Paper Makers in Japan. *Resources, Conservation and Recycling*, **23**, 193-208.

Klecka, W. (1980). *Discriminant Analysis*. Beverly Hills: Sage Publications.

Kleinbaum, D., Kupper, L., Muller, K. & Nizam, A. (1998). *Applied Regression Analysis and Other Multivariable Methods*. Pacific Grove: Duxbury Press.

Klockars A. & Sax, G. (1986). *Multiple Comparisons*. Newbury Park: Sage Publications.

Knight, K. (2000). *Mathematical Statistics*. Boca Raton: Chapman & Hall/CRC.

Knoke, D. & Burke, P. (1980). *Log-Linear Models*. Newbury Park: Sage Publications.

Koehler, K. (1986). Goodness-of-Fit Tests for Log-Linear Models in Sparse Contingency Tables. *Journal of the American Statistical Association*, **81**(394), 483-493.

Kosinski, A. (1999). A Procedure for the Detection of Multivariate Outliers. *Computational Statistics & Data Analysis*, **29**, 145-161.

Kraemer, H. C. & Thiemann, S. (1987). *How Many Subjects? Statistical Power Analysis in Research*. Newbury Park: Sage Publications, Inc.

Kramer, S. & Rosental, R. (1999). Effect Sizes and Significance Levels in Small-Sample Research. In R. Hoyle (Ed.), *Statistical Strategies for Small Sample Research* (pp. 59-79). Thousand Oakes: Sage Publications, Inc.

Kritzer, B. (1996). *Surviving Statistical Spitting Matches*. A Professional Development Seminar Presentation for Senior Staff of the National Conference of State Legislatures, Madison, Wisconsin, October 10, 1996. Διαθέσιμο στην ιστοσελίδα: http://www.polisci.wisc.edu/~kritzer/misc/ legstaff/legstaff.htm

Kroke, A. *et al*. (2001). Assignment to Menopausal Status and Estimation of Age at Menopause for Women with Missing or Invalid Data-A Probabilistic Approach with Weighting Factors in a Large-Scale Epidemiological Study. *Maturitas*, **40**, 39-46.

Kroonenberg, P. & Lombardo, R. (1999). Nonsymmetric Correspondence Analysis: A Tool for Analysing Contingency Tables With a Dependence Structure. *Multivariate Behavioral Research*, **34**(3), 367-396.

Kruskal, J. & Shepard, R. (1974). A Nonmetric Variety of Linear Factor Analysis. *Psychometrika*, **39**(2), 123-157.

Kruskal, J. & Wish, M. (1978). *Multidimensional Scaling*. Newbury Park: Sage Publications.

Krzanowski, W. J. & Radley, D. (1989). Nonparametric Confidence and Tolerance Regions in Canonical Variate Analysis. *Biometrics*, **45**(4), 1163-1173.

Krzanowski, W. J. (1987α). Cross-Validation in Principal Components Analysis. *Biometrics*, **43**(3), 575-584.

Krzanowski, W. J. (1987β). Selection of Variables to Preserve Multivariate Data Structure, Using Principal Components. *Applied Statistics*, **36**(1), 22-33.

Krzanowski, W. J. (1993). Attribute Selection in Correspondence Analysis of Incidence Matrices. *Applied Statistics*, **42**(3), 529-541.

Kuehl, R. (2000). *Designs of Experiments: Statistical Principles of Research Design and Analysis*. Pacific Grove: Duxbury Thomson Learning.

Kuriki, S. (2005). Asymptotic Distribution of Inequality-Restricted Canonical Correlation With Application to Tests for Independence in Ordered Contingency Tables. *Journal of Multivariate Analysis*, **94**, 420-449.

Kutner, M., Nachtsheim, C., Neter, J. & Li, W. (2005). *Applied Linear Statistical Models*. Singapore: McGraw-Hill, Inc.

Lachin, J. (1977). Sample Size Determinations for rxc Comparative Trials. *Biometrics*, **33**(2), 315-324.

Lal Saxena, K. M. & Alam, K. (1982). Estimation of the Non-Centrality Parameter of a Chi Squared Distribution. *The Annals of Statistics*, **10**(3), 1012-1016.

Lancaster, H. O. (1957). Some Properties of the Bivariate Normal Distribution Considered in the Form of a Contingency Table. *Biometrika*, **44**(1/2), 289-292.

Lancaster, H. O. (1963α). Canonical Correlations and Partitions of χ^2. *Q. J. Math.*, **14**, 220-224.

Lancaster, H. O. (1963β). Correlations and Canonical Forms of Bivariate Distributions. *The Annals of Mathematical Statistics*, **34**(2), 532-538.

Lancaster, H. O. (1969). *The Chi-Squared Distribution*. New York: John Wiley, Inc.

Landis, R. & Koch, G. (1977). A One-Way Components of Variance Model for Categorical Data. *Biometrics*, **33**(4), 671-679.

Langeheine, R., Pannekoek, J. & Van de Pol, F. (1996). Bootstrapping Goodness-of – Fit Measures in Categorical Data Analysis. *Sociological Methods & Research*, **24**(4), 492-516.

Langley, R. (1971). *Practical Statistics Simply Explained*. New York: Dover Publications, Inc.

Lauro, C. & Balbi, S. (1999). The Analysis of Structured Qualitative Data. *Applied Stochastic Models and Data Analysis*, **15**, 1-27.

Lauro, C. & Siciliano, R. (1989). Exploratory Methods and Modeling for Contingency Tables Analysis: An Integrated Approach. *Statistica Applicata*, **1**(1), 5-32.

Lavit, C., Escoufier, Y., Sabatier, R. & Traissac, P. (1994). The ACT (STATIS method). *Computational Statistics & Data Analysis*, **18**, 97-119.

Le Roux, B. & Chiche, J. (2004). Specific Multiple Correspondence Analysis. *Τετράδια Ανάλυσης Δεδομένων*, **4**/04, 30-41.

Le Roux, B. & Rouanet, H. (1998). Interpreting Axes in Multiple Correspondence Analysis: Method of the Contributions of Points and Deviations. In J. Blasius and M. Greenacre (Eds), *Visualization of Categorical Data*, (pp. 197-220). San Diego: Academic Press.

Le Roux. B. & Rouanet, H. (2004). *Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis*. Dordrecht: Kluwer Academic Publishers.

Lebart, L. & Mirkin, B. G. (1993). Correspondence Analysis and Classification. In C. M. Cuadras & C. R. Rao (Eds), *Multivariate Analysis: Future Directions 2*, (pp. 341-357). Amsterdam: North-Holland.

Lebart, L. (1976). The Significance of Eigenvalues Issued from Correspondence Analysis. In *Proceedings in Computational Statistics (COMPSTAT)*, (pp. 38-45). Vienna: Physica – Verlag.

Lebart, L. (1994). Complementary Use of Correspondence Analysis and Cluster Analysis. In M. Greenacre and J. Blasius (Eds), *Correspondence Analysis in the Social Sciences*, (pp. 162-178). London: Academic Press.

Lebart, L. (2005). Validation Techniques in Simple and Multiple Correspondence Analysis. *Τετράδια Ανάλυσης Δεδομένων*, **6**/05, 10-25.

Lebart, L., Morineau, A. & Piron, M. (2000). *Statistique Exploratoire Multidimensionnelle*. Paris: Dunod.

Lebart, L., Morineau, A. & Tabard, N. (1977). *Techniques de la Description Statistique: Méthodes et Logiciels pour l'Analyse des Grands Tableaux*. Paris: Dunod.

Lebart, L., Morineau, A. & Warwick, K. M. (1984). *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. New York: John Willey, Inc.

Lee, S. & Zelen, M. (2000). Clinical Trials and Sample Size Considerations: Another Perspective. *Statistical Science*, **15**(2), 95-103.

Leik, R. & Gove, W. (1971). Integrated Approach to Measuring Association. *Sociological Methodology*, **3**, 279-301.

Lenth, R. (2001). Some Practical Guidelines for Effective Sample Size Determination. *The American Statistician*, **55**(3), 187-193.

Levesque, R. (2005). *SPSS Programming and Data Management: A Guide for SPSS and SAS Users*. Chicago: SPSS Inc.

Lewis, G., Mathieu, D. & Phan-Tan-Luu, R. (1999). *Pharmaceutical Experimental Design*. New York: Marcel Dekker, Inc.

Lewis, R. (2000). *Power Analysis and Sample Size Determination: Concepts and Software Tools*. Paper Presented at the 2000 Annual Meeting of the Society for Academic Emergency Medicine (SAEM) in San Francisco, California.

Liao, T.-F. (1994). *Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models*. Thousand Oakes: Sage Publications.

Light, R. & Margolin, B. (1971). An Analysis of Variance for Categorical Data. *Journal of the American Statistical Association*, **66**(335), 534-544.

Lindelöf, B. *et al*. (1991). PUVA and Cancer: A Large Scale Epidemiological Study. *The Lancet*, **338**(8759), 91-93.

Lindley, D. (1997). The Choice of Sample Size. *The Statistician*, **46**(2), *Special Issue: Sample Size Determination*, 129-138.

Lipkovich, I. & Smith, E. P. (2002). Biplot and Singular Value Decomposition Macros for Excel©. *Journal of Statistical Software*, **7**(5), 1-15. Διαθέσιμο στην ιστοσελίδα: http://www.jstatsoft.org.

Lipschutz, S. & Lipson, M. (2003). *Γραμμική Άλγεβρα*. Θεσσαλονίκη: Εκδόσεις Τζιόλα.

Lipsey, M. (1990). *Design Sensitivity: Statistical Power for Experimental Research*. Newbury Park: Sage Publications, Inc.

Lipsitz, S., Laird, N. & Harrington, D. (1990). Finding the Design Matrix for the Marginal Homogeneity Model. *Biometrika*, **77**(2), 353-358.

Lobry, J. & Chessel, D. (2003). Internal Correspondence Analysis of Codon and Amino-Acid Usage in Thermophilic Bacteria. *J. Appl. Genet.*, **44**(2), 235-261.

Loftus, R., (1991). On the Tyranny of Hypothesis Testing in the Social Sciences. *Contemporary Psychology*, **36**, 102-105.

Lohninger, H., (1999). *Teach Me Data Analysis: Single User Edition*, [Computer program manual]. New York: Springer.

Lohr, S. (1999). *Sampling: Design and Analysis*. Pacific Grove: Duxbury Press.

López-Blázquez, F. (2000). Unbiased Estimation in the Non-Central Chi-Square Distribution. *Journal of Multivariate Analysis*, **75**, 1-12.

Loslever, P. & Bouilland, S. (1999). Marriage of Fuzzy Sets and Multiple Correspondence Analysis: Examples with Subjective Interval Data and Biomedical Signals. *Fuzzy Sets and Systems*, **107**, 255-275.

MacCallum, R. & Hong, S. (1997). Power Analysis in Covariance Structure Modeling Using GFI and AGFI. *Multivariate Behavioral Research*, **32**(2), 193-210.

MacCallum, R. C., Widaman, K. F., Zhang, S. & Hong, S. (1999). Sample Size in Factor Analysis. *Psychological Methods*, **4**, 84-99.

MacCallum, R., Widaman, K., Preacher, K. & Hong, S. (2001). Sample Size in Factor Analysis: The Role of Model Error. *Multivariate Behavioral Research*, **36**(4), 611-637.

Magnussen, P. (2003). Treatment and Re-treatment Strategies for Schistosomiasis Control in Different Epidemiological Settings: A Review of 10 Years' Experience. *Acta Tropica*, **86**, 243-254.

Makarenkov, V. & Legendre, P. (2001). Optimal Variable Weighting for Ultrametric and Additive Trees and K-Means Partitioning: Methods and Software. *Journal of Classification*, **18**, 245-271.

Malhotra, N. K. (1996). *Marketing Research. An Applied Orientation*. Englewood Cliffs: Prentice Hall.

Malinvaud, E., (1987). Data Analysis in Applied Socio-Economic Statistics With Special Consideration of Correspondence Analysis. *Marketing Science Conference Proceedings, HEC-ISA*. Joy en Josas.

Malmgren, B., Oviatt, C., Gerber, R. & Jeffries, P. (1978). Correspondence Analysis: Applications to Biological Oceanographic Data. *Estuarine and Coastal Marine Science*, **6**, 429-437.

Manfredi, R. (2004). HIV Infection and Advanced Age Emerging Epidemiological, Clinical, and Management Issues. *Ageing Research Reviews*, **3**, 31-54.

Manly, B. (1994). *Multivariate Statistical Methods. A Primer*. Chapman & Hall, London.

Marchand, P. & T. Holland (2002). *Graphics and GUIs with MATLAB*. Chapman & Hall/CRC.

Mardia, K., Kent, J. & Bibby, J. (2003). *Multivariate Analysis*. London: Academic Press.

Margolin, B. & Light, R. (1974). An Analysis for Variance for Categorical Data, II: Small Sample Comparisons with Chi Square and Other Competitors. *Journal of the American Statistical Association*, **69**(347), 755-764.

Markus, M. (1994α). *Bootstrap Confidence Regions in Nonlinear Multivariate Analysis*. Leiden University Leiden: DSWO Press.

Markus, M. (1994β). Bootstrap Confidence Regions for Homogeneity Analysis; the Influence of Rotation on Coverage Percentages. In R. Dutter and W. Grossmann (Eds), *Proceedings: Computational Statistics COMPSTAT*, (pp. 337-342).

Martens, B. (1994). Analyzing Event History Data by Cluster Analysis and Multiple Correspondence Analysis: An Example Using Data about Work and Occupation of Scientists and Engineers. In M. Greenacre and J. Blasius (Eds), *Correspondence Analysis in the Social Sciences. Recent Developments and Applications*, (pp. 233-251). London: Academic Press.

Martin, J.-F. (1988). On Probability Coding. In J. Van Rijckevorsel & J. De Leeuw (Eds), *Component and Correspondence Analysis. Dimension Reduction by Functional Approximation*, (pp. 103-114). Chichester: John Willey & Sons Ltd.

Martín-Rodríguez, J. *et al*. (2002). Comparison and Integration of Subspaces from a Biplot Perspective. *Journal of Statistical Planning and Inference*, **102**, 411-423.

Mathew, T. & Nordstrom, K. (1997). Wishart and Chi-Square Distributions Associated with Matrix Quadratic Forms. *Journal of Multivariate Analysis*, **61**, 129-143.

McDonald, R. (1981). The Dimensionality of Tests and Items. *British Journal of Mathematical and Statistical Psychology*, **34**(1), 100-117.

McEwan, J. & Schlich, P. (1992). Correspondence Analysis in Sensory Evaluation. *Food Quality and Preference*, **3** (1991/92), 23-36.

McKinlay, S. (1975). The Design and Analysis of the Observational Study - A Review. *Journal of the American Statistical Association*, **70**(351), 503-520.

Mead, R. & Curnow, R. N. (1990). *Statistical Methods in Agriculture and Experimental Biology*. London: Chapman and Hall.

Mehta C. & Patel R. (1996). *SPSS Exact Tests 7.0 for Windows*. Chicago: SPSS Inc.

Mellinger, M. (1987). Correspondence Analysis: The Method and Its Application. *Chemometrics and Intelligent Laboratory Systems*, **2**, 61-77.

Meloun, M., Čapec, J., Milkšik, P. & Brereton, R. (2000). Critical Comparison of Methods Predicting the Number of Components in Spectroscopic Data. *Analytica Chimica Acta*, 20736, 1-18.

Mendenhall, W. & Sincich, T. (1996). *A Second Course in Statistics: Regression Analysis*. New Jersey: Prentice Hall, Inc.

Menexes, G. (1998). *An Investigation into Theories of the Development of Concepts of Probability*. Dissertation Submitted in Part Fulfilment of the Degree of Master of Arts in Education Studies of the University of Surrey.

Meng, R. & Chapman, D. (1966). The Power of Chi Square Tests for Contingency Tables. *Journal of the American Statistical Association*, **61**(316), 965-975.

Mertens, D. (1998). *Research Methods in Education and Psychology: Integrating Diversity with Quantitative & Qualitative Approaches*. Thousand Oaks, Sage Publications, Inc.

Merz, C. (1997). Using Correspondence Analysis to Combine Classifiers. *Machine Learning*, **0**, 1-26.

Meulman J. & Heiser, W. (2001). *SPSS Categories 11.0*. Chicago: SPSS Inc.

Meulman J. & Heiser, W. (2004). *SPSS Categories 13.0*. Chicago: SPSS Inc.

Meulman, J. & Heiser, W. (1998). Visual Display of Interaction in Multiway Contingency Tables by Use of Homogeneity Analysis: The 2×2×2×2 Case. In J. Blasius & M. Greenacre (Eds), *Visualization of Categorical Data*, (pp. 277-296). San Diego: Academic Press.

Meulman, J., (1999). *Optimal Scaling Methods for Multivariate Categorical Data Analysis*. SPSS White Paper. Διαθέσιμο στην ιστοσελίδα: http://www.spss.com/cool/papers/optimal_scaling.html

Meyer, C. (2000). *Matrix Analysis and Applied Linear Algebra*. Philadelphia: Society for Industrial and Applied Mathematics.

Meyer, T. & Mark, M. (1996). Statistical Power and Implications of Meta-Analysis for Clinical Research in Psychosocial Oncology. *Journal of Psychosomatic Research*, **41**(5), 409-413.

Michailidis, G. & De Leeuw, J. (1997). *A Regression Model for Multilevel Homogeneity Analysis*. UCLA Statistics Series, #212.

Michailidis, G. & De Leeuw, J. (1998). The Gifi System of Descriptive Multivariate Analysis. *Statistical Science*, **13**(4), 307-336.

Michailidis, G. & De Leeuw, J. (2000). Multilevel Homogeneity Analysis With Differential Weighting. *Computational Statistics & Data Analysis*, **32**, 411-442.

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

Michailidis, G. (1996). *Multilevel Homogeneity Analysis*. Διδακτορική διατριβή που υποβλήθηκε στο Τμήμα Μαθηματικών του Πανεπιστημίου της Καλιφόρνια στο Λος Άντζελες.

Micheloud, F.-X., (1997). *Jean Paul Benzécri's Correspondence Analysis*. Διαθέσιμο στην ιστοσελίδα: http://www.micheloud.com/FXM/COR/E/index.htm.

Milian, L. & Whittaker, J. (1995). Application of the Parametric Bootstrap to Models that Incorporate a Singular Value Decomposition. *Applied Statistics*, **44**(1), 31-49.

Miller, J., Daly, J., Wood, M., Roper, M. and Brooks, A., (1997). Statistical Power and its Subcomponents - Missing and Misunderstood Concepts in Empirical Software Engineering Research. *Information and Software Technology*, **39**, 285-295.

Mirkin, B. (2001). Eleven Ways to Look at the Chi-Squared Coefficient for Contingency Tables. *The American Statistician*, **55**(2), 111-120.

Miyake, S., Loslever, P. & Hancock, P.A. (2001). Individual Differences in Tracking. *Ergonomics*, **44**(12), 1056-1068.

Mohr, L. (1990). *Understanding Significance Testing*. Thousand Oakes: Sage Publications.

Mone, M., Mueller, G. & Mauland, W. (1996). The Perceptions and Usage of Statistical Power in Applied Psychology and Management Research. *Personnel Psychology*, **49**, 103-120.

Montgomery, D. (1997). *Design and Analysis of Experiments*. New York: John Willey & Sons, Inc.

Montgomery, D. (1999). Experimental Design for product and Process Design and Development. *The Statistician*, **48**(2), 159-177.

Mooney, C. & Duval, R. (1993). *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Newbury Park: Sage Publications.

Moran, M. A. & Gornbein, J. (1988). *CA-Correspondence Analysis*. Technical Report #87, BMDP Statistical Software, Inc., Los Angeles.

Morgan, B. (1981). Three Applications of Methods of Cluster Analysis. *The Statistician*, **30**(3), 205-223.

Morrison, D. (1967). Measurement Problems in Cluster Analysis. *Management Science*, **13**(12), *Series B, Managerial*, B775-B780.

Moussa, M. A. A. & Ouda, B. A. (1988). Correspondence Analysis of Contingency Tables. *Computer Methods and Programs in Biomedicine*, **27**, 111-119.

Muller, K., LaVange, L., Landersman-Ramey, S. and Ramey, C. (1992). Power Calculations for General Linear Multivariate Models Including Repeated Measures Applications. *Journal of the American Statistical Association*, **87**(420), 1209-1226.

Mumby, P. (2002). Statistical Power of Non-Parametric Tests: A Quick Guide for Designing Sampling Strategies. *Marine Pollution Bulletin*, **44**, 85-87.

Murphy, K. & Myors, B. (1998). *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests*. New Jersey: Lawrence Erlbaum Associates, Inc.

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

Nadarajah, S. & Kotz, S. (2005). Sampling Distributions Associated With the Multivariate *t* Distribution. *Statistica Neerlandica*, **59**(2), 214-234.

Nakayama, T. (2001). Tests for Redundancy of Some Variables in Correspondence Analysis. *Hiroshima Math. J.*, **31**, 1-34.

Nathan, G. (1972). On the Asymptotic Power of Tests for Independence in Contingency Tables from Stratified Samples. *Journal of the American Statistical Association*, **67**(340), 917-920.

Nemec, A. (1991). *Power Analysis Handbook for the Design and Analysis for Forestry Trials*. Biometrics Information Handbook Series, no 2. British Columbia: Ministry of Forests.

Neter, J., Kutner, M., Nachtsheim, C. & Wasserman, W. (1996). *Applied Linear Regression Models*. Chicago: Irwin, Inc.

Ngai, E. W. T. & Cheng, T. C. E. (1997). Identifying Potential Barriers to Total Quality Management Using Principal Components Analysis and Correspondence Analysis. *International Journal of Quality & Reliability Management*, **14**(4), 391-408.

Nicewander, A. & Price, J. (1997). A Consonance Criterion for Choosing Sample Size. *The American Statistician*, **51**(4), 311-317.

Nishisato, S. & Clavel, J. (2003). A Note on Between-Set Distances in Dual Scaling and Correspondence Analysis. *Behaviormetrika*, **30**(1), 87-98.

Nishisato, S. (1978). Optimal Scaling of Paired Comparison and Rank Order Data: An Alternative to Guttman's Formulation. *Psychometrik*a, **43**(2), 263-271.

Nishisato, S. (1980). *Analysis of Categorical Data: Dual Scaling and its Applications*. Torondo: University of Torondo Press.

Nishisato, S. (1990). Dual Scaling of Designed Experiments. In M. Schader & W. Gaul (Eds), *Knowledge, Data and Computer-Assisted Decitions*, (NATO ASI Series F: Computer and Systems Science Vol. **61**), (pp. 115-125). Berlin: Springer-Verlag.

Nishisato, S. (1993). On Quantifying Different Types of Categorical Data. *Psychometrika*, **58**(4), 617-629.

Nishisato, S. (1994). *Elements of Dual Scaling: An Itroduction to Practical Data Analysis*. New Jersey: Lawrence Erlbaum Associates, Publishers.

Nishisato, S. (1995). Graphical Representation of Quantified Categorical Data: Its Inherent Problems. *Journal of Statistical Planning and Inference*, **43**, 121-132.

Nishisato, S. (1996). Gleaning in the Field of Dual Scaling. *Psychometrika*, **61**(4), 559-599.

Nishisato, S. (1998). Graphing is Believing. In J. Blasius and M. Greenacre (Eds), *Visualization of Categorical Data*, (pp. 185-196). San Diego: Academic Press.

Nishsisato, S. & Sheu, W.-J. (1980). Piecewise Method of Reciprocal Avarages for Dual Scaling of Multiple-Choice Data. *Psychometrika*, **45**(4), 467-478.

Nix, T. & Barnett, J. J. (1998). The Data Analysis Dilemma: Ban or Abandon. A Review of Null Hypothesis Significance Testing. *Research in the Schools*, **5**(2), 3-14.

Noma, E. & Smith, R. (1985). Scaling Sociomatrices by Optimizing an Explicit Function: Correspondence Analysis of Binary Single Response Sociomatrices. *Multivariate Behavioral Research*, **20**, 179-197.

Norusis, M. (1992α). *SPSS Professional Statistics 6.1*. Chicago: SPSS Inc.

Norusis, M. (1992β). *SPSS for Windows Advanced Statistics Release 5*. Chicago: SPSS Inc.

Novak, T. & Hoffman, D. (1990). Residual Scaling: An Alternative to Correspondence Analysis for the Graphical Representation of Residuals from Log-Linear Models. *Multivariate Behavioral Research*, **25**(3), 351-370.

Nowak, E. & Bar-Hen, A. (2005). Influence Function and Correspondence Analysis. *Journal of Statistical Planning and Inference*, **134**, 26-35.

Nunnally, J. C. (1978). *Psychometric Theory*. New York: McGraw Hill Book Co.

Nutahara, H. *et al* (2001). A Simple Computerized Program for the Calculation of the Required Sample Size Necessary to Ensure Statistical Accuracy in Medical Experiments. *Computer Methods and Programs in Biomedicine*, **65**, 133-139.

O'Brien, R. (1979). The Use of Pearson's R with Ordinal Data. *American Sociological Review*, **44**(5), 851-857.

O'Neil, M. E. (1978α). Asymptotic Distributions of the Canonical Correlations from Contingency Tables. *Aust. J. Statist.*, **20**, 75-82.

O'Neil, M. E. (1978β). Distributional Expansions for Canonical Correlations from Contingency Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **40**(3), 303-312.

O'Neil, M. E. (1980). The Distribution of Higher-Order Interactions in Contingency Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **42**(3), 357-365.

O'Neil, M. E. (1981). A Note on the Canonical Correlations from Contingency Tables. *Aust. J. Statist.*, **23**, 58-66.

Oakes, M. (1986). *Statistical Inference: A Commentary for the Social and Behavioral Sciences.* Chichester: John Wiley & Sons, Inc.

Ocerin, C., Mohedano, E. & Segador, G. (1999). Automatic Aggregation of Categories in Multivariate Contingency Tables Using Information Theory. *Computational Statistics & Data Analysis*, **29**, 285-294.

Oja, H. & Randles, R. (2004). Multivariate Nonparametric Tests. *Statistical Science*, **19**(4), 598-605.

Osmond, C. (1985). Biplot Models to Canser Mortality Rates. *Applied Statistics*, **34**(1), 63-70.

Pack, P. & Jolliffe, I. T. (1992). Influence in Correspondence Analysis. *Appl. Statist.*, **41**(2), 365-380.

Pagano, M. & Gauvreau, K. (2000). *Αρχές Βιοστατιστικής*. Περιστέρι-Αθήνα: Εκδόσεις ΕΛΛΗΝ.

Pagès, J. (2004). Multiple Factor Analysis: Main Features and Application to Sensory Data. *Τετράδια Ανάλυσης Δεδομένων*, **4**/04, 7-29.

Pan, W. (2001). Sample Size and Power Calculations With Correlated Data. *Controlled Clinical Trials*, **22**, 211-227.

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

Papadimitiou, I. (1987). Decomposition d' une Matrice de Leontief par l' Analyse des Correspondances. *Les Cahiers de l' Analyse des Données*, Vol. **XII**, No 2, 147-168. Paris: Dunod.

Parkhurst, F. (1985). Interpreting Failure to Reject a Null Hypothesis. *Bulletin of the Ecological Society of America*, **66**, 301-302.

Patnaik, P. (1949). The Non-Central chi^2 and F Distribution and Their Applications. *Biometrika*, **36**(1/2), 202-232.

Pearce, S. (1979). Experimental Design: R. A. Fisher and Some Modern Rivals. *The Statistician*, **28**(3), 153-161.

Pearson, E. S. & Hartley, H. O. (Eds) (1972). *Biometrika Tables for Statisticians, Vol. II*. London: Cambridge University Press.

Pearson, K. (1906). On Certain Poits Connected With Scale Order in the Case of Correlations of Two Characters Which for Some Arrangement Give a Linear Regression Line. *Biometrika*, **5**, 176-178.

Peay, E. (1988). Multidimensional Rotation and Scaling of Configurations to Optimal Agreement. *Psychometrika*, **53**(2), 199-208.

Penny, K. (1996). Appropriate Critical Values When Testing for a Single Multivariate Outlier by Using the Mahalanobis Distance. *Applied Statistics*, **45**(1), 73-81.

Peres-Neto, P., Jackson, D. & Somers, K. (2005). How Many Principal Components? Stopping Rules for Determining the Number of Non-Trivial Axes Revisited. *Computational Statistics & Data Analysis*, **49**, 974-997.

Perreault, W. & Young, W. (1980). Alternating Least Squares Optimal Scaling: Analysis of Nonmetric Data in Marketing Research. *Journal of Marketing Research*, **XVII**, 1-13.

Perrière, G. & Thioulouse, J. (2003). Use of Correspondence Discriminant Analysis to Predict the Subcellular Location of Bacterial Proteins. *Computer Methods and Programs in Biomedicine*, **70**, 99-105.

Pfeiffer, P. (1978). *Concepts of Probability*. New York: Dover Publications, Inc.

Polgar, S. & Thomas, S. (1992). *Introduction to Research in the Health Sciences*. London: Churchill Livingstone.

Poon, W.-Y. & Hung, H.-Y. (1996). Analysis of Square Tables With Ordered Categories. *Computational Statistics & Data Analysis*, **22**, 303-322.

Pratt, W. (1976). A Discussion of the Wuestion: For What Use Are Tests of Hypotheses And Tests of Significance. *Communications in Statistics, Series A* 5, 779-787.

Preece, D. A. (1990). R. A. Fisher and Experimental Design: A Review. *Biometrics*, **46**(4), 925-935.

Proschan, F. (1953). Confidence And Tolerance Intervals for the Normal Distribution. *Journal of the American Statistical Association*, **48**(263), 550-564.

Purwins, H., Graepel, T., Blankertz, B. & Obermayer, K. (2003). Correspondence Analysis for Visualizing Inteplay of Pitch Class, Key, and Composer. In E. Luis-Puebla, G. Mazzola and T. Noll (Eds), *Perspectives in Mathematical Music Theory*, (pp. 1-23). Osnabrück: Epos-Verlag.

Qian, G., Gabor, G. & Gupta, R. P. (1994). Principal Components Selection by the Criterion of the Minimum Mean Difference Complexity. *Journal of Multivariate Analysis*, **49**, 55-75.

Queiros, C. E., Gelsema, E. S. & Timmers, T. (1983). Correspondence Analysis in the Context of Pattern Recognition. *Pattern Recognition Letters*, **1**(4), 229-236.

Raktoe, B. L. & Federer, W. T. (1970). Characterization of Optimal Saturated Main Effects Plans of the 2^n Factorial. *The Annals of Mathematical Statistics*, **41**(1), 203-206.

Raktoe, B. L. & Federer, W. T. (1973). Balanced Optimal Saturated Main Effect Plans of the 2^n Factorial and Their Relation to (v, k, λ) Configurations. *The Annals of Statistics*, **1**(5), 924-932.

Rao, C. R. (1980). Matrix Approximations and Reduction of Dimensionality in Multivariate Analysis. In P. R. Krishnaiah (Ed.), *Multivariate Analysis – V, Proceedings of the Fifth International Symposium on Multivariate Analysis*, (pp. 3-22). Amsterdam: North – Holland Publishing Company.

Rao, C. R. (1995). The Use of Hellinger Distance in Graphical Displays of Contingence Table Data. In E.-M. Tiit, T. Kolo and H. Niemi (Eds), *New Trends in Probability and Statistics, Volume 3, Multivariate Statistics and Matrices in Statistics*, (pp. 143-161). Zeist: VSP BV, TEV Ltd.

Rao, C. R. (2002). *Linear Statistical Inference and its Applications*. New York, John Willey & Sons.

Raykov, T. & Little, T. (1999). A Note on Procrustean Rotation in Exploratory Factor Analysis: A Computer Intensive Approach to Goodness-of-Fit Evaluation. *Educational and Psychological Measurement*, **59**(1), 47-57.

Reed, K. (2002). The Use of Correspondence Analysis to Develop a Scale to Measure Workplace Morale from Multi-Level Data. *Social Indicators Research*, **57**, 339-351.

Remenyi, D. (1992). Researching Information Systems: Data Analysis Methodology Using Content and Correspondence Analysis. *Journal of Information Technology*, **7**, 76-86.

Rencher, A. (2000). *Linear Models in Statistics*. New York: John Willey & Sons, Inc.

Reynolds, H. T. (1984). *Analysis of Nominal Data*. Newbury Park: Sage Publications.

Ringrose, T. J. (1992). Bootstrapping and Correspondence Analysis in Archaeology. *Journal of Archaeological Science*, **19**, 615-629.

Ritov, Y. & Gilula, Z. (1993). Analysis of Contingency Tables by Correspondence Models Subject to Order Constraints. *Journal of the American Statistical Association*, **88**(424), 1380-1387.

Roberts, J. (1996). Alternative Approaches to Correspondence Analysis of Sociomatrices. *Journal of Mathematical Psychology*, **21**(4), 359-368.

Roberts, J. (2000). Correspondence Analysis of Two-Mode Network Data. *Social Networks*, **22**, 65-72.

Rocke, D. & Woodruff, D. (1996). Identification of Outliers in Multivariate Data. *Journal of the American Statistical Association*, **91**(435), 1047-1061.

Romney, K., Moore, C. & Brazill, T. (1998). Correspondence Analysis as a Multidimensional Scaling Technique for Nonfrequency Similarity Matrices. In J. Blasius and M. Greenacre (Eds), *Visualization of Categorical Data*, (pp. 329-345). San Diego: Academic Press.

Rousseeuw, P. & Van Zomeren, B. (1990). Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Association*, **85**(411), 633-639.

Rovan, J. (1994). Visualizing Solutions in More than Two Dimensions. In J. Blasius and M. Greenacre (Eds), *Correspondence Analysis in the Social Sciences*, (pp. 210-229). London: Academic Press.

Rozeboom, W., (1960). The Fallacy of the Null-Hypothesis Significance Test. *Psychological Bulletin*, **57**, 416-428.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse Data in Surveys*. New York: John Willey & Sons, Inc.

Rummel. R. J. (1970). *Applied Factor Analysis*. Evanston: Northwestern University Press.

Sackrowitz, H. & Samuel-Cahn, E. (1999). *P* Values as Random Variables-Expected *P* Values. *The American Statistician*, **53**(4), 326-331.

Sands, R. & Young, F. (1980). Components Models fro Three-Way Data: An Alternating Least Squares Algorithm with Optimal Scaling Features. *Psychometrika*, **45**(1), 39-67.

Sankaran, M. (1963). Approximations to the Non-Central Chi-Square Distribution. *Biometrika*, **50**(1/2), 199-204.

Sapiro, G. (2002). The Structure of the French Literary Field During the German Occupation (1940-1944). A Multiple Correspondence Analysis. *Poetics*, **30**, 387-402.

Saporta, G. & Tambrea, N. (1993). About the Selection of the Number of Components in Correspondence Analysis. In J. Jansen and C. Skiadas (Eds), *Applied Stochastic Models and Data Analysis*, (pp. 846-856). Singapore: World Scientific Publishing Co. Pte. Ltd.

SAS Institute, (1990). *SAS/STAT User's Guide, Version 6*, 4th ed. (Vol. 1). Cary, NC: SAS Institute, Inc.

SAS Institute, (1999). *SAS/STAT User's Guide Version 8*. Cary, NC: SAS Institute, Inc.

Schabenberger, O., Gregoire, T. & Kong, F. (2000). Collections of Simple Effects and Their Relationship to Main Effects and Interactions in Factorials. *The American Statistician*, **54**(3), 210-214.

Schaffer, C. & Green, P. (1996). An Empirical Comparison of Variables Standardization Methods in Cluster Analysis. *Multivariate Behavioral Research*, **31** (2), 149-167.

Schmidt, L., (1996). Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers. *Psychological Methods*, **1**(2), 115-129.

Schönemann, P. (1966). A Generalized Solution of the Orthogonal Procrustes Problem. *Psychometrika*, **31**(1), 1-10.

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

Scott, S. (1989). PC-MDS: *A Multivariate Statistics Package*. Provo, Utah: Brigham Young University.

Sharma, S. (1996). *Applied Multivariate Techniques*. New York: John Willey & Sons, Inc.

Sheppard, C. (1999). How Large Should My Sample Be? Some Quick Guides to Sample Size and the Power of Tests. *Marine Pollution Bulletin*, **38**(6), 439-447.

Shih, W. J. & Zhao, P.-L. (1997). Design for Sample Size Re-Estimation with Interim Data for Double-Blind Clinical Trials with Binary Outcomes. *Statistics in Medicine*, **16**, 1913-1923.

Sibson, R. (1978). Studies in the Robustness of Multidimensional Scaling: Procrustes Statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, **40**(2), 234-238.

Siciliano, R. & Mola, F. (1998). Ternary Classification Trees: A Factorial Approach. In J. Blasius and M. Greenacre (Eds), *Visualization of Categorical Data*, (pp. 311-323). San Diego: Academic Press.

Sieber, T. N., Petrini, O. & Greenacre, M. (1998). Correspondence Analysis as a Tool in Fungal Taxonomy. *Systemic and Applied Microbiology*, **21**, 433-441.

Smith, K. (1918). On the Standard Deviations of Adjusted and Interpolated Values of an Observed Polynomial Function and its Constants and the Guidance they Give Towards a Proper Choice of the Distribution of Observations. *Biometrika*, **12**(1/2), 1-85.

Snee, R. (1974). Graphical Display of Two-Way Contingency Tables. *The American Statistician*, **28**(1), 9-12.

Snelders, H. M. I. J. (Dirk) & Stokmans, M. (1994). Product Perception and Preference in Consumer Decision Making. In M. Greenacre and J. Blasius (Eds), *Correspondence Analysis in the Social Sciences. Recent Developments and Applications*, (pp. 324-349). London: Academic Press.

Spector, P. E. (1992). *Summated Rating Scale Construction: An Introduction*. Newbury Park: Sage Publications.

Sprent, P. (1973). Frank Yates and Experimental Design-Reflections Inspired by his Selected Papers. *The Statistician*, **22**(2), *Geophysical Statistical Symposium*, 151-158.

SPSS Inc. (1997). *SPSS 7.5 Statistical Algorithms*. Chicago: SPSS Inc.

SPSS Inc. (1998α). *SPSS Categories 8.0*. Chicago: SPSS, Inc.

SPSS Inc. (1998β). *SPSS Tables 8.0*. Chicago: SPSS, Inc.

SPSS Inc. (2002α). *SPSS 11.5 Syntax Reference Guide Base System Advanced Models Regression Models*. Chicago: SPSS, Inc.

SPSS Inc. (2002β). *SPSS Tables 11.5*. Chicago: SPSS, Inc.

SPSS Inc. (2004α). *SPSS 13 Statistical Algorithms*. Chicago: SPSS Inc.

SPSS Inc. (2004β). *SPSS Regression Models 13.0*. Chicago: SPSS Inc.

Srikantan, K. S. (1970). Canonical Association Between Nominal Measurements. *Journal of the American Statistical Association*, **65**(329), 284-292.

Srinivasan, V. & Basu, A. (1989). The Metric Quality of Ordered Categorical Data. *Marketing Science*, **8**(3), 205-230.

Srivastava, M. S. (2002). *Methods of Multivariate Statistics*. New York: John Wiley & Sons, Inc.

Stapleton, J. (1995). *Linear Statistical Models*. New York: John Willey & Sons, Inc.

Stark, A. (1990). An Algorithm for Computing Standard Errors from Categorical Data. *Computational Statistics & Data Analysis*, **10**, 293-296.

Steel, R. & Torrie, J. (1986). *Principles and Procedures of Statistics: A Biometrical Approach*. Singapore: McGraw-Hill Book Company.

Stevens, J. (2002). *Applied Multivariate Statistics for the Social Sciences*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Stoddard, A. (1979). Standardization of Measures Prior to Cluster Analysis. *Biometrics*, **35**(4), 765-773.

Stokes, L. & Plummer, J. (2004). Using Spreadsheet Solvers in Sample Design. *Computational Statistics & Data Analysis*, **44**, 527-546.

Strang, G., (2001). *Γραμμική Άλγεβρα και Εφαρμογές*. Ηράκλειο: Πανεπιστημιακές Εκδόσεις Κρήτης.

Strub, M. (2000). Reliability and Generalizability Theory. In L. Grimm and P. Yarnold (Eds), *Reading and Understanding More Multivariate Statistics*, (pp. 23-66). Washington: American Psychological Association.

Szczesny, W. (2002). Grade Correspondence Analysis Applied to Contingency Tables and Questionnaire Data. *Intelligent Data Analysis*, **6**, 17-51.

Tabachnick, B. & Fidell, L. (1989). *Using Multivariate Statistics*. New York: Harper & Row Publishers.

Tacq, J. (1997). *Multivariate Analysis Techniques in Social Science Research*. London: Sage Publications.

Takane, Y. & Shibayama, T. (1991). Principal Components Analysis with External Information on Both Subjects and Variables. *Psychometrika*, **56**(4), 97-120.

Takane, Y., Yanai, H. & Mayekawa, S. (1991). Relationships Among Several Methods of Linearly Constrained Correspondence Analysis. *Psychometrika*, **56**(4), 667-684.

Takane, Y., Young, F. & De Leeuw, J. (1977). Nonmetric Individual Differences Multidimentional Scaling: An Alternating Least Squares Methods With Optimal Scaling Features. *Psychometrika*, **42**(1), 7-67.

Tateneni, K. & Browne, M. (2000). A Noniterative Method of Joint Correspondence Analysis. *Psychometrika*, **65**(2), 157-165.

Taylor, J. & Yu, M. (2002). Bias and Efficiency Loss Due to Categorizing an Explanatory Variable. *Journal of Multivariate Analysis*, **83**, 248-263.

Teil, H. (1975). Correspondence Factor Analysis: An Outline of its Method. *Mathematical Geology*, **7**(1), 3-12.

Ten Berge, J. & Knol, D. (1984). Orthogonal Rotations to Maximal Agreement for Two or More Matrices of Different Column Orders. *Psychometrika*, **49**(1), 49-55.

Ten Berge, J. (1977). Orthogonal Procrustes Rotation for Two or More Matrices. *Psychometrika*, **42**(2), 267-276.

Tenenhaus, M. & Young, F. (1985). An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis and Other Methods for Quantifying Categorical Multicariate Data. *Psychometrika*, **50**(1), 91-119.

Ter Braak, C. (1985). Correspondence Analysis of Incidence and Abundance Data: *Properties in Terms of Unimodal Response Model*. *Biometrics*, **41**(4), 859-873.

Ter Braak, C. (1986). Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis. *Ecology*, **67**(5), 1167-1179.

Ter Braak, C. (2002). Ordination. In R. Jongman, C. Ter Braak and O. Van Tongeren (Eds), *Data Analysis in Community and Landscape Ecology*, (pp. 91-173). Cambridge: Cambridge University Press.

Thiessen, V., Rohlinger, H. & Blasius, J. (1994). The 'Significance' of Minor Changes in Panel Data: A Correspondence Analysis of the Division of Household Tasks. In M. Greenacre and J. Blasius (Eds), *Correspondence Analysis in the Social Sciences. Recent Developments and Applications*, (pp. 252-266). London: Academic Press.

Thomas, G. E. & Kiwanga, S. S. (1993). Use of Ranking and Scoring Methods in the Analysis of Ordered Categorical Data from Factorial Experiments. *The Statistician*, **42**(1), 55-67.

Thomas, L. & Juanes, F. (1996). The Importance of Statistical Power Analysis: An Example From Animal Behaviour. *Anim. Behav.*, **52**, 856-859.

Thomas, L. & Krebs, C. (1997). A Review of Statistical Power Analysis Software. *Bulletin of the Ecological Society of America*, **78**(2). Διαθέσιμο στην ιστοσελίδα: http://sustain.forestry.ubc.ca/cacb/power/review/powrev.html

Tofallis, C. (1999). Model Building with Multiple Dependent Variables and Constraints. *The Statistician*, **48**(3), 371-378.

Toothaker, L (1993). *Multiple Comparison Procedures*. Newbury Park: Sage Publications, Inc.

Torres, A. & Greenacre, M. (2002). Dual Scaling and Correspondence Analysis of Preferences, Paired Comparisons and Ratings. *Intern. J. of Research in Marketing*, **19**, 401-405.

Torres, A. & Van de Velden, M. (2007). Perceptual Mapping of Multiple Variable Batteries by Plotting Supplementary Variables in Correspondence Analysis of Rating Data. *Food Quality and Preference*, **18**, 121-129.

Traub, R. (1994). *Reliability for the Social Sciences.* Thousand Oaks: Sage Publications, Inc.

Trendafilov, N. & Lippert, R. (2002). The Multimode Procrustes Problem. *Linear Algebra and its Applications*, **349**, 245-264.

Tryfos, P. (1996). *Sampling Methods for Applied Research: Text and Cases*. New York: John Wiley & Sons, Inc.

Tukey, J. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics*, **33**(1), 1-67.

Tukey, J. (1977). *Exploratory Data Analysis*. Menlo Park: Addison-Wesley Publishing Company.

Tukey, J. (1979). Methodology, and the Statistician's Responsibility for BOTH Accuracy AND Relevance. *Journal of the American Statistical Association*, **74**(368), 786-793.

Tukey, J. (1980). We Need Both Exploratory and Confirmatory. *The American Statistician*, **34**(1), 23-25.

Udina, F. (2005). Interactive Biplot Construction. *Journal of Statistical Software*, **13**(5), 1-16. Διαθέσιμο στην ιστοσελίδα: http://www.jstatsoft.org.

Unwin, A., Volinsky, C. & Winkler, S. (2003). Parallel Coordinates for Exploratory Modelling Analysis. *Computational Statistics & Data Analysis*, **43**, 553-564.

Valle, S., Li, W. & Qin, J. (1999). Selection of the Number of Principal Components: The Variance of the Recontruction Error Criterion with a Comparison to Other Methods. *Ind. Eng. Chem. Res.*, **38**, 4389-4401.

Van Buuren, S. & De Jeeuw, J. (1992). Equality Constraints in Multiple Correspondence Analysis. *Multivariate Behavioral Research*, **27**(4), 567-583.

Van de Geer, J. P. (1993α). *Multivariate Analysis of Categorical Data: Theory*. Thousand Oakes: Sage Publications, Inc.

Van de Geer, J., (1993β). *Multivariate Analysis of Categorical Data: Applications*. Thousand Oakes: Sage Publications, Inc.

Van de Velden, M. & Neudecker, H. (2000). On an Eigenvalue Property Relevant in Correspondence Analysis and Related Methods. *Linear Algebra*, **321**, 347-364.

Van de Velden, M. (2000). *Topics in Correspondence Analysis*. Amsterdam: Tinbergen Institute Research Series, No 238.

Van der Burg, E. & De Leeuw, J. (1983). Non-linear Canonical Correlation. *British Journal of Mathematical and Statistical Psychology*, **36**, 54-80.

Van der Burg, E. & De Leeuw, J. (1988). Homogeneity Analysis with k sets of Variables: An Alternating Least Squares Method with Optimal Scaling Features. *Psychometrika*, **53**(2), 177-197.

Van der Burg, E. & De Leeuw, J. (1988). Use of the Multinomial Jack-Knife and Bootstrap in Generalized Non-Linear Canonical Correlation Analysis. *Applied Stochastic Models and Data Analysis*, **4**, 159-172.

Van der Burg, E. & De Leeuw, J. (1994). OVERALS: Non Linear Canonical Correlation with *k* Sets of Variables. *Computational Statistics & Data Analysis*, **18**, 141-163.

Van der Heijden, P. & De Leeuw, J. (1985). Correspondence Analysis Used Complementary to Loglinear Analysis. *Psychometrika*, **50**(4), 429-447.

Van der Heijden, P. & De Leeuw, J. (1989). Correspondence Analysis, With Special Attention to the Analysis of Panel Data and Event History Data. *Sociological Methodology*, **19**, 43-87.

Van der Heijden, P. & Worsley, K. (1988). Comment on "Correspondence Analysis used Complementary to Loglinear Analysis". *Psychometrika*, **53**(2), 287-291.

Van der Heijden, P. G. M. & Escofier, B. (1988). *Multiple Correspondence Analysis with Missing Data*. I.N.R.I.A., Rapport de Recherche, No 902.

Van der Heijden, P., De Falguerolles, A. & De Leeuw, J. (1989). A Combined Approach to Contingency Table Analysis Using Correspondence Analysis and Log-Linear Analysis. *Appl. Statist*., **38**(2), 249-292.

Van der Heijden, P., De Vries, H. & Van Hoof, J. (1990). Correspondence Analysis of Transition Matrices, With Special Attention to Missing Entries and Assymetry. *Animal Behaviour*, **40**, 49-64.

Van der Heijden, P., Gilula, Z. & Van der Ark, A. (1999). An Extended Study into the Relationship Between Correspondence Analysis and Latent Class Analysis. *Sociological Methodology*, **29**, 147-186.

Van der Heijden, P., Mooijaart, A., & Takane, Y. (1994). Correspondence Analysis and Contingency Tables Models. In M. Greenacre and J. Blasius (Eds), *Correspondence Analysis in the Social Sciences. Recent Developments and Applications*, (pp. 79-111). London: Academic Press.

Van der Heijden, P., Teunissen, J. & Van Orlé, C. (1997). Multiple Correspondence Analysis as a Tool for Quantification or Classification of Career Data. *Journal of Educational and Behavioral Statistics*, **22**(4), 447-477.

Van der Kooij, A. J. & Meulman, J. (1997). MURALS: Multiple Regreesion and Optimal Scoring using Alternating Least Squares. In W. Bandilla and F. Faulbaum (Eds), *SoftStat ´97 Advances in Statistical Software 6*, (pp. 99-106). Stuttgart: Lucius & Lucius.

Van Meter, K., Schiltz, M.-A., Cibois, F. & Mounier, L. (1994). Correspondence Analysis: A History and French Sociological Perspective, In J. Blasius and M. Greenacre (Eds), *Visualization of Categorical Data*, (pp. 128-137). San Diego: Academic Press.

Van Rijckevorsel, J. & De Leeuw, J. (Eds) (1988). *Component and Correspondence Analysis. Dimension Reduction by Functional Approximation*, (pp. 103-114). Chichester: John Willey & Sons Ltd.

Van Rijckevorsel, J. (1987). *The Application of Fuzzy Coding and Horseshoes in Multiple Correspondence Analysis*. Leiden: DSWO Press.

Van Rijckevorsel, J. (1988). Fuzzy Coding and B-Splines. In J. Van Rijckevorsel and J. De Leeuw (Eds), *Component and Correspondence Analysis. Dimension Reduction by Functional Approximation*, (pp. 33-54). Chichester: John Willey & Sons Ltd.

Velleman, P. F. & Wilkinson, L. (1993). Nominal, Ordinal, Interval, and Ratio Typologies are Misleading. *American Statistician*, **47**(1), 65-72.

Verkuilen, J. (2001). *Measuring Fuzzy Set Membership Functions: A Dual Scaling Approach*. Paper Presented at the Annual Meeting of the APSA, San Francisco, CA, August 30-September 2, 2001.

Verma, R. & Goodale, J. (1995). Statistical Power in Operations Management Research. *Journal of Operations Management*, **13**, 139-152.

Vivien, M. & Sabatier, R. (2004). A Generalization of STATIS-ACT Strategy: DO-ACT for Two Multiblocks Tables. *Computational Statistics & Data Analysis*, **46**, 115-171.

Von Eye, A. & Spiel, C. (1996). Standard and Nonstandard Log-Linear Symmetry Models for Measuring Change in Categorical Variables. *The American Statistician*, **50**(4), 300-305.

Wald, A. & Wolfowitz, J. (1946). Tolerance Limits for a Normal Distribution. *The Annals of Mathematical Statistics*, **17**(2), 208-215.

Walsh, S. & Diamond, D. (1995). Non-Linear Curve Fitting Using Microsoft EXCEL *SOLVER*. *Talanta*, **42**(4), 561-572.

Ward, J. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, **58**(301), 236-244.

Wasserman, S. & Faust, K. (1989). Canonical Analysis of the Composition and Structure of Social Networks. *Sociological Methodology*, **19**, 1-42.

Weinberg, S., Carroll, D. & Cohen, H. (1984). Confidence Regions for INDSCAL Using the Jackknife and Bootstrap Techniques. *Psychometrika*, **49**(4), 475-491.

Weir, B. (1996). *Genetic Data Methods for Discrete Population Genetic Data*. Sunderland: Sinauer Associates, Inc. Publishers.

Weisberg, H., Krosnick, J. & Bowen, B. (1996). *Introduction to Survey Research, Polling, and Data Analysis*. Thousand Oakes: Sage Publications, Inc.

Weller, S. & Romney A. K. (1990). *Metric Scaling: Correspondence Analysis*. Newbury Park: Sage Publications.

Westfall, P. & Young, S. (1989). *P* Value Adjustments for Multiple Tests in Multivariable Binomial Models. *Journal of the American Statistical Association*, **84**(407), 780-786.

Wilks, S. S. (1942). Statistical Prediction With Special Reference to the Problem of Tolerance Regions. *The Annals of Mathematical Statistics*, **13**(4), 400-409.

Williams, A. (1950). On the Choice of the Number and Width of Classes for the Chi-Square Test of Goodness of Fit. *Journal of the American Statistical Association*, **45**(249), 77-86.

Williams, E. J. (1952). Use of Scores for the Analysis of Association in Contingency Tables. *Biometrika*, **39**(3/4), 274-289.

Williams, W. (1971). Principles of Clustering. *Annual Review of Ecology and Systematics*, **2**, 303-326.

Winsberg, S. (1988). Two Techniques: Monotone Spline Transformations for Dimension Reduction in PCA and Easy-to-Generate Metrics for PCA of Sampled Functions. In J. Van Rijckevorsel and J. De Leeuw (Eds), *Component and Correspondence Analysis. Dimension Reduction by Functional Approximation*, (pp. 115-135). Chichester: John Willey & Sons Ltd.

Wolf, F. (1986). *Meta-Analysis: Quantitative Methods for Research Synthesis*. Newbury Park: Sage Publications.

Wolter, K. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Wuebben, P. (1968). Experimental Design, Measurement, and Human Subjects: A Neglected Problem of Control. *Sociometry*, **31**(1), 89-101.

Yates, F. (1951). The Influence of Statistical Methods for Research Workers on the Development of the Science of Statistics. *Journal of the American Statistical Association*, **46**, 19-34.

Yick, J. & Lee, A. (1998). Unmasking Outliers in Two-Way Contingency Tables. *Computational Statistics and Data Analysis*, **29**, 69-79.

Yoccuz, G. (1991). Use, Overuse, and Misuse of Significance Tests in Evolutionary Biology and Ecology. *Bulletin of the Ecological Society of America*, **72**, 106-111.

Young, F. (1981). Quantitative Analysis of Qualitative Data. *Psychometrika*, **46**(4), 357-388.

Young, F. (1992). *Vista: The Visual Statistics System*. UNC Psychometric Laboratory, Chapel Hill NC.

Young, F. Valero-Mora, P. & Ladesma-Mouripo, R. D. (2000). Visualizing Categorical Data in Vista. In Proceedings *I Congreso de Metodos Numericos en Ciencias Sociales*, (pp. 196-207). Barcelona.

Young, F., De Leeuw, J. & Takane, Y. (1976). Regression With Qualitative and Quantitative Variables: An Alternating Least Squares Methods With Optimal Scaling Features. *Psychometrika*, **41**(4), 505-529.

Young, F., Faldowski, R. & McFarlane, M. (1993). Multivariate Statistical Visualization. In C. R. Rao (Ed.), *Computational Statistics. Handbbok of Statistics*, Vol. **9**, (pp. 959-998). Amsterdam: Elsevier Science.

Young, F., Takane, Y. & De Leeuw, J. (1978). The Principal Components of Mixed Measurment Level Multivariate Data: An Alternating Least Squares Methods With Optimal Scaling Features. *Psychometrika*, **43**(2), 279-281.

Zar, J. (1996). *Biostatistical Analysis*. New Jersey: Prentice-Hall International, Inc.

Zelen, M. (1991). Multinomial Response Models. *Computational Statistics & Data Analysis*, **12**, 249-254.

Ελληνική

Αθανασιάδης, Η. (1995). *Παραγοντική Ανάλυση Αντιστοιχιών και Ιεραρχική Ταξινόμηση*. Αθήνα: Εκδόσεις ΝΕΩΝ ΤΕΧΝΟΛΟΓΙΩΝ.

Αθανασιάδης, Η. (2002). Η Γεωγραφική Κατανομή της Κοινωνικής Προέλευσης των Φοιτητών. *Τετράδια Ανάλυσης Δεδομένων*, **2**/02, 95-102.

Αθανασιάδης, Η. (2005). Οι Σπουδαστές των Τ.Ε.Ι. και το Επίπεδο Εκπαίδευσης των Γονέων τους. *Τετράδια Ανάλυσης Δεδομένων*, **6**/05, 101-113.

Αθανασίου, Γ. & Παπαδημητρίου, Γ. (2002). Ανάλυση Κειμένων με Μεθόδους Ανάλυσης Δεδομένων (Αφορμή: Ο «Εραστής της Μ. Duras»). *Τετράδια Ανάλυσης Δεδομένων*, **1**/02, 107-117.

Αναστασιάδου, Σ. & Καράκος, Α. (2005). Ανάλυση Δεδομένων σε Θέματα Αλλαγής Στάσεων των Φοιτητών. *Τετράδια Ανάλυσης Δεδομένων*, **6**/05, 138-151.

Αναστασιάδου, Σ. & Κοσμά, Α. (2004). Συγκριτική Μελέτη Αντιμετώπισης του Παραμυθιού από τα Παιδιά της Προσχολικής Ηλικίας. *Τετράδια Ανάλυσης Δεδομένων*, **4**/04, 136-149.

Αναστασιάδου, Σ. & Παπαδημητρίου, Γ. (2001α). Χρήση Μεθόδων της Πολυδιάστατης Στατιστικής Ανάλυσης για τον Προσδιορισμό των Διαθέσεων των Φοιτητών προς τη Στατιστική. *Πρακτικά, Τέταρτο Παγκύπριο Συνέδριο Μαθηματικής Παιδείας και Συμπόσιο Αστροναυτικής και Διαστήματος*, (σ.σ. 327-335). Λάρνακα.

Αναστασιάδου, Σ. & Παπαδημητρίου, Γ. (2001β). Εγκυρότητα της Ελληνικής Προσαρμογής του Ερωτηματολογίου Μέτρησης Διαθέσεων προς τη Στατιστική. *Πρακτικά, Τέταρτο Παγκύπριο Συνέδριο Μαθηματικής Παιδείας και Συμπόσιο Αστροναυτικής και Διαστήματος*, (σ.σ. 359-367). Λάρνακα.

Αναστασιάδου, Σ. (2000). *Προσδιορισμός των Διαθέσεων των Φοιτητών προς τη Στατιστική με Μεθόδους της Πολυδιάστατης Στατιστικής Ανάλυσης*. Διδακτορική Διατριβή που υποβλήθηκε στο Τμήμα Εφαρμοσμένης Πληροφορικής του Πανεπιστημίου Μακεδονίας.

Βαζιργιάννης, Μ. & Χαλκίδη, Μ. (2003). *Εξόρυξη Γνώσης από Βάσεις Δεδομένων*. Αθήνα: Τυπωθήτω-ΓΙΩΡΓΟΣ ΔΑΡΔΑΝΟΣ.

Βασιλείου, Π. Χ. (1985). *Στοχαστικές Μέθοδοι στις Επιχειρησιακές Έρευνες*. Θεσσαλονίκη: Εκδόσεις Χριστοδουλίδη.

Γιαλαμάς, Β. & Κασιμάτη, Α. (2004). Εφαρμογή των Μεθόδων Ανάλυσης Δεδομένων στις Στάσεις των Εκπαιδευτικών Όσον Αφορά το Σύστημα Αξιολόγησής τους. *Τετράδια Ανάλυσης Δεδομένων*, **4**/04, 90-102.

Γναρδέλλης, Χ. & Κουλιεράκης, Γ. (2002). Τυπολογία Ομάδων Υψηλού Κινδύνου Μόλυνσης από τον HIV στις Ελληνικές Φυλακές Ανδρών. Ο Παράγοντας της Χρήσης Ενδοφλέβιων Ναρκωτικών. *Τετράδια Ανάλυσης Δεδομένων*, **1**/02, 52-64.

Δαφέρμος, Β. (2005). *Κοινωνική Στατιστική με το SPSS*. Θεσσαλονίκη: Εκδόσεις Ζήτη.

Δερμάνης, Α. (1986). *Συνορθώσεις Παρατηρήσεων και Θεωρία Εκτίμησης, Τόμος 1*. Θεσσαλονίκη: Εκδόσεις Ζήτη.

Δερμάνης, Α. (1998). *Γραμμική Άλγεβρα και Θεωρία Πινάκων*. Θεσσαλονίκη: Εκδόσεις Ζήτη.

Δημητρόπουλος, Ε. (1994). *Εισαγωγή στη Μεθοδολογία της Επιστημονικής Έρευνας*. Περιστέρι: Εκδόσεις Έλλην.

Δρόσος, Γ. (2005). *Στατιστική Ανάλυση Δεδομένων Γλωσσικών Πληροφοριών*. Διδακτορική Διατριβή που υποβλήθηκε στο Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών στις Επιστήμες της Γλώσσας και της Επικοινωνίας του Αριστοτέλειου Πανεπιστήμιου Θεσσαλονίκης.

Ζαχαροπούλου, Χ. (1994). *Παλινδρόμηση - Συσχέτιση: Θεωρία και Πράξη*. Θεσσαλονίκη: Προσωπική Έκδοση.

Κάλλας, Γ. (2002). *Ζητήματα Σχεδιασμού Εμπειρικών Ερευνών: Αξιοποίηση Μεθόδων της Πληροφορικής Τεχνολογίας*. Αθήνα: Εθνικό Κέντρο Κοινωνικών Ερευνών, Εκδόσεις Νεφέλη.

Καραγεώργος, Δ. (2002). *Μεθοδολογία Έρευνας στις Επιστήμες της Αγωγής: Μια Διδακτική Προσέγγιση*. Αθήνα: Εκδόσεις Σαββάλα.

Καράκος, Α. (2003). ΠΡΑΞΙΤΕΛΗΣ: Πρόγραμμα Αξιοποίησης Τεχνικών Ανάλυσης Δεδομένων. *Τετράδια Ανάλυσης Δεδομένων*, **3**/03, 135-152.

Καράκος, Α., Μαβίδης, Α. & Παπαβασιλείου, Α. (1996). Εφαρμογή Μεθόδων Ανάλυσης Δεδομένων σε Θέματα της Επιστήμης Φυσικής Αγωγής. *Πρακτικά, 9ο Πανελλήνιο Συνέδριο Στατιστικής*, (σ.σ. 148-155). Ελληνικό Στατιστικό Ινστιτούτο.

Καρακώστας, Κ. (1993). *Παλινδρόμηση και Ανάλυση Διακύμανσης*. Εκδόσεις Πανεπιστημίου Ιωαννίνων.

Καραπιστόλης, Δ. (1996). *Δημιουργία Λογισμικού για την Κατάρτιση Φερέγγυου Χαρτοφυλακίου με Μεθόδους της Ανάλυσης Δεδομένων*. Διδακτορική Διατριβή που υποβλήθηκε στο Τμήμα Εφαρμοσμένης Πληροφορικής του Πανεπιστημίου Μακεδονίας.

Καραπιστόλης, Δ. (1999). *Ανάλυση Δεδομένων και Έρευνα Αγοράς*. Θεσσαλονίκη: Εκδόσεις ΑΝΙΚΟΥΛΑ.

Καραπιστόλης, Δ. (2001). *Ανάλυση Δεδομένων και Έρευνα Αγοράς*. Θεσσαλονίκη: Εκδόσεις ΑΝΙΚΟΥΛΑ.

Καραπιστόλης, Δ. (2002). Το Λογισμικό MAD. *Τετράδια Ανάλυσης Δεδομένων*, **2**/02, 133-147.

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

Κάτος, Α. (1986). *Στατιστική*. Θεσσαλονίκη: Παρατηρητής.

Κιοσέογλου, Γ. & Δικαίου, Μ. (1993). Εφαρμογή της Παραγοντικής Ανάλυσης των Αντιστοιχιών και των Λογαριθμικών Γραμμικών Μοντέλων στην Ανάλυση των Προβλημάτων των Εφήβων και των Τρόπων Παροχής Βοήθειας από Κατηγορίες Ατόμων Στήριξης. *Πρακτικά, 6^{ο} Πανελλήνιο Συνέδριο Στατιστικής*, (σ.σ. 179-187). Ελληνικό Στατιστικό Ινστιτούτο.

Κιοσέογλου, Γ. (1997). Διερεύνηση των Προβλημάτων των Εφήβων και των Στρατηγικών Αντιμετώπισής τους Μέσω της Παραγοντική Ανάλυσης των Αντιστοιχιών και της Ιεραρχικής Ανάλυσης Συστάδων. *Πρακτικά, 10ο Πανελλήνιο Συνέδριο Στατιστικής*, (σ.σ. 203-219). Ελληνικό Στατιστικό Ινστιτούτο.

Κιοσέογλου, Γ. (2002). Η Ανάλυση Δεδομένων στην Ψυχολογική Έρευνα. *Τετράδια Ανάλυσης Δεδομένων*, **2**/02, 5-14.

Κιοσέογλου, Γ. (2003). Προσδιορισμός του Αριθμού των Παραγόντων που Χρήζουν Ερμηνείας στην Παραγοντική Ανάλυση των Αντιστοιχιών μέσω Παλινδρομήσεων στο Διάγραμμα των Ιδιοτιμών. *Τετράδια Ανάλυσης Δεδομένων*, **3**/03, 5-14.

Κίτσος, Χ. (1994). *Στατιστική Ανάλυση Πειραματικών Σχεδιασμών*. Αθήνα: Εκδόσεις Νέων τεχνολογιών.

Κολυβά-Μαχαίρα, Φ. & Μπόρα-Σέντα, Ε. (1996). *Στατιστική: Θεωρία Εφαρμογές*. Θεσσαλονίκη: Εκδόσεις Ζήτη.

Κουνιάς, Σ. & Καλπαζίδου, Σ. (1985). *Πιθανότητες ΙΙ: Θεωρία και Ασκήσεις*. Θεσσαλονίκη: Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Έκδοση: Υπηρεσία Δημοσιευμάτων.

Κουνιάς, Σ. & Μωυσιάδης, Χ. (1985). *Πιθανότητες Ι: Θεωρία και Ασκήσεις*. Θεσσαλονίκη: Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Έκδοση: Υπηρεσία Δημοσιευμάτων.

Κουνιάς, Σ., Κολυβά-Μαχαίρα, Φ., Μπαγιάτης, Κ. & Μπόρα-Σέντα, Ε. (1985). *Εισαγωγή στη Στατιστική*. Θεσσαλονίκη.

Κουτσουπιάς, Ν. & Παπαδημητρίου, Γ. (1999). Βοήθεια Ερμηνείας Αποτελεσμάτων των Μεθόδων Ανάλυσης Δεδομένων. *Πρακτικά, 12ο Πανελλήνιο Συνέδριο Στατιστικής*, (σ.σ. 318-326). Ελληνικό Στατιστικό Ινστιτούτο.

Κουτσουπιάς, Ν. (1999α). *Στατιστική Ανάλυση της Ασφαλιστικής Αγοράς Β. Ελλάδας με Χρήση Σχεσιακών Βάσεων Δεδομένων (RDBMS) και Δομημένης Γλώσσας Ερωτοαποκρίσεων (SQL)*. Διδακτορική Διατριβή που υποβλήθηκε στο Τμήμα Εφαρμοσμένης Πληροφορικής του Πανεπιστημίου Μακεδονίας.

Κουτσουπιάς, Ν. (1999β). *Λογισμικό Στατιστικής Ανάλυσης Δεδομένων*. Θεσσαλονίκη: Παρατηρητής.

Κουτσουπιάς, Ν. (2002). S-PRO: Εφαρμογή Υλοποίησης Μεθόδων της Ανάλυσης Δεδομένων. *Τετράδια Ανάλυσης Δεδομένων*, **1**/02, 118-128.

Κουτσουπιάς, Ν. (2005). *Εφαρμογές Ανάλυσης Δεδομένων*. Θεσσαλονίκη: Εκδοτικός Οίκος Αντ. Σταμούλη.

Κυριαζή, Ν. (1998). *Η Κοινωνιολογική Έρευνα: Κριτική Επισκόπηση των Μεθόδων και Τεχνικών*. Αθήνα: Ελληνικές Επιστημονικές Εκδόσεις.

Κωνσταντινίδης, Γ. (2002). «Scripta manent»: Προεκλογικά Κείμενα και Ιδεολογίες στη Μεταπολιτευτική Ελλάδα. *Τετράδια Ανάλυσης Δεδομένων*, **2**/02, 30-40.

Λελάκης, Γ. (1987). *Εκπαιδευτική Έρευνα*. Αθήνα: Οργανισμός Εκδόσεων Διδακτικών Βιβλίων.

Λούκας, Δ. & Παπαδημητρίου, Γ. (2005). Εισαγωγή της Παραγοντικής Ανάλυσης των Αντιστοιχιών στην Επίλυση Πολυκριτηρίων Προβλημάτων Απόφασης. *Τετράδια Ανάλυσης Δεδομένων*, **6**/05, 62-75.

Λούκας, Δ. (2004). *Πολυκριτήρια Υποστήριξη Αποφάσεων με τη Βοήθεια της Παραγοντικής Ανάλυσης των Αντιστοιχιών*. Διδακτορική Διατριβή που υποβλήθηκε στο Τμήμα Εφαρμοσμένης Πληροφορικής του Πανεπιστημίου Μακεδονίας.

Μάλλιαρη, Α. (2004). Ανάκτηση Πληροφοριών από τον Αυτοματοποιημένο Κατάλογο Ακαδημαϊκής Βιβλιοθήκης. *Τετράδια Ανάλυσης Δεδομένων*, **4**/04, 123-135.

Μάλλιαρη, Α. (2005). *Ανάκτηση Πληροφοριών από Αυτοματοποιημένα Συστήματα Βιβλιοθηκών και Στατιστική Ανάλυση των Δεδομένων*. Διδακτορική Διατριβή που υποβλήθηκε στο Τμήμα Εφαρμοσμένης Πληροφορικής του Πανεπιστημίου Μακεδονίας.

Μάρκος, Α. & Παπαδημητρίου, Γ. (2003). Οι Χαρακτηριστικές Ρίζες των Πινάκων και το Παρεξηγημένο Ποσοστό Ερμηνείας των Παραγοντικών Αξόνων στην Παραγοντική Ανάλυση των Αντιστοιχιών. *Πρακτικά, 16^{ο} Πανελλήνιο Συνέδριο Στατιστικής*, (σ.σ. 417-425). Ελληνικό Στατιστικό Ινστιτούτο.

Μάρκος, Α. (2006). *Βοήθεια στην Ερμηνεία των Αποτελεσμάτων της Παραγοντικής Ανάλυσης των Αντιστοιχιών & Αλγόριθμοι Κατασκευής και Ανάλυσης Ειδικών Πινάκων Εισόδου: Η Περίπτωση του Λογισμικού CHIC Analysis*. Διδακτορική Διατριβή που υποβλήθηκε στο Τμήμα Εφαρμοσμένης Πληροφορικής του Πανεπιστημίου Μακεδονίας.

Μάρκος, Α., Μενεξές, Γ. & Γ. Παπαδημητρίου (2005). Προσέγγιση Μεθόδου Επιλογής Υποπίνακα Συμπτώσεων από το Γενικευμένο Πίνακα Burt. *Τετράδια Ανάλυσης Δεδομένων*, (έγινε δεκτή προς δημοσίευση).

Μάρκος, Α., Μενεξές, Γ. & Παπαδημητρίου, Γ. (2005). Αλγόριθμος Επιλογής Υποπίνακα με την Πλησιέστερη Απεικόνιση μέσω της AFC στο Γενικευμένο Πίνακα Συμπτώσεων. *Πρακτικά, 18ο Πανελλήνιο Συνέδριο Στατιστικής*, (σ.σ. 247-256). Ελληνικό Στατιστικό Ινστιτούτο.

Μασούρα, Ε. & Παπαδημητρίου, Γ. (2002). Δημόσιες Δαπάνες για την Απασχόληση και Σχετική Θέση των 15 Χωρών-Μελών της Ε.Ε. με τις Μεθόδους AFC και CAH. *Τετράδια Ανάλυσης Δεδομένων*, **2**/02, 41-51.

Μασούρα, Ε. (2005). *Διαχρονική Εξέλιξη (1974-2004) Πολιτικών των Εργασιακών Σχέσεων και της Κοινωνικής Ασφάλισης στην Ελλάδα με Μεθόδους της Ανάλυσης Δεδομένων*. Διδακτορική Διατριβή που υποβλήθηκε στο Τμήμα Εφαρμοσμένης Πληροφορικής του Πανεπιστημίου Μακεδονίας.

Μαυρομάτης, Γ. (1999). Οργάνωση Πληροφοριών στην Καλαθοσφαίριση Μέσω της Ανάλυσης των Αντιστοιχιών. *Πρακτικά, 12^{ο} Πανελλήνιο Συνέδριο Στατιστικής*, (σ.σ. 381-390). Ελληνικό Στατιστικό Ινστιτούτο.

Μαυρομάτης, Γ. (1999). *Στατιστικά Μοντέλα και Μέθοδοι Ανάλυσης Δεδομένων*. Θεσσαλονίκη: University Studio Press.

Μεϊμάρης, Μ. (2002). Εξισώσεις Μεταφοράς και Μεταφορά Εξισώσεων. *Τετράδια Ανάλυσης Δεδομένων*, **1**, 12-19.

Μεϊμάρης, Μ. (2005). Το 3^{ο} Πανελλήνιο Συνέδριο Ανάλυσης Δεδομένων με Διεθνή Συμμετοχή 15-18/9/2005 στο Porto Carras της Χαλκιδικής. *Τετράδια Ανάλυσης Δεδομένων*, **6**/05, 7-9.

Μενεξές, Γ. & Οικονόμου, Α. (2002). Σφάλματα και Παρανοήσεις στους Στατιστικούς Ελέγχους Υποθέσεων: Υπέρβαση μέσω της Ανάλυσης Δεδομένων. *Τετράδια Ανάλυσης Δεδομένων*, **2**, 52-64.

Μενεξές, Γ. (2001). Η Παραγοντική Ανάλυση των Αντιστοιχιών Μέσω του SPSS. *Πρακτικά, 14^{ο} Πανελλήνιο Συνέδριο Στατιστικής, Α' Τόμος*, (σ.σ. 357-364). Ελληνικό Στατιστικό Ινστιτούτο.

Μενεξές, Γ. (2002). Ανάλυση Ισχύος των Στατιστικών Ελέγχων: Μία Πρώτη Προσέγγιση. *Πρακτικά, 15^{ο} Πανελλήνιο Συνέδριο Στατιστικής, Β' Τόμος*, (σ.σ. 481-490). Ελληνικό Στατιστικό Ινστιτούτο.

Μοσχίδης, Ο. & Παπαδημητρίου, Γ. (2002). Κλίμακες Αξιολόγησης (Υποκειμενικότητα – Μετασχηματισμός των Αρχικών Βαθμών σε Ποσοστά). *Τετράδια Ανάλυσης Δεδομένων*, **1**/02, 42-51.

Μοσχίδης, Ο. (2003α). *Συμβολή στη Μελέτη Κλιμάκων Αξιολόγησης με Μεθόδους της Πολυδιάστατης Ανάλυσης Δεδομένων*. Διδακτορική Διατριβή που υποβλήθηκε στο Τμήμα Εφαρμοσμένης Πληροφορικής του Πανεπιστημίου Μακεδονίας.

Μοσχίδης, Ο. (2003β). Πρόταση Κωδικοποίησης Κλιμάκων Αξιολόγησης για Επεξεργασία με την Παραγοντική Ανάλυση Πολλαπλών Αντιστοιχιών. *Πρακτικά, 16^{ο} Πανελλήνιο Συνέδριο Στατιστικής*, (σ.σ. 337-346). Ελληνικό Στατιστικό Ινστιτούτο.

Μοσχίδης, Ο., Παπαδημητρίου, Γ. & Χατζηπαντελής, Θ. (2005). Πρόταση «Εξισορρόπησης» Κατηγορικών Μεταβλητών στην Παραγοντική Ανάλυση Πολλαπλών Αντιστοιχιών. *Τετράδια Ανάλυσης Δεδομένων*, **5**/05, 35-41.

Μπαγιάτης, Β. & Παπαδημητρίου, Γ. (2002). Οι Ευρωπαϊκές Χώρες Όπως Αυτές Φαίνονται Μέσα από τη Διαχρονική τους Εξέλιξη στους Δείκτες της Πραγματικής Οικονομίας (1981-1998). *Τετράδια Ανάλυσης Δεδομένων*, **2**/02, 124-132.

Μπαγιάτης, Β. (2004). *Συνέπειες για τις Ευρωπαϊκές Χώρες από την Προσπάθειά τους για Σύγκλιση και Παραμονή στην ΟΝΕ*. Διδακτορική Διατριβή που υποβλήθηκε στο Τμήμα Εφαρμοσμένης Πληροφορικής του Πανεπιστημίου Μακεδονίας.

Μπαγιάτης, Κ. (1997). *Μεθοδολογία Έρευνας στη Φυσική Αγωγή*. Θεσσαλονίκη: Εκδόσεις Χριστοδουλίδη.

Μπεχράκης, Θ. (1999). *Πολυδιάστατη Ανάλυση Δεδομένων: Μέθοδοι και Εφαρμογές*, Αθήνα: Εκδόσεις ΝΕΑ ΣΥΝΟΡΑ-Α. Α. ΛΙΒΑΝΗΣ.

Μπεχράκης, Θ. (2003). Στατιστική Ανάλυση Κειμένων: Χριστός, Βούδας, Μωάμεθ. *Τετράδια Ανάλυσης Δεδομένων*, **3**/03, 15-26.

Μπόρα-Σέντα, Ε. & Μωϋσιάδης, Χ. (1992). *Εφαρμοσμένη Στατιστική*. Θεσσαλονίκη: Εκδόσεις Ζήτη.

Παπαδημητρίου, Γ. & Αθανασίου, Γ. (2003). Η Τηλεοπτική Ταυτότητα των Ελλήνων Τηλεθεατών μέσα από το Φαινόμενο Big Brother. (Η Συμπληρωματικότητα της AFC με την CAH). *Τετράδια Ανάλυσης Δεδομένων*, **3**/03, 41-52.

Παπαδημητρίου, Γ. & Μάρκος, Α. (2004). Θέσεις Πολιτών για την ΕΧΡΟ 2008 στη Θεσσαλονίκη. *Τετράδια Ανάλυσης Δεδομένων*, **4**/04, 150-162.

Παπαδημητρίου, Γ. & Φλώρου, Γ. (1992). Στατιστική Επεξεργασία του Τρισδιάστατου Πίνακα Δεδομένων των Πρωτοετών Φοιτητών του Πανεπιστημίου Μακεδονίας με Μεθόδους της Ανάλυσης Δεδομένων. *Πρακτικά, 5^{ο} Πανελλήνιο Συνέδριο Στατιστικής*, (σ.σ.144-153). Ελληνικό Στατιστικό Ινστιτούτο.

Παπαδημητρίου, Γ. & Φλώρου, Γ. (1995). Ανασκόπηση των Αποστάσεων για Ταξινόμηση Μεταβλητών. *Πρακτικά, 8^{ο} Πανελλήνιο Συνέδριο Στατιστικής*, (σ.σ. 198-206). Ελληνικό Στατιστικό Ινστιτούτο.

Παπαδημητρίου, Γ. & Φλώρου, Γ. (1999). Νέος Δείκτης Ερμηνείας Αποτελεσμάτων στις Μεθόδους Ανάλυσης Δεδομένων. *Πρακτικά, 12^{ο} Πανελλήνιο Συνέδριο Στατιστικής*, (σ.σ. 434-443). Ελληνικό Στατιστικό Ινστιτούτο.

Παπαδημητρίου, Γ. & Φλώρου, Γ. (1999). Νέος Δείκτης Ερμηνείας Αποτελεσμάτων στις Μεθόδους της Ανάλυσης Δεδομένων. *Πρακτικά, 12^{ο} Πανελλήνιο Συνέδριο Στατιστικής*, (σ.σ. 434-443). Ελληνικό Στατιστικό Ινστιτούτο.

Παπαδημητρίου, Γ. (1990). *Μέθοδοι Επεξεργασίας Ερωτηματολογίων*. Θεσσαλονίκη: Παρατηρητής.

Παπαδημητρίου, Γ. (1991). Στατιστική Επεξεργασία Τριδιάστατων Πινάκων με την Παραγοντική Ανάλυση των Αντιστοιχιών. *Πρακτικά, 4^{ο} Πανελλήνιο Συνέδριο Στατιστικής*, (σ.σ.160-169). Ελληνικό Στατιστικό Ινστιτούτο.

Παπαδημητρίου, Γ. (1994). *Μέθοδοι Ανάλυσης Δεδομένων: Πανεπιστημιακές Παραδόσεις*. Θεσσαλονίκη: Έκδοση Πανεπιστήμιου Μακεδονίας Οικονομικών και Κοινωνικών Επιστημών

Παπαδημητρίου, Γ. (2001). *Περιγραφική Στατιστική*. Θεσσαλονίκη: Παρατηρητής.

Παπαδημητρίου, Γ. (2002). Η Ανάλυση Δεδομένων στην Ελλάδα. *Τετράδια Ανάλυσης Δεδομένων*, **1**/02, 5-11.

Παπαδημητρίου, Γ. (2004). *Πολυμεταβλητή Στατιστική Ανάλυση: Πανεπιστημιακές Παραδόσεις*. Θεσσαλονίκη: Έκδοση Πανεπιστήμιου Μακεδονίας Οικονομικών και Κοινωνικών Επιστημών.

Παπαδημητρίου, Γ. (2006). *Η Ανάλυση Δεδομένων. Παραγοντική Ανάλυση των Αντιστοιχιών, Ιεραρχική Ταξινόμηση και άλλες Μέθοδοι*. Αθήνα: Εκδόσεις τυπωθήτω, Γ. Δαρδανός.

Περσίδης, Δ. (1997). *Εφαρμοσμένη Στατιστική στην Τεχνολογία Τροφίμων*. Θεσσαλονίκη: Εκδοτική ΟΜΗΡΟΣ.

Πρίπορας, Κ. Β. & Μενεξές, Γ. (2005). Μέτρηση Εικόνας Καταστήματος με Μεθόδους της Ανάλυσης Δεδομένων: Μια Εμπειρική Έρευνα στην Περιοχή της Θεσσαλονίκης. *Τετράδια Ανάλυσης Δεδομένων*, **5**, 42-61.

Σιάρδος, Γ. (1999). *Μέθοδοι Πολυμεταβλητής Στατιστικής Ανάλυσης: Με την Επίλυση Ασκήσεων Μέσω του Στατιστικού Προγράμματος SPSS, Μέρος Πρώτο: Διερεύνηση Σχέσεων Μεταξύ Μεταβλητών*. Θεσσαλονίκη: Εκδόσεις Ζήτη.

Σιάρδος, Γ. (2000). *Μέθοδοι Πολυμεταβλητής Στατιστικής Ανάλυσης: Με την Επίλυση Ασκήσεων Μέσω του Στατιστικού Προγράμματος SPSS, Μέρος Δεύτερο: Διερεύνηση Εξάρτησης Μεταξύ Μεταβλητών*. Θεσσαλονίκη: Εκδόσεις Ζήτη.

Ταγαράς, Γ. (2001). *Στατιστικός Έλεγχος Ποιότητας*. Θεσσαλονίκη: Εκδόσεις Ζήτη.

Τζήμος, Χ. & Παπαδημητρίου, Γ. (2004). Διερεύνηση των Διακλαδικών Σχέσεων της Ελληνικής Οικονομίας. *Πρακτικά, 17^{ο} Πανελλήνιο Συνέδριο Στατιστικής*, (σ.σ. 423-430). Ελληνικό Στατιστικό Ινστιτούτο.

Τσάντας, Ν. & Βασιλείου, Π.-Χ. (2000). *Εισαγωγή στην Επιχειρησιακή Έρευνα*. Θεσσαλονίκη: Εκδόσεις Ζήτη.

Τσάντας, Ν., Μωϋσιάδης, Χ., Μπαγιάτης, Ν. & Χατζηπαντελής, Θ. (1999). *Ανάλυση Δεδομένων με τη Βοήθεια Στατιστικών Πακέτων*. Θεσσαλονίκη: Εκδόσεις Ζήτη.

Τσοπάνογλου, Α. (2000). *Μεθοδολογία της Επιστημονικής Έρευνας και Εφαρμογές της στην Αξιολόγηση της Γλωσσικής Κατάρτισης*. Θεσσαλονίκη: Εκδόσεις Ζήτη.

Φαρμάκης, Ν. (1987). *Ειδικές Κατασκευές Βέλτιστων Πειραματικών Σχεδιασμών*. Διδακτορική Διατριβή που υποβλήθηκε στο Τμήμα Μαθηματικών του Α.Π.Θ.

Φαρμάκης, Ν. (1994). *Εισαγωγή στη Δειγματοληψία*. Θεσσαλονίκη: Εκδόσεις Κ. Χριστοδουλίδη.

Φαρμάκης, Ν. (2003). *Δημοσκοπήσεις και Δεοντολογία*. Θεσσαλονίκη: Εκδόσεις Χριστοδουλίδη.

Φίλιας, Β. (1996). *Εισαγωγή στη Μεθοδολογία και τις Τεχνικές των Κοινωνικών Ερευνών*. Αθήνα: Gutenberg.

Φλώρου, Γ. (1997). *Προσδιορισμός της Ιδανικότερης Μετρικής Απόστασης και Καλύτερου Τρόπου Ομαδοποίησης στις Διάφορες Μεθόδους της Αυτόματης Ταξινόμησης κατά Αύξουσα Ιεραρχία*. Διδακτορική Διατριβή που υποβλήθηκε στο Τμήμα Εφαρμοσμένης Πληροφορικής του Πανεπιστημίου Μακεδονίας.

Χάλκος, Γ. (2000). *Στατιστική Θεωρία Εφαρμογές & Χρήση Στατιστικών Προγραμμάτων σε Η/Υ*. Αθήνα: Τυπωθήτω - ΓΙΩΡΓΟΣ ΔΑΡΔΑΝΟΣ.

Χατζηνικολάου, Δ. (2002). *Στατιστική για Οικονομολόγους*. Ιωάννινα: Προσωπική Έκδοση.