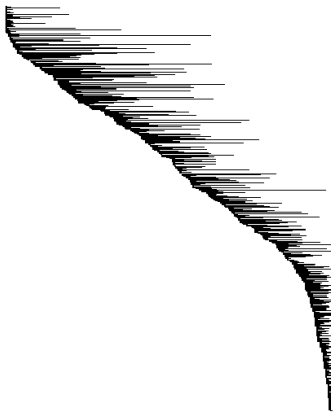
ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
"ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΑΝΑΛΟΓΙΣΤΙΚΑ – ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΑ ΜΑΘΗΜΑΤΙΚΑ"

ΤΟΠΟΛΟΓΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Διπλωματική Εργασία

ΣΥΚΟΒΑΡΙΔΟΥ ΣΟΦΙΑ



Επιβλέπων Καθηγητής
Ξανθόπουλος Στυλιανός

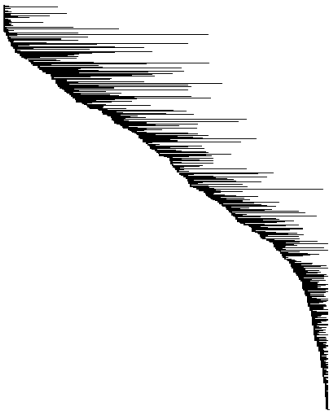
Σάμος, Ιούνιος 2018

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
"ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΑΝΑΛΟΓΙΣΤΙΚΑ – ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΑ ΜΑΘΗΜΑΤΙΚΑ"

ΤΟΠΟΛΟΓΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Διπλωματική Εργασία

ΣΥΚΟΒΑΡΙΔΟΥ ΣΟΦΙΑ



Τριμελής Επιτροπή

Ξανθόπουλος	Καραγρηγορίου	Τσαπόγας
Στυλιανός	Αλέξανδρος	Γεώργιος
Αν. Καθηγητής	Καθηγητής	Καθηγητής

Σάμος, Ιούνιος 2018

Ευχαριστίες

Ευχαριστώ θερμά τον επιβλέποντα καθηγητή κύριο Στέλιο Ξανθόπουλο για τη δυνατότητα που μου προσέφερε να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα, καθώς και την πολύτιμη καθοδήγησή του κατά την εκπόνηση της παρούσας μεταπτυχιακής εργασίας.

Ιδιαίτερες ευχαριστίες θέλω να απευθύνω στον φίλο και συνάδελφο Δημήτριο Δουλφή για το ειλικρινές ενδιαφέρον του και τη σημαντική βοήθειά που μου προσέφερε σε όλα τα στάδια της εργασίας.

Τέλος, θέλω να ευχαριστήσω την οικογένεια μου για τη στήριξη και την ενθάρρυνση που μου προσέφεραν μέχρι τελευταία στιγμή.

Περίληψη

Συνήθως, τα δεδομένα είναι υπό μορφή πεπερασμένου συνόλου θορυβωδών σημείων, με δείγμα από έναν άγνωστο χώρο και ενσωματωμένο σε ένα χώρο μεγάλης διάστασης. Κατά την Τοπολογική Ανάλυση Δεδομένων γίνεται προσπάθεια ανάκτησης του χώρου δειγματοληψίας. Κύριος στόχος της παρούσας διπλωματικής αποτελεί η παρουσίαση δύο βασικών μεθόδων της Τοπολογικής Ανάλυσης Δεδομένων, της Εμμένουσας Ομολογίας και του αλγορίθμου Mapper. Η Εμμένουσα Ομολογία υπολογίζει τοπολογικά χαρακτηριστικά του χώρου δειγματοληψίας κάνοντας χρήση αλγεβρικής τοπολογίας, ενώ ο αλγόριθμος Mapper αποτελεί μία απεικονιστική μέθοδο που παρέχει πληροφορίες για το σύνολο των δεδομένων συμπιέζοντας ένα σύνολο δεδομένων μεγάλων διαστάσεων και κατασκευάζοντας τοπολογικά δίκτυα.

Λέξεις – Κλειδιά : Τοπολογική Ανάλυση Δεδομένων, Πλεγματικά Σύμπλοκα, Εμμένουσα Ομολογία, Barcodes, Mapper

Abstract

Data is often in the form of a finite set of noisy points, sampled from an unknown space and embedded in a high dimensional space. TDA recovers the structure of the sampled space. Main purpose of the present master thesis is the presentation of two basic methods of Topological Data Analysis, Persistent Homology and Mapper Algorithm. Persistent homology is an algebraic tool for measuring topological features of the sampled space, while mapper algorithm is a visualization method that provides information for data, compresses a set of large dimension data and constructs topological networks.

Keywords: Topological Data Analysis, Simplicial Complexes, Persistent Homology, Barcodes, Mapper

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΙΣΑΓΩΓΗ

Τις τελευταίες δεκαετίες, η ευρεία διαθεσιμότητα συσκευών μέτρησης και εργαλείων προσομοίωσης (simulation tools) έδωσε τη δυνατότητα συλλογής μεγάλου όγκου δεδομένων σε υψηλότερες διαστάσεις σε όλους σχεδόν τους τομείς της επιστήμης, της βιομηχανίας, της οικονομίας ακόμη και την καθημερινή ζωή. Το γεγονός αυτό κατέστησε επιτακτική την ανάγκη ανάπτυξης νέων μεθόδων για την αποτελεσματικότερη ανάλυση των δεδομένων. Μία σύγχρονη οπτική του θέματος αποτελεί η Τοπολογική Ανάλυση Δεδομένων (Topological Data Analysis ή TDA).

Στην Τοπολογική Ανάλυση Δεδομένων, τα δεδομένα αντιμετωπίζονται ως αντικείμενα μεγάλων διαστάσεων. Στόχος της είναι να συναγάγει σχετικές, ποιοτικές και ποσοτικές τοπολογικές δομές απευθείας από τα δεδομένα [6]. Συνήθως, τα επιστημονικά δεδομένα είναι υπό μορφή πεπερασμένου συνόλου θορυβώδους σημείων, με δείγμα από έναν άγνωστο χώρο και ενσωματωμένο σε ένα χώρο μεγάλης διάστασης. Η ανάλυση των τοπολογικών δεδομένων επικεντρώνεται στην ανάκτηση της τοπολογίας του χώρου δειγματοληψίας [39].

Οι δύο πιο δημοφιλείς προσεγγίσεις στην Τοπολογική Ανάλυση Δεδομένων είναι η Εμμένουσα Ομολογία (Persistent Homology) [14] και ο αλγόριθμος Mapper [32]. Όσον αφορά τον αλγόριθμο Mapper είναι μια απεικονιστική μέθοδος που διατηρεί την τοπολογική δομή, ενώ η Εμμένουσα Ομολογία παρέχει ένα πλαίσιο καθώς και αποτελεσματικούς αλγορίθμους για την κωδικοποίηση της εξέλιξης της τοπολογίας του σχήματος από μικρή σε μεγάλη κλίμακα [26].

Αξιοσημείωτες επιτυχίες της Τοπολογικής Ανάλυσης Δεδομένων αποτελούν η κατανόηση της τοπολογίας μια ομάδας γυναικών με καρκίνο του μαστού [28], η κατανόηση της τοπολογίας ορθοδοντικών δεδομένων [21] και η ανίχνευση γονιδίων με περιοδικό προφίλ [11].

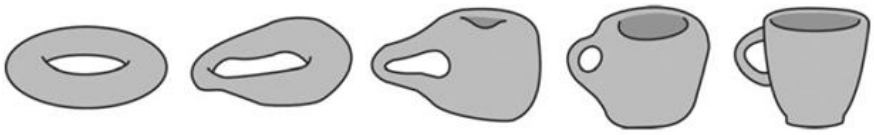
Κεφάλαιο 1

Βασικές Τοπολογικές Έννοιες

Η Τοπολογική Ανάλυση Δεδομένων αναφέρεται σ' ένα σύνολο μεθόδων εύρεσης τοπολογικών δομών απευθείας από τα δεδομένα. Συνεπώς, κρίνεται χρήσιμο στο πρώτο κεφάλαιο της παρούσας διπλωματικής να γίνει μια σύντομη αναφορά σε βασικές τοπολογικές έννοιες, οι οποίες θα φανούν ιδιαιτέρως χρήσιμες στην καλύτερη κατανόηση της TDA. Για την επίτευξη του παραπάνω σκοπού χρησιμοποιήθηκαν οι πηγές [2], [20], [35], [38].

Η συνολοθεωρητική Τοπολογία εκφράζει οτιδήποτε έχει σχέση με εγγύτητα, «γειτνίαση», σύγκλιση σ' οποιαδήποτε επιστήμη. Εμφανίσθηκε, αρχικά στα Μαθηματικά, ως μια μελέτη θέσεως, γι' αυτό χρησιμοποιήθηκε ο όρος Τοπολογία από την ελληνική λέξη τόπος, και στις αρχικές εργασίες ο Poincaré έδωσε τον λατινικό τίτλο Analysis Situs. Γρήγορα, ενδιαφέρθηκε για ιδιότητες, οι οποίες παρότι έχουν γεωμετρικό, άρα μεταβαλλόμενο, χαρακτήρα, μένουν σταθερές π.χ. στη «σύνθλιψη» και στην «έκταση» των σχημάτων.

Αυτό μπορεί να γίνει καλύτερα αντιληπτό μέσω του παρακάτω παραδείγματος. Έστω δύο αντικείμενα που είναι κατασκευασμένα από ελαστικό καουτσούκ. Στην τοπολογία επιτρέπεται να τεντώσουμε το αντικείμενο, να το περιστρέψουμε και οι τοπολογικές του ιδιότητες να διατηρούνται. Αυτό, όμως, που δεν επιτρέπεται είναι ο διαχωρισμός ή η ένωση σημείων του αντικειμένου. Έτσι, το ντόνατ είναι τοπολογικά ισοδύναμο με ένα φλιτζάνι καφέ, διότι στον μετασχηματισμό από το πρώτο στο δεύτερο τηρούνται οι παραπάνω περιορισμοί (σχήμα 1.1).



Σχήμα 1.1 Μετασχηματισμός ντόνατ σε φλυτζάνι καφέ [24]

Βέβαια, η ιδέα της τοπολογικής ισοδυναμίας δεν περιορίζεται μόνο σε φυσικούς χώρους, αλλά επεκτείνεται σε γενικούς χώρους X.

1.1 Τοπολογία

Ορισμός 1.1.1 : Έστω ένα σύνολο X. Μια ***τοπολογία T του X*** είναι μία οικογένεια υποσυνόλων του X που ικανοποιούν τις ακόλουθες ιδιότητες :

i. Τα σύνολα $\emptyset, X$ ανήκουν στο T.
ii. Εάν $S_1,\ S_2 \in T$, τότε $S_1 \cap S_2 \in$ T.
iii. Εάν $\{S_j \ / \ j \in J\}$, τότε $\bigcup_{j\in J} S_j \in$ T.

Ένα σύνολο X για το οποίο ορίζεται μια τοπολογία T καλείται ***τοπολογικός χώρος*** και συμβολίζεται (X, T). Τα στοιχεία της T λέγονται ***ανοιχτά*** σύνολα του X. Ένα σύνολο $S \subset X$ λέγεται ***κλειστό***, όταν το συμπλήρωμα του $X - S$ είναι ανοιχτό.

Διαισθητικά, ένας τοπολογικός χώρος είναι ένα σύνολο σημείων, καθένα από τα οποία γνωρίζει το γείτονά του.

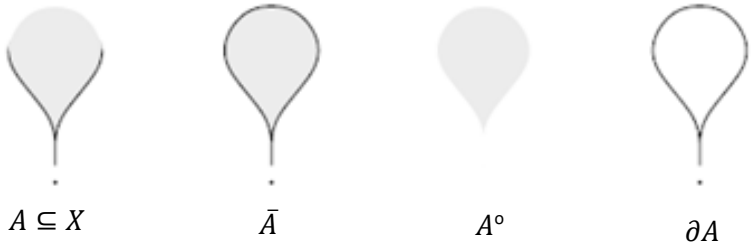
Ορισμός 1.1.2 : Ένα υποσύνολο $A \subseteq X$ με ***επαγόμενη τοπολογία*** $T_A = \{S \cap A \mid S \in T\}$ είναι ένας ***τοπολογικός υποχώρος*** A του X.

Από τον ορισμό της τοπολογίας, σε συνδυασμό με τους ορισμούς ανοιχτού και κλειστού συνόλου, προκύπτουν οι παρακάτω ιδιότητες :

i. Το $\emptyset$ είναι ανοιχτό σύνολο.
ii. Η ένωση οποιασδήποτε οικογένειας ανοιχτών υποσυνόλων του X είναι ανοιχτό υποσύνολο του X.
iii. Η τομή πεπερασμένης οικογένειας ανοιχτών υποσυνόλων του X είναι ανοιχτό υποσύνολο του X.

Γίνεται αντιληπτό ότι ένα σύνολο μπορεί να είναι μόνο κλειστό, μόνο ανοιχτό, ανοιχτό και κλειστό, ή τίποτα από τα δύο. Για παράδειγμα, το $\emptyset$ είναι ανοιχτό και κλειστό εξ' ορισμού.

Ορισμός 1.1.3 : Το ***εσωτερικό*** σύνολο $\boldsymbol{A^o}$ του συνόλου $A \subseteq X$ είναι η ένωση όλων των ανοιχτών συνόλων που περιέχονται στο A. Η ***κλειστότητα*** $\boldsymbol{\overline{A}}$ του συνόλου $A \subseteq X$ είναι η τομή όλων των κλειστών συνόλων που περιέχουν το σύνολο A. Το ***σύνορο*** ενός συνόλου A είναι $\boldsymbol{\partial A}$ = $\overline{\text{A}} - \text{A}^{\text{o}}$ (σχήμα 1.1.1).



Σχήμα 1.1.1 Κλειστότητα, εσωτερικό και σύνορο του συνόλου A

Ορισμός 1.1.4 : Μια ***γειτονιά*** του $x \in X$ είναι ένα σύνολο $A \subseteq X$ τέτοιο ώστε $x \in A^o$.

Η γειτονιά ενός σημείου είναι μία από τις βασικότερες τοπολογικές έννοιες και χρησιμοποιείται για να τυποποιήσει την «εγγύτητα» άλλων σημείων ως προς αυτό.

1.2 Ομοιομορφισμοί

Δοθέντος ενός συνόλου τοπολογικών χώρων, είναι σημαντικό να γίνει ορθός διαχωρισμός των χώρων σε σύνολα χώρων που συνδέονται με τον ίδιο τρόπο. Για την επίτευξη του παραπάνω στόχου γίνεται χρήση της έννοιας του ομοιομορφισμού.

Ορισμός 1.2.1 : Μια απεικόνιση $f : X \to Y$ καλείται ***συνάρτηση*** όταν αντιστοιχίζει κάθε στοιχείο του τοπολογικού χώρου X με ένα μοναδικό στοιχείο του τοπολογικού χώρου Y.

Ορισμός 1.2.2 : Μία συνάρτηση $f : X \to Y$ καλείται ***μονομορφισμός*** εάν για κάθε ζεύγος στοιχείων $x_1, x_2 \in X$ τέτοια ώστε $x_1 \neq x_2$, να ισχύει $f(x_1) \neq f(x_2)$. Η f καλείται, επίσης, $1-1$.

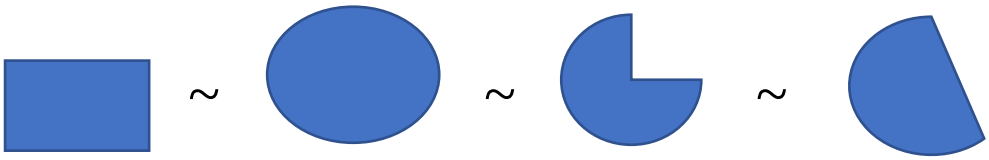
Ορισμός 1.2.3 : Μία συνάρτηση $f : X \to Y$ καλείται ***επιμορφισμός*** εάν για κάθε στοιχείο $y \in Y$ υπάρχει ακριβώς ένα στοιχείο $x \in X$ τέτοιο ώστε $f(x) = y$.

Ορισμός 1.2.4 : Μια συνάρτηση $f : X \to Y$ είναι ***συνεχής***, αν για κάθε ανοιχτό υποσύνολο A του Y, το $f^{-1}(A)$ είναι ανοιχτό υποσύνολο του X.

Ορισμός 1.2.5 : Έστω X, Y τοπολογικοί χώροι. Μια συνάρτηση $f : X \to Y$ καλείται ***ομοιομορφισμός*** εάν η f είναι 1 – 1, επί, συνεχής και έχει συνεχή αντίστροφη συνάρτηση.

Ορισμός 1.2.6 : Δύο τοπολογικοί χώροι X, Y είναι ***ομοιομορφικοί ή τοπολογικά ισοδύναμοι***, αν υπάρχει ομοιομορφισμός f με $f : X \to Y$.

Οι ομοιομορφισμοί αποτελούν σχέση ισοδυναμίας πάνω στους τοπολογικούς χώρους. Οπότε δύο τοπολογικά ισοδύναμοι χώροι X και Y θα έχουν τον ίδιο τοπολογικό τύπο που συμβολίζεται με $X \approx Y$. Ο τοπολογικός τύπος είναι ο ορθότερος τρόπος ταξινόμησης σε μία τοπολογία. Παρακάτω παρουσιάζονται τοπολογικά ισοδύναμοι χώροι (σχήμα 1.2.1). Να παρατηρήσουμε ότι μπορούμε να μεταβούμε από τον έναν χώρο στον άλλον, απλά μέσω συνθλίψεων ή εκτάσεων.



Σχήμα 1.2.1 Τοπολογικά ισοδύναμοι χώροι

Η έννοια του ομοιομορφισμού είναι καθοριστικής σημασίας για την Τοπολογική Ανάλυση Δεδομένων, καθώς κύριο στόχο της μεθόδου αποτελεί η ορθή προσέγγιση του χώρου δειγματοληψίας με κάποιον ομοιομορφικό του.

1.3 Τοπολογικά αναλλοίωτες

Οι ομοιομορφισμοί παρέχουν τον καλύτερο τρόπο ταξινόμησης δύο τοπολογικών χώρων. Στο πρόβλημα του ομοιομορφισμού αναζητείται λύση για το αν δύο τοπολογικοί χώροι X, Y είναι ομοιομορφικοί. Για να δειχθεί ότι δύο τοπολογικοί χώροι δεν είναι ισοδύναμοι, αρκεί να βρεθούν τοπολογικές αναλλοίωτες που διακρίνουν τον έναν χώρο από τον άλλον, αφού εάν οι τοπολογικοί χώροι έχουν διαφορετικές αναλλοίωτες δεν μπορούν να είναι ομοιομορφικοί.

Ορισμός 1.3.1 : Μια ***τοπολογική αναλλοίωτη*** είναι μια συνάρτηση f που αποδίδει το ίδιο αντικείμενο σε ομοιομορφικούς χώρους, δηλαδή :

$$X \approx Y \Rightarrow f(X) = f(Y)$$

Δηλαδή, οι τοπολογικές αναλλοίωτες είναι ιδιότητες που δεν αλλάζουν κάτω από συνεχείς μετασχηματισμούς.

Ωστόσο, η συνάρτηση αυτή θεωρείται χρήσιμη μόνο μέσω αντιθετοαντιστροφής :

$$f(X) \neq f(Y) \Rightarrow X \neq Y$$

και χρησιμεύει για να αποδειχθεί ότι δύο τοπολογικοί χώροι δεν είναι τοπολογικά ισοδύναμοι.

Επομένως, μια αναλλοίωτη (trivial invariant) που αποδίδει το ίδιο αντικείμενο σε όλους τους χώρους δεν βοηθάει στην εξαγωγή κάποιου χρήσιμου συμπεράσματος. Από την άλλη πλευρά, μια αναλλοίωτη (complete invariant) που αποδίδει διαφορετικά αντικείμενα σε δύο τοπολογικούς χώρους, λύνει το πρόβλημα ομοιομορφισμού, καθώς οδηγεί στο συμπέρασμα ότι οι δύο χώροι δεν είναι τοπολογικά ισοδύναμοι. Οι περισσότερες αναλλοίωτες (incomplete invariant) εμπίπτουν στο φάσμα μεταξύ αυτών των δύο άκρων [39].

Οι τοπολογικά αναλλοίωτες χωρίζονται σε τρεις βασικές κατηγορίες :

1. Η ***Αναλλοίωτη των Συντεταγμένων*** : λαμβάνει υπόψιν μόνο την ιδιότητα του σχήματος (πέρα από τις συντεταγμένες που έχει) και την περιστροφή που μπορεί να υποστεί.

Σχήμα 1.3.1 Αναλλοίωτη των Συντεταγμένων

2. Η ***Αναλλοίωτη της Παραμόρφωσης*** : δεν αλλάζει την ταυτότητα του σχήματος εάν το τεντώσουμε ή το συνθλίψουμε, με την προϋπόθεση ότι δεν το σκίσουμε ή δεν το ενώσουμε. Για παράδειγμα, ένας κύκλος είναι πανομοιότυπος με έναν πατημένο κύκλο που μοιάζει με έλλειψη.

Σχήμα 1.3.2 Αναλλοίωτη της Παραμόρφωσης

3. Η ***Συμπιεσμένη Αναπαράσταση*** : επικεντρώνεται στη συνδεσιμότητα και τη συνέχεια. Αυτή η ιδιότητα επιτρέπει στην Τοπολογία να αναγνωρίσει συσχετιζόμενα μοτίβα δεδομένων. Για παράδειγμα, ένας κύκλος έχει άπειρες κορυφές ενώ το εξάγωνο είναι όμοιο με τον κύκλο, παρόλο που έχει έξι κόμβους και έξι κορυφές.

Σχήμα 1.3.3 Συμπιεσμένη Αναπαράσταση

1.4 Ομοτοπία

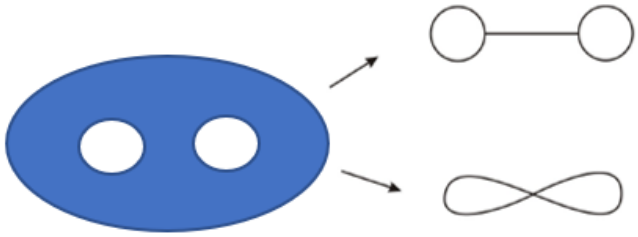
Μία ασθενέστερη σχέση μεταξύ τοπολογικών χώρων η οποία, όμως, διατηρεί πολλές τοπολογικά αναλλοίωτες είναι η ομοτοπική ισοδυναμία. Κατά κάποιον τρόπο, δύο τοπολογικοί χώροι είναι ομοτοπικοί, εάν ο ένας μπορεί να μετασχηματιστεί στον άλλον μέσω μιας συνεχούς παραμόρφωσης.

Ορισμός 1.4.1 : Αν X και Y είναι δύο τοπολογικοί χώροι και $f_0, f_1 : X \to Y$ είναι συνεχείς συναρτήσεις, τότε η f_0 είναι ***ομοτοπική*** με την f_1, αν υπάρχει συνεχής συνάρτηση $F : X \times [0,1] \to Y$, τέτοια ώστε :

$$F(x,0) = f_0(x) \text{ και}$$
$$F(x,1) = f_1(x) \quad \text{για κάθε } x \in X$$

Μια τέτοια συνάρτηση F λέγεται Ομοτοπία.

Ορισμός 1.4.2 : Δύο τοπολογικοί χώροι είναι ***ομοτοπικά ισοδύναμοι***, αν υπάρχουν συνεχείς συναρτήσεις $f : X \to Y$ και $g : X \to Y$, τέτοιες ώστε $g \circ f = id_X$ και $f \circ g = id_Y$. Στην περίπτωση αυτή θα γράφουμε $X \approx Y$.



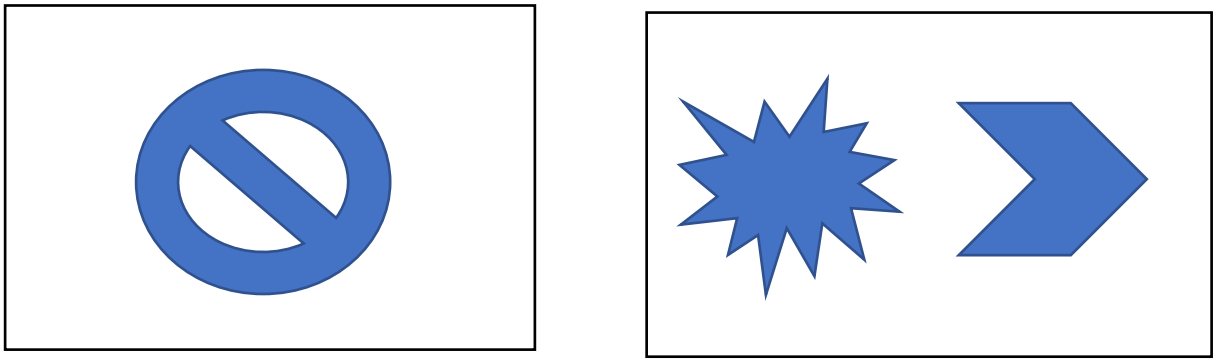
Σχήμα 1.4.1 Ομοτοπικά ισοδύναμοι χώροι

Ορισμός 1.4.3 : Ένας τοπολογικός χώρος είναι ***συσταλτός***, αν είναι ομοτοπικά ισοδύναμος με ένα σημείο.

Διαισθητικά, ένας τοπολογικός χώρος είναι συσταλτός, αν μπορεί να συρρικνωθεί με συνεχή τρόπο σε ένα σημείο. Για παράδειγμα, αποδεικνύεται ότι ο δίσκος είναι ομοτοπικά ισοδύναμος με ένα σημείο (σχήμα 1.4.2). Να παρατηρήσουμε ότι στο δίσκο δεν υπάρχουν καθόλου «τρύπες».

Σχήμα 1.4.2 Ο δίσκος είναι συσταλτός χώρος

Από την άλλη, οι παρακάτω δύο χώροι δεν είναι συσταλτοί (σχήμα 1.4.3). Ο πρώτος χώρος έχει δύο «τρύπες», ενώ ο δεύτερος αποτελείται από δύο υποχώρους, οι οποίοι δεν μπορούν να συρρικνωθούν σε ένα σημείο. Βέβαια, καθένας από αυτούς μπορεί να συρρικνωθεί σε σημείο. σημείο.



Σχήμα 1.4.3 Μη συσταλτοί χώροι

1.5 Πολλαπλότητες

Η πολλαπλότητα αποτελεί τοπολογικό χώρο που τοπικά, κοντά σε κάθε σημείο, μοιάζει με Ευκλείδειο χώρο. Πιο συγκεκριμένα, κάθε σημείο μιας n – διάστατης πολλαπλότητας έχει μία γειτονιά που είναι ομοιομορφική με τον Ευκλείδειο χώρο διάστασης n. Παραδείγματα μονοδιάστατης πολλαπλότητας αποτελούν η γραμμή και ο κύκλος, ενώ δισδιάστατης το επίπεδο, η σφαίρα και η σαμπρέλα. Δηλαδή, οτιδήποτε δημιουργείται χωρίς να τέμνει τον εαυτό του.

Η διάσταση της πολλαπλότητας είναι μια τοπολογική ιδιότητα, που σημαίνει ότι οποιαδήποτε πολλαπλότητα που είναι ομοιομορφική με μια n – πολλαπλότητα έχει, επίσης, διάσταση n. Όπως προκύπτει από την αναλλοίωτη μια n – πολλαπλότητα δεν μπορεί να είναι ομοιομορφική με μια m – πολλαπλότητα για $n \neq m$. Μια πολλαπλότητα διάστασης ένα συχνά καλείται καμπύλη, ενώ μια πολλαπλότητα διάστασης δύο ονομάζεται επιφάνεια. Πολλαπλότητες μεγαλύτερης διάστασης συνήθως ονομάζονται n – πολλαπλότητες.

Ορισμός 1.5.1 : Ένα ***διάγραμμα*** στο $p \in X$ είναι μια συνάρτηση $\varphi : U \to \mathbb{R}^d$, όπου $U \subseteq X$ είναι ανοιχτό σύνολο που περιέχει το p και φ είναι ένας ομοιομορφισμός σε ένα ανοιχτό υποσύνολο του $\mathbb{R}^d$. Η διάσταση του διαγράμματος φ είναι d.

Ορισμός 1.5.2 : Ένας τοπολογικός χώρος X είναι ***Hausdorff***, εάν για $κάθε\ x, y \in X$, με $x \neq y$, υπάρχουν γειτονιές U, V του x, y αντίστοιχα, τέτοιες ώστε $U \cap V = \emptyset$.

Ορισμός 1.5.3 : Ένας τοπολογικός χώρος X είναι ***διαχωρίσιμος***, εάν έχει μετρήσιμη βάση γειτονιών.

Ορισμός 1.5.4 : Ένας διαχωρίσιμος χώρος Hausdorff X καλείται (τοπολογική) ***d – πολλαπλότητα*** εάν υπάρχει διάγραμμα d – διάστασης σε κάθε σημείο $x \in X$, δηλαδή εάν το $x \in X$ έχει γειτονιά ομοιομορφική με τον $\mathbb{R}^d$. Επιπλέον, καλείται ***d – πολλαπλότητα με σύνορο*** εάν το x ∈ X έχει γειτονιά ομοιομορφική με τον $\mathbb{R}^d$ ή τον ευκλείδειο υποχώρο $H^d = \{x \in \mathbb{R}^d / x_1 \geq 0\}$. Το σύνορο του X είναι το σύνολο των σημείων με γειτονιές ομοιομορφικές με τον H^d. Η πολλαπλότητα είναι διάστασης d.

Θεώρημα 1.5.1 : Το ***σύνορο μιας d – πολλαπλότητας*** με σύνορο είναι μία $(d-1)$ – πολλαπλότητα χωρίς σύνορο.

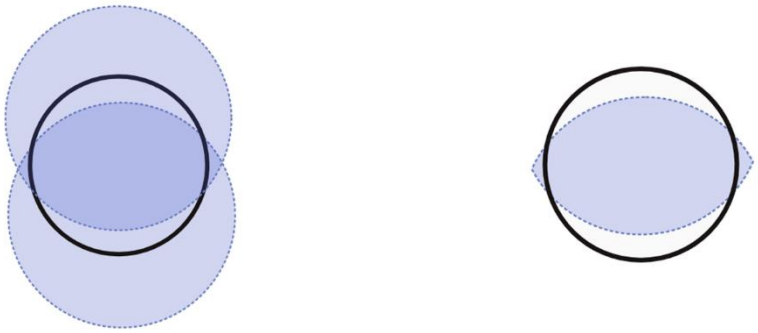
Το μεγαλύτερο μέρος ενασχόλησης με πολλαπλότητες είναι στους Ευκλείδειους χώρους. Συνεπώς, πάντα σκεφτόμαστε μία πολλαπλότητα ενσωματωμένη σε έναν Ευκλείδειο χώρο. Είναι σημαντικό, όμως, να θυμόμαστε ότι μια πολλαπλότητα υπάρχει ανεξάρτητα από οποιαδήποτε ενσωμάτωση. Για παράδειγμα, μια σφαίρα δεν πρέπει να 'καθίσει' στον $\mathbb{R}^3$ για να είναι σφαίρα [38].

1.6 Καλύμματα

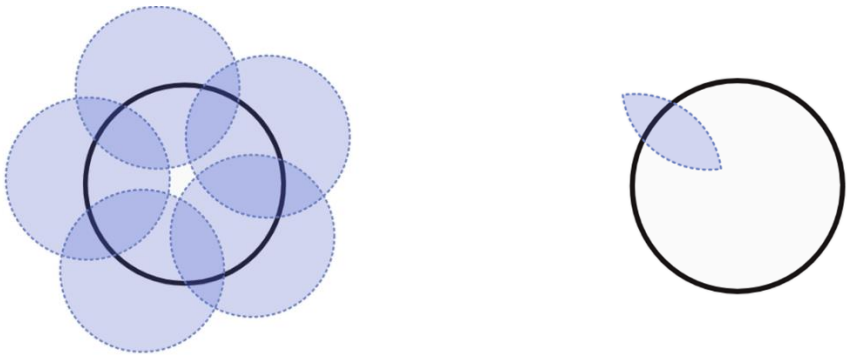
Μια βασική ιδέα στην Τοπολογική Ανάλυση Δεδομένων είναι η προσέγγιση του χώρου δειγματοληψίας X τοπικά, χρησιμοποιώντας κομμάτια του χώρου ενσωμάτωσης Y. Η προσέγγιση αυτή επιτυγχάνεται με τη χρήση καλυμμάτων.

Ορισμός 1.6.1 : ***Κάλυμμα*** ενός συνόλου $A \subseteq X$ είναι μια οικογένεια συνόλων $\{U_i \mid i \in I\}$, τέτοια ώστε $A \subseteq \bigcup_{i \in I} U_i$. Ένα ***ανοιχτό κάλυμμα*** είναι ένα κάλυμμα που αποτελείται από ανοικτά σύνολα. ***Υποκάλυμμα*** του $\{U_i \mid i \in I\}$ είναι ένα κάλυμμα $\{C_k \mid k \in K\}$, όπου $K \subseteq I$. Ένα σύνολο $A \subseteq X$ είναι ***συμπαγές*** εάν κάθε ανοιχτό κάλυμμα του A έχει ένα πεπερασμένο υποκάλυμμα. Ένα κάλυμμα U είναι ***καλό***, εάν όλα τα Ui είναι συσταλτά όπως επίσης και οι μη κενές πεπερασμένες τομές τους.

Για την καλύτερη κατανόηση της έννοιας του καλύμματος και του καλού καλύμματος παρουσιάζονται τα παρακάτω παραδείγματα. Στην πρώτη περίπτωση, ο κύκλος καλύπτεται μόνο με δύο ανοιχτά σύνολα (όχι καλό κάλυμμα), ενώ στη δεύτερη με περισσότερα (καλό κάλυμμα). Να παρατηρήσουμε ότι στην πρώτη περίπτωση, η τομή των δύο ανοιχτών συνόλων με τον κύκλο οδηγεί σε μη συσταλτό χώρο (σχήμα 1.6.1). Αντιθέτως, στη δεύτερη η τομή δύο οποιονδήποτε ανοιχτών συνόλων με τον κύκλο δίνει ένα πολύ μικρό μέρος του κύκλου, που είναι συσταλτός χώρος (σχήμα 1.6.2). Επομένως, το κάλυμμα είναι καλό και μας δίνει περισσότερη πληροφορία για το πως είναι το σχήμα τοπικά.



Σχήμα 1.6.1 Όχι καλό κάλυμμα



Σχήμα 1.6.2 Καλό κάλυμμα

Ορισμός 1.6.2 : ***Νεύρο*** ενός καλύμματος U είναι το σύνολο N για το οποίο ισχύουν τα εξής:

i. $\emptyset \in N$ και

ii. Εάν $\bigcap_{j\in J} U_j \neq \emptyset$, για $J \subseteq I$, τότε J$\in N$

Η ένωση των συνόλων σε ένα ανοιχτό κάλυμμα είναι η συνήθης προσέγγιση που χρησιμοποιείται για τον προσδιορισμό του χώρου δειγματοληψίας X, καθώς όλη η ενδιαφέρουσα τοπολογία του καλύμματος εκτίθεται στο νεύρο.

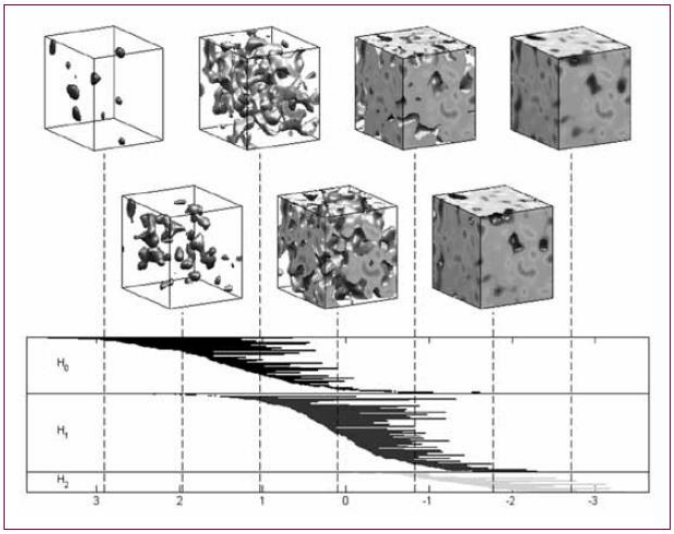
Σύμφωνα με το Nerve Lemma του Leray, το νεύρο ενός καλού καλύμματος είναι ομοτοπικά ισοδύναμο με το κάλυμμα, δηλαδή την ένωση των συνόλων στο κάλυμμα [4], [31]. Αυτό το λήμμα αποτελεί βάση αρκετών μεθόδων για τη συνδυαστική αναπαράσταση ενός συνόλου σημείων [39].

Κεφάλαιο 2

Persistent Homology

Στο παρόν κεφάλαιο παρουσιάζεται το θεωρητικό πλαίσιο μιας από τις πιο δημοφιλείς μεθόδους της Τοπολογικής Ανάλυσης Δεδομένων. Συγκεκριμένα, πρόκειται για την Εμμένουσα Ομολογία (Persistent Homology), προσέγγιση κατά την οποία ‘κωδικοποιείται’ η εξέλιξη της τοπολογίας του σχήματος που προκύπτει από τα δεδομένα.

Η βασική προσέγγιση της εμμένουσας ομολογίας είναι η αντικατάσταση του συνόλου των δεδομένων με οικογένειες από simplicial complexes, έτσι ώστε τα δεδομένα να μετατραπούν σε τοπολογικά αντικείμενα, ακολουθεί ο υπολογισμός της ομολογίας του χώρου που προκύπτει και τέλος καταγράφεται η αλλαγή της ομολογίας κατά τη διάρκεια ενός φιλτραρίσματος. Το κύριο χαρακτηριστικό της ομολογίας είναι η δυνατότητα της να περιγραφεί μέσω γραφημάτων που καλούνται barcodes. Οι μπάρες σε ένα barcode αναπαριστούν τη διάρκεια ζωής των χαρακτηριστικών εντός ενός φιλτραρίσματος [25]. Η χρήση των barcodes δίνει τη δυνατότητα μέτρησης της σημασίας των ομολογιακών χαρακτηριστικών μέσω της διάρκειας ζωής τους. Με αυτόν τον τρόπο μπορούμε να δούμε σε ένα point cloud τα στατιστικώς σημαντικά στοιχεία, όπως τους βρόγχους ή τα κενά [24].



Σχήμα 2.1 Barcodes [1]

2.1 Τοπολογική Ανάλυση Δεδομένων

Η Τοπολογική Ανάλυση Δεδομένων αναφέρεται σ' ένα σύνολο μεθόδων εύρεσης σχετικών ποιοτικών και ποσοτικών τοπολογικών δομών απευθείας από τα δεδομένα [6]. Δηλαδή, τα δεδομένα αντιμετωπίζονται ως τοπολογικά αντικείμενα υψηλών διαστάσεων, τα οποία φέρουν κάποια γεωμετρική δομή που αντανακλά σημαντικές ιδιότητες των χώρων από τους οποίους έχουν προέλθει.

Ορισμός 2.1.1 : Δοθέντος ενός συνόλου (θορυβωδών) σημείων $S \subseteq Y$, τα οποία προήλθαν από έναν άγνωστο χώρο X, η ***Τοπολογική Ανάλυση Δεδομένων*** ανακτά την Τοπολογία του X, υποθέτοντας ότι ο X και ο Y είναι τοπολογικοί χώροι.

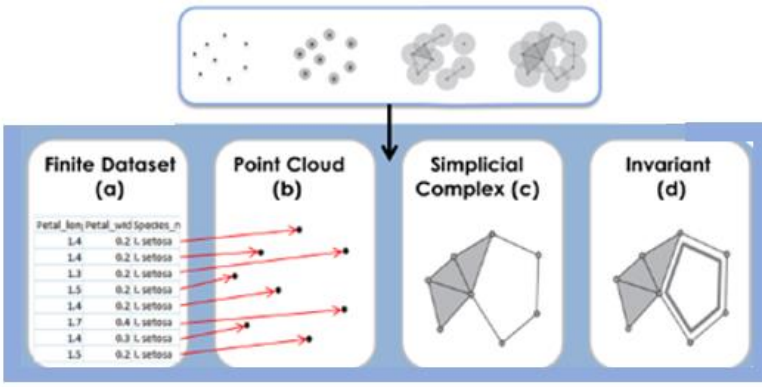
Έστω ένα σύνολο σημείων S (PCD*), το οποίο είναι ενσωματωμένο σε κάποιον d – διάστατο χώρο Y. Υποθέτουμε ότι αυτά τα δεδομένα προέρχονται από έναν άγνωστο k – διάστατο υποχώρο $X \subseteq Y$, όπου $k \leq d$. Τα δεδομένα είναι ένα πεπερασμένο σύνολο (θορυβώδους) δείγματος. Κατά τη διάρκεια της δειγματοληψίας, τόσο η γεωμετρία όσο και η τοπολογία του σχήματος χάθηκαν. Επομένως, σκοπός είναι η ανάκτηση πληροφοριών σχετικά με το X μέσα από το σύνολο των σημείων S. Λαμβάνοντας υπόψη ότι ιδιότητες του χώρου ενσωμάτωσης Y είναι εξωτερικές, ενώ οι ιδιότητες του άγνωστου χώρου X εσωτερικές, κατά την ανάλυση των δεδομένων γίνεται προσπάθεια ανάκτησης των εσωτερικών πληροφοριών, με χρήση των εξωτερικών [39]. Δηλαδή, για ένα Y αρκετά «κοντά» στο X, αναμένεται ότι η τοπολογία του X μπορεί να βρεθεί από το Y.

Βέβαια, τα point clouds δεν φέρουν καμία μη τετριμμένη τοπολογική ή γεωμετρική ιδιότητα. Επομένως, είναι απαραίτητο να θεωρηθούν γεωμετρικές δομές στα point clouds με σκοπό να ανακτηθούν πληροφορίες σχετικά με τα σχήματα που προσεγγίζουν. Στην ιδανικότερη των περιπτώσεων, είναι δυνατό να προσδιοριστεί μία τέτοια προσέγγιση Y του αντικειμένου X που είναι ομοιομορφική με το X [26].

* Ένα point cloud data είναι απλά ένα πεπερασμένο σύνολο σημείων X εφοδιασμένο με μία μετρική d.

Ο πιο απλός τρόπος για την προσέγγιση του άγνωστου χώρου *X* είναι η μετατροπή της συλλογής δεδομένων ενός μετρικού χώρου σε ένα αντικείμενο, χρησιμοποιώντας το point cloud ως τις κορυφές ενός σχήματος του οποίου οι πλευρές καθορίζονται από την εγγύτητα (κορυφές που απέχουν κάποια συγκεκριμένη απόσταση) [19]. Σ' αυτήν την περίπτωση, η τοπολογία του *X* μπορεί να περιγραφεί από τις τοπολογικά αναλλοίωτες του Υ.

Συνήθως, αναζητούνται αναλλοίωτες που είναι αποτελεσματικά υπολογίσιμες, ώστε να γίνεται ευκολότερα η αποθήκευση τους σε υπολογιστές. Προφανώς, όσο πιο ισχυρή μία αναλλοίωτη, τόσο πιο δύσκολο είναι να την υπολογίσουμε [39]. Συνεπώς, η τοπολογία ενός χώρου είναι δυνατό να περιγραφεί από τοπολογικά αναλλοίωτες του, όπως για παράδειγμα η ομολογία, ο χαρακτηριστικός αριθμός του Euler και τα Betti number.



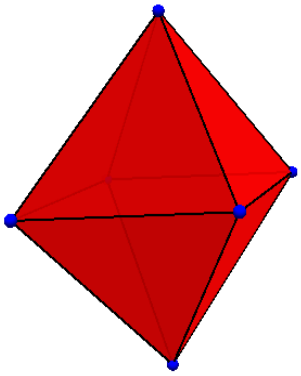
Σχήμα 2.1.1 Μετατροπή ενός συνόλου δεδομένων σε τοπολογικό αντικείμενο και εύρεση των αναλλοίωτών του [23]

Συνοπτικά, πρώτο βήμα στην Τοπολογική Ανάλυση Δεδομένων αποτελεί η αντικατάσταση του συνόλου των δεδομένων με οικογένειες από simplicial complexes, έτσι ώστε τα δεδομένα να μετατραπούν σε τοπολογικά αντικείμενα. Στη συνέχεια, μελετώνται τα simplicial complexes με χρήση αλγεβρικής τοπολογίας και πιο συγκεκριμένα, με τον υπολογισμό τοπολογικά αναλλοίωτων. Το παρόν κεφάλαιο εστιάζει στον υπολογισμό της εμμένουσας ομολογίας. Η διαδικασία ολοκληρώνεται με αναπαράσταση της εμμένουσας ομολογίας μέσω των barcodes. Τα barcodes αποτυπώνουν την εμμένουσα ομολογία ενός συνόλου δεδομένων με τη μορφή μιας παραμετρικοποιημένης εκδοχής των Betti numbers.

2.2 Πλεγματικά Σύμπλοκα – Simplicial Complexes

Τα Πλεγματικά Σύμπλοκα (Simpicial Complexes) είναι ιδιαίτερα δημοφιλή στην Τοπολογική Ανάλυση Δεδομένων εξαιτίας της απλής δομής τους. Μέσω αυτών, γίνεται ευκολότερη η μελέτη μεγάλων συνόλων δεδομένων και ο προσδιορισμός των χώρων στους οποίους ζουν.

Βασική ιδέα πίσω από τα simplicial complexes είναι η επιλογή ενός πλήθους σημείων, πλευρών, τριγώνων και υψηλότερων διαστάσεων τριγώνων και η ένωση τους κατά μήκος των πλευρών τους για την κατασκευή αντικειμένων. Για παράδειγμα, μια δισκόμπαλα, η οποία είναι φτιαγμένη από γυάλινα τρίγωνα, μπορεί να θεωρηθεί ως ένα simplicial complex (όπως ένα κούφιο οχτάεδρο) που προσεγγίζει μία σφαίρα.



Σχήμα 2.2.1
Κούφιο οχτάεδρο

Παρακάτω, παρουσιάζεται η έννοια του Πλεγματικού Συμπλόκου με χρήση εννοιών της Γεωμετρίας και της Τοπολογίας.

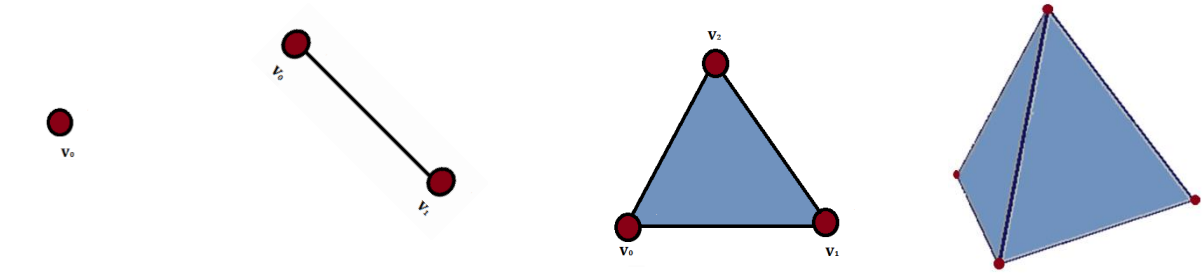
Ορισμός 2.2.1 : Έστω $S = \{v_0, v_1, \dots, v_k\} \subseteq \mathbb{R}^d$. Ένας ***γραμμικός συνδυασμός*** είναι ένα άθροισμα $x = \sum_{i=0}^{k} \lambda_i v_i$, όπου $\lambda_i \in \mathbb{R}$. Ένας ***αφινικός συνδυασμός*** είναι ένας γραμμικός συνδυασμός με $\sum_{i=0}^{k} \lambda_i = 1$. Ένας ***κυρτός συνδυασμός*** είναι ένας αφινικός συνδυασμός με $\lambda_i \geq 0$ για κάθε i. Το σύνολο όλων των κυρτών συνδυασμών ονομάζεται ***κυρτό περίβλημα***.

Για παράδειγμα, το κυρτό περίβλημα ενός συνόλου σημείων P του Ευκλείδειου χώρου $\mathbb{R}^n$ είναι το ελάχιστο κυρτό σύνολο που περιέχει όλα τα σημεία του P.

Ορισμός 2.2.2 : Ένα σύνολο S καλείται ***γραμμικά ανεξάρτητο*** εάν κανένα σημείο του S δεν είναι αφινικός συνδυασμός των άλλων σημείων του S.

Ορισμός 2.2.3 : Ένα ***k – simplex*** γεωμετρικά είναι ένα κυρτό περίβλημα από $k + 1$ γραμμικά ανεξάρτητα σημεία του $\mathbb{R}^d$, όπου $d \geq k$. Ως τοπολογικός χώρος ονομάζεται πολύεδρο και συμβολίζεται $|K|$.

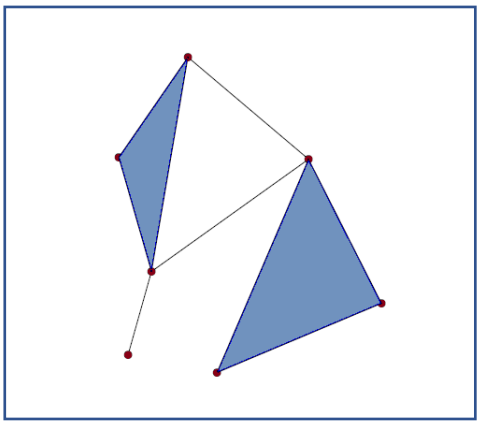
Με λίγα λόγια, ένα k – simplex φτιάχνεται από $k + 1$ σημεία, τα οποία ενώνονται μεταξύ τους με ευθύγραμμα τμήματα. Για $k = 0, 1, 2, 3$ έχουμε σύμπλοκα χαμηλότερων διαστάσεων, τα οποία είναι οι κορυφές, οι έδρες, τα τρίγωνα και τα τετράεδρα.



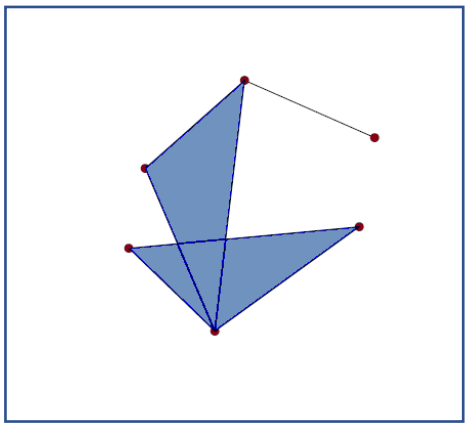
Σχήμα 2.2.2 Σύμπλοκα διαστάσεων 0,1,2 και 3

Ορισμός 2.2.4 : Ένα ***γεωμετρικό πλεγματικό σύμπλοκο (geometric simplicial complex)*** K είναι ένα σύνολο από simplices τέτοιο ώστε :

i. Όλες οι όψεις ενός simplex του K είναι, επίσης, στο K.
ii. Η τομή οποιονδήποτε δύο simplices $\sigma, \tau \in K$ είναι είτε όψη των σ, τ είτε το κενό σύνολο.



Σχήμα 2.2.3 Γεωμετρικό πλεγματικό σύμπλοκο [26]



Σχήμα 2.2.4 Σύνολο από simplices [26]

Συνεπώς, ένα πλεγματικό σύμπλοκο K μπορεί να ενσωματωθεί στον Ευκλείδειο χώρο ως ένωση simplices τα οποία τέμνονται μόνο σε κοινές όψεις. Αυτή η ένωση είναι ο τοπολογικός χώρος $|K| = \bigcup_{\sigma \in K} K_\sigma$ του K [39].

Ο ορισμός ενός πλεγματικού συμπλόκου χρησιμοποιεί τη γεωμετρία με ένα θεμελιώδη τρόπο, καθώς η αντίληψή μας για τη διάσταση των simplices πηγάζει από την ικανότητα μας να αντιλαμβανόμαστε ένα n – simplex γεωμετρικά ως έναν n – διάστατο υποχώρο του $\mathbb{R}^d$, όπου $d \geq n$.

Ωστόσο, σε περιπτώσεις που δεν μας ενδιαφέρει η γεωμετρική εικόνα του συμπλόκου, δηλαδή οι αποστάσεις ή οι θέσεις των σημείων, αλλά ενδιαφερόμαστε για τη συνδυαστική φύση της προκείμενης αναπαράστασης, γίνεται χρήση του παρακάτω ορισμού :

Ορισμός 2.2.5 : Ένα ***αφηρημένο πλεγματικό σύμπλοκο (abstract simplicial complex)*** Κ στο σύνολο των κορυφών $\{v_0, v_1, \ldots, v_k\}$ είναι μια συλλογή υποσυνόλων του $\{v_0, v_1, .., v_k\}$, που είναι κλειστή ως προς τη σχέση εγκλεισμού, δηλαδή αν $\sigma \in K$ και $\tau \subset \sigma$, τότε $\tau \in K$. Τα στοιχεία του K ονομάζονται ***όψεις (faces)*** του K.

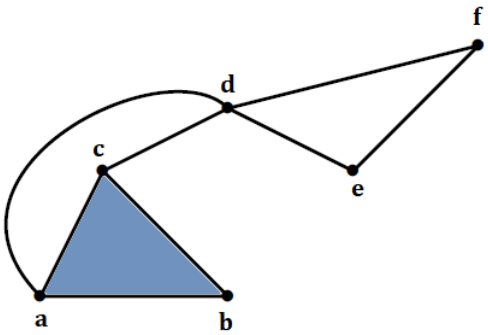
Ορισμός 2.2.6 : Ένα στοιχείο $\sigma \in K$ λέγεται ***μεγιστική όψη*** **(facet)**, αν δεν υπάρχει όψη $\tau \in K$, τέτοια ώστε $\sigma \subset \tau$. Το σύνολο των μεγιστικών όψεων του K συμβολίζεται με $\mathcal{F}(K)$. Είναι φανερό ότι το πλεγματικό σύμπλοκο Κ καθορίζεται πλήρως από το σύνολο $\mathcal{F}(K)$.

Ορισμός 2.2.7 : Αν η μοναδική μεγιστική όψη του K είναι το σύνολο των κορυφών $\{v_0, v_1, .., v_k\}$, τότε το K ονομάζεται ***πλήρες πλεγματικό σύμπλοκο*** **(*simplex*)**. Τα πλήρη πλεγματικά σύμπλοκα είναι συσταλτοί τοπολογικοί χώροι.

Πιο απλά, μεγιστική όψη είναι ένα υποσύνολο του K, για το οποίο δεν γίνεται να βρεθεί μεγαλύτερη όψη. Με βάση τη μεγιστική όψη ορίζεται και η διάσταση ενός πλεγματικού συμπλόκου, ως η διάσταση της μεγαλύτερης μεγιστικής του όψης.

Ορισμός 2.2.8 : Η ***διάσταση μιας όψης*** σ είναι $|\sigma| - 1$. Η ***διάσταση του K*** ορίζεται να είναι $dim(K) = max\{dim(\sigma) \mid \sigma \in \mathcal{F}(K)\}$. Το κενό σύνολο είναι πάντα μία όψη του K, εκτός αν το K είναι το κενό πλεγματικό σύμπλοκο, το οποίο δεν έχει καμία όψη. Το $\emptyset$ είναι η μοναδική όψη του K διάστασης -1. Το κενό πλεγματικό σύμπλοκο ορίζεται να έχει διάσταση $-\infty$.

Παράδειγμα 2.2.1 : Έστω το πλεγματικό σύμπλοκο K στο σύνολο των κορυφών $\{a, b, c, d, e, f\}$ με μεγιστικές όψεις $\mathcal{F}_1 = \{a, b, c\}$, $\mathcal{F}_2 = \{a, d\}$, $\mathcal{F}_3$ = {c,d}, $\mathcal{F}_4$ = {d,e}, $\mathcal{F}_5 = \{d, f\}$, $\mathcal{F}_6$ = $\{e, f\}$. Ισχύει ότι $dim(\mathcal{F}_1) = 2$ και $dim(\mathcal{F}_j) = 1$, για κάθε j ∈ {2, … ,6} . Άρα η διάσταση του K είναι 2. Η γεωμετρική εικόνα του K είναι η εξής :

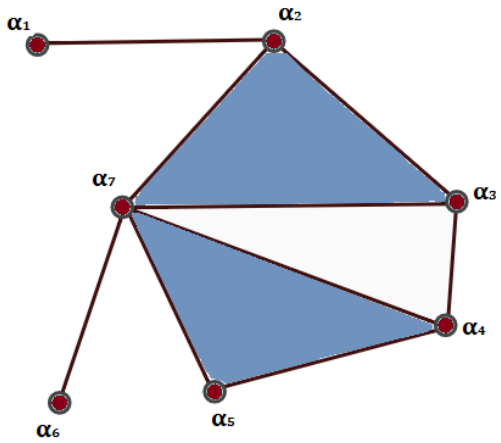


Σχήμα 2.2.5 Πλεγματικό σύμπλοκο διάστασης 2

Ορισμός 2.2.9 : Τα υποσύνολα του K που είναι επίσης πλεγματικά σύμπλοκα λέγονται ***υποσύμπλοκα* (*subcomplexes*)** του K.

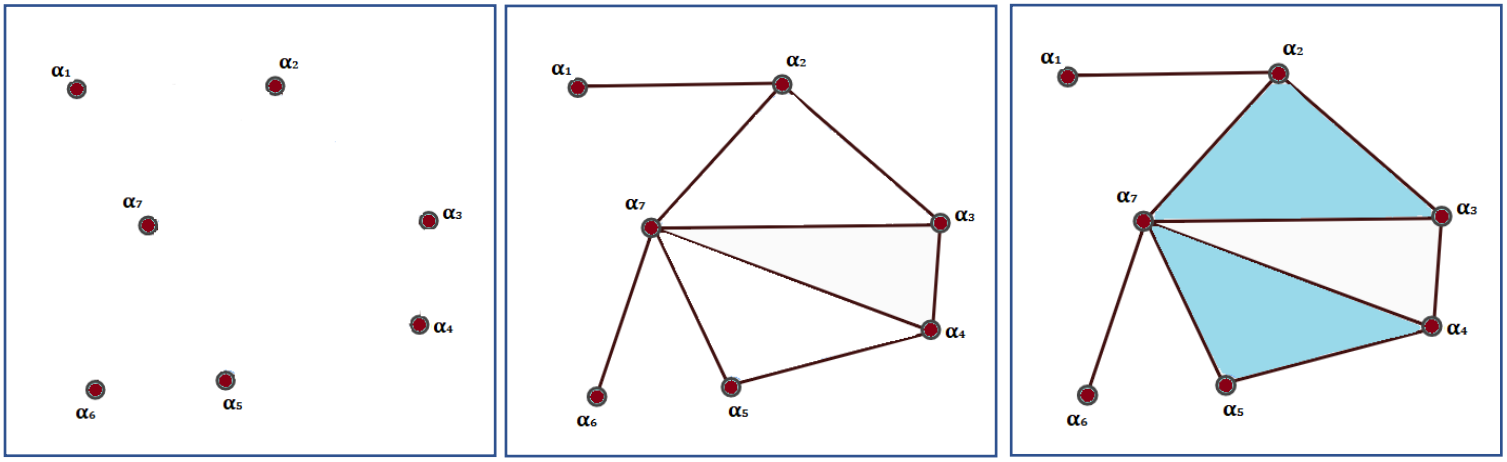
Ένα σημαντικό υποσύμπλοκο είναι ο ***n – skeleton*** που αποτελείται από τα πλεγματικά υποσύμπλοκα του K που έχουν διάσταση μικρότερη ή ίση του n. Να παρατηρήσουμε ότι ο *1 – skeleton* ενός simplicial complex είναι πάντα ένα γράφημα.

Παράδειγμα 2.2.2 : Έστω το πλεγματικό σύμπλοκο Σ στο σύνολο των κορυφών $\{a_i, i \in \{1,2,3,4,5,6,7\}\}$, με μεγιστικές όψεις $\mathcal{F}_1 = \{a_i, i \in \{1,2\}\}$, $\mathcal{F}_2 = \{a_i, i \in \{3,4\}\}$, $\mathcal{F}_3 = \{a_i, i \in \{6,7\}\}$, $\mathcal{F}_4 = \{a_i, i \in \{2,3,7\}\}$, $\mathcal{F}_5 = \{a_i, i \in \{4,5,7\}\}$, του οποίου η γεωμετρική εικόνα φαίνεται στο παρακάτω γράφημα :



Σχήμα 2.2.6 Πλεγματικό σύμπλοκο

Ένας n – σκελετός μιας τέτοιας δομής είναι η συλλογή που περιέχει όλα τα αντικείμενα τα οποία είναι το πολύ διάστασης n. Συνεπώς, ο 0 – σκελετός είναι το σύνολο των σημείων, ο 1 – σκελετός είναι το σύνολο των πλευρών και των σημείων, ο 2 – σκελετός είναι το σύνολο των τριγώνων, των πλευρών και των σημείων. Τα αντίστοιχα γραφήματα του συγκεκριμένου παραδείγματος παρουσιάζονται παρακάτω :



Σχήμα 2.2.7 0,1 και 2 σκελετός

Ορισμός 2.2.10 : Έστω ένα πεπερασμένο σύνολο από πλεγματικά σύμπλοκα $\{K_i\}_{i\in A}$, όπου A το σύνολο των κορυφών. Το ***νεύρο*** του $\{K_i\}_{i\in A}$ είναι το πλεγματικό σύμπλοκο N, το οποίο έχει όψεις τα σύνολα :

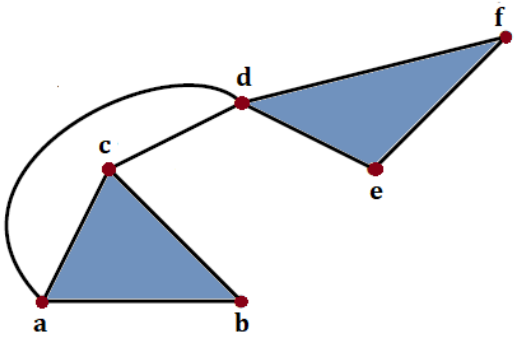
$$\{\sigma \subset A \,/\, \bigcap_{i\in\sigma} K_i \neq \emptyset\}$$

Σημειώνουμε ότι κάθε στοιχείο του συνόλου των κορυφών A ανήκει στο νεύρο N.

Θεώρημα 2.2.1 : Έστω K ένα πλεγματικό σύμπλοκο και $\{\mathrm{K_i}\}_{\mathrm{i}\in\mathrm{A}}$ ένα πεπερασμένο σύνολο από υποσύμπλοκα, τέτοιο ώστε $\cup_{i\in A}\ \mathrm{K_i} = K$. Έστω N νεύρο του συνόλου $\{\mathrm{K_i}\}_{\mathrm{i}\in\mathrm{A}}$. Αν για κάθε όψη $\sigma \subset A$ του N ισχύει ότι η μη κενή τομή $\cap_{i\in\sigma}\ \mathrm{K_i}$ είναι συσταλτός τοπολογικός χώρος, τότε το πλεγματικό σύμπλοκο K και το νεύρο N είναι ομοτοπικά ισοδύναμοι τοπολογικοί χώροι.

Απόδειξη. Βλέπε [29], Nerve Theorem 36.11

Παράδειγμα 2.2.3 : Έστω το πλεγματικό σύμπλοκο K στο σύνολο των κορυφών $\{a,b,c,d,e,f\}$ με μεγιστικές όψεις $\mathcal{F}_1 = \{a,b,c\}$, $\mathcal{F}_2 = \{d,e,f\}$, $\mathcal{F}_3$ = {a,d}, $\mathcal{F}_4$ = {c,d}, όπως φαίνεται στο παρακάτω σχήμα :

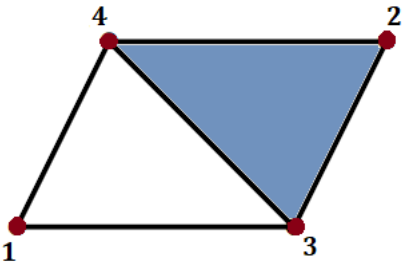


Σχήμα 2.2.8 Πλεγματικό σύμπλοκο

Θεωρούμε τα υποσύμπλοκα :

$$K_1 = <\{a,b,c\}>,\ K_2 = <\{d,e,f\}>,\ K_3 = <\{\mathrm{a,d}\}>,\ K_4 = <\{\mathrm{c,d}\}>$$

Θέτουμε Σ = $\{K_i\ /\ \mathrm{i}\in\{1,2,3,4\}\}$. Το νεύρο N (σχήμα 2.2.9) του Σ είναι το πλεγματικό σύμπλοκο στο σύνολο των κορυφών $\{1,2,3,4\}$, που παράγεται από τις όψεις $\{2,3,4\},\{1,3\},\{1,4\}$.



Σχήμα 2.2.9 Νεύρο του πλεγματικού συμπλόκου

Είναι εύκολο να ελέγξουμε ότι $\bigcup_{i\in\{1,2,3,4\}} K_i = K$ και ότι για κάθε όψη $\sigma \subset \{1,2,3,4\}$ του N η τομή $\bigcap_{\iota\in\sigma} K_i$ είναι σημείο και άρα συσταλτός τοπολογικός χώρος. Επομένως, σύμφωνα με το προηγούμενο θεώρημα, $N \cong K$.

2.3 Γεωμετρικά Σύμπλοκα – Geometric Complexes

Έστω ένα πεπερασμένο σύνολο σημείων (PCD) $S \subseteq Y$. Σ' αυτή την ενότητα, κατασκευάζουμε simplicial complexes K που προσεγγίζουν τον άγνωστο χώρο X, από τον οποίο προήλθε το δείγμα S. Για το σκοπό αυτό χρησιμοποιήθηκαν οι πηγές : [6], [7], [24], [26] και [39].

Δοθέντος ενός συνόλου (θορυβωδών) σημείων $S \subseteq Y$, όπου (Y, d_y) μετρικός χώρος, τα οποία προήλθαν από έναν άγνωστο χώρο X, σκοπός είναι να κατασκευάσουμε ένα simplicial complex, στο σύνολο των κορυφών του S, του οποίου ο ομολογιακός ή ο ομοτοπικός τύπος είναι όμοιος με του X [7]. Δηλαδή, είναι επιθυμητό να κατασκευαστούν simplicial complexes που υπολογίζουν την ομολογία του άγνωστου χώρου X, ή τουλάχιστον έχουν μια ισχυρή σχέση με αυτόν. Υπάρχουν ποικίλα simplicial complexes που μπορούν να κατασκευαστούν, βασικές κατασκευές αποτελούν το **Cech Complex, το Vietoris – Rips και το Witness**.

Αρχικά, θυμίζουμε ότι :

Ορισμός 2.3.1 : Ένας ***μετρικός χώρος*** είναι ένα ζεύγος (X, d_x), όπου X είναι ένα σύνολο εφοδιασμένο με μία μετρική $d_x : X \times X \to \mathbb{R}$ που ικανοποιεί τις ακόλουθες ιδιότητες για κάθε $x, y, z \in X$:

i. $d_x(x,y) \geq 0$ *και* $d_x(x,y) = 0 \Leftrightarrow x = y$
ii. $d_x(x,y) = d_x(y,x)$

iii. $d_x(x,z) \leq d_x(x,y) + d_x(y,z)$

Ορισμός 2.3.2 : Έστω (X, d_x) μετρικός χώρος, $x \in X$ και $\varepsilon > 0$. Θέτουμε $B_\varepsilon(x) = \{y \in X \mid d_x(x,y) < \varepsilon\}$. Το σύνολο $B_\varepsilon(x)$ ονομάζεται ***ανοιχτή μπάλα*** με κέντρο x και ακτίνα ε ή **ε – περιοχή του x**.

Čech Complex

Η πιο απλή μέθοδος κατασκευής ενός simplicial complex, το οποίο παρέχει πληροφορία για την ομολογία του άγνωστου χώρου X, είναι μέσω του νεύρου του καλύμματος $U_\varepsilon(x) = \{B_\varepsilon(x) \,/\, x \in S\}$. Αφού οι μπάλες είναι κυρτές και τα κυρτά σύνολα είναι συσταλτά, το κάλυμμα είναι καλό και το νεύρο του αιχμαλωτίζει την τοπολογία του χώρου.

Ορισμός 2.3.3 : Έστω ο μετρικός χώρος (X, d_x). Για $\varepsilon > 0$ ορίζουμε το simplicial complex $\check{\boldsymbol{C}}\boldsymbol{ech}_\varepsilon(\boldsymbol{X}) = \boldsymbol{C}_\varepsilon(\boldsymbol{X})$ στο σύνολο των κορυφών του X από την ακόλουθη συνθήκη :

$$[x_0, x_1, .., x_k] \in \check{C}ech_\varepsilon(X) \Leftrightarrow \bigcap_{i=0}^{k} B_\varepsilon(x_i) \neq \emptyset,$$

όπου $\boldsymbol{C_0}(\boldsymbol{X}) = \emptyset$ και $\boldsymbol{C}_\infty(\boldsymbol{X})$ είναι ένα $(|X| - 1)$ – simplex. Κάθε σημείο $\bar{x}$ του συνόλου της τομής $\bigcap_\iota B_\varepsilon(x_i)$ καλείται ***ε – κέντρο*** του simplex $[x_0, x_1, .., x_k]$. Αναφέρουμε ότι υπάρχει ένας φυσικός εγκλεισμός $C_a(X) \subseteq C_b(X)$, για $a \leq b$. Έτσι, το simplicial complex $\check{C}\text{ech}\,(X)$ σε συνδυασμό με τις σχέσεις εγκλεισμού ορίζουν ένα filtered complex στο X, το $\check{C}\text{ech}$ complex.

Στην πράξη, το $\check{C}\text{ech}$ Complex είναι δύσκολο να υπολογιστεί λόγω της υπολογιστικής πολυπλοκότητάς του. Επειδή ένα n – simplex έχει 2^{n+1} όψεις, το complex μπορεί να γίνει τεράστιο σε υψηλές κλίμακες. Δηλαδή, το $\check{C}\text{ech}$ complex μπορεί να έχει πολύ μεγαλύτερη διάσταση από το χώρο ενσωμάτωσης [39].

Μια ιδέα για την αντιμετώπιση αυτού του προβλήματος είναι να κατασκευάσουμε ένα simplicial complex το οποίο μπορεί να ανακτηθεί μόνο από τις πληροφορίες που παρέχουν οι έδρες. Αυτό υποδηλώνει την ακόλουθη παραλλαγή της κατασκευής $\check{C}\text{ech}$, που αναφέρεται ως ***Vietoris – Rips*** complex. Το ***Vietoris – Rips*** complex (ή απλά Rips complex) είναι ιδιαίτερα χρήσιμο, είναι εύκολο να υπολογιστεί και έχει καλές ιδιότητες προσέγγισης.

Vietoris – Rips Complex

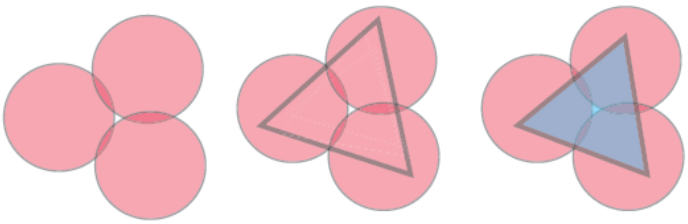
Ορισμός 2.3.4 : Έστω ο μετρικός χώρος (X, d_x). Για $\varepsilon > 0$ ορίζουμε το simplicial complex $\boldsymbol{Rips_\varepsilon(X)}$ στο σύνολο των κορυφών του X από την ακόλουθη συνθήκη :

$$[x_0, x_1, .., x_k] \in Rips_\varepsilon(X) \Leftrightarrow d_x(x_i, x_j) \leq \varepsilon, \text{ για κάθε } i, j$$

Αναφέρουμε ότι υπάρχει ένας φυσικός εγκλεισμός $Rips_a(X) \subseteq Rips_b(X)$, για $a \leq b$. Έτσι, το simplicial complex $Rips_\varepsilon(X)$ σε συνδυασμό με τις σχέσεις εγκλεισμού ορίζουν ένα filtered complex στο X, το Vietoris – Rips complex.

Παράδειγμα 2.3.1

Θεωρούμε ένα σύνολο κορυφών. Το Σχήμα 2.3.1 παρουσιάζει τις διαφορές μεταξύ των δύο προαναφερθέντων complexes.



Σχήμα 2.3.1 Κάλυμμα, Čech και Vietoris – Rips complex [6]

Το πρώτο γράφημα του Σχήματος 2.3.1 δείχνει το κάλυμμα, το δεύτερο το Čech complex και το τρίτο παριστάνει το Vietoris – Rips complex. Να παρατηρηθεί ότι η παράμετρος για τον υπολογισμό του Čech complex αντιστοιχεί στην ακτίνα των μπαλών, ενώ η παράμετρος για το Vietoris – Rips complex είναι η απόσταση μεταξύ των κέντρων των μπαλών.

Να σημειωθεί εδώ ότι στο παραπάνω παράδειγμα χρησιμοποιήθηκε ένας μικρός αριθμός σημείων. Από την άλλη πλευρά, εάν προσπαθήσουμε να δημιουργήσουμε ένα Vietoris – Rips complex, για παράδειγμα, για έναν κύκλο (ο οποίος έχει άπειρα σημεία) θα χρειαστούμε άπειρα 0 – simplices. Πιο ρεαλιστικά, εάν είχαμε ένα μεγάλο αριθμό σημείων δεδομένων, θα

χρειαζόταν να αποθηκεύσουμε ένα μεγάλο αριθμό 0 – simplices, πράγμα που είναι υπολογιστικά ακριβό, ειδικά αν θέλουμε μόνο την ουσία των δεδομένων. Από τη στιγμή που τα VR complexes μπορεί να είναι ογκώδη, προσπαθούμε να τα προσεγγίσουμε με μικρότερο αριθμό κορυφών.

Witness Complex

Έστω δύο σύνολα L, W ('$landmarks$' και '$witnesses$') και η συνάρτηση $\Lambda : L \times W \to \mathbb{R}$. Για κάθε πεπερασμένο υποσύνολο $\sigma \subseteq L$, και για κάθε $w \in W$, $\varepsilon > 0$, το w είναι ένας ε – witness για το simplex σ αν και μόνο αν :

$$\Lambda(l, w) \leq \Lambda(l', w) + \varepsilon, \text{ για κάθε } l \in \sigma \text{ και } l' \in L \backslash \sigma$$

Δοθέντος των συνόλων L, W και της συνάρτησης Λ, ορίζουμε για κάθε $\varepsilon > 0$ το simplicial complex $Wit_{\varepsilon}(L, W)$ από την ακόλουθη συνθήκη :

$$\sigma \in Wit_{\varepsilon}(L, W) \Leftrightarrow \forall \tau \subseteq \sigma, \exists\, w \in W \text{ τέτοιο ώστε το w να είναι } \varepsilon - \text{witness για το } \tau$$

Υπάρχει ένας φυσικός εγκλεισμός $Wit_a(L, W) \subseteq Wit_b(L, W)$, για $a \leq b$, αφού ένας a – witness είναι προφανώς ένας b – witness. Έτσι, το simplicial complex $Wit_{\varepsilon}(L, W)$ σε συνδυασμό με τις σχέσεις εγκλεισμού ορίζουν ένα filtered complex στο σύνολο των κορυφών του L, το Witness Complex Filtration.

Έστω ο μετρικός χώρος (X, d_x). Για κάθε πεπερασμένο υποσύνολο $\sigma \subseteq L$, και για κάθε $w \in W$ και $\varepsilon > 0$, το w είναι ένας *strong* – witness για το simplex σ αν και μόνο αν :

$$d_x(l, w) \leq m_w + \varepsilon, \text{ για κάθε } l \in \sigma$$

όπου m_w είναι η απόσταση του σημείου w από το σύνολο L, ενώ το w είναι ένας $\varepsilon - weak$ *witness* για το simplex σ αν και μόνο αν :

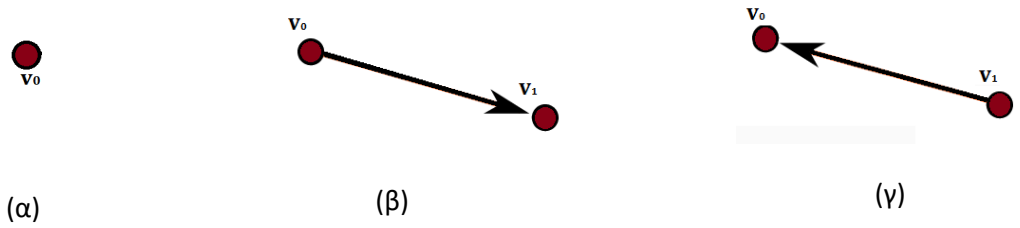
$$d_x(l_i, w) \leq d_x(l, w) + \varepsilon, \text{ για κάθε } l_i \in \sigma \text{ και } l \in L \backslash \sigma$$

Τα αντίστοιχα complexes καλούνται strong witness complex και weak witness complex.

2.4 Αλυσιδωτά Σύμπλοκα – Chain Complexes

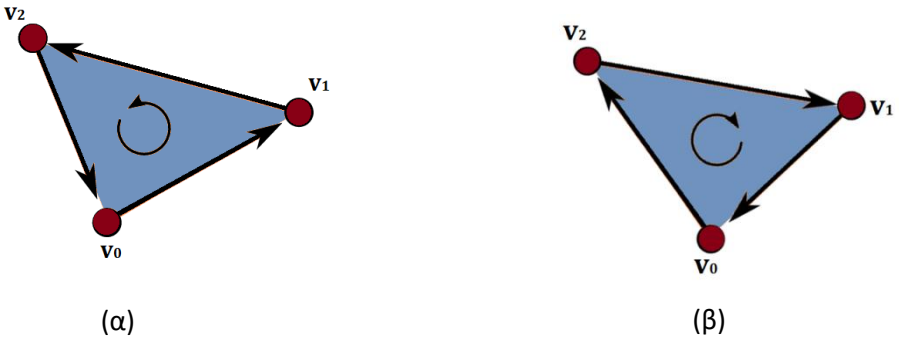
Ορισμός 2.4.1 : ***Προσανατολισμός*** ενός k – simplex σ, όπου $\sigma = \{v_0, v_1, \dots, v_k\}$, είναι ***μία διατεταγμένη κλάση ισοδυναμίας*** στο σύνολο των κορυφών του σ, όπου $(v_0, v_1, \dots, v_k)$ και $(v_{\tau(0)}, v_{\tau(1)}, \dots, v_{\tau(k)})$ είναι ισοδύναμα εάν το πλήθος των μεταθέσεων τ είναι άρτιος αριθμός. Ένα ***προσανατολισμένο simplex*** δηλώνεται με $[\sigma]$.

Παρακάτω παρουσιάζονται τα προσανατολισμένα k – simplices για $k = 0, 1, 2$. Για παράδειγμα, το προσανατολισμένο 1 – simplex $[v_0, v_1]$ περιγράφεται από το σχήμα 2.4.1 (β), ενώ στο σχήμα 2.4.1 (γ) φαίνεται το προσανατολισμένο 1 – simplex $[v_1, v_0]$.



Σχήμα 2.4.1 Προσανατολισμένα simplices

Αντίστοιχα, τα προσανατολισμένα 2 – simplices $[v_0, v_1, v_2]$ και $[v_2, v_1, v_0]$ παρουσιάζονται στο σχήμα 2.4.2 (α) και (β) αντίστοιχα.



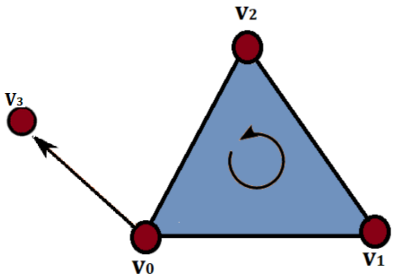
Σχήμα 2.4.2 Προσανατολισμένα 2 – simplices

Παρατηρούμε ότι επειδή το 0 – simplex (σχήμα 2.4.1 (α)) έχει μόνο ένα στοιχείο στο σύνολο του, έχει και μόνο έναν προσανατολισμό. Γενικά, κάθε n – simplex, εκτός από τα 0 – simplices, έχει δύο διαφορετικούς προσανατολισμούς.

Ορισμός 2.4.2 : Η ***k – αλυσιδωτή (αβελιανή) ομάδα (k – chain (abelian) group)*** $\mathrm{C_k}(K)$ του K είναι η ελεύθερη αβελιανή ομάδα πάνω στο σύνολο των προσανατολισμένων k – simplices, όπου $[\sigma] = -[\tau]$ εάν $\sigma = \tau$ και τα σ και τ έχουν διαφορετικό προσανατολισμό. Ένα στοιχείο $c \in \mathrm{C_k}$ είναι μία ***k – αλυσίδα (k – chain)***, $c = \sum_i n_i [\sigma_i]$, όπου $\sigma_i \in K$ και $n_i \in \mathbb{Z}$.

Οι αλυσιδωτές ομάδες ορίζονται για κάθε ακέραιο k, αλλά για παράδειγμα στον $\mathbb{R}^3$ μη μηδενικές είναι για ακεραίους $k \in [0,3]$.

Παράδειγμα 2.4.1 : Έστω το πλεγματικό σύμπλοκο K που καθορίζεται πλήρως από τις μεγιστικές του όψεις $\mathcal{F}_1 = [\mathrm{v_0}, \mathrm{v_1}, \mathrm{v_2}]$ και $\mathcal{F}_2 = [\mathrm{v_0}, \mathrm{v_3}]$. Η γεωμετρική εικόνα του K είναι η εξής :



Σχήμα 2.4.3 Προσανατολισμένο πλεγματικό σύμπλοκο

Τότε, σύμφωνα με τον παραπάνω ορισμό, για $n_{i,i\in[1,4]} \in \mathbb{Z}$ θα είναι :

$$\mathrm{C_0}(K) = n_1[\mathrm{v_0}] + n_2[\mathrm{v_1}] + n_3[\mathrm{v_2}] + n_4[\mathrm{v_3}]$$

$$\mathrm{C_1}(K) = n_1[\mathrm{v_0},\mathrm{v_1}] + n_2[\mathrm{v_1},\mathrm{v_2}] + n_3[\mathrm{v_2},\mathrm{v_0}] + n_4[\mathrm{v_0},\mathrm{v_3}]$$

$$\mathrm{C_2}(K) = n_1[\mathrm{v_0},\mathrm{v_1},\mathrm{v_2}]$$

Επειδή δεν υπάρχει $\sigma \in K$ με $dim(\sigma) \geq 3$, έχουμε ότι $\mathrm{C_n}(K) = 0$, για κάθε $n \geq 3$.

Για οποιοδήποτε στοιχείο της αλυσίδας $C_0(K)$, πρέπει να καθορίσουμε μόνο τα $(n_1, n_2, n_3, n_4) \in \mathbb{Z}^4$. Το $\mathbb{Z}^4$ μας πληροφορεί ότι πρέπει να καθορίσουμε ένα διατεταγμένο σύνολο που περιέχει 4 στοιχεία, το καθένα από το $\mathbb{Z}$. Εφόσον οποιοδήποτε στοιχείο αυτού του τύπου καθορίζει μια 0 – αλυσίδα, προκύπτει ότι :

$C_0(K) \cong \mathbb{Z}^4$

Με παρόμοιο σκεπτικό προκύπτουν οι ισομορφίες :

$C_1(K) \cong \mathbb{Z}^4$

$C_2(K) \cong \mathbb{Z}$

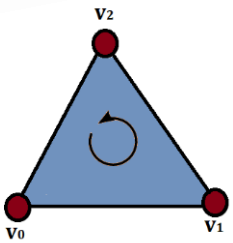
Ορισμός 2.4.3 : Ο ***συνοριακός τελεστής (boundary operator)*** $\partial_k : C_k \to C_{k-1}$ είναι ένας ομοιομορφισμός που ορίζεται γραμμικά πάνω σε μια αλυσίδα c μέσω της δράσης του σε οποιοδήποτε simplex $\sigma = [v_0, v_1, \dots, v_k] \in c$:

$$\partial_k(\sigma) = \sum_i (-1)^i [v_0, v_1, \dots, \hat{v}_i, \dots, v_k]$$

όπου το $\hat{v}_i$ δηλώνει ότι το v_i διαγράφεται από την ακολουθία και $\partial_0 \equiv 0$.

Με λίγα λόγια, το σύνορο $\partial_k(\sigma)$ ενός k – simplex σ είναι η συλλογή όλων των $(k-1)$ – διάστασης όψεων του, οι οποίες αποτελούν μία $(k-1)$ – αλυσίδα. Το σύνορο μιας k – αλυσίδας είναι το άθροισμα των συνόρων των simplices του, $\partial_k(C) = \sum_{\sigma \in c} \partial_k(\sigma)$.

Για παράδειγμα, το σύνορο ενός προσανατολισμένου k – simplex, για k = 0, 1, 2 είναι:

- 0 – simplex $= [v_0]$  με $\partial_0(v_0) = 0$
- 1 – simplex $= [v_0, v_1] = e$  με $\partial_1(e) = v_1 - v_0$
- 2 –simplex $= [v_0, v_1, v_2] = f$  με $\partial_2(f) = [v_1, v_2] - [v_0, v_2] + [v_0, v_1]$

Σε περιπτώσεις που δεν μας ενδιαφέρει ο προσανατολισμός, μπορούμε να αυξήσουμε την ταχύτητα υπολογισμού των συνόρων, έχοντας μερική απώλεια πληροφορίας, εργαζόμενοι στον $\mathbb{Z}_2$. Το σύνορο ενός μη προσανατολισμένου k – simplex, για $k = 0, 1, 2$ είναι :

- 0 – simplex $= \{v_0\}$ με $\partial_0(v_0) = 0$
- 1 – simplex $= \{v_0, v_1\}$ με $\partial_1(e) = v_1 - v_0 = v_1 + v_0$
- 2 – simplex $= \{v_0, v_1, v_2\}$ με $\partial_2(f) = \{v_1, v_2\} + \{v_0, v_2\} + \{v_0, v_1\}$

Ορισμός 2.4.4 : Ο συνοριακός τελεστής συνδέει τις αλυσιδωτές ομάδες σε ένα ***αλυσιδωτό σύμπλοκο (chain complex)*** C_* :

$$\dots \to C_{k+1} \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} C_{k-1} \to \dots$$

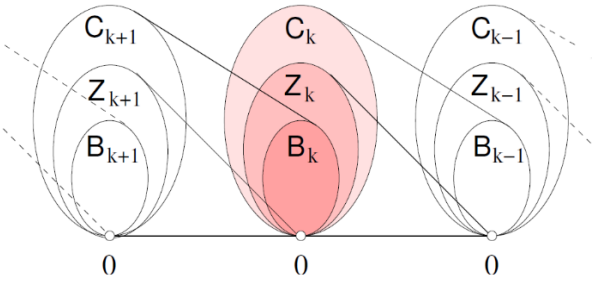
Ο πυρήνας του ∂_k είναι η συλλογή όλων των k – αλυσίδων με κενό σύνορο και η εικόνα του ∂_k είναι η συλλογή όλων των $(k-1)$ – αλυσίδων που είναι σύνορα k – αλυσίδων. Δηλαδή :

$$ker\partial_k = \{c \in C_k \,/\, \partial_k(c) = 0\}$$

$$im\partial_k = \{d \in C_{k-1} \,/\, \exists c \in C_k : d = \partial_k(c)\}$$

Τα στοιχεία του $ker\partial_k$ ονομάζονται ***k – κύκλοι (k – cycles)*** και τα στοιχεία του $im\partial_{k+1}$ ονομάζονται ***k – σύνορα (k – boundaries)***. Επίσης, με τη βοήθεια του συνοριακού τελεστή ορίζονται οι υποομάδες του C_k, τα ***k – οστά κυκλίματα (k – cycle group)*** $Z_k = ker\,\partial_k$ και τα ***k – οστά σύνορα (k – boundary group)*** $B_k = im\,\partial_{k+1}$.

Μία σημαντική ιδιότητα του συνοριακού τελεστή είναι ότι το σύνορο κάθε συνόρου είναι πάντα το κενό σύνολο, δηλαδή $\partial_k \partial_{k+1} = 0$. Από την προηγούμενη ιδιότητα σε συνδυασμό με τους ορισμούς προκύπτει ότι : $B_k \subseteq Z_k \subseteq C_k$ (σχήμα 2.4.4). Επιπλέον, αξίζει να αναφερθεί ότι κάθε 0 – αλυσίδα είναι και 0 – κύκλος ($Z_0 = C_0$), καθώς το σύνορο κάθε κορυφής ταυτίζεται με το κενό σύνολο.



Σχήμα 2.4.4 Αλυσίδες, κύκλοι, σύνορα καθώς και η εικόνα τους μέσω του συνοριακού τελεστή [37]

2.5 Simplicial Homology

Η ομολογία είναι μία μαθηματική έννοια που χρησιμοποιείται για την περιγραφή της συνδεσιμότητας των αντικειμένων, δηλαδή παρέχει πληροφορίες σχετικά με τον αριθμό των συνιστωσών, τις τρύπες, τις σήραγγες και τα κενά σε έναν τοπολογικό χώρο. Η αναλλοίωτη αυτή λειτουργεί εκχωρώντας μια ομάδα σ' έναν τοπολογικό χώρο, αντί να υπολογίζει απλά έναν ακέραιο αριθμό. Η ομολογία είναι αρκετά δημοφιλής στην τοπολογική ανάλυση δεδομένων, καθώς μπορεί να υπολογιστεί αποτελεσματικά.

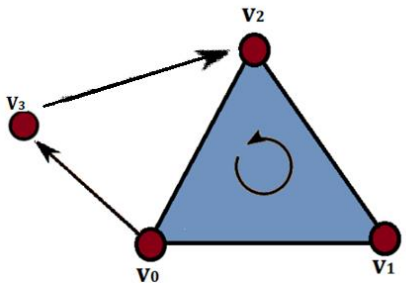
Ορισμός 2.5.1 : Έστω C_* ένα Αλυσιδωτό Σύμπλοκο. Η ***k – οστή Ομάδα Ομολογίας (k – th Homology Group)*** H_k είναι το k – οστό Cycle Group που παράγεται από το k – οστό Boundary Group, δηλαδή $H_k = Z_k / B_k$. Τα στοιχεία της H_k ονομάζονται ***Κλάσεις Ομολογίας***. Ουδέτερο στοιχείο είναι το B_k. Δύο κύκλοι της ίδιας κλάσης ομολογίας ονομάζονται ***Ομόλογοι (Homologous)***.

Διαισθητικά, δύο p – κύκλοι είναι ομόλογοι, εάν υπάρχει συνεχής μετασχηματισμός από τον έναν στον άλλον. Επιπλέον, οι p – κύκλοι μπορούν να ομαδοποιηθούν σε κλάσεις ισοδύναμων p – κύκλων, όπου κάθε κλάση μπορεί να αναπαρασταθεί από ένα μοναδικό αντιπρόσωπο (γεννήτορα).

Η Ομολογία αποτελεί αναλλοίωτη για δύο Ομοτοπικά Ισοδύναμους χώρους. Δηλαδή, για τα πλεγματικά σύμπλοκα K_1 , K_2 ισχύει ότι :

$$|K_1| \approx |K_2| \Rightarrow H_k(K_1) = H_k(K_2), για\ κάθε\ k \geq 0$$

Παράδειγμα 2.5.1 : Έστω το πλεγματικό σύμπλοκο K που καθορίζεται πλήρως από τις μεγιστικές όψεις $\mathcal{F}_1 = [v_0, v_1, v_2]$, $\mathcal{F}_2 = [v_0, v_3]$, $\mathcal{F}_3 = [v_3, v_2]$, όπως φαίνεται και στο παρακάτω σχήμα :



Σχήμα 2.5.1 Πλεγματικό σύμπλοκο

Τότε, οι k – αλυσιδωτές ομάδες, για $n_{i,i\in[1,5]} \in \mathbb{Z}$, είναι :

$\mathrm{C}_0(K) = n_1[\mathrm{v}_0] + n_2[\mathrm{v}_1] + n_3[\mathrm{v}_2] + n_4[\mathrm{v}_3] \cong \mathbb{Z}^4$

$\mathrm{C}_1(K) = n_1[\mathrm{v}_0,\mathrm{v}_1] + n_2[\mathrm{v}_1,\mathrm{v}_2] + n_3[\mathrm{v}_2,\mathrm{v}_0] + n_4[\mathrm{v}_0,\mathrm{v}_3] + n_5[\mathrm{v}_3,\mathrm{v}_2] \cong \mathbb{Z}^5$

$\mathrm{C}_2(K) = n_1[\mathrm{v}_0,\mathrm{v}_1,\mathrm{v}_2] \cong \mathbb{Z}$

Επειδή δεν υπάρχει $\sigma \in K$ με $dim(\sigma) \geq 3$, έχουμε ότι $\mathrm{C_n}(K) = 0$, για κάθε $n \geq 3$. Σύμφωνα με τα παραπάνω προκύπτει το αλυσιδωτό σύμπλοκο $C_2 \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0$.

Στη συνέχεια, υπολογίζουμε τις ομάδες ομολογίας H_0 και H_1 . Θυμίζουμε ότι : $H_k = Z_k / B_k$, όπου :

$$\mathrm{Z_k} = \ker\partial_k = \{\mathrm{c} \in C_k / \ \partial_k(\mathrm{c}) = \emptyset\}$$

$$\mathrm{B_k} = \mathrm{im}\partial_{k+1} = \{\mathrm{d} \in C_k / \ \exists c \in C_{k+1} : \mathrm{d} = \partial_{k+1}(\mathrm{c})\}$$

Για την ομάδα ομολογίας H_0 έχουμε :

$\mathrm{Z}_0 = \ker\partial_0 = \{\mathrm{c} \in C_0 / \ \partial_0(\mathrm{c}) = 0\} = \mathrm{C}_0(K) \cong \mathbb{Z}^4$

$\mathrm{B}_0 = \mathrm{im}\partial_1 = \{\mathrm{d} \in \mathrm{C}_0 / \ \exists c \in C_1 : \mathrm{d} = \partial_1(\mathrm{c})\} = \mathrm{C}_0(K) \cong \mathbb{Z}^4$

Άρα, $H_0 = Z_0 / B_0 \cong \mathbb{Z}^4 / \mathbb{Z}^4 = \{0\}$

Για την ομάδα ομολογίας H_1 έχουμε :

$Z_1 = \ker\partial_1 = \{c \in C_1 / \partial_1(c) = 0\}$

$\partial_1(n_1[v_0,v_1] + n_2[v_1,v_2] + n_3[v_2,v_0] + n_4[v_0,v_3] + n_5[v_3,v_2]) =$

$= n_1\partial_1[v_0,v_1] + n_2\partial_1[v_1,v_2] + n_3\partial_1[v_2,v_0] + n_4\partial_1[v_0,v_3] + n_5\partial_1[v_3,v_2] =$

$= n_1(v_1 - v_0) + n_2(v_2 - v_1) + n_3(v_0 - v_2) + n_4(v_3 - v_0) + n_5(v_2 - v_3) =$

$= (-n_1 + n_3 - n_4)v_0 + (n_1 - n_2)v_1 + (n_2 - n_3 + n_5)v_2 + (n_4 - n_5)v_3,\ n_{i,i\in[1,5]} \in \mathbb{Z}$

$n_1[v_0,v_1] + n_2[v_1,v_2] + n_3[v_2,v_0] + n_4[v_0,v_3] + n_5[v_3,v_2]$ είναι κύκλος ⇔
$(-n_1 + n_3 - n_4)v_0 + (n_1 - n_2)v_1 + (n_2 - n_3 + n_5)v_2 + (n_4 - n_5)v_3 = 0$

Οπότε,
$Z_1 = \ker\partial_1 =$
$=\{c \in C_1 / \partial_1(c) = 0\} = \{n_1[v_0,v_1] + n_1[v_1,v_2] + (n_1 + n_4)[v_2,v_0] + n_4[v_0,v_3] + n_4[v_3,v_2]\} =$
$=\{n_1([v_0,v_1] + [v_1,v_2] + [v_2,v_0]) + n_4([v_2,v_0] + [v_0,v_3] + [v_3,v_2])\} \cong \mathbb{Z} \oplus \mathbb{Z}$

$B_1 = \mathrm{im}\partial_2 = \{d \in C_1 / \exists c \in C_2 : d = \partial_2(c)\} = \{d = \partial_2(n_1[v_0,v_1,v_2])\} \cong \mathbb{Z}$

Άρα, $H_1 = Z_1 / B_1 \cong \mathbb{Z} \oplus \mathbb{Z} / \mathbb{Z} \cong \mathbb{Z}$. ■

Επομένως, η ομολογία προκύπτει από το σύνολο των κύκλων παραβλέποντας αυτούς που προέρχονται από σύνορα.

Για να γίνει ευκολότερα αντιληπτή η δομή μιας αλγεβρικής αναλλοίωτης, όπως η ομολογία ενός simplicial complex, μπορούμε να χρησιμοποιήσουμε την ακόλουθη προσέγγιση:

Αντιστοίχιση → Ταξινόμηση → Παραμετρικοποίηση

Στο πρώτο βήμα, αναγνωρίζουμε την αλγεβρική δομή. Στο δεύτερο βήμα, αποκτούμε μία πλήρη ταξινόμηση της δομής, έως έναν ισομορφισμό. Στο τρίτο βήμα, παραμετρικοποιούμε την ταξινόμηση. Πιο αναλυτικά :

α) Αντιστοίχιση : Η k – οστή ομολογία H_k ενός simplicial complex είναι ομάδα, ή ισοδύναμα ένας $\mathbb{Z}$ – module, όπου $\mathbb{Z}$ ο δακτύλιος των συντελεστών. Βέβαια, αυτή η οπτική επιτρέπει τη χρήση διαφορετικών δακτυλίων συντελεστών, συμπεριλαμβανομένων και σωμάτων. Οπότε, αντ' αυτού, μπορούν να κατασκευαστούν modules πάνω από άλλους δακτυλίους R. Όσο το complex K είναι πεπερασμένο, η H_k γίνεται ένας πεπερασμένα παραγόμενος R – module.

β) Ταξινόμηση : Υποθέτουμε ότι R είναι ένας δακτύλιος κυρίων ιδεωδών (ΔΚΙ), όπως το $\mathbb{Z}$. Οποιοδήποτε πεπερασμένα παραγόμενο R – module αποσυντίθεται με μοναδικό τρόπο στη μορφή :

$$(\bigoplus_{i=1}^{\beta_n} R) \oplus (\bigoplus_{j=1}^{m} R/t_j R)$$

για ακεραίους $\beta_n \geq 0$ και μη μηδενικά και μη μοναδιαία στοιχεία $t_j \in R$, τέτοια ώστε t_j / t_{j+1}.

γ) Παραμετρικοποίηση : Το αριστερό άμεσο ευθύ άθροισμα είναι ο ελεύθερος υπο – module και χαρακτηρίζεται από τον αριθμό Betti $\beta_n = rank H_n$. Το δεξί ευθύ άθροισμα είναι ο υπο – module στρέψης και χαρακτηρίζεται από τους συντελεστές στρέψης του t_j. Το σύνολο των $m + 1$ στοιχείων $\{\beta_n\} \cup \{t_j\}$ είναι η παραμετρικοποίηση. Πάνω από ένα σώμα k συντελεστών, όπως το $\mathbb{R}$, το $\mathbb{Q}$, ή το $\mathbb{Z}_p$ για p πρώτο, η H_n απλοποιείται σ΄ έναν k – διανυσματικό χώρο διάστασης $\beta_n = dim H_n$, έτσι ώστε η παραμετρικοποίηση να είναι απλώς ο ακέραιος β_n, δηλαδή ο υπομόδιος στρέψης εξαφανίζεται.

Υπάρχει αντιστοιχία ένα προς ένα μεταξύ της παραπάνω παραμετρικοποίησης και των πεπερασμένα παραγόμενων R – modules, οπότε αυτή η παραμετρικοποίηση είναι μια πλήρης αμετάβλητη. Έχουμε έναν πλήρη χαρακτηρισμό της ομολογίας, υπό την προϋπόθεση ότι την υπολογίζουμε πάνω από έναν ΔΚΙ [39].

Υπολογίζοντας τον αριθμό των γεννητόρων μιας ομάδας ομολογίας εισάγεται ταυτόχρονα η έννοια της τοπολογικής αμετάβλητης Betti numbers. Τα Betti numbers είναι μία μέθοδος αναπαράστασης των ομολογιακών ομάδων διερευνώντας τις ιδιότητες του τοπολογικού χώρου. Διακρίνουν τους τοπολογικούς χώρους με βάση τη συνδεσιμότητα των n – διάστατων simplicial complexes [23].

Ορισμός 2.5.2 : Το ***k – οστό Betti number*** του K είναι η τάξη της k – οστής ομάδας ομολογίας, δηλαδή $\beta_k = rankH_k$. Όμως, $H_k = Z_k / B_k$, οπότε $\beta_k = rankZ_k - rankB_k$.

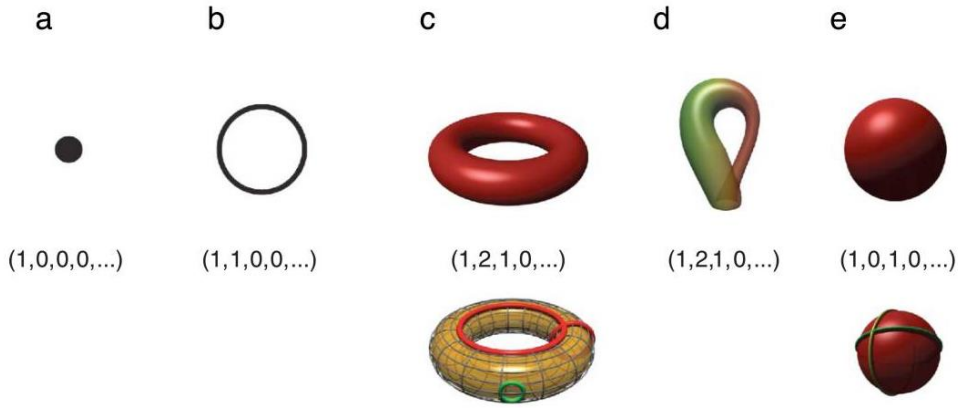
Η ομολογία αναφέρεται σε k – διάστατες τρύπες που εμφανίζονται σε n – διάστατα σύνολα. Οι ομάδες ομολογίας μπορούν να περιγράψουν τα αλγεβρικά ανάλογα των τρυπών σε μία πολλαπλότητα σε διάφορες διαστάσεις. Το k – οστό Betti number αναφέρεται στον αριθμό των k – διάστατων τρυπών μιας τοπολογικής επιφάνειας.

Για παράδειγμα, στον $\mathbb{R}^3$, το Betti number ενός simplicial complex K είναι :
– β_0 ο αριθμός των στοιχείων του K,

– β_1 η τάξη μιας οποιασδήποτε βάσης των μη συσταλτών κλειστών καμπυλών του K,

– β_2 η τάξη μιας οποιασδήποτε βάσης των μη συσταλτών κλειστών επιφανειών του K.

Συνεπώς, τα Betti numbers παρέχουν μία ταυτότητα της τοπολογίας του σχήματος. Στο παρακάτω σχήμα παρουσιάζεται το διάνυσμα των Betti numbers (β_0, β_1, β_2, …) του (a) σημείου, (b) του κύκλου, (c) της σαμπρέλας, (d) του *Klein bottle* και (e) της σφαίρας.



Σχήμα 2.5.2 Betti numbers βασικών σχημάτων [33]

Όσον αφορά τη σαμπρέλα, έχει τρεις βρόχους στην επιφάνειά της. Οι κόκκινοι βρόχοι δεν μπορούν να συρρικνωθούν σε σημείο, ούτε μπορούν να μετασχηματιστούν ο ένας στον άλλον χωρίς να σκιστούν. Από την άλλη, ο πράσινος βρόχος μπορεί να μετασχηματιστεί σε σημείο εύκολα. Έτσι, έχουμε $b_1 = 2$. Στην περίπτωση της σφαίρας, όλοι οι βρόχοι μπορούν να

συρρικνωθούν σε σημεία, οπότε $b_1 = 0$. Τόσο η σφαίρα όσο και η σαμπρέλα έχουν $b_2 = 1$, διότι και οι δύο επιφάνειες περικλείουν ένα μέρος του χώρου.

Παράδειγμα 2.5.2 : Σε συνέχεια του Παραδείγματος 2.5.1, τα Betti numbers του πλεγματικού συμπλόκου είναι :

$$\beta_0 = rankH_0 = rank\{0\} = 1$$
$$\beta_1 = rankH_1 = rank\mathbb{Z} = 1$$

Παρατήρηση : Σύμφωνα με το Universal Coefficient Theorem for Homology [27], για τα complexes του $\mathbb{R}^3$, ο αριθμός Betti κάτω από το $\mathbb{Z}_2$ είναι ο ίδιος με αυτόν του $\mathbb{Z}$.

2.6 Χαρακτηριστικός Αριθμός Euler

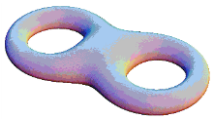
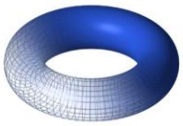
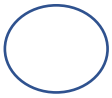
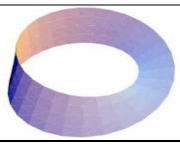
Ορισμός 2.6.1 : Έστω K ένα πλεγματικό σύμπλοκο. Ο ***Χαρακτηριστικός Αριθμός Euler*** $\chi(K)$ ορίζεται ως εξής :

$$\chi(K) = \sum_{\sigma \in K}(-1)^{\dim\sigma} = \sum_{n=0}^{\dim K}(-1)^n c_n$$

όπου c_n είναι το πλήθος των n – διάστατων simplices του K.

Ο (Χαρακτηριστικός) Αριθμός Euler είναι μία ακέραιη αναλλοίωτη του $|K|$ για ομοτοπικά ισοδύναμους χώρους. Επομένως, διαφορετικά complexes των οποίων οι χώροι είναι ομοτοπικά ισοδύναμοι έχουν τον ίδιο αριθμό Euler.

Παρακάτω, παρουσιάζονται οι Χαρακτηριστικοί Αριθμοί Euler βασικών τοπολογικών χώρων (Πίνακας 2.6.1).

Τοπολογικός Χώρος		$\chi(X)$	Τοπολογικός Χώρος		$\chi(X)$
Solid double torus		**-1**	**Solid torus**		**0**
Double torus	**Boundary of solid double torus**	**-2**	**Ball**		**1**
circle		**0**	**disk**		**1**
Annulus		**0**	**Closed interval**	[——]	**1**
Mobius band		**0**	**sphere**	**Boundary of ball**	**2**

Πίνακας 2.6.1 Χαρακτηριστικοί Αριθμοί Euler βασικών τοπολογικών χώρων

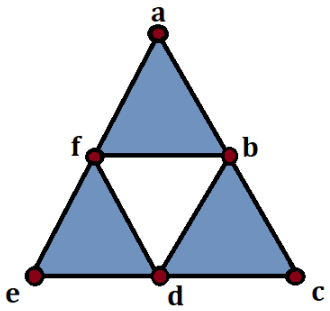
Παρατήρηση : Αντικείμενα με τον ίδιο Αριθμό Euler δεν είναι απαραίτητα τοπολογικά ισοδύναμα.

Το γεγονός ότι ο Αριθμός Euler είναι αναλλοίωτος μπορεί, επίσης, να συναχθεί από το γεγονός ότι η ομολογία είναι αναλλοίωτη. Για έναν τοπολογικό χώρο X, ο τύπος του Euler – Poincare είναι ο εξής :

$$\chi(X) = \sum_n (-1)^n \beta_n$$

Ο παραπάνω τύπος δίνει έμφαση στο γεγονός ότι ο Αριθμός Euler μπορεί να οριστεί καθαρά μέσω της ομολογίας και βασίζεται μόνο στον τύπο ομοτοπίας του X. Δηλαδή, ο $\chi(X)$ είναι ανεξάρτητος από την επιλογή του complex που αναπαριστά το X.

Παράδειγμα 2.6.1 : Έστω το πλεγματικό σύμπλοκο K στο σύνολο των κορυφών $\{a, b, c, d, e, f\}$ που καθορίζεται πλήρως από τις μεγιστικές όψεις $\mathcal{F}_1 = \{a, b, f\}$, $\mathcal{F}_2 = \{b, c, d\}$, $\mathcal{F}_3 = \{d,e,f\}$, όπως φαίνεται και στο παρακάτω σχήμα :



Σχήμα 2.6.1 Πλεγματικό σύμπλοκο

Τότε, $\chi(K) = \sum_{n=0}^{\dim K}(-1)^n c_n = (-1)^0 * 6 + (-1)^1 * 9 + (-1)^2 * 3 = 0$

ή $\chi(K) = \sum_n(-1)^n\beta_n = (-1)^0 * 1 + (-1)^1 * 1 = 0$

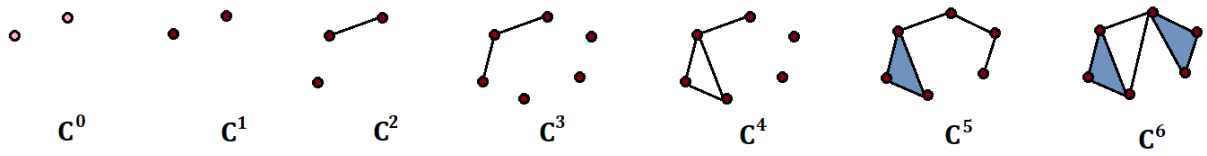
2.7 Εμμένουσα Ομολογία – Persistent Homology

Για την μετατροπή ενός PCD σε ένα τοπολογικό αντικείμενο απαιτείται η επιλογή μιας παραμέτρου ε. Για ε ικανοποιητικά μικρό, το complex είναι ένα διακεκριμένο σύνολο, ενώ για ε αρκετά μεγάλο, το complex είναι ένα simplex υψηλών διαστάσεων.

Επομένως, η ομολογία ενός complex που συνδέεται με ένα PCD με συγκεκριμένο ε είναι ανεπαρκής. Είναι λάθος να ρωτήσουμε ποια τιμή του ε είναι η βέλτιστη. Ούτε επαρκεί να γνωρίζουμε μία απλή μέτρηση του αριθμού και του είδους των τρυπών που εμφανίζονται σε κάθε τιμή της παραμέτρου. Τα Betti numbers δεν είναι αρκετά. Πρέπει να αποσαφηνιστεί ποιες τρύπες είναι σημαντικές και ποιες μπορούν να αγνοηθούν. Η τυπική τοπολογική κατασκευή της ομολογίας και της ομοτοπίας δεν προσφέρουν τέτοιες δυνατότητες.

Εκείνα τα τοπολογικά χαρακτηριστικά που επιμένουν σε ένα σημαντικό εύρος της τιμής της παραμέτρου θεωρούνται σήμα, ενώ τα υπόλοιπα εκλαμβάνονται ως θόρυβος. Κατά την τοπολογική ανάλυση δεδομένων, μας ενδιαφέρουν εκείνα τα χαρακτηριστικά που επιμένουν.

Ορισμός 2.7.1 : Ένα φιλτράρισμα ενός simplicial complex C είναι μία αύξουσα ακολουθία από σύμπλοκα $\emptyset = C^0 \subseteq C^1 \subseteq \ldots \subseteq C^m = C$. Το C καλείται ***filtered complex***.



Σχήμα 2.7.1 Filtered complex

Με λίγα λόγια σε ένα filtered complex, κάθε καινούργιο complex που προκύπτει περιέχει το προηγούμενο.

Για τον υπολογισμό τον ομολογιακών χαρακτηριστικών που επιμένουν κατά τη διάρκεια ενός φιλτραρίσματος, γίνεται χρήση της εμμένουσας ομολογίας.

Ορισμός 2.7.2 : Έστω C ένα filtered complex, $C^0 \subseteq C^1 \subseteq \ldots \subseteq C^m$. Η ***p − εμμένουσα k − οστή ομολογιακή ομάδα (p − persistent k − th homology group)*** του C^i ορίζεται ως :

$$H_k^{i,p} = Z_k^i / (B_k^{i+p} \cap Z_k^i)$$

η οποία είναι καλώς ορισμένη καθώς $B_k^{i+p}, Z_k^i \subseteq C_k^{i+p}$ άρα $(B_k^{i+p} \cap Z_k^i) \subseteq Z_k^i$.

Ορισμός 2.7.3 : Η τάξη της ελεύθερης αβελιανής ομάδας $H_k^{i,p}$ ονομάζεται ***p − persistent k − th Betti number*** του C^i. Τα ***p − persistent k − th homology groups*** μπορούν ακόμα να οριστούν με τη βοήθεια του ισομορφισμού :

$$\eta_k^{i,p} : H_k^i \to H_k^{i+p}$$

που απεικονίζει μία ομολογιακή κλάση σε μία που την περιέχει. Η εικόνα του ισομορφισμού είναι ισόμορφη με το p − persistent k − th homology group, δηλαδή $im\, \eta_k^{i,p} \approx H_k^{i,p}$.

Στη συνέχεια, επεκτείνουμε τον παραπάνω ορισμό για μια οικογένεια αλυσιδωτών συμπλόκων.

Ορισμός 2.7.4 : Ένα ***Επίμονο Σύμπλοκο (Persistence Complex)*** είναι μια οικογένεια αλυσιδωτών συμπλόκων $\boldsymbol{C} = \{C_*^i\}_{i\geq 0}$ πάνω από τον R, μαζί με το chain maps $f^i : C_*^i \to C_*^{i+1}$, έτσι ώστε να προκύπτει το ακόλουθο διάγραμμα :

$$
\begin{array}{ccccccc}
\downarrow\partial_3 & & \downarrow\partial_3 & & \downarrow\partial_3 & & \\
C_2^0 & \xrightarrow{f^0} & C_2^1 & \xrightarrow{f^1} & C_2^2 & \xrightarrow{f^2} & \dots \\
\downarrow\partial_2 & & \downarrow\partial_2 & & \downarrow\partial_2 & & \\
C_1^0 & \xrightarrow{f^0} & C_1^1 & \xrightarrow{f^1} & C_1^2 & \xrightarrow{f^2} & \dots \\
\downarrow\partial_1 & & \downarrow\partial_1 & & \downarrow\partial_1 & & \\
C_0^0 & \xrightarrow{f^0} & C_0^1 & \xrightarrow{f^1} & C_0^2 & \xrightarrow{f^2} & \dots
\end{array}
$$

Το φιλτράρισμα αυξάνεται οριζοντίως προς τα δεξιά κάτω από το chain maps f^{i}, και η διάσταση μειώνεται καθέτως προς τα κάτω μέσω των συνοριακών τελεστών. Από τη στιγμή που τα complexes αυξάνουν με το ε, τα chain maps f^{i} έχουν σχέσεις εγκλεισμού.

Ορισμός 2.7.5 : Για $i < j$, η $(\boldsymbol{i,j})$ – ***εμμένουσα ομάδα ομολογίας (persistence homology group)*** του $\boldsymbol{C}$, $H_*^{i\to j}(C)$, ορίζεται ως η εικόνα του επαγόμενου ομοιομορφισμού $f_*^{i\to j}$: $H_*\left(C_*^i\right) \to H_*\left(C_*^j\right)$.

Ο παραπάνω ορισμός επεκτείνεται και για ΔΚΙ, όπως και στην απλή ομολογία. Η δομή τους περιγράφεται αναλυτικά στο [39].

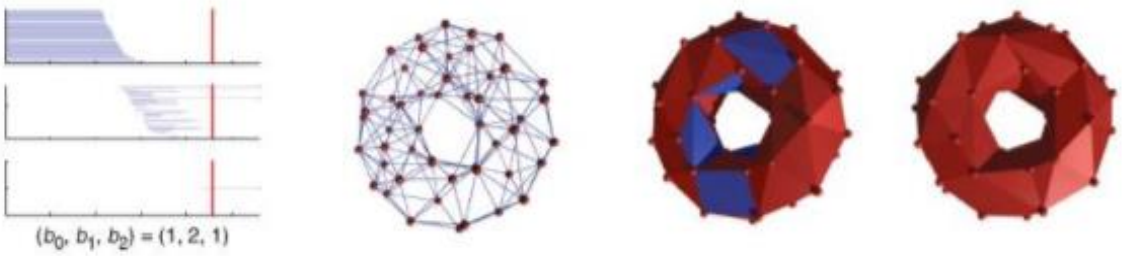
2.8 Barcodes

Τελικό βήμα της μεθόδου που αναπτύχθηκε αποτελεί η αναπαράσταση της εμμένουσας ομολογίας μέσω των barcodes. Τα barcodes αποτυπώνουν την εμμένουσα ομολογία ενός συνόλου δεδομένων με τη μορφή μιας παραμετρικοποιημένης εκδοχής των Betti numbers, απεικονίζοντας τη διάρκεια ζωής μιας ομολογιακής κλάσης εντός ενός φιλτραρίσματος ε, με τη μορφή ενός συνόλου διαστημάτων. Δηλαδή, παρέχουν τη δυνατότητα μέτρησης της σημασίας των ομολογιακών χαρακτηριστικών μέσω της διάρκειας ζωής τους. Τα barcodes

έχουν εφαρμοστεί επιτυχώς σε ένα πλήθος κλάδων, συμπεριλαμβανομένης της αναγνώριση σχήματος [10], της βιοφυσικής [22] και του computer vision [5].

Ορισμός 2.8.1 : Το *γράφημα* ***Barcode*** είναι μία οπτικοποίηση της ομολογιακής ομάδας που αναπαρίσταται από μια συλλογή οριζόντιων ευθυγράμμων τμημάτων. Ο άξονας των x αναπαριστά την παράμετρο ε και ο άξονας των y αντιπροσωπεύει μία αυθαίρετη διάταξη των ομολογιακών γεννητόρων.

Η ιδιαίτερη χρησιμότητα του γραφήματος Barcode είναι ότι φιλτράρει τον τοπολογικό θόρυβο και αιχμαλωτίζει μόνο τα σημαντικά χαρακτηριστικά. Οι μικρές γραμμές θεωρούνται θόρυβος, επειδή δεν διαρκούν αρκετά. Αντιθέτως, οι μεγάλες γραμμές αναπαριστούν σημαντικά χαρακτηριστικά των δεδομένων [23]. Με αυτόν τον τρόπο είναι δυνατόν να εντοπιστούν σε ένα point cloud τα στατιστικώς σημαντικά στοιχεία.



Σχήμα 2.8.1 Κατασκευή του Rips complex και δημιουργία barcodes για 50 τυχαία σημεία ενός torus [33].

Κεφάλαιο 3

Εφαρμογές σε πραγματικά δεδομένα

Μετά τη μαθηματική περιγραφή της εμμένουσας ομολογίας, σ' αυτό το κεφάλαιο, συνοψίζονται οι τοπολογικές πληροφορίες που παρέχουν τα δεδομένα μέσα από τα βήματα που αναπτύχθηκαν στη θεωρία. Δηλαδή, δοθέντος ενός συνόλου πραγματικών δεδομένων, επικεντρωνόμαστε στην ανάκτηση της τοπολογίας του χώρου δειγματοληψίας, με τον υπολογισμό της ομολογίας, την παράθεση των αντίστοιχων γραφημάτων και τη σχετική ερμηνεία τους.

Η διαδικασία υπολογισμού της εμμένουσας ομολογίας ολοκληρώνεται σε δύο βασικά βήματα :

1. Κατασκευή ενός φιλτραρισμένου simplicial complex δοθέντος ενός γεωμετρικού συνόλου δεδομένων, με την παράλληλη μεταβολή του ε.
2. Υπολογισμός της εμμένουσας ομολογίας του φιλτραρισμένου complex και αναπαράστασή του μέσω γραφικών απεικονίσεων.

Συγκεκριμένα, οι δύο βασικές γραφικές απεικονίσεις, που συνίστανται στην αναπαράσταση των τοπολογικών δομών, αποτελούν το *Barcode* και το *διάγραμμα Persistence*. Το *Barcode*, όπως αναπτύχθηκε και στην Ενότητα 2.8, απεικονίζει τη διάρκεια ζωής μιας ομολογιακής κλάσης εντός ενός φιλτραρίσματος ε, με τη μορφή ενός συνόλου διαστημάτων. Από την άλλη, το *διάγραμμα Persistence* παρουσιάζει το παραπάνω σύνολο διαστημάτων $(birth, death)$, ως σημεία στο επίπεδο με συντεταγμένες $x = birth$ και $y = death$.

Πιο αναλυτικά, εντός ενός φιλτραρίσματος ε, τοπολογικά χαρακτηριστικά (αριθμός συνιστωσών, τρύπες, σήραγγες κτλ. ενός τοπολογικού χώρου) θα γεννηθούν και στη συνέχεια θα πεθάνουν. Κάθε χαρακτηριστικό μπορεί να αναπαρασταθεί από ένα ζεύγος $z = (birth, death)$ που δείχνει το χρόνο γέννησης και το χρόνο θανάτου αυτού του χαρακτηριστικού. Έτσι, ένα *διάγραμμα Persistence* είναι μια γραφική παράσταση των σημείων z. Συμπερασματικά, προκύπτει ότι χαρακτηριστικά τα οποία δεν επιμένουν κατά τη διάρκεια του φιλτραρίσματος, έχουν χρόνους θανάτου κοντά στους χρόνους γέννησής τους. Με άλλα λόγια, τα χαρακτηριστικά αυτά αντιστοιχούν σε σημεία z κοντά στη διαγώνιο. Συνεπώς, ο

τοπολογικός θόρυβος αντιστοιχεί σε σημεία κοντά στη διαγώνιο και το σήμα αντιστοιχεί σε σημεία μακριά από τη διαγώνιο [26].

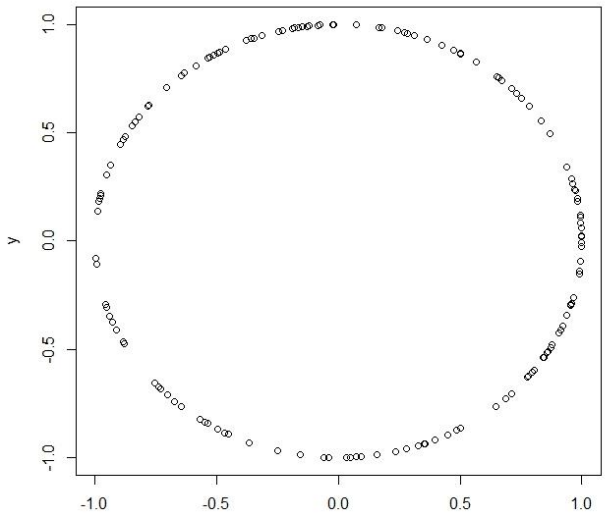
Αξιοποιώντας τα παραπάνω στάδια, γίνεται εφαρμογή της μεθόδου σε σύνολα δεδομένων που παράγονται από δείγματα αντικειμένων με γνωστή τοπολογία, ώστε να επιβεβαιωθεί η αποτελεσματικότητα της μεθόδου. Στη συνέχεια, εφαρμόζεται στα σύνολα δεδομένων *'Iris flower data set' και 'German Credit data set'*, ώστε να προσδιοριστεί ο χώρος στον οποίο ζουν τα δεδομένα.

Για την υλοποίηση της παραπάνω διαδικασίας χρησιμοποιήθηκε η γλώσσα προγραμματισμού ανοιχτού κώδικα R, και συγκεκριμένα το πακέτο "pHom", το οποίο υπολογίζει την εμμένουσα ομολογία ενός τοπολογικού χώρου. Αυτό το πακέτο δίνει τη δυνατότητα κατασκευής δύο διαφορετικών filtered complexes ενός μετρικού χώρου (X, *d*), του Vietoris – Rips και του Lazy – Witness. Όσον αφορά την επιλογή μετρικής απόστασης, παρέχει δυνατότητα επιλογής ανάμεσα στις "euclidean", "maximum", "manhattan", "canberra", "binary", "minkowski". Εναλλακτικά, μπορεί να γίνει χρήση του πακέτου της R "TDA" ή του JavaPlex στη MatLab.

3.1 Εφαρμογή σε σύνολα δεδομένων με γνωστή τοπολογία

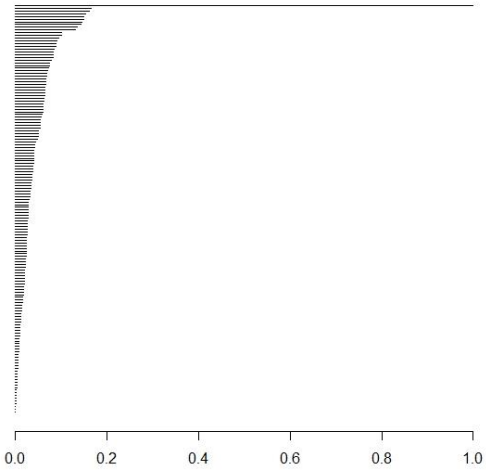
Κύκλος

Έστω ένα τυχαίο δείγμα 150 δεδομένων μέσα από το σύνολο $\{(\cos(t), \sin(t)) \,/\, t \in [0, 2\pi]\}$. Για τη δημιουργία του PCD (*Σχήμα 3.1.1*) θεωρούμε το σύνολο των δεδομένων ως υποσύνολο του χώρου $(\mathbb{R}^2, d_{\mathbb{R}})$, όπου $d_{\mathbb{R}}$ η Ευκλείδεια μετρική.

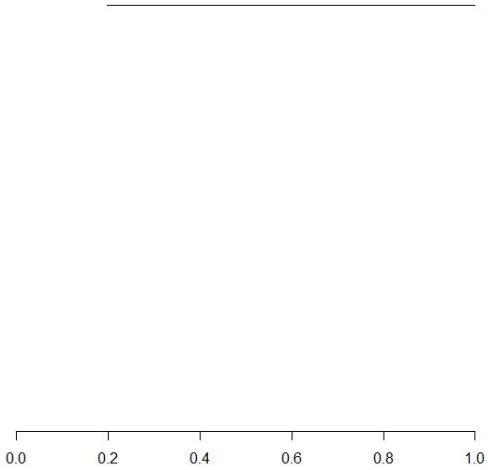


Σχήμα 3.1.1 Δειγματοληψία από μοναδιαίο κύκλο

Παρακάτω, παρουσιάζονται οι πληροφορίες για τα ομολογιακά χαρακτηριστικά του συνόλου των δεδομένων, μέσα από τα γραφήματα *Barcodes* (για H_0, H_1) και το *διάγραμμα Persistence*, όπως προκύπτουν με τη δημιουργία των Vietoris Rips complexes, καθώς εξελίσσεται η τοπολογία του σχήματος για $0 < \varepsilon < 1$.

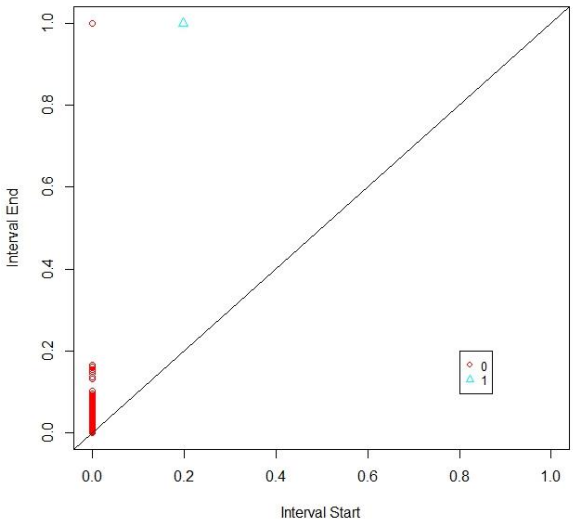


Σχήμα 3.1.2 (α) Barcode για H_0



Σχήμα 3.1.2 (β) Barcode για H_1

Στο σχήμα *3.1.2 (α)*, τα οριζόντια τμήματα (μπάρες) αναπαριστούν τους κύκλους μηδενικής διάστασης. Όλα τα τμήματα ξεκινούν από το χρόνο μηδέν, αλλά μπορούν να λήξουν σε διαφορετικούς χρόνους καθώς μεταβάλλεται το ε. Στο συγκεκριμένο σχήμα διαπιστώνεται ότι μόνο μία μπάρα έχει μεγάλη διάρκεια ζωής. Στο σχήμα *3.1.2 (β)*, οι μπάρες αναπαριστούν τους μονοδιάστατους κύκλους. Υπάρχει μία μόνο μπάρα, η οποία μάλιστα επιμένει.



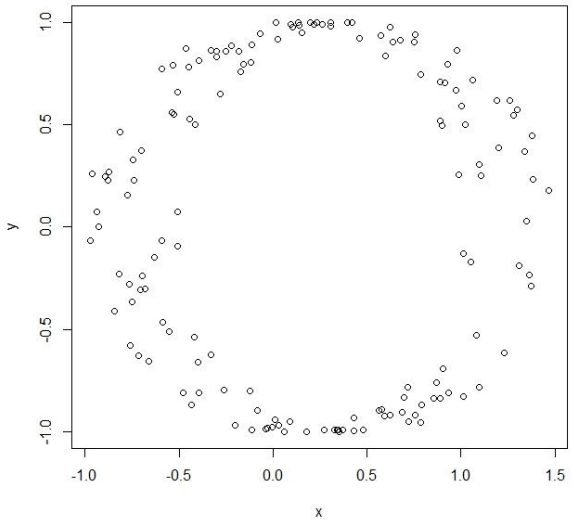
Σχήμα 3.1.3 Διάγραμμα Persistence

Το *διάγραμμα Persistence* στο *Σχήμα 3.1.3* παρουσιάζει σημεία που αντιπροσωπεύουν τη γέννηση και το θάνατο κύκλων στο H_i, $για\ i\ =\ 0,1$. Ένα κόκκινο σημείο είναι ένας κύκλος στο H_0, ενώ ένα πράσινο τρίγωνο είναι ένας κύκλος στο H_1. Η πλειονότητα των κύκλων στο H_0 συσσωρεύονται κοντά στη διαγώνιο, εκτός από έναν που απέχει αισθητά από αυτήν. Στο H_1 εμφανίζεται ένας μοναδικός κύκλος, μακριά από τη διαγώνιο.

Επομένως, ο χώρος του δείγματος αποτελείται από μία συνιστώσα και έναν βρόγχο, δηλαδή μπορεί να περιγραφεί από το $b = (1,1,0,..,0)$.

Κύκλος με θόρυβο

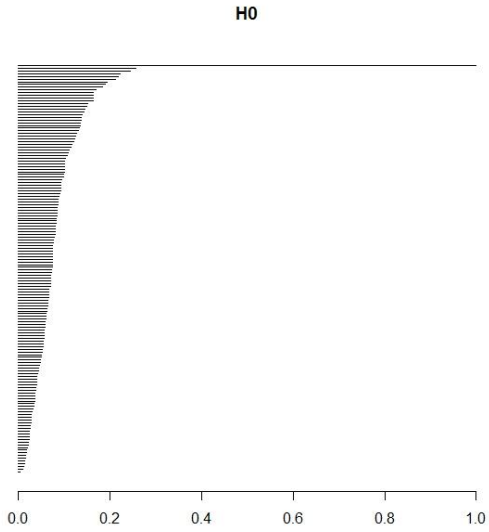
Έστω ένα τυχαίο δείγμα 150 δεδομένων μέσα από το σύνολο $\{(\cos(t) + \frac{1}{2} * s, \sin(t))\ /\ t\ \in\ [0, 2\pi],\ s \in\ [0, 1]\}$. Για τη δημιουργία του PCD (*Σχήμα 3.1.4*) θεωρούμε το σύνολο των δεδομένων ως υποσύνολο του χώρου $(\mathbb{R}^2, d_{\mathbb{R}})$, όπου $d_{\mathbb{R}}$ η Ευκλείδεια μετρική.



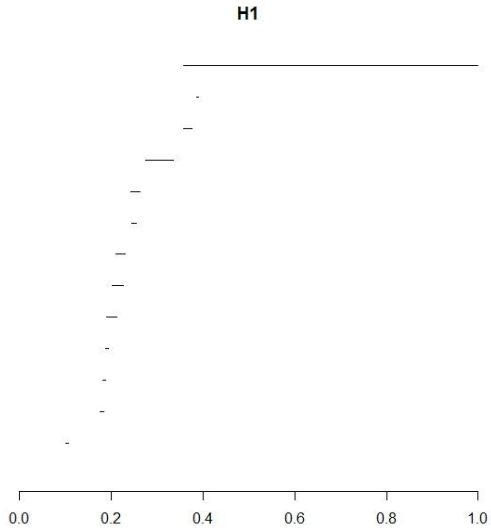
Σχήμα 3.1.4 Δειγματοληψία από κύκλο με θόρυβο

Στα παρακάτω σχήματα παρουσιάζονται τα *Barcodes* για H_0 και H_1, που προκύπτουν με τη δημιουργία των Vietoris Rips complexes, καθώς εξελίσσεται η τοπολογία του σχήματος

για $0 < \varepsilon < 1$. Όπως φαίνεται παρακάτω, και τα δύο γραφήματα έχουν από μία μόνο μπάρα που επιμένει.

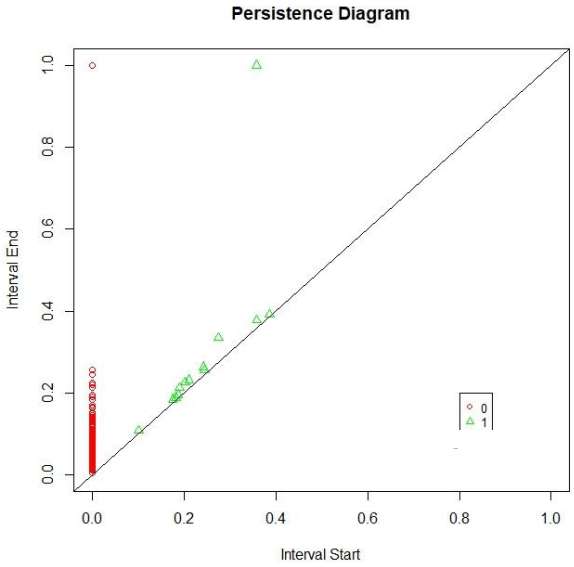


Σχήμα 3.1.5 (α) Barcode για H_0



Σχήμα 3.1.5 (β) Barcode για H_1

Αντίστοιχα αποτελέσματα προκύπτουν και από το διάγραμμα Persistence, δηλαδή ο χώρος δειγματοληψίας περιγράφεται από το διάνυσμα $b = (1,1,0,..,0)$.

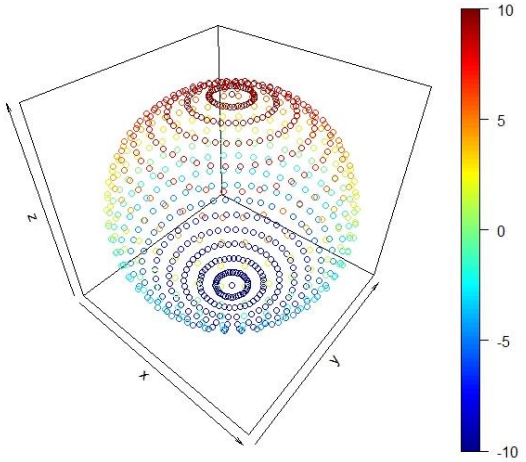


Σχήμα 3.1.6 Διάγραμμα Persistence

Παρατηρείται ότι παρά το θόρυβο που προστέθηκε, με χρήση της μεθόδου είναι δυνατό να εντοπιστεί ο χώρος δειγματοληψίας, αφού και πάλι προκύπτει μία συνιστώσα και ένας βρόγχος. Δηλαδή, τα *Barcodes* και το *Διάγραμμα Persistence* είναι σε θέση να διαχωρίσουν επαρκώς τον τοπολογικό θόρυβο από τα τοπολογικά χαρακτηριστικά.

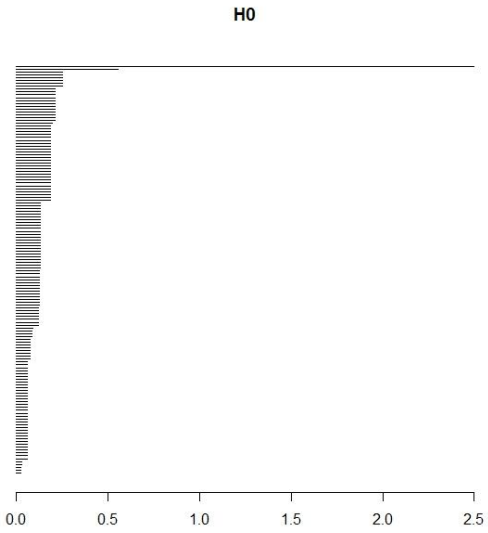
Σφαίρα

Έστω ένα τυχαίο δείγμα 1681 δεδομένων μέσα από το σύνολο $\{(10 * cos(p) * cos(t), 10 * cos(p) * sin(t), 10 * \sin(t)) \ / \ t, p \in [0, 2\pi]\}$. Για τη δημιουργία του PCD (*Σχήμα 3.1.7*) θεωρούμε το σύνολο των δεδομένων ως υποσύνολο του χώρου $(\mathbb{R}^3, d_{\mathbb{R}})$, όπου $d_{\mathbb{R}}$ η Ευκλείδεια μετρική.

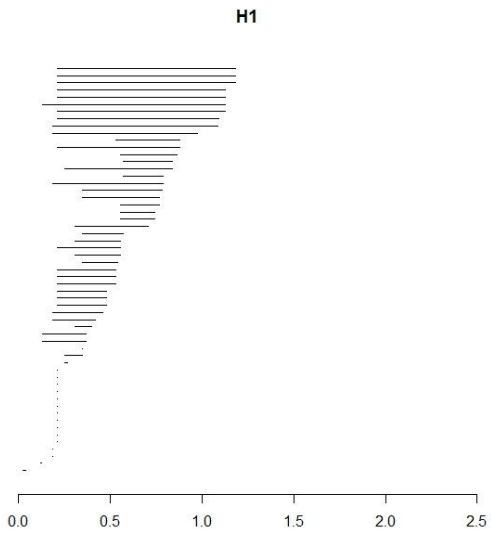


Σχήμα 3.1.7 Δειγματοληψία από σφαίρα

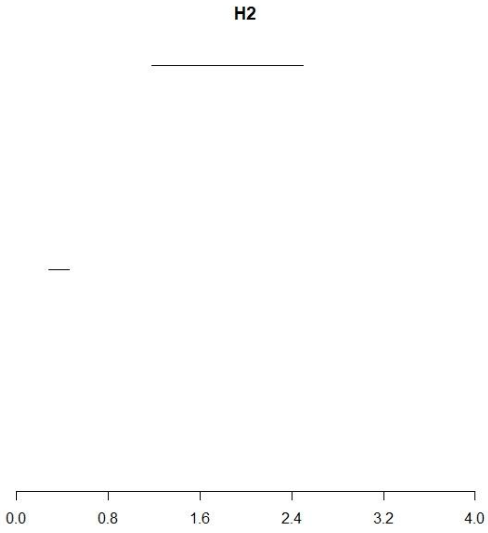
Παρακάτω, εμφανίζονται τα *Barcodes* για H_0, H_1 και H_2 που προκύπτουν με τη δημιουργία των *Low – Witness* complexes, καθώς εξελίσσεται η τοπολογία του σχήματος για $0 < \varepsilon < 2.5$.



Σχήμα 3.1.8 (α) Barcode για H_0

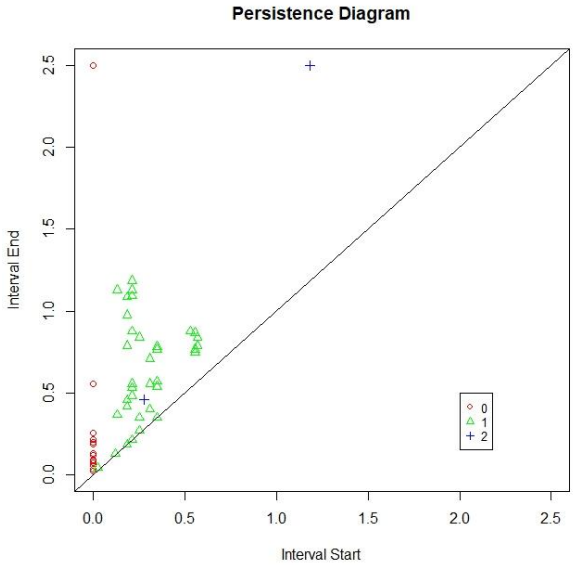


Σχήμα 3.1.8 (β) Barcode για H_1



Σχήμα 3.1.8 (γ) Barcode για H_2

Διαπιστώνεται ότι οι περισσότεροι 0 – κύκλοι έχουν πολύ μικρή διάρκεια ζωής. Πιο συγκεκριμένα, ο μέσος όρος ζωής είναι 0.14796. Ωστόσο, παρατηρείται ένας 0 – διάστασης κύκλος, ο οποίος επιμένει, με τη διάρκεια ζωής του να ανέρχεται στο 2.5. Οι κύκλοι στο H_1 δεν επιμένουν, ενώ στο H_2 υπάρχει ένας επίμονος κύκλος.

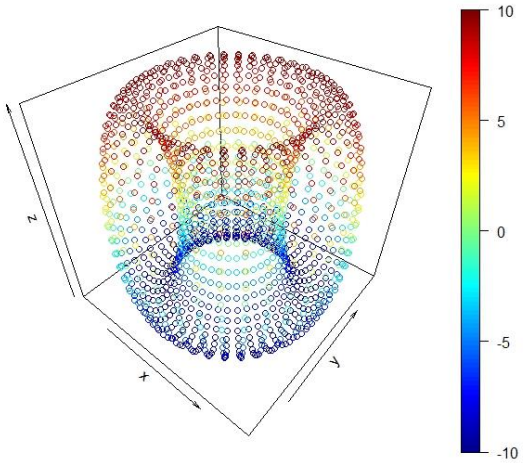


Σχήμα 3.1.9 Διάγραμμα Persistence

Το *διάγραμμα Persistence* στο *Σχήμα 3.1.9* παρουσιάζει σημεία που αντιπροσωπεύουν τη γέννηση και το θάνατο κύκλων στο H_i, $για\ i\ =\ 0,1,2$. Όπως προηγουμένως, ένα κόκκινο σημείο είναι ένας κύκλος στο H_0, ένα πράσινο τρίγωνο είναι ένας κύκλος στο H_1, ενώ το μπλε σύμβολο είναι ένας κύκλος στο H_2. Παρατηρείται ότι οι κύκλοι στο H_0 και H_1 συσσωρεύονται κοντά στη διαγώνιο, πέρα από έναν κύκλο της H_0 που απέχει αισθητά. Η H_2 αποτελείται από έναν κύκλο μικρής διάρκειας ζωής και έναν κύκλο μεγάλης διάρκειας και ίσης με 1.31475. Οι κύκλοι στο H_1 δεν απομακρύνονται σε μεγάλο βαθμό από τη διαγώνιο. Συνοπτικά, ο χώρος δειγματοληψίας περιγράφεται πλήρως από το διάνυσμα $b = (1,0,1,..,0)$.

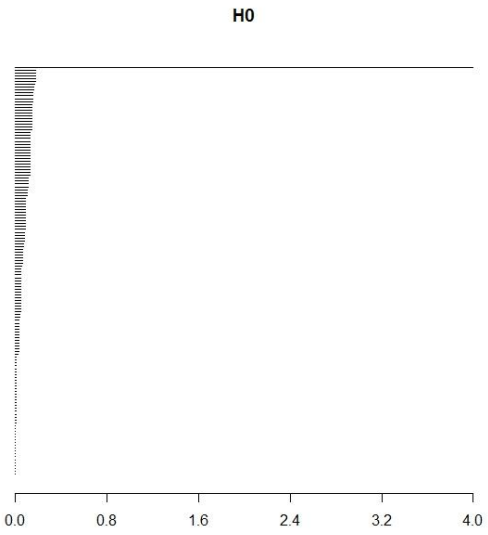
Torus

Έστω ένα τυχαίο δείγμα 1681 δεδομένων μέσα από το σύνολο $\{(25 * \cos(t) + 10 * cos(p) * cos(t), 25 * \sin(t) + 10 * cos(p) * sin(t), 10 * \sin(t))\ /\ t, p\ \in\ [0, 2\pi]\}$. Για τη δημιουργία του PCD (*Σχήμα 3.1.10*) θεωρούμε το σύνολο των δεδομένων ως υποσύνολο του χώρου $(\mathbb{R}^3, d_{\mathbb{R}})$, όπου $d_{\mathbb{R}}$ η Ευκλείδεια μετρική.

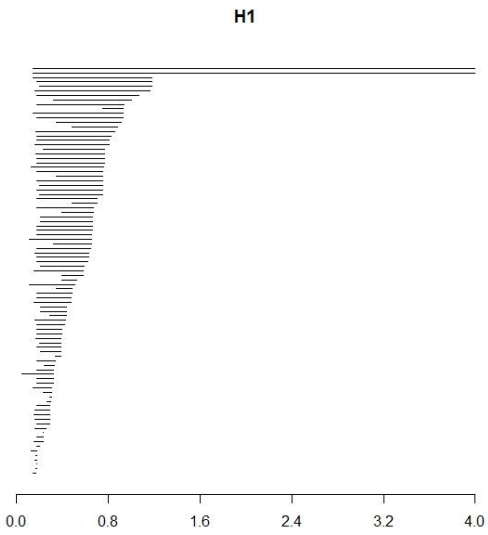


Σχήμα 3.1.10 Δειγματοληψία από torus

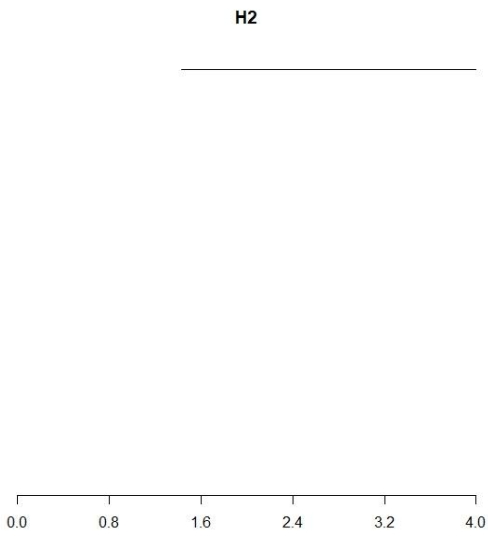
Στα σχήματα παρουσιάζονται τα *Barcodes* για H_0, H_1 και H_2 που προκύπτουν με τη δημιουργία των *Low – Witness* complexes, καθώς εξελίσσεται η τοπολογία του σχήματος για $0 < \varepsilon < 4$.



Σχήμα 3.1.11 (α) Barcode για H_0

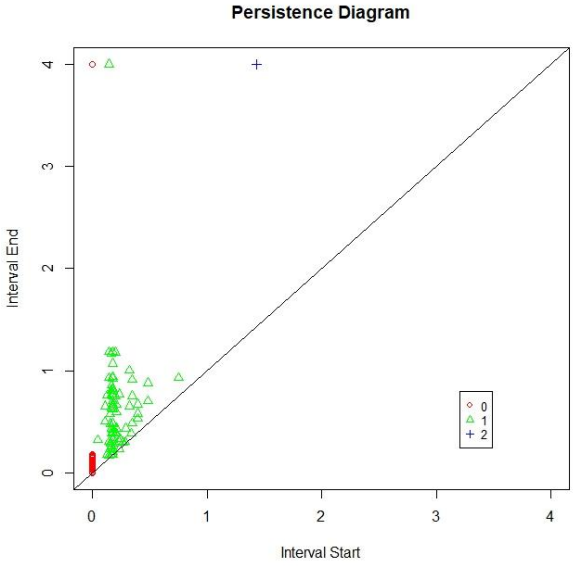


Σχήμα 3.1.11 (β) Barcode για H_1



Σχήμα 3.1.11 (γ) Barcode για H_2

Είναι εμφανές ότι στο H_0 επιμένει ένας κύκλος, στο H_1 δύο κύκλοι και στο H_2 ένας. Παρόμοια συμπεράσματα προκύπτουν και με την ερμηνεία του *διαγράμματος Persistence.* Οπότε, ο χώρος δειγματοληψίας μπορεί να περιγραφεί από το $b = (1,2,1,..,0)$.



Σχήμα 3.1.12 Διάγραμμα Persistence

Στα παραπάνω παραδείγματα αποτυπώθηκε η θεωρία της εμμένουσας ομολογίας σε point cloud data με γνωστή τοπολογία μέσω των *Barcodes* και του *Διαγράμματος Persistence*. Έτσι, έγινε σαφές ότι η εμμένουσα ομολογία είναι ένα ισχυρό εργαλείο για τη μέτρηση των τοπολογικών πληροφοριών των point clouds χωρίς να επηρεάζεται από τις διαταραχές.

3.2 Εφαρμογή στο Iris Data Set

Το *Iris flower data* set ή *Fisher's Iris data set* είναι ένα σύνολο δεδομένων πολλαπλών μεταβλητών που πρωτοχρησιμποποίησε ο βρετανός στατιστικός και βιολόγος Ronald Fisher σε δημοσίευση του το 1936 [18] και αποτελεί κλασσικό παράδειγμα για τις πλείστες τεχνικές πολυμεταβλητής ανάλυσης. Σκοπό έχει να ποσοτικοποιήσει τη μορφολογική διακύμανση των λουλουδιών Iris τριών σχετικών ειδών (Iris setosa, Iris virginica και Iris versicolor).

Το σύνολο των δεδομένων αποτελείται από 50 δείγματα καθενός από τα τρία είδη Iris (setosa, versicolor, virginica). Από κάθε δείγμα μετρήθηκαν τέσσερα χαρακτηριστικά : το μήκος και το πλάτος των κάλυκων και των πετάλων σε εκατοστά. Με βάση το συνδυασμό αυτών των τεσσάρων χαρακτηριστικών, δημιουργείται ένα σύνολο 150 τετραδιάστατων δεδομένων, στα οποία δεν συμπεριλαμβάνεται το είδος του λουλουδιού.

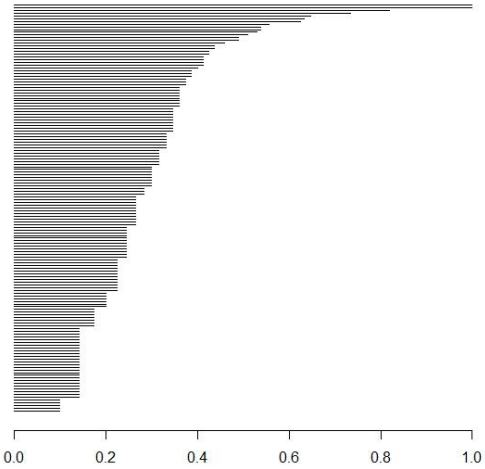
Παρακάτω, παρουσιάζεται συνοπτικός πίνακας με μέρος των δεδομένων :

Είδος	**Μήκος κάλυκα**	**Πλάτος κάλυκα**	**Μήκος πετάλου**	**Πλάτος πετάλου**
Setosa	5.1	3.5	1.4	0.2
Setosa	4.9	3.0	1.4	0.2
…	…	…	…	…
Versicolor	6.4	3.2	4.5	1.5
Versicolor	6.9	3.2	4.9	1.5
..	…	…	…	…
Virginica	6.3	3.3	6	2.5
Virginica	5.8	2.7	5.1	1.9

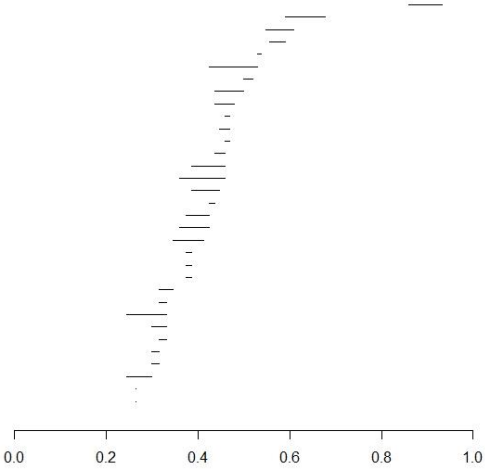
Πίνακας 3.2.1 Συνοπτικός πίνακας των δεδομένων

Για τη δημιουργία του PCD θεωρούμε το σύνολο των δεδομένων ως υποσύνολο του χώρου $(\mathbb{R}^4, d_{\mathbb{R}})$, όπου $d_{\mathbb{R}}$ η Ευκλείδεια μετρική. Στη συνέχεια, παρουσιάζονται οι πληροφορίες για τα ομολογιακά χαρακτηριστικά του συνόλου των δεδομένων, μέσα από τα γραφήματα *Barcodes* (για H_0, H_1 και H_2) και *διάγραμμα Persistence*, που προκύπτουν με τη δημιουργία των Vietoris Rips complexes, καθώς εξελίσσεται η τοπολογία του σχήματος για $0 < \varepsilon < 1$.

Να σημειωθεί ότι το Barcode για H_3, όπως υπολογίστηκε με χρήση της συνάρτησης pHom, δεν περιέχει διαστήματα με μέτρο μεγαλύτερο του μηδενός.



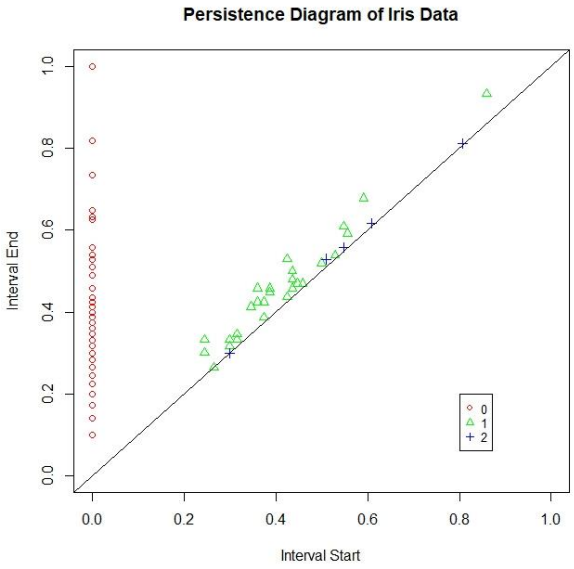
Σχήμα 3.2.1 (α) Barcode για H_0



Σχήμα 3.2.1 (β) Barcode για H_1

Σχήμα 3.2.1 (γ) Barcode για H_2

Οι περισσότεροι 0 – κύκλοι έχουν μικρή διάρκεια ζωής. Πιο συγκεκριμένα, ο μέσος όρος ζωής είναι 0.2945212 με τυπική απόκλιση 0.1541176. Ωστόσο, παρατηρούνται δύο 0 – διάστασης κύκλοι οι οποίοι επιμένουν για ε > 0.82, δηλαδή έχουν διάρκεια ζωής τέσσερις τυπικές αποκλίσεις μεγαλύτερη από το μέσο όρο. Οι κύκλοι μεγαλύτερης διάστασης δεν επιμένουν.



Σχήμα 3.2.2 Διάγραμμα Persistence

Το *διάγραμμα Persistence* στο Σχήμα 3.2.2 παρουσιάζει σημεία που αντιπροσωπεύουν τη γέννηση και το θάνατο κύκλων στο H_i, $για\ i = 0,1,2$. Οι κύκλοι στο H_1 και H_2 συσσωρεύονται κοντά στη διαγώνιο, ενώ οι κύκλοι του H_0 απέχουν αρκετά από τη διαγώνιο, με δύο κύκλους να απέχουν αισθητά.

Συνεπώς, από τη συνολική εικόνα που παρέχουν τα γραφήματα, δύναται να υποθέσουμε ότι δημιουργούνται δύο complexes χωρίς καμία τρύπα. Παρόμοια αποτελέσματα προκύπτουν με χρήση κλασσικών μεθόδων διαχωριστικής ανάλυσης. Για την ακρίβεια, εμφανίζεται το είδος Iris setosa ως ένα cluster, ενώ δεν υπάρχει διαχωρισμός μεταξύ των άλλων δύο.

3.3 Εφαρμογή σε German Credit Data Set

Σ’ αυτήν την ενότητα γίνεται εφαρμογή της μεθόδου σε μια πιο σύνθετη μορφή δεδομένων, για την ακρίβεια πρόκειται για τα German Credit δεδομένα, που προέκυψαν από

έρευνα του καθηγητή Hans Hofmann του τμήματος Οικονομικών του Αμβούργο. Τα δεδομένα αποτελούνται από 1000 παρατηρήσεις και 20 μεταβλητές από τις οποίες οι 7 είναι αριθμητικές και οι 13 κατηγορικές. Παρακάτω, εμφανίζεται συνοπτικός πίνακας με μέρος των δεδομένων :

Balance of current account	Duration in months	Credit history	…	Telephone	Foreign Worker
A11	6	A34		A192	A201
A12	48	A32		A191	A201
A14	12	A34		A191	A201
A11	42	A32		A191	A201

Πίνακας 3.3.1 Συνοπτικός πίνακας των δεδομένων

Κατά την Τοπολογική Ανάλυση Δεδομένων δεν υπάρχει δυνατότητα επεξεργασίας αυτούσιου του συγκεκριμένου συνόλου δεδομένων, καθώς η TDA δεν είναι σε θέση να μετρήσει συσχετίσεις μεταβλητών διαφορετικού τύπου (π.χ. αριθμητική με κατηγορική), συνεπώς κρίνεται απαραίτητη η χρήση του επεξεργασμένου συνόλου δεδομένων. Η διαδικασία μορφοποίησης τροποποιεί τις υπάρχουσες τιμές των δεδομένων, χωρίς ωστόσο να αλλάζει το νόημά τους. Για παράδειγμα, οι τιμές τις μεταβλητής 'Foreign Worker' μετατράπηκαν σε αριθμητικές θέτοντας την τιμή 'A201' ίση με 1 και την τιμή 'A202' ίση με 2, δηλαδή χαρακτηριστικά που ταξινομούνται ως κατηγορικά, κωδικοποιήθηκαν με τη βοήθεια ακέραιων αριθμών. Κατά την τροποποίηση αυτή έχουν προστεθεί επιπλέον μεταβλητές, έτσι καταλήγουμε με 25 αριθμητικές μεταβλητές. Ενδεικτικά παρουσιάζεται ο πίνακας :

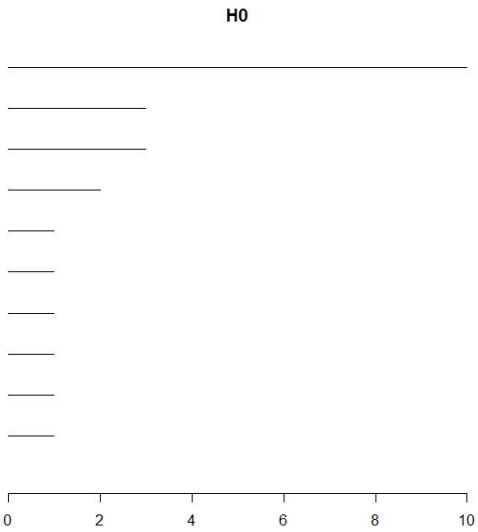
Balance of current account	Duration in months	Credit history	…	Telephone	Foreign Worker
1	6	4		2	1
2	48	2		1	1
4	12	4		1	1
1	42	2		1	1

Πίνακας 3.3.2 Συνοπτικός πίνακας των μορφοποιημένων δεδομένων

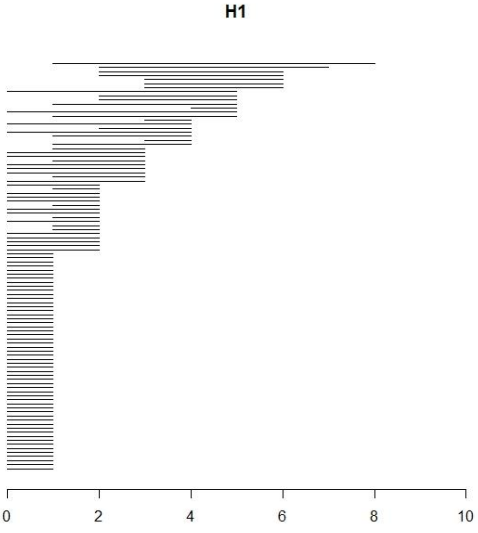
Περισσότερες λεπτομέρειες σχετικά με το σύνολο των δεδομένων είναι διαθέσιμες στον διαδικτυακό τόπο :

ftp://ftp.ics.uci.edu/pub/machine-learning-databases/statlog/german/german.doc

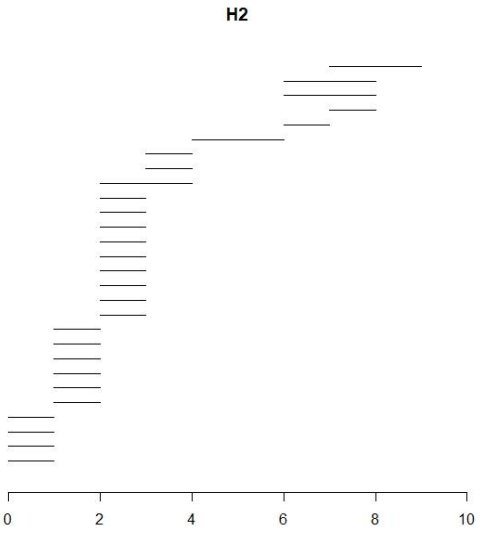
Για τη δημιουργία του PCD θεωρούμε το σύνολο των δεδομένων ως υποσύνολο του χώρου ($\mathbb{R}^{25}$, d), όπου d η απόσταση Manhattan. Στη συνέχεια, παρουσιάζονται οι πληροφορίες για τα ομολογιακά χαρακτηριστικά του συνόλου των δεδομένων, μέσα από τα γραφήματα *Barcodes* (για H_0, H_1 και H_2) και *διάγραμμα Persistence*, που προκύπτουν με τη δημιουργία των *Low – Witness complexes*, καθώς εξελίσσεται η τοπολογία του σχήματος για $0 < \varepsilon < 10$. Τα *Barcodes* μεγαλύτερης διάστασης υπολογίστηκαν εντός του φιλτραρίσματος $0 < \varepsilon < 5$, λόγω υπολογιστικής πολυπλοκότητας. Συγκεκριμένα, παρουσιάζεται το γράφημα για H_3, ενώ *Barcodes* μεγαλύτερης διάστασης, όπως υπολογίστηκε με χρήση της συνάρτησης pHom, δεν περιέχουν διαστήματα με μέτρο μεγαλύτερο του μηδενός.



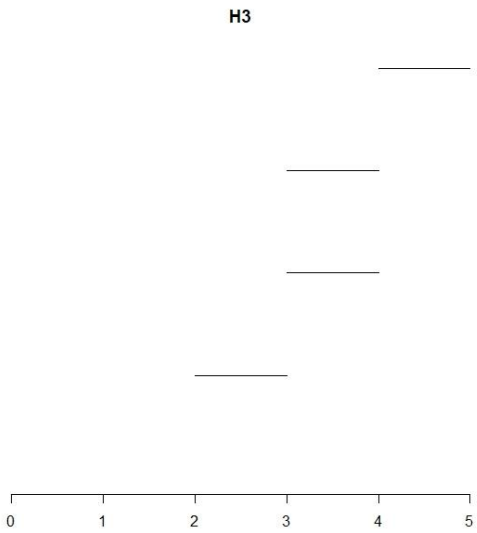
Σχήμα 3.3.1 (α) Barcode για H_0



Σχήμα 3.3.1 (β) Barcode για H_1

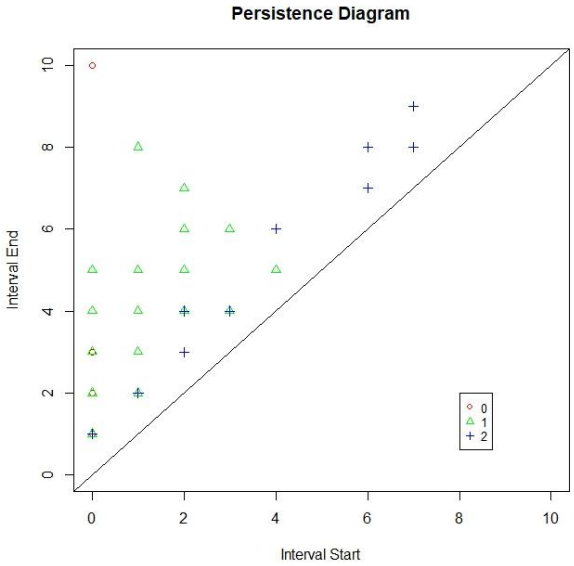


Σχήμα 3.3.1 (γ) Barcode για H_2



Σχήμα 3.3.1 (δ) Barcode για H_3

Είναι εμφανές ότι κατά τη διάρκεια του φιλτραρίσματος επιμένει ένας μόνο κύκλος 0 – διάστασης. Όσον αφορά τους 1 – διάστατους κύκλους, έχουν μέση διάρκεια ζωής 1.533981 και τυπική απόκλιση 1.017644, ενώ παρατηρούνται δύο 1 – διάστασης κύκλοι οι οποίοι επιμένουν με διάρκεια που ξεπερνά το 5. Οι κύκλοι μεγαλύτερης διάστασης δεν επιμένουν.



Σχήμα 3.3.2 Διάγραμμα Persistence

Το *διάγραμμα Persistence* στο Σχήμα 3.3.2 παρουσιάζει σημεία που αντιπροσωπεύουν τη γέννηση και το θάνατο κύκλων στο H_i, $για\ i = 0,1,2$. Οι κύκλοι στο H_3 δεν συμπεριλαμβάνονται στο διάγραμμα, καθώς υπολογίστηκαν για διαφορετικό φιλτράρισμα. Παρατηρείται ένας κύκλος στο H_0 να απομακρύνεται σε σημαντικό βαθμό από τη διαγώνιο. Η πλειονότητα των κύκλων στο H_1 συσσωρεύονται κοντά στη διαγώνιο, εκτός από δύο που απέχουν αισθητά από αυτήν. Στο H_2 όλοι οι κύκλοι συσσωρεύονται γύρω από τη διαγώνιο. Συνεπώς, ο χώρος δειγματοληψίας μπορεί να περιγραφεί από το $b = (1,2,0,..,0)$.

Να επισημανθεί ότι το παραπάνω σύνολο δεδομένων αναλύθηκε ακόμη με τη χρήση άλλων μετρικών ("euclidean", "minkowski") και complexes (Vietoris – Rips) και προέκυψαν αντίστοιχα αποτελέσματα σε διαφορετικό φιλτράρισμα. Γενικότερα, οι μέθοδοι Τοπολογικής Ανάλυσης Δεδομένων παρουσιάζουν ανθεκτικότητα σε αλλαγές της μετρικής [6].

Κεφάλαιο 4

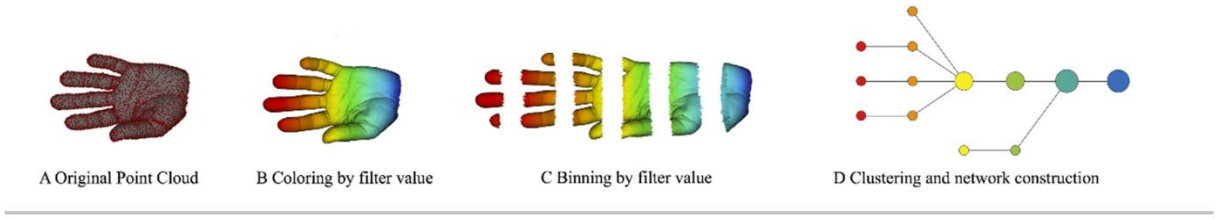
Mapper

Όπως αναπτύχθηκε στα προηγούμενα κεφάλαια, η βασική ιδέα πίσω από την Τοπολογική Ανάλυση Δεδομένων είναι ότι κάθε σύνολο δεδομένων φέρει κάποια γεωμετρική δομή που αντανακλά σημαντικές ιδιότητες του χώρου από τον οποίο έχει προέλθει. Κατά συνέπεια, η TDA αποτελείται από μεθόδους που χρησιμοποιούν την τοπολογία, ή τη μελέτη του σχήματος, για να εξαγάγουν πληροφορίες από τα δεδομένα. Αυτή η προσέγγιση επιτρέπει στους ερευνητές την ελάττωση του όγκου αλλά και των διαστάσεων των δεδομένων, χωρίς ωστόσο να 'θυσιάζει' τις βασικές τοπολογικές ιδιότητες.

Μέχρι στιγμής αναλύθηκε η δυνατότητα της TDA, με χρήση ομολογίας, να εξάγει πληροφορίες για το σχήμα των δεδομένων. Υπάρχουν, ωστόσο, και άλλες διευθύνσεις για οπτικοποίηση και ποσοτική αναπαράσταση των PCD, που επιτρέπουν την εξαγωγή ζωτικής σημασίας πληροφοριών για τη φύση των δεδομένων, όπως ο αλγόριθμος Mapper.

Το Mapper αποτελεί μία απεικονιστική μέθοδο που συνδυάζει τη βασική ιδέα της Θεωρίας Μορς και κλασσικές μεθόδους ομαδοποίησης των δεδομένων για την αποτελεσματική ανάλυση (υψηλών διαστάσεων) δεδομένων σ' έναν αυθαίρετο μετρικό χώρο. Προτάθηκε από τους Gurjeet Sight, Memolli και Carlsson ως ένα γεωμετρικό εργαλείο για την ανάλυση και οπτικοποίηση ενός συνόλου δεδομένων, ενώ αποτελεί τον πυρήνα μιας από τις πιο επιτυχημένες εφαρμογές του [23].

Στη συνέχεια, ακολουθεί η κεντρική ιδέα του αλγορίθμου mapper μέσω ενός παραδείγματος. Έστω ένα πεπερασμένο σύνολο δεδομένων εφοδιασμένο με μία μετρική d (PCD), όπως φαίνεται στο σχήμα 4.1 A. Μία filter function εφαρμόζεται στο PCD και το αντικείμενο χρωματίζεται από τις τιμές των filter values (σχήμα 4.1 B). Το σύνολο των δεδομένων διαιρείται σε επικαλυπτόμενα σύνολα (σχήμα 4.1 C). Κάθε σύνολο ομαδοποιείται, ενώ η διαδικασία ολοκληρώνεται με την κατασκευή ενός δικτύου (σχήμα 4.1 D). Το δίκτυο που προκύπτει αποτελεί μία τοπολογική σύνοψη του χώρου από τον οποίο λήφθηκε το δείγμα. Δηλαδή, το Mapper συμπιέζει ένα σύνολο δεδομένων μεγάλων διαστάσεων και κατασκευάζει τοπολογικά δίκτυα που είναι διακριτά και συνδυαστικά αντικείμενα [23].



Σχήμα 4.1 Αλγόριθμος Mapper [25]

Ορισμός 4.1 : Έστω X και Z τοπολογικοί χώροι και $f : X \to Z$ μία συνεχής συνάρτηση. Έστω $U = \{U_\alpha\}_{\alpha \in A}$ ένα πεπερασμένο ανοιχτό κάλυμμα του Z. Θεωρούμε το κάλυμμα $f^*(U) = \{f^{-1}(U_\alpha)\}_{\alpha \in A}$ του X, το οποίο καλείται *pull-back του U μέσω της f*. Η *κατασκευή του mapper* που προκύπτει από αυτά τα δεδομένα ορίζεται να είναι το νεύρο του καλύμματος $f^*(U)$, δηλαδή $M(U, f) := N(f^*(U))$. Η συνάρτηση f καλείται *filter function*, ενώ ο τοπολογικός χώρος Z μαζί με το ανοιχτό κάλυμμα καλείται *παραμετρικός χώρος*.

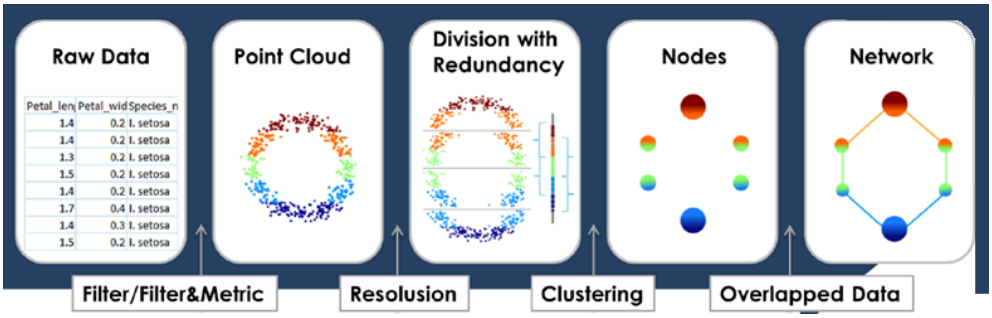
Πιο απλά, για την προσέγγιση ενός τοπολογικού χώρου X επιλέγουμε μία συνεχή απεικόνιση $f : X \to Z$, όπου Z γνωστός τοπολογικός χώρος, με σκοπό την ανάκτηση πληροφοριών σχετικά με τον χώρο X μέσω του χώρου Z. Στη συνέχεια, μέσω της f αποκτούμε ανοιχτά καλύμματα του X από ανοιχτά καλύμματα του Z, και βλέπουμε τι πληροφορίες μπορούμε να αποκομίσουμε σχετικά με το X από το νεύρο αυτών των καλυμμάτων. Η παραπάνω κατασκευή μπορεί να γίνει πιο αποτελεσματική εφαρμόζοντας αλγόριθμο ομαδοποίησης στα πλεονάζοντα data points που προκύπτουν από $f^*(U)$ πριν την κατασκευή του τοπολογικού δικτύου.

Με την επιλογή του αλγορίθμου mapper για την ανάλυση δεδομένων εξασφαλίζεται μία πληρέστερη κατανόηση της δομής τους, ενώ παράλληλα προκύπτουν χρήσιμες γεωμετρικές πληροφορίες. Επίσης, δίνεται η δυνατότητα για άμεση οπτικοποίηση της τοπολογίας των δεδομένων, με την ταυτόχρονη έκθεση σχέσεων μεταξύ ομάδων, μεταβλητών κλπ. Τέλος, ο αλγόριθμος είναι εύκολα υπολογίσιμος ακόμα και για μεγάλα σύνολα δεδομένων.

4.1 Ο Αλγόριθμος Mapper

Όπως προκύπτει από τον *ορισμό 4.1* για την εφαρμογή του αλγορίθμου mapper σ ’ένα σύνολο δεδομένων $S \subseteq X$ αρκεί να προσδιορίσουμε έναν μετρικό χώρο (X, d_x), μία filter

function $f : X \rightarrow Z$ (συνήθως $Z = \mathbb{R}$), ένα πεπερασμένο ανοιχτό κάλυμμα U του Z (όπου μία τυπική επιλογή για το κάλυμμα στην περίπτωση που $Z = \mathbb{R}$ είναι μία πεπερασμένη ακολουθία επικαλυπτόμενων διαστημάτων) και τέλος να επιλέξουμε μία μέθοδο ομαδοποίησης.



Σχήμα 4.1.1 Μετατροπή ενός συνόλου δεδομένων σε τοπολογικό δίκτυο με χρήση του αλγορίθμου mapper [23]

Στη συνέχεια παρουσιάζονται εν συντομία τα βήματα που ακολουθούνται για την εφαρμογή του αλγόριθμου mapper σε ένα σύνολο δεδομένων.

4.1.1 Μετρική Απόσταση

Πρώτο βήμα της μεθόδου αποτελεί η δημιουργία ενός PCD εφοδιάζοντας το σύνολο των δεδομένων με μία μετρική. Η απόσταση αποτελεί μία βασική έννοια στην Ανάλυση Δεδομένων, καθώς μετρά πόσο απέχουν δύο παρατηρήσεις, προσδιορίζει δηλαδή το βαθμό ομοιότητας τους. Η επιλογή της μετρικής καθορίζεται από το πείραμα που παράγει τα δεδομένα καθώς και από το τι μας ενδιαφέρει να καταλάβουμε για τα δεδομένα. Τυπικά, δεν υπάρχει καμία καλή επιλογή μετρικής και είναι αποδεκτό να πραγματοποιείται η ανάλυση με διάφορες μετρικές και να συγκρίνονται τα αποτελέσματα.

Παρακάτω, παρουσιάζονται βασικές μετρικές, χωρισμένες σε ομάδες, ανάλογα με το είδος των δεδομένων στο οποία δύναται να εφαρμοσθούν. Για την επίτευξη του παραπάνω σκοπού χρησιμοποιήθηκαν [13], [23], [24] και [34].

Μετρικές για συνεχή δεδομένα

Όσον αφορά στα συνεχή δεδομένα, οι μετρικές που χρησιμοποιούνται συχνότερα είναι η **Ευκλείδεια** Απόσταση, η Απόσταση **Mahalanobis**, η Απόσταση **Manhattan** καθώς επίσης και η Ομοιότητα κατά **Συνημίτονο**. Για όλες τις παραπάνω μετρικές θεωρούμε δύο μη μηδενικά διανύσματα x = $(x_1, x_2,\ldots, x_n)$ και y = $(y_1, y_2,\ldots, y_n)$, όπου τα x και y είναι δύο παρατηρήσεις και οι συντεταγμένες τους αποτελούν τα χαρακτηριστικά των συγκεκριμένων παρατηρήσεων.

1) Ευκλείδεια Απόσταση

Ο τύπος της Ευκλείδειας Απόστασης ορίζεται να είναι ο εξής :

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Η παραπάνω συνάρτηση μετράει τη συνήθη ευκλείδεια απόσταση μεταξύ δύο σημείων στο n – διάστατο επίπεδο κάνοντας επανειλημμένη χρήση του Πυθαγόρειου θεωρήματος. Η Ευκλείδεια απόσταση αποτελεί την πιο απλή και ταυτόχρονα την πιο γνωστή απόσταση ανάμεσα σε συνεχή δεδομένα.

2) Απόσταση Mahalanobis ή Κανονικοποιημένη Ευκλείδεια Απόσταση

Η Απόσταση Mahalanobis δίνεται από την παρακάτω σχέση :

$$d(x, y) = \frac{\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}}{s_i}$$

όπου s_i είναι η τυπική απόκλιση.

Η παραπάνω απόσταση προκύπτει από την Ευκλείδεια Απόσταση, αν διαιρέσουμε με την τυπική απόκλιση. Καθώς διαιρεί κάθε μεταβλητή με την τυπική της απόκλιση, κάθε μεταβλητή αναφέρεται σε μονάδες τυπικής απόκλισης. Επομένως, δημιουργείται μία συγκρίσιμη κλίμακα σε περιπτώσεις που τα δεδομένα περιέχουν ετερογενής κλίμακας μεταβλητές (για παράδειγμα, βάρος, ύψος και ηλικία).

3) Απόσταση Manhattan

Η Απόσταση Manhattan ορίζεται ως εξής :

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

Η Απόσταση Manhattan μοιάζει πολύ με την Ευκλείδεια Απόσταση με τη διαφορά ότι αντί για τετραγωνικές αποκλίσεις χρησιμοποιούνται απόλυτες αποκλίσεις.

4) Ομοιότητα κατά Συνημίτονο

Η Ομοιότητα κατά Συνημίτονο δίνεται από τον παρακάτω τύπο :

$$d(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2}\sqrt{\sum_{i=1}^{n} {y_i}^2}}$$

Σε αντίθεση με τις παραπάνω αποστάσεις, η Ομοιότητα κατά Συνημίτονο τείνει να αναγνωρίζει καλύτερα πότε δύο δείγματα είναι συγγραμμικά. Για παράδειγμα, η Ομοιότητα κατά Συνημίτονο μπορεί να εφαρμοστεί στις αγοραστικές συνήθειες μιας ομάδας ανθρώπων με σκοπό την ανίχνευση κοινών συμπεριφορών [23].

Μετρικές για κατηγορικά δεδομένα

Όσον αφορά στα κατηγορικά δεδομένα, οι πιο σημαντικές μετρικές είναι η Ομοιότητα κατά **Συνημίτονο**, η Ανομοιότητα **Jaccard**, καθώς επίσης και η Απόσταση **Hamming**.

1) Ομοιότητα κατά Συνημίτονο

Ο τύπος της μετρικής Ομοιότητα κατά **Συνημίτονο** για κατηγορικά δεδομένα είναι ο ίδιος όπως για συνεχή δεδομένα, όμως προηγουμένως απαιτείται μια προετοιμασία των δεδομένων. Ανάλογα με τη συχνότητα της κάθε στήλης, τα κατηγορικά δεδομένα μετατρέπονται σε αριθμητικά.

2) Ανομοιότητα Jaccard

$$d(x,y) = 1 - \frac{\text{set(X)} \cap \text{set(Y)}}{\text{set(X)} \cup \text{set(Y)}}$$

Η μετρική Ανομοιότητα Jaccard υπολογίζει την ανομοιότητα ασύμμετρων πληροφοριών σε μη δυαδικά δεδομένα. Η Ανομοιότητα Jaccard συχνά χρησιμοποιείται για να καθορίσουμε την ομοιότητα των λέξεων σε μία σειρά.

3) Απόσταση Hamming

Η Απόσταση Hamming ορίζεται για δύο λέξεις ίσου μήκους. Η συγκεκριμένη απόσταση καθορίζει το πλήθος των θέσεων των δύο λέξεων, στις οποίες τα αντίστοιχα σύμβολα είναι διαφορετικά. Για παράδειγμα, η Απόσταση Hamming μεταξύ του 4**860**1519 και του 4**725**1519 είναι 3.

4.1.2 Filter Function

Στο δεύτερο βήμα του αλγορίθμου Mapper πραγματοποιείται ο υπολογισμός της filter function. Η συνάρτηση αυτή λειτουργεί απεικονίζοντας κάθε πολυδιάστατο σημείο του συνόλου των δεδομένων σε μια μεμονωμένη τιμή του συνόλου Z (όπου συχνά $Z = \mathbb{R}$). Παράλληλα, μπορεί να χρησιμοποιηθεί αυτόνομα ή να χρησιμοποιήσει το αποτέλεσμα μιας μετρικής ως είσοδο.

Η filter function συνοψίζει σχετικές πληροφορίες από το πεπερασμένο σύνολο του (θορυβώδους) δείγματος. Κάθε filter function είναι δυνατόν να βασίζεται σε τυχαία, γεωμετρικά ή στατιστικά χαρακτηριστικά, συνοψίζοντας τα δεδομένα με διαφορετικό τρόπο. Γι' αυτό και κάθε σύνοψη είναι σε θέση να διατηρήσει διαφορετικές πληροφορίες για το σύνολο των δεδομένων. Βέβαια, υπάρχει δυνατότητα χρήσης πολλαπλών φίλτρων ανεβάζοντας, ωστόσο, τη διάσταση του παραμετρικού χώρου [23].

Η ορθή επιλογή της filter function είναι ιδιαιτέρως σημαντική, καθώς μπορεί να συμβάλλει στην εξαγωγή σημαντικών γεωμετρικών πληροφοριών σχετικά με το σύνολο των δεδομένων. Υπάρχει πληθώρα διαθέσιμων filter functions, όπως η συνάρτηση εκκεντρότητας,

η Kernel Density function, η Distance to measure Function, SVD filters κτλ. Στη συνέχεια, παρουσιάζονται εν συντομία οι συχνότερα χρησιμοποιούμενες filter functions.

1) Συνάρτηση Εκκεντρότητας

Δοθέντος ενός συνόλου σημείων $S \subseteq X$, όπου (X, d_x) μετρικός χώρος, η συνάρτηση *p – οστής* εκκεντρότητας του σημείου $x \in S$ δίνεται από τη σχέση :

$$E_p(x) = \left(\frac{\sum_{y \in S} d_x(x,y)^p}{|S|}\right)^{\frac{1}{p}}$$

όπου p ∈ $[1, +\infty)$.

Η συνάρτηση εκκεντρότητας αποτελεί έναν τρόπο ποσοτικοποίησης της εγγύτητας στο κέντρο του συνόλου των δεδομένων. Δηλαδή, «μετράει» πόσο κεντρικά ή περιφερειακά είναι τα δεδομένα. Τα σημεία που βρίσκονται μακριά από το κέντρο λαμβάνουν υψηλές τιμές εκκεντρότητας. Να παρατηρηθεί, όμως, ότι τα PCD δεν έχουν ένα προκαθορισμένο κέντρο, οπότε έχουμε να κάνουμε με ένα σχετικό μέτρο. Επομένως, μία ουσιώδης έννοια εκκεντρότητας παράγεται από τις κατά ζεύγη αποστάσεις.

Σε περίπτωση που το p τείνει στο άπειρο, η συνάρτηση δίνεται από τη σχέση :

$$E_p(x) = max_{y \in S} d_x(x,y)$$

2) Kernel Density function

Δοθέντος ενός συνόλου σημείων $S \subseteq X$, όπου (X, d_x) μετρικός χώρος και $x \in S$, η **Kernel Density function** ορίζεται ως εξής:

$$f_\varepsilon(x) = C_\varepsilon \sum_{y \in S} \exp\left(\frac{-d_x(x,y)^2}{\varepsilon}\right)$$

όπου C_ε είναι μία κανονικοποιημένη σταθερά και $\varepsilon > 0$ καθορίζει το πόσο λεία είναι η συνάρτηση. Όσο μεγαλύτερο είναι το ε, τόσο πιο λεία είναι. Η πυκνότητα ενός σημείου x ενός PCD S είναι ένα μέτρο της εγγύτητας του x στα γειτονικά του σημεία. Σημεία που έχουν μακρινούς γείτονες έχουν χαμηλές τιμές πυκνότητας και αντίστροφως.

3) kNN – distance Function

Η απόσταση από τον k – πλησιέστερο γείτονα αποτελεί ένα (αντίστροφο) μέτρο πυκνότητας. Δοθέντος ενός συνόλου σημείων $S \subseteq X$, όπου (X, d_x) μετρικός χώρος και $x \in S$, η k – *th* **nearest neighbor density** ορίζεται ως εξής :

$$d_k(x) = \frac{k}{|S| d_x(x, x_k)}$$

Προφανώς, η συνάρτηση δεν έχει νόημα για k = 1, αφού ο πρώτος πλησιέστερος γείτονας ενός σημείου δεδομένων είναι πάντα το ίδιο το σημείο και συνεπώς το αποτέλεσμα είναι σταθερό και ίσο με το μηδέν. Για k = 2 η συνάρτηση υπολογίζει την απόσταση από το x_i στο πλησιέστερο σημείο δεδομένων εξαιρώντας το ίδιο το x_i.

4) Distance to measure Function

Δοθέντος ενός συνόλου σημείων $S \subseteq X$, όπου (X, d_x) μετρικός χώρος και $x \in S$, η **Distance to measure Function** δίνεται από τον παρακάτω τύπο :

$$f(x) = \sqrt{\frac{1}{k} \sum_{j=1}^{k} d_x \left(x, v_j(x)\right)^2}$$

όπου $v_j(x), j \in \{1,2,..,k\}$ είναι οι k – πλησιέστεροι γείτονες του x στο σύνολο των δεδομένων.

5) Distance Matrix Eigenvector (SVD Filter)

Το SVD φίλτρο (Singular Value Decomposition ή Παραγοντοποίηση ιδιαζουσών τιμών) αποτελεί το συνηθέστερα χρησιμοποιούμενο, ως προς τις εφαρμογές, προβολικό (projection) φίλτρο, καθώς θεωρείται ένα αξιόλογο μαθηματικό ‘τέχνασμα’ για την ελάττωση των διαστάσεων των δεδομένων.

Η SVD εκφράζει την παραγοντοποίηση ενός (όχι κατ’ ανάγκη πραγματικού) mxn πίνακα A στη μορφή UDV^t , όπου U και V είναι ορθοκανονικοί πίνακες mxm και nxn, αντίστοιχα και ο D ένας mxn «ορθογώνια διαγώνιος» πίνακας.

Οπότε, δοθέντος ενός συνόλου δεδομένων εφαρμόζεται παραγοντοποίηση ιδιαζουσών τιμών για τον υπολογισμό του k – οστού ιδιάζοντος διανύσματος του πίνακα των αποστάσεων. Η προβολή των δεδομένων στο k – οστό ιδιάζον διάνυσμα μπορεί να χρησιμοποιηθεί ως filter function, ενώ ταυτόχρονα παρέχει μία σύνοψη του τοπολογικού χώρου.

4.1.3 Resolution

Εφόσον υπολογιστούν οι τιμές του φιλτραρίσματος ακολουθεί ο προσδιορισμός ενός καλύμματος για το εύρος του συνόλου Z και εν συνεχεία η αλληλοεπικάλυψή του. Η επιλογή του αριθμού των καλυμμάτων καθώς και το ποσοστό της αλληλοεπικάλυψης προσαρμόζεται ανάλογα με το επιθυμητό αποτέλεσμα. Οι πληροφορίες που παρέχουν τα τελικά δίκτυα επηρεάζονται άμεσα από το ποσοστό αλληλοεπικάλυψης [23].

Στις περισσότερες των περιπτώσεων επιλέγεται κάλυμμα για το σύνολο $[a, b] \subseteq \mathbb{R}$. Το κάλυμμα αυτό αποτελείται από n ισομήκη διαστήματα όπου το p τοις εκατό ενός διαστήματος επικαλύπτεται με το επόμενο διάστημα. Μικρές τιμές του p δίνουν μικρότερη επικάλυψη και άρα λιγότερες συνδέσεις στο τελικό δίκτυο. Μικρές τιμές του n δημιουργούν μεγαλύτερα καλύμματα και επομένως λιγότερους κόμβους.

4.1.4 Clustering

Τέταρτο βήμα του αλγόριθμου Mapper αποτελεί η εφαρμογή κλασσικών μεθόδων ομαδοποίησης στο σύνολο των δεδομένων με σκοπό τη δημιουργία κόμβων. Οι κόμβοι είναι μία ομαδοποιημένη αναπαράσταση του συνόλου των δεδομένων.

Με την πάροδο των ετών έχουν αναπτυχθεί ποικίλοι μέθοδοι ομαδοποίησης των δεδομένων, όπως η μέθοδος K-means και αλγόριθμος Dbscan. Μία από τις πιο κλασσικές μεθόδους αποτελεί η Ιεραρχική Ανάλυση Συστάδων (συσσωρευτική), όπου αρχικά κάθε παρατήρηση θεωρείται ως μια ξεχωριστή συστάδα και σε κάθε επανάληψη του αλγορίθμου εκτελούνται διαδοχικές συγχωνεύσεις συστάδων (οι δύο πλησιέστερες συστάδες συνενώνονται) εωσότου όλες οι παρατηρήσεις μπουν σε μία ομάδα. Παρακάτω παρουσιάζονται εν συντομία τα τρία πιο συχνά χρησιμοποιούμενα μέτρα εγγύτητας μεταξύ των ομάδων:

1) **Single Linkage**

Σύμφωνα με την μέθοδο Single Linkage (κοντινότερου γείτονα) η απόσταση μεταξύ δύο συστάδων C_1, C_2 ορίζεται ως:

$$d(C_1, C_2) = \min_{x_a \in C_1, x_b \in C_2} d(x_a, x_b)$$

2) Complete Linkage

Σύμφωνα με την μέθοδο Complete Linkage (μακρύτερου γείτονα) η απόσταση μεταξύ δύο συστάδων C_1, C_2 ορίζεται ως:

$$d(C_1, C_2) = \max_{x_a \in C_1, x_b \in C_2} d(x_a, x_b)$$

3) Average Linkage

Σύμφωνα με την μέθοδο Average Linkage η απόσταση μεταξύ δύο συστάδων C_1, C_2 ορίζεται ως:

$$d(C_1, C_2) = \frac{\sum_{x_a \in C_1} \sum_{x_b \in C_2} d(x_a, x_b)}{N_{C_1}, N_{C_2}} \ ,$$

όπου N_{C_1}, N_{C_2} το πλήθος των συστάδων C_1, C_2 αντίστοιχα.

4.1.5 Connection

Η διαδικασία ολοκληρώνεται με την κατασκευή του τοπολογικού δικτύου αξιοποιώντας τα πλεονάζοντα **data points**. Οι κόμβοι του τοπολογικού δικτύου αντιπροσωπεύουν σύνολα όμοιων data points, ενώ οι έδρες συνδέουν κόμβους με κοινά σημεία. Εάν ο κόμβος δεν περιέχει κοινά σημεία με άλλον κόμβο, παραμένει **μονοσύνολο**. Τέλος, συνηθίζεται ο χρωματισμός του δικτύου από τις filter values, με σκοπό την καλύτερη κατανόηση του συνόλου των δεδομένων.

4.2 Εφαρμογή

Με σκοπό να γίνει ευκολότερα αντιληπτός ο τρόπος λειτουργίας του αλγορίθμου mapper, σ' αυτή την ενότητα πραγματοποιείται εφαρμογή της μεθόδου σε σύνολα δεδομένων που παράγονται από δείγματα αντικειμένων με γνωστή τοπολογία, σε 3D σχήματα καθώς και στο σύνολο δεδομένων *'Iris flower data set'*.

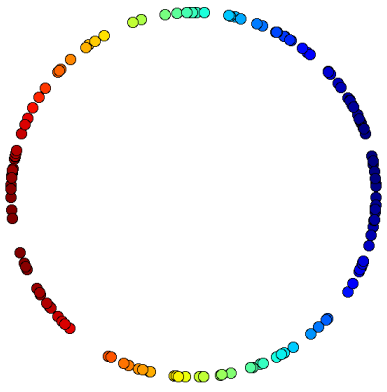
Για την επίτευξη του παραπάνω στόχου χρησιμοποιήθηκε η γλώσσα προγραμματισμού ανοιχτού κώδικα Python, και συγκεκριμένα το πακέτο Python Mapper (PM). Το εν λόγω πακέτο επεξεργάζεται τα δεδομένα με βάση το δοσμένο φίλτρο, τη μετρική, το κάλυμμα και τη μέθοδο ομαδοποίησης, και εξάγει ως αποτέλεσμα το τοπολογικό δίκτυο χρωματισμένο κατάλληλα με τις τιμές του filter value. Εναλλακτικά, μπορεί να χρησιμοποιηθεί η γλώσσα προγραμματισμού R και συγκεκριμένα τα πακέτα "TDA_mapper" και "mappeR".

4.2.1 Εφαρμογή σε σύνολα δεδομένων με γνωστή τοπολογία

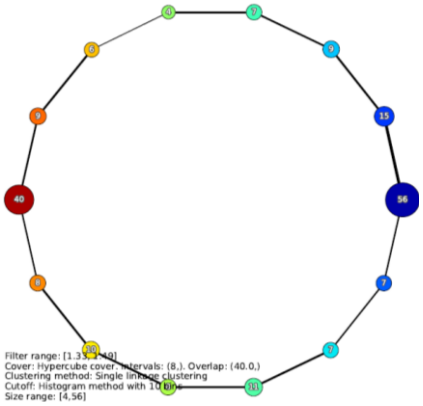
Κύκλος

Έστω ένα τυχαίο δείγμα 150 δεδομένων μέσα από το σύνολο $\{(\cos(t), \sin(t)) \,/\, t \in [0, 2\pi]\}$. Για τη δημιουργία του PCD θεωρούμε το σύνολο των δεδομένων ως υποσύνολο του χώρου $(\mathbb{R}^2, d_{\mathbb{R}})$, όπου $d_{\mathbb{R}}$ η Ευκλείδεια μετρική.

Παρακάτω, παρουσιάζεται το PCD χρωματισμένο κατάλληλα με βάση την filter function (Σχήμα 4.2.1.1 (α)) καθώς και το τοπολογικό δίκτυο (Σχήμα 4.2.1.1 (β)) που προκύπτει μετά από εφαρμογή του αλγορίθμου mapper επιλέγοντας ως φίλτρο τη συνάρτηση εκκεντρότητας με εκθέτη 2, ως τύπο καλύμματος το uniform 1 – d 8 διαστημάτων και ποσοστό αλληλοεπικάλυψης 40, ενώ ως μέθοδος ομαδοποίησης επιλέχθηκε η single linkage.



Σχήμα 4.2.1.1 (α)

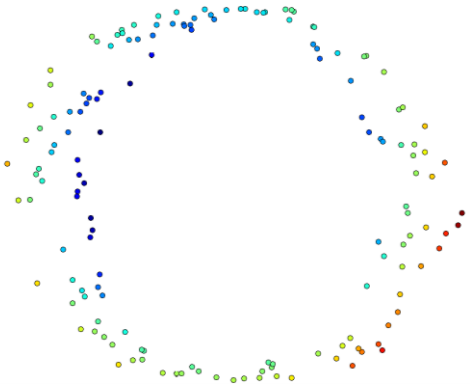


Σχήμα 4.2.1.1 (β)

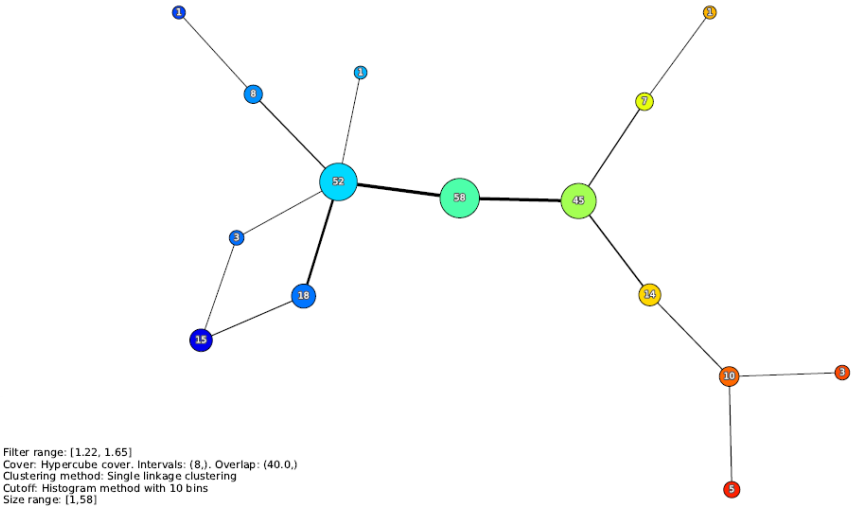
Κύκλος με θόρυβο

Έστω ένα τυχαίο δείγμα 150 δεδομένων μέσα από το σύνολο $\{(\cos(t) + \frac{1}{2} * s, \sin(t)) \,/\, t \in [0, 2\pi],\ s \in [0, 1]\}$. Για τη δημιουργία του PCD θεωρούμε το σύνολο των δεδομένων ως υποσύνολο του χώρου $(\mathbb{R}^2, d_{\mathbb{R}})$, όπου $d_{\mathbb{R}}$ η Ευκλείδεια μετρική.

Παρακάτω, παρουσιάζεται το PCD χρωματισμένο κατάλληλα με βάση την filter function (Σχήμα 4.2.1.2 (α)) καθώς και το τοπολογικό δίκτυο (Σχήμα 4.2.1.2 (β)) που προκύπτει μετά από εφαρμογή του αλγορίθμου mapper επιλέγοντας ως φίλτρο τη συνάρτηση εκκεντρότητας με εκθέτη 2, ως τύπο καλύμματος το uniform 1 – d 8 διαστημάτων και ποσοστό αλληλοεπικάλυψης 40, ενώ ως μέθοδος ομαδοποίησης επιλέχθηκε η single linkage.



Σχήμα 4.2.1.2 (α)



Σχήμα 4.2.1.2 (β)

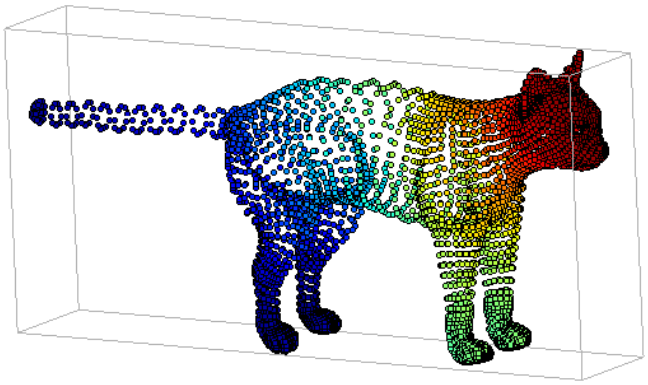
Παρατηρείται ότι παρά το θόρυβο που προστέθηκε, με χρήση της μεθόδου είναι δυνατό να εντοπιστεί ο βρόγχος.

4.2.2 Εφαρμογή σε 3D σχήματα

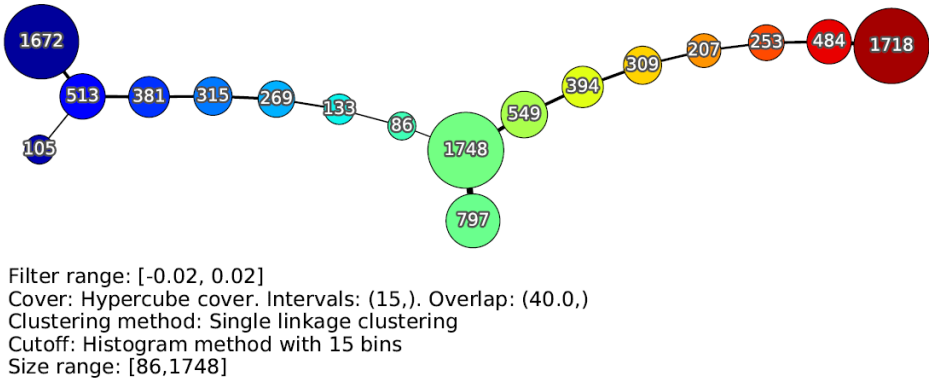
Στη συνέχεια, εφαρμόζουμε τον αλγόριθμο mapper σε 3D σχήματα, κάνοντας χρήση έτοιμων συνόλων δεδομένων που παρέχει η Python.

Στα ακόλουθα σχήματα εμφανίζεται το PCD (7207 τυχαία σημεία) χρωματισμένο κατάλληλα με βάση την filter function (Σχήμα 4.2.2.1 (α)) καθώς και το τοπολογικό δίκτυο

(Σχήμα 4.2.2.1 (β)) που προκύπτει μετά από εφαρμογή του αλγορίθμου mapper επιλέγοντας την Ευκλείδεια μετρική, ως φίλτρο το SVD, ως τύπο καλύμματος το uniform 1 – d 15 διαστημάτων και ποσοστό αλληλοεπικάλυψης 40, ενώ ως μέθοδος ομαδοποίησης επιλέχθηκε η single linkage.



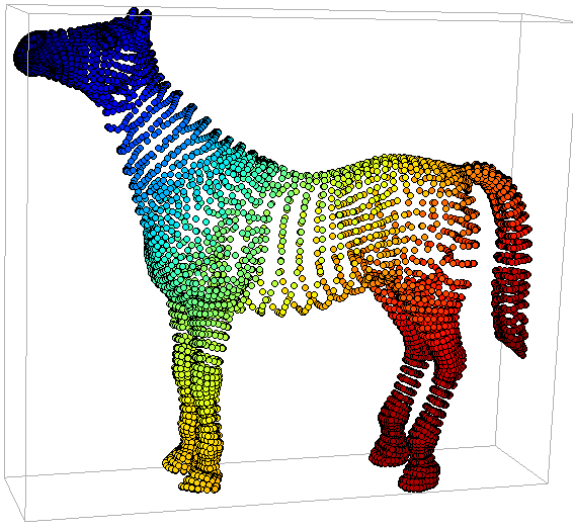
Σχήμα 4.2.2.1(α) Δειγματοληψία από γάτα



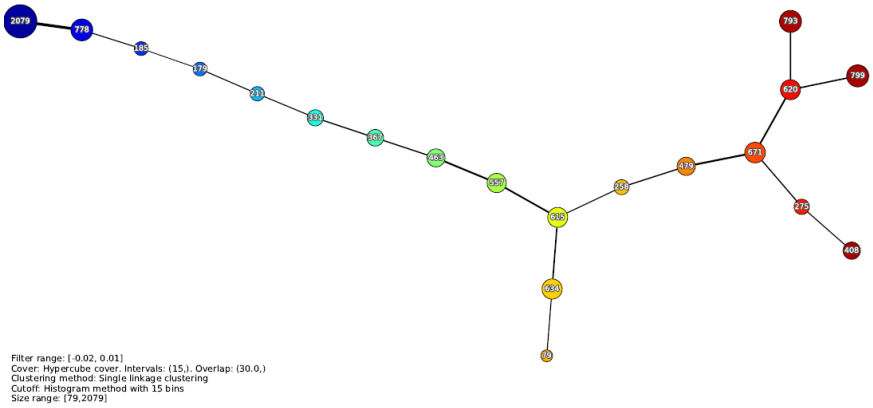
Σχήμα 4.2.2.1(β) Δειγματοληψία από γάτα

Στη συνέχεια εμφανίζεται το PCD (8431 τυχαία σημεία) χρωματισμένο κατάλληλα με βάση την filter function (Σχήμα 4.2.2.2 (α)) καθώς και το τοπολογικό δίκτυο (Σχήμα 4.2.2.2 (β)) που προκύπτει μετά από εφαρμογή του αλγορίθμου mapper επιλέγοντας την ευκλείδεια μετρική ως φίλτρο τη συνάρτηση εκκεντρότητας με εκθέτη 2, ως τύπο καλύμματος το uniform

1 – d 8 διαστημάτων και ποσοστό αλληλοεπικάλυψης 40, ενώ η ομαδοποίηση γίνεται με την μέθοδο single linkage.



Σχήμα 4.2.2.2 (α) Δειγματοληψία από Άλογο

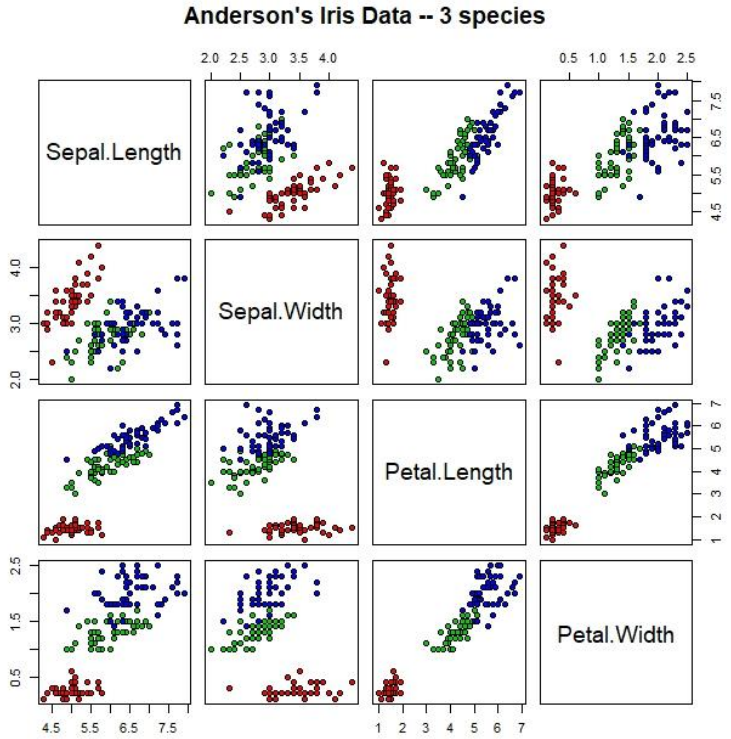


Σχήμα 4.2.2.2 (β)Δειγματοληψία από Άλογο

Παρατηρούμε ότι με χρήση του αλγορίθμου mapper είναι δυνατό να εντοπιστεί με ακρίβεια ο σκελετός των σχημάτων.

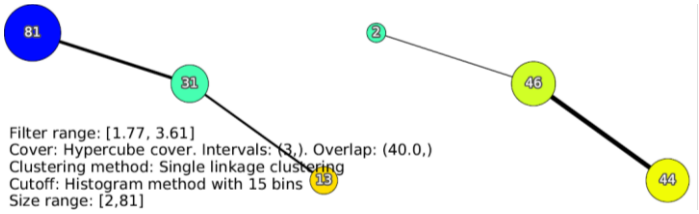
4.2.3 Εφαρμογή στο Iris Data Set

Στα παραπάνω παραδείγματα παρουσιάστηκε η αποτελεσματικότητα του mapper στην αναγνώριση σχήματος και χαρακτηριστικών παρά το θόρυβο. Ωστόσο, υπάρχει η δυνατότητα εφαρμογής και σε περιπτώσεις που είναι επιθυμητή η ομαδοποίηση των δεδομένων. Ένα χαρακτηριστικό σύνολο δεδομένων στο οποίο δοκιμάζονται τεχνικές ομαδοποίησης είναι το *Iris flower data* set.

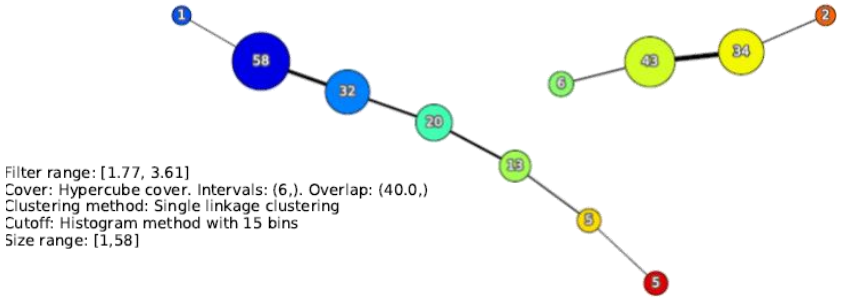


Σχήμα 4.2.3.1 Προβολές των Iris data σε κάθε επίπεδο του $\mathbb{R}^3$. Διαφορετικά χρώματα αντιστοιχούν σε διαφορετικά είδη των iris flowers.

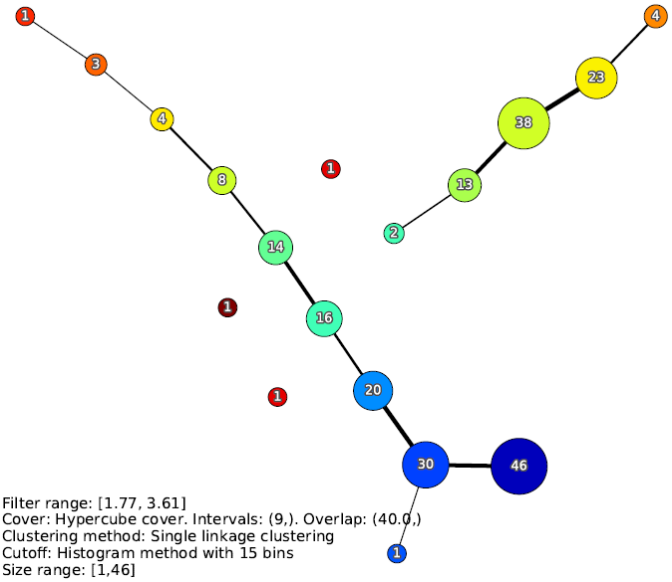
Παρακάτω, παρουσιάζονται (Σχήμα 4.2.3.2 (α),(β) και (γ)) τα τοπολογικά δίκτυα που προκύπτουν μετά από εφαρμογή του αλγορίθμου mapper επιλέγοντας την Ευκλείδεια μετρική, ως φίλτρο τη συνάρτηση εκκεντρότητας με εκθέτη 2, ως τύπο καλύμματος το uniform 1 – d 3, 6 και 9 διαστημάτων αντίστοιχα και ποσοστό αλληλοεπικάλυψης 40, ενώ η ομαδοποίηση γίνεται με την μέθοδο single linkage.



Σχήμα 4.2.3.2 (α)

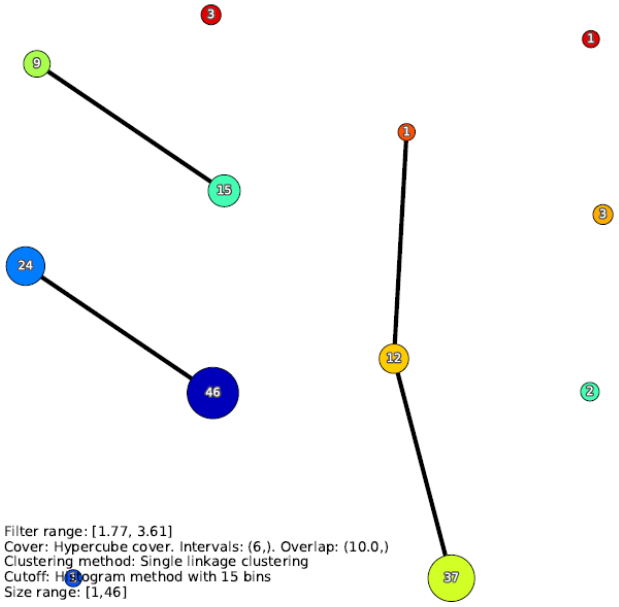


Σχήμα 4.2.3.2 (β)

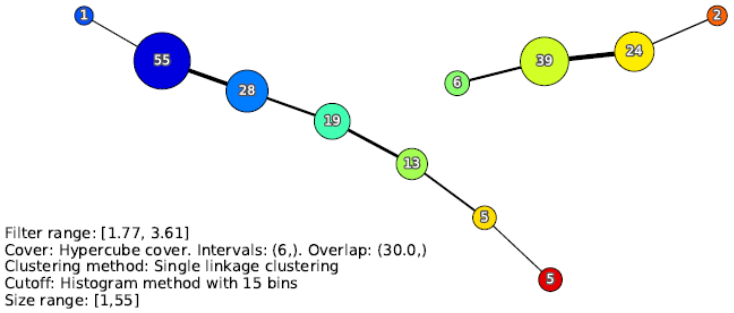


Σχήμα 4.2.3.2 (γ)

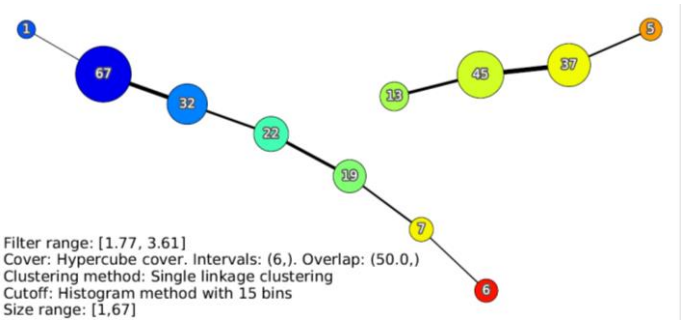
Στη συνέχεια εφαρμόζουμε τον αλγόριθμο mapper διατηρώντας σταθερό τον αριθμό των διαστημάτων, και ίσο με 6, ενώ μεταβάλουμε το ποσοστό αλληλοεπικάλυψης σε 10%, 30% και 50% αντίστοιχα.



Σχήμα 4.2.3.3 (α)



Σχήμα 4.2.3.3 (β)



Σχήμα 4.2.3.3 (γ)

Παρατηρούμε ότι παίρνοντας μικρότερο αριθμό διαστημάτων δημιουργούνται λιγότεροι κόμβοι, ενώ τα μικρότερα ποσοστά αλληλοεπικάλυψης δημιουργούν λιγότερες συνδέσεις στο τελικό δίκτυο.

Όσον αφορά την ομαδοποίηση των δεδομένων, από τη συνολική εικόνα που παρέχουν τα δίκτυα, είναι εμφανές ότι προκύπτουν δύο clusters. Με χρήση του πακέτου mappeR στην R βρίσκουμε ότι εμφανίζεται το είδος Iris setosa ως ένα cluster, ενώ δεν υπάρχει διαχωρισμός μεταξύ των άλλων δύο.

Βιβλιογραφία

[1] Adler R. (2014). TOPOS, and why you should care about it. Bulletin of the American Mathematical Society, 45(2), 4-5.

[2] Armstrong, M. A. (2013). Basic topology. Springer Science & Business Media.

[3] Balakrishnan, S., Fasy, B., Lecci, F., Rinaldo, A., Singh, A., & Wasserman, L. (2013). Statistical inference for persistent homology.

[4] Bott, R., & Tu, L. W. (2013). Differential forms in algebraic topology (Vol. 82). Springer Science & Business Media.

[5] Carlsson, G., Ishkhanov, T., De Silva, V., & Zomorodian, A. (2008). On the local behavior of spaces of natural images. International journal of computer vision, 76(1), 1-12.

[6] Carlsson, G. (2009). Topology and data. Bulletin of the American Mathematical Society, 46(2), 255-308.

[7] Chazal, F., De Silva, V., & Oudot, S. (2014). Persistence stability for geometric complexes. Geometriae Dedicata, 173(1), 193-214.

[8] Chazal, F., Fasy, B., Lecci, F., Michel, B., Rinaldo, A., & Wasserman, L. (2015, June). Subsampling methods for persistent homology. In International Conference on Machine Learning (pp. 2143-2151).

[9] Chazal, F., De Silva, V., Glisse, M., & Oudot, S. (2016). The structure and stability of persistence modules. Springer.

[10] Collins, A., Zomorodian, A., Carlsson, G., & Guibas, L. J. (2004). A barcode shape descriptor for curve point cloud data. Computers & Graphics, 28(6), 881-894.

[11] Dequeant, M. L., Ahnert, S., Edelsbrunner, H., Fink, T. M., Glynn, E. F., Hattem, G. & Pachter, L. (2008). Comparison of pattern detection methods in microarray time series of the segmentation clock. PLoS One, 3(8), e2856.

[12] De Silva, V. (2004). Geometry and Topology of Point-Cloud Data Sets: A Statement Of My Research Interests.

[13] Ding, W. (2017). Experiment of Mapper Algorithm on High-Dimensional Data in Microseismic Monitoring (Doctoral dissertation).

[14] Edelsbrunner, H., Letscher, D., & Zomorodian, A. (2002). Topological persistence and simplification. Discrete and Computational Geometry 28:511–533.

[15] Edelsbrunner, H., & Harer, J. (2008). Persistent Homology - a Survey. Contemporary Mathematics Series, vol. 453, pp. 257–282. Amer Mathematical Soc., Providence.

[16] Farahani, R. Z., & Hekmatfar, M. (Eds.). (2009). Facility location: concepts, models, algorithms and case studies. Springer Science & Business Media.

[17] Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., & Singh, A. (2014). Confidence sets for persistence diagrams. The Annals of Statistics, 42(6), 2301-2339.

[18] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of human genetics, 7(2), 179-188.

[19] Ghrist, R. (2008). Barcodes: the persistent topology of data. Bulletin of the American Mathematical Society, 45(1), 61-75.

[20] Hatcher, A. (2002). Algebraic Topology. Cambridge University Press, Cambridge.

[21] Heo, G., Gamble, J., & Kim, P. T. (2012). Topological analysis of variance and the maxillary complex. Journal of the American Statistical association, 107(498), 477-492.

[22] Kasson, P. M., Zomorodian, A., Park, S., Singhal, N., Guibas, L. J., & Pande, V. S. (2007). Persistent voids: a new structural metric for membrane fusion. Bioinformatics, 23(14), 1753-1759.

[23] Kim, H. E. (2015). Evaluating Ayasdi's Topological Data Analysis for Big Data (Master's thesis), Offenburg University of Applied Sciences.

[24] Kraft, R. (2016). Illustrations of Data Analysis Using the Mapper Algorithm and Persistent Homology (Master's thesis), KTH Royal Institute of Technology.

[25] Lum, P. Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., ... & Carlsson, G. (2013). Extracting insights from the shape of complex data using topology. Scientific reports, 3, srep01236.

[26] Michel, B. (2015). A Statistical Approach to Topological Data Analysis (Doctoral dissertation, UPMC Université Paris VI).

[27] Munkres, J. R. (1984) Elements of Algebraic Topology. Addison- Wesley, Redwood City, California.

[28] Nicolau M., Levine A.J., Carlsson G. (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proceedings of the National Academy of Sciences of the United States of America 108(17), 7265–7270.

[29] Peeva, I. (2010). Graded syzygies (Vol. 14). Springer Science & Business Media.

[30] Reininghaus, J., Huber, S., Bauer, U., & Kwitt, R. (2015). A stable multi-scale kernel for topological machine learning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4741-4748).

[31] Rotman, J. J. (1988) An introduction to algebraic topology, Springer-Verlag, New York.

[32] Singh, G., Memoli, F., Carlsson, G. (2007). Topological methods for the analysis of high dimensional data sets and 3d object recognition. In: Botsch, M., Pajarola, R. (eds.) Eurographics Symposium on Point-Based Graphics, vol. 22, pp. 91–100. Euro Graphics.

[33] Singh, G., Memoli, F., Ishkhanov, T., Sapiro, G., Carlsson, G., & Ringach, D. L. (2008). Topological analysis of population activity in visual cortex. Journal of vision, 8(8), 11-11.

[34] Stovner, R. B. (2012). On the mapper algorithm: A study of a new topological method for data analysis (Master's thesis, Institutt for matematiske fag).

[35] Tierny, J. (2017). Introduction to topological data analysis (Doctoral dissertation). France. <cel-01581941>

[36] Wasserman, L. (2016). Topological data analysis. Annual Review of Statistics and Its Application.

[37] Zomorodian, A., & Carlsson, G. (2005). Computing persistent homology. Discrete & Computational Geometry, 33(2), 249-274.

[38] Zomorodian, A. J. (2005). Topology for computing (Vol. 16). Cambridge University Press.

[39] Zomorodian, A. (2012). Topological data analysis. Advances in applied and computational topology, 70, 1-39.