ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

του

ΓΑΛΛΟΥ ΙΩΑΝΝΗ

Μεταπτυχιακού φοιτητή στο ΔΠΜΣ << ΕΦΑΡΜΟΣΜΕΝΗ ΜΗΧΑΝΙΚΗ >>

Συγκριτική ανάλυση μεθόδων μείωσης διάστασης και εκμάθησης πολλαπλοτήτων για την ταξινόμηση δεδομένων Λειτουργικής Απεικόνισης Μαγνητικού Συντονισμού (FMRI)

Τριμελής Εξεταστική Επιτροπή

Σιέττος Κωνσταντίνος, Αναπληρωτής Καθηγητής University of Naples, Federico Naples II (Επιβλέπων)

Κομίνης Ιωάννης, Επίκουρος Καθηγητής Ε.Μ.Π

Θεοτόκογλου Ευστάθιος, Καθηγητής Ε.Μ.Π

Αθήνα, Ιούλιος 2019

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή Κωνσταντίνο Σιέττο που μου έδωσε τη δυνατότητα πραγματοποίησης της εργασίας αυτής και με καθοδήγησε μεθοδικά στην προσπάθεια μου και τους προβληματισμούς μου. Επίσης, θα ήθελα να ευχαριστήσω την οικογένεια μου που με στήριξε όλα αυτά τα χρόνια υλικά και πνευματικά αλλά και τους συμφοιτητές μου, με τους οποίους συνεργαστήκαμε και κάναμε καλή παρέα καθ' όλη τη διάρκεια των σπουδών.

Περίληψη

Η εν λόγω μεταπτυχιακή εργασία πραγματεύεται την συγκριτική ανάλυση μεθόδων εκμάθησης πολλαπλοτήτων (manifold learning) για την μείωση της διάστασης (data reduction) μεγάλης κλίμακας δεδομένων αλλά και μεθόδων εκμάθησης μηχανών (machine learning) για την ταξινόμηση δεδομένων. Η συγκριτική ανάλυση πραγματοποιείται σε δεδομένα απεικόνισης μαγνητικού συντονισμού (FMRI) σε υγιείς και ασθενείς με σχιζοφρένεια. Η εξαγωγή του σήματος ενδιαφέροντος από τα FMRI γίνεται με χρήση της μεθόδου ανάλυσης ανεξάρτητων συνιστωσών (ICA) ξεχωριστά για κάθε υποκείμενο. Οι μέθοδοι εκμάθησης πολλαπλοτήτων που χρησιμοποιούνται και αξιολογούνται είναι η πολυδιάστατη κλιμακοποίηση (Multi-Dimensional Scaling), η απεικόνιση ισομετρικών χαρακτηριστικών (ISOMAP) και οι απεικονίσεις διάχυσης (Diffusion Maps). Αρχικά γίνεται μία εισαγωγή σε δεδομένα Νευροαπεικόνισης και κυρίως στα δεδομένα FMRI που εφαρμόζουμε τις παραπάνω μεθόδους, ενώ γίνεται και μία πολύ σύντομη εισαγωγή στην θεωρία δικτύων. Στο 2^{ο} κεφάλαιο της εργασίας ορίζεται το πρόβλημα της μείωσης διάστασης ενώ αναλύονται μία προς μία οι μέθοδοι εκμάθησης πολλαπλοτήτων που πρόκειται να γίνουν αργότερα αντικείμενο της συγκριτικής ανάλυσης. Στο 3^{ο} Κεφάλαιο όπου αποτελεί και το κύριο μέρος της έρευνας, παρουσιάζεται το πρόβλημα με τα δεδομένα FMRI (74 υγιείς και 72 ασθενείς με σχιζοφρένεια) και στη συνέχεια περιγράφονται οι μέθοδοι εκμάθησης πολλαπλοτήτων και τα βήματα τα οποία ακολουθήθηκαν με σκοπό την αξιολόγηση των μεθόδων με δύο διαφορετικά μέτρα ομοιότητας, ένα μέτρο βασιζόμενο στην ευκλείδεια απόσταση και ένα άλλο βασιζόμενο στην πολλαπλή συσχέτιση με καθυστερήσεις. Στο 4^{ο} και 5^{ο} κεφάλαιο θα γίνει η παρουσίαση των αποτελεσμάτων και η αξιολόγηση αυτών αντίστοιχα. Η παρούσα εργασία προτείνει την χρήση μετρικών βασιζόμενων στην πολλαπλή συσχέτιση αντί αυτών που βασίζονται στην ευκλείδεια νόρμα όσων αφορά τα δεδομένα FMRI. Ακόμα προτείνει ότι η μέθοδος Diffusion Maps λειτουργεί καλύτερα από άλλες παρόμοιες μεθόδους αφού σημείωσε το υψηλότερο ποσοστό επιτυχούς ταξινόμησης μεταξύ των υπόλοιπων μεθόδων. Οι μέθοδοι εκμάθησης μηχανών που δίνουν συγκριτικά τα πιο αξιόπιστα ποσοστά είναι τα νευρωνικά δίκτυα και οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines) με ακτινικές συναρτήσεις βάσης (radial basis functions).

Abstract

In this thesis, we conduct a comparative analysis among manifold learning techniques for the dimensionality reduction of big data and methods of machine learning for classification. This comparative study is performed on Functional Magnetic Resonance Imaging (FMRI) data consisting of Schizophrenia patients and healthy controls. The extraction of the signal of interest is done using Independent Component Analysis (ICA) separately for each subject. Methods of manifold learning that are used and evaluated are the Multi-Dimensional Scaling (MDS), the Isometric Feature Mapping (ISOMAP) and the Diffusion Maps. At first, there is an introduction to Neuroimaging data and especially the FMRI data to which we apply the above methods, along with a very brief introduction to Network Theory. The 2nd chapter of thesis defines the problem of dimensionality reduction while one by one the manifold learning methods are discussed and analyzed. In the 3rd chapter, which is the core of this dissertation, the main problem and dataset (74 controls and 72 schizophrenia patients) are introduced and then follows the description methods and steps taken to evaluate manifold learning methods with two different similarity measures, one based on the Euclidean norm and another based on the lagged cross-correlation. Finally, 4th and 5th chapter present the results and the discussion of them respectively. This study suggests that the use of similarity measures for FMRI brain signals should be lagged cross correlation based, instead of euclidian. Also, Diffusion Maps performed better than any other similar method for dimensionality reduction giving the best classification rate among the other methods (namely MDS, ISOMAP). Finally, classification algorithms that gave (relatively) the most reliable classification rates were the Neural Nets and Support Vector Machines with a Radial Basis Functions (RBF) kernel.

ΠΕΡΙΕΧΟΜΕΝΑ

1 ΕΙΣΑΓΩΓΗ

Στην παρούσα μεταπτυχιακή εργασία θα μας απασχολήσει το πρόβλημα της λειτουργικής διασυνδεσιμότητας εγκεφαλικών σημάτων με τη χρήση θεωρίας πολύπλοκων δικτύων. Για τη συστηματική μελέτη του προβλήματος αυτού χρησιμοποιείται γενικά η ανάλυση δεδομένων νευροαπεικόνισης (πολλές φορές με πολύ μεγάλο μέγεθος (big data)) για τη μελέτη της διασυνδεσιμότητας ενός η περισσότερων περιοχών του εγκεφάλου. Απώτερος σκοπός της διαδικασίας, είναι αφενός μεν να κατανοήσουμε τη σχέση κάποιας λειτουργίας (νοητικής ή αισθησιοκινητικής) με μία ή και περισσότερες εγκεφαλικές περιοχές, και αφετέρου να μελετήσουμε την αλληλεπίδραση αυτών των εγκεφαλικών περιοχών μεταξύ τους. Στα πλαίσια της εργασίας αυτής, θα επικεντρώσουμε τη προσοχή μας συγκεκριμένα στα δεδομένα Λειτουργικής Απεικόνισης Μαγνητικού Συντονισμού (Functional Magnetic Resonance Imaging). Η FMRI είναι μία μέθοδος νευροαπεικόνισης που παράγει δεδομένα τεσσάρων διαστάσεων και είναι ικανή να αναδείξει διαφορές σε ένα σήμα που είναι εξαρτώμενο από το επίπεδο οξυγόνωσης του αίματος (Blood Oxygen Level Dependent). Το σήμα αυτό συνδέεται έμμεσα με την νευρωνική ενεργοποίηση. Η νευρωνική δραστηριότητα συνεπάγεται αύξηση στη ροή αίματος (λειτουργία γνωστή ως αιμοδυναμική απόκριση) (Kim et al. 1999) που συνοδεύεται με μία φυσική αύξηση της συγκέντρωσης οξυαιμοσφαιρίνης (η οποία τρέφει με οξυγόνο τους νευρώνες) γεγονός που εντοπίζεται στο σήμα BOLD.

Ειδικότερα, στη παρούσα εργασία, μας ενδιαφέρει η ανάλυση εγκεφαλικών σημάτων και η ανίχνευση ψυχικών ασθενειών βασιζόμενοι στις διαφοροποιήσεις του σήματος BOLD των FMRI δεδομένων μέσα και ανάμεσα σε εγκεφαλικά δίκτυα. Για την αντιμετώπιση του προβλήματος θα χρησιμοποιήσουμε τη μέθοδο της λειτουργικής διασυνδεσιμότητας δικτύου (Functional Connectivity Network). Η μέθοδος αυτή στην ουσία χρησιμοποιεί κάποια μεγέθη λειτουργικής διασυνδεσιμότητας που δείχνουν αν υπάρχουν διαφορές στη διασυνδεσιμότητα του εγκεφάλου που μπορεί να οφείλονται σε κάποια ασθένεια η κάποια ψυχική διαταραχή (Biswal et al. 1995, Van De Ven et al. 2004). Η μέθοδος αυτή δηλαδή ερευνά αν υπάρχουν διαφορές στη διασυνδεσιμότητα ανατομικών περιοχών του εγκεφάλου ή λειτουργικά οριζόμενων δικτύων. Στα πλαίσια αυτής της εργασίας για την κατασκευή των δικτύων λειτουργικής διασυνδεσιμότητας χρησιμοποιούνται γραμμικές και μη γραμμικές μέθοδοι εκμάθησης πολλαπλοτήτων (manifold learning) για την μείωση της διάστασης των FMRI δεδομένων όσο και για την ανάδειξη χαρακτηριστικών που δεν είναι εύκολο να εντοπιστούν σε χώρους υψηλής διάστασης. Πραγματοποιείται για πρώτη φορά η κατασκευή δικτύων διασυνδεσιμότητας με απεικονίσεις διάχυσης και συγκρίνεται η αποτελεσματικότητα τους σε σχέση με άλλες μεθόδους εκμάθησης πολλαπλοτήτων. Επίσης πραγματοποιείται συγκριτική ανάλυση μεθόδων εκμάθησης μηχανών (machine learning) με στόχο την επιτυχή ταξινόμηση ανάμεσα σε υγιείς και ασθενείς με σχιζοφρένεια.

Ωστόσο επειδή το θέμα της εργασίας είναι διεπιστημονικό και χρειάζεται ένα υπόβαθρο σε επιστημονικά πεδία όπως τα εφαρμοσμένα μαθηματικά (γραμμική άλγεβρα, συναρτησιακή ανάλυση, θεωρία γράφων), η στατιστική και η νευροεπιστήμη, κρίθηκε σκόπιμο να υπάρξει ένα εισαγωγικό κεφάλαιο τόσο για την κατανόηση της φύσης των δεδομένων όσο και για κάποιες εισαγωγικές έννοιες από τα επιμέρους πεδία. Τα δεδομένα FMRI είναι κατά κανόνα μεγάλα σε μέγεθος, θορυβώδη (χρειάζονται προ-επεξεργασία) και η εξαγωγή του σήματος ενδιαφέροντος είναι σχεδόν ένα αυτόνομο πεδίο ερευνητικά.

Αρχικά λοιπόν θα αναφέρουμε κάποια στοιχεία ξεχωριστά για την νευροαπεικόνιση η οποία αποτελεί ένα πολύ σημαντικό μέρος της παρούσας εργασίας, το ποιες είναι οι κύριες μέθοδοι και πως καταγράφονται τα δεδομένα. Επικεντρωνόμαστε κυρίως στη μέθοδο και τα δεδομένα FMRI. Εξετάζουμε τα ποιοτικά τους χαρακτηριστικά, την συνήθη προ-επεξεργασία

που γίνεται, την ερμηνεία τους και τα συγκριτικά πλεονεκτήματα και μειονεκτήματα που παρουσιάζουν σε σύγκριση με άλλες μεθόδους. Στο τέλος της εισαγωγής γίνεται μία σύντομη αναφορά στη θεωρία πολύπλοκων δικτύων και σε έννοιες της θεωρίας γράφων, σε κάποια από τα σημαντικότερα γραφοθεωρητικά μέτρα και δύο από τις σημαντικότερες κλάσεις δικτύων που αποτέλεσαν και σημείο αναφοράς για τη σύγχρονη θεωρία.

Στο 2ο κεφάλαιο της εργασίας παρουσιάζεται το γενικό πρόβλημα της μείωσης διάστασης. Εξετάζονται μία προς μία οι μέθοδοι εκμάθησης πολλαπλοτήτων που στοχεύουν στην μείωση της διάστασης που χρησιμοποιούμε στην μετέπειτα συγκριτική ανάλυση. Οι μέθοδοι εκμάθησης πολλαπλοτήτων που χρησιμοποιούνται στην εργασία είναι η κλασική μέθοδος της πολυδιάστατης κλιμακοποίησης (Multi-Dimensional Scaling) (Douglas Carroll και Arabie 1998), η Ισομετρική απεικόνιση χαρακτηριστικών (Isometric feature Mapping) (Tenenbaum et al. 2000) και η μέθοδος των απεικονίσεων διάχυσης (Diffusion Maps) (Coifman et al. 2005).

Στο 3ο και κύριο μέρος της εργασίας γίνεται μία εκτενής αναφορά στις μέχρι σήμερα ερευνητικές προσπάθειες στο πεδίο της διερεύνησης εγκεφαλικής διασυνδεσιμότητας. Η αναφορά αυτή περιλαμβάνει όχι μόνο μεθόδους που χρησιμοποιήθηκαν και διερευνήθηκαν τα προηγούμενα χρόνια αλλά και προσπάθειες που αφορούν την ικανότητα διαχωρισμού και εντοπισμού ανωμαλιών στην εγκεφαλική λειτουργία που σχετίζονται με ψυχικές νόσους όπως η σχιζοφρένεια, το Alzheimer κτλ. Στη συνέχεια, γίνεται αναφορά στη χρήση μεθόδων μείωσης διάστασης και αναγνώρισης προτύπων που έχουν εφαρμοστεί με επιτυχία σε δεδομένα νευροαπεικόνισης και ειδικότερα σε FMRI. Ακολουθεί η παρουσίαση των δεδομένων FMRI με καταγραφή των χαρακτηριστικών και δημογραφικών στοιχείων δείγματος 146 ατόμων από τα οποία οι 72 είναι ασθενείς διαγνωσμένοι με σχιζοφρένεια και 74 από αυτούς είναι υγιείς. Στηριζόμενοι στο δείγμα αυτό και βασιζόμενη στη μέθοδο FCN παρουσιάζουμε βήμα-βήμα την μεθοδολογία που ακολουθήθηκε για την συγκριτική ανάλυση μεθόδων μείωσης διάστασης χρησιμοποιώντας δύο διαφορετικά μέτρα ομοιότητας εγκεφαλικών σημάτων, ένα βασιζόμενο στην πολλαπλή συσχέτιση με καθυστερήσεις (Lagged cross-correlation) και ένα βασιζόμενο στην ευκλείδεια απόσταση (euclidian distance). Η σύγκριση αυτή θα είναι πολυεπίπεδη και αφορά τόσο την ικανότητα των μεθόδων να μειώνουν τις διαστάσεις δεδομένων μεγάλης κλίμακας όσο και την απόδοση τους όσον αφορά την ταξινόμηση και την κατακράτηση της χρήσιμης πληροφορίας η οποία φαίνεται ως μοτίβο διασυνδεσιμότητας με τη βοήθεια γραφοθεωρητικών μέτρων στους υπό εξέταση γράφους.

Στο 4ο κεφάλαιο γίνεται η καταγραφή των αποτελεσμάτων σε κάθε επίπεδο της συγκριτικής ανάλυσης με τη βοήθεια γραφημάτων.

Στο 5ο κεφάλαιο γίνεται η συζήτηση των αποτελεσμάτων με αξιολόγηση των μεθόδων και σύγκριση με τη βιβλιογραφία. Καταγράφονται τα συμπεράσματα καθώς και προτάσεις για ευρύτερη έρευνα στην επιστημονική περιοχή.

Στο 6ο και τελευταίο μέρος της εργασίας θα παραθέσουμε αναλυτικά την βιβλιογραφία που χρησιμοποιήσαμε σε όλα τα κεφάλαια ταξινομημένη σε αλφαβητική σειρά για την διευκόλυνση του αναγνώστη.

1.1 ΝΕΥΡΟΑΠΕΙΚΟΝΙΣΗ

Με τον όρο νευροαπεικόνιση εννοούμε τη χρήση διαφόρων τεχνικών για την έμμεση ή άμεση απεικόνιση της δομής ή λειτουργίας του νευρικού συστήματος του εγκεφάλου. Είναι μία σχετικά νέα επιστημονική γνώση που εμπίπτει σε επιστήμες όπως η Φαρμακολογία, η Νευροεπιστήμη και η Ψυχολογία. Γενικά, η Νευροαπεικόνιση χωρίζεται σε δύο κατηγορίες :

- Την δομική απεικόνιση, που ασχολείται με την απεικόνιση της δομής του νευρικού συστήματος και τη διάγνωση των -μεγάλης κλίμακας- ενδοκρανιακών νόσων (όπως όγκων) και τραυματισμών.

- Την λειτουργική απεικόνιση, που ασχολείται με τη διάγνωση διαταραχών μεταβολισμού και ομοιόστασης (όπως η νόσος Alzheimer, απεικονίζοντας αλλοιώσεις σε μία πιο λεπτή κλίμακα) αλλά και την παροχή δεδομένων στην έρευνα σε πολλά πεδία της επιστήμης όπως η Νευροεπιστήμη και η γνωστική ψυχολογία.

Στην εργασία μας απασχολεί η δεύτερη κατηγορία Νευροαπεικόνισης, εκείνη της λειτουργικής απεικόνισης.

Με τον όρο λειτουργική Νευροαπεικόνιση εννοούμε της χρήση της τεχνολογίας της Νευροαπεικόνισης για να μετρήσουμε/εντοπίσουμε κάποιον παράγοντα της εγκεφαλικής λειτουργίας ή και για να καταλάβουμε τη σχέση ανάμεσα σε συγκεκριμένες περιοχές του εγκεφάλου και συγκεκριμένες πνευματικές ή σωματικές λειτουργίες.

Κάποιες από τις πιο συνηθισμένες σύγχρονες μεθόδους Νευροαπεικόνισης είναι (Otte and Halsband 2006):

- Τομογραφία εκπομπής ποζιτρονίων (PET)
- Λειτουργική απεικόνιση μαγνητικού συντονισμού (FMRI)
- Ηλεκτροεγκεφαλογράφημα (EEG)
- μαγνητικό εγκεφαλογράφημα (MEG)

Από τις παραπάνω μεθόδους οι PET και FMRI μπορούν να μετρήσουν τοπικές αλλαγές της εγκεφαλικής αιματικής ροής που σχετίζονται με τη νευρωνική δραστηριότητα. Οι αλλαγές αυτές αναφέρονται ως ενεργοποιήσεις. Οι περιοχές του εγκεφάλου που ενεργοποιούνται όταν ένα άτομο εκτελεί μια συγκεκριμένη εργασία μπορεί να παίζουν ένα σημαντικό ρόλο για τον εντοπισμό των νευρώνων που συμβάλλουν στη συμπεριφορά . Για παράδειγμα, η ευρεία ενεργοποίηση του ινιακού λοβού τυπικά εμφανίζεται σε διεργασίες που αφορούν την οπτική διέγερση. Αυτό είναι το τμήμα του εγκεφάλου που λαμβάνει σήματα από τον αμφιβληστροειδή και γενικά γνωρίζουμε σήμερα ότι παίζει ένα σημαντικό ρόλο στην οπτική αντίληψη (Warrington 2013).

Άλλες μέθοδοι Νευροαπεικόνισης περιλαμβάνουν την καταγραφή των ηλεκτρικών τάσεων ή μαγνητικών πεδίων όπως για παράδειγμα οι EEG και MEG αντίστοιχα. Διαφορετικές μέθοδοι έχουν συγκριτικά πλεονεκτήματα/μειονεκτήματα και η επιλογή τους γίνεται ανάλογα με το πρόβλημα που προσπαθούμε να λύσουμε. Για παράδειγμα, η MEG μετρά τη δραστηριότητα του εγκεφάλου με υψηλή χρονική ανάλυση (στο επίπεδο του χιλιοστού δευτερολέπτου), αλλά περιορίζεται στην ικανότητά της να εντοπίσει την εν λόγω δραστηριότητα στην ακριβή τοποθεσία στον εγκέφαλο. Η FMRI αντίθετα κάνει μια πολύ καλύτερη δουλειά στον εντοπισμό της εγκεφαλικής δραστηριότητας στη χωρική ανάλυση

(~1-2 χιλιοστά μέγεθος voxel (ελάχιστο κυβικό στοιχείο σε μία 3D φωτογραφία)), αλλά με πολύ χαμηλότερη ανάλυση χρόνου (~ κάποια δευτερόλεπτα)(Otte and Halsband 2006).

Από την άλλη μεριά οι μέθοδοι PET και FMRI μετρούν τις αλλαγές στη σύνθεση του αίματος σε περιοχές των εγκεφάλου. Όμως επειδή οι μετρήσιμες μεταβολές στο αίμα είναι αργές (της τάξης των δευτερολέπτων) (Cohen και Bookheimer 1994), αυτές οι μέθοδοι είναι πολύ χειρότερες σε μετρήσεις της χρονικής εξέλιξης των νευρωνικών ενεργοποιήσεων, αλλά είναι γενικά καλύτερες στον εντοπισμό των περιοχών που συμβαίνουν οι ενεργοποιήσεις.

Στην παρούσα εργασία η μελέτη αφορά ένα benchmark πρόβλημα με δεδομένα λειτουργικής απεικόνισης μαγνητικού συντονισμού (FMRI) και όπως αναλύεται εκτενέστερα στο κεφάλαιο 3 αφορά 74 υγιείς ανθρώπους καθώς και 72 ασθενείς που έχουν διαγνωσθεί με σχιζοφρένεια (άνδρες και γυναίκες σε ηλικίες από 18-65) (Anderson και Cohen 2013). Ο όγκος των δεδομένων αυτών είναι πολύ μεγάλος αφού τα δεδομένα αποτελούνται από μία χρονοσειρά καθώς και 3D εικόνα με μέγεθος εικονοστοιχείου (voxel) 3x3x4 mm^3.

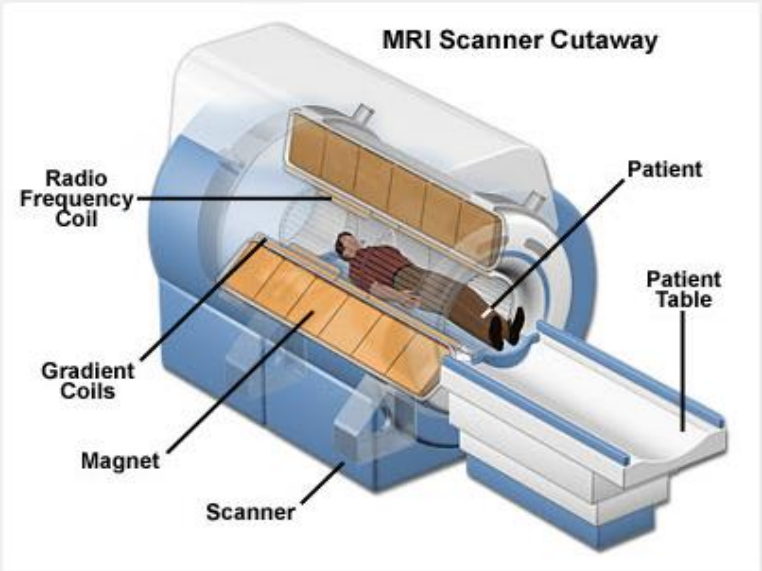
Τα δεδομένα αυτά είναι resting state, δηλαδή δεδομένα που συγκεντρώθηκαν σε κατάσταση ηρεμίας του υποκειμένου, χωρίς αυτό δηλαδή να εκτελεί κάποιο ρητό στόχο (task-related). Εν γένει, αυτού του τύπου τα δεδομένα FMRI μας βοηθούν να κατανοήσουμε τις περιφερειακές αλληλεπιδράσεις που συμβαίνουν στο υποκείμενο όταν αυτό δεν κάνει μία συγκεκριμένη δραστηριότητα. Το υποκείμενο λοιπόν αναπαύεται και παρατηρούνται οι διαφοροποιήσεις στο σήμα κατανάλωσης οξυγόνου στο αίμα σε περιοχές του εγκεφάλου οι οποίες ανιχνεύονται μέσω της μεθόδου FMRI (λεπτομέρειες σχετικά με την μέθοδο δίνονται στην αμέσως επόμενη ενότητα). Επειδή η δραστηριότητα του εγκεφάλου είναι παρούσα ακόμη και εν απουσία μιας συγκεκριμένης δραστηριότητας, κάθε περιοχή του εγκεφάλου έχει αυθόρμητη διακύμανση στο σήμα BOLD. Η προσέγγιση της μεθόδου αυτής (resting state) είναι χρήσιμη για να διερευνηθεί η λειτουργική οργάνωση του εγκεφάλου ή και να εξεταστεί αν υπάρχουν νευρολογικής ή ψυχιατρικής φύσεως ασθένειες σε ασθενείς (Jafri et al. 2008, Yu et al. 2011). Ήδη, η μέθοδος αυτή έχει αποκαλύψει μια σειρά από δίκτυα του εγκεφάλου τα οποία βρίσκονται συνεχώς σε υγιή άτομα και αφορούν π.χ. τα διάφορα στάδια της συνείδησης (Heine et al. 2012) (ακόμα και μεταξύ άλλων ειδών) και αντιπροσωπεύει συγκεκριμένο μοτίβο της σύγχρονης ερευνητικής δραστηριότητας.

Στο επόμενο υποκεφάλαιο λοιπόν θα αναλύσουμε συγκεκριμένα τη μέθοδο FMRI και τα δεδομένα που καταγράφονται.

1.2 Μεθοδος και δεδομενα FMRI

Η Λειτουργική Απεικόνιση Μαγνητικού Συντονισμού (Functional Magnetic Resonance Imaging) απεικονίζει την αιμοδυναμική αντίδραση που σχετίζεται με τη νευρωνική δραστηριότητα στον εγκέφαλο και το Νωτιαίο Μυελό. Είναι σχετικά πρόσφατη μέθοδος Νευροαπεικόνισης και βασίζεται αφενός στο φαινόμενο πυρηνικού μαγνητικού συντονισμού (Nuclear Magnetic Resonance) όπου ένας πυρήνας μπορεί να απορροφήσει και να εκπέμψει ηλεκτρομαγνητική ακτινοβολία σε μία συγκεκριμένη συχνότητα συντονισμού (resonant frequency) και αφετέρου στην θεμελιώδη υπόθεση ότι αύξηση στην νευρωνική δραστηριότητα συμπίπτει με αύξηση στην εγκεφαλική αιματική ροή (Kim et al. 1999).

Από τη δεκαετία του 1890, χάρη στους Roy και Sherrington (Roy και Sherrington 1890) ήταν γνωστό ότι αλλαγές στην αιμοδυναμική απόκριση του εγκεφάλου συνδέονται στενά με τη νευρωνική δραστηριότητα. Οι νευρώνες καταναλώνουν το οξυγόνο που μεταφέρουν οι αιμοσφαιρίνες (πρωτεΐνες του αίματος) των ερυθροκυττάρων (ερυθρά αιμοσφαίρια) από γειτονικά τριχοειδή αγγεία. Η τοπική αντίδραση σε αυτή την κατανάλωση οξυγόνου είναι μια αύξηση της ροής αίματος στις περιοχές αυξημένης νευρωνικής δραστηριότητας (Kim et al. 1999), που συμβαίνει με καθυστέρηση περίπου 1-5 δευτερολέπτων. Αυτή η αιμοδυναμική απόκριση γίνεται μέγιστη μετά τα 4-5 δευτερόλεπτα και στη συνέχεια επιστρέφει στην φυσιολογική τιμή (συχνά ξεπερνώντας την ελαφρώς) κάτι που θα αναλύσουμε εκτενέστερα σε επόμενο υποκεφάλαιο. Αυτή η απόκριση έχει ως συνέπεια τοπικές αλλαγές στις σχετικές συγκεντρώσεις οξυαιμοσφαιρίνης (αιμοσφαιρίνη ενωμένη με οξυγόνο) και δεοξυαιμοσφαιρίνης (αιμοσφαιρίνη που δεν έχει δεσμευμένο οξυγόνο). Οι αλλαγές αυτές επιφέρουν εκ νέου μεταβολή στον όγκο του αίματος στην περιοχή, γεγονός που επιφέρει και μεταβολή στη ροή του αίματος (Cohen και Bookheimer 1994). Ο κυλινδρικός σωλήνας του MRI σκάνερ (Εικόνα 1.1) έχει στο εσωτερικό του έναν πανίσχυρο ηλεκτρομαγνήτη (Magnet). Ένα τυπικό ερευνητικό σκάνερ τέτοιου τύπου έχει ένα μαγνητικό πεδίο στα 3 Tesla (T), περίπου 50,000 φορές μεγαλύτερο από το βαρυτικό πεδίο της γης. Το μαγνητικό πεδίο μέσα στο σκάνερ επηρεάζει το μαγνητικό πυρήνα των ατόμων. Σε τυπικές συνθήκες οι πυρήνες των ατόμων είναι τυχαία οριοθετημένοι αλλά υπό την επιρροή του μαγνητικού πεδίου οι πυρήνες ευθυγραμμίζονται με την κατεύθυνση του πεδίου. Όσο πιο δυνατό είναι το μέτρο του πεδίου τόσο μεγαλύτερος ο βαθμός ευθυγράμμισης. Αφού επιτευχθεί η ευθυγράμμιση των πυρήνων, εφαρμόζεται και άλλο ένα μαγνητικό πεδίο από πηνία κλίσης (gradient coils), το μαγνητικό πεδίο κλίσης (gradient field),



Εικόνα 1.1 *Αναπαράσταση ενός μαγνητικού τομογράφου και των δομικών μερών του για την παραγωγή δεδομένων FMRI [www.cs.washington.edu].*

που εφαρμόζεται για να εντοπίσει χωρικά διαφορετικούς πυρήνες. Έτσι υπάρχει και ένα δεύτερο μαγνητικό πεδίο που δεν είναι ίσου μέτρου κατά μήκος του εγκεφάλου. Ο λόγος είναι ότι όπως γνωρίζουμε από το φαινόμενο NMR, αν προκαλέσουμε κάποια ραδιοκύματα με συγκεκριμένη συχνότητα ίση με τη συχνότητα συντονισμού (resonant frequency, που είναι ανάλογη με τη δύναμη του μαγνητικού πεδίου) για κάθε πυρήνα, τότε αυτός μπορεί να τα απορροφήσει. Όταν ωστόσο σταματήσουμε την εκπομπή εκείνος μπορεί να εκπέμψει πίσω

σαν μία ηχώ (αυτός είναι ο λόγος που συνήθως ένας FMRI σαρωτής λέγεται echo-planar) ραδιοκύματα με συχνότητα που βασίζεται στην δύναμη του μαγνητικού πεδίου. Έτσι αν είχαμε μόνο ένα μαγνητικό πεδίο δεν θα μπορούσαμε να πετύχουμε την εκπομπή όλων των πυρήνων σε μία δεδομένη συχνότητα *X*, αφού η αιμοσφαιρίνη είναι διαμαγνητική ενώ η δεοξυαιμοσφαιρίνη παραμαγνητική (Cohen και Bookheimer 1994). Η παρουσία οποιασδήποτε ουσίας εντός ενός μαγνητικού πεδίου μεταβάλλει το μαγνητικό πεδίο της περιοχής σε ένα συγκεκριμένο βαθμό. Επομένως σε περιοχές όπου θα υπήρχε μεγαλύτερη συγκέντρωση δεοξυαιμοσφαιρίνης οι πυρήνες δεν θα εξέπεμπαν πίσω τα ραδιοκύματα στη συχνότητα *X* . Άρα, χρησιμοποιώντας το μαγνητικό πεδίο κλίσης καταφέρνουμε να αλλάξουμε το μέτρο του μαγνητικού πεδίου κατά μήκος του εγκεφάλου. Αν λοιπόν στη συνέχεια εκπέμψουμε ένα λευκό θόρυβο σε όλο το εύρος των συχνοτήτων (RF pulse) με τη βοήθεια του πηνίου ραδιοσυχνοτήτων (radio frequency coil) αλλάζουμε τις οριοθετήσεις των πυρήνων και από τις ανταποκρίσεις των τελευταίων γνωρίζουμε την ακριβή τοποθεσία τους αφού η συχνότητα που εκπέμπουν δίνεται από τη μαγνητική δύναμη στην οποία υπόκεινται. Επομένως, μετά μπορούμε με ακρίβεια να υπολογίσουμε την τοποθεσία της κάθε ανταπόκρισης που είναι συνάρτηση της δύναμης του μαγνητικού πεδίου.

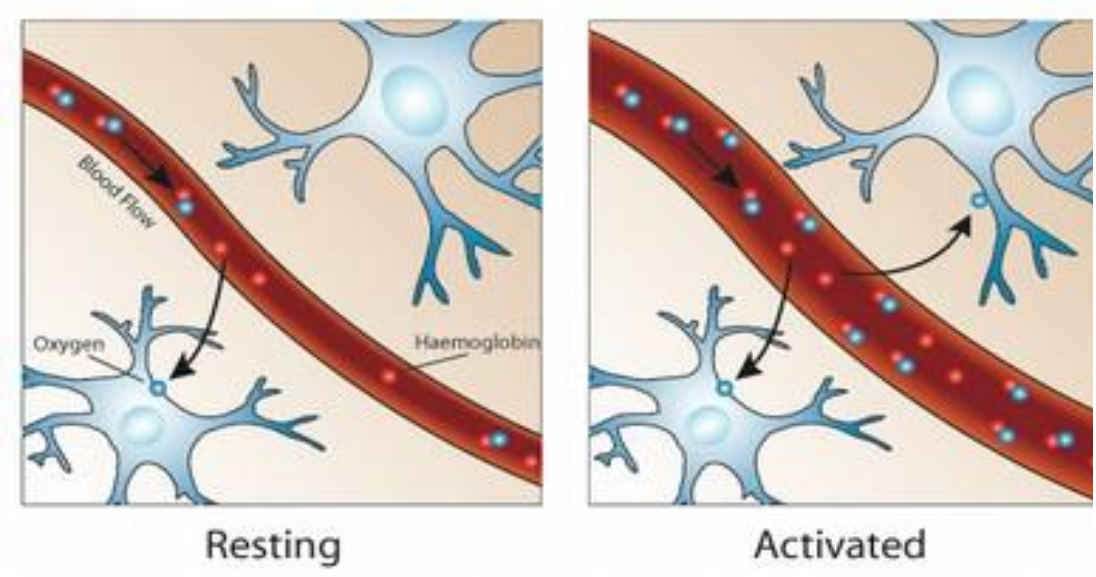
Στην παραπάνω διαδικασία είναι το μαγνητικό σήμα από πυρήνες υδρογόνου που εντοπίζεται. Το γεγονός αυτό οφείλεται στο ότι το ανθρώπινο σώμα αποτελείται σε μεγάλο βαθμό από νερό. Όπως αναφέραμε και παραπάνω η παρουσία οποιασδήποτε ουσίας εντός ενός μαγνητικού πεδίου μεταβάλλει το μαγνητικό πεδίο σε ένα συγκεκριμένο βαθμό. Η διαφορά στις μαγνητικές ιδιότητες οδηγεί σε μικρές διαφορές στα MR σήματα του αίματος που εξαρτώνται από το βαθμό οξυγόνωσης της περιοχής (Cohen και Bookheimer 1994). Συγκεκριμένα περιοχές στις οποίες υπάρχει υψηλή συγκέντρωση σε δεοξυαιμοσφαιρίνη έχουν μεγαλύτερο βαθμό εξασθένησης του δείκτη T_2^* και τα MR σήματα είναι πιο εξασθενημένα.

Γενικά ο μαγνητικός τομογράφος ανάλογα με τις ρυθμίσεις του μπορεί να λάβει διαφορετικά δεδομένα. Η λογική είναι ότι ρυθμίζουμε έτσι τον τομογράφο ώστε να μεγιστοποιήσουμε τις διαφορές σε συγκεκριμένο δείκτη που εξετάζουμε με απώτερο σκοπό να πάρουμε πληροφορίες για διαφορετικούς ιστούς . Σημαντικότεροι δείκτες, είναι ο δείκτης T_1 που αφορά το χρόνο διαμήκους χαλάρωσης των πυρήνων (μετά την απόκλιση από το μαγνητικό πεδίο λόγω των ραδιοκυμάτων), ο δείκτης T_2 που αφορά την εγκάρσια χαλάρωση (ουσιαστικά μετράει το πόσο γρήγορα οι πυρήνες εκπέμπουν ενέργεια μέχρι να επιτευχθεί και πάλι ισορροπία) και ο δείκτης T_2^* που σχετίζεται στενά με τον T_2 αλλά αφορά την ανίχνευση ανομοιογένειας στο τοπικό μαγνητικό πεδίο των πυρήνων αλλά και τις αλληλεπιδράσεις μεταξύ των μορίων.

Βάσει αυτών των δεικτών ο μαγνητικός τομογράφος παράγει εικόνες με διαφορετική στόχευση την κάθε φορά. Για παράδειγμα, αν μας ενδιαφέρει να εντοπίσουμε ανατομικές ανωμαλίες στον εγκέφαλο χρησιμοποιούμε εικόνες (δομικές εικόνες του εγκεφάλου) βασιζόμενοι στον δείκτη T_1. Ο δείκτης που σχετίζεται με τα FMRI είναι κυρίως ο T_2^* και δηλώνει το βαθμό εξασθένησης του MR σήματος (κυρίως λόγω ανομοιογένειας στο τοπικό μαγνητικό πεδίο των πυρήνων) ανιχνεύοντας έτσι έμμεσα την νευρωνική δραστηριότητα μίας εγκεφαλικής περιοχής. Αφού η οξυγόνωση του αίματος (BOLD) διαφέρει ανάλογα με τα επίπεδα της νευρωνικής δραστηριότητας, οι διαφορές στα MR σήματα ανάλογα με τις περιοχές (εξασθενημένο σήμα $\Longrightarrow$ περιοχή με υψηλή συγκέντρωση δεοξυαιμοσφαιρίνης και το αντίθετο) αυτές μπορούν να χρησιμοποιηθούν για να ανιχνεύεται εγκεφαλική δραστηριότητα. Οξυγόνο διανέμεται στους νευρώνες από την αιμοσφαιρίνη σε τριχοειδή ερυθρών αιμοσφαιρίων. Όταν η νευρωνική δραστηριότητα

αυξάνεται τότε υπάρχει αύξηση στη ροή του αίματος (Εικόνα 1.2) στις περιοχές με αυξημένη νευρωνική δραστηριότητα (Kim et al. 1999).

Επομένως για μία περιοχή με εξασθενημένο MR σήμα αυτό θα σημαίνει ότι υπάρχει υψηλή συγκέντρωση δεοξυαιμοσφαιρίνης (παραμαγνητικές ιδιότητες μεγαλύτερη εξασθένηση για το T_2^*) άρα αυτό θα κάνει την απεικόνιση στο συγκεκριμένο εικονοστοιχείο να είναι πιο σκοτεινή ενώ σε αντίθετη περίπτωση πιο λαμπερή. Με τον τρόπο αυτό διαμορφώνονται πολλές εικόνες ανάλογα με το πείραμα που γίνεται για κάποια συγκεκριμένη ενέργεια του υποκειμένου με αισθητικοκινητικές λειτουργίες (task related) ή και σε κατάσταση ηρεμίας (resting state).



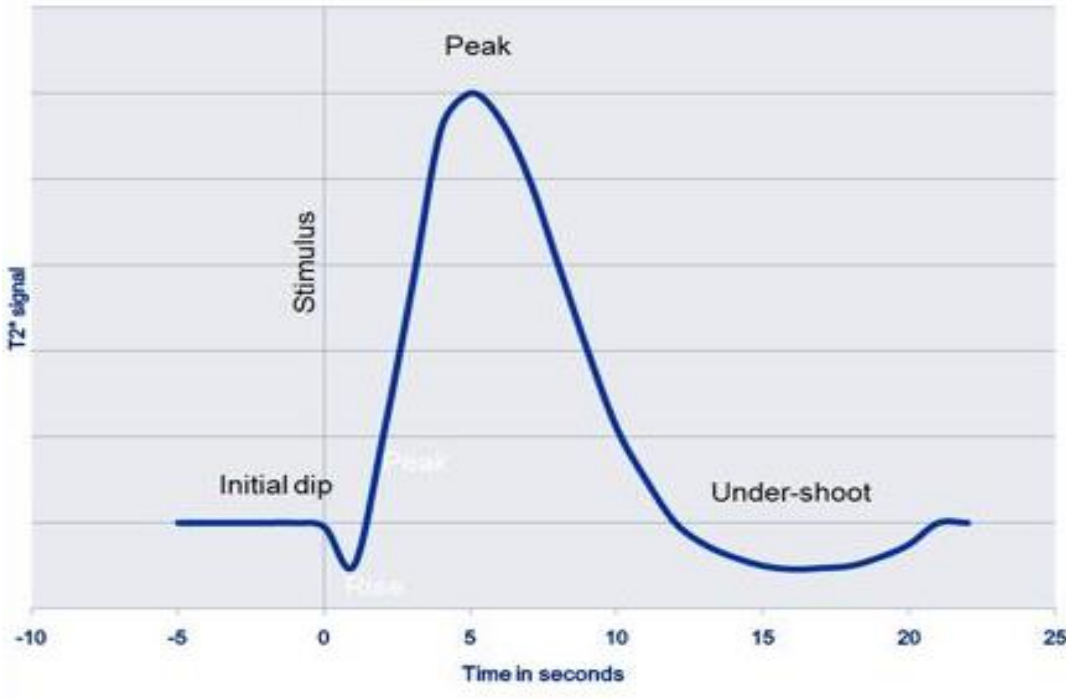
Εικόνα 1.2 *Αναπαράσταση μίας εγκεφαλικής περιοχής κατά τη χαλάρωση και ενεργοποίηση των νευρώνων και ο αντίκτυπος στη ροή και την οξυγόνωση του αίματος [www.psychcentral.com].*

1.2.1 Συμπεριφορά του σήματος BOLD

Αξιοσημείωτη επίσης είναι η κατεύθυνση της μεταβολής οξυγόνωσης του αίματος (Εικόνα 1.3) για μία έντονη δραστηριότητα. Ενώ κανείς θα μπορούσε ενστικτωδώς να περιμένει να μειωθεί η οξυγόνωση με την μείωση της δραστηριότητας (δεδομένου ότι οι νευρώνες «τρέφονται» με οξυγόνο), στην πραγματικότητα τα πράγματα είναι πιο πολύπλοκα. Υπάρχει λοιπόν μία στιγμιαία μείωση της οξυγόνωσης του αίματος ακριβώς μετά από αυξημένη νευρωνική δραστηριότητα, καλούμενη και ως “the initial dip” (αρχική βουτιά) στην αιμοδυναμική απόκριση. Αυτή η μείωση ακολουθείται στη συνέχεια από μία περίοδο 4 περίπου δευτερολέπτων όπου η ροή του αίματος αυξάνεται, όχι μόνο σε επίπεδο όπου ικανοποιείται η ανάγκη σε οξυγόνο, αλλά σε επίπεδο που υπερκαλύπτεται η ζήτηση αυτή. Πρακτικά, αυτό σημαίνει ότι η οξυγόνωση του αίματος αυξάνεται πραγματικά μετά την ενεργοποίηση των νευρώνων. Η ροή του αίματος φτάνει στη μέγιστη τιμή της μετά από περίπου 6 δευτερόλεπτα και στη συνέχεια επιστρέφει πίσω στην αρχική της κατάσταση (συχνά ξεπερνώντας την ελαφρώς), συχνά συνοδευόμενη από μία μετά-διεγερτική υποτίναξη (post-stimulus undershoot).

1.2.2 Ποιοτικά και ποσοτικά χαρακτηριστικά των FMRI δεδομένων

Τα αποτελέσματα του BOLD μετρώνται κάνοντας ογκομετρική λήψη εικόνων με αντίθεση σταθμισμένη (ρυθμίσεις τομογράφου) με βάση τις παραμέτρους χαλάρωσης T_2 και T_2^* . Οι εικόνες αυτές έχουν συνήθως από μέτρια έως πολύ καλή ανάλυση. Πρόσφατες πρόοδοι στην τεχνική, με χρήση υψηλότερων μαγνητικών πεδίων και πολυκαναλικών ανιχνευτών ραδιοκυμάτων έχουν ανεβάσει την χωρική διακριτότητα μέχρι και το επίπεδο του χιλιοστού. Γενικά οι εικόνες FMRI λαμβάνονται κάθε 1-4 δευτερόλεπτα και κάθε στοιχείο όγκου της εικόνας (voxel) έχει μέγεθος 1-6 χιλιοστά σε κάθε ακμή. Παρόλο που μπορούν να διαχωριστούν χρονικά ερεθίσματα μέχρι και 2 δευτερόλεπτα, ο ολικός χρόνος της απόκρισης BOLD σε ένα σύντομο ερέθισμα διαρκεί περίπου 15 δευτερόλεπτα (για να θεωρούνται τυχαίες οι συνθήκες για την επόμενη μέτρηση). Αφού συλλεχθούν όλες οι φωτογραφίες, οι οποίες συλλέγονται σε φέτες του εγκεφάλου (slices) με πολλαπλές σαρώσεις, αποθηκεύονται στον 2D χώρο Fourier (στο χώρο των συχνοτήτων) σαν δεδομένα από ραδιοκύματα επί το χρόνο. Όσο καλύτερη γίνεται η ανάλυση της φωτογραφίας τόσο περισσότερες σαρώσεις πρέπει να γίνουν σε περισσότερες «φέτες» του εγκεφάλου. Στη συνέχεια μετά από επεξεργασία αυτών των ακατέργαστων (raw) δεδομένων γίνεται η μεταφορά των εικόνων από το χώρο των συχνοτήτων στον καρτεσιανό χώρο.



Εικόνα 1.3 *Αναπαράσταση της συμπεριφοράς του σήματος BOLD στο χρόνο για μία δραστηριότητα του υποκειμένου μετά από αρχικό ερέθισμα [www.radiopaedia.org].*

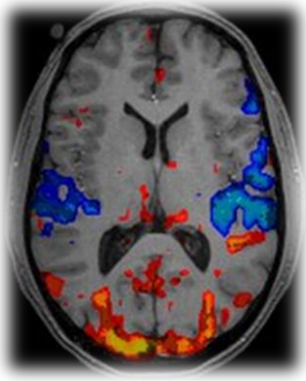
1.2.3 Επεξεργασία των δεδομένων FMRI

Σκοπός της ανάλυσης δεδομένων FMRI είναι η ανίχνευση συσχετίσεων (στο σήμα BOLD) μέσα σε μία ή και ανάμεσα σε εγκεφαλικές περιοχές κατά τη διάρκεια αισθησιοκινητικών λειτουργιών ή ήρεμης κατάστασης. Επειδή όμως το σήμα της FMRI είναι ασθενές και οι διαφορές του BOLD που εντοπίζονται είναι επίσης ασθενείς και μικρές, ο θόρυβος είναι ένα μεγάλο πρόβλημα στα FMRI δεδομένα που μπορεί να αποπροσανατολίσει εύκολα την μετέπειτα ανάλυση.

Σε μια τυπική σάρωση FMRI με χρήση απεικόνισης echo-planar (EPI) που είναι και μία από τις τεχνικές για γρήγορη συλλογή των εικόνων, ο όγκος της κεφαλής απεικονίζεται κάθε 1 ή 2 δευτερόλεπτα, με μέχρι και μερικές χιλιάδες εικόνες. Οι εικόνες λαμβάνονται στο χώρο των συχνοτήτων, και πρέπει να μετασχηματιστούν αντίστροφα στον 3D χώρο για ανάλυση όπως προαναφέρθηκε και παραπάνω. Η φύση της ανίχνευσης και ατέλειες στον ανιχνευτή όπως θερμική κίνηση, αλλά και η κίνηση και η αναπνοή του υποκειμένου παράγουν παραμορφώσεις.

Μετά την ανακατασκευή των εικόνων, η έξοδος του ανιχνευτή είναι μια σειρά 3D εικόνων του εγκεφάλου. Οι συνήθεις παρεμβάσεις που γίνονται στις εικόνες αυτές είναι η διόρθωση κίνησης, θερμικού θορύβου, εξαίρεση τυχαίων ενεργοποιήσεων νευρώνων, η διόρθωση φυσιολογικών συνεπειών (αναπνοή του υποκειμένου κτλ.) καθώς και αφαίρεση σημάτων που δεν προέκυψαν από εγκεφαλικούς ιστούς κτλ. Με χρήση χρονικών φίλτρων επίσης μπορεί να επιχειρηθεί η από-συνέλιξη της απόκρισης BOLD για την ανάκτηση του σήματος στην περίπτωση που αναμένεται νευρωνική δραστηριότητα ενισχύοντας το επιθυμητό σήμα. Το τελευταίο μέρος της επεξεργασίας συχνά περιλαμβάνει μείωση διάστασης των δεδομένων με κλασσικές τεχνικές όπως η ανάλυση σε κύριες συνιστώσες (PCA) (Jolliffe 2002) λόγο του τεράστιου και θορυβώδους όγκου των δεδομένων. Τέλος, μπορεί κανείς να ακολουθήσει ανάλυση σε ανεξάρτητες συνιστώσες (ICA) (Hyvärinen et al. 2000), να χρησιμοποιήσει στατιστικούς παραμετρικούς χάρτες (Friston et al. 1994) (μέθοδος βασιζόμενη στο γενικό γραμμικό μοντέλο (General Linear Model)) και πολλές άλλες μεθόδους για να ανακαλύψει και να απομονώσει περιοχές των δεδομένων που θέλει να μελετήσει ως περιοχές ενδιαφέροντος (Regions Of Interest).

1.2.4 FMRI scan (Χάρτες ενεργοποίησης)



Εικόνα 1.4. *FMRI scan (Χάρτης ενεργοποίησης). [www.psychcentral.com]*

Όπως αναφέραμε και παραπάνω ένας από τους κυριότερους σκοπούς της ανάλυσης FMRI είναι η ανίχνευση λειτουργικής διασυνδεσιμότητας του εγκεφάλου. Η εικόνα 1.4 που φαίνεται στα δεξιά είναι το αποτέλεσμα ενός απλού πειράματος FMRI. Ενώ το υποκείμενο βρίσκεται σε μαγνητικό τομογράφο παρακολουθεί μία οθόνη η οποία δείχνει κάθε 30 δευτερόλεπτα (στο ενδιάμεσο είναι σκοτεινή) ένα οπτικό ερέθισμα. Εν τω μεταξύ, ο μαγνητικός τομογράφος παρακολουθεί το σήμα σε όλο τον εγκέφαλο. Σε περιοχές του εγκεφάλου που ανταποκρίνονται στο οπτικό ερέθισμα περιμένουμε το σήμα να έχει αυξομειώσεις αφού το ερέθισμα ενεργοποιείται και απενεργοποιείται, αν και ελαφρώς θολό από την καθυστέρηση στην απόκριση της ροής του αίματος που είδαμε παραπάνω. Οι ερευνητές λοιπόν, βλέπουν τη δραστηριότητα σε σάρωση, σε εικονοστοιχεία - ή στοιχειώδη εικονοστοιχεία όγκου (voxel). Το voxel είναι το μικρότερο διακριτό σχήμα κουτιού που είναι μέρος μιας τρισδιάστατης εικόνας. Η δραστηριότητα σε ένα voxel συνήθως ορίζεται ως το πόσο κοντά είναι η χρονική σειρά του σήματος λόγω του οπτικού ερεθίσματος με εκείνη που αντιστοιχεί στην τοπική εγκεφαλική δραστηριότητα πριν το ερέθισμα. Voxel του οποίου το σήμα αντιστοιχεί σε μεγάλο βαθμό του δίνεται μια υψηλή βαθμολογία για ενεργοποίηση, ενώ voxel του οποίου το σήμα δεν δείχνει κάποια σημαντική συσχέτιση του δίνεται χαμηλή βαθμολογία και voxel που δείχνει το αντίθετο (απενεργοποίηση) του δίνεται ένα αρνητικό αποτέλεσμα. Βέβαια, για την αποφυγή λάθος

εκτιμήσεων για την ενεργοποίηση των voxel, έχουν κατασκευαστεί πολλά μοντέλα με πολλαπλές συγκρίσεις για την μείωση του στατιστικού λάθους και την εξασφάλιση μίας σίγουρης ενεργοποίησης. Τελικά, αφού έχουμε για κάθε voxel κάποιο βαθμό ενεργοποίησης /απενεργοποίησης, αυτά μπορούν να μετατραπούν σε χάρτες ενεργοποίησης. Οι χάρτες αυτοί, αναπαριστούν τη διέγερση των εγκεφαλικών περιοχών κατά τη διάρκεια μίας συγκεκριμένης λειτουργίας ή και την αλληλεπίδραση μεταξύ των περιοχών αυτών σε κατάσταση ηρεμίας (resting state).

1.2.5 Πλεονεκτήματα και μειονεκτήματα της μεθόδου FMRI

Όπως αναφέραμε και πιο πάνω, καμία μέθοδος Νευροαπεικόνισης δεν είναι ιδανική για όλες τις περιπτώσεις. Η επιλογή της μεθόδου νευροαπεικόνισης που πρόκειται να χρησιμοποιηθεί είναι σε άμεση συνάρτηση με το αντικείμενο και τη φύση της μελέτης που πρόκειται να πραγματοποιηθεί. Ας δούμε λοιπόν παρακάτω συγκεντρωτικά τα χαρακτηριστικά της μεθόδου FMRI (Otte and Halsband 2006):

- Δεν χρειάζονται ραδιενεργοί ανιχνευτές, γεγονός που κάνει τη μέθοδο αυτή μη επεμβατική (non-invasive) για το υποκείμενο και ασφαλής (δεν υπάρχουν γνωστές παρενέργειες μέχρι σήμερα), αντίθετα με άλλες μεθόδους όπως η PET.

- Υψηλή χωρική ανάλυση στα 1-6 χιλιοστά στις περισσότερες εφαρμογές, που θεωρείται πραγματικά πολύ ικανοποιητική για τον εντοπισμό της ακριβής τοποθεσίας της νευρωνικής δραστηριότητας.

- Αρκετά γρήγορη χρονική ανάλυση της τάξης των δευτερολέπτων, ώστε να γίνει διάκριση μεταξύ των δοκιμών (να θεωρηθούν δηλαδή οι συνθήκες τυχαίες για την επόμενη μέτρηση,event related design randomization).

- Χρονική ανάλυση που δεν είναι αρκετά γρήγορη για να γίνει διάκριση μεταξύ προτύπων ενεργοποίησης που συνδέονται με τα διάφορα στάδια της επεξεργασίας ερεθίσματος, αντίθετα με άλλες μεθόδους όπως η EEG.

- Η μέθοδος FMRI μετράει έμμεσα την νευρωνική δραστηριότητα βασιζόμενη στην υπόθεση ότι αύξηση στη νευρωνική δραστηριότητα συνεπάγεται μία αύξηση της αιματικής ροής. Πρακτικά λοιπόν μπορεί να μελετήσει μόνο την αιματική ροή και όχι ατομικά, τα εγκεφαλικά κύτταρα.

- Η μέθοδος FMRI είναι σχετικά ακριβή γεγονός που βάζει περιορισμούς στην ευρεία χρήση τις μεθόδου.

- Κατά τη διάρκεια του πειράματος το υποκείμενο πρέπει να μένει κυριολεκτικά ακίνητο (πολλές φορές μία αλλά και δύο ώρες) κάτι που στην πραγματικότητα είναι αδύνατο, δημιουργώντας τον κίνδυνο θορύβου που είναι ικανός να παραμορφώσει τα δεδομένα σε μεγάλο βαθμό και να αποπροσανατολίσει τη μετέπειτα ανάλυση.

- Δεν μπορεί ένα υποκείμενο να υποβληθεί στο πείραμα αν φέρει εσωτερικά μεταλλικά στοιχεία (όπως ο βηματοδότης και άλλα εμφυτεύματα). Γεγονός που ισχύει και για την απλή απεικόνιση μαγνητικού συντονισμού (μαγνητικός τομογράφος, λόγω του πανίσχυρου μαγνητικού πεδίου).

1.3 ΕΙΣΑΓΩΓΗ ΣΤΗ ΘΕΩΡΙΑ ΔΙΚΤΥΩΝ (NETWORK THEORY)

Ο όρος δίκτυο είναι δυνατόν να επιδέχεται πολλές σημασίες από διαφορετικούς ανθρώπους αφού έχει χρησιμοποιηθεί ευρέως για να περιγράψει συστήματα όπως κοινωνικά δίκτυα, ηλεκτρικά κυκλώματα, οικονομικά δίκτυα, εγκεφαλικά δίκτυα κ.α. Παρόλο που όλα αυτά τα συστήματα μπορεί να μοιάζουν διαφορετικά μεταξύ τους, είναι σημαντικό να κατανοηθεί ότι τα πεδία αυτά μοιράζονται ισχυρά μεθοδολογικά θεμέλια και έχουν κοινές μεθόδους ανάλυσης, μοντελοποίησης και κατανόησης. Όλα αυτά λοιπόν τα δίκτυα, μπορεί να αναπαριστάνονται και να μοντελοποιούνται με το απλό μαθηματικό μοντέλο ενός γράφου.

1.3.1 Βασικά στοιχεία θεωρίας γράφων

Ένας γράφος είναι ένα μαθηματικό αντικείμενο που είναι χρήσιμο για τη λύση πολλών ειδών προβλημάτων. Στην ουσία, ένας γράφος αποτελείται από ένα σύνολο κόμβων και ένα σύνολο ακμών όπου ακμή είναι η σύνδεση που ενώνει δύο κόμβους μεταξύ τους. Μαθηματικά λοιπόν συμβολίζουμε ένα γράφο ως ένα ζευγάρι (V, E) όπου V είναι ένα πεπερασμένο σύνολο και E είναι ένα μία δυαδική σχέση που αντιστοιχείται στο V . V ονομάζεται το σετ των κόμβων και τα στοιχεία του ονομάζονται κόμβοι. E ονομάζεται μία συλλογή ακμών, όπου μία ακμή είναι ένα ζευγάρι (u, v) με $u, v \in V$.Ακολουθούν κάποιες βασικές έννοιες στη θεωρία γραφημάτων που θα βοηθήσουν τον αναγνώστη να παρακολουθήσει τα υπόλοιπα κεφάλαια:

- Γράφος $G = (V, E)$ είναι μία συλλογή κόμβων V που συνδέονται από E συνδέσεις. Ο ορισμός του γράφου είναι κάπως αφηρημένος από πολλές απόψεις. Δεν μας υποδεικνύει τι συμβολίζει κάθε κόμβος ή ακμή. Θα μπορούσαν οι κόμβοι να είναι εγκεφαλικά σήματα, πόλεις ή ακόμα και αεροδρόμια, ενώ οι ακμές να συμβολίζουν αντίστοιχα τη δύναμη του σήματος, την χιλιομετρική απόσταση και τις ώρες πτήσης.

- Μονοπάτι: Ένα μονοπάτι είναι ένας απλός γράφος οπού οι κόμβοι του μπορούν να ταξινομηθούν με τέτοιο τρόπο έτσι ώστε δύο κόμβοι να γειτνιάζουν αν και μόνο αν ο ένας διαδέχεται αμέσως τον άλλον.

- Μη κατευθυνόμενος γράφος: Ένας γράφος για τον οποίο κάθε ακμή συμβολίζει μια μη οριοθετημένη, μεταβατική σχέση μεταξύ δύο κόμβων. Τέτοιες ακμές συμβολίζονται σαν γραμμές που ενώνουν δύο κόμβους χωρίς κατεύθυνση.

- Κατευθυνόμενος γράφος: Ένας γράφος για τον οποίο κάθε ακμή συμβολίζει οριοθετημένη μη ελεύθερη μεταβατική κατάσταση μεταξύ δυο κόμβων. Τέτοιες ακμές με κατεύθυνση, συμβολίζονται συνήθως με βέλη από έναν κόμβο σε έναν άλλο.

- Βαθμός: Ο αριθμός των ακμών που ενώνονται με έναν κόμβο. Πόσες δηλαδή ακμές προσπίπτουν σε έναν κόμβο.

- Ακμή χωρίς βάρος: Οι ακμές που συμβολίζουν σχέσεις μεταξύ κόμβων για τις οποίες θεωρούμε ότι έχουν ίση δυναμικότητα. Τέτοιες ακμές συμβολίζονται συνήθως με απλές γραμμές που έχουν ίσο πάχος.

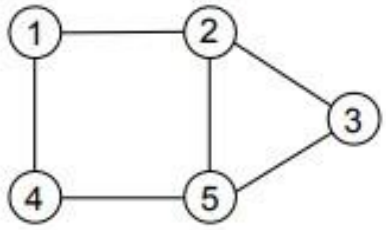
- Ακμές με βάρος: Οι ακμές που συμβολίζουν σχέσεις μεταξύ κόμβων για τις οποίες θεωρούμε ότι έχουν μία συγκεκριμένη τιμή που διαφοροποιείται από άλλες ακμές. Για παράδειγμα, το βάρος κάθε ακμής μπορεί να προκύπτει από μία μέτρηση όπως απόσταση ή χρονική μονάδα. Επομένως αν οι ακμές έχουν βάρος τότε τοποθετούμε τα βάρη $w: E \rightarrow R$

- Δέντρο: Ένας μη κατευθυνόμενος γράφος T για τον οποίο δεν υπάρχουν κύκλοι. Υπάρχει μόνο ένα απλό μονοπάτι μεταξύ οποιονδήποτε κόμβων του γράφου.

- Απλό μονοπάτι: Ένα απλό μονοπάτι είναι ένα μονοπάτι όπου όλοι οι κόμβοι είναι διακεκριμένοι.

1.3.2 Γραφοθεωρητικά μέτρα Γράφου

Για έναν γράφο χωρίς βάρη w και μη κατευθυνόμενο, ο πίνακας γειτνίασης (Εικόνα 1.5) είναι ένας τετραγωνικός πίνακας A ο οποίος αποτελείται από στοιχεία $a_{ij} = a_{ji}$ τα οποία είναι είτε 1 είτε 0 ανάλογα με το αν υπάρχει σύνδεση μεταξύ ενός κόμβου i με έναν κόμβο j. Παρακάτω θα δούμε κάποια από τα πιο σημαντικά γραφοθεωρητικά μέτρα που θα μας χρειαστούν στη συνέχεια της εργασίας.Παρακάτω θεωρούμε τα βάρη w ως δύναμη της σύνδεσης, επομένως μεγαλύτερο βάρος συνεπάγεται μία δυνατότερη/πιθανότερη σύνδεση μεταξύ δύο κόμβων.



$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

Εικόνα 1.5 *Πίνακας γειτνίασης ενός μη κατευθυνόμενου γράφου με $n(V) = 5$ και $n(E) = 6$ ακμές. Η γραμμή i περιγράφει την το μοτίβο διασυνδεσιμότητας του κόμβου i.Για παράδειγμα, η πρώτη γραμμή μας δείχνει ότι ο κόμβος 1 συνδέεται με τους κόμβους 2 και 4. [Hernandez and Van Mieghem 2011]*

Μέσο μήκος μονοπατιού (Average path length)

Ένα τυπικό πρόβλημα που αφορά τα δίκτυα είναι η εύρεση του ελάχιστου μονοπατιού (shortest path length) ως η πιο αποδοτική μετάβαση από ένα σημείο σε ένα άλλο μέσα στο δίκτυο. Σε πολλές πραγματικές καταστάσεις διάφορα χαρακτηριστικά λαμβάνονται υπόψιν και στη συνέχεια λύνονται προβλήματα βελτιστοποίησης που αφορούν το πρόβλημα ελάχιστου μονοπατιού (με την έννοια του ελάχιστου κόστους της μέγιστης ωφέλειας κτλ.). Για το λόγο αυτό δημιουργήθηκαν αποδοτικοί αλγόριθμοι οι οποίοι μπορούν να λύσουν ένα

τέτοιο πρόβλημα γρήγορα και αξιόπιστα όπως είναι ο αλγόριθμος των Dijkstra (Dijkstra 1959), Floyd Warshal (Floyd 1962) κτλ.. Με χρήση των αλγορίθμων αυτών κατασκευάζουμε έναν πίνακα D_G με στοιχεία

$$\mathrm{d_{ij}} = \min \sum_{\mathrm{ij}} \left(\frac{1}{\mathrm{w_{ij}}}\right).$$

Στην περίπτωση που ο γράφος μας δεν έχει βάρη *w* τότε όλα τα βάρη των ακμών είναι ίσα με τη μονάδα και το ελάχιστο μονοπάτι έγκειται στον ελάχιστο αριθμό αλμάτων που χρειαζόμαστε για να προσεγγίσουμε έναν κόμβο j ξεκινώντας από έναν κόμβο i. Το μέτρο που βασίζεται στο πρόβλημα της εύρεσης ελάχιστου μονοπατιού και δίνει πληροφορία για την συνδεσιμότητα και την αποδοτικότητα της ροής πληροφορίας σε έναν γράφο ονομάζεται μέσο μήκος μονοπατιού και συμβολίζεται με L_p.

Το μέσο μήκος μονοπατιού $\mathrm{L_p}$ μπορεί να οριστεί ως η μέση τιμή των ελάχιστων μονοπατιών από κάθε κόμβο ενός γράφου σε έναν άλλο και δίνεται από τη σχέση

$$L_p = \frac{1}{n(V)(n(V)-1)} \sum_{\mathrm{i \neq j}} \mathrm{d_{ij}}.$$

Βάρος ή δύναμη της σύνδεσης

Το βάρος ή δύναμη της σύνδεσης είναι ένα μέτρο που προκύπτει από την γενίκευση του βαθμού του κόμβου ενός γράφου. Έτσι για παράδειγμα μπορεί ένας κόμβος να συνδέεται με πολλές ακμές και να έχει μεγαλύτερο βαθμό από έναν άλλο που συνδέεται με λιγότερες. Τι γίνεται όμως αν αυτές οι λιγότερες συνδέσεις είναι πιο σημαντικές; Σε κάθε περίπτωση και πρόβλημα η ερμηνεία των βαρών ενός γράφου είναι διαφορετική και μας ενδιαφέρει άλλοτε να βρούμε τον κόμβο που δέχεται το μέγιστο βάρος ενώ άλλοτε μας ενδιαφέρει να εντοπίσουμε τον κόμβο με το μικρότερο βάρος. Η εύρεση ενός τέτοιου κόμβου θα μπορούσε ενδεχομένως να μας δώσει πληροφορία σχετικά με το ποιο μέρος του συστήματος που μελετάμε δέχεται μεγαλύτερη ροή πληροφορίας, ποιο μέρος δέχεται τη λιγότερη, πως μοιράζεται η πληροφορία κτλ.

Επομένως, ο βαθμός ενός κόμβου ή το βάρος ενός κόμβου (σε γράφο με βάρη *w*) δίνεται από τη σχέση

$$s_{ij} = \sum_{\mathrm{j}=1}^{n(V)} \mathrm{a_{ij} w_{ij}}$$

όπου a_{ij} είναι στοιχεία του πίνακα γειτνίασης A ενώ τα w_{ij} είναι τα βάρη που αντιστοιχούν σε κάθε ακμή του γράφου. Στην περίπτωση που τα βάρη είναι όλα ίσα με τη μονάδα τότε το s_{ij} ταυτίζεται με το βαθμό του κόμβου.

Συντελεστής συγκρότησης (Clustering Coefficient)

Ο συντελεστής συγκρότησης στη θεωρία γραφημάτων είναι ένα από τα σημαντικότερα μέτρα που περιγράφουν το βαθμό στον οποίο οι κόμβοι ενός γράφου ομαδοποιούνται (σχηματίζουν δομή). Εμπειρικές μελέτες δείχνουν ότι στα περισσότερα δίκτυα του πραγματικού κόσμου και ειδικότερα στα κοινωνικά δίκτυα, οι κόμβοι έχουν την τάση να σχηματίζουν κλειστές υπό-ομάδες που χαρακτηρίζονται από πολύ πυκνές ενώσεις μεταξύ τους.

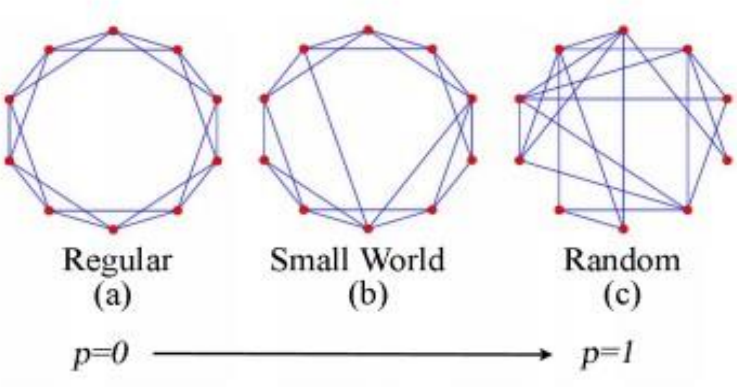
Για έναν γράφο χωρίς βάρη η ποσότητα αυτή μπορεί να υπολογιστεί απλώς σαν τον αριθμό των τριγώνων ενός γράφου προς τον μέγιστο αριθμό των τριγώνων που δύναται να υπάρξουν σε έναν γράφο με κόμβους V. Ωστόσο στη γενική περίπτωση για γράφους με βάρη η χωρίς, ο συντελεστής συγκρότησης ενός γράφου G δίνεται από τη σχέση:

$$C_g = \frac{\sum_{t_c} w_c}{\sum_t w},$$

όπου t_c είναι μία κλειστή τριπλέτα του γράφου ενώ t είναι μία ανοικτή. Κλειστή τριπλέτα είναι μία τριάδα κόμβων η οποία είναι συνδεδεμένη έτσι ώστε να δημιουργεί ένα κλειστό κύκλωμα(τρίγωνο), σε αντίθετη περίπτωση μία ανοικτή τριπλέτα είναι μία τριάδα συνδεδεμένη έτσι ώστε να είναι ανοικτή (σαν ένα τρίγωνο που του λείπει μία ακμή). Το w_c αναφέρεται στον αριθμητικό μέσο των βαρών των ακμών που αποτελούν μία κλειστή τριπλέτα και αντίστοιχα το w για τις ανοικτές. Ο ορισμός αυτός των Opsahl και Panzarasa 2009 είναι μία γενίκευση του μέτρου για γράφους χωρίς βάρη και στην περίπτωση που όλα τα βάρη είναι ίσα με τη μονάδα το μέτρο εκπίπτει στον αρχικό ορισμό.

1.3.3 Δίκτυα μικρού κόσμου και δίκτυα ελεύθερα-κλίμακας (Small-world and Scale-free Networks)

Από τα παραπάνω 3 γραφοθεωρητικά στοιχεία, αναδείχθηκαν 2 από τις πιο σημαντικές κλάσεις δικτύων, τα δίκτυα μικρού κόσμου και τα δίκτυα ελεύθερα-κλίμακας. Ένα δίκτυο μικρού κόσμου όπως ορίστηκε από τους Watts και Strogatz (1998), είναι ένα δίκτυο με μεγάλο συντελεστή συγκρότησης και μικρό μέσο μήκος μονοπατιού. Ένα δίκτυο ελεύθερο-κλίμακας όπως ορίστηκε από τους Barabási και Albert (1999), είναι ένα δίκτυο όπου η κατανομή του βαθμού των κόμβων του ακολουθεί έναν νόμο δύναμης (power law).Τα μοντέλα αυτά αναπτύχθηκαν με σκοπό να εξηγήσουμε πως αυτών των τύπων τα δίκτυα εμφανίζονται στον πραγματικό κόσμο.

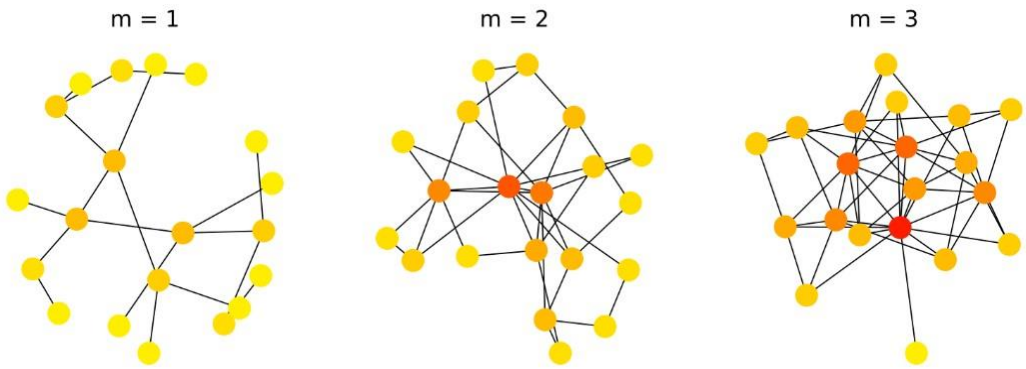


Εικόνα 1.6 *Από ένα κανονικό σε ένα τυχαίο δίκτυο, όπου τυχαία επανασύνδεση κάποιων ακμών σε ένα καθορισμένο δίκτυο παράγει ένα δίκτυο μικρού κόσμου με υψηλό συντελεστή συγκρότησης και μικρό μέσο μήκος μονοπατιού.[Zaidi 2013]*

Ας δούμε λοιπόν το μοντέλο δικτύου μικρού κόσμου όπως αυτό προτάθηκε από τους Watts και Strogatz (1998). Ξεκινάμε με ένα δαχτυλίδι n κόμβων για τους οποίους ισχύει ότι κάθε κόμβος συνδέεται με τους k κοντινότερους γείτονες, για ένα δοσμένο k .Αυτή η διαδικασία σχηματίζει έναν κανονικό (Regular) γράφο όπως φαίνεται στην εικόνα 1.6 (α).Στη

συνέχεια, κάθε ακμή επανατοποθετείται με μία δοσμένη πιθανότητα p για να συνδεθεί με οποιονδήποτε κόμβο στο γράφο. Σε έναν καθορισμένο γράφο, από τη στιγμή που οι γείτονες είναι συνδεδεμένοι ο ένας με τον άλλον, ο συντελεστής συγκρότησης είναι υψηλός. Από την άλλη, το μέσο μήκος μονοπατιού, είναι χαμηλό αφού οι κόμβοι είναι συνδεδεμένοι με τους γείτονες τους. Η τυχαία επανασύνδεση κάποιων κόμβων που ενώνονται με πιο μακρινούς κόμβους μικραίνει περαιτέρω το μέσο μήκος μονοπατιού. Από τη στιγμή που αρκετοί κόμβοι είναι συνδεδεμένοι με τους γείτονες τους, ο συντελεστής συγκρότησης παραμένει υψηλός ενώ το μέσο μήκος μονοπατιού μειώνεται εκ νέου δίνοντας τις ιδιότητες μικρού κόσμου (εικόνα 1.6 (b)).Αν η διαδικασία της τυχαίας επανασύνδεσης συνεχίσει έχουμε τελικά ένα τυχαίο δίκτυο (εικόνα 1.6 (c)) και οι κόμβοι παύουν να συνδέονται μόνο με τους γείτονες τους.

Οι Barabási και Albert (1999) εξήγησαν πως τα ελεύθερα-κλίμακας δίκτυα εμφανίζονται στον πραγματικό κόσμο με ένα άλλο μοντέλο. Αρχικά, έχουμε n κόμβους και καμια ακμή να τους ενώνει. Σε κάθε χρονικό σημείο t ,ένας νέος κόμβος u με m ακμές εισάγεται στο δίκτυο. Αυτές οι ακμές συνδέονται με τους υπάρχοντες κόμβους με πιθανότητα ανάλογη του βαθμού των κόμβων του δικτύου. Στην αρχή που οι κόμβοι δεν είναι συνδεδεμένοι με καμιά ακμή η πιθανότητα είναι ίδια για όλους τους κόμβους. Καθώς το δίκτυο μεγαλώνει, σταδιακά κάποιοι κόμβοι αρχίζουν να έχουν μεγαλύτερο βαθμό από άλλους και επομένως μεγαλύτερη πιθανότητα να συνδεθούν με έναν καινούργιο κόμβο που εισάγεται στο δίκτυο. Επομένως νέοι κόμβοι ‘’προτιμούν’’ να συνδέονται με αυτούς που είναι συνδεδεμένοι με τους περισσότερους κόμβους. Ένα παράδειγμα τριών γράφων με 20 κόμβους (με διαφορετική παράμετρο m για τον καθένα) όπως αυτοί προκύπτουν από το μοντέλο Barabasi-Albert φαίνονται στην εικόνα 1.7. Παρόλο που αυτές οι δύο κλάσεις δικτύων περιγράφονται ξεχωριστά, τα περισσότερα πραγματικά δίκτυα ανήκουν και στις δύο κλάσεις ταυτόχρονα. Ο πιο γενικός όρος πολύπλοκα δίκτυα χρησιμοποιείται συχνά και αναφέρεται σε δίκτυα που ανήκουν σε μία ή και στις δύο κλάσεις ταυτόχρονα. Δίκτυα που δεν είναι ούτε κανονικά, αλλά ούτε και τυχαία.



Εικόνα 1.7 *Παραδείγματα δικτύων ελεύθερων κλίμακας που παρήχθησαν μέσω του μοντέλου των Barabasi και Albert με διαφορετικό αριθμό m συνδέσεων για κάθε νέο κόμβο στο δίκτυο. Η χρωματική διαφορά στους κόμβους αποδίδει το μέγεθος του βαθμού του κάθε κόμβου(με κόκκινο χρώμα ο μεγάλος βαθμός και κίτρινο ο μικρότερος βαθμός όπως φαίνεται στο σχήμα, η κλίμακα είναι ίδια και για τις τρεις περιπτώσεις). [https://en.wikipedia.org/wiki/Barabási–Albert_model]*

2 ΜΕΘΟΔΟΙ ΕΚΜΑΘΗΣΗΣ ΠΟΛΛΑΠΛΟΤΗΤΩΝ: ΤΟ ΠΡΟΒΛΗΜΑ ΤΗΣ ΜΕΙΩΣΗΣ ΔΙΑΣΤΑΣΗΣ (DIMENSIONALITY REDUCTION PROBLEM)

Το πρόβλημα της μείωσης διάστασης μπορεί να οριστεί ως εξής. Ας υποθέσουμε ότι έχουμε ένα σύνολο δεδομένων σε έναν πίνακα $\boldsymbol{X}$ διάστασης $n \times D$ που αποτελείται από n διανύσματα x_i $(i \in \{1,2, \dots, n\}$ με διάσταση D.

Ας υποθέσουμε ακόμη ότι η πραγματική (intrinsic) διάσταση των δεδομένων μας είναι d (όπου $d < D$ και συχνά $d \ll D$).

Εδώ, είναι σημαντικό να σημειωθεί ότι με μαθηματικούς όρους, λέγοντας πραγματική διάσταση εννοούμε ότι τα σημεία μας στο σύνολο δεδομένων $\boldsymbol{X}$ βρίσκονται πάνω ή κοντά σε μία πολλαπλότητα (manifold) με διάσταση d η οποία είναι ενσωματωμένη (embedded) σε έναν χώρο διάστασης D .Οι τεχνικές μείωσης διάστασης μετασχηματίζουν ένα σύνολο δεδομένων $\boldsymbol{X}$ με διάσταση D σε ένα καινούργιο σύνολο δεδομένων $\boldsymbol{Y}$ με διάσταση d $(d < D)$, ενώ παράλληλα διατηρούν την γεωμετρική δομή που υπάρχει στα δεδομένα όσο περισσότερο αυτό είναι δυνατό.

Εν γένει, δεν μας είναι γνωστή η γεωμετρική δομή της πολλαπλότητας (manifold) των δεδομένων ούτε η πραγματική τους διάσταση d. Έτσι, το πρόβλημα της μείωσης διάστασης είναι ένα πρόβλημα που δεν είναι καλά ορισμένο (ill-posed problem) και τις περισσότερες φορές λύνεται με βάση συγκεκριμένες υποθέσεις και ιδιότητες των δεδομένων (όπως πιθανές γεωμετρικές δομές που περιμένουμε να παρουσιάζουν).Παρακάτω ακολουθεί μία παρουσίαση βασικών στοιχείων άλγεβρας και στατιστικής όπου θα βοηθήσει τον αναγνώστη στην καλύτερη και πιο εύκολη ανάγνωση του κεφαλαίου και των μεθόδων που πρόκειται να παρουσιαστούν.

2.1 ΣΤΟΙΧΕΙΑ ΓΡΑΜΜΙΚΗΣ ΑΛΓΕΒΡΑΣ

Έστω πίνακας $\boldsymbol{A}$ διαστάσεων $n \times m$ με $\boldsymbol{A} \in R^{n \times m}$ και $\boldsymbol{A}^T$ ο ανάστροφος του. Το επόμενο θεώρημα είναι ένα από τα βασικότερα θεωρήματα στη γραμμική άλγεβρα.

Θεώρημα 1. Αν $\boldsymbol{A}$ συμμετρικός πίνακας ($\boldsymbol{A}^T = \boldsymbol{A}$) τότε ο $\boldsymbol{A}$ είναι ορθογώνια διαγωνοποιήσιμος και έχει μόνο πραγματικές ιδιοτιμές. Έτσι, υπάρχουν $\lambda_1, \dots, \lambda_n$ ιδιοτιμές και ορθογώνια μη μηδενικά ιδιοδιανύσματα $u_1, \dots, u_n$ τέτοια ώστε για κάθε $i = 1,2, \dots, n$:

$$\boldsymbol{A}u_i = \lambda_i u_i \tag{2.1}$$

Παρατήρηση 1. Πίνακας $\boldsymbol{A}$ διαστάσεων $n \times m$ με $\boldsymbol{A} \in R^{n \times m}$ τότε ο $m \times m$ πίνακας $\boldsymbol{A}\boldsymbol{A}^T$ και ο $n \times n$ πίνακας $\boldsymbol{A}^T\boldsymbol{A}$ είναι συμμετρικοί πίνακες.

Πρόταση 1.Οι πίνακες $\boldsymbol{A}\boldsymbol{A}^T$ και $\boldsymbol{A}^T\boldsymbol{A}$ έχουν τις ίδιες μη μηδενικές ιδιοτιμές

Απόδειξη. Έστω u ένα μη μηδενικό ιδιοδιάνυσμα του πίνακα $\boldsymbol{A}\boldsymbol{A}^T$ με μία μη μηδενική ιδιοτιμή λ.Αυτό σημαίνει ότι :

$$(\boldsymbol{A}^T\boldsymbol{A})u = \lambda u \tag{2.2}$$

Αν πολλαπλασιάσουμε (από αριστερά) και τα δύο μέλη με $\boldsymbol{A}$ έχουμε:

$$\boldsymbol{A}\boldsymbol{A}^T(\boldsymbol{A}u) = \lambda(\boldsymbol{A}u) \tag{2.3}$$

Έχουμε λοιπόν ότι $\boldsymbol{A}u$ είναι ιδιοδιάνυσμα του πίνακα $\boldsymbol{A}\boldsymbol{A}^T$ με ιδιοτιμή λ.

Πρόταση 2.Οι πίνακες $\boldsymbol{A}\boldsymbol{A}^T$ και $\boldsymbol{A}^T\boldsymbol{A}$ έχουν τις ίδιες μη αρνητικές ιδιοτιμές.

Απόδειξη. Έστω u ένα ιδιοδιάνυσμα του πίνακα $\boldsymbol{A}^T\boldsymbol{A}$ με μία ιδιοτιμή λ υπολογίζουμε λοιπόν το μέτρο του $\boldsymbol{A}u$:

$$\left|\left|\boldsymbol{A}u\right|\right|^2 = (\boldsymbol{A}u)^T(\boldsymbol{A}u) = u^T(\boldsymbol{A}^T\boldsymbol{A})u = \lambda u^T u = \lambda\left|\left|u\right|\right|^2 \tag{2.4}$$

Έτσι λοιπόν δείχνουμε ότι το λ είναι θετικός αριθμός ή μηδενική ιδιοτιμή.Αν αντικαταστήσουμε το $\boldsymbol{A}$ με $\boldsymbol{A}^T$ παίρνουμε όμοιο αποτέλεσμα για τον πίνακα $\boldsymbol{A}\boldsymbol{A}^T$.

2.2 ΣΤΟΙΧΕΙΑ ΒΑΣΙΚΗΣ ΣΤΑΤΙΣΤΙΚΗΣ

Αν υποθέσουμε ότι μετράμε μία συγκεκριμένη μεταβλητή A (όπως για παράδειγμα το ύψος από ένα τυχαίο δείγμα ανθρώπων) n φορές. Θα συμβολίσουμε αυτές τις n μετρήσεις σαν $a_1, \dots, a_n$.Η πιο βασική ποσότητα στην στατιστική είναι η μέση τιμή μιας μεταβλητής Α.Ωστόσο στην πράξη σπάνια γνωρίζουμε αυτήν την τόσο σημαντική ποσότητα και έτσι υπολογίζουμε την μέση τιμή μέσω του δειγματικού μέσου (εκτιμητής της μέσης τιμής):

$$E(A) = \mu_A = \frac{1}{n}(a_1 + \cdots + a_n) \tag{2.5}$$

Η μέση τιμή μας δείχνει πού συγκεντρώνονται οι μετρήσεις μας, ενώ η διασπορά μας δείχνει πόσο αυτές οι τιμές διασπείρονται γύρω από τη μέση τιμή. Συνήθως όμως, δεν γνωρίζουμε ούτε την διασπορά της μεταβλητής A και επομένως εκτιμούμε την πραγματική της τιμή μέσω της δειγματικής διασποράς (εκτιμητής της διασποράς):

$$Var(A) = E\{(A^2 - E\{A\})^2\} = \frac{1}{n-1}((a_1 - \mu_A)^2 + \cdots + (\alpha_n - \mu_A)^2) \tag{2.6}$$

Η τυπική απόκλιση είναι συγγενικό μέγεθος της διασποράς και ορίζεται ως $Sdev(A) = \sqrt{Var(A)}$ και ενώ εκφράζει με παρόμοιο τρόπο το πόσο τα δεδομένα απέχουν από την μέση τιμή διατηρεί το πλεονέκτημα απέναντι στη διασπορά να εκφράζεται στις ίδιες μονάδες με τη μεταβλητή Α.

Στην περίπτωση που έχουμε να συγκρίνουμε ανάμεσα σε δύο τυχαίες μεταβλητές A, B είναι φυσικό να αναρωτηθούμε για την σχέση (εδώ γραμμική) που μπορεί να έχουν μεταξύ τους (Για παράδειγμα, θα περίμενε κανείς μία σημαντική σχέση ανάμεσα στο ύψος των ανθρώπων και το βάρος τους. Ένας τρόπος να βρούμε αυτή τη σχέση μεταξύ δύο μεταβλητών είναι η χρήση της δειγματικής συνδιακύμανσης (εκτιμητής της συνδιακύμανσης):

$$Cov(A,B) = E\{A,B\} - E\{A\}E\{B\} \\ = \frac{1}{n-1}\big((a_1-\mu_A)(b_1-\mu_B) + \cdots + (\alpha_n-\mu_A)(b_n-\mu_B)\big) \tag{2.7}$$

Αν η συνδιακύμανση μεταξύ των δύο μεταβλητών είναι αρνητική, αυτό μας δείχνει ότι για παράδειγμα αν η μεταβλητή Α μεγαλώνει τότε αντίστοιχα η μεταβλητή Β τείνει να μικραίνει. Επίσης παρατηρούμε ότι $Cov(A,B) = Cov(B,A)$. Από την τιμή της συνδιακύμανσης μεταξύ δύο μεταβλητών είναι εύκολο να βρούμε και τον συντελεστή συσχέτισης ρ (Pearson correlation coefficient) που ουσιαστικά αποτελεί μία συνάρτηση της συνδιακύμανσης και οι τιμές του είναι κανονικοποιημένες. Ουσιαστικά μας δείχνει το μέτρο γραμμικής συσχέτισης που παρουσιάζουν δύο μεταβλητές με το 1 ή -1 να είναι απόλυτα γραμμική συσχέτιση (είτε αρνητική είτε θετική) ενώ το 0 καθόλου γραμμική συσχέτιση. Δίνεται από:

$$\rho = \frac{Cov(A,B)}{Sdev_A Sdev_B} \tag{2.8}$$

όπου $-1 \leq p \leq 1$.

2.3 ΓΡΑΜΜΙΚΕΣ ΜΕΘΟΔΟΙ ΕΚΜΑΘΗΣΗΣ ΠΟΛΛΑΠΛΟΤΗΤΩΝ

2.3.1 Ανάλυση σε κύριες συνιστώσες (Principal Component Analysis)

Η ανάλυση σε κύριες συνιστώσες είναι αδιαμφισβήτητα μία από τις πιο δημοφιλείς και δοκιμασμένες γραμμικές μεθόδους μείωσης διάστασης (Jolliffe 2002).Παρακάτω θα δούμε τη βασική ιδέα της μεθόδου.

Βασική Ιδέα

Ας υποθέσουμε ότι έχουμε ένα τυχαίο διάνυσμα $\boldsymbol{x}$ με m μεταβλητές και n παρατηρήσεις.Πως θα μπορούσαμε ενδεχομένως να περιγράψουμε την κατανομή του $\boldsymbol{x}$ με απλούς όρους; Ένας προφανής τρόπος να χαρακτηρίσουμε σε ένα πρώτο επίπεδο την κατανομή θα ήταν ο υπολογισμός της μέσης τιμής για κάθε μεταβλητή. Ωστόσο, η μέση τιμή είναι ροπή πρώτης τάξης. Για να εκτιμήσουμε την εναπομένουσα πληροφορία θα μπορούσαμε να υπολογίσουμε τις διασπορές των μεταβλητών. Παρόλα αυτά, η διαδικασία αυτή δεν χαρακτηρίζει ικανοποιητικά μία πολυπαραγοντική κατανομή (multivariate distribution). Στην Εικόνα 2.1 φαίνεται αυτό που διαισθητικά θα θέλαμε να περιγράψουμε για τα δεδομένα μας. Αν δηλαδή τα σημεία μας σχηματίζουν κάποια δομή και πως αυτά επιμηκύνονται ή κατευθύνονται σε αυτή. Η μέθοδος PCA είναι μία απάντηση στο πρόβλημα αυτό.

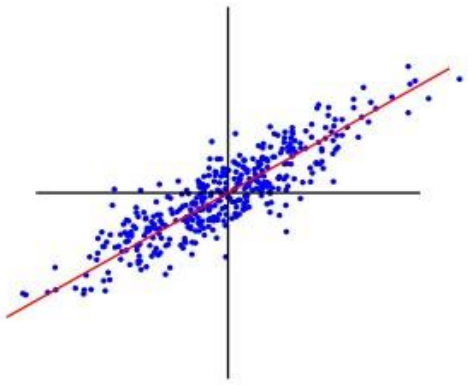
Ας αφαιρέσουμε λοιπόν τη μέση τιμή από το τυχαίο διάνυσμα $\boldsymbol{x}$ έτσι ώστε κάθε x_i να έχει μηδενική μέση τιμή. Μπορούμε να αποθηκεύσουμε σε ένα διάνυσμα $\boldsymbol{\mu}$ τους δειγματικούς μέσους των μεταβλητών μας:

$$\mu = \frac{1}{n}(x_1 + \cdots + x_n)\,, \tag{2.9}$$

Στη συνέχεια αφαιρούμε το μέσο όρο μ από κάθε μεταβλητή x_i. Με τον τρόπο αυτό μεταφέρουμε τον πυρήνα των δεδομένων ώστε το "κέντρο μάζας" να είναι το μηδέν. Η διαδικασία αυτή ονομάζεται κεντροποίηση (data centering).Έστω λοιπόν πίνακας $\boldsymbol{B}$ με μέγεθος $n \times m$ τέτοιος ώστε η $i-οστή$ στήλη να είναι $x_i - \mu$:

$$\boldsymbol{B} = [\,|x_i - \mu| \ldots |x_n - \mu|\,] \tag{2.10}$$

Με τον τρόπο αυτό μπορούμε να συγκεντρωθούμε στην δομή που υπάρχει στο διάνυσμα εκτός του μέσου όρου. Παρακάτω θα θεωρούμε ότι όλες οι μεταβλητές έχουν μηδενική μέση τιμή. Στην πράξη, θα έχουμε γραμμικούς συνδυασμούς των μεταβλητών όπου θα έχουν και αυτοί, μηδενική μέση τιμή.



Εικόνα 2.1 Απόδοση της PCA. Η κύρια συνιστώσα δισδιάστατων δεδομένων είναι η ευθεία που φαίνεται στο σχήμα (κόκκινο χρώμα).Η προβολή των δεδομένων στην κύρια συνιστώσα εξηγεί την περισσότερη διασπορά των δεδομένων από οποιονδήποτε άλλον άξονα.

Οι κύριες συνιστώσες είναι γραμμικοί συνδυασμοί $s = \sum_i w_i x_i$ όπου εγκλωβίζουν (ή εξηγούν) την περισσότερη δυνατή διασπορά από τα αρχικά δεδομένα. Φαίνεται τελικά πως η ποσότητα της διασποράς που εξηγείται σχετίζεται άμεσα με τη διασπορά της κάθε συνιστώσας και αυτό θα εξηγηθεί παρακάτω. Έτσι, η πρώτη κύρια συνιστώσα μπορεί διαισθητικά να ορισθεί σαν εκείνος ο γραμμικός συνδυασμός των μεταβλητών ο οποίος έχει τη μεγαλύτερη διασπορά. Η ιδέα λοιπόν είναι να βρούμε τον κύριο άξονα της δομής των σημείων, όπως απεικονίζεται στην Εικόνα 2.1.

Ωστόσο πρέπει να θέσουμε και κάποιους περιορισμούς στα βάρη w_i τα οποία αποκαλούμε βάρη της κύριας συνιστώσας. Αν δεν υπήρχαν αυτοί οι περιορισμοί, το μέγιστο της διασποράς θα επιτυγχάνονταν όταν όλα τα βάρη w_i θα γίνονταν απείρως μεγάλα (και το ελάχιστο όταν αυτά θα γίνονταν μηδέν.)Στην πράξη, αν πολλαπλασιάζαμε όλα τα βάρη με το 2,θα παίρναμε διασπορά που θα ήταν μεγαλύτερη 4 φορές,ενώ διαιρώντας τους συντελεστές με το 2,η διασπορά θα μίκραινε στο ¼.

Αυτό που είναι φυσικό να κάνουμε είναι να περιορίσουμε τη νόρμα του διανύσματος $\boldsymbol{w} = (w_1, \ldots, w_n)$:

$$||w|| = \sqrt{\sum_i w_i^2} = 1 \tag{2.11}$$

Για λόγους απλότητας, περιορίζουμε τη νόρμα έτσι ώστε να ισούται με 1, αλλά οποιαδήποτε άλλη τιμή δίνει τα ίδια αποτελέσματα. Με τον ορισμό που έχουμε δώσει μέχρι αυτό το σημείο, η διαδικασία θα έδινε μία μόνο κύρια συνιστώσα. Ένας τρόπος για να πάρουμε περισσότερες κύριες συνιστώσες είναι με μία διαδικασία ενός επιπλέον περιορισμού, αυτού της ορθογωνιότητας: Αφού λοιπόν υπολογίσουμε την πρώτη κύρια συνιστώσα, θέλουμε να βρούμε έναν γραμμικό συνδυασμό τέτοιο ώστε να εγκλωβίζει την μέγιστη δυνατή διασπορά ενώ ταυτόχρονα θα είναι ορθογώνιος ως προς τον πρώτο ($< s_1, s_2 >= 0$) .Αυτός ο γραμμικός συνδυασμός λέγεται η δεύτερη κύρια συνιστώσα.Η διαδικασία μπορεί να επαναληφθεί ώστε να γίνει εξαγωγή τόσων κύριων συνιστωσών όσος είναι και ο αριθμός των διαστάσεων των δεδομένων. Με μαθηματικό φορμαλισμό, αν υποθέσουμε ότι έχουμε υπολογίσει κ κύριες συνιστώσες, που δίνονται από διανύσματα βαρών $\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_k$τότε η k+1-οστή κύρια συνιστώσα δίνεται από τη σχέση :

$$\arg\max_{\mathbf{w}} Var\left(\sum_i w_i x_i\right) \tag{2.12}$$

Με τους περιορισμούς,

$$||\boldsymbol{w}|| = \sqrt{\sum_i w_i^2} = 1 \tag{2.13}$$

$$\sum_i w_{ji} w_i = 0 \ \ \mu\varepsilon \ \ \mathrm{j} = 1, \dots, \mathrm{k} \tag{2.14}$$

Μείωση διάστασης με την PCA

Μία εφαρμογή όπου η PCA έχει αποδειχθεί πολύ χρήσιμη είναι εκείνη της μείωσης διάστασης έτσι ώστε το μεγαλύτερο μέρος της διασποράς να διατηρείται μετά από αυτήν. Αντιμετωπίζοντας λοιπόν το γενικό πρόβλημα πρόβλημα ας υποθέσουμε ότι έχουμε ένα μεγάλο αριθμό από τυχαίες μεταβλητές $x_1, \dots, x_m$. Οι ενδεχόμενοι υπολογισμοί για ένα τόσο μεγάλο σύνολο δεδομένων θα μπορούσε σε ορισμένες περιπτώσεις να είναι πολύ δύσκολοι αν όχι και απαγορευτικοί. Θέλουμε λοιπόν να μειώσουμε τη διάσταση των δεδομένων σε έναν μικρότερο αριθμό, ας πούμε n μεταβλητών όπου και θα συμβολίσουμε με $z_1, \dots, z_n$:

$$z_i = \sum_i w_{ij} x_j = 0 \ \ \mu\varepsilon \ \ \mathrm{j} = 1, \dots, \mathrm{n} \tag{2.15}$$

Ο αριθμός των νέων μεταβλητών n μπορεί να είναι μόνο 10% ή ακόμα και 1% του αριθμού m των αρχικών μεταβλητών. Θέλουμε λοιπόν να βρούμε νέες μεταβλητές που να διατηρούν όσο γίνεται μεγαλύτερη ποσότητα της πληροφορίας των δεδομένων γίνεται. Αυτή όμως η ‘’διατήρηση πληροφορίας’’ πρέπει να οριστεί επακριβώς. Ο πιο δημοφιλής ορισμός λοιπόν είναι να κοιτάξει κανείς το τετραγωνικό σφάλμα που παίρνουμε όταν προσπαθούμε να

επαναπαράξουμε τα δεδομένα χρησιμοποιώντας μόνο τις μεταβλητές z_i. Αυτό γίνεται με την επαναπαραγωγή των x_j σαν ένας γραμμικός συνδυασμός $\sum_i a_{ji} z_i$ ελαχιστοποιώντας το μέσο σφάλμα ως εξής:

$$E\left\{\sum_j \left(x_j - \sum_i w_{ij} x_j\right)^2\right\} = E\left\{\left\|x - \sum_i a_i z_i\right\|^2\right\} = 1, \ldots, \mathrm{n} \tag{2.16}$$

Όπου τα a_{ij} υπολογίζονται έτσι ώστε να ελαχιστοποιείται το σφάλμα. Για λόγους απλότητας θα ασχοληθούμε μόνο με μετασχηματισμούς για τους οποίους τα βάρη είναι ορθογώνια και έχουν νόρμα ίση με 1.

$$\sum_j w_{ji}^2 = 1\,,\ για\ όλα\ τα\ i \tag{2.17}$$

$$\sum_i w_{ij} w_{kj} = 0\ \ για\ όλα\ τα\ \ i \neq \mathrm{k} \tag{2.18}$$

Ο καλύτερος τρόπος για να μειώσουμε τη διάσταση των δεδομένων μας είναι να πάρουμε τις πρώτες n κύριες συνιστώσες ($z_i\ με\ i = 1, \ldots, n$). Τα βάρη a_i της σχέσης 2.16 δίνονται επίσης από τις κύριες συνιστώσες z_i.

Σημειώνουμε εδώ ότι η λύση δεν ορίζεται μοναδικά αφού οποιοσδήποτε ορθογώνιος μετασχηματισμός των z_i θα ήταν το ίδιο βέλτιστος. Αυτό είναι κατανοητό αφού κάθε μετασχηματισμός των z_i περιέχει την ίδια πληροφορία.

Λύση της PCA χρησιμοποιώντας αποσύνθεση πίνακα σε ιδιάζουσες τιμές (Eigenvalue Decomposition)

Οι διασπορές και οι συνδιακυμάνσεις των στοιχείων ενός τυχαίου διανύσματος x συχνά συλλέγονται σε έναν πίνακα συνδιακύμανσης (Covariance Matrix) του οποίου το i,j-οστό στοιχείο είναι απλώς η συνδιακύμανση της μεταβλητής x_i και x_j.Ο πίνακας λοιπόν έχει την εξής μορφή:

$$\boldsymbol{C}(\boldsymbol{x}) = \begin{bmatrix} Cov(x_1, x_1) & \cdots & Cov(x_1, x_n) \\ \vdots & \ddots & \vdots \\ Cov(x_n, x_1) & \cdots & Cov(x_n, x_n) \end{bmatrix} \tag{2.19}$$

Σημειώνουμε εδώ ότι η συνδιακύμανση της μεταβλητής x_iμε τον εαυτό της είναι απλώς η διασπορά του x_i. Επομένως η διαγώνιος του πίνακα είναι οι διασπορές των μεταβλητών x_i. Ο πίνακας συνδιακύμανσης αποτελεί γενίκευση της διασποράς σε τυχαία διανύσματα με παραπάνω από μία τυχαία μεταβλητή.

Σε γραφή πινάκων ο πίνακας συνδιακύμανσης δίνεται από τη σχέση:

$$\boldsymbol{C}(\boldsymbol{x}) = E\left\{\boldsymbol{x}\boldsymbol{x}^T\right\} - E\{\boldsymbol{x}\}E\{\boldsymbol{x}\}^{\mathrm{T}} \tag{2.20}$$

Στην περίπτωση μας και επειδή έχουμε αφαιρέσει τη μέση τιμή από τις μεταβλητές x_i ο όρος $E\{x\}E\{x\}^T$ είναι ίσος με μηδέν. Όσον αφορά τον πίνακα διασποράς σημειώνουμε ότι αν οι

μεταβλητές μας είναι ασυσχέτιστες περιμένουμε πως όλα τα στοιχεία του πίνακα συνδιακύμανσης θα είναι 0 πλην της διαγωνίου. Αν επιπλέον οι μεταβλητές μας είναι και κανονικοποιημένες με διασπορά ίση με τη μονάδα, τότε ο πίνακας συνδιακύμανσης ισούται με τον μοναδιαίο πίνακα $\mathbf{I}$.

Μία ακόμα χρήσιμη παρατήρηση είναι ότι η διασπορά ενός οποιουδήποτε γραμμικού συνδυασμού μπορεί να υπολογιστεί χρησιμοποιώντας τον πίνακα συνδιακύμανσης των δεδομένων. Ας πάρουμε για παράδειγμα έναν γραμμικό συνδυασμό $\boldsymbol{w}^T\boldsymbol{x} = \sum_i \boldsymbol{w}_i\boldsymbol{x}_i$ μπορούμε να υπολογίσουμε την διασπορά ως εξής:

$$E\{(\boldsymbol{w}^T\boldsymbol{x})^2\} = E\{(\boldsymbol{w}^T\boldsymbol{x})(\boldsymbol{x}^T\boldsymbol{w})\} = E\{\boldsymbol{w}^T(\boldsymbol{x}\boldsymbol{x}^T\}\boldsymbol{w}\} = \boldsymbol{w}^T E\{\boldsymbol{x}\boldsymbol{x}^T\}\boldsymbol{w} = \boldsymbol{w}^T\boldsymbol{C}\boldsymbol{w} \qquad (2.21)$$

Πράγματι από τη σχέση 2.20 προκύπτει ότι $C = E\{xx^T\}$. Επομένως το βασικό πρόβλημα της μεθόδου PCA μπορεί να περιγραφεί ως:

$$\arg \max_{\mathbf{w}:||\mathbf{w}||=1} \boldsymbol{w}^T\boldsymbol{C}\boldsymbol{w} \qquad (2.22)$$

Με τη χρήση λοιπόν του θεωρήματος 1 και της παρατήρησης 1 (βλέπε Στοιχεία γραμμικής άλγεβρας στην αρχή του κεφαλαίου ενότητα 2.1) ο πίνακας συνδιακύμανσης C μπορεί να εκφραστεί στη μορφή:

$$\boldsymbol{C} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T \qquad (2.23)$$

Όπου $\boldsymbol{U}$ είναι ένας ορθογώνιος πίνακας και $\boldsymbol{D} = diag(\lambda_1, \dots, \lambda_m)$ είναι η διαγώνιος. Οι στήλες του $\boldsymbol{U}$ ονομάζονται ιδιοδιανύσματα και τα λ_i ιδιοτιμές.

Τώρα μπορούμε να λύσουμε το πρόβλημα της PCA εύκολα. Κάνοντας αλλαγή μεταβλητής έχουμε ότι $\boldsymbol{v} = \boldsymbol{U}^T\boldsymbol{w}$ και επομένως έχουμε :

$$\boldsymbol{w}^T\boldsymbol{C}\boldsymbol{w} = \boldsymbol{w}^T\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{w} = \boldsymbol{v}^T\boldsymbol{D}\boldsymbol{v} = \sum_i u_i^2\lambda_i \qquad (2.24)$$

Επειδή το $\boldsymbol{U}$ είναι ορθογώνιο, $||\boldsymbol{v}|| = ||\boldsymbol{w}||$ άρα και ο περιορισμός είναι ο ίδιος για το $\boldsymbol{v}$ όπως και για το $\boldsymbol{w}$.Αν κάνουμε μία ακόμα αλλαγή μεταβλητής, $m_i = v_i^2$ ο περιορισμός ότι το άθροισμα του m_i θα πρέπει να ισούται με 1 (επίσης θα πρέπει να είναι θετικό αφού είναι στο τετράγωνο).Τότε το πρόβλημα μετασχηματίζεται σε:

$$\arg \max_{m_i \geq 0, \sum_i m_i = 1} \sum_i m_i\lambda_i \qquad (2.25)$$

Εδώ είναι φανερό ότι το μέγιστο βρίσκεται όταν η m_i που αντιστοιχεί στην μεγαλύτερη ιδιοτιμή λ_i είναι ίση με 1 και οι άλλες με 0. Στο σημείο αυτό θα συμβολίζουμε με i^* τους δείκτες των μέγιστων ιδιοτιμών. Γυρίζοντας πίσω στα $\boldsymbol{w}$ παρατηρούμε ότι αντιστοιχεί στο i-οστό ιδιοδιάνυσμα που είναι και η i-οστή στήλη του πίνακα $\boldsymbol{U}$**.**

Αφού τα ιδιοδιανύσματα ενός συμμετρικού πίνακα είναι ορθογώνια, ο υπολογισμός της δεύτερης κύριας συνιστώσας σημαίνει μεγιστοποίηση της διασποράς έτσι ώστε το u_{i^*} να είναι μηδέν. Αυτό όμως είναι ισοδύναμο με το να απαιτήσουμε το νέο $\boldsymbol{w}$ να είναι ορθογώνιο στο πρώτο ιδιοδιάνυσμα. Επομένως για το δεύτερο πρόβλημα μεγιστοποίησης έχω ακριβώς το ίδιο πρόβλημα προσθέτοντας μία συνθήκη επιπλέον, ότι το $m_{i^*} = 0$. Με τον τρόπο αυτό εξασφαλίζουμε ότι το $\boldsymbol{w}$ θα είναι ίσο με το ιδιοδιάνυσμα που θα αντιστοιχεί στη δεύτερη κύρια συνιστώσα. Τελικά, όλες οι κύριες συνιστώσες μπορούν να βρεθούν ταξινομώντας τα ιδιοδιανύσματα $\boldsymbol{u}_i, i = 1, \dots, m$ στο $\boldsymbol{U}$ έτσι ώστε οι αντίστοιχες ιδιοτιμές να είναι σε φθίνουσα διάταξη.Επομένως η i-οστή κύρια συνιστώσα s_iθα είναι ίση με:

$$s_i = \boldsymbol{u}_i^T \boldsymbol{x} \tag{2.26}$$

Από την πρόταση 2 (βλέπε Στοιχεία γραμμικής άλγεβρας στην αρχή του κεφαλαίου ενότητα 2.1) σημειώνουμε ότι οι ιδιοτιμές του πίνακα C είναι $\lambda_i \geq 0$. Ακόμη, χρησιμοποιώντας την αποσύνθεση πίνακα σε ιδιάζουσες τιμές μπορούμε να αποδείξουμε και μία άλλη ιδιότητα της PCA, ότι οι κύριες συνιστώσες είναι μεταξύ τους ασυσχέτιστες. Αν λοιπόν το διάνυσμα των κύριων συνιστωσών είναι $s = U^T x$ τότε έχουμε:

$$E\{\boldsymbol{s}\boldsymbol{s}^T\} = E\{\boldsymbol{U}^T\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{U}\} = \boldsymbol{U}^T E\{\boldsymbol{x}\boldsymbol{x}^T\}\boldsymbol{U} = \boldsymbol{U}^T(\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T)\boldsymbol{U} = (\boldsymbol{U}^T\boldsymbol{U})\boldsymbol{D}(\boldsymbol{U}^T\boldsymbol{U}) = \boldsymbol{D} \tag{2.27}$$

Επομένως ο πίνακας συνδιακύμανσης των κύριων συνιστωσών είναι διαγώνιος πίνακας. Το γεγονός αυτό υποδεικνύει ότι οι κύριες συνιστώσες είναι και ασυσχέτιστες μεταξύ τους.
Τέλος, από την σχέση 2.27 βλέπουμε ότι οι διασπορές των κύριων συνιστωσών είναι ίσες με τα λ_i.

Μοναδικότητα της PCA

Το γεγονός ότι οι διασπορές των κύριων συνιστωσών είναι ίσες με λ_i έχει έναν σημαντικό αντίκτυπο στην μοναδικότητα του μετασχηματισμού της μεθόδου. Αν δύο από τις ιδιοτιμές είναι ίσες, τότε η διασπορά των κύριων συνιστωσών είναι ίσες. Ως επακόλουθο, οι κύριες συνιστώσες δεν είναι καλά ορισμένες διότι μπορούμε να κάνουμε μία στροφή αυτών των συνιστωσών χωρίς να υπάρξει διαφορά στις διασπορές. Αυτό συμβαίνει διότι αν z_i και z_{i+1} έχουν ίση διασπορά, τότε οι γραμμική συνδιασμοί όπως $\sqrt{1/2}z_i + \sqrt{1/2}z_{i+1}$και $\sqrt{1/2}z_i - \sqrt{1/2}z_{i+1}$ θα έχουν την ίδια διασπορά.Όλες οι συνθήκες (μοναδιαία διασπορά και ορθογωνιότητα) πληρούνται, επομένως δεν μπορεί να γίνει ταξινόμηση μεταξύ των δύο. Στην γραμμική άλγεβρα είναι γνωστό ότι μοναδικά ορισμένη είναι μία αποσύνθεση πίνακα στην οποία οι ιδιοτιμές είναι όλες διακεκριμένες. Ωστόσο ο υπόχωρος των κύριων συνιστωσών περιέχει όλους τους πιθανούς γραμμικούς συνδυασμούς των n πρώτων κύριων συνιστωσών. Επομένως στην πραγματικότητα τις περισσότερες φορές ο n-διάστατως υπόχωρος των κύριων συνιστωσών ορίζεται μοναδικά παρόλο που κάποιες κύριες συνιστώσες μπορεί να έχουν ίσες διασπορές. Μπορεί να συμβεί βέβαια η n-οστή και η $n+1$-οστή κύρια συνιστώσα να έχουν ίσες διασπορές και εκεί πράγματι δεν θα μπορούσαμε να αποφασίσουμε ποια από τις δυο θα έπρεπε να συμπεριληφθεί στον υπόχωρο. Τελικά, σε πραγματικές εφαρμογές η επίδραση που έχει αυτή η περίπτωση όπου δύο ή περισσότερες κύριες συνιστώσες έχουν την ίδια διασπορά είναι συνήθως πολύ μικρή και στην πράξη μπορεί να αγνοηθεί.

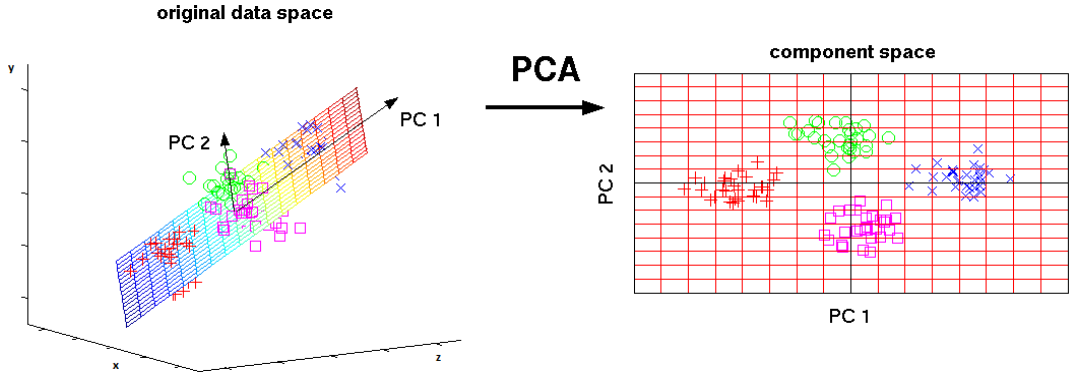
Ποσοστό επεξήγησης διασποράς από τις συνιστώσες

Παρατηρήσαμε παραπάνω ότι οι ιδιοτιμές του πίνακα συνδιακύμανσης δίνουν την διασπορά της κάθε συνιστώσας. Το γεγονός αυτό μεταφράζεται ως η ποσότητα της συνολικής διασποράς των δεδομένων που επεξηγεί/εγκλωβίζει μία κύρια συνιστώσα. Αν εξάγουμε τις m πρώτες κύριες συνιστώσες, όλες μαζί έχουν συνολική διασπορά $\sum_i^m \lambda_i$. Συνήθως όμως η διασπορά που εγκλωβίζεται από τις m κύριες συνιστώσες παρουσιάζεται σαν ποσοστό επί της συνολικής διασποράς των δεδομένων:

$$\textit{Proportion of variance explained} = \frac{\sum_i^m \lambda_i}{\sum_i^n \lambda_i} \tag{2.28}$$

Όπου n είναι η αρχική διάσταση των δεδομένων. Ο παρονομαστής περιέχει τη συνολική διασπορά των δεδομένων του $\boldsymbol{x}$.

Ένα παράδειγμα σε ένα αρχικά τρισδιάστατο χώρο και η προβολή των δεδομένων σε δύο κύριες συνιστώσες φαίνεται παρακάτω στην εικόνα 2.2 για την εποπτεία της μεθόδου.



Εικόνα 2.2 *Οπτική απεικόνιση της μεθόδου PCA.Στα αριστερά βλέπουμε τα δεδομένα στον αρχικό χώρο και τις δύο κύριες συνιστώσες που κατά σειρά εγκλωβίζουν τη μεγαλύτερη διασπορά των δεδομένων. Από αριστερά βλέπουμε τον μετασχηματισμό των δεδομένων στον καινούργιο χώρο χαρακτηριστικών που εκφράζονται μέσω των δύο κύριων χαρακτηριστικών.[Scholz 2006].*

2.3.2 Μέθοδος της πολυδιάστατης Κλιμακοποίησης (Multidimensional Scaling)

Η μέθοδος της πολυδιάστατης κλιμακοποίησης είναι μία μέθοδος μείωσης διάστασης που μας βοηθά να οπτικοποιήσουμε το επίπεδο ομοιότητας κάποιων μεταβλητών m (σε ένα σύνολο δεδομένων $\boldsymbol{X} \in R^{n\times m}$) χρησιμοποιώντας τις αποστάσεις μεταξύ των n παρατηρήσεων. Δοθέντος ενός πίνακα απόστασης $\boldsymbol{D_X} \in R^{m\times m}$ με στοιχεία d_{ij} όπου $i,j = 1,\dots,m$ ο στόχος της μεθόδου είναι να βρει έναν πίνακα σχηματισμού (configuration matrix) των σημείων m $\boldsymbol{Y} \in R^{m\times p}$ σε έναν χαμηλότερης διάστασης χώρο $p < m$ όπου τα σημεία θα έχουν έναν πίνακα απόστασης $\boldsymbol{D_Y} \in R^{m\times m}$ (ευκλείδειο η μη) που θα είναι όσο το δυνατόν πιο κοντά στον αρχικό πίνακα απόστασης.

Θα αναπτύξουμε λοιπόν παρακάτω μία απόδειξη του κλασικού αλγορίθμου πολυδιάστατης κλιμακοποίησης (με ευκλείδιες αποστάσεις) βήμα-βήμα:

Ας υποθέσουμε ότι για ένα σύνολο δεδομένων $\boldsymbol{X} \in R^{n\times m}$ (με x_i ($i \in [1,n]$) να αποτελεί μία στήλη του πίνακα $\boldsymbol{X}$) δεν γνωρίζουμε τον $\boldsymbol{X}$ αλλά ούτε και τη διάσταση των x_i. Ωστόσο γνωρίζουμε τον ευκλείδειο πίνακα απόστασης μεταξύ των σημείων (pairwise distances) $\boldsymbol{D}_X \in R^{m\times m}$. Επομένως κάθε στοιχείο του ευκλείδειου πίνακα μπορεί να εκφραστεί ως:

$$d_{ij}^2 = (x_i - x_j)(x_i - x_j)^T = x_i^T x_i - 2x_i^T x_j + x_j^T x_j \tag{2.29}$$

Η γενική ιδέα της κλασικής μεθόδου MDS είναι :

1. Μετασχηματισμός του πίνακα $\boldsymbol{D_X}$ των ευκλείδειων (ή μη) αποστάσεων στο τετράγωνο σε μορφή εσωτερικού γινομένου.
2. Υπολογισμός της αποσύνθεσης πίνακα σε ιδιάζουσες τιμές του εσωτερικού αυτού γινομένου.

Παρακάτω θα δούμε την υλοποίηση των δύο αυτών βημάτων δοθέντος ενός πίνακα ευκλείδειων αποστάσεων $\boldsymbol{D_X}$:

Ορίζουμε τον πίνακα εσωτερικού γινομένου:

$$\boldsymbol{K} = \boldsymbol{X}^T\boldsymbol{X} \tag{2.30}$$

Ορίζω ως $k = diag(K_{ij}) \in R^m$ και $\mathbf{1} = [1,1 \dots ,1]^T \in R^m$. Επομένως η σχέση 2.29 γράφεται :

$$\boldsymbol{D_X} = \left(d_{ij}^2\right) = k\,\mathbf{1}^T - 2\boldsymbol{K} + \mathbf{1}\,k^{\boldsymbol{T}} \tag{2.31}$$

Στη συνέχεια θα ορίσουμε την πράξη κεντροποίησης πινάκων μέσω ενός πίνακα $\boldsymbol{H} \in R^{m\times m}$**.**Πολλαπλασιάζοντας τον πίνακα $\boldsymbol{H}$ με τον αρχικό πίνακα μπορώ να αφαιρέσω το μέσο όρο από κάθε στήλη ή γραμμή του αρχικού πίνακα $\boldsymbol{D}_X$ ανάλογα με την πράξη του πολλαπλασιασμού από δεξιά ή αριστερά. Πιο συγκεκριμένα ο πολλαπλασιασμός από αριστερά αφαιρεί το μέσο όρο από κάθε στήλη και ο πολλαπλασιασμός από δεξιά αφαιρεί το μέσο όρο από κάθε γραμμή. Ο πίνακας $\boldsymbol{H}$ έχει τη μορφή:

$$\boldsymbol{H} = \boldsymbol{I}_m - \frac{1}{m}\mathbf{1}\mathbf{1}^T \tag{2.32}$$

Πολλαπλασιάζοντας τον αρχικό πίνακα $\boldsymbol{X}$ με τον πίνακα κεντροποίησης $\boldsymbol{H}$ έχω:

$$\boldsymbol{X'} = \boldsymbol{H}\boldsymbol{X} = \boldsymbol{X} - \frac{1}{m}\,\mathbf{1}\,\mathbf{1}^T\boldsymbol{X} \tag{2.33}$$

Ο πίνακας $\boldsymbol{X'}$ είναι ο αρχικός πίνακας έχοντας αφαιρέσει τους μέσους όρους από κάθε στήλη. Η σχέση 2.30 γίνεται:

$$\boldsymbol{K'} = \boldsymbol{X'}^T\boldsymbol{X'} = \left(\boldsymbol{X} - \tfrac{1}{m}\boldsymbol{X}\,\mathbf{1}\,\mathbf{1}^T\right)^T\left(\boldsymbol{X} - \tfrac{1}{m}\boldsymbol{X}\,\mathbf{1}\,\mathbf{1}^T\right) =$$
$$\boldsymbol{K} - \frac{1}{m}\boldsymbol{K}\,\mathbf{1}\,\mathbf{1}^T - \frac{1}{m}\,\mathbf{1}\,\mathbf{1}^T\,\boldsymbol{K} + \frac{1}{m^2}\,\mathbf{1}\,\mathbf{1}^T\,\boldsymbol{K}\,\mathbf{1}\,\mathbf{1}^T \tag{2.34}$$

Αν θέσουμε $\boldsymbol{B} = -\frac{1}{2}\boldsymbol{H}\boldsymbol{D_X}\boldsymbol{H}^T$ τότε ο πίνακας $\boldsymbol{B}$ ισούται με:

$$\boldsymbol{B} = -\frac{1}{2}\boldsymbol{H}\,(k\,\mathbf{1}^T + \mathbf{1}\,k^T - 2\boldsymbol{K})\boldsymbol{H}^T \tag{2.35}$$

Αφού,

$k\,\mathbf{1}^T H^T = k\,\mathbf{1}\,\left(\boldsymbol{I} - \frac{1}{m}\,\mathbf{1}\,\mathbf{1}^T\right) = k\,\mathbf{1} - k\left(\frac{\mathbf{1}^T\mathbf{1}}{m}\right)\mathbf{1} = \mathbf{0}$

έχουμε ότι,

$\boldsymbol{H}\,k\,\mathbf{1}\,\boldsymbol{H}^T = \boldsymbol{H}\,\mathbf{1}\,k^T\,\boldsymbol{H}^T = 0$

Επομένως,

$$\begin{gathered}\boldsymbol{B} = \boldsymbol{H}\boldsymbol{K}\,\boldsymbol{H}^T = \left(\boldsymbol{I} - \frac{1}{m}\,\mathbf{1}\,\mathbf{1}^T\right)\boldsymbol{K}\,\left(\boldsymbol{I} - \frac{1}{m}\,\mathbf{1}\,\mathbf{1}^T\right) = \\ \boldsymbol{K} - \frac{1}{m}\,\mathbf{1}\,\mathbf{1}^T\boldsymbol{K} - \frac{1}{m}\boldsymbol{K}\,\mathbf{1}\,\mathbf{1}^T + \frac{1}{m^2}\,\mathbf{1}(\mathbf{1}^T\boldsymbol{K}\,\mathbf{1})\mathbf{1}^T = \boldsymbol{K}'\end{gathered} \tag{2.36}$$

Όπου από την σχέση 2.34 έχω:

$$\boldsymbol{B_X} = -\frac{1}{2}\boldsymbol{H}\boldsymbol{D_X}\boldsymbol{H}^T = \boldsymbol{X'^T}\boldsymbol{X'} \tag{2.37}$$

Ο πίνακας $\boldsymbol{X'}$ είναι ο αρχικός πίνακας που του έχουμε αφαιρέσει το μέσο όρο κατά στήλη (είναι καλό να σημειωθεί ότι αφού οι αποστάσεις είναι ευκλείδιες μεταξύ των σημείων αυτό δεν επηρεάζει την απόσταση ''σαν μέτρο'' μεταξύ τους).

Ας θυμηθούμε ότι ο στόχος της μεθόδου είναι να βρούμε ένα σύνολο m σημείων ($\boldsymbol{Y}$) σε p διαστάσεις έτσι ώστε ο αντίστοιχος πίνακας απόστασης $\boldsymbol{D}_Y \in R^{m\times m}$ να είναι μία καλή εκτίμηση του αρχικού πίνακα $\boldsymbol{D}_X \in R^{m\times m}$. Το πρόβλημα μπορεί να περιγραφεί και ως:

$$\boldsymbol{D}_Y = argmin_{rank(\boldsymbol{D}_Y \le p)}\left|\left|\boldsymbol{D}_X - \boldsymbol{D}_Y\right|\right|^2 \tag{2.38}$$

Μετά από την διπλή κεντροποίηση των πινάκων $\boldsymbol{X}$ και $\boldsymbol{Y}$ η σχέση 2.38 γίνεται:

$$\boldsymbol{B_Y} = argmin_{rank(\boldsymbol{B}_Y \le p)}\left|\left|\boldsymbol{B}_X - \boldsymbol{B}_Y\right|\right|^2 = \left|\left|\boldsymbol{X'}^T\boldsymbol{X'} - \boldsymbol{Y'}^T\boldsymbol{Y'}\right|\right|^2 \tag{2.39}$$

Το παραπάνω πρόβλημα είναι ένα γνωστό πρόβλημα βελτιστοποίησης που λύνεται με αποσύνθεση πίνακα με ιδιάζουσες τιμές (Singular Value Decomposition) του πίνακα $\boldsymbol{B_X}$. Η δεύτερη ισότητα στην (2.39) οφείλεται στο γεγονός ότι κάναμε διπλή κεντροποίηση και στον πίνακα **Y** που ψάχνουμε. Με αυτόν τον τρόπο γνωρίζουμε ότι η ενσωμάτωση (embedding) που θα πάρουμε θα είναι και αυτή κεντροποιημένη**.**

Επομένως,

$$\boldsymbol{B}_X \approx \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T = \left(\left(\boldsymbol{U}\boldsymbol{D}_{\boldsymbol{e}}^{\frac{1}{2}}\right)\left(\boldsymbol{D}_{\boldsymbol{e}}^{\frac{1}{2}}\boldsymbol{U}^T\right)\right) = \boldsymbol{Y}'^T\boldsymbol{Y}' \tag{2.40}$$

Όπου $\boldsymbol{U}$ είναι ένας πίνακας διάστασης $m \times p$ και ο $\boldsymbol{D}_{\boldsymbol{e}}$ είναι ένας διαγώνιος πίνακας με τις p μεγαλύτερες ιδιοτιμές στη διαγώνιο και ο $\boldsymbol{Y}' = \boldsymbol{D}_{\boldsymbol{e}}^{\frac{1}{2}}\boldsymbol{U}^T$ είναι ένας $k \times N$ πίνακας. Τελικά έχουμε μία ενσωμάτωση m μεταβλητών σε $p < m$ διαστάσεις σαν στήλες του με την ιδιότητα ότι έχουν μηδενική μέση τιμή η κάθε μία. Η ενσωμάτωση αυτή με διάσταση p διατηρεί όσο το δυνατόν καλύτερα τις ευκλείδιες αποστάσεις μεταξύ των σημείων στη αρχική μορφή $\boldsymbol{D}_X$.

Παρατηρήσεις:

- Η μέθοδος της πολυδιάστατης κλιμάκωσης λειτουργεί βάση της θεωρίας όταν ο πίνακας των αποστάσεων είναι ευκλείδειος. Αν ο αρχικός πίνακας δεν είναι ευκλείδειος τότε ο πίνακας $\boldsymbol{Y}$ της μεθόδου δεν μπορεί ακριβώς να αναπαράγει τον πίνακα $\boldsymbol{X}$.
- Αν στο τελευταίο βήμα του αλγορίθμου και κατά την αποσύνθεση του πίνακα σε ιδιάζουσες τιμές παρατηρήσουμε αρνητικές ιδιοτιμές τότε αυτό θα σημαίνει ότι οι αρχικές αποστάσεις μας δεν είναι ευκλείδειες. Αυτό οφείλεται στο ότι για να είναι ο πίνακας **D** ευκλείδειος θα πρέπει να ισχύει ότι ο **B** είναι θετικά ημιορισμένος (positive semidefinite). Ωστόσο στην περίπτωση που υπάρχουν αρνητικές ιδιοτιμές, αν υπάρχουν έστω και δύο μεγάλες θετικές ιδιοτιμές μπορούμε να προσεγγίσουμε τον αρχικό πίνακα χρησιμοποιώντας αυτές.
- Μπορούμε να μετρήσουμε την καλή προσαρμογή της μεθόδου μέσω δεικτών όπως είναι το τετράγωνο του μέσου όρου του σφάλματος (mean square error) ανάμεσα στον αρχικό πίνακα $\boldsymbol{D}_X$ και $\boldsymbol{D}_Y$ κτλ.

Υπάρχουν πολλές προεκτάσεις και μέθοδοι που ανήκουν στην οικογένεια των μεθόδων πολυδιάστατης κλιμακοποίησης. Όλες ενώ βασίζονται στην ίδια λογική παρουσιάζουν διαφορές ως προς τη λύση του τελικού προβλήματος βελτιστοποίησης χρησιμοποιώντας διαφορετικά εργαλεία και μέτρα προσαρμογής. Στην παρούσα μεταπτυχιακή εργασία θα μείνουμε στην κλασική μέθοδο πολυδιάστατης κλιμακοποίησης.

2.3.3 Ισομετρική απεικόνιση χαρακτηριστικών (ISOMAP)

Η μέθοδος ISOMAP είναι μία μη γραμμική μέθοδος εκμάθησης πολλαπλοτήτων που προτάθηκε από τους (Tenenbaum et al. 2000).

Η μέθοδος αυτή βασίζεται στην κλασική γραμμική μέθοδο Multi-Dimensional Scaling (MDS) αλλά στοχεύει στη διατήρηση της εγγενής γεωμετρικής δομής των δεδομένων διατηρώντας την γεωδαισιακή απόσταση μεταξύ τους. Στην ουσία υπολογίζει γεωδαισιακές αποστάσεις μεταξύ μακρινών σημείων έχοντας μόνο ως δεδομένη την απόσταση πριν τον μετασχηματισμό, στον αρχικό χώρο (input space). Για τα γειτονικά σημεία που βρίσκονται πολύ κοντά μεταξύ τους η ευκλείδεια απόσταση που γνωρίζουμε είναι μία καλή εκτίμηση της γεωδαισιακής απόστασης, όμως για τα μακρινά σημεία, η γεωδαισιακή απόσταση μπορεί να υπολογιστεί προσθέτοντας στην ακολουθία «μικρά άλματα» από γειτονικά σημεία. Αυτές οι

εκτιμήσεις υπολογίζονται με ελαφρύ υπολογιστικό κόστος προσδιορίζοντας το πιο σύντομο μονοπάτι μεταξύ δύο οποιονδήποτε σημείων σε έναν γράφο με κόμβους τα σημεία και ακμές τις αποστάσεις μεταξύ τους.

Ο αλγόριθμος του ISOMAP έχει τρία βασικά βήματα.

Δοθέντος ενός συνόλου δεδομένων $\boldsymbol{X} \in R^{n\times m}$ με m μεταβλητές και n παρατηρήσεις, στο πρώτο βήμα απλώνουμε τα δεδομένα στον αρχικό χώρο (input space) και επιλέγουμε τη συνθήκη ή τη μέθοδο που θα αποφασίζει αν και πότε δύο σημεία θα θεωρούνται γειτονικά (δηλαδή επιτρέπεται η μετακίνηση από το ένα στο άλλο), βασιζόμενοι σε αποστάσεις $d_X(i,j)$ μεταξύ κάθε ζευγαριού σημείων. Έτσι, κατασκευάζουμε έναν πίνακα διασυνδεσιμότητας (connectivity) $\boldsymbol{D}_X \in R^{m\times m}$. Δύο απλές μέθοδοι για την εκπλήρωση του βήματος αυτού είναι είτε η ένωση κάθε σημείου με οποιοδήποτε άλλο σημείο που βρίσκεται σε σταθερή ακτίνα ε ή ενώνοντας κάθε σημείο με τους k κοντινότερους γείτονες (βασιζόμενοι στον κανόνα k-NN (k - Nearest Neighbours rule)).

Οι σχέσεις γειτνίασης μεταξύ των σημείων εκφράζονται μέσω ενός γράφου $G = (V, E)$. Οι κόμβοι του γράφου G προσδιορίζουν τα σημεία και οι ακμές προσδιορίζουν τις ευκλείδιες αποστάσεις $d_X(i,j)$ μεταξύ των σημείων (σημειώνεται ότι μπορεί να χρησιμοποιηθεί οποιαδήποτε μετρική για την εύρεση των γειτονικών σημείων). Επομένως μία ακμή μεταξύ δύο κόμβων δημιουργείται αν $d(x_i, x_j) = \|x_i - x_j\|_{l_2} < \varepsilon, \forall i \neq j$.Επομένως το βάρος/δύναμη της συνδεσης w_{ij} που αντιστοιχεί μεταξύ των κόμβων i και j θα είναι $w_{ij} = \frac{1}{d(x_i,x_j)}$.Αν δεν υπάρχει σύνδεση μεταξύ i και j κόμβου έχω $w_{ij} = 0$.

Κατά το δεύτερο βήμα του αλγορίθμου εκτιμώνται οι γεωδαισιακές αποστάσεις $d_M(i,j)$ μεταξύ των σημείων του προτύπου M και υπολογίζονται οι κοντινότερες αποστάσεις σημείου από σημείο $d_G(i,j)$ στον γράφο G. Ένας απλός αλγόριθμος που βρίσκει τις ελάχιστες αποστάσεις μεταξύ των κόμβων ενός γράφου είναι για παράδειγμα ο αλγόριθμος του Dijkstra (αν έχουμε αρνητικά βάρη μπορούμε να χρησιμοποιήσουμε τον αλγόριθμο Floyd-Warshall $O(V^3)$) με υπολογιστική πολυπλοκότητα $O(V^2)$ όπου V αριθμός των κόμβων του γράφου G. Περισσότεροι αλγόριθμοι μπορούν να βρεθούν από (Grama et al. 2003).Η διαδικασία παράγει έναν πίνακα $\boldsymbol{D}_G \equiv d_G(x_i, x_j) = \min\{d_{ij}, d_{i,k} + d_{k,j}\}, k = 1,2, \dots, m, k \neq i, j$

Στο τρίτο και τελευταίο βήμα, ο αλγόριθμος της μεθόδου ISOMAP εφαρμόζει την κλασική πολυδιάστατη κλιμακοποίηση (MDS) για τον πίνακα με τις ελάχιστες αποστάσεις σημείου από σημείο του γραφήματος G. Κατασκευάζεται δηλαδή μία ενσωμάτωση του $\boldsymbol{D}_G = \{ d_G(i,j) \}$ για τα δεδομένα σε έναν p-διάστατο ($p < m$) Ευκλείδειο χώρο Y που διατηρεί βέλτιστα (κατά το δυνατό) την εγγενή δομή της μορφής-προτύπου των δεδομένων.

Οι συντεταγμένες $\boldsymbol{y}_i$ για τα σημεία στον χώρο Y επιλέγονται με βάση την ελαχιστοποίηση της συνάρτησης κόστους :

$$\boldsymbol{E} = argmin_{rank(\boldsymbol{D}_Y \leq p)} ||\tau(\boldsymbol{D}_G) - \tau(\boldsymbol{D}_Y)||_{L^2} \quad (2.41)$$

Όπου, $\boldsymbol{D}_Y$ είναι ο πίνακας των Ευκλείδειων αποστάσεων

$d_Y(i,j) = \{ ||y_i - y_j|| \},$

και $||\boldsymbol{A}||_{L^2}$ η L^2 νόρμα πίνακα $\sqrt{\sum_{i,j} A_{i,j}^2}$.

Ο τελεστής τ μετατρέπει τις αποστάσεις σε εσωτερικά γινόμενα όπως είδαμε στην μέθοδο MDS. Πιο συγκεκριμένα,

$\tau(\boldsymbol{D}) = -\frac{\boldsymbol{HSH^T}}{2}$ όπου $\boldsymbol{S}$ είναι ο πίνακας τετραγωνισμένων αποστάσεων $\{ S_{i,j} = D_{i,j}^2 \}$ και ο $\boldsymbol{H}$ είναι ο πίνακας κεντροποίησης (centering matrix) (Mardia et al., 1979).

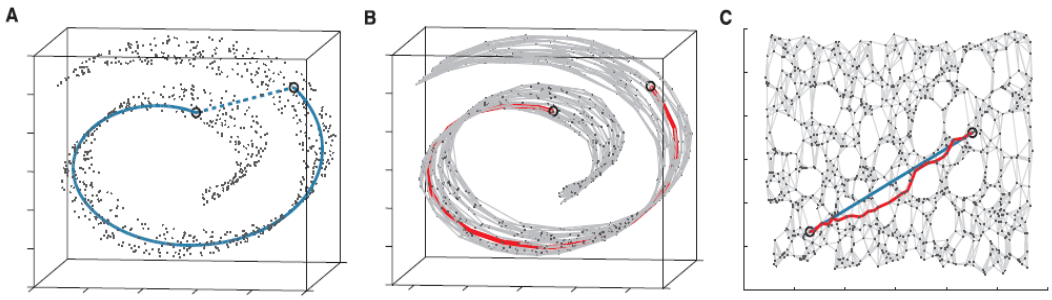
Με τον τρόπο αυτό η MDS προσαρμόζει την γεωμετρία των δεδομένων σε χαμηλότερη διάσταση με μορφή τέτοια ώστε να διατηρεί αποδοτικά την εγγενή γεωμετρία τους. Το ολικό ελάχιστο της συνάρτησης κόστους επιτυγχάνεται θέτοντας συντεταγμένες $\boldsymbol{y}_i$ για τα σημεία με βάση τα p επιλεγμένα ιδιοδιανύσματα του πίνακα $\tau(D_G)$ που αντιστοιχούν στις υψηλότερες ιδιοτιμές (Mardia et al., 1979). Επομένως οι νέες συντεταγμένες $[\boldsymbol{y}_1 \quad \cdots \quad \boldsymbol{y}_m]$ δίνονται από τη σχέση:

$$[\boldsymbol{y}_1 \quad \cdots \quad \boldsymbol{y}_m] = \boldsymbol{\Sigma}_{pxp}\mathbf{V}_{pxm}^T \tag{2.42}$$

Όπου, $\boldsymbol{\Sigma}$ είναι οι p μεγαλύτερες ιδιοτιμές και $\mathbf{V}$ τα ιδιοδιανύσματα του πίνακα $-\frac{\boldsymbol{HSH^T}}{2}$.

Όπως στις μεθόδους PCA/MDS, έτσι και στην μέθοδο ISOMAP η πραγματική διάσταση των δεδομένων μετά την εφαρμογή της μεθόδου μπορεί να εκτιμηθεί με μία μείωση του λάθους καθώς η διάσταση αυξάνεται. Ένα παράδειγμα-εφαρμογή της μεθόδου είναι η ελβετική «κουλούρα» γνωστή και ως «Swiss roll» (Εικόνα 2.3), όπου όπως θα περιμέναμε οι μέθοδοι MDS και PCA αποτυγχάνουν να αναγνωρίσουν αφού αυτή η γεωμετρική δομή είναι μη γραμμική.

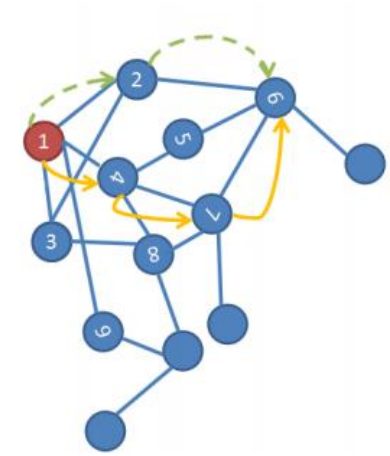
Όπως η Swiss roll, αυτές οι δομές δεδομένων, των οποίων η γεωμετρία είναι αυτή των κυρτών περιοχών του Ευκλείδειου χώρου αλλά και των οποίων η περιβάλλουσα γεωμετρία ανήκει σε υψηλής διάστασης χώρο, μπορεί να είναι ιδιαίτερα διπλωμένη, στριμμένη ή κυρτή. Για τις μη-Ευκλείδειες δομές, όπως ένα ημισφαίριο ή η επιφάνεια ενός ντόνατ, η ISOMAP παράγει μια βέλτιστη και χαμηλή σε διαστάσεις Ευκλείδεια αναπαράσταση.



Εικόνα 2.3 *Οπτική αναπαράσταση της εφαρμογής του ISOMAP για το Swiss roll. Βλέπουμε αρχικά (A) ποια είναι η γεωδαισιακή απόσταση (με συνεχόμενη γραμμή-καμπύλη) και ποια είναι η ευκλείδεια απόσταση (με συγκεκομμένη γραμμή). Δίπλα ακριβώς (B) βλέπουμε μια αναπαράσταση του γραφήματος αποστάσεων με ακμές μεταξύ σημείων που γειτνιάζουν μεταξύ τους. Με κόκκινη καμπύλη φαίνεται η γεωδαισιακή απόσταση δύο σημείων. Στην τελευταία αναπαράσταση (C) βλέπουμε τη διαφορά που έχουν γεωδαισιακή(κόκκινο) και ευκλείδεια (μπλε) απόσταση μεταξύ τους σε μία απόδοση δύο διαστάσεων του Swiss roll.[Tenenbaum et al., 2000]*

2.3.4 Απεικονίσεις Διάχυσης (Diffusion Maps)

Οι απεικονίσεις διάχυσης(Coifman et al. 2005) είναι όπως και η ISOMAP μία μέθοδος που ανήκει στην οικογένεια των μη γραμμικών μεθόδων εκμάθησης πολλαπλοτήτων. Στόχος της μεθόδου είναι η μείωση διάστασης αναδιατάσωντας τα δεδομένα σύμφωνα με παραμέτρους της εγγενούς γεωμετρίας τους. Η σύνδεση ενός συνόλου δεδομένων, χρησιμοποιώντας ένα μέτρο ομοιότητας (similarity measure) οδηγεί στη δημιουργία μίας χρονικά εξαρτώμενης διαδικασία διάχυσης. Καθώς η διάχυση κλιμακώνεται, αιχμαλωτίζει την τοπική γεωμετρία των σημείων ανακαλύπτοντας γεωμετρικές δομές σε διαφορετικές κλίμακες. Ορίζοντας λοιπόν μία χρονικά εξαρτώμενη παράμετρο διάχυσης μπορούμε να μετρήσουμε την ομοιότητα ανάμεσα σε δύο σημεία για μία συγκεκριμένη κλίμακα. Ένας χάρτης διάχυσης ενσωματώνει τα δεδομένα σε έναν χώρο μικρότερης διάστασης απο τον αρχικό έτσι ώστε η ευκλείδεια απόσταση μεταξύ των σημείων να προσεγγίζει την απόσταση διάχυσης στον αρχικό χώρο που βρίσκονται. Η διάσταση του χώρου διάχυσης αποφασίζεται απο την εγγενή γεωμετρική δομή των δεδομένων καθώς και το πόσο καλά η μέθοδος προσεγγιζει την απόσταση διάχυσης.



Εικόνα 2.4 *Ένας τυχαίος περίπατος σε ένα σετ δεδομένων. Κάθε άλμα έχει μία πιθανότητα που σχετίζεται με αυτό. Το διακεκομμένο μονοπάτι ανάμεσα στους κόμβους 1 και 6 χρειάζεται δύο άλματα(για παράδειγμα δύο χρονικές μονάδες) με την πιθανότητα να είναι p(κόμβου 1→ 2) * p(κόμβου 2 →κόμβου 6).[Porte and Herbst 2008]*

Απόσταση διάχυσης

Ας υποθέσουμε ότι κάνουμε έναν τυχαίο περίπατο στα δεδομένα μας, κάνοντας άλματα ανάμεσα σε σημεία στον αρχικό χώρο (Εικόνα 2.4). Η μεταφορά από ένα σημείο σε ένα άλλο κοντινό σημείο, είναι πιο πιθανό να συμβεί από μια μεταφορά σε ένα σημείο μακρύτερα. Η παρατήρηση αυτή μας δίνει μία σχέση ανάμεσα στην απόσταση και την πιθανότητα μετάβασης από ένα σημείο σε ένα άλλο σε ένα βήμα ενός τυχαίου περιπάτου. Η σύνδεση c μεταξύ των σημείων x και y ορίζεται αρχικά:

$$c(x,y) = p(x,y)\,, \tag{2.43}$$

όπου για την ποσοτικοποίηση αυτής της πιθανότητας χρησιμοποιούμε τον γνωστό Γκαουσιανό πυρήνα

$$k(x,y) = \exp\left(-\frac{|x-y|^2}{\sigma}\right), \tag{2.44}$$

η συμπεριφορά του πυρήνα είναι τέτοια που αν απομακρυνθούμε από μία γειτονιά ενός σημείου γρήγορα αυτή κάνει την τιμή του k να προσεγγίζει το 0.

Η γειτονιά του x μπορεί να οριστεί σαν όλα εκείνα τα στοιχεία y όπου ισχύει $k(x,y) \geq \varepsilon$ με $0 > \varepsilon \gg 1$. Έτσι ορίζεται η περιοχή εκείνη μέσα στην οποία ορίζουμε την σύνδεση μεταξύ δύο σημείων την οποία θα προσπαθήσουμε να προσεγγίσουμε στον χαμηλότερης διάστασης χώρο. Είναι σημαντικό να σημειωθεί ότι η απόσταση μεταξύ x και y δεν χρειάζεται να είναι απαραίτητα η ευκλείδεια απόσταση αλλά ένα οποιοδήποτε μέτρο ομοιότητας. Με την κατάλληλη επιλογή του σ όπου είναι παράμετρος διάχυσης επιλέγουμε το μέγεθος της περιοχής, βασιζόμενοι σε κάποιο μαθηματικό σχήμα η προηγούμενη γνώση για την δομή των δεδομένων και την πυκνότητα της κατανομής στον αρχικό χώρο. Ο πυρήνας διάχυσης ικανοποιεί τις παρακάτω ιδιότητες:

- $k(x,y) = k(y,x)$ (Συμμετρία)
- Η σύνδεση κ είναι $k(x,y) \geq 0$ (Θετικά ημιορισμένος)

Οι ιδιότητες του πυρήνα είναι σημαντικές αν σκεφτεί κανείς ότι η πρώτη ιδιότητα είναι απαραίτητη για την φασματική ανάλυση ενός πίνακα αποστάσεων $\boldsymbol{D}_X$ με στοιχεία $D_X(i,j) = d(x_i, x_j)$ όπου περιέχει τις αποστάσεις μεταξύ των σημείων στον αρχικό χώρο. Η δεύτερη ιδιότητα είναι σημαντική για να θεωρήσουμε την τιμή του k ως μία κανονικοποιημένη πιθανότητα (η τιμή της οποίας οφείλει να είναι θετική) έτσι ώστε:

$$\frac{1}{d_X}\sum_{y\in X} K(x,y) = 1 \tag{2.45}$$

Έτσι η σχέση μεταξύ του πυρήνα και της σύνδεσης μεταξύ των σημείων γίνεται:

$$c(x,y) = p(x,y) = \frac{1}{d_X}K(x,y) \tag{2.46}$$

με το $\frac{1}{d_X}$ να αποτελεί μία σταθερά κανονικοποίησης.

Ορίζουμε λοιπόν έναν κανονικοποιημένο (κατά γραμμή) πίνακα Διάχυσης $\boldsymbol{P}$, με στοιχεία

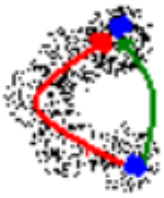
$$\mathrm{P}_{ij} \;=\; \mathrm{p}(x_{\mathrm{i}}\,,\mathrm{x}_j)\,.$$

Κάθε στοιχείο εκφράζει την σύνδεση μεταξύ δύο σημείων $\mathrm{x}_{\mathrm{i}}\,,\mathrm{x}_j$ και εξηγεί την απόσταση τους μέσω της πιθανότητας μετάβασης. Σε αναλογία με έναν τυχαίο περίπατο (random walk), ο πίνακας δίνει τις πιθανότητες για κάθε βήμα από ένα σημείο σε ένα άλλο. Υψώνοντας δυνάμεις του πίνακα διάχυσης απλώς αυξάνουμε τον αριθμό των βημάτων στη διαδικασία. Για παράδειγμα σε έναν απλό πίνακα διάχυσης 2 × 2 Έχουμε:

$$\boldsymbol{P} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \tag{2.47}$$

όπου το p_{ij} συμβολίζει την πιθανότητα να μεταπηδήσουμε από το σημείο i στο j.Αν υψώσουμε τον πίνακα P στην αμέσως επόμενη δύναμη έχουμε:

$$\boldsymbol{P}^2 = \begin{bmatrix} p_{11}p_{11} + p_{12}p_{21} & p_{12}p_{22} + p_{11}p_{12} \\ p_{21}p_{12} + p_{22}p_{21} & p_{22}p_{22} + p_{21}p_{12} \end{bmatrix} \tag{2.48}$$



Εικόνα 2.5 *Μονοπάτι που ακολουθεί την πραγματική γεωμετρία των δεδομένων (με κόκκινο χρώμα) και έχει μεγαλύτερη πιθανότητα να ακολουθηθεί σε έναν τυχαίο περίπατο.[Porte and Herbst 2008]*

Το άθροισμα των δύο πιθανοτήτων $p_{11}p_{11} + p_{12}p_{21}$ αντιστοιχεί στην πιθανότητα να μείνουμε στο σημείο 1 και στην πιθανότητα να μετακινηθούμε απο το 1 στο 2 και να επιστρέψουμε πίσω. Για έναν τυχαίο περίπατο με δύο μεταβάσεις, αυτά είναι όλα τα δυνατά μονοπάτια από το i στο j. Όμοια, το P_{ij}^t αθροίζει όλα τα μονοπάτια με μήκος t απο το i στο j.

Καθώς υπολογίζουμε τις πιθανότητες $\boldsymbol{P}^t$ για αυξανόμενες τιμές του t παρατηρούμε το σύνολο δεδομένων σε διαφορετικές κλίμακες (χρονικά). Αυτή είναι μία διαδικασία διάχυσης, μέσω της οποίας παρατηρούμε την τοπική διασυνδεσιμότητα σημείων ώστε να εξάγουμε συμπεράσματα για την ολική διασυνδεσιμότητα των σημείων. Για αυξανόμενες τιμές του t (όσο δηλαδή η διαδικασία διάχυσης ‘’τρέχει’’ στο χρόνο), η πιθανότητα να ακολουθήσουμε ένα μονοπάτι αυξάνει αποκαλύπτοντας έτσι μία γεωμετρική δομή. Αυτό συμβαίνει διότι πάνω στη γεωμετρική δομή, τα σημεία είναι πυκνά και επομένως όπως είδαμε παραπάνω υπάρχει μεγάλη σύνδεση μεταξύ τους. Τα μονοπάτια στη γεωμετρική δομή είναι πιθανότερο να ακολουθηθούν από άλλα εκτός της δομής διότι περιέχουν άλματα μικρής πιθανότητας κάνοντας έτσι το συνολικό μονοπάτι πολύ λιγότερο πιθανό να ακολουθηθεί (Εικόνα 2.5).

Η απόσταση μεταξύ σημείων κατά τη διαδικασία που περιγράφηκε παραπάνω ονομάζεται απόσταση διάχυσης. Η μετρική αυτή μετράει την ομοιότητα που παρουσιάζουν δύο σημεία στον χώρο ως προς την πιθανότητα μετάβασης από το ένα στο άλλο σε μία χρονική κλίμακα t. Η απόσταση διάχυσης συνδέεται με τον στοχαστικό πίνακα $\boldsymbol{P}$ και δίνεται από τη σχέση:

$$D_t(X_i, X_j)^2 = \sum\nolimits_{u \in X} \left|p_t(X_i, u) - p_t(X_j, u)\right|^2 = \sum\nolimits_k \left|P_{ik}^t - P_{kj}^t\right|^2 \tag{2.49}$$

Από τη σχέση 2.49 παρατηρούμε ότι η απόσταση διάχυσης είναι μικρή αν υπάρχουν πολλά μονοπάτια υψηλής πιθανότητας μεταξύ δύο σημείων. Σε αντίθεση με την ISOMAP που εκτιμά τη γεωδαισιακή απόσταση μεταξύ των σημείων, η μετρική διάχυσης είναι περισσότερο αξιόπιστη(robust) στο θόρυβο που μπορεί να παρουσιάζουν οι μετρήσεις αφού προσθέτει όλα τα πιθανά μονοπάτια μήκους t μεταξύ δύο σημείων.

Η πιθανότητα $p_t(X_i, u)$ είναι η πιθανότητα μετάβασης από το x στο u (ένα οποιοδήποτε u στο σύνολο δεδομένων) σε t χρονικές μονάδες, προσθέτοντας τις πιθανότητες από όλα τα πιθανά μονοπάτια μηκους t.

Απεικόνιση Διάχυσης

Χαμηλής διάστασης δεδομένα είναι συχνά ενσωματωμένα σε ένα χώρο μεγαλύτερης διάστασης. Τα δεδομένα μπορεί να σχηματίζουν γεωμετρικές δομές που συχνά μπορεί να είναι μη γραμμικές. Παραπάνω, βρήκαμε μία μετρική η οποία μπορεί να εκτιμήσει τις αποστάσεις πάνω στην πραγματική γεωμετρική δομή στην οποία βρίσκονται τα δεδομένα. Παρακάτω, θα ψάξουμε μία απεικόνιση των δεδομένων σε έναν ευκλείδειο χώρο η οποία θα αποδίδει όσο το δυνατόν καλύτερα τις αποστάσεις διάχυσης μεταξύ των σημείων. Η απόσταση διάχυσης γίνεται τελικά ευκλείδεια απόσταση σε έναν νέο χώρο διάχυσης.

Μία απεικόνιση διάχυσης, απεικονίζει συντεταγμένες ανάμεσα στα δεδομένα και τον χώρο διάχυσης και έχει στόχο την αναδιάταξη των σημείων σύμφωνα με την μετρική διάχυσης. Θα χρησιμοποιήσουμε λοιπόν αυτή τη διαδικασία για να πετύχουμε μία μείωση της διάστασης των δεδομένων και να ανακαλύψουμε γεωμετρικές δομές χρήσιμες για την καλύτερη κατανόηση των δεδομένων σε χώρο χαμηλότερης διάστασης.

Η ερώτηση όμως που ανακύπτει, είναι πόσες διαστάσεις θα χρειαστούμε, η μάλλον πόσες διαστάσεις θα θεωρήσουμε αμελητέες έτσι ώστε να αποδώσουμε όσο το δυνατόν καλύτερα τις αποστάσεις διάχυσης (άρα και τη γεωμετρική δομή) με αντίστοιχες ευκλείδειες σε έναν χώρο διάχυσης χαμηλότερης διάστασης.

Επομένως θα εξετάσουμε την απεικόνιση:

$$y_i := \begin{bmatrix} p_t(x_i, x_1) \\ p_t(x_i, x_2) \\ \dots \\ p_t(x_i, x_N) \end{bmatrix} = P^t_{i*} \tag{2.50}$$

Για την απεικόνιση της 2.50 η ευκλείδεια απόσταση μεταξύ δύο σημείων y_i και y_j είναι η ακόλουθη:

$$\left|\left|y_i - y_j\right|\right|_E^2 = \sum_{u \in X} \left|p_t(x_i, u) - p_t(x_j, u)\right|^2 = \sum_k \left|P^t_{ik} - P^t_{kj}\right|^2 = D_t\left(x_i, x_j\right)^2 \tag{2.51}$$

Η οποία είναι η απόσταση διάχυσης ανάμεσα σε δύο σημεία x_i , x_j. Η απεικόνιση της 2.50 μας παρέχει την αναδιάταξη των δεδομένων που ψάχνουμε. Ας σημειώσουμε όμως εδώ ότι δεν έχουμε πετύχει ακόμα κάποια μείωση διάστασης. Η μείωση διάστασης επιτυγχάνεται

θεωρώντας αμελητέες συγκεκριμένες διαστάσεις του χώρου διάχυσης. Ας πάρουμε λοιπόν τον κανονικοποιημένο πίνακα διάχυσης,

$$\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{K} \tag{2.52}$$

Όπου $\boldsymbol{D}$ είναι ένας διαγώνιος πίνακας που αποτελείται από τα αθροίσματα των γραμμών του $\boldsymbol{K}$. Αποδεικνύεται ότι οι αποστάσεις διάχυσης μπορούν να εκφραστούν με όρους ιδιοδιανυσμάτων και ιδιοτιμών του στοχαστικού πίνακα $\boldsymbol{P}$:

$$Y'_i := \begin{bmatrix} \lambda_1\psi_1(i) \\ \lambda_2\psi_2(i) \\ \dots \\ \lambda_N\psi_N(i) \end{bmatrix} \tag{2.53}$$

Όπου $\psi_1(i)$ είναι το i-οστό στοιχείο του πρώτου ιδιοδιανύσματος του $\boldsymbol{P}$ ενώ λ_1 είναι η μεγαλύτερη ιδιοτιμή του πίνακα $\boldsymbol{P}$ κ.ο.κ.

Τελικά, η ευκλείδεια απόσταση ανάμεσα στα απεικονισθέντα σημεία Y'_i και Y'_j είναι η απόσταση διάχυσης. Το σύνολο των ορθογώνιων αριστερών ιδιοδιανυσμάτων του $\boldsymbol{P}$ φτιάχνουν μία βάση στον χώρο διάχυσης, που σχετίζεται με τις ιδιοτιμές λ_i που μας δείχνουν το πόσο σημαντική είναι η κάθε διάσταση.

Η μείωση διάστασης επιτυγχάνεται με την διατήρηση m διαστάσεων που συνδέονται με τις μεγαλύτερες ιδιοτιμές και τα αντίστοιχα ιδιοδιανύσματα.

Μαθηματική Θεμελίωση των απεικονίσεων διάχυσης

Παρακάτω θα επιχειρηθεί μία μαθηματική απόδειξη της μεθόδου των απεικονίσεων διάχυσης. Στην πραγματικότητα θα αποδείξουμε ότι αν εισάγουμε ένα νέο σύστημα συντεταγμένων σε ένα χώρο διάχυσης έτσι ώστε οι συντεταγμένες να είναι ίσες με κλιμακοποιημένα στοιχεία (scaled components) των ιδιοδιανυσμάτων του πίνακα διάχυσης, τότε η απόσταση διάχυσης μεταξύ δύο σημείων στον αρχικό χώρο θα είναι ίση με την ευκλείδεια απόσταση τους στον χώρο διάχυσης. Για την παρακάτω απόδειξη θα θεωρήσουμε την περίπτωση όπου ο αρχικός χώρος είναι ο R^n.Η απόδειξη θα χωριστεί σε 2 Λήμματα.

Λήμμα 1. Υποθέτουμε ότι ο $\boldsymbol{K}$ είναι ένας $n \times n$ συμμετρικός πίνακας πυρήνα (Kernel matrix) τέτοιος ώστε $K[i,j] = k(i,j)$ και D ένας διαγώνιος πίνακας που κανονικοποιεί τις γραμμές του $\boldsymbol{K}$ έτσι ώστε ο πίνακας διάχυσης να είναι:

$$\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{K} \tag{2.54}$$

Τότε, ο πίνακας $\boldsymbol{P'}$ που ορίζεται ως:

$$\boldsymbol{P'} = \boldsymbol{D}^{\frac{1}{2}}\boldsymbol{P}\boldsymbol{D}^{-\frac{1}{2}} \tag{2.55}$$

Είναι:

1. Συμμετρικός
2. Έχει τις ίδιες ιδιοτιμές με τον πίνακα P
3. Τα ιδιοδιανύσματα x'_k του πίνακα $\boldsymbol{P'}$ πολλαπλασιάζονται με $\boldsymbol{D^{-\frac{1}{2}}}$ και $\boldsymbol{D^{\frac{1}{2}}}$ αντίστοιχα για να δώσουν τα αριστερά ($e^T P = \lambda e^T$) και δεξιά($Pv = \lambda v$) ιδιοδιανύσματα του πίνακα $\boldsymbol{P}$ αντίστοιχα**.**

Απόδειξη. Αντικαθιστούμε την σχέση 2.54 στην 2.55 και προκύπτει:

$$\boldsymbol{P' = D^{-\frac{1}{2}} K D^{-\frac{1}{2}}} \tag{2.56}$$

Αφού ο πίνακας $\boldsymbol{K}$ είναι συμμετρικός τότε και ο $\boldsymbol{P'}$ θα είναι συμμετρικός. Λύνω την σχέση 2.55 ως προς $\boldsymbol{P}$ και έχω:

$$\boldsymbol{P = D^{-\frac{1}{2}} P' D^{\frac{1}{2}}} \tag{2.57}$$

Αφού ο πίνακας $\boldsymbol{P'}$ είναι συμμετρικός, τότε θα υπάρχει ένα ορθοκανονικό σύνολο ιδιοδιανυσμάτων του $\boldsymbol{P'}$ τέτοιο ώστε:

$$\boldsymbol{P' = S \Lambda S^T} \tag{2.58}$$

Όπου, $\boldsymbol{\Lambda}$ είναι ένας διαγώνιος πίνακας που περιέχει τις ιδιοτιμές του $\boldsymbol{P'}$ και $\boldsymbol{S}$ είναι ένας πίνακας με τα ορθοκανονικά ιδιοδιανύσματα του $\boldsymbol{P'}$ ως στήλες. Αντικαθιστώντας την σχέση 2.58 στην 2.57 έχουμε:

$$\boldsymbol{P = D^{-\frac{1}{2}} S \Lambda S^T D^{\frac{1}{2}}} \tag{2.59}$$

Αφού ο $\boldsymbol{S}$ είναι ένας ορθογώνιος πίνακας,

$$\boldsymbol{P = D^{-\frac{1}{2}} S \Lambda S^T D^{\frac{1}{2}} = \left(D^{-\frac{1}{2}} S\right) \Lambda \left(D^{-\frac{1}{2}} S\right)^{-1}} \tag{2.60}$$

$$\boldsymbol{= Q \Lambda Q^{-1}} \tag{2.61}$$

Επομένως οι ιδιοτιμές των $\boldsymbol{P'}$ και $\boldsymbol{P}$ είναι οι ίδιες.Επιπλέον, τα δεξιά ιδιοδιανύσματα του $\boldsymbol{P}$ είναι οι στήλες του πίνακα,

$$\boldsymbol{Q = D^{-\frac{1}{2}} S} \tag{2.62}$$

ενώ τα αριστερά ιδιοδιανύσματα είναι οι γραμμές του αντιστρόφου του πίνακα $\boldsymbol{Q}$,

$$\boldsymbol{Q^{-1} = S^T D^{\frac{1}{2}}} \tag{2.63}$$

από τη σχέση 2.62 παρατηρούμε ότι η εξίσωση για τα ιδιοδιανύσματα του $\boldsymbol{P}$ μπορεί να εκφραστεί σε όρους των ιδιοδιανυσμάτων x'_k του πίνακα $\boldsymbol{P'}$. Επομένως, τα δεξιά ιδιοδιανύσματα του $\boldsymbol{P}$ θα είναι:

$$v_k = \boldsymbol{D^{-\frac{1}{2}}} x'_k \tag{2.64}$$

και τα αριστερά ιδιοδιανύσματα θα είναι:

$$e_k = \boldsymbol{D}^{\frac{1}{2}} x'_k \tag{2.65}$$

Από τη σχέση 2.60 παίρνουμε την ιδιοαποσύνθεση,

$$P = \sum_{\kappa} \lambda_\kappa v_\kappa e_k^T \tag{2.66}$$

Αν εξετάσουμε προσεκτικά την παραπάνω ιδιοαποσύνθεση, παρατηρούμε κάτι ενδιαφέρον, η σχέση 2.66 εκφράζει κάθε γραμμή του πίνακα διάχυσης σε όρους μίας νέας βάσης, της e_k, των αριστερών ιδιοδιανυσμάτων του πίνακα διάχυσης.

Σε αυτό το νέο σύστημα αξόνων στον R^n, μία γραμμή i του $\boldsymbol{P}$ παριστάνεται από το σημείο:

$$M_i := \begin{bmatrix} \lambda_1 v_1(i) \\ \lambda_2 v_2(i) \\ \dots \\ \lambda_n v_n(i) \end{bmatrix} \tag{2.67}$$

Όπου $v_n(i)$ είναι το i-οστό στοιχείο του v-οστού δεξιού ιδιοδιανύσματος.Παρόλα αυτά, ο πίνακας $\boldsymbol{P}$ δεν είναι συμμετρικός, και έτσι το σύστημα συντεταγμένων δεν θα είναι ορθοκανονικό. Για παράδειγμα $e_k^T e_k \neq 1$. Αυτό είναι αποτέλεσμα της κλιμακοποίησης που εφαρμόστηκε στον $\boldsymbol{P}'$στην σχέση 2.64 και 2.65 . Αυτή η κλιμακοποίηση μπορεί να αντιμετωπιστεί με χρησιμοποιώντας μία διαφορετική μετρική $\boldsymbol{Q}$ τέτοια ώστε $e_k^T \boldsymbol{Q} e_k = 1$, όπου $\boldsymbol{Q}$ θα πρέπει να είναι θετικά ορισμένος , συμμετρικός πίνακας. Επιλέγουμε:

$$\boldsymbol{Q} = \boldsymbol{D}^{-1} \tag{2.68}$$

Όπου $\boldsymbol{D}$ είναι ένας διαγώνιος πίνακας κανονικοποίησης. Αυτός ικανοποιεί τις δύο προϋποθέσεις που θέσαμε παραπάνω και χρησιμοποιώντας την 2.65 οδηγεί:

$$e_k^T \boldsymbol{Q} e_k = e_k^T \left(\boldsymbol{D}^{-\frac{1}{2}}\right)\left(\boldsymbol{D}^{-\frac{1}{2}}\right) e_k = {x'_k}^T x'_k = 1 \tag{2.69}$$

Επομένως τα αριστερά ιδιοδιανύσματα του πίνακα διάχυσης σχηματίζουν ένα ορθοκανονικό σύστημα αξόνων στον R^n ο οποίος είναι εφοδιασμένος με την μετρική D^{-1}. Παρακάτω ορίζουμε τον R^n με τη μετρική D^{-1} ως χώρο διάχυσης όπου και θα συμβολίσουμε με $l\text{^}2\ (R^n, \boldsymbol{D}^{-1})$.Για παράδειγμα, η ευκλείδεια απόσταση μεταξύ δύο διανυσμάτων, α και a' ορίζεται κανονικά,

$$d(a, a')_{l_2} = d(a, a')_{l_2(R^n, I)} = (a - a')^T (a - a')$$

Ενώ στον $l^2\ (R^n, D^{-1})$ γίνεται:

$$d(a, a')_{l_2(R^n, D^{-1})} = (a - a')^T \boldsymbol{D}^{-1}\ (a - a')$$

Λήμμα 2 Αν επιλέξουμε τις συντεταγμένες διάχυσης όπως στην 2.53, τότε η απόσταση διάχυσης μεταξύ των σημείων στον αρχικό χώρο (ο οποίος μετράται με την μετρική D^{-1}) είναι ίση με την ευκλείδεια απόσταση στον χώρο διάχυσης.

Απόδειξη. Αρκεί να δείξουμε ότι:

$$D_t(x_i, x_j)^2 = \left|\left|p_t(x_i, .) - p_t(x_j, .)\right|\right|^2_{l_2(R^n, D^{-1})} \tag{2.70}$$

$$= \left|\left|M_i - M_j\right|\right|^2_{l_2(R^n, I)} \tag{2.71}$$

$$= \sum_{\kappa} \lambda_\kappa^{2t} \left(v_\kappa(i) - v_k(j)\right)^2 \tag{2.72}$$

Εδώ, $p_t(x_i, x_j)^2 = P_{ij}$ είναι οι πιθανότητες οι οποίες σχηματίζουν τα στοιχεία του πίνακα διάχυσης.Για λόγους απλότητας θα υποθέσουμε ότι $t = 1$. Τότε:

$$D(x_i, x_j)^2 = \left|\left|p(x_i, .) - p(x_j, .)\right|\right|^2_{l_2(R^n, D^{-1})} = \left|\left|P(i, .) - P(j, .)\right|\right|^2_{l_2(R^n, D^{-1})} \tag{2.73}$$

Σύμφωνα με την ιδιοαποσύνθεση της 2.66, η 2.73 γίνεται:

$$\left|\sum_{\kappa} \lambda_\kappa v_\kappa(i) e_k^T - \sum_{\kappa} \lambda_\kappa v_\kappa(j) e_k^T\right|^2 = \left|\sum_{\kappa} \lambda_\kappa e_k^T \left(v_k(i) - v_k(j)\right)\right|^2$$

$$= \left|\sum_{\kappa} \lambda_\kappa x'^T_k D^{\frac{1}{2}} \left(v_k(i) - v_k(j)\right)\right|^2$$

$$= \left|\sum_{\kappa} \lambda_\kappa x'^T_k \left(v_k(i) - v_k(j)\right) D^{\frac{1}{2}}\right|^2$$

Στον $l^2\ (R^n, D^{-1})$ η απόσταση αυτή είναι:

$$\left(\sum_{\kappa} \lambda_\kappa x'^T_k \left(v_k(i) - v_k(j)\right) D^{\frac{1}{2}}\right) D^{-1} \left(\sum_{m} \lambda_m x'^T_m \left(v_m(i) - v_m(j)\right) D^{\frac{1}{2}}\right)^T$$

$$= \left(\sum_{\kappa} \lambda_\kappa x'^T_k \left(v_k(i) - v_k(j)\right) D^{\frac{1}{2}}\right) D^{-1} \left(D^{\frac{1}{2}} \sum_{m} \lambda_m x'^T_m \left(v_m(i) - v_m(j)\right)\right)$$

$$= \sum_{k} \lambda_\kappa x'^T_k \left(v_k(i) - v_k(j)\right) \sum_{m} \lambda_m x'^T_m \left(v_m(i) - v_m(j)\right)$$

Αφού $\{x'_k\}$ είναι ένα ορθοκανονικό σύνολο, $x'^T_m x'_k = 0$ για $m \neq k$.

Επομένως,

$$D(x_i, x_j)^2 = \sum_{\kappa} \lambda_\kappa^2 \left(v_\kappa(i) - v_k(j)\right)^2 \tag{2.74}$$

Παραπάνω δείξαμε ότι η απόσταση διάχυσης στον αρχικό χώρο, $D_t(x_i, x_j)^2$, είναι η ευκλείδεια απόσταση μεταξύ των απεικονισθέντων σημείων ,M_iκαι M_j στον χώρο διάχυσης.

3 ΣΥΓΚΡΙΤΙΚΗ ΑΝΑΛΥΣΗ ΜΕΘΟΔΩΝ ΕΚΜΑΘΗΣΗΣ ΠΟΛΛΑΠΛΟΤΗΤΩΝ ΣΕ ΔΕΔΟΜΕΝΑ ΛΕΙΤΟΥΡΓΙΚΗΣ ΑΠΕΙΚΟΝΙΣΗΣ ΜΑΓΝΗΤΙΚΟΥ ΣΥΝΤΟΝΙΣΜΟΥ (FMRI)

Στο 3ο και κύριο μέρος της εργασίας αυτής γίνεται σύγκριση γραμμικών και μη γραμμικών μεθόδων εκμάθησης πολλαπλοτήτων για την μείωση διάστασης δεδομένων λειτουργικής απεικόνισης μαγνητικού συντονισμού (FMRI) με σκοπό να ελεγχθεί η δυνατότητα τους να συνεισφέρουν στην ανάλυση λειτουργικής διασυνδεσιμότητας δικτύου και την ταξινόμηση των δεδομένων. Ακόμη, εξετάζονται δυο μετρικές (η μία βασίζεται στην πολλαπλή συσχέτιση (με καθυστερήσεις) και η άλλη στην ευκλείδεια νόρμα) με σκοπό να δούμε ποια μετρική λειτουργεί καλύτερα σε δεδομένα FMRI. Για την ταξινόμηση των δεδομένων γίνεται χρήση δημοφιλών αλγορίθμων ταξινόμησης όπως είναι ο Support Vector Machines (SVM),ο ταξινομητής των k κοντινότερων γειτόνων (Knn Classifier) και τα νευρωνικά δίκτυα (Neural Nets).

Η συγκριτική αυτή ανάλυση συνεισφέρει σημαντικά στο να ερευνήσουμε ποια μέθοδος μείωσης διάστασης συμπεριφέρεται καλύτερα, πως αυτή επηρεάζεται από τη μετρική που χρησιμοποιούμε και τέλος πως συνδράμει στην ταξινόμηση των δεδομένων. Στην εργασία αυτή θα επικεντρωθούμε σε δεδομένα που είναι διαθέσιμα σε ανοικτή βάση δεδομένων στο διαδίκτυο στη διεύθυνση: http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html.

Τα δεδομένα αυτά έχουν αποτελέσει αντικείμενο έρευνας σε αρκετές εργασίες όπως (Anderson και Cohen 2013, Tomasi και Volkow 2014, Qureshi et al. 2017, Zeng et al. 2018) και αφορούν FMRI σε ήρεμη κατάσταση υγειών και ασθενών διαγνωσμένων με σχιζοφρένεια.

Μετά από μία βασική προ-επεξεργασία στα δεδομένα FMRI χρησιμοποιήσαμε την μέθοδο ICA (ICA AROMA (Pruim et al. 2015)) για να καθαρίσουμε το σήμα μας από παράγοντες όπως η κίνηση, τα υψηλοσυχνοτικά σήματα κτλ.

Στη συνέχεια και για την εξαγωγή του σήματος ενδιαφέροντος, χρησιμοποιήσαμε τη μέθοδο RAICAR (Yang et al. 2008) (Ranking and Averaging of Independent Components by Reproducibility) η οποία τρέχει επαναληπτικά την μέθοδο ICA με σκοπό να παράγει ανεξάρτητες συνιστώσες τις οποίες και ταξινομεί ανάλογα με την τάση της κάθε συνιστώσας να επανεμφανίζεται και πάλι ως ανεξάρτητη ανάμεσα σε συνεχόμενες εφαρμογές της μεθόδου.

Με τον τρόπο αυτό χωρίσαμε το σήμα σε ανεξάρτητους χωρικούς χάρτες και τις χρονοσειρές που τους διέπουν. Με τη χρήση των δύο μετρικών που χρησιμοποιούνται στην εργασία κατασκευάσαμε για κάθε υποκείμενο πίνακες απόστασης μεταξύ των χρονοσειρών με σκοπό να συγκρίνουμε ποιες από αυτές τις χρονοσειρές είναι όμοιες με άλλες (όμοιες σε σχέση με τη μετρική που χρησιμοποιείται).

Επειδή όμως αυτοί οι πίνακες έχουν υψηλή διάσταση εφαρμόζουμε γραμμικές και μη γραμμικές μεθόδους μείωσης διάστασης διατηρώντας έναν μικρό αριθμό διαστάσεων χρησιμοποιώντας μεθόδους όπως η πολυδιάστατη κλιμακοποίηση(MDS), η ισομετρική απεικόνιση χαρακτηριστικών (ISOMAP) και οι απεικονίσεις διάχυσης (Diffusion Maps). Κοινό χαρακτηριστικό των παραπάνω μεθόδων είναι η διατήρηση των αποστάσεων (distance-preserving) μεταξύ των σημείων των δεδομένων παρέχοντας μία χαμηλής διάστασης ενσωμάτωση. Στην πρώτη περίπτωση, η MDS ψάχνει μία χαμηλής διάστασης ενσωμάτωση των δεδομένων διατηρώντας βέλτιστα τις ευκλείδιες αποστάσεις μεταξύ των σημείων, η

ISOMAP τις γεωδαισιακές, ενώ η Diffusion Maps τις αποστάσεις διάχυσης.Η διατήρηση της απόστασης μεταξύ των σημείων εννοείται διαφορετικά από κάθε μέθοδο, η οποία αντιλαμβάνεται την απόσταση με διαφορετικό τρόπο, όπως αυτό επεξηγήθηκε εκτενώς στο προηγούμενο κεφάλαιο.

Τελικά, οι χρονοσειρές για το κάθε υποκείμενο απεικονίζονται σε έναν χώρο χαμηλότερης διάστασης όπου με τη χρήση πολλαπλών τιμών κατωφλιού (thresholding) σχηματίζονται γράφοι εγκεφαλικής διασυνδεσιμότητας. Για κάθε έναν από τους γράφους αυτούς υπολογίζονται κάποια γραφοθεωρητικά μέτρα (Μέσο μήκος μονοπατιού, συντελεστής συγκρότησης και μέσο βάρος/δύναμη των συνδέσεων) με σκοπό την αξιολόγηση της διασυνδεσιμότητας του γράφου. Τα μέτρα αυτά δίνονται ως είσοδοι σε αλγορίθμους ταξινόμησης με σκοπό τον διαχωρισμό των υποκειμένων σε σχιζοφρενείς και υγιείς. Ο βαθμός στον οποίο κάθε μέθοδο διαχωρίζει σωστά τα δεδομένα, αποτελεί εδώ και το κριτήριο καλής προσαρμογής αφενός μεν της μεθόδου μείωσης διάστασης αλλά και της μετρικής που χρησιμοποιήθηκε.

3.1 Μια επισκοπηση για τα FMRI δεδομενα και αναφορες για τις μεχρι σημερα προσπαθειες αναλυσης λειτουργικης διασυνδεσιμοτητας (Functional Connectivity)

Η λειτουργική απεικόνιση μαγνητικού συντονισμού είναι μία τεχνολογική μέθοδος που μετράει την εγκεφαλική λειτουργία στο χρόνο όπως αυτή αναλύθηκε και στην εισαγωγή της εργασίας.

Το σήμα του επιπέδου εξάρτησης οξυγόνωσης αίματος (Blood Oxygen Level Dependent) είναι μία έμμεση αντανάκλαση της δραστηριότητας των νευρώνων που ανιχνεύεται κατά την διάρκεια μίας FMRI σάρωσης και η ανάλυση του διεξάγεται υπό την υπόθεση ότι η δραστηριότητα των νευρώνων συμπίπτει με την αυξημένη ροή του αίματος. Η αύξηση στη ροή του αίματος κατά την δραστηριοποίηση των νευρώνων είναι γνωστή ως αιμοδυναμική απόκριση (Kim et al. 1999).

Όταν υπάρξει ενεργοποίηση σε μία περιοχή του εγκεφάλου, οξυγονωμένη αιμοσφαιρίνη ρέει στην περιοχή για να αυξήσει τοπικά την συγκέντρωση οξυγόνου. Η δεοξυαιμοσφαιρίνη (αιμοσφαιρίνη που δεν έχει δεσμευμένο οξυγόνο) έχει ένα γρηγορότερα εξασθενούμενο MR σήμα (μικρότερο T_2^*) σε σχέση με την οξυαιμοσφαιρίνη (Cohen και Bookheimer 1994). Ως αποτέλεσμα, εγκεφαλικά σήματα από καλά οξυγονωμένες περιοχές εμφανίζουν ένα πιο δυνατό MR σήμα σε σχέση με περιοχές που παρουσιάζουν μικρότερη συγκέντρωση σε οξυγόνο. Συμπερασματικά, περιοχές με αυξημένη νευρωνική δραστηριότητα δίνουν ισχυρότερο MR σήμα που αντικατοπτρίζει έμμεσα την δυναμική της νευρωνικής δραστηριότητας.

Η τεσσάρων διαστάσεων FMRI σάρωση μπορεί να χρησιμοποιηθεί για ανακάλυψη ανατομικών περιοχών που είναι υπαίτιες για συγκεκριμένες διαδικασίες όπως η αντίληψη της γλώσσας (language processing) (Bookheimer 2002), η αναγνώριση προσώπου (Gauthier et al. 1999) ή ακόμα και για διάγνωση ελλαττωματικής περιοχής που σχετίζεται με γνωστικές διαταραχές όπως η νόσος Alzheimer, τραυματική εγκεφαλική βλάβη ή η σχιζοφρένεια (Anderson et al. 2010, Ford et al. 2010). Τέτοιες μελέτες τυπικά αναλύουν τη ροή του αίματος σε περιοχές του εγκεφάλου για να ανακαλύψουν περιοχές που ενεργοποιούνται κατά τη διάρκεια μίας διαδικασίας (task) ή να συγκρίνουν τα σήματα BOLD δύο ομάδων, όπως για

παράδειγμα ασθενείς με Alzheimer και υγιείς ανθρώπους, αναδεικνύοντας διαφορές που θα μπορούσαν να αποτελούν αιτίες για τη γνωστική βλάβη.

3.1.1 Μέθοδοι ανάλυσης FMRI

Παραδοσιακά η ανάλυση FMRI βασίζεται σε προσεγγίσεις οι οποίες βασίζονται σε κάποιο μοντέλο (model-based). Στην περίπτωση αυτή, ψάχνουμε συνήθως για καταστάσεις σύνδεσης βασιζόμενοι κυρίως σε συσχετίσεις ανάμεσα σε εικονοστοιχεία, περιοχές εικονοστοιχείων κτλ. Αυτού του τύπου η ανάλυση ωστόσο, δεν μπορεί να ξεπεράσει τη φύση, τις υποθέσεις και τους περιορισμούς στους οποίους βασίζεται.

Οι περισσότερες από αυτές τις μεθόδους βασιζόμενες στο μοντέλο (Model-based), όπως είναι οι στατιστικοί παραμετρικοί χάρτες (Statistical Parametric Maps) (Friston et al. 1994), στηρίζονται στο Γενικό Γραμμικό Μοντέλο (General Linear Model) μαζί με άλλα μοντέλα ενεργοποίησης όπως η συνάρτηση αιμοδυναμικού μοντέλου απόκρισης (Hemodynamic Response Function), οι παράγωγοι του (HRF derivatives) κτλ.

Αυτή η πολύ δημοφιλής, δοκιμασμένη και αξιόπιστη προσέγγιση χρησιμοποιήθηκε με επιτυχία στην αναγνώριση εγκεφαλικών δικτύων για αρκετά χρόνια από την εργασία των (Biswal et al. 1995) όπου προτάθηκε ως μέθοδος η ανάλυση λειτουργικής διασυνδεσιμότητας (Functional Connectivity). Χρησιμοποιώντας τον ‘’συμβατικό’’ τρόπο, χρειάζεται να έχουμε κάποια a priori γνώση της λειτουργικής οριοθέτησης της εγκεφαλικής λειτουργίας, ενώ ψάχνοντας για πιθανές συσχετίσεις μεταξύ ‘’ενεργοποιημένων’’ εικονοστοιχείων μπορεί να παρεμβάλλονται νευροβιολογικές συσχετίσεις ανάμεσα σε λειτουργικά συνδεδεμένες περιοχές του εγκεφάλου οι οποίες αποπροσανατολίζουν την ανάλυση.

Από τη στιγμή που χρειάζεται μία περιοχή εικονοστοιχείων του εγκεφάλου ως σημείο αναφοράς για να βρεθεί κάποια ενδεχόμενη συσχέτιση με άλλα εικονοστοιχεία (Seed to Voxel) (ή ακόμα και άλλες περιοχές εικονοστοιχείων (Seed to Seed)), προβλήματα ανακύπτουν και στο επίπεδο της ανάλυσης που αφορά ομάδες υποκειμένων. Για το λόγο αυτό, η έρευνα προχώρησε σε καινούργιες μεθόδους, όπως αυτές της συγκρότησης (clustering methods) και εν γένει σε προσεγγίσεις που βασίζονται απευθείας στα εμπειρικά δεδομένα (Goutte et al. 1999, Venkataraman et al. 2009) (Data-driven).

Ωστόσο, ένας αλγόριθμος συγκρότησης, θα είναι όσο ακριβής θα είναι και η επιλογή της μετρικής που θα γίνει για την εξήγηση του φυσικού προβλήματος. Όπως θα περίμενε κανείς, διαφορετικά μέτρα ομοιότητας (similarity measures) προτάθηκαν, τόσο γραμμικά, εκ των οποίων τα περισσότερα βασίζονται στην απλή συσχέτιση (Jafri et al. 2008, Liu et al. 2012, Anderson και Cohen 2013) και την ευκλείδεια απόσταση (Mezer et al. 2009), όσο και μη γραμμικά, όπως η μερική συσχέτιση (Liu et al. 2009) και η αμοιβαία πληροφορία (Benjaminsson 2010).

Τα τελευταία χρόνια, διαφορετικές γραμμικές και μη γραμμικές μέθοδοι βασιζόμενες σε εμπειρικά δεδομένα έκαναν την εμφάνιση τους στην ανάλυση FMRI. Για παράδειγμα, η διάσημη γραμμική μέθοδος PCA έγινε μέρος της συμβατικής ανάλυσης των FMRI. Ο ρόλος της μεθόδου PCA σήμερα, αφορά κυρίως το στάδιο της προ-επεξεργασίας των FMRI και είναι ένας αποδοτικός τρόπος να μειώσουμε την διάσταση των δεδομένων μας αλλά και να καθαρίσουμε το σήμα από τις διάφορες πηγές θορύβου (κίνηση ανομοιογένειες στο μαγνητικό πεδίο κτλ.). Ωστόσο η PCA μεταξύ άλλων, εφαρμόστηκε και για την εξαγωγή περιοχών ενδιαφέροντος (Region Of Interest) από FMRI σαρώσεις από τους (Viviani et al. 2005).

Μια ακόμα γραμμική μέθοδος, η MDS (Kruskal 1964), έχει χρησιμοποιηθεί για την μείωση διάστασης και την οπτικοποίηση των FMRI στην προσπάθεια να βρούμε μία χαμηλότερης διάστασης απεικόνιση των δεδομένων (που συχνά είναι μεγάλης διάστασης), διατηρώντας τις ευκλείδειες αποστάσεις μεταξύ των σημείων (Βλέπε κεφάλαιο 2) στον νέο χώρο. Μία εφαρμογή της μεθόδου σε FMRI σε κατάσταση ηρεμίας (resting state) έγινε από τους (Benjaminsson 2010) εξάγοντας με επιτυχία περιοχές ενδιαφέροντος (ROIs) και παρέχοντας παράλληλα μία οπτικοποίηση των δεδομένων.

Μία άλλη πολύ δημοφιλής και επιτυχημένη μέθοδος βασιζόμενη στα εμπειρικά δεδομένα είναι η μέθοδος Ανάλυσης σε Ανεξάρτητες Συνιστώσες (ICA) (Hyvärinen 1999). Σε αντίθεση με την PCA όπου μετασχηματίζει γραμμικά τα δεδομένα στην κατεύθυνση του εγκλωβισμού της περισσότερης δυνατής διασποράς από τις κύριες συνιστώσες, η ICA ψάχνει έναν γραμμικό μετασχηματισμό με την μεγαλύτερη δυνατή (στατιστική) ανεξαρτησία ανάμεσα στα εξαγόμενα χαρακτηριστικά. Πολλές εργασίες κάνουν χρήση της μεθόδου ως έναν τρόπο να εξάγουν το σήμα ενδιαφέροντος (Van De Ven et al. 2004, Jafri et al. 2008) ή να βρουν χαρακτηριστικά που οφείλονται σε θόρυβο και στη συνέχεια να τα αγνοήσουν στη μετέπειτα ανάλυση (Salimi-Khorshidi et al. 2014, Pruim et al. 2015). Ειδικότερα, οι (Smith et al. 2009) δημοσίευσαν μία εργασία στην οποία συμμετείχαν σχεδόν 30.000 άνθρωποι όπου χρησιμοποιώντας την ICA αποκάλυψαν την λειτουργική αρχιτεκτονική του εγκεφάλου και τα γνωστά σήμερα ανεξάρτητα δίκτυα του εγκεφάλου όπως είναι το αισθησιοκινητικό (sensory-motor) δίκτυο, το προεπιλεγμένο (default mode) δίκτυο και το ακουστικό (auditory). Τα ευρήματα αυτά οδήγησαν σε εντυπωσιακά αποτελέσματα και ανακαλύψεις σε πολλά από τα πεδία της Νευροεπιστήμης.

Ωστόσο όταν μέθοδοι όπως η PCA, η ICA και η κλασσική MDS δεν είναι ικανές να αποκαλύψουν χρήσιμες αναπαραστάσεις και χαμηλής διάστασης ενσωματώσεις για τα δεδομένα, μη γραμμικές αντίστοιχες τεχνικές χρησιμοποιούνται για ανεύρεση χαμηλής διάστασης ενσωματωμένων δικτύων, οπτικοποίηση, εξαγωγή χαρακτηριστικών και αναλύσεις συγκρότησης (clustering analysis).

Μη γραμμικές μέθοδοι όπως η Τοπικά Γραμμική Ενσωμάτωση (LLE) (Roweis και Saul 2000) και η ISOMAP (Tenenbaum et al. 2000) έχουν εφαρμοστεί σε δεδομένα FMRI τόσο σε κατάστασης ηρεμίας (resting state) όσο και βασιζόμενα σε ρητό καθήκον (task related) (Van De Ven et al. 2004, Shen et al. 2010, Mannfolk et al. 2010). Οι Απεικονίσεις διάχυσης (diffusion maps) που προτάθηκαν σχετικά πρόσφατα από τους Coifman et al. (2005) έχουν επίσης χρησιμοποιηθεί σε δεδομένα FMRI βασιζόμενα σε ρητό καθήκον (task related) (Shen και Meyer 2011) .Συγκεκριμένα, μία εφαρμογή των Diffusion Maps έγινε από τους (Sipola et al. 2013) για την ανάλυση συγκρότησης (clustering) fMRI χωρικών χαρτών όπου η μέθοδος λειτούργησε το ίδιο καλά με τις παραδοσιακές μεθόδους και σε κάποιες περιπτώσεις καλύτερα από αυτές.

Ειδικότερα, στην εργασία των (Anderson και Cohen 2013) με ένα σχετικά μεγάλο δείγμα 146 υποκειμένων (από τους οποίους 74 υγιείς και 72 σχιζοφρενείς), η ISOMAP χρησιμοποιήθηκε για την ανάλυση λειτουργικής διασυνδεσιμότητας (functional connectivity) σε εγκεφαλικά δίκτυα σε κατάσταση ηρεμίας (resting state) σε έναν χώρο χαμηλής διάστασης. Παρόλο που τα δεδομένα αυτά προ-επεξεργάστηκαν σε ένα πολύ στοιχειώδες επίπεδο (δεν έγινε ποιοτικός έλεγχος των ανεξάρτητων συνιστωσών), ανιχνεύτηκαν διαφορές σε ιδιότητες μικρού κόσμου (small world properties) ανάμεσα σε σχιζοφρενείς και υγιείς. Οι διαφορές αυτές αφορούσαν γραφοθεωρητικά στοιχεία γράφων που δημιουργήθηκαν από την εφαρμογή της μεθόδου ISOMAP. Μια εκτενέστερη αναφορά και ανάλυση μεθόδων στατιστικής εκμάθησης μηχανών και αναγνώρισης προτύπων για την ανάλυση λειτουργικής

διασυνδεσιμότητας μπορεί να βρεθεί στις εργασίες των (Richiardi et al. 2013, Siettos και Starke 2016).

3.1.2 Μέθοδος της λειτουργικής διασυνδεσιμότητας δικτύου και ταξινόμηση FMRI

Η μέθοδος της λειτουργικής διασυνδεσιμότητας δικτύου (Functional Connectivity Network) που θα εφαρμόσουμε και στην ανάλυση μας παρακάτω χρησιμοποιείται για να εξεταστεί αν ο συγχρονισμός ανάμεσα σε ανατομικά-οριζόμενες εγκεφαλικές περιοχές ή λειτουργικά - υποτιθέμενα δίκτυα διαφέρει λόγω ηλικίας, ασθένειας κτλ. ή αν η διαφορά αυτή σχετίζεται με μία διαδικασία που διεξάγεται. Διαφορές στη διασυνδεσιμότητα μεταξύ εγκεφαλικών δικτύων θεωρείται από την επιστημονική κοινότητα ότι μπορεί να είναι υπεύθυνες για αρκετές διαταραχές και νόσους όπως είναι ο αυτισμός (Koshino et al. 2005) και η σχιζοφρένεια (Liang et al. 2006).

Πιο γενικά, η χρονική διασυνδεσιμότητα μεταξύ εγκεφαλικών περιοχών έχει χρησιμοποιηθεί ευρέως σε δεδομένα FMRI στην προσπάθεια ανακάλυψης σχέσεων επιρροής ανάμεσα σε νευρωνικούς πληθυσμούς ((Roebroeck et al. 2005)) κάνοντας χρήση της αιτιότητας Granger αλλά και για να εντοπιστούν οι διαφορές ανάμεσα σε σχιζοφρενείς και υγιείς ανθρώπους (Garrity et al. 2007, Jafri et al. 2008, Anderson et al. 2010, Yu et al. 2011), χρησιμοποιώντας μέτρα πολλαπλής συσχέτισης για την ανάλυση των χρονοσειρών.

Συγκεκριμένα για τη σχιζοφρένεια, αλλοιωμένες ιδιότητες small-world δικτύων (η τυπική απόσταση L μεταξύ δύο τυχαίων κόμβων μεγαλώνει αναλογικά με το λογάριθμο του αριθμού των κόμβων N του δικτύου) βρέθηκαν μετά από σύγκριση σχιζοφρενών-υγιών ανάμεσα σε 90 φλοιώδεις και υποφλοιώδεις περιοχές του εγκεφάλου (Liu et al. 2008).

Ενισχυμένη τοπική λειτουργική διασυνδεσιμότητα (σε διάστημα 0.06-0.125 Hz) μαζί με μειωμένη ένταση του σήματος βρέθηκε στη σχιζοφρένεια από τους (Bassett et al. 2008). Στο default mode network (ένα δίκτυο αλληλεπιδρώντων περιοχών του εγκεφάλου με δραστηριότητα σε μεγάλο βαθμό συσχετιζόμενη μεταξύ αυτών και διακριτή από άλλα δίκτυα του εγκεφάλου), βρέθηκαν ασυνήθιστα μεγάλη λειτουργική διασυνδεσιμότητα και αλλοιωμένη χρονική συχνότητα (Garrity et al. 2007, Whitfield-Gabrieli et al. 2009). Οι ασθενείς με σχιζοφρένεια είχαν μεγαλύτερη συσχέτιση ανάμεσα σε επτά επιλεγμένα δίκτυα ήρεμης κατάστασης (resting state) σε σχέση με υγιείς (Jafri et al. 2008) ενώ διαφορές σε τοπολογικά μέτρα βρέθηκαν ανάμεσα σε δίκτυα ήρεμης κατάστασης που εξήχθησαν με τη μέθοδο group-ICA (ICA σε ομάδα) (Yu et al. 2011). Συλλογικά, οι μελέτες αυτές αλλά και πολλές άλλες, προτείνουν ότι όσον αφορά τη σχιζοφρένεια, μέτρα λειτουργικής διασυνδεσιμότητας μπορούν να χρησιμοποιηθούν για τον εντοπισμό γνωρισμάτων που είναι χαρακτηριστικά της ίδιας της ασθένειας.

Επειδή η λειτουργική διασυνδεσιμότητα εξαρτάται από τον τρόπο με τον οποίο ορίζονται δίκτυα ή περιοχές καθώς και το πως οι γραφικές ιδιότητες των περιοχών αυτών μπορούν να μετρηθούν (Toppi et al. 2012, Zalesky et al. 2012), είναι ζωτικής σημασίας για τους ερευνητές να μπορούν να προσαρμόσουν την ανάλυση διασυνδεσιμότητας στα δικά τους δεδομένα έτσι ώστε να επιτρέπουν στην επιστημονική κοινότητα να κρίνει τόσο την προ-υπάρχουσα γνώση τους όσο και το αποτέλεσμα της ανάλυσης αυτής καθ' αυτής. Για παράδειγμα, οι (Sato et al. 2010), εφάρμοσαν ανάλυση λειτουργικής διασυνδεσιμότητας ανάμεσα σε περιοχές ενδιαφέροντος (ROIs), τη στιγμή που οι (Chu et al. 2011) ανέλυσαν τη διασυνδεσιμότητα ανάμεσα σε εικονοστοιχεία. Όμοια, οι (Yu et al. 2011) εφάρμοσαν τη μέθοδο λειτουργικής διασυνδεσιμότητας δικτύου σε ανεξάρτητες συνιστώσες ομάδων

υποκειμένων (Group-ICA) ενώ οι (Anderson et al. 2010) σε ανεξάρτητες συνιστώσες ανά υποκείμενο (Single subject ICA).

Η μεγαλύτερη πρόκληση στην ταξινόμηση των FMRI είναι η αφθονία των παρατηρήσεων σε μία σάρωση, οι περισσότερες από τις οποίες είναι συσχετισμένες σε μεγάλο βαθμό τόσο ως προς το χρόνο όσο και το χώρο. Παρόλο που πολλά από τα εικονοστοιχεία είναι άδεια, αυτά δεν θα είναι συστηματικά άδεια για όλα τα υποκείμενα σαν αποτέλεσμα των διαφορών στο μέγεθος και το σχήμα του εγκεφάλου. Επειδή πολλά από τα σημεία στα δεδομένα είναι πλεονάζοντα, χρησιμοποιούνται τεχνικές μείωσης διάστασης δημιουργώντας στατιστικά δελτία ατομικά για τα εικονοστοιχεία (τ-έλεγχοι, έλεγχοι συσχέτισης), απομονώνοντας «περιοχές του ενδιαφέροντος» (ROI) ή εφαρμόζοντας κλασσικές μεθόδους μείωσης διάστασης - όπως η PCA -για να αναλυθεί ολόκληρη η σάρωση σε ορθογώνιες πηγές σήματος στο χρόνο. Νεότερες μέθοδοι όπως αυτή του ICA (Hyvärinen et al. 2000) μιμούνται την προσέγγιση του PCA με τη διαφορά ότι επιβάλλουν πρόσθετες υποθέσεις όπως είναι η στατιστική ανεξαρτησία στην περίπτωση αυτή (ICA).

Με την ολοκλήρωση των βημάτων εξαγωγής των κυρίων χαρακτηριστικών και της μείωσης διάστασης, τα μειωμένα σε διάσταση δεδομένα μπορούν να οδηγηθούν σε ταξινομητές όπως για παράδειγμα είναι ο SVM, τα νευρωνικά δίκτυα (Neural Nets) και οι αλγόριθμοι ώθησης (boosting algorithms). Αυτές οι τεχνικές ταξινόμησης έχουν χρησιμοποιηθεί πρόσφατα για να διαχωριστούν σαρώσεις PET (Positron Emission Tomography) με ασθενείς που είναι θετικοί στον HIV από υγιείς ανθρώπους (Liow et al. 2000), για να διαχωριστούν σαρώσεις FMRI με ασθενείς εθισμένους σε ναρκωτικές ουσίες από υγιείς (Lei Zhang et al. 2005) καθώς και για να διαχωριστούν ασθενείς με Alzheimer, σχιζοφρένεια, ψυχιατρικές ασθένειες και εγκεφαλικά τραύματα από υγιείς ανθρώπους του εκάστοτε δείγματος (Ford et al. 2003).

Ωστόσο αλγόριθμοι σαν τους παραπάνω κατασκευάζονται και προσαρμόζονται έτσι ώστε να ταιριάζουν σε κάθε σύνολο δεδομένων το οποίο καλούνται να διαχωρίσουν. Επομένως, είναι δύσκολο επικυρώσει κανείς αποτελέσματα ταξινόμησης ανάμεσα σε ομάδες ασθενών και υγειών. Επιπλέον, οι μελέτες αυτές αφορούν συνήθως πολύ μικρά δείγματα με περίπου $n \approx 20$ υποκείμενα. Όπως είναι εύκολα αντιληπτό, η επαναληψιμότητα τέτοιων αποτελεσμάτων είναι συχνά μη εφικτή, αφήνοντας ανοικτό το ενδεχόμενο κριτικής όσον αφορά μία καλή επίδοση στην ταξινόμηση. Η επίδοση αυτή, θα μπορούσε να οφείλεται κυρίως στο μοντέλο και όχι σε πραγματικές λειτουργικές και ανατομικές διαφορές μεταξύ των ομάδων του δείγματος. Ακόμα, είναι δύσκολο να χειριστεί κανείς τα δεδομένα που παράγονται στα εργαστήρια αφού οι παράμετροι σάρωσης, η ανάλυση εικόνας και η αλληλουχία των εικόνων πρέπει να είναι περίπου η ίδια. Για τους λόγους αυτούς, η αξιολόγηση των μοντέλων που χρησιμοποιούνται για την ανάλυση ή την ταξινόμηση θα πρέπει να αξιολογούνται με βάση διαφορετικά σύνολα δεδομένων ώστε να είμαστε όσο το δυνατόν πιο σίγουροι για την εγκυρότητα των αποτελεσμάτων. Μιας και τα μοντέλα που χρησιμοποιούνται είναι μοντέλα για την ανάλυση όμοιων ομάδων ασθενών θα πρέπει να παράγουν παρόμοια αποτελέσματα για διαφορετικά δείγματα του πληθυσμού.

Παρακάτω θα αναλυθούν εκτενώς τα βήματα και οι μέθοδοι που χρησιμοποιήθηκαν για την ανάλυση λειτουργικής διασυνδεσιμότητας συμπεριλαμβανομένου και της προσπάθειας ταξινόμησης των δύο ομάδων των δεδομένων μας (σχιζοφρενείς και υγιείς).Το ποσοστό επιτυχούς ταξινόμησης θα είναι και το κριτήριο για τη συγκριτική ανάλυση ανάμεσα στις γραμμικές και μη γραμμικές μεθόδους μείωσης διάστασης όπως και των μέτρων ομοιότητας που θα χρησιμοποιήσουμε.

3.2 ΠΑΡΟΥΣΙΑΣΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ FMRI ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ

Τα δεδομένα τα οποία θα αναλύσουμε αποτελούν πειραματικά δεδομένα λειτουργικής απεικόνισης μαγνητικού συντονισμού (FMRI) σε κατάσταση ηρεμίας (resting state) και είναι ανοικτά για το κοινό και την επιστημονική κοινότητα στην διεύθυνση :

http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html.

Τα δεδομένα αυτά παράχθηκαν και εν συνεχεία δόθηκαν στο κοινό και την επιστημονική κοινότητα από ένα Κέντρο Βιοϊατρικής Έρευνας Αριστείας (Center of biomedical research excellence). Συνοπτικά, τα Κέντρα Ιατροβιολογικών Ερευνών Αριστείας (COBRE) είναι κέντρα υπό την ηγεσία και χρηματοδότηση του Εθνικού Ιδρύματος Υγείας των Η.Π.Α . Χρηματοδοτούν ερευνητές με εμπειρία και ειδίκευση στο θέμα της πρότασης επιχορήγησης και υποστηρίζουν θεματικά, διεπιστημονικά κέντρα που στοχεύουν στην ενίσχυση της θεσμικής ικανότητας της βιοϊατρικής έρευνας.

Τα δεδομένα αφορούν FMRI σε κατάσταση ηρεμίας (resting state) ανάμεσα σε 72 ασθενείς διαγνωσμένους με κάποιο τύπο σχιζοφρένειας και 74 υγιείς ανθρώπους που συμμετείχαν στο πείραμα. Όλοι οι συμμετέχοντες στην πειραματική διαδικασία είχαν ηλικία από 18-65 χρονών.

Παρακάτω και συγκεκριμένα στην εικόνα 3.1 βλέπουμε τον δημογραφικό πίνακα με στοιχεία του δείγματος.

	N	Age (*SD*)	% Female	% Right-Handed
Schizophrenia	72	38.16 (13.89)	0.19	0.83
Patients	74	35.82 (11.58)	0.31	0.96

Εικόνα 3.1 *Δημογραφικός πίνακας που μας παρουσιάζει το πλήθος N των υγιών και σχιζοφρενών του δείγματος, τη μέση ηλικία (Age) και τυπική απόκλιση των δειγμάτων εντός της παρένθεσης (SD) καθώς και ποιο ποσοστό αυτών είναι αντίστοιχα γυναίκες και δεξιόχειρες [Anderson και Cohen, 2013].*

Όλοι οι συμμετέχοντες που έλαβαν μέρος στο πείραμα για την εξαγωγή των δεδομένων πέρασαν από μαγνητικό τομογράφο. Αντίθετα, εξαιρέθηκαν αυτοί οι οποίοι είχαν κάποιο ιστορικό νευρολογικής πάθησης, νοητικής καθυστέρησης, σοβαρό εγκεφαλικό τραύμα με περισσότερο από πέντε λεπτά απώλεια των αισθήσεων, χρήση ναρκωτικών ουσιών ή εξάρτηση έως και 12 μήνες πριν από τη συμμετοχή τους στο πείραμα. Οι πληροφορίες για την διάγνωση των συμμετεχόντων έγινε με χρήση της δομημένης κλινικής συνέντευξης για την ταξινόμηση σύμφωνα με το Διαγνωστικό και Στατιστικό Εγχειρίδιο των Ψυχικών Διαταραχών (DSM).

Τεχνικά Χαρακτηριστικά

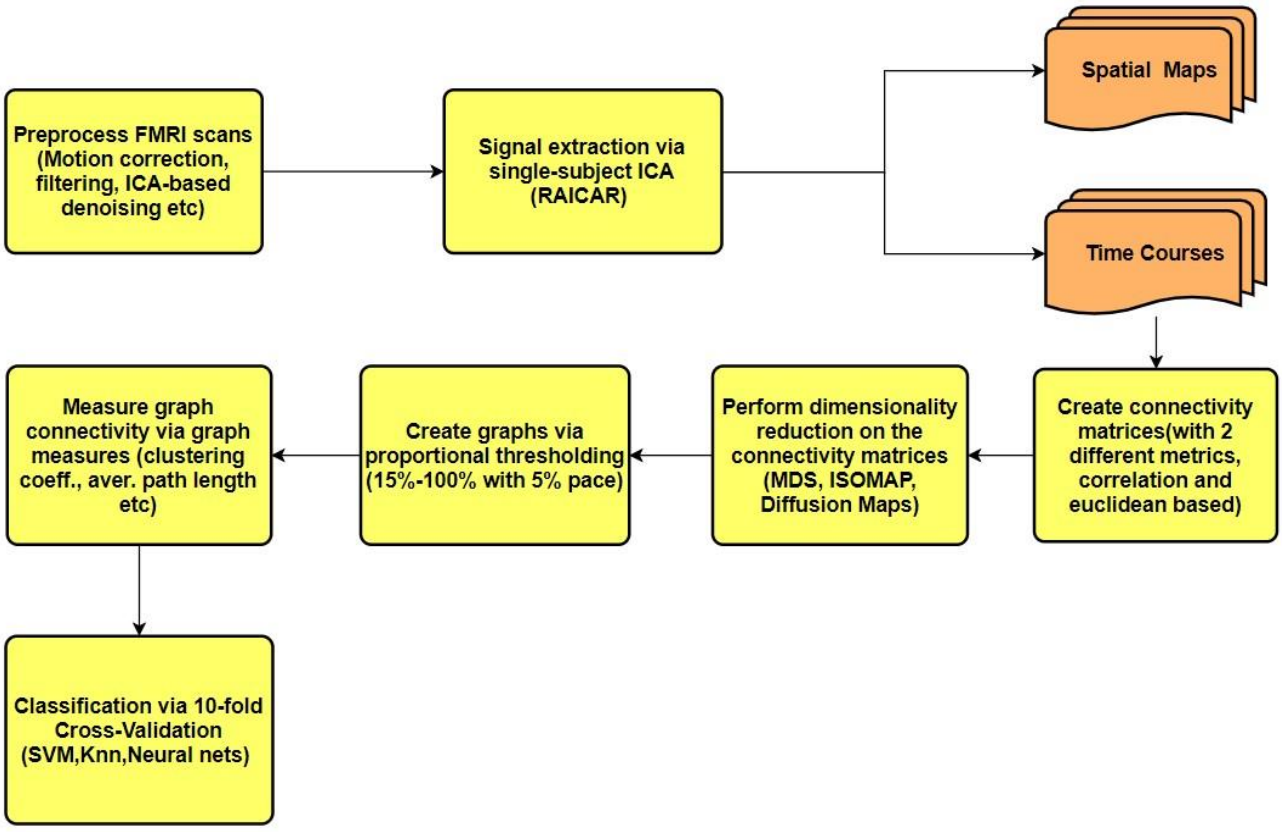
Τα δεδομένα συλλέχθηκαν σε κατάσταση ηρεμίας του υποκειμένου (resting state) με ενιαία λήψη, με τη μέθοδο *k* -space echo-planar imaging (EPI), με διόρθωση αναβάθμισης δειγματοληψίας χρησιμοποιώντας διασυνδετική γραμμή (intercomissural line) (AC - PC).

Για τις βασικές παραμέτρους του μαγνήτη κατά τη διαδικασία λήψης: TR: 2 s, TE : 29 ms, το μέγεθος πίνακα : 64x64 σε 32 φέτες του εγκεφάλου και το μέγεθος του εικονοστοιχείου (voxel): 3x3x4 mm^3 .

Η παράμετρος TE (echo time) αντιπροσωπεύει το χρονικό διάστημα που μεσολαβεί ανάμεσα στην εκπομπή των RF (radio frequency) παλμών και τη μέτρηση του σήματος που εκπέμπουν οι πυρήνες πίσω λόγω του φαινομένου του πυρηνικού συντονισμού (βλέπε Εισαγωγή).Η παράμετρος TR (repetition time) αφορά το χρονικό διάστημα που μεσολαβεί από την εκπομπή ενός RF παλμού μέχρι την επόμενη εκπομπή.

3.3 ΜΕΘΟΔΟΛΟΓΙΑ ΓΙΑ ΤΗΝ ΑΝΑΛΥΣΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Στο υποκεφάλαιο αυτό της εργασίας, αναλύονται επ' ακριβώς οι μέθοδοι και τα βήματα που έγιναν για την ανάλυση λειτουργικής διασυνδεσιμότητας δικτύου ανάμεσα στις δύο ομάδες του δείγματος. Σκοπός μας ήταν η συγκριτική ανάλυση μεθόδων εκμάθησης πολλαπλοτήτων (χρησιμοποιώντας δύο μέτρα ομοιότητας) για την μείωση διάστασης και την



Εικόνα 3.2 *Αναπαράσταση της διαδικασίας που ακολουθήθηκε για την ανάλυση των δεδομένων σε διάγραμμα ροής (flow chart).*

ταξινόμηση των δεδομένων. Η συγκριτική ανάλυση πραγματοποιείται σε δεδομένα απεικόνισης μαγνητικού συντονισμού (FMRI) σε υγιείς και ασθενείς με σχιζοφρένεια. Η εξαγωγή του σήματος ενδιαφέροντος από τα FMRI γίνεται με χρήση της μεθόδου ανάλυσης ανεξάρτητων συνιστωσών (ICA) ξεχωριστά για κάθε υποκείμενο (Single-subject

ICA).Παρακάτω θα κάνουμε μία περίληψη της ανάλυσης που πραγματοποιήσαμε όπως αυτή φαίνεται σε διάγραμμα ροής στην Εικόνα 3.2.

Αρχικά, έγινε η προ-επεξεργασία των δεδομένων (διόρθωση κίνησης κτλ.) και στη συνέχεια κάναμε χρήση της γραμμικής μεθόδου ICA (για κάθε υποκείμενο ξεχωριστά) για να αναλύσουμε κάθε σάρωση σε ένα σύνολο από χωρικά ανεξάρτητα εγκεφαλικά δίκτυα που μεταβάλλονται στο χρόνο σύμφωνα με τις αντίστοιχες χρονοσειρές (παρακάτω επικεντρωνόμαστε στις χρονοσειρές αυτές).

Στη συνέχεια φτιάξαμε πίνακες αποστάσεων για κάθε υποκείμενο ξεχωριστά χρησιμοποιώντας 2 μετρικές (ως μέτρα ομοιότητας ανάμεσα στις χρονοσειρές). Η πρώτη βασίστηκε στην πολλαπλή συσχέτιση με καθυστερήσεις (time lags) ενώ η δεύτερη στην ευκλείδεια νόρμα. Στη συνέχεια πραγματοποιήσαμε μείωση διάστασης στους πίνακες αυτούς (2 πίνακες για κάθε υποκείμενο) χρησιμοποιώντας τρεις μεθόδους, την MDS ,την ISOMAP και την Diffusion Maps απεικονίζοντας τα δεδομένα σε ένα χαμηλής διάστασης χώρο (από 2 έως 5 διαστάσεις).

Έπειτα, υπολογίσαμε τις αποστάσεις των σημείων μεταξύ τους στον νέο χώρο χαμηλότερης διάστασης και με μία προσέγγιση πολλαπλού αναλογικού κατωφλιού (proportional thresholding) δημιουργήσαμε γράφους όπου οι συνδέσεις μεταξύ των σημείων αυξάνονταν αναλογικά από το 15% των δυνατότερων (πιο κοντινών σημείων στο χώρο) συνδέσεων μέχρι και το 100% με βήμα 5%. Στο 100% έχουμε απόλυτα συνδεδεμένο γράφο με κάθε σημείο να συνδέεται με όλα τα υπόλοιπα σημεία. Με την προσέγγιση αυτή μπορέσαμε να οπτικοποιήσουμε ολικά το μοτίβο διασυνδεσιμότητας χωρίς να πρέπει να διαλέξουμε μία αυθαίρετη τιμή κατωφλιού (threshold).

Για κάθε γράφημα μετρήσαμε κάποια από τα πιο γνωστά γραφοθεωρητικά μέτρα όπως είναι το μέσο μήκος μονοπατιού και ο συντελεστής συγκρότησης. Τα μέτρα αυτά δόθηκαν σε αλγόριθμους ταξινόμησης για το διαχωρισμό των υποκειμένων σε υγιείς και ασθενείς με σχιζοφρένεια.

Τέλος, πραγματοποιήσαμε στατιστικούς ελέγχους, για να εκτιμήσουμε τη διασυνδεσιμότητα (connectivity) των γράφων και να εξάγουμε συμπεράσματα που αφορούν ποιοτικές διαφορές ανάμεσα σε ασθενείς με σχιζοφρένεια και υγιείς ανθρώπους του δείγματος ενώ παράλληλα πραγματοποιήσαμε ανάλυση διασποράς (ANOVA) 2 παραγόντων για την αξιολόγηση των μεθόδων εκμάθησης πολλαπλοτήτων και των αλγορίθμων ταξινόμησης που χρησιμοποιήθηκαν.

3.3.1 Προ-επεξεργασία των δεδομένων

Το κύριο μέρος της προ-επεξεργασίας των δεδομένων πραγματοποιήθηκε κάνοντας χρήση του λογισμικού επεξεργασίας δεδομένων MRI FSL (FMRIB's Software Library,www.fmrib.ox.ac.uk/fsl). Οι ακόλουθες διαδικασίες προ-επεξεργασίας εφαρμόστηκαν κατά σειρά (μέσω του λογισμικού FEAT που αποτελεί μέρος του μεγαλύτερου πακέτου FSL): διόρθωση κίνησης χρησιμοποιώντας την συνάρτηση MCFLIRT ((Jenkinson et al. 2002)), χρονική διόρθωση της λήψης των φετών του εγκεφάλου πραγματοποιώντας αλλαγή φάσης των χρονοσειρών σε χώρο Φουριέ. Αφαίρεση μη-εγκεφαλικών ιστών από τα δεδομένα χρησιμοποιώντας την συνάρτηση BET (Smith 2002), χωρική εξομάλυνση με χρήση Γκαουσιανού πυρήνα (Full Width Half Maximum) 5 χιλιοστών, κανονικοποίηση ολικού μέσου όρου (grand-mean) της έντασης του σήματος όλων των εικονοστοιχείων με έναν απλό πολλαπλασιαστικό παράγοντα.

Στη συνέχεια έγινε χρήση της μεθοδολογίας ICA AROMA (Pruim et al., 2015) για την εύρεση ανεξάρτητων συνιστωσών που σχετίζονταν με θόρυβο ο οποίος και αφαιρέθηκε. Τέλος, εφαρμόστηκε ένα φίλτρο υψηλών συχνοτήτων στα 100 Hz για να καθαρίσει τα δεδομένα από θόρυβο των χαμηλών συχνοτήτων.

Για τη διαδικασία της προ-επεξεργασίας χρησιμοποιήθηκε το ελεύθερο λογισμικό FSL που είναι ένα εργαλείο ειδικευμένο για την ανάλυση δεδομένων από MRI απεικονίσεις, ενώ για την μέθοδο ICA AROMA χρησιμοποιήθηκε η έκδοση που διατίθεται ελεύθερα από τους δημιουργούς στη διεύθυνση: https://github.com/maartenmennes/ICA-AROMA.

3.3.2 Ανάλυση ανεξάρτητων συνιστωσών - εξαγωγή του σήματος ενδιαφέροντος

Για την εξαγωγή των περιοχών ενδιαφέροντος χρησιμοποιήσαμε τη μέθοδο RAICAR (Yang et al. 2008) (Ranking and Averaging of Independent Components by Reproducibility) η οποία τρέχει επαναληπτικά την μέθοδο ICA με σκοπό να παράγει ανεξάρτητες συνιστώσες τις οποίες στη συνέχεια ταξινομεί ανάλογα με την τάση της κάθε συνιστώσας να επανεμφανίζεται και πάλι ως ανεξάρτητη ανάμεσα σε συνεχόμενες εφαρμογές της μεθόδου (ICA). Ο λόγος που χρησιμοποιήσαμε την μέθοδο RAICAR είναι το γνωστό πλέον πρόβλημα των αλγορίθμων που υλοποιούν την ICA, οι οποίοι χρησιμοποιούν γεννήτριες τυχαίων πινάκων και στη συνέχεια το πρόβλημα βελτιστοποίησης για την εξαγωγή των συνιστωσών παρουσιάζει μία μεταβλητότητα στα αποτελέσματα ακόμα και με ίδιες παραμέτρους (Himberg et al. 2004). Επομένως, η χρήση της μεθόδου RAICAR έχει να κάνει με τον καλύτερο εντοπισμό των περιοχών ενδιαφέροντος που θα κάνει τα τελικά αποτελέσματά της εργασίας μας περισσότερο αξιόπιστα και αναπαραγώγιμα με μεγάλη ακρίβεια.

Όπως είδαμε και παραπάνω, η μέθοδος ICA έχει αποκτήσει ιδιαίτερη φήμη για την εφαρμογή της σε δεδομένα FMRI. Αυτή η μέθοδος, μπορεί να απομονώσει δίκτυα ανάλογα με την δραστηριότητα των νευρώνων, γεγονός που έχει επικυρωθεί με την ανίχνευση κοινών χαρακτηριστικών που υπάρχουν στο εύρος των υποκειμένων (πολλά διαφορετικά δείγματα) και των σαρώσεων που αντιστοιχούν σε λειτουργικά ανιχνεύσιμα εγκεφαλικά δίκτυα (McKeown et al. 2003, Anderson et al. 2011).Παρακάτω θα δούμε πως λειτουργεί η μέθοδος συγκεκριμένα για τα δεδομένα FMRI.

Υπό την υπόθεση ότι η δραστηριότητα του εγκεφάλου είναι κατασκευασμένη σε ανατομικά δίκτυα που αλληλοεπιδρούν για να παράγουν σημαντικές ψυχοσυμπεριφορικές καταστάσεις, η συνολική δραστηριότητα αναλύεται σε υποσυνιστώσες με τη βοήθεια της ICA. Στην ουσία, ο χώρος «ξετυλίγεται». Η τεσσάρων διαστάσεων σάρωση (3 διαστάσεις χώρου και 1 χρόνου) μετασχηματίζεται σε έναν πίνακα με διαστάσεις χώρου επί το χρόνο, έτσι ώστε μία σάρωση διάστασης *(x,y,z,t)* να γίνει ένας πίνακας διάστασης *(t (χρόνος), $x \times y \times z$ (χώρος)).*

Έτσι μία σάρωση χρονικής διάρκειας T και χωρικής διάστασης S μπορεί να εκφραστεί σαν γραμμικός συνδυασμός από $M < T$ συνιστώσες και αντίστοιχες χρονοσειρές :

$$X_{ts} = \sum_{\mu=1}^{M} A_{t\mu} C_{\mu s} \tag{3.1}$$

Όπου με X_{ts} συμβολίζεται η ένταση της σάρωσης τη χρονική στιγμή $t < T$ και χωρικής τοποθεσίας $s < S$, $A_{t\mu}$ είναι το εύρος της συνιστώσας μ σε χρόνο t, και $C_{\mu s}$ είναι το χωρικό μέγεθος για τη συνιστώσα μ στη χωρική τοποθεσία s .

Οι συνιστώσες *C* εκτιμώνται έτσι ώστε να είναι όσο το δυνατόν στατιστικά (χωρικά) ανεξάρτητες λύνοντας ουσιαστικά το αντίστροφο πρόβλημα -στην περίπτωση μας- με τον αλγόριθμο FAST-ICA. Για την εκτίμηση των ανεξάρτητων πηγών *c* στο γενικευμένο πρόβλημα $x = Ac$ λύνεται το αντίστροφο πρόβλημα $y = w^T x$ όπου w είναι μία γραμμή του A^{-1} (αντίστροφος του πίνακα μίξης). Στη συνέχεια,

$$y = w^T x \Rightarrow \ y = w^T Ac \tag{3.2}$$

και αντικαθιστώντας $z = A^T w$ στην 3.2 έχουμε :

$$y = z^T c \tag{3.3}$$

Το *w* βελτιστοποιείται έτσι ώστε το σήμα $y = w^T x = z^T c$ να είναι όσο το δυνατόν λιγότερο Γκαουσιανή πηγή, πράγμα που συνεπάγεται και οδηγεί σε ακόμα λιγότερο Γκαουσιανές πηγές *c* λόγω του Κεντρικού Οριακού Θεωρήματος.Η Μεγιστοποίηση της κύρτωσης, η ελαχιστοποίηση της εντροπίας και η μεγιστοποίηση της αρνητικής εντροπίας (negentropy) για το *w* είναι όλες μέθοδοι για να βρει κανείς τις λιγότερο Γκαουσιανές $y = w^T x = z^T c$.

Ωστόσο, εμείς θα ασχοληθούμε με την αρνητική εντροπία (negentropy), την οποία υπολογίζει η Fast-ICA ώστε να αναλύσει τα δεδομένα σε ανεξάρτητες συνιστώσες. Έχουμε λοιπόν :

$$Negentropy = J(y) = H(y\_Gauss\ ) - H(y) \tag{3.4}$$

Όπου,

$$H(y) = -\sum_i P(y = a_i)\log[P(y = a_i)] \tag{3.5}$$

Ενώ για τη συνεχή περίπτωση η σχέση 3.5 γίνεται:

$$H(y) = -\int f(y)\log\big(f(y)\big)\,dy \tag{3.6}$$

Με διεξαγωγή 1000 το πολύ επαναλήψεων για τον αλγόριθμο Fast-ICA χρησιμοποιούμε την συνάρτηση:

$$G(u) = \frac{1}{a}\log[\cosh(au)] \tag{3.7}$$

Όπου, $\alpha \in [1,2]$ και είναι η σταθερά που χρησιμοποιείται για να εκτιμηθεί η αρνητική εντροπία.

Ο τρόπος με τον οποίο η Fast-ICA κάνει τον υπολογισμό του αριθμού των ανεξάρτητων συνιστωσών που πρόκειται να υπολογίσει είναι βασισμένη συνήθως στην εύρεση των ιδιοτιμών του πίνακα διασποράς για το κάθε υποκείμενο (χρησιμοποιεί δηλαδή μία PCA και απορρίπτει τις διαστάσεις που σχετίζονται με εξαιρετικά χαμηλές ιδιοτιμές). Το θέμα αυτό είναι αρκετά επίμαχο αφού πολλές φορές αυτός ο αριθμός μπορεί να επηρεάσει τα συνολικά αποτελέσματα της μεθόδου (Ma et al. 2007).

Ως επακόλουθο, καταλαβαίνουμε ότι κάθε υποκείμενο θα έχει διαφορετικό αριθμό ανεξάρτητων συνιστωσών για τη μετέπειτα ανάλυση. Το γεγονός αυτό είναι λογικό αν σκεφτεί κανείς ότι τα δεδομένα είναι θορυβώδη και ανάμεσα στα υποκείμενα υπάρχουν ανατομικές διαφορές κτλ.

Με τη βοήθεια της μεθόδου RAICAR ωστόσο, είμαστε όσο το δυνατόν βέβαιοι ότι οι ανεξάρτητες συνιστώσες για κάθε υποκείμενο είναι πράγματι ανεξάρτητες σε έναν συνεχόμενο αριθμό αποσυνθέσεων (decomposition) ICA, αγνοώντας εκείνες τις συνιστώσες που έχουν μικρό αναπαραγώγιμο βαθμό.

3.3.3 Κατασκευή πινάκων λειτουργικής διασυνδεσιμότητας

Για την κατασκευή πινάκων λειτουργικής διασυνδεσιμότητας (δύο πίνακες για κάθε υποκείμενο) χρησιμοποιήθηκαν δύο μετρικές από τις οποίες, η μία βασίστηκε στη πολλαπλή συσχέτιση χρονοσειρών (με καθυστερήσεις) -που αντιστοιχούν σε χωρικούς χάρτες του εκάστοτε υποκειμένου- και η άλλη στην ευκλείδεια νόρμα (παρακάτω θα συμβολίζουμε απλώς με A_i την χρονοσειρά που προκύπτει απευθείας από την ICA). Για την πρώτη, η συνάρτηση απόστασης $d(A_{i,}A_j)$ που παρουσιάζεται στη σχέση 3.8 είναι ένας μετασχηματισμός από μέγιστες - σε απόλυτη τιμή πολλαπλές συσχετίσεις (για όλα τα lags) - μεταξύ δύο χρονοσειρών. Ο υπολογισμός αυτός γίνεται για κάθε πιθανό ζευγάρι χρονοσειρών ενός υποκειμένου και έτσι μετασχηματίζουμε τα δεδομένα FMRI σε έναν πίνακα. Αυτή η μετρική, είναι ένα μέτρο για τη λειτουργική διασυνδεσιμότητα ανάμεσα σε δύο συνιστώσες αλλά είναι απλώς ένα από τα πιθανά (μέτρα) που θα μπορούσαμε να χρησιμοποιήσουμε (π.χ. συνοχή (coherence), αμοιβαία πληροφορία (mutual information) κ.α.) .

Η συνάρτηση πολλαπλής συσχέτισης (Cross correlation function) μεταξύ των χρονοσειρών υπολογίζεται στο εύρος των χρονικών καθυστερήσεων (temporal lags). Αφαιρούμε λοιπόν τη μέγιστη απόλυτη τιμή της πολλαπλής συσχέτισης από τη μονάδα για να φτιάξουμε τη μετρική, $d(A_i, A_j)$ που δίνεται από τη σχέση :

$$d(A_i, A_j) = 1 - max[\ |CCF(A_i, A_j, l)|\ ] \tag{3.8}$$

Η συνάρτηση πολλαπλής συσχέτισης CCF δίνεται από τη σχέση 3.9 :

$$\mathrm{CCF}(A_i, A_j, \mathrm{l}) = \frac{\mathrm{E}[(A_i(t+l) - \overline{A}_i)(A_j(t) - \overline{A}_j)]}{\sqrt{\mathrm{E}\left[\left(A_i(t) - \overline{\mathrm{A}}_{\mathrm{i}}\right)^2\right] \mathrm{E}\left[\left(A_j(t) - \overline{A}_j\right)^2\right]}} \tag{3.9}$$

Όπου *l* είναι η χρονική καθυστέρηση που χωρίζει δύο χρονοσειρές, A_i και A_j, ο δείκτης αναφέρεται στην ανεξάρτητη συνιστώσα (χωρικός χάρτης που αντιστοιχεί η χρονοσειρά) με την οποία δουλεύουμε, και $\bar{A}_i$ είναι ο μέσος όρος της $A_i = (a(t_1), a(t_2), \dots \dots, a(t_n))$ όπου για τα δεδομένα που θα αναλύσουμε παρακάτω είναι n=150 σημεία.

Οι χρονοσειρές υπολογίζονται σε χρονικές καθυστερήσεις από 0 έως 3 χρονικά σημεία που αντιστοιχούν σε περίπου 6 δευτερόλεπτα (TR=2 δευτερόλεπτα). Μεγαλύτερες χρονικές καθυστερήσεις έχουν ως αποτέλεσμα λιγότερα χρονικά σημεία να υπολογίζουν τη συσχέτιση, κάνοντας μία θορυβώδη εκτίμηση που εκτός των άλλων, δεν θα ήταν και λογική από νευροβιολογική άποψη.

Η δεύτερη μετρική που χρησιμοποιήθηκε είναι η συνήθης ευκλείδεια απόσταση που δίνεται από τη σχέση:

$$L_2(A_i, A_j) = \sqrt{\sum_{t=1}^{n} \left(A_i(t) - A_j(t)\right)^2} \tag{3.10}$$

Ωστόσο, οι πίνακες λειτουργικής διασυνδεσιμότητας που προκύπτουν από τη χρήση των μετρικών ανάμεσα στις χρονοσειρές είναι δύσκολο να ερμηνευθούν αφού αποτελούν απλώς αναπαραστάσεις ενός συνόλου συνδεδεμένων αντικειμένων σε υψηλής διάστασης χώρο. Ο πρωταρχικός μας στόχος ήταν να μετρήσουμε διασυνδεσιμότητα, όχι μόνο το πόσο στενά συνδεδεμένα είναι τα αντικείμενα αλλά επίσης το πως αυτή η διασυνδεσιμότητα αλλάζει αντίστοιχα με την ομάδα που μελετάμε (σχιζοφρενείς-υγιείς).Αυτό που θα επιχειρήσουμε να κάνουμε είναι αρχικά να μειώσουμε τη διάσταση τους και στη συνέχεια να δημιουργήσουμε γράφους στον χώρο ενσωμάτωσης (embedded space).

Ο τρόπος με τον οποίο γίνεται η κατασκευή των γράφων είναι με το να θέσουμε κάποιες ανώτατες τιμές (threshold) απόστασης και να συγκρίνουμε αν δυο χρονοσειρές βρίσκονται κάτω από το όριο αυτής της τιμής, αν ναι, τότε θα θεωρούμε ότι οι συγκεκριμένες χρονοσειρές είναι συνδεδεμένες. Στην εργασία μας ακολουθήθηκε η προσέγγιση της πολλαπλής αναλογικής τιμής κατωφλιού (multiple proportional thresholding).

Με τον τρόπο αυτό μπορούμε να κατασκευάσουμε δίκτυα ,να μετρήσουμε τις ιδιότητες τους και να απαντήσουμε σε ερευνητικά ερωτήματα. Για παράδειγμα, αλληλεπιδρά κάθε δίκτυο με όλα τα υπόλοιπα; Υπάρχουν υπογράφοι που είναι αποκομμένοι από τον αρχικό γράφο; Διαφέρει ο αριθμός των βημάτων για να πάμε από έναν κόμβο σε έναν άλλο; Για να απαντηθούν αυτές οι ερωτήσεις θα πρέπει να μετατρέψουμε τους πίνακες διασυνδεσιμότητας σε γράφους και μετά να συνεχίσουμε με ανάλυση διασυνδεσιμότητας δικτύου-γράφου. Παρακάτω αναλύεται η μεθοδολογία της μείωσης διάστασης και της κατασκευής γράφων.

3.3.4 Μείωση διάστασης των πινάκων απόστασης και κατασκευή γράφων με τη μέθοδο του πολλαπλής αναλογικής τιμής κατωφλιού (multiple proportional thresholding)

Οι πίνακες λειτουργικής διασυνδεσιμότητας είναι τετραγωνικοί συμμετρικοί πίνακες που έχουν υψηλή διάσταση και είναι δύσκολο να συγκριθούν μεταξύ τους (Anderson and Cohen 2013) . Θα προσπαθήσουμε λοιπόν να μειώσουμε τη διάσταση τους με τις μεθόδους που αναλύθηκαν εκτενώς στο κεφάλαιο 2. Οι διαστάσεις ενσωμάτωσης d που εξετάστηκαν για κάθε μέθοδο ήταν $2 \leq d \leq 5$.

Όπως είδαμε σε προηγούμενο κεφάλαιο, η MDS είναι μία μέθοδος που στόχο έχει να διατηρήσει τις αποστάσεις σημείων μεταξύ τους (στον αρχικό χώρο) σε έναν ευκλείδειο χώρο χαμηλότερης διάστασης από τον αρχικό. Έτσι στην περίπτωση μας η είσοδος στη μέθοδο είναι ένα σύνολο συνδεδεμένων αντικειμένων σε έναν υψηλής διάστασης χώρο που στη μία περίπτωση οι αποστάσεις μεταξύ των σημείων είναι βασισμένες στην πολλαπλή συσχέτιση με καθυστερήσεις και στην άλλη περίπτωση οι αποστάσεις είναι ευκλείδειες. Επομένως αυτό που περιμένουμε από τη μέθοδο είναι μία χαμηλής διάστασης ενσωμάτωση η οποία θα διατηρεί τις ευκλείδειες αποστάσεις μεταξύ των σημείων (στον αρχικό χώρο) κατά τον βέλτιστο δυνατό τρόπο (βλέπε κεφάλαιο 2). Αν το πρότυπο/μοτίβο των συνδεδεμένων σημείων δεν εμφανίζει έντονες μη γραμμικότητες ευελπιστούμε σε μία απεικόνιση που δεν έχει χαμένη πληροφορία και μπορεί ίσως να μας αποκαλύψει ιδιότητες που σε έναν υψηλής διάστασης χώρο δεν θα μπορούσαμε ενδεχομένως να αντιληφθούμε. Για την κατασκευή γράφων εφαρμόστηκε η διαδικασία πολλαπλών αναλογικών τιμών κατωφλιού στο νέο χώρο χαμηλής διάστασης από το 15% των μικρότερων αποστάσεων μέχρι το 100% με βήμα 5%.

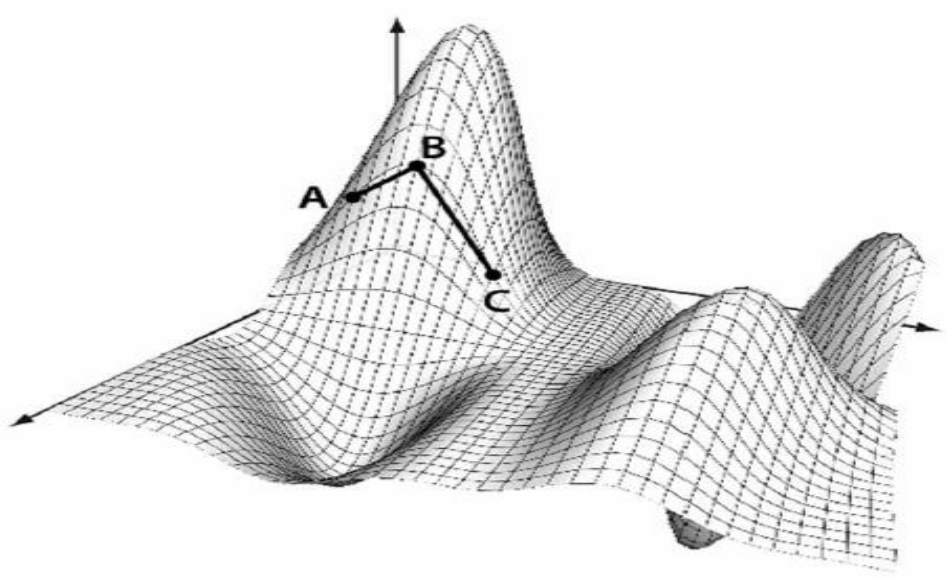
Τέλος, η εφαρμογή της μεθόδου είναι απλή αφού δεν χρειάζεται να ψάξουμε για κάποια ρυθμιστική παράμετρο σε αντίθεση με τις δύο επόμενες μη γραμμικές μεθόδους.

Εννοιολογικά, κάθε σύνολο σημείων που περιέχεται στον πίνακα αποστάσεων διάστασης D που υπολογίσαμε, μπορεί να ενσωματωθεί σε ένα χώρο διάστασης $D - 1$ χωρίς να έχουμε απώλεια πληροφορίας (πρακτικά όλες οι αποστάσεις μπορούν να διατηρηθούν). Συνήθως ένας τέτοιος μετασχηματισμός υποθέτει ότι ο χώρος στον οποίο βρίσκονται τα δεδομένα είναι γραμμικός. Ωστόσο μία τέτοια υπόθεση μπορεί να είναι λανθασμένη.

Για παράδειγμα, αν προσπαθούσαμε να μετρήσουμε την απόσταση από το Σακραμέντο στην Αμερική, μέχρι την Σανγκάη στην Κίνα -χρησιμοποιώντας τις συντεταγμένες της κάθε πόλης στο τρις ορθογώνιο σύστημα αξόνων (x, y, z) - η γραμμική απόσταση μεταξύ των πόλεων ενώ θα ήταν υπολογίσιμη, θα θεωρούσε ότι το σωστό και συντομότερο μονοπάτι περνάει από τον πυρήνα της γης, κάτι που είναι πρακτικά αδύνατο.

Αντί αυτού λοιπόν, ένας πιο ορθολογικός τρόπος να μετρήσουμε την απόσταση θα ήταν μέσω διασυνδετικών σημείων όπου θα μπορούσαμε να πετάξουμε με αεροπλάνο και να φτάσουμε π.χ. από το Σακραμέντο στο Λος Άντζελες, από το Λος Άντζελες στο Τόκυο και τελικά από το Τόκυο στη Σανγκάη. Αυτή η μέθοδος μέτρησης της απόστασης είναι γνωστή ως

γεωδαισιακή απόσταση ή απόσταση μεταξύ συνδετικών σημείων. Υποθέτει δηλαδή, ότι το ταξίδι ανάμεσα σε μακρινά σημεία, συχνά απαιτεί πορεία μέσω ενδιάμεσων κόμβων όπως



Εικόνα 3.3 *Η απόσταση ανάμεσα στο A και το C υπολογίζεται σαν το μονοπάτι από το A στο B και τελικά στο C, αντί της γραμμική απόστασης απευθείας από το A στο C. Το γεγονός αυτό παραβιάζει την υπόθεση ότι τα σημεία βρίσκονται σε ένα γραμμικό χώρο χρησιμοποιώντας ευκλείδεια απόσταση [Anderson και Cohen, 2013].*

φαίνεται παρακάτω στην Εικόνα 3.3 .

Αυτή λοιπόν την προσέγγιση θα χρησιμοποιήσουμε κι εμείς για να «εξαφανίσουμε» τις πολύ αδύναμες συνδέσεις, αντικαθιστώντας αυτές με διασυνδετικά μονοπάτια δημιουργώντας γράφους από τους πίνακες διασυνδεσιμότητας που κατασκευάσαμε. Μετατρέπουμε λοιπόν κάθε πίνακα σε μία δομή γράφου χρησιμοποιώντας αρχικά μία από τις δύο βασικές προσεγγίσεις για την εφαρμογή της ISOMAP (Tenenbaum et al. 2000).

Οι προσεγγίσεις αυτές έχουν να κάνουν - όπως προαναφέρθηκε και στο δεύτερο κεφάλαιο - με τη συνθήκη γειτνίασης μεταξύ των κόμβων μας. Είτε ορίζουμε μία ακτίνα *ε,* βάση της οποίας κάθε σημείο που βρίσκεται σε απόσταση ίση με την ακτίνα αυτή θεωρείται ότι γειτνιάζει με το σημείο αυτό, είτε εφαρμόζουμε την μέθοδο κοντινότερων γειτόνων (*k*-nearest neighbours, *k*-NN).

Με τον τρόπο αυτό καταφέρνουμε να ξεσκαρτάρουμε τις πολύ αδύναμες συνδέσεις αντικαθιστώντας τις αποστάσεις αυτές (μεταξύ μη συνδεδεμένων κόμβων) με την ελάχιστη απόσταση μεταξύ συνδεδεμένων κόμβων. Οι ελάχιστες αποστάσεις από κάθε σημείο σε οποιοδήποτε άλλο σημείο του γράφου υπολογίζονται μέσω αλγορίθμων όπως του Dijkstra ή Floyd - Warsal κτλ. Συνεχίζοντας την εφαρμογή της ISOMAP σαν τελευταίο βήμα έχουμε την εφαρμογή της γραμμικής μεθόδου MDS που ψάχνει μία βέλτιστη προσαρμογή σε χάρτη χαμηλότερης διάστασης, των συνδεδεμένων κόμβων - ανεξάρτητων συνιστωσών του κάθε υποκειμένου.

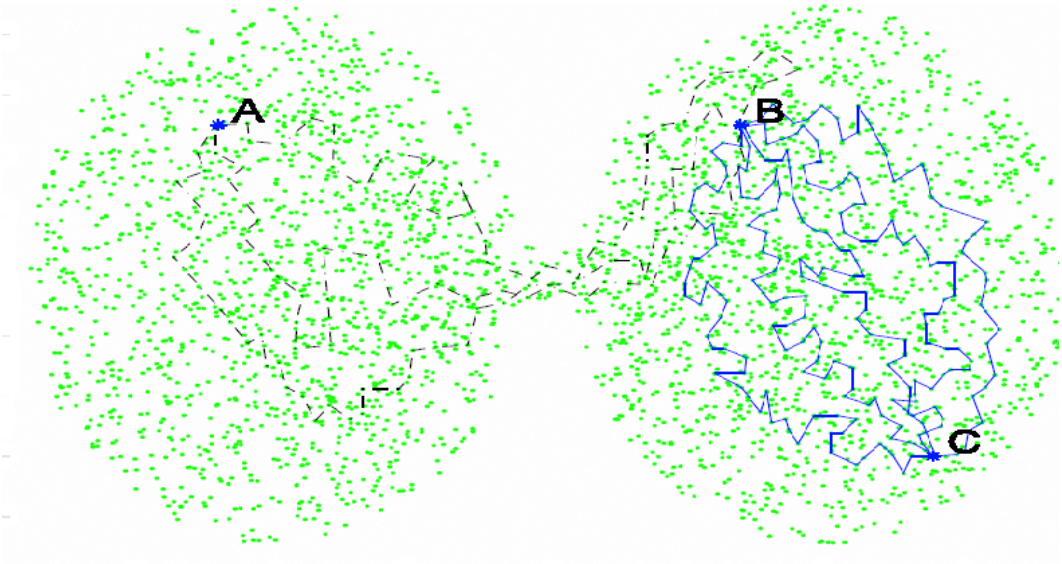
Για τη συνθήκη γειτνίασης της μεθόδου ISOMAP στην παρούσα εργασία χρησιμοποιήθηκε η απόσταση ε. Η μέθοδος επιστρέφει ένα γράφο με αποτέλεσμα να μην χρειάζεται όπως σε άλλες μεθόδους να θέσουμε τιμή κατωφλιού αργότερα στο χώρο ενσωμάτωσης. Η τιμή του ε επιλέχθηκε σε ένα εύρος τιμών από το 15% των χαμηλότερων

αποστάσεων στον πίνακα απόστασης έως το 100% με βήμα 5% για να υπάρχει αντιστοιχία και με τις άλλες μεθόδους τέτοια ώστε να είναι δυνατή μία σύγκριση αργότερα. Με αυτόν τον τρόπο η ISOMAP δημιουργεί γράφους σε διαφορετικά επίπεδα σύνδεσης ώστε να μπορούμε να παρατηρήσουμε ένα ολόκληρο μοτίβο διασυνδεσιμότητας.

Τέλος η Τρίτη μέθοδος που χρησιμοποιήθηκε ήταν η μέθοδος των απεικονίσεων διάχυσης η οποία σε αντίθεση με τη μέθοδο ISOMAP που στοχεύει στη διατήρηση της γεωδαισιακής απόστασης μεταξύ των σημείων, προσπαθεί να προσεγγίσει σε ένα χαμηλότερης διάστασης χώρο τις αποστάσεις διάχυσης μεταξύ των σημείων. Σε αντίθεση με τη γεωδαισιακή απόσταση, η απόσταση διάχυσης μοιάζει να είναι πιο ανθεκτική και αξιόπιστη (robust) στο θόρυβο, αφού λαμβάνει υπόψη όλα τα πιθανά μονοπάτια από ένα σημείο σε ένα άλλο (Εικόνα 3.4) όπως επεξηγήθηκε αναλυτικά στο 2^{ο} κεφάλαιο.

Ωστόσο, η μέθοδος των απεικονίσεων διάχυσης απαιτεί ρύθμιση δύο παραμέτρων. Η πρώτη, αφορά μία παράμετρο διάχυσης σ η οποία προσαρμόζει έναν πυρήνα (εδώ Γκαουσιανό) στα δεδομένα. Η παράμετρος αυτή στην παρούσα εργασία υπολογίστηκε με το λογαριθμικό σχήμα που υπέδειξαν οι (Singer et al. 2009) για κάθε υποκείμενο ξεχωριστά. Επιλέξαμε εκείνη την τιμή που βρίσκεται στην μέση της γραμμικής περιοχής του λογαριθμικού σχήματος κλίμακας. Ας σημειωθεί ότι θα μπορούσαμε να πάρουμε οποιαδήποτε τιμή στην γραμμική περιοχή αλλά επιλέξαμε να πάρουμε μία τιμή στην μέση για να είμαστε συνεπής με την ανάλυση ανάμεσα στα υποκείμενα.

Η δεύτερη παράμετρος t που αφορά τη χρονική κλίμακα τέθηκε ίση με τη μονάδα. Τελικά η ενσωμάτωση στον χαμηλότερης διάστασης χώρο μετατρέπεται σε γράφο με την μέθοδο πολλαπλών αναλογικών τιμών κατωφλιού από 15% των μικρότερων αποστάσεων



Εικόνα 3.4 *Απεικόνιση της απόστασης διάχυσης μέσω τυχαίων περιπάτων .Ένα μονοπάτι από το B στο C είναι πιο πιθανό να ακολουθηθεί από ένα μονοπάτι από το B στο A. Επομένως η απόσταση διάχυσης είναι μικρότερη από το B στο C έναντι του B στο A.[Gao and Vanschoren 2011]*

στο 100% με βήμα 5%.

3.3.5 Μέτρηση γραφοθεωρητικών στοιχείων

Στο σημείο αυτό, κάθε εγκεφαλική σάρωση κάθε υποκειμένου έχει μετασχηματιστεί σε μία γραφική δομή, όπου κάθε κόμβος αναπαριστά ένα εγκεφαλικό δίκτυο και η διασυνδεσιμότητα μεταξύ των κόμβων αναπαριστά την ομοιότητα στην δραστηριότητα των δικτύων (με διαφορετικούς όρους ανάλογα τη μετρική που χρησιμοποιήθηκε). Κάθε γράφος μπορεί τώρα να συνοψιστεί από τις ιδιότητες που αφορούν την διασυνδεσιμότητα του. Υπάρχουν πολλά γραφοθεωρητικά μέτρα που εξηγούν τα χαρακτηριστικά και τη διασυνδεσιμότητα ενός γράφου. Παρακάτω, θα δούμε κάποια από αυτά και συγκεκριμένα αυτά που υπολογίσαμε για τα γραφήματα των υποκειμένων μας.

Τα γραφοθεωρητικά μέτρα που είναι όλα συσχετιζόμενα μεταξύ τους, δίνουν ποσοτικές μετρήσεις για τη διασυνδεσιμότητα της δομής. Τα μέτρα αυτά μπορούν να χαρακτηρίσουν ένα γράφο ποιοτικά. Όπως είδαμε και παραπάνω, μπορούν να χρησιμοποιηθούν για να αναδειχθούν διαφορές ανάμεσα σε εγκεφαλικά δίκτυα που συνδέονται με ασθένειες και ψυχικές διαταραχές. Ωστόσο, τα μέτρα αυτά και η προσέγγιση της θεωρίας πολύπλοκων δικτύων έχει εφαρμοστεί σε πολλά πεδία (εκτός της Νευροεπιστήμης) των επιστημών με αξιοσημείωτη επιτυχία (βλέπε Εισαγωγή). Τα παρακάτω γραφοθεωρητικά μέτρα για κάθε γράφημα μπορούν να συγκριθούν άμεσα μεταξύ των ομάδων (σχιζοφρενείς-υγιείς) με στατιστικούς τ-ελέγχους (με κατάλληλη διόρθωση της τιμής ελέγχου p για πολλαπλές συγκρίσεις).

Τα γραφοθεωρητικά στοιχεία που υπολογίσαμε για την ανάλυση λειτουργικής διασυνδεσιμότητας των γράφων-δικτύων με τη χρήση του πακέτου igraph (Csardi and Nepusz 2006) σε γλώσσα R είναι τα εξής :

- Μέσο μήκος μονοπατιού (Average path length) : Είναι ο υπολογισμός του μέσου μήκους των ελαχίστων διαδρομών από έναν κόμβο σε έναν άλλο.

- Πυκνότητα γράφου (Graph density) : Είναι αποτέλεσμα του λόγου των κόμβων με τον αριθμό των πιθανών ακμών (αριθμός κόμβων διαιρούμενος με τον αριθμό όλων των δυνατών ακμών στο γράφημα) .

- Μέσος βαθμός ή βάρος κόμβου (Median degree/strength) : Η μέση τιμή του αριθμού των ακμών που οδηγούν σε ένα κόμβο, ή η μέση τιμή του βάρους που συλλέγει ένας κόμβος από συνδέσεις που προσπίπτουν πάνω του.

- Αριθμός των κόμβων (Vertex count) : Αριθμός των κόμβων σε ένα γράφο.

- Αριθμός των ακμών (Edges count) : Αριθμός των ακμών σε ένα γράφο.

- Συντελεστής συγκρότησης (clustering coefficient) : πιθανότητα δύο κόμβοι να συνδέονται μεταξύ τους ή με άλλα λόγια μία ποσότητα που μας δείχνει σε ποιο βαθμό υπάρχει ή όχι μία δομή στο γράφο.

3.3.6 Διαχωρισμός των υποκειμένων χρησιμοποιώντας μεθόδους εκμάθησης μηχανών

Αφού μετασχηματίστηκε κάθε FMRI σάρωση για το κάθε υποκείμενο σε γράφημα -όπου οι κόμβοι του αναπαριστούν λειτουργικά δίκτυα σε κατάσταση ηρεμίας (resting state networks) και οι αποστάσεις μεταξύ τους αναπαριστούν την ομοιότητα στη δραστηριότητα των δικτύων αυτών (ανάλογα με την μετρική που χρησιμοποιήθηκε)-, αναλύθηκε η λειτουργική διασυνδεσιμότητα των γραφημάτων (Functional Network Connectivity).

Όλοι οι γράφοι για κάθε μέθοδο και μετρική, ελέγχθηκαν για διαφορές όσον αφορά τον αριθμό των ακμών, των κόμβων και την πυκνότητα του γράφου. Μεταξύ υγειών και σχιζοφρενών δεν βρέθηκε στατιστικά σημαντική διαφορά (για καμία τιμή κατωφλιού). Ο έλεγχος αυτός κρίθηκε απαραίτητος αφού διαφορές σε οποιοδήποτε από τα βασικά αυτά μέτρα θα επηρέαζε καθοριστικά επόμενες συγκρίσεις ανάμεσα στους γράφους (Van Wijk et al. 2010).

Για την ταξινόμηση των υποκειμένων τελικά χρησιμοποιήσαμε 3 χαρακτηριστικά διασυνδεσιμότητας γράφου, το μέσο μήκος μονοπατιού, τον μέσο όρο του βάρους των κόμβων και το συντελεστή συγκρότησης.

Η γλώσσα R έχει έναν τεράστιο αριθμό από βιβλιοθήκες που είναι διαθέσιμες για ταξινόμηση. Τελικά, στην παρούσα εργασία χρησιμοποιήθηκαν ένας απλός SVM (Support Vector Machine) ταξινομητής, ένας ταξινομητής των κ κοντινότερων γειτόνων και νευρωνικά δίκτυα με απλή αρχιτεκτονική (με ένα κρυμμένο επίπεδο νευρώνων). Η ρύθμιση των παραμέτρων για τον κάθε αλγόριθμο ταξινόμησης έγινε με αναζήτηση πλέγματος (grid search) και 10-fold cross-validation σχήμα. Οι ταξινομητές αυτοί υλοποιήθηκαν μέσω του πακέτου caret (Kuhn 2008) σε γλώσσα R.

3.3.6.1 Support Vector Machines (SVM)

Ο αλγόριθμος SVM ψάχνει ένα υπερεπίπεδο (hyperplane) που διαχωρίζει με βέλτιστο τρόπο διαφορετικές ομάδες, χρησιμοποιώντας μόνο σημεία που περιέχονται στο περιθώριο (margin) ή περιοχή επικάλυψης (region of overlap). Για ένα σύνολο σημείων (x_i, y_i) όπου $(x_i) \in \mathbb{R}^n$ είναι το σύνολο χαρακτηριστικών (εδώ τα γραφοθεωρητικά στοιχεία) για το γράφο G_i που αντιστοιχεί στο υποκείμενο i, ένα μέλος της κλάσης $y_i \in (-1,1)$ (π.χ. ασθενής ή υγιής), ο SVM θα προσαρμόσει ένα υπερεπίπεδο έτσι ώστε να διαχωρίζει βέλτιστα τις κλάσεις στο $(-1,1)$. Ένα υπερεπίπεδο περιγράφεται ως:

$$w * x_i - b = 0 \tag{3.11}$$

Όπου, w είναι τα κανονικά διανύσματα του υπερεπιπέδου. Τα παράλληλα υπερεπίπεδα που διαχωρίζουν τις παρατηρήσεις μπορούν να ορισθούν σαν

$$w * x_i - b \geq -1 \; \gamma\iota\alpha \;\; y_i = 1 \tag{3.12}$$

Και

$$w * x_i - b \leq -1 \; \gamma\iota\alpha \;\; y_i = -1 \tag{3.13}$$

Πρόβλημα βελτιστοποίησης γίνεται τώρα η μεγιστοποίηση της απόστασης ανάμεσα στα επίπεδα, $\frac{2}{||W||}$, τέτοια ώστε:

$$y_i(w * x_i - b) \geq -1 \quad (3.14)$$

Ωστόσο, σε πραγματικές εφαρμογές χρησιμοποιείται μία παράμετρος c η οποία ρυθμίζει την απόδοση της ταξινόμησης με αντάλλαγμα μία μικρότερη απόσταση ανάμεσα στα επίπεδα των ομάδων που διαχωρίζονται.

Παρόλα αυτά, υπάρχουν περιπτώσεις που τα δεδομένα μας δεν είναι γραμμικώς διαχωρίσιμα στον αρχικό n-διάστατο χώρο και γι αυτο μπορεί κανείς να χρησιμοποιήσει έναν πυρήνα ακτινικών συναρτήσεων βάσης (RBF) για να απεικονίσει τα δεδομένα σε έναν υψηλότερης διάστασης χώρο όπου εκεί τα δεδομένα μπορεί να είναι γραμμικώς διαχωρίσιμα.

Στην παρούσα εργασία χρησιμοποιήθηκαν ο κλασικός SVM γραμμικός ταξινομητής (Linear SVM) και ο SVM με RBF πυρήνα (Radial SVM) όπου εισάγει μία ακόμα παράμετρο σ που ρυθμίζει την κλίμακα του πυρήνα. Οι τιμές των σ και c που διερευνήθηκαν με αναζήτηση πλέγματος (grid search) ήταν $\sigma = (0.001,0.01,0.1)$ και $c = (0.001,0.01,0.1,1,10)$.

3.3.6.2 Ταξινομητής των κ κοντινότερων γειτόνων (KNN classifier)

Ο αλγόριθμος των κ κοντινότερων γειτόνων είναι ένας από τους πιο απλούς αλγορίθμους ταξινόμησης. Έχοντας ένα σύνολο σημείων (x_i, y_i) -όπου $x_i \in \mathbb{R}^n$ είναι το σύνολο χαρακτηριστικών (εδώ τα γραφοθεωρητικά στοιχεία) για το γράφο G_i που αντιστοιχεί στο υποκείμενο i και y_i η ομάδα που ανήκει το συγκεκριμένο υποκείμενο- χρησιμοποιεί την ευκλείδεια απόσταση και εντοπίζει τους κ κοντινότερους γείτονες στο χώρο των χαρακτηριστικών (feature space) για να διεξάγει ένα σύστημα ψηφίσματος.

Η παράμετρος κ είναι μία παράμετρος που αποφασίζει το πόσους γείτονες θα ‘’συμβουλευτεί’’ κάθε σημείο για να κατατάξει τον εαυτό σου σε μία ομάδα και παίζει ένα πολύ σημαντικό ρόλο στην απόδοση του αλγορίθμου.

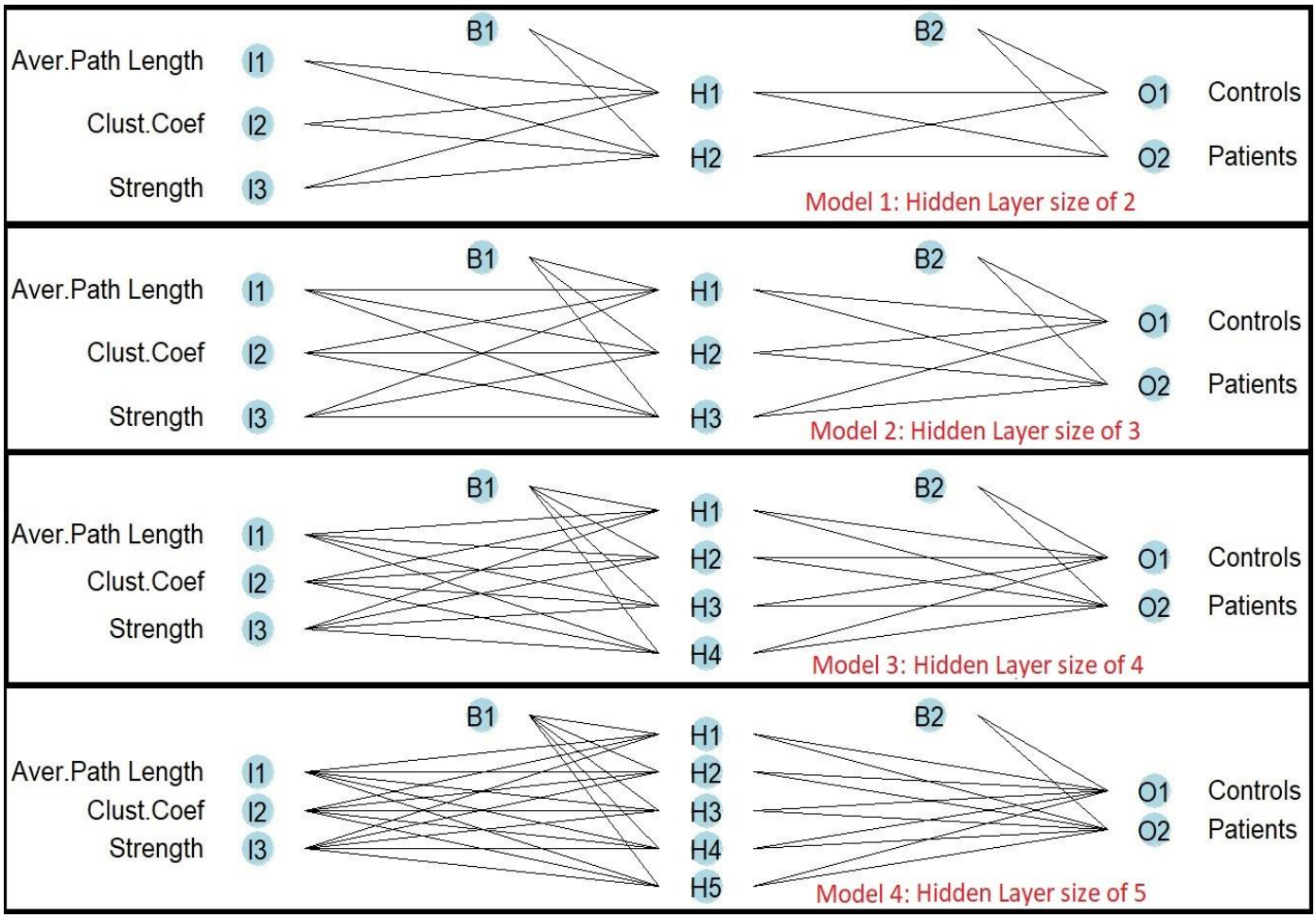
Με αυτόν τον τρόπο κάθε σημείο στο χώρο χαρακτηριστικών κατατάσσεται σαν υγιές ή σχιζοφρενές (στην περίπτωση μας) αν οι περισσότεροι από τους γείτονές του ανήκουν σε μία από τις δύο ομάδες.

Είναι σημαντικό να σημειώσουμε ότι οι τιμές του κ που επιλέχθηκαν για διερεύνηση ήταν όλες περιττές για να αποφευχθούν τυχόν ισοπαλίες στο σύστημα ψηφίσματος. Οι τιμές που διερευνήθηκαν ήταν οι $\mathrm{k} = (1,3,5,7,9)$.

3.3.6.3 Νευρωνικά δίκτυα (Neural Nets)

Feed-forward back-propagation Νευρωνικά δίκτυα αποτελούμενα από ένα μόνο κρυμμένο επίπεδο νευρώνων (1-hidden layer) χρησιμοποιήθηκαν στην εργασία για την ταξινόμηση σε υγιείς και σχιζοφρενείς. Πιο συγκεκριμένα:

Η είσοδος (1 επίπεδο εισόδου) στον αλγόριθμο ήταν τα τρία γραφοθεωρητικά μέτρα που υπολογίσαμε για κάθε γράφο. Η αρχιτεκτονική που επιλέχθηκε για διερεύνηση ήταν απλή αφού και το μέγεθος του δείγματος ήταν σχετικά μικρό. Στην Εικόνα 3.4 φαίνονται καθαρά τα μοντέλα που διερευνήθηκαν. Για το κρυμμένο επίπεδο (1 κρυμμένο επίπεδο) χρησιμοποιήθηκαν από 2 έως 5 νευρώνες συν τον όρο προκατάληψης (bias term) ενώ το



Εικόνα 3.4 *Μοντέλα νευρωνικών δικτύων που διερευνήθηκαν. Το επίπεδο εισόδου (Input layer) αποτελείται από τα 3 χαρακτηριστικά (γραφοθεωρητικά μέτρα).Το κρυμμένο επίπεδο (Hidden layer) αποτελείται από 2 έως και 5 νευρώνες ενώ προστίθεται και ένας όρος προκατάληψης (Bias term).Το επίπεδο εξόδου (Output layer) αποτελείται από δύο κόμβους αφού το πρόβλημα αφορά δυαδική ταξινόμηση(σχιζοφρενείς/υγιείς).*

επίπεδο εξόδου (1 επίπεδο εξόδου) αποτελούνταν από δύο κόμβους αφού το πρόβλημα αφορά μία δυαδική ταξινόμηση (υγιής/σχιζοφρενής).Η συνάρτηση ενεργοποίησης που χρησιμοποιήθηκε για όλους τους νευρώνες ήταν η λογιστική μεταβατική συνάρτηση (logistic transfer function) (Ripley and Hjort 1996).

Η διαδικασία εκπαίδευσης έγινε με τη μέθοδο back-propagation (Hecht-Nielsen 1988) χρησιμοποιώντας ένα 10-fold cross validation σχήμα. Τελικά, μία ακόμα παράμετρος χρησιμοποιήθηκε που αφορούσε το ρυθμό α με τον οποίο τα βάρη μειώνονται. Ο λόγος ήταν η αποφυγή της υπερπροσαρμογής και η βελτίωση της γενικότητας του αποτελέσματος (Krogh και Hertz 1992). Οι τιμές του α που διερυνήθηκαν ήταν $\alpha = (0.001, 0.01, 0.1)$.Ο αλγόριθμος που χρησιμοποιήθηκε ανήκει στο πακέτο nnet (Ripley and Venables 2011) σε γλώσσα R.

3.3.7 Έλεγχοι υποθέσεων για τα γραφοθεωρητικά στοιχεία

Στην προσπάθεια επικύρωσης και ανάδειξης των διαφορών στα γραφοθεωρητικά στοιχεία των δύο ομάδων (ασθενών-υγιών) πραγματοποιήθηκαν τυπικοί t-έλεγχοι υπόθεσης (formal hypothesis testing).Γενικά, μπορούμε να συγκρίνουμε ατομικά ένα-ένα τα χαρακτηριστικά για τις δύο ομάδες ή να συγκρίνουμε όλα τα χαρακτηριστικά μαζί, άμεσα.

Ωστόσο, από τη στιγμή που τα χαρακτηριστικά αυτά είναι συσχετιζόμενα και συγγενικά μεταξύ τους, θα πρέπει να προσαρμόσουμε τον έλεγχο μας για πολλαπλές συγκρίσεις χρησιμοποιώντας την διόρθωση Bonferonni (Bonferonni correction) όπου θεωρεί σημαντικές μόνο p-τιμές ελέγχου κάτω από $^{0.05}/_{\nu}$ (όπου ν είναι ο αριθμός των συγκρίσεων που πραγματοποιούνται).

Στην εργασία μας, για τους στατιστικούς ελέγχους των γραφοθεωρητικών στοιχείων έγινε πράγματι διόρθωση Bonferonni για ν = 3 ($p_{corrected} = \frac{0.05}{3} = 0.016$), όπου είναι και ο αριθμός των βασικών μέτρων διερεύνησης (μέσο μήκος μονοπατιού, μέσος βαθμός/βάρος κόμβου και συντελεστής συγκρότησης).

3.3.8 Επιλογή διαστάσεων για κάθε μέθοδο και αξιολόγηση της κάθε μεθόδου και μετρικών

Οι διαστάσεις που κρατήσαμε για κάθε μέθοδο μείωσης διάστασης έγινε με την επιλογή εκείνης της διάστασης που έδωσε το καλύτερο μέσο ποσοστό ταξινόμησης σε όλα τα επίπεδα κατωφλιού (15%-95%, στο 100% το μέσο μήκος μονοπατιού είναι απλώς το βάρος της σύνδεσης ενώ ο συντελεστής συγκρότησης είναι ίσος με τη μονάδα).

Μετά την επιλογή της διάστασης του χώρου ενσωμάτωσης για κάθε μέθοδο εκμάθησης πολλαπλοτήτων πραγματοποιήσαμε ανάλυση διασποράς (ANOVA) με δύο παράγοντες. Η ανάλυση αυτή μας οδηγεί στο να βγάλουμε συμπεράσματα σχετικά με το αν υπάρχει διαφοροποίηση στην ταξινόμηση των υποκειμένων λόγω των διαφορετικών μεθόδων εκμάθησης πολλαπλοτήτων αλλά και το αν υπάρχει διαφοροποίηση στο ποσοστό ταξινόμησης εξαιτίας των αλγορίθμων ταξινόμησης.

Τελικά, πραγματοποιήθηκαν Tuckey-post hoc τεστ (τα οποία λαμβάνουν υπόψη και το σφάλμα πολλαπλών συγκρίσεων (Family-wise error)) για να επικυρώσουμε ποια μέθοδος συμπεριφέρεται καλύτερα (σε σχέση με άλλες) αλλά και ποιοι αλγόριθμοι ταξινόμησης είναι αποδοτικότεροι (σε σχέση με άλλους). Η ανάλυση αυτή έγινε ξεχωριστά για κάθε μετρική που χρησιμοποιήθηκε για την ανάλυση.

4 ΑΠΟΤΕΛΕΣΜΑΤΑ

Παρακάτω παρουσιάζονται τα αποτελέσματα ξεχωριστά για τη χρήση των δύο μετρικών που χρησιμοποιήθηκαν για την ανάλυση των εγκεφαλικών δικτύων ηρεμίας (resting state networks). Στην παρούσα ανάλυση εξαιρέθηκαν υποκείμενα με λιγότερες από 20 ανεξάρτητες συνιστώσες.Το δείγμα των υποκειμένων που συμμετείχαν στην ανάλυση ήταν τελικά 47 ασθενείς με σχιζοφρένεια και 57 υγιείς.

4.1 ΜΕΤΡΙΚΗ ΒΑΣΙΖΟΜΕΝΗ ΣΤΗΝ ΠΟΛΛΑΠΛΗ ΣΥΣΧΕΤΙΣΗ ΜΕ ΚΑΘΥΣΤΕΡΗΣΕΙΣ

4.1.1 Σύγκριση των γραφοθεωρητικών στοιχείων ανάμεσα σε σχιζοφρενείς και υγιείς πριν και μετά την μείωση διάστασης με τη χρήση των μεθόδων MDS,ISOMAP και Diffusion Maps.

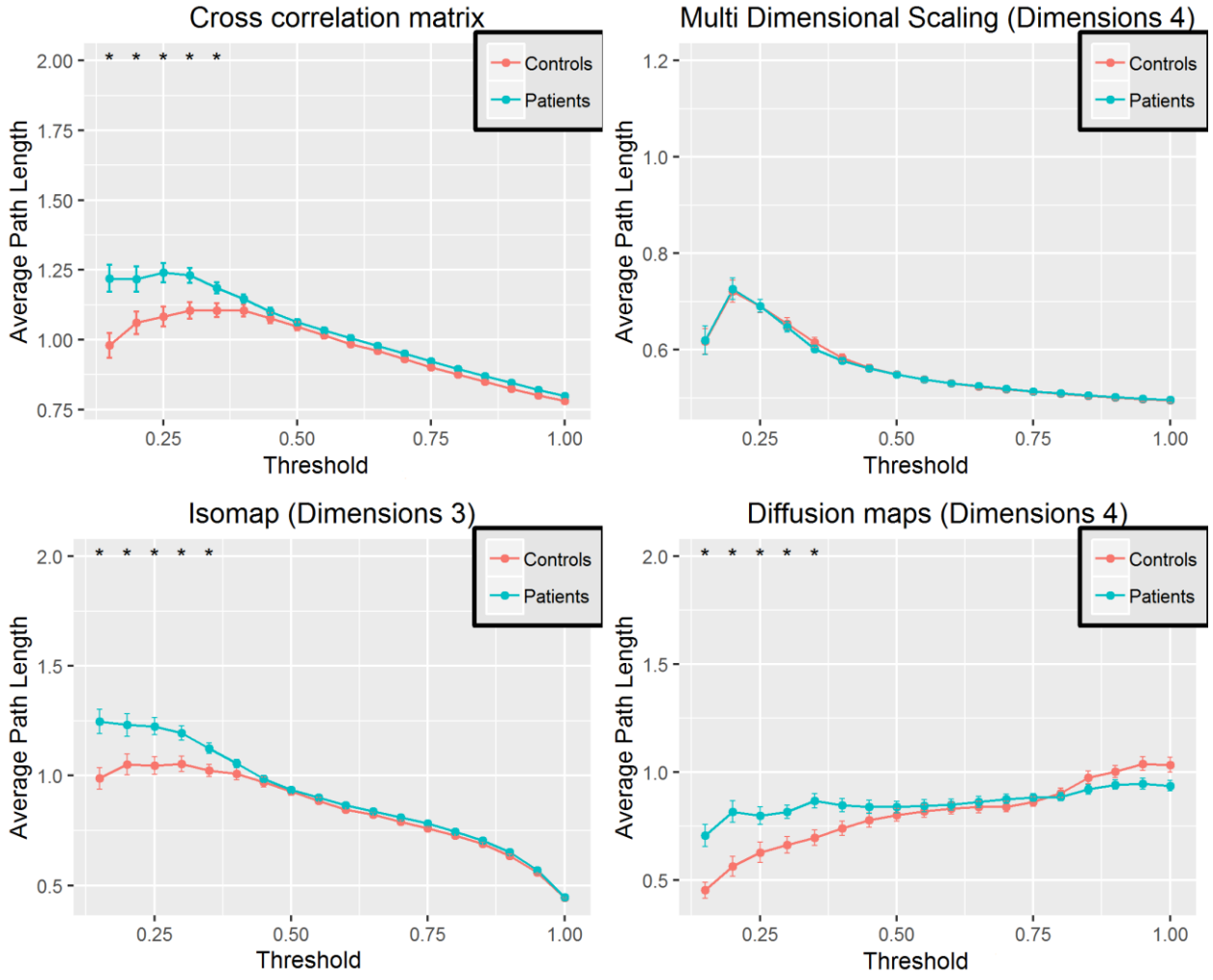
Χρησιμοποιώντας τη μετρική πολλαπλής συσχέτισης με καθυστερήσεις για την ανάλυση της διασυνδεσιμότητας των ανεξάρτητων συνιστωσών δημιουργήθηκαν γράφοι με

		Αριθμός διαστάσεων ενσωμάτωσης			
		2	3	4	5
Μέθοδοι	**Ταξινομητές**	**Ποσοστό επιτυχούς ταξινόμησης (%)**			
MDS	Linear SVM	55	55	54	57
	KNN	58	56	57	59
	Radial SVM	6	58	**61**	61
	Neural Nets	6	56	59	60
ISOMAP	Linear SVM	61	61	62	62
	KNN	60	61	60	60
	Radial SVM	63	**64**	64	63
	Neural Nets	62	62	63	62
DIFFUSION MAPS	Linear SVM	59	62	66	65
	KNN	62	59	61	62
	Radial SVM	61	63	**68**	68
	Neural Nets	61	62	66	66
		Όλες οι διαστάσεις			
Πίνακας Διασυνδεσιμότητας	Linear SVM	61			
	KNN	57			
	Radial SVM	62			
	Neural Nets	62			

Πίνακας 1 *Μέσο ποσοστό ταξινόμησης για κάθε μέθοδο και ταξινομητή με τη μετρική που βασίζεται στην πολλαπλή συσχέτιση (με καθυστερήσεις).Σε πράσινο φόντο είναι η επιλεχθείσα διάσταση για την εκάστοτε μέθοδο. Επιπλέον σημειώνονται τα αποτελέσματα του πίνακα διασυνδεσιμότητας (πίνακα πολλαπλής συσχέτισης) χωρίς μείωση διάστασης.*

εφαρμογή των μεθόδων μείωσης διάστασης και εκμάθησης πολλαπλοτήτων. Για την επιλογή των διαστάσεων του χώρου ενσωμάτωσης υπολογίσαμε το μέσο ποσοστό επιτυχούς ταξινόμησης (μέσος όρος σε όλα τα επίπεδα κατωφλιού από 15%-95%) σε κάθε διάσταση $2 \leq d \leq 5$ και για κάθε μέθοδο και ταξινομητή. Στον πίνακα 1 φαίνεται το ποσοστό επιτυχούς ταξινόμησης για κάθε μέθοδο μείωσης διάστασης και αλγόριθμο ταξινόμησης. Με έντονα γράμματα σημειώνεται η καλύτερη επίδοση για κάθε μέθοδο βάση της οποίας επιλέχθηκαν οι διαστάσεις της τελικής ενσωμάτωσης.

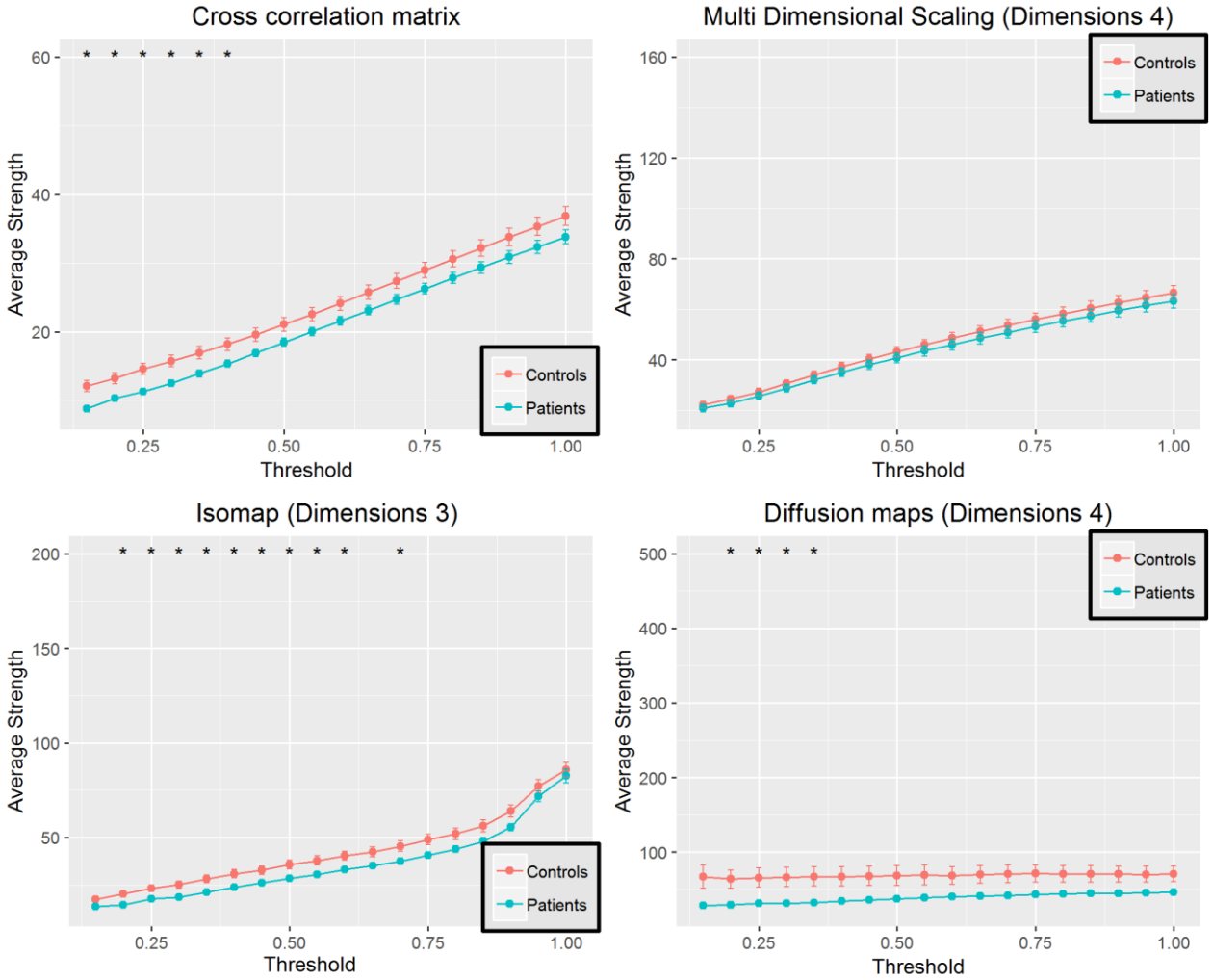
Παρακάτω, απεικονίζονται τα αποτελέσματα των γραφοθεωρητικών στοιχείων ανάμεσα σε υγιείς και σχιζοφρενείς για κάθε μία από τις μεθόδους συμπεριλαμβανομένου και του πίνακα απόστασης για τον οποίο δεν έχει γίνει μείωση διάστασης. Για το μέσο μήκος μονοπατιού (Εικόνα 4.1), για το μέσο βάρος/δύναμη κόμβου(Εικόνα 4.2) και το συντελεστή συγκρότησης (Εικόνα 4.3).



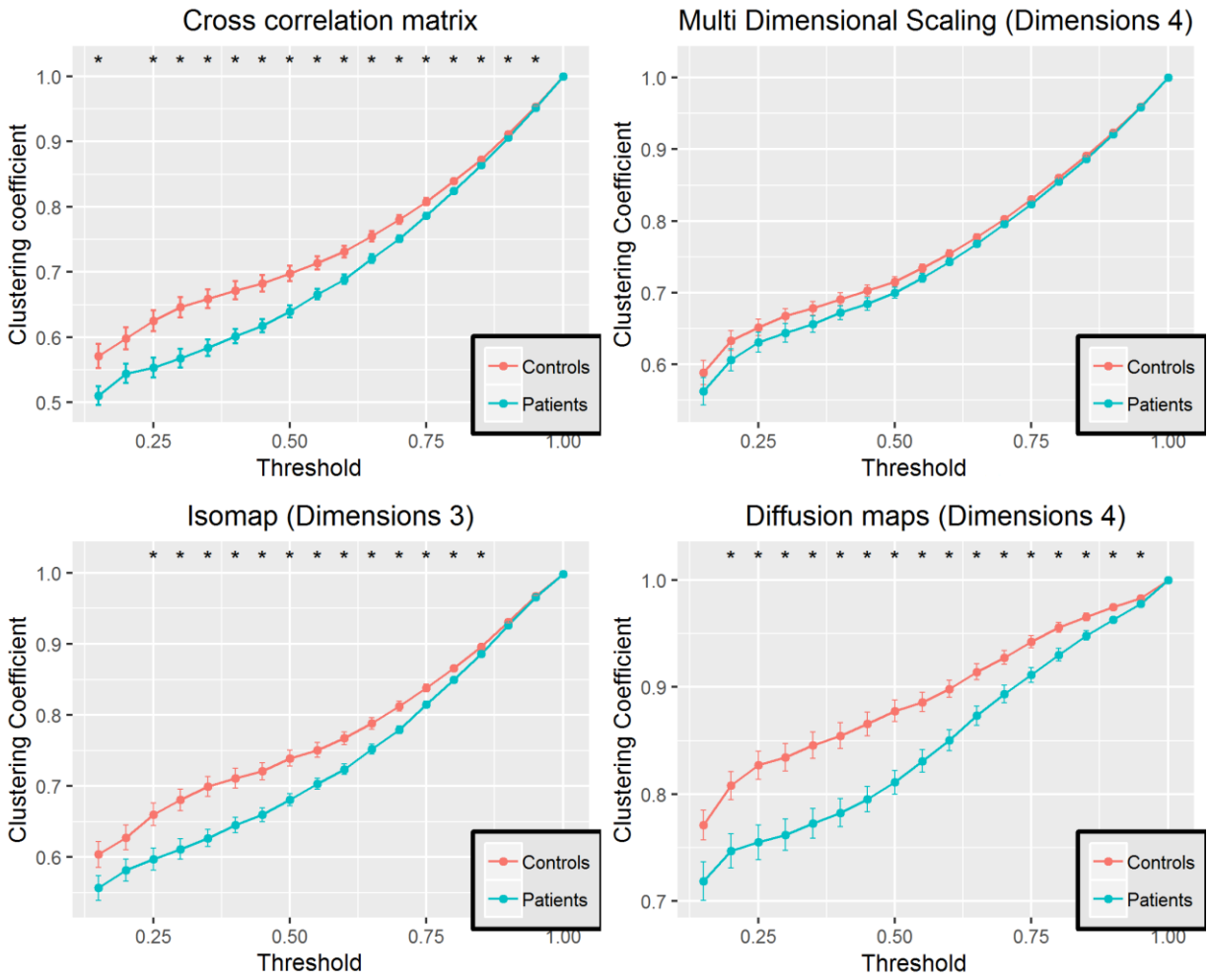
Εικόνα 4.1 *Μοτίβο διασυνδεσιμότητας με βάση το μέσο μήκος μονοπατιού (Average path length) όπως αυτό υπολογίστηκε στο διάστημα 15% -100% (με βήμα 5%) των δυνατότερων συνδέσεων των εγκεφαλικών δικτύων για κάθε μέθοδο(σε παρένθεση οι διαστάσεις που επιλέχθηκαν).Παρουσιάζονται ακόμη τα αποτελέσματα του πίνακα πολλαπλής συσχέτισης (Cross correlation matrix) χωρίς μείωση διάστασης (πάνω αριστερά).Τέλος, με αστερίσκο σημειώνεται η στατιστικά σημαντική διαφορά στη μέση τιμή (p¬corrected < 0.05) ανάμεσα σε υγιείς (κόκκινο) και σχιζοφρενείς (μπλε).*

Με αστερίσκο σημειώνεται επάνω από κάθε γράφημα σε ποιο κατώφλι η διαφορά στη μέση τιμή είναι στατιστικά σημαντική με επίπεδο σημαντικότητας το οποίο έχουμε διορθώσει με Bonferonni διόρθωση ($p_{corrected} < 0.05$) όπως περιγράφεται στο προηγούμενο κεφάλαιο.

Η μπάρα σφάλματος μας δίνει το τυπικό σφάλμα της μέσης τιμής και τελικά ο αριθμός των διαστάσεων της ενσωμάτωσης αναγράφεται για κάθε μέθοδο στον τίτλο του κάθε υπογραφήματος.



Εικόνα 4.2 *Μοτίβο διασυνδεσιμότητας με βάση το μέσο βάρος/δύναμη (Average strength) των συνδέσεων όπως αυτές υπολογίστηκαν στο διάστημα 15% -100% (με βήμα 5%) των δυνατότερων ακμών των εγκεφαλικών δικτύων για κάθε μέθοδο (σε παρένθεση οι διαστάσεις που επιλέχθηκαν).Παρουσιάζονται ακόμη τα αποτελέσματα του πίνακα πολλαπλής συσχέτισης Cross correlation matrix χωρίς μείωση διάστασης (πάνω αριστερά).Τέλος, με αστερίσκο σημειώνεται η στατιστικά σημαντική διαφορά στη μέση τιμή ($p_{corrected} < 0.05$) ανάμεσα σε υγιείς (κόκκινο) και σχιζοφρενείς (μπλε).*



Εικόνα 4.3 *Μοτίβο διασυνδεσιμότητας με βάση το συντελεστή συγκρότησης (Clustering coefficient) όπως αυτός υπολογίστηκε στο διάστημα 15% -100% (με βήμα 5%) των δυνατότερων συνδέσεων των εγκεφαλικών δικτύων για κάθε μέθοδο(σε παρένθεση οι διαστάσεις που επιλέχθηκαν).Παρουσιάζονται ακόμη τα αποτελέσματα του πίνακα πολλαπλής συσχέτισης (Cross correlation matrix) χωρίς μείωση διάστασης (πάνω αριστερά).Τέλος, με αστερίσκο σημειώνεται η στατιστικά σημαντική διαφορά στη μέση τιμή ($p_{corrected} < 0.05$) ανάμεσα σε υγιείς (κόκκινο) και σχιζοφρενείς (μπλε).*

4.1.2 Ανάλυση ANOVA 2 παραγόντων για την εκτίμηση της κάθε μεθόδου και τη συμπεριφορά των ταξινομητών

Για τις επιλεχθείσες διαστάσεις σε κάθε μέθοδο πραγματοποιήσαμε ανάλυση ANOVA 2 παραγόντων εξετάζοντας την μέση επιτυχή ταξινόμηση σαν συνάρτηση των μεθόδων μείωσης διάστασης (1ος παράγοντας) αλλά και των αλγορίθμων ταξινόμησης (2ος παράγοντας). Παρακάτω και συγκεκριμένα στον Πίνακα 2 παρουσιάζονται τα αποτελέσματα της ανάλυσης διασποράς 2 παραγόντων για τη μετρική πολλαπλής συσχέτισης με καθυστερήσεις. Παρατηρούμε τα F στατιστικά είναι μεγάλα τόσο για τις μεθόδους όσο και τους ταξινομητές. Αυτό σημαίνει ότι η επιτυχής ταξινόμηση των υποκειμένων σε σχιζοφρενείς και υγιείς εξαρτάται σημαντικά και από τους δύο παράγοντες.

Στη συνέχεια διεξάγουμε Tuckey post-hoc τεστ για να αποφανθούμε ειδικότερα ποια μέθοδος συμπεριφέρεται καλύτερα (Πίνακας 3) αλλά και ποιοι από τους ταξινομητές (Πίνακας 4) συμπεριφέρονται καλύτερα στο πρόβλημα αυτό. Ουσιαστικά πραγματοποιούμε διαδοχικούς τ-ελέγχους ανά ζευγάρια μεθόδων στην πρώτη περίπτωση και ζευγάρια

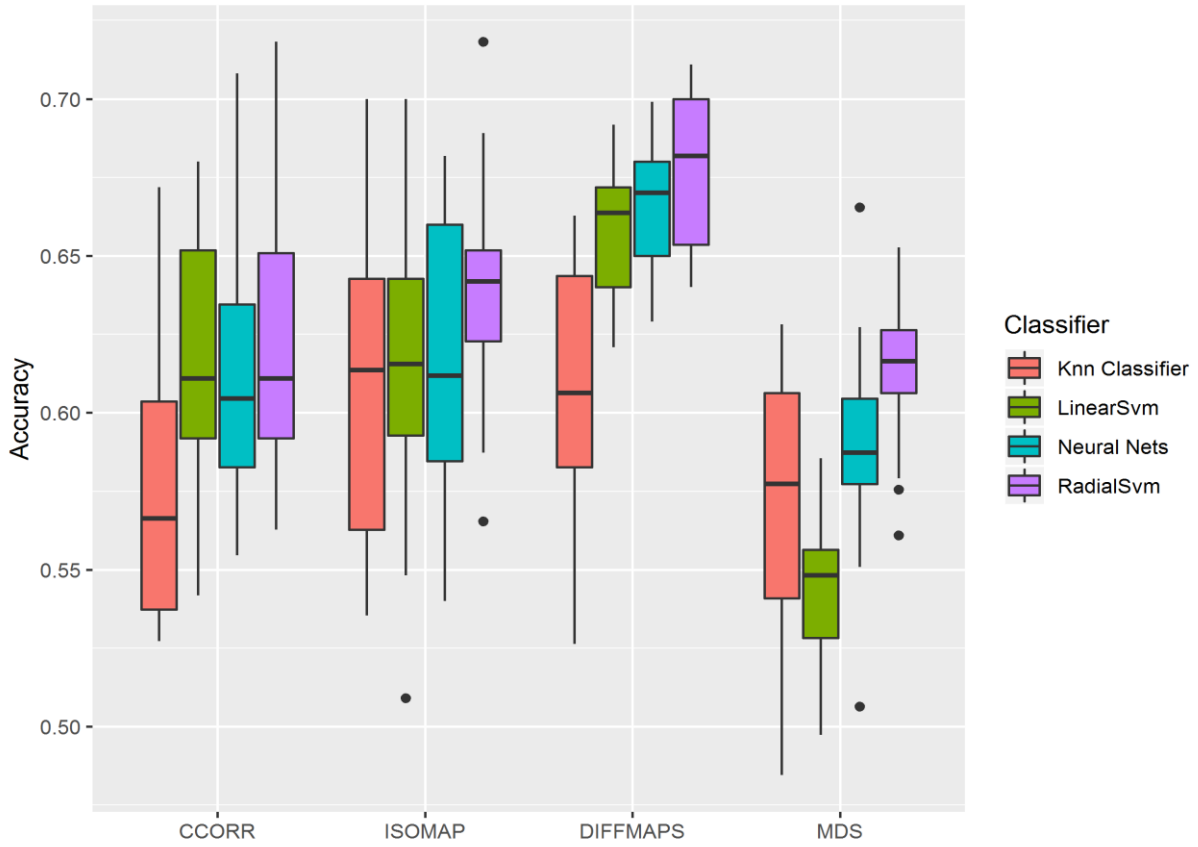
αλγορίθμων ταξινόμησης στην δεύτερη. Τελικά, παρουσιάζονται τα αποτελέσματα για κάθε μέθοδο και ταξινομητή σε μορφή boxplots στην Εικόνα 4.4 .

Πηγή (Source)	Βαθμοί ελευθερίας (Degrees of freedom	Τετραγωνικά αθροίσματα (Squared Sums)	Μέσο Άθροισμα (Mean Sums)	F στατιστικό	Πιθανότητα (> F)
Ταξινομητές	3	0.0845	0.02816	18.27	**< e-10**
Μέθοδοι	3	0.1820	0.06068	39.35	**< 2e-16**
Υπόλοιπο	265	0.4086	0.00154		

Πίνακας 2 *Πίνακας ανάλυσης διασποράς δύο παραγόντων για τη μετρική βασιζόμενη στην πολλαπλή συσχέτιση (με καθυστερήσεις). Με έντονα γράμματα σημειώνονται στατιστικά σημαντικοί παράγοντες της ανάλυσης.*

Tuckey Post-hoc (για τον παράγοντα Μέθοδοι, t-έλεγχοι)	Διάστημα εμπιστοσύνης (95%)	Στατιστική διαφορά στη μέση τιμή	p- τιμή ελέγχου (FWE corrected)
ISOMAP-Cross Corr Matrix	(-0.005,0.029)	0.011	0.3035479
Diffusion Maps-Cross Corr Matrix	(0.027,0.062)	0.045	**< e-10**
MDS-Cross Corr Matrix	(-0.044 -0.009)	-0.027	**0.0004**
Diffusion Maps-ISOMAP	(0.016,0.050)	0.033	**< 8e-8**
MDS-ISOMAP	(-0.056 -0.021)	-0.038	**< e-08**
MDS-Diffusion Maps	(-0.089 -0.054)	-0.072	**< e-10**

Πίνακας 3 *Post-hoc Tuckey test για τον ρόλο που παίζει η εκάστοτε μέθοδος όσον αφορά το μέσο ποσοστό ταξινόμησης. Με έντονα γράμματα σημειώνονται στατιστικά σημαντικές διαφορές(με διόρθωση πολλαπλών συγκρίσεων).*



Εικόνα 4.4 *Συγκεντρωτικά αποτελέσματα του μέσου ποσοστού ταξινόμησης για κάθε μέθοδο (καθώς και του πίνακα πολλαπλής συσχέτισης (CCORR) χωρίς μείωση διάστασης) και ταξινομητή. Η οριζόντια γραμμή είναι η διάμεσος της κατανομής ενώ σαν σημεία αναγράφονται ακραίες τιμές της κατανομής.*

Tuckey Post-hoc (για τον παράγοντα Ταξινομητές, t-έλεγχοι)	Διάστημα εμπιστοσύνης (95%)	Στατιστική διαφορά στη μέση τιμή	p- τιμή ελέγχου (FWE corrected)
LSVM-KNN	(-0.001,0.033)	0.016	0.0814155
NNETS-KNN	(0.013,0.048)	0.031	**0.0000339**
RSVM-KNN	(0.030,0.064)	0.047	**<e-10**
NNETS-LVSM	(-0.002,0.032)	0.015	0.1153998
RSVM-LSVM	(0.014,0.048)	0.031	**0.0000289**
RSVM-NNETS	(-0.001,0.033)	0.016	0.0746884

Πίνακας 4 *Post-hoc Tuckey test για τον ρόλο που παίζει ο εκάστοτε αλγόριθμος ταξινόμησης όσον αφορά το μέσο ποσοστό ταξινόμησης. Με έντονα γράμματα σημειώνονται στατιστικά σημαντικές διαφορές (με διόρθωση πολλαπλών συγκρίσεων).*

4.2 ΜΕΤΡΙΚΗ ΒΑΣΙΖΟΜΕΝΗ ΣΤΗΝ ΕΥΚΛΕΙΔΕΙΑ ΝΟΡΜΑ

4.2.1 Σύγκριση των γραφοθεωρητικών στοιχείων ανάμεσα σε σχιζοφρενείς και υγιείς πριν και μετά την μείωση διάστασης με τη χρήση των μεθόδων MDS,ISOMAP και Diffusion Maps.

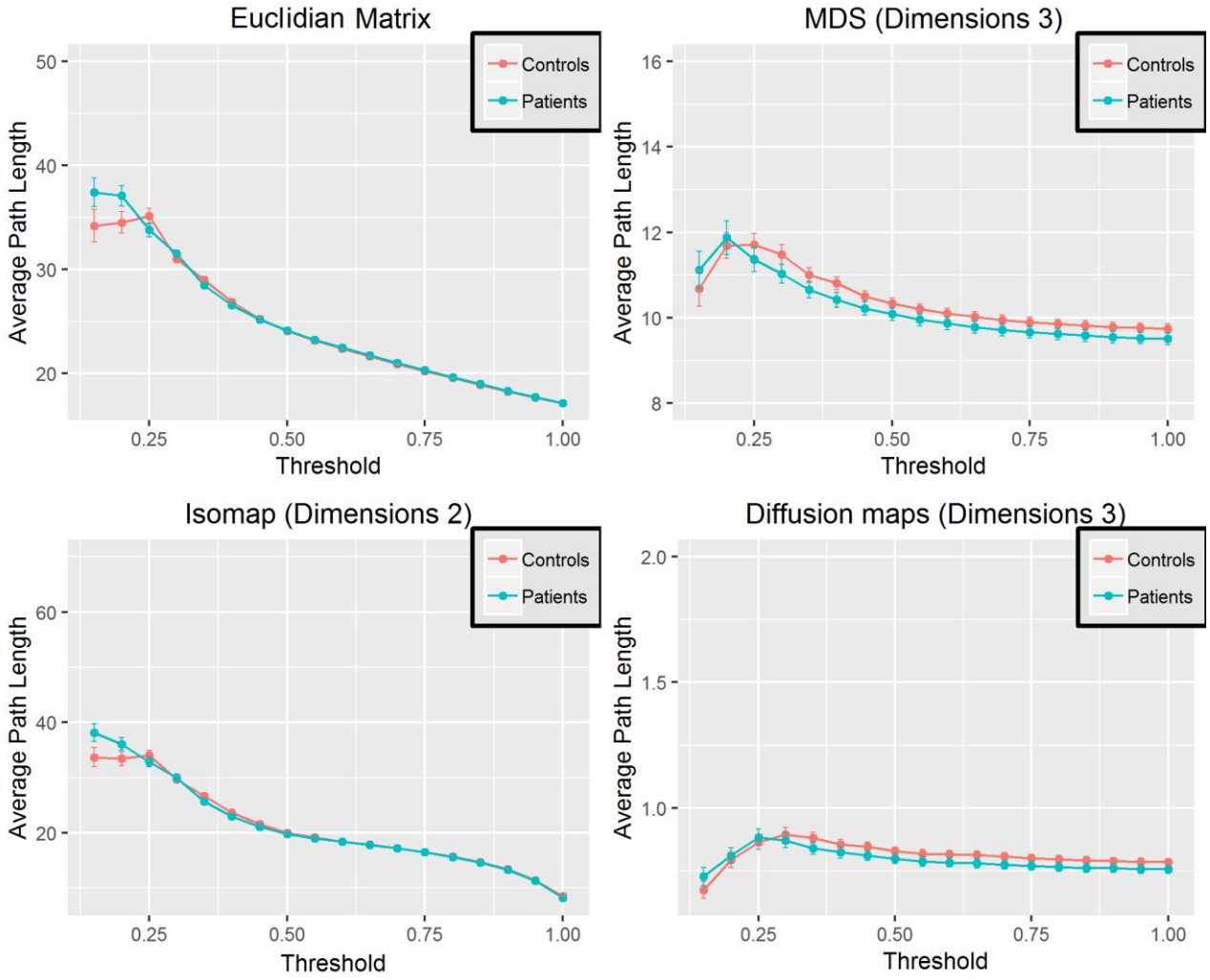
Χρησιμοποιώντας την ευκλείδεια νόρμα ως μετρική για την ανάλυση της διασυνδεσιμότητας των ανεξάρτητων συνιστωσών δημιουργήθηκαν γράφοι με εφαρμογή μεθόδων μείωσης διάστασης και εκμάθησης πολλαπλοτήτων. Για την επιλογή των διαστάσεων του χώρου ενσωμάτωσης υπολογίσαμε το μέσο ποσοστό επιτυχούς ταξινόμησης (μέσος όρος σε όλα τα επίπεδα κατωφλιού από 15%-95%). Για την επιλογή των διαστάσεων του χώρου ενσωμάτωσης υπολογίσαμε το μέσο ποσοστό επιτυχούς ταξινόμησης (μέσος όρος σε όλα τα επίπεδα κατωφλιού) σε κάθε διάσταση $2 \leq d \leq 5$ και για κάθε μέθοδο μείωσης διάστασης και ταξινομητή. Στον πίνακα 5 φαίνεται το ποσοστό επιτυχούς ταξινόμησης για κάθε μέθοδο και αλγόριθμο ταξινόμησης που χρησιμοποιήθηκε. Με έντονα γράμματα σημειώνεται η καλύτερη επίδοση για κάθε μέθοδο βάση της οποίας επιλέχθηκαν οι

		Αριθμός διαστάσεων ενσωμάτωσης			
		2	3	4	5
Μέθοδοι	**Ταξινομητές**	**Ποσοστό επιτυχούς ταξινόμησης (%)**			
MDS	Linear SVM	54	53	53	53
	KNN	57	57	54	53
	Radial SVM	56	**58**	56	56
	Neural Nets	56	57	55	54
ISOMAP	Linear SVM	54	55	56	56
	KNN	56	57	58	56
	Radial SVM	58	57	58	58
	Neural Nets	**60**	58	59	59
DIFFUSION MAPS	Linear SVM	0.56	54	54	59
	KNN	0.59	60	54	57
	Radial SVM	0.58	**61**	57	59
	Neural Nets	0.58	58	58	59
		Όλες οι διαστάσεις			
Πίνακας Διασυνδεσιμότητας	Linear SVM	53			
	KNN	56			
	Radial SVM	57			
	Neural Nets	58			

Πίνακας 5 *Μέσο ποσοστό ταξινόμησης για κάθε μέθοδο και ταξινομητή για τη μετρική που βασίζεται στην ευκλείδεια νόρμα. Σε πράσινο φόντο είναι η επιλεχθείσα διάσταση για την εκάστοτε μέθοδο. Επιπλέον σημειώνονται τα αποτελέσματα του πίνακα διασυνδεσιμότητας (ευκλείδειου πίνακα αποστάσεων) χωρίς μείωση διάστασης.*

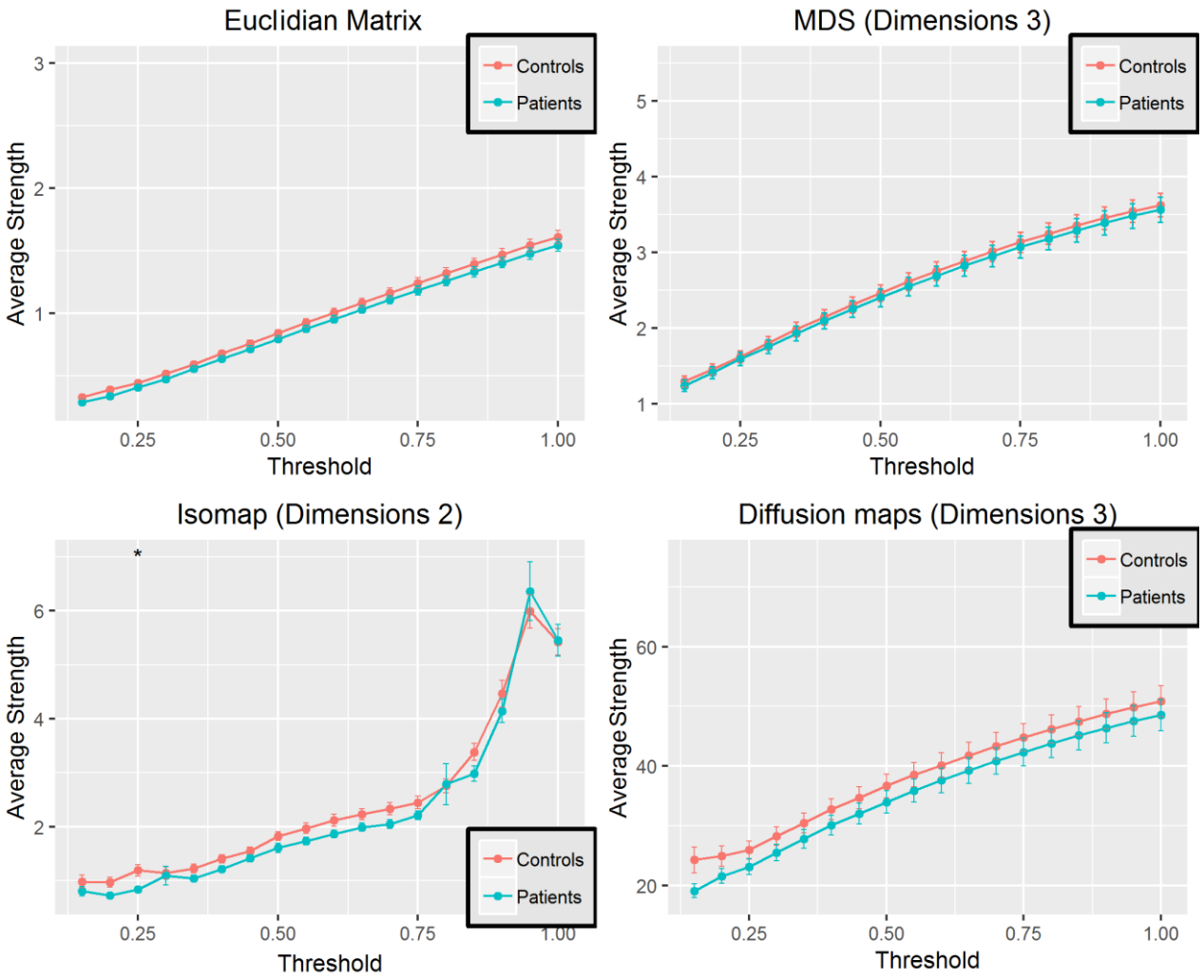
διαστάσεις της τελικής ενσωμάτωσης (σε πράσινο φόντο τα ποσοστά επιτυχούς ταξινόμησης για κάθε ταξινομητή στην επιλεχθείσα ενσωμάτωση).

Παρακάτω, βλέπουμε τη συμπεριφορά των γραφοθεωρητικών στοιχείων ανάμεσα σε υγιείς και σχιζοφρενείς για κάθε μία από τις μεθόδους συμπεριλαμβανομένου και του πίνακα απόστασης για τον οποίο δεν έχει γίνει μείωση διάστασης. Για το μέσο μήκος μονοπατιού (Εικόνα 4.5),για το μέσο βάρος/δύναμη κόμβου (Εικόνα 4.6) αλλά και το συντελεστή συγκρότησης (Εικόνα 4.7). Με αστερίσκο σημειώνεται επάνω σε κάθε υπογράφημα σε ποιο κατώφλι η διαφορά στη μέση τιμή είναι στατιστικά σημαντική με επίπεδο σημαντικότητας το οποίο έχουμε διορθώσει με Bonferonni διόρθωση ($p_{corrected} < 0.05$) όπως αυτή περιγράφηκε στο προηγούμενο κεφάλαιο. Η μπάρα σφάλματος μας δίνει



Εικόνα 4.5 *Μοτίβο διασυνδεσιμότητας με βάση το μέσο μήκος μονοπατιού (Average path length) όπως αυτό υπολογίστηκε στο διάστημα 15% -100% (με βήμα 5%) των δυνατότερων συνδέσεων των εγκεφαλικών δικτύων για κάθε μέθοδο(σε παρένθεση οι διαστάσεις που επιλέχθηκαν).Παρουσιάζονται ακόμη τα αποτελέσματα του πίνακα ευκλείδειας απόστασης (Euclidean matrix) χωρίς μείωση διάστασης (πάνω αριστερά).Τέλος, με αστερίσκο σημειώνεται η στατιστικά σημαντική διαφορά στη μέση τιμή (p¬corrected < 0.05) ανάμεσα σε υγιείς (κόκκινο) και σχιζοφρενείς (μπλε).*

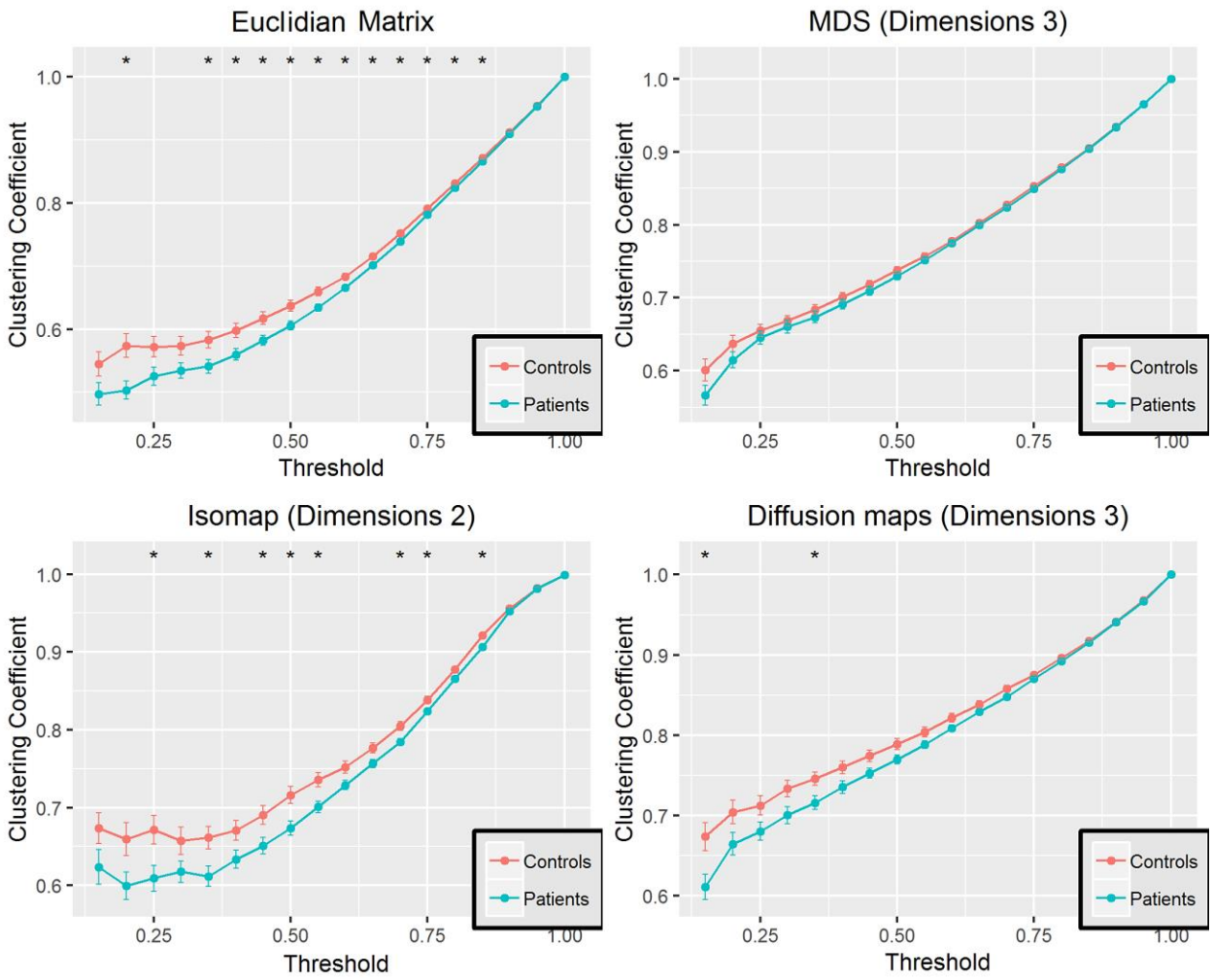
το τυπικό σφάλμα της μέσης τιμής και τελικά ο αριθμός των διαστάσεων της ενσωμάτωσης αναγράφεται για κάθε μέθοδο στον τίτλο του κάθε υπογραφήματος (σε παρένθεση).



Εικόνα 4.6 *Μοτίβο διασυνδεσιμότητας με βάση το μέσο βάρος/δύναμη (Average strength) των συνδέσεων όπως αυτές υπολογίστηκαν στο διάστημα 15% -100% (με βήμα 5%) των δυνατότερων ακμών των εγκεφαλικών δικτύων για κάθε μέθοδο (σε παρένθεση οι διαστάσεις που επιλέχθηκαν).Παρουσιάζονται ακόμη τα αποτελέσματα του πίνακα ευκλείδειας απόστασης (Euclidean matrix) χωρίς μείωση διάστασης (πάνω αριστερά).Τέλος, με αστερίσκο σημειώνεται η στατιστικά σημαντική διαφορά στη μέση τιμή ($p_{corrected} < 0.05$) ανάμεσα σε υγιείς (κόκκινο) και σχιζοφρενείς (μπλε).*

4.2.2 Ανάλυση ANOVA 2 παραγόντων για την εκτίμηση της κάθε μεθόδου και τη συμπεριφορά των ταξινομητών

Για τις επιλεχθείσες διαστάσεις σε κάθε μέθοδο πραγματοποιήσαμε ανάλυση ANOVA 2 παραγόντων εξετάζοντας την μέση επιτυχή ταξινόμηση σαν συνάρτηση των μεθόδων μείωσης διάστασης (1ος παράγοντας) αλλά και των αλγορίθμων ταξινόμησης (2ος παράγοντας). Παρακάτω και συγκεκριμένα στον πίνακα 6 παρουσιάζονται τα αποτελέσματα της ανάλυσης διασποράς 2 παραγόντων για τη μετρική βασιζόμενη στην



Εικόνα 4.7 *Μοτίβο διασυνδεσιμότητας με βάση το συντελεστή συγκρότησης (Clustering coefficient) όπως αυτός υπολογίστηκε στο διάστημα 15% -100% (με βήμα 5%) των δυνατότερων συνδέσεων των εγκεφαλικών δικτύων για κάθε μέθοδο(σε παρένθεση οι διαστάσεις που επιλέχθηκαν).Παρουσιάζονται ακόμη τα αποτελέσματα του πίνακα ευκλείδειας απόστασης (Euclidean matrix) χωρίς μείωση διάστασης (πάνω αριστερά).Τέλος, με αστερίσκο σημειώνεται η στατιστικά σημαντική διαφορά στη μέση τιμή ($p_{corrected} < 0.05$) ανάμεσα σε υγιείς (κόκκινο) και σχιζοφρενείς (μπλε).*

ευκλείδεια νόρμα. Παρατηρούμε ότι τα F στατιστικά είναι μεγάλα τόσο για τις μεθόδους όσο και τους ταξινομητές. Αυτό σημαίνει ότι η επιτυχής ταξινόμηση των υποκειμένων σε σχιζοφρενείς και υγιείς εξαρτάται σημαντικά και από τους δύο παράγοντες.

Στη συνέχεια διεξάγουμε Tuckey post-hoc τεστ για να αποφανθούμε ειδικότερα ποια μέθοδος συμπεριφέρεται καλύτερα (Πίνακας 7) αλλά και ποιοι από τους ταξινομητές συμπεριφέρονται καλύτερα (Πίνακας 8) στο πρόβλημα αυτό. Τελικά παρουσιάζονται τα αποτελέσματα για κάθε μέθοδο και ταξινομητή σε μορφή boxplots Εικόνα 4.8.

Πηγή (Source)	Βαθμοί ελευθερίας	Τετραγωνικά αθροίσματα	Μέσο Άθροισμα	F στατιστικό	Πιθανότητα (> F)

	(Degrees of freedom	(Squared Sums)	(Mean Sums)		
Ταξινομητές	3	0.10771	0.03590	30.825	**< e-16**
Μέθοδοι	3	0.02343	0.00781	6.706	**0.000223**
Υπόλοιπο	265	0.30867	0.00116		

Πίνακας 6 *Πίνακας ανάλυσης διασποράς δύο παραγόντων για την μετρική βασιζόμενη στην ευκλείδεια νόρμα. Με έντονα γράμματα σημειώνονται στατιστικά σημαντικοί παράγοντες της ανάλυσης.*

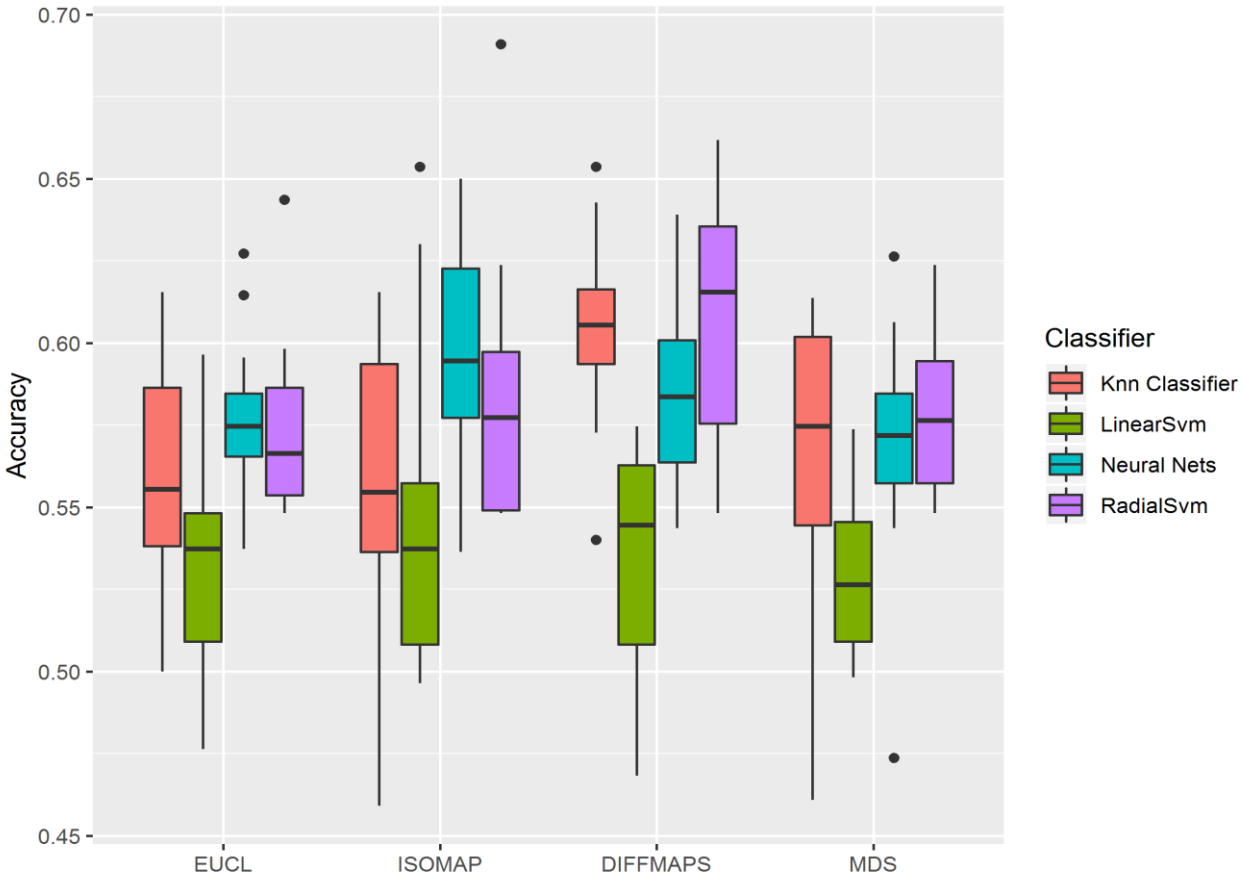
Tuckey Post-hoc (για τον παράγοντα Μέθοδοι, t-έλεγχοι)	Διάστημα εμπιστοσύνης (95%)	Στατιστική διαφορά στη μέση τιμή	p- τιμή ελέγχου (FWE corrected)
ISOMAP-Euclidian Matrix	(-0.006,0.023)	0.008	0.4607292
Diffusion Maps-Euclidian Matrix	(0.007,0.037)	0.022	**0.0007102**
MDS-Euclidian Matrix	(-0.014,0.015)	0.000	0.9999882
Diffusion Maps-ISOMAP	(0.000,0.029)	0.014	0.0735025
MDS-ISOMAP	(-0.023,0.006)	-0.008	0.4804291
MDS-Diffusion Maps	(-0.037,-0.007)	-0.022	**0.0008028**

Πίνακας 7 *Post-hoc Tuckey test για τον ρόλο που παίζει η εκάστοτε μέθοδος όσον αφορά το μέσο ποσοστό ταξινόμησης. Με έντονα γράμματα σημειώνονται στατιστικά σημαντικές διαφορές (με διόρθωση πολλαπλών συγκρίσεων).*

Tuckey Post-hoc (για τον παράγοντα Ταξινομητές, t-έλεγχοι)	Διάστημα εμπιστοσύνης (95%)	Στατιστική διαφορά στη μέση τιμή	p- τιμή ελέγχου (FWE corrected)
LSVM-KNN	(-0.051, -0.020)	-0.036	**< e-10**
NNETS-KNN	(0.004,0.025)	0.010	0.2789575
RSVM-KNN	(0.000,0.029)	0.014	0.0714779
NNETS-LVSM	(0.031,0.061)	0.046	**< e-10**

RSVM-LSVM	(0.035,0.065)	0.050	< e-10
RSVM-NNETS	(-0.011,0.018)	-0.003	0.9151462

Πίνακας 8 *Post-hoc Tuckey test για τον ρόλο που παίζει ο εκάστοτε αλγόριθμος ταξινόμησης όσον αφορά το μέσο ποσοστό ταξινόμησης. Με έντονα γράμματα σημειώνονται στατιστικά σημαντικές διαφορές (με διόρθωση πολλαπλών συγκρίσεων).*



Εικόνα 4.8 *Συγκεντρωτικά αποτελέσματα του μέσου ποσοστού ταξινόμησης για κάθε μέθοδο (καθώς και του πίνακα ευκλείδειων αποστάσεων (EUCL) χωρίς μείωση διάστασης) και ταξινομητή. Η οριζόντια γραμμή είναι η διάμεσος της κατανομής ενώ σαν σημεία αναγράφονται ακραίες τιμές της κατανομής.*

5 ΣΥΖΗΤΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Στην παρούσα μεταπτυχιακή εργασία, πραγματοποιήθηκε μία συγκριτική ανάλυση μεθόδων εκμάθησης πολλαπλοτήτων για την μείωση της διάστασης δεδομένων στην προσπάθεια της ανάλυσης λειτουργικής διασυνδεσιμότητας εγκεφαλικών δικτύων με στόχο την ταξινόμηση των υποκειμένων σε υγιείς και σχιζοφρενείς.

Στα πλαίσια της συγκριτικής ανάλυσης χρησιμοποιήθηκαν δύο μετρικές οι οποίες μας έδωσαν και το μέτρο ομοιότητας ανάμεσα στα εγκεφαλικά σήματα.

Η μία μετρική βασίστηκε στην συνήθη πολλαπλή συσχέτιση ενώ η δεύτερη στην ευκλείδεια νόρμα. Είναι καλό να τονίσουμε σε αυτό το σημείο ότι στόχος της εργασίας δεν ήταν η καλύτερη δυνατή ταξινόμηση των υποκειμένων αλλά η σύγκριση των μεθόδων μείωσης διάστασης, το κατά πόσο δηλαδή αυτές διατηρούν τη χρήσιμη πληροφορία του πίνακα ομοιότητας σε έναν χαμηλότερης διάστασης χώρο, κατά πόσο αυτό φαίνεται να επιτυγχάνεται στο επίπεδο του μοτίβου διασυνδεσιμότητας όσων αφορά τα γραφοθεωρητικά μέτρα που υπολογίσαμε και τέλος στο επίπεδο της επιτυχούς ταξινόμησης ανάμεσα στις δύο ομάδες του δείγματος (υγιείς και σχιζοφρενείς).

Σύμφωνα με τα αποτελέσματα αλλά και τις δύο μετρικές που χρησιμοποιήθηκαν παρατηρήσαμε μία γενικότερη τάση για διαφορές των εγκεφαλικών δικτύων στις ιδιότητες μικρού κόσμου αλλά και το βάρος των συνδέσεων.

Οι σχιζοφρενείς παρουσίασαν μεγαλύτερο μέσο μήκος μονοπατιού, μικρότερο μέσο βάρος/δύναμη σύνδεσης αλλά και μικρότερο συντελεστή συγκρότησης κάτι που πρακτικά σημαίνει ότι τα εγκεφαλικά δίκτυα των σχιζοφρενών είναι περισσότερο αποσυνδεδεμένα και λιγότερο συγχρονισμένα (Liu et al. 2008, Anderson και Cohen 2013).

Τα ευρήματα αυτά έρχονται σε πλήρη αντιστοιχία με ευρήματα τις βιβλιογραφίας που δείχνουν εν γένει ότι η σχιζοφρένεια μπορεί να οφείλεται σε μοτίβο αποσυνδεσιμότητας των εγκεφαλικών δικτύων(Friston and Frith 1995; Friston 1998).

Το μοτίβο αυτό φαίνεται πράγματι καλύτερα χρησιμοποιώντας την μετρική της πολλαπλής συσχέτισης (με καθυστερήσεις) τόσο στον βαθμό που οι τ έλεγχοι διαχωρίζουν τις δύο ομάδες αλλά και όσον αφορά το ποσοστό της επιτυχούς ταξινόμησης. Τα αποτελέσματα της εργασίας αυτής δείχνουν ότι δεν θα πρέπει να γίνεται χρήση της ευκλείδειας νόρμας για την διερεύνηση της λειτουργικής διασυνδεσιμότητας εγκεφαλικών σημάτων. Μία εξήγηση για την κακή απόδοση της ευκλείδειας μετρικής είναι ότι κάποια σήματα στον εγκέφαλο μπορεί όντως να συσχετίζονται με μία φάση υστέρησης (phase lag), κάτι που η ευκλείδεια μετρική δεν μπορεί να αποδώσει.

Όσον αφορά τις μεθόδους μείωσης διάστασης, η ανάλυση διασποράς με δύο παράγοντες μας δείχνει ότι και στις δύο περιπτώσεις μετρικών ο παράγοντας της μεθόδου μείωσης διάστασης παίζει ρόλο στη διαφοροποίηση του μέσου ποσοστού ταξινόμησης. Από τα post-hoc τεστ, φαίνεται πως οι μη γραμμικές μέθοδοι όπως το ISOMAP και το Diffusion Maps λειτουργούν καλύτερα από την γραμμική (που εδώ είναι η MDS). Πράγματι, οι μη γραμμικές μέθοδοι εγκλωβίζουν το μοτίβο διασυνδεσιμότητας του πίνακα ομοιότητας σε καλύτερο βαθμό από ότι η γραμμική μέθοδος στις ίδιες περίπου διαστάσεις.

Ωστόσο αυτό φαίνεται καλύτερα για την μετρική της πολλαπλής συσχέτισης, αφού για την ευκλείδεια απόσταση οι διαφορές στο μέσο μήκος μονοπατιού και το μέσο βάρος

σύνδεσης δείχνουν να εξαφανίζονται ενώ οι διαφορές στον συντελεστή συγκρότησης μοιάζουν να αμβλύνονται. Σε επίπεδο ταξινόμησης και πάλι οι μη γραμμικές μέθοδοι ISOMAP και Diffusion Maps διαχώρισαν τα δεδομένα σε καλύτερο ποσοστό από ότι η MDS ενώ αξιοσημείωτο είναι το γεγονός ότι παρόλη τη μείωση διάστασης και την αναπόφευκτη απώλεια πληροφορίας η Diffusion Maps διαχώρισε τα δεδομένα σε στατιστικά σημαντικά καλύτερο βαθμό από οποιαδήποτε άλλη μέθοδο συμπεριλαμβανομένου και του πίνακα ομοιότητας χωρίς μείωση διάστασης.

Αναφορικά με τους ταξινομητές η ανάλυση διασποράς και στις δύο περιπτώσεις των μετρικών μας δείχνει ότι υπάρχει διαφοροποίηση του μέσου ποσοστού ταξινόμησης ανάλογα με τον ταξινομητή που χρησιμοποιείται. Από τα post-hoc τεστ, καλύτερα φαίνεται να συμπεριφέρονται στην περίπτωσή μας ο SVM με Radial Basis Functions και τα νευρωνικά δίκτυα. Όσον αφορά την ευκλείδεια απόσταση χειρότερα αποτελέσματα συγκριτικά με άλλους αλγόριθμους ταξινόμησης έδωσε ο απλός γραμμικός SVM, ενώ στην περίπτωση της πολλαπλής συσχέτισης ο KNN ταξινομητής.

Χρησιμοποιώντας σχετικά απλές μεθόδους και απλούς ταξινομητές αναδείξαμε ένα μέσο ποσοστό ταξινόμησης 68% σε έναν χώρο ενσωμάτωσης τεσσάρων διαστάσεων για ένα σύνολο δεδομένων (47 σχιζοφρενείς και 57 υγιείς) όπου ένας τυχαίος ταξινομητής θα απέδιδε περίπου 50.4% ενώ με ένας αυστηρός κανόνας της πλειοψηφίας (majority rule) θα απέδιδε 54.8%. Το ποσοστό αυτό, ενώ φαίνεται να έχει χαμηλή αξιοπιστία είναι κάτι σημαντικό αν σκεφτεί κανείς ότι το μοντέλο μας ήταν πολύ απλό αφού δεν λάβαμε υπόψη πληροφορίες που θα μπορούσαν να αυξήσουν το ποσοστό επιτυχίας όπως δημογραφικά χαρακτηριστικά των υποκειμένων π.χ. ηλικία, φύλο κτλ. Επίσης τα χαρακτηριστικά που δώσαμε στους ταξινομητές για τον διαχωρισμό των υποκειμένων ήταν μόνο 3 βασικά γραφοθεωρητικά χαρακτηριστικά.

Ακόμα θα μπορούσαμε να έχουμε ένα πιο ισορροπημένο δείγμα όσον αφορά το φύλο την ηλικία κτλ. των υποκειμένων. Για παράδειγμα στο δείγμα το οποίο αναλύσαμε οι άνδρες του δείγματος ήταν περισσότερο πιθανό να είναι σχιζοφρενείς από ότι οι γυναίκες. Αν σκεφτεί κανείς λοιπόν την απλοϊκότητα του μοντέλου, το αποτέλεσμά μας είναι κάτι αξιοσημείωτο. Ωστόσο η ανάλυση αυτή, μπορεί να χρησιμοποιηθεί γενικά και σε άλλες περιπτώσεις με κάποιες παραλλαγές που θα βελτιώσουν δραματικά την απόδοση της. Το συμπέρασμα που προκύπτει είναι ότι ως μέθοδος η ανάλυση λειτουργικής διασυνδεσιμότητας δικτύου μπορεί να βοηθήσει στην ταξινόμηση δεδομένων FMRI και το διαχωρισμό ψυχικών ασθενειών.

Συμπερασματικά λοιπόν, δείξαμε πως η λειτουργική διασυνδεσιμότητα μπορεί να μετρηθεί από μία γραφοθεωρητική σκοπιά, χρησιμοποιώντας τις ιδιότητες μικρού κόσμου (small-world properties) εφαρμόζοντας γραμμικές και μη γραμμικές μεθόδους για τον μετασχηματισμό των 4-διαστάσεων δεδομένων μας (FMRI scans) σε γράφους με κόμβους τις ανεξάρτητες συνιστώσες των υποκειμένων και αποστάσεις που εκφράζουν το μέτρο ομοιότητας των χρονοσειρών ανάλογα με τη μετρική που χρησιμοποιείται. Ωστόσο υπάρχουν κι άλλες μέθοδοι για την εφαρμογή της ανάλυσης λειτουργικής διασυνδεσιμότητας όπως για παράδειγμα η μέθοδος της Granger αιτιότητας (Granger causality) (Roebroeck et al. 2005) ανάμεσα σε περιοχές ενδιαφέροντος (ROIs) όπως παρουσίασαν οι (Sato et al. 2010), όπου αυτή η εργασία εφαρμόζει λειτουργική διασυνδεσιμότητα μέσω συσχέτισης σε λειτουργικά δίκτυα που εξάγονται από την ICA. Ακόμα υπάρχουν και άλλες μετρικές αποστάσεις που θα μπορούσαμε να χρησιμοποιήσουμε αντί της πολλαπλής συσχέτισης (Cross-correlation) και της ευκλείδειας απόστασης αφού σύμφωνα με πρόσφατα ευρήματα από τους (Zalesky et al. 2012) η χρήση μέτρων συσχέτισης για τον υπολογισμό απόστασης αυτόματα οδηγεί σε μη-τυχαίες γραφικές δομές.

Τέλος, παρότι η ανάλυση αυτή έγινε για FMRI δεδομένα πιο γενικά μπορεί να εφαρμοστεί για προβλήματα όπου η σχέση στις πηγές των σημάτων μπορεί να αποφασίσει την κατηγορία στην οποία ανήκει το υποκείμενο μελέτης. Η κοινή συμπεριφορά των πηγών σημάτων (ανεξάρτητες συνιστώσες) παρατηρήθηκε σαν γράφημα όπου αποστάσεις ανάμεσα στις πηγές ανέδειξαν ομοιότητα στη συμπεριφορά. Η λειτουργική διασυνδεσιμότητα λοιπόν, είναι μία τεχνική που θα πρέπει να εκτιμηθεί από διάφορες σκοπιές.

6 ΑΝΑΦΟΡΕΣ

Anderson A, Bramen J, Douglas PK, et al (2011) Large sample group independent component analysis of functional magnetic resonance imaging using anatomical atlas-based reduction and bootstrapped clustering. Int J Imaging Syst Technol 21:223–231. doi: 10.1002/ima.20286

Anderson A, Cohen MS (2013) Decreased small-world functional network connectivity and clustering across resting state networks in schizophrenia: an fMRI classification tutorial. Front Hum Neurosci 7:520. doi: 10.3389/fnhum.2013.00520

Anderson A, Dinov ID, Sherin JE, et al (2010) Classification of spatially unaligned fMRI scans. Neuroimage 49:2509–2519. doi: 10.1016/j.neuroimage.2009.08.036

Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science (80-) 286:509–512. doi: 10.1126/science.286.5439.509

Bassett DS, Bullmore E, Verchinski BA, et al (2008) Hierarchical Organization of Human Cortical Networks in Health and Schizophrenia. J Neurosci 28:9239–9248. doi: 10.1523/jneurosci.1929-08.2008

Benjaminsson (2010) A novel model-free data analysis technique based on clustering in a mutual information space: application to resting-state fMRI. Front Syst Neurosci 4:34. doi: 10.3389/fnsys.2010.00034

Biswal B, Zerrin Yetkin F, Haughton VM, Hyde JS (1995) Functional connectivity in the motor cortex of resting human brain using echo-planar mri. Magn Reson Med 34:537–541. doi: 10.1002/mrm.1910340409

Bookheimer S (2002) Functional MRI of language: new approaches to understanding the cortical organization of semantic processing. Annu Rev Neurosci 25:151–188

Chu C, Handwerker DA, Bandettini PA, Ashburner J (2011) Measuring the consistency of global functional connectivity using kernel regression methods. In: Proceedings - International Workshop on Pattern Recognition in NeuroImaging, PRNI 2011. pp 41–44

Cohen MS, Bookheimer SY (1994) Localization of brain function using magnetic resonance imaging. Trends Neurosci 17:268–277

Coifman RR, Lafon S, Lee AB, et al (2005) Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. Proc Natl Acad Sci U S A 102:7426–7431

Csardi G, Nepusz T (2006) The igraph software for complex network research. InterJournal, Complex Syst 1695:1–9

Dijkstra EW (1959) A note on two problems in connexion with graphs. Numer Math 1:269–271. doi: 10.1007/BF01386390

Douglas Carroll J, Arabie P (1998) Multidimensional Scaling. Meas Judgm Decis Mak 11:179–250. doi: 10.1016/b978-012099975-0.50005-1

Floyd RW (1962) Algorithm 97: Shortest path. Commun ACM 5:345. doi: 10.1145/367766.368168

Ford J, Farid H, Makedon F, et al (2003) Patient Classification of fMRI Activation Maps. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp 58–65

Friston KJ (1998) The disconnection hypothesis. Schizophr Res 30:115–125. doi: 10.1016/S0920-9964(97)00140-0

Friston KJ, Frith CD (1995) Schizophrenia: a disconnection syndrome? Clin Neurosci 3:89–97

Friston KJ, Holmes AP, Worsley KJ, et al (1994) Statistical parametric maps in functional imaging: A general linear approach. Hum Brain Mapp 2:189–210. doi: 10.1002/hbm.460020402

Gao B, Vanschoren J (2011) Visualizations of machine learning behavior with dimensionality reduction

techniques. In: Benelearn 2011. Proceedings of the Twentieth Belgian Dutch Conference on Machine Learning 35-42). pp 35–42

Garrity AG, Pearlson GD, McKiernan K, et al (2007) Aberrant “default mode” functional connectivity in schizophrenia. Am J Psychiatry 164:450–457. doi: 10.1176/ajp.2007.164.3.450

Gauthier I, Tarr MJ, Anderson AW, et al (1999) Activation of the middle fusiform ’face area’ increases with expertise in recognizing novel objects. Nat Neurosci 2:568–573

Goutte C, Toft P, Rostrup E, et al (1999) On clustering fMRI time series. Neuroimage 9:298–310. doi: 10.1006/nimg.1998.0391

Grama A, Gupta A, Karypis G, Kumar V (2003) Introduction to Parallel Computing (Second Edition) , Addison-Wesley

Hecht-Nielsen R (1988) Theory of the backpropagation neural network. Neural Networks 1:445. doi: 10.1016/0893-6080(88)90469-8

Heine L, Soddu A, Gómez F, et al (2012) Resting state networks and consciousness. Front Psychol 3:295. doi: 10.3389/fpsyg.2012.00295

Hernandez JM, Van Mieghem P (2011) Classification of graph metrics. Delft Univ Technol Mekelweg, Netherlands 1–20

Himberg J, Hyvärinen A, Esposito F (2004) Validating the independent components of neuroimaging time series via clustering and visualization. Neuroimage 22:1214–1222. doi: 10.1016/j.neuroimage.2004.03.027

Hyvärinen A (1999) Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans Neural Networks 10:626–634. doi: 10.1109/72.761722

Hyvärinen A, Karhunen J, Oja E (2000) Independent component analysis: algorithms and applications. Neura Networks 13:411–430

Jafri MJ, Pearlson GD, Stevens M, Calhoun VD (2008) A method for functional network connectivity among spatially independent resting-state components in schizophrenia. Neuroimage 39:1666–1681. doi: 10.1016/j.neuroimage.2007.11.001

Jenkinson M, Bannister P, Brady M, Smith S (2002) Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17:825–841

Jolliffe IT (2002) Principal Component Analysis. Second Edition

Kim S-G, Rostrup E, Larsson HBW, et al (1999) Determination of relative CMRO 2 from CBF and BOLD changes: significant increase of oxygen consumption rate during visual stimulation. Magn Reson Med 41:1152–1161

Koshino H, Carpenter PA, Minshew NJ, et al (2005) Functional connectivity in an fMRI working memory task in high-functioning autism. Neuroimage 24:810–821. doi: 10.1016/j.neuroimage.2004.09.028

Krogh A, Hertz JAJA (1992) A Simple Weight Decay Can Improve Generalization. In: Advances in Neural Information Processing Systems 4: Proceedings of the 1991 Conference (NIPS 1991). pp 950–957

Kruskal JB (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29:1–27. doi: 10.1007/BF02289565

Kuhn M (2008) caret Package. J Stat Softw 28:1–26

Lei Zhang, Samaras D, Tomasi D, et al (2005) Machine Learning for Clinical Diagnosis from Functional Magnetic Resonance Imaging. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp 1211–1217

Liang M, Zhou Y, Jiang T, et al (2006) Widespread functional disconnectivity in schizophrenia with resting-state

functional magnetic resonance imaging. Neuroreport 17:209–213. doi: 10.1097/01.wnr.0000198434.06518.b8

Liow JS, Rehm K, Strother SC, et al (2000) Comparison of voxel- and volume-of-interest-based analyses in FDG PET scans of HIV positive and healthy individuals. J Nucl Med 41:612–21

Liu P, Zhang Y, Zhou G, et al (2009) Partial correlation investigation on the default mode network involved in acupuncture: An fMRI study. Neurosci Lett 462:183–187. doi: 10.1016/j.neulet.2009.07.015

Liu X, Zhu XH, Qiu P, Chen W (2012) A correlation-matrix-based hierarchical clustering method for functional connectivity analysis. J Neurosci Methods 211:94–102. doi: 10.1016/j.jneumeth.2012.08.016

Liu Y, Liang M, Zhou Y, et al (2008) Disrupted small-world networks in schizophrenia. Brain 131:945–961

Ma L, Wang B, Chen X, Xiong J (2007) Detecting functional connectivity in the resting brain: a comparison between ICA and CCA. Magn Reson Imaging 25:47–56. doi: 10.1016/j.mri.2006.09.032

Mannfolk P, Wirestam R, Nilsson M, et al (2010) Dimensionality reduction of fMRI time series data using locally linear embedding. Magn Reson Mater Physics, Biol Med 23:327–338. doi: 10.1007/s10334-010-0204-0

McKeown MJ, Hansen LK, Sejnowsk TJ (2003) Independent component analysis of functional MRI: What is signal and what is noise? Curr. Opin. Neurobiol. 13:620–629

Mezer A, Yovel Y, Pasternak O, et al (2009) Cluster analysis of resting-state fMRI time series. Neuroimage 45:1117–1125. doi: 10.1016/j.neuroimage.2008.12.015

Opsahl T, Panzarasa P (2009) Clustering in weighted networks. Soc Networks 31:155–163. doi: 10.1016/j.socnet.2009.02.002

Otte A, Halsband U (2006) Brain imaging tools in neurosciences. J Physiol Paris 99:281–292. doi: 10.1016/j.jphysparis.2006.03.011

Porte JD La, Herbst B (2008) An introduction to diffusion maps. In: Proceedings of the 19th Symposium of the Pattern Recognition Association of South Africa. pp 15–25

Pruim RHR, Mennes M, van Rooij D, et al (2015) ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. Neuroimage 112:267–277

Qureshi MNI, Oh J, Cho D, et al (2017) Multimodal Discrimination of Schizophrenia Using Hybrid Weighted Feature Concatenation of Brain Functional Connectivity and Anatomical Features with an Extreme Learning Machine. Front Neuroinform 11:59. doi: 10.3389/fninf.2017.00059

Richiardi J, Achard S, Bunke H, Van De Ville D (2013) Machine learning with brain graphs: predictive modeling approaches for functional imaging in systems neuroscience. IEEE Signal Process Mag 30:58–70

Ripley B, Hjort N (1996) Pattern recognition and neural networks. Cambridge University press.

Ripley B, Venables W (2011) nnet: Feed-forward neural networks and multinomial log-linear models. R Packag version 7:5

Roebroeck A, Formisano E, Goebel R (2005) Mapping directed influence over the brain using Granger causality and fMRI. Neuroimage 25:230–242. doi: 10.1016/j.neuroimage.2004.11.017

Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science (80-) 290:2323–2326. doi: 10.1126/science.290.5500.2323

Roy CS, Sherrington CS (1890) On the Regulation of the Blood-supply of the Brain. J Physiol 11:85–158. doi: 10.1113/jphysiol.1890.sp000321

Salimi-Khorshidi G, Douaud G, Beckmann CF, et al (2014) Automatic denoising of functional MRI data: Combining independent component analysis and hierarchical fusion of classifiers. Neuroimage 90:449–468. doi:

10.1016/j.neuroimage.2013.11.046

Sato JR, Fujita A, Cardoso EF, et al (2010) Analyzing the connectivity between regions of interest: An approach based on cluster Granger causality for fMRI data analysis. Neuroimage 52:1444–1455. doi: 10.1016/j.neuroimage.2010.05.022

Scholz M (2006) Approaches to analyse and interpret biological profile data. PhD thesis, Univ Potsdam, Fac Math Nat Sci 101

Shen H, Wang L, Liu Y, Hu D (2010) Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fMRI. Neuroimage 49:3110–3121. doi: 10.1016/j.neuroimage.2009.11.011

Shen X, Meyer FG (2011) Analysis of Event-Related fMRI Data Using Diffusion Maps. In: Biennial International Conference on Information Processing in Medical Imaging. pp 652–663

Siettos C, Starke J (2016) Multiscale modeling of brain dynamics: from single neurons and networks to mathematical tools. Wiley Interdiscip Rev Syst Biol Med 8:438–458

Singer A, Erban R, Kevrekidis IG, Coifman RR (2009) Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps. Proc Natl Acad Sci 106:16090–16095. doi: 10.1073/pnas.0905547106

Sipola T, Cong F, Ristaniemi T, et al (2013) Diffusion map for clustering fMRI spatial maps extracted by independent component analysis. In: IEEE International Workshop on Machine Learning for Signal Processing, MLSP. pp 1–6

Smith SM (2002) Fast robust automated brain extraction. Hum Brain Mapp 17:143–55. doi: 10.1002/hbm.10062

Smith SM, Fox PT, Miller KL, et al (2009) Correspondence of the brain's functional architecture during activation and rest. Proc Natl Acad Sci 106:13040–13045. doi: 10.1073/pnas.0905267106

Tenenbaum JB, De Silva V, Langford JC (2000a) A global geometric framework for nonlinear dimensionality reduction. Science (80-) 290:2319–2323. doi: 10.1126/science.290.5500.2319

Tomasi D, Volkow ND (2014) Mapping small-world properties through development in the human brain: Disruption in schizophrenia. PLoS One 9:e96176. doi: 10.1371/journal.pone.0096176

Toppi J, De Vico Fallani F, Vecchiato G, et al (2012) How the Statistical Validation of Functional Connectivity Patterns Can Prevent Erroneous Definition of Small-World Properties of a Brain Connectivity Network. Comput Math Methods Med 2012:1–13. doi: 10.1155/2012/130985

Van De Ven VG, Formisano E, Prvulovic D, et al (2004) Functional connectivity as revealed by spatial independent component analysis of fMRI measurements during rest. Hum Brain Mapp 22:165–178. doi: 10.1002/hbm.20022

Van Wijk BCM, Stam CJ, Daffertshofer A (2010) Comparing brain networks of different size and connectivity density using graph theory. PLoS One 5:e13701. doi: 10.1371/journal.pone.0013701

Venkataraman A, Van Dijk KRA, Buckner RL, Golland P (2009) Exploring functional connectivity in fMRI via clustering. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. pp 441–444

Viviani R, Grön G, Spitzer M (2005) Functional principal component analysis of fMRI data. Hum Brain Mapp 24:109–129. doi: 10.1002/hbm.20074

Warrington EK (2013) Visual Deficits Associated with Occipital Lobe Lesions in Man. In: Experimental Brain Research Supplementum. pp 247–261

Watts DJ, Strogatz SH (1998) Collective dynamics of "small-world" networks. Nature 393:440–2. doi: 10.1038/30918

Whitfield-Gabrieli S, Thermenos HW, Milanovic S, et al (2009) Hyperactivity and hyperconnectivity of the default network in schizophrenia and in first-degree relatives of persons with schizophrenia. Proc Natl Acad Sci 106:1279–1284

Yang Z, LaConte S, Weng X, Hu X (2008) Ranking and averaging independent component analysis by reproducibility (RAICAR). Hum Brain Mapp 29:711–725. doi: 10.1002/hbm.20432

Yu Q, Sui J, Rachakonda S, et al (2011) Altered topological properties of functional network connectivity in schizophrenia during resting state: A small-world brain Network study. PLoS One 6:e25423. doi: 10.1371/journal.pone.0025423

Zaidi F (2013) Small world networks and clustered small world networks with random connectivity. Soc Netw Anal Min 3:51–63. doi: 10.1007/s13278-012-0052-1

Zalesky A, Fornito A, Bullmore E (2012) On the use of correlation as a measure of network connectivity. Neuroimage 60:2096–2106. doi: 10.1016/j.neuroimage.2012.02.001

Zeng LL, Wang H, Hu P, et al (2018) Multi-Site Diagnostic Classification of Schizophrenia Using Discriminant Deep Learning with Functional Connectivity MRI. EBioMedicine 30:74–85. doi: 10.1016/j.ebiom.2018.03.017