Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΧΡΙΣΤΟΠΟΥΛΟΣ ΚΛΕΦΤΟΓΙΑΝΝΗΣ ΠΑΝΑΓΙΩΤΗΣ

Υπολογιστική Ανάλυση της Γεωμετρικής Συμμεταβολής Χρηματο-οικονομικών Χρονοσειρών Χαρτοφυλακίων με Αλγορίθμους Εκμάθησης Πολλαπλοτήτων

Επιβλέπων Καθηγητής : Κωνσταντίνος Σιέττος, Αναπληρωτής Καθηγητής

Αθήνα, Σεπτέμβριος 2017

Ευχαριστίες

Πρωτίστως θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή της σχολής Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών Κωνσταντίνο Σιέττο για την ανάθεση, την απόλυτη εμπιστοσύνη και την πλήρη στήριξη που μου παρείχε κατά τη διάρκεια της παρούσας διπλωματικής. Έπειτα θα ήθελα να ευχαριστήσω τον διδακτορικό φοιτητή και φίλο Γιάννη Γάλλο καθώς οι συμβουλές του ήταν παραπάνω από χρήσιμες. Επίσης θα ήθελα να ευχαριστήσω τον φίλο Απόστολο Παραδέλλη για την καθολική βοήθεια που μου παρείχε στην ανεύρεση των δεδομένων που χρησιμοποιήθηκαν στην διπλωματική. Τέλος ένα μεγάλο ευχαριστώ οφείλω στην οικογένεια μου που με στηρίζει σε κάθε απόφαση της ζωής μου με όλα τα μέσα που διαθέτει καθώς και στους φίλους - συμφοιτητές μου που σημαίνουν τόσα για μένα.

Περίληψη

Ο σκοπός της παρούσας διπλωματικής εργασίας είναι η μελέτη και εφαρμογή σε χρηματοοικονομικά δεδομένα μεθόδων – αλγορίθμων μείωσης διαστάσεων υψηλού όγκου δεδομένων. Οι αλγόριθμοι εκμάθησης πολλαπλοτήτων (manifold learning), όπως αλλιώς ονομάζονται, που διέπουν τα δεδομένα που μελετιούνται απασχολούν την τελευταία δεκαετία σε πολύ μεγάλο βαθμό, την μαθηματική και ευρύτερη επιστημονική κοινότητα. Η συνεισφορά τους, βρίσκει μεγάλο αντίκτυπο σε τομείς όπως αυτούς των νευροεπιστημών, της βιολογίας και των χρηματοοικονομικών. Συγκεκριμένα στον τομέα των χρηματοοικονομικών, η ανεύρεση χώρων χαμηλότερων διαστάσεων, κρίνεται σε πολλές περιπτώσεις αναγκαία για την περαιτέρω ανάλυση και οπτικοποίηση του συνόλου των διαθέσιμων δεδομένων. Σε πρώτη φάση, δίνεται μία εισαγωγική προ - επισκόπηση στις γραμμικές μεθόδους ανεύρεσης πολλαπλοτήτων PCA και MDS καθώς και στις μη γραμμικές ISOMAP και Diffusion Maps. Επίσης παρουσιάζεται και η εφαρμογή τους στον τομέα των χρηματοοικονομικών. Κατόπιν η παρούσα διπλωματική επικεντρώνεται στο μαθηματικό υπόβαθρο που διέπει αυτές τις μεθόδους οι οποίες αναλύονται εκτενώς. Για την πλήρη κατανόηση των μεθόδων χρησιμοποιούνται παραδείγματα δεδομένων, με συγκεκριμένη γεωμετρική δομή, με στόχο την ανάδειξη της χρησιμότητας και περιορισμών της κάθε τεχνικής. Σε επόμενη φάση γίνεται μία συγκριτική μελέτη αυτών, σε δεδομένα που η γεωμετρική δομή τους παρουσιάζεται μέσω μίας Τοροειδής Έλικας (Toroidal Helix). Στο τελευταίο κεφάλαιο παρουσιάζεται η δημοσίευση του Phoa, 2012 και η αναπαραγωγή των αποτελεσμάτων αυτής. Η εφαρμογή της μεθόδου των απεικονίσεων διάχυσης βρίσκει μεγάλο αντίκτυπο στην ανάλυση δεδομένων (μετοχές) και στην εξαγωγή χρήσιμων ποσοτικών μέτρων, για την ανεύρεση της διαφοροποίησης που διέπει ένα χαρτοφυλάκιο και τις τοπικές κρίσης που επηρεάζουν αυτό. Έν τέλει δίνεται ένα παράρτημα, που περιλαμβάνει τους κώδικες που χρησιμοποιούνται για την εφαρμογή των αποτελεσμάτων και των διαγραμμάτων, που παράχθηκαν στα πλαίσια αυτής της διπλωματικής καθώς και η σχετική βιβλιογραφία.

Abstract

This thesis is dealing with the representation of dimensionality reduction techniques; methods applied in big volume data sets. These algorithms, known as manifold learning techniques are considered nowadays to be one of the hottest topics in the world of mathematical and computer science community, as they play a crucial role in scientific areas such as neuroscience, biology and finance. In particular with regards to finance, finding lower dimension spaces is necessary for further analysis examination and visualization of each financial data set. The thesis begins with the description of several manifold learning techniques (linear: PCA and MDS, nonlinear: ISOMAP and Diffusion Maps) and their implementation in the area of finance. Moving on, the focus is given on the mathematical background and the deep analysis of each method. Moreover, some examples (toy problems) are provided, giving prominence to each method. Following that, a comparison of the techniques efficiency is attempted, by applying them on the toroidal helix benchmark problem. The last chapter focuses in the publication of W. Phoa, 2012. In particular, the reproduction of the publication is attempted; though with the inclusion of alternative and fewer data. It can be concluded that the method of diffusion maps gives significant results in the visualization of financial data (assets). The method gives also the chance to extract important measures for the portfolio diversification, along with local shocks that could affect it. Finally, the thesis latest sections include the appendix, with all related codes used for measures and plots and the bibliography that we used.

Πίνακας περιεχομένων

Πίνακας σχημάτων

Εισαγωγή

Πολλά αντικείμενα στον σημερινό κόσμο μπορούν να παρουσιαστούν σε ψηφιακή μορφή στο διαδίκτυο αντιπροσωπεύοντας υψηλής διάστασης δεδομένων (σήματα ομιλίας, εικόνες, ηλεκτρονικά κείμενα, μετοχές κ.α.). Πολύ συχνά κρίνεται απαραίτητη η ανάλυση, η επεξεργασία και η κατανόηση των μεγάλων αυτών ποσοτήτων δεδομένων που αντιπροσωπεύουν χωροχρονικές μεταβολές φυσικών, χημικών, βιολογικών, οικονομικών μεγεθών. Λόγω των μεγάλων διαστάσεων που αντιπροσωπεύουν τέτοιου είδους δεδομένα η απευθείας επεξεργασία, καθιστά την εξαγωγή χρήσιμης πληροφορίας ιδιαίτερα πολύπλοκη. Η επίλυση του προβλήματος αυτού έρχεται να δοθεί από τη χρησιμοποίηση μεθόδων-τεχνικών μείωσης διαστάσεων (Dimensionality Reduction methods). Οι τεχνικές αυτές λόγω της σπουδαιότητας τους που παρουσιάζουν στην ανάλυση δεδομένων χρησιμοποιούνται σε πολλούς τομείς όπως στην αναγνώριση προτύπων (pattern recognition), εξόρυξη δεδομένων (data mining), συμπίεση δεδομένων (data compression), μηχανική μάθηση (machine learning) κ.α.

Οι μέθοδοι μείωσης διαστάσεων έχουν ως απώτερο στόχο την παρουσίαση δεδομένων υψηλών διαστάσεων, μέσω της εμφύτευσης (embedding) τους σε χώρο μικρότερων διαστάσεων με τέτοιο τρόπο ώστε η εγγενής γεωμετρία των δεδομένων να αναδεικνύεται με το καλύτερο δυνατό τρόπο. Αυτές οι τεχνικές οδηγούν στην άμεση χρησιμοποίηση τους για να πραγματοποιηθούν κάποιες από τις ακόλουθες ενέργειες:

- *Μείωση διαστάσεων μεγάλης κλίμακας (Data dimensionality reduction)*: Παραγωγή μίας συμπαγούς χαμηλότερης διάστασης αποκωδικοποίησης ενός συνόλου δεδομένων που χαρακτηρίζεται από πολλές, σε πλήθος, διαστάσεις.
- *Οπτικοποίηση δεδομένων (Data visualization)*:
  Εξαγωγή χρήσιμων πληροφοριών μέσω του μειωμένης διάστασης χώρου, μεταφράζοντας την δομή του αρχικού ως προς τους βαθμούς ελευθερίας που τον διακατέχουν.
- *Μείωση διαστάσεων και η χρησιμοποίηση τους για επιβλεπόμενη μάθηση (Preprocessing for supervised learning)*:
  Απλοποίηση, μείωση και καθαρισμός των δεδομένων με απώτερο σκοπό την εισχώρησή τους σε αλγορίθμους επιβλεπόμενης μάθησης.

Κλασικοί μέθοδοι μείωσης διαστάσεων όπως η PCA και η cMDS βασίζονται σε γραμμικά μοντέλα δεδομένων, οι οποίες θεωρούν πως τα δεδομένα εκτείνονται πάνω σε γραμμικούς υπό-χώρους χαμηλότερων διαστάσεων. Η αποτελεσματικότητα τους πάνω σε δεδομένα που δεν βρίσκονται πάνω σε υπερ-επίπεδα, ανέδειξε τον δρόμο για την έρευνα εύρεσης μη-γραμμικών τεχνικών μείωσης διαστάσεων, γνωστές και ως τεχνικές εύρεσης της πολλαπλότητας (manifold learning). Πρακτικά αυτές οι μέθοδοι υποθέτουν πως τα δεδομένα υψηλών διαστάσεων εκτείνονται πάνω σε μια χαμηλότερης διάστασης πολλαπλότητα (manifold). Μια μη-γραμμική μέθοδος ''μαθαίνει'' την υποβόσκουσα πολλαπλότητα από τα δεδομένα του εκάστοτε προβλήματος και τα αναπαριστά μέσω των χαμηλών διαστάσεων των συντεταγμένων τους **(Wang, 2012)**. Τέτοιες μέθοδοι που θα μελετηθούν στην παρούσα διπλωματική είναι αυτές των Isomap και Diffusion Maps.

Η μείωση διαστάσεων μπορεί να οριστεί ως η διαδικασία εξόρυξης του συνόλου των βαθμών ελευθερίας, οι οποίες χρησιμοποιούνται ώστε να αναπαράγουν το μεγαλύτερο μέρος της μεταβλητότητας ενός συνόλου δεδομένων **(Ghodsi, 2006)**.

Σε αυτό το σημείο κρίνεται ικανό να δοθεί το μαθηματικό υπόβαθρό πίσω από τις μεθόδους μείωση διαστάσεων. Δοθέντος λοιπόν ενός συνόλου δεδομένων $X = \{x_1, x_2, \dots, x_N\}$, $x_i \in \mathbb{R}^D$ όπου N το συνολικό νούμερο δειγμάτων (samples) και D το συνολικό νούμερο χαρακτηριστικών (features), η μείωση διαστάσεων μπορεί να χαρακτηριστεί ως, η εύρεση μιας συνάρτησης απεικόνισης $F : x \to y$ η οποία μεταμορφώνει τα $x \in \mathbb{R}^D$ στην επιθυμητή μειωμένης διάστασης παρουσίαση $y \in \mathbb{R}^d$ **(Chang, 2012)**.

Εδώ και έναν αιώνα περίπου οι ερευνητές ανά το κόσμο στόχευαν στην παρουσίαση των δεδομένων, με τέτοιους τρόπους σώστε να γίνονται αντιληπτές οι δομές και η εγγενής γεωμετρία που διακατέχουν τα εκάστοτε δεδομένα. Σε πολλούς τομείς όπως αυτοί τις τεχνητής νοημοσύνης, της βιοπληροφορικής και των χρηματοοικονομικών οι μέθοδοι εύρεσης πολλαπλότητας βρίσκουν πρόσφορο έδαφος ανάπτυξής και εφαρμογής τους.

Ιδιαίτερα στον τομέα των χρηματοοικονομικών οι μέθοδοι μείωσης διαστάσεων προσφέρονται άμεσα για οπτικοποίηση, εξερεύνηση και εξαγωγή συμπερασμάτων των δεδομένων καθώς συνήθως οι χρηματοοικονομικοί δείκτες είναι πολυπληθής και τα δεδομένα που περιγράφουν μια οικονομία διαμορφώνουν πολύπλοκες δομές **(Seng & Lee, 2000)**.

Ειδικότερα τα τελευταία χρόνια λόγω της αύξησης των δεδομένων και της καλυτέρευσης των υπολογιστικών συστημάτων, οι ερευνητές ανά το κόσμο εφάρμοσαν τέτοιες τεχνικές, βασιζόμενοι στην δομή κλασσικών και μη τεχνικών εύρεσης πολλαπλότητας. Αρκετές δημοσιεύσεις την τελευταία δεκαετία έβγαλαν πολύ σημαντικά αποτελέσματα στο τομέα των χρηματοοικονομικών είτε χρησιμοποιώντας τεχνικές μείωσης διαστάσεων για οπτικοποίηση και εξαγωγή διάφορων μοτίβων είτε για την εφαρμογή τους σε τεχνικές επιβλεπόμενης μάθησης.

Η μέθοδος της ανάλυσης κύριων συνιστωσών(Principal Component Analysis:PCA) είναι η πιο γνωστή μέθοδος λόγω της απλότητας της και η πρώτη που χρησιμοποιήθηκε στην πολυμεταβλητή ανάλυση. Το εύρος των εφαρμογών της εξαπλώθηκε τα τελευταία 50 χρόνια, καθώς όπως αναφέρθηκε η ανάπτυξη του κλάδου της πληροφορικής, βοήθησε την χρησιμοποίηση της σε ποικίλους τομείς. Η ικανότητα της να μετατρέπει συσχετισμένες μεταβλητές σε ασυσχέτιστα συστατικά την κάνει προσοδοφόρα στο τομέα των χρηματοοικονομικών αγορών. Έχει εφαρμοστεί στην αναπαραγωγή νέων δεικτών αγοράς (market indices) **(Feeney and Hester, 1967)** και στην αναγνώριση κοινών παραγόντων στις αποδόσεις ομολόγων (bond returns) **(Driesson et al., 2003; Périgdon et al., 2007)**. Πιο πρόσφατα, η ανάπτυξή της επικεντρώθηκε στην εκμάθηση της συμμεταβολής της αγοράς και του συστημικού κινδύνου **(Billioand et al., 2012; Kritzman et al., 2011; Zheng et al., 2012)**. Παρά την σπουδαιότητα της και την ευρεία χρησιμοποίηση της στο τομέα των χρηματοοικονομικών, η μέθοδος ανιχνεύει μόνο γραμμικότητες στα δεδομένα κάτι που τι καθιστά αρκετά αναποτελεσματική σε δεδομένα που εκτείνονται πάνω σε μη γραμμικούς χώρους.

Το 2000, ο μαθηματικός Joshua B. Tenenbaum και οι συνεργάτες του, άνοιξαν το δρόμο της εξερεύνησης μη γραμμικών τεχνικών μείωσης διαστάσεων, δημοσιεύοντας τον αλγόριθμο του Isomap (Isometric Feature Mapping) μια γενίκευση του αλγορίθμου MDS (multidimensional scaling). Στόχος της μεθόδου είναι, η ανίχνευση της εγγενούς γεωμετρίας των δεδομένων, προσπαθώντας να την υπολογίσει μέσω γεωδαισιακών διαδρομών που συνδέουν τα δεδομένα. Η χρησιμοποίηση της μεθόδου σε χρηματοοικονομικά στοιχεία εφαρμόζεται ραγδαία τα τελευταία χρόνια, καθώς η χρησιμοποίηση της μεθόδου για ανίχνευση της πολλαπλότητας, που υπάρχει σε μεγάλους διαστάσης χώρους δεδομένων, βοηθάει στην εύρεση χαμηλής διάστασης γεωμετρικών δομών χρηματοοικονομικών δεδομένων για την ομαδοποίηση τους ή και την ανάπτυξη συστημάτων πρόβλεψης. Η μέθοδος έχει εφαρμοστεί στην κατηγοριοποίηση μετοχών στο χώρο και στην κατάλληλη ομαδοποίηση τους με βάση τις αποδόσεις τους **(Liu et al., 2012)**, στην πρόβλεψη μελλοντικής αποτυχίας επιχείρησης (business failure), χρησιμοποιώντας την στην μείωση μεγάλου όγκου δεδομένων και την εισαγωγή τους σε τεχνικές επιβλεπόμενης μάθησης **(Lin, 2010)**.

Η πιο νέα μέθοδος και από τις τρείς που θα μελετηθούν στην εν λόγω διπλωματική είναι αυτή των απεικονίσεων διάχυσης (Diffusion Maps). Το 2006 οι μαθηματικοί Ronald R. Coifman, Stéphane Lafon δημοσίευσαν μία εργασία με τίτλο 'Diffusion Maps' ορίζοντας ένα πλαίσιο βασισμένο στη πιθανοτική διαδικασία διάχυσης για εύρεση σημαντικών γεωμετρικών δομών που διέπουν σύνολα δεδομένων. Το μεγαλύτερο πλεονέκτημα τις μεθόδου είναι η αποδοτικότητα της σε δεδομένα που τα διέπουν θόρυβος. Η δημοσίευση, χρησιμοποίησης της μεθόδου σε χρηματοοικονομικά στοιχεία, πραγματοποιήθηκε από τον Phoa το 2013. Στόχος της ήταν η εξαγωγή μοτίβων μέσω της οπτικοποίησης μετοχών και η εξαγωγή ποσοτικών μέτρων, για την διαφοροποίηση που διέπει ένα χαρτοφυλάκιο **(Phoa,2013)**.

ΜΕΡΟΣ ΠΡΩΤΟ

Τεχνικές Μείωσης Διαστάσεων

Κεφάλαιο 1

Ανάλυση κυρίων συνιστωσών (PCA : principal components analysis)

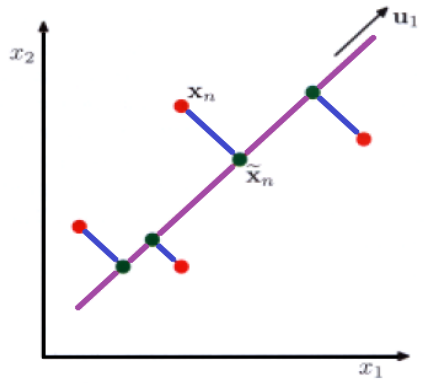
Η ανάλυση κυρίων συνιστωσών (PCA) είναι ένα πολύ χρήσιμο εργαλείο στην μοντέρνα ανάλυση δεδομένων, σε πολλούς επιστημονικούς κλάδους. Οι λόγοι που την καθιστούν τόσο διάσημη και χρήσιμη, είναι αυτοί της απλότητας και της μη χρησιμοποίησης παραμέτρων για την εξαγωγή χρήσιμης πληροφορίας, από μεγάλα σε όγκο δεδομένα. Με πολύ μικρό υπολογιστικό κόστος παρέχει έναν "χάρτη" για τη μείωση διαστάσεων, που στοχεύει στη αποκάλυψη σημαντικών δομών που μπορεί να βρίσκονται στα δεδομένα που αναλύονται **(Shlens, 2009).**

1.1 Η κεντρική ιδέα της PCA

Η κεντρική ιδέα της PCA **(Jolliffe, 2002)** είναι η μείωση διαστάσεων δεδομένων, τα οποία εκφράζονται μέσω μεγάλου αριθμού αλληλοσυσχετισμένων μεταβλητών, διατηρώντας όσο το δυνατό περισσότερο τη συνολική διασπορά που παρατηρείται στα δεδομένα. Αυτή η μείωση επιτυγχάνεται διαμορφώνοντας ένα σύνολο καινούργιων κύριων συνιστωσών (*principal components*), οι οποίες είναι ασυσχέτιστες μεταξύ τους και διατεταγμένες με τέτοιο τρόπο ώστε να διατηρείται το μεγαλύτερο ποσοστό της διασποράς των αρχικών. Όπως θα διαπιστωθεί αυτό επιτυγχάνεται υπολογιστικά μέσω της επίλυσης ενός προβλήματος ιδιοτιμών - ιδιοδιανυσμάτων, ενός συμμετρικού θετικά ημι-ορισμένου πίνακα.

Σε γενικότερα πλαίσια είναι αποδεκτό ότι η πιο πρώιμη αναφορά της σημερινής γνωστής τεχνικής PCA είχε δοθεί από τον **Pearson (1901)** και μεταγενέστερα από τον **Hoteling (1933)**. Οι δύο δημοσιεύσεις υιοθέτησαν διαφορετικές μεθοδολογίες για την προσέγγιση της μεθόδου. Ο πρώτος όρισε πως η μέθοδος μπορεί να περιγραφεί μέσω της ελαχιστοποίησης του μέσου κόστους προβολών, ορίζοντας τες ως τη μέση

τετραγωνική απόσταση μεταξύ σημείων και των αντίστοιχων προβολών τους. Αντίστοιχα ο δεύτερος, όρισε πως η λύση θα προέλθει από την ορθογώνια προβολή των δεδομένων, σε χαμηλότερης διάστασης γραμμικό υπόχωρο, γνωστός και ως κύριος υπόχωρος (*principal subspace*), στον οποίο οι διασπορές που διατηρούνται υπό προβολή στους άξονες μεγιστοποιούνται. Στο σχήμα 1.1 διαφαίνεται σχηματικά και οι δυο προσεγγίσεις. Η PCA αναζητεί έναν χώρο χαμηλότερης διάστασης γνωστός ως principal subspace, ο οποίος υποδεικνύεται από την μωβ γραμμή, στον οποίο οι ορθογώνιες προβολές, των σημείων στο χώρο (κόκκινα σημεία), σε αυτόν του υποχώρου επιτυγχάνεται η μεγιστοποίηση της διασποράς τους (πράσινα) **(Hoteling, 1933)**. Διαφορετικός ορισμός της PCA βασίζεται στην ελαχιστοποίηση του αθροίσματος των τετραγωνικών σφαλμάτων των προβολών των σημείων(μπλε γραμμές) **(Pearson, 1901)** :



Σχήμα 1.1 : Αναπαράσταση της μεθόδου PCA, όπως ορίστηκε από τον Hoteling, 1933 και από τον Pearson, 1901 **(Bishop, 2006).**

1.2 Η μαθηματική προσέγγιση της PCA (maximum variance formulation)

1.2.1 Πρόλογος

Σε αυτό το κεφάλαιο, θα εξετασθεί η μαθηματική προσέγγιση που έδωσε ο **Hotelling** στη μέθοδο της PCA και η εξαγωγή του αποτελέσματος που είναι ένας υπόχωρος μειωμένης διάστασης. Θεωρείται λοιπόν, ένα σύνολο δεδομένων από παρατηρήσεις $\{x_n\}$ όπου $n = 1, \dots, N$ με $x_n \in \mathbb{R}^D$ και $X = [x_1, x_2, \dots, x_n]$ ένας πίνακας με στήλες τα διανύσματα $\{x_n\}$. Στόχος της μεθόδου όπως αναφέρθηκε πριν είναι να προβληθούν τα δεδομένα σε ένα χώρο διαστάσεων $M < D$, έτσι ώστε να μεγιστοποιείται η διασπορά των δεδομένων υπό προβολή.

1.2.2 Ορισμός προβλήματος

Ξεκινώντας, για να απλοποιηθεί το πρόβλημα της επίλυσης της μεθόδου της PCA, θα εξεταστεί το πρόβλημα από τη θεώρηση της προβολής των δεδομένων σε 1- διάστασης χώρου $(M = 1)$.

Ορίζεται η διεύθυνση του χώρου, χρησιμοποιώντας ένα D -διάστασης διάνυσμα u_1 το οποίο ορίζεται εξίσου να έχει μέτρο ίσο με ένα $(u_1^T u_1 = 1)$ και σαν συνέπεια αυτού κάθε σημείο x_n προβάλλεται στο $u_1^T \cdot x_n$. 'Οπως θα δειχθεί μετέπειτα ο περιορισμός του διανύσματος u_1 να έχει σταθερό μέτρο είναι κρίσιμος για την εξαγωγή της λύσης. Ακολούθως ορίζεται η μέση τιμή των δεδομένων υπό προβολή ως $u_1^T \bar{x}$ όπου $\bar{x}$ η μέση τιμή του δείγματος του συνόλου των δεδομένων :

$$\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n \tag{1.1}$$

Ακολούθως και η διασπορά του των δεδομένων υπό προβολή ως :

$$\frac{1}{N} \sum_{n=1}^{N} \{u_1^T x_n - u_1^T \bar{x}\}^2 =$$

$$u_1^T\left[\frac{1}{N}\sum_{n=1}^{N}(x_n-\bar{x})(x_n-\bar{x})^T\right]u_1 =$$

$$u_1^T\mathbf{S}u_1 \tag{1.2}$$

Καθώς $\mathbf{S}$ είναι ο πίνακας διασποράς των δεδομένων ο οποίος ορίζεται ως εξής :

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\bar{x})(x_n-\bar{x})^T \tag{1.3}$$

Όπως αναφέρθηκε στο προηγούμενο κεφάλαιο στόχος είναι η μεγιστοποίηση της διασποράς των δεδομένων υπό προβολή. Συμπεριλαμβανομένου τον περιορισμό που τέθηκε, στόχος είναι η άμεση εύρεση του u_1. Επομένως το πρόβλημα βελτιστοποίησης που προκύπτει είναι το ακόλουθο :

$$\underset{\substack{u_1^T u_1=1 \\ u_1 \text{ unknown}}}{\text{maximize}} \quad u_1^T\,\mathbf{S}u_1 \tag{1.4}$$

1.2.3 Η επίλυση του προβλήματος βελτιστοποίησης

Από τη θεωρία για ακρότατα συναρτήσεων υπό συνθήκη, είναι γνωστό πως αν υπάρχει συνάρτηση $f(x,y)$ με ακρότατο το σημείο (x_0,y_0) υπό τη δέσμευση $g(x,y)=0$ τότε :

$$\nabla f(x_0,y_0) = \lambda\nabla g(x_0,y_0) \tag{1.5}$$

οπού λ είναι ο πολλαπλασιαστής Lagrange.

Επομένως η (1.4) μπορεί να μεταφραστεί ως, η εύρεση ακροτάτου της συνάρτησης $f(u_1)=u_1^TSu_1$ υπό τη δέσμευση $g(u_1)=u_1^Tu_1-1=0$. Σύμφωνα με την (1.5) τότε το πρόβλημα διαμορφώνεται ως :

$$\nabla\left(u_1^T\mathbf{S}u_1\right) = \lambda_1\nabla\left(u_1^Tu_1-1\right) \Rightarrow \tag{1.6}$$

$$\nabla\left[\left(u_1^T\mathbf{S}u_1\right)-\lambda_1\left(u_1^Tu_1-1\right)\right]=0 \tag{1.7}$$

Ορίζοντας, $$L(u_1,\lambda) = \left(u_1^T\mathbf{S}u_1\right)-\lambda_1\left(u_1^Tu_1-1\right) \tag{1.8}$$

Προκύπτει, $$\frac{\partial L}{\partial u_1} = 2\mathbf{S}u_1 - 2\lambda_1u_1 = 0 \Rightarrow \tag{1.9}$$

$$\mathbf{S}u_1 = \lambda_1u_1 \tag{1.10}$$

Η εξίσωση (1.10) υποδεικνύει πως η επίλυση του προβλήματος βελτιστοποίησης ανάγεται στην εύρεση του ιδιοδιανύσμάτος u_1 του πίνακα διασποράς **S.**

Πολλαπλασιάζοντας από αριστερά με u_1^T τη σχέση (1.10) και χρησιμοποιώντας το περιορισμό $u_1^T u_1 = 1$ παρατηρείται πως $u_1^T \mathbf{S} u_1 = \lambda_1$. Επομένως η μεγιστοποίηση της διασποράς των σημείων υπό προβολή επιτυγχάνεται όταν η ιδιοτιμή λ_1 είναι η μέγιστη με το ιδιοδιάνυσμα αυτής u_1 να θεωρείται ως η πρώτη κύρια συνιστώσα (1st principal component).

Εν τέλει μπορεί να οριστούν και επιπλέον κύριες συνιστώσες, με τα ιδιοδιανύσματα του πίνακα $\mathbf{S}$ να αντιστοιχούν σε αυτές, προσθέτοντας τον περιορισμό πως πρέπει να είναι ασυσχέτιστα μεταξύ τους. Η επιλογή τους γίνεται διατάσσοντας τις ιδιοτιμές που αντιστοιχούν σε αυτά κατά φθίνουσα σειρά, διατηρώντας όσο το δυνατόν περισσότερο τη συνολική διασπορά των δεδομένων. Η υπόθεση αυτή γίνεται, καθώς αποδεικνύεται πως συνολική διασπορά των κυρίων συνιστωσών είναι ίδια με αυτή των αρχικών μεταβλητών, πριν πραγματοποιηθεί κάποια μείωση διαστάσεων :

$$\text{var}\left(u_i^T x\right) = u_i^T \mathbf{S} u_i = \lambda_i \tag{1.11}$$

$$\sum_{i=1}^{D} \text{var}\left(u_i^T x\right) = \sum_{L=1}^{D} \lambda i = \text{Tr}(\mathbf{S}) = \text{var}(X) \tag{1.12}$$

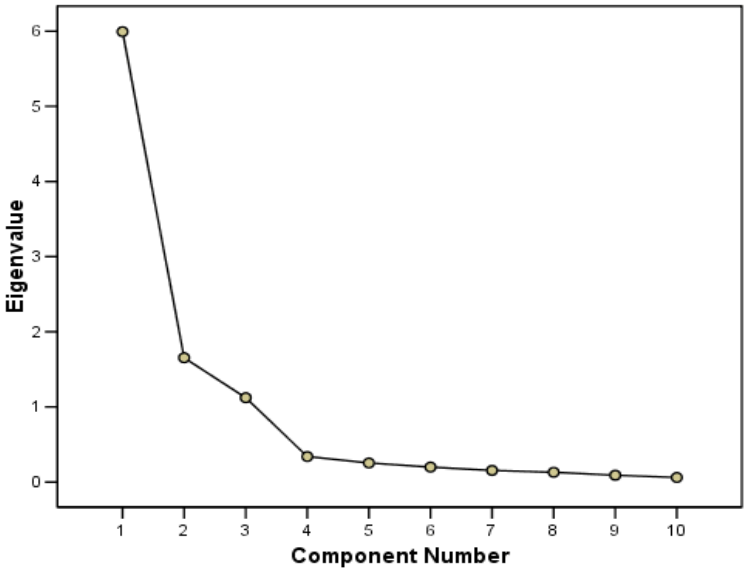
Σύμφωνα λοιπόν με τη παρατήρηση αυτή που έγινε, μπορεί να οριστεί η ποσότητα

$$\frac{\lambda_i}{\sum_{i=1}^{D} \lambda_i}$$

Η οποία δείχνει το ποσοστό στης συνολικής διασποράς που εξηγεί η i συνιστώσα. Είναι ευνόητο πως αν γίνει η επιλογή διατήρησης όλων των συνιστωσών τότε η συνολική διασπορά τους θα ισούται με αυτή των αρχικών μεταβλητών. Αντιθέτως αν γίνει επιλογή μερικών, το ποσοστό της συνολικής διασποράς μειώνεται. Για την επίτευξη μείωσης διαστάσεων η μερική επιλογή συνιστωσών είναι αναγκαία και μάλιστα πρέπει να γίνει με γνώμονα του ποσοστού των διασπορών που προσφέρουν. Με αυτό τον τρόπο μπορεί να αποφανθούν οι διαστάσεις, στις οποίες μπορούν να μειωθούν τα δεδομένα που βρίσκονται υπό επεξεργασία.

Μια επιπλέον μέθοδος επιλογής διαστάσεων, είναι το scree plot (διάγραμμα διαλογής) που παρουσιάστηκε από τον **Cattel (1966).** Το scree plot είναι απλώς μία παρουσίαση ενός διαγράμματος, όπου στον άξονα x παρουσιάζονται με φθίνουσα σειρά οι κύριες συνιστώσες με βάση τις ιδιοτιμές τους και στον άξονα y το άθροισμα των ιδιοτιμών. Αυτό που προτείνεται είναι να επιλεγούν οι συνιστώσες εκεί όπου το γράφημα τείνει να γίνει επίπεδο. Δηλαδή εκεί που αλλάζει κλίση.

Για παράδειγμα στο σχήμα 1.2 διαφαίνεται πως οι τέσσερίς διαστάσεις είναι αρκετές για να πραγματοποιηθεί μια ικανοποιητική μείωση διαστάσεων.



Σχήμα 1.2 : Αναπαράσταση ενός τυχαίου scree plot στο οποίο φαίνεται η μεταβολή του αθροίσματος των ιδιοτιμών σε σχέση με τον αριθμό των συνιστωσών, που όπως διαφαίνεται η επιλογή τεσσάρων διαστάσεων είναι ικανοποιητική **(www.ibm.com).**

1.3 Η αλγεβρική προσέγγιση της PCA

1.3.1 Πρόλογος

Πριν πραγματοποιηθεί η γεωμετρική προσέγγιση της PCA είναι σκόπιμο να γίνει αναφορά σε δύο έννοιες από τις οποίες θα αντληθούν τα αποτελέσματα :

- **Λόγος σήματος προς θόρυβο SNR (signal-to-noise ratio) :** $\mathbf{SNR} = \frac{\sigma^2_{\mathbf{signal}}}{\sigma^2_{\mathbf{noise}}}$
  Ο λόγος αυτός φανερώνει, το πόσο μεγάλη είναι η ακρίβεια που παρατηρείται στα δεδομένα που εξετάζονται. Μεγάλες τιμές φανερώνουν μεγάλη ακρίβεια στα δεδομένα, ενώ μικρές υποδεικνύουν μη-αξιόπιστη τη χρήση αυτών.
- **Πλεονασμός (redundancy) :**
  Η έννοια του πλεονασμού φανερώνει το κατά πόσο συσχετισμένες ή μη είναι δύο μεταβλητές και επομένως κατά πόσο μεγάλος η μικρός είναι ο πλεονασμός που υπάρχει στα δεδομένα του προβλήματος.

Σε αυτό το σημείο λοιπόν δίνεται ο στόχος της PCA που είναι η εύρεση ενός γραμμικού υποχώρου έτσι ώστε να επιτυγχάνεται:

i. Η μεγιστοποίηση του σήματος.
ii. Η ελαχιστοποίηση του πλεονασμού μεταξύ των μεταβλητών.

1.3.2 Ορισμός προβλήματος

Η λύση στον τελικό ορισμό του προβλήματος που τέθηκε πριν, καλείται να δώσει η χρησιμοποίηση του πίνακα διασποράς $C_X = \frac{1}{N} XX^T$, ενός πίνακα X διαστάσεων $d \times n$ με τις $\{x_n\}$ να αποτελούν παρατηρήσεις διαστάσεων d και με τις ακόλουθες ιδιότητες :

i. Ο C_X είναι τετραγωνικός.
ii. Οι όροι στη διαγώνιο του εκφράζουν την διασπορά των μεταβλητών.
iii. Οι όροι εκτός διαγώνιου εκφράζουν την συν διασπορά μεταξύ των μεταβλητών.

Κατά αυτόν τον τρόπο γίνεται αντιληπτό η αναδιαμόρφωση του στόχου της PCA ο οποίος είναι η εύρεση γραμμικού υποχώρου των δεδομένων έτσι ώστε να επιτυγχάνεται η

μεγιστοποίηση των διασπορών και οι συν διακυμάνσεις να είναι μηδενικές. Δηλαδή η εύρεση ενός ορθοκανονικού πίνακα P έτσι ώστε :

$$\mathrm{Y} = \mathrm{PX} \tag{1.13}$$

με τον πίνακα διασπορών $\mathrm{C_Y}$ να είναι διαγώνιος και οι γραμμές του P να αποτελούν τις κύριες συνιστώσες του X **(Shlens, 2009)**.

1.3.3 Η επίλυση του προβλήματος

Αναλύοντας τον πίνακα διασπορών του πίνακα Y προκύπτει το εξής αποτέλεσμα:

$$\begin{aligned}\mathrm{C_Y} &= \frac{1}{\mathrm{n}}\mathrm{YY^T}\\ &= \frac{1}{\mathrm{n}}\mathrm{PX(PX)^T}\\ &= \frac{1}{\mathrm{n}}\mathrm{PXX^TP^T}\\ &= \mathrm{PC_XP^T}\end{aligned} \tag{1.14}$$

Καθώς ο $\mathrm{C_X}$ συμμετρικός πίνακας τότε μπορεί να αναλυθεί με βάση την αποσύνθεση ιδιοτιμών του ως $\mathrm{C_X} = \mathrm{ADA^T}$ όπου D διαγώνιος πίνακας με στοιχεία τις ιδιοτιμές του $\mathrm{C_X}$ και A ορθογώνιος πίνακας που περιέχει στις στήλες του τα ιδιοδιανύσματα του πίνακα $\mathrm{C_X}$. Επομένως η εξίσωση (1.14) γίνεται :

$$\mathrm{C_Y} = \mathrm{P\,ADA^TP^T} \tag{1.15}$$

Ορίζοντας P να είναι ο πίνακας που περιέχει τα ιδιοδιανύσματα του πίνακα $\mathrm{C_X}$ στις γραμμές του, θα ισχύει $\mathrm{P^T} = \mathrm{A}$, επομένως :

$$\begin{aligned}\mathrm{C_Y} &= \mathrm{P\,P^TDPP^T}\\ &= \mathrm{P\,P^{-1}DPP^{-1}}\\ &= \mathrm{D}\end{aligned} \tag{1.16}$$

Εν κατακλείδι αυτό που αποδείχθηκε, είναι πως η επιλογή του P να έχει στις γραμμές του τα ιδιοδιανύσματα του πίνακα $\mathrm{C_X}$ ικανοποιεί τον στόχο της PCA.

1.4 Η αλγεβρική προσέγγιση της PCA μέσω της μεθόδου SVD

1.4.1 Πρόλογος

Μια επιπλέον προσέγγιση της μεθόδου της PCA μπορεί να δοθεί μέσω της αποσύνθεσης μοναδικών τιμών (Singular-Value-Decomposition-SVD). Σε πρώτη φάση θα αναλυθεί το θεωρητικό υπόβαθρο της SVD και κατ' επέκταση η συσχέτιση της με την PCA.

1.4.2 SVD (θεωρητικό υπόβαθρο)

Ορίζεται αυθαίρετος πίνακας X διαστάσεων $n \times d$ (σε αυτό το κεφάλαιο αντιστρέφονται οι δείκτες του πίνακα από $d \times n$ σε $n \times d$ καθώς θα βοηθήσει περισσότερο στην κατανόηση της μεθόδου) και ο $X^T X$ τάξης r, τετραγωνικός συμμετρικός $d \times d$ πινακας.Σε αυτό το σημείο μπορούν να καθοριστούν οι ποσότητες που μπορούν να αντληθούν από τους παραπάνω πίνακες **(Shlens, 2009)** :

- Το $\{\hat{v}_1, \hat{v}_2, \dots, \hat{v}_r\}$ είναι το σύνολο των ορθοκανονικών διανυσμάτων $m \times 1$ με τις εκάστοτε ιδιοτιμές $\{\lambda_1, \lambda_2 \dots, \lambda_r\}$ για τον $X^T X$ πίνακα, δηλαδή:

$$(X^T X)\hat{v}_i = \lambda_i \hat{v}_i \tag{1.17}$$

- Οι $\sigma_i = \sqrt{\lambda_i}$ είναι οι θετικές πραγματικές και ονομάζονται singular values.
- Το $\{\hat{u}_1, \hat{u}_2, \dots, \hat{u}_r\}$ είναι το σύνολο των ορθοκανονικών $n \times 1$ διανυσμάτων που ορίζονται ως $\hat{u}_i = \frac{1}{\sigma_i} X\hat{v}_i$.

Αποδεικνύεται ακόμα πως,

- $\hat{u}_i \hat{u}_j = \delta_{ij}$
- $\|X\hat{v}_i\| = \sigma_i$

Με βάση λοιπόν την εξίσωση :

$$\hat{u}_i = \frac{1}{\sigma_i} X\hat{v}_i \quad \Leftrightarrow \quad X\hat{v}_i = \hat{u}_i \sigma_i \tag{1.18}$$

εξάγονται πολύ χρήσιμες πληροφορίες :

Ο αυθαίρετος πίνακας X πολλαπλασιασμένος με ένα ιδιοδιάνυσμα του πίνακα $X^T X$ ισούται με ένα άλλο διάνυσμα πολλαπλασιασμένο με μια παράμετρο σ καθώς και τα σετ των ιδιοδιανυσμάτων $\{\hat{v}_i\},\{\hat{u}_i\}$ είναι ορθοκανονικά. με βάση λοιπόν την εξίσωση αυτή είναι αρκετά πιο σαφής η διαμόρφωση της σε εξίσωση πινάκων η οποία θα δώσει και τη ολική μορφή της SVD.

Ξεκινώντας λοιπόν με την κατασκευή του διαγώνιου πίνακα Σ όπου $\sigma_{\tilde{1}} \geq \sigma_{\tilde{2}} \geq \cdots \geq \sigma_{\tilde{r}}$ το ταξικά διατεταγμένο σύνολο των singular values.

$$\mathbf{\Sigma} \equiv \begin{bmatrix} \sigma_{\tilde{1}} & & & & & 0 \\ & \ddots & & & & \\ & & \sigma_{\tilde{r}} & & & \\ & & & 0 & & \\ & & & & \ddots & \\ 0 & & & & & 0 \end{bmatrix} \tag{1.19}$$

Κατά τον ίδιο τρόπο κατασκευάζονται και οι ορθογώνιοι πίνακες V και U, όπου :

$$V = \left[\hat{v}_{\tilde{1}}, \hat{v}_{\tilde{2}}, \dots, \hat{v}_{\tilde{d}}\right] \tag{1.20}$$

$$U = \left[\hat{u}_{\tilde{1}}, \hat{u}_{\tilde{2}}, \dots, \hat{u}_{\tilde{n}}\right] \tag{1.21}$$

Οι οποίοι έχουν επί πρόσθετα (m-r) και (n-r) ορθοκανονικά διανύσματα ώστε οι πίνακες V και U να γεμίσουν. Αναλύοντας λοιπόν σχηματικά την εξίσωση (1.18) το αποτέλεσμα που αντλείται βρίσκεται στο σχήμα 1.3, το οποίο την παρέχει γραφική αναπαράσταση για το πώς όλα τα κομμάτια ταιριάζουν μεταξύ τους για να σχηματίσουν την έκδοση των πινάκων της αποσύνθεσης μοναδικών τιμών.

$$XV = U\Sigma \tag{1.22}$$

όπου κάθε στήλη των V και U εκτελεί την εκδοχή της αποσύνθεσης. Επειδή ο V είναι ορθογώνιος, πολλαπλασιάζοντας και τις δύο πλευρές με $V^{-1} = V^T$ αναδεικνύεται η τελική μορφή της αποσύνθεσης.

$$X = U\Sigma V^T \tag{1.23}$$

Αυτή η αποσύνθεση είναι αρκετά ισχυρή. Η εξίσωση ορίζει ότι κάθε αυθαίρετος πίνακας X μπορεί να μετατραπεί σε έναν ορθογώνιο πίνακα, ένα διαγώνιο πίνακα και έναν άλλο ορθογώνιο πίνακα (ή μια περιστροφή, μια επέκταση και μια δεύτερη περιστροφή).

Η συγκεκριμένη αποσύνθεση περιέχει πολύ σημαντικές ιδιότητες που είναι σημαντικό να αναφερθούν για την πλήρη κατανόηση της. Η (1.22) μπορεί να διαμορφωθεί ως εξής:

$$X = U\Sigma V^T \quad \Leftrightarrow \tag{1.24}$$

$$U^T X = \Sigma V^T \quad \Leftrightarrow \tag{1.25}$$

$$U^T X = Z \tag{1.26}$$

Κατά αυτόν τον τρόπο παρατηρείται πως συγκρίνοντας την εξίσωση (1.26) με την εξίσωση (1.13) τα διανύσματα του πίνακα U έχουν τον ίδιο ρόλο όπως αυτά του πίνακα P.Στην ουσία δηλαδή ο U^T είναι μια αλλαγή βάσης από τον X στον Z.όπως και πριν αυτό που πραγματοποιείται είναι μια μετατροπή διανυσμάτων στηλών. Με λίγα λόγια, το γεγονός πως μια ορθοκανονική βάση U^T(P) μετατρέπει τα διανύσματα στηλών σημαίνει ότι η U^T αποτελεί μια βάση η οποία καλύπτει τις στήλες του X.

Με την αντίστοιχη συμμετρία του SVD που προκύπτει όπως θα δειχθεί ορίζεται μια αντίστοιχη ποσότητα αυτή του χώρου των γραμμών.

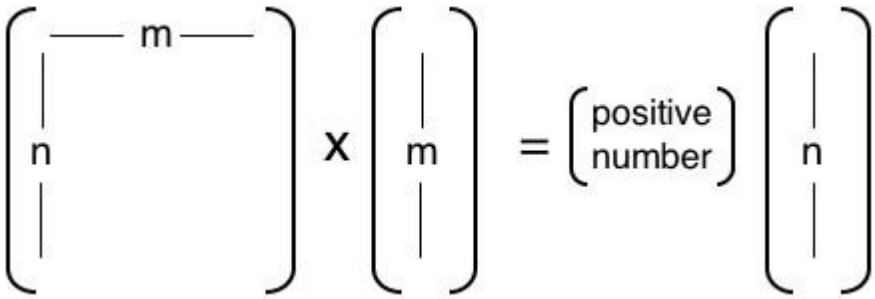
$$X = U\Sigma V^T \quad \Leftrightarrow \tag{1.27}$$

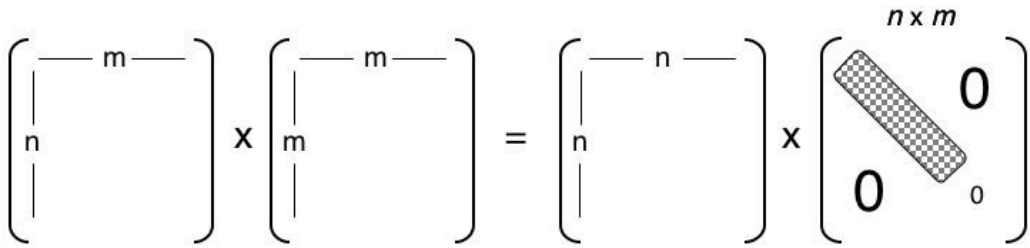
$$XV = U\Sigma \quad \Leftrightarrow \tag{1.28}$$

$$(XV)^T = (U\Sigma)^T \quad \Leftrightarrow \tag{1.29}$$

$$V^T X^T = Z \tag{1.30}$$

Ξανά οι γραμμές του V^T είναι μια ορθοκανονική βάση που μετατρέπει τον X^T στον Z. Λόγω της αναστροφής του X συμπεραίνεται πως ο V αποτελεί μια ορθοκανονική βάση η οποία καλύπτει τον χώρο γραμμών του X.





Σχήμα 1.3 : Αναπαράσταση του αποτελέσματος της μεθόδου SVD γραφικά εικονιζόμενή με πίνακες $XV = U\Sigma$ **(Shlens, 2009)**.

1.4.3 Η επίλυση της PCA μέσω της SVD

Χρησιμοποιώντας πάλι τον αρχικό πίνακα X διαστάσεων $\mathrm{d} \times \mathrm{n}$ και ορίζοντας τον πίνακα $\mathrm{Y} = \frac{1}{\sqrt{\mathrm{n}}}\mathrm{X}^{\mathrm{T}}$ διαστάσεων $\mathrm{n} \times \mathrm{d}$ όπου κάθε στήλη του έχει μηδενικό μέσο παρατηρείται πως :

$$\begin{aligned} \mathrm{Y}^{\mathrm{T}}\mathrm{Y} &= \left(\frac{1}{\sqrt{\mathrm{n}}}\mathrm{X}^{\mathrm{T}}\right)^{\mathrm{T}}\left(\frac{1}{\sqrt{\mathrm{n}}}\mathrm{X}^{\mathrm{T}}\right) \\ &= \frac{1}{\mathrm{n}}\mathrm{X}\mathrm{X}^{\mathrm{T}} \\ &= \mathrm{C}_{\mathrm{X}} \end{aligned} \tag{1.31}$$

Δοθέντος λοιπόν τον πίνακα Y και εφαρμόζοντας SVD σε αυτόν η αποσύνθεση του θα είναι η εξής :

$$Y = USV^T \tag{1.32}$$

Παρατηρείται πως :

$$Y^TY = VSU^TUSV^T = VS^2V^T \tag{1.33}$$

Δηλαδή τα διανύσματα που βρίσκονται στις στήλες του πίνακα V της μεθόδου SVD στην ανάλυση του Y είναι τα ιδιοδιανύσματα του πίνακα του Y^TY και κατε επ'έκταση τα ιδιοδιανύσματα του C_X τα οποία είναι και οι κύριες συνιστώσες της PCA.

Κατά αυτόν τον τρόπο παρατηρείται ακόμα, πως δοθέντος του πίνακα $Y = \frac{1}{\sqrt{n}}X$ διαφορετικών διαστάσεων $d \times n$

$$\begin{aligned} YY^T &= \left(\frac{1}{\sqrt{n}}X\right)\left(\frac{1}{\sqrt{n}}X\right)^T \\ &= \frac{1}{n}XX^T \\ &= C_X \end{aligned} \tag{1.34}$$

και εφαρμόζοντας σε αυτόν SVD η αποσύνθεση του θα ήταν η εξής :

$$Y_{d\times n} = U_{d\times d}S_{d\times n}V^T_{n\times n} \tag{1.35}$$

Παρατηρείται πως :

$$YY^T = USV^TVSU^T = US^2U^T \tag{1.36}$$

Δηλαδή τα ιδιοδιανύσματα που βρίσκονται στη στήλες του πίνακα U της μεθόδου SVD στην ανάλυση του Y είναι τα ιδιοδιανύσματα του πίνακα του YY^T και κατε επέκταση τα ιδιοδιανύσματα του C_X τα οποία είναι και οι κύριες συνιστώσες της PCA.

Συμπερασματικά δοθέντος πίνακα X διαστάσεων $d \times n$ και εφαρμόζοντας SVD είτε στον $Y = \frac{1}{\sqrt{n}}X$ είτε στον $Y = \frac{1}{\sqrt{n}}X^T$ με την κατάλληλη επιλογή των U ή V η λύση οδηγεί σε αυτήν της PCA.

1.5 Αλγόριθμος μεθόδου PCA

Ψευδοκώδικας

Είσοδος :

- Πίνακας X διαστάσεων $\mathrm{d} \times \mathrm{n}$ όπου στις γραμμές του βρίσκονται οι μεταβλητές.
- Το νούμερο των μειωμένων διαστάσεων p που θα κρατηθούν.

1. Αφαίρεσε το μέσο κάθε μεταβλητής από όλες της μετρήσεις για τη συγκεκριμένη μεταβλητή.
2. Υπολόγισε τον πίνακα διασποράς διαστάσεων $\mathrm{d} \times \mathrm{d}$, $\mathrm{C_X} = \frac{1}{\mathrm{n}-1}\mathrm{XX^T}$.
3. Βρες τα ορθοκανονικά ιδιοδιανύσματα $\{\hat{\mathrm{u}}_\mathrm{i}\}$ και τις ιδιοτιμές λ_i του πίνακα διασποράς $\mathrm{C_X}$.
4. Ταξινόμησε κατά φθίνουσα σειρά τα ιδιοδιανύσματα $\{\hat{\mathrm{u}}_\mathrm{i}\}$ με βάση τις ιδιοτιμές λ_i.
5. Υπολόγισε πίνακα $\mathrm{Y} = \mathrm{U^T X}$ όπου U πίνακας με στήλες τα $\{\hat{\mathrm{u}}_\mathrm{i}\}$.

Έξοδος :

Πίνακας Y μειωμένων διαστάσεων $\mathrm{p} \times \mathrm{n}$.

Ψευδοκώδικας PCA μέσω SVD

Είσοδος :

- Πίνακας X διαστάσεων $\mathrm{d} \times \mathrm{n}$ όπου στις γραμμές του βρίσκονται οι μεταβλητές.
- Το νούμερο των μειωμένων διαστάσεων p που θα κρατηθούν.

1. Αφαίρεσε το μέσο κάθε μεταβλητής από όλες της μετρήσεις για τη συγκεκριμένη μεταβλητή.
2. Υπολόγισε SVD του πίνακα $\mathrm{Y} = \frac{1}{\sqrt{\mathrm{n}-1}}\mathrm{X^T}$ (ή $\mathrm{Y} = \frac{1}{\sqrt{\mathrm{n}-1}}\mathrm{X}$).
3. Χρησιμοποίησε τον πίνακα V (ή πίνακα U) που εξάγεται από την SVD και υπολόγισε:

$$\mathrm{Y} = \mathrm{V^T X} \quad (\text{η'}\ \mathrm{Y} = \mathrm{U^T X})$$

Έξοδος :

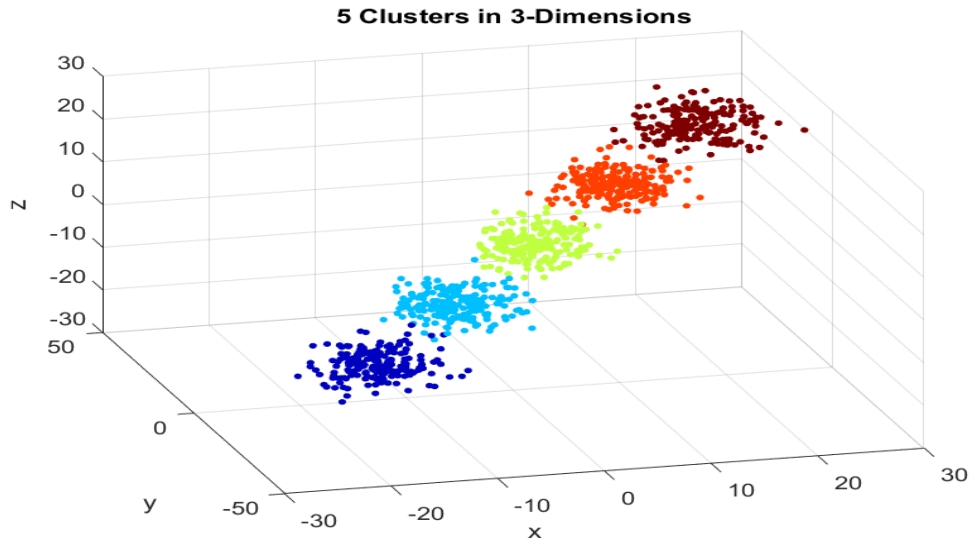
- Πίνακας Y μειωμένων διαστάσεων $\mathrm{p} \times \mathrm{n}$.

1.6 Εφαρμογή της μεθόδου PCA πάνω σε δεδομένα

1.6.1 Εισαγωγή

Για την πειραματική εφαρμογή της μεθόδου σε δεδομένα, παράχθηκαν πέντε σύννεφα-ομάδες (clusters) αποτελούμενα από $n = 200$ σημεία το καθένα, τα οποία σε μια κλίμακα μεσαίου μεγέθους σχηματίζουν μία ευθεία με μία μόνο παράμετρο, αυτής της θέσης των σημείων πάνω σε αυτό. Η παράμετρος αποκρυπτογραφείται από τα χρώματα που έχουν δοθεί στις ομάδες (σχήμα 1.4).

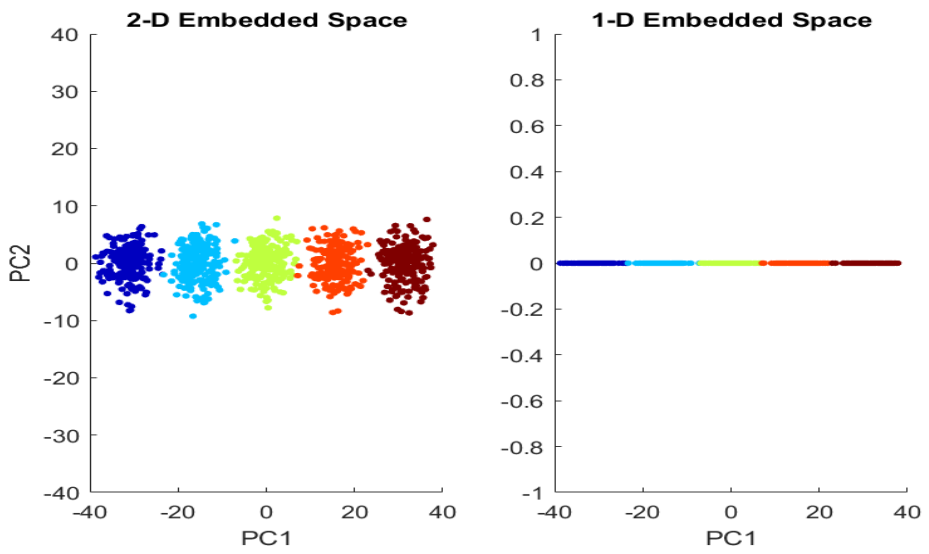
Λόγω της γραμμικότητας που περιγράφει η δομή των δεδομένων στον αρχικό χώρο η PCA αναμένεται να ανιχνεύσει τη δομή αυτή και να προβάλει τα δεδομένα, σε μικρότερης διάστασης χώρους με μεγάλη επιτυχία. Διατηρώντας την παράμετρο της θέσης των σημείων (διατήρηση της σειράς των χρωμάτων) με τον καλύτερο δυνατό τρόπο. Επίσης εφαρμόζοντας το scree test πάνω στα δεδομένα αυτά αναμένεται η βέλτιστη επιλογή μείωσης διαστάσεων να είναι αυτή των δύο.



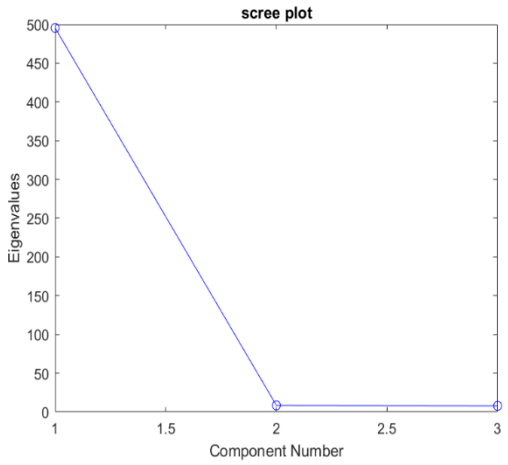
Σχήμα 1.4 : Απεικόνιση N=1000 σημείων σε τρισδιάστατο χώρο, σχήματος ευθείας, αποτελούμενο από 5 ομάδες (clusters) συνολικού αριθμού n=200 τυχαίων σημείων το καθένα αντίστοιχα.

1.6.2 Αποτελέσματα

Στο σχήμα 1.5 παρουσιάζονται τα αποτελέσματα της μεθόδου PCA, προβάλλοντας τα δεδομένα σε χώρους δύο διαστάσεων και μίας αντίστοιχα. Διαφαίνεται πως η γεωμετρία των δεδομένων στον αρχικό χώρο έχει διατηρηθεί με ελάχιστες ως και μηδαμινές αλλοιώσεις στον μειωμένο δύων διαστάσεων χώρο. Επίσης η παράμετρος της θέσης των σημείων πάνω και στους δύο αυτούς τους χώρους έχει διατηρηθεί με τον καλύτερο δυνατό τρόπο. Στο σχήμα 1.6 παρουσιάζεται το αποτέλεσμα του scree plot στο οποίο διαφαίνεται ξεκάθαρα πως η επιλογή των δύο διαστάσεων είναι η καλύτερη δυνατή.



Σχήμα 1.5 : Απεικόνιση 1000 τυχαίων σημείων (που εκτείνονται σε γραμμικό χώρο), σε χώρους μειωμένων διαστάσεων δύο και μίας αντίστοιχα, μετά από εφαρμογή τη PCA.



Σχήμα 1.6 : Απεικόνιση Scree plot για τα τυχαία δεδομένα 1000 γραμμικών σημείων.

Κεφάλαιο 2

Πολυδιάστατη Κλιμάκωση (MDS : Multidimensional Scaling)

Μια εναλλακτική προοπτική στο πρόβλημα της μείωσης διαστάσεων μεγάλου όγκου δεδομένων προσφέρεται μέσω της γνωστής μεθόδου MDS (multidimensionality scaling). Ένα μεγάλο πλεονέκτημα εξίσου και αυτής της μεθόδου είναι η μη-χρησιμοποίηση παραμέτρων για την εξαγωγή χρήσιμων πληροφορίων. Δοθέντος, μόνο ενός πίνακα ομοιοτήτων ή ανομοιοτήτων, μεταξύ των αντικειμένων που μελετιούνται, παρέχει έναν ''χάρτη'' που υποδεικνύει σημαντικές πληροφορίες μεταξύ των δεδομένων όπως αυτής της ομοιότητας τους. Όπως θα δειχτεί και μετέπειτα η MDS είναι μια γραμμική μέθοδος μείωσης διαστάσεων της οποίας η λύση στην απλή της περίπτωση συνδέεται ακράδαντα με αυτή της PCA.

2.1 Η κεντρική ιδέα της MDS

Η κεντρική ιδέα της MDS είναι να αντιστοιχίσει έναν χώρο μεγάλων διαστάσεων σε έναν μικρότερων διαστάσεων, διατηρώντας όσο το δυνατόν περισσότερο τις κατά ζεύγη αποστάσεις των σημείων που βρίσκονται στους δύο χώρους. Στη περίπτωση της κλασσικής πολυδιάστατης κλιμάκωσης (cMDS) οι κατά ζεύγη αποστάσεις που προσπαθούν να διατηρηθούν είναι οι ευκλείδειες.

Το κίνητρο για την χρησιμοποίηση της είναι να εικονιστούν τα δεδομένα σε δύο ή τρεις διαστάσεις, έτσι ώστε ένας ερευνητής να κατανοήσει τη δομή τους και να εξάγει χρήσιμα μοτίβα και συμπεράσματα **(Groenen, 2013)**.

2.2 Η μαθηματική προσέγγιση της cMDS

2.2.1 Πρόλογος

Ξεκινώντας, ορίζεται ένας πίνακας D διαστάσεων $n \times n$ ο οποίος καλείται πίνακας αποστάσεων ή πίνακας συνάφειας όπου $d_{ij} = |x_i - x_j|^2$ αν και μόνο αν ικανοποιεί τις παρακάτω συνθήκες :

i. $d_{ii} = 0$
ii. $d_{ij} > 0 \quad \forall i, j \ : i \neq j$
iii. $d_{ij} = d_{ji}$
iv. $d_{ij} + d_{jk} \geq d_{ik}$ (τριγωνική ανισότητα)

Δοθέντος λοιπόν ενός πίνακα αποστάσεων $D^{(x)}$, $n \times n$ διαστάσεων, η MDS προσπαθεί να βρει n σημεία $\{y_i\}$ διαστάσεων d $\left(y_i \in \mathbb{R}^d\right)$ όπου $i = 1,2, \dots, n$ έτσι ώστε $D^{(Y)} \cong D^{(X)}$ όπου $D^{(Y)}$ πίνακας αποστάσεων με στοιχεία τις ευκλείδειες αποστάσεις των σημείων y_i, y_j όπως επισημάνθηκε πριν.

2.2.2 Ορισμός προβλήματος

Το πρόβλημα της MDS θα εξεταστεί μελετώντας την μετρική πολυδιάστατη κλιμάκωση η οποία προσπαθεί να ανάξει το πρόβλημα στην εξής βάση **(Cox & Cox, 2001)**:

$$\underset{Y \text{ unknown}}{\text{minimize}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(d_{ij}^{(X)} - d_{ij}^{(Y)}\right)^2 \tag{2.1}$$

Όπου $d_{ij}^{(X)} = \|x_i - x_j\|^2$ και $d_{ij}^{(Y)} = \|y_i - y_j\|^2$ αντίστοιχα.

Χρήσιμη πληροφορία στην επίλυση του προβλήματος βελτιστοποίησης είναι η χρησιμοποίηση της ιδιότητας της νόρμας του Forbenius για πίνακα $A \in C^{mxn}$ που δίνεται από την εξής εξίσωση **(Mayer, 2000)** :

$$\|A\|_F^2 = \sum_{i,j} |\alpha_{ij}|^2 = \text{trace}\left(A^T A\right) \tag{2.2}$$

Με βάση λοιπόν την (2.2) το πρόβλημα μεταβάλλεται ως :

$$\underset{Y \text{ unknown}}{\text{minimize}} \ \left|D^{(X)} - D^{(Y)}\right|_F^2 \tag{2.3}$$

<u>Ισχυρισμός</u>:

Ο πίνακας αποστάσεων D που εξάγεται μέσω ενός πίνακα $X = [x_1, x_2, \dots, x_n]$, με στήλες τα διανύσματα $\{x_i\}$ όπου $x_i \in \mathbb{R}^d$, μετατρέπεται σε πίνακα πυρήνα εσωτερικού γινομένου $K = X^T X$ ως εξής : $K = -\frac{1}{2} HDH$ όπου $H = I - \frac{1}{n} ee^T$ με e διάνυσμα – στήλη με όλα τα στοιχεία του ίσα με 1 **(Cox & Cox, 2001).**

<u>Απόδειξη ισχυρισμού</u>:

Έστω λοιπόν σημεία $\{x_1, x_2, \dots, x_n\}$ σε ένα ευκλείδιο χώρο d διαστάσεων έτσι ώστε για $x_i \in \mathbb{R}^d$ ορίζονται τα σημεία $x_i = (x_{i1}, x_{i2}, \dots, x_{id})^T$ με $i = 1, \dots, n$ τότε η ευκλείδεια απόσταση που ορίζεται μεταξύ i -οστού και j -οστού σημείου είναι η

$$d_{ij}^2 = (x_\iota - x_j)^T (x_i - \chi_j) = x_i^T x_i + x_j^T x_j - 2x_i^T x_j \tag{2.4}$$

Ακολούθως ορίζεται και ο πίνακας εσωτερικού γινομένου $[K]_{ij} = k_{ij} = x_i^T x_j$.

Για να ξεπεραστεί το πρόβλημα του απροσδιορίστου της λύσης θεωρείται η κεντροποίηση (centering) των σημείων στην αρχή των αξόνων και κατά αυτόν τον τρόπο :

$$\sum_{i=1}^{n} x_{ik} = 0, \quad \forall \ k = 1, \dots, d \tag{2.5}$$

Καθώς επίσης προκύπτει και το ακόλουθο σύστημα εξισώσεων :

$$\begin{cases} \frac{1}{n}\sum_{i=1}^{n} dij^2 = \frac{1}{n}\left[\sum_{i=1}^{n} x_i^T x_i + \sum_{i=1}^{n} x_j^T x_j - 2\sum_{i=1}^{n} x_i^T x_j\right] = \frac{1}{n}\sum_{i=1}^{n} x_i^T x_i + x_j^T x_j \\ \frac{1}{n}\sum_{j=1}^{n} dij^2 = \frac{1}{n}\left[\sum_{j=1}^{n} x_i^T x_i + \sum_{j=1}^{n} x_j^T x_j - 2\sum_{j=1}^{n} x_i^T x_j\right] = \frac{1}{n}\sum_{i=1}^{n} x_j^T x_j + x_i^T x_i \\ \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} dij^2 = \frac{1}{n^2}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} x_i^T x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} x_j^T x_j - 2\sum_{i=1}^{n}\sum_{j=1}^{n} x_i^T x_j\right] = \frac{2}{n}\sum_{i=1}^{n} x_i^T x_i \end{cases} \tag{2.6}$$

Όπου, σύμφωνα με την εξίσωση (2.5) :

$$\sum_{i=1}^{n} x_i^T x_j = \sum_{i=1}^{n}\sum_{\kappa=1}^{d} x_{ik} x_{jk} = \sum_{k=1}^{d} x_{jk} \sum_{i=1}^{n} x_{ik} = 0 \tag{2.7}$$

Επομένως το σύστημα (2.6) διαμορφώνεται ως :

$$\begin{cases} x_j^T x_j = \frac{1}{n}\sum_{i=1}^{n} dij^2 - \frac{1}{n}\sum_{i=1}^{n} x_i^T x_i \\ x_i^T x_i = \frac{1}{n}\sum_{j=1}^{n} dij^2 - \frac{1}{n}\sum_{i=1}^{n} x_j^T x_j \\ \sum_{i=1}^{n} x_i^T x_i = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n} dij^2 \end{cases} \Leftrightarrow \tag{2.8}$$

$$\begin{cases} x_j^T x_j = \frac{1}{n}\sum_{i=1}^{n} dij^2 - \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} dij^2 \\ x_i^T x_i = \frac{1}{n}\sum_{j=1}^{n} dij^2 - \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} dij^2 \end{cases} \tag{2.9}$$

Χρησιμοποιώντας την εξίσωση (2.4) και το σύστημα (2.9) :

$$d_{ij}^2 = x_i^T x_i + x_j^T x_j - 2x_i x_j^T \Leftrightarrow \tag{2.10}$$

$$x_i x_j^T = -\frac{1}{2}\left[d_{ij}^2 - \frac{1}{n}\sum_{i=1}^{n} dij^2 - \frac{1}{n}\sum_{j=1}^{n} dij^2 + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} dij^2\right] \Leftrightarrow \tag{2.11}$$

$X^T X = -\frac{1}{2} HDH$ όπου $H = I - \frac{1}{n} ee^T$ με e διάνυσμα στήλη με όλα τα στοιχεία του ίσα με 1.

<u>Τέλος απόδειξης ισχυρισμού</u>

Επομένως το πρόβλημα της βελτιστοποίησης (2.3) παίρνει την τελική του μορφή με βάση τον προηγούμενο ισχυρισμό και ανάγεται στο εξής πρόβλημα :

$$\underset{\text{Y unknown}}{\text{minimize}} \ \left|K^{(X)} - K^{(Y)}\right|_F^2 \tag{2.12}$$

2.2.3 Η επίλυση του προβλήματος βελτιστοποίησης

Η διαδικασία για την εύρεση του κατάλληλου πίνακα Y είναι η ακόλουθη :

$$\underset{\text{Y unknown}}{\text{minimize}} \ \left|K^{(X)} - K^{(Y)}\right|_F^2 \tag{2.13}$$

$$\rightarrow \underset{\text{Y unknown}}{\text{minimize}} \ \left|X^T X - Y^T Y\right|_F^2 \tag{2.14}$$

Χχρησιμοποιώντας τη μέθοδο της SVD για πίνακες X, Y, $X^T X = V\Lambda V^T$ και $Y^T Y = Q\widehat{\Lambda}Q^T$ με V, Q να αποτελούν πίνακες με στήλες τα ορθοκανονικά ιδιοδιανύσματα του πίνακα $X^T X$ και $Y^T Y$ αντίστοιχα καθώς και $\Lambda, \widehat{\Lambda}$ πίνακες διαγώνιοι με στοιχεία τα ιδιοδιανύσματα των πινάκων $X^T X$ και $Y^T Y$ αντίστοιχα.

$$\xrightarrow{(2.14)} \underset{\widehat{\Lambda},Q \text{ unknown}}{\text{minimize}} \ \left|V\Lambda V^T - Q\widehat{\Lambda}Q^T\right|_F^2 \tag{2.15}$$

$$\xrightarrow{(2.2)} \underset{\widehat{\Lambda},Q \text{ unknown}}{\text{minimize}} \ \text{Trace}\left[V\Lambda V^T - Q\widehat{\Lambda}Q^T\right]^2 \tag{2.16}$$

$$\longrightarrow \underset{\widehat{\Lambda},Q \text{ unknown}}{\text{minimize}} \ \text{Trace}\left[\Lambda - V^T Q\widehat{\Lambda}Q^T V\right]^2 \tag{2.17}$$

Θέτοντας

$$V^T Q = G \tag{2.18}$$

$$\rightarrow \underset{\widehat{\Lambda},G \text{ unknown}}{\text{minimize}} \ \text{Trace}\left[\Lambda - G\widehat{\Lambda}G^T\right]^2 \tag{2.19}$$

Φιξάροντας το $\widehat{\Lambda}$ Η παράσταση ελαχιστοποιείται για $G = I$.

$$\rightarrow \underset{\widehat{\Lambda},\text{unknown}}{\text{minimize}} \ \text{Trace}\left[\Lambda - \widehat{\Lambda}\right]^2 \tag{2.20}$$

Επομένως για να γίνει η παράσταση αυτή όσο ελάχιστη γίνεται θα πρέπει $\Lambda \cong \widehat{\Lambda}$ όσο το δυνατόν περισσότερο. Αυτό επιτυγχάνεται διαλέγοντας τα στοιχεία του πίνακα $\widehat{\Lambda}$ να είναι τα $p \leq d$ μεγαλύτερα στοιχεία του πίνακα Λ. Και εν κατακλείδι η εύρεση του πίνακα Y δίνεται μέσω της $V^T Q = G$ και του γεγονός πώς $G = I$, επομένως:

$$V = Q \tag{2.21}$$

Καθώς και από την αποσύνθεση που έγινε προηγουμένως:

$$Y^T Y = Q\widehat{\Lambda}Q^T \Leftrightarrow$$

$$Y^T Y = Q\widehat{\Lambda}^{\frac{1}{2}}\widehat{\Lambda}^{\frac{1}{2}}Q^T \Leftrightarrow$$

$$Y^T Y = (\widehat{\Lambda}^{\frac{1}{2}}Q^T)^T\, \widehat{\Lambda}^{\frac{1}{2}}Q^T \Leftrightarrow$$

$$Y = \widehat{\Lambda}^{\frac{1}{2}}Q^T \Leftrightarrow$$

$$Y = \widehat{\Lambda}^{\frac{1}{2}} V^T \tag{2.22}$$

Επομένως, η επίλυση του προβλήματος βελτιστοποίησης και κατ' επέκταση η λύση της MDS βρίσκεται στην εύρεση του πίνακα Y όπου V τα ιδιοδιανύσματα του πίνακα $X^T X$ που αντιστοιχούν στις p μεγαλύτερες ιδιοτιμές του $X^T X$ **(Ghodsi, 2006)**.

2.2.4 cMDS και PCA

Άξιο παρατήρησης είναι το γεγονός πως αν και μόνο αν οι αποστάσεις που χρησιμοποιούνται στην μέθοδο της MDS είναι ευκλείδειες τότε και μόνο τότε η λύση που προκύπτει είναι ακριβώς η ίδια με αυτήν της PCA.

Η λύση που δίνεται μετά την εφαρμογή της μεθόδου της PCA σε έναν πίνακα X διαστάσεων $m \times n$, όπως αποδείχθηκε στο προηγούμενο κεφάλαιο, είναι η $Y = U^T X$ όπου U πίνακας με στήλες τα ιδιοδιανύσματα του πίνακα διασποράς C_x.

Εφαρμόζοντας SVD στον πίνακα X :

$$X_{m \times n} = U_{m \times m} S_{m \times n} V_{n \times n}^T \tag{2.23}$$

Και παρατηρώντας ότι :

$$X^T X = V S U^T U S V^T = V S^2 V^T = V \Lambda V^T \tag{2.24}$$

Εξάγεται το συμπέρασμα πώς ο πίνακας V έχει στις στήλες του τα ιδιοδιανύσματα του πίνακα $X^T X$ και ο πίνακας $S^2 = \Lambda$ τα ιδιοδιανύσματα του πίνακα $X^T X$ στη διαγώνιο του.

Με βάση την (2.23) η λύση της PCA μπορεί να διαμορφωθεί ως :

$$Y = U^T X = U^T U S V^T = S V^T = \Lambda^{\frac{1}{2}} V^T \tag{2.25}$$

Επομένως, η λύση της PCA οδηγεί σε ακριβώς ίδια λύση με αυτήν της cMDS.

2.3 Αλγόριθμος μεθόδου cMDS

Ψευδοκώδικας

Είσοδος :

- Πίνακας X διαστάσεων $d \times n$ όπου στις γραμμές του βρίσκονται οι μεταβλητές.
- Το νούμερο των μειωμένων διαστάσεων p που θα κρατηθούν.

1. Κατασκεύασε τον πίνακα των τετράγωνων των αποστάσεων D^X.
2. Υπολόγισε τον κεντροποιημένο πίνακα $H = I - \frac{1}{n}ee^T$.
3. Υπολόγισε τον πίνακα εσωτερικών γινομένων $K = -\frac{1}{2}HDH$.
4. Υπολόγισε τα ιδιοδιανύσματα $\{v_i\}$ και τις ιδιοτιμές $\{\lambda i\}$ του πίνακα K.
5. Ταξινόμησε κατά φθίνουσα σειρά τα ιδιοδιανύσματα με βάση της ιδιοτιμές.
6. Υπολόγισε τον πίνακα $Y = \Lambda^{\frac{1}{2}}V^T$ όπου Λ πίνακας διαγώνιος με στοιχεία $\{\lambda i\}$ τα ιδιοδιανύσματα του K και V πίνακας με στήλες $\{v_i\}$ τα ιδιοδιανύσματα του K.

Έξοδος :

- Πίνακας Y μειωμένων διαστάσεων $p \times n$.

Κεφάλαιο 3

Απεικόνιση γεωμετρικών συνιστωσών (ISOMAP)

Ο αλγόριθμος Isomap **(Tenenbaum et al., 2000)** αποτελεί μια μη γραμμική μέθοδο μείωσης διαστάσεων που στόχο έχει την εύρεση μη γραμμικών βαθμών ελευθερίας που βρίσκονται σε πολλές και πολύπλοκες φυσικές παρατηρήσεις, εν αντίθεση με τις μεθόδους της PCA και cMDS. Έτσι λοιπόν ο στόχος είναι η διατήρηση όσο το δυνατόν περισσότερο της εγγενής γεωμετρίας που διέπουν τα δεδομένα που εκτείνονται στο αρχικό χώρο (original space).

3.1 Η κεντρική ιδέα του Isomap

Η κεντρική ιδέα του Isomap είναι η εφαρμογή της μεθόδου της cMDS που αναλύθηκε προηγουμένως, σε έναν διαφορετικό κατά τρόπο χώρο που δεν εκφράζεται από τις ευκλείδειες αποστάσεις των σημείων στον αρχικό χώρο. Ο χώρος που θα μελετηθεί και θα εφαρμοστεί μετά πάνω σε αυτόν η μέθοδος της cMDS, είναι αυτός των γεωδαισιακών αποστάσεων.

Οι γεωδαισιακές αποστάσεις αναπαριστούν τα συντομότερα μονοπάτια που ενώνουν δύο σημεία στο χώρο κατά μήκος της γεωμετρίας που εκφράζει τα δεδομένα τα οποία θεωρούνται ότι εκτείνονται πάνω σε μία χαμηλής διάστασης πολλαπλότητα (manifold). Αυτή η απόσταση μπορεί να υπολογιστεί κάνοντας πολλά στο πλήθος μικρά βήματα ανάμεσα σε γειτονικά σημεία των δεδομένων.

Εφαρμόζοντας έπειτα τη μέθοδο της cMDS στις γεωδαισιακές αποστάσεις και όχι στις ευκλείδειες, στόχος είναι η διατήρηση όσο το δυνατόν περισσότερο των γεωδαισιακών μονοπατιών τα οποία εκφράζονται πλέον μέσω της ευκλείδειας μετρικής στο χώρο μειωμένων διαστάσεων.

3.2 Η υπολογιστική προσέγγιση της μεθόδου Isomap

Η υπολογιστική προσέγγιση της μεθόδου θα αναδείξει με τον καλύτερο τρόπο την λειτουργία της μεθόδου καθώς όπως αναφέρθηκε η μέθοδος τροποποιεί κατά κάποιο τρόπο την μέθοδο της MDS αντικαθιστώντας τον πίνακα ευκλείδειων αποστάσεων με τον πίνακα των γεωδαισιακών. Επομένως δεν υπάρχει κάποιο μαθηματικό υπόβαθρο να αναλυθεί.

Ο αλγόριθμος του Isomap δέχεται τις αποστάσεις $d_X(i,j)$ (ευκλείδειες κατά προτίμηση) μεταξύ όλων των σημείων N που βρίσκονται στον αρχικό χώρο μελέτης. Στόχος του είναι η εξαγωγή διανυσμάτων συντεταγμένων για το κάθε σημείο σε έναν d-διαστάσεων ευκλείδειο χώρο, τα οποία είναι ικανά να αντιπροσωπεύσουν επάξια την γεωμετρική δομή του αρχικού. Ακολουθώντας τα εξής βήματα ο αλγόριθμος Isomap έχει ως εξής :

<u>Πρώτο βήμα</u>:

Αρχικά καθορίζονται ποια σημεία στο χώρο που μελετάται θεωρούνται γειτονικά με βάση τις αποστάσεις $d_X(i,j)$ που έχουν οριστεί σαν είσοδο στον αλγόριθμο. Αυτό πραγματοποιείται επιλέγοντας ένα από τους δύο ακόλουθους τρόπους:

1. Ορίζεται ακτίνα ε και κέντρο κάθε σημείο i του χώρου. Όσα σημεία j περιέχονται μέσα στην ακτίνα που έχει οριστεί θεωρούνται γειτονικά του i.
2. Ορίζεται παράμετρος k που εκφράζει το πλήθος των πλησιέστερων γειτονικών σημείων j κάθε σημείου στο χώρο, i.

Κατά αυτόν τον τρόπο οι αποστάσεις του γράφου τροποποιούνται ως εξής :

$d_G(i,j) = d_X(i,j)$, αν τα σημεία i, j είναι συνδεδεμένα με μία ακμή

$d_G(i,j) = \infty$, αλλιώς

<u>Δεύτερο βήμα</u>:

Με βάση τον αλγόριθμο του Floyd **(Cormen, 1990)** ο οποίος στοχεύει στην ανεύρεση σύντομων μονοπατιών σε ένα γράφημα, εκτιμιούνται οι γεωδαισιακές αποστάσεις μεταξύ των σημείων. Ο αλγόριθμος αυτός στη πράξη για κάθε μία από τις τιμές τις παραμέτρου $k = 1, \dots, N$ διαμορφώνει τις καταχωρήσεις του $d_G(i,j)$ ως εξής :

$$d_G(i,j) = \min\{d_G(i,j), d_G(i,k) + d_G(k,j)$$

Κατά αυτόν τον τρόπο όταν τελειώσουν όλες οι επιλογές του k και οι καταχωρήσεις στον στοιχείων,ο πίνακας $D_G = \{d_G(i,j)\}$ θα περιέχει τα συντομότερα μονοπάτια όλων των σημείων στο χώρο.

<u>Τρίτο βήμα</u>:

Εφαρμόζοντας την cMDS στον πίνακα D_G κατασκευάζεται μία εμφύτευση στον ευκλείδειο χώρο και τα σημεία σε αυτό αντιπροσωπεύουν με τον καλύτερο τρόπο την γεωμετρία του αρχικού. Τα διανύσματα y_i επιλέγονται ώστε να ελαχιστοποιείται το ακόλουθο πρόβλημα:

$$\|\tau(D_G) - \tau(D_Y)\|_F$$

όπου D_Y ο πίνακας ευκλείδειων αποστάσεων $d_Y(i,j) = \|y_i - y_j\|$, η νόρμα F όπως έχει οριστεί στο κεφάλαιο 2 και εν τέλει τ ο τελεστής που μετατρέπει της αποστάσεις σε γινόμενα.

Το πρόβλημα ελαχιστοποίησης κατά κύριο λόγο είναι ακριβώς το ίδιο με αυτό που χρησιμοποιήθηκε στην cMDS μόνο με τη διαφορά πως αντί του D_X χρησιμοποιείται ο πίνακας γεωδαισιακών αποστάσεων μεταξύ των σημείων, D_G.

Επομένως όπως και στην cMDS στόχος είναι η ανεύρεση των ιδιοδιανύσμάτων u_i και των ιδιοτιμών λ_i του πίνακα $\tau(D_G)$ οι οποίες θα δώσουν τις νέες συντεταγμένες στα σημεία του νέου, μειωμένης διάστασης χώρου κατά p. Επιλέγοντας κατά φθίνουσα σειρά τις ιδιοτιμές και αντίστοιχα τα ιδιοδιανύσματα τους δημιουργούνται ακολούθως οι εξής συντεταγμένες των σημείων :

$$y_i = \begin{bmatrix} \sqrt{\lambda_1} u_1[i] \\ \sqrt{\lambda_2} u_2[i] \\ \vdots \\ \sqrt{\lambda_p} u_p[i] \end{bmatrix} \tag{3.1}$$

3.3 Επιλογή της παραμέτρου k

Η επιλογή της παραμέτρου k η οποία εκφράζει το πλήθος των γειτονικών σημείων που ορίζεται σαν είσοδος από το χρήστη στον αλγόριθμο παίζει σπουδαίο ρόλο στην ανάδειξη της γεωμετρίας των δεδομένων σε χαμηλότερης διάστασης χώρου. Η μη σωστή επιλογή της οδηγεί σε παραπλανητικά αποτελέσματα ανεύρεση της πολλαπλότητας του αρχικού χώρου.

Μία μέθοδος εκτίμησης της μπορεί να πραγματοποιηθεί από την εύρεση ενός μέτρου που θα υποδεικνύει την ''ποιότητα'' της απεικόνισης από τον αρχικό χώρο στον χώρο μειωμένων διαστάσεων. Δηλαδή πόσο καλά αντιπροσωπεύεται η γεωμετρική δομή των δεδομένων στο αρχικό χώρο από των αυτώ μειωμένων διαστάσεων.

Ορίζεται λοιπόν το μέτρο διακύμανσης υπολοίπων (residual variance) ως εξής :

$$1 - \rho_{\widehat{D}_X(k)D_Y} \tag{3.2}$$

Όπου ο $\rho_{\widehat{D}_X(k)D_Y}$ ο πλυθησμιακός συντελεστής γραμμικής συσχέτισης του Pearson μεταξύ όλων των στοιχείων του πίνακα $\widehat{D}_X$ και του D_Y. Με τους ιδίους να εκφράζουν τις γεωδαισιακές αποστάσεις που ορίζονται από τα συντομότερα μονοπάτια του αρχικού χώρου X τα οποία είναι συναρτήσει της παραμέτρου k και τις ευκλείδιες αποστάσεις στον χαμηλότερων διαστάσεων χώρο Y αντίστοιχα.

Όσο χαμηλότερο είναι η διακύμανση υπολοίπων τόσο καλύτερη είναι η αναπαράσταση του αρχικού χώρου στον χώρο μειωμένων διαστάσεων και επομένως τόσο καταλληλότερη η επιλογή του k. Η επιλογή του επομένως μπορεί να οριστεί ως :

$$k_{opt} = \underset{k}{\operatorname{argmin}}\left(1 - \rho_{\widehat{D}_X(k)D_Y}\right) \tag{3.3}$$

Επομένως τρέχοντας τον αλγόριθμο του Isomap για διάφορες τιμές του k δημιουργούνται διάφορες τιμές των διακυμάνσεων υπολοίπων για τις διαφορετικές διαστάσεις που διατηρούνται. Δημιουργώντας ένα γράφημα με τις διαστάσεις στον ένα άξονα και τις τιμές των διακυμάνσεων υπολοίπων στον άλλον, η επιλογή της χαμηλότερης τιμής των διακυμάνσεων υπολοίπων δίνει και την κατάλληλη επιλογή του και για τις διαστάσεις που επιθυμείτε να διατηρηθούν **(Jing & Shao, 2011)**.

3.4 Αλγόριθμος της Isomap

Ψευδοκώδικας

Είσοδος :

- Πίνακας X διαστάσεων $n \times p$ όπου στις γραμμές του βρίσκονται οι μεταβλητές.
- Παράμετρος k-nearest neighbors (ή ακτίνα (radius) ε).
- Το νούμερο των μειωμένων διαστάσεων p που θα κρατηθούν.

1. Κατασκεύασε τον γράφο γειτνίασης σημείων $G = (V, E)$ χρησιμοποιώντας παράμετρο k ή ε.
2. Εφάρμοσε τον αλγόριθμο του Floyd για την κατασκευή σύντομων μονοπατιών.
3. Εφάρμοσε cMDS στον πίνακα D_G :

a) Υπολόγισε τον κεντροποιημένο πίνακα $H = I - \frac{1}{n}ee^T$.

b) Υπολόγισε τον πίνακα εσωτερικών γινομένων $K_G = -\frac{1}{2}HD_GH$.

c) Υπολόγισε τα ιδιοδιανύσματα $\{v_i\}$ και τις ιδιοτιμές $\{\lambda i\}$ του πίνακα K_G.

d) Ταξινόμησε κατά φθίνουσα σειρά τα ιδιοδιανύσματα με βάση της ιδιοτιμές.

e) Υπολόγισε τον πίνακα $Y = \Lambda^{\frac{1}{2}}V^T$ όπου Λ πίνακας διαγώνιος με στοιχεία $\{\lambda i\}$ τα ιδιοδιανύσματα του K_G και V πίνακας με στήλες $\{v_i\}$ τα ιδιοδιανύσματα του K_G.

Έξοδος :

Πίνακας Y μειωμένων διαστάσεων $n \times p$.

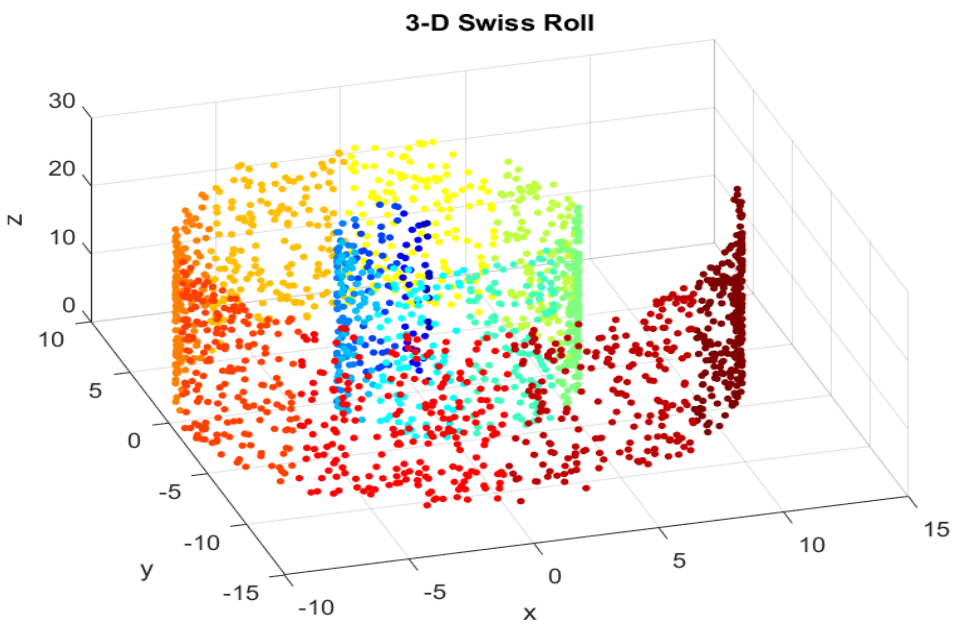
3.5 Εφαρμογή της μεθόδου Isomap σε δεδομένα

3.5.1 Εισαγωγή

Για την πειραματική εφαρμογή της μεθόδου σε δεδομένα παράχθηκαν $\mathrm{N} = 2000$ σημεία τα οποία σε έναν χώρο τριών διαστάσεων εκτείνονται πάνω σε ένα ελικοειδές σχήμα το οποίο είναι γνωστό ως ελβετική κουλούρα (swiss roll). Οι εξισώσεις που χρησιμοποιήθηκαν για την παραγωγή της είναι οι ακόλουθες :

$$\begin{cases} x_i = (t_i + 1)\cos(t_i) \\ y_i = (t_i + 1)\sin(t_i) \\ z_i = 8\pi a_i \end{cases} \tag{3.4}$$

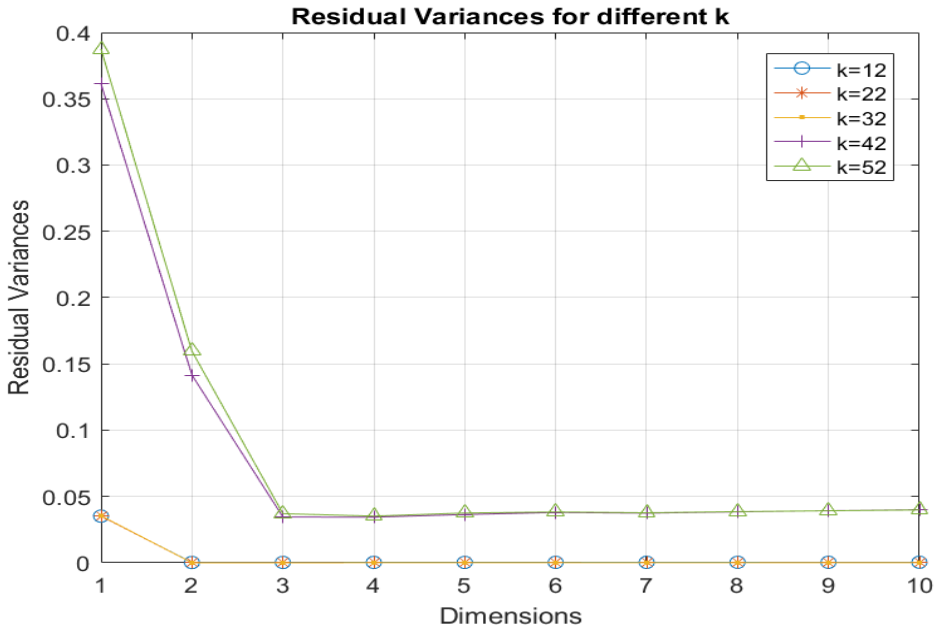
Όπου $t_i = 4\pi\sqrt{r_i}\,, i = 1, \dots, 2000$ με r_i, a_i να είναι ένα σύνολο ομοιόμορφων κατανεμημένων τυχαίων αριθμών στο διάστημα $[0,1]$. Η παράμετρος της θέσης των σημείων ορίζεται από τα χρώματα που έχουν δοθεί στα σημεία τα οποία ακολουθούν την γεωμετρική δομή της κουλούρας όπως διαφαίνεται στο σχήμα 3.1.



Σχήμα 3.1 : Απεικόνιση της ελβετικής κουλούρας (swiss roll) αποτελούμενη από N=2000 σημεία.

3.5.2 Εύρεση της παραμέτρου k

Πριν γίνει η εφαρμογή της μεθόδου Isomap πάνω στα δεδομένα του swiss roll πρωταρχικός στόχος είναι η εύρεση της παραμέτρου k. Η μεθοδολογία που ακολουθείται είναι αυτής που παρουσιάστηκε στο κεφάλαιο 3.3. Ο αλγόριθμος έτρεξε για τιμές του $\mathrm{k} = 12,22,32,42,52$ και τα αποτελέσματα της γραφικής παράστασης των διακυμάνσεων υπολοίπων παρουσιάζονται στο σχήμα 3.2. Η καταλληλότερη επιλογή του k φανερώνει πώς είναι αυτές των τιμών από 12 έως 32 για την μείωση των διαστάσεων σε δύο, ενώ από τη επιλογή για $\mathrm{k} > 42$ τα αποτελέσματα δίνουν υψηλές τιμές στις διακυμάνσεις υπολοίπων.



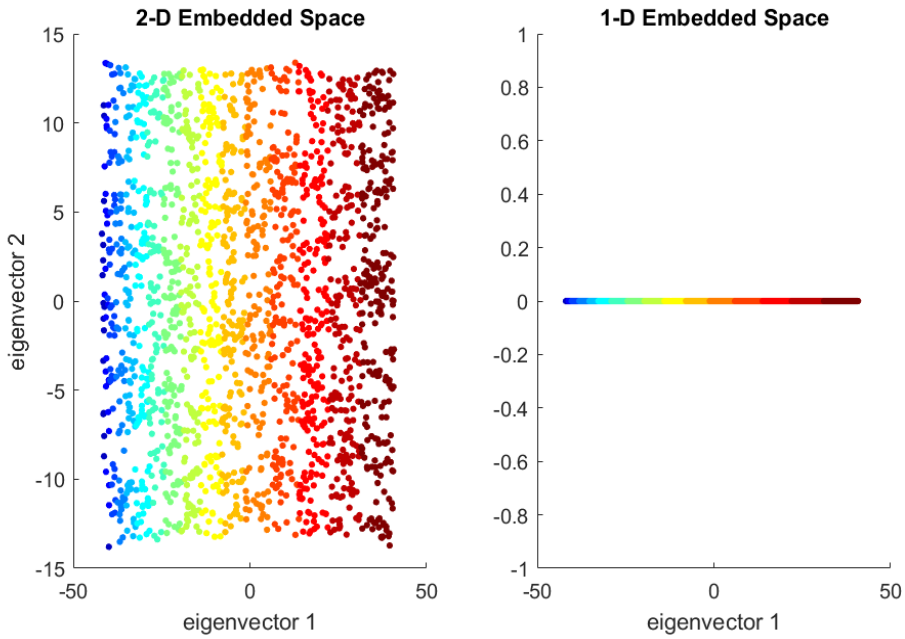
Σχήμα 3.2 : Απεικόνιση του διαγράμματος, διακυμάνσεων υπολοίπων για διάφορες τιμές της παραμέτρου k, πάνω στα δεδομένα της ελβετικής κουλούρας.

3.5.3 Αποτελέσματα

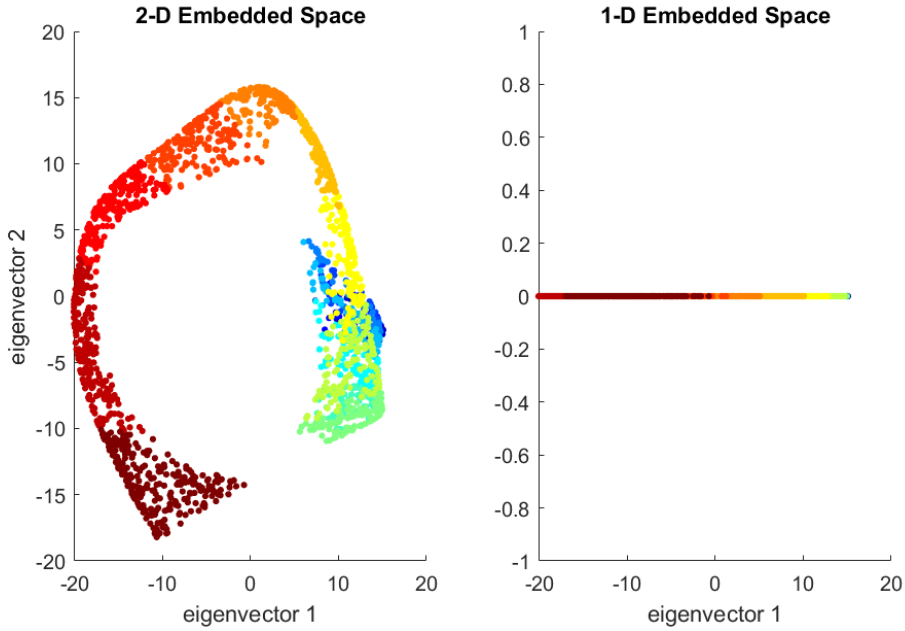
Στο σχήμα 3.3 παρουσιάζονται τα αποτελέσματα μείωσης διαστάσεων της μεθόδου Isomap πάνω στα δεδομένα που παράχθηκαν και δημιούργησαν την ελβετική κουλούρα. Στην πρώτη εικόνα είναι εμφανές πως η μείωση των διαστάσεων σε δύο (για $k = 12$) ανέδειξαν την πολλαπλότητα του αρχικού χώρου που εκτείνονταν τα δεδομένα και ''ξεδίπλωσαν'' την κουλούρα στον χώρο των δύο διαστάσεων διατηρώντας τη σειρά των σημείων (σειρά του χρώματος) στο έπακρο. Στην δεύτερη εικόνα του σχήματος 3.3 διαφαίνεται πως ακόμα και στη μείωση σε μία διάσταση το αποτέλεσμα είναι άκρως ικανοποιητικό, καθώς η παράμετρος της θέσης των σημείων πάνω στο swiss roll έχει διατηρηθεί. Επομένως η Isomap κρίνεται πλέον αποτελεσματική για την εύρεση πολλαπλοτήτων που σχηματίζονται σε μη γραμμικούς χώρους.

Η καταλληλόλητα της παραμέτρου k όπως επισημάνθηκε παίζει καθοριστικό ρόλο στην ανάδειξη της εγγενούς γεωμετρίας των δεδομένων του αρχικού χώρου σε αυτούς των μειωμένων διαστάσεων. Στο σχήμα 3.4 παρουσιάζονται οι μειωμένες διαστάσεις των δεδομένων του swiss roll με την επιλογή παραμέτρου $k = 52$ η οποία δείχνει πως η μέθοδος αποτυγχάνει πλήρως στην αναδίπλωση του swiss roll σε δύο διαστάσεις και η παράμετρος της θέσης των σημείων (χρώμα) έχει χάσει τη σειρά της. Αυτό αιτιολογεί κατά μεγάλο βαθμό την απόρριψη της επιλογής αυτής που έγινε από την γραφική παράσταση των διακυμάνσεων υπολοίπων.

Άξιο παρατήρησης είναι ακόμα το γεγονός πως όσο αυξάνεται η παράμετρος k τόσο περισσότερο η αναδίπλωση του swiss roll που προσπαθείτε να επιτευχθεί αποτυγχάνει. Αυτό συμβαίνει, διότι όσο το k μεγαλώνει, τόσο περισσότερα σημεία στον χώρο ενώνονται με γείτονές τους. Επομένως όσο πιο κοντά το k φτάνει το ολικό πλήθος των σημείων στον χώρο τόσο περισσότερα σημεία έχουν ενωθεί μεταξύ τους με ευκλείδειες αποστάσεις. Μια τέτοια διαδικασία οδηγεί, στην απλή εφαρμογή της γραμμικής μεθόδου cMDS (όλα τα σημεία στον χώρο ενώνονται με ευκλείδειες αποστάσεις) η οποία αποτυγχάνει πλήρως να αναγνωρίσει και να αναδείξει τις γεωμετρικές δομές του συγκεκριμένου γραμμικού προβλήματος.



Σχήμα 3.3 : Απεικόνιση 1000 σημείων του swiss roll σε χώρους μειωμένων διαστάσεων δύων και μίας αντίστοιχα, μετά από εφαρμογή της Isomap (k=12).



Σχήμα 3.4 : Απεικόνιση 1000 σημείων του swiss roll σε χώρους μειωμένων διαστάσεων δύων και μίας αντίστοιχα, μετά από εφαρμογή της Isomap (k=52).

Κεφάλαιο 4

Απεικονίσεις Διάχυσης (Diffusion Maps)

Η πιο νέα κατά κύριο λόγο μέθοδος μείωσης διαστάσεων, από αυτές που μελετιούνται στη συγκεκριμένη εργασία είναι αυτή των απεικονίσεων διάχυσης (diffusion maps). Κύριος στόχος της μη γραμμικής μεθόδου των απεικονίσεων διάχυσης, είναι η ανεύρεση της πολλαπλότητας (manifold) χαμηλότερων διαστάσεων στην οποία εικάζεται ότι εκτείνονται τα δεδομένα **(Porte et al., 2008).** Κατά αυτό τον τρόπο η μέθοδος στοχεύει στην εύρεση σημαντικής πληροφορίας μέσω μελέτης της τοπικής γεωμετρίας των δεδομένων και όχι της ολικής. Με αυτόν τον τρόπο η επισήμανση της διαφοροποίησης της από τις γραμμικές μεθόδους PCA και MDS είναι αναπόφευκτη, καθώς θεωρεί πως σε πολλές εφαρμογές οι μακρινές αποστάσεις είναι άνευ σημασίας και εν τέλει η διατήρηση τους σε ένα χώρο μικρότερων διαστάσεων καθίσταται μη αναγκαία **(Coifmman and Lafon, 2006).** Η ανάδειξη όμως της ολικής γεωμετρίας εν τέλει επιτυγχάνεται σταδιακά, στοχεύοντας αρχικά στην διατήρηση της τοπικής γεωμετρίας που παρατηρείται στα σημεία που βρίσκονται κοντά.

4.1 Η κεντρική ιδέα των Απεικονίσεων Διάχυσης

Κατά κύριο λόγω η μέθοδος των απεικονίσεων διάχυσης έχει ως στόχο, τη μείωση διαστάσεων, αναδιοργανώνοντας τα δεδομένα που μελετιούνται, σύμφωνα με τις παραμέτρους που αναδεικνύουν την γεωμετρία τους.

Η σύνδεση (connectivity) του συνόλου των δεδομένων, πραγματοποιείται χρησιμοποιώντας μέτρα τοπικής ομοιότητας, που χρησιμοποιείται για τη δημιουργία μίας χρονικής εξαρτημένης διαδικασίας διάχυσης. Αυτή με τη σειρά της καθώς πορεύεται ενσωματώνει την τοπική γεωμετρία για την ανάδειξη γεωμετρικών δομών του συνόλου των δεδομένων σε διαφορετικές κλίμακες. Ορίζοντας έτσι ένα μέτρο διάχυσης χρονικά εξαρτημένο μπορεί να οριστεί η ομοιότητα μεταξύ δύο σημείων σε συγκεκριμένο χρόνο βασιζόμενη στη γεωμετρία που έχει αναδειχθεί.

Οπότε οι απεικονίσεις διάχυσης, μεταμορφώνουν-αναδιαμορφώνουν τα δεδομένα σε ένα χαμηλότερης διάστασης χώρο έτσι ώστε η ευκλείδεια απόσταση μεταξύ των σημείων σε αυτόν, να αντιστοιχούν όσο το δυνατόν περισσότερο με αυτήν της απόστασης διάχυσης στον αρχικό χώρο. Οι διαστάσεις του χώρου διάχυσης (του μειωμένου διαστάσεων χώρου) ορίζονται με βάση τη γεωμετρική δομή των δεδομένων και με την ακρίβεια που προσεγγίζεται η απόσταση διάχυσης.

4.2 Η μαθηματική προσέγγιση των Απεικονίσεων Διάχυσης

4.2.1 Εισαγωγή

Συνδεσιμότητα (Connectivity)

Υποθέτοντας πως ένας τυχαίος περίπατος λαμβάνει μέρος στα δεδομένα υπό επεξεργασία το άμεσο αποτέλεσμα που αντλείται είναι πως η μετάβαση σε κοντινά σημεία από το σημείο εκκίνησης του περιπάτου είναι πιθανότερη σε σχέση με αυτά που είναι απομακρυσμένα. Αυτή η παρατήρηση, καθιστά απαραίτητη τη συσχέτιση της απόστασης στον χώρο και της έννοιας της πιθανότητας. Κατά αυτόν τον τρόπο η συνδεσιμότητα δύο σημείων στο χώρο ορίζεται ως η πιθανότητα μετάβασης από το x στο y σε ένα βήμα τυχαίου περιπάτου :

$$\mathrm{connectivity}(x, y) = p(x, y) \tag{4.1}$$

Είναι βοηθητικό να εκφραστεί η συνδεσιμότητα αυτή με μια μη κανονικοποιημένη συνάρτηση πιθανοφάνειας $k(x, y)$ γνωστή ως πυρήνας διάχυσης, ο οποίος περιγράφει με τον καλύτερο τρόπο αυτή τη σύνδεση.

$$\mathrm{conectivity}(x, y) \propto k(x, y) \tag{4.2}$$

Ο Gaussian kernel αποτελεί την πιο συχνή επιλογή πυρήνα, ο οποίος ορίζεται ως :

$$k(x, y) = e^{\left(-\frac{|x-y|^2}{2\sigma^2}\right)} \tag{4.3}$$

και υποδεικνύει την συνδεσιμότητα των δεδομένων με τέτοιο τρόπο, όπου ορίζει τοπικά μέτρα ομοιότητας μεταξύ των σημείων, μέσα σε συγκεκριμένες περιοχές (''γειτονιές'') και ακολουθεί τις εξής βασικές ιδιότητες :

i. Συμμετρικός : $k(x, y) = k(y, x)$
ii. Μη αρνητικός: $k(x, y) \geq 0$
iii. Τοπικότητα μεταξύ των σημείων, δοθέντος παραμέτρου $\sigma > 0$:
$k(x, y) \to 1$ για τα $\|x - y\| \ll \sigma$
$k(x, y) \to 0$ για τα $\|x - y\| \gg \sigma$

Η πρώτη ιδιότητα όπως θα δειχθεί είναι αναγκαίο να υπάρχει καθώς η συμμετρικότητα του πίνακα $K_{ij} = k(x_i, x_j)$ παίζει καθοριστικό ρόλο στην εξαγωγή αποτελεσμάτων.

Η δεύτερη ιδιότητα είναι εξίσου κρίσιμη καθώς η αναγωγή του πίνακα K σε ένα πίνακα πιθανοτήτων καθιστά αναγκαία την μη ύπαρξη αρνητικών ποσοτήτων σε αυτόν. Ερμηνεύοντας λοιπόν το $k(x, y)$ ως κλιμακωτή πιθανότητα,

$$\frac{1}{d_X}\sum_{y\in X} k(x, y) = 1 \tag{4.4}$$

Κατά αυτόν τον τρόπο η συνδεσιμότητα ταυτίζεται με τον πυρήνα ως εξής :

$$\text{connectivity}(x, y) = p(x, y) = \frac{k(x, y)}{d_X} \tag{4.5}$$

Όπου $\frac{1}{d_X}$ η σταθερά κανονικοποίησης.

Πολύ βασική σημασία πρέπει να δοθεί στην τρίτη ιδιότητα που ορίζει την τοπικότητα του πυρήνα διάχυσης. Η επιλογή της παραμέτρου σ είναι ένα θέμα κρίσιμο που έχει μελετηθεί κατά καιρούς σε πολλές εργασίες, καθώς η λανθασμένη εκτίμησή της εγκυμονεί κινδύνους για παραπλανητική εκτίμηση γραφημάτων και μπορεί να οδηγήσει σε μη αναμενόμενα αποτελέσματα. Μια μικρής κλίμακας παράμετρος σ αναδεικνύει με επιτυχία τη τοπικότητα ενός γραφήματος αλλά μπορεί παράλληλα να μειώσει τη συνδεσιμότητα του. Εν αντίθεσή, η επιλογή μεγάλης κλίμακας παραμέτρου σ μπορεί να δημιουργήσει ένα απόλυτα συνδεδεμένο γράφημα αλλά να καθιστά μη-ευαίσθητες τις διακυμάνσεις των αποστάσεων και κατ' επέκταση την έλλειψη ανάδειξης της τοπικής γεωμετρίας των δεδομένων **(Talmon, 2013).** Στην ενότητα 4.3 παρουσιάζονται κάποιες μεθοδολογίες επιλογής της παραμέτρου σ.

Διαδικασία διάχυσης (Diffusion Process)

Ορίζεται λοιπόν ένας πίνακας διάχυσης P με $P_{ij} = p(x_i, x_j)$ ο οποίος δίνει τη πληροφορία σύνδεσης μεταξύ δυο σημείων x_i, x_j στα πλαίσια της τοπικότητας που παρέχει ο πυρήνας διάχυσης. Αναλογικά με τον τυχαίο περίπατο ο πίνακας παρέχει τις πιθανότητες ενός βήματος από το σημείο i στο σημείο j και παρουσιάζεται ως :

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \tag{4.6}$$

Περνώντας δυνάμεις στον πίνακα P,αυξάνονται στην ουσία ο αριθμός των βημάτων στον τυχαίο περίπατο και κατά αυτόν τον τρόπο παρατηρείται το εξής :

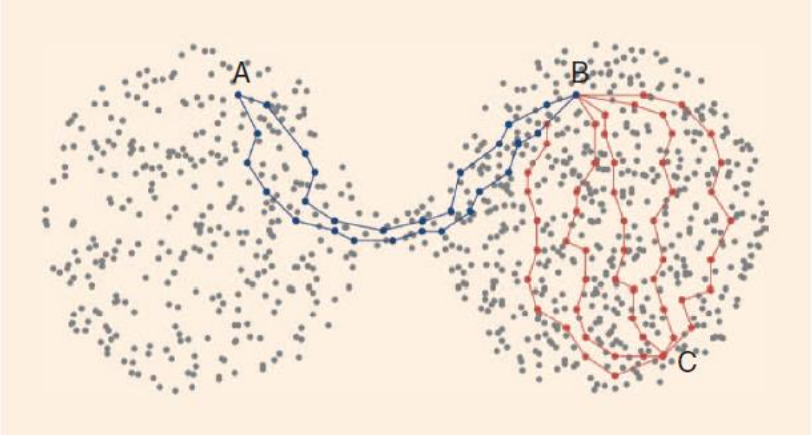
$$P^2 = \begin{bmatrix} p_{11}p_{11} + p_{12}p_{21} & p_{12}p_{22} + p_{11}p_{12} \\ p_{21}p_{12} + p_{22}p_{21} & p_{22}p_{22} + p_{21}p_{22} \end{bmatrix} \tag{4.7}$$

Το P_{11} αθροίζει δύο πιθανότητες αυτή του να παραμείνει ο τυχαίος περίπατος στο σημείο 1 και να μεταφερθεί στο σημείο 2 και πάλι πίσω. Για δύο μεταπηδήσεις αυτά είναι όλα τα μονοπάτια από το i στο j.

Κατά αυτόν τον τρόπο υπολογίζονται οι πίνακες πιθανότητας P^t όπου για αυξημένες τιμές του βήματος t παρατηρείται το σύνολο δεδομένων σε διαφορετικές κλίμακες. Αυτή είναι η διαδικασία διάχυσης η οποία προσφέρει μέσω της δομής των τοπικών συνδέσεων να παρουσιάσει την ολική συνδεσιμότητα τους συνόλου των δεδομένων.

Καθώς λοιπόν το t αυξάνεται (η διαδικασία διάχυσης εξελίσσεται) η πιθανότητα να ακολουθουθεί ένα μονοπάτι πάνω στην γεωμετρική δομή που σχηματίζουν τα δεδομένα μεγαλώνει. Αυτό συμβαίνει λόγω πυκνότητας των σημείων που εκτείνονται πάνω στην γεωμετρική δομή του συνόλου και επομένως η σύνδεση του είναι ψηλή. Τα μονοπάτια αυτά λοιπόν περιέχουν μικρές - ψηλής πιθανότητας μεταπηδήσεις (jumps). Αντιθέτως τα μονοπάτια που δεν ακολουθούν τη γεωμετρική δομή των δεδομένων περιλαμβάνουν μία ή περισσότερες, μεγάλες-μικρής πιθανότητας μεταπηδήσεις (jumps) το οποίο έχει αντίκτυπο στην σύνδεση μεταξύ σημείων **(de la Porte et al., 2008)**.

Όπως παρατηρείται στο σχήμα 4.1 όσο αυξάνεται το t τόσο μεγαλύτερες οι πιθανότητες μεταπηδήσεων που δημιουργούνται και άρα τόσο πιο εύκολο είναι να φτάσει το A σημείο στο B. Καθώς η διαδικασία διάχυσης τρέχει στο χρόνο είναι πιο εύκολο να φτάσει το σημείο B στο C παρότι στο A, λόγω του πλήθους των υψηλών πιθανοτήτων μονοπατιών που σχηματίζονται. Κατά αυτόν τον τρόπο το σημείο B ομαδοποιείται κατά μία έννοια περισσότερο με το σημείο C παρότι με το σημείο A **(Gao, 2015)**.



Σχήμα 4.1 : Απεικόνιση δύο διαφορετικών ομοιόμορφων ομάδων δεδομένων και η αναπαράσταση μονοπατιών πιθανοτήτων που σχηματίζονται μεταξύ σημείων A,B,C καθώς η διαδικασία διάχυσης εξελίσσεται **(Gao, 2015).**

Απόσταση διάχυσης (Diffusion Distance)

Βασισμένο λοιπόν στη δομή που παρουσιάστηκε, ορίζεται ένα μέτρο διάχυσης που στόχο έχει την μέτρηση της ομοιότητας δύο σημείων στο χώρο που μελετάται. Στόχο έχει την ανάδειξη της συνδεσιμότητας μεταξύ τους και ορίζεται ως απόσταση διάχυσης :

$$D_t(X_i, X_j) = \sum_{u \in X} \left| p_t(X_i, u) - p_t(X_j, u) \right|^2 \tag{4.8}$$

Η απόσταση διάχυσης είναι μικρή αν υπάρχουν πολλά υψηλής πιθανότητας μονοπάτια μεταξύ δύο σημείων και καθώς αθροίζει μεταξύ όλων των μονοπατιών που σχηματίζονται, την καθιστά πολύ ισχυρή στην αντιμετώπιση θορύβου στα δεδομένα.

Κατά αυτόν τον τρόπο καθώς η διαδικασία διάχυσης προχωράει στο χρόνο, κομβικό ρόλο στον υπολογισμό της απόστασης διαδραματίζουν τα μονοπάτια που σχηματίζονται μεταξύ των σημείων.

Εξετάζοντας καλύτερα λοιπόν την εξίσωση διάχυσης και παρατηρώντας τον όρο $p_t(X_i, u)$ ο οποίος εκφράζει το άθροισμα όλων των πιθανοτήτων όλων των πιθανών μονοπατιών σε χρόνο t μεταξύ του σημείου X_i και οποιουδήποτε σημείου u οδηγούμαστε σε πολύ κρίσιμο συμπέρασμα. Για να παραμείνει η απόσταση διάχυσης μικρή όσο το δυνατόν γίνεται πρέπει η πιθανότητα σύνδεσης μεταξύ του X_i και του u να είναι ίδια με αυτή του X_j και του u και αυτό συνδέεται όταν το X_i συνδέεται με το X_j μέσω του u. Με άλλα λόγια δύο σημεία είναι όμοια αν διαχέονται μαζί και επηρεάζουν το γράφημα με τον ίδιο ακριβώς τρόπο **(Fouss et.al, 2016)**.

4.2.2 Ορισμός προβλήματος

Απεικόνιση Διάχυσης (Diffusion Map)

Οπως έχει προαναφερθεί και στην εισαγωγή του κεφαλαίου τα δεδομένα που βρίσκονται υπό μελέτη μπορεί να μην εκτείνονται σε γραμμικό χώρο αλλά σε κάποια μη-γραμμική πολλαπλότητα. Προηγουμένως μελετήθηκε η απόσταση διάχυσης η οποία είναι ικανή να υπολογίσει τις αποστάσεις που εκτείνονται σε μια πολλαπλότητα.

Ο υπολογισμός όλων των αποστάσεων διάχυσης είναι υπολογιστικά χρονοβόρος και για αυτό το λόγο είναι αρκετά βολικό να απεικονίσουμε τα σημεία (δεδομένα) μέσω μίας απεικόνισης, σε έναν ευκλείδειο χώρο σύμφωνα με το μέτρο διάχυσης που έχει οριστεί. Η απόσταση διάχυσης έτσι πολύ απλά γίνεται η ευκλείδεια απόσταση σε ένα νέο χώρο που καλείται χώρος διάχυσης (diffusion space).

Η απεικόνιση διάχυσης λοιπόν που απεικονίζει τα σημεία από τον χώρο παρατηρήσεων στο χώρο διάχυσης στόχος της είναι η κατά κάποιο τρόπο η αναδιοργάνωση των δεδομένων σύμφωνα με τη μετρική διάχυσης που ορίστηκε. Κατά αυτόν τον τρόπο έρχεται η εκμετάλλευση αυτής της αναδιοργάνωσης για τη μείωση των διαστάσεων των δεδομένων που αναζητείται.

Μια λοιπόν απεικόνιση διάχυσης έχει ως στόχο μέσω της απεικόνισης από τον αρχικό χώρο στο χώρο διάχυσης, να διατηρήσει όσο γίνεται την εγγενή γεωμετρία των δεδομένων. Κατ' επέκταση καθώς η απεικόνιση μετράει αποστάσεις στον νέο χώρο δομής χαμηλότερων διαστάσεων, αναμένεται η χρησιμοποίηση λιγότερων συντεταγμένων για τα δεδομένα που εξετάζονται.

Παρατηρώντας τώρα τον ακόλουθο χάρτη :

$$Y_i = \begin{bmatrix} p_t(X_i, X_1) \\ p_t(X_i, X_2) \\ \vdots \\ p_t(X_i, X_N) \end{bmatrix} = P_{i*}^T \tag{4.9}$$

η ευκλείδεια απόσταση μεταξύ των σημείων απεικόνισης Y_i και Y_j είναι :

$$\left|\left|Y_i - Y_j\right|\right|_E^2 = \left[p_t(X_i, X_1) - p_t(X_j, X_1)\right]^2 + \cdots + \left[p_t(X_i, X_N) - p_t(X_j, X_N)\right]^2 = \sum_{u \in x} |p_t(X_i, u) - p_t(X_j, u)|^2 \tag{4.10}$$

η οποία είναι η απόσταση διάχυσης που έχει οριστεί μεταξύ των σημείων X_i και X_j. Αυτό υποδεικνύει την αναδιοργάνωση που ήταν υπό αναζήτηση σύμφωνα με τη απόσταση διάχυσης. Καμία μείωση διαστάσεων δεν έχει πραγματοποιηθεί ακόμα καθώς η διάσταση των απεικονισμένων σημείων Y_i είναι ακόμα N διαστάσεων. Η μείωση των διαστάσεων θα επιτευχθεί όπως θα δειχθεί στο επόμενο κεφάλαιο μέσω της αναδιατύπωσης των Y_i, εκφράζοντας τα συναρτήσει των ιδιοδιανύσμάτων και ιδιοτιμών του πίνακα P. Μελετώντας περαιτέρω την απόσταση αυτών στο χώρο διάχυσης προέρχεται και η υπόδειξη για την μείωση των διαστάσεων.

4.2.3 Επίλυση του προβλήματος των Απεικονίσεων Διάχυσης

Όπως ορίστηκε και πριν, έστω ένας πίνακας $K_{N\times N}$ πυρήνας διάχυσης που είναι συμμετρικός έτσι ώστε $K[i,j] = k(i,j)$, ένας διαγώνιος πίνακας D κανονικοποιεί τις γραμμές του ώστε να παραχθεί ένας πίνακας διάχυσης :

$$P = D^{-1}K \tag{4.11}$$

Έστω ότι ορίζεται σε αυτό το σημείο ένας πίνακας P' όπου :

$$P' = D^{\frac{1}{2}}PD^{-\frac{1}{2}} \tag{4.12}$$

Εύκολα διαπιστώνεται πως ο πίνακας αυτός είναι συμμετρικός καθώς εμφωλεύοντας την (4.11) στην (4.12) λαμβάνεται το εξής :

$$P' = D^{\frac{1}{2}}D^{-1}KD^{-\frac{1}{2}} = D^{-\frac{1}{2}}KD^{-\frac{1}{2}} \tag{4.13}$$

και καθώς K συμμετρικός P' συμμετρικός. Έτσι λοιπόν από τη συμμετρικότητα του P' υπάρχουν ορθογώνια ιδιοδιανύσματα του P' και ιδιοτιμές του έτσι ώστε :

$$P' = S\Lambda S^{T} \tag{4.14}$$

Όπου S πίνακας με στήλες τα ιδιοδιανύσματα και Λ πίνακας διαγώνιος με τις ιδιοτιμές του P'.

Επομένως πολλαπλασιάζοντας από αριστερά με $D^{-\frac{1}{2}}$, από δεξιά με $D^{\frac{1}{2}}$ προκύπτει η σχέση :

$$P = D^{-\frac{1}{2}}P'D^{\frac{1}{2}} \tag{4.15}$$

και εμφωλεύοντας στη σχέση (4.15) τη σχέση (4.14) ο πίνακας P γράφεται ως :

$$P = D^{-\frac{1}{2}}S\Lambda S^{T}D^{\frac{1}{2}} \tag{4.16}$$

Και καθώς S ορθογώνιος με $SS^{T} = I \Leftrightarrow S^{T} = S^{-1}$:

$$P = D^{-\frac{1}{2}}S\Lambda S^{-1}D^{\frac{1}{2}} \Leftrightarrow \tag{4.17}$$

$$P = D^{-\frac{1}{2}}S\Lambda(D^{-\frac{1}{2}}S)^{-1} \Leftrightarrow \tag{4.18}$$

$$P = Q\Lambda Q^{-1} \tag{4.19}$$

Όπου

$$Q = D^{-\frac{1}{2}}S \tag{4.20}$$

Κατά αυτόν τον τρόπο παρατηρείται πως ο πίνακας P με τον πίνακα P' έχουν τις ίδιες ιδιοτιμές. Παρατηρείται ακόμα πως η σχέση (4.19) δίνει και κάποιες επιπλέον ιδιότητες που έχουν να κάνουν με τα ιδιοδιανύσματα του P συναρτήσει των ιδιοδιανύσμάτων του P'. Η σχέση λοιπόν (4.19) μπορεί να γραφτεί και ως :

$$\mathrm{PQ} = \mathrm{Q}\Lambda \quad \text{όπως και} \quad \mathrm{Q}^{-1}\mathrm{P} = \Lambda\mathrm{Q}^{-1}$$

Που αυτό σημαίνει ότι οι στήλες του Q αποτελούν τα δεξιά ιδιοδιανύσματα του P και οι γραμμές του Q^{-1} αποτελούν τα αριστερά ιδιοδιανύσματα του P. Κατά αυτόν τον τρόπο ορίζοντας $\mathrm{u_k}$ τα δεξιά ιδιοδιανύσματα $(\mathrm{Pu_k} = \lambda\mathrm{u_k})$, $\mathrm{v_k}$ τα αριστερά ιδιοδιανύσματα $(\mathrm{v_k^T P} = \mathrm{v_k^T}\lambda)$ του P και εν τέλει $\mathrm{x_k'}$ τα ιδιοδιανύσματα του P', εξάγονται με βάση την (4.20) οι ακόλουθες σχέσεις :

$$\mathrm{u_k} = \mathrm{D}^{-\frac{1}{2}}\mathrm{x_k'} \qquad \mathrm{D}^{\frac{1}{2}}\mathrm{u_k} = \mathrm{x_k'} \tag{4.21}$$

$$\mathrm{v_k} = \mathrm{D}^{\frac{1}{2}}\mathrm{x_k'} \qquad \mathrm{D}^{-\frac{1}{2}}\mathrm{v_k} = \mathrm{x_k'} \tag{4.22}$$

Επειδή ο P' είναι συμμετρικός θα έχει την ακόλουθη ιδιοσύνθεση όπως έχει προαναφερθεί :

$$\mathrm{P}' = \mathrm{S}\Lambda\mathrm{S}^{\mathrm{T}} \Leftrightarrow \tag{4.23}$$

$$\mathrm{P}' = \sum_{\mathrm{k}} \lambda_{\mathrm{k}}\mathrm{x_k'}\mathrm{x_k'}^{\mathrm{T}} \tag{4.24}$$

Και καθώς $\mathrm{P} = \mathrm{D}^{-\frac{1}{2}}\mathrm{P}'\mathrm{D}^{\frac{1}{2}}$,

$$\mathrm{P} = \mathrm{D}^{-\frac{1}{2}}\left(\sum_{\mathrm{k}} \lambda_{\mathrm{k}}\mathrm{x_k'}\mathrm{x_k'}^{\mathrm{T}}\right)\mathrm{D}^{\frac{1}{2}} \Leftrightarrow$$

$$\mathrm{P} = \sum_{\mathrm{k}} \lambda_{\mathrm{k}}\left(\mathrm{D}^{-\frac{1}{2}}\mathrm{x_k'}\right)\left(\mathrm{x_k'}^{\mathrm{T}}\mathrm{D}^{\frac{1}{2}}\right) \Leftrightarrow$$

$$\mathrm{P} = \sum_{\mathrm{k}} \lambda_{\mathrm{k}}\mathrm{u_k}\mathrm{v_k}^{\mathrm{T}} \tag{4.25}$$

Εφαρμόζοντας PCA στον P' παρατηρείται πως οι συντεταγμένες των γραμμών του πίνακα $\mathrm{P}' = \mathrm{S}\Lambda\mathrm{S}^{\mathrm{T}}$ θα είναι οι $\{\lambda_{\mathrm{k}}\mathrm{x_k'}[\mathrm{i}]\}$ στο νέο ορθογώνιο σύστημα συντεταγμένων που ορίζεται από τα ορθοκανονικά ιδιοδιανύσματα $\mathrm{x_k'}$ του πίνακα P'. Δηλαδή οι

$$\mathrm{M_i} = \begin{bmatrix} \lambda_1\mathrm{x_1'}[\mathrm{i}] \\ \lambda_2\mathrm{x_2'}[\mathrm{i}] \\ \vdots \\ \lambda_{\mathrm{N}}\mathrm{x_N'}[\mathrm{i}] \end{bmatrix} \tag{4.26}$$

Αποτελούν τις συντεταγμένες των γραμμών του P'. Κατά αυτόν τον τρόπο ορίζονται λόγω των εξισώσεων (4.20), (4.21) και οι

$$Y_i' = \begin{bmatrix} \lambda_1 u_1[i] \\ \lambda_2 u_2[i] \\ \vdots \\ \lambda_N u_N[i] \end{bmatrix} \tag{4.27}$$

οι οποίες αποτελούν τις συντεταγμένες των γραμμών του P σε ένα νέο μη ορθογώνιο σύστημα συντεταγμένων που ορίζεται από τα αριστερά ιδιοδιανύσματα του πίνακα P. Αυτό συμβαίνει καθώς ${v_k}^T v_k \neq 1$ ή αντίστοιχα ${v_k}^T v_l \neq 0$ για $l \neq k$.

Χρησιμοποιώντας λοιπόν ένα διαφορετικό μέτρο έστω τη ποσότητα $Q_0 = D^{-1}$ τότε ικανοποιούνται και οι δύο προδιαγραφές για την ορθογωνιότητα των διανυσμάτων, ${v_k}^T D^{-1} v_k = {x_\kappa'}^T x_\kappa' = 1$ και αντίστοιχα ${v_k}^T D^{-1} v_l = {x_\kappa'}^T x_l' = 0$ για τα $l \neq k$.

Έτσι λοιπόν τα αριστερά ιδιοδιανύσματα δημιουργούν ένα ορθοκανονικό σύστημα συντεταγμένων στον $\mathbb{R}^N$ με τη μετρική D^{-1}. Ορίζεται λοιπόν ο $\mathbb{R}^N$ με τη μετρική D^{-1}ως ο χώρος διάχυσης ο οποίος περιγράφεται από το $l_2(\mathbb{R}^N, D^{-1})$.

Οπότε η ευκλείδειά απόσταση μεταξύ δύο διανυσμάτων α, α' είναι η :

$$d(\alpha, \alpha')_{l_2}^2 = d(\alpha, \alpha')_{l_2(\mathbb{R}^N, I)}^2 = (a - a')^T (a - a') \tag{4.28}$$

και αντίστοιχα η ευκλείδεια απόσταση με τη μετρική D^{-1} είναι η :

$$d(\alpha, \alpha')_{l_2}^2 = d(\alpha, \alpha')_{l_2(\mathbb{R}^N, D^{-1})}^2 = (a - a')^T D^{-1} (a - a') \tag{4.29}$$

Ισχυρισμός : Αν γίνει η επιλογή των συγκεκριμένων συντεταγμένων διάχυσης όπως στην (4.27) τότε η απόσταση διάχυσης μεταξύ των σημείων στον χώρο παρατηρήσεων (χρησιμοποιώντας την μετρική D^{-1}) είναι ίση με την ευκλείδεια απόσταση στον χώρο διάχυσης.

Απόδειξη ισχυρισμού : (Για απλούστευση χρησιμοποιείται t=1)

$$D_t(x_i, x_j)^2 = \left\|p(x_i,\cdot) - p(x_j,\cdot)\right\|^2_{l_2(\mathbb{R}^N, D^{-1})}$$

$$= \|P[i,\cdot] - P[j,\cdot]\|^2_{l_2(\mathbb{R}^N, D^{-1})} \tag{4.30}$$

Με βάση την (4.25) η παράσταση γίνεται :

$$\left|\sum_k \lambda_k u_k[i] v_k - \sum_k \lambda_k u_k[j] v_k\right|^2$$

$$= \left|\sum_k \lambda_k v_k (u_k[i] - u_k[j])\right|^2$$

$$= \left|\sum_k \lambda_k D^{\frac{1}{2}} x_k{}' (u_k[i] - u_k[j])\right|^2$$

$$= \left|\sum_k D^{\frac{1}{2}} \lambda_k x_k{}' (u_k[i] - u_k[j])\right|^2 \tag{4.31}$$

Στην $l_2(\mathbb{R}^N, D^{-1})$,αυτή η απόσταση γίνεται :

$$\left[\sum_k D^{\frac{1}{2}} \lambda_k x_k{}' (u_k[i] - u_k[j])\right]^T D^{-1} \left[\sum_l D^{\frac{1}{2}} \lambda_l x_l{}' (u_l[i] - u_l[j])\right]$$

$$= \sum_k \lambda_k x_k'^T (u_k[i] - u_k[j]) D^{\frac{1}{2}} D^{-1} D^{\frac{1}{2}} \sum_l \lambda_l x_l{}' (u_l[i] - u_l[j])$$

$$= \sum_k \lambda_k x_k'^T (u_k[i] - u_k[j]) \sum_l \lambda_l x_l{}' (u_l[i] - u_l[j]) \tag{4.32}$$

με $x_\kappa'^T x_\kappa' = 1$ και $x_\kappa'^T x_l' = 0$ η (4.32) γίνεται :

$$\sum_k \lambda_k^2 (u_k[i] - u_k[j])^2 \tag{4.33}$$

$$= \left||Y_i' - Y_j'|\right|^2_{l_2(\mathbb{R}^N, I)} \tag{4.34}$$

Τέλος απόδειξης ισχυρισμού

Χωρίς την απλοποίηση που πραγματοποιήθηκε (t=1) το τελικό αποτέλεσμα θα είναι :

$$D_t(x_i, x_j)^2 = \sum_k \lambda_k^{2t}(u_k[i] - u_k[j])^2 = \left|\left|Y_i' - Y_j'\right|\right|^2_{l_2(\mathbb{R}^N, I)} \qquad (4.35)$$

Επομένως η απόσταση διάχυσης είναι η ευκλείδεια απόσταση μεταξύ των σημείων που έχουν προβληθεί στον χώρο της απεικόνισης διάχυσης. Στην ουσία η απόδειξη αυτή δίνει μία πολύ ισχυρή τοποθέτηση για τη μείωση των διαστάσεων. Οι ιδιοτιμές του πίνακα P μπαίνουν σε φθίνουσα σειρά ώστε όσο αυξάνεται το k η απόσταση διάχυσης να έχει μειωμένες συνεισφορές. Η μείωση δηλαδή των διαστάσεων επιτυγχάνεται διατηρώντας όσο το δυνατόν τα κυρίαρχα ιδιοδιανύσματα που προέρχονται από τις κυρίαρχες ιδιοτιμές του πίνακα P, έτσι ώστε η $\left|\left|Y_i' - Y_j'\right|\right|$ να είναι ταυτόσημη με τη απόσταση διάχυσης $D_t(x_i, x_j)$, όσο το δυνατόν περισσότερο.

Σημαντική παρατήρηση επίσης είναι πως λόγω του γεγονότος ότι ο πίνακας P είναι πίνακας πιθανοτήτων τότε η πρώτη ιδιοτιμή του θα ισούται με 1 ($\lambda_0 = 1$) και επομένως το ιδιοδύανυσμα που αντίστιχει σε αυτην είναι ίσο με το e ($u_1 \propto e$), άρα η συνεισφορά του στο άθροισμα είναι μηδενική **(Fouss, 2016)**.

Επομένως η μείωση των διαστάσεων των δεδομένων μέσω της μεθόδου των diffusion maps έρχεται υπολογίζοντας τα ιδιοδιανύσματα και τις ιδιοτιμές του πίνακα P, ταξινομώντας τις ιδιοτιμές και αντίστοιχα τα ιδιοδιανύσματα με φθίνουσα σειρά. Η επιλογή των μεγαλύτερων p ιδιοτιμών και αντίστοιχα p ιδιοδιανύσμάτων δίνει τις μειωμένες συντεταγμένες των σημείων :

$$Y_i' = \begin{bmatrix} \lambda_1^t u_1[i] \\ \lambda_2^t u_2[i] \\ \vdots \\ \lambda_p^t u_p[i] \end{bmatrix} \qquad (4.36)$$

με p<N.

4.3 Επιλογή της παραμέτρου ε

Στις απεικονίσεις διάχυσης όπως προαναφέρθηκε η παράμετρος $\varepsilon = 2\sigma^2$ είναι πολύ κρίσιμη στον υπολογισμό του πυρήνα $K_{ij} = e^{\left(-\frac{|x_i - x_j|^2}{2\sigma^2}\right)} = e^{\left(-\frac{|x_i - x_j|^2}{\varepsilon}\right)}$. Αυτό που πρέπει να τονιστεί είναι πως η επιλογή του ε είναι πλήρως εξαρτώμενη από τα ίδια τα δεδομένα. Κατά καιρούς πολλές τεχνικές έχουν εφαρμοστεί για τον υπολογισμό της παραμέτρου αυτής, σε πολλές εργασίες μελέτης των diffusion maps. Στην παρούσα εργασία θα εξεταστούν δύο τεχνικές επιλογής της παραμέτρου.

1η τεχνική

Ο **Lafon (2006)** στη διατριβή του, πρότεινε η επιλογή του ε να γίνει υπολογίζοντας τον μέσο όρο των μικρότερων, μη μηδενικών τετραγωνικών αποστάσεων $|x_i - x_j|^2$:

$$\varepsilon = \frac{1}{N} \sum_{i}^{N} \min_{j: x_i \neq x_j} |x_i - x_j|^2 \tag{4.37}$$

Ο κυρίαρχος στόχος αυτής της επιλογής του ε είναι η διασφάλιση σίγουρης πραγματοποίησης μιας διάχυσης στα δεδομένα που εξετάζονται.

2η τεχνική

Η δεύτερη τεχνική επιλογής ε που θα αναλυθεί θεωρείται πως δίνει πιο εύρωστα αποτελέσματα από τη πρώτη και προτάθηκε από τον **Singer (2007)**. Η συγκεκριμένη τεχνική δίνει μια κατά κάποιο τρόπο φυσική παρουσίαση, χρησιμοποιώντας τη λογική, για τις ακραίες τιμές επιλογής της παραμέτρου ε.

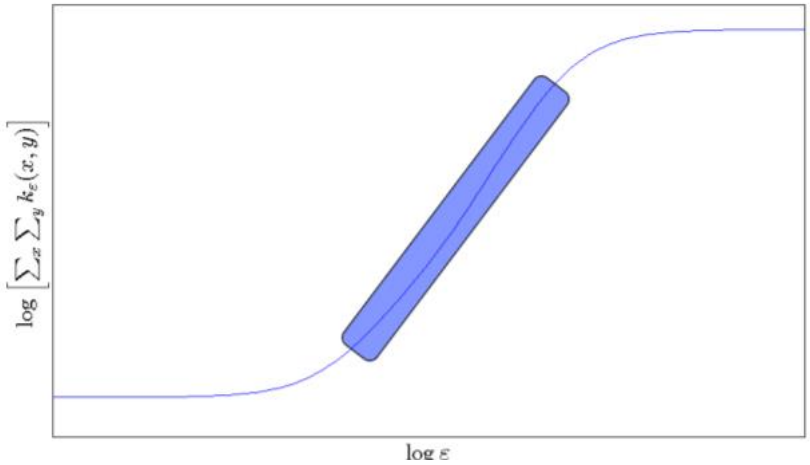
Η αντίληψη είναι πως όταν το ε παίρνει υψηλές τιμές σε σχέση με τις τετραγωνικές αποστάσεις των δεδομένων τα στοιχεία του K θα πλησιάζουν την μονάδα. Αντιθέτως όταν το ε παίρνει χαμηλές τιμές σε σχέση με τετραγωνικές αποστάσεις τα στοιχεία του K θα τείνουν στον 0. Η πρώτη επιλογή υποδεικνύει πως η διάχυση θα καταλάβει πολύ μεγάλο μέρος στα δεδομένα ενώ στην δεύτερη περίπτωση η διάχυση θα είναι μηδαμινή.

Οι επιλογές του ε λοιπόν θα ήταν ιδανικό να λαμβάνονται τιμές ανάμεσα σε αυτές τις ακραίες περιπτώσεις που αναφέρθηκαν. Κατά αυτόν τον τρόπο η τεχνική εφαρμόζεται κατά τα ακόλουθα βήματα :

1. Για διάφορες τιμές του ε υπολόγισε τον πίνακα $K = K(\varepsilon)$.
2. Υπολόγισε τη συνάρτηση $L(\varepsilon)$, η οποία είναι ίση με το ολικό άθροισμα των στοιχείων του K:

$$L(\varepsilon) = \sum_{i=1}^{N} \sum_{j=1}^{N} K_{ij}(\varepsilon) \tag{4.38}$$

3. Δημιούργησε ένα γράφημα $L(\varepsilon)$ σε λογαριθμική κλίμακα. Σε αυτό το γράφημα παρατηρούνται δύο ασύμπτωτες όταν το $\varepsilon \to 0$ και $\varepsilon \to \infty$.
4. Επέλεξε το ε εκεί όπου το λογαριθμικό γράφημα εμφανίζεται να είναι γραμμικό (σχήμα 4.2).



Σχήμα 4.2 : Απεικόνιση του διαγράμματος επιλογής παραμέτρου ε για την μέθοδο των diffusion maps. Η κατάλληλη επιλογή ε επιτυγχάνεται εκεί που η καμπύλη γίνεται γραμμική (μπλε παραλληλόγραμμο).

4.4 Αλγόριθμος μεθόδου Diffusion Maps

Ψευδοκώδικας

Είσοδος :

- Πίνακας X διαστάσεων $n \times p$ όπου στις γραμμές του βρίσκονται οι μεταβλητές.
- Παράμετρος πυρήνα διάχυσης ε.
- Παράμετρος χρόνου t.
- Το νούμερο των μειωμένων διαστάσεων p που θα κρατηθούν.

1. Υπολόγισε τον πυρήνα διάχυσης $K[x_i, x_j] = e^{\left(-\frac{|x_i - x_j|^2}{\varepsilon}\right)}$.
2. Υπολόγισε τον διαγώνιο πίνακα D όπου τα στοιχεία στη διαγώνιο του είναι τα αθροίσματα των γραμμών του πίνακα K.
3. Υπολόγισε τον πίνακα $P = D^{-1}K$.
4. Βρες τα ιδιοδιανύσματα u_i του πίνακα P και τις αντίστοιχες ιδιοτιμές λ_i.
5. Αγνόησε την πρώτη ιδιοτιμή $(\lambda_0 = 1)$ και ιδιοδύανυσμα (u_1) του πίνακα P.
6. Υπολόγισε $Y = U\Lambda$ όπου U πίνακας με στήλες τα ιδιοδιανύσματα u_i και Λ πίνακας διαγώνιος με στοιχεία τα λ_i^t.

Έξοδος :

- Πίνακας Y μειωμένων διαστάσεων $n \times p$.

4.5 Εφαρμογή της μεθόδου των απεικονίσεων διάχυσης σε δεδομένα

4.5.1 Εισαγωγή

Για την πειραματική εφαρμογή της μεθόδου σε δεδομένα παράχθηκαν πέντε σύννεφα-ομάδες (clusters) αποτελούμενα από 200 σημεία το καθένα, τα οποία σε μια κλίμακα μεσαίου μεγέθους σχηματίζουν ένα διάσπαρτο ημι-ελικοειδές C-σχήμα με μία μόνο παράμετρο, αυτής της θέσης των σημείων πάνω σε αυτό. Η παράμετρος αποκρυπτογραφείται από τα χρώματα που έχουν δοθεί στις ομάδες (σχήμα 4.3).

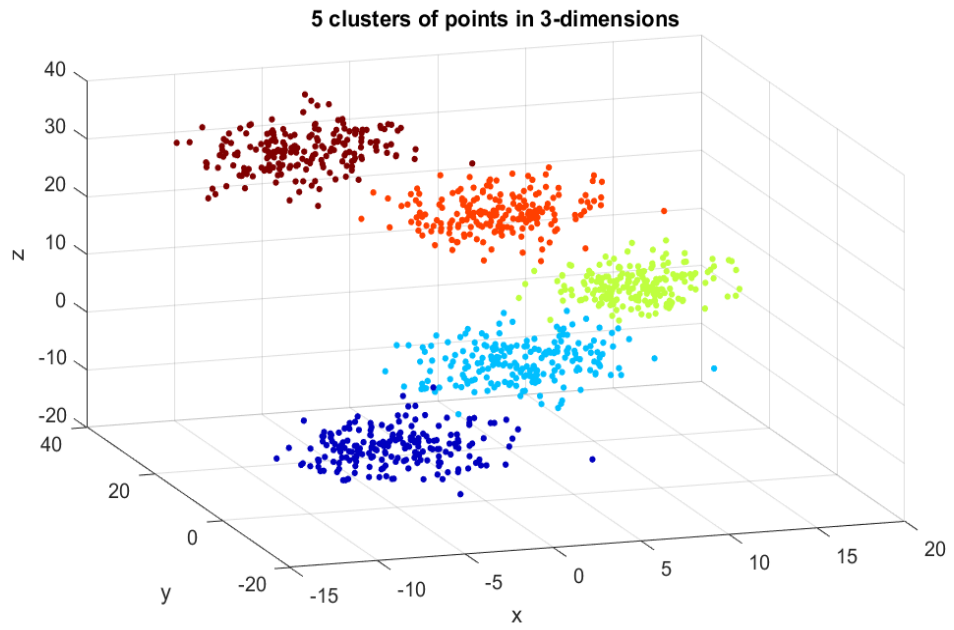
Οι εξισώσεις που το παρήγαγαν είναι οι ακόλουθες :

- Για τη πρώτη ομάδα : $\begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} = -9 + 5.2 \begin{pmatrix} r1_i \\ r2_i \\ r3_i \end{pmatrix} + 2.5 \begin{pmatrix} noise1_i \\ noise2_i \\ noise3_i \end{pmatrix}$ (4.39)
- Για τη δεύτερη ομάδα : $\begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} = 5.2 \begin{pmatrix} r1_i \\ r2_i \\ r3_i \end{pmatrix} + 2.5 \begin{pmatrix} noise1_i \\ noise2_i \\ noise3_i \end{pmatrix}$ (4.40)
- Για τη τρίτη ομάδα : $\begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} = 8 + 5.2 \begin{pmatrix} r1_i \\ r2_i \\ r3_i \end{pmatrix} + 1.8 \begin{pmatrix} noise1_i \\ noise2_i \\ noise3_i \end{pmatrix}$ (4.41)
- Για τη τέταρτη ομάδα : $\begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} = 2 + 5.2 \begin{pmatrix} r1_i \\ r2_i+16 \\ r3_i+16 \end{pmatrix} + 2.5 \begin{pmatrix} noise1_i \\ noise2_i \\ noise3_i \end{pmatrix}$ (4.42)
- Για τη πέμπτη ομάδα : $\begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} = -3 + 5.2 \begin{pmatrix} r1_i-5 \\ r2_i+26 \\ r3_i+32 \end{pmatrix} + 2.5 \begin{pmatrix} noise1_i \\ noise2_i \\ noise3_i \end{pmatrix}$ (4.43)

Όπου $r1_i, r2_i, r3_i, noise1_i, noise2_i, noise3_i$ σύνολο ομοιόμορφων κατανεμημένων τυχαίων αριθμών στο διάστημα $[0,1]$ για $i = 1, \dots, 200$ για την κάθε ομάδα.

Στόχος του πειράματος σε πρώτη φάση είναι η μείωση των διαστάσεων σε δύο διαστάσεις και μετέπειτα σε μία. Αναμένεται λόγω της μη γραμμικότητας που παρουσιάζουν τα δεδομένα η μέθοδος των diffusion maps να καταφέρει να μειώσει τις διαστάσεις, διατηρώντας την γεωμετρία του σχήματος στο νέο χώρο διάχυσης, διατηρώντας όσο το δυνατόν γίνεται τη παράμετρο της θέσης πάνω στο σχήμα. Δηλαδή την διατήρηση της σειράς των χρωμάτων όπως παρατηρούνται στον αρχικό χώρο.

Σε δεύτερη φάση στόχος του πειράματος είναι η ανάδειξη διαφορετικών γεωμετρικών δομών των δεδομένων καθώς το βήμα t αυξάνεται και επομένως μεγαλύτερα σε μήκος μονοπάτια σχηματίζονται μεταξύ τους. Αναμένεται επομένως οι αποστάσεις διάχυσης να μειωθούν και σα συνέπεια οι χώροι οι διάχυσης να εμφανίσουν τα δεδομένα σε ποιο συμπυκνωμένες δομές.

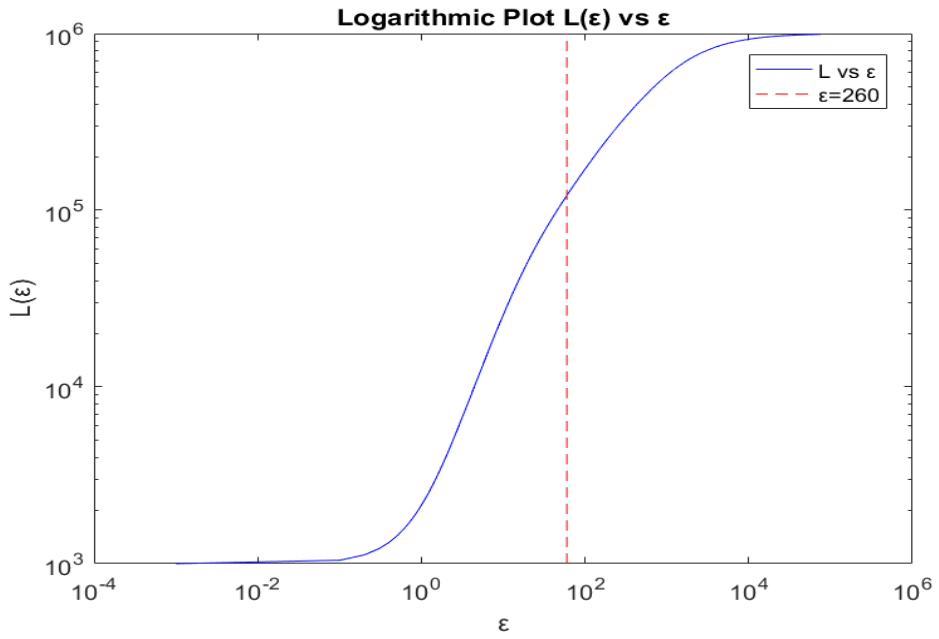


Σχήμα 4.3: Απεικόνιση 1000 σημείων σε τρισδιάστατο χώρο, σχήματος τύπου C-shape, αποτελούμενο από 5 ομάδες (clusters) συνολικού αριθμού 200 τυχαίων σημείων το καθένα αντίστοιχα.

4.5.2 Εύρεση παραμέτρου ε

Πριν γίνει η εφαρμογή της μεθόδου πρωταρχικός στόχος αποτελεί η εύρεση της παραμέτρου ε του πυρήνα διάχυσης. Σύμφωνα με την μεθοδολογία που αναπτύχθηκε στο κεφάλαιο 4 τα αποτελέσματα που πάρθηκαν συγκεντρώνονται στο σχήμα 4.4 όπου δείχνει μια επιλογή του ε (διακεκομμένη κόκκινη γραμμή) που έγινε μέσα από το εύρος των τιμών που υποδεικνύουν τα σημεία όπου η (μπλε) καμπύλη γίνεται γραμμική ανάμεσα στις δύο ασύμπτωτες. Η επιλογή του έγινε μέσω πειραματικών δοκιμών για τις τιμές αυτές που παρατηρείται η γραμμικότητα και ήταν ε = 260.

Είναι άξιο αναφοράς επίσης πως η επιλογή του ε με βάση την πρώτη τεχνική που αναλύθηκε στο κεφάλαιο 4 για την σωστή επιλογή παραμέτρου έδωσε μη ικανοποιητικά αποτελέσματα. Για την επιλογή ε = 1.456, η δομή των δεδομένων είχε χαθεί εξ ολοκλήρου εμφανίζοντας τα σημεία στον μειωμένο χώρο διάχυσης τυχαία.



Σχήμα 4.4 : Απεικόνιση λογαριθμικού σχεδιαγράμματος του L(ε) σε σχέση με το ε. χρησιμοποιώντας τον πίνακα πυρήνα W διαστάσεων 1000×1000 των δεδομένων του C-shape σχήματος.

4.5.3 Αποτελέσματα

Στη πρώτη εικόνα (σχήμα 4.5) παρουσιάζονται τα αποτελέσματα των δύο χώρων διάχυσης που έχει επιτευχθεί μείωση διαστάσεων σε δύο και μία αντίστοιχα. Όπως διαφαίνεται στην πρώτη περίπτωση οι δύο διαστάσεις έχουν διατηρήσει τη δομή του C-σχήματος και της εγγενούς γεωμετρίας των δεδομένων καθώς και την σειρά των χρωμάτων δηλαδή τις θέσεις των σημείων σε σχέση με τα υπόλοιπα. Στη δεύτερη εικόνα (σχήμα 4.5) παρατηρείται πως η μέθοδος είναι εξίσου ικανή να διατηρήσει τι σειρά των ομάδων (clusters) δηλαδή των χρωμάτων σε μια διάσταση με εξίσου πολύ μεγάλη επιτυχία.

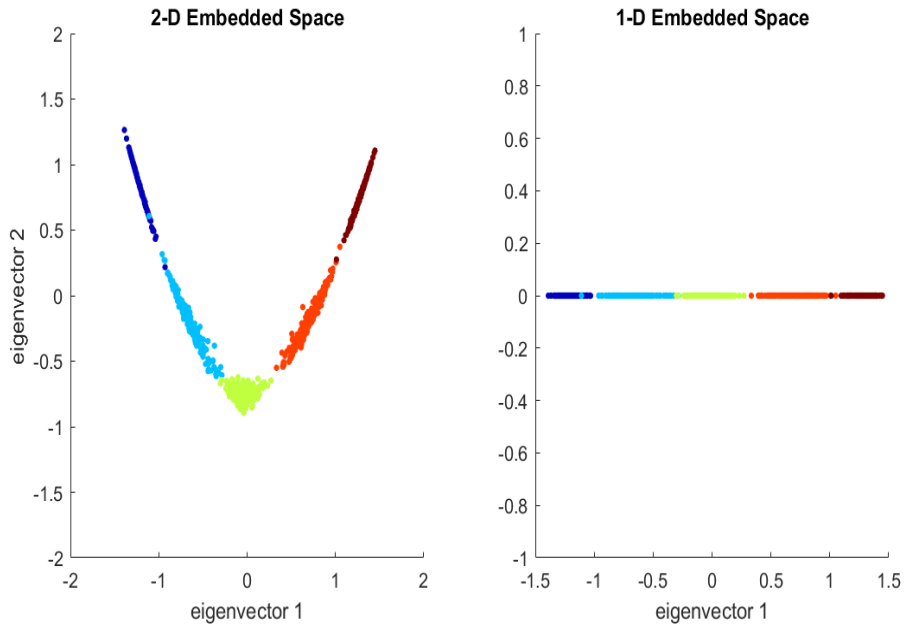
Σε δεύτερη φάση για την ανακάλυψη διαφορετικών γεωμετρικών δομών των diffusion maps χρησιμοποιήθηκαν οι χρόνοι t=1,3,10,25 και τα αποτελέσματα τους φαίνονται στο σχήμα 4.6.

Για $t = 1$, λαμβάνεται ο χώρος διάχυσης που εμφανίστηκε και πριν. Από την σκοπιά της μελέτης του βήματος t η επισήμανση που μπορεί να γίνει είναι πως η τοπική γεωμετρική δομή των ομάδων(clusters) διατηρείται όπως στον αρχικό χώρο, με τα clusters να διατηρούν τις αποστάσεις τους το ένα από το άλλο. Αυτό οφείλεται στο ότι το βήμα t

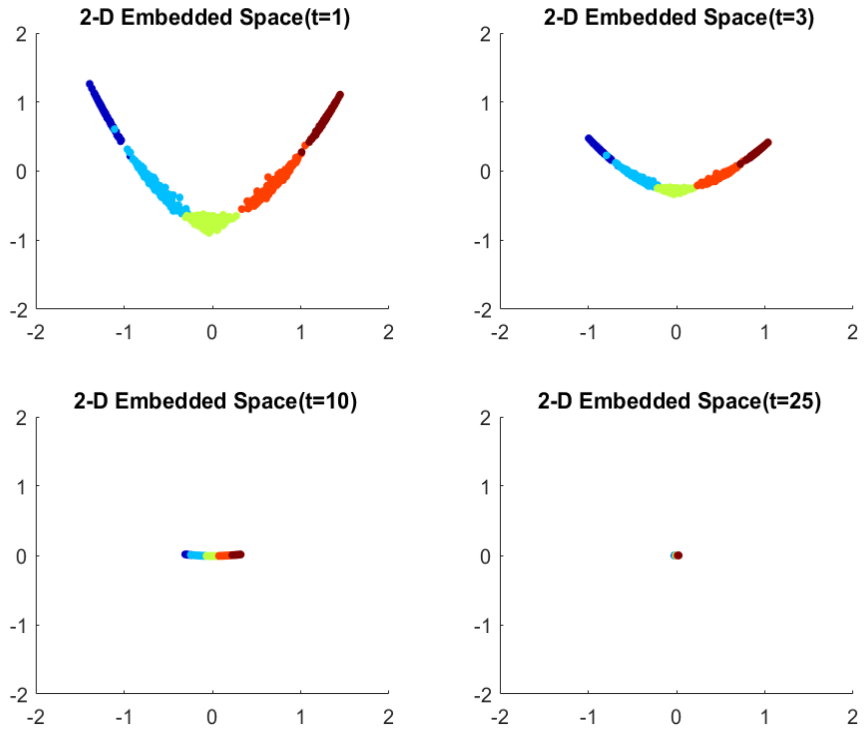
είναι πολύ μικρό, δημιουργώντας έτσι μικρής πιθανότητας μονοπάτια μεταξύ των σημείων.

Για $t = 10$, οι ομάδες (clusters) συνδέονται μεταξύ τους και σχηματίζουν μια ενιαία ομάδα (cluster) που μπορεί να χαρακτηριστεί κι από μία μόνο διάσταση καθώς τα σημεία πλέον έχουν σχηματίσει μία γραμμή έντονου πάχους. Αυτό συμβαίνει καθώς τα μονοπάτια πλέον για t=3 αυξήθηκαν και οι πιθανότητες μετάβασης από το ένα σημείο στο άλλο μεγάλωσαν αισθητά. Οι αποστάσεις διαχύσεις μεταξύ των ομάδων μειώθηκαν αντιστρόφως.

Για t=25, μία τρίτη γεωμετρική δομή εμφανίστηκε. Οι πέντε ομάδες σχημάτισαν μια ύπερ -ομάδα (super-cluster). Σε αυτή τη κλίμακα όλα τα σημεία στον αρχικό χώρο είναι συνδεδεμένα απόλυτα και σαν συνέπεια αυτό να δείχνει τις αποστάσεις διάχυσης πολύ μικρές. Κάτι που στον χώρο διάχυσης σημαίνει πως οι ευκλείδειες αποστάσεις μεταξύ των σημείων είναι μηδενικές και άρα σχηματίζουν ένα σημείο.



Σχήμα 4.5 : Απεικόνιση 1000 σημείων του σχήματος C-shape σε χώρους μειωμένων διαστάσεων δύο και μίας αντίστοιχα, μετά από εφαρμογή των diffusion maps (ε=260,t=1).



Σχήμα 4.6 : Απεικόνιση1000 σημείων σχήματος C-shape σε χώρους δύο διαστάσεων για χρόνους t=1,3,10,25 (ε=260) από αριστερά στα δεξιά.

Κεφάλαιο 5

Συγκριτική Ανάλυση Μεθόδων

Σε αυτό το κεφάλαιο, θα πραγματοποιηθεί μια συγκριτική ανάλυση μεταξύ των αλγορίθμων που παρουσιάστηκαν μέχρι στιγμής. Η ανάλυση τους θα γίνει πάνω σε ένα σύνολο δεδομένων που σχηματίζει, σε ένα τρισδιάστατο χώρο μία τοροειδή έλικα (Toroidal Helix), (σχήμα 5.1). Οι εξισώσεις που το παρήγαγαν είναι οι ακόλουθες :

$$\begin{cases} x_i = (0.2\sin(8t_i) + 0.3)\cos(t_i) \\ y_i = (0.2\sin(8t_i) + 0.3)\sin(t_i) \\ z_i = 0.2\cos(t_i) \end{cases}$$

Όπου $t_i \in [0,2\pi],\ i = 1, \dots, 1571$

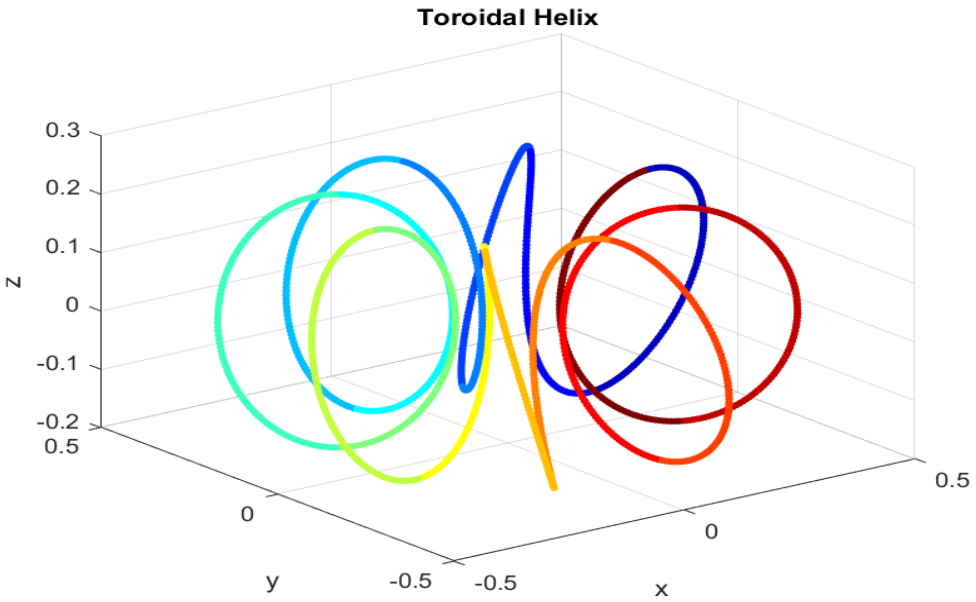
Η αποτελεσματικότητα της κάθε μεθόδου θα κριθεί, με βάση την εύρεση της γεωμετρικής δομής, στον μειωμένο χώρο δύων διαστάσεων, με απώτερο σκοπό τον σχηματισμό ενός κύκλου.

Σε πρώτη φάση εφαρμόστηκε η μέθοδος της PCA στα δεδομένα του παραδείγματος. Στο σχήμα 5.2 διαφαίνεται η απεικόνιση του μειωμένου χώρου δύο διαστάσεων της μεθόδου. Είναι σαφές, πως η μέθοδος απέτυχε εξ ολοκλήρου να σχηματίσει το σχήμα του κύκλου που αναζητείται εμφανίζοντας ένα αστεροειδές σχήμα. Οι μεγαλύτερες διασπορές του αρχικού χώρου παρατηρούνται στους άξονες x, y και για αυτό αποτελούν τις κύριες συνιστώσες (principal components) του χώρου εμφύτευσης (embedding). (Η μέθοδος της CMDS εφαρμόστηκε και αυτή με τα αποτελέσματα της, να παρουσιάζονται ακριβώς τα ίδια με αυτά της PCA. Για αυτό το λόγω μελετάται μόνο το παράδειγμα της δεύτερης).

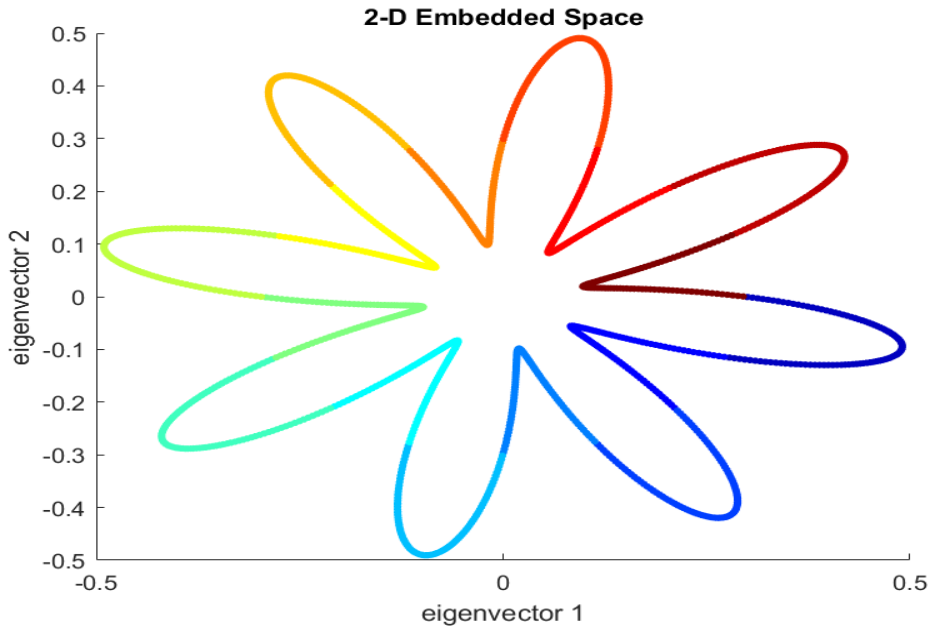
Σε δεύτερη φάση εφαρμόστηκε η μέθοδος ISOMAP. Αρχικά παράχθηκε το διάγραμμα διακύμανσης υπολοίπων για τις τιμές της παραμέτρου : k=6,12,24,48. Παρατηρώντας το σχήμα 5.3 είναι αρκετά ισχυρό το συμπέρασμα, πως αναμένεται η χρησιμοποίηση των τιμών 6, 12 να αποφέρουν ικανοποιητικά αποτελέσματα για την μείωση του χώρου σε δύο διαστάσεις. Όντως όπως διαφαίνεται στους χώρους μειωμένων διαστάσεων (σχήμα 5.4), η μέθοδος κατάφερε με πολύ μεγάλη επιτυχία να αναγνωρίσει την πολλαπλότητα των δεδομένων για k=6, k=12 και παρήγαγε κύκλους στους δύο διαστάσεων χώρους εμφύτευσης. Καθώς το k αυξάνεται η μέθοδος δείχνει να έχει πρόβλημα στην ανίχνευση

(k=24, k=48) και να προσεγγίζει όλο και περισσότερο την λύση της PCA. Αυτό όπως έχει τονιστεί και ποιο πριν, συμβαίνει καθώς όσο μεγαλώνει η παράμετρος k τόσα περισσότερα σημεία συνδέονται και άρα τόσο περισσότερο τα γεωδεσιακά μονοπάτια τείνουν να γίνουν, οι απλές ,ευκλείδειες αποστάσεις.

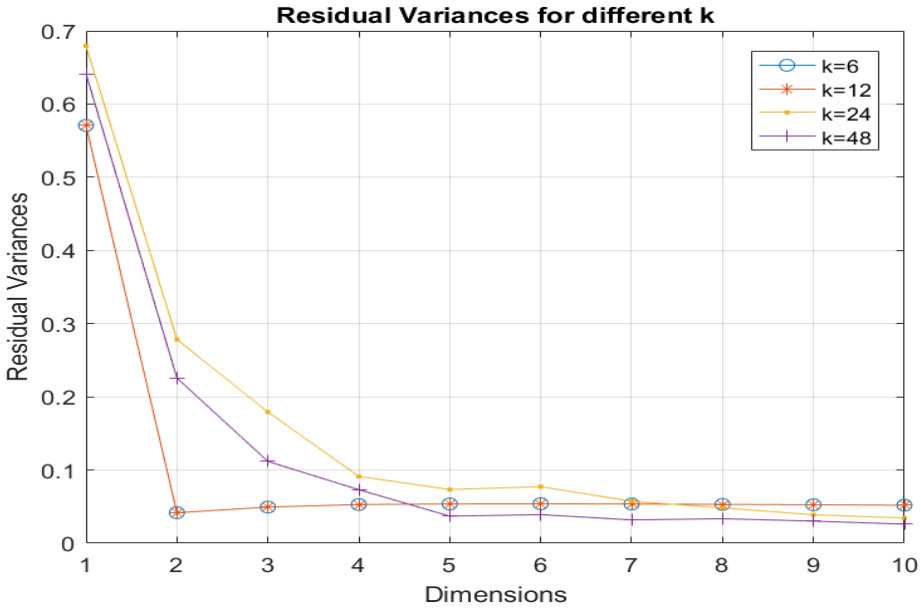
Τέλος, χρησιμοποιήθηκε η μέθοδος των απεικονίσεων διάχυσης για την εύρεση της πολλαπλότητας των δεδομένων που εξετάζονται. Σε πρώτη φάση, έγινε η μελέτη για την σωστή επιλογή της παραμέτρου ε και παράχθηκε το διάγραμμα επιλογής της, με βάση τη μεθοδολογία που χρησιμοποίησε ο **Singer (2007)** . Όπως διαφαίνεται στο σχήμα 5.5 η παράμετρος ε πήρε την τιμή 0.00065. Τα αποτελέσματα που πάρθηκαν ήταν άκρως ικανοποιητικά για την τιμή αυτή, όπως φανερώνονται στο σχήμα 5.6 καθώς η αναδίπλωση της τοροειδής έλικας πραγματοποιήθηκε με επιτυχία, διαμορφώνοντας στον χώρο διάχυσης δύο διαστάσεων, έναν κύκλο. Σημαντική παρατήρηση είναι, πως στα δεδομένα που εξετάζονται, η ακριβής επιλογή της παραμέτρου παίζει πολύ μεγάλή βαρύτητα, καθώς μικρές αποκλίσεις της παραμέτρου, δεν φέρνουν τα αναμενόμενα αποτελέσματα (σχήμα 5.6)



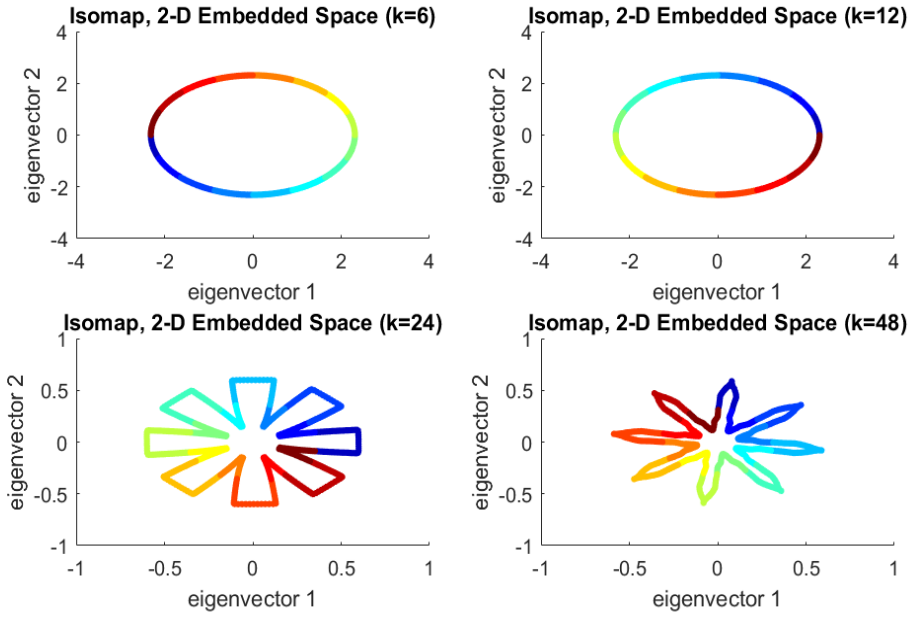
Σχήμα 5.1 : Απεικόνιση τοροειδής έλικας (Toroidal Helix) αποτελούμενη από N=1571 σημεία.



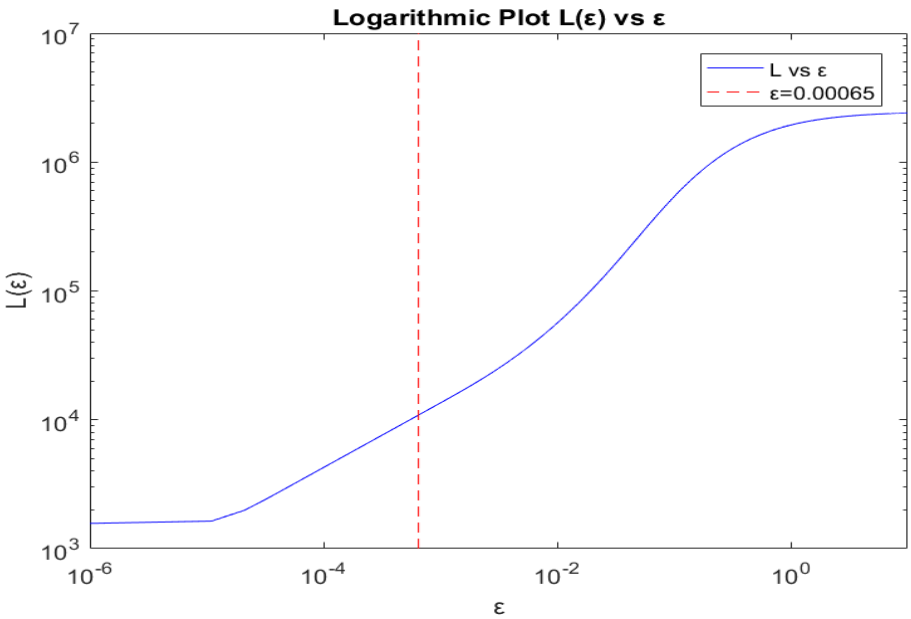
Σχήμα 5.2 : Απεικόνιση 1571 σημείων της τοροειδής έλικας σε χώρο δύο μειωμένων διαστάσεων , μετά από εφαρμογή της PCA.



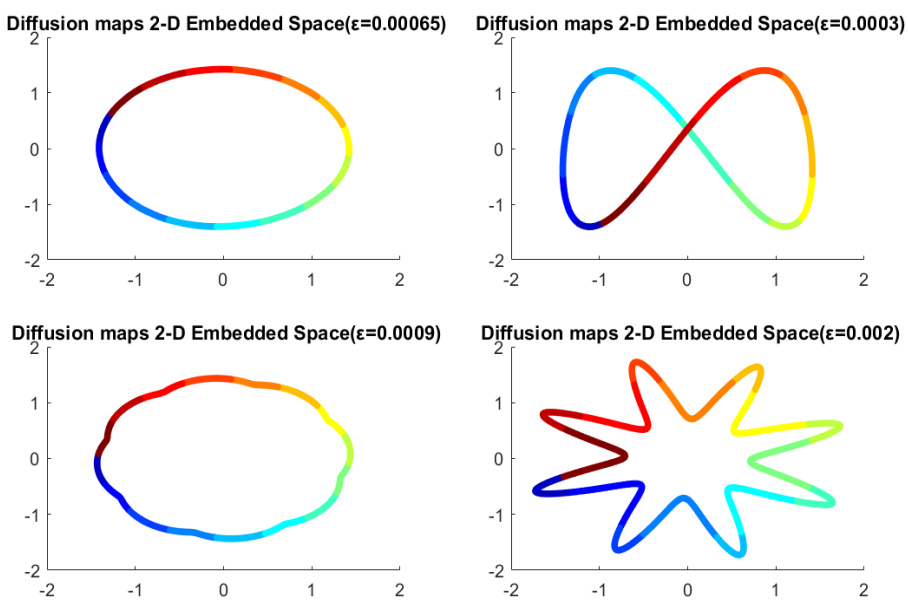
Σχήμα 5.3 : Απεικόνιση του διαγράμματος, διακυμάνσεων υπολοίπων για διάφορες τιμές της παραμέτρου k, πάνω στα δεδομένα της τοροειδής έλικας.



Σχήμα 5.4 : Απεικόνιση 1571 σημείων του της τοροειδής έλικας, σε χώρους δύο μειωμένων διαστάσεων, μετά από εφαρμογή των γεωμετρικών συνιστωσών, για παράμετρο k=6,12,24,48 από αριστερά στα δεξιά.



Σχήμα 5.5 : Απεικόνιση λογαριθμικού διαγράμματος του L(ε) σε σχέση με το ε χρησιμοποιώντας τον πίνακα πυρήνα W, διαστάσεων 1571×1571 των δεδομένων της τοροειδής έλικας.



Σχήμα 5.6 : Απεικόνιση 1571 σημείων της τοροειδής έλικας σε χώρους δύο μειωμένων διαστάσεων, μετά από εφαρμογή των απεικονίσεων διάχυσης, για παράμετρο ε=0.00065, 0.0003, 0.0009, 0.002 από αριστερά στα δεξιά.

ΜΕΡΟΣ ΔΕΥΤΕΡΟ

Εφαρμογή της μεθόδου των απεικονίσεων διάχυσης (Diffusion Maps) σε χρηματοοικονομικά δεδομένα

Κεφάλαιο 6

Χρησιμοποίηση της μεθόδου των Diffusion Maps, για την ανάδειξη συγκεντρώσεων χαρτοφυλακίου και της γεωμετρίας της συμμεταβολής των μετοχών του.

6.1 Πρόλογος

Στόχος του δεύτερου μέρους της εργασίας, είναι να αναδειχθεί η τεχνική μείωσης διαστάσεων των απεικονίσεων διάχυσης, χρησιμοποιώντας την σε πραγματικά δεδομένα μετοχών. Πέρα από τη εύρεση των ίδιων χώρων διάχυσης (χώροι μειωμένων διαστάσεων) και την εξαγωγή χρήσιμων συμπερασμάτων μέσω της εξερεύνησης τους και οπτικοποίησης των μετοχών, αναδεικνύεται ακόμα η χρησιμοποίησή τους, έτσι ώστε να παρθούν χρήσιμα ποσοτικά οικονομικά μέτρα. Για το σκοπό αυτό, έγινε ολική αναπαραγωγή (με διαφορετικής μορφής δεδομένα), της δημοσίευσης του *Wesley Phoa «Portfolio Concentration and the Geometry of Co-Movement (May 29, 2012). The Journal of Portfolio Management, Forthcoming.*

Available at SSRN: https://ssrn.com/abstract=2108339.

6.2 Εισαγωγή

Η κατανόηση της συμμεταβολής διαφόρων μετοχών, συμβάλλοντας στην κατασκευή χαρτοφυλακίων καθώς και στον έλεγχο του ρίσκου, διαδραματίζει σπουδαίο ρόλο στον τομέα τις διαχείρισης χαρτοφυλακίων. Για μικρές ποσότητες μετοχών η λύση του προβλήματος αποκτά μεγάλη απλότητα, καθώς μελετώντας και αναλύοντας τον πίνακα συσχετίσεων (correlation matrix), η απευθείας εξαγωγή συμπερασμάτων για την συμμεταβολή μεταξύ τους είναι απολύτως εφικτή.

Στην αντίθετη περίπτωση, όταν ο αριθμός των μετοχών αυξάνεται ραγδαία παρόλο που η κατασκευή χαρτοφυλακίων μέσω κάποιας στατιστικής ανάλυσης είναι εφικτή, η πλήρης κατανόηση του πίνακα συσχετίσεων και οι γεωμετρικές δομές που κρύβονται σε αυτόν, καθιστά τον ιδίων αναξιόπιστη πηγή.

Κατά αυτόν τον τρόπο η χρήση μεθόδων μείωσης διαστάσεων για την οπτικοποίηση των δεδομένων, καθίσταται αναγκαία. Η εξαγωγή διαγραμμάτων, με τα σημεία να συμβολίζουν μετοχές και η μελέτη αυτών θα οδηγήσει στην άμεση κατανόηση των γεωμετρικών δομών που διέπουν τα δεδομένα. Τη λύση σε αυτό το πρόβλημα δίνει η μέθοδος των diffusion maps η οποία μεταφράζει τις συσχετίσεις των μετοχών στον αρχικό χώρο με τις ευκλείδειες αποστάσεις στον χώρο διάχυσης. Όσο δύο μετοχές είναι θετικά συσχετισμένες, τόσο πιο κοντά θα βρίσκονται στον χώρο μειωμένων διαστάσεων.

Η σημαντικότητα της μεθόδου δε παραμένει μόνο στην ομαδοποίηση κατά τρόπο τινά των μετοχών αλλά επεκτείνεται και περαιτέρω. Μέσω των χώρων διάχυσης, όπως θα δειχθεί στη συνέχεια, η εξαγωγή ποσοτικών μέτρων αναδεικνύουν μη ορατές συγκεντρώσεις ρίσκου στο χαρτοφυλάκιο που μελετάται.

6.3 Περιγραφή δεδομένων

Πριν γίνει η αναφορά στα δεδομένα, κρίνεται απαραίτητη η προεπισκόπηση κάποιων χρηματοοικονομικών όρων που θα συναντηθούν στην συνέχεια της εργασίας για την διευκόλυνση της κατανόησης των αποτελεσμάτων που θα εξαχθούν.

6.3.1 Χρηματοοικονομικοί Όροι

<u>Αποδόσεις μετοχών</u>

Οι περισσότεροι ερευνητές ανά το κόσμο χρησιμοποιούν αποδόσεις παρά τιμές μετοχών. Οι δύο κυριότεροι λόγοι που δόθηκαν από τους Campbell, Lo και McKinley (1997) είναι πως το μεγαλύτερο ποσοστό των ερευνητών, χρησιμοποιούν τις αποδόσεις διότι χρήζουν καλύτερων στατιστικών συμπερασμάτων και κατά δεύτερων θεωρούνται μια ολοκληρωμένη, χωρίς κλίμακα σύνοψη μιας επενδυτικής ευκαιρίας.

Έστω P_t η τιμή μιας μετοχής τη χρονική περίοδο t, τότε ορίζονται ακολούθως :

Απόδοση σε μία περίοδο (One Period Single Return):

Κρατώντας μια μετοχή από την ημερομηνία $t-1$ έως t, το αποτέλεσμα θα είναι ένα single gross Return :

$$1 + R_t = \frac{P_t}{P_{t-1}} \Leftrightarrow P_t = P_{t-1}(1 + R_t) \tag{6.1}$$

Και το αντίστοιχο one period simple net Return :

$$R_t = \frac{P_t}{P_{t-1}} - 1 = \frac{P_t - P_{t-1}}{P_{t-1}} \tag{6.2}$$

Continuously Compounded Return :

Ο φυσικός λογάριθμος των single gross Return καλείται Continuously compounded returns η απλά log-Return :

$$r_t = \ln(1 + R_t) = \ln\frac{P_t}{P_{t-1}} = p_t - p_{t-1} \quad \text{με } p_t = \ln(P_t) \tag{6.3}$$

Τα log-Return διακατέχονται από κάποιες ισχυρές ιδιότητες σε αντίθεση με τα single gross Return και είναι προτιμότερη η χρησιμοποίηση τους.

Απόδοση χαρτοφυλακίου (Portfolio Return) :

Η καθαρή απόδοση (net return) ενός χαρτοφυλακίου αποτελούμενο από N το πλήθος μετοχών είναι ο σταθμισμένος μέσος των καθαρών αποδόσεων των μετοχών του χαρτοφυλακίου με τα βάρη κάθε μετοχής να είναι το ποσοστό που διακατέχει η εκάστοτε μετοχή στο χαρτοφυλάκιο.

$$E(R_t) = \sum_{i}^{N} w_i R_{it} \tag{6.4}$$

Συντελεστής συσχετισμού(Correlation Coefficient) :

Ο συντελεστής συσχέτισης μεταξύ δύων τυχαίων μεταβλητών X, Y ορίζεται ως:

$$\rho(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}} = \frac{E(X - \mu_X)E(Y - \mu_Y)}{\sqrt{E(X - \mu_X)^2 E(Y - \mu_Y)^2}} \tag{6.5}$$

Όπου μ_X, μ_Y οι μέσοι των X, Y αντίστοιχα. Ο συντελεστής συσχέτισης μετράει την δύναμη της γραμμικής εξάρτησης μεταξύ δύο μεταβλητών και δέχεται τιμές στο διάστημα [-1,1] καθώς και ικανοποιεί την εξίσωση $\rho(X, Y) = \rho(Y, X)$. Δύο τυχαίες μεταβλητές είναι ανεξάρτητες αν και μονο αν $\rho(X, Y) = 0$.

Αντίστοιχα ο δειγματικός συντελεστής συσχέτισης για δείγμα $\{(x_t, y_t)\}_{t=1}^{T}$ είναι ο ακόλουθος :

$$\hat{\rho}(x, y) = \frac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{T}(x_t - \bar{x})^2 \sum_{t=1}^{T}(y_t - \bar{y})^2}} \tag{6.6}$$

Όπου $\bar{x} = \frac{\sum_{t=1}^{T} x_t}{T}$, $\bar{y} = \frac{\sum_{t=1}^{T} y_t}{T}$ οι δειγματικοί μέσοι του x, y αντίστοιχα.

<u>Κεφαλαιοποίηση Αγοράς (Market Capitalization)</u>

Κεφαλαιοποίηση αγοράς είναι η χρηματιστηριακή αξία μιας επιχείρησης που προκύπτει, εάν πολλαπλασιαστεί ο εκάστοτε αριθμός των μετοχών της (Market Value) επί την τρέχουσα χρηματιστηριακή αξία (τιμή) της μετοχής :

$$MK_t = MV * P_t \tag{6.7}$$

Η συνολική κεφαλαιοποίηση ενός χρηματιστηρίου προκύπτει, εάν προστεθούν όλες οι χρηματιστηριακές αξίες των μετοχών, όλων των εισηγμένων εταιριών.

<u>Standard & Poor's 500 δείκτης (Index) (S&P 500)</u>

Ο Standard & Poor's 500 Index (S&P 500) είναι ένας δείκτης που περιλαμβάνει 500 κορυφαίες αμερικάνικες μετοχές (υψηλής κεφαλαιοποίησης) και η διακύμανση της τιμής του θεωρείται ως μια σημαντική ένδειξη για τις μετοχές των ΗΠΑ και κατ' επέκταση της παγκόσμιας οικονομίας συνολικά.

Θεωρείται ευρέως, ως η πιο ακριβής μέτρηση απόδοσης της μεγάλης κεφαλαιοποίησης των αμερικανικών μετοχών. Ο S&P θεωρείται αντιπροσωπευτικός δείκτης της αγοράς, διότι περιλαμβάνει ένα σημαντικό τμήμα της συνολικής αξίας της.

Οι 500 εταιρείες που περιλαμβάνονται σε αυτόν έχουν επιλεγεί από την Επιτροπή του Δείκτη, μια ομάδα αναλυτών και οικονομολόγων της Standard & Poor's (του γνωστού Οίκου αξιολόγησης). Αυτοί οι εμπειρογνώμονες εξετάζουν διάφορα στοιχεία για να συμπεριλάβουν μια μετοχή στο δείκτη, συμπεριλαμβανομένου του μεγέθους της αγοράς, της ρευστότητας και της βιομηχανίας που δραστηριοποιείται η εταιρεία.

6.3.2 Δεδομένα

Για τη χρησιμοποίηση της μεθόδου των diffusion maps πάνω σε χρηματοοικονομικά δεδομένα, αναλύθηκαν οι αποδόσεις των συστατικών του δείκτη S&P 100 και S&P 500 αντίστοιχα για τη χρονική περίοδο από τον Ιανουάριο 2002 έως και τον Μάρτιο 2012. Οι τιμές των μετοχών για εκείνη την περίοδο προσφέρθηκαν μέσω της διαδικτυακής σελίδας https://finance.yahoo.com και τα αντίστοιχα Market Caps της εκάστοτε μετοχής μέσω της http://siblisresearch.com. Οι αντίστοιχοι τομείς των μετοχών πάρθηκαν από την διαδικτυακή σελίδα http://www.nasdaq.com.

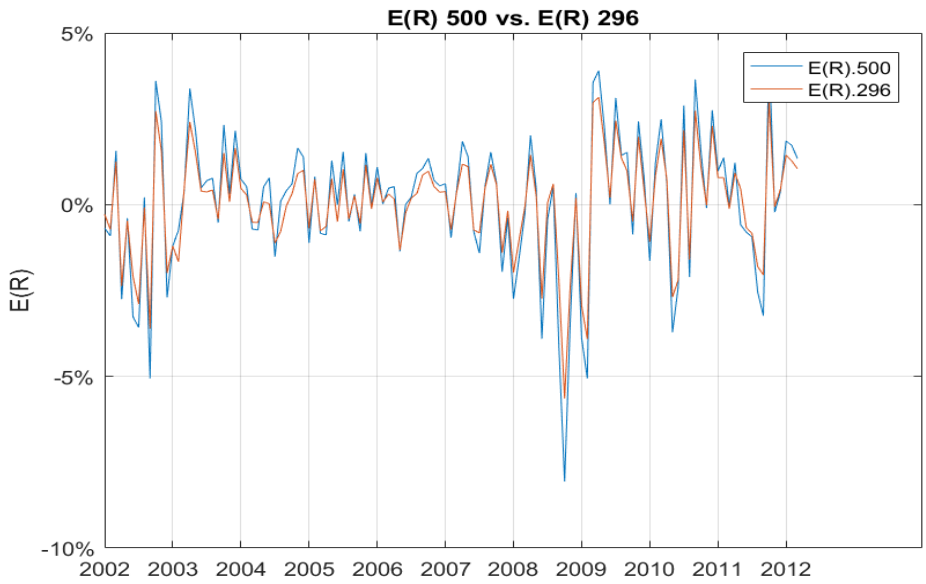
Μέσω των δεδομένων αυτών αντλήθηκαν και τα υπόλοιπα στοιχεία που κρίνονται απαραίτητα για την μετέπειτα εξαγωγή κάποιων μέτρων που θα προκύψουν, όπως είναι αυτά των λογαριθμικών αποδόσεων και των βαρών κάθε μετοχής, για τον εκάστοτε μήνα. Αυτά προκύπτουν από την απλή εφαρμογή του τύπου (3.3) και της εξίσωσης :

$$\mathrm{w_i} = \frac{\mathrm{MK_i}}{\mathrm{MK_{total}}} \tag{6.8}$$

Για τον δείκτη του S&P 100 οι μετοχές που μετείχαν σε αυτόν, κατά τη διάρκεια τής δεκαετίας που μελετιέται, όπως γίνεται αντιληπτό θα είναι παραπάνω από 100 σε μέγεθος λόγω της εισαγωγής και εξαγωγής μετοχών. Επομένως οι μετοχές αυτού του δείκτη θα υπερβαίνουν τις 100 στην εκάστοτε εφαρμογή τους.

Λόγω πολλών ελλιπών παρατηρήσεων μετοχών η ανάλυση των συστατικών του S&P 500 περιορίστηκε σε αυτές οι οποίες έδιναν πλήρη στοιχεία για την περίοδο που μελετιέται. Για την ακρίβεια συμπεριλήφθηκαν οι μετοχές που είχαν στο σύνολο τους πάνω από 120 μήνες παρατηρήσεις και το πλήθος τους ανερχόταν στις 296. Ο λόγος που πραγματοποιήθηκε αυτή η αρκετά μεγάλη μείωση των μετοχών, είναι διότι η κατάλληλη εξαγωγή μέτρων εξαρτάται άμεσα από το πλήθος τους που πρέπει να διατηρείται σταθερό κατά τη διάρκεια του χρόνου και από την ευαισθησία του δείκτη συσχέτισης στις ελλιπής τιμές. Στην ουσία οι μετοχές του δείκτη του S&P 500 μεταμορφώθηκαν σε ένα κατά τρόπο μικρότερο χαρτοφυλάκιο που οι μετοχές του ίδιου του χαρτοφυλακίου μετείχαν ως συστατικά του δείκτη, καθ' όλη τη διάρκεια της δεκαετίας.

Όπως διαφαίνεται στο σχήμα 6.1 οι αποδόσεις του περιορισμένου χαρτοφυλακίου, τείνουν σε αρκετά ικανοποιητικό βαθμό τις αποδόσεις του ολικού. Επίσης ο δειγματικός συντελεστής συσχέτισης για αυτές τις δύο χρονοσειρές ισούται με $\hat{\rho} = 0.9867$. Κατά αυτόν τον τρόπο η αντιπροσώπευση του δείκτη μέσω του μειωμένου χαρτοφυλακίου κρίνεται ικανοποιητική.



Σχήμα 6.1 : Οι καθαρές αποδόσεις του δείκτη S&P 500 και αυτές του μειωμένου χαρτοφυλακίου, αποτελούμενο από συστατικά του S&P 500 ($\hat{\rho} = 0.9867$).

6.4 Από τον πίνακα συσχετίσεων στους χώρους διάχυσης.

6.4.1 Πίνακας συσχετίσεων (correlation matrix)

Όπως έγινε αντιληπτό από τη μεθοδολογία των απεικονίσεων διάχυσης, δοθέντος ενός πίνακα πυρήνα ο οποίος ικανοποιεί την ιδιότητά της συμμετρίας και έχοντας στοιχεία θετικά, μπορεί να παρέχει μέσω της μεθόδου, απεικονίσεις δύο ή τριών διαστάσεων των δεδομένων που μελετιούνται. Ένα από τα ποιο κοινά και διαδεδομένα μέτρα για τον προσδιορισμό της συμμεταβολής μεταξύ μετοχών είναι αυτό του συντελεστή συσχέτισης μεταξύ τους. Όσο μεγαλύτερος είναι ο συντελεστής τόσο μεγαλύτερη η συμμεταβολή μεταξύ των μετοχών και όσο μικρότερος είναι, τόσο μικρότερη αντίστοιχα. Για την εφαρμογή λοιπόν του πίνακα συσχέτισης (correlation matrix) και την θεώρησή του ως πυρήνα το κριτήριο της θετικότητας των στοιχείων του δεν ικανοποιείται. Για το σκοπό αυτό δημιουργείται ο πυρήνας $\mathrm{K}(\mathrm{x},\mathrm{y}) = 1 + \rho(\mathrm{x},\mathrm{y})$, ο οποίος πληρεί και το κριτήριο της συμμετρικότητας και τα στοιχεία του είναι θετικά εξ ολοκλήρου, καθώς οι τιμές του θα κυμαίνονται στο διάστημα $[0,2]$.

Κατά αυτόν τον τρόπο υπολογίζοντας τον νέο $\mathrm{K}(\mathrm{x},\mathrm{y})$ και εφαρμόζοντας την μέθοδο των diffusion maps, η δημιουργία διαγραμμάτων με σημεία τις μετοχές στον χώρο επιτυγχάνεται. Όσο μεγαλύτερός ο συντελεστής συσχέτισης τόσο πιο μικρή η ευκλείδεια

απόσταση στο νέο αυτό χώρο διάχυσης και αντίστροφα. Η διαδικασία ανεύρεσης των διαγραμμάτων ακολουθεί το πρότυπο αλγορίθμου που δίνεται στο κεφάλαιο (4.7).

6.4.2 Χώροι διάχυσης

Σε πρώτη φάση αναλύθηκαν οι μετοχές (συστατικά) του δείκτη S&P 100 καθόλη τη διάρκεια της δεκαετίας. Στο σχήμα 6.2 φαίνεται, ο δύο διαστάσεων χώρος διάχυσης των δεδομένων αυτών, λαμβάνοντας υπόψη τις πρώτες δύο συντεταγμένες της εκάστοτε μετοχής, που δίνονται από τα αποτελέσματα του αλγορίθμου diffusion maps. Ο δύο διαστάσεων χάρτης, ενώ δίνει ικανοποιητικά αποτελέσματα και παρέχει την κατάλληλη πληροφόρηση για την γεωμετρία της συν μετακίνησης των μετοχών, μπορεί να οδηγήσει σε παραπλανητικά αποτελέσματα.

Κατά αυτόν τον τρόπο δημιουργήθηκαν οι τριών διαστάσεων απεικονίσεις χρησιμοποιώντας τις πρώτες τρείς συντεταγμένες της εκάστοτε μετοχής και διαφαίνονται στα σχήματα 6.3 και 6.4 με διαφορετική οπτική. Είναι πασιφανές πως οι απλές εικονίσεις αυτές δίνουν τον τρόπο που μπορούν να εξαχθούν συμπεράσματα για την εν λόγω εργασία. Η πλήρης κατανόηση της συμμεταβολής των μετοχών τους, προέρχεται από την δια δραστικότητα αυτών των χαρτών εξερευνώντας τους από όλες τις μεριές οπτικής. Η κατηγοριοποίηση των μετοχών επιτυγχάνεται δίνοντας διαφορετικά χρώματα με βάση τομέα που δραστηριοποιείται η εκάστοτε εταιρεία και παρουσιάζονται στο Σχήμα 6.5.

6.4.3 Αποτελέσματα

Τα αποτελέσματα των σχημάτων (6.2), (6.3), (6.4) υποδεικνύουν, κάποια εμφανή και κάποια όχι τόσο, αποτελέσματα τα οποία αναφέρονται ακολούθως :

- Εταιρείες που προέρχονται από τον ίδιο τομέα συχνά σχηματίζουν ομάδες μεταξύ τους, όπως για παράδειγμα οι εταιρείες ενέργειας Devon Energy (DVN), Occidental Petroleum (OXY), Apache (APA), αριστερά κάτω του σχήματος 6.2 Ιατροφαρμακευτικές εταιρείες όπως η Merck (MRK), Baxter International (BAX), Abbott Laboratories (ABT) κ.α. παρατηρούνται στη κορυφή του σχήματος καθώς και κάποιες στο τομέα των χρηματοοικονομικών στο δεξί, όπως οι Bank of America (BAC), Wells Fargo (WFC) και US Bancorp (USB). Εν τέλει κάποιες του τεχνολογικού τομέα καταλαμβάνουν θέσεις στο κέντρο του διαγράμματος όπως οι Texas Instruments (TXN), Dell (DELL), Ibm (IBM).

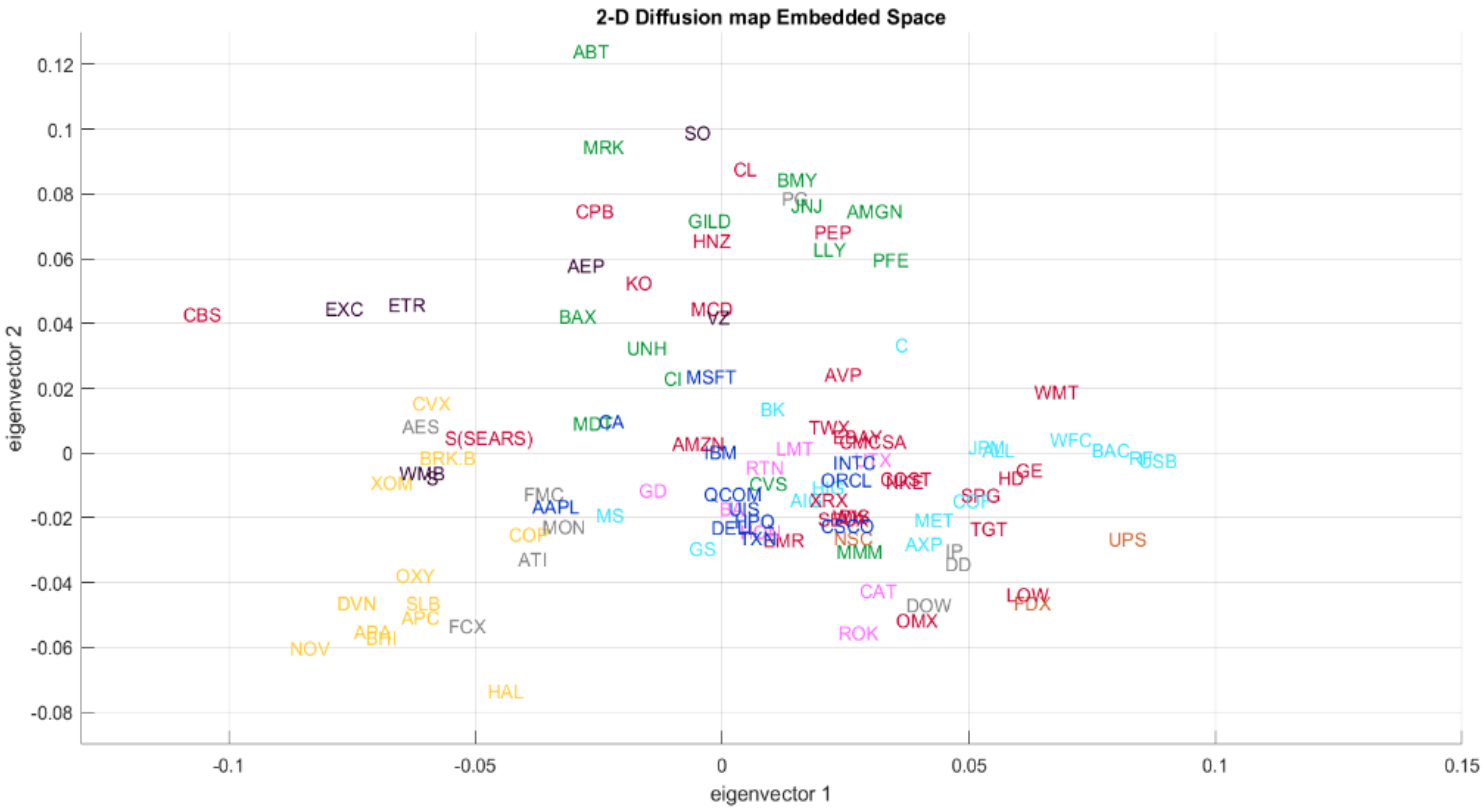
- Οι ομάδες των μετοχών έχουν διαφορετικές τοποθεσίες με βάση το κέντρο του διαγράμματος. Για παράδειγμα οι εταιρείες που δραστηριοποιούνται στο τομέα της τεχνολογίας καταλαμβάνουν θέση στο κέντρο, ενώ οι εταιρείες ενέργειας σε σχέση με της χρηματοοικονομικές διαχωρίζονται εξ ολοκλήρου. Οι μεν καταλαμβάνουν το αριστερό μέρος και οι δε το δεξί.

- Κάποιες εταιρείες που δραστηριοποιούνται στον ίδιο τομέα, παρατηρείται πώς δεν σχηματίζουν απαραίτητα ομαδοποιήσεις. Για παράδειγμα η εταιρεία Apple (AAPL) φαίνεται να έχει σχετικά μεγάλη απόσταση από την εταιρεία Gilead Sciences (GILD).

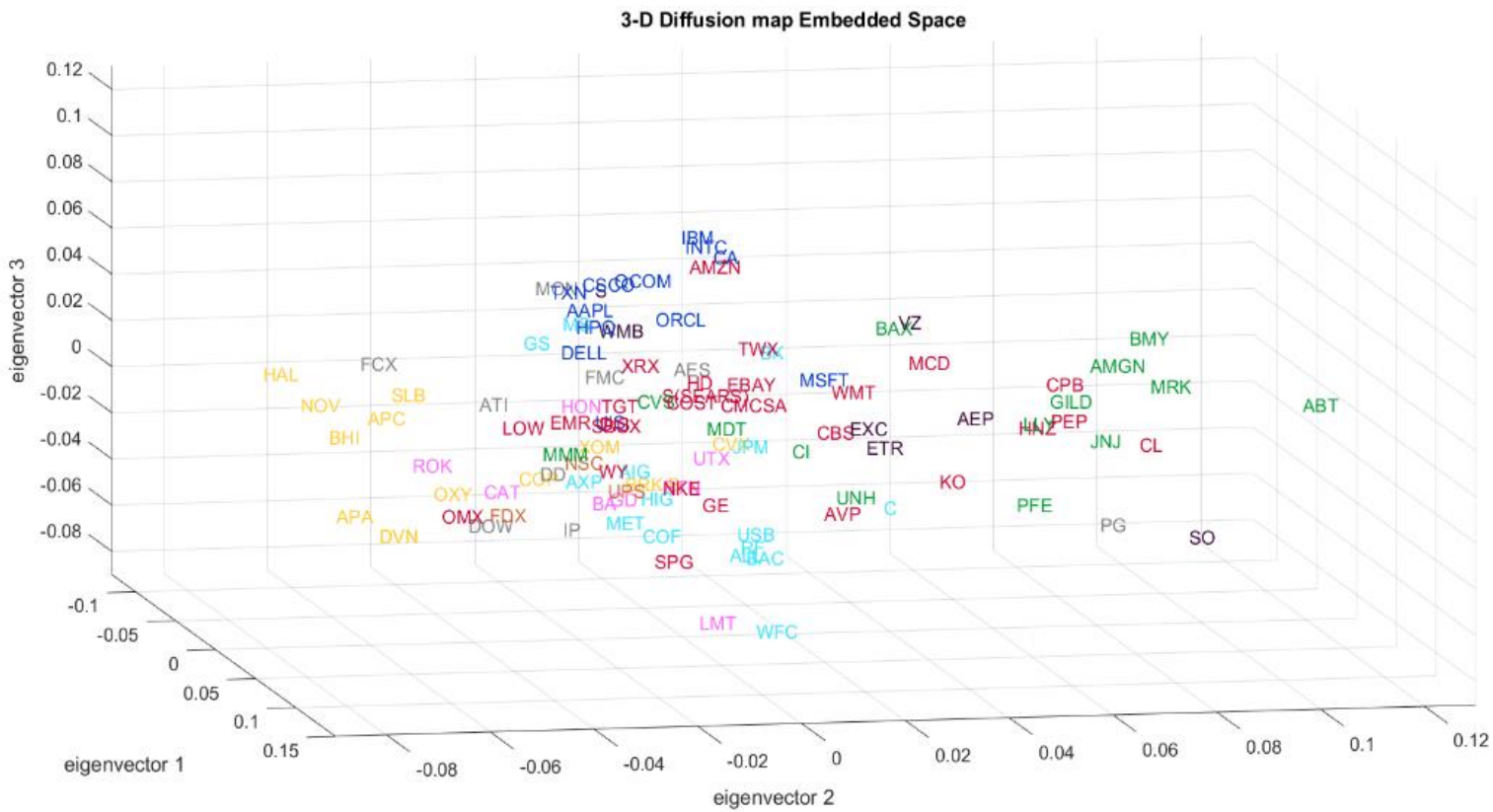
- Από την άλλη μεριά εταιρείες από διαφορετικούς τομείς σχηματίζουν ομάδες. Γνωρίζοντας που ακριβώς δραστηριοποιείται η κάθε εταιρεία μπορούν να εξαχθούν πολύ σημαντικά συμπεράσματα. Για παράδειγμα οι εταιρείες Lowe's (LOW) και Home Depot (HD) οι οποίες δραστηριοποιούνται στο τομέα της επίπλωσης και ανακατασκευής σπιτιών είναι κοντά στις τράπεζες οι οποίες παρέχουν στεγαστικά δάνεια. Αν αναλογιστεί κανείς τη περίοδο που μελετιέται (2002-2012) η συσχέτιση αυτή βγάζει τεράστιο νόημα λόγω της αύξησης των στεγαστικών δανείων και της οικονομικής κρίσης λόγω αυτών. Ένα ακόμα παράδειγμα είναι η διαπίστωση πως η Colgate-Palmolive (CL) εταιρεία που παράγει αγαθά προς κατανάλωση και συγκεκριμένα υλικά περιποίησης βρίσκεται κοντά στις ιατροφαρμακευτικές εταιρείες.

- Στα σχήματα (6.3), (6.4) διαπιστώνεται γιατί η επεξεργασία των διαγραμμάτων τριών διαστάσεων είναι προτιμότερη από αυτή των δύο. Πολλές φορές οι δομές που σχηματίζονται, λόγω οπτικής στα διαγράμματα των δύο, οδηγούν σε λανθασμένα συμπεράσματα. Για παράδειγμα το εξαγόμενο συμπέρασμα πως οι μετοχές των Pfizer (PFE) και Amgen (AMGN) είναι κοντά στον χώρο διάχυσης δύο διαστάσεων είναι λανθασμένο. Παρατηρώντας το χώρο διάχυσης τριών διαστάσεων οι δύο αυτές μετοχές έχουν αρκετά μεγάλη απόσταση μεταξύ τους.

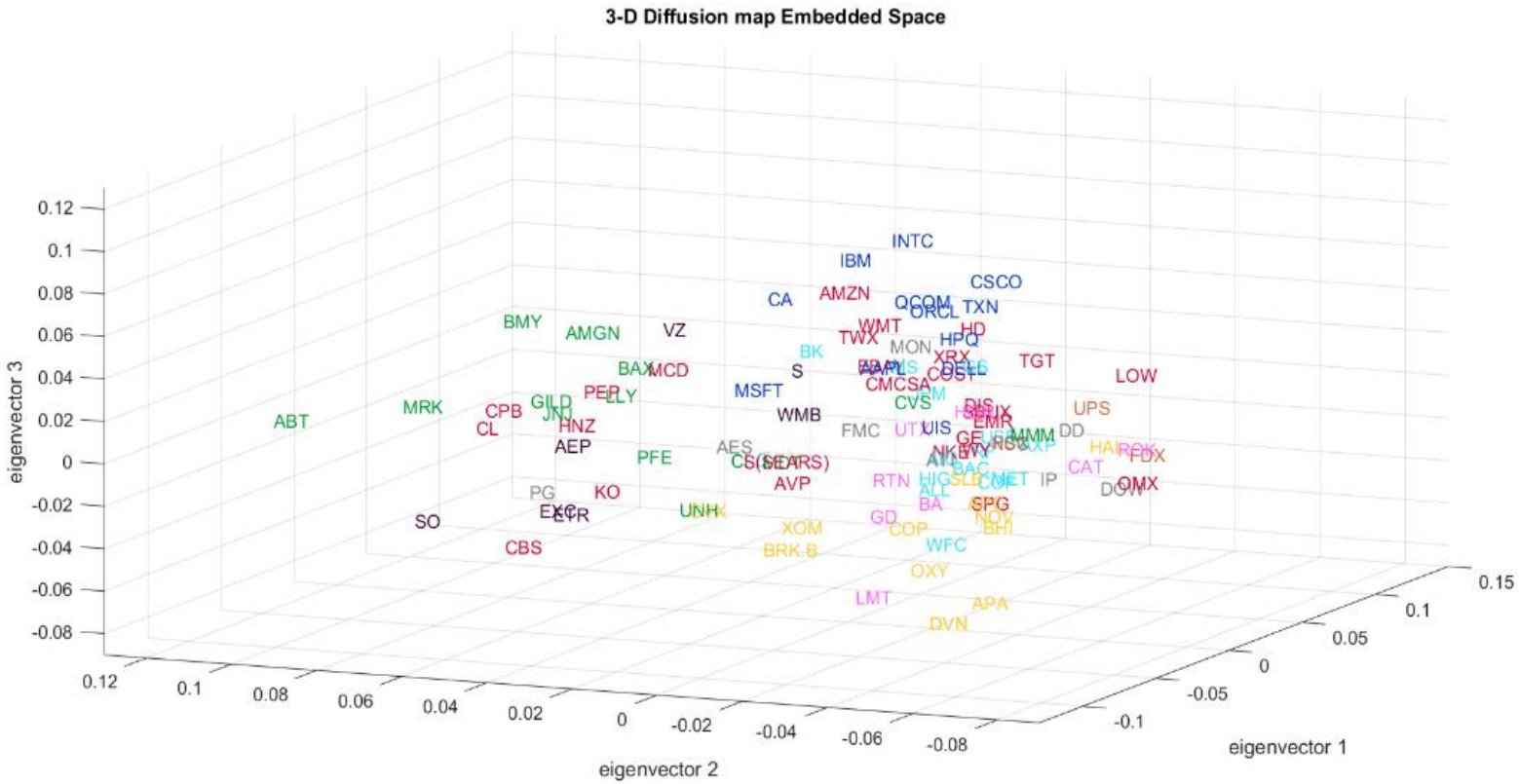
- Εξερευνώντας καλύτερα τα σχήματα πού προκύπτουν από τους χώρους διάχυσης τριών διαστάσεων προκύπτει και ένα ακόμα συμπέρασμα που δεν διαφαίνεται από την απεικόνιση των δύο. Κάποιες μετοχές που ανήκουν στον τομέα των ιατροφαρμακευτικών και κάποιες στον τομέα των υλικών αγαθών σχηματίζουν ομάδα μακριά από τον κεντρικό όγκο του ''σύννεφου'' υποδεικνύοντας ένα μη ομογενές και μη-συμπαγές ''σύννεφο'' όπως υποδηλώνουν οι δύο διαστάσεις. Κατά αυτόν τον τρόπο η παρατήρηση των τριών διαστάσεων δίνει μια ολική εικόνα για τις μετοχές του εκάστοτε χαρτοφυλακίου.



Σχήμα 6.2 : Απεικόνιση 108 μετοχών του δείκτη S&P 100 για τη περίοδο 2002-2012, σε δύο διαστάσεων χώρο διάχυσης, μέσω της μεθόδου diffusion maps χρησιμοποιώντας τον πίνακα συσχετίσεων.



Σχήμα 6.3 : Απεικόνιση 108 μετοχών του δείκτη S&P 100 για τη περίοδο 2002-2012, σε τριών διαστάσεων χώρο διάχυσης, μέσω της μεθόδου diffusion maps χρησιμοποιώντας τον πίνακα συσχετίσεων.



Σχήμα 6.4 : Απεικόνιση 108 μετοχών του δείκτη S&P 100 για τη περίοδο 2002-2012, σε τριών διαστάσεων χώρο διάχυσης, μέσω της μεθόδου diffusion maps χρησιμοποιώντας τον πίνακα συσχετίσεων.

- Capital Goods Companies
- Technology Companies
- Health Care Companies
- Finance Companies
- Consumer Services Companies
- Transportation Companies
- Public Utilities Companies
- Basic Industries Companies
- Energy Companies

Σχήμα 6.5 : Χρωματική κατηγοριοποίηση μετοχών, αναλογικά με το τον τομέα δραστηριοποίησης τους.

6.5 Μέτρηση της ολικής διαφοροποίησης χαρτοφυλακίου μέσω της μεθόδου των Diffusion Maps

6.5.1 Εισαγωγή

Όπως διαπιστώθηκε πρωτύτερα, η χρησιμοποίηση της μεθόδου των απεικονίσεων διάχυσης με βάση τις συσχετίσεις των μετοχών, ανέδειξε σημαντικές δομές μέσω της οπτικοποίησης των χώρων διάχυσης. Τα συμπεράσματα που προέκυψαν μέσω αυτών, μπορεί να ερμηνευτούν ως την προσπάθεια της μεθόδου να ομαδοποιήσει τις μετοχές, διατηρώντας την τοπικότητα των δεδομένων αγνοώντας πλήρως την ολική συνδεσιμότητα του γραφήματος.

Αυτό όπως θα φανεί είναι μια αντιμετώπιση εντελώς λανθασμένη καθώς ο χώρος εμφύτευσης που εξερευνάται αποκτά ολικό νόημα. Το γράφημα εκτός από την ιδιότητα της ομαδοποίησης (clustering) μεταξύ των μετοχών έχει και την ιδιότητα σύγκρισης μεταξύ απομακρυσμένων σημείων. Το γεγονός πως η BAC και η DELL έχουν μικρότερη απόσταση από ότι η BAC και η OXY είναι ένα ζήτημα που έχει οντότητα, παρόλο που οι αποστάσεις είναι μεγάλες.

Αυτή η τοποθέτηση, οδηγεί στο συμπέρασμα πως τα γραφήματα που μπορούν να διαμορφωθούν για διαφορετικές χρονικές περιόδους δεν είναι μόνο χρήσιμα για οπτικοποίηση δεδομένων. Όπως θα δειχθεί, το γεγονός πως οι αποστάσεις μεταξύ όλων των μετοχών έχουν γενική σημασία, επιτρέπει μέσω των χώρων διάχυσης να εξαχθούν μεγάλης σημασίας ποσοτικά μέτρα όπως αυτό της διαφοροποίησης (diversification).

6.5.2 Διαφοροποίηση χαρτοφυλακίου (portfolio diversification)

Ένα από τα πιο σημαντικά ζητήματα που πρέπει να διασφαλίζει ένας επενδυτής είναι η προστασία του χαρτοφυλακίου του από τους εκάστοτε κινδύνους της αγορά (συστημικούς, πολιτικούς, πιστωτικούς, κ.α.). Ο πιο διαδεδομένος τρόπος προστασίας είναι αυτός της διαφοροποίησης (diversification). Εν συντομία είναι ο στόχος επιλογής διαφορετικών τύπων επενδύσεων (μετοχές, ομόλογα, αμοιβαία κεφάλαια) αποτελούμενα από διαφορετικά χαρακτηριστικά και διαφορετική προέλευση.

Η διαδικασία βέλτιστης διαφοροποίησης γνωστή και ως Markowitz diversification (λόγω του γνωστού οικονομολόγου- νομπελίστα που υιοθέτησε τον όρο της), αναφέρει πως για τη δημιουργία διαφοροποιημένων χαρτοφυλακίων στόχος είναι, η εύρεση μετοχών των οποίων οι συσχετίσεις μεταξύ τους είναι μη ικανοποιητικές ή ακόμα και αρνητικές. Ένας τρόπος λοιπόν για να διατηρηθεί μία επένδυση ασφαλής στην πάροδο του χρόνου είναι η επιλογή μετοχών από διαφορετικούς κλάδους χαρακτηριστικών. Κατά αυτόν τον τρόπο η επιρροή μίας υποτιθέμενης κρίσης, δε θα έχει μεγάλο αντίκτυπο σε όλο το χαρτοφυλάκιο **(Markowitz, 1952)**.

Το ερώτημα που μπορεί να προκύψει, είναι πόσες και ποιες μετοχές είναι αυτές που θα δημιουργήσουν ένα διαφοροποιημένο χαρτοφυλάκιο; Η απάντησή στην επιλογή του πλήθους, της επιλογής των μετοχών έχει εξεταστεί εκτενώς στο παρελθόν. Οι **Evans και Archer (1968)** ανέφεραν πως η επιλογή 10 τυχαίων μετοχών καθιστούν ένα χαρτοφυλάκιο πλήρως διαφοροποιημένο. Εν αντίθεση ο **Statman (1987)** σύγκρινε το κόστος και τα πλεονεκτήματα της διαφοροποίησης και κατέληξε πως ένα χαρτοφυλάκιο πρέπει να διέπετε από 30 στο πλήθος μετοχές. Η απάντηση στο ποιες μετοχές μπορούν να σχηματίσουν ένα διαφοροποιημένο χαρτοφυλάκιο θα προέλθει στο επόμενο κεφάλαιο. Με βάση λοιπόν την επιλογή ενός χαρτοφυλακίου θα ήταν πολύ χρήσιμο η ανεύρεση ενός μέτρου που θα υποδεικνύει το πόσο διαφοροποιημένο είναι.

Δοθέντος λοιπόν του ορισμού της διαφοροποίησης και η σωστή κατανόηση της οδηγεί στο εξής ερώτημα. Μπορεί η μέθοδος των απεικονίσεων διάχυσης να εξάγει χρήσιμα ολικά μέτρα για την εκτίμηση της σε ένα χαρτοφυλάκιο; Δημιουργώντας το ερώτημα αυτό η απάντηση θεωρείται πως θα έρθει, από το κατά πόσο τα γραφήματα που μπορούν να προκύψουν, χρησιμοποιώντας τη μέθοδο αυτή, για διαφορετικές περιόδους, καταφέρνουν να εξάγουν μία τέτοιου είδους πληροφορία.

Στο σχήμα 6.6 εμφανίζονται τρία στο πλήθος γραφήματα (σε τρισδιάστατους χώρους διάχυσης) για τρείς διαφορετικές περιόδους (χρησιμοποιώντας πλέον τα δεδομένα των αποδόσεων των μετοχών του δείκτη S&P 500) :

1. Τη χρονική περίοδο Μάιο 2005 - Ιούνιο 2007, η οποία αντιστοιχεί στην περίοδο της πιστωτικής επέκτασης (Credit Boom period).

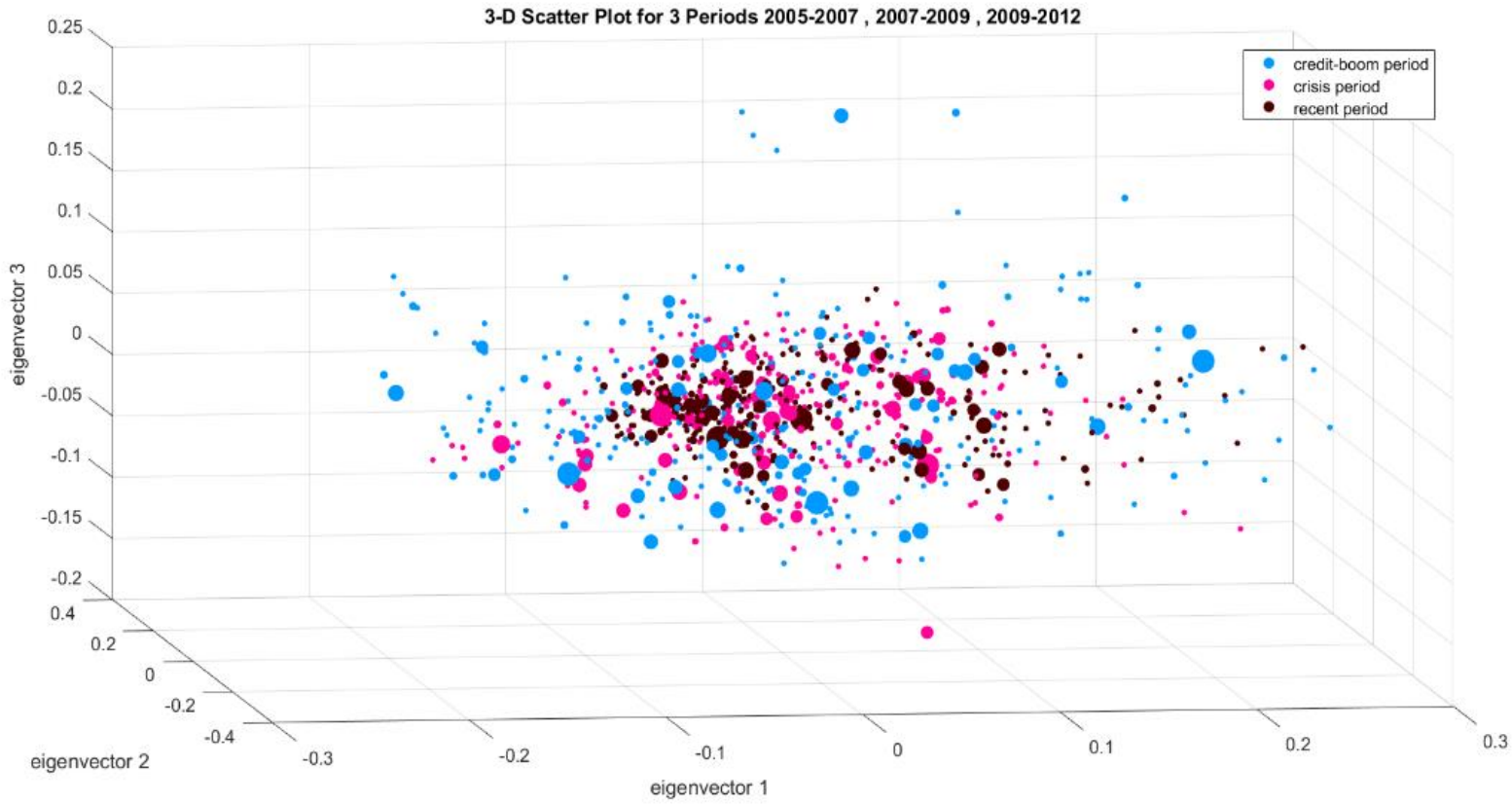
2. Τη χρονική περίοδο Ιούλιο 2007 - Σεπτέμβριο 2009, η οποία αντιστοιχεί στην περίοδο της χρηματοπιστωτικής κρίσης(Crisis Period).

3. Τη χρονική περίοδο Οκτώβριο 2009 - Απρίλιο 2012, η οποία αντιστοιχεί στην μετέπειτα περίοδο κρίσης, περιλαμβάνοντας και τις επιπτώσεις της ευρωπαϊκής κρίσης στο τέλος της (Recent Period).

Για καλύτερη ερμηνεία των δεδομένων τα ονόματα των μετοχών, δεν περιλήφθηκαν υπόψη και οι όγκοι των συμβόλων αντικατοπτρίζουν το μέσο βάρος της εκάστοτε μετοχής του δείκτη για την εκάστοτε περίοδο.

Το ‘’σύννεφο’’ όπως διαφαίνεται στο σχήμα 6.6 αποκτά διαφορετικές γεωμετρικές δομές κατά τη διάρκεια των διαφορετικών περιόδων που εξετάζονται, κάτι που ενισχύει τη θεώρηση που έγινε πρωτύτερα για της ολικής σημασίας ρόλο που παίζουν οι αποστάσεις των μετοχών στους χώρους διάχυσης. Κατά τη διάρκεια της πιστωτικής επέκτασης των τράπεζων (την περίοδο 05’-07’) το ‘’σύννεφο’’ παρουσιάζεται αρκετά διασκορπισμένο. Εν συνεχεία, άρχισε να συρρικνώνεται προς το κέντρο, κατά τη διάρκεια της κρίσης (περίοδος 07’-09’) και εν τέλει έγινε πιο συμπαγές την περίοδο μετρά κρίσης-recent period (09’-12’), λαμβάνοντας υπόψιν τα βάρη των συστατικών του δείκτη.

Με βάση το διάγραμμα λοιπόν αντιλαμβάνεσαι κανείς πως οι απεικονίσεις διάχυσης ανιχνεύουν πλήρως την συμμεταβολή των μετοχών κατά το πέρας του χρόνου. Συμμεταβολές μετοχών στη πλειοψηφία του δείκτη υποδηλώνουν συμπαγή ‘’σύννεφα’’ ενώ αντίθετα ανεξάρτητες κινήσεις μετοχών στο χρόνο υποδηλώνουν αραιά, μη-συμπαγή ‘’σύννεφα’’. Κατά αυτόν τον τρόπο λοιπόν μέτρα για την ολική διαφοροποίηση των μετοχών του δείκτη μπορούν να αντληθούν μέσο των χώρων διάχυσης.



Σχήμα 6.6 : Απεικόνιση χώρων διάχυσης τριών διαστάσεων των μετοχών του δείκτη S&P 500 για τις χρονολογίες 2005-2007, 2007-2009, 2009-2012.

6.5.3 Αλγόριθμος εύρεσης μέτρου διαφοροποίησης

Όπως ειπώθηκε και πριν, το μέγεθος του ''σύννεφου'' ή η έκταση που καταλαμβάνει στον χώρο, υποδεικνύει την διαφοροποίηση που αντιλαμβάνεται κάποιος στο σύνολο των μετοχών. Το κατά πόσο δηλαδή τείνουν να κινηθούν μαζί ή το κατά πόσο διαφοροποιημένες είναι οι συμμεταβολές τους. Ένα απλό, άμεσο μέτρο για την διαφοροποίηση που ορίζεται κάθε χρονική περίοδο δίνεται μέσω της μεθόδου. Τα βήματα υπολογισμού του δίνονται ακολούθως :

1. Οι απεικονίσεις διάχυσης (diffusion maps) θεωρούν το σύνολο των μετοχών ως ένα ''σύννεφο'' σημείων στον χώρο διάχυσης.
2. Το ''σύννεφο'' γενικά θα έχει μη-κανονικό σχήμα. Αγνοώντας αυτό υπέθεσε πως είναι ένα δείγμα προερχόμενο από μια πολυμεταβλητή κανονική κατανομή.
3. Κατά αυτόν τον τρόπο οι παράμετροι αυτής της κατανομής υπολογίζονται άμεσα. Υπολόγισε τον πίνακα σταθμισμένων διασπορών (weighted covariance matrix) : Σ_{w} όπου w το βάρος της εκάστοτε μετοχής. Ο Σ_{w} περιγράφει την έκταση του ''σύννεφου'', μεγαλύτερες διασπορές ισοδυναμούν με μεγαλύτερα εκτάσεως ''σύννεφα''.
4. Το μέτρο ολικής συγκέντρωσης (Global Concentration Measure), εν συντομία G.C.M, είναι :

$$G.C.M = (\mathrm{tr}\Sigma_{\mathrm{w}})^{-\frac{1}{2}} \tag{6.9}$$

Όσο μεγαλύτερο το μέτρο της ολικής συγκέντρωσης, τόσο μικρότερης έκτασης ''σύννεφο'' αντιστοιχείται και άρα μικρότερη η παρουσία διαφοροποίησης στο χαρτοφυλάκιο. Όσο μικρότερο το μέτρο ολικής συγκέντρωσης τόσο σε μεγαλύτερο ''σύννεφο'' ανταποκρίνεται και άρα μεγαλύτερη η παρουσία της διαφοροποίησης στις μετοχές του χαρτοφυλακίου.

5.5.4 Εφαρμογή αλγορίθμου και αποτελέσματα

Για να εξεταστεί η λειτουργία του αλγορίθμου και την εύρεση του ολικού μέτρου της διαφοροποίησης μεταξύ των μετοχών που εξετάζονται, για διάφορες χρονικές περιόδους εφαρμόστηκε ένα rolling-window (κινούμενο παράθυρο) 12-μηνών και βήματος ένα. Με τη διαδικασία του rolling window θα κατανοηθεί πλήρως οι διακυμάνσεις που μπορούν να παρατηρηθούν στις τιμές του μέτρου διαφοροποίησης G.C.M. για διαφορετικές περιόδους.

Στην ουσία το rolling window που εφαρμόστηκε στη συγκεκριμένη εργασία είναι η επαναλαμβανόμενη διαδικασία εξαγωγής μέτρων G.C.M. Η διαδικασία που ακολουθείται, είναι η συνεχομένη εισχώρηση σαν είσοδο στον αλγόριθμο των diffusion maps, συνεχόμενων πινάκων συσχετίσεων που προέρχονται από μήκους 12-μηνών πινάκων αποδόσεων, προσθέτοντας κάθε φόρα βήμα $\mathrm{t} = 1$ (στην συγκεκριμένη περίπτωση $\mathrm{t} =$ μήνας) μέχρι το πέρας της χρονολογίας που μελετιούνται τα δεδομένα. Μετέπειτα, μέσω των χώρων διάχυσης που προκύπτουν (στη συγκεκριμένη περίπτωση διαστάσεων $\mathrm{p} = 5$) υπολογίζεται το μέτρο G.C.M για κάθε χώρο ξεχωριστά. Στην πράξη, επί παραδείγματι, ξεκινώντας από τον Ιανουάριο του 2002 μέχρι τον Ιανουάριο του 2003,υπολογίζεται ο πίνακας συσχετίσεων των μετοχών, έπειτα προκύπτει άμεσα ο χώρος διάχυσης πέντε διαστάσεων, μέσω της μεθόδου των diffusion maps και εν τέλει υπολογίζεται το G.C.M. Εν συνεχεία ακολουθείτε η ίδια διαδικασία, για την περίοδο Φεβρουάριο του 2002 έως Φεβρουάριο του 2003 και συνεχίζει μέχρι το τέλος. Εν τέλει το αποτέλεσμα που θα προκύψει, είναι 112 ολικά ποσοτικά μέτρα διαφοροποίησης τα οποία υποδηλώνουν μια χρονολογική σειρά. Το σχήμα 6.7 παρέχει μια σχηματική απεικόνιση του πώς λειτουργεί η διαδικασία του rolling window.

Σημαντική παρατήρηση που πρέπει να δοθεί σε αυτό το σημείο, είναι πως για να επιτευχθούν μετρήσεις με νόημα, θα πρέπει κατά τη διάρκεια των περιόδων που πραγματοποιείται η διαδικασία του rolling window να παραμένουν σταθεροί οι ακόλουθοι παράμετροι :

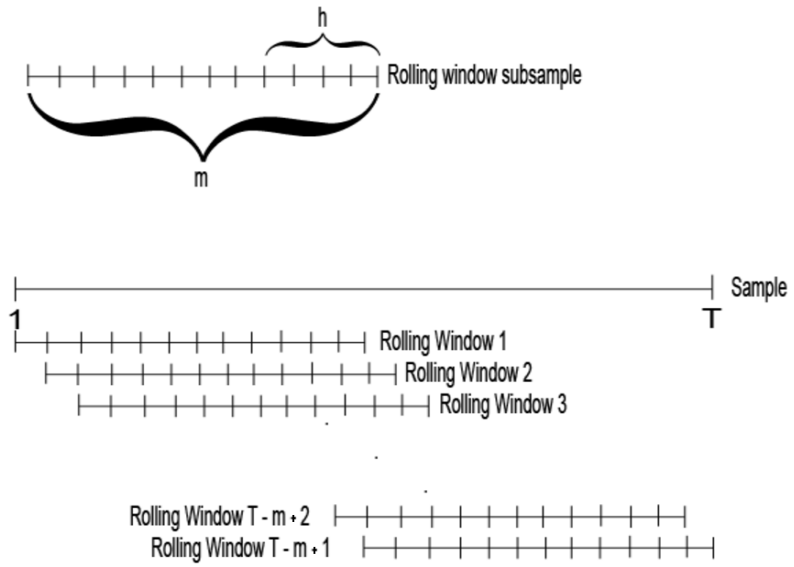
- Ο δειγματικός πληθυσμός των μετοχών (assets) που λαμβάνονται υπόψη στη διαδικασία.
- Το μήκος των αποδόσεων των χρονοσειρών, που χρησιμοποιούνται για τη διαδικασία του rolling window.
- Ο αριθμός των διαστάσεων των χώρων διάχυσης.

Τα αποτελέσματα που πάρθηκαν από αυτήν την διαδικασία που περιεγράφηκε, φαίνονται στο σχήμα (6.8) όπου αναπαρίσταται μία χρονολογική σειρά του ολικού μέτρου συγκέντρωσης, κατά τη διάρκεια της δεκαετίας. Το διάγραμμα δείχνει τα ακόλουθα :

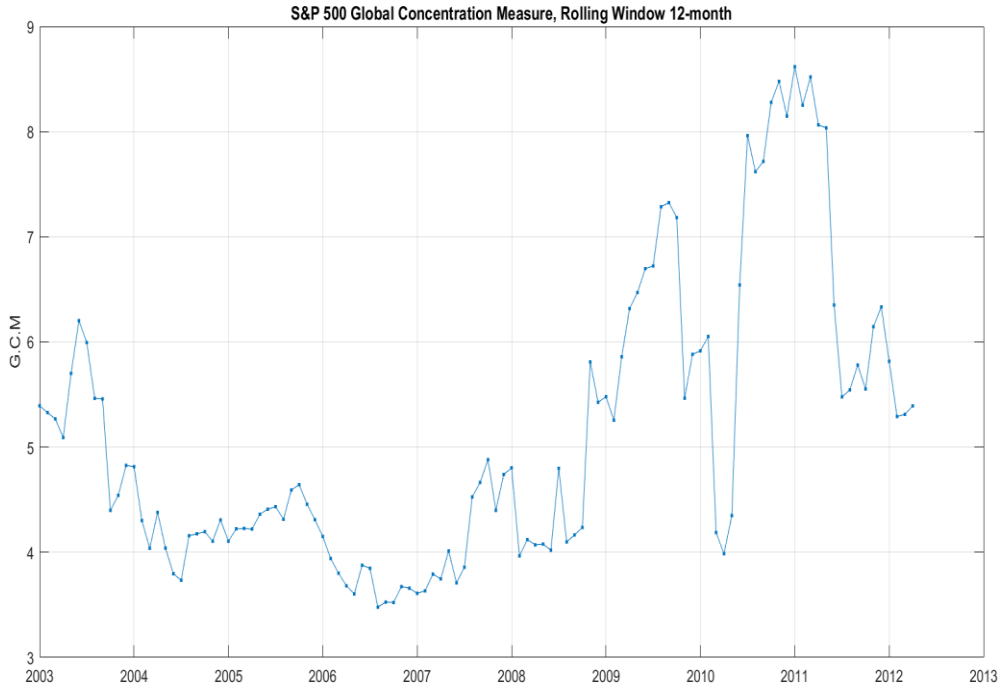
- Ο δείκτης στα μέσα της δεκαετίας περίπου σημειώνει τη χαμηλότερη τιμή ολικής συγκέντρωσης. Αυτό συνέβη καθώς εκείνη τη περίοδο οι μετοχές προσπάθησαν να κινηθούν ανεξάρτητα η μία από την άλλη, σε σχέση του σχετικού, οικονομικού, σταθερού πλαισίου που περιλάμβανε εκείνη τη περίοδο.

- Όταν όμως η οικονομική κρίση άρχισε να επιδρά στην αγορά στα μέσα του 2007, το μέτρο της ολικής συγκέντρωσης άρχισε να αυξάνεται ραγδαία. Η περισσότερη συν μετακίνηση διαπιστώνεται στο έτος του 2009. Αμέσως μετά ο δείκτης δείχνει να έχει μέτρα σε φθίνουσα πορεία και η μικρότερη τιμή του λαμβάνεται στις αρχές του 2010. Το μέτρο όμως, της ολική συγκέντρωσης ανέπτυξε και πάλι αστραπιαίες αύξουσες τάσεις, λόγω αυξήσεως των market caps μέσω της μετοχικής ανάπτυξής (equity rally) που πραγματοποιήθηκε. Αμέσως μετά διαπιστώνεται άλλη μια κάθοδος του μέτρου, την οποία ακολουθεί μία λιγότερο εκτενής άνοδος, λόγω του αντίκτυπου της κρίσης της ευρωπαϊκής ένωσης το 2011. Όσο για τον Απρίλη όπου σταματάει το rolling window, η τιμή του μέτρου είναι αρκετά υψηλή σε σχέση με τις τιμές που έλαβε ο δείκτης την περίοδο της πιστωτική επέκτασης.

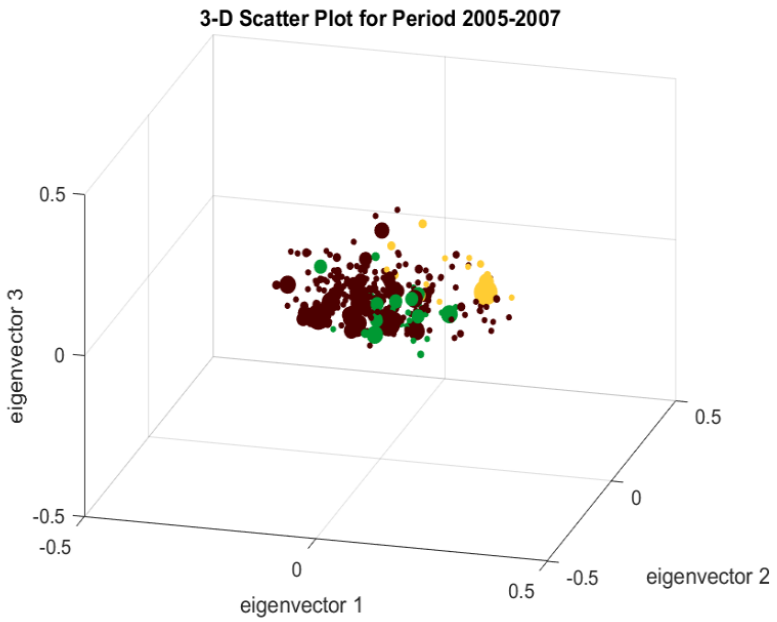
Δοθέντος λοιπόν του των ολικών μέτρων συγκέντρωσης (G.C.M), η εκτίμηση της διαφοροποίησης μεταξύ των μετοχών εκτιμάτε σε πιο γενικά πλαίσια καθώς αποτελεί μια γενική σύνοψη της γεωμετρίας και της συμμεταβολής των δεδομένων στη διάρκεια του χρόνου. Η πληροφορία για το ποιοι είναι οι αληθινοί παράγοντες, που ευθύνονται στην αλλαγή τη γεωμετρίας των ‘’σύννεφων’’ και κατ’ επέκταση της διαφοροποίησης των μετοχών, αντλείται από τους ίδιους τους χώρους διάχυσης. Για παράδειγμα το γεγονός πώς κατά τη περίοδο της πιστωτικής επέκτασης η ολική διαφοροποίηση ήταν υψηλή, έγκειται στο γεγονός πως οι μετοχές που δραστηριοποιούνται στο χώρο της ενέργειας και στον χώρο των ιατροφαρμακευτικών, σχημάτισαν ομάδες μακριά από το κέντρο του ‘’σύννεφου’’. Ενώ μέρος ευθύνης της αύξησης του μέτρου G.C.M φέρει το γεγονός πως οι ίδιες πορεύονταν προς το κέντρο του ‘’σύννεφου’’ (σχήμα 6.9).

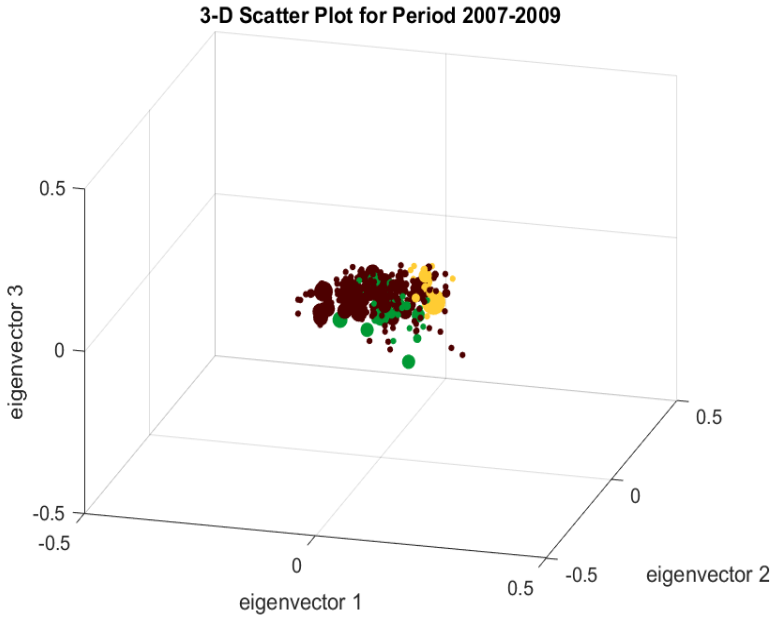


Σχήμα 6.7 : Σχηματική αναπαράσταση της μεθοδολογίας του rolling window (στην εφαρμογή της εργασίας: m=12, T=123, t=1) **(https://uk.mathworks.com).**



Σχήμα 6.8 : Απεικόνιση της χρονοσειράς του μέτρου ολικής συγκέντρωσης (G.C.M.) χρησιμοποιώντας rolling window 12-μηνών βήματος 1 για τα δεδομένα των συστατικών του δείκτη S&P 500.





Σχήμα 6.9 : Απεικόνιση χώρων διάχυσης τριών διαστάσεων, των ιατροφαρμακευτικών και ενεργειακών μετοχών, χρώματος πράσινου και κίτρινου αντίστοιχα, για τις χρονολογίες 2005-2007, 2007-2009.

6.6 Μέτρηση τοπικών συγκεντρώσεων σε χαρτοφυλάκιο μέσω των diffusion maps

6.6.1 Εισαγωγή

Όπως προαναφέρθηκε στο προηγούμενο κεφάλαιο η εύρεση του μέτρου ολικής συγκέντρωσης ενός χαρτοφυλακίου υπέδειξε την γενική εικόνα διαφοροποίησης που το διέπει στο σύνολο του. Εκτός όμως από την ολική διαφοροποίηση ενδιαφέρον παρουσιάζει η εύρεση τοπικών συγκεντρώσεων κατά τη διάρκεια διαφόρων χρονικών περιόδων. Υποθετικά οι μετοχές μπορούν να δεχτούν τοπικές ή ιδιοσυγκρασιακές κρίσεις (shock's) οι οποίες επηρεάζουν ορισμένες μόνο περιοχές του αυθαίρετου αρχικού χώρου. Το ερώτημα που προκύπτει είναι ποιες περιοχές είναι ποιο σημαντικές. Στην ουσία, το ερώτημα που προκύπτει είναι σε ποιες περιοχές οι μετοχές είναι πιο συγκεντρωμένες.

Η εκτίμηση τέτοιων κρίσεων στο οποιοδήποτε χαρτοφυλάκιο μπορεί να αποφέρει τεράστια πλεονεκτήματα και στην ανάλυση ρίσκου αλλά και στην κατασκευή ενός χαρτοφυλακίου. Οι μετοχές που βρίσκονται κοντά σε περιοχές μεγάλων τοπικών συγκεντρώσεων θα πρέπει να μελετηθούν περαιτέρω καθώς η επιλογή τους στη διαμόρφωση ενός χαρτοφυλακίου μπορεί να αποφέρει αρνητικά αποτελέσματα για την ολική διαφοροποίηση του.

Ακόμα η εξαγωγή διαγραμμάτων που θα περιέχουν ποσοτική πληροφορία και θα δίνουν τα προφίλ των τοπικών συγκεντρώσεων (κρίσεων) για διαφορετικές περιόδους, μπορεί να δώσουν τις κατάλληλες απαντήσεις για την διαφοροποίηση του χαρτοφυλακίου και την στόχευση σε συγκεκριμένες περιόδους εύρεσης τοπικών συγκεντρώσεων.

6.6.2 Μεθοδολογία εύρεσης τοπικών συγκεντρώσεων

Για την υλοποίηση της συγκεκριμένης ιδέας, ακολουθούνται τα εξής βήματα :

1. Θεώρησε μία συνάρτηση που θα περιγράφει ένα τοπικό σοκ. Θα ήταν σκόπιμη η λήψη μίας συμμετρικής συνάρτησης πυκνότητας πιθανότητας, η οποία θα λαμβάνει ως μέγιστη τιμή τη μονάδα. Αυτή περιγράφει μια τοπική κρίση (shock) που θα έχει μέγιστο σε ένα σημείο του χώρου και θα πέφτει ραγδαία μετά το μέγιστο.

2. Χρησιμοποιώντας τη παράμετρο **ε** της κανονικής κατανομής, ορίζεται το εύρος των μετοχών που λαμβάνονται υπόψη από το σοκ. Με μεγάλες τιμές του **ε,** το εύρος αυτό μεγαλώνει και το σοκ επηρεάζει περισσότερες μετοχές.

3. Λαμβάνοντας υπόψη τις ζυγισμένες στο χώρο μετοχές (με βάση τα Market Caps), για συγκεκριμένο σημείο στο χώρο που λαμβάνει μέρος ένα σοκ, το μέτρο που δίνει το μέγεθος του σοκ είναι ένα πεπερασμένο άθροισμα: Οι μετοχές κοντά στο σοκ, έχοντας υπόψη και τα αντίστοιχα Market Caps τους, συμμετέχουν σε μεγαλύτερο βαθμό στο άθροισμα. Το σημείο που μεγιστοποιείται το άθροισμα αυτό είναι και η μέγιστη - **ε** τοπική συγκέντρωση στον χώρο (L.C.M).

6.6.3 Αλγόριθμος εύρεσης προφίλ τοπικής συγκέντρωσης

Στόχος αυτής της ενότητας, είναι να δοθεί ο αλγόριθμος υπολογισμών του προφίλ τοπικής συγκέντρωσης. Η εύρεση δηλαδή των k μεγαλύτερων τοπικών συγκεντρώσεων κατά φθίνουσα σειρά.

1. Προσδιόρισε την παράμετρο **ε.**
2. Βρες τη μεγαλύτερη συγκέντρωση (όπως προαναφέρθηκε).
3. Για κάθε μετοχή υπολόγισε: $market\ cap$ = $market\ cap$ (1−$value\ of\ local\ shock\ function$).
4. Επέστρεψε στο βήμα 2 και υπολόγισε ξανά για όσες k τοπικές συγκεντρώσεις επιθυμείτε να βρεθούν.

6.6.4 Εφαρμογή αλγορίθμου και αποτελέσματα

Σε πρώτη φάση χρησιμοποιώντας, πάλι τις μηνιαίες αποδόσεις των μετοχών για την περίοδο μετά-κρίσης, που επιλέχθηκαν σαν αντιπροσώπευση του δείκτη S&P 500 εφαρμόστηκε ο αλγόριθμος εύρεσης τοπικών συγκεντρώσεων για την γραφική αναπαράσταση των θέσεων τους στο χώρο διάχυσης και παράχθηκε το διάγραμμα που φαίνεται στο σχήμα 6.10. Η παραγωγή του οποίου έγινε χρησιμοποιώντας πάλι τις πρώτες πέντε διαστάσεις καθώς θεωρείτε πως καλύπτουν μεγαλύτερο ποσοστό της σχετικής πληροφορίας των δεδομένων του αρχικού χώρου.

Για τον εντοπισμό των πέντε μεγαλύτερων τοπικών συγκεντρώσεων (χρώματος γκρι στο διάγραμμά) χρησιμοποιήθηκε η παράμετρος **ε** = 0.05 καθώς θεωρήθηκε πως μια τέτοια τιμή ενσωματώνει αρκετές μετοχές για τον προσδιορισμό της εκάστοτε τοπικής κρίσης (shock). Τα μεγέθη των συμβόλων του διαγράμματος αναφέρονται στα σχετικά Market Caps της εκάστοτε μετοχής.

Σημαντική παρατήρηση που πρέπει να δοθεί σε αυτό το σημείο είναι πως ο ορισμός του **ε** πρέπει να γίνει με σύνεση καθώς μικρό **ε** επιφέρει απλά την φθίνουσα επιστροφή των ζυγισμένων μετοχών καθώς η συνάρτηση υπολογισμού τοπικών κρίσεων θα επιλέγει μόνο μια μετοχή σε κάθε βήμα υπολογισμού.

Όπως διαφαίνεται στο σχήμα 6.10 η μεγαλύτερη τοπική συγκέντρωση βρίσκεται κοντά στο κέντρο του ''σύννεφου'' όπου η συνάρτηση εύρεσης τοπικών κρίσεων έχει εντοπίσει κάποιες μετοχές με σχετικά μεγάλα Market Caps. Κάποιες από αυτές όπως διαφαίνεται είναι οι Orcl company (ORCL) και η Sisco company (CSCO). Η δεύτερη μεγαλύτερη τοπική συγκέντρωση φαίνεται να είναι πολύ κοντά στην πρώτη και να περιλαμβάνει μετοχές όπως οι Exon mob. Company (XOM) και η Campbell Soup company (CPB). Η τρίτη φαίνεται πως βρίσκεται πολύ κοντά στην Emc corp. (EMC).

Σε δεύτερη φάση παράχθηκε το ολικό προφίλ των τοπικών συγκεντρώσεων για τις περιόδους που παρουσιάστηκαν στο σχήμα 6.6 της ενότητας 6.5 (credit boom period, crisis period, recent period) και εμφανίζονται στο σχήμα 6.11. Για ακόμα μια φορά η παράμετρος **ε** πήρε τιμή 0.05. Το διάγραμμα δείχνει τα ακόλουθα:

- Στην περίοδο προ-κρίσης και στην τωρινή (μετά κρίσης), η μεγαλύτερη τοπική συγκέντρωση είναι σημαντικά μεγαλύτερη σε σχέση με τις άλλες 15.

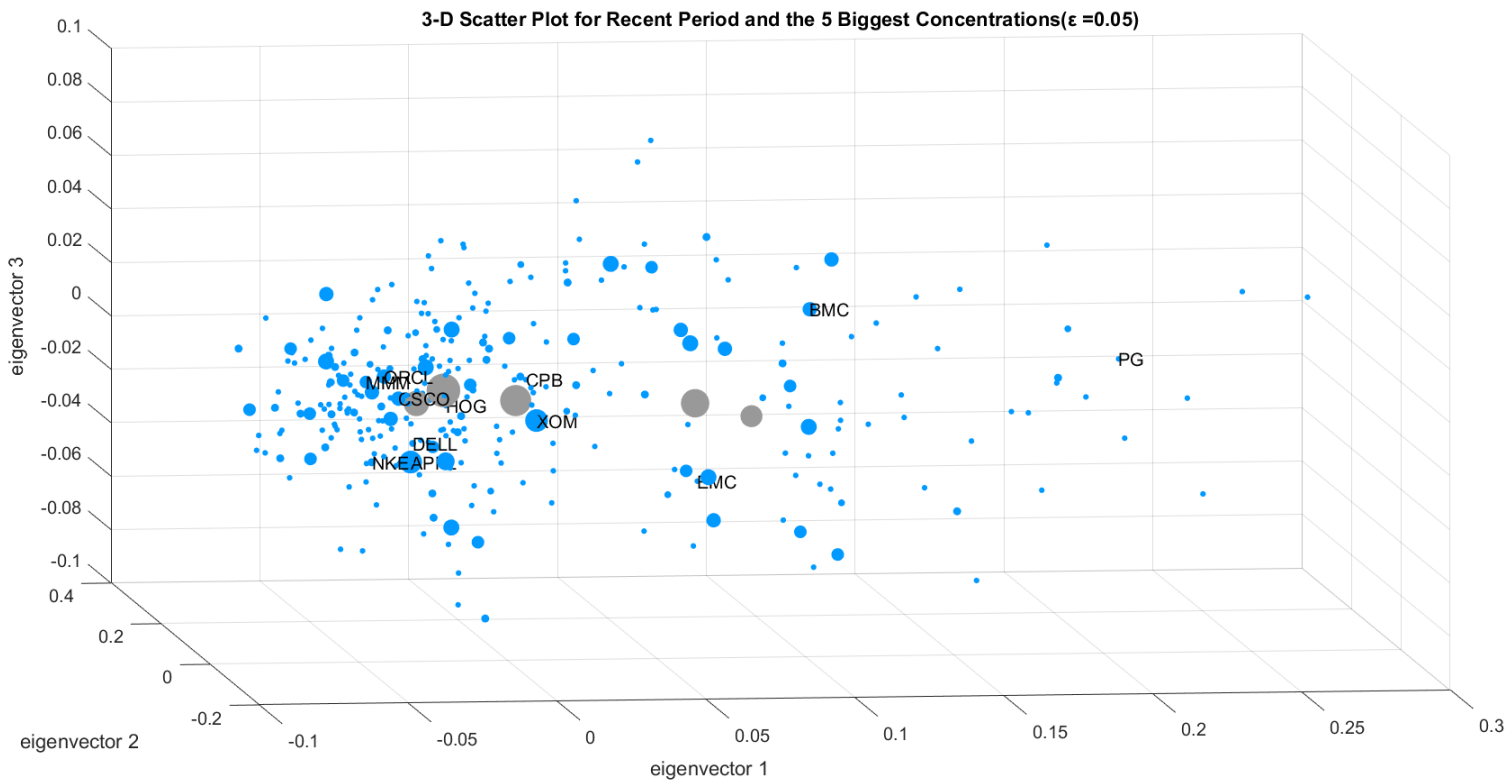
- Στην περίοδο κρίσης, η μεγαλύτερη τοπική συγκέντρωση είναι σημαντικά πιο μικρή από τις άλλες 2 περιόδους καθώς και πιο μικρή στη συντριπτική πλειοψηφία των μετρήσεων. Μία εξήγηση που μπορεί να δοθεί είναι ότι οι τοπικές συγκεντρώσεις στη κρίση δεν παίζουν τόσο μεγάλο ρόλο, γιατί σε αντίθεση επηρεάζονται αρκετά από την ολική τάση των μετοχών να κινηθούν μαζί.

- Παρατηρείται μία μεγαλύτερη πτώση στα μέτρα τις τοπικής συγκέντρωσης κατά την τωρινή περίοδο(περίοδος μετά-κρίσης), σε σχέση με τη περίοδο της πιστωτικής επέκτασης.

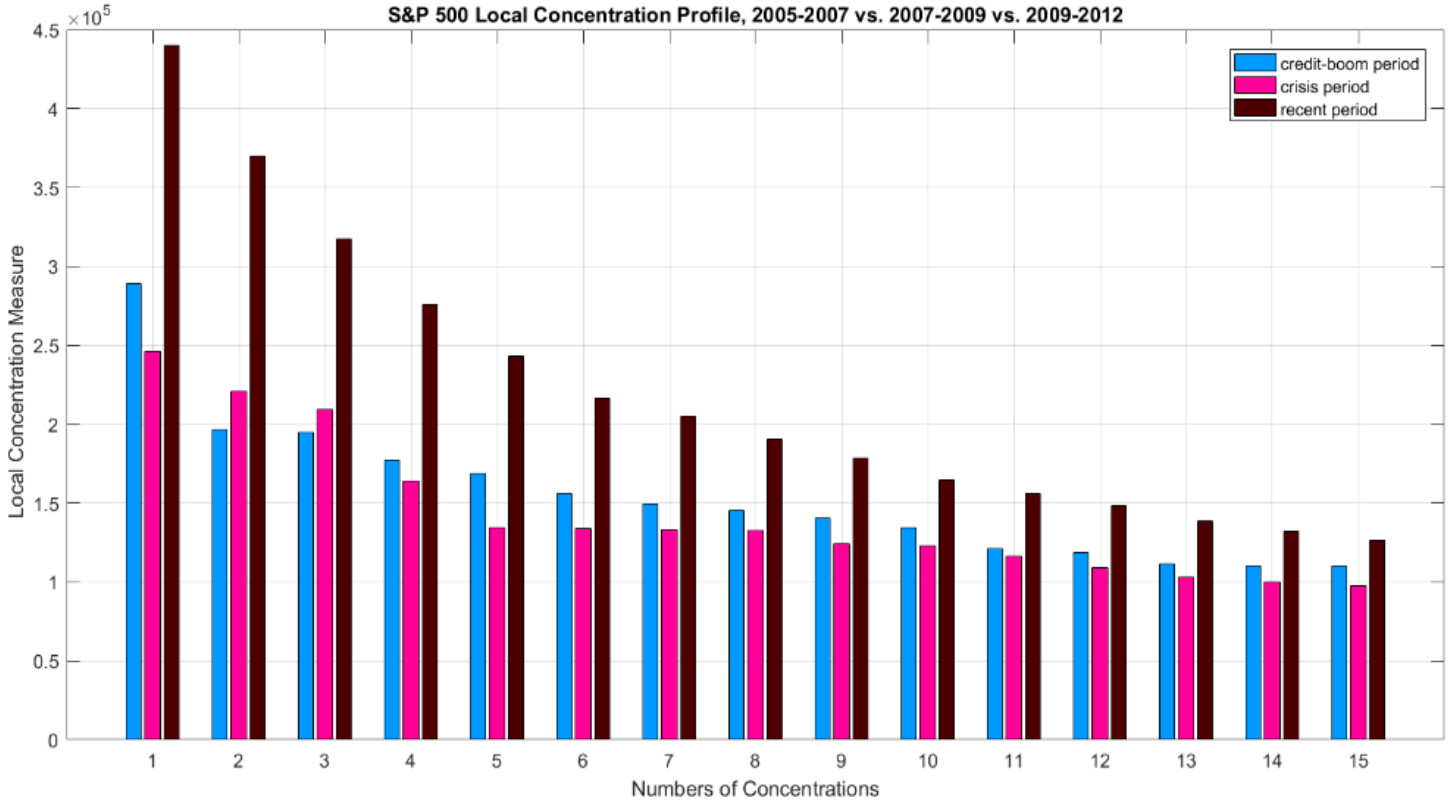
- Οι τελευταίες μετρήσεις τοπικών συγκεντρώσεων όλων των περιόδων που εξετάζονται φαίνονται να είναι ίδιες. Αυτό συμβαίνει καθώς στην πλειοψηφία τους βρίσκονται στις μετοχές που βρίσκονται απομακρυσμένες από το κέντρο του ''σύννεφου'' και άρα έχουν μικρές αποκλίσεις.

Παρόλο που το μέτρο της ολικής συγκέντρωσης παρουσιάζεται να είναι το ίδιο την περίοδο της κρίσης και της πρόσφατης οι λόγοι που συμβαίνουν αυτό το γεγονός φαίνονται μέσω του προφίλ των τοπικών συγκεντρώσεων. Κατά τη διάρκεια της κρίσης το ''σύννεφο'' ήταν πιο ομοιόμορφα συσσωρευμένο (όλες οι συσχετίσεις ήταν υψηλές), ενώ στη διάρκεια της πρόσφατης υπάρχουν μεγαλύτερες συγκεντρώσεις μετοχών που έχουν υψηλά Market Caps και οι σημαντικές τοπικές συγκεντρώσεις φαίνονται να συσσωρεύονται μαζί.

Η απόφαση να εξερευνηθεί ο χώρος διάχυσης και οι τοπικές συγκεντρώσεις που σχηματίζονται σε αυτόν, για την περίοδο μετά κρίσης πριν την παραγωγή του προφίλ τοπικών συγκεντρώσεων ήταν καθαρά για λόγους ροής της παρούσας διπλωματικής. Η άμεση εξερεύνηση των χώρων για οποιαδήποτε περίοδο έχει ολικό νόημα μετά από την εξαγωγή των προφίλ τοπικών συγκεντρώσεων.



Σχήμα 6.10 : Απεικόνιση χώρου διάχυσης τριών διαστάσεων των μετοχών του δείκτη S&P 500, για τη περίοδο μετά κρίσης και η υπόδειξη με γκρι χρώμα των 5 μεγαλύτερων τοπικών συγκεντρώσεων (ε=0.05).



Σχήμα 6.11 : Απεικόνιση του προφίλ 15 τοπικών συγκεντρώσεων για τις περιόδους 2005-2007, 2007-2009, 2009-2012.

Κεφάλαιο 7

Συμπεράσματα

Στην παρούσα διπλωματική εργασία δόθηκε έμφαση στις μεθόδους μείωσης διαστάσεων και πώς η μέθοδος των απεικονίσεων διάχυσης βρίσκει μεγάλη εφαρμογή σε χρηματοοικονομικά δεδομένα. Η χρησιμοποίηση τους αποφέρει μεγάλη χρησιμότητα στις οπτικοποίηση και κατανόηση της συμμεταβολής των μετοχών, ενός χαρτοφυλακίου καθώς και στην εξαγωγή χρήσιμων ποσοτικών μέτρων για το ίδιο. Κάποιες χρήσιμες παρατηρήσεις που πρέπει να τονιστούν είναι οι ακόλουθες:

Η εφαρμογή των μεθοδολογιών που αναπτύχθηκαν στην εύρεση ποσοτικών μέτρων, όπως αυτή της εύρεσης μέτρου ολικής συγκέντρωσης και των μέτρων τοπικών συγκεντρώσεων δεν περιορίζεται μόνο για τον δείκτη του S&P 500, αλλά μπορούν να χρησιμοποιηθούν και για ενεργά χαρτοφυλάκια μετοχών και να διεξαχθεί μια συγκριτική μελέτη.

Στην περίπτωση του μέτρου της ολικής συγκέντρωσης μπορούν να εξαχθούν μέτρα για το εκάστοτε χαρτοφυλάκιο χρησιμοποιώντας τις μετοχές (οι οποίες πρέπει να είναι συστατικά του δείκτη S&P 500) αλλά και τα εκάστοτε βάρη του ιδίου χαρτοφυλακίου. Ο λόγος μεταξύ του μέτρου ολικής συγκέντρωσης του εκάστοτε χαρτοφυλακίου και του δείκτη S&P 500 μπορεί να θεωρηθεί ως το σχετικό μέτρο ολικής συγκέντρωσης (relative global concentration measure). Χρησιμοποιώντας το νέο αυτό μέτρο υποδεικνύεται ο τρόπος σύγκρισης μεταξύ του χαρτοφυλακίου που μελετάται και αυτού του δείκτη.

Η ίδια συγκριτική μελέτη μπορεί να διεξαχθεί και για το μέτρο της τοπικής συγκέντρωσης ενός ενεργού χαρτοφυλακίου. Χρησιμοποιώντας πάλι τις μετοχές και τα εκάστοτε βάρη τους στο χαρτοφυλάκιο μπορεί να εξαχθεί το προφίλ τοπικής συγκέντρωσης του και η περαιτέρω συγκριτική μελέτη του με αυτή του προφίλ του δείκτη S&P 500.

Η μέθοδος των diffusion maps όπως έχει τονιστεί στο κεφάλαιο 4 του πρώτου μέρους είναι αρκετά ανθεκτική στον θόρυβο δεδομένων σε σχέση με ανταγωνιστικές της όπως είναι αυτή του Isomap η οποία επηρεάζεται σε μεγάλο βαθμό από αυτόν. Αυτό είναι πολύ σημαντικό όταν μελετιούνται χρηματοοικονομικά δεδομένα στα οποία η εύρεση θορύβου είναι πολύ συχνό φαινόμενο. Ένα μειονέκτημα που πρέπει να αναφερθεί είναι πως οι συντεταγμένες των μετοχών στους χώρους διάχυσης δεν έχουν κάποιο οικονομικό νόημα. Αυτό συμβαίνει καθώς οι χώροι διάχυσης εκφράζουν τα δεδομένα σε

τέτοιο βαθμό ώστε να διατηρείται η απόσταση διάχυσης και μόνο αυτή, με τον καλύτερο δυνατό τρόπο. Ακριβώς και για αυτό το λόγο η περιστροφή και αναδίπλωση του εκάστοτε χώρου διάχυσης περιέχει την ίδια πληροφορία.

Αντίστοιχα με τον πυρήνα συσχετίσεων που χρησιμοποιήθηκε στην παρούσα διπλωματική είναι εφικτό να εφαρμοστούν και άλλοι πυρήνες οι οποίοι εκφράζουν συσχετίσεις μεταξύ των μετοχών. Τέτοιοι πυρήνες για να χρησιμοποιηθούν στη μέθοδο των diffusion maps θα πρέπει να έχουν στοιχεία τα οποία όσο μεγαλύτερες τιμές έχουν τόσο μεγαλύτερη είναι η συσχέτιση μεταξύ δύο αντικειμένων. Επίσης θα πρέπει να τονιστεί ξανά πως τα στοιχεία του θα πρέπει να είναι θετικά και ο πίνακας του πυρήνα συμμετρικός. Τέτοιοι πυρήνες είναι οι ακόλουθη :

- R^2kernel: Ο $K(x, y)$ είναι ο συντελεστης R^2 της γραμμικής παλινδρόμηση ελαχίστων τετραγώνων μεταξύ των μετοχών x, y.
- Angles kernel: Θεωρώντας τις μετοχές ως διανύσματα σε χώρο n διαστάσεων, τα στοιχεία του πυρήνα, θεωρούνται οι γωνίες μεταξύ των διανυσμάτων αυτών. Ποιο συγκεκριμένα η ομοιότητα μεταξύ δύο μετοχών x, y θεωρείται η γωνία που σχηματίζουν τα διανύσματα αυτά, θεωρώντας ως συντεταγμένες τους τις αποδόσεις τους.
- The distance kernel: Ο πυρήνας διάχυσης που μελετήθηκε στο κεφάλαιο των diffusion maps είναι ένας τέτοιος πυρήνας.

Κεφάλαιο 8

Παράρτημα

Σε αυτό το κεφάλαιο δίνονται εκτενώς οι κώδικες που χρησιμοποιήθηκαν για την παραγωγή των αποτελεσμάτων του έκτου κεφαλαίου. Η ανάπτυξη των αλγορίθμων επιτευχθεί με την γλώσσα προγραμματισμού MATLAB_R (2016a).

Οι τιμές των μετοχών φορτώθηκαν αρχικά μέσω του του προγράμματος excel και τροποποιήθηκαν έτσι ώστε να προκύψουν οι μηνιαίες αποδόσεις και τα εκάστοτε βάρη όπως και οι καθαρές μηνιαίες αποδόσεις του χαρτοφυλακίου. Η ανάλυση και η εξαγωγή μέτρων, διαγραμμάτων πραγματοποιήθηκε μέσω του προγράμματος MATLAB_R(2016a).

Για την διευκόλυνση του αναγνώστη υπάρχουν επεξηγηματικά σχόλια σε όλα τα scripts και σε όλες τις συναρτήσεις τα οποία ακολουθούνται μετά από το σύμβολο σχολιασμού %. Ακόμα οι κώδικες δίνονται με τη σειρά που παρουσιάζονται τα αποτελέσματα τους στη παρούσα εργασία.

Πριν παρατεθούν, είναι σημαντικό να αναφερθεί πως ο κώδικας υπολογισμού των diffusion maps στο τελευταίο κεφάλαιο πάρθηκε από αυτόν που δίνεται στη δημοσίευση των Ann B Lee and Larry Wasserman (Spectral connectivity.March 2009). Η παραλλαγή που υπέστη η συνάρτηση ήταν αυτή του πυρήνα καθώς χρησιμοποιήθηκε ο $\mathrm{K}(\mathrm{x},\mathrm{y}) = 1 + \mathrm{r}(\mathrm{x},\mathrm{y})$ και μετονομάστηκε σε diffuse_co_Corp.

Οι βασικοί πίνακες δεδομένων που χρησιμοποιήθηκαν για την παραγωγή των διαγραμμάτων είναι αυτοί, των λογαριθμικών αποδόσεων του δείκτη S&P 100, των λογαριθμικών αποδόσεων, κεφαλαιοποιήσεων, βαρών και καθαρών αποδόσεων για τις μετοχές του δείκτη S&P 500. Τα αντίστοιχα ονόματα τους είναι τα ακόλουθα: R100, R500,MARKETCAP500, weights500, ER_500. Επίσης χρησιμοποιήθηκαν και τα κελιά χαρακτήρων (cell arrays) που περιείχαν τα ονόματα και τους τομείς που προέρχονται οι εκάστοτε μετοχές και για τους δύο δείκτες, με αντίστοιχα ονόματα : names100, names500, sector100, sector500.

Σχήμα 2.1,2.2,2.3 (textplot_for_stocks_of_S&P 100)

```
color1= [1 0.4 1]; %purple, capital goods
color2= [0 0.2 0.8]; %blue, technology
color3= [0 0.6 0.2]; %green, health care
color4= [0.2 0.9 1]; %blue light, finance
color5= [0.8 0 0.2]; %red, consumer services
color6= [0.8 0.4 0.2]; % brown, transportation
color7= [0.2 0 0.2]; %black, public utilities
color8= [0.5 0.5 0.5]; %grey, basic industries
color9= [1 0.8 0.2]; %yellow, energy
%give colors to sectors
 for i= 1: size(sector100);
    if strcmp(sector100(i),'Capital Goods')
       sector100{i}=color1;
   elseif strcmp(sector100(i),'Technology')
       sector100{i}=color2;
   elseif strcmp(sector100(i),'Health Care')
       sector100{i}=color3;
   elseif strcmp(sector100(i),'Finance')
       sector100{i}=color4;
   elseif strcmp(sector100(i),'Consumer Services')
       sector100{i}=color5;
   elseif strcmp(sector100(i),'Consumer Non-Durables')
       sector100{i}=color5;
   elseif strcmp(sector100(i),'Miscellaneous')
       sector100{i}=color5;
   elseif strcmp(sector100(i),'Transportation')
       sector100{i}=color6;
   elseif strcmp(sector100(i),'Public Utilities')
       sector100{i}=color7;
   elseif strcmp(sector100(i),'Basic Industries')
       sector100{i}=color8;
   else sector100{i}=color9;
    end
end
diff100=diffuse_co_corp(R100,3,1); %Compute embedded space for S&P 100
figure
xlim ([-0.13 0.15]); ylim ([-0.09 0.13]);zlim ([-0.09 0.13])
for K = 1: 108
text(diff100(K,1), diff100(K,2), diff100(K,3), names100{K}, 'color', sector100{K});
end
grid on
title ('3-D Diffusion Map Embedded Space')
%title ('2-D Diffusion Map Embedded Space')
 xlabel ('eigenvector 1'); ylabel ('eigenvector 2'); zlabel ('eigenvector 3')
```

Σχήμα 2.5 (three_periods_scatter_plot)

```
subperiod1=R500 (:,41:65); %log-Returns for May 05 - June 07 ''credit boom period''
subperiod2=R500 (:,66:92); %log-Returns for July 07 - Sept 09 ''crisis period''
subperiod3=R500 (:,93:123); %log-Returns for oct 09 - April 12 ''recent period''


DIFF1=diffuse_co_corp(subperiod1,3,1); % Compute embedded space for first period
DIFF2=diffuse_co_corp(subperiod2,3,1); % Compute embedded space for second period
DIFF3=diffuse_co_corp(subperiod3,3,1); % Compute embedded space for third period
market1=mean (MARKETCAP500 (:,41:65),2)./100000;% Compute the average market
caps  for first period for visualization purposes
market2=mean (MARKETCAP500 (:,66:92),2)./100000;% Compute the average market
caps  for second period for visualization purposes
market3=mean (MARKETCAP500 (:,93:123),2)./100000;% Compute the average market
caps for third period for visualization purposes
%%%%%%visualize stocks, according to the average of their market caps%%%%%%%%
for l=1:296
   if market1(l)<=0.25;
   size1(l)=10;
   elseif market1(l)>=0.25 && market1(l)<=0.30;
   size1(l)=15;
   elseif market1(l)>=0.30 && market1(l)<=0.50;
   size1(l)=20;
   elseif market1(l)>=0.50 && market1(l)<=1;
   size1(l)=50;
   elseif market1(l)>=1 && market1(l)<=1.5;
   size1(l)=65;
   elseif market1(l)>=1.5 && market1(l)<=2;
   size1(l)=80;
   elseif market1(l)>=2 && market1(l)<=2.5;
   size1(l)=100;
   else
   size1(l)=160;
   end
end
for l=1:296
   if market2(l)<=0.25;
   size2(l)=10;
   elseif market2(l)>=0.25 && market2(l)<=0.30;
   size2(l)=15;
   elseif market2(l)>=0.30 && market2(l)<=0.50;
   size2(l)=20;
   elseif market2(l)>=0.50 && market2(l)<=1;
   size2(l)=50;
   elseif market2(l)>=1 && market2(l)<=1.5;
```

```
  size2(l)=65;
  elseif market2(l)>=1.5 && market2(l)<=2;
  size2(l)=80;
  elseif market2(l)>=2 && market2(l)<=2.5;
  size2(l)=100;
  else
  size2(l)=160;
  end
end
for l=1:296
  if market3(l)<=0.25;
  size3(l)=10;
  elseif market3(l)>=0.25 && market3(l)<=0.30;
  size3(l)=15;
  elseif market3(l)>=0.30 && market3(l)<=0.50;
  size3(l)=20;
  elseif market3(l)>=0.50 && market3(l)<=1;
  size3(l)=50;
  elseif market3(l)>=1 && market3(l)<=1.5;
  size3(l)=65;
  elseif market3(l)>=1.5 && market3(l)<=2;
  size3(l)=80;
  elseif market3(l)>=2 && market3(l)<=2.5;
  size3(l)=100;
  else
  size3(l)=160;
  end
end
 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
figure;
%give colors to stocks for different periods
aa= [0.0,0.6,1.0];
bb= [1.0,0,0.6];
cc= [0.3,0.0,0.0];
hold on;
for i=1:296
y1=scatter3(DIFF1(i,1), DIFF1(i,2), DIFF1(i,3), size1(i), aa,' filled'); %scatter plot for the first    period
y2=scatter3(DIFF2(i,1), DIFF2(i,2), DIFF2(i,3), size2(i), bb, 'filled'); %scatter plot for the second period
y3=scatter3(DIFF3(i,1), DIFF3(i,2), DIFF3(i,3), size3(i), cc, 'filled'); %scatter plot for the third period
legend([y1,y2,y3],'credit-boom period','crisis period','recent period');
end
grid on;
title ('3-D Scatter Plot for 3 Periods 2005-2007, 2007-2009, 2009-2012')
xlabel('eigenvector 1');ylabel('eigenvector 2');zlabel('eigenvector 3')
```

Σχήμα 2.7(rolling_window)

```
%create 112 matrices (length 12 month) of log-returns and weights for rolling window
purposes.
for I = 1:112;
  Xrolling{i} = R500(:,i:i+11);
   Wrolling{i} = W500(:,i:i+11);
    k = 0;
end
for time=1:112;
%compute 5 dimensions embedded space and the average weight for each stock
[DiffR{time},lamda] = diffuse_co_corp(Xrolling{time},5,1)
Weights{time} = mean(Wrolling{time},2);
%compute global concentration measure for each period
CGC(time)=(trace(weightedcov(DiffR{time},Weights{time}))).^-(1/2);
end
figure
x = 2003:1/12:2012.3
plot (x,CGC,'.-')
set (gca, 'XTick', 2003:20012.3)
grid on
title ('S&P 500 Global Concentration Measure, Rolling Window 12-month')
ylabel ('G.C.M')
```

Σχήμα 2.8 (2_period_scatter_plot_for_energyandpharmaceutical-healthcare_comp.)

```
%3-D scatter to show energy and pharmaceutical-healthcare companies in
%credit boom period and crisis period.
%Use the same data results as in three_periods_scatter script (DIFF1, DIFF2,size1,size2)
figure;
color1 = [0 0.6 0.2]; %green --- health care
color2 = [1 0.8 0.2]; %yellow ---energy
color3 = [0.3,0.0,0.0]; % black --- others
for i = 1: size(sector500);
   if strcmp(sector500(i),'Health Care')
      sector500{i} = color1;
   elseif strcmp(sector500(i),'Energy')
      sector500{i} = color2;
   else sector500{i} = color3;
   end
end
%embedded space for credit boom period with energy and pharmaceutical companies
colored
  figure
```

```
    for K = 1:296
y1 = scatter3(DIFF1(K,1), DIFF1(K,2),DIFF1(K,3),size1(K),sector500{K},'filled');
hold on
 end
xlim([-0.5 0.5]);ylim([-0.5 0.5]);zlim([-0.5 0.5])
title ('3-D Scatter Plot for Period 2002-2005')
xlabel('eigenvector 1');ylabel('eigenvector 2');zlabel('eigenvector 3')

%embedded space for crisis period with energy and pharmaceutical companies colored
figure
    for K=1:296
y2 = scatter3(DIFF2(K,1), DIFF2(K,2), DIFF2(K,3), size2(K), sector500{K}, 'filled');
hold on
    end
xlim([-0.5 0.5]);ylim([-0.5 0.5]);zlim([-0.5 0.5])
title ('3-D Scatter Plot for Period 2005-2007')
xlabel('eigenvector 1');ylabel('eigenvector 2');zlabel('eigenvector 3')
```

<u>ΣΥΝΑΡΤΗΣΗ ΕΥΡΕΣΗΣ ΜΕΤΡΩΝ ΚΑΙ ΘΕΣΕΩΝ ΣΤΟ ΧΩΡΟ ΤΟΠΙΚΩΝ ΣΥΓΚΕΝΤΡΩΣΕΩΝ</u>

```
function [measure, positions] = local_concentration(X,Y,r,epsilon,dimensions,numbers)
%local concentration function
% output: measure of local concentration and positions of points in space in
%addition
%X: matrix of returns for specific period
%Y: market cap of stocks for specific period
%r: random points in space for example
%r = -0.3 + (0.3+0.3) *rand (100000,3);
%epsilon: give scalar e
%dimensions: dimensions for the diffusion map
%numbers=numbers of local values
LOCAL = diffuse_co_corp(X, dimensions,1);
MARKETCAP = mean(Y,2);
Sigma = sqrt(epsilon/2)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%figure
%scatter3(r(:,1),r(:,2),r(:,3),10,'b','filled');
%hold on
%scatter3(LOCAL(:,1),LOCAL(:,2),LOCAL(:,3),10,'r','filled');
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% It's convenient to take a symmetrical normal density
% function, rescaled to have unit maximum.
% a: constant to give value function unit maximum
% f(x)=exp(-0.5 * ((x-a)./sigma).^2)./ (sqrt(2*pi).* sigma))
% i want f(x)<1 then (x-a) ^2 > -epsilon*ln(sqrt(epsilon*pi))
a = sqrt(-epsilon*log(sqrt(epsilon*pi)))
for k = 1: numbers;
```

```
 for i = 1: size(r,1);
  % compute distances between random points and stocks in embedded space
   QQ = [r(i,:);LOCAL];
   Distances = squareform (pdist (QQ,'euclidean'));
   x{i} = Distances (1,2: end);
   % compute prices for shock function which belongs in [0,1]
   value_function{i} = (exp(-0.5 * ((x{i}+a)./sigma).^2)./ (sqrt(2*pi).* sigma));
   % compute measures of local concentration for each random point i in space (sum
prices of schock functions with Market Caps)
   m(i) = value_function{i}*MARKETCAP;
 end
 % find the maximum global concentration measure and its position
 [M, I] = max(m);
 measure(k) = M;
 positions(k) = I;
 % re-compute Market Caps to find less important global concentration measures
 MARKETCAP = MARKETCAP. *(1-value_function{I})';
end
 return;
```

Σχήμα 2.9 (recent_period_scatter_plot_with_5_L.C.M's)

```
subperiod1 = R(:,93:123);% log-Returns for the period  Oct 09 - April 12 ''recent period''
DIFF1 = diffuse_co_corp(subperiod1,3,1); %compute embedded space
market1 = mean(MARKETCAP(:,93:123),2)./100000; %compute average market cap for
each stock
%%%%%%visualize stocks, according to the average of their market caps%%%%%%
for l=1:296
   if market1(l)< = 0.25;
   size1(l) = 10;
   elseif market1(l)> = 0.25 && market1(l)< = 0.30;
   size1(l) = 15;
   elseif market1(l)> = 0.30 && market1(l)< = 0.50;
   size1(l) = 20;
   elseif market1(l)> = 0.50 && market1(l)< = 1;
   size1(l) = 50;
   elseif market1(l)> = 1 && market1(l)< = 1.5;
   size1(l) = 65;
   elseif market1(l)> = 1.5 && market1(l)< = 2;
   size1(l) = 80;
   elseif market1(l)> = 2 && market1(l)< = 2.5;
   size1(l) = 100;
   else
   size1(l) = 160;
   end
end
```

```
r = -0.25 + (0.25+0.25) *rand (500000,3); % random points in embedded space
% compute measures and positions of local concentrations for recent period
[measure0, positions0] = local_concentration (subperiod1,market1, r, 0.05,3,5)
figure;
aa = [0.0,0.6,1.0];
FF = [0.6,0.6,0.6];
hold on;
% scatter plot of stocks in embedded space with their market caps
for i=1:296
y1 = scatter3(DIFF1(i,1), DIFF1(i,2), DIFF1(i,3), size1(i), aa, 'filled');
end
hold on;
% scatter plot of five biggest concentrations in embedded space
scatter3(r(positions0(1),1), r(positions0(1),2), r(positions0(1),3), 350, FF, 'filled');
scatter3(r(positions0(2),1), r(positions0(2),2), r(positions0(2),3), 300, FF, 'filled');
scatter3(r(positions0(3),1), r(positions0(3),2), r(positions0(3),3), 250, FF, 'filled');
scatter3(r(positions0(4),1), r(positions0(4),2), r(positions0(4),3), 200, FF, 'filled');
scatter3(r(positions0(5),1), r(positions0(5),2), r(positions0(5),3), 150, FF, 'filled');
hold on

% scatter plot of random stocks in embedded space near L.C. M's

text(DIFF1(77,1),DIFF1(77,2),DIFF1(77,3), 'CSCO'); % Sisco systems
text(DIFF1(291,1),DIFF1(291,2),DIFF1(291,3), 'XOM');% Exxon mob. comp.
text(DIFF1(99,1),DIFF1(99,2),DIFF1(99,3), 'EMC');% Emc corp
text(DIFF1(3,1),DIFF1(3,2),DIFF1(3,3), 'APPL'); % Apple
text(DIFF1(87,1),DIFF1(87,2),DIFF1(87,3), 'DELL');% Dell
text(DIFF1(196,1),DIFF1(196,2),DIFF1(196,3), 'NKE');% Nike
text(DIFF1(183,1),DIFF1(183,2),DIFF1(183,3), 'MMM'); % 3M comapny
text(DIFF1(133,1),DIFF1(133,2),DIFF1(133,3), 'HOG'); % harley davinson
text(DIFF1(205,1),DIFF1(205,2),DIFF1(205,3), 'ORCL'); % Orcl company
text(DIFF1(79,1),DIFF1(79,2),DIFF1(79,3), 'CPB'); % Campell soup company
text(DIFF1(220,1),DIFF1(220,2),DIFF1(220,3), 'PG'); % Procter and gample company
text(DIFF1(53,1),DIFF1(53,2),DIFF1(53,3), 'BMC'); % Bmc software company


grid on;
title ('3-D Scatter Plot for Credit Boom Period and the 5 Biggest Concentrations')
xlabel('eigenvector 1');ylabel('eigenvector 2');zlabel('eigenvector 3')
```

Σχήμα 2.10 (profile_of_local_concentrations_for_3_periods)

```
subperiod1R = R(:,41:65);%july 03 - june 07 ''credit boom''
subperiod2R = R(:,66:92);%july 07 - june 09 ''crisis period''
subperiod3R = R(:,93:123);%july 09 - april 12 ''recent period''
subperiod1M = MARKETCAP(:,41:65);%july 05 - june 07 ''credit boom''
subperiod2M = MARKETCAP(:,66:92);%july 07 - june 09 ''crisis period''
subperiod3M = MARKETCAP(:,93:123);%july 10 - april 12 ''recent period''
r = -0.25 + (0.25+0.25) *rand(500000,3);
[measure1, positions1 ] = local_concentration(subperiod1R,subperiod1M,r,0.05,5,20);
[measure2, positions2 ] = local_concentration(subperiod2R,subperiod2M,r,0.05,5,20);
[measure3, positions3 ] = local_concentration(subperiod3R,subperiod3M,r,0.05,5,20);

Barplot = [measure1; measure2; measure3]';
xdata = [1 2 3 4 5 6 7 8 9 10 11 12 13 14 15];
ydata = barplot;
clr = [0 0.6 1.0;
1.0 0.0 0.6;
0.3 0.0 0.0];
colormap(clr);
H = bar (xdata, ydata);
l = cell (1,3);
l{1} = 'credit-boom period' ;  l{2} = 'crisis period' ;  l{3}='recent period';
legend(H,l);
grid on;
xlabel ('Numbers of Concentrations')
ylabel (' Local Concentration Measure')
title ('S&P 500 Local Concentration Profile, 2005-2007 vs. 2007-2009 vs. 2009-2012')
```

Κεφάλαιο 9

Βιβλιογραφία

Bah, B. (2008). Diffusion maps: analysis and applications. PhD thesis, University of Oxford,U.K.

Belkin, M. & Niyogi, P. (2001). Laplacian eigenmaps and spectral techniques for embedding and clustering. 'NIPS', 14, 585–591.

Belkin, M. & Niyogi, P. (2008). Towards a theoretical foundation for laplacian-based manifold methods. Journal of Computer and System Sciences, 74(8), 1289–1308.

Belkin, M., Niyogi, B. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, 15(6), 1373-1396.

Billio, M., Getmansky, M., Lo, A. W., Pelizzon L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. Journal of Financial Economics, 104, 535–559.

Bishop, C.M. (2006). Pattern Recognition and Machine Learning. Springer.

Cattell, R.B. (1966). The scree test for the number of factors, Multivariate Behavioral Research, 1, 245–276.

Chang, Y. (2014). Graph Embedding and Extensions: A GeneralFramework for Dimensionality Reduction. Technical Report.

Chen, H.C. (2014). Visualization of financial time series by linear principal component analysis and nonlinear principal component analysis. Master thesis, University of Leicester,U.K.

Coifman, R.R., Lafon, S. (2006). Diffusion Maps. Appl. Comput. Harmonic Anal, 21, 5-30.

Cormen, T., Leiserson, C., Rivest, R. (1990). Introduction to Algorithms. MIT Press and McGraw-Hill. First Edition.

Cox, T., Cox. M. (2001) Multidimensional Scaling. Chapman Hall, Boca Raton, 2nd edition.

de la Porte, J., Herbst, B., Hereman, W., van der Walt, S. (2008). An Introduction to Diffusion Maps. Proceedings of the Nineteenth Annual Symposium of the Pattern

Recognition Association of South Africa (PRASA), 125-130.

Driesson, J., Melenberg, B., Nijman,T. (2003). Common factors in international bond returns. Journal of International Money and Finance, 22, 629–656.

Feeney, G. J., Hester, D. D. (1967). Risk Aversion and Portfolio Choice. New York: Wiley. Financial performance analysis, 14(2), 148-167.

Fouss, F., Saerens, M., Shimbo, M. (2016). Algorithms and Models for Network Data and Link Analysis. Cambridge university press.

Gao, B., Berendt, B. (2011). Visual data mining for higher-level patterns: discrimination-aware data mining and beyond. In: Proceedings of the 20th machine learning conference of Belgium and The Netherlands.

Ghodsi, A. (2007). Department of Statistics and Actuarial Science University of Waterloo, STAT 442 / 890, CM 462, Lecture 3,4,5,10,11.

Groenen, J.F., Borg,I. (2013) The Past, Present, and Future of Multidimensional Scaling Econometric Institute Report EI.

Hargreaves, C.A., Mani, C.K. (2015). The Selection of Winning Stocks Using Principal Component Analysis. American Journal of Marketing Research, 1(3), 183-188.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology. 24, 417–441, and 498–520.

Huang, Y., Kou, G. (2014). A kernel entropy manifold learning approach for financial data analysis. Decision Support Systems, 64, 31-42.

Huang, Y., Kou, G., Peng, Y. (2017). Nonlinear manifold learning for early warnings in financial markets. European Journal of Operational Research, 258(2), 692–702.

Jolliffe, I. (2002). Principal component analysis. Wiley Online Library.

Kritzman, M.,Li, Y., Page, S., Rigobon, R. (2011). Principal Components as a measure of systemic risk. Journal of Portfolio Management, 37, 112–126.

Lee, A.B., Wasserman, L. (2010). Spectral Connectivity Analysis. Journal of the American Statistical Association, 105(491), 1241-1255.

Lerman, G. (2005). University of Minnesota. Course on geometric data analysis. MANI: Manifold Learning Toolkit.

Lin, F., Yeh, C.C., Lee, M.Y. (2011). The use of hybrid manifold learning and support vector machines in the prediction of business failure. Knowledge-Based Systems, 24(1), 95-101.

Lindsay, S. (2002). A tutorial on principal component analysis. Details available at /csnet.otago.ac.nz/cosc453/student_tutorials/principal_ components.

Liu, R., Cai, H., Luo, C. (2012). Cluster Analysis of Stocks of CSI 300 Index Based on Manifold Learning. Journal of Intelligent Learning Systems and Applications, 4, 120-126.

Meyer, C.D. (2000). Matrix analysis and applied linear algebra. SIAM.

Nadler, B., Lafon, S., Coifman, R.R., Kevrekidis, I. (2006). Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. Applied and Computational Harmonic Analysis, 21(1), 113-127.

Pe'rignon, C., Smith, D.R., Villa, C. (2007). Why common factors in international bond returns are not so common. Journal of International Money and Finance, 26, 284–304.

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine, 2(11), 559–572.

Phoa, W. (2012). Portfolio Concentration and the Geometry of Co-Movement. The Journal of Portfolio Management.

S. Lafon. (2004) Diffusion maps and geometric harmonics. PhD thesis, Yale University, U.S.A.

Samko, O., Marshall, A., Rosin, P. (2006). Selection of the optimal parameter value for the Isomap algorithm. Pattern Recognit Lett, 27(9), 968–79.

Sarlin, P. (2013). Data and dimension reduction for visual financial performance analysis Information Visualization, 14(2), 148-167.

Shlens, J. (2009). A Tutorial on Principal Component Analysis. Cornell University Library Computer Science Learning, available at: http://arxiv.org/abs/1404.1100.

Singer, A., Erban, R., Kevrekidis, I., Coifman, R.R. (2007). Detecting the slow manifold by anisotropic diffusion maps. PNAS, 106(38), 16090-16095.

Talmon, R., Cohen, I., Gannot, S., Coifman, R. R. (2013). Diffusion maps for signal processing: A deeper look at manifold-learning techniques based on kernels and graphs. IEEE Signal Process. Mag., 30(4), 75–86.

Tenenbaum, J. B., De Silva, V., Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. Science, 290(5500), 2319–2323.

Tsay, S.R. (2005). Analysis of Financial Time Series. John Wiley & Sons, Second Edition.

Vaidya, U., Hagen, G., Banaszuk, A., Lafon S., Mezic I., Coifman R. R. (2005). Comparison of systems using diffusion maps in Decision and Control. 44th IEEE Conference,7931–7936.

von Luxburg, U. (2007). A Tutorial on Spectral Clustering. Statistics and Computing, 17(4), 395-416.

Wang, J. (2011). Geometric Structure of High-Dimensional Data and Dimensionality Reduction. Springer.

Yang, L., W. S. Rea, A. Rea (2015). Stock Selection with Principal Component Analysis. https://ideas.repec.org/p/cbt/econwp/15-03.html.

Yen, L., Saerens, M., Fouss, F. (2010). A Link Analysis Extension of Correspondence Analysis for Mining Relational Databases. IEEE transaction on Knowledge and Data Engineering, 23, 481-495.

Zhang, J.P., Huang, H., J. Wang, J. (2010). Manifold Learning for Visualizing and Analyzing High-dimensional Data. IEEE Intelligent Systems, 25(4), 54-61.

Zheng, Z., Podobnik, B., Feng, L., Li, B. (2012). Changes in cross-correlations as an indicator for systemic risk. Scientific Reports, 2, 888.