ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ
Δ.Π.Μ.Σ. «ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ»

ΠΟΛΥΔΙΑΣΤΑΤΙΚΗ ΚΛΙΜΑΚΩΣΗ ΣΤΗΝ ΠΟΛΥΜΕΤΑΒΛΗΤΗ ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Νικόλαος Χρ. Πούλιος

Επιβλέπων:
Κοκολάκης Γεώργιος, Ομ. Καθηγητής

Αθήνα, 2016

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ
Δ.Π.Μ.Σ. «ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ»

ΠΟΛΥΔΙΑΣΤΑΤΙΚΗ ΚΛΙΜΑΚΩΣΗ ΣΤΗΝ ΠΟΛΥΜΕΤΑΒΛΗΤΗ ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Νικόλαος Χρ. Πούλιος

Επιβλέπων:
Κοκολάκης Γεώργιος, Ομ. Καθηγητής

Υποβολή διπλωματικής εργασίας για την εκπλήρωση των απαιτούμενων για το Μεταπτυχιακό Δίπλωμα Ειδίκευσης (MSc) του Διατμηματικού Μεταπτυχιακού Προγράμματος Σπουδών του Τομέα Μαθηματικών της Σχολής Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών του Εθνικού Μετσοβίου Πολυτεχνείου.

Αθήνα, 2016

Περίληψη

Στο πλαίσιο της παρούσας διπλωματικής εργασίας, μελετάται η τεχνική της Πολυδιαστατικής κλιμάκωσης (Multidimensional Scaling - MDS) για την ανάλυση δεδομένων που περιγράφουν ομοιότητες ή ανομοιότητες για ένα σύνολο αντικειμένων. Τέτοιου είδους δεδομένα μπορεί να είναι αξιολογήσεις για το πόσο όμοιοι είναι οι πολιτικοί υποψήφιοι ή οι δείκτες του εμπορίου για ένα σύνολο χωρών. Ο σκοπός αυτής της τεχνικής είναι να παρουσιάσει τα δεδομένα αυτά ως αποστάσεις μεταξύ των σημείων σε ένα γεωμετρικό χώρο. Ο σκοπός που ακολουθούμε αυτή τη διαδικασία είναι επειδή θέλουμε να έχουμε μια γραφική απεικόνιση της δομής που έχουν τα δεδομένα καθώς επίσης να πάρουμε μια καλύτερη διαίσθηση των δεδομένων από το να τα βλέπαμε αυτά μέσα από έναν πίνακα. Έτσι μετά από μια ιστορική αναδρομή για την MDS τεχνική στο πρώτο κεφάλαιο, περνάμε στο δεύτερο, το οποίο είναι ένα από τα πιο σημαντικά μιας και αναφερόμαστε στις πιο βασικές και στοιχειώδεις γνώσεις επάνω στη γραμμική άλγεβρα που είναι απαραίτητες για να πραγματοποιηθεί η παραπάνω τεχνική. Στο επόμενο κεφάλαιο αναφερόμαστε στο σύνολο των τεχνικών Αξιωμάτων της MDS τεχνικής. Στη συνέχεια στο τέταρτο κεφάλαιο αναπτύσουμε τη συνεισφορά της μεθόδου των Κανονικών Ελαχίστων Τετραγώνων για την περίπτωση των ποσοτικών δεδομένων. Τέλος, στα δύο τελευταία κεφάλαια αναπτύσονται το MDS μοντέλο, το οποίο αναφέρεται στο μη σταθμισμένο μοντέλο απόστασης και στο εναπομένον κεφάλαιο δίνεται έμφαση στους πιο σημαντικούς και γνωστούς αλγορίθμους που χρησιμοποιούνται στην MDS, όπως οι αλγόριθμοι Torgerson's, Kruskal, INDSCAL κλπ.

Abstract

In the following pages, I show the Multidimensional Scaling (MDS) technique for analysis of similarity or dissimilarity data on a set of objects. Such a data may be ratings of similarity on political candidates, or trade indices for a set of countries. The scope of MDS, is to represent such data as distances among points in a geometric space. We are working to this direction, because we want a graphical display of the structure of the data and another good reason is, to give us a more understandable perception than to try to focus an array with data. Thus, after some historic information's about the MDS, in the first section, the second section is one of the most important chapter in this dissertation. I mention the most basic and fundamental parts of matrix algebra, that I need through out with those techniques. In the next section, I am writing and set the technical Axioms of MDS. Furthermore, the next chapter I am giving the contribution of Ordinary Least Squares method, in case of quantitative data. In the following two sections, I mention the multidimensional scaling model, wich covers the unweighted (simple) distance model, while in the second one chapter, I am giving the most important and famous algorithm's which are used in MDS, like Torgerson's, Kruskal, INDSCAL etc.

Αφιερώνεται στους πολυαγαπημένους μου γονείς!
Χωρίς την στίριξή τους, δεν θα είχε ολοκληρωθεί η προσπάθεια αυτή!

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Λίστα συντομογραφιών και συμβόλων

Σύμβολα

$\mathbf{A}$	Πίνακας A
a_{ij}	Στοιχείο του πίνακα A
$\mathbf{A}^\top$ **ή** $\mathbf{A}'$	Ανάστροφος του πίνακα A
$\mathbf{A}^{-1}$	Αντίστροφος πίνακας
$\mathbf{I}$	Μοναδιαίος πίνακας
$\mathbb{1}$	Διάνυσμα στήλη με όλα τα στοιχεία του ίσα με τη μονάδα
$\mathbf{trA}$	Το ίχνος του πίνακα A
$A\vec{B} \equiv \vec{a}$	Διάνυσμα από το A στο B
x_{he}^{+}	Η τιμή της συντεταγμένης της επόμενης επανάληψης
$\mathbf{A}^{+}$	Ψευδοαντίστροφος πίνακας A
$e \circ x$	Πολλαπλασιασμός διανύσματος με αριθμό (βαθμωτό)
d_{ab}	Η απόσταση μεταξύ δύο σημείων a και b
$\mathcal{A}$	Σύνολο A
$C(a, b)$	Το σύνολο των συνεχών συναρτήσεων πάνω στο $[a, b]$
$\hat{b}$	Συνήθης εκτιμητής ελαχίστων τετραγώνων

Συντομογραφίες

MDS	Multidimensional Scaling
ALSCAL	Alternating Least Squares Scaling
SMACOF	Scaling by MAjoring a COmplicated Function
PD	Positive Definite
PSD	Positive Semidefinite
PCA	Principal Component Analysis
RDA	Redundancy Analysis
CMDS	Classical Multidimensional Scaling
CMDU	Classical Multidimensional Scaling Unfolding
RMDU	Rectangular Multidimensional Unfolding
RMDS	Replicated Multidimensional Scaling

GEM General Euclidean Model

INDSCAL Individual Differences Scaling

SVD Singular Value Decomposition

iff if and only if

ανν αν και μόνο αν

τ.ω. τέτοιο ώστε

ε.ω. έτσι ώστε

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια του Διατμηματικού Προγράμματος Μεταπτυχιακών Σπουδών «*Εφαρμοσμένες Μαθηματικές Επιστήμες*» *(MSc in Applied Mathematical Sciences)*, της Σχολής Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών, του Εθνικού Μετσόβιου Πολυτεχνείου.

Πριν την παρουσίαση της διπλωματικής εργασίας αισθάνομαι έντονα την υποχρέωση να ευχαριστήσω θερμά ορισμένους από τους ανθρώπους που γνώρισα, συνεργάστηκα, διδάχθηκα πάρα πολλά και έπαιξαν καθοριστικό ρόλο στην ολοκλήρωση αυτής καθώς και στον τρόπο μου σκέψης.

Πρώτο απ᾽ όλους θέλω να ευχαριστήσω θερμά τον επιβλέποντα της διπλωματικής εργασίας κ. Κοκολάκη Γεώργιο, Ομότιμο Καθηγητή για την πολύτιμη καθοδήγησή του, την εκτίμησή του που μου έδειξε, την ευκαιρία να ασχοληθώ με ένα πολύ ενδιαφέρον αντικείμενο καθώς και την τεράστια υπομονή και συμπαράστασή του καθ᾽ όλη τη διάρκεια της εκπόνησης. Επίσης, θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες προς όλα τα μέλη του διδακτικού ερευνητικού προσωπικού του ανωτέρω μεταπτυχιακού προγράμματος, που με δίδαξαν και με καθοδήγησαν όλα αυτά τα χρόνια.

Τέλος, δε θα μπορούσα να παραλείψω την οικογένειά μου και κυρίως τους γονείς μου Χρήστο και Ιωάννα για τη μεγάλη υπομονή, τη δύναμη και την ηθική συμπαράστασή τους στις επιλογές μου και θα συνεχίσω να τους προσφέρω επιτυχίες.

1 Εισαγωγή

Είναι σύνηθες φαινόμενο στη Στατιστική επιστήμη, να επεξεργαζόμαστε δεδομένα πολλών διαστάσεων. Το πρόβλημα έγκειται στο γεγονός, όχι μόνο ως προς την παρουσίαση των δεδομένων αλλά και η κατανόηση της μορφής που έχουν αυτά. Ένα σύνολο μαθηματικών τεχνικών που επιτρέπει τον ερευνητή να κατανοήσει και να προβάλει τη δομή των πολυδιάστατων δεδομένων, λέγεται πολυμεταβλητή κλίμακα (Multivariate Scaling). Πιο συγκεκριμένα, είναι η παρουσίαση των δεδομένων, συνήθως στον Ευκλείδειο χώρο, σε αρκετά χαμηλή διάσταση. Βασικά τα δεδομένα στον MDS είναι μετρήσεις των προσεγγίσεων (proximities) μεταξύ ζευγαριών αντικειμένων. Ένα μέτρο προσέγγισης (proximity measure) είναι ένας δείκτης ο οποίος εκφράζει το βαθμό στον οποίο, δύο αντικείμενα είναι όμοια. Τα πιο συνήθη παραδείγματα μέτρων προσεγγίσεων είναι ο συντελεστής συσχέτισης και η από κοινού πιθανότητα. Μεγάλο είναι το εύρος των επιστημονικών τομέων που εφαρμόζουν MDS.Ένας από αυτούς είναι οι ανθρωπιστικές επιστήμες και κυρίως η ψυχολογία, όπου έχει οδηγήσει στην ολοένα και περισσότερη κατανόηση των σύνθετων ψυχολογικών φαινομένων. Επιπλέον σχετικά με την παραπάνω τεχνική αποτελεί ιδαίτερο ενδιαφέρον, η εφαρμογή της στο χώρο των πολιτικών επιστημών. Για παράδειγμα, έστω ότι διεξάγεται μία έρευνα στα πλαίσια της ανάληψης της προεδρίας της Ευρωπαϊκής Ένωσης, από 28 υποψήφιους προέδρους μελών. Σε αυτό το σημείο, δημιουργούνται κάποια ενδιαφέροντα ερωτήματα. Υπάρχουν κάποιες ομοιότητες ή διαφορές μεταξύ αυτών των υποψήφιων προέδρων; Ή ένα άλλο ενδιαφέρον ερώτημα είναι, υπάρχουν κάποια σπουδαία χαρακτηριστικά γνωρίσματα που μπορούν να βρούν οι ψηφοφόροι στους παραπάνω υποψήφιους και να τους καθοδηγήσουν στην τελική τους απόφαση; Με αυτού του είδους τις ερωτήσεις, η μέθοδος της MDS βοηθά στο να πάρουμε τις απαντήσεις με το να τοποθετήσουν τις απόψεις των πολιτών για τους υποψηφίους, πάνω σε μία χωρική διαμόρφωση ή αλλιώς χάρτη. Από τη στιγμή που οι διάφορες απόψεις των ψηφοφόρων, έχουν εκφρασθεί ως διάφορα σημεία σε αυτόν το «χάρτη», αυτό που απομένει είναι να προσδιοριστεί η «κρυφή» δομή των δεδομένων αυτών. Έχοντας έναν πίνακα δεδομένων, μπορεί να εφαρμοστεί η μέθοδος της MDS, αρκεί τα δεδομένα του πίνακα να εκφράζουν ένα βαθμό της σχετικότητας μεταξύ των αντικειμένων (objects) και των γεγονότων (events), τα οποία παρουσιάζονται από τις γραμμές και τις στήλες του πίνακα δεδομένων αντίστοιχα. Συνήθως τέτοιου είδους δεδομένων, καλούνται ως «σχεσιακά» δεδομένα (relational data), ή αποστάσεις, προσεγγίσεις, ομοιότητες, πίνακας προτιμήσεων κλπ. Ο τύπος των πινάκων προτιμήσεων μπορεί να είναι είτε συμμετρικοί είτε ασύμμετροι, δύο ή πολλών κατευθύνσεων (one-way or multi-way matrices) και διάφοροι άλλοι τύποι πινάκων. Τέλος, διάφορες επιστήμες που χρησιμοποιούν την τεχνική της MDS είναι για παράδειγμα η ανθρωπολογία η οποία ενδιαφέρεται για τη σύγκριση των διαφορετικών γκρουπ κουλτούρας που βασίζονται στη θρησκεία, στη γλώσσα και τα πιστεύω, όπως επίσης και η οικονομολογία που ενδιαφερόμαστε για τις καταναλωτικές προτιμήσεις για μία ευρεία γκάμα προϊόντων και αρκετές άλλες.

1.1 Ιστορικό

Στα μισά του 20ου αιώνα, πρωτοεμφανίστηκε η μέθοδος του Multidimensional Scaling. Το 1952 ορίστηκε το πρώτο Multidimensional Scaling πρόβλημα από τον Torgerson και υπέδειξε την πρώτη μετρική λύση. Στην επόμενη δεκαετία, βρίσκονται από τη μια μεριά οι Shepards (1962) [2] και Kruskal (1964) [2], δουλέυοντας επάνω σε μη μετρικούς multidimensional scaling και από την άλλη η πολύ σημαντική δουλειά του Coombs (1964) [2] επάνω στη θεωρία δεδομένων. Στην τρίτη δεκαετία βρίσκονται οι Karol και Chang (1970) [2] εργαζόμενοι στις ατομικές διαφορές (individual differences) MDS. Τέλος, τις τελευταίες δεκαετίες, υπήρξε η ανάπτυξη της μεγίστης πιθανοφάνειας MDS, όπως ερμινεύτηκε από τους Ramsey (1982) [2] και Takane (1980a,1980b) [2].

1.2 Μετρικό Multidimensiona Scaling

1.2.1 Torgerson

Ο Torgerson, θεωρήθηκε ως ο «πατέρας» της μεθόδου MDS, μιας και ήταν ο πρώτος που εισήγαγε το γενικό θέμα της MDS. Επίσης, μία εκτενέστατη περιγραφή για τη μέθοδο αυτή, δόθηκε μέσω μιας δημοσίευσης από

τον ίδιο, η οποία ήταν και η πρώτη στο χώρο αυτό. Επιλέον, ο Klingberg (1941) [1] είχε εκτελέσει τη μέθοδο MDS, σε δεδομένα που αφορούσαν το βαθμό εχθρικότητας μεταξύ κρατών, η οποία εργασία ήταν βασισμένη στη δουλειά που είχε κάνει ο Richardson (1938) [2]. Δυστυχώς έχουμε μόνο μία περίληψη από την εργασία του τελευταίου, η οποία αναφέρεται στη μελέτη του MDS στην ψυχοφυσική και συγκεκριμένα στην αντίληψη χρωμάτων. Επίσης, οι Torgerson, Klingberg, χρησιμοποίησαν ένα θεώρημα, το οποίο αποδείχθηκε από τους Young και Housholder (1938) [2], όπου έδειξαν ότι από έναν πίνακα με error-free distances, είναι πιθανό να προσδιορισθούν:

1. Η διάσταση των αποστάσεων
2. Ένας Ευκλείδειος χώρος, που να έχει για σημεία αυτές τις αποστάσεις

Επιπροσθέτως, ο Torgerson (1952) [2], υπέδειξε την πρώτη συμμετρική διαδικασία, για τον προσδιορισμό ενός πολυδιάστατου χάρτη σημείων από πλήρως λανθασμένες ενδιάμεσες αποστάσεις.

1.2.2 Shepard

Μία από τις πιο σημαντικές δημοσιεύσεις επάνω στη μέθοδο του MDS ήταν αυτό που εισήχθει από τον Shepard (1962) [2], αυτό που αναφερόμαστε σήμερα ως μή-μετρικό MDS (nonmetric MDS). Η σπουδαιότητα αυτής της δημοσίευσης, έγκειται στο γεγονός, όχι μόνο στην ικανότητα να ληφθεί ένας πολυδιάστατος χάρτης από αριθμούς που εκφράζουν αποστάσεις που ορίζονται μόνο σε κανονικό επίπεδο, αλλά και στην τεράστια επίδραση που άσκησε στη μετέπειτα έρευνα. Ακόμη, ήταν ο επιστήμονας όπου με τη δουλειά του, πυροδότησε την έρευνα για το MDS και έτσι έκανε τη μέθοδο αυτή πολύ διάσημη σε πολλές άλλες ερευνητικές περιοχές, όπως η αρχιτεκτονική, γεωγραφία, πολιτικές επιστήμες, ψυχολογία και τη διοίκηση επιχειρήσεων. Τέλος, εισήγαγε την έννοια της μη-μετρικής στη γενικότερη έννοια της ανάλυση δεδομένων, που ξεπερνούσαν τα όρια του MDS.

1.2.3 Kruskal

Ύστερα από μία μεγάλη προσφορά στην MDS από τον Shepard, τη σκυτάλη πήρε ο Kruskal (1964) [2]. Ο Kruskal υπήρξε και αυτός ένας από τους μεγάλους ερευνητές και ακολούθησε μία διαφορετική προσέγγιση για το MDS, η οποία υιοθετήθηκε από πολλούς μεταγενέστερους ερευνητές. Η διαφορά μεταξύ του Kruskal και των άλλων επιστημόνων, ήταν σημαντική. Πιο συγκεκριμένα, όσο ο Shepard, στη διαδικασία που χρισημοποίησε, λάμβανε μία παρουσίαση των n σημείων σε $n-1$ διαστάσεις, χωρίς ωστόσο να περιλαμβάνει τη μείωση διαστάσεων, πρωτού γίνει η παρουσίαση αυτών των σημείων. Σε αντίθεση με τον Kruskal, ο οποίος στη διαδικασία του, χρησημοποίησε την καλύτερη δυνατή παρουσίαση των n αυτών σημείων σε r-διάστατο χώρο, όπου το r ήταν πολύ μικρότερο από αυτό του n και μάλιστα το προσδιόριζε ο χρήστης πριν γίνει η ανάλυση αντί μετά. Αξίζει να αναφερθεί, ότι ο Kruskal (όπως και ο Torgerson), η διαδικασία η οποία προτάθηκε, είχε την ανάλυση όπως αυτή της ανάλυσης παραγόντων (factor analysis). Επιπλέον ο Kruskal, επηρεασμένος από τη στατιστική επιστήμη, προσέγγισε το πρόβλημα ως πρόβλημα βελτιστοποίησης. Βέβαια, μία από τις πιο σπουδαίες συνεισφορές του Kruskal στην έρευνα ήταν η εξής. Έθεσε το πρόβλημα του Shepard, σε μια πιο σταθερή θεμελείωση της αριθμητικής ανάλυσης, με το να προτείνει ένα δείκτη ελαχιστοποίησης των κανονικών σφαλμάτων και μετά να ορίσει τη διαδικασία που θα ακολουθήσει για την ελαχιστοποίηση. Έτσι, επέλεξε να εφαρμόσει ελάχιστα τετράγωνα, μεταξύ μονοτονικού μετασχηματισμού των δεδομένων και του πολυδιάστατου χώρου (multidimensional space), γι᾽ αυτό όρισε και ένα μέτρο απόκλισης από το σημείο της τέλειας προσαρμογής. Στη συνέχεια, υποστήριξε ότι θα πρέπει να ορισθεί μία διαδικασία βελτιστοποίησης, με το να παρθούν οι μερικές παράγωγοι της συνάρτησης ως προς κάθε παράμετρο του μοντέλου και εν συνεχεία να χρησιμοποιηθούν αυτές οι παράγωγοι ως μια διαδικασία βελτιστοποίησης, γνωστή ως απότομης καθόδου διαδικασία, (steepest descent procedure). Επίσης, σπουδαία συνεισφορά του Kruskal, θεωρείται πως ήταν, όχι μόνο μία γενική προσέγγιση για την επίλυση προβλημάτων, μη μετρικών δεδομένων, αλλά και η ανάπτυξη του *μονοτονικού μετασχηματισμού ελαχίστων τετραγώνων*, μία μέθοδο για τη λήψη ενός συνόλου αριθμών

το οποίο είναι μονοτονικά σχεσιακό σε ένα δεύτερο σύνολο αριθμών, τα δεδομένα, και είναι προσαρμοσμένο υπό την έννοια των ελαχίστων τετραγώνων σε ένα τρίτο σύνολο αριθμών, δηλαδή των αποστάσεων. Επίσης, άλλη μία σημαντική προσφορά του Kruskal ήταν ότι είχε συμπεριλάβει στην ανάλυσή του, πίνακες που δεν ήταν ολοκληρωμένοι, δηλαδή περιείχαν ελλειπείς τιμές (missing values), όπου κατά το παρελθόν οι τεχνικές που εφαρμόζονταν από άλλους, απαιτούνταν να μην υπήρχε τέτοιο πρόβλημα. Επίσης, ήταν εκείνος ο οποίος, εισήγαγε τον πρώτο αλγόριθμο, οποίος είχε τη δυνατότητα να πραγματοποιεί την παρουσίαση των δεδομένων, σε μη-Ευκλείδειους χώρους, το οποίο φυσικά και ήταν ένα πολύ σημαντικό πλεονέκτημα. Τέλος, εξίσου σημαντική θεωρείται η συνεισφορά του για την εισαγωγή του πρώτου αλγορίθμου, όπου υπήρχε η δυνατότητα της δημιουργίας ενός χάρτη σημείων σε οποιοδήποτε Minkowski χώρο, χωρίς να απαιτείται ο προσδιορισμός των διαστάσεων εκ των προτέρων, πράγμα που αυτό δε μπορούσε να πραγματοποιηθεί σε προηγούμενη εργασία επάνω σε μη-Ευκλείδειους χώρους.

1.2.4 Θεωρία Δεδομένων και MDS

Χρειάστηκαν να περάσουν 14 χρόνια συνεχούς έρευνας επάνω στις κοινωνικές και ανθρωπιστικές επιστήμες, για να εκδοθεί ένα μνημειώδες βιβλίο, από τον Coombs (1964) [2] επάνω στη θεωρία δεδομένων. Επίσης, μία στοιχειώδη υπόθεση που έγραψε και στο βιβλίο του, ήταν ότι θα πρέπει να αντιλαμβανόμαστε τα δεδομένα ως μια σχέση μεταξύ σημείων και χώρου. Από την πλευρά του υποστήριξε, ότι υπάρχουν τεσσάρων κατηγοριών δεδομένων τα οποία είναι:

1. Δεδομένα με βάση την επιλογή προτίμησης (preferential choice data).

   Πιο συγκεκριμένα, στην κατηγορία αυτή ανήκουν τα δεδομένα εκείνα τα οποία τα άτομα, θέτουν ποιο και σε τι βαθμό από τα δύο αντικείμενα προτιμούν. Για παράδειγμα, ποιον από τους δύο Η/Υ, προτιμάτε περισσότερο και σε ποιο βαθμό;

2. Δεδομένα ενός χαρακτηριστικού γνωρίσματος (single-stimulus data).

   Στην κατηγορία αυτή, ανήκουν δεδομένα όπου αποτυπώνεται η αντίδραση του ατόμου, στο άκουσμα ενός μόνο χαρακτηριστικού γνωρίσματος του αντικειμένου. Για παράδειγμα, αν το άτομο αρέσκεται ή όχι σε μία συγκεκριμένη επωνυμία Η/Υ.

3. Δεδομένα σύγκρισης χαρακτηριστικών (stimulus comparison data).

   Η κατηγορία αυτή αναφέρεται στο άτομο το οποίο καλείται να κρίνει πιο από τα δύο χαρακτηριστικά έχει κάτι περισσότερο από το άλλο. Για παράδειγμα, ποιος από τους δύο επεξεργαστές είναι ταχύτερος;

4. Δεδομένα που έχουν ομοιότητες (similarities data).

   Στη περίπτωση αυτή, ζητείται από το άτομο να κρίνει το κατά πόσο δύο χαρακτηριστικά είναι όμοια. Για παράδειγμα, πόσο όμοιοι είναι οι δύο επεξεργαστές διαφορετικής επωνυμίας Η/Υ;

   Στην συνέχεια ο Coombs, προσπάθησε να προσδιορίσει πως θα μπορούσαν οι παραπάνω τέσσερις τύποι δεδομένων, να τους ενοποιήσει στη θεωρία δεδομένων. Όσο μάλιστα αφορά την πρώτη κατηγορία, σκέφτηκε ότι θα πρέπει τα χαρακτηριστικά γνωρίσματα και τα άτομα, να παρουσιάζονται ως σημεία και να βρίσκονται στον ίδιο χώρο. Έτσι ο χώρος αυτός, ονομάστηκε από τον ίδιο ως «από κοινού» χώρος, διότι ενσωματώνει στον ίδιο χώρο, τόσο τα άτομα (individuals), όσο και τα χαρακτηριστικά γνωρίσματα (stimuli). Τα σύνολα των σημείων κατανέμονται στο χώρο, έτσι ώστε ένα άτομο, να βρίσκεται κοντά στο σημείο του χαρακτηριστικού γνωρίσματος, εκφράζοντας την προτίμησή του για αυτό το χαρακτηριστικό, ενώ στην αντίθετη πλευρά, το σημείο του ατόμου, θα βρίσκεται μακρύτερα από το σημείο του χαρακτηριστικού γνωρίσματος, εκφράζοντας έτσι τη λιγότερη προτίμησή του. Έτσι ο Coombs, καταλήγει ότι οι αποστάσεις των σημείων του ατόμου, είναι αντιστρόφως ανάλογες με τα σημεία που δηλώνουν προτιμήσεις. Επιπλέον, ο Coombs, σκέφτηκε και έδωσε τον τίτλο του ιδανικού σημείου, εκείνο το οποίο το άτομο πιθανώς να βρίσκεται στην προτιμητέα πλευρά του χώρου των χαρακτηριστών γνωρισμάτων. Όσο αφορά τη δεύτερη κατηγορία δεδομένων, το άτομο θα εκφράζει την αρέσκειά του

ως προς ένα χαρακτηριστικό γνώρισμα, όταν το σημείο αυτό βρίσκεται κοντά στο ιδανικό σημείο ενώ αντιθέτως όσο απομακρίνεται από αυτό το ιδανικό σημείο, θα εκφράζει την απαρέσκειά του για αυτό το χαρακτηριστικό. Επιπροσθέτως, τόσο για την τρίτη κατηγορία δεδομένων, όσο και για την τέταρτη, ο Coombs υποστήριξε ότι τα χαρακτηριστικά γνωρίσματα θα αποτελούν τον κυρίαρχο ρόλο και γι᾽ αυτό θα πρέπει να παρουσιάζονται στο χώρο.

1.2.5 Κλίμακα μέτρησης και μορφή δεδομένων

Οι κλίμακες μέτρησης που συναντάμε στους κανόνες ανάλυσης των δεδομένων είναι η τακτική ανάλυση, η κλίμακα ισοδιαστημάτων και η αναλογική. Αυτές οι τρεις κλίμακες μετρήσεων εντάσσονται σε δύο κατηγορίες που ορίζουν δύο τύπων δεδομένων και δύο τύπους ανάλυσης. Από τη μια πλευρά έχουμε την πρώτη μέτρηση σε τακτική κλίμακα και αφορά ποιοτικά χαρακτηριστικά δεδομένα και άρα η ανάλυσή τους είναι μη μετρική. Αντιθέτως η δεύτερη μέτρηση που περιλαμβάνει την κλίμακα ισοδιαστημάτων και αναλογική, η οποία αναφέρεται σε ποσοτικά δεδομένα και επομένως η ανάλυσή τους είναι μετρική. Εκτός από τις κλίμακες μέτρησης των δεδομένων μας, αυτό που έχει επίσης μεγάλη σημασία είναι η μορφή που έχουν αυτά. Έτσι τα δεδομένα μπορεί να είναι τετραγωνικής ή ορθογώνιας μορφής, όπου η πρώτη μπορεί να είναι είτε συμμετρικής είτε ασύμμετρης μορφής. Λέγοντας συμμετρική χαρακτηρίζουμε εκείνη τη μήτρα των δεδομένων όπου ο αριθμός των παρατηρήσεων είναι ίδιος με τον αριθμό των μεταβλητών. Αντιθέτως ασυμμετρική χαρακτηρίζουμε εκείνη τη μήτρα για την οποία ο αριθμός των παρατηρήσεων είναι ίδιος με αυτόν των μεταβλητών αλλά διαφέρουν οι τιμές εκατέρωθεν της κύριας διαγωνίου. Τα δεδομένα ορθογώνιας μορφής είναι εκείνα που υποδηλώνουν την ανομοιότητα όλων των υποκειμένων μιας ομάδας προς όλα τα υποκείμενα μιας άλλης ομάδας, δηλαδή η μήτρα όπου τόσο οι σειρές όσο και οι στήλες δεδομένων παριστάνουν διαφορετικά θέματα (αντικείμενα). Επίσης, ένα άλλο σημαντικό χαρακτηριστικό των δεδομένων είναι ο αριθμός των κατευθύνσεων (number of ways). Με τον χαρακτηρισμό αυτό εννοούμε τον αριθμό των κατευθύνσεων για ένα σύνολο δεδομένων που αντιστοιχούν στον αριθμό πειραματικών καταστάσεων που χειρίζεται ο πειραματιστής. Για παράδειγμα έστω ότι ένας πειραματιστής ρωτάει ένα άτομο για να κρίνει την ομοιότητα που υπάρχει για όλα τα ζεύγη για ένα σύνολο n αυτοκινήτων. Στην περίπτωση αυτή τα δεδομένα θα είναι δύο κατευθύνσεων, οι οποίες είναι τα αυτοκίνητα και τα αυτοκίνητα και αυτό γιατί το κάθε αυτοκίνητο αντιστοιχίζεται με κάθε άλλο αυτοκίνητο. Δηλαδή, αυτό το πείραμα σχεδίασης (experimental design) θα περιέχει το καρτεσιανό γινόμενο $< I \times I >$ του συνόλου I των n_i αυτοκινήτων, όπου n_i το πλήθος των επιπέδων ανά πειραματική κατάσταση i. Τέλος, μια άλλη εξίσου σημαντική διάκριση που χρησιμοποιούμε στα δεδομένα είναι αυτή του αριθμού των τρόπων (number of modes). Η βοήθεια που μας παρέχεται είναι στο να μπορούμε να διακρίνουμε τον τύπο δεδομένων με βάση το καρτεσιανό του γινόμενο. Δηλαδή, με βάση το αμέσως παραπάνω παράδειγμα, τα δεδομένα μας θα χαρακτηριζόνταν ως one mode σε αντίθεση αν είχαμε για παράδειγμα το καρτεσινό γινόμενο $< I \times J >$, όπου στην περίπτωση αυτή θα μιλούσαμε για two mode. Σε κάθε πρίπτωση ο αριθμός των τρόπων (modes) δε μπορεί να ξεπερνά τον αριθμό των κατευθύνσεων (ways).

1.2.6 Ατομικές διαφορές MDS

Ένας από τους περιορισμούς που υπήρχε σε διάφορες διαδικασίες MDS, ήταν ότι ο ερευνητής μπορούσε να κάνει ανάλυση δεδομένων από έναν πίνακα. Οπότε αμέσως τέθηκε το ερώτημα, τι θα γινόταν στην περίπτωση που υπάρξουν αρκετοί πίνακες δεδομένων; Στο ερώτημα αυτό, οι δυνατότητες που υπήρχαν ήταν ή ο ερευνητής να πάρει το μέσο όρο από όλους τους πίνακες και να οδηγηθεί σε ένα μόνο πίνακα και να κάνει την ανάλυσή του κατά τα γνωστά ή να κάνει την ανάλυση σε κάθε έναν πίνακα ξεχωριστά. Φυσικό και επόμενο ήταν, να μην είναι εφικτό να πραγματοποιηθεί τίποτα από τα προηγούμενα. Έτσι για να ξεπεραστεί αυτό το εμπόδιο, προτάθηκε από τους ερευνητές η ανάπτυξη ατομικών διαφορών MDS, μια διαδικασία που θα επέτρεπε την ταυτόχρονη ανάλυση πινάκων δεδομένων, δίχως να χρειάζεται να γίνει ο μέσος όρος των διάφορων αυτών πινάκων δεδομένων.

1.2.7 Ενοποίηση

Σε όλο αυτό το διάστημα που γινόταν κοπιαστική και λεπτομερέστατη δουλειά επάνω στους αλγορίθμους, ήρθε μια νέα προσέγγιση από τους Takane, Young και de Leeuw. Έτσι ο Young, το 1972, ανέπτυξε ένα πολύ γενικό πλαίσιο αλγορίθμου, ο οποίος ήταν θεωρητικός εφαρμόσιμος τόσο στο σταθμισμένο Ευκλείδειο μοντέλο όσο και στο απλό και το όνομα αυτού ήταν POLYCON. Παρ΄ όλα αυτά, όταν δημοσιεύθηκε, δεν περιελάμβανε το σταθμισμένο μοντέλο αν και ο Young, αφιέρωσε πάρα πολύ μεγάλη προσπάθεια στο να το συμπεριλάβει. Ωστόσο, σύμφωνα με τη μέθοδο Monte Carlo, έδειξε ότι ο αλγόριθμος είχε κάποιο πρόβλημα, και πιο συγκεκριμένα δεν μπορούσε να βρει με μεγάλη ακρίβεια τα βάρη στο σταθμισμένο μοντέλο. Στη συνέχεια, ο Young, αποφάσισε να εφαρμόσει κάτι διαφορετικό. Αυτό που έκανε, ήταν να βελτιστοποιήσει μια συνάρτηση, η οποία βασίζονταν στην προσαρμογή των τετραγωνικών αποστάσεων επάνω στις ανομοιότητες (dissimilarities). Ο λόγος που οδηγήθηκε σε αυτή την κατεύθυνση ήταν πως υπέθεσε ότι μέσω αυτής της προσέγγισης, θα ήταν πιθανό ο υπολογισμός των βαρών, που αφορούσε το σταθμισμένο μοντέλο, μέσω πολλαπλής παλινδρόμισης αντί της αρνητικής κλίσης (negative gradients). Σε αυτό το σημείο ο αλγόριθμος ήταν αρκετά καλός ως προς την εκτίμηση των βαρών και εδώ εισέρχεται ο Takane ο οποίος παρήγαγε μία νέα μέθοδο για να έχει στην κατοχή του τις εκτιμήσεις των ελαχίστων τετραγόνων για τις συντεταγμένες των χαρακτηριστικών γνωρισμάτων. Έτσι, στηρίχτηκε στη λύση της κυβικής εξίσωσης αντί της χρησιμοποίησης των κλίσεων. Οπότε ο Takane μέσω του συνδυασμού της δικής του προσέγγισης, καθώς και με τη συνεισφορά των προσεγγίσεων των Young's και Kruskal's, de Leeuw, οδηγήθηκε στη δημοσίευση του ALSCAL αλγορίθμου, ο οποίος επέστρεφε πολύ καλά αποτελέσματα σε σχέση με αυτά που έδινε η μέθοδος του Monte Carlo.

2 Άλγεβρα για MDS

2.1 Στοιχειώδεις πράξεις με πίνακες

Για να εφαρμόσουμε MDS θα πρέπει να κάνουμε χρήση κάποιων στοιχειωδών γνώσεων από τη γραμμική άλγεβρα. Όταν λέμε πίνακα εννοούμε μία ορθογώνια διάταξη αριθμών, συμβόλων ή εκφράσεων διαταγμένων σε σειρές και στείλες. Έτσι ένας πίνακας δεδομένων, μπορεί να περιέχει μετρήσεις για ένα πλήθος n ατόμων για m διαφορετικά αντικείμενα. Συνήθως σε έναν πίνακα δεδομένων, οι γραμμές αντιστοιχούν στα άτομα ενώ οι στήλες στα αντικείμενα. Τον πίνακα τον συμβολίζουμε με κεφαλαίο έντονο γράμμα, και τα στοιχεία του με μικρό γράμμα μέσα σε παρένθεση, δηλαδή $\mathbf{A} = (a_{ij})$. Επίσης, ο αριθμός των γραμμών και στηλών, i και j, ορίζουν την τάξη του πίνακα, οπότε αν το $i = 0$ ή $j = 0$, τότε ο πίνακας καλείται και διάνυσμα. Επιπλέον, ένας πίνακας λέγεται συμμετρικός αν ισχύει $a_{ij} = a_{ji}$, για όλα τα i, j ή ισοδύναμα αν $\mathbf{A}^\top = \mathbf{A}$. Στους παρακάτω πίνακες παρουσιάζονται κάποιες βασικές ιδιότητες πρόσθεσης και πολλαπλασιασμού των πινάκων.

Πίνακας 1: Βασικές ιδιότητες πρόσθεσης και βαθμωτού πολλαπλασιασμού πινάκων.

$\mathbf{A} = \mathbf{B}$	$a_{ij} = b_{ij}$ για όλα τα i, j
$\mathbf{A} + \mathbf{B} = \mathbf{C}$	$c_{ij} = a_{ij} + b_{ij}$ για όλα τα i, j
$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$	αντιμεταθετική ιδιότητα
$(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$	προσεταιριστική ιδιότητα
$c\mathbf{A}$	έχει στοιχεία ca_{ij} για όλα τα i, j
$c\,(k\mathbf{A}) = (ck)\,\mathbf{A} = (kc)\,\mathbf{A} = k\,(c\mathbf{A})$	προσεταιριστική ιδιότητα
$c\,(\mathbf{A} + \mathbf{B}) = c\mathbf{A} + c\mathbf{B}$	επιμεριστική ιδιότητα για πίνακες
$(c + k)\,\mathbf{A} = c\mathbf{A} + k\mathbf{A}$	επιμεριστική ιδιότητα για βαθμωτό
$\mathbf{A} + \mathbf{0} = \mathbf{A}$	πρόσθεση μηδενικού πίνακα

Πίνακας 2: Βασικές ιδιότητες πολλαπλασιασμού πινάκων, ανάστροφος και αντίστροφος πίνακας.

$\mathbf{A_{n\times r}B_{r\times m}} = \mathbf{C_{n\times m}}$ ανν $c_{ij} = \sum_{k=1}^{r} a_{ik}b_{kj}$
$(\mathbf{AB})\,\mathbf{C} = \mathbf{A}\,(\mathbf{BC})$
$\mathbf{AA} = \mathbf{A}^2$
$(\mathbf{A} + \mathbf{B})\,(\mathbf{C} + \mathbf{D}) = \mathbf{A}\,(\mathbf{C} + \mathbf{D}) + \mathbf{B}\,(\mathbf{C} + \mathbf{D})$
$\left(\mathbf{A^T}\right)^T = \mathbf{A}$
$(\mathbf{AB})^T = \mathbf{B^T A^T}$
$(\mathbf{ABC})^T = \mathbf{C^T B^T A^T}$
$(\mathbf{A} + \mathbf{B})^T = \mathbf{A^T} + \mathbf{B^T}$
$\mathbf{IA} = \mathbf{A} = \mathbf{AI}$
$\mathbf{B} = \mathbf{A}^{-1}$ ανν $\mathbf{BA} = \mathbf{I} = \mathbf{AB}$
$(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
$\left(\mathbf{A^T}\right)^{-1} = (\mathbf{A}^{-1})^T$
$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$

Επίσης, αξίζει να αναφερθεί, ότι γνωρίζουμε από τη γραμμική άλγεβρα, ότι δίνοντας δύο διανύσματα, ίδιας τάξης το εσωτερικό τους γινόμενο δίνεται από την παρακάτω σχέση:

$$< x, y >= x_1y_1 + \cdots + x_ny_n \tag{2.1}$$

καθώς και αν ισχύει $x'y = 0$, τότε τα διανύσματα λέγονται ορθογώνια. Τέλος, άλλη μία αρκετά σπουδαία έννοια στους πίνακες είναι το ίχνος (trace), όπου παρακάτω δίνεται η σχέση καθώς και κάποιες ιδιότητές του. Έτσι, έστω ένας πίνακας $\mathbf{A}_{n\times n}$, τότε το ίχνος θα δίνεται από τη σχέση:

$$tr\mathbf{A} = \sum_{i=1}^{n} a_{ii} \tag{2.2}$$

δηλαδή το άθροισμα των στοιχείων της κύριας διαγωνίου.

Πίνακας 3: Βασικές ιδιότητες της συνάρτησης του ίχνους.

$tr\mathbf{A}$=$\sum_{i=1}^{n} a_{ii}$	Ορισμός του ίχνους
$tr\mathbf{A}$=$tr\mathbf{A}'$	Αμετάβλητο ως προς τον ανάστροφο του $\mathbf{A}$
$tr\mathbf{ABC}$=$tr\mathbf{CAB}$=$tr\mathbf{BCA}$	Αμετάβλητο ως προς τις κυκλικές μεταθέσεις
$tr\left(\mathbf{A}'\mathbf{B}\right)$=$tr\left(\mathbf{A}'\mathbf{B}\right)'$=$tr\mathbf{B}'\mathbf{A}$=$tr\mathbf{A}\mathbf{B}'$	Συνδυασμός των δύο προηγούμενων ιδιοτήτων
$tr\mathbf{a}\mathbf{b}'$=$\mathbf{a}'\mathbf{b}$	
$tr\left(\mathbf{A}+\mathbf{B}\right)$=$tr\mathbf{A}$+$tr\mathbf{B}$	Προσθετικός κανόνας

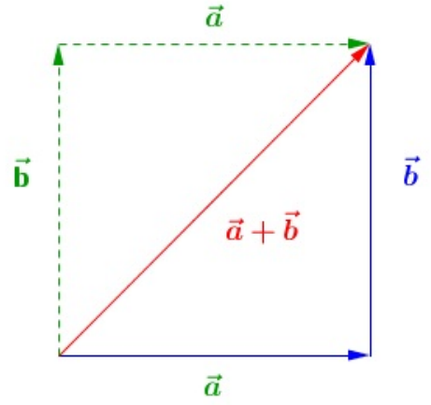
Πίνακας 4: Κανόνες παραγώγισης του ίχνους ενός πίνακα, ως προς έναν άγνωστο πίνακα $\mathbf{X}$.

$\partial tr(\mathbf{A})/\partial\mathbf{X} = 0$	
$\partial tr(\mathbf{AX}) = \mathbf{A}' = \partial tr\left[\left((\mathbf{AX})\right)'\right]/\partial\mathbf{X}$	
$\partial tr\left(\mathbf{X}'\mathbf{AX}\right)/\left(\mathbf{A}+\mathbf{A}'\right)X$	
$\partial tr\left(\mathbf{X}'\mathbf{AX}\right)/\partial\mathbf{X} = 2\mathbf{AX}$	αν ο $\mathbf{A}$ είναι συμμετρικός
$\partial tr\left(\mathbf{U}+\mathbf{V}\right)/\partial\mathbf{X} = \partial tr\left(\mathbf{U}\right)/\partial\mathbf{X} + \partial tr\left(\mathbf{V}\right)/\partial\mathbf{X}$	
$\partial tr\left(\mathbf{UVW}\right)/\partial\mathbf{W} = \partial tr\left(\mathbf{WUV}\right)/\partial\mathbf{X} = \partial tr\left(\mathbf{VUW}\right)/\partial\mathbf{X}$	Αμετάβλητο ως προς τις κυκλικές μεταθέσεις
$\partial tr\left(\mathbf{UV}\right)/\partial\mathbf{X} = \partial tr\left(\mathbf{U_c V}\right)/\partial\mathbf{X} + \partial tr\left(\mathbf{U V_c}\right)/\partial\mathbf{X}$	Κανόνας γινομένου: $\mathbf{U_c}$ και $\mathbf{V_c}$ λαμβάνεται ως σταθερός πίνακας στην παραγώγιση.

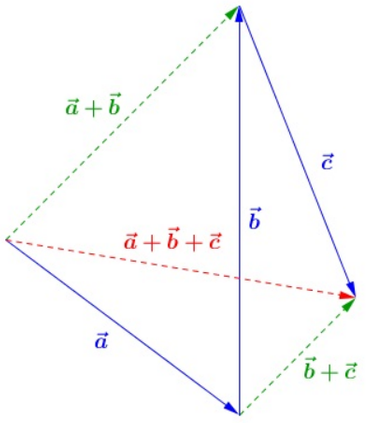
Στον παραπάνω Πίνακα 4, ισχύει ότι ο πίνακας $\mathbf{A}$ είναι σταθερός. Οι δε πίνακες $\mathbf{U}, \mathbf{V}, \mathbf{W}$ είναι συναρτήσεις του $\mathbf{X}$ (Schönemann).

2.2 Διανυσματικοί χώροι και Νόρμες

Ο r-διάστατος χώρος $\mathbf{R}^r$ αποτελεί μία γενίκευση της γραμμής ($\mathbf{R^1}$), του επιπέδου ($\mathbf{R^2}$) και του χώρου των τριών διαστάσεων $\mathbf{R^3}$, οι οποίες έννοιες αυτών, είναι γνωστές σε όλους τους μαθηματικούς και όχι μόνο. Ο χώρος στον οποίο ζούμε, αντιστοιχεί στον $\mathbf{R^3}$, όπου μπορούμε να βρούμε το επίπεδο ($\mathbf{R^2}$). Διανύσματα σε αυτούς τους χώρους, είναι το σύνολο όλων των πιθανών μετατοπίσεων τα οποία είναι προσανατολισμένα ευθύγραμμα τμήματα, χωρίς να παίζει σπουδαίο ρόλο η αρχή. Έτσι συνηθίζεται να γράφουμαι $\overrightarrow{AB} \equiv \vec{a}$ για το διάνυσμα και εννοούμε τη μετατόπιση από το σημείο Α στο σημείο Β. Επίσης, δύο διανύσματα είναι ίσα, αν έχουν ίσα μέτρα και ίδια κατεύθυνση (διεύθυνση και φορά) και θα γράφουμε $\overrightarrow{AB} = \overrightarrow{CD}$. Βέβαια, το μήκος του διανύσματος το συμβολίζουμε με $|\vec{a}|$ και είναι θετικός αριθμός. Τέλος, οι πράξεις της πρόσθεσης και του πολλαπλασιασμού των διανυσμάτων με έναν πραγματικό αριθμό είναι πολύ σημαντικές, διότι ορίζουν ένα διανυσματικό χώρο και ικανοποιούν τις παρακάτω 8 ιδιότητες, για οποιαδήποτε διανύσματα $\vec{a}, \vec{b}$ και για



(α′) Αντιμεταθετική



(β′) Προσεταιριστική

Σχήμα 1: Πρόσθεση διανυσμάτων

$\forall \alpha, \beta \in \mathbb{R}$ και μάλιστα όταν οι ιδιότητες αυτές χρησιμοποιούνται για να ορίσουν το διανυσματικό χώρο, τότε αναφέρονται και ως αξιώματα.

- Ιδιότητες πρόσθεσης διανυσμάτων:

$$\vec{a} + \vec{b} = \vec{b} + \vec{a} \tag{2.3}$$
$$\vec{a} + (\vec{b} + \vec{c}) = (\vec{a} + \vec{b}) + \vec{c} \tag{2.4}$$
$$\vec{a} + \vec{0} = \vec{a} = \vec{0} + \vec{a} \tag{2.5}$$
$$\vec{a} + (-\vec{a}) = 0 \tag{2.6}$$

- Ιδιότητες βαθμωτού πολλαπλασιασμού:

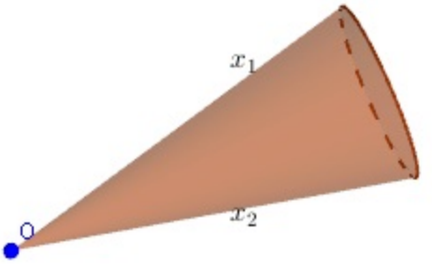
$$(\alpha + \beta)\vec{a} = \alpha\vec{a} + \beta\vec{a} \tag{2.7}$$
$$(\alpha\beta)\vec{a} = \alpha(\beta\vec{a}) \tag{2.8}$$
$$\alpha(\vec{a} + \vec{b}) = \alpha\vec{a} + \alpha\vec{b} \tag{2.9}$$
$$1 \cdot \vec{a} = \vec{a} \tag{2.10}$$

2.3 Κυρτά σύνολα

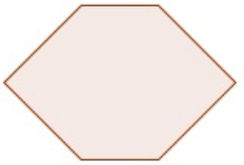
Η έννοια των κυρτών συνόλων είναι πολύ σημαντική τόσο στα μαθηματικά όσο και στη μέθοδο MDS. Ένα υποσύνολο του διανυσματικού χώρου, θα λέμε ότι είναι κυρτό, αν το ευθύγραμμο τμήμα περιέχεται μέσα σε αυτό. Με άλλα λόγια, ένα σύνολο C είναι κυρτό, αν για οποιαδήποτε $x, y \in C$, τότε το ευθύγραμμο τμήμα $\{ax + (1-a)y : a \in [0,1]\}$, περιέχεται στο C. Μέσω της επαγωγής, μπορούμε να πούμε ότι ένα σύνολο είναι κυρτό, ανν για κάθε πεπερασμένο υποσύνολο $\{x_1, x_2, \ldots, x_n\}$ του C και μη αρνητικά βαθμωτά $\{a_1, a_2, \ldots, a_n\}$ που ισχύει ότι $\sum_{i=1}^{n} a_i = 1$, τότε ο γραμμικός συνδυασμός $\sum_{i=1}^{n} a_i x_i$, βρίσκεται στο C. Τέλος, το κυρτό σύνολο έχει αρκετές ενδαφέρουσες ιδιότητες, τις οποίες παραθέτουμε παρακάτω.

Ιδιότητες:
Σε οποιδήποτε διανυσματικό χώρο:

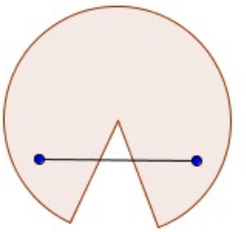
1. Το άθροισμα δύο κυρτών συνόλων, είναι κυρτό σύνολο.



Σχήμα 2: Κυρτός κώνος



(α′) Κυρτό σύνολο



(β′) Μη κυρτό σύνολο

Σχήμα 3: Παραδείγματα κυρτών και μη κυρτών συνόλων.

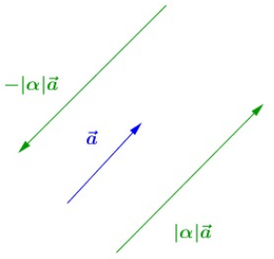
2. Ο βαθμωτός πολλαπλασιασμός ενός κυρτού συνόλου, είναι κυρτό σύνολο.

3. Η τομή μιας οικογένειας κυρτών συνόλων, είναι κυρτό σύνολο.

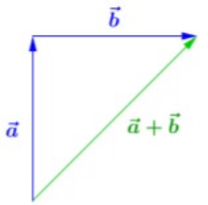
4. Ένα κυρτό σύνολο που περιέχει το μηδέν, είναι κυκλικό ανν είναι συμμετρικό.

5. Σε έναν τοπολογικό διανυσματικό χώρο, τόσο το εσωτερικό (*interior*) όσο και το κλειστό (*closure*) του κυρτού συνόλου, είναι κυρτό σύνολο.

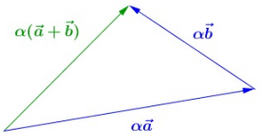
6. Ένα σύνολο $\mathcal{C}$ είναι κυρτό ανν $a\mathcal{C} + b\mathcal{C} = (a + b)\mathcal{C}$ για όλα τα a και b, μη-αρνητικά βαθμωτά.



(α′)



(β′)



(γ′)

Σχήμα 4: Ιδιότητα της επιμεριστικότητας

3 Αξιωματική θεμελείωση της μεθόδου MDS

Σε αυτή την ενότητα θα αναπτυχθούν κάποιες σημαντικές αξιωματικές αρχές που είναι χρήσιμες για την εφαρμογή της μεθόδου MDS. Έτσι μία πολύ σπουδαία έννοια είναι αυτή του μέτρου ή απόστασης. Οπότε μπορούμε να αναρωτηθούμε τι εννούμε όταν λέμε μέτρο; Εδώ είναι προτιμότερο να αναπτύξουμε μία εισαγωγή, με τη μορφή ενός παραδείγματος, πρωτού δωθεί ένας αυστηρότερος μαθηματικός ορισμός. Ας αναλογιστούμε, ότι κάποιο άτομο σκοπεύει να πραγματοποιήσει ένα ταξίδι από το Γκέτινγκεν (*Göttingen*) στη Χαϊδελβέργη (*Heidelberg*), της Γερμανίας. Τότε αυτά που μας έρχονται με μιας στο μυαλό είναι:

1. Η χιλιομετρική απόσταση που συνδέει το Γκέντινγκεν με τη Χαϊδελβέργη, από την άποψη μιας ευθείας γραμμής.
2. Η χιλιομετρική απόσταση μεταξύ των δύο αυτών πόλεων, μέσω του οδικού δικτύου.
3. Ο χρόνος σε λεπτά, της κοντινότερης διαδομής, μέσω του σιδηροδρομικού δικτύου.
4. Το κόστος σε ευρώ, της φθηνότερης διαδρομής των ανωτέρω δύο πόλεων, μέσω του σιδηροδρόμου.

Την απάντηση στα παραπάνω, μπορούμε να την πάρουμε από την Τοπολογία. Θα πρέπει να αναφερθεί, ότι είναι γνωστό ότι ένα σημείο είναι μία ακριβή θέση στην επιφάνεια του επίπεδου. Δηλαδή όταν βάζουμε μία τελεία για να υποδηλώσουμε τη θέση ενός σημείου, αυτό στην πραγματικότητα, δε θα έχει κάποια διάμετρο, ανεξαρτήτου το πόσο θα το μεγενθύνουμε-ζουμάρουμε. Επομένως, αφού το σημείο είναι μία θέση, δεν έχει διαστάσεις. Οπότε μετά από αυτή τη διευκρίνιση, μπορούν να αναφερθούν πιο αυστηροί ορισμοί και ιδιότητες των μετρικών χώρων. Οπότε, για κάθε ζευγάρι σημείων, θα υπάρχει μία και μόνο μία απόσταση. Πιο συγκεκριμένα, ας υποθέσουμε ότι έχουμε ένα σύνολο $\mathcal{A}$, το οποίο περιέχει κάποια σημεία και έστω ότι αυτό το σύνολο έχει ως στοιχεία τα $a \in \mathcal{A}$ και $b \in \mathcal{A}$. Τότε, για κάθε ζευγάρι

$$(a \times b) \in < \mathcal{A} \times \mathcal{A} > \tag{3.1}$$

θα υπάρχει ένα μοναδικό στοιχείο του συνόλου $\mathbb{R}$, των πραγματικών αριθμών. Αυτό το στοιχείο, δε θα είναι τίποτα άλλο, παρά η απόσταση μεταξύ των σημείων a, b και συμβολίζεται ως d_{ab}. Επιπλέον, η συνάρτηση απόστασης (distance function) είναι εκείνη η οποία θα προσδιορίσει τον πραγματικό αριθμό που θα αντιστοιχεί σε ένα συγκεκριμένο ζευγάρι σημείων και ο χώρος ο οποίος περιέχει αυτά τα σημεία είναι ο μετρικός χώρος.

Ορισμός 3.1. *Μέτρο ή απόσταση σε ένα σύνολο $\mathcal{X}$ είναι μία συνάρτηση d: $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$, που ικανοποιεί τα παρακάτω αξιώματα:*

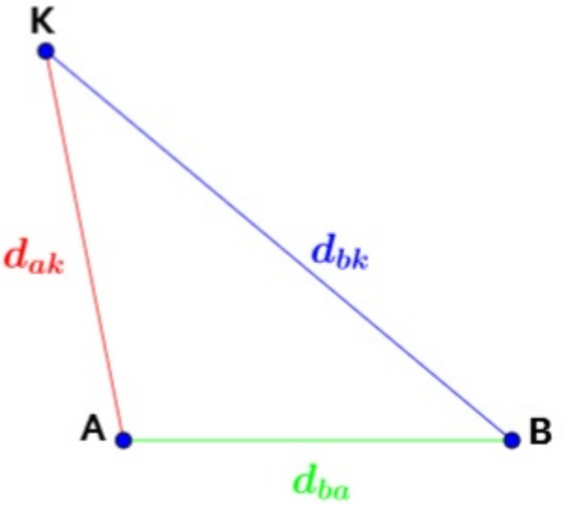
1. *Διακριτή (Discrimination): $d_{ab} = 0 \quad \Rightarrow \quad a = b$*
2. *Μη-Αρνητική (Non-Negativity): $d_{ab} \geq 0$ και $d_{aa} = 0$, για κάθε $a, b \in \mathcal{X}$*
3. *Συμμετρική (Symmetric): $d_{ab} = d_{ba}$, για κάθε $a, b \in \mathcal{X}$*
4. *Τριγωνική ανισότητα (Triangle Inequality): $d_{ab} \leq d_{ak} + d_{kb}$, για κάθε $a, b, k \in \mathcal{X}$*

Τότε καταλήγουμε ότι η d είναι μετρική στο $\mathcal{X}$ και το ζευγάρι $(\mathcal{X}, d)$ ονομάζεται μετρικός χώρος (metric space). Επιπλέον, ημί-μετρική d στο χώρο $\mathcal{X}$, είναι η πραγματική συνάρτηση στο $\mathcal{X} \times \mathcal{X}$, για την οποία ισχύει η θετικότητα, συμμετρικότητα, καθώς και ικανοποιείται η συνθήκη $d_{aa} = 0$ και για κάθε a ικανοποιείται και η τριγωνική ανισότητα $d_{bk} \leq d_{ba} + d_{ak}$. Αξίζει να αναφερθεί ότι η μετρική είναι ημι-μετρική μιας και έχει την ιδιότητα ότι $d_{ab} = 0$ που συνεπάγεται $a = b$. Ακόμη, είναι πολύ χρήσιμο να αναφερθεί η έννοια της ανοικτής σφαίρας (μπάλας)(open ball), η οποία παίζει σπουδαίο ρόλο στους μετρικούς χώρους. Έτσι, έστω $(\mathcal{X}, d)$ μετρικός χώρος. Ισχύει ότι για κάθε $x \in \mathcal{X}$ και $\rho \in \mathbb{R}$, με $\rho > 0$, θα συμβολίζουμε και θα γράφουμε με

$\mathcal{B}(x, \rho)$ την ανοικτή σφαίρα του $\mathcal{X}$ με κέντρο x και ακτίνα ρ και θα γράφουμε:

$$\mathcal{B}(x, \rho) = \{y \in \mathcal{X} | d(x, y) < \rho\} \tag{3.2}$$

Αν θέλουμε να δώσουμε μία περιγραφική ανάλυση των παραπάνω τεσσάρων αξιωμάτων, που αναφέρονται στον ορισμό 3.1, θα μπορούσαμε να πούμε τα εξής:
Το μεν πρώτο αξίωμα είναι πολύ ξεκάθαρο και μας λέει ότι η απόσταση μεταξύ δύο ταυτόσημων σημείων a και b, θα είναι πάντα μηδέν, καθώς και από την μια πλευρά είναι αδύνατο για δύο διαφορετικά σημεία, να βρίσκονται στην ίδια θέση και από την άλλη, ένα μοναδικό σημείο δεν μπορεί να βρίσκεται σε περισσότερες από μία θέσεις. Όσο για την τρίτη ιδιότητα, αυτή της συμμετρικότητας, δεν έχει σημασία η απόσταση από το a στο b ή το αντίστροφο, ή ακόμη καλύτερα δεν υπάρχει διαφορά στην απόσταση για έναν παρατηρητή, που βλέπει την απόσταση μεταξύ του a και b ή το αντίστροφο. Τέλος, όσο αφορά σχετικά με την τέταρτη ιδιότητα, μπορεί να βοηθήσει στην κατανόηση, το Σχήμα 5.



Σχήμα 5: Τριγωνική ανισότητα

Αυτό που μας υποδηλώνει το Σχήμα 5, είναι ότι η άμεση απόσταση από το b στο k, είναι η μικρότερη απόσταση αν οδηγούμασταν, πάλι στο ίδιο σημείο, αλλά αυτή τη φορά μέσω του σημείου A. Αυτό το συγκεκριμένο αξίωμα, σχετίζεται με το γεγονός ότι στον Ευκλείδειο χώρο, η μικρότερη απόσταση μεταξύ δύο σημείων, είναι η ευθεία γραμμή. Πράγματι, αυτό ισχύει, ότι δηλαδή η απόσταση μέσω της έμμεσης γραμμής, θα είναι πάντοτε μεγαλύτερη από την ευθεία γραμμή, εκτός και αν το τρίτο αυτό σημείο, βρίσκεται επάνω στην ευθεία που ενώνει τα δύο αυτά σημεία, όπως αποτυπώνεται αυτό στο Σχήμα 6.

Σχήμα 6: Οποιαδήποτε άλλη γραμμή η οποία πηγαίνει έμμεσα έστω ενός τρίτου σημείου, δε μπορεί να είναι μικρότερη από την ευθεία γραμμή.

Βέβαια, το προηγούμενο εξαιρείται στην περίπτωση όπου το τρίτο αυτό σημείο εμπίπτει στην γραμμή που εννώνει τα δύο αυτά σημεία. Επίσης, ημιμετρική στον $\mathcal{X}$ (semimetric), είναι μία συνάρτηση $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, που ικανοποιεί τα αξιώματα (2),(3) και (4). Είναι προφανές ότι κάθε μετρικός χώρος είναι και ημιμετρικός.

Αν η d είναι μετρική σε ένα χώρο $\mathcal{X}$, τότε το ζευγάρι $(\mathcal{X}, d)$ καλείται μετρικός χώρος και ομοίως, αν η d είναι ημιμετρική, τότε το ζευγάρι $(\mathcal{X}, d)$ είναι ένας ημιμετρικός χώρος. Για να ορίσουμε τα χαρακτηριστικά ενός μετρικού χώρου, δεν είναι απαραίτητο να ικανοποιούνται όλα τα αξιώματα αλλά αρκεί μόνο δύο από αυτά. Αυτό που θα πρέπει να ακολουθήσουμε είναι να ξαναγράψουμε την τριγωνική ανισότητα με διαφορετικό τρόπο από ότι είχε δωθεί στο παραπάνω ορισμό 3.1 καθώς και το αξίωμα της μη-αρνητικότητας (nonnegativity). Οπότε έχουμε:

$$d_{bk} \leq d_{ab} + d_{ak} \tag{3.3}$$

Όπως προαναφέρθηκε, η εξίσωση (3.3), δεν είναι ίδια με εκείνη που εμφανίζεται στον ορισμό (3.1). Οπότε τώρα είναι δυνατό να παραχθεί η συμμετρικότητα και η θετικότητα, κάνοντας χρήση μόνο της σχέσης (1), από τον ορισμό 3.1 και την σχέση (3.3). Αν θέσουμε $k = b$, τότε η σχέση (3.3), δίνει:

$$\begin{aligned} d_{bb} &\leq d_{ak} + d_{ak} \\ d_{bb} &\leq 2d_{ab} \end{aligned} \tag{3.4}$$

Αλλά από το αξίωμα (1), είναι γνωστό ότι $d_{bb} = 0$, οπότε η σχέση (3.4) γίνεται:

$$0 \leq 2d_{ab} \tag{3.5}$$

οπότε είναι φανερό από την ανισότητα (3.5), ότι η d_{ab} δεν είναι αρνητική (positivity). Από την άλλη πλευρά, αν θέσουμε $k = a$, τότε η ανισότητα (3.3), δίνει:

$$\begin{aligned} d_{ba} &\leq d_{ab} + d_{aa} \\ d_{ba} &\leq d_{ab} \end{aligned} \tag{3.6}$$

Ωστόσο, μιας και τα σημεία a, b είναι τυχαία επιλεγμένα, οδηγούμαστε στην εξής παρατήρηση:

$$\begin{aligned} d_{ab} &\leq d_{ba} + d_{bb} \\ d_{ab} &\leq d_{ba} \end{aligned} \tag{3.7}$$

Οπότε από τις σχέσεις (3.6) και (3.7), συνεπάγεται ότι d_{ab} πρέπει να είναι ίσο με d_{ba}. Έτσι αντιλαμβανόμαστε ότι είναι εφικτό να εξάγουμε τη συνθήκη της συμμετρικότητας από την σχέση (3.3) και να παράγουμε την συνθήκη της θετικότητας, η οποία παρήχθει από το αξίωμα (1) και την εξίσωση (3.3). Τέλος, η συμμετρικότητα μας δίνει το δικαίωμα να λάβουμε το αξίωμα (4), από την σχέση (3.3). Επομένως, το αξίωμα (1) και η σχέση (3.3), είναι ικανά για να ορισθούν τα αξιώματα των μετρικών χώρων που δίνονται από τα (1) έως το (4), του ορισμού 3.1. Στην συνέχεια αξίζει να αναφερθούν μερικά χαρακτηριστικά παραδείγματα μετρικών χώρων.

α′) **Η ευθεία γραμμή**
Ας υποθέσουμε ότι ο $X = \mathbb{R}$ είναι ένα σύνολο πραγματικών αριθμών με μετρική: $d(x, y) = |x - y|$

β′) **Ο Ευκλείδειος χώρος**
Έστω $x = (\xi_1, \ldots, \xi_m)$ είναι το σύνολο όλων των διαταγμένων m-πλειάδων από πραγματικούς αριθμούς, με Ευκλείδεια μετρική. Αν $x = (\xi_1, \ldots, \xi_m)$ και $y = (\eta_1, \ldots, \eta_m)$ ανήκει στο $\mathbb{R}^{\mathrm{m}}$, τότε ορίζουμε:

$$d(x, y) = \sqrt{(\xi_1 - \eta_1)^2 + \cdots + (\xi_m - \eta_m)^2} \tag{3.8}$$

γ′) **Ο ακολουθιακός χώρος $\mathbf{l}_\infty$**
Ο χώρος $X = \mathbf{l}_\infty$ περιέχει όλες τις φραγμένες ακολουθίες των πραγματικών αριθμών: η ακολουθία $x = (\xi_1, \ldots, \xi_m)$ (σε συντομογραφία γράφουμε $x = (\xi_k)$) ανήκει στο X αν υπάρχει $M_x > 0$, το οποίο εξαρτάται από την ακολουθία, τ.ω. $\forall\ k \in \mathbb{N}$, $|\xi_k| \leq M_x$. Ισοδύναμα, $x = (\xi_k) \in X \Leftrightarrow Sup\{|\xi_k| : k \in \mathbf{N}\} < +\infty$
Ορίζουμε τις αποστάσεις μεταξύ δύο φραγμένων ακολουθιών $x = (\xi_k), y = (\eta_k)$ ως:

$$d(x, y) = Sup\{|\xi_k - \eta_k| : k \in \mathbb{N}\} \tag{3.9}$$

δ′) **Ο συναρτησιακός χώρος $\mathbf{C}[\mathbf{a}, \mathbf{b}]$**
Έστω $[\alpha, \beta]$ ένα κλειστό διάστημα στο $\mathbb{R}$. Ο χώρος $X = C[a, b]$ περιέχει όλες τις συνεχείς συναρτήσεις f: $[a, b] \to \mathbb{R}$. Την απόσταση μεταξύ δύο σημείων του χώρου, μπορεί να ορισθεί ως:
Αν f, g: $[a, b] \to \mathbb{R}$ είναι συνεχείς συναρτήσεις, θέτουμε $d(f, g) = \max_{t\in[a,b]} |f(t) - g(t)|$, το μέγιστο είναι καλά ορισμένο: $|f - g|$ είναι συνεχές στο $[a, b]$ και έκτοτε λαμβάνει τη μέγιστη τιμή. Ένας τέτοιος χώρος, που μόλις προαναφέρθηκε, συμβολίζεται με C [a, b] και τον αποκαλούμε *χώρο των συνεχών συναρτήσεων στο* $[\alpha, \beta]$.

ε′) **Η διακριτή μετρική**
Υποθέτουμε ένα μη κενό σύνολο $\mathcal{X}$ και $\forall\ x, y \in \mathcal{X}$ ορίζουμε:

$$d(x, y) = \begin{cases} 1, & if \quad x \neq y \\ 0, & if \quad x = y \end{cases}$$

d να είναι η διακριτή μετρική στο σύνολο $\mathcal{X}$.

3.1 Ευκλείδειοι χώροι

3.1.1 Εσωτερικό γινόμενο και Ευκλείδειοι χώροι

Το όλο πλαίσιο του διανυσματικού χώρου μας δίνει την ευκαιρία να πραγματευόμαστε με σχέσεις διανυσμάτων και γραμμικούς συνδυασμούς, αλλά δεν υπάρχει η δυνατότητα να εκφραστεί η έννοια του μήκους ενός ευθήγραμμου τμήματος ή για να αναφερθούμε περί ορθογωνιότητας των διανυσμάτων. Αντ᾽ αυτού η Ευκλείδεια δομή μας δίνει το δικαίωμα να δουλεύουμε με έννοιες μετρικές, όπως η ορθογωνιότητα και το μήκος (ή απόσταση). Αρχικά δίνεται ο ορισμός της Ευκλείδειας δομής, ενός διανυσματικού χώρου.

Ορισμός 3.2. *Ένας πραγματικός διανυσματικός χώρος E, θα είναι Ευκλείδειος χώρος αν είναι εφοδιασμένος με τον τύπο της συμμετρικότητας φ: $E \times E \to \mathbb{R}$, η οποία είναι και θετικά ορισμένη, πράγμα που σημαίνει ότι: $\varphi(u, u) > 0$, για κάθε $u \neq 0$.*

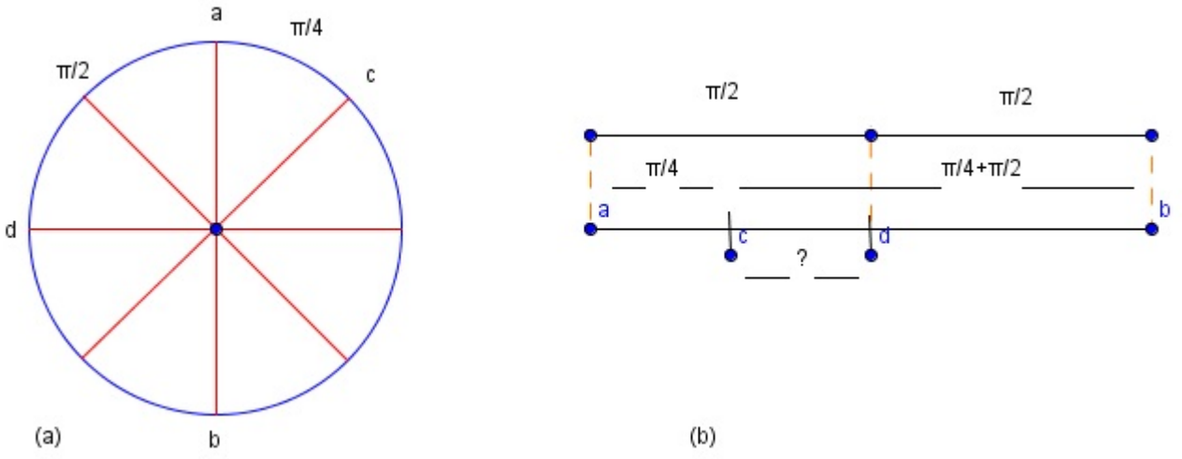
Πιο συγκεκριένα, η φ: $E \times E \to \mathbb{R}$ ικανοποιεί τα παρακάτω αξιώματα:

$$\begin{aligned} \varphi(u_1 + u_2, v) &= \varphi(u_1, v) + \varphi(u_2, v) \\ \varphi(u, v_1 + v_2) &= \varphi(u, v_1) + \varphi(u, v_2) \\ \varphi(\lambda u, v) &= \lambda\varphi(u, v) \\ \varphi(u, \lambda v) &= \lambda\varphi(u, v) \\ \varphi(u, v) &= \varphi(v, u) \\ u \neq 0 &\Rightarrow \varphi(u, u) > 0 \end{aligned} \tag{3.10}$$

Ο πραγματικός αριθμός $\varphi(u, v)$, καλείται επίσης και εσωτερικό γινόμενο (ή βαθμωτό γινόμενο) του u και v. Επίσης, ορίζεται η τετραγωνική μορφή που σχετίζεται με την φ ως συνάρτηση Φ: $E \to \mathbb{R}_+$, τ.ω. $\Phi(u) = \varphi(u, u)$

για όλα τα $u \in E$. Από τη στιγμή που η φ, είναι γραμμική ως προς όλες τις μεταβλητές (bilinear), έχουμε ότι $\varphi(0,0) = 0$ και μιας και είναι θετικά ορισμένη, έχουμε το ισχυρότερο γεγονός ότι $\varphi(u,u) = 0$ ανν $u = 0$, δηλαδή $\Phi(u) = 0$ ανν $u = 0$. Οστώσο, οι έννοιες «μετρικός» και «μη μετρικός», όπως χρησιμοποιούνται στην ψυχομετρεία και στο MDS, έχουν εντελώς διαφορετική σημασία.

3.2 Απόσταση σε κύκλο



Σχήμα 7: (α) Ακτινικές αποστάσεις μεταξύ $a,\ldots,d$ και (β) πώς ερμηνεύονται ως Ευκλείδειες αποστάσεις.

Το Σχήμα 7 απεικονίζει, μια διάταξη τεσσάρων σημείων σε ένα κύκλο. Για να γίνει ο προσδιορισμός των αποστάσεων, συνήθως χρησιμοποιούμε ένα χάρακα. Αυτό που θα έχει επιτευχθεί, θα έχει ως αποτέλεσμα την παραγωγή Ευκλείδειων αποστάσεων. Οστώσο, θα πρέπει να επισημανθεί ότι όλες οι συναρτήσεις, δε σημαίνει οτι εκφράζουν μία απόσταση (distance function). Και αυτό γιατί, για να την χαρακτηρίζει (τη συνάρτηση) ως απόσταση, θα πρέπει αυτή να αντιστοιχεί έναν πραγματικό αριθμό για κάθε ζευγάρι πραγματικών αριθμών καθώς επίσης και να ικανοποιούνται τα τέσσερα αξιώματα, που έχουν αναφερθεί στον ορισμό 3.1. Επιπλέον, οι Ευκλείδειες αποστάσεις σχετίζονται με ιδότητες όπως αυτές των βαθμωτών γινομένων. Δύο από αυτές τις ιδιότητες αντιστοιχούν σε γενικότερα αξιώματα των αποστάσεων, όπως η συμμετρικότητα (symmetry) και η μη αρνητικότητα (nonnegatitivity). Βέβαια, υπάρχει και άλλη μία ιδιότητα, αυτή της γραμμικότητας (linearity), η οποία εισάγει σε πιο ειδικού τύπου ιδιότητες των Ευκλείδειων αποστάσεων, η οποία δίνεται από τη σχέση:

$$b(s \circ u + t \circ v, w) = s \cdot b(u,w) + t \cdot b(v,w) \tag{3.11}$$

για οποιαδήποτε διανύσματα u,v,w και βαθμωτό s,t. Στην παραπάνω σχέση (3.11), οι συμβολισμοί «·» και «∘» συμβολίζουν, το συνήθη πολλαπλασιασμό πραγματικών αριθμών και τον πολλαπλασιασμό ενός διανύσματος με αριθμό (scalar) αντίστοιχα. Επίσης, με το συμβολισμό + , θεωρείται η πρόσθεση διανυσμάτων και όχι η συνήθη πρόσθεση μεταξύ αριθμών.[1] Φυσικά θα πρέπει να υπάρχουν κάποιοι κανόνες, κάτω από τους οποίους, τόσο ο βαθμωτός πολλαπλασιασμός όσο και η πρόσθεση διανυσμάτων να έχουν νόημα. Όσο αυτοί δεν ορίζονται, η έννοια της γραμμικότητας δεν έχει νόημα. Οι κανόνες που συλλέγονται σε ένα σύστημα από αξιώματα είναι γνωστό ως *Αβελιανός διανυσματικός χώρος*. Σε αυτό περιλαμβάνεται το πεδίο (field) και η ομάδα (group). Στο μεν πεδίο, συνήθως αποτελείται από ένα σύνολο πραγματικών αριθμών, μαζί με τις δύο πράξεις της πρόσθεσης και του πολλαπλασιασμού, στη δε ομάδα στην οποία περιέχονται το σύνολο των στοιχείων με μία πράξη και ικανοποιούν τα παρακάτω αξιώματα:

[1] Ένας διαφορετικός συνδυασμός, για να αποφεύγεται η σύγχυση με το σύμβολο της πρόσθεσης + των αριθμών, είναι το σύμβολο ⊕.

1. για οποιαδήποτε από τα τρία στοιχεία x, y και z, $(x + y) + z = x + (y + z)$
2. υπάρχει το μηδενικό στοιχείο, z τ.ω. $x + z = x$, για κάθε x
3. υπάρχει ένα αντίστροφο στοιχείο $x^{(i)}$ για κάθε x, τ.ω. $x + x^{(i)} = z$
4. (μόνο για τις Αβελιανές ομάδες) για κάθε ζεύγος στοιχείων x, y, $x + y = y + x$

Έτσι ο διανυσματικός χώρος, «δένει» το πεδίο και την ομάδα μέσω της πράξης «$\circ$», τ.ω.:

1. $k \circ (x + y) = k \circ x + c \circ y$
2. $(s + t) \circ x = s \circ x + t \circ x$
3. $s \circ (t \circ x) = (s \cdot t) \circ x$
4. $e \circ x = x$

όπου s, t, e είναι βαθμωτά, e είναι το ουδέτερο στοιχείο του πεδίου και x, y είναι οποιαδήποτε στοιχεία της ομάδας. Αυτό που μας διδάσκουν τα παραπάνω είναι ότι, όταν αναφερόμαστε σε Ευκλείδειες αποστάσεις, θα πρέπει να έχουμε στο νού μας ότι πρόκειτε για μια πλούσια δομή. Δηλαδή, Ευκλείδειες αποστάσεις μπορούν να οριστούν μόνο σε συστήματα, που αναπτύχθηκαν παραπάνω. Μπορεί να οριστεί για ένα σύνολο σημείων u, v, w, εφόσον αυτά τα σημεία συνδέονται αρχικά με τα αντίστοιχα στοιχεία u, v, w του διανυσματικού χώρου, σε αντίθεση με τις αποστάσεις γενικότερα, που δεν απαιτούν τέτοια δομή. Έτσι, για να ελέγξουμε τις ιδιότητες του διανυσματικού χώρου, δε μπορούμε να το πραγματοποιήσουμε για οποιοδήποτε πεπερασμένο σύνολο διανυσμάτων και αυτό γιατί θα πρέπει να ισχύει για κάθε βαθμωτά s και t και έκτοτε να περιλαμβάνει όλα τα διανύσματα του χώρου. Για αυτό όταν δίνεται ένα σύνολο αριθμών για να διαπιστώσουμε αν είναι Ευκλείδειες αποστάσεις, αυτό που θα πρέπει να διασταυρώσουμε, είναι αν αυτοί οι αριθμοί μπορούν να ενσωματωθούν σε αποστάσεις του Ευκλείδειου χώρου.

3.3 Ειδικοί χώροι

Σε αυτή την παράγραφο αναπτύσουμε τριών ειδών ειδικών χώρων, πέραν αυτών των γενικών χώρων που παρουσιάστηκαν προηγουμένως. Δύο από αυτούς είναι οι υπέρμετρικοι και τμηματικοί προσθετικοί χώροι αντίστοιχα (ultrametric and the additive segment spaces), οι οποίοι είναι ειδικοί τύποι μετρικών χώρων και ο τρίτος λέγεται γενικευμένος μετρικός χώρος. Ξεκινώντας με τον ultrametric μπορούμε να αναφέρουμε αρχικά τον ορισμό του.

Ορισμός 3.3. *Ένας μετρικός χώρος (X, d) θα λέγεται ultrametric, αν το μέτρο ικανοποιεί την ισχυρή τριγωνική ανισότητα, δηλαδή:*

$$d_{bk} \leq max\{d_{ba}, d_{ak}\} \tag{3.12}$$

για όλα τα $a, b, k \in X$. Ισοδύναμα μπορούμε να πούμε ότι ένας μετρικός χώρος (X, d), θα λέγεται ότι είναι ultrametric ανν δοθέντων τριών σημείων στον X, μπορούμε να αναδιατάξουμε τα a, b, k έτσι ώστε να ισχύει:

$$d_{bk} \leq d_{ba} = d_{ak}$$

Η παραπάνω σχέση που αναφέρεται στον ορισμό, κρίνεται αναγκαία συνθήκη για διάφορους τύπους ιεραρχικών συστάδων μοντέλων (hierarchical clustering models), τα οποία έχουν γίνει αρκετά δημοφιλή στις κοινωνικές επιστήμες και στη βιολογία. Αυτό που υποθέτουμε στους ultrametric χώρους είναι ότι η απόσταση μεταξύ δύο σημείων, δεν είναι μεγαλύτερη από τη μακρύτερη απόσταση ενός εκ των δύο αυτών σημείων, ως

προς ένα τρίτο. Επιπλέον, αξίζει να αναφερθεί, ότι αν a, b, k είναι τρία διακριτά σημεία σε έναν ultrametric χώρο (X, d), τότε αυτά δημιουργούν τις κορυφές ενός ισοσκελούς τριγώνου, όπου το (μετρικό) μήκος της βάσης δεν ξεπερνά το μήκος των πλευρών. Γι᾽ αυτό καμιά φορά στη βιβλιογραφία, τον ultrametric, μπορούμε να το συναντήσουμε και ως *ισοσκελή ή μη-Αρχιμήδεια*. Ενδιαφέροντα παραδείγματα ultrametric χώρων, αποτελούν οι δακτύλιοι $\mathbb{Z}_p$ των p-adic ακεραίων, ο Baire χώρος $B_{\aleph}$, η μη-Αρχιμήδεια νόρμα πεδίου [2] κλπ. Δεν είναι δύσκολο να διαπιστωθεί ότι ο ultrametric χώρος, είναι μία ειδική μορφή του μετρικού. Αυτό μπορεί να το αντιληφθεί κάποιος αν από τη σχέση:

$$\max(d_{ba}, d_{ak}) \leq d_{ba} + d_{ak} \tag{3.13}$$

σε συνδυασμό με τη σχέση (3.12), μας δίνει την τέταρτη εξίσωση του ορισμού (3.1), χωρίς να ισχύει και το αντίστροφο. Επιπλέον, εκτός του ultrametric χώρου, ένα άλλο είδος ειδικού χώρου είναι αυτό του κατά τμήμα προσθετικού χώρου (additive segment metric), ο οποίος αποτελεί εξίσου ενδιαφέρον στη τεχνική του MDS και μάλιστα εισήχθει από τους Beals, Krantz και Tversky (1968) [2]. Αυτού του είδους μετρικού χώρου, έχει όλες τις ιδιότητες του μετρικού χώρου που αναφέρονται στη πρώτη εξίσωση του ορισμού 3.1 και στη σχέση (3.3), καθώς επίσης μιας ακόμη ιδιότητας που αφορά την κατά τμήμα προσθετικότητα. Πιο γενικά, η επιπρόσθετη αυτή ιδιότητα απαιτεί ότι για οποιαδήποτε δύο σημεία a και b, θα υπάρχει ένα τρίτο σημείο k, το οποίο θα είναι ενδιάμεσο στα a, b και έτσι οι δύο μικρότερες αποστάσεις προστίθενται στη μεγαλύτερη. Έτσι για κάθε δύο σημεία, θα υπάρχει η δυνατότητα να συνδεθούν μέσω ενός ευθύγραμμου τμήματος, όπου οι αποστάσεις τους μπορούν να προστεθούν. Πιο επίσημα, το αξίωμα αυτό θα μπορούσε να διατυπώθει και με πιο επίσημο τρόπο ως εξής:

Ορισμός 3.4. *Για οποιαδήποτε δύο σημεία a και k για τα οποία ισχύει $d_{ak} > 0$, θα υπάρχει ένα σύνολο $\mathcal{S}$ σημείων τα οποία θα απεικονίζονται ένα-προς-ένα και επί, σε ένα διάστημα των πραγματικών αριθμών τ.ω. a και k θα είναι τα τελικά σημεία αυτού του διαστήματος τ.ω.*

$$d_{as} + d_{sk} = d_{ak} \tag{3.14}$$

για όλα τα $s \in \mathcal{S}$

Επίσης, οι Beals, Krantz και Tversky, παρουσίασαν ένα παράδειγμα μετρικών χώρων όπου στο μεν ένα περιελάμβαναν την ιδιότητα της κατά τμήμα προσθετικότητας και στο δε άλλον όχι. Η σκέψη ήταν να υποθέσουν διάφορα σημεία στην περιφέρεια του κύκλου τα οποία μπορούμε να τα δούμε ότι διαμορφώνουν δύο μέτρα. Ένα εξ᾽ αυτών είναι η Ευκλείδεια απόσταση σε δύο διαστάσεις, η οποία είναι επίσης το μήκος της χορδής που συνδέει οποιαδήποτε δύο σημεία, που βρίσκοντα επάνω στην περιφέρεια του κύκλου. Το άλλο είναι το μήκος του μικρότερου τόξου, που συνδέει οποιαδήποτε δύο σημεία. Αυτοί οι δύο μετρικοί μπορεί να παράγουν διαφορετικές αποστάσεις, αλλά η τάξη είναι η ίδια και για τους δύο. Γι᾽ αυτό οι μετρικοί είναι κανονικοί ισομετρικοί αλλά όχι αυστηρά ισομετρικοί [3] . Τέλος, για την τρίτη ειδική περίπτωση μετρικού, μπορούμε να πούμε ότι ενώ από τη μια ο υπερμετρικός (ultrametric), που είναι ειδική περίπτωση του μετρικού είναι μετρικός, από την άλλη πλευρά η περίπτωση του γενικευμένου μετρικού, δεν είναι μετρικός. Ο γενικευμένος μετρικός είναι αποτέλεσμα των αξιωμάτων 1,2 και 4 αλλά όχι του τρίτου, που αναφέρονται στον ορισμό 3.1. Επομένως, κάθε συνάρτηση που εκφράζει απόσταση και είναι μετρική, είναι επίσης και γενικευμένη μετρική. Ωστόσο, η συνάρτηση απόστασης (distance function) για την οποία ισχύει:

$$d_{as} \neq d_{sa} \tag{3.15}$$

[2] Μία νόρμα $\|\cdot\|$ θα λέγεται μη-Αρχιμήδεια αν $\|x + y\| \leq \max(\|x\|, \|y\|)$.

[3] Δοθέντος ενός μετρικού χώρου, θα λέμε ισομετρία ένα μετασχηματισμό ο οποίος απεικονίζει τα στοιχεία στον ίδιο ή σε άλλο μετρικό χώρο, ε.ω. οι αποστάσεις μεταξύ των στοιχείων της απεικόνισης στο νέο μετρικό χώρο, να είναι ίσες με τις αποστάσεις μεταξύ των στοιχείων του αρχικού μετρικού χώρου $d(a, b) = d(f(a), f(b))$.

που παρ᾽ όλο ικανοποιεί τα αξιώματα για τον γενικευμένο χώρο, δεν είναι όμως μετρικός. Τέτοιου είδους συναρτήσεις απόστασης, (distance function) λέγονται ασσυμετρικές συναρτήσεις απόστασης (assymetric distance function). Γι᾽ αυτό, ναί μεν μία ασσυμετρική συνάρτηση απόστασης να μην είναι μετρική, μπορεί όμως να είναι μία γενικευμένη μετρική.

Για την αξιωματική θεμελίωση της μεθόδου MDS είναι επιτακτική ανάγκη να αναφερθούν και κάποιες άλλες πολύ χρήσιμες και στοιχειώδεις έννοιες, για τους μετρικούς χώρους όπως οι έννοιες της πληρότητας (completeness) ενός χώρου (πλήρεις χώροι) και ισομετρίας. Φυσικά πολύ σημαντική έννοια είναι επίσης και αυτή της σύγκλισης μιας ακολουθίας σε ένα όριο. Πιο συγκεκριμένα, έστω ότι έχουμε ένα μετρικό χώρο $\mathcal{E}$ και μία ακολουθία από σημεία στο χώρο αυτόν, που τα συμβολίζουμε με $a_1, a_2, \ldots, a_n$. Θα λέμε ότι αυτή η ακολουθία συγκλίνει στον αριθμό, στο σημείο, $A \in \mathcal{E}$, αν η απόσταση μεταξύ των σημείων της ακολουθίας και του σημείου A (του ορίου δηλαδή), γίνεται ολοένα και μικρότερη από κάθε άλλον προηγούμενο επιλεχθέντα θετικό αριθμό r. Επιπλέον, μια ακολουθία $\{a_n\}$, θα λεγεται ότι είναι Cauchy σε ένα μετρικό χώρο (X, d), αν για κάθε $\varepsilon > 0$, υπάρχει κάποιο n_0 (που εξαρτάται από το ε) που ικανοποιεί ότι $d(a_n, a_m) < \varepsilon$, για όλα τα $n, m \geq n_0$ ή ισοδύναμα, αν $\lim_{n,m\to\infty} d(a_n, a_m) = 0$ ή ισοδύναμα $\lim_{n\to\infty} diam\{a_n, a_{n+1}, \ldots\} = 0$. Έτσι ο μετρικός χώρος (X, d) είναι πλήρης, αν κάθε Cauchy ακολουθία στο X συγκλίνει στο X, στην οποία περίπτωση θα μπορούμε να πούμε ότι το d είναι πλήρης μετρικός στο X. Επίσης, μια βασική και εξίσου σημαντική έννοια στους μετρικούς χώρους, είναι αυτή της ισομετρίας (Isometry).

Ορισμός 3.5. *Ισομετρία μεταξύ μετρικών χώρων (X, d) και (Y, d) είναι η ένα-προς-ένα συνάρτηση φ, που απεικονίζει το X στο Y και ικανοποιεί την ακόλουθη σχέση:*

$$d(x, y) = \rho(\phi(x), \phi(y)) \tag{3.16}$$

για όλα τα $x, y \in X$. Αν επιπροσθέτως, η συνάρτηση φ είναι και επί, τότε οι (X, d) και (Y, d) είναι ισομετρικοί.

4 Διανυσματική Παλινδρόμηση

Η ανάλυση παλινδρόμησης είναι πολύ χρήσιμο και σύνηθες εργαλείο στη στατιστική και μας βοηθά να ερμηνεύσουμε τους εκτιμητές που λαμβάνουμε ως λύση όταν παλινδρομούμε την εξαρτημένη μεταβλητή, έστω Y στις ανεξάρτητες ή επεξηγηματικές μεταβλητές, έστω X. Επίσης, στην περίπτωση που έχουμε ποσοτικοποιημένα δεδομένα, είναι εφικτό να πραγματοποιήσουμε πολλαπλή παλινδρόμηση συνηθισμένων ελαχίστων τετραγώνων. Αντιθέτως, στην περίπτωση που έχουμε ποιοτικά δεδομένα, τότε ακολουθούμε βέλτιστη πολλαπλή παλινδρόμηση. Παρακάτω γίνεται μία υπενθύμιση του μοντέλου της πολλαπλής παλινδρόμησης, κάνοντας χρήση των πινάκων.

4.1 Μοντέλο πολλαπλής παλινδρόμησης με μορφή πινάκων

Έστω η εξίσωση παλινδρόμησης:

$$y = b_0 + b_1 x_1 + \cdots + b_k x_k + \varepsilon \tag{4.1}$$

και έστω ότι έχουμε N παρατηρήσεις για τις μεταβλητές μας $y, x_1, \ldots, x_k$, που συμβολίζονται με δείκτη $t = 1, \ldots, N$. Τότε, οι N πραγματοποιήσεις της παραπάνω σχέσης μπορούν να γραφούν με τη βοήθεια των πινάκων όπως στην παρακάτω σχέση:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & & \vdots & \vdots \\ 1 & x_{N1} & \ldots & x_{Nk} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix} \tag{4.2}$$

με τον πίνακα Y να έχει διάσταση $(n \times 1)$, τον πίνακα των επεξηγηματικών μεταβλητών X με διάσταση $(n \times k)$, ο πίνακας των εκτιμητών b, με διάσταση $(k \times 1)$ και τέλος ο πίνακας των υπολοίπων, με διάσταση $(n \times 1)$. Η προηγούμενη σχέση (4.2), μπορεί να γραφεί σε μια πιο συντετμημένη μορφή ως εξής:

$$y = \mathbf{X}b + \varepsilon \tag{4.3}$$

Σκοπός της παλινδρόμησης είναι να εξαχθεί μία έκφραση για τον εκτιμητή $\underset{\sim}{\hat{b}}$ μέσω της μεθόδου των ελαχίστων τετραγώνων. Ο σκοπός αυτής της τεχνικής είναι να ελαχιστοποιήσουμε το άθροισμα των τετραγωνικών υπολοίπων, το οποίο μπορεί να γραφεί με τη μορφή πινάκων ως εξής:

$$S(b) = (y - \mathbf{X}b)'(y - \mathbf{X}b) = y'y - y'\mathbf{X}b - b'\mathbf{X}'y + b'\mathbf{X}'\mathbf{X}b = y'y - 2y'\mathbf{X}b + b'\mathbf{X}'\mathbf{X}b \tag{4.4}$$

Αν πάρουμε τις συνθήκες πρώτης τάξης, ως προς το διάνυσμα του εκτιμητή $\underset{\sim}{\hat{b}}$ και θέσουμε μετά την παράσταση ίση με μηδέν, θα λάβουμε την εξής εξίσωση:

$$\frac{\partial S}{\partial b} = -2y'\mathbf{X} + 2b'\mathbf{X}'\mathbf{X} = 0 \tag{4.5}$$

Έτσι για να πάρουμε μια έκφραση που θα περιέχει το διάνυσμα του εκτιμητή, αρκεί ο πίνακας $X'X$, να είναι μη ιδιάζων, για να υπάρχει ο αντίστροφος, και να πάρουμε το επιθυμητό αποτέλεσμα, δηλαδή τις *κανονικές εξισώσεις (normal equations)*, πιο συγκεκριμένα:

$$\mathbf{X}'\mathbf{X}b = \mathbf{X}'y \tag{4.6}$$

Όπως προαναφέραμε, με την υπόθεση ότι ο αντίστροφος πίνακας $\mathbf{X}'\mathbf{X}$ υπάρχει, τότε η σχέση (4.6) επιστρέφει μοναδική λύση, η οποία είναι το διάνυσμα του εκτιμητή των ελαχίστων τετραγώνων:

$$\underset{\sim}{\hat{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y. \tag{4.7}$$

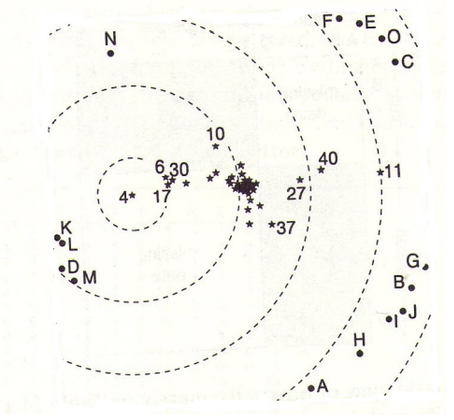
Στην συνέχεια γίνεται αναφορά σε μοντέλα, που χρησιμοποιούνται συχνά, τα οποία βοηθούν στην επεξήγηση των αποτελεσμάτων, ύστερα από την πραγματοποίηση της MDS ανάλυσης. Τα μοντέλα αυτά μας επιτρέπουν

να προσδιορίσουμε εκείνες τις ιδιότητες ή χαρακτηριστικά, τα οποία είναι πολύ κοντά σε εκείνα που λαμβάνουν υπόψη τους οι ερωτηθέντες, όταν οι τελευταίοι ασκούν την κριτική τους για τις ομοιότητες, που αφορά ένα σύνολο ερεθισμάτων (stimuli). Διάφοροι είναι εκείνοι οι τρόποι, με τους οποίους μπορούμε να εφαρμόσουμε παλλινδρόμηση. Ωστόσο, τα μοντέλα αυτά δεν μπορούν να μας πληροφορήσουν με απόλυτη βεβαιότητα, αν ο ανταποκρινόμενος χρησιμοποιεί αυτά τα χαρακτηριστικά που του θέτουν, όταν ο ίδιος εκφράζει προσωπικά την κρίση του για το θέμα που ερωτάται. Αντιθέτως, αυτά τα μοντέλα μας πληροφορούν, για το αν κάποιο γνώρισμα, δεν χρησιμοποιήθηκε από τον ερωτηθέντα. Έτσι έχουμε δύο κατηγορίες μοντέλων που μπορούμε να συναντήσουμε και να σχετίζονται με την κρίση των γνωρισμάτων. Κάθ᾽ ένα από αυτά, διαφέρει ως προς τη βασική υπόθεση που κάνει ως προς τη συνεισφορά στην κρίση της ομοιότητας των γνωρισμάτων που συμπεριλαμβάνονται στην έρευνα. Έτσι το πρώτο από τα δύο μοντέλα είναι η διανυσματική πολυμεταβλητή παλινδρόμηση. Σκοπός του μοντέλου αυτού είναι να βρούμε την κατεύθυνση του διανύσματος στον χώρο, η οποία αυξάνεται όσο αρέσει περισσότερο το χαρακτηριστικό. Έτσι τα χαρακτηριστικά ή οι ιδιότητες παρουσιάζονται ως διάφορες κατευθύνσεις στον πολυδιάστατο χώρο. Επιπλέον το μοντέλο αυτού του τύπου, είναι πιο σημαντικό όταν το σύνολο των ερεθισμάτων περιέχει εκείνα τα ερεθίσματα τα οποία δε διαθέτουν ούτε πάρα πολλά χαρακτηριστικά αλλά ούτε και πολύ λίγα, αντίστοιχα. Αυτό που θα πρέπει να μας μείνει στο μυαλό για το διανυσματικό μοντέλο είναι ότι στην ουσία ο τύπος του μοντέλου αυτού δεν είναι τίποτα άλλο παρά η εκτέλεση πολλαπλής παλινδρόμησης των χαρακτηριστικών στο διαβαθμισμένο χώρο (scaling space). Με άλλα λόγια το διανυσματικό μοντέλο βρίσκει την κατεύθυνση μέσω του διαβαθμισμένου χώρου, η οποία αντιστοιχεί στην αυξημένη ποσότητα για το χαρακτηριστικό αυτό. Το χαρακτηριστικό διάνυσμα είναι η ευθεία γραμμή στο χώρο, στον οποίο οι προβολές του ερεθίσματος αντιστοιχούν όσο είναι πιθανό πιο κοντά με την κλίμακα που έχει το χαρακτηριστικό για το ερέθισμα. Επομένως, όταν κάνουμε πολλαπλή παλινδρόμηση του παραπάνω τύπου μοντέλου, η μεταβλητή που εκφράζει το γνώρισμα ή την προτίμηση είναι η μεταβλητή απόκρισης (response variable), σε μία διαδικασία ανάλυσης παλινδρόμησης όπου χρησιμοποιεί τις διαστάσεις του πολυδιάστατου χώρου ως μεταβλητές πρόβλεψης. Έτσι πραγματοποιώντας τη διαδικασία της πολλαπλής παλινδρόμησης, αυτό που θέλουμε να πετύχουμε είναι να προσαρμόσουμε το διάνυσμα κατά τέτοιο τρόπο ώστε η ευθεία γραμμή να διέρχεται όσο είναι εφικτό πλησιέστερα στα σημεία να έχει δηλαδή την ελάχιστη απόσταση από αυτά. Ένα κριτήριο για το πόσο η ευθεία μας, περιγράφει καλύτερα το μοντέλο μας με βάση τις παρατηρήσεις που έχουμε, είναι ο συντελεστής πολλαπλής συσχέτισης (multiple correlation coefficient) R^2, χωρίς ωστόσο αυτό το κριτήριο να είναι το μοναδικό και το καλύτερο. Επομένως, όσο πιο μεγάλος είναι ο αριθμός αυτός και συγκεκριμένα, όσο πιο κοντά στη μονάδα είναι, τόσο καλύτερα προσαρμόζεται και περιγράφει καλύτερα το μοντέλο τα δεδομένα μας. Οι τιμές που λαμβάνει ο συντελεστής συσχέτισης, είναι από μηδέν έως τη μονάδα, δηλαδή $0 \leq R^2 \leq 1$. Στη συνέχεια, μιας και έχουμε ολοκληρώσει την παλινδρόμηση, είναι σχετικά εύκολο να παρουσιάσουμε το χαρακτηριστικό γνώρισμα, με τη μορφή ενός διανύσματος. Έτσι, έχοντας λάβει τους συντελεστές παλινδρόμησης, από την ανάλυσή μας, τους θεωρούμε ως συντεταγμένες του διανύσματος χαρακτηριστικών στον $r-$διάστατο χώρο. Οπότε, για να σχεδιάσουμε το κάθε διανυσματικό γνώρισμα-προτίμηση, κάνουμε χρήση του ακατέργαστου συντελεστή παλινδρόμησης (raw regression coefficients), ως συντεταγμένες ενός σημείου του MDS χώρου. Τέλος, φέρουμε ένα διάνυσμα από το ενδιάμεσο σημείο του χώρου ή από την αρχή των αξόνων, στην περίπτωση εκείνη που οι διαστάσεις είναι κεντραρισμένες ως προς το σημείο αυτό, βάζοντας ένα βέλος στο τελικό αυτό σημείο. Η κατεύθυνση που θα βρίσκεται το διάνυσμα έχει να κάνει με το πόσο προτιμάται αυτό το χαρακτηριστικό γνώρισμα. Αξίζει να αναφερθεί ότι το μήκος του διανύσματος αντιστοιχεί στη συσχέτιση. Όσο πιο μικρό είναι αυτό το μήκος τόσο πιο μικρή είναι η συσχέτιση καθώς και όχι τόσο ισχυρή η σχέση μεταξύ του χαρακτηριστικού γνωρίσματος-προτίμησης με τον MDS χώρο.

★ Παλινδρόμηση Ιδανικού Σημείου - Ideal Point Regression ★

Η παλινδρόμηση ιδανικού σημείου είναι μια ειδική περίπτωση της πολλαπλής παλινδρόμησης την οποία συναντάμε στο unfolding μοντέλο, το οποίο είναι κατάλληλο για την επιλογή προτιμήσεων. Το μοντέλο αυτό υποθέτει ότι τα άτομα αντιλαμβάνονται κατά τον ίδιο τρόπο το πλήθος των χαρακτηριστικών που έχουν να επιλέξουν, αλλά σε αυτό που τους διαφοροποιεί είναι το ως προς τι θεωρούν ως ιδανικό συνδυασμό των

χαρακτηριστικών που έχουν στη διάθεσή τους για επιλογή. Τα δεδομένα στο παραπάνω μοντέλο συνήθως έχουν μια μορφή σειράς κατάταξης των ατόμων που έχουν να επιλέξουν από ένα σύνολο χαρακτηριστικών. Μπορούμε να φανταστούμε, αυτού του είδους την παλινδρόμηση, ως ένα περίγραμμα (contour), όπου κάθε άτομο θα αποτυπώνεται ως ένα ιδανικό σημείο. Πιο συγκεκριμένα, έστω ότι έχουμε 100 άτομα και ζητάμε να κατατάξουν τους 300 βουλευτές, από το 1 (αυτόν που προτιμούν περισσότερο), έως 20 (αυτόν που προτιμούν ελάχιστα). Τα δεδομένα θα μπορούσαν να παρουσιαστούν σε έναν πίνακα διπλής εισόδου (two-mode)[4] και να κάνουμε το περίγραμμα (contour). Έτσι θα παρατηρούσαμε ότι το γράφημα, σαν αυτό που απεικονίζεται στο Σχήμα[5] 8, θα είχε πολλούς ομόκεντρους κύκλους όπου μέσα σε αυτούς παρουσιάζονται τόσο τα άτομα όσο και τα χαρακτηριστικά που έχουν να επιλέξουν. Κάθε άτομο παρουσιάζεται ως ένα ιδανικό σημείο και όσο ένα χαρακτηριστικό σημείο βρίσκεται πολύ κοντά στο ιδανικό, αυτό σημαίνει ότι τόσο περισσότερο προτιμάται από το άτομο αυτό. Δηλαδή, από το Σχήμα 8 το άτομο 4 προτιμά ελαφρώς περισσότερο τους βουλευτές K, L σε σχέση με τους αντίστοιχους D, M ενώ αντιθέτως για τους βουλευτές A, H, I, J, B, G, C, O, F δείχνει την απάθειά του, με τους κύκλους γύρω από το άτομο 4 να παρουσιάζουν τις ισο-προτιμήσεις περιγραμμάτων. Ισχύει ότι όσα χαρακτηριστικά παρουσιάζονται επάνω στον ίδιο κύκλο αυτό σημαίνει ότι για το άτομο είναι ίσης προτίμησης. Βέβαια, ανάλογα με τον τύπο του μοντέλου, θα αναφερθούν παρακάτω, το διάγραμμα αυτό που περιγράφει τα χαρακτηριστικά γνωρίσματα-προτιμήσεις (iso-attribute/preference), μπορεί να μην είναι μόνο κύκλος αλλά και έλλειψη, ανάλογα τον τύπο του μοντέλου που δουλεύουμε κάθε φορά. Κατ᾽ αυτό τον



Σχήμα 8: Οι κύκλοι δείχνουν τις ισο-προτιμήσεις για ένα άτομο, σε ένα διάγραμμα (contour).

τρόπο, οι προτιμήσεις του κάθε ατόμου, μοντελοποιούνται σχετίζοντας τα ιδανικά σημεία στα σημεία που παρουσιάζονται οι επιλογές των χαρακτηριστικών (π.χ. οι 300 βουλευτές) και με αυτό τον τρόπο ορίζουμε το *εξωτερικό «ξετύλιγμα» μοντέλου του ιδανικού σημείου (ideal-point model of external unfolding).* Αυτού του είδους μοντέλα υποθέτουν ότι οι ομοιότητες για τις επιλογές των χαρακτηριστικών θεωρούνται δεδομένα μέσω μιας MDS ανάλυσης που έχει προηγηθεί. Αν έχουμε δεδομένα προτιμήσεων για αυτά τα χαρακτηριστικά για ένα ή περισσότερα άτομα τότε το external unfolding τοποθετεί ένα σημείο για κάθε άτομο στο χώρο, έτσι ώστε όσο πιο κοντά βρίσκεται αυτό το σημείο σε εκείνο το σημείο που παρουσιάζει τις επιλογές των χαρακτηριστικών τόσο περισσότερο θα προτιμάτε αυτό το χαρακτηριστικό από το άτομο αυτό σε σχέση με τα υπόλοιπα που έχει στην διάθεσή του για επιλογή. Φυσικά υπάρχουν αρκετές τεχνικές για τον εντοπισμό εκείνου του ιδανικού σημείου. Μία από αυτές τις μεθόδους είναι αυτή του *χάρτη προτιμήσεων* (Preference Mapping or PREFMAP), που εισήχθει από τον Carroll το 1972 [5] στα εργαστήρια της Bell, η οποία υποθέτει ότι η προτίμηση είναι αντιστόφως ανάλογη στην τετραγωνική απόσταση από το ιδανικό σημείο. Η μέθοδος αυτή παρέχει «εξωτερικής» ανάλυσης σε δεδομένα προτίμησης. Δηλαδή, αναζητά να σχετίσει αντικείμενα προτιμήσεων για ένα σύνολο ερεθισμάτων σε μια υπάρχουσα διαμόρφωση των σημείων των ερεθισμάτων με τη βοήθεια των μέσων, μέσα από τέσσερα διαφορετικά μοντέλα. Στην ουσία αυτά τα μοντέλα είναι οι τέσσερις

[4]Λέμε έναν πίνακα διπλής εισόδου two-mode, όταν ο πίνακας αυτός έχει διαφορετικά στοιχεία εισόδου, στις γραμμές και στήλες αντίστοιχα.

[5]Πηγή: *Modern Multidimensional Scaling, Theory and Applications, Springer*, σελ.234

φάσεις του προγράμματος, οι οποίες είναι «εμφωλευμένες» με την έννοια ότι κάθε φάση είναι ειδική περίπτωση της φάσης που προηγήθηκε. Σε όλες τις τέσσερις φάσεις το πρόγραμμα προσπαθεί να αναζητήσει και να παρουσιάσει την πληροφορία προτίμησης ανάλογα με τα κριτήρια που του έχουμε ορίσει. Όσο προχωράμε από την πρώτη στην τέταρτη φάση οι περιορισμοί γίνονται πιο αυστηροί. Έτσι στην πρώτη φάση συναντάμε το γενικευμένο μοντέλο (unfolding model), ενώ στη δεύτερη φάση έχουμε το σταθμισένο μοντέλο (weighted unfolding model). Τέλος, στην τρίτη και τέταρτη φάση συναντάμε το απλό (simple unfolding model) και το διανυσματικό μοντέλο (vector model), αντίστοιχα. Επιπλέον, η ρουτίνα αυτή του *χάρτη προτιμήσεων*, δίνει τη δυνατότητα της επιλογής της μη-μετρικότητας, η οποία επιτρέπει, τα δεδομένα προτιμήσεων να υφίστανται ένα μονοτονικό μετασχηματισμό, με αντίκρυσμα την καλύτερη προσαρμογή στην περίπτωση της γραμμικότητας. Κάτω από αυτή την επιλογή, οι στατιστικοί έλεγχοι για τη σύγκριση της προσαρμογής των μοντέλων γίνονται καλύτεροι. Μία δεύτερη ρουτίνα που μπορεί να χρησιμοποιηθεί στη τοποθέτηση του ιδανικού σημείου σε ένα χώρο είναι ο *γραμμικός χάρτης* (Linear Mapping-LINMAP, Shocker and Srinivasan, 1974 [6]). Η ρουτίνα αυτή προέρχεται από την ίδια υπόθεση που συναντήσαμε στη δεύτερη φάση της μεθόδου PREFMAP δηλαδή, η προτίμηση είναι αντιστρόφως ανάλογη στο τετράγωνο της απόστασης από το ιδανικό σημείο σε ένα χώρο όπου επιτρέπεται το διαφορικό τέντωμα (differential stretching) των αξόνων για κάθε αντικείμενο. Αντί να χρησιμοποιεί την τετραγωνική παλινδρόμηση για να προσδιορίσει τη θέση των ιδανικών σημείων η LINMAP χρησιμοποιεί γραμμικό προγραμματισμό. Αυτή η προσέγγιση επιτρέπει να τεθούν οι περιορισμοί σε βάρη στους άξονες, έτσι ώστε να μην είναι αρνητικά. Ένα αρνητικό βάρος συνεπάγεται την ύπαρξη ενός αντι-ιδανικού σημείου, ούτως ώστε το επικρατέστερο σημείο της τετραγωνικής επιφάνειας να είναι το λιγότερο παρά το περισσότερο προτιμητέο. Ακόμη, υποθέτει ότι ο ερωτηθείς (decision maker) έχει ένα ιδανικό σημείο (σε όρους μέτρου σταθμισμένης Ευκλείδειας απόστασης), που υποδηλώνει την περιοχή που προτιμά περισσότερο σε ένα n-διάστατο χώρο χαρακτηριστικών. Εναλλακτικά, αντικείμενα που βρίσκονται πιο κοντά στο ιδανικό σημείο θεωρούνται ότι είναι υψηλής προτίμησης.

Στην περίπτωση που το περίγραμμα των προτιμήσεων απεικονίζεται από κύκλους, τότε λέμε ότι έχουμε την περίπτωση της κυκλικής παλινδρόμησης ιδανικού σημείου. Το μοντέλο του ιδανικού σημείου περιλαμβάνει την ίδια πολλαπλή παλινδρόμηση με εκείνη που συναντάμε στο διανυσματικό μοντέλο, μιας και το τελευταίο αποτελεί μια γενίκευση του ιδανικού μοντέλου. Πιο αναλυτικά η μεταβλητή απόκρισης είναι οι κλιμακούμενες προτιμήσεις (atributes/preferences), ενώ οι μεταβλητές πρόβλεψης αποτελούνται από τις διαστάσεις του χώρου ερεθισμάτων (stimulus space dimensions) καθώς επίσης και μία ή περισσότερες έξτρα μεταβλητή-ες (dummy variable), η οποία αποτελείται από το άθροισμα των τετραγώνων της ψευδομεταβλητής. Τέλος, όπως και στο διανυσματικό μοντέλο, οι συντελεστές παλινδρόμησης είναι εκείνες οι τιμές για τις οποίες μεγιστοποιείται η συσχέτιση μεταξύ της πραγματικής τιμής και της πρόβλεψης του χαρακτηριστικού. Η μεταβλητή πρόβλεψης δίνεται από την παρακάτω σχέση:

$$p_i = b_0 + \sum_{a=1}^{r} b_a x_{ia} + c \left(\sum_{g=1}^{r} x_{ig}^2 \right) \tag{4.8}$$

όπου p_i είναι το γνώρισμα ή η προτίμηση για το ερέθισμα i και x_{ia} είναι οι συντεταγμένες του σημείου i στη διάσταση a του $r-$διάστατου χώρου. Επίσης, οι συντελεστές παλινδρόμησης είναι οι b_a (για τις διαστάσεις) και με c συμβολίζουμε το άθροισμα τετραγώνων της μεταβλητής. Η εύρεση στον MDS χώρο, του ιδανικού σημείου που παρουσιάζει το χαρακτηριστικό γνώρισμα, δίνεται από τον λόγο:

$$y_{ia} = -\frac{b_a}{2c} \tag{4.9}$$

όπου y_{ia} είναι η $i-$στη συντεταγμένη του ιδανικού σημείου στον $r-$διάστατο κλιμακωτό χώρο, με b_a συμβολίζουμε το συντελεστή παλινδρόμησης για την a διάσταση. Η σχέση (4.9) περιγράφεται αναλυτικότερα από την επόμενη παράγραφο και συγκεκριμένα από τις σχέσεις (4.10) έως (4.14).

Ο σχεδιασμός αυτού του ιδανικού σημείου γίνεται με την εύρεση της θέσης των συντεταγμένων. Τέλος, στην περίπτωση που έχουμε το συντελεστή παλινδρόμησης της λανθάνουσας μεταβλητής (dummy variable)

να έχει αρνητικό πρόσημο, τότε το ιδανικό σημείο το ερμηνεύουμε ως το ελάχιστο προτιμητέο σημείο για αυτό το χαρακτηριστικό γνώρισμα.

★ Παραγωγή του Κυκλικού Ιδανικού Σημείου και ο τύπος για τις συντεταγμένες αυτού-Circular Ideal Point Formula for the Coordinates of the Ideal Point ★

Στην περίπτωση της σημειακής παλινδρόμησης, παλινδρομούμε τη y πάνω στα x. Αυτό μπορούμε να το αποτυπώσουμε ως εξής:

$$y = b_1 x_1 + b_2 x_2 + c\left(x_1^2 + x_2^2\right) \tag{4.10}$$

αν διαιρέσουμε την προηγούμενη σχέση με c τότε έχουμε:

$$\frac{y}{c} = x_1^2 + \frac{b_1}{c} x_1 + x_2^2 + \frac{b_2}{c} x_2 \tag{4.11}$$

προσθέτουμε και αφαιρούμε το $(\frac{b_1}{2c})^2$ κάνουμε χρήση δηλαδή ανάπτυγμα τετραγώνων για να οδηγηθούμε σε ταυτότητα:

$$\frac{y}{c} + \left(\frac{b_1}{2c}\right)^2 + \left(\frac{b_2}{2c}\right)^2 = \left[x_1 + \frac{b_1}{2c}\right]^2 + \left[x_2 + \frac{b_2}{2c}\right]^2 \tag{4.12}$$

Η παραπάνω σχέση (4.12) εκφράζει την εξίσωση του κύκλου με μεταβλητή ακτίνα μιας και γνωρίζουμε ότι η εξίσωση του κύκλου με κέντρο (g, h) και ακτίνα r δίνεται από τη σχέση:

$$r^2 = (x_1 - g)^2 + (x_2 - h)^2 \tag{4.13}$$

Επομένως, η εξίσωσή μας είναι κύκλος με κέντρο και ακτίνα αντίστοιχα τα:

$$\left(\frac{-b_1}{2c}, \frac{-b_2}{2c}\right), \sqrt{\frac{y}{c} + \left(\frac{b_1}{2c}\right)^2 + \left(\frac{b_2}{2c}\right)^2} \tag{4.14}$$

με την ακτίνα να μεταβάλλεται ανάλογα την τιμή της y.

Η άλλη περίπτωση που μπορεί να συναντήσουμε στο περίγραμμα των προτιμήσεων εκτός του κύκλου, είναι αυτή των ελλείψεων, όπου οι άξονες είναι παράλληλοι των διαστάσεων του χώρου. Οι μεταβλητές πρόβλεψης είναι οι διαστάσεις καθώς υπάρχει και μια επιπλέον μεταβλητή για κάθε διάσταση, η οποία περιέχει τα τετράγωνα των συντεταγμένων. Πιο συγκεκριμένα η μεταβλητή αυτή δίνεται από την παρακάτω σχέση, η οποία μοιάζει αρκετά με εκείνη που είχαμε συναντήσει νωρίτερα στην περίπτωση του κύκλου:

$$p_i = b_0 + \sum_{a=1}^{r} b_a x_{ia} + \sum_{g=1}^{r} c_g x_{ig}^2 \tag{4.15}$$

Τέλος, η εύρεση στον MDS χώρο, του ιδανικού σημείου που παρουσιάζει το χαρακτηριστικό γνώρισμα, δίνεται από το λόγο:

$$y_{ia} = -\frac{b_a}{2c_a} \tag{4.16}$$

όπου y_{ia} είναι η $i-$στη συντεταγμένη του ιδανικού σημείου στον $r-$διάστατο κλιμακωτό χώρο, με b_a και c_g συμβολίζουμε το συντελεστή παλινδρόμησης και την έξτρα μεταβλητή που περιέχει τα τετράγωνα των συντεταγμένων για την a διάσταση, αντίστοιχα. Η σχέση (4.16) περιγράφεται αναλυτικότερα από την επόμενη παράγραφο και συγκεκριμένα από τις σχέσεις (4.17) έως την (4.21).

★ Παραγωγή του παράλληλου Ελλειπτικού Ιδανικού Σημείου και ο τύπος για τις συντεταγμένες αυτού-Parallel Elliptical Ideal Point Formula for the Coordinates of the Ideal Point ★

Στη περίπτωση της σημειακής παλινδρόμησης παλινδρομούμε τη y πάνω στα x. Αυτό γράφεται ως εξής:

$$y = b_1x_1 + b_2x_2 + c_1x_1^2 + c_2x_2^2 \tag{4.17}$$

αν διαιρέσουμε την προηγούμενη σχέση με c_1c_2 έχουμε:

$$\frac{y}{c_1c_2} = \frac{b_1}{c_1c_2}x_1 + \frac{b_2}{c_1c_2}x_2 + \frac{c_1}{c_1c_2}x_1^2 + \frac{c_2}{c_1c_2}x_2^2 \tag{4.18}$$

ύστερα από την αναγωγή όμοιων όρων, βγάζουμε κοινό παράγοντα από το δεύτερο μέλος το $\frac{1}{c_2}$ και το $\frac{1}{c_1}$, αντίστοιχα και παίρνουμε:

$$\frac{y}{c_1c_2} = \frac{x_1^2 + \frac{b_1}{c_1}x_1}{c_2} + \frac{x_2^2 + \frac{b_2}{c_2}x_2}{c_1} \tag{4.19}$$

μετά από ανάπτυγμα τετραγώνων οδηγούμαστε στην παρακάτω εξίσωση:

$$\frac{y}{c_1c_2} + \left(\frac{b_1}{2c_1}\right)^2 + \left(\frac{b_2}{2c_2}\right)^2 = \frac{\left[x_1 + \frac{b_1}{2c_1}\right]^2}{c_2} + \frac{\left[x_2 + \frac{b_2}{2c_2}\right]^2}{c_1} \tag{4.20}$$

Η σχέση (4.20) εκφράζει την εξίσωση της έλλειψης με κέντρο και ακτίνα:

$$\left(\frac{-b_1}{2c_1}, \frac{-b_2}{2c_2}\right), \sqrt{\frac{y}{c_1c_2} + \left(\frac{b_1}{2c_1}\right)^2 + \left(\frac{b_2}{2c_2}\right)^2} \tag{4.21}$$

μιας και εκείνη όπως γνωρίζουμε στη γενική της μορφή με άξονες παράλληλους στις διαστάσεις και με κέντρο (g, h) και ακτίνα r δίνεται από τη σχέση:

$$r^2 = \frac{(x_1 - g)^2}{a^2} + \frac{(x_2 - h)^2}{b^2} \tag{4.22}$$

με την ακτίνα να κυμαίνεται ανάλογα την τιμή του y.

4.2 Κανονική Ανάλυση Παλινδρόμησης (Canonical Regression Analysis)

Η κανονική ανάλυση παλινδρόμησης (canonical analysis regression) αναπτύχθηκε από τον Bartlet (1938), ως επέκταση της κανονικής ανάλυσης συσχέτισης (canonical correlation analysis) από τον Hotteling (1935,1936). Στην κανονική ανάλυση παλινδρόμησης ένα σύνολο από κοινού εξαρτημένων μεταβλητών τοποθετούνται αριστερά της εξίσωσης ως ένας γραμμικός συνδυασμός, όπως αντίστοιχα συναντάμε το γραμμικό συνδυασμό των παλινδρομούντων μεταβλητών στη δεξιά πλευρά της εξίσωσης. Οπότε σκοπός της ανάλυσης αυτής είναι ο ταυτόχρονος προσδιορισμός της καλύτερης πρόβλεψης γραμμικού συνδυασμού στο δεξιό μέρος και την καλύτερη πρόβλεψη του γραμμικού συνδυασμού για το αριστερό μέρος της εξίσωσης. Η κανονική παλινδρόμηση είναι αποτέλεσμα της κανονικής ανάλυσης συσχέτισης και εφαρμόζεται σε μηδενικού μέσου ή μεταβλητές που τους έχουν αφαιρεθεί ο μέσος. Οπότε μέσω της κανονικής ανάλυσης συσχέτισης θα καταλάβουμε το τρόπο λειτουργείας των μεθόδων αυτών.

4.2.1 Κανονική Ανάλυση Συσχέτισης (Canonical Correlation Analysis)

Γενικότερα όταν έχουμε να κάνουμε με μονομεταβλητά δεδομένα (univariate data) είναι στιγμές που θέλουμε να μετρήσουμε τη γραμμική σχέση μεταξύ πραγμάτων. Η πιο απλή περίπτωση είναι εκείνη που έχουμε δύο μεταβλητές και ενδιαφερόμαστε να μετρήσουμε τη γραμμική τους σχέση. Αυτό που θα πρέπει να κάνουμε είναι η διμεταβλητή συσχέτιση (bivariate correlation). Μια άλλη περίπτωση είναι να έχουμε πολλαπλή παλινδρόμηση, δηλαδή να έχουμε αρκετές ανεξάρτητες μεταβλητές και μία εξαρτημένη μεταβλητή, όπου θα χρησιμοποιήσουμε το συντελεστή πολλαπλής παλινδρόμησης (R^2). Οπότε θα ήταν χρήσιμο να μπορούσαμε να επεκτείνουμε την

ιδέα των παραπάνω περιπτώσεων σε μια άλλη περίπτωση, όπου θα είχαμε πολλές y και x μεταβλητές. Στη περίπτωση της διμεταβλητής συσχέτισης (y, x), η οποία περιγράφει το βαθμό στον οποίο η μία μεταβλητή σχετίζεται (μπορεί να προβλέψη) με την άλλη. Όσο πιο ισχυρή είναι η συσχέτιση τόσο περισσότερα θα γνωρίζουμε για τη y με τη γνώση που έχουμε ήδη για τη x. Από την άλλη η πολυμεταβλητή συσχέτιση, όπου θα έχουμε μια μεταβλητή y και πολλές μεταβλητές x και θα θέλουμε να δούμε το πόσο καλά το σύνολο των μεταβλητών x μπορούν να προβλέψουν τη x, αρκεί να υπολογίσουμε την ευθεία παλινδρόμησης. Οπότε μέσω της ευθείας παλινδρόμησης μπορούμε να υπολογίσουμε την εκτιμημένη $\hat{y}$ και να τη συγκρίνουμε με τη y, ώστε να υπολογίσουμε την απλή συσχέτιση μιας και έχουμε δύο μεταβλητές τις y και $\hat{y}$. Η κανονική συσχέτιση όμως ενδιαφέρεται για τη σχέση που υπάρχει μεταξύ πολλών μεταβλητών x και y και επομένως, αφού έχουμε ένα σύνολο μεταβλητών x και y τότε δεν είμαστε σε θέση να υπολογίσουμε την απλή συσχέτιση. Έτσι είναι κατανοητό ότι θα πρέπει να ορίσουμε δύο γραμμικούς συνδυασμούς έναν για το σύνολο των x μεταβλητών (b_1) και έναν για το σύνολο $y(a_1)$. Η πρώτη κανονική συσχέτιση περιγράφει τη συσχέτιση μεταξύ αυτών των δύο νέων μεταβλητών (b_1x και a_1y). Ο τρόπος με τον οποίο επιλέγουμε τους γραμμικούς συνδυασμούς (b_1, a_1) είναι τέτοιος ώστε η συσχέτιση αυτών των δύο νέων μεταβλητών να είναι μέγιστη, όπως στην περίπτωση της πολλαπλής παλινδρόμησης. Βέβαια, θα πρέπει να έχουμε στο νού μας ότι, δεν αρκεί ένας γραμμικός συνδυασμός για να αποτυπώσει όλη την πληροφορία που παρουσιάζουν τα δεδομένα μας και έτσι θα πρέπει να αναλογιστούμε πόσοι γραμμικοί συνδυασμοί απαιτούνται, όπου μεταξύ τους θα είναι ασυσχέτιστοι, για να πάρουμε περισσότερη πληροφορία. Επομένως αυτό που πρέπει να κάνουμε είναι να ορίσουμε περισσότερα σύνολα γραμμικών συνδυασμών, b_i και a_i, με $i = 1, \ldots, n$ με $n = \min(p, q)$ όπου p, q είναι ο αριθμός των μεταβλητών στα σύνολα x και y, αντίστοιχα. Επομένως καθένας από τους γραμμικούς συνδυασμούς αυτό που κάνει είναι να μεγιστοποιεί τη συσχέτιση μεταξύ των νέων μεταβλητών υπό τον περιορισμό ότι είναι ασυσχέτιστοι με όλους τους προηγούμενους. Όσο αφορά τον τρόπο υπολογισμού των κανονικών συσχετίσεων αρχικά έχουμε στο νού μας τον αμέσως επόμενο πίνακα συνδιακυμάνσεων.

	x_1 x_2 ... x_n	y
x_1 x_2 ⋮ x_n	$\mathbf{C_{xx}}$	$\mathbf{C_{xy}}$
y	$\mathbf{C_{xy}}'$	$\mathbf{C_{yy}}$

Από αυτόν τον πίνακα ορίζουμε τέσσερις υποπίνακες, όπου θα προκύψουν τα ζητούμενα. Αρχικά μπορούμε να ορίσουμε την πολλαπλή τετραγωνική συσχέτιση (R^2) ως γινόμενο τεσσάρων πινάκων συσχετίσεων, δηλαδή του πίνακα εξαρτημένων μεταβλητών $(\mathbf{C_{yy}^{-1}})$, των ανεξάρτητων μεταβλητών πίνακα $(\mathbf{C_{xx}^{-1}})$ και το γινόμενο των εξαρτημένων και ανεξάρτητων μεταβλητών πινάκων $(\mathbf{C_{yx}}, \mathbf{C_{xy}})$. Δηλαδή το προηγούμεο γράφεται:

$$R_M^2 = \mathbf{C_{yy}^{-1}} \mathbf{C_{yx}} \mathbf{C_{xx}^{-1}} \mathbf{C_{xy}} \tag{4.23}$$

Στη συνέχεια υπολογίζουμε την τετραγωνική ρίζα των ιδιοτιμών $(r_1, r_2, \ldots, r_n)$ και των ιδιοδιανυσμάτων $(a_1, a_2, \ldots, a_n)$ της σχέσης (4.23). Ομοίως υπολογίζουμε την τετραγωνική ρίζα των ιδιοτιμών $(r_1, r_2, \ldots, r_n)$ και των ιδιοδιανυσμάτων $(b_1, b_2, \ldots, b_n)$ της σχέσης:

$$R_M^2 = \mathbf{C_{xx}^{-1}} \mathbf{C_{xy}} \mathbf{C_{yy}^{-1}} \mathbf{C_{yx}} \tag{4.24}$$

Ιδανικά οι ιδιοτιμές για τις σχέσεις (4.23) και (4.24) είναι ίσες και μάλιστα μεταξύ του μηδενός και της μονάδας. Τέλος, από τα ιδιοδιανύσματα έχουν προσδιορισθεί οι γραμμικοί μετασχηματισμοί για τους νέους γραμμικούς συνδυασμούς.

4.3 Παλινδρόμηση Κυρίων Συνιστωσών - Principal Components Regression (PCR)

Η ανάλυση κυρίων συνιστωσών είναι μια τεχνική η οποία έχει αφετηρία από τους Pearson και Hotteling 1901 και 1933, αντίστοιχα. Στην παλινδρόμηση αυτή, πρώτα εφαρμόζουμε ανάλυση κύριων συνιστωσών στα

πραγματικά δεδομένα και στη συνέχεια εφαρμόζουμε μείωση διαστάσεων με το να επιλέξουμε m αριθμό κύριων συνιστωσών κάνοντας χρήση cross-validation και τέλος να εφαρμόσουμε παλινδρόμηση χρησιμοποιόντας τις m αυτές διαστάσεις. Με άλλα λόγια η ανάλυση αυτή είναι μια παλιά τεχνική μείωσης δεδομένων με απώτερο ενδιαφέρον το χώρο που ορίζει η συνολική διακύμανση των μεταβλητών. Οι αρχικές μεταβλητές μειώνονται σε ένα μικρότερο αριθμό συνιστωσών, λαμβάνοντας υπόψη τη διακύμανση για τις μεταβλητές στο σύνολό τους. Το πρόβλημα με τη μέθοδο αυτή είναι ο μετασχηματισμός των αρχικών μεταβλητών σε νέες μεταβλητές (συνιστώσες), οι οποίες θα ερμηνεύουν όσο γίνεται περισσότερο πιθανό τις παρατηρούμενες μεταβλητές. Με άλλα λόγια, στόχος μας είναι η όσο το δυνατόν σε μεγαλύτερο εύρος επεξήγησης της συστηματικής διακύμανσης των παρατηρηθέντων μεταβλητών στα δεδομένα μας, με όσο το δυνατό λιγότερες συνιστώσες. Έτσι με τη μέθοδο αυτή, θα έχουμε καταφέρει να παράξουμε ένα σύνολο συνιστωσών από τα αρχικά μας δεδομένα, για τα οποία θα είναι από κοινού ασυσχέτιστα και οι διακυμάνσεις τους μέγιστες. Δηλαδή η πρώτη συνιστώσα θα έχει τη μεγαλύτερη διακύμανση, η δεύτερη συνιστώσα θα έχει πιθανότατα τη μεγαλύτερη διακύμανση μεταξύ όλων των υπολοίπων συνιστωσών, οι οποίες είναι ασυσχέτιστες με την πρώτη κ.ο.κ. Το άθροισμα των διακυμάνσεων όλων των κύριων συνιστωσών ισούται με το άθροισμα των διακυμάνσεων των αρχικών μεταβλητών.

4.4 Η μέθοδος της Ανάλυσης Κυρίων Συνιστωσών - PCA Procedure

Έστω $\mathbf{X}$ ένα τυχαίο διάνυσμα διάστασης p:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} \tag{4.25}$$

με πίνακα διακύμανσης - συνδιακύμανσης στον πληθυσμό:

$$Var(\mathbf{X}) = \mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{pmatrix} \tag{4.26}$$

ο οποίος πίνακας είναι τετραγωνικός και συμμετρικός. Έστω οι γραμμικοί συνδυασμοί:

$$Y_1 = e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p$$

$$Y_2 = e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p$$

$$\vdots$$

$$Y_p = e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p$$

Τις παραπάνω εξισώσεις μπορούμε να τις θεωρούμε ως γραμμική παλινδρόμηση της Y_i πάνω στα X_i, για $i = 1, 2, \dots, p$, χωρίς τομή και τα στοιχεία $e_{i1}, e_{i2}, \dots, e_{ip}$ μπορεί να θεωρηθούν ως συντελεστές παλινδρόμησης. Επιπλέον θα πρέπει να λάβουμε υπόψη ότι η Y_i είναι τυχαία μεταβλητή, μιας και είναι συνάρτηση τυχαίων μεταβλητών. Η πληθυσμιακή διακύμανση θα είναι:

$$Var(Y_i) = \sum_{k=1}^{p}\sum_{l=1}^{p} e_{ik}e_{il}\sigma_{kl} \tag{4.27}$$

ή σε μορφή πινάκων:

$$Var(Y_i) = \underset{\sim}{\boldsymbol{e_i}}' \sum \underset{\sim}{\boldsymbol{e_i}} \tag{4.28}$$

καθώς και πληθυσμιακή συνδιακύμανση:

$$Cov(Y_i, Y_j) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{ik} e_{il} \sigma_{kl} \tag{4.29}$$

ή σε μορφή πινάκων:

$$Cov(Y_i, Y_j) = \underset{\sim}{\boldsymbol{e_i}}' \sum \underset{\sim}{\boldsymbol{e_j}} \tag{4.30}$$

με $\underset{\sim}{\boldsymbol{e_i}}$ να είναι το διάνυσμα εκείνο που περιέχει τους συντελεστές, δηλαδή:

$$\underset{\sim}{\boldsymbol{e_i}} = \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{ip} \end{bmatrix} \tag{4.31}$$

★ Εξαγωγή 1ης κύριας συνιστώσας (PCA1): Y_1 ★

Η 1η κύρια συνιστώσα είναι ο γραμμικός συνδυασμός των x μεταβλητών, οι οποίες έχουν τη μέγιστη δυνατή διακύμανση μεταξύ όλων των γραμμικών συνδυασμών με σκοπό να αποτυπώνει όσο είναι περισσότερο δυνατό, τη μέγιστη διακύμανση των δεδομένων. Πιο συγκεκριμένα ορίζουμε τους συντελεστές $e_{i1}, e_{i2}, \dots e_{ip}$ για τη συνιστώσα κατά τέτοιον τρόπο ώστε να μεγιστοποιείται η διακύμανση, με την προυπόθεση ότι να ισχύει ένας περιορισμός. Ο περιορισμός αυτός μας λέει ότι το άθροισμα των τετραγώνων των συντελεστών ισούται με τη μονάδα. Ο λόγος για τον οποίο θέτουμε αυτόν τον περιορισμό είναι για να πάρουμε μοναδική λύση. Ο προηγούμενος περιορισμός μπορεί να περιγραφεί ως εξής, να επιλέξουμε εκείνα τα $e_{i1}, e_{i2}, \dots e_{ip}$, για τα οποία μεγιστοποιείται η διακύμανση:

$$Var(Y_1) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{1k} e_{1l} \sigma_{kl} = \underset{\sim}{\boldsymbol{e_1}}' \underline{\sum} \underset{\sim}{\boldsymbol{e_1}} \tag{4.32}$$

υπό τον περιορισμό:

$$e_1' e_1 = \sum_{j=1}^{p} e_{1j}^2 = 1. \tag{4.33}$$

★ Εξαγωγή 2ης κύριας συνιστώσας (PCA2): Y_2 ★

Στη συνέχεια η 2η κύρια συνιστώσα είναι εκείνος ο γραμμικός συνδυασμός των x μεταβλητών, ο οποίος αποτυπώνει όσο γίνεται περισσότερη εναπομένουσα διακύμανση, με τον περιορισμό ότι η συσχέτιση μεταξύ της 1ης συνιστώσας και της 2ης, θα είναι μηδέν. Με άλλα λόγια, επιλέγουμε $e_{21}, e_{22}, \dots, e_{2p}$ για τα οποία μεγιστοποιείται η διακύμανση της νέας αυτής συνιστώσας:

$$Var(Y_2) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{2k} e_{2l} \sigma_{kl} = \underset{\sim}{\boldsymbol{e_2}}' \underline{\sum} \underset{\sim}{\boldsymbol{e_2}} \tag{4.34}$$

υπό τον αντίστοιχο περιορισμό της 1ης συνιστώσας, δηλαδή το άθροισμα των τετραγώνων των συντελεστών να ισούται με τη μονάδα. Δηλαδή $e_2' e_2 = \sum_{j=1}^{p} e_{2j}^2 = 1$. Φυσικά, στην περίπτωση της 2ης συνιστώσας, θα ισχύει και ακόμη ένας περιορισμός, αυτός της ασυσχέτισης, δηλαδή θα πρέπει οι δύο συνιστώσες να είναι ασυχέτιστες μεταξύ τους και έτσι γράφουμε.

$$Cov(Y_1, Y_2) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{1k} e_{2l} \sigma_{kl} = \underset{\sim}{\boldsymbol{e_1}}' \underline{\sum} \underset{\sim}{\boldsymbol{e_2}} = 0. \tag{4.35}$$

Φυσικά και όλες οι υπόλοιπες συνιστώσες που θα ακολουθήσουν, θα έχουν την ίδια ιδιότητα, τόσο για τη μέγιστη πιθανή αποτύπωση της εναπομένουσας διακύμανσης όσο και για τη μεταξύ τους ασυσχετικότητα. Επομένως στη γενικότερη αυτή περίπτωση έχουμε:
επιλέγουμε εκείνα τα $e_{i1}, e_{i2}, \ldots, e_{ip}$ για τα οποία μεγιστοποιούν την Y διακύμανση, δηλαδή:

$$Var(Y_i) = \sum_{k=1}^{p}\sum_{l=1}^{p} e_{ik}e_{il}\sigma_{kl} = \underset{\sim}{\boldsymbol{e_i}}'\underline{\sum}\underset{\sim}{\boldsymbol{e_i}} \tag{4.36}$$

υπό τον περιορισμό ότι το άθροισμα τετραγώνων των συντελεστών θα αθροίζονται στη μονάδα καθώς και στον έτερο περιορισμό που αναφέρονται στη μεταξύ τους ασυσχετικότητα όλων των προηγηθέντων συνιστωσών. Οπότε στη γενική μορφή για τη συνιστώσα Y_i, γράφουμε:

$$\begin{aligned}
e_i'e_i &= \sum_{j=1}^{p} e_{ij}^2 = 1 \\
Cov(Y_1, Y_i) &= \sum_{k=1}^{p}\sum_{l=1}^{p} e_{1k}e_{il}\sigma_{kl} = \underset{\sim}{\boldsymbol{e_1}}'\underline{\sum}\underset{\sim}{\boldsymbol{e_i}} = 0 \\
Cov(Y_2, Y_i) &= \sum_{k=1}^{p}\sum_{l=1}^{p} e_{2k}2_{2l}\sigma_{kl} = \underset{\sim}{\boldsymbol{e_2}}'\underline{\sum}\underset{\sim}{\boldsymbol{e_i}} = 0 \\
&\vdots \\
Cov(Y_{i-1}, Y_i) &= \sum_{k=1}^{p}\sum_{l=1}^{p} e_{i-1,k}e_{il}\sigma_{kl} = \underset{\sim}{\boldsymbol{e_{i-1}}}'\underline{\sum}\underset{\sim}{\boldsymbol{e_i}} = 0
\end{aligned} \tag{4.37}$$

Πράγμα που σημαίνει ότι όλες οι κύριες συνιστώσες είναι ασυσχέτιστες μεταξύ τους.

4.5 Ανάλυση πλεονασμού - Redundancy Analysis (RDA)

Η μέθοδος RDA είναι κατάλληλη για καταστάσεις στις οποίες έχουμε δύο σύνολα πολλαπλών μεταβλητών. Επιπλέον, η μέθοδος αυτή μας βοηθά στο να καταλάβουμε τι μέρος της μεταβλητότητας του ενός συνόλου μεταβλητών, επεξηγείται από κάποιο άλλο σύνολο μεταβλητών και είναι ανάλογο της απλής γραμμικής παλινδρόμησης. Επιπλέον, η κύρια ιδέα πίσω από τη μέθοδο αυτή είναι να εφαρμόσουμε γραμμική παλινδρόμηση έτσι ώστε να παρουσιάσουμε το Y ως γραμμική συνάρτηση του X και στη συνέχεια να κάνουμε χρήση της μεθόδου PCA με σκοπό να παρουσιάσουμε το αποτέλεσμα. Έτσι ο τρόπος «σκέψης», είναι ότι μεταξύ ποιων συνιστωσών του Y, μπορεί να ερμηνευθεί καλύτερα από τα X και επέλεξε εκείνες τις συνιστώσες οι οποίες παρουσιάζουν τη μεγαλύτερη διακύμανση.

5 Μη Σταθμισμένα Μοντέλα Απόστασης-Unweighted Distance Models

Όταν αναφερόμαστε στα μη σταθμισμένα μοντέλα απόστασης, εννοούμε τα απλά μοντέλα απόστασης και αναφέρονται έτσι για να διακρίνονται από τα σταθμισμένα που είναι πιο σύνθετα. Τόσο τα μη σταθμισμένα όσο και τα σταθμισμένα μοντέλα απόστασης, είναι αλγεβρικές εξισώσεις με γεωμετρική παρουσίαση, που στη μεν περίπτωση της μη σταθμισμένης είναι λιγότερο σύνθετη τόσο το αλγεβρικό κομμάτι όσο και η γεωμετρική παρουσίαση, σε σχέση με το σταθμισμένο. Και στις δύο περιπτώσεις χρησιμοποιείται το μοντέλο της πολυδιαστατικής κλιμάκωσης (Multidimensional Scaling-MDS), δηλαδή ένα μοντέλο το οποίο περιγράφεται τόσο από μια αλγεβρική εξίσωση, όσο και από τη γεωμετρική του παρουσίαση. Ο λόγος για τον οποίο ακολουθούμε αυτή την ανάλυση, είναι επειδή γίνεται πιο εύκολη η παρατήρηση και η κατανόηση των δεδομένων ως μια εικόνα, παρά ως σκέτα δεδομένα. Αξίζει να αναφερθεί, ότι δε θα πρέπει να συγχαίουμε την έννοια του MDS μοντέλου και αυτή της MDS ανάλυσης. Έτσι όταν αναφερόμαστε για μοντέλο, εννοούμε την αλγεβρική εξίσωση μόνο και τίποτα δεν αναφέρουμε για τα δεδομένα. Όταν μιλάμε για ανάλυση, τότε εφαρμόζουμε το μοντέλο αυτό επάνω στα δεδομένα με σκοπό να τα καταλάβουμε καλύτερα αυτά.

5.1 Ειδικές μορφές MDS Minkowski μοντέλα

Ένα από τα γενικότερα μοντέλα είναι αυτό του Minkowski, του οποίου η σχέση περιγράφεται από:

$$d_{ij}^p = \sum_{a=1}^{r} |x_{ia} - x_{ja}|^p, \quad p \geq 1, \quad x_i \neq x_j \tag{5.1}$$

όπου r είναι οι διαστάσεις, x_{ia} συμβολίζουμε τη συντεταγμένη του σημείου i στη διάσταση a και τέλος το x_i είναι το r-οστό στοιχείο του διανύσματος γραμμής, της i-οστής γραμμής του πίνακα $\mathbf{X_{n\times r}}$, ο οποίος περιέχει τις συντεταγμένες x_{ia} όλων των n σημείων και όλων των r διαστάσεων. Η δύναμη p συνήθως αναφέρεται ως δύναμη Minkowski και μπορεί να δέχεται τιμές όχι μικρότερες της μονάδας. Βέβαια, τριών ειδών μοντέλα είναι τα επικρατέστερα από άποψη ενδιαφέροντος στο χώρο της στατιστικής επιστήμης, ανάλογα τον αριθμό που λαμβάνει ο εκθέτης p. Έτσι το πιο δημοφιλές μοντέλο είναι αυτό το οποίο ο εκθέτης λαμβάνει τον αριθμό 2, το οποίο λέγεται Ευκλείδειο μοντέλο και δίνεται από τον παρακάτω τύπο:

$$d_{ij}^2 = \sum_{a=1}^{r} (x_{ia} - x_{ja})^2 \tag{5.2}$$

Έτσι, η σχέση (5.2), είναι η r-διάστατη έκδοση της απόστασης d_{ij}, η οποία εκφράζεται ως την τετραγωνική ρίζα του αθροίσματος των τετραγωνικών διαφορών μεταξύ συντεταγμένων. Επίσης, στην περίπτωση αυτή είναι πιο εύκολη η ερμηνεία από το ανθρώπινο μάτι, μιας και είναι μικρότερη των τεσσάρων διαστάσεων καθώς και όταν οι ερευνητές δε γνωρίζουν την διαδικασία κατά την οποία παρήχθησαν τα δεδομένα που έχουν στη διάθεσή τους και γι᾽ αυτό άλλωστε είναι και το πιο διαδεδομένο μοντέλο με τη μέθοδο MDS. Στην περίπτωση που ο εκθέτης p, πάρει την τιμή 1, τότε μιλάμε για το *city block*, μοντέλο το οποίο δίνεται από την παρακάτω σχέση:

$$d_{ij} = \sum_{a=1}^{r} |x_{ia} - x_{ja}| \tag{5.3}$$

και είναι το άθροισμα των απολύτων διαφορών των συντεταγμένων, μεταξύ δύο σημείων. Τέλος, η τρίτη κατηγορία που συναντάται συχνά, είναι το κυρίαρχο μοντέλο (dominance model), όπου το p λαμβάνει απείρως μεγάλες τιμές και το οποίο δίνεται από τη σχέση:

$$d_{ij} = \max_{a}^{r} |x_{ia} - x_{ja}| \tag{5.4}$$

με το $\max_{a}^{r}$, να προσδιορίζει τη μέγιστη απόλυτη διαφορά, ως προς όλες τις διαστάσεις $a = 1, \ldots, r$.

5.2 Μη σταθμισμένη ανάλυση

Στην περίπτωση των μη σταθμισμένων μοντέλων διακρίνουμε τέσσερις διαφορετικούς τύπους ανάλυσης, οι οποίοι βασίζονται στο Minkowski μοντέλο. Το συγκεκριμένο μοντέλο, το οποίο πρωτάθηκε από τον Torgerson το 1952 [2], χρησιμοποιεί τα δεδομένα που υπάρχουν σε έναν τετραγωνικό και υπό συνθήκη πίνακα.[6]Ο πρώτος από τους τέσσερις τύπους MDS ανάλυσης, ονομάζεται κλασικός MDS, (CMDS), από τους Schiffman et al. και χρησιμοποιεί το μοντέλο Minkowski, για να περιγράψει την πληροφορία σε έναν πίνακα που περιέχει τα τετράγωνα των δεδομένων και λέγεται υπό συνθήκη πίνακας. Έστω ότι έχουμε έναν τετραγωνικό και συμμετρικό πίνακα δεδομένων. Ο παραπάνω τύπος ανάλυσης, παρουσιάζει τα χαρακτηριστικά γνωρίσματα ως σημεία σε ένα Minkowski χώρο, όπου τόσο η διαστατικότητα όσο και ο εκθέτης Minkowski προσδιορίζονται πριν ακολουθήσει η ανάλυση. Στην ουσία με αυτόν τον τύπο ανάλυσης, αποτυπώνονται τα σημεία στο χώρο κατά τέτοιον τρόπο, ώστε οι αποστάσεις να αντιστοιχούν όσο γίνεται περισσότερο στα δεδομένα. Για παράδειγμα, εάν ένα χαρακτηριστικό γνώρισμα έχει χαρακτηρισθεί ως αρκετά όμοιο (π.χ. το μαρούλι με το σπανάκι), τότε θα πρέπει τα σημεία για αυτό το ζεύγος να βρίσκονται αρκετά κοντά ενώ στην αντίθετη περίπτωση (π.χ. η πατάτα με τον τόνο), θα πρέπει να βρίσκονται αρκετά μακρυά. Ο δεύτερος τύπος ανάλυσης προσδιορίσθηκε από τους Schiffman et al. και καλείται αναπαραγόμενος MDS, (RMDS). Στην ουσία αυτός ο τύπος ανάλυσης είναι όμοιος με εκείνον της κλασσικής, αλλά η διαφορά έγκειται ως προς τον αριθμό των πινάκων. Έτσι η ανάλυση αυτή, κάνοντας χρήση του μοντέλου Minkowski, προσπαθεί και περιγράφει την πληροφορία σε αρκετούς τετραγωνικούς και υπό συνθήκη πίνακες. Στη συνέχεια ο τρίτος τύπος ανάλυσης καλείται Classical Multidimensional Unfolding (CMDU) και πρωτάθηκε από τον Coombs το 1964 [2], χρησιμοποιώντας το μοντέλο Minkowski, για να περιγράψει την πληροφορία σε έναν πίνακα τετραγωνικών δεδομένων, ο οποίος είναι δεσμευμένος ανά γραμμή. Αυτός ο τύπος ανάλυσης παρουσιάζει από κοινού τόσο τα ερεθίσματα όσο και τα αντικείμενα ως σημεία του χώρου Minkowski. Επομένως, στην περίπτωση της ανάλυσης αυτής, αυτό που κάνει είναι να εντοπίζει εκείνα τα σημεία, έτσι ώστε η απόσταση μεταξύ δύο σημείων ενός συνόλου, να αντικατοπτρίζει την προτίμηση των ατόμων, όσο περισσότερο γίνεται. Έτσι έχουμε, την από κοινού απεικόνιση τόσο των σημείων των ερεθισμάτων, όσο και των αντικειμένων και όσο πιο κοντά βρίσκεται ένα άτομο σε αυτά τα σημεία, τόσο περισσότερο εκφράζει την προτίμησή του, ενώ όσο απομακρύνεται από αυτά συμβαίνει το αντίθετο. Τέλος, ο τέταρτος τύπος ανάλυσης είναι ο επαναλαμβανόμενος MDU (RMDU) και πρωτάθηκε από τους Young και Lewyckyj (1979a) [2]. Η ανάλυση αυτού του τύπου, δεν έχει και πολύ μεγάλη διαφορά από την ανάλυση (CMDU), απλά στην προκειμένη περίπτωση έχουμε περισσότερους ορθογώνιους πίνακες και δεδομένα τα οποία είναι δεσμευμένα ανά γραμμή.

5.3 Οι χώροι Minkowski είναι μετρικοί

Στη συνέχεια γίνεται μια αναφορά στο ότι ο χώρος Minkowski, είναι μετρικός. Όπως έχουμε αναφέρει για να είναι μια συνάρτηση απόστασης, θα πρέπει να πληρεί τα τέσσερα αξιώματα που είχαν αναφερθεί στον ορισμό 3.1. Περιληπτικά το πρώτο αξίωμα μας λέει ότι η απόσταση μεταξύ δύο διακριτών σημείων, είναι διάφορη του μηδενός καθώς και η απόσταση μεταξύ ενός σημείου και του εαυτού του, πρέπει να είναι μηδέν. Φυσικά είναι ξεκάθαρο, ότι η εξίσωση (5.1) είναι μονίμως μηδέν για $i = j$, μιας και διαπιστώνουμε ότι οι διαφορές μέσα στα αθροίσματα θα είναι πάντα μηδέν, ενώ για $i \neq j$ ή $x_i \neq x_j$ συμβαίνει το αντίθετο. Ένα από τα αξιώματα που χρήζει ιδιαίτερης προσοχής από την άποψη του μέτρου, είναι το τέταρτο αξίωμα, το οποίο είναι η τριγωνική ανισότητα και θέλει περαιτέρω διερεύνηση. Έχει αποδειχθεί ότι, η ανισότητα αυτή ισχύει για όλους τους χώρους Minkowski και έτσι έχουμε:

$$\left(\sum_{a}^{r}|p_a + q_a|^p\right)^{1/p} \leq \left(\sum_{a}^{r}|p_a|^p\right)^{1/p} + \left(\sum_{a}^{r}|q_a|^p\right)^{1/p} \tag{5.5}$$

για $p \geq 1$. Αν θέσουμε

$$p_a = (x_{ia} - x_{ka}) \tag{5.6}$$

[6]Ένας πίνακας θα λέγεται υπό συνθήκη (conditional matrix), ανάλογα με βάση πιο ερέθισμα αποτιμάται από τα άτομα κάθε φορά.

και

$$q_a = (x_{ka} - x_{ja}) \tag{5.7}$$

τότε έχουμε:

$$\left(\sum_a^r |x_{ia} - x_{ka} + x_{ka} - x_{ja}|\right)^{1/p} \leq \left(\sum_a^r |x_{ia} - x_{ka}|^p\right)^{1/p} + \left(\sum_a^r |x_{ka} - x_{ja}|^p\right)^{1/p} \tag{5.8}$$

ή σε πιο απλοποιημένη μορφή:

$$d_{ij} \leq d_{ik} + d_{kj} \tag{5.9}$$

Οπότε, δοθέντος ότι η ανισότητα Minkowski ισχύει, έπεται ότι όλοι οι χώροι Minkowski ικανοποιούν το αξίωμα 4 που δίνεται στον ορισμό 3.1.

★Τρεις σημαντικές ανισότητες★

Λήμμα 5.1. *Έστω* $0 < \lambda < 1$, *τότε*

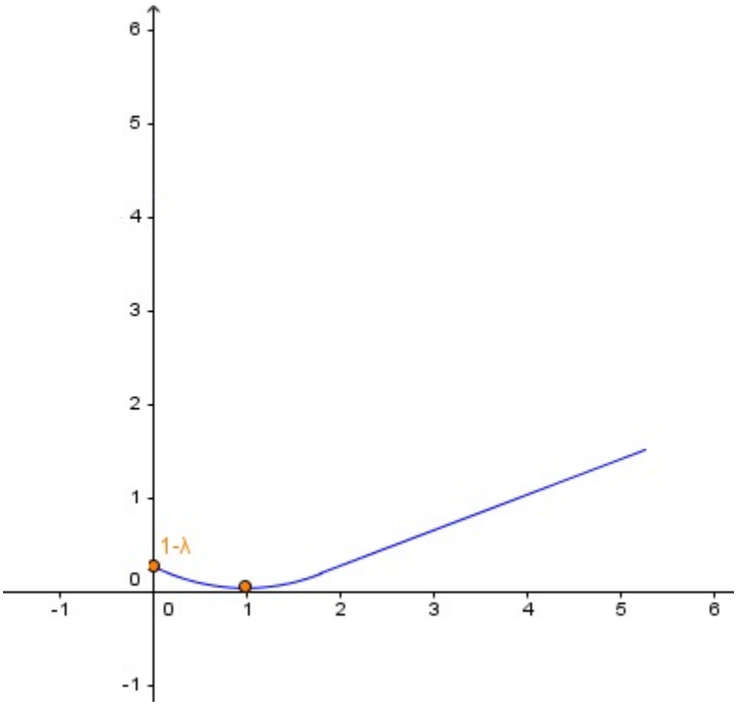
$$t^\lambda \leq 1 - \lambda + \lambda t \quad for \quad all \quad t \geq 0 \tag{5.10}$$

για την προηγούμενη ανισότητα, μετατρέπεται σε ισότητα όταν το $t = 1$.

Απόδειξη. Ορίζουμε την $f : [0, \infty] \to \mathbb{R}$ ως εξής:

$$f(t) := 1 - \lambda + \lambda t - t^\lambda \tag{5.11}$$

το γράφημα το οποίο απεικονίζεται στο Σχήμα 9.



Σχήμα 9: Γράφημα της $f(t) := 1 - \lambda + \lambda t - t^\lambda$.

Παίρνοντας την παράγωγο ως προς t, λαμβάνουμε:

$$f'(t) = \lambda - \lambda t^{\lambda - 1} = \lambda \left(1 - \frac{1}{t^{1-\lambda}}\right) \tag{5.12}$$

Οπότε:

$$f'(t) = \begin{cases} > 0 & \text{αν} \quad t > 1 \\ < 0 & \text{αν} \quad 0 < t < 1 \end{cases} \tag{5.13}$$

με ελάχιστη τιμή το μηδέν, για $t = 1$. Συμπερασματικά το $0 = f(1)$, είναι η ελάχιστη τιμή της f. Έκτοτε, $f(t) \geq 0$ για όλα τα $t \geq 0$ και με την ισότητα να ισχύει ανν $t = 1$. Οπότε έχουμε:

$$\begin{aligned} t^\lambda &\leq 1 - \lambda + \lambda t \quad \text{για} \quad \text{όλα τα} \quad t \geq 0 \\ t^\lambda &= 1 - \lambda + \lambda t \qquad \text{ανν} \qquad t = 1 \end{aligned} \tag{5.14}$$

■

Ορισμός 5.1. *(Συζυγής εκθέτης) Οι θετικοί πραγματικοί αριθμοί p, q για τους οποίους ισχύει:*

$$\frac{1}{p} + \frac{1}{q} = 1 \tag{5.15}$$

θα λέγονται συζυγείς εκθέτες. Το ζευγάρι 1,∞ θεωρείται εξίσου ως ζευγάρι συζυγών εκθετών, διότι καθώς το $p \to 1$ συνεπάγεται ότι $q \to \infty$. Στην περίπτωση που p, q είναι ακέραιοι, το μόνο ζευγάρι συζυγών εκθετών είναι το 2,2.

Λήμμα 5.2. *Έστω ότι p, q είναι συζυγείς εκθέτες με $1 < q < \infty$. Τότε*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q} \quad , \forall \quad a, b \geq 0 \tag{5.16}$$

Η ισότητα ισχύει ανν $a^p = b^q$.

Παρακάτω αναπτύσσονται δύο ακόμη αποδείξεις όπου το αξίωμα 4 πού αναφέρεται στον ορισμό 3.1, ισχύει για τους χώρους Minkowski, στην περίπτωση που οι εκθέτες λάβουν τη μονάδα και το ∞. Πιο συγκεκριμένα, όταν το $p = 1$ (χώρος city-block), τότε έχουμε την εξίσωση (5.3), και αν εισάγουμε ένα τρίτο σημείο x_k λαμβάνουμε:

$$d_{ij} = \sum_a^r |x_{ia} - x_{ka} + x_{ka} - x_{ja}| \tag{5.17}$$

Στην συνέχεια έχουμε:

$$d_{ij} = |x_{it} - x_{jt}| = |x_{it} - x_{kt} + x_{kt} - x_{jt}| \leq |x_{it} - x_{kt}| + |x_{kt} - x_{jt}| \tag{5.18}$$

η οποία είναι η επιθυμητή τριγωνική ανισότητα. Άρα με το προηγούμενο γίνεται αντιληπτό, ότι το αξίωμα 4 του ορισμού 3.1, ισχύει για τον $r-$διάστατο χώρο city-block. Επίσης, έχει αποδειχθεί, ότι η τριγωνική ανισότητα ισχύει και για τους dominance χώρους, όπως ορίζεται από την σχέση (5.4). Ας υποθέσουμε ότι η απόλυτη διαφορά για τις συντεταγμένες των σημείων x_i και x_j είναι μεγαλύτερη, όταν προβάλεται στη διάσταση t. Πιο συγκεκριμένα:

$$d_{ij} = \max_a^r |x_{ia} - x_{ja}| = |x_{it} - x_{jt}| \tag{5.19}$$

Στη συνέχεια έχουμε:

$$d_{ij} = |x_{it} - x_{jt}| = |x_{it} - x_{kt} + x_{kt} - x_{jt}| \leq |x_{it} - x_{kt}| + |x_{kt} - x_{jt}| \tag{5.20}$$

επίσης είναι σαφή τα παρακάτω:

$$|x_{it} - x_{kt}| \leq \max_{a}^{r} |x_{ia} - x_{ka}| \leq d_{ik} \tag{5.21}$$

$$|x_{kt} - x_{jt}| \leq \max_{a}^{r} |x_{ka} - x_{ja}| \leq d_{jk} \tag{5.22}$$

Αν αντικαταστήσουμε τις σχέσεις (5.22) και (5.21) στη σχέση (5.20), έπεται ότι:

$$d_{ij} \leq d_{ik} + d_{jk} \tag{5.23}$$

η οποία είναι η επιθυμητή τριγωνική ανισότητα. Επομένως, ο $r-$διάστατος κυρίαρχος χώρος (dominance space), ικανοποιεί το αξίωμα 4. Επίσης, το αξίωμα 2, μας λέει ότι αν μια συνάρτηση καλείται συνάρτηση απόστασης, τότε δε μπορεί η απόσταση αυτή μεταξύ των δύο σημείων, να είναι αρνητική. Αυτή η ιδιότητα βέβαια, αποτυπώνεται στους χώρους Minkowski, μιας και από την εξίσωση (5.1), παρατηρούμε ότι πρόκειται για τη θετική ρίζα του αθροίσματος μη αρνητικών ορισμάτων. Επιπλέον, το αξίωμα 3 (συμμετρικότητα), ισχύει μιας και να αλλάξουμε τη σειρά των δεικτών δεν επηρεάζεται η τιμή των απόλυτων διαφορών. Έτσι οδηγήθηκαν στο συμπέρασμα ότι και τα τέσσερα αξιώματα του μέτρου ικανοποιούνται από τους χώρους Minkowski και επομένως οι χώροι αυτοί, είναι μετρικοί.

5.4 Γεωμετρικές ιδιότητες

Το πιο βασικό μοντέλο αποστάσεων από γεωμετρικής άποψης, είναι το Ευκλείδειο μοντέλο το οποίο δίνεται από τη σχέση:

$$d_{ij}^2 = \sum_{a}^{r} |x_{ia} - x_{ja}|^2 \tag{5.24}$$

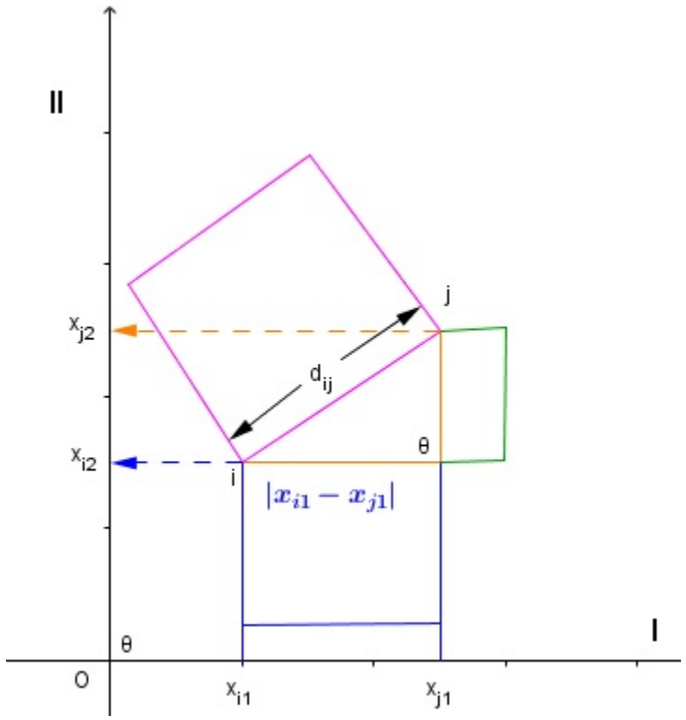
και η οποία μας λέει ότι, η τετραγωνική απόσταση μεταξύ δύο σημείων είναι απλά το άθροισμα των τετραγωνικών διαφορών των συντεταγμένων των διανυσμάτων αυτών, όπως αποικονίζεται στο Σχήμα[7] 10.

Από το Σχήμα 10 παρατηρούμε, ότι αν πάρουμε τα τετράγωνα και αθροίσουμε αυτές τις δύο πλευρές $|x_{i1} - x_{j1}|$ και $|x_{i2} - x_{j2}|$(δηλαδή εφαρμόσουμε το Πυθαγόρειο θεώρημα), αυτό που λαμβάνουμε είναι το τετράγωνο της υποτίνουσας, δηλαδή την απόσταση d_{ij}. Έχει αποδειχθεί, ότι οι σχέσεις διατηρούνται ανεξαρτήτως τον προσανατολισμό (orientation) και τη θέση των αξόνων (translated). Δηλαδή αν τις διαστάσεις I, II του Σχήματος 10 τις περιστρέψουμε και μετατοπίσουμε την αρχή των αξόνων, η σχέση (5.24) θα ισχύει. Επομένως, αυτό που αποδείξαν είναι ότι στους Ευκλείδειους χώρους, οι αποστάσεις μεταξύ των σημείων, είναι αμετάβλητες-αναλοίωτες σε μία περιστροφή, μετάθεση, αντανάκλαση, μετατροπή ή και διαστολή των διαστάσεων. Το προηγούμενο απεικονίζεται στο Σχήμα[8] 11. Ένα άλλο μοντέλο απόστασης το οποίο έχει ενδιαφέρουσες γεωμετρικές ιδιότητες, είναι αυτό του city-block , όπου δίνεται από τη σχέση:

$$d_{ij} = \sum_{a}^{r} |x_{ia} - x_{ja}| \tag{5.25}$$

Η εξίσωση (5.25), μας λέει ότι η απόσταση d_{ij} είναι απλά το άθροισμα της απόλυτης διαφοράς των προβολών των σημείων i, j επάνω στις διαστάσεις του χώρου, όπως απεικονίζεται στο Σχήμα [9]12. Ο λόγος που ονομάζεται έτσι αυτού του είδους μοντέλου, είναι ότι οι αποστάσεις μεταξύ των σημείων παρομοιάζονται με τις αντίστοιχες αποστάσεις όταν περπατάμε σε μια πόλη ανάμεσα στα οικοδομικά τετράγωνα. Βέβαια, οι διαφορά στην προηγούμενη παρομοίωση είναι στη μεν περίπτωση των οικοδομικών τετραγώνων μιας πόλης είναι πεπερασμένος ο αριθμός των δρόμων που μπορεί να επιλέξει κάποιος, ενώ στην περίπτωση των city-block χώρων είναι άπειρος ο αριθμός των γραμμών που είναι παράλληλες στους άξονες και μπορούν να χρησιμοποιηθούν

[7]Πηγή: Multidimensional Scaling History, Theory and Applications, LEA 1987, σελ. 95
[8]Πηγή: Multidimensional Scaling History, Theory and Applications, LEA 1987, σελ. 96
[9]Πηγή: Multidimensional Scaling History, Theory and Applications, LEA 1987, σελ. 97



Σχήμα 10: Ευκλείδεια απόσταση και το Πυθαγόρειο θεώρημα.

για την επιλογή της συντομότερη απόστασης. Αυτό τελικά που αποδείχθηκε για αυτό το χώρο είναι ότι δεν είναι περιστρέψιμος. Δηλαδή, αν περιστραφούν οι άξονες του συγκεκριμένου χώρου, τότε οι αποστάσεις μεταξύ των σημείων θα διαφοροποιηθούν. Τέλος, ο τρίτος χώρος που παρουσιάζει ενδιαφέροντα χαρακτηριστικά από γεωμετρικής άποψης, είναι το κυρίαρχο μοντέλο (dominance model), ο οποίος δίνεται από τη σχέση:

$$d_{ij} = \max_{a}^{r} |x_{ia} - x_{ja}| \tag{5.26}$$

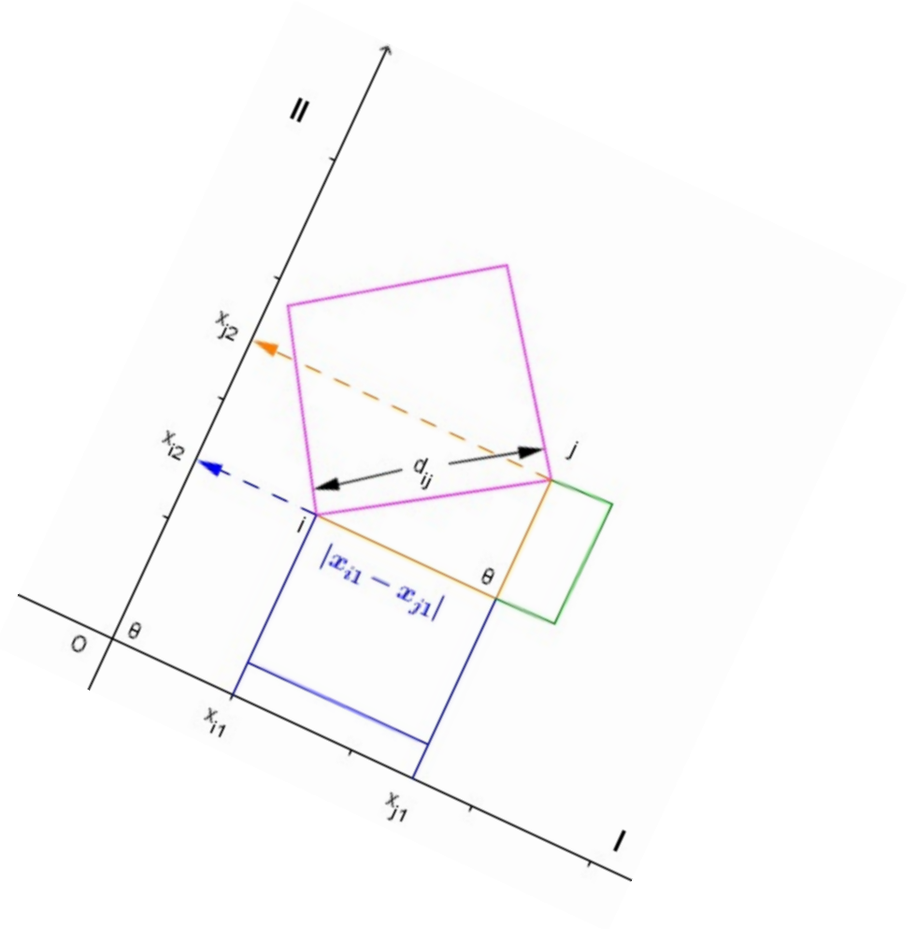
και αποτελεί ειδική περίπτωση του μοντέλου Minkowski, όταν ο εκθέτης πάρει την τιμή του απείρου. Ο συγκεκριμένος τύπος χώρου μας λέει ότι η απόσταση μεταξύ δύο σημείων, είναι ίση με τη μέγιστη διαφορά των προβολών των σημείων επάνω στις διαστάσεις του χώρου, όπως αυτό απεικονίζεται στο Σχήμα[10] 13. Ομοίως και σε αυτή την περίπτωση μοντέλου, ο χώρος σε περίπτωση περιστροφής, οι αποστάσεις του θα διαφοροποιηθούν.

5.5 Αλγεβρικές ιδότητες

5.5.1 Μετασχηματισμός ομοιότητας - Similarity Transformation

Με τον όρο similarity transformation εννοείται εκείνος ο ένα-προς-ένα μετασχηματισμός ο οποίος αλλάζει τον έναν χώρο σε έναν άλλον, ο οποίος είναι όμοιός του. Πιο συγκεκριμένα, δύο χώροι θα λέγονται όμοιοι όταν οι αποστάσεις μεταξύ όλων των ζευγών σημείων του ενός χώρου είναι κατά αναλογία s ως προς όλες τις αποστάσεις μεταξύ των αντίστοιχων ζευγαριών των σημείων του άλλου χώρου. Έτσι μιας και οι αποστάσεις είναι ορισμένες κατά ένα μέτρο αναλογίας, μια αλλαγή στην απόσταση κατά μια σταθερά πολλαπλάσια του s, η αλλαγή που θα προκαλέσει είναι μόνο ως προς τη μονάδα μέτρησης της απόστασης. Βέβαια έχει αποδειχθεί ότι ο similarity transformation, αποτελεί περιορισμένο γραμμικό μετασχηματισμό του πίνακα των συντεταγμένων των σημείων του Ευκλείδειου χώρου. Είναι αποδεδειγμένο ότι οι ακόλουθες πράξεις:

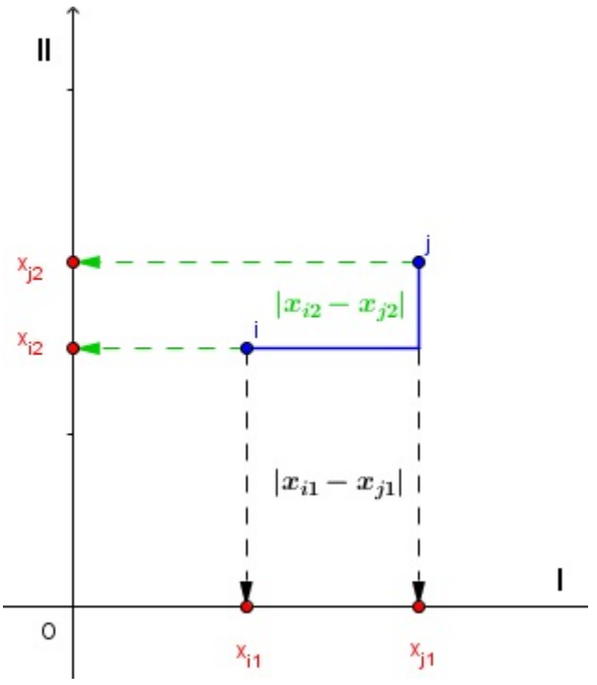
[10]Πηγή: Multidimensional Scaling History, Theory and Applications, LEA 1987, σελ. 99



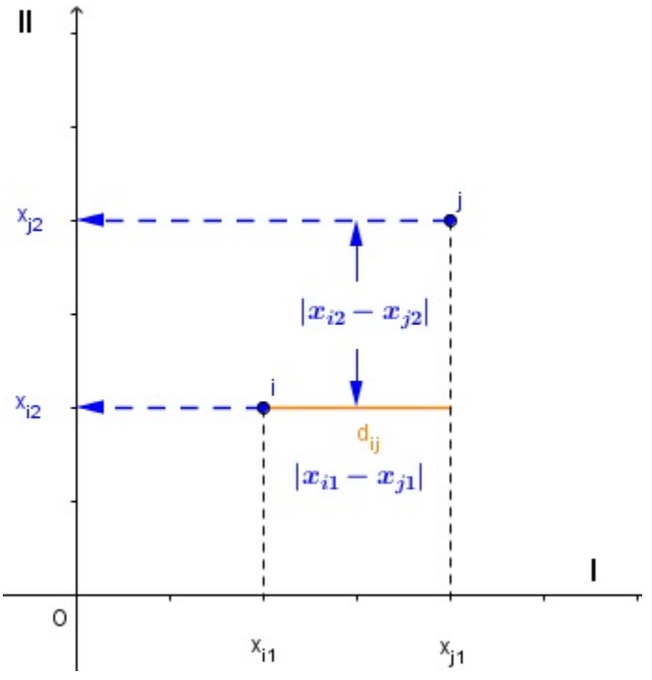
Σχήμα 11: Η περιστροφή δεν επηρεάζει τις Ευκλείδειες αποστάσεις.

1. Ορθογώνια περιστροφή (Orthogonal Rotation)
2. Μετατόπιση (Translation)
3. Μετάθεση (Permutation)
4. Αντανάκλαση (Reflection)
5. Κεντρική διαστολή (Central Dilation)

είναι επιτρεπόμενες για similarity transformation του Ευκλείδειου χώρου. Αξίζει να αναφέρουμε δυο λόγια για τις παραπάνω επιτρεπόμενες πράξεις, σχετικά με το τι είναι η κάθε μια. Η πρώτη πράξη όπως το λέει και το όνομά της, περιστρέφει τις διαστάσεις της γύρω από την αρχή των αξόνων, με τις διαστάσεις να παραμένουν ορθογώνιες. Επιπλέον μετατόπιση είναι εκείνη η πράξη όπου όταν πραγματοποιείται η διαμόρφωση των σημείων, γίνεται κατά τέτοιο τρόπο όπου έχουμε τη μετακίνηση από ένα μέρος του χώρου σε έναν άλλον. Ενναλακτικά μπορούμε να αντιληφθούμε τη μετατόπιση του χώρου, ως μια γραμμή που ενώνει δύο σημεία πριν τη μετατόπιση αυτού και την παράλληλη μεταφορά αυτής της γραμμής για τα ίδια σημεία. Από την άλλη η μετάθεση, από γεωμετρικής άποψης, δεν είναι τίποτε άλλο από την αναδιάταξη των διαστάσεων. Η αντανάκλαση αφορά εκείνη την πράξη που συμβαίνει όταν τα πρόσημα όλων των συντεταγμένων αλλάξουν. Το προηγούμενο αλγεβρικά πραγματοποιείται με τον πολλαπλασιασμό του πίνακα των συνταταγμένων με ένα διαγώνιο πίνακα ο οποίος έχει -1 για εκείνες τις διαστάσεις που θέλουμε να αντανακλαστούν και 1 για εκείνες που δεν επιθυμούμε. Τέλος, η τελευταία πράξη αυτή της κεντρικής διαστολής, από γεωμετρικής σκοπιάς έχει να κάνει με την ολική διαμόρφωση του μεγέθους του χώρου, είτε να γίνει πιο μεγάλος είτε πιο μικρός. Αλγεβρικά αυτό γίνεται



Σχήμα 12: Απόσταση City-block.



Σχήμα 13: Κυρίαρχο μοντέλο απόστασης.

με το να πολλαπλασιάσουμε όλες τις συντεταγμένες με μία θετική σταθερά, όπου για σταθερά μεγαλύτερη της μονάδας, έχουμε διαστολή του χώρου ενώ στην περίπτωση μικρότερης της μονάδας συμβαίνει το αντίθετο. Εκτός από τα μη σταθμισμένα μοντέλα απόστασης, υπάρχουν και τα σταθμισμένα (Weighted Distance Models), τα οποία είναι πιο πολύπλοκα σε σχέση με τα μη σταθμισμένα από άποψη άλγεβρας και γεωμετρικής παρουσίασης, με το χώρο συνήθως να είναι ο Ευκλείδειος. Το Ευκλείδειο γενικό μοντέλο (GEM) (Young, 1984 [2]) το οποίο χρησιμοποιείται συνήθως δίνεται από τη σχέση:

$$d^2_{ijk} = (y_i - x_j)\, V_i W_k\, (y_i - x_j)' \tag{5.27}$$

όπου d^2_{ijk} είναι η τετραγωνική απόσταση μεταξύ των σημείων i, j τροποποιημένες από τα βάρη του W_k. Επίσης όπου V_i και W_k πρόκειται για τετραγωνικούς, συμμετρικούς, τάξεως r, θετικά ημιορισμένους πίνακες που σχετίζονται με το αντικείμενο γραμμή i και τον πίνακα των βαρών, αντίστοιχα. Τέλος, με x_j και y_i είναι το $r-$οστό σημείο του διανύσματος γραμμής των συντεταγμένων, που προσδιορίζουν τη θέση του σημείου στον $r-$διάστατο Ευκλείδειο χώρο για τους πίνακες συντεταγμένων X και Y, αντίστοιχα.

6 Αλγόριθμοι κατά τη διαδικασία MDS

6.1 Αλγόριθμος Torgerson

Χάρη στους Torgerson (1952, 1958) [1] και Gower (1966) [1], έχουμε την πρώτη διαθέσιμη τεχνική για MDS την κλασική κλιμάκωση (classical scaling) γνωστή και ως Torgerson scaling και Torgerson-Gower scaling, όπου βασίζεται στα θεωρήματα των Eckart και Young (1936) [1] καθώς και των Young και Householder (1938) [1]. Η κεντρική ιδέα του κλασικού scaling, βασίστηκε στο να υποθέσουν ότι οι ανομοιότητες είναι αποστάσεις και να βρούνε τις συντεταγμένες που αντιστοιχούν σε αυτές, έτσι ώστε να μπορούν να τις ερμηνεύουν. Με άλλα λόγια είναι δυνατό να κατασκευασθεί ένας πίνακας $\mathbf{X}$ που περιέχει τις καρτεσιανές συντεταγμένες των σημείων του Ευκλείδειου χώρου, αν προηγουμένως είναι γνωστό ο πίνακας $\mathbf{D}$ που περιέχει τις αποστάσεις, μεταξύ αυτών των σημείων. Το ζητούμενο αποτέλεσμα μπορεί να προκύψει με τα δύο βήματα που ακολουθούν, και αυτά είναι:

1. Αρχικά χρησιμοποιώντας τον νόμο των συνημιτόνων για τη μετατροπή του πίνακα $\mathbf{D}$ σε έναν πίνακα $\mathbf{\Delta} = \mathbf{X}\mathbf{X}'$ *βαθμωτού γινομένου (scalar product).*

2. Δημιουργία του $\mathbf{X}$ πίνακα, αν εφαρμοσθεί η μέθοδο της SVD στον πίνακα $\mathbf{\Delta}$.

Στο μοντέλο του Torgerson, η βασική υπόθεση που κάνει είναι να υποθέτει τις ανομοιότητες ίσες με τις αποστάσεις στον πολυδιαστατικό Ευκλείδειο χώρο και εν συνεχεία να βρίσκει τις συντεταγμένες που θα τις περιγράφουν. Έστω δ_{ij} είναι η παρατηρηθείσα ανομοιότητα μεταξύ των ερεθισμάτων i και j. Επίσης, έστω x_{ik} και x_{jk} με $i = 1, \ldots, I; j = 1, \ldots, J; I = J; k = 1, \ldots, K$ να είναι οι συντεταγμένες των ερεθισμάτων i και j αντίστοιχα ως προς τη διάσταση k. Αξίζει να επισημανθεί ότι, τόσο ο αριθμός των γραμμών I όσο και των στηλών J στον πίνακα που περιέχει τις ανομοιότητες είναι ίσος μιας και ανιστοιχούν στα ίδια προς κρίση ερεθίσματα. Οπότε η βασική αρχή της υπόθεσης του Torgerson, μας λέει ότι:

$$\delta_{ij} = d_{ij} = [\sum_{k} (x_{ik} - x_{jk})^2)]^{1/2} \tag{6.1}$$

Χωρίς να επηρεάζεται το τελικό αποτέλεσμα, ο Torgerson υπέθεσε ότι η μέση συντεταγμένη κάθε διάστασης ερεθίσματος, είναι μηδέν:

$$\sum_{i} x_{ik} = \sum_{j} x_{jk} = 0 \tag{6.2}$$

Ο Torgerson αυτό που σκέφτηκε ήταν να δημιουργήσει έναν double-center πίνακα $\mathbf{\Delta}^*$ με στοιχεία δ^*_{ij} υπολογιζόμενος από τον πίνακα δεδομένων. Ο πίνακας αυτός, είναι εκείνος για τον οποίο τόσο ο μέσος των στοιχείων της κάθε γραμμής όσο και εκείνος των στηλών είναι μηδέν. Με άλλα λόγια για να προκύψει αυτού του είδους ο πίνακας θα πρέπει να αφαιρεθούν τα τετράγωνα των μέσων όρων των γραμμών και στηλών αντιστοίχως, από τα στοιχεία που βρίσκονται στον πίνακα των ανομοιοτήτων και να προσθέσουμε το τετράγωνο του μέγιστου μέσου όρου καθώς και να πολλαπλασιάσουμε επί $-1/2$. Επομένως τα νέα στοιχεία που θα περιέχει ο «νέος» πίνακας $\mathbf{\Delta}^*$ θα είναι:

$$\delta^*_{ij} = -\frac{1}{2}\left(\delta^2_{ij} - \delta^2_{i.} - \delta^2_{.j} + \delta^2_{..}\right) \tag{6.3}$$

όπου $\delta^2_{i.}, \delta^2_{.j}$ και $\delta^2_{..}$ είναι:

$$\delta^2_{i.} = \frac{1}{J}\sum_{J} \delta^2_{ij} \tag{6.4}$$

$$\delta^2_{.j} = \frac{1}{I}\sum_{i} \delta^2_{ij} \tag{6.5}$$

$$\delta_{\cdot\cdot}^2 = \frac{1}{IJ}\sum_i\sum_j \delta_{ij}^2 \tag{6.6}$$

με δ_{ij} οι παρατηρούμενες ανομοιότητες μεταξύ των γνωρισμάτων i και j. Τέλος, ο Torgerson απέδειξε ότι στην περίπτωση που τα δεδομένα ικανοποιούν τη σχέση (6.1) τότε κάθε στοιχείο στο νέο πίνακα $\mathbf{\Delta}^*$, θα είναι της μορφής:

$$\delta_{ij}^* = \sum_k x_{ik}x_{jk} \tag{6.7}$$

Επιπλέον η σχέση (6.7) αποτελεί το θεμελειώδες θεώρημα γύρω από το οποίο ο Torgerson σχεδίασε τον αλγόριθμό του. Υπό μορφή πινάκων η σχέση (6.7) γράφεται ως εξής:

$$\mathbf{\Delta}^* = \mathbf{X}\mathbf{X}' \tag{6.8}$$

με το $\mathbf{\Delta}^*$ να καλείται συχνά ως πίνακας βαθμωτού γινομένου (scalar product matrix) και τον πίνακα $\mathbf{X}$ διάστασης $(I \times K)$ να είναι ο πίνακας των συντεταγμένων ερεθισμάτων. Επίσης ένα άλλο ερώτημα που απασχόλησε τον Torgerson, είναι το πως μπορούσε να φτάσει σε ένα βαθμωτό πίνακα έστω $\mathbf{B}$ δίνοντας αρχικά τις τετραγωνικές αποστάσεις $\mathbf{D^{(2)}}$. Οι τετραγωνικές αποστάσεις μπορούν να υπολογιστούν από τον πίνακα $\mathbf{X}$ ως εξής:

$$\mathbf{D^{(2)}} = \underset{\sim}{\boldsymbol{c}}\mathbb{1}' + \mathbb{1}\underset{\sim}{\boldsymbol{c}}' - \mathbf{2XX}' = \underset{\sim}{\boldsymbol{c}}\mathbb{1}' + \mathbb{1}\underset{\sim}{\boldsymbol{c}}' - \mathbf{2B} \tag{6.9}$$

όπου $\underset{\sim}{\boldsymbol{c}}$ είναι το διάνυσμα των στοιχείων του πίνακα $\mathbf{XX}'$. Αν πολλαπλασιαστεί από δεξιά και αριστερά με τον κεντραρισμένο πίνακα $\mathbf{J} = \mathbf{I} - \mathbf{n}^{-1}\mathbb{1}\mathbb{1}'$ καθώς και με τον παράγοντα $-\frac{1}{2}$ δίνει:

$$\begin{aligned}
-\frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{JD^{(2)}J} &= -\frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{J}(\underset{\sim}{\boldsymbol{c}}\mathbb{1}' + \mathbb{1}\underset{\sim}{\boldsymbol{c}}' - \mathbf{2}\boldsymbol{XX'})\boldsymbol{J} \\
&= -\frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{J}\underset{\sim}{\boldsymbol{c}}\mathbb{1}'\boldsymbol{J} - \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{J}\mathbb{1}\underset{\sim}{\boldsymbol{c}}'\boldsymbol{J} + \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{J}(\mathbf{2}\boldsymbol{B})\boldsymbol{J} \\
&= -\frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{J}\underset{\sim}{\boldsymbol{c}}\mathbf{0}' - \frac{\mathbf{1}}{\mathbf{2}}\mathbf{0}\underset{\sim}{\boldsymbol{c}}'\boldsymbol{J} + \boldsymbol{JBJ} = \boldsymbol{B}
\end{aligned} \tag{6.10}$$

οι δύο πρώτοι όροι της σχέσης (6.10) είναι μηδέν διότι αν «κεντράρουμε» ένα διάνυσμα με μονάδες, τότε αυτό θα επιστρέψει ένα διάνυσμα με μηδενικά ($\mathbb{1}'\boldsymbol{J} = \mathbf{0}$) και η διαδικασία αυτή ονομάζεται double centering. Επομένως για να βρεθούν οι συντεταγμένες από το $\mathbf{B}$, τον παραγοντοποιούμε μέσω της μεθόδου Singular Value Decomposition (SVD). Η μέθοδος SVD είναι εκείνη η οποία μας λέει ότι για κάθε πίνακα $\mathbf{A_{n\times m}}$ μπορεί να αναλυθεί σε:

$$\mathbf{A} = \mathbf{P\Phi Q}' \tag{6.11}$$

όπου ο $\mathbf{P}$ ένας $n \times m$ πίνακας για τον οποίο ισχύει $\mathbf{PP}' = \mathbf{I}$, $\mathbf{\Phi}$ είναι $m \times m$ διαγώνιος πίνακας με ιδιάζουσες τιμές (singular values) στην κύρια διαγώνιο $\phi_i \geq 0$ και ο $\mathbf{Q}$ ένας $m \times m$ πίνακας για τον οποίο ισχύει $\mathbf{Q'Q} = \mathbf{I}$. Επιπλέον η αποσύνθεση (eigendecomposition) του $\mathbf{A}'\mathbf{A}$ δίνει:

$$\mathbf{A}'\mathbf{A} = \mathbf{Q\Phi P}'\mathbf{P\Phi Q}' = \mathbf{Q\Phi^2 Q}' \tag{6.12}$$

το οποίο αποδεικνύει ότι οι ιδιοτιμές του $\mathbf{A}'\mathbf{A}$ είναι όλες μη αρνητικές μιας και περιέχουν το ϕ_i^2 και τέλος χρησιμοποιόντας την ορθοκανονικότητα του $\mathbf{Q}$ και το διαγώνιο $\mathbf{\Phi}$ μας δίνει τον πίνακα $\mathbf{P}$:

$$\begin{aligned}
\boldsymbol{A} &= \boldsymbol{P\Phi Q'} \\
\boldsymbol{AQ} &= \boldsymbol{P\Phi Q'Q} = \boldsymbol{P\Phi} \\
\boldsymbol{AQ\Phi^{-1}} &= \boldsymbol{P\Phi\Phi^{-1}} = \boldsymbol{P}.
\end{aligned} \tag{6.13}$$

Τέλος, στην εύρεση ενός πίνακα συντεταγμένων δοθέντος ενός πίνακα βαθμωτού γινομένου π.χ. $\mathbf{B} = \mathbf{XX}'$, η μέθοδος του κλασικού scaling διαφέρει στο αντί να ληφθεί υπόψη ο πίνακας των τετραγωνικών αποστάσεων, τώρα συμπεριλαμβάνεται ο πίνακας των τετραγωνικών ονομοιοτήτων (dissimilarities) $\mathbf{\Delta^{(2)}}$ και με αυτόν εφαρμόζεται η διαδικασία που περιγράφει η σχέση (6.10). Στη συνέχεια εφαρμόζεται double centering στον πίνακα $\mathbf{B_\Delta} = \mathbf{-1/2 J \Delta^{(2)} J}$ καθώς και ο υπολογισμός της αποσύνθεσης (eigendecomposition) $\mathbf{B_\Delta} = \mathbf{QUQ}'$, όπου $\mathbf{U}$ ο πίνακας ιδιοτιμών. Έστω m να είναι η διάσταση της λύσης. Τότε στον πίνακα $\mathbf{U}_+$ έχουμε τις πρώτες m θετικές ιδιοτιμές και στον $\mathbf{Q}_+$ τις πρώτες m στήλες του $\mathbf{Q}$. Τότε ο πίνακας συντεταγμένων του κλασικού scaling δίνεται από: $\mathbf{X} = \mathbf{Q}_+\mathbf{U}_+^{\mathbf{1/2}}$.

6.2 Αλγόριθμος Kruskal

Ο αλγόριθμος του Kruskal (1964) [3] είχε να κάνει με τη μελέτη της προσαρμογής των δεδομένων, τα οποία είναι μονοτονικά σχεσιακά με τις αποστάσεις στο γενικό χώρο Minkowski. Με άλλα λόγια ο αλγόριθμος αυτός επιτρέπει στο χρήστη την εκτίμηση των συντεταγμένων των ερεθισμάτων από τα δεδομένα και θα έχουν την παρακατω μορφή:

$$\delta_{ij} = f(d_{ij}) = f\left[\left(\sum_k |x_{ik} - x_{jk}|^p\right)^{1/p}\right] \tag{6.14}$$

όπου f μία μονότονη συνάρτηση που ικανοποιεί την παρακάτω σχέση:

$$d_{ij} < d_{i^*j^*} \rightarrow f(d_{ij}) < f(d_{i^*j^*}) \tag{6.15}$$

για όλα τα i, i^*, j, j^* και το p θα δίνεται από το χρήστη ύστερα από την προσαρμογή των δεδομένων σε μοντέλα απόστασης Minkowski και τάσσεται υπέρ εκείνου του μοντέλου και άρα εκείνο το p, για το οποίο έχει την καλύτερη προσαρμογή στα δεδομένα. Συνήθως βέβαια η επιλογή του δείκτη p είναι εκείνη για $p = 2$, δηλαδή για το Ευκλείδειο μοντέλο, μιας και ο χρόνος που απαιτείται από ένα υπολογιστικό πρόγραμμα αυξάνεται αρκετά όταν αντί για την περίπτωση αυτή, έχουμε την προσαρμογή ενός μη-Ευκλείδειου, δηλαδή για $p \neq 2$. Ο Kruskal όρισε ένα μέτρο προσαρμογής με όνομα STRESS το οποίο περιελάμβανε τόσο τις ανομοιότητες όσο και τις αποστάσεις και δίνεται από τους παρακάτω τύπους:

$$S_1 = \left[\frac{\sum_i \sum_j \left(\hat{\delta_{ij}} - \hat{d_{ij}}\right)^2}{\sum_i \sum_j \hat{d_{ij}^2}}\right]^{1/2} \quad S_2 = \left[\frac{\sum_i \sum_j \left(\hat{\delta_{ij}} - \hat{d_{ij}}\right)^2}{\sum_i \sum_j (\hat{d_{ij}} - \hat{d_{..}})^2}\right]^{1/2} \tag{6.16}$$

όπου $\hat{d_{..}} = \frac{1}{IJ}\sum_i \sum_j \hat{d_{ij}}$, ο αριθμιτικος μέσος των εκτιμημένων αποστάσεων και με τους δύο τύπους S_1, S_2 να διαφέρουν μόνο ως προς τη σταθερά κανονικοποίησης που βρίσκεται στον παρονομαστή. Από τα πιο γνωστά μη-μετρικά MDS προγράμματα, θεωρούνται τα M-D-SCAL, TORSCA (Young and Torgerson, 1967) [3] και το KYST (Kruskal et al., 1973) [3] , τα οποία παράγουν τις εκτιμημένες εκείνες συντεταγμένες που ελαχιστοποιούν το μέτρο STRESS, με το τελευταίο να θεωρείται το καλύτερο από τα άλλα δύο. Ένα από τα πλεονεκτήματα του KYST έναντι των υπολοίπων είναι ότι δίνει τη δυνατότητα στο χρήστη να επιλέξει μεταξύ των δύο STRESS τύπων ως συνάρτηση που θα ελαχιστοποιηθεί από τις εκτιμημένες συντεταγμένες. Αρκετοί ερευνητές κατέληξαν στο συμπέρασμα πως όταν έχουν ένα συμμετρικό πίνακα δεδομένων από ανομοιότητες ή ομοιότητες (dissimilarities or similarities), είναι καταλληλότερος σαν μέτρο ο S_1, λόγω του ότι μειώνει αρκετά τα υπολογιστικά προβλήματα, ενώ αν τα δεδομένα αποτελούνται από προτιμήσεις τότε η καλύτερη επιλογή είναι αυτή του τύπου S_2, της σχέσης (6.16) αντίστοιχα.

6.3 Αλγόριθμος ALSCAL

Ο αλγόριθμος ALSCAL είναι ένα πρόγραμμα μη-μετρικό το οποίο χρησιμοποιεί εναλλακτικά ελάχιστα τετράγωνα (Alternating Least Squares), ο οποίος είχε προταθεί από τους Takane, Young και de Leeuw [2] [3] το 1977. Με τον όρο μη-μετρικό εννοούμε εκείνο το πρόγραμμα που τα δεδομένα που επεξεργάζεται μετρούνται σε τακτική κλίμακα (π.χ. τόπος γέννησης, μορφωτικό επίπεδο κ.τ.λ.) ενώ αν μετρούνται σε κλίμακα

ισοδιαστημάτων ή αναλογική (π.χ. χρόνοι απόκρισης, βαθμολογίες εξετάσεων κ.τ.λ.) τότε το πρόγραμμα θα χαρακτηριζόταν ως μετρικό. Ο αλγόριθμος αυτός είναι κατάλληλος για κάθε τύπο δεδομένων δύο ή τριών κατευθύνσεων δεδομένων (two or three way data) τα οποία μπορεί να είναι ορθογώνια ή τετραγωνικά, συμμετρικά ή μη συμμετρικά, επαναλαμβανόμενα ή οχι, με ή χωρίς να λείπουν δεδομένα (missing data). Όπως και στον προηγούμενο αλγόριθμο τον Kruskal, έτσι και εδώ, ο αλγόριθμος αυτός προσαρμόζει ένα μέτρο το οποίο μοιάζει με το $STRESS$ και λέγεται $S-STRESS$. Οι τύποι που δίνουν αυτό το μέτρο είναι οι εξής:

$$SS_1 = \left[\frac{\sum_i \sum_j \left(\hat{\delta}_{ij}^2 - \hat{d}_{ij}^2\right)^2}{\sum_i \sum_j \left(\hat{d_{ij}}^2\right)^2}\right]^{1/2} \quad SS_2 = \left[\frac{\sum_i \sum_j \left(\hat{\delta}_{ij}^2 - \hat{d}_{ij}^2\right)^2}{\sum_i \sum_j \left(\hat{d}_{ij}^2 - \hat{d}_{..}^2\right)^2}\right]^{1/2} \tag{6.17}$$

Η διαφορά του τύπου $STRESS$ της σχέσης (6.16) με αυτή του $S-STRESS$ της σχέσης (6.17) έγκειται στο γεγονός ότι ο $S-STRESS$ περιλαμβάνει τόσο τα τετράγωνα των αποστάσεων όσο και των ανομοιοτήτων. Βέβαια αντίστοιχα με τη σχέση (6.16), στη σχέση (6.17) όπου $\hat{d}_{..} = \frac{1}{IJ}\sum_i \sum_j \hat{d}_{ij}^2$ ο αριθμητικός μέσος των τετραγωνικών εκτιμημένων αποστάσεων. Επιπλέον, όσο για το ποια από τις δύο σχέσεις είναι η καταλληλότερη αυτό αφήνεται στον τελικό χρήστη αφού είναι στην ευχέρειά του να επιλέξει ανάλογα με το τι δεδομένα έχει στη διάθεσή του κάθε φορά. Έτσι ο μεν πρώτος τύπος (SS_1), της (6.17) είναι προτιμότερος όταν ο χρήστης έχει ως δεδομένα ανομοιότητες ή ομοιότητες και όχι προτιμήσεις.

6.4 Αλγόριθμος INDSCAL

6.4.1 Σταθμισμένο Ευκλείδειο Μοντέλο

Η περίπτωση του σταθμισμένου Ευκλείδειου μοντέλου διαφέρει από το απλό μοντέλο στο ότι το μεν πρώτο λαμβάνει υπόψη του διαφορετικό βάρος για κάθε διάσταση ενώ στην περίπτωση του δευτέρου μοντέλου περιλαμβάνει ένα μόνο βάρος για όλες τις διαστάσεις. Με άλλα λόγια στο σταθμισμένο, τα βάρη ποικίλουν ανά διάσταση και μπορούν να πάρουν διάφορες τιμές ενώ στο απλό μοντέλο δεν υπάρχει αυτή η δυνατότητα. Το όλο ενδιαφέρον γύρω από την περίπτωση του σταθμισμένου Ευκλείδειου μοντέλου είναι να παρουσιαστούν οι ανομοιότητες μεταξύ των αντικειμένων i, j όπως αυτές περιγράφηκαν από τα άτομα k, μέσω τις απόστασης d_{ijk}:

$$d_{ijk}\left(GW_k\right) = \left[\sum_{a=1}^{m}\left(w_{aak}g_{ia} - w_{aak}g_{ja}\right)^2\right]^{1/2} = \left[\sum_{a=1}^{m} w_{aak}^2\left(g_{ia} - g_{ja}\right)^2\right]^{1/2} \tag{6.18}$$

όπου οι δείκτες i, j να παίρνουν τις τιμές $1, \cdots, n$ και $k = 1, \cdots, K$, $a = 1, \cdots, m$ ο W_k να είναι ένας $(m \times m)$ διαγώνιος θετικός πίνακας που περιέχει τα βάρη w_{aak} για κάθε διάσταση a και άτομο k καθώς και ο G να αποτελεί τον πίνακα συντεταγμένων του χώρου ερεθισμάτων του G (group stimulus space or common space). Τέλος, η σχέση (6.18) ονομάζεται σταθμισμένη Ευκλείδεια απόσταση.

Ο αλγόριθμος αυτός αποσκοπεί τόσο στην ανάλυση όσο και στη σύγκριση των διαφορών των λύσεων που λαμβάνουμε μέσω της μεθόδου MDS που αφορά στα ίδια n ερεθίσματα για τα k άτομα. Επιπλέον ο αλγόριθμος αυτός βοηθά στην επίλυση σταθμισμένων μοντέλων, τα οποία βασίζονται σε ένα βαθμωτού γινομένου πίνακα, ο οποίος μοιάζει με αυτό στην περίπτωση του κλασσικού scaling. Έστω $B_{\Delta_k} = -\frac{1}{2}J\Delta_k^{(2)}J$, ο $(n \times n)$ πίνακας βαθμωτού γινομένου για το άτομο k, η σχέση που προέκυψε από τις αποστάσεις μέσω της (6.10). Εισάγοντας τα βάρη στον αλγόρθμο INDSCAL, και με τον πίνακα B_{Δ_k} γνωστό, δημιουργείται η συνάρτηση απώλειας:

$$L_{IND}\left(G, W_1, \cdots, W_k\right) = \sum_{k=1}^{K} \|B_{\Delta_k} - GW_k^2G'\|^2 = \sum_{k=1}^{K}\sum_{i,j}\left(b_{ijk} - \sum_{a=1}^{m} g_{ia}g_{ja}w_{aak}^2\right)^2 \tag{6.19}$$

Η παραπάνω σχέση (6.19) λύνεται ως προς τα δύο σύνολα παραμέτρων G και W_k, $(k = 1, \cdots, K)$ αντίστοιχα. Επειδή δεν είναι εφικτό να δωθεί αναλυτική λύση για την προηγούμενη σχέση, ο αλγόριθμος INDSCAL χρησιμοποιεί μια διαφορετική προσέγγιση, με την ανανέωση του G κρατώντας σταθερό τον πίνακα W_k και μετά

γίνεται η ανανέωση του W_k με σταθερό τον πίνακα G. Αυτή η συνεχής ανανέωση πραγματοποιείται έως ότου υπάρξει σύγκλιση. Δηλαδή ο INDSCAL προσεγγίζει κάθε κεντραρισμένο πίνακα B_{Δ_k} μέσω $B_{\Delta_k} \approx GW_k^2G'$ και αναζητά λύση για τα $(G, W_1^2, W_2^2, \cdots, W_k^2)$ τέτοια ώστε το μοντέλο να προσαρμόζεται στα δεδομένα υπό την έννοια των ελαχίστων τετραγώνων. Έτσι ο INDSCAL ελαχιστοποιεί την σχέση (6.19), με τον πίνακα G να έχει μέγιστη τάξη στηλών (full column rank) R και ο W_k^2 να είναι ο διαγώνιος και μη αρνητικός $(R \times R)$ πίνακας. Επίσης, ο αλγόριθμος INDSCAL μπορεί να δώσει μία καλύτερη προσέγγιση του πίνακα G με γνωστό τον πίνακα W_k^2 με το να ελαχιστοποιήσει την παρακάτω συνάρτηση:

$$L_{IND}\,(G,H) = \sum_{k=1}^{K} \|B_{\Delta_k} - HW_k^2G'\|^2 = \sum_{k=1}^{K} trB_{\Delta_k}^2 + tr\left(G\left[\sum_k W_k^2H'HW_k^2\right]G\right)' - 2tr\left(G\left[\sum_k W_k^2H'B_{\Delta_k}\right]\right) \tag{6.20}$$

ως προς τους πίνακες G και H αντίστοιχα, όπου με την τεχνική αυτή ονομάζεται αλγόριθμος Candecomp Caroll & Chang, 1970 [1]. Μετά από τη σύγκλιση του αλγορίθμου οι πίνακες G και H μπορεί και να είναι ίσοι. Παραγογίζοντας την (6.20) ως προς G και εξισώνοντας με το μηδέν επιστρέφει την ανανέωση του G:

$$G = \left(\sum_k B_{\Delta_k}HW_k^2\right)\left(\sum_k W_k^2H'HW_k^2\right)^{-1} \tag{6.21}$$

με την ίδια παραπάνω διαδικασία πραγματοποιείται και για την ανανέωση του H, αντιστρέφοντας τους ρόλους με τον G και επαναλαμβάνεται η διαδικασία αυτή μέχρις ότου υπάρξει σύγκλιση και αυτή θα είναι η ολική βέλτιστη λύση για τους W και G, αντίστοιχα.

7 Συμπεράσματα - Σχόλια

Στην παρούσα διπλωματική εργασία αναπτύχθηκαν η θεωρία αλλά και οι αλγόριθμοι που χρησιμοποιούνται ευρέως στη μέθοδο πολυδιαστατικής κλιμάκωσης. Πιο συγκεκριμένα αναπτύχθηκαν οι απαραίτητες αξιωματικές αρχές που είναι απαραίτητες για τη μελέτη μοντέλων που χρησιμοποιούν τη μέθοδο της πολυδιαστατικής κλιμάκωσης (MDS). Επίσης, βασική υπόθεση στην παραπάνω μέθοδο είναι η ανάλυση των δεδομένων που εκφράζουν την ομοιότητα ή την ανομοιότητα των αντικειμένων ως προς ένα χαρακτηριστικό γνώρισμα ή δεδομένα τα οποία η μέτρησή τους στηρίζεται στις μεταξύ τους αποστάσεις με σκοπό τη δημιουργία ενός γραφήματος με τα σημεία αυτών. Ένα από τα σημαντικότερα πλεονεκτήματα αυτής της μεθόδου είναι όταν έχουμε πολλές μήτρες με πολλαπλά δεδομένα, όπου στην περίπτωση αυτή μπορεί να υπολογισθεί για καθεμιά μήτρα ξεχωριστή μήτρα ανομοιοτήτων και να αναλυθεί. Επιπλέον έγιναν αναφορές στα βασικά μοντέλα απόστασης που αντιστοιχούν στους τρόπους μέτρησης των δεδομένων με το πιο βασικό να είναι εκείνο του Ευκλείδειου μοντέλου και να ακολουθούν τα μοντέλα των city-block (Manhattan) καθώς και του κυρίαρχου μοντέλου που χρησιμοποιεί ως μέτρο την απόσταση Chebychev.

Στη συνέχεια έγιναν αναφορές σε αλγορίθμους που χρησιμοποιούνται στη μέθοδο MDS με τον πιο γνωστό αυτό του Torgerson [1]. Η βασική ιδέα που κάνει ο παραπάνω αλγόριθμος είναι να υποθέτει τις ανομοιότητες ότι είναι αποστάσεις και στόχος του είναι να βρει τις συνεταγμένες που θα τις περιγράψει. Ένας άλλος αλγόριθμος που αναπτύχθηκε ήταν ο INDSCAL, [11] [1] όπου παρουσιάζει τα δεδομένα για κάθε αντικείμενο με μεγαλύτερη ακρίβεια. Έτσι υποθέτει το γενικευμένο σταθμισμένο Ευκλείδειο μετρικό μοντέλο με διαφορετικά βάρη για κάθε αντικείμενο. Επίσης, ο αλγόριθμος αυτός έχει μια πολύ σημαντική ιδιότητα αυτή της «μοναδικότητας διάστασης» πράγμα που σημαίνει ότι παρ᾽ όλο που είναι ένα Ευκλείδειο μοντέλο απόστασης οι διαστάσεις του ή οι άξονες προσδιορίζονται με μοναδικότητα και δε χρειάζεται να γίνει κάποια περιστοφή και αποτελεί μεγάλο πλεονέκτημα στην MDS ερμηνεία. Τέλος, ένας άλλος αλγόριθμος που αναφέρθηκε και είναι εξίσου σημαντικός είναι ο ALSCAL που πρωτάθηκε απο τον Forrest W. Young [21] [2] [3] και είναι ελεύθερης διανομής λογισμικό. Με τον αλγόριθμο αυτόν μπορούμε να αναλύσουμε έναν ή περισσότερους πίνακες ομοιοτήτων ή ανομοιοτήτων δεδομένων. Έχει τη δυνατότητα να παρουσιάζει τις γραμμές και τις στήλες του πίνακα δεδομένων ως σημεία στον Ευκλείδειο χώρο και όσο πιο όμοιοι είναι οι γραμμές και οι στήλες τόσο πιο κοντά θα είναι τα αντίστοιχα σημεία τους με το αντίθετο να συμβαίνει όταν οι γραμμές και οι στήλες είναι ανόμοιες.

Αναφορές

[1] Ingwer Borg, Patrick Groenen, *"Modern Multidimensional Scaling Theory and Applications"*, Springer, 1997.

[2] Forrest W. Youg, Robert M. Hamer, *"Multidimensional Scaling History, Theory and Applications"*, LEA, 1987.

[3] Mark L. Davison, *"Multidimensional Scaling"*, Wiley, 1983.

[4] T. W. *Körner*, *"Metric and Topological Spaces"*, Σημειώσεις Cambridge, 2014.

[5] Carroll, J.D. (1972a) *"Individual differences and multidimensional scaling"*, in R.N. Shepard, A.K. Romney and S.B. Nerlove (eds), Multidimensional scaling: Theory and applications in the behavioral sciences, vol.1, Academic Press, New York.

[6] Allan D. Shocker and V. Srinivasan, *"A Consumer-Based Methodology for the Identification of New Product Ideas,"* Management Science, 20,6(February, 1974), 921-937.

[7] Joseph B. Kruskal and Myron Wish, *"Multidimensional Scaling"*, Bell Laboratories, Series: Quantitative Applications in the Social Sciences, SAGE Publications, ninth printing, 1981.

[8] Γεώργιος Κ. Σιάρδος *«Μέθοδοι Πολυμεταβλητής Στατιστικής Ανάλυσης»*, 3η Έκδοση Βελτιωμένη & Συμπληρωμένη, Εκδόσεις ΖΗΤΗ, Θεσσαλονίκη, 2004.

[9] David J. Bartholomew, Fiona Steele, Irini Moustaki, Jane I. Galbraith, *«Ανάλυση Πολυμεταβλητών Τεχνικών Στις Κοινωνικές Επιστήμες»*, Δεύτερη αμερικανική έκδοση, Εκδόσεις Κλειδάριθμος, 2011.

[10] J.Douglas Carroll, *"Multidimensional Scaling", Bell Telephone Laboratories, Murray Hill, New Jersey, 1980.*

[11] J.D. Carroll & J.J. Chang, (1970): *"Analysis of Individual Differences in Multidimensional scaling via an N-way generalization of "Eckart-Young" Decomposition"*. Psychometrika 35: 283−319.

[12] Alvin C. Rencher, *"Methods of Multivariate Analysis"*, Second Edition, A JOHN WILEY & SONS, INC. PUBLICATION, 2002.

[13] William Martens and Nick Zacharov, *"Multidimensional Perceptual Unfolding of Spatially Processed Speech I: Deriving Stimulus Space Using INDSCAL"*, AES 109^{th} Convention, Audio Engineering Society, Inc., 2000.

[14] `https://onlinecourses.science.psu.edu/stat505/node/49`

[15] `https://onlinecourses.science.psu.edu/stat505/node/63`

[16] `https://onlinecourses.science.psu.edu/stat505/node/159`

[17] `http://documents.software.dell.com/Statistics/Textbook/Canonical-Analysis`

[18] `http://www.acrwebsite.org/volumes/9188/volumes/v03/NA-03`

[19] `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3555222/`

[20] `http://forrest.psych.unc.edu/`

[21] `http://forrest.psych.unc.edu/research/alscal.html`