ΤΟ ΘΕΩΡΗΤΙΚΟ ΠΛΑΙΣΙΟ ΤΗΣ ΤΟΠΟΛΟΓΙΚΗΣ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ

Πτυχιακή εργασία στο πλαίσιο του Π.Π.Σ. του τμήματος
Στατιστικής και Αναλογιστικών - Χρηματοοικονομικών Μαθηματικών

Ελένη Θεοχάρη Παϊπούρογλου

Το θεωρητικό πλαίσιο της Τοπολογικής Ανάλυσης Δεδομένων

Ελένη Θεοχάρη Παϊπούρογλου

Ιούνιος 2022

Περιεχόμενα

Τριμελής Επιτροπή:

Στυλιανός Ξανθόπουλος, Αναπληρωτής Καθηγητής στο Τμήμα Στατιστικής και Αναλογιστικών - Χρηματοοικονομικών Μαθηματικών του Πανεπιστημίου Αιγαίου (Επιβλέπων)

Παντελής Λάππας, Επίκουρος Καθηγητής στο Τμήμα Στατιστικής και Αναλογιστικών - Χρηματοοικονομικών Μαθηματικών του Πανεπιστημίου Αιγαίου

Ελευθέριος Ταχτσής, Καθηγητής στο Τμήμα Στατιστικής και Αναλογιστικών - Χρηματοοικονομικών Μαθηματικών του Πανεπιστημίου Αιγαίου

Κεφάλαιο 1

Λίγα λόγια για την Ανάλυση Δεδομένων

Η ανάλυση δεδομένων είναι μία διαδικασία επιθεώρησης, καθαρισμού, μετασχηματισμού και μοντελοποίησης δεδομένων. Σκοπός είναι η εύρεση χρήσιμης πληροφορίας, η εξαγωγή συμπερασμάτων σχετικών με τα δεδομένα και με αυτόν τον τρόπο η υποστήριξη της διαδικασίας λήψης αποφάσεων. Η ανάλυση δεδομένων έχει πολλές όψεις και προσεγγίσεις και αξιοποιείται μεγάλη ποικιλία τεχνικών. Χρησιμοποιείται ευρέως σε διάφορους τομείς όπως οι θετικές, οικονομικές και κοινωνικές επιστήμες. [0]

Σχετικά με τις εφαρμογές της στην Στατιστική, η ανάλυση δεδομένων χωρίζεται στις εξής τρεις κατηγορίες:

- **Περιγραφική Στατιστική**
- **Διερευνητική Ανάλυση Δεδομένων (Exploratory Data Analysis (EDA))**, που επικεντρώνεται στην ανακάλυψη νέων χαρακτηριστικών των δεδομένών
- **Επιβεβαιωτική Ανάλυση Δεδομένων (Confirmatory Data Analysis (CDA))**, που εστιάζει στην απόρριψη ή μη προϋπαρχόντων υποθέσεων

Από την άλλη μεριά, έχουμε:

- **Ανάλυση Προγνωστικών (predictive analysis)**, στοχεύει στην εφαρμογή στατιστικών μοντέλων για την πρόβλεψη ή κατηγοριοποίηση
- **Ανάλυση Κειμένου (text analytics)**, εφαρμόζει γλωσσικές, στατιστικές και κατασκευαστικές τεχνικές για την εξόρυξη και ταξινόμηση δεδομένων από πηγές κειμένου, που αποτελούν ένα είδος αδόμητων δεδομένων.

Όλα τα παραπάνω είναι μορφές της ανάλυσης δεδομένων. Ιστορικά η ενοποίηση δεδομένων (Data Integration) είναι ο πρόγονος της ανάλυσης δεδομένων η οποία συνδέεται πολύ στενά με την οπτικοποίηση δεδομένων (Data Visualization) και τη διασπορά δεδομένων (Data Dissemination). [0]

1.1 Η διαδικασία της ανάλυσης δεδομένων

Η ανάλυση ως όρος χρησιμοποιείται για να καταδείξει τη διαίρεση του όλου στα συστατικά του στοιχεία, από τα οποία προσπαθούμε να "καταλάβουμε" τις ιδιότητες που χαρακτηρίζουν τα δεδομένα μας. Η ανάλυση δεδομένων, είναι μία διαδικασία κατά την οποία επεξεργαζόμαστε ακατέργαστα δεδομένα με στόχο από αυτά να εξάγουμε χρήσιμη πληροφορία. Με τον όρο χρήσιμη πληροφορία εννοούμε ότι αυτή σε μεγάλο ή μικρό βαθμό θα συμβάλει στη λήψη αποφάσεων από τους χρήστες. Ως διαδικασία προέκυψε από την ανάγκη να απαντηθούν ερωτήματα, να ελεγχθούν υποθέσεις ή να απορριφθούν θεωρίες, με βάση τα όσα μπορούμε να γνωρίσουμε από ένα σύνολο δεδομένων. [0]

Σύμφωνα με τον John Tukey, το 1961, η ανάλυση δεδομένων αποτελείται από:

¨Διαδικασίες για την ανάλυση δεδομένων, τεχνικές για την ερμηνεία των αποτελεσμάτων, τρόπους σχεδιασμού της συλλογής δεδομένων έτσι ώστε η ανάλυση να γίνει πιο εύκολη, πιο ακριβής, (είτε με την έννοια ότι τα δεδομένα είναι πιο κοντά στην πραγματικότητα με κριτήριο την αντίστοιχη τιμή p-value, είτε με την έννοια ότι είναι πιο κοντά μεταξύ τους) και ολόκληρο το "οπλοστάσιο" και τα αποτελέσματα της (μαθηματικής) στατιστικής που εφαρμόζει στην ανάλυση δεδομένων' [0]

Παρακάτω θα αναφερθούμε στις διάφορα στάδια της ανάλυσης δεδομένων που μπορούμε να διακρίνουμε. Τα στάδια αυτά επαναλαμβάνονται και με αυτόν τον τρόπο η ανατροφοδότηση που παίρνουμε από τις τελευταίες φάσεις οδηγεί σε επιπλέον επεξεργασία στις αρχικές. [0]

Προδιαγραφές Δεδομένων

Τα δεδομένα είναι απαραίτητα ως καταχωρήσεις στην ανάλυση, που γίνεται ποιο εξειδικευμένη με βάση τις προδιαγραφές που θέτουν οι αναλυτές ή εκείνοι που θα αναλύσουν το τελικό αποτέλεσμα. Ο φορέας πάνω στον οποίο γίνεται η συλλογή των δεδομένων θα ονομάζεται πειραματική μονάδα. Επιπλέον θα προσδιορίζουμε και θα επιλέγουμε συγκεκριμένες μεταβλητές που αφορούν στον πληθυσμό, χωρίς να μας δημιουργείται πρόβλημα είτε αυτές είναι ποσοτικές είτε ποιοτικές. [0]

Συλλογή Δεδομένων

Τα δεδομένα συλλέγονται από διάφορες πηγές. Οι απαιτήσεις μπορεί να διατυπώνονται από τους αναλυτές στους συλλέκτες των δεδομένων. Συλλέκτες δεδομένων μπορεί να είναι για παράδειγμα το προσωπικό τεχνολογίας πληροφοριών κάποιας εταιρίας. Τα δεδομένα μπορεί να συλλεχθούν ακόμα και από αισθητήρες στο περιβάλλον, όπως κάμερες κυκλοφορίας, συσκευές ηχογράφησης κλπ. Επιπλέον, μπορούμε να πάρουμε δεδομένα μέσω συνεντεύξεων, λήψεων από διαδικτυακές πηγές ή από διάφορα έγγραφα. [0]

Επεξεργασία Δεδομένων

Αφού συλλέξουμε τα δεδομένα, θα πρέπει να τα επεξεργαστούμε ή με κάποιο τρόπο να τα οργανώσουμε ώστε να τα αναλύσουμε. Ένας τέτοιος τρόπος οργάνωσης των δεδομένων μπορεί να είναι η γνωστή διάταξη σε γραμμές και στήλες, δηλαδή σε πίνακες, συχνά με τη χρήση υπολογιστικών φύλλων ή στατιστικών λογισμικών. [0]

Καθαρισμός Δεδομένων

Μετά την επεξεργασία που έχουν υποστεί τα δεδομένα μπορεί να είναι ατελή, να εμφανίζονται παραπάνω από μία φορές ή να περιέχουν ακραίες παρατηρήσεις. Από αυτά τα προβλήματα που εμφανίζονται κατά την εισαγωγή και ταξινόμηση των δεδομένων, είναι που αναδεικνύουν την ανάγκη καθαρισμού των δεδομένων. Ο καθαρισμός είναι η διαδικασία που ακολουθούμε για να προλάβουμε ή να διορθώσουμε αυτά τα λάθη. Στις συνήθεις πρακτικές συγκαταλέγονται αντιστοιχίσεις καταγραφών, εξακρίβωση της ακρίβειας των δεδομένων, αξιολόγηση της συνολικής ποιότητας των υπαρχόντων δεδομένων, η διαγραφή των καταχωρήσεων που εμφανίζονται για δεύτερη φορά και η κατάτμηση σε στήλες. [0]

Τέτοια προβλήματα με τα δεδομένα μπορούν να εντοπιστούν χρησιμοποιώντας διάφορες αναλυτικές τεχνικές. Στα χρηματοοικονομικά, για παράδειγμα, τα σύνολα συγκεκριμένων μεταβλητών μπορούν να συγκριθούν με διαφορετικά δημοσιευμένα δεδομένα που θεωρούνται αξιόπιστα. [0] Είναι πιθανό να μελετήσουμε και ασυνήθιστες ποσότητες, με την έννοια ότι μπορεί να βρίσκονται πολύ παραπάνω ή αντίστοιχα πολύ παρακάτω από κάποια προκαθορισμένα όρια (κατώφλια). Υπάρχουν διάφοροι τρόποι καθαρισμού δεδομένων, που εξαρτώνται από το είδος των μεταβλητών. Οι μέθοδοι για τον εντοπισμό ακραίων παρατηρήσεων που αφορούν σε ποσοτικά δεδομένα, χρησιμοποιούνται για να

εξαιρέσουμε τα δεδομένα που έχουν τη μεγαλύτερη πιθανότητα να οδηγήσουν σε εσφαλμένα συμπεράσματα.

Διερευνητική Ανάλυση Δεδομένων

Εφόσον έχουμε ολοκληρώσει τον καθαρισμό των δεδομένων, μπορούμε να προχωρήσουμε στην ανάλυσή τους. Οι αναλυτές δεδομένων αξιοποιούν πληθώρα τεχνικών, στις οποίες αναφερόμαστε ως διερευνητική ανάλυση δεδομένων, για να καταφέρουν να αποκωδικοποιήσουμε την πληροφορία που περιέχεται στα δεδομένα που έχουν συλλεχθεί. Η διερεύνηση των δεδομένων είναι διαδικασία που μπορεί να οδηγήσει εκ νέου στον καθαρισμό δεδομένων ή στη δημιουργία καινούριων απαιτήσεων σχετικών με τα δεδομένα. Εν τέλει, η όλη διαδικασία επάγει επαναλήψεις κάποιων διαδικασιών. Αξιοποιώντας την περιγραφική στατιστική, στοιχεία όπως η μέση τιμή ή η διάμεσος, μπορούν να λειτουργήσουν βοηθητικά στην κατανόηση των δεδομένων. Μία άλλη συχνά αξιοποιούμενη τεχνική είναι η οπτικοποίηση, που δίνει τη δυνατότητα στον αναλυτή να εξετάσει τα δεδομένα ως γραφήματα με σκοπό να εξάγει επιπρόσθετα συμπεράσματα σχετικά με τα δεδομένα. [0]

Μοντελοποίηση και Αλγόριθμοι

Διάφοροι αλγόριθμοι, μπορούν να εφαρμοστούν στα δεδομένα, με σκοπό να προσδιοριστούν οι σχέσεις μεταξύ των μεταβλητών, όπως για παράδειγμα η συσχέτιση και η αιτιότητα. Σε γενικές γραμμές, κατασκευάζονται μοντέλα με τέτοιο τρόπο ώστε να εκτιμούν την επιρροή που ασκούν σε μία ορισμένη μεταβλητή, άλλες μεταβλητές που περιέχονται στο σύνολο δεδομένων που διαθέτουμε, συνυπολογίζοντας και κάποιο υπολειπόμενο σφάλμα που εξαρτάται από την ακρίβεια του μοντέλου που εφαρμόζουμε. (Σχηματικά: Δεδομένα = Μοντέλο + Σφάλμα) [0]

Η διερευνητική στατιστική, περιλαμβάνει την αξιοποίηση τεχνικών, οι οποίες ποσοτικοποιούν τις σχέσεις μεταξύ ορισμένων μεταβλητών. Η ανάλυση παλινδρόμησης για παράδειγμα, μπορεί να χρησιμοποιηθεί σ΄ αυτήν την κατεύθυνση, όπου το εκάστοτε μοντέλο ¨οφείλει' να ελαχιστοποιεί το σφάλμα (τα κατάλοιπα στην προκειμένη περίπτωση), και να εξηγεί με την επίδραση της μεταβολής της ανεξάρτητης μεταβλητής στην εξαρτημένη μεταβλητή. Οι αναλυτές καλούνται, σε πολλές περιπτώσεις, να κατασκευάσουν μοντέλα που περιγράφουν τα δεδομένα, με σκοπό να απλοποιήσουν την ανάλυση και να καταλήξουν σε αποτελέσματα. [0]

Προϊόν Δεδομένων

Ένα προϊόν δεδομένων, είναι μία εφαρμογή στην οποία καταχωρούνται δεδομένα (inputs) και εξάγει αποτελέσματα (outputs), τα οποία επιστρέφει στο περιβάλλον. Είναι δυνατό να βασίζεται σε κάποιο μοντέλο ή αλγόριθμο. Για παράδειγμα, ένα προϊόν δεδομένων είναι μία εφαρμογή στην οποία καταχωρείται το ιστορικό αγορών των πελατών μίας επιχείρησης και εξάγει εξατομικευμένες προτάσεις αγορών για τους καταναλωτές. [0]

Επικοινωνία

Αφού ολοκληρωθεί η ανάλυση των δεδομένων, τα αποτελέσματά της περιλαμβάνονται σε διάφορων μορφών αναφορές έτσι ώστε να ικανοποιούνται τα αιτήματα των αναλυτών. Ωστόσο, η ανατροφοδότηση που λαμβάνουν οι αναλυτές συχνά οδηγεί σε επιπρόσθετη ανάλυση. Κατά αυτόν τον τρόπο ο κύκλος της ανάλυσης είναι επαναληπτικός. [0]

Κατά τον προσδιορισμό του τρόπου επικοινωνίας των αποτελεσμάτων, οι αναλυτές είναι πιθανό να αξιοποιήσουν τη δυνατότητα οπτικοποίησης των δεδομένων με χρήση ποικίλων τεχνικών, με σκοπό να επικοινωνήσουν τα αποτελέσματα της ανάλυσης σύντομα και με σαφήνεια. Η οπτικοποίηση των δεδομένων στηρίζεται στην απεικόνιση πληροφοριών, με γραφήματα, πίνακες κτλ. [0]

1.2 Εισαγωγή στην Τοπολογική Ανάλυση Δεδομένων

Εν γένει στην Ανάλυση Δεδομένων αποτελεί πρόκληση η εξαγωγή συμπερασμάτων από σύνολα που είναι πολυδιάστατα, ατελή και έχουν θόρυβο. Η Τοπολογική Ανάλυση Δεδομένων φαίνεται να δίνει μία απάντηση σε αυτά τα προβλήματα. Όπως προδίδει και η ονομασία της, το νέο που έχει αυτή να εισφέρει είναι η αξιοποίηση τεχνικών της τοπολογίας επί του συνόλου δεδομένων που έχουμε προς ανάλυση. Το γενικό πλαίσιο στο οποίο κινείται δίνει τη δυνατότητα, ανάλογα με τη μετρική που επιλέγουμε, να ελαχιστοποιεί τη διάσταση και την ευρωστία του θορύβου, να γίνεται η ανάλυση πιο ευπροσάρμοστη σε νέα μαθηματικά εργαλεία.

Η Τοπολογική Ανάλυση Δεδομένων βασίζεται στην ιδέα ότι το σχήμα του συνόλου δεδομένων περιέχει κάποια σχετική πληροφορία. Στην πραγματικότητα, τα πολυδιάστατα δεδομένα είναι συνήθως διασκορπισμένα, και παρουσιάζουν ιδιότητες μόνο σε λιγότερες διαστάσεις. Αυτό αποτελεί ένα από τα ζητήματα στα οποία καλείται να δώσει απαντήσεις η Τοπολογική Ανάλυση Δεδομένων. Για να γίνει λίγο πιο εύληπτη αυτή η προσέγγιση, θεωρούμε ότι έχουμε ένα απλό σύστημα θηρευτή – θηράματος Lotka – Voltera. Η τροχιά του δημιουργεί ένα κλειστό κύκλο σε ένα χώρο καταστάσεων. Η Τοπολογική Ανάλυση Δεδομένων παρέχει τα εργαλεία για την αναγνώριση και ποσοτικοποίηση αυτή της περιοδικής κίνησης.

Η διαδικασία της Τοπολογικής Ανάλυσης Δεδομένων (όπως και στην Ανάλυση Δεδομένων) προϋποθέτει την επιλογή κάποιων παραμέτρων και ως εκ τούτου γίνεται πιο περίπλοκη. Η εύρεση των σωστών παραμέτρων είναι μία αρκετά δύσκολη διαδικασία, η οποία στην περίπτωση μας προσεγγίζεται μέσω της Εμμένουσας Ομολογίας. Το σκεπτικό της Εμμένουσας Ομολογίας είναι να χρησιμοποιήσουμε την πληροφορία που λαμβάνουμε από όλες τις τιμές που μπορεί να δοθούν στις παραμέτρους και να κωδικοποιήσουμε αυτήν την τεράστια σε ποσότητα πληροφορία σε μία πιο εύκολα κατανοητή και ερμηνεύσιμη φόρμα. Ουσιαστικά, με τη Τοπολογική Ανάλυση Δεδομένων υπάρχει μία μαθηματική ερμηνεία, όπου η πληροφορία είναι μία ομάδα ομολογίας. Σε γενικές γραμμές, υποθέτουμε ότι τα χαρακτηριστικά που εμμένουν σε μία ευρεία κλίμακα παραμέτρων είναι και τα "πραγματικά" χαρακτηριστικά του συνόλου δεδομένων. Ενώ τα χαρακτηριστικά που εμμένουν για ένα πιο περιορισμένο εύρος παραμέτρων θεωρούνται θόρυβος. Για την τελευταία παρατήρηση δεν υπάρχει σαφής θεωρητική απόδειξη.

Κεφάλαιο 2

Τοπολογική Ανάλυση Δεδομένων

Η Τοπολογική Ανάλυση Δεδομένων στοχεύει στην παροχή καλά θεμελιωμένων μαθηματικών, στατιστικών και αλγοριθμικών μεθόδων με τις οποίες θα συμβάλει στην ανάλυση και την εξαγωγή συμπερασμάτων εκμεταλλευόμενη τις περίπλοκες τοπολογικές και γεωμετρικές ιδιότητες που υποβόσκουν στα δεδομένα, οι οποίες μάλιστα συχνά αναπαρίστανται μέσω νεφών σημείων μέσα στον Ευκλείδειο ή και άλλους μετρικούς χώρους.

Η Τοπολογική Ανάλυση Δεδομένων βασίζεται κατά κύριο λόγο στην ιδέα ότι η τοπολογία και η γεωμετρία δίνουν τη δυνατότητα μίας προσέγγισης που οδηγεί σε εύρωστες ποιοτικές, και κάποιες φορές ποσοτικές πληροφορίες για τη δομή των δεδομένων (π.χ. Chazal (2017).

Η Τοπολογική Ανάλυση Δεδομένων (Topological Data Analysis (TDA)) είναι ένα σύγχρονο πεδίο που προέκυψε από διάφορες εργασίες στην εφαρμοσμένη αλγεβρική τοπολογία και την υπολογιστική γεωμετρία κατά την πρώτη δεκαετία του $21^{ου}$ αιώνα. Παρ΄ ότι υπήρξαν γεωμετρικές προσεγγίσεις στην ανάλυση δεδομένων πολύ παλαιότερα, η Τοπολογική Ανάλυση Δεδομένων ουσιατικά εμφανίστηκε στης καινοτόμες εργασίες των Edelbrunner et al. (2002) και Zomorodian and Carlsson (2005) στην Εμμένουσα Ομολογία, οι οποίες δημοσιεύτηκαν σε μια δημοσίευση που αποτελεί σημείο αναφοράς το 2009 (Carlsson (2009)).

Τα τελευταία χρόνια έχει γίνει σημαντική προσπάθεια να παρασχεθούν εύρωστες και αποδοτικές δομές δεδομένων και αλγόριθμοι για την Τοπολογική Ανάλυση Δεδομένων. Τα παραπάνω είναι πλέον προσβάσιμα και είναι εύκολο να χρησιμοποιηθούν μέσω βιβλιοθηκών, όπως η βιβλιοθήκη Gudhi, με τη χρήση των γλωσσών προγραμματισμού C++ και Python [Maria et al. (2014)] και του στατιστικού πακέτου R [Fasy et al. (2014a)]. Παρά την γρήγορη ανάπτυξη και εξέλιξη του αντικειμένου, η Τοπολογική Ανάλυση Δεδομένων ήδη παρέχει ένα σύνολο ολοκληρωμένων και αποτελεσματικών εργαλείων που μπορούν να αξιοποιηθούν σε συνδυασμό ή συμπληρωματικά με άλλα εργαλεία της ανάλυσης δεδομένων.

Οι Βασικές Υποθέσεις της Τοπολογικής Ανάλυσης Δεδομένων

Η Τοπολογική Ανάλυση Δεδομένων τα τελευταία χρόνια εξελίσσεται προς διάφορες κατευθύνσεις και πεδία εφαρμογής. Υπάρχει ήδη μεγάλη ποικιλία μεθόδων που έχουν προκύψει από τοπολογικές και γεωμετρικές προσεγγίσεις. Χωρίς να μπαίνουμε σε λεπτομερείς καταγραφές αυτών των προσεγγίσεων, θα παρουσιάσουμε σύντομα τους βασικούς όρους και προϋποθέσεις που τις χαρακτηρίζουν και έτσι εισάγουμε τις πρώτες υποθέσεις που είναι απαραίτητες για την κατανόηση της διαδικασίας που τυπικά ακολουθείται. Πιο συγκεκριμένα:

1. Τα εισαγόμενα δεδομένα υποθέτουμε ότι είναι ένα πεπερασμένο σύνολο σημείων, εφοδιασμένα με την έννοια κάποιας απόστασης ή ομοιότητας μεταξύ τους. Αυτή η απόσταση μπορεί να επάγεται από μία μετρική σε έναν περιβάλλοντα χώρο. Για παράδειγμα, μία τέτοια μετρική θα μπορούσε να είναι η Ευκλείδεια όταν τα δεδομένα είναι ενσωματωμένα στον Ευκλείδειο χώρο ($\mathbb{R}^d$). Θα μπορούσε, επίσης, να χρησιμοποιείται μία εσωτερική μετρική, ορισμένη από κάποιον πίνακα προσδιορισμού της απόστασης ανά ζεύγη. Ο ορισμός της μετρικής που εφαρμόζεται στα δεδομένα συνήθως αποτελεί δεδομένο που εισάγεται στο πρόγραμμα με το οποίο θα γίνει ή ανάλυση ή, εναλλακτικά, δίνονται οι απαραίτητες εντολές για την καθοδήγηση του υπολογισμού της απόστασης. Είναι, ωστόσο, ύψιστης σημασίας, λόγω του αντίκτυπου στο αποτέλεσμα, η επιλογή της μετρικής, καθώς μπορεί να αποκαλύψει κρίσιμες ή/και ενδιαφέρουσες τοπολογικές και γεωμετρικές ιδιότητες των δεδομένων.

2. Στη συνέχεια επί των δεδομένων κατασκευάζεται ένας ¨συνεχής' χώρος, με σκοπό να δοθεί έμφαση στην υποβόσκουσα τοπολογία ή γεωμετρία. Συχνά προκύπτει ένα πλεγματικό σύμπλεγμα (simplicial complex) ή μία εμφωλευμένη οικογένεια πλεγματικών συμπλεγμάτων, που ονομάζεται δι-

ήθηση, η οποία αντανακλά τη δομή των δεδομένων υπό διαφορετικές κλίμακες οπτικής. Τα πλεγματικά συμπλέγματα μπορούμε να τα δούμε ως υψηλότερης διάστασης γενικεύσεις των γειτονικών γραφημάτων που κλασικά κατασκευάζονται επί των δεδομένων σε πληθώρα τυπικών αλγορίθμων που αξιοποιούνται στην ανάλυση δεδομένων ή την εκπαίδευση αλγορίθμων. Η δυσκολία σε αυτό το σημείο έγκειται στον ορισμό μίας τέτοιας δομής, που αποδεδειγμένα αντανακλά, σχετικές με τη δομή των δεδομένων, πληροφορίες. Οι δομές αυτές όμως θα πρέπει να παρουσιάζουν κάποια ευκολία στην κατασκευή και αποτελεσματικότητα στην πράξη.

3. Από τις δομές που κατασκευάζονται επί των δεδομένων εξάγονται πληροφορίες σχετικές με τις τοπολογικές και γεωμετρικές ιδιότητες που αυτά φέρουν. Αυτή η μελέτη μπορεί να οδηγήσει σε ολική ανακατασκευή, τυπικά μία τριγωνοποίηση του σχήματος που υποβόσκει στα δεδομένα από τα οποία εύκολα εξάγονται οι τοπολογικές ή/και γεωμετρικές πληροφορίες ή από ακατέργαστα συγκεντρωτικά αρχεία ή προσεγγίσεις για τα οποία η εξαγωγή των σχετικών πληροφοριών απαιτεί συγκεκριμένες μεθόδους, όπως είναι για παράδειγμα η εμμένουσα ομολογία. Πέρα, όμως, από την ταυτοποίηση των τοπολογικών ή γεωμετρικών πληροφοριών που παρουσιάζουν ενδιαφέρον, την οπτικοποίησή και την ερμηνεία τους, συνιστά δυσκολίες σε αυτό το σημείο η ανάδειξη της σχετικότητας, της σταθερότητας, λαμβάνοντας υπόψιν διαταραχές, ή την παρουσία θορύβου στα εισαγόμενα δεδομένα. Προκύπτει, με αυτόν τον τρόπο, η ανάγκη κατανόησης της στατιστικής συμπεριφοράς των συναγόμενων ιδιοτήτων.

4. Οι τοπολογικές και γεωμετρικές πληροφορίες που εξάγονται αναδεικνύουν πιθανές τοπολογικές αναλλοίωτες που παρέχουν νέες οπτικές περιγραφής των δεδομένων. Αυτές μπορούν να αξιοποιηθούν στη βελτίωση της κατανόησης των δεδομένων, ειδικά μέσω της οπτικοποίησής τους, ή σε συνδυασμό με άλλα είδη ιδιοτήτων των δεδομένων να προβούμε σε περαιτέρω ανάλυση και εργασίες μηχανικής μάθησης. Στόχος προς κατάκτηση, κάθε φορά, είναι η απόδειξη της βελτίωσης της ανάλυσης και της συμπληρωματικότητας, σε σχέση με άλλα χαρακτηριστικά, των πληροφοριών που παρέχονται από τη χρήση εργαλείων της Τοπολογικής Ανάλυσης Δεδομένων.

Τοπολογική Ανάλυση Δεδομένων και Στατιστική

Μέχρι πρόσφατα, οι θεωρητικές πλευρές της Τοπολογικής Ανάλυσης Δεδομένων και των τοπολογικών συμπερασμάτων βασίζονταν κατά κύριο λόγο

σε αιτιοκρατικές προσεγγίσεις. Δυστυχώς, οι αιτιοκρατικές προσεγγίσεις αδυνατούν να συνυπολογίσουν την επίδραση της τυχαιότητας που διέπει τα δεδομένα και της εσωτερικής μεταβλητότητάς των τοπολογικών ιδιοτήτων που αυτή συνεπάγεται. Κατά συνέπεια, οι περισσότερες μέθοδοι παραμένουν επεξηγηματικές, χωρίς να παρέχουν την δυνατότητα να γίνεται αποτελεσματική διάκριση μεταξύ της πληροφορίας και αυτού που συχνά αναφέρεται ως ¨τοπολογικός θόρυβος¨.

Η στατιστική προσέγγιση της Τοπολογικής Ανάλυσης Δεδομένων βασίζεται στην ιδέα ότι τα δεδομένα προέρχονται από κάποια άγνωστη κατανομή, όπως επίσης και στο ότι οι συναγόμενες τοπολογικές ιδιότητες από της μεθόδους της Τοπολογικής Ανάλυσης Δεδομένων θεωρούνται εκτιμήτριες των τοπολογικών ποσοτήτων που περιγράφουν κάποιο υποβόσκον αντικείμενο. Υπό αυτήν την προσέγγιση, το άγνωστο αντικείμενο, συνήθως το άγνωστο υποβόσκον αντικείμενο σχετίζεται με το στήριγμα της κατανομής των δεδομένων, ή είναι τουλάχιστον κοντά στο στήριγμά της. Ωστόσο, αυτό το στήριγμα δεν έχει πάντοτε φυσική υπόστασή. Για παράδειγμα, οι γαλαξίες του σύμπαντος φαίνεται να είναι οργανωμένοι σε ένα πλέγμα, το οποίο όμως δεν υπάρχει.

Ο βασικός στόχος της στατιστικής προσέγγισης στην ανάλυση δεδομένων εμπεριέχει κάποιες δυσκολίες. Στην παρακάτω λίστα συνοψίζονται αυτά τα προβλήματα:

1. Η εύρεση περιοχών εμπιστοσύνης για τις τοπολογικές ιδιότητες και ο έλεγχος της σημαντικότητας των εκτιμημένων τοπολογικών ποσοτήτων.
2. Η επιλογή σχετικών κλιμάκων στις οποίες τα τοπολογικά φαινόμενα θα μπορούν να θεωρούνται ως συναρτήσεις των παρατηρούμενων δεδομένων.
3. Η αντιμετώπιση των ακραίων παρατηρήσεων και η αξιοποίηση εύρωστων μεθόδων για την Τοπολογική Ανάλυση Δεδομένων.

Η Τοπολογική Ανάλυση Δεδομένων στην Επιστήμη των Δεδομένων

Από άποψη εφαρμογής, πολλά πρόσφατα επιτυχή και υποσχόμενα αποτελέσματα έχουν αναδείξει το ενδιαφέρον που παρουσιάζουν οι τοπολογικές και γεωμετρικές προσεγγίσεις σε ένα συνεχώς αυξανόμενο αριθμό πεδίων. Ονομαστικά κάποιες από αυτές της εφαρμογές είναι:

- στην επιστήμη των υλικών (Kramar et al. (2013), Nakamura et al. (2015))

- στην ανάλυση τρισδιάστατων σχημάτων (Skraba et al. (2010), Turner et al. (2014b))
- στην ανάλυση πολυμεταβλητών χρονοσειρών (Seversky et al. (2016))
- στη βιολογία (Yao et al. (2009))
- τη χημεία (Lee et al. (2017))

Από την άλλη πλευρά, η επιτυχία των αποτελεσμάτων της Τοπολογικής Ανάλυσης Δεδομένων στηρίζεται και στον συνδυασμό της με άλλες τεχνικές ανάλυσης ή εκπαίδευσης, όπως θα συζητηθεί παρακάτω στην εργασία. Επομένως η αποσαφήνιση του τρόπου και της συμβατότητας της Τοπολογικής Ανάλυσης Δεδομένων σε σχέση με την ανάλυση μέσω άλλων προσεγγίσεων και εργαλείων στην επιστήμη των δεδομένων είναι επίσης ένα πολύ σημαντικό πεδίο συζήτησης, το οποίο επί του παρόντος ερευνάται.

2.1 Μετρικοί Χώροι, Καλύμματα και Πλεγματικά Συμπλέγματα

Καθώς οι τοπολογικές και γεωμετρικές ιδιότητες συνήθως συσχετίζονται με συνεχείς χώρους, ενώ τα δεδομένα παρουσιάζονται ως πεπερασμένα σύνολα παρατηρήσεων, οι τοπολογικές πληροφορίες δεν αποκαλύπτονται απευθείας. Ένας απλός τρόπος να αναδείξουμε την τοπολογική δομή των δεδομένων είναι να κατά κάποιο τρόπο να ενώσουμε τα σημεία στα οποία απεικονίζονται τα δεδομένα που βρίσκονται κοντά το ένα στο άλλο έτσι ώστε να εκθέσουμε το καθολικό συνεχές υποβόσκων σχήμα των δεδομένων. Η ποσοτικοποίηση της έννοιας της εγγύτητας μεταξύ των σημείων στα οποία απεικονίζονται τα δεδομένα συνήθως επιτυγχάνεται με τη χρήση κάποιας μετρικής ή κάποιο μέτρο ανομοιότητας και είναι συνήθως βολικό στην Τοπολογική Ανάλυση Δεδομένων να θεωρούμε τα σύνολα δεδομένων ως διακριτούς μετρικούς χώρους ή ως δείγματα μετρικών χώρων.

Μετρικοί Χώροι

Υπενθυμίζουμε τον ορισμό του μετρικού Χώρου:

2.1.1 Ορισμός

Έστω ένα σύνολο M και μία συνάρτηση ρ. Η ρ καλείται μετρική αν για κάθε $x, y, z \in M$ ικανοποιεί τις εξής ιδιότητες:

(i) $\rho(x, y) \geq 0$ και $\rho(x, y) = 0$ αν και μόνο άν $x = y$.

(ii) $\rho(x, y) = \rho(y, x)$ και

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

Τότε το ζεύγος (M, ρ) ονομάζεται μετρικός χώρος.

Δοθέντος ενός μετρικού χώρου (M, ρ), το σύνολο των συμπαγών υποσυνόλων του, έστω $\mathcal{K}(M)$, μπορεί να εφοδιαστεί με τη λεγόμενη απόσταση Hausdorff. Πιο αναλυτικά, δεδομένων δύο συμπαγών υποσυνόλων, έστω $A, B \subseteq M$ η απόσταση Hausdorff $d_H(A, B)$ μεταξύ των A και B ορίζεται

ως ο ελάχιστος μη αρνητικός αριθμός δ τέτοιος ώστε για κάθε $a \in A$ υπάρχει $b \in B$ τέτοιο ώστε $\rho(a,b) \leq \delta$ και για κάθε $b \in B$, υπάρχει $a \in A$ τέτοιο ώστε $\rho(a,b) \leq \delta$, όπως φαίνεται παρακάτω στο Σχήμα 1. Εναλλακτικά, αν για κάθε συμπαγές υποσύνολο $C \subseteq M$, συμβολίζουμε τη συνάρτηση που δίνει την απόσταση από το C με $d(.,C) : M \to \mathbb{R}_+$, η οποία ορίζεται ως $d(x,C) := inf_{c \in C}\rho(x,c)$, για κάθε $x \in M$, τότε μπορούμε να αποδείξουμε ότι η απόσταση Hausdorff μεταξύ των A και B ορίζεται από οποιαδήποτε από τις παρακάτω ισότητες:

$$\begin{aligned} d_H(A,B) &= max\{sup_{b\in B}d(b,A), sup_{a\in A}d(a,B)\} \\ &= sup_{x\in M}|d(x,A) - d(x,B)| = ||d(.,A) - d(.,B)||_\infty \end{aligned}$$

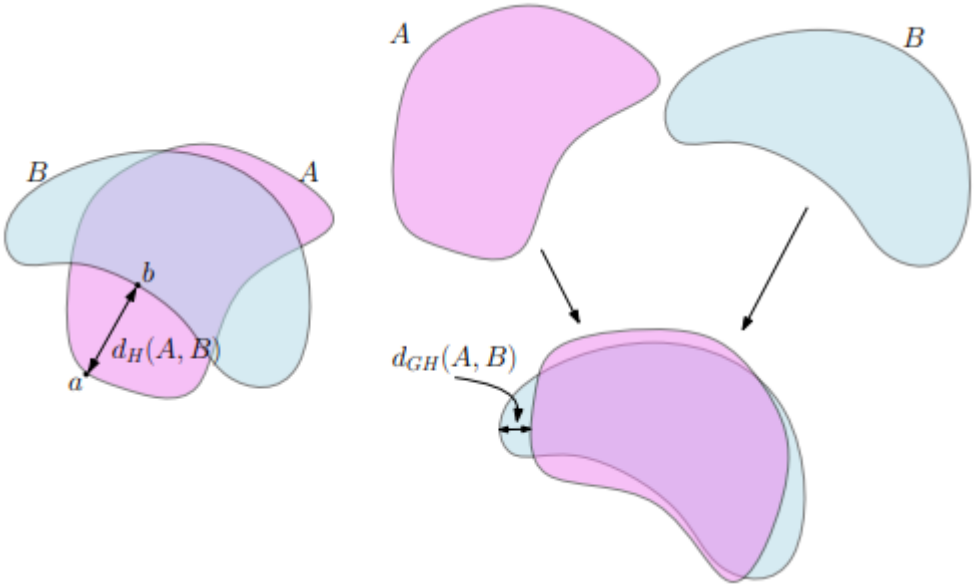
Κλασικό αποτέλεσμα αποτελεί το ότι η απόσταση Hausdorff είναι όντως απόσταση μεταξύ των συμπαγών υποσυνόλων ενός μετρικού χώρου. Υπό το πρίσμα της Τοπολογικής Ανάλυσης Δεδομένων, παρέχει έναν εύχρηστο τρόπο να ποσοτικοποιούμε την εγγύτητα μεταξύ διαφορετικών συνόλων δεδομένων που δεν έχουν συντεθεί από στοιχεία του ίδιου περιβάλλοντα χώρου. Ευτυχώς, η έννοια της απόστασης Hausdorff μπορεί να γενικευθεί στη σύγκριση οποιωνδήποτε δύο συμπαγών μετρικών χώρων, που δημιουργεί το έδαφος για την έννοια της απόστασης Gromov - Hausdorff.

Δύο συμπαγείς μετρικοί χώροι, έστω (M_1, ρ_1) και (M_2, ρ_2), καλούνται ισομετρικοί αν υπάρχει μία αμφιμονοσήμαντη συνάρτηση $\varphi : M_1 \to M_2$, τέτοια ώστε να διατηρεί τις αποστάσεις, δηλαδή να ισχύει ότι $\rho_2(\varphi(x), \varphi(y)) = \rho_1(x,y)$, για κάθε $x, y \in M_1$. Η απόσταση Gromov-Hausdorff μετρά πόσο μακριά από ισομετρικοί είναι δύο μετρικοί χώροι.

2.1.2 Ορισμός

Η απόσταση Gromov-Hausdorff $d_{GH}(M_1, M_2)$ μεταξύ δύο συμπαγών μετρικών χώρων είναι το infimum των πραγματικών αριθμών $r \geq 0$ τέτοιων ώστε να υπάρχει ένας μετρικός χώρος (M, r) και οι συμπαγείς υπόχωροι $C_1, C_2 \subset M$ που είναι ισομετρικοί με τους M_1 και M_2 αντίστοιχα και τέτοιοι ώστε $d_H(C_1, C_2) \leq r$.

Η απόσταση Gromov-Hausdorff θα χρησιμοποιηθεί παρακάτω στην εργασία για τη μελέτη των ιδιοτήτων σταθερότητας των εμμενόντων διαγραμμάτων.



Σχήμα 2.1: **Αριστερά:** Η απόσταση Gromov-Hausdorff μεταξύ των υποσυνόλων A και B. Το A μπορεί να περιστραφεί. Αυτό αποτελεί μία ισομετρική ενσωμάτωση του A στο επίπεδο. Η περιστροφή γίνεται με σκοπό να μειωθεί η απόσταση Hausdorff από το B και κατ΄ επέκταση $d_{GH}(A,B) \leq d_H(A,B)$. **Δεξιά:** Η απόσταση Hausdorff μεταξύ των υποσυνόλων A και B του επιπέδου. Σε αυτό το παράδειγμα ως $d_H(A,B)$ ορίζεται η απόσταση μεταξύ των σημείων $a \in A$ και $b \in B$, διότι το σημείο a είναι το πιο ¨απομακρυσμένο' στο A σημείο από το b. [0]

Η σύνδεση ζευγών γειτονικών σημείων δεδομένων με μία ακμή οδηγεί στη συνήθη έννοια του γραφήματος γειτνίασης από το οποίο οι συνδεσιμότητα των δεδομένων μπορεί να αναλυθεί, για παράδειγμα χρησιμοποιώντας αλγορίθμους ομαδοποίησης. Για να προχωρήσουμε πέρα από τη συνδεσιμότητα, η κεντρική ιδέα της Τοπολογικής Ανάλυσης Δεδομένων είναι να κατασκευάσουμε ένα υψηλότερης διάστασης ισοδύναμου του γραφήματος γειτνίασης συνδέοντας, όμως, όχι μόνο τα ζεύγη σημείων αλλά και τα $(k+1)$-διάστατα γειτονικά σημεία. Τα αντικείμενα στα οποία καταλήγουμε ονομάζονται πλεγματικά συμπλέγματα και μας επιτρέπουν να αναγνωρίζουμε νέες τοπολογικές ιδιότητες, όπως κύκλους, κενά και τα υψηλότερης διάστασης αντίστοιχα σημεία τους.

2.1.1 Γεωμετρικά και αφηρημένα πλεγματικά συμπλέγματα

Τα πλεγματικά συμπλέγματα μπορούν να θεωρηθούν ως γενίκευση των γραφημάτων σε ανώτερες διαστάσεις. Είναι αντικείμενα με τοπολογική και συνδυαστική υπόσταση, ιδιότητα που τα καθιστά ιδιαίτερα χρήσιμα στην Τοπολογική Ανάλυση Δεδομένων.

Δοθέντος ενός συνόλου $\mathbb{X} = \{x_0, x_1, ..., x_k\} \subset \mathbb{R}^d$ των $k+1$ αφινικά ανεξάρτητων σημείων, το k-διάστατο σύμπλεγμα $\sigma = [x_0, x_1, ..., x_k]$ που επεκτείνεται από το $\mathbb{X}$ είναι η κυρτή καμπύλη που περικλείει το $\mathbb{X}$. Σημεία του $\mathbb{X}$ ονομάζονται τα *διανύσματα του* σ και τα πλέγματα που επεκτείνονται από υποσύνολα του $\mathbb{X}$ καλούνται *όψεις του* σ. Ένα γεωμετρικό πλεγματικό σύμπλεγμα K στον $\mathbb{R}^d$ είναι μία συλλογή πλεγμάτων για την οποία:

i) Κάθε όψη ενός πλέγματος K είναι ένα πλέγμα του K,

ii) Η τομή οποιωνδήποτε δύο πλεγμάτων του K είναι είτε το κενό, είτε μία όψη.

Η ένωση των πλεγμάτων του K είναι ένα υποσύνολο του $\mathbb{R}^d$ ονομάζεται υποβόσκοντας χώρος του K και κληρονομεί από την τοπολογία του $\mathbb{R}^d$. Επομένως, μπορούμε να δούμε το K και ως έναν τοπολογικό χώρο μέσω του υποβόσκοντα χώρου του. Σε αυτό το σημείο σημειώνεται ότι όταν τα διανύσματά του είναι γνωστά, το K είναι πλήρως καθορισμένο από τη συνδυαστική περιγραφή μίας συλλογής πλεγμάτων που ικανοποιεί κάποιους κανόνες περιπτώσεων.

Δεδομένου ενός συνόλου V, ένα *αφηρημένο πλέγματικό σύμπλεγμα* με σύνολο κορυφών V είναι ένα σύνολο $\tilde{K}$ που αποτελείται από πεπερασμένα υποσύνολα του V, τέτοιο ώστε τα στοιχεία του V να ανήκουν στο $\tilde{K}$ και για κάθε $\sigma \in \tilde{K}$ κάθε υποσύνολο του σ ανήκει στο $\tilde{K}$. Τα στοιχεία του $\tilde{K}$ καλούνται όψεις ή πλέγματα του $\tilde{K}$. Η διάσταση ενός αφηρημένου πλέγματος είναι η πληθικότητα του μείον 1 και η διάσταση του $\tilde{K}$ είναι η μεγαλύτερη από τις διαστάσεις των πλεγμάτων του. Επιπλέον, παρατηρούμε ότι τα πλεγματικά συμπλέγματα διάστασης 1 είναι γραφήματα.

Η συνδυαστική περιγραφή κάθε γεωμετρικού πλέγματος K δίνει τη δυνατότητα δημιουργίας ενός αφηρημένου πλεγματικού συμπλέγματος $\tilde{K}$. Το αντίστροφό του επίσης ισχύει. Μπορούμε πάντα να συσχετίσουμε ένα αφηρημένο πλεγματικό σύμπλεγμα $\tilde{K}$ με ένα τοπολογικό χώρο $|\tilde{K}|$ τέτοιο ώστε αν

K είναι ένα γεωμετρικό πλέγμα, η συνδυαστική περιγραφή του οποίου είναι ίδια με αυτή του $\tilde{K}$, τότε ο υποβόσκων χώρος του K είναι ομοιομορφικός με τον $|\tilde{K}|$. Ένα τέτοιο K θα ονομάζεται *γεωμετρική πραγματοποίηση* του K. Κατά συνέπεια, τα αφηρημένα πλεγματικά συμπλέγματα μπορούμε να τα δούμε ως γεωμετρικές πραγματοποιήσεις των υποβοσκόντων συνδυαστικών δομών. Με αυτόν τον τρόπο, μπορούμε να θεωρήσουμε ότι τα πλεγματικά συμπλέγματα είναι ταυτόχρονα συνδυαστικά αντικείμενα, που είναι κατάλληλα κατασκευασμένα για αποδοτικούς υπολογισμούς, και τοπολογικοί χώροι, τους οποίους μπορούμε να μελετήσουμε ως προς τις τοπολογικές τους ιδιότητες.

2.1.2 Κατασκευή πλεγματικών συμπλεγμάτων από δεδομένα

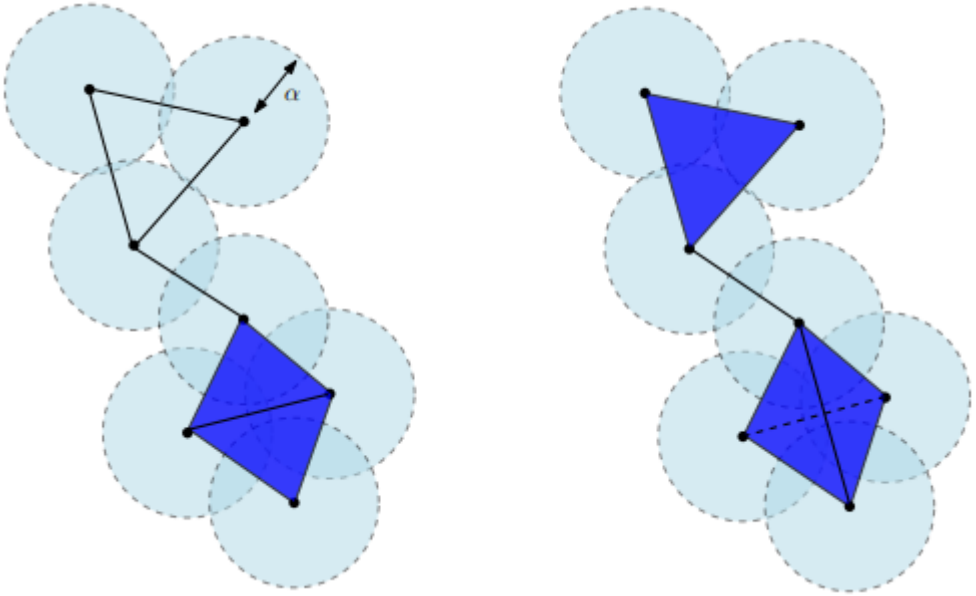
Δεδομένου ενός συνόλου δεδομένων, η γενικότερα ενός τοπολογικού ή μετρικού χώρου, υπάρχουν πολλοί τρόποι να κατασκευαστούν πλεγματικά συμπλέγματα. Παρουσιάζουμε παρακάτω μερικά κλασικά παραδείγματα που είναι ευρέως χρησιμοποιούμενα για εξάσκηση.

Το πρώτο παράδειγμα είναι μία άμεση επέκταση της έννοιας του a-γειτονικού γραφήματος. Υποθέτουμε ότι δοθέντος ενός συνόλου σημείων του $\mathbb{X}$ σε ένα μετρικό χώρο (M, ρ) και έναν πραγματικό αριθμό $a \geq 0$, το πλέγμα Vierotis-Rips $Rips_a(\mathbb{X})$ είναι ένα σύνολο πλεγμάτων $\{x_0, x_1, ..., x_k\}$ τέτοιο ώστε $d_{\mathbb{X}}(x_i, x_j) \leq a$, για κάθε (i, j). Από τον ορισμό προκύπτει άμεσα ότι αυτό αποτελεί ένα αφηρημένο πλεγματικό σύμπλεγμα. Ωστόσο, σε γενικές γραμμές, ακόμη και όταν $\mathbb{X}$ είναι ένα πεπερασμένο σύνολο του $\mathbb{R}^d$, το $Rips_a(\mathbb{X})$ δεν επάγει κάποια γεωμετρική πραγματοποίηση στον $\mathbb{R}^d$. Αυτό προκύπτει από το ότι η διάστασή του θα είναι μεγαλύτερη του d.

Το σύμπλεγμα Vietoris-Rips είναι στενά συνδεδεμένο με το σύμπλεγμα Čech, το οποίο συμβολίζουμε με $Cech_a(\mathbb{X})$, που ορίζεται από το σύνολο των πλεγμάτων $[x_0, x_1, ..., x_k]$. τέτοιο ώστε με τη $(k+1)$-διάστατη κλειστή μπάλα $B(x_i, a)$ να έχουν μη κενή τομή. Σημειώνεται ότι αυτά τα δύο συμπλέγματα σχετίζονται ως εξής:

$$Rips_a(\mathbb{X}) \subseteq Cech_a(\mathbb{X}) \subseteq Rips_{2a}(\mathbb{X})$$

και το ότι αν $\mathbb{X} \subset \mathbb{R}^d$ τότε τα $Cech_a(\mathbb{X})$ και $Rips_{2a}(\mathbb{X})$ έχουν τον ίδιο μονοδιάστατο σκελετό, δηλαδή έχουν τα ίδια διανύσματα και ακμές.



Σχήμα 2.2: Το συμπλέγμα Čech, $Cech_a(\mathbb{X})$, (αριστερά) και το Vietoris-Rips, $Rips_{2a}(\mathbb{X})$, (δεξιά) ενός πεπερασμένου νέφους σημείων στο επίπεδο $\mathbb{R}^d$. Το κάτω μέρος του $Cech_a(\mathbb{X})$ είναι η ένωση των δύο γειτονικών τριγώνων, ενώ το κάτω μέρος του $Rips_{2a}(\mathbb{X})$ είναι ένα τετράεδρο που επεκτείνεται από τέσσερα διανύσματα και όλες τις όψεις τους. Το σύμπλεγμα Čech είναι δισδιάστατο ενώ το Vietoris-Rips είναι τρισδιάστατο. Επισημαίνεται ότι το Vietoris-Rips δεν είναι ενσωματωμένο στον $\mathbb{R}^d$. [0]

2.1.3 Το θεώρημα νεύρου

Το σύμπλεγμα Čech είναι μία ιδιαίτερη περίπτωση μίας οικογένειας συμπλεγμάτων που συνδέονται με καλύμματα. Δοθέντος ενός καλύμματος $\mathcal{U} = (U_i)_{i\in I}$ του $\mathbb{M}$, δηλαδή μία οικογένεια συνόλων U_i, τέτοιο ώστε $\mathbb{M} = \bigcup_{i\in I} U_i$, το νεύρο του $\mathcal{U}$ είναι ένα αφηρημένο πλεγματικό σύμπλεγμα $C(\mathcal{U})$, του οποίου τα διανύσματα είναι τα U_i, για τα διάφορα $i \in I$, και τέτοια ώστε:

$$\sigma = [U_{i_0}, U_{i_1}, ..., U_{i_k}] \text{ αν και μόνο αν } \bigcap_{j=0}^{k} U_{i_j} \neq \emptyset.$$

Δεδομένου ενός καλύμματος ενός συνόλου δεδομένων, όπου κάθε σύνολο του καλύμματος μπορεί να είναι για παράδειγμα μία τοπική συστάδα ή κάποια ομαδοποίηση σημείων των δεδομένων με κάποια κοινά χαρακτηριστικά. Τότε το

νεύρο του παρέχει μία συμπαγή και συνδυαστική περιγραφή της σχέσης μεταξύ αυτών των συνόλων μέσω των μοτίβων της τομής τους (βλέπε Σχήμα 2.3).

Ένα θεμελιώδες θεώρημα της αλγεβρικής τοπολογίας συνδέει, διατυπώνοντας πρώτα κάποιες υποθέσεις, την τοπολογία του νεύρου του καλύμματος με την τοπολογία των ενώσεων των συνόλων του καλύμματος. Για τη διατύπωση του θεωρήματος θα χρειαστεί να εισάγουμε κάποιες έννοιες.

Δύο τοπολογικοί χώροι, έστω X και Y, θεωρούνται συνήθως όμοιοι από τοπολογικής άποψης αν είναι ομοιομορφικοί, δηλαδή αν υπάρχουν δύο συνεχείς αμφιμονοσήμαντες συναρτήσεις $f : X \to Y$ και $g : Y \to X$ τέτοιες ώστε οι $f \circ g$ και $g \circ f$ να είναι οι ταυτοτικές απεικονίσεις του Y και του X αντίστοιχα. Σε πολλές περιπτώσεις το να είναι ομοιομορφικοί οι X και Y είναι πολύ ισχυρότερο αίτημα από αυτό που απαιτείται για να δείξουμε ότι οι X και Y έχουν κοινές τοπολογικές ιδιότητες που ενδιαφέρουν την Τοπολογική Ανάλυση Δεδομένων.

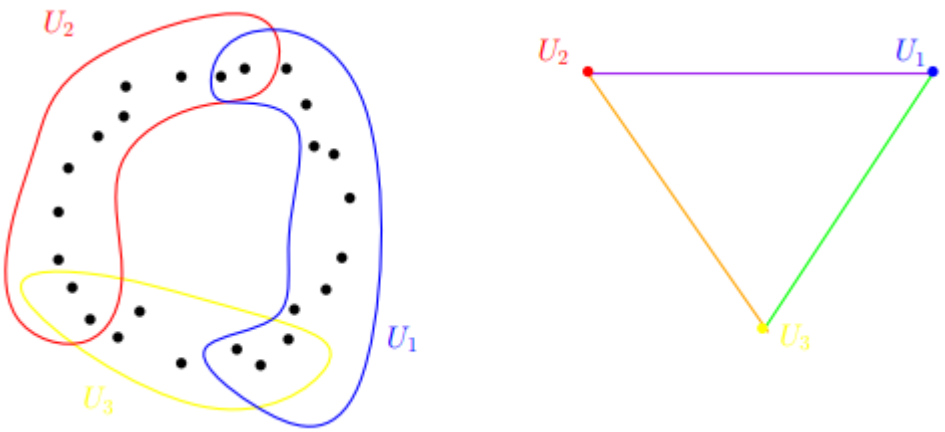
Δύο συνεχείς απεικονίσεις, έστω $f_0, f_1 : X \to Y$ θα λέγονται ομοτοπικές αν υπάρχει συνεχής απεικόνιση $H : X \times [0, 1] \to Y$ τέτοια ώστε για κάθε $x \in X$, $H(x, 0) = f_0(x)$ και $H(x, 1) = g(x)$. Τότε η χώροι X και Y θα ονομάζονται ομοτοπικά ισοδύναμοι. Η έννοια της ομοτοπικής ομοιότητας είναι ασθενέστερη της έννοιας του ομοιομορφικότητας. Αν οι X και Y είναι ομοιομορφικοί τότε θα είναι και ομοτοπικά ισοδύναμοι, για να είμαστε πιο ακριβείς θα έχουν την ίδια ομολογία όπως θα δούμε και παρακάτω.

Ένας χώρος καλείται συσταλτός αν είναι ομοτοπικά ισοδύναμος με ένα σημείο. Κάποια κλασικά παραδείγματα συσταλτών χώρων είναι οι μπάλες και γενικότερα τα κυρτά σύνολα στον $\mathbb{R}^d$. Τα ανοικτά καλύμματα τα στοιχεία των οποίων και οι τομές τους είναι συσταλτά έχουν την εξής ιδιότητα:

2.1.3 Θεώρημα (Θεώρημα Νεύρου)

Έστω $\mathcal{U} = (U_i)_{i \in I}$ ένα κάλυμμα ενός τοπολογικού χώρου X με ανοικτά σύνολα τέτοια ώστε η τομή οποιονδήποτε U_i να είναι είτε κενή είτε συσταλτός χώρος. Τότε το X και το νεύρο $C(\mathcal{U})$ είναι ομοτοπικά ισοδύναμα.

Εύκολα εξακριβώνουμε ότι τα κυρτά υποσύνολα του Ευκλείδειου χώρου είναι συσταλτά. Κατά συνέπεια, αν $\mathcal{U} = (U_i)_{i \in I}$ είναι μία συλλογή κυρτών υποσυνόλων του $\mathbb{R}^d$, τότε τα $C(\mathcal{U})$ και $\bigcup_{i \in I} U_i$ είναι ομοτοπικά ισοδύναμοι. Πιο συγκεκριμένα, αν το $\mathbb{X}$ είναι ένα σύνολο σημείων του $\mathbb{R}^d$, τότε το σύμπλεγμα Čech, $Cech_a(\mathbb{X})$, είναι ομοτοπικά ισοδύναμο με την ένωση των μπαλών



Σχήμα 2.3: Το νεύρο ενός καλύμματος ενός συνόλου σημείων ενός δείγματος στο επίπεδο. [0]

$\bigcup_{x\in\mathbb{X}} B(x,a)$.

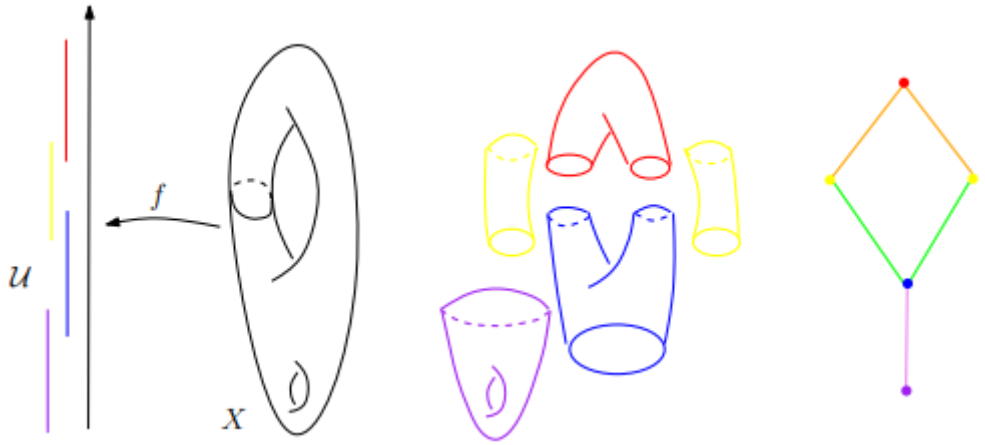
Το Θεώρημα Νεύρου είναι πολύ σημαντικό για την Τοπολογική Ανάλυση Δεδομένων. Ακριβέστερα, παρέχει έναν τρόπο κωδικοποίησης της τοπολογίας συνεχών χώρων σε αφηρημένες συνδυαστικές δομές που είναι κατάλληλες για το σχεδιασμό αποδοτικών δομών και αλγορίθμων δεδομένων.

2.2 Χρήση καλυμμάτων και νεύρων για τη διερευνητική ανάλυση δεδομένων και την οπτικοποίηση: Ο Αλγόριθμος Mapper

Η χρήση του νεύρου των καλυμμάτων ως μέσο για να συνοψίσουμε και να εξερευνήσουμε τα δεδομένα αποτελεί μία ιδέα που προκύπτει με φυσικό τρόπο και που ήταν η πρώτη που προτάθηκε για την Τοπολογική Ανάλυση Δεδομένων στο Singh et al. (2007), που δημιούργησε τη βάση για τον λεγόμενο αλγόριθμο Mapper.

2.2.1 Ορισμός

Έστω $f : X \to \mathbb{R}^d$, $d \geq 1$, μία συνεχής συνάρτηση πραγματικής μεταβλητής και $\mathcal{U} = (U_i)_{i \in I}$ κάποιο κάλυμμα του $\mathbb{R}^d$. Το κάλυμμα pull-back του X που επάγεται από το $(f, \mathcal{U})$ είναι η συλλογή των ανοικτών συνόλων της μορφής $(f^{-1}(U_i))_{i \in I}$. Τότε, το επεξεργασμένο pull-back είναι η συλλογή των ενωμένων τμημάτων των ανοικτών συνόλων $f^{-1}(U_i)$, $i \in I$.



Σχήμα 2.4: Το επεξεργασμένο κάλυμμα του pull-back της συνάρτησης ύψους στην επιφάνεια $\mathbb{R}^3$ και το νεύρο του. [0]

Εν γένει η ιδέα του αλγορίθμου Mapper δεδομένου ενός συνόλου δεδομένων $\mathbb{X}$ και μίας καλά ορισμένης συνάρτησης πραγματικής μεταβλητής $f : \mathbb{X} \to \mathbb{R}^d$,είναι να συνοψίσει το $\mathbb{X}$ μέσω του νεύρου του επεξεργασμένου pull-

back του καλύμματος $\mathcal{U}$ του $f(\mathbb{X})$. Για καλώς επιλεγμένα καλύμματα $\mathcal{U}$, όπως θα δούμε παρακάτω, το νεύρο αποτελεί γράφημα που παρέχει ένα εύληπτο τρόπο οπτικοποίησης της σύνοψης των δεδομένων. Στο Σχήμα 2.4 φαίνεται ένα απλό τέτοιο παράδειγμα.

Ο αλγόριθμος Mapper είναι σχετικά απλός, ωστόσο εγείρει αρκετά ερωτήματα σχετικά με της διάφορες επιλογές που αφήνονται στο χρήστη. Αυτό θα συζητηθεί συνοπτικά παρακάτω.

Η επιλογή της συνάρτηση f

Η επιλογή της συνάρτησης f, που συνήθως καλείται συνάρτηση φίλτρου ή φακού, εξαρτάται σε μεγάλο βαθμό από τα χαρακτηριστικά των δεδομένων που θέλουμε να τονίσουμε. Παρακάτω παρατίθενται τα συνηθέστερα στη βιβλιογραφία:

- *Εκτιμήσεις πυκνότητας:* Το σύμπλεγμα mapper μπορεί να συμβάλλει στην κατανόηση της δομής και συνεκτικότητας μίας περιοχής με μεγάλη πυκνότητα (συστάδες)

- *PCA συντεταγμένες ή συναρτήσεις συντεταγμένων που προκύπτουν από την τεχνική της μη γραμικής διαστατικής μείωσης, ιδιοσυναρτήσεις γραφημάτων Laplace:* Μπορεί να συμβάλλουν στην ανακάλυψη και κατανόηση κάποιων ασαφιών που προκύπτουν κατά τη χρήση μη γραμμικών διαστατισκών μειώσεων

- *Η συνάρτηση κεντρικότητας* $f(x) = \sum_{y \in \mathbb{X}} d(x, y)$ *και η συνάρτηση κεντρικότητας* $f(x) = max_{y \in \mathbb{X}} d(x, y)$*:* αποτελούν καλές επιλογές σε κάποιες περιπτώσεις που δεν προϋποτίθενται συγκεκριμένες γνώσεις για τα δεδομένα

- Για δεδομένα που εξάγονται από δείγματα μονοδιάστατων δομών νημάτων, η συνάρτηση απόστασης σε ένα ορισμένο σημείο επιτρέπει την εύρεση της υποβόσκουσας τοπολογίας των δομών νημάτων (Chazal et al. (2015c))

Παρακάτω παρατίθενται ένα παράδειγμα αλγόριθμου mapper και το σχήμα 2.5, στο οποίο αποτυπώνεται η διαδικασία του αλγορίθμου mapper στην περίπτωση που τα δεδομένα φαίνεται να σχηματίζουν κύκλο.

Αλγόριθμος 1 (Αλγόριθμος Mapper)

Input	Ένα σύνολο $\mathbb{X}$ δεδομένων με μία μετρική ή ένα μέτρο ανομοιότητας μεταξύ σημείων των δεδομένων, μία συνάρτηση $f : \mathbb{X} \to \mathbb{R}^d$ και ένα κάλυμμα $\mathcal{U}$ του $f(\mathbb{X})$
Output	· Ένα πλεγματικό σύμπλεγμα, το νεύρο (συχνά είναι ένα γράφημα για καλά επιλεγμένα καλύμματα για να είναι εύκολη η οπτικοποίηση) · Ένα διάνυσμα $v_{U,i}$ για κάθε συστάδα $C_{U,i}$ · Μία ακμή μεταξύ των διανυσμάτων $v_{U,i}$ και $v_{U',i}$, αν και μόνο αν $C_{U,i} \cap C_{U',i} \neq \emptyset$

Η επιλογή του καλύμματος $\mathcal{U}$

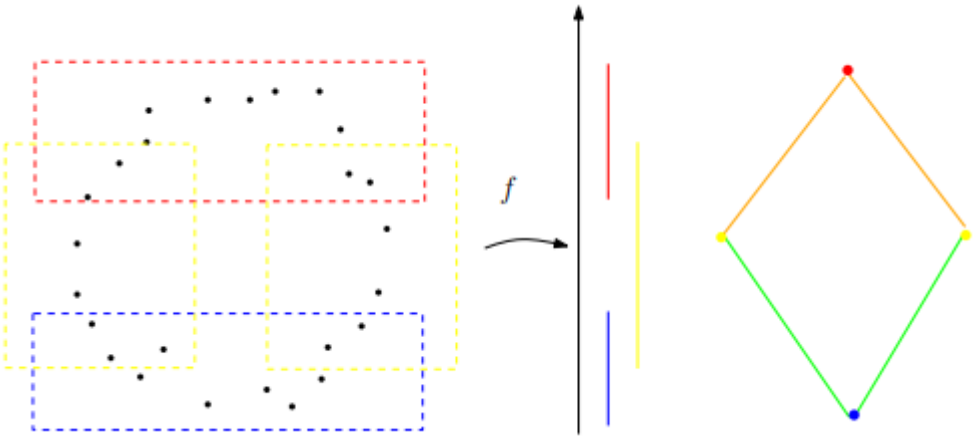
Έστω f μία πραγματική συνάρτηση. Τότε μία συνήθης επιλογή για το κάλυμμα $\mathcal{U}$ είναι ένα σύνολο διαστημάτων κανονικών αποστάσεων ίσου μήκους $r > 0$ που καλύπτει το σύνολο $f(\mathbb{X})$. Ο πραγματικός αριθμός r καλείται κάποιες φορές *διάλυση (resolution)* του καλύμματος. Επιπλέον το ποσοστό g επικάλυψης μεταξύ διαδοχικών διαστημάτων ονομάζεται *ωφέλεια (gain)* του καλύμματος (βλέπε Σχήμα 2.6).

Επισημαίνεται ότι αν η ωφέλεια g επιλέγεται να είναι μικρότερη του 50

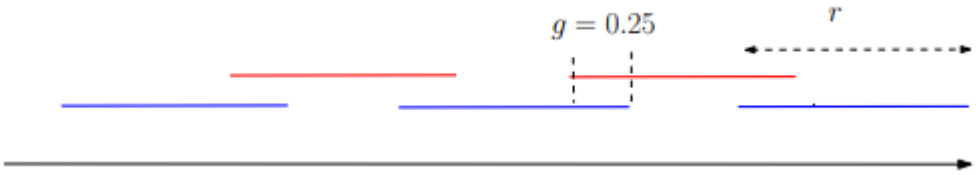
Η επιλογή των συστάδων

Ο αλγόριθμος Mapper προϋποθέτει τη συσταδοποίηση των προεικόνων των ανοικτών συνόλων $U \in \mathcal{U}$. Υπάρχουν δύο στρατηγικές προσεγγίσεις για τη συσταδοποίηση, οι οποίες παρατίθενται συνοπτικά παρακάτω:

1. Εφαρμόζουμε για κάθε $U \in \mathcal{U}$ έναν αλγόριθμο συσταδοποίησης, ο οποίος επιλέγεται από το χρήστη, στην προεικόνα $f^{-1}(U)$.
2. Κατασκευάζουμε ένα γράφημα γειτνίασης επί του συνόλου δεδομένων $\mathbb{X}$



Σχήμα 2.5: Ο αλγόριθμος Mapper σε ένα νέφος σημείων που σχηματίζουν κύκλο. [0]



Σχήμα 2.6: Παράδειγμα καλύμματος της πραγματικής ευθείας, με διάλυση r και ωφέλεια $g = 25\%$. [0]

για κάθε $U \in \mathcal{U}$, κρατώντας τα ενωμένα συστατικά στοιχεία του υπογραφήματος με το σύνολο κορυφών $f^{-1}(U)$. Η δεύτερη προσέγγιση είναι και η πιο διαδεδομένη.

Η θεωρητική και στατιστική πλεύρα του αλγορίθμου Mapper

Βασισμένοι στα αποτελέσματα για τη σταθερότητα και τη δομή του αλγορίθμου Mapper στα οποία κατέληξαν οι Carriere and Oudot (2015), έχουν πρόσφατα γίνει βήματα προς μία καλά θεμελιωμένη στατιστικά εκδοχή του αλγορίθμου Mapper Carriere et al (2017). Όπως είναι αναμενόμενο η απόδοση αποτελεσμάτων του αλγορίθμου Mapper εξαρτάται από το δείγμα που θα επιλεγεί καθώς και από τη συνάρτηση φίλτρου. Επιπρόσθετα, μπορούμε να προτείνουμε στρατηγικές για τη λήψη μικρότε-

ρων του αρχικού δειγμάτων με σκοπό να επιλεγεί ένα σύμπλεγμα σε ένα φιλτράρισμα Rips σε κάποια βολική για την ανάλυση κλίμακα, όπως επίσης και η διάλυση και η ωφέλεια για τον προσδιορισμό του γραφήματος Mapper. Άλλες προσεγγίσεις για την αντιμετώπιση της αστάθειας του αλγορίθμου Mapper έχουν προταθεί για μελέτη από τους Dey et al (2016, 2017).

Ανάλυση δεδομένων με τη χρήση του αλγορίθμου Mapper

Ο αλγόριθμος Mapper έχει επιτυχώς χρησιμοποιηθεί ως εργαλείο διερεύνησης στην ανάλυση δεδομένων για τη συσταδοποίηση και την επιλογή χαρακτηριστικών προς μελέτη. Η κεντρική ιδέα είναι να ανγνωρίσουμε συγκεκριμένες δομές στο γράφημα (ή σύμπλεγμα) Mapper και πιο συκεκριμένα βρόχους και ανοίγματα. Αυτές οι δομές θα χρησιμοποιηθούν στη συνέχεια για την αναγνώριση συστάδων που παρουσιάζουν ενδιαφέρον για την ανάλυση ή για την επιλογή χαρακτηριστικών ή μεταβλητών που βοηθούν στη βέλτιστη διάκριση των δεδομένων σε αυτές τις δομές. Για εφαρμογές σε πραγματικά δεδομένα, στις οποίες παρουσιάζονται αυτές οι τεχνικές μπορούμε να ανατρέξουμε για παράδειγμα στα έργα των Lum et al (2013) και Yao et al (2009).

2.3 Γεωμετρική ανακατασκευή και ομολογιακή συμπερασματολογία

Ένας άλλος τρόπος να κατασκευάσουμε καλύμματα και να χρησιμοποιήσουμε τα νεύρα τους για να εκθέσουμε την τοπολογική δομή των δεδομένων είναι να θεωρήσουμε μία ένωση μπαλών, το κέντρο των οποίων θα είναι τα σημεία που αναπαριστούν τα δεδομένα. Σε αυτήν την ενότητα, υποθέτουμε ότι το $\mathbb{X}_n = \{x_0, x_1, ..., x_n\}$ αποτελεί υποσύνολο του $\mathbb{R}^d$, των οποίων η κατά κατανομή δειγματοληψία έγινε με βάση κάποιο μέτρο πιθανότητας μ, με συμπαγές στήριγμα $M \subset \mathbb{R}^d$. Η γενική στρατηγική για τη εξαγωγή τοπολογικής πληροφορίας σχετικά με το M από το μ αποτελείται από μία διαδικασία δύο βημάτων, η οποία περιγράφεται συνοπτικά παρακάτω:

1. Το $\mathbb{X}_n$ καλύπτεται από μία ένωση μπαλών σταθερής ακτίνας με κέντρο τα x_i. Δεδομένων κάποιων υποθέσεων για την κανονικότητα του M, μπορούμε να συσχετίσουμε την τοπολογία αυτής της ένωσης μπαλών με αυτή του M.

2. Από πρακτικής και αλγοριθμικής άποψης, οι τοπολογικές ιδιότητες του M εξάγονται από το νεύρο της ένωσης μπαλών, αξιοποιώντας το Θεώρημα Νεύρου.

Σε αυτό το πλαίσιο, είναι πιθανό να συγκρίνουμε χώρους μέσω της *ισοδυναμίας ισοτοπίας*. Η ισοδυναμία ισοτοπίας αποτελεί ισχυρότερη έννοια του ομοιομορφισμού.

2.3.1 Ορισμός

Έστω δύο χώροι $X \subseteq \mathbb{R}^d$ και $Y \subseteq \mathbb{R}^d$. Οι X και Y θα λέγονται *ισοτοπικοί* αν υπάρχει συνεχής οικογένεια ομοιομορφισμών $H : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$, η H είναι συνεχής έτσι ώστε για κάθε $t \in [0, 1]$, με $H_\tau = H(t, \dot{}) : \mathbb{R}^d \to \mathbb{R}^d$ να είναι ομοιομορφισμός, H_0 να είναι η ταυτοτική απεικόνιση στον $\mathbb{R}^d$ και $H_1(X) = Y$.

Παρατήρηση 2.3.1.1

Αν οι X και Y είναι ισοτοπικοί τότε θα είναι και ομοιομορφικοί.

Παρατήρηση 2.3.1.2

Το αντίστροφο της παρατήρησης 2.3.1.1 δεν ισχύει. Θα δώσουμε ένα αντιπαράδειγμα. Έστω ότι έχουμε δύο κύκλους στον $\mathbb{R}^3$, με τον έναν να έχει κόμπο και τον άλλο όχι. Οι κύκλοι

Ένωση μπαλών και συναρτήσεις απόστασης

Δεδομένου ενός συμπαγούς υποσυνόλου K του $\mathbb{R}^d$ και ενός μη αρνητικού πραγματικού αριθμού, έστω r, η ένωση των μπαλών ακτίνας r με κέντρο στο K, έστω $K^r = \cup_{x\in K} B(x,r)$, η οποία καλείται r-μετατόπιση του K, είναι το r-υποεπίπεδο σύνολο της συνάρτησης απόστασης $d_K : \mathbb{R}^d \to \mathbb{R}$ που ορίζεται από την $d_k(x) = \inf_{y\in K} ||x-y||$. Με άλλα λόγια, $K^r = d_k^{-1}([0,r])$.

Αυτή η παρατήρηση μας επιτρέπει να χρησιμοποιούμε διαφορετικές ιδιότητες της συνάρτησης απόστασης και να συγκρίνουμε την τοπολογία των μετατοπίσεων συμπαγών συνόλων που βρίσκονται κοντά μεταξύ τους και σέβονται την απόσταση Hausdorff. Σε αυτό το σημείο θα χρειαστεί να υπενθυμίσουμε τον ορισμό της απόστασης Hausdorff.

2.3.2 Ορισμός

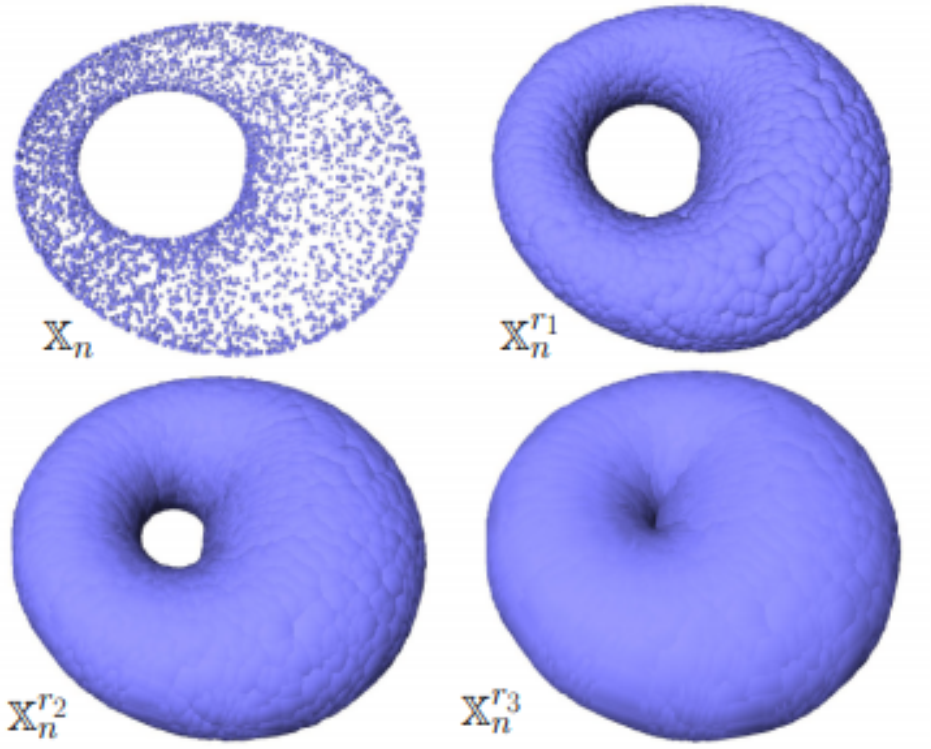
Έστω K και K' δύο συμπαγή υποσύνολα του $\mathbb{R}^d$. Ορίζουμε την *απόσταση Hausdorff* μεταξύ των συμπαγών K και K' ως εξής

$$d_H(K,K') = ||d_K - d_{K'}||_\infty = \sup_{x\in\mathbb{R}^d} |d_K - d_{K'}|.$$

Στην περίπτωσή μας, τα προαναφερθέντα συμπαγή σύνολα είναι τα σύνολα δεδομένων $\mathbb{X}_n$ και M το στήριγμα του μέτρου πιθανότητας μ. Όταν το M είναι λεία συμπαγής πολλαπλότητα, υπό κάποιες συνθήκες σχετικά με την απόσταση Hausdorff $d_H(\mathbb{X}, M)$, για κάποιο καλώς επιλεγμένο r, οι μετατοπίσεις του $\mathbb{X}_n$ είναι ομοτοπικά ισοδύναμες με το στήριγμα M. [0]

Στην ακόλουθη εικόνα βλέπουμε μία απεικόνιση των παραπάνω [0]

Αυτά τα αποτελέσματα επεκτείνονται σε μεγαλύτερες κλάσεις συμπαγών συνόλων γεγονός το οποίο οδηγεί σε δυνατότερα αποτελέσματα σχετικά με τα συμπεράσματα που εξάγονται από τον τύπο ισοτοπίας των μετατοπίσεων του M. [0] Επιπλέον, τα παραπάνω οδηγούν σε αποτελέσματα σχετικά με την εκτίμηση των λοιπών γεωμετρικών και διαφορικών ποσοτήτων όπως κανονικών σχημάτων [0], καμπυλοτήτων [0] ή μέτρα συνόρων



Σχήμα 2.7: Το παράδειγμα ενός νέφους σημείων $\mathbb{X}_n$ που έχουν επιλεγεί από την επιφάνεια ενός τόρου στον $\mathbb{R}^d$ (επάνω αριστερά) και τις μετατοπίσεις του για διάφορες τιμές της ακτίνας r_1, r_2, r_3, με $r_1 \leq r_2 \leq r_3$. [0]

[0] υπό ορισμένες υποθέσεις για την απόσταση Hausdorff μεταξύ του υποβόσκοντος σχήματος και του συνόλου δεδομένων.

Τα παραπάνω στηρίζονται στην 1-ημικοιλότητα του τετραγώνου της συνάρτησης απόστασης d_K^2, έτσι ώστε η κυρτότητα της συνάρτησης $x \to ||x||^2 - d_K^2(x)$ και μπορούν με φυσικό τρόπο να διατυπωθούν μέσω του ακόλουθου γενικού πλαισίου.

2.3.3 Ορισμός

Μια συνάρτηση θα ονομάζεται πλήρης αν η προεικόνα κάθε συμπαγούς συνόλου στο $\mathbb{R}$ είναι συμπαγές σύνολο στον $\mathbb{R}^d$.

2.3.4 Ορισμός

Μία συνάρτηση $\varphi : \mathbb{R}^d \to \mathbb{R}_+$ είναι σχεδόν-απόσταση αν είναι πλήρης και η $x \to ||x||^2 - \varphi^2(x)$ είναι κυρτή.

Λόγω της ημικοιλότητας της, μία συνάρτηση σχεδόν-απόστασης φ έχει μία καλά ορισμένη, αλλά όχι συνεχή βαθμίδα (gradient) $\nabla\varphi : \mathbb{R}^d \to \mathbb{R}^d$ που μπορεί να ενσωματωθεί σε μία συνεχή ροή που επιτρέπει την παρακολούθηση της εξέλιξης της τοπολογίας των συνόλων υποεπιπέδων και τη σύγκρισή της με αυτά των συνόλων υποεπιπέδων των κοντινών συναρτήσεων σχεδόν-απόστασης.

2.3.5 Ορισμός

Έστω φ είναι μία συνάρτηση σχεδόν-απόστασης και έστω $\varphi^r = \varphi^{-1}([0, r])$ ένα σύνολο r-υποεπιπέδων της φ. Θα ονομάζουμε ένα σημείο $x \in \mathbb{R}^d$ α-κρίσιμο αν $||\nabla_x\varphi|| \leq \alpha$.

2.3.6 Ορισμός

Έστω φ είναι μία συνάρτηση σχεδόν-απόστασης και έστω $\varphi^r = \varphi^{-1}([0, r])$ ένα σύνολο r-υποεπιπέδων της φ. Το μέγεθος αδύναμων χαρακτηριστικών (weak feature size) της φ στο r είναι το ελάχιστο $r' > 0$ το οποίο δεν έχει καμία κρίσιμη τιμή μεταξύ των r και $r + r'$. Συμβολίζουμε με $wfs_\varphi(r)$. Για κάθε $0 < \alpha < 1$ ορίζουμε ως α-έκταση της φ το μέγιστο r για το οποίο το $\varphi^{-1}((0, r])$ δεν περιέχει κανένα α-κρίσιμο σημείο.

2.3.7 Λήμμα (Grove (1993))

Έστω φ μία συνάρτηση σχεδόν-απόστασης και r_1, r_2 δύο θετικοι αριθμοί με $r_1 < r_2$, τέτοια ώστε η φ να μην έχει 0-κρίσιμα σημεία, δηλαδή σημεία x για τα οποία $\nabla\varphi(x) = 0$, στο υποσύνολο $\varphi^{-1}([r_1, r_2])$. Τότε όλα τα σύνολα υποεπιπέδων $\varphi^{-1}([0, r])$ είναι ισοτοπικά για κάθε $r \in [r_1, r_2]$. [0]

Παρατήρηση 2.3.7

Άμεση συνέπεια του παραπάνω λήμματος είναι ότι όλα τα σύνολα υποεπιπέδων της φ μεταξύ των r και $r + wfs_\varphi(r)$ έχουν την ίδια τοπολογία.

Το παρακάτω θεώρημα παρέχει μία σύνδεση μεταξύ της τοπολογίας των συνόλων των υποεπιπέδων των κοντινών συναρτήσεων σχεδόν-απόστασης. [0]

2.3.8 Θεώρημα

Έστω φ και ψ δύο συναρτήσεις σχεδόν-απόστασης τέτοιες ώστε $||\varphi - \psi||_\infty < \varepsilon$ με έκταση, $reach_a(\varphi) \geq R$, για κάποιους θετικούς αριθμούς ε και α. Τότε για κάθε $r \in \left[\frac{4\varepsilon}{\alpha^2}, R - 3\varepsilon\right]$ και κάθε $\eta \in (0, R)$, τα σύνολα των υποεπιπέδων ψ^r και φ^η είναι ομοτοπικά ισοδύναμα όταν:

$$\varepsilon \leq \frac{R}{5 + \frac{4}{\alpha^2}}$$

Διατηρώντας της παραπάνω υποθέσεις, αν αυτές είναι ελαφρά πιο απαιτητικές από τεχνικής άποψης, το Θεώρημα Ανακατασκευής (Reconstruction Theorem) μπορεί να επεκταθεί με τέτοιο τρόπο ώστε να αποδεικνύει ότι τα σύνολα υποεπιπέδων είναι όντως ομοιομορφικά και ακόμα ισοτοπικά. [0], [0]

Επιστρέφοντας στις υποθέσεις μας, αν επιπλέον θεωρήσουμε την $\varphi = d_M$ και την $\psi = d_{\mathbb{X}_n}$, της συναρτήσεις δηλαδή απόστασης στο στήριγμα M του μέτρου μ και στο σύνολο δεδομένων $\mathbb{X}_n$, υπό την υπόθεση ότι $reach_a(d_M) \geq R$ μπορεί να ερμηνευτεί ως υπόθεση κανονικότητας επί του M. Για παράδειγμα αν το M είναι μία λεία συμπαγής πολυπλοκότητα τότε το $reach_O(\emptyset)$ είναι πάντοτε θετικό και γνωστό ως έκταση του M. [0] Το Θεώρημα Ανακατασκευής συνδυαστικά με το Θεώρημα Νεύρου μας δίνουν ότι, για καλά επιλεγμένες τιμές των r και η, τα η-μετατοπίσεις του M είναι ομοτοπικά ισοδύναμες με το νεύρο της ένωσης μπαλών ακτίνας r με κέντρο στο $\mathbb{X}_n$, με άλλα λόγια το σύμπλεγμα Cech, το οποίο συμβολίζεται με $Cech_r(\mathbb{X}_n)$.

Από στατιστικής άποψης, το κύριο πλεονέκτημα αυτών των αποτελεσμάτων συμπεριλαμβανομένης της απόστασης Hausdorff είναι ότι η εκτίμηση των τοπολογικών ποσοτήτων που έχουμε θεωρήσει, μπορεί να υποστηρίξει την εκτίμηση απαντήσεων σε ερωτήματα που έχουν ευρέως μελετηθεί.

Τα παραπάνω αποτελέσματα παρέχουν ένα καλά θεμελιωμένο μαθηματικά πλαίσιο για την εξαγωγή της τοπολογίας των σχημάτων από κάποιο πλεγματικό σύμπλεγμα που έχει δομηθεί επί ενός προσεγγιστικού πεπερασμένου δείγματος. Ωστόσο, αν το εξετάσουμε πιο πρακτικά δύο ζητήματα αναδύονται. Πρώτον, το Θεώρημα Ανακατασκευής απαιτεί την υπόθεση της κανονικότητας μέσω μίας προϋπόθεσης για την α-έκταση που μπορεί να μην ικανοποιείται πάντα και την επιλογή μια ακτίνας r για την μπάλα που θα χρησιμοποιήσουμε για την κατασκευή του Cech συμπλέγματος, $Cech_r(\mathbb{X}_n)$. Δεύτερον, το σύμπλεγμα $Cech_r(\mathbb{X}_n)$ παρέχει μία τοπολογικά συνεπή περίληψη των δεδομένων, μέσω ενός πλεγματικού συμπλέγματος του οποίου η συνήθης χρήση δεν ταιριάζει στην επεξεργασία των δεδομένων. Υπό αυτό το πρίσμα, συχνά ψάχνουμε για εύκολα αξιοποιήσιμες περιγραφές τοπολογιών, κυρίων αριθμητικές, οι οποίες μπορούν εύκολα να υπολογιστεί από το σύμπλεγμα. Το πρώτο ζήτημα που προκύπτει θα μελετηθεί εκτενέστερα στην εισαγωγή στην Εμμένουσα Ομολογία (Persistent Homology) που θα γίνει σε επόμενη ενότητα, ενώ το δεύτερο ζήτημα που προκύπτει αντιμετωπίζεται θεωρώντας την ομολογία των πλεγματικών συμπλεγμάτων που θα θεωρήσουμε στην επόμενη παράγραφο.

Ομολογία σε ένα κέλυφος

Η ομολογία είναι μία κλασική έννοια στην αλγεβρική τοπολογία που μας δίνει ένα ισχυρό εργαλείο για την τυποποίηση και το χειρισμό των εννοιών των τοπολογικών ιδιοτήτων ενός τοπολογικού χώρου ή ενός πλεγματικού συμπλέγματος από αλγεβρικής άποψης. Για κάθε διάσταση k, οι k-διάστατες ¨τρύπες' αναπαρίστανται από ένα διανυσματικό χώρο, έστω H_k, του οποίου η διάσταση είναι διαισθητικά ο αριθμός των ανεξάρτητων χαρακτηριστικών. Για παράδειγμα, η 0-διάστατη ομάδα ομολογίας H_0 αναπαριστά τα ενωμένα συστατικά του συμπλέγματος, η 1-διάστατη ομάδα ομολογίας H_1 αναπαριστά του μονοδιάστατους κόμβους, η 2-διάστατη ομάδα ομολογίας H_2 αναπαριστά τις δισδιάστατες κοιλότητες κ.ο.κ.

Για την αποφυγή τεχνικών λεπτομερειών και δυσκολιών, στο εξής θα περιορίσουμε την εισαγωγή στην ομολογία στο ελάχιστο που είναι απαραίτητο για την παρούσα εργασία. Πιο συγκεκριμένα, θα περιοριστούμε σε ομολογίες με συντελεστές στο $\mathbb{Z}_2$, δηλαδή στο πεδίο που περιέχει δύο μόνο στοιχεία, τα 0 και 1 για τα οποία ισχύει ότι $1 + 1 = 0$, που φαίνεται

να είναι το πιο απλό διαισθητικά από γεωμετρικής άποψης. Από την άλλη πλευρά, όλες οι έννοιες και τα αποτελέσματα που θα παρουσιαστούν στη συνέχεια με φυσικό τρόπο επεκτείνονται σε ομολογία με συντελεστές σε κάθε πεδίο. Για την πιο ολοκληρωμένη και κατανοητή εισαγωγή στην ομολογία αναφερόμαστε στον Hatcher (2001) [0] και για μία πλήρη εισαγωγή στην εφαρμοσμένη αλγεβρική τοπολογία και τη σύνδεσή της με την ανάλυση δεδομένων στον Ghrist (2017) [0].

Έστω K ένα πεπερασμένο πλεγματικό σύμπλεγμα και k ένας μη αρνητικός ακέραιος αριθμός. Ο χώρος των k-αλυσίδων στον K. Τότε $C_k(K)$ είναι το σύνολο του οποίου τα στοιχεία είναι τυπικά (πεπερασμένα) αθροίσματα k-πλεγμάτων του K. Για να είμαστε πιο ακριβείς, αν $\{\sigma_1, \sigma_2, \ldots, \sigma_p\}$ είναι το σύνολο των k-πλεγμάτων του K, τότε οποιαδήποτε k-αλυσίδα μπορεί για $\varepsilon_i \in \mathbb{Z}_2$ να γράφεται ως εξής:

$$c = \sum_{i=1}^{p} \varepsilon_i \sigma_i$$

Επιπλέον αν $c' = \sum_{i=1}^{p} \varepsilon'_i \sigma_i$ κάποια άλλη k-αλυσίδα και $\lambda \in \mathbb{Z}_2$, τότε:

1. Το άθροισμα δύο k-αλυσίδων ορίζεται ως εξής:

$$c + c' = \sum_{i=1}^{p} (\varepsilon_i + \varepsilon'_i)\sigma_\iota$$

2. Ο βαθμωτός πολλαπλασιασμός ορίζεται ως εξής:

$$\lambda c = \sum_{i=1}^{p} (\lambda \varepsilon_i)\sigma_i$$

Με αυτόν τον τρόπο συμπεραίνουμε ότι ο $C_k(K)$ είναι διανυσματικός χώρος με συντελεστές στο $\mathbb{Z}_2$. Γεωμετρικά μπορούμε να δούμε μία k-αλυσίδα ως μία πεπερασμένη συλλογή k-πλεγμάτων και το άθροισμα k-αλυσίδων ως τη συμμετρική διαφορά των δύο συλλογών αντίστοιχα. Υπενθυμίζουμε σε αυτό το σημείο ότι η συμμετρική διαφορά δύο συνόλων, έστω A και B είναι το σύνολο $A \triangle B = (A \setminus B) \cup (B \setminus A)$.

Το σύνορο ενός k-πλέγματος $\sigma = [v_0, v_1, \ldots, v_k]$ είναι η $k-1$-αλυσίδα:

$$\partial_k(\sigma) = \sum_{i=0}^{k}(-1)^i[v_0, v_1, \ldots, \hat{v}_i, \ldots, v_k]$$

όπου $[v_0, v_1, \ldots, \hat{v}_i, \ldots, v_k]$ είναι το k-πλέγμα που επάγεται από όλα τα διανύσματα εκτός από το v_i. Σημειώνεται επίσης ότι επειδή βρισκόμαστε στο $\mathbb{Z}_2$ ισχύει ότι $-1 = 1$ και έτσι $(-1)^i = 1$, για κάθε i. Τα k-πλέγματα δημιουργούν μία βάση για το $C_k(K)$ και το σύνορο ∂_k επεκτείνεται ως μία γραμμική απεικόνιση από το $C_k(K)$ στο $C_{k-1}(K)$ και ονομάζεται τελεστής συνόρου. Ο πυρήνας $Z_k(K) = \{c \in C_k(K) : \partial_k = 0\}$ του συνόρου ∂_k ονομάζεται χώρος των k-κύκλων του K και η εικόνα

$$B_k(K) = \{c \in C_k(K) : \exists c' \in C_{k+1}(K), \partial_{k+1}(c') = c\}$$

του ∂_{k+1} θα ονομάζεται χώρος των k-συνόρων του K. Οι τελεστές συνόρων ικανοποιούν τη θεμελιώδη ιδιότητα που διατυπώνεται παρακάτω:

$$\partial_{k-1} \circ \partial_k \equiv 0, \forall k \geq 1$$

Με άλλα λόγια, κάθε k-σύνορο είναι ένας k-κύκλος και έτσι $B_k(K) \subseteq Z_k(K) \subseteq C_k(K)$. Οι προαναφερθείσες έννοιες απεικονίζονται στο Σχήμα 2.8.

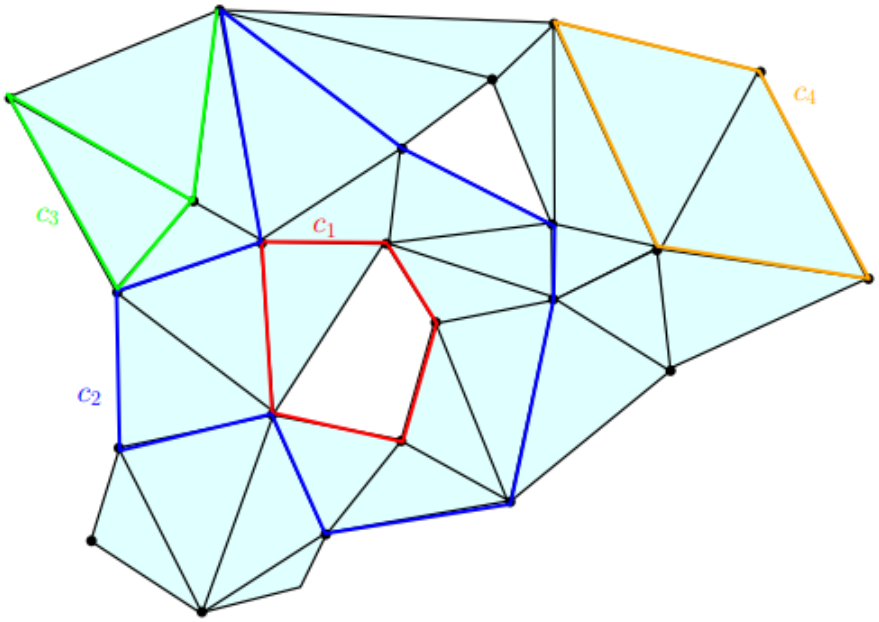
2.3.9 Ορισμός (Ομάδα Πλεγματικής Ομολογίας)

Η ομάδα k-οστής (πλεγματικής) ομολογίας του K ορίζεται ώς:

$$H_k(K) = Z_k(K)/B_k(K)$$

Παρατήρηση 2.3.9

Όταν δουλεύουμε στο $\mathbb{Z}_2$ (ή σε κάθε σώμα $\mathbb{Z}_p$, όπου p πρώτος αριθμός) τότε μπορεί κανείς να δεί ότι αυτή η ομάδα ομολογίας έχει τη δομή διανυσματικού χώρου, αφού είναι πηλίκο διανυσματικών χώρων.



Σχήμα 2.8: Για κάποια παραδείγματα αλυσίδων, κύκλων και συνόρων σε ένα δισδιάστατο σύμπλεγμα K: τα c_1, c_2 και c_4 είναι 1-κύκλοι, το c_3 είναι 1-αλυσίδα αλλά όχι 1-κύκλος, το c_4 είναι 1-σύνορο, πιο συγκεκριμένα το σύνορο των 2-αλυσίδων που προκύπτει από το άθροισμα των δύο τριγώνων που περιβάλλονται γύρω από το c_4. Οι κύκλοι c_1 και c_2 επάγονται από το ίδιο στοιχείο του $H_1(K)$ και η διαφορά τους είναι μία 2-αλυσίδα που αναπαρίσταται από την ένωση των τριγώνων που περιβάλλονται από την ένωση των c_1 και c_2. [0]

2.3.10 Ορισμός (Αριθμοί Betti)

Ο k-οστός αριθμός Betti του K είναι η διάσταση $\beta_k(K) = dim H_k(K)$ του διανυσματικού χώρου $H_k(K)$.

Δύο κύκλοι c και $c' \in Z_k(K)$ είναι ομόλογοι αν διαφέρουν κατά ένα σύνορο, δηλαδή αν υπάρχει μία $k+1$-αλυσίδα, έστω d, τέτοια ώστε $c = c' + \partial_{k+1}(d)$. Δύο ομόλογοι κύκλοι αναδεικνύουν το ίδιο στοιχείο του H_k. Με άλλα λόγια , τα στοιχεία του $H_k(K)$ αποτελούν κλάσεις ισοδυναμίας ομόλογων κύκλων.

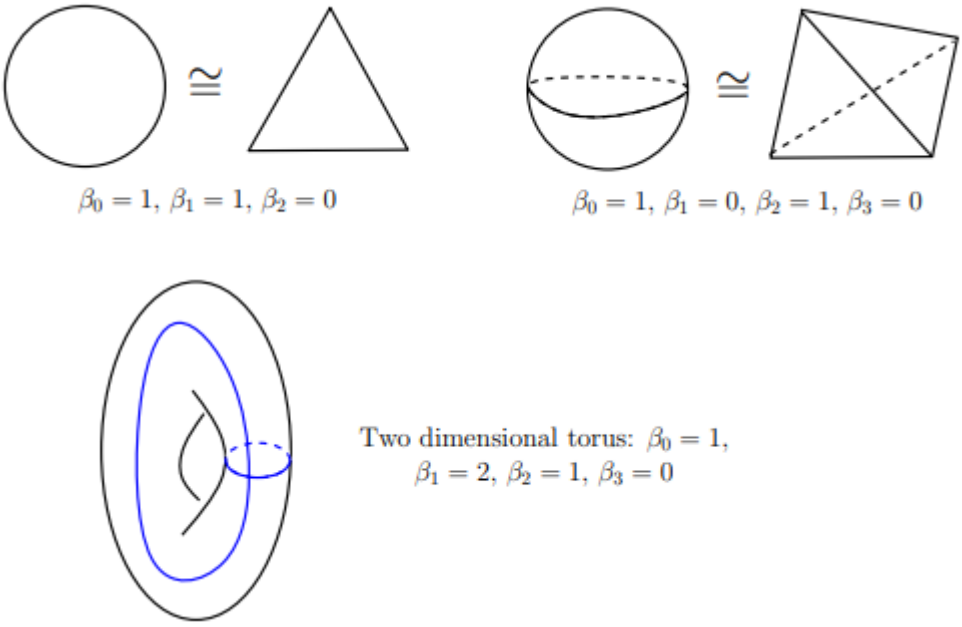
Οι ομάδες πλεγματικής ομολογία και οι αριθμοί Betti είναι τοπολογικές

αναλλοίωτες. Αν K, K' είναι δύο πλεγματικά συμπλέγματα των οποίων οι γεωμετρικές αναπαραστάσεις είναι ομοτοπικά ισοδύναμες, τότε οι ομάδες ομολογίας είναι ισομορφικές και οι αριθμοί Betti, που τους αντιστοιχούν, είναι οι ίδιοι.

Η *μοναδιαία ομολογία* είναι μία διαφορετική έννοια ομολογίας που επιτρέπει τη μελέτη μεγαλύτερων κλάσεων τοπολογικών χώρων. Ορίζεται για οποιονδήποτε τοπολογικό χώρο X με παρόμοιο τρόπο όπως στην πλεγματική ομολογία. Όμως η έννοια του πλέγματος αντικαθίσταται από το μοναδιαίο πλέγμα που είναι μία συνεχής απεικόνιση $\sigma : \Delta_k \to X$, όπου Δ_k είναι το τυπικό k-διάστατο πλέγμα. Ο χώρος των k-αλυσίδων είναι ένας διανυσματικός χώρος που επάγεται από τα k-διάστατα μοναδιαία πλέγματα και το σύνορο του ενός πλέγματος, έστω σ, το οποίο ορίζεται ως το (εναλλακτικό) άθροισμα των περιορισμών του σ στο $(k-1)$-διάστατων πλευρών του Δ_k. Μία σημαντική παρατήρηση σχετικά με τη μοναδιαία ομολογία είναι το γεγονός ότι συμπίπτει με την πλεγματική ομολογία όταν ο X είναι ομοιομορφικός με τη γεωμετρική αναπαράσταση ενός πλεγματικού συμπλέγματος. Το παραπάνω μας επιτρέπει στη συνέχεια να συζητάμε χωρίς να διαχωρίζουμε την πλεγματική από τη μοναδιαία ομολογία για τοπολογικούς χώρους και πλεγματικά συμπλέγματα.

Παρατηρούμε, επίσης, ότι αν η $f : X \to Y$ είναι μία συνεχής απεικόνιση, τότε για οποιοδήποτε μοναδιαίο πλέγμα $\sigma : \Delta_k \to X$ στον X, η $f \circ \sigma : \Delta_k \to Y$ είναι ένα μοναδιαίο πλέγμα στον Y. Εύκολα καταλήγουμε στο συμπέρασμα ότι οι συνεχείς απεικονίσεις μεταξύ τοπολογικών χώρων εισάγουν με κανονικό τρόπο ομοιομορφισμούς μεταξύ των ομάδων ομολογίας. Πιο συγκεκριμένα, αν η f είναι ένας ομοιομορφισμός ή μία ομοτοπική ισοδυναμία, τότε αυτή εισάγει έναν ισομορφισμό μεταξύ των $H_k(X)$ και $H_k(Y)$, για κάθε μη αρνητικό ακέραιο αριθμό k. Για παράδειγμα, έπεται από το Θεώρημα Νεύρου ότι κάθε σύνολο σημείων του $X \subset \mathbb{R}^d$ και κάθε κάθε $r > 0$, η r-μετατόπιση X^r και το σύμπλεγμα $Cech$, $Cech_r(X)$, έχουν ισομορφικές ομάδες ομολογία και τους ίδιους αριθμούς Betti.

Κατ΄ επέκταση, το Θεώρημα Ανακτασκευής μας οδηγεί στο ακόλουθο συμπέρασμα για την εκτίμηση των αριθμών Betti.



Σχήμα 2.9: Οι αριθμοί Betti του κύκλου (πάνω αριστερά) η δισδιάστατη σφαίρα (πάνω δεξιά) και ο δισδιάστατος τόρος (κάτω). Οι μπλέ καμπύλες στον τόρο αναπαριστούν δύο ανεξάρτητους κύκλους των οποίων η κλάση ομολογίας αποτελεί βάση για την μονοδιάστατη ομάδα ομολογίας του. [0]

2.3.11 Θεώρημα

Έστω $M \subset \mathbb{R}^d$ είναι ένα συμπαγές σύνολο τέτοιο ώστε η έκταση $reach_a d_M \geq R \geq 0$, για κάποιο $a \in (0, 1)$ και έστω $\mathbb{X}$ ένα πεπερασμένο σύνολο σημείων για τα οποία ισχύει ότι:

$$d_H(M, \mathbb{X}) = \varepsilon < \frac{R}{5 + \frac{4}{a^2}}.$$

Τότε για κάθε $r \in [4\varepsilon/a^2, R - 3\varepsilon]$ και κάθε $\eta \in (0, R)$, οι αριθμοί Betti του $Cech_r(\mathbb{X})$ είναι οι ίδιοι με αυτούς του M^n. Συγκεκριμένα, αν το M είναι μία λεία m-διάστατη υποπολλαπλότητα του $\mathbb{R}^d$, τότε $\beta_k(Cech_r()\mathbb{X}) = \beta_k(M)$, για κάθε $k = 0, 1, \ldots, m$.

Από μία πιο πρακτική άποψη, αυτό το αποτέλεσμα φέρνει στο επίκεντρο τρεις δυσκολίες. Πρώτον, η υπόθεση της κανονικότητας που περιλαμβάνει

η α-έκταση του M μπορεί να είναι υπέρ το δέον περιοριστική. Δεύτερον, ο υπολογισμός του νεύρου μίας ένωσης μπαλών απαιτεί τη χρήση ενός δύσκολου κατηγορηματικού ελέγχου της κενότητας μίας πεπερασμένης ένωσης μπαλών. Τρίτον, η εκτίμηση των αριθμών Betti βασίζεται στην παράμετρο κλίμακας r, της οποίας η επιλογή μπορεί να είναι αρκετά δύσκολη.

Για να ξεπεράσουμε αυτά τα ζητήματα οι Chazal και Oudot (2008) παρουσιάζουν το ακόλουθο αποτέλεσμα που λύνει τα πρώτα δύο προβλήματα. [0]

{2.3.12 **Θεώρημα**

Έστω $M \subset \mathbb{R}^d$ ένα σύνολο τέτοιο ώστε $wfs(M) = wfs_{d_M}(0) \geq R > 0$. Επιπλέον, έστω $\mathbb{X}$ ένα πεπερασμένο σύνολο σημείων τέτοιο ώστε $d_H(M, \mathbb{X}) = \varepsilon < \frac{1}{9} wfs(M)$. Τότε για κάθε $r \in [2\varepsilon, \frac{1}{4}(wfs(M) - \varepsilon)]$ και κάθε $\eta \in (0, R)$, έχουμε:

$$\beta_k(X^\eta) = rk(H_k(Rips_r(\mathbb{X})) \to H_k(Rips_{4r}(\mathbb{X})))$$

όπου $rk(H_k(Rips_r(\mathbb{X})) \to H_k(Rips_{4r}(\mathbb{X})))$ συμβολίζει την τάξη του ομοιομορφισμού που επάγεται από τη συνεχή κανονική εμφύτευση $Rips_r(\mathbb{X}) \hookrightarrow Rips_{4r}(\mathbb{X})$.

Παρόλο που αυτό το αποτέλεσμα αφήνει αναπάντητο το ερώτημα της επιλογής της παραμέτρου κλίμακας r, αποδεικνύεται από τους Chazal και Oudot (2008) ότι υπάρχει μια πολυεπίπεδη στρατηγική που μπορεί να αξιοποιηθεί, ωστόσο αυτή είναι πέρα από τους σκοπούς της παρούσας εργασίας. [0]

2.3.1 Στατιστικές πλευρές της Ομολογιακής Συμπερασματολογίας

Σύμφωνα με τα αποτελέσματα για τη σταθερότητα που έχουν παρουσιαστεί, μια στατιστική προσέγγιση στην ομολογιακή συμπερασματολογία συνδέεται ισχυρά με το πρόβλημα της εκτίμησης του στηρίγματος και της εκτίμησης των συνόλων επιπέδων υπό την μετρική Hausdorff. Ένα μεγάλο πλήθος μεθόδων και αποτελεσμάτων είναι διαθέσιμα για την εκτίμηση του στηρίγματος της κατανομής στη στατιστική. Για παράδειγμα, ο εκτιμητής Devroye κ αι Wise (2008) ([0]) που ορίζεται επί ενός δείγματος $\mathbb{X}_n$ είναι επίσης μια μετατόπιση του $\mathbb{X}_n$. Οι αναλογίες σύγκλισης για το $\mathbb{X}_n$ και τον εκτιμητή Devroye και Wise στο στήριγμα της κατανομής για την απόσταση Hausdorff στον $\mathbb{R}^d$. Η εκτίμηση των minimax αναλογιών σύγκλισης μίας πολλαπλότητας για την απόσταση Hausdorff, που σχετίζεται ισχυρά με την τοπολογική συμπερασματολογία, έχει μελετηθεί από τους Genovese et al (2012). [0] Υπάρχει επίσης εκτενής βιβλιογραφία σχετικά με την εκτίμηση των συνόλων επιπέδων για διάφορες μετρικές ([0], [0], [0]). Όλες αυτές οι εργασίες για την εκτίμηση των στηριγμάτων και των συνόλων επιπέδων, δίνουν πληροφορίες για τη στατιστική ανάλυση των διαδικασιών της στατιστικής συμπερασματολογίας.

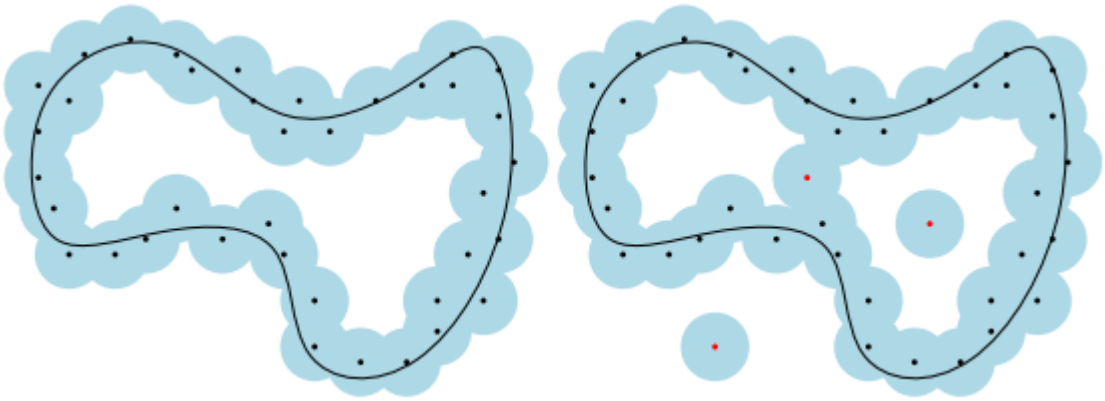
Στην εργασία των Niyogi et al. (2008), φαίνεται ότι ο τύπος ομοτοπίας για τις πολλαπλότητες Riemann, που είναι μεγαλύτερος από κάποια σταθερά που εξάγεται με μεγάλη πιθανότητα από τις μετατοπίσεις του δείγματος επάνω ή έστω πολύ κοντά στην πολλαπλότητα. Αυτή η εργασία αποτελεί πιθανά την πρώτη προσπάθεια να σκεφτούμε με όρους πιθανοτήτων το πρόβλημα της στατιστικής συμπερασματολογίας. Το αποτέλεσμα αυτής της εργασίας προκύπτει από τη συζήτηση σχετικά με την ανάκληση συστολής και συνδέεται ισχυρά με τον αριθμό ¨στοιβών' της πολλαπλότητας, με σκοπό να ελέγξουμε την απόσταση Hausdorff μεταξύ της πολλαπλότητας και του παρατηρούμενου νέφους σημείων. Η στατιστική συμπερασματιολογία είναι μία αρκετά πολύπλοκη περίπτωση. Υπό την έννοια της κατανομής των παρατηρήσεων που είναι συγκεντρωμένη γύρω από την πολλαπλότητα υπάρχει θόρυβος, κάτι που έχει επίσης μελετηθεί από τους Niyogi et al. (2008, 2011). ([0], [0]) Η υπόθεση ότι το γεωμετρικό αντικείμενο είναι μία λεία πολλαπλότητα Riemann χρησιμοποιείται στη εργασία με στόχο τον έλεγχο της απόστασης Hausdorff μεταξύ του δείγματος και της πολλαπλότητας από άποψη πιθανότητας. Αυτό δεν είναι ωστόσο απαραίτητο για το τοπολογικό μέρος του αποτελέσμτος. Λαμβάνουμε υπόψην και τα τοπολογικά αποτελέσματα που υ-

πάρχουν στις εργασίες των Chazal et al. (2009δ) και Chazal ανδ Lieutier (2008β), στο συγκεκριμένο πλαίσιο των πολλαπλοτήτων Riemann. ([0], [0] Ξεκινώντας από το αποτέλεσμα των Niyogi et al. (2008), οι minimax αναλογίες του τύπου ομοτοπίας έχουν μελετηθεί από τους Balakrishna et al. (2012) για διάφορα μοντέλα, για πολλαπλότητες Riemannian με έκταση μεγαλύτερη από μία σταθερά σε αντίθεση με κάποια στατιστική εκδοχή της εργασίας των Chaza et al. (2009δ) η οποία δεν έχει ακόμα προταθεί. ([0], [0], [0])

Οι Niyogi et al. (2008) και Bobrowski et al. (2014) πρότειναν έναν εύρωστο εκτιμητή ομολογίας για τα σύνολο επιπέδων των συναρτήσεων πυκνότητας και παλινδρόμησης, θεωρώντας ότι η συνάρτηση ένθεσης μεταξύ των εμφωλευμένων ζευγών των εκτιμώμενων συνόλων επιπέδων, με την έννοια του Θεωρήματος 2.3.12, μπορεί να βρεθεί με τη βοήθεια της προσέγγισης σύνδεσης (plug-in) από τους εκτιμητές του πυρήνα. ([0], [0])

2.3.2 Απόσταση σε μέτρο

Είναι γνωστό ότι οι μέθοδοι της Τοπολογικής Ανάλυσης Δεδομένων που βασίζονται στην απόσταση μπορεί να αποτύχουν παταγωδώς αν υπάρχουν ακραίες παρατηρήσεις. Πράγματι, όπως βλέπουμε και στο Σχήμα 2.10, με την προσθήκη ακόμα και μίας ακραίας παρατήρησης μπορεί να προκύψουν σημαντικές μεταβολές στη συνάρτηση απόστασης. Για να αντιμετωπιστεί αυτό το πρόβλημα, οι Chazal et al. (2011β) εισήγαγαν μία εναλλακτική συνάρτηση απόστασης που είναι εύρωστη στο θόρυβο, την οποία καλούν *απόσταση σε μέτρο*.



Σχήμα 2.10: Η επίδραση των ακραίων παρατηρήσεων στα σύνολα επιπέδων των συναρτήσεων απόστασης. Αν προσθέσουμε μερικές ακόμα ακραίες παρατηρήσεις σε ένα νέφος σημείων, είναι πιθανό η συνάρτηση απόστασης και η τοπολογία των μετατοπίσεών του να αλλάξουν δραματικά. [0]

Δεδομένης κάποιας κατανομής πιθανότητας, έστω P στον $\mathbb{R}^d$ και μία πραγματική παράμετρο, έστω u, για την οποία ισχύει $0 \leq u \leq 1$, η έννοια της απόστασης από το στήριγμα P μπορεί να γενικευτεί στην παρακάτω συνάρτηση:

$$\delta_{P,u} : x \in \mathbb{R}^d \mapsto inf\{t > 0, P(B(x,t)) \geq u\}$$

όπου $B(x,t)$ είναι μία κλειστή Ευκλείδεια μπάλα με κέντρο το x και ακτίνα r. Για να αποφύγουμε ζητήματα που προκύπτουν από τα σημεία ασυνέχειας της απεικόνισης $P \to \delta_{P,u}$, η συνάρτηση απόστασης σε μέτρο, έστω DTM (Distance-To-Measure), με την παράμετρο $m \in [0,1]$ και εκθέτη $r \geq 1$ ορίζεται ως εξής:

$$d_{P,m,r}(x) : x \in \mathbb{R}^d \mapsto \left(\frac{1}{m} \int_0^m \delta_{P,u}^r(x) du \right)^{1/r} \tag{2.1}$$

Μία ιδιότητα της DTM, που έχει αποδειχθεί από τους Chazal et al. (2011β), είναι η σταθερότητά της, που σέβεται τις διαταραχές του P για τη μετρική Wasserstein. [0] Πιο συγκεκριμένα, η απεικόνιση $P \mapsto \delta_{P,u}$ είναι $m^{1/r}$-Lipschitz, ή με άλλα λόγια αν P και $\tilde{P}$ είναι δύο κατανομές πιθανότητας στον $\mathbb{R}^d$, τότε έχουμε:

$$||d_{P,m,r} - d_{\tilde{P},m,r}||_\infty \leq m^{-1/r} W_r(P, \tilde{P}) \tag{2.2}$$

όπου W_r είναι η απόσταση Wasserstein για την Ευκλείδεια μετρική στον $\mathbb{R}^d$ με εκθέτη r. Αυτή η ιδιότητα, μας δίνει ότι η DTM που σχετίζεται με κοντινές κατανομές με την μετρική Wasserstein έχουν κοντινά σύνολα επιπέδων. Επιπλέον, όταν $r = 2$, Η συνάρτηση $d_{P,M,2}^2$ είναι ημικοίλη, το οποίο διασφαλίζει ισχυρές ιδιότητες κανονικότητας επί της γεωμετρίας των συνόλων υποεπιπέδων της. Αποδεικνύεται από τους Chazal et al. (2011β) ότι χρησιμοποιώντας αυτές τις ιδιότητες, υπό κάποιες γενικές υποθέσεις, αν η $\tilde{P}$ είναι μία κατανομή πιθανότητας που προσεγγίζει το P, τότε τα σύνολα υποεπιπέδων της συνάρτησης $d_{\tilde{P},m,2}$ παρέχουν μία τοπολογικά ορθή προσέγγιση στο στήριγμα του P.

Στην πράξη το μέτρο P είναι γνωστό συνήθως μόνο στην περίπτωση που το σύνολο των παρατηρήσεων είναι πεπερασμένο, έστω $Xn = \{X_1, X_2, \ldots, X_n\}$ που περιέχει στοιχεία από το P, συνιστά στην ανάγκη διερευνήσουμε την προσέγγιση του DTM. Με φυσικό τρόπο προκύπτει η ιδέα να εκτιμήσουμε την DTM του X_n είναι να συνδέσουμε το εμπειρικό μέτρο P_n αντί του P από τον ορισμό της DTM. Αυτή η στρατηγική ¨σύνδεσης' αξιοποιείται στον υπολογισμό της απόστασης του εμπειρικού μέτρου ($DTEM$). Για $m = \frac{k}{n}$, η $DTEM$ ικανοποιεί το εξής:

$$d_{P_n,\frac{k}{n},r}^r(x) := \frac{1}{k} \sum_{j=1}^{k} ||x - \mathbb{X}_n||_{(j)}^r$$

όπου με $||x - \mathbb{X}_n||_{(j)}$ συμβολίζουμε την απόσταση μεταξύ του x και του j-οστού γειτονικού σημείου στο $\{X_1, X_2, \ldots, X_n$. Αυτή είναι μία εύκολα υπολογίσιμη ποσότητα στην πράξη αφού απαιτούνται οι αποστάσεις μεταξύ του x και των σημείων του δείγματος. Η σύγκλιση του $DTEM$

στο DTM μελετήθηκε από τους Chazal et al. (2014a) και Chazal et al. (2016b). ([0], [0])

Η εισαγωγή του DTM οδήγησε εργασίες και εφαρμογές σε διάφορους τομείς όπως μεταξύ άλλων η Τοπολογική Ανάλυση Δεδομένων ([0]), η ανάλυση εντοπισμού GPS ([0]), η εκτίμηση πυκνότητας ([0]), οι έλεγχοι υποθέσεων ([0]), η συσταδοποίηση ([0]). Επιπλέον, υπάρχουν εργασίες σχετικά με προσεγγίσεις, γενικεύσεις και μεταβλητές της DTM. ([0], [0], [0]).

2.4 Εμμένουσα Ομολογία (Persistent Homology)

Η Εμμένουσα Ομολογία είναι ένα πολύ ισχυρό εργαλείο για τον υπολογισμό, τη μελέτη και την κωδικοποίηση με επαρκή τρόπο των πολύπλοκων τοπολογικών ιδιοτήτων των εμφωλευμένων οικογενειών πλεγματικών συμπλεγμάτων και τοπολογικών χώρων. Μάλιστα, όχι μόνο παρέχει αποδοτικούς αλγορίθμους για τον υπολογισμό των αριθμών Betti κάθε συμπλέγματος στις οικογένειες που έχουμε θεωρήσει, όπως έχουμε δει ότι απαιτείται για τη στατιστική συμπερασματολογία, αλλά επιπλέον κωδικοποιεί την εξέλιξη των ομάδων ομολογίας των εμφωλευμένων συμπλεγμάτων στις διάφορες κλίμακες.

2.4.1 Διηθήσεις

Η διήθηση ενός πλεγματικού συμπλέγματος K είναι μία εμφωλευμένη οικογένεια υποσυμπλεγμάτων $(K_r)_{r\in T}$, όπου $T \subseteq \mathbb{R}$, τέτοιο ώστε για κάθε $r, r' \in T$, με $r \leq r'$, έχουμε $K_r \subseteq K_{r'}$ και $K = \cup_{r\in T} K_r$. Το υποσύνολο T μπορεί να είναι είτε πεπερασμένο, είτε άπειρο. Γενικότερα, μία διήθηση ενός τοπολογικού χώρου M είναι μία εμφωλευμένη οικογένεια υπόχωρων $(M_r)_{r\in T}$ όπου $T \subseteq \mathbb{R}$, τέτοιο ώστε για κάθε $r, r' \in T$, με $r \leq r'$, έχουμε $M_r \subseteq M_{r'}$ και $M = \cup_{r\in T} M_r$. Για παράδειγμα αν $f : \mathbb{M} \to \mathbb{R}$ είναι μία συνάρτηση. Τότε η οικογένεια $M_r = f^{-1}((-\infty, r])$, όπου $r \in \mathbb{R}$ ορίζει μία διήθηση που ονομάζεται διήθηση συνόλου υποεπιπέδων της f.

Πρακτικά, η παράμετρος $r \in T$ μπορεί συχνά να ερμηνευθεί ως μία παράμετρος κλίμακας. Επιπλέον οι διηθήσεις που παραδοσιακά χρησιμοποιούνται στην Τοπολογική Ανάλυση Δεδομένων συνήθως ανήκουν σε μία από της παρακάτω οικογένειες:

Κατασκευή διηθήσεων επί των δεδομένων

Έστω ένας συμπαγής μετρικός χώρος (M, ρ), ένα υποσύνολό του $\mathbb{X}$ οι οικογένειες των πλεγμάτων Rips – Vietoris, που συμβολίζεται με $(Rips_r((X))_{r\in\mathbb{R}}$ και τα συμπλέγματα Cech, που συμβολίζονται με $(Cech_r((X))_{r\in\mathbb{R}}$, είναι διηθήσεις. Σημειώνεται ότι κάνουμε την παραδοχή ότι για $r \leq 0$, $Rips_r(\mathbb{R}) = Cech_r(\mathbb{X}) = \emptyset$. Εδώ η παράμετρος r μπορεί να ερμηνευθεί

ως μία ανάλυση με βάση την οποία θεωρούμε τα δεδομένα.

Για παράδειγμα, αν το $\mathbb{X}$ είναι ένα νέφος δεδομένων στον $\mathbb{R}^d$, βασιζόμενοι στο Θεώρημα Νεύρου, διήθηση $(Cech_r((X))_{r\in\mathbb{R}}$ κωδικοποιεί την τοπολογία ολόκληρης της οικογένειας των ενώσεων των μπαλών, δηλαδή $\mathbb{X}^r = \cup_{x\in\mathbb{X}} B(x,r)$, καθώς πηγαίνει από το 0 στο ∞. Εν γένει η έννοια της διήθησης είναι αρκετά ευέλικτη, και κατ' επέκταση διάφορες διηθήσεις έχουν θεωρηθεί στη βιβλιογραφία και μπορούν να κατασκευαστούν επί των δεδομένων, όπως είναι για παράδειγμα το λεγόμενο σύμπλεγμα του μάρτυρα witness complex είναι αρκετά δημοφιλές στην Τοπολογική Ανάλυση Δεδομένων. [0]

Διηθήσεις συνόλων υποεπιπέδων

Οι συναρτήσεις που ορίζονται πάνω στις ακμές των πλεγματικών συμπλεγμάτων ωθεί στην παρατήρηση ενός άλλου σημαντικού παραδείγματος διήθησης, το οποίο και θα αναλύσουμε. Έστω K ένα πλεγματικό σύμπλεγμα με σύνολο κορυφών V και μία $f : V \to \mathbb{R}$. Τότε η f μπορεί να επεκταθεί σε όλα τα πλέγματα του K ως εξής $f([v_0, v_1, \ldots, v_k]) = max\{f(v_i) : i = 1, 2, \ldots, k\}$ για κάθε πλέγμα $\sigma = [v_0, v_1, \ldots, v_k] \in K$. Τότε η ομάδα υποπλεγμάτων $K_r = \{\sigma \in K : f(\sigma) \leq r\}$ ορίζει μία διήθηση που καλείται *διήθηση του συνόλου υποεπιπέδων της* f. Με παρόμοιο τρόπο ορίζουμε και τη διήθηση του συνόλου υπερεπιπέδων της f.

Στην πράξη, ακόμα και αν το σύνολο δεικτών είναι άπειρο, όλες οι σχετικές διηθήσεις κατασκευάζονται πάνω σε πεπερασμένα σύνολα και είναι και οι ίδιες πεπερασμένες. Για παράδειγμα, όταν το σύνολο δεδομένων $\mathbb{X}$ είναι πεπερασμένο, τότε το σύμπλεγμα Vietoris - Rips, $Rips_r(\mathbb{X})$ μεταβάλλεται μόνο για ένα πεπερασμένο αριθμό δεικτών r. Με αυτόν τον τρόπο είναι εύκολος ο χειρισμός τους μέσω κάποιας αλγοριθμικής προσέγγισης. [0]

2.4.2 Αρχικά παραδείγματα

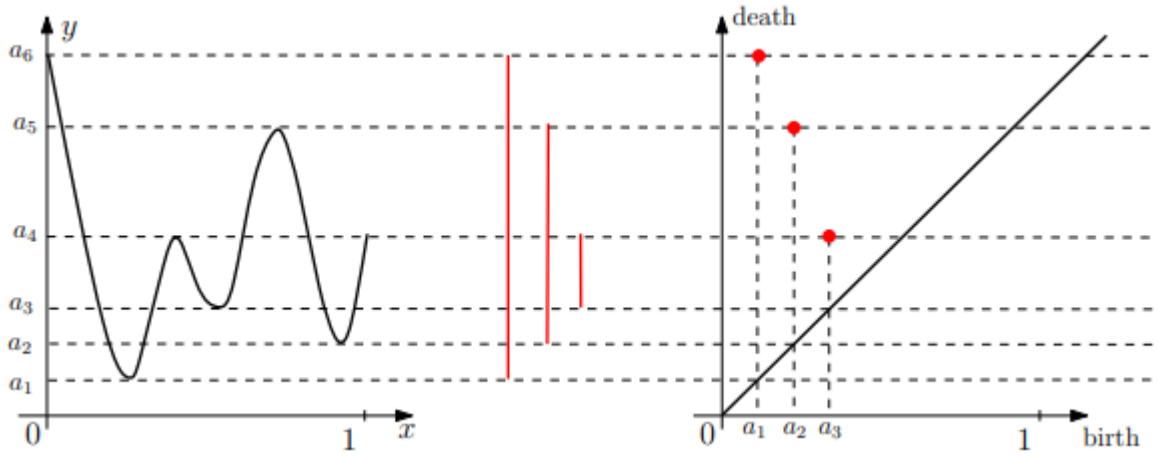
Δεδομένης κάποιας διήθησης, έστω $Filt = (F_r)_{r \in T}$ ενός πλεγματικού συμπλέγματος ή ενός τοπολογικού χώρου, η ομολογία των F_r μεταβάλλεται καθώς το r αυξάνεται, ως εξής: Μπορεί να εμφανιστούν νέες ενωμένες συνιστώσες, μπορεί να συγχωνευτούν ήδη υπάρχουσες συνιστώσες και μπορεί να εμφανιστούν ή να "γεμίσουν" κόμβοι ή κοιλότητες κ.ο.κ. Η Εμμένουσα Ομολογία (Persistent Homology) ανιχνεύσει αυτές τις αλλαγές, αναγνωρίζει τις ιδιότητες που προκύπτουν και σχετίζει κάποιο χρόνο ζωής με αυτές. Η πληροφορία στην οποία καταλήγουμε είναι κωδικοποιημένη ως ένα σύνολο διαστημάτων το οποίο ονομάζουμε barcode ή, ισοδύναμα, ως ένα πολυσύνολο σημείων στον $\mathbb{R}^2$, όπου οι συντεταγμένες κάθε σημείου είναι τα άκρα του διαστήματος αντίστοιχα.

Πριν προχωρήσουμε στην αυστηρή διατύπωση ορισμών θα παρατεθούν εισαγωγικά κάποια παραδείγματα:

Παράδειγμα 1

Έστω $f : [0, 1] \to \mathbb{R}$ μία συνάρτηση της εικόνας 11 που βρίσκεται παρακάτω. Έστω $(F_r = f^{-1}((-\infty, r]))_{r \in R}$ η διήθηση συνόλου υποεπιπέδων της f. Όλα τα σύνολα υποεπιπέδων της f είναι είτε κενά είτε ενώσεις διαστημάτων, επομένως η μόνη μη τετριμμένη τοπολογική πληροφορία που αυτά φέρουν είναι η 0-διάστατη τοπολογία τους, ή, ισοδύναμα, ο αριθμός που εκφράζει το πλήθος των ενωμένων συνιστωσών του. Για $r < \alpha_1$, το F_r είναι το κενό σύνολο. Όμως αν $r = \alpha_1$ εμφανίζεται η πρώτη ένωση συνιστωσών στην F_{α_1}. Σύμφωνα με την Εμμένουσα Ομολογία τα α_1 καταγράφεται ως η ώρα γέννησης μία ενωμένης συνιστώσας και ξεκινά να καταγράφει τις αλλαγές του, αναπαριστώντας ένα διάστημα που ξεκινά από το α_1. Τότε, το F_r παραμένει ενωμένο έως ότου το r φτάσει την τιμή α_2, όπου η μία δεύτερη συνιστώσα εμφανίζεται. Η Εμμένουσα Ομολογία αρχίζει και πάλι να καταγράφει τις αλλαγές δημιουργώντας ένα δεύτερο διάστημα που ξεκινά από το α_2. Με παρόμοιο τρόπο λειτουργεί προχωρώντας στο α_3 και η Εμμένουσα Ομολογία δημιουργεί άλλο ένα διάστημα που ξεκινά από το α_3. Ωστόσο, όταν το r φτάσει το α_4, οι δύο ενωμένες συνιστώσες που δημιουργήθηκαν στα α_1 και α_3 συγχωνεύονται σε μία μεγαλύτερη συνιστώσα. Σε αυτό το σημείο η Εμμένουσα Ομολογία θεωρεί πως αυτή είναι η πιο πρόσφατα εμφανιζόμενη συνιστώσα στη διήθηση που ¨πεθαίνει¨. Επομένως το διάστημα που ξεκινά από το α_3, τελειώνει

στο α_4 και αυτό αποτελεί το πρώτο εμμένον διάστημα που κωδικοποιεί τη διάρκεια ζωής της συνιστώσας που δημιουργήθηκε όταν εμφανίστηκε το α_3. Όταν το r φτάνει το α_5, όπως και στην προηγούμενη περίπτωση, η συνιστώσα που δημιουργείται στα α_2 'πεθαίνει' και το εμμένον διάστημα (α_2, α_5) δημιουργείται. Το διάστημα που δημιουργείται στα α_1 παραμένει έως ότου η διήθηση δίνει το εμμένον διάστημα (α_1, α_6), αν η διήθηση σταματά στο α_6, ή στο (α_1, ∞) αν το r πηγαίνει στο ∞. Στη δεύτερη περίπτωση σημειώνεται ότι η διήθηση παραμένει σταθερή για $r \geq \alpha_6$. Το σύνολο διαστημάτων που έχει παραχθεί κωδικοποιεί τη διάρκεια ζωής των διάφορων ομολογικών ιδιοτήτων που συναντάμε κατά τη διήθηση ονομάζονται *εμμένοντα barcode* της f. Κάθε διάστημα της μορφής (α, α') μπορεί να παρασταθεί από ένα σημείο με συντεταγμένες (α, α'). Το σύνολο σημείων στο οποίο καταλήγουμε καλείται *εμμένον διάγραμμα της* f. Σημειώνεται ότι μία συνάρτηση μπορεί να έχει πλήθος αντιγράφων του ίδιου διαστήματος στο εμμένον barcode και κατά συνέπεια το εμμένον διάγραμμα της f είναι ένα πολυσύνολο, στο οποίο κάθε σημείο έχει κάποια ακέραια πολλαπλότητα. Τέλος, για τεχνικούς λόγους που θα γίνουν πιο κατανοητοί στην επόμενη ενότητα, προσθέτουμε στην εμμένουσα όλα τα σημεία ης διαγωνίου, έστω $\Delta = \{(b, d) : b = d\}$.



Σχήμα 2.11: Το εμμένον Barcode και το εμμένον διάγραμμα μίας συνάρτησης $f : [0, 1] \to \mathbb{R}$ [0]

Παράδειγμα 2

Έστω $f : M \to \mathbb{R}$ Η συνάρτηση που απεικονίζεται στο Σχήμα 12, όταν

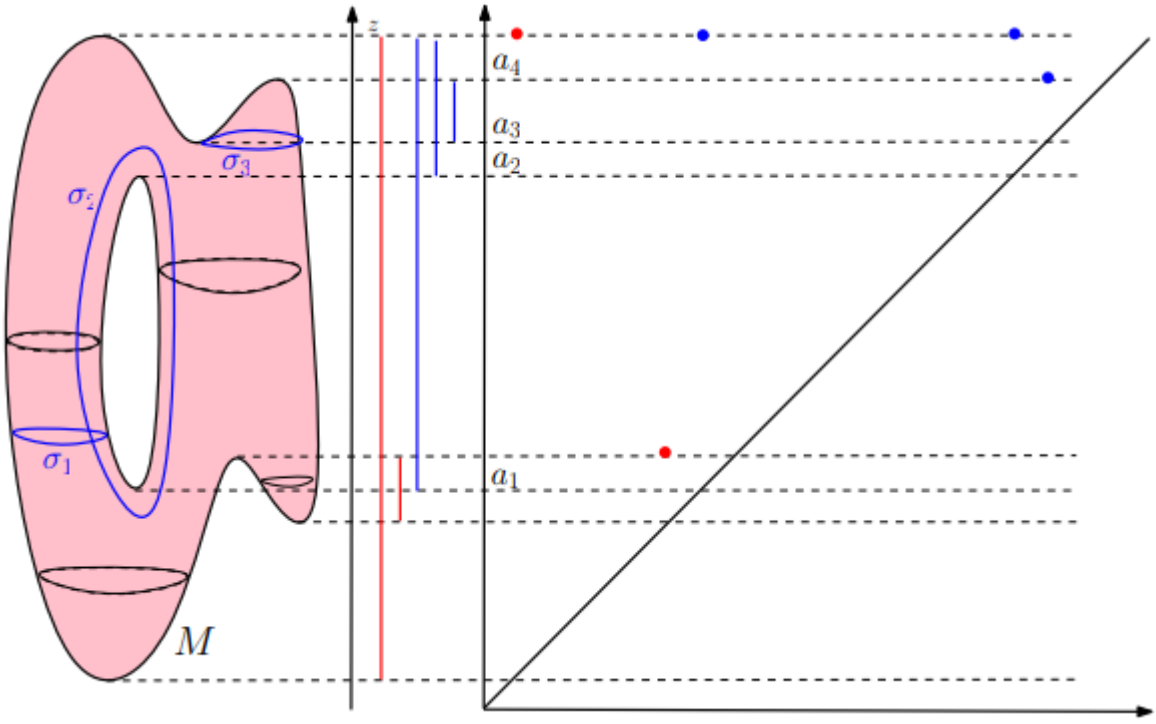
το M είναι μία 2-διάστατη επιφάνεια ομοιομορφική με έναν τόρο. Έστω $(F_r = f^{-1}((-\infty, r])_{r\in\mathbb{R}}$ η διήθηση του συνόλου υποεπιπέδων της f. Η 0-διάστατη εμμένουσα ομολογία, που υπολογίζεται όπως και στο Παράδειγμα 1, δίνοντας κόκκινες γραμμές στο barcode. Τα σύνολα υποεπιπέδων επίσης φέρουν 1-διάστατες ομολογικές ιδιότητες. Καθώς το r ξεπερνά το ύψος a_1 τα σύνολα υποεποεπιπέδων F_r που είναι ομοιομορφικά με δύο δίσκους γίνονται ομοιομορφικά με την κατατμημένη ένωση ενός δίσκου και ενός αντικειμένου σε σχήμα δακτυλίου, δημιουργώντας έναν πρώτο κύκλο ομολογικό με τον σ_1, όπως φαίνεται στο Σχήμα 12. Παρατηρούμε τότε ότι ένα διάστημα (με μπλε) αναπαριστά τη γέννηση αυτού του 1-κύκλου, το οποίο θα ξεκινά από το a_1. Με την ίδια λογική, όταν το r ξεπερνά το ύψος α_2 ένας νέος κύκλος, αυτή τη φορά ομοιομορφικός με το σ_2 δημιουργείται, και έτσι ξεκινά από εκείνο το σημείο ένα νέο εμμένον διάστημα. Οι δύο προαναφερθέντες κύκλοι δε θα ¨γεμίσουν' ποτέ και τα αντίστοιχα διαστήματα παραμένουν έως το τέλος της διήθησης. Όταν το r ξεπεράσει το a_3 δημιουργείται ένας νέος κύκλος, ο οποίος όμως ¨γεμίζει' αι έτσι ¨πεθαίνει' στο a_4, δημιουργώντας έτσι το εμμένον διάστημα (a_3, a_4). Επομένως πλέον η διήθηση του συνόλου υποεπιπέδων της f παράγει δύο barcodes ένα για τη 0-διάστατη ομολογία και ένα για την 1-διάστατη ομολογία, που φαίνονται στο Σχήμα 12 με κόκκινο και μπλε αντίστοιχα. Όπως και προηγουμένως, αυτά τα δύο barcodes μπορούν ισοδύναμα να παρασταθούν ως διαγράμματα στο πεδίο.

Παράδειγμα 3

Σε αυτό το παράδειγμα θα θεωρήσουμε τη διήθηση που δίνεται από μία ένωση αυξανόμενων μπαλών με κέντρο στο πεπερασμένο σύνολο σημείων $\mathcal{C}$ του Σχήματος 13. Παρατηρούμε ότι το σύνολο υποσυνόλων διήθησης της συνάρτησης απόστασης στο $\mathcal{C}$, και λόγω του Θεωρήματος Νεύρου, αυτή η διήθηση είναι ομοτοπικά ισοδύναμη με τη διήθηση Cech που κατασκευάζεται επί του $\mathcal{C}$. Στο Σχήμα 13 απεικονίζονται διάφορα σύνολα επιπέδων της διήθησης, τα οποία παρουσιάζονται παρακάτω:

i. Για την ακτίνα $r = 0$, η ένωση μπαλών περιορίζεται στο αρχικό σύνολο σημείων, καθένα από τα οποία αντιστοιχίζεται σε κάποιο 0-διάστατο χαρακτηριστικό, δηλαδή μία ενωμένη συνιστώσα. Έτσι δημιουργείται ένα διάστημα για τη ¨γέννηση' καθενός από αυτά τα χαρακτηριστικά στο $r = 0$.

ii. Παρατηρούμε ότι κάποιες μπάλες φαίνεται να επικαλύπτονται, το οποίο οδηγεί στο ¨θάνατο' κάποιων συνδεδεμένων συνιστωσών που



Σχήμα 2.12: Το εμμένονbarcode και το εμμένον διάγραμμα της συνάρτησης ύψους, δηλαδή η προβολή στον άξονα z που ορίζεται επί μίας επιφάνειας του $\mathbb{R}^3$. [0]

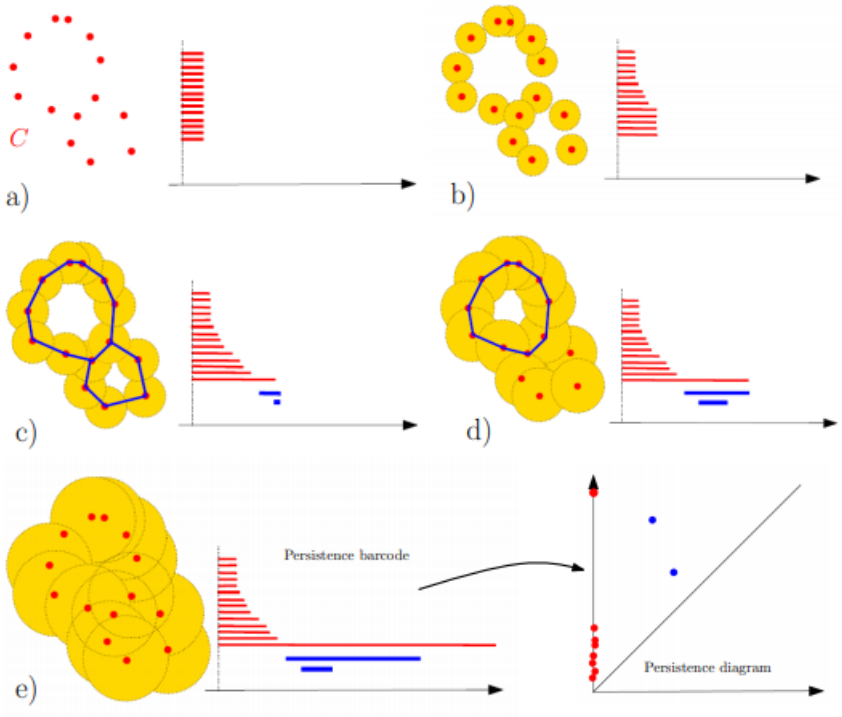
συνενώνονται. Το εμμένον διάγραμμα αποδελτιώνει αυτούς τους θανάτους, θέτοντας ένα σημείο για τη λήξη των αντίστοιχων διαστημάτων καθώς αυτά εξαφανίζονται.

iii. Νέες συνιστώσες συνενώνονται δίνοντας έτσι μία συνδεδεμένη συνιστώσα και με αυτόν τον τρόπο όλα τα διαστήματα που σχετίζονται με τα 0-διάστατα χαρακτηριστικά έχουν τελειώσει, με εξαίρεση εκείνο που αντιστοιχίζεται στην εναπομένουσα συνιστώσα. Τότε εμφανίζονται δύο καινούρια 1-διάστατα χαρακτηριστικά, που με τη σειρά τους οδηγούν σε δύο νέα διαστήματα, που στο Σχήμα 13 απεικονίζονται με μπλε, εκκινώντας την κλίμακα ¨γέννησής' τους.

iv. Ένας από τους δύο 1-διάστατους κύκλους έχει γεμίσει, επομένως ¨πεθαίνει' στη διήθηση και το οδηγώντας στο τελικό σημείο του αντίστοιχου (μπλε) διαστήματος.

v. Όλα τα 1-διάστατα χαρακτηριστικά έχουν ¨πεθάνει' πλέον και απομένει μόνο η μακριά κόκκινη γραμμή που απεικονίζει το διάστημα που

δεν ¨πεθαίνει ποτέ¨. Όπως και στα προηγούμενα παραδείγματα, το τελικό barcode μπορεί επίσης να παρασταθεί ισοδύναμα ως εμμένον διάγραμμα, στο οποίο κάθε διάστημα (a, b) αναπαρίσταται από ένα σημείο με συντεταγμένες $(a, b) \in \mathbb{R}^2$. Διαισθητικά καταλαβαίνουμε ότι όσο μεγαλύτερο είναι ένα διάστημα στο barcode, ή ισοδύναμα όσο πιο μακριά βρίσκεται από τη διαγώνιο το αντίστοιχο σημείο του διαγράμματος, τόσο πιο εμμένουσα και με αυτόν τον τρόπο σχετική είναι η αντίστοιχη ομολογική ιδιότητα στη διήθηση. Σημειώνεται επιπλέον ότι για δοθείσα ακτίνα r, ο k-οστός αριθμός Betti της αντίστοιχης ένωσης μπαλών είναι ίσος με το πλήθος των εμμενόντων διαστημάτων που αντιστοιχούν στις k-διάστατες ομολογικές ιδιότητες και περιέχουν το r. Επομένως, το εμμένον διάγραμμα μπορούμε να το δούμε ως μία πολλαπλής κλίμακας τοπολογική υπογραφή που κωδικοποιεί την ομολογία της ένωσης μπαλών για κάθε ακτίνα, όπως επίσης την ανταπόκρισή της στις μεταβολές των τιμών της ακτίνας r.



Σχήμα 2.13: Η διήθηση συνόλου υποεπιπέδων της συνάρτησης απόστασης σε ένα νέφος δεδομένων και η ¨κατασκευή¨ του αντίστοιχου εμμένοντος barcode καθώς η ακτίνα των μπαλών αυξάνεται. [0]

Λίγα λόγια για τα εμμένοντα διαγράμματα και τα Barcodes

Τα εμμένοντα barcodes βοηθούν στην οπτικοποίηση των n-διάστατων ομάδων ομολογίας μέσω των γεννητόρων τους. Κάθε μπάρα του barcode αναπαριστά ένα γεννήτορα της ομάδας της εμμένουσας ομολογίας. Πιο απλά, η κάθε μπάρα δείχνει τη ¨διαρκεια ζωής' των εκαστοτε χαρακτηριστικών καθώς η ακτίνα r μεγαλώνει.

Το εμμένον διάγραμμα, είναι ένα δισδιάστατο γράφημα που δείχνει τη ¨γέννηση' και το ¨θάνατο' κάθε χαρακτηριστικου ως εξής:

- Στον άξονα των x σημειώνεται η ¨γέννηση' του χαρακτηριστικού και
- Στον άξονα των y σημειώνεται ο ¨θάνατός' του

Το πόσο ¨εμμένει' ένα χαρακτηριστικό γίνεται αντιληπτό από το μήκος της αντίστοιχης μπάρας του barcode και, αντίστοιχα, της απόστασης μεταξύ του σημείου που αναπαριστά το συγκεκριμένο χαρακτηριστικό στο εμμένον διάγραμμα και της διαγωνίου. Με άλλα λόγια, ένα χαρακτηριστικό που απεικονίζεται με μία μπάρα μεγάλου μήκους στο barcode και (ισοδύναμα) το αντίστοιχο σημείο στο οποίο αντιστοιχεί στο εμμένον διάγραμμα βρίσκεται μακριά από τη διαγώνιο, θεωρούμε ότι περιέχει σημαντική τοπολογική πληροφορία, σχετικά με τα δεδομένα και το υποβόσκων σχήμα τους.

2.4.3 Εμμένοντα Τοπία

Προηγούμενα μέσα στην εργασία έγινε μία εισαγωγή των εμμενόντων διαγραμμάτων και barcodes. Τα εργαλεία αυτά δεν είναι όμως εύκολα συμβατά με άλλες τεχνικές, όπως η μηχανική μάθηση. Μία εναλλακτική προσέγγιση είναι τα εμμένοντα τοπία. Το βασικό τεχνικό πλεονέκτημα που αυτά παρουσιάζουν είναι ότι αποτελούν συναρτήσεις και έτσι μπορούμε να αξιοποιήσουμε τον χώρο συναρτήσεων στον οποίο ζουν. Τα εμμένοντα τοπία απεικονίζουν τα εμμένοντα διαγράμματα σε ένα χώρο συναρτήσεων. Ο χώρος αυτός είναι συχνά χώρος Hilbert. [0]

Για να ορίσουμε την έννοια ενός χώρου Hilbert πρέπει αρχικά να ορίσουμε την έννοια του εσωτερικού γινομενου σε ενα διανυσματικό χώρο καθώς και κάποιες σχετικές ιδιότητες και προτάσεις διανυσματικών χώρων με εσωτερικό γινόμενο που θα μας χρειαστούν στη συνέχεια, χωρίς τις αποδείξεις τους.

2.4.3.1 Ορισμός

Θεωρούμε έναν διανυσματικό χώρο, έστω V επί ενός σώματος $\mathbb{C}$. Μία απεικόνιση $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{C}$, που για κάθε $x, y, z \in V$ και $\lambda \in \mathbb{C}$ ικανοποιεί τις παρακάτω ιδιότητες:

(i) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$

(ii) $\langle \lambda x, y \rangle = \lambda \langle x, y \rangle$

(iii) $\langle x, y \rangle = \overline{\langle y, x \rangle}$

(iv) $\langle x, x \rangle \geq 0$ και αν $\langle x, x \rangle = 0$, τότε $x = 0_V$

θα λέγεται *εσωτερικό γινόμενο* στον V.

Ο V με την απεικόνιση $\langle \cdot, \cdot \rangle$ ονομάζεται διανυσματικός χώρος με εσωτερικό γινόμενο.

2.4.3.2 Πρόταση

Έστω ένας διανυσματικός χώρος με εσωτερικό γινόμενο $(V, \langle \cdot, \cdot \rangle)$. Η απεικόνιση $\| \cdot \| : V \times V \to \mathbb{R}$, όπου $\|x\| = \langle x, x \rangle^{1/2}$ είναι μία νόρμα.

Παρατήρηση 2.4.3.2

Από την πιο πάνω πρόταση, συμπεραίνουμε ότι κάθε απεικόνιση εσωτερικού γινομένου επάγει και μια νόρμα, δηλαδή καθε χώρος με εσωτερικό γινόμενο είναι και χώρος με νόρμα.

2.4.3.3 Ορισμός

Χώρος Banach ονομάζεται ένας πλήρης διανυσματικός χώρος με νόρμα.

2.4.3.4 Ορισμός

Ένας χώρος με εσωτερικό γινόμενο, έστω $H, \langle \cdot, \cdot \rangle$, ονομάζεται *χώρος Hilbert*, αν είναι χώρος Banach ως προς τη νόρμα που ορίζει το εσωτερικό γινόμενο $\langle \cdot, \cdot \rangle$.

Στην περίπτωση, λοιπόν, που ο χώρος συναρτήσεων που δημιουργείται από το εμμένον τοπίο είναι χώρος Hilbert είναι τότε μπορούν να εφαρμοστούν εργαλεία της στατιστικής και της μηχανικής μάθησης.

Η απεικόνιση ενός εμμένοντος διαγράμματος σε ένα εμμένον τοπίο σε κάποιες περιπτώσεις να αντιστραφεί. Δηλαδή, από ένα εμμένον τοπίο μπορούμε, κάποιες φορές, να ανακατασκευάσουμε το αντίστοιχο εμμένον διάγραμμα.

Το εμμένον τοπίο έχει παρουσιαστεί αρχικά ως εναλλακτική αναπαράσταση των εμμενόντων διαγραμμάτων. [0] Αυτή η προσέγγιση στοχεύει στην αναπαράσταση τοπολογικών πληροφοριών κωδικοποιημένων με εμμένοντα διαγράμματα ως στοιχεία ενός χώρου Hilbert, για τον οποίο μπορούν άμεσα να εφαρμοστούν μέθοδοι στατιστικής εκμάθησης.

Το εμμένον τοπίο είναι μία συλλογή συνεχών τμηματικά συναρτήσεων $\lambda : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$, η οποία συνοψίζει ένα εμμένον διάγραμμα dgm, όπως φαίνεται στο Σχήμα 14. Το τοπίο ορίζεται θεωρώντας το σύνολο των συναρτήσεων που δημιουργείται μετασχηματίζοντας κάθε σημείο στην παρακάτω μορφή:

$$p = (x, y) = \left(\frac{a_{birth} + a_{death}}{2}, \frac{a_{death} - a_{birth}}{2} \right)$$

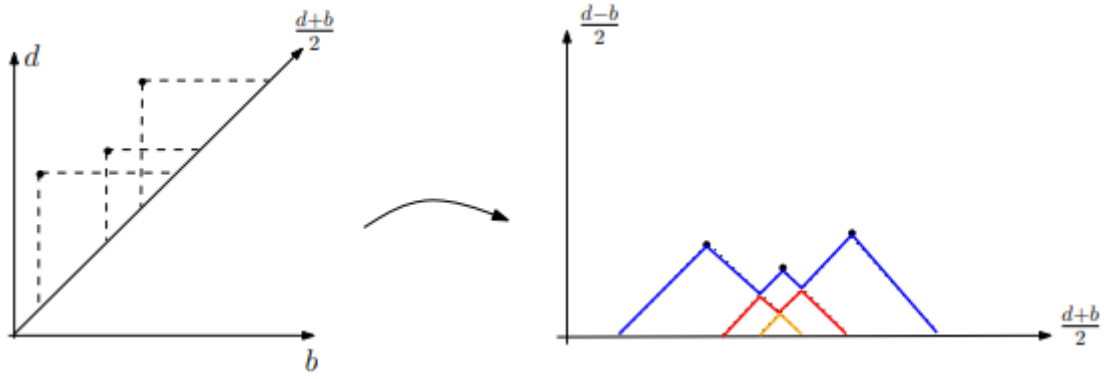
Το παραπάνω αναπαριστά ένα ζεύγος $(a_{birth}, a_{death}) \in dgm$ ως εξής:

$$\Lambda_p(t) = \left\{ \begin{array}{ll} t - x + y, & \text{όπου } t \in [x-y, x] \\ x + y - t, & \text{όπου } t \in (x, x+y] \\ 0, & \text{διαφορετικά} \end{array} \right\} = \left\{ \begin{array}{ll} t - a_{birth}, & \text{όπου } t \in \left[a_{birth}, \frac{a_{birth}+a_{death}}{2}\right] \\ a_{death} - t, & \text{όπου } t \in \left(\frac{a_{birth}+a_{death}}{2}, a_{death}\right] \\ 0, & \text{διαφορετικά} \end{array} \right\}$$

Το εμμένον τοπίο του dgm είναι μία μορφή οργανωμένης σύνοψης των επί μέρους γραμμικών καμπυλών που παίρνουμε από την επικάλυψη γραφημάτων των συναρτήσεων $\{\Lambda_p\}_p$. Για να είμαστε πιο τυπικοί, το εμμένον τοπίο του dgm της συλλογής των συναρτήσεων:

$$\lambda_{dgm}(k,t) = kmax_p \Lambda_p(t), t \in [0,T], k \in \mathbb{N}$$

όπου $kmax$ είναι η k-οστή μεγαλύτερη τιμή στο σύνολο. Ωστόσο, το $1max$ είναι η συνήθης μέγιστη συνάρτηση. Δοθέντος ενός $k \in \mathbb{N}$, η συνάρτηση $\lambda_{dgm}(k,.) : \mathbb{R} \to \mathbb{R}$ ονομάζεται k-οστό τοπίο του dgm. Για την απεικόνιση που συνδέεται με κάθε εμμένον διάγραμμα, μπορούμε εύκολα να διαπιστώσουμε ότι το αντίστοιχο τοπίο του είναι ενθετικό. Συμπεραίνουμε έτσι ότι όταν έναν εμμένον διάστημα αναπαρίσταται, το εμμένον τοπίο του δε χάνει πληροφορία.



Σχήμα 2.14: Στα δεξιά βλέπουμε το εμμένον τοπίο που συνδέεται με το εμμένον διάγραμμα στα αριστερά. Το πρώτο τοπίο αναπαρίσταται από τη μπλε γραμμή, το δεύτερο αναπαρίσταται από την κόκκινη γραμμή και το τελευταίο από την πορτοκαλί. Όλα τα υπόλοιπα τοπία είναι μηδενικά. [0]

Το πλεονέκτημα της αναπαράστασης του εμμένοντος τοπίου είναι μία 2-πολλαπλότητα. Αρχικά, τα εμμένοντα διαγράμματα απεικονίζονται ως

στοιχεία ενός χώρου συναρτήσεων. Αυτό μας δίνει τη δυνατότητα να χρησιμοποιήσουμε μεγάλη ποικιλία εργαλείων της στατιστικής και της ανάλυσης δεδομένων για την περαιτέρω επεξεργασία των τοπολογικών ιδιοτήτων. Θεμελιώδες από θεωρητικής άποψης είναι το γεγονός ότι τα εμμένοντα τοπία έχουν τις ίδιες ιδιότητες σταθερότητας όπως τα εμμένοντα διαγράμματα. Με το ίδιο σκεπτικό, έχουν προταθεί και άλλες εναλλακτικές για τα εμμένοντα διαγράμματα, όπως για παράδειγμα οι εμμένουσες εικόνες. [0]

2.4.4 Μετρικές στους χώρους των εμμενόντων διαγραμάτων

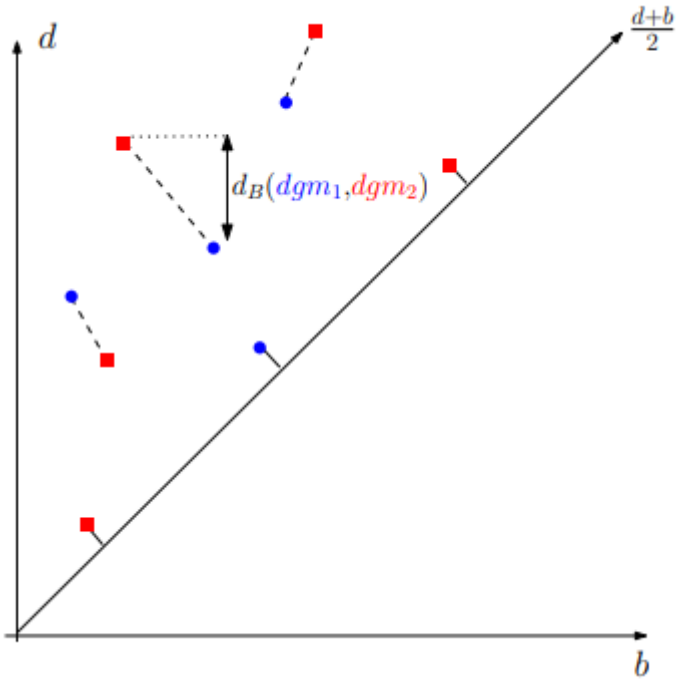
Για να έχουμε τη δυνατότητα να εξερευνήσουμε τις τοπολογική πληροφορία και τις τοπολογικές ιδιότητες που συνάγονται από την εμμένουσα ομολογία, είναι απαραίτητο να μπορούμε να συγκρίνουμε εμμένοντα διαγράμματα. Με άλλα λόγια, να εφοδιάσουμε το χώρο των εμμενόντων διαγραμμάτων με μία μετρική δομή. Για αυτό το σκοπό μπορούμε να αξιοποιήσουμε μία σειρά μετρικών, ωστόσο, η συνήθης (βιβλιογραφικά θεμελιώδης) μετρική που χρησιμοποιούμε είναι η απόσταση bottleneck.

Θυμίζουμε ότι τα εμμένοντα διαγράμματα είναι η ένωση ενός διακριτού πολυσυνόλου που είναι ένα ημιεπίπεδο επάνω από τη διαγώνιο Δ και, για τεχνικούς λόγους που θα δούμε στη συνέχεια, όπου τα σημεία του Δ θεωρούμε ότι είναι σημεία άπειρης πολλαπλότητας. Στο Σχήμα 15, που φαίνεται παρακάτω, βλέπουμε μία αντιστοίχιση (matching) μεταξύ δύο διαγραμμάτων, έστω dgm_1 και dgm_2, είναι ένα υποσύνολο $m \subseteq dgm_1 \times dgm_2$, τέτοιο ώστε κάθε σημείο στα $dgm_1\Delta$ και $dgm_2\Delta$ εμφανίζεται ακριβώς μία φορά m. Δηλαδή, για κάθε $p \in dgm_1\Delta$ και κάθε $q \in dgm_2\Delta$, καθένα από τα $(\{p\} \times dgm_2) \cap m$ και $(dgm_1 \times \{q\}) \cap m$ περιέχει ένα ζεύγος. Η απόσταση Bottleneck μεταξύ των dgm_1 και dgm_2 ορίζεται τότε ως εξής:

$$d_b(dgm_1, dgm_2) = inf_{matching m} max_{(p,q)\in m} ||p - q||_\infty$$

Ο υπολογισμός, πρακτικά, της απόστασης Bottleneck συμπυκνώνεται ην εύρεση μίας τέλειας αντιστοιχίας σε ένα διμερές γράφημα του οποίου οι κλασικοί αλγόριθμοι μπορούν να χρησιμοποιηθούν.

Η μετρική Bottleneck είναι όμοια με μία L_∞ μετρική. Η μετρική αυτή λειτουργεί με φυσικό τρόπο στην έκφραση ιδιοτήτων σταθερότητας των εμμενόντων διαγραμμάτων που θα παρουσιαστεί αναλυτικότερα στην επόμενη ενότητα. Έχει όμως τα ίδια 'ελαττώματα' με τις L_∞ νόρμες, δηλαδή καθορίζεται απόλυτα από μία μεγαλύτερη απόσταση μεταξύ των ζευγών και δε λαμβάνει υπόψη την εγγύτητα των εναπομεινάντων ζευγών σημείων. Για να υπερκεράσουμε αυτό το ζήτημα, θεωρούμε κάποιες φορές μία μεταβλητή, τη λεγόμενη απόσταση Wasserstein μεταξύ των διαγραμμάτων. Η απόσταση αυτή, δεδομένου ενός $p \geq 1$, ορίζεται ως εξής:



Σχήμα 2.15: Μία τέλεια αντίστοιχη και η απόσταση Bottleneck μεταξύ των ενός κόκκινου και ενός μπλε διαγράμματος. Σημειώνεται επίσης ότι κάποια σημεία και από τα δύο διαγράμματα αντιστοιχίζονται σε σημεία της διαγωνίου. [0]

$$W_p(dgm_1, dgm_2)^p = inf_{matching m} \sum_{(p,q \in m)} ||p - q||_\infty^p$$

Βιβλιογραφικά συναντάμε χρήσιμα αποτελέσματα για τη σταθερότητα της εμμονής (persistence) της W_p μετρικής, ([0]) βασίζονται όμως σε υποθέσεις που επηρεάζουν τα αποτελεσμάτα για τη σταθερότητα στη μετρική Bottleneck.

2.4.5 Στατιστικές πλευρές της εμμένουσας ομολογίας

Η εμμένουσα ομολογία δε λαμβάνει υπόψη την τυχαία φύση των δεδομένων και την εγγενή μεταβλητότητα της τοπολογικής ποσότητας που αυτά συνεπάγονται. Θα παρουσιάσουμε σε αυτό το σημείο μία στατιστική προσέγγιση της εμμένουσας ομολογίας, το οποίο σημαίνει ότι θα θεωρήσουμε ότι τα δεδομένα προέρχονται από μία άγνωστη κατανομή. Θα ξεκινήσουμε με διάφορα αποτελέσματα σχετικά με τη συνοχή της συμπερασματολογίας της εμμένουσας ομολογίας.

Εκτίμηση της εμμένουσας ομολογίας ενός μετρικού χώρου

Υποθέτουμε ότι έχουμε n παρατηρήσεις $(X_1, X_2, \ldots, X_n)$ σε ένα μετρικό χώρο (M, ρ) ανεξάρτητα και ισόνομα από κάποιο άγνωστο μέτρο πιθανότητας μ, με στήριγμα $\mathbb{X}_\mu$. Η απόσταση Gromov-Hausdorff μας επιτρέπει να συγκρίνουμε το στήριγμα $\mathbb{X}_\mu$ με συμπαγείς μετρικούς χώρους που δεν είναι απαραίτητα ενσωματωμένοι στον M. Στη συνέχεια, ένας εκτιμητής $\hat{\mathbb{X}}$ του $\mathbb{X}_\mu$ είναι μία συνάρτηση των X_1, X_2, $\ldots$, X_n που παίρνει τιμές στο σύνολο των συμπαγών μετρικών χώρων και είναι μετρήσιμο για την άλγεβρα Borel ου επάγεται από την d_{GH}.

Έστω $Filt(\mathbb{X}_\mu)$ και $Filt(\hat{\mathbb{X}})$ δύο διηθήσεις που ορίζονται επί του $\mathbb{X}_\mu$ και του $\hat{\mathbb{X}}$ αντίστοιχα. Θα ξεκινήσουμε από το Θεώτημα 2.4.5.5. Μια στρατηγική, που προκύπτει φυσικά για την εκτίμηση της εμμένουσας ομολογίας στην $Filt(\mathbb{X}_\mu)$, στηρίζεται στην εκτίμηση του στηρίγματος $\mathbb{X}_\mu$. Σε κάποιες περιπτώσεις ο χώρος M μπορεί να είναι άγνωστος και οι παρατηρήσεις X_1, X_2, $\ldots$, X_n είναι τα μόνα γνωστά στοιχεία μέσω των ανά δύο αποστάσεών τους $\rho(X_i, X_j)$, $i, j = 1, 2, \ldots, n$. Η χρήση της απόστασης Gromov-Hausdorff μας επιτρέπει να θεωρήσουμε αυτό το σύνολο παρατηρήσεων ως ένα αφηρημένο μετρικό χώρο με n το πλήθος στοιχεία, ανεξάρτητα από τον τρόπο που αυτό ενσωματώνεται στο M. Αυτό το γενικό πλαίσιο περιέχει μία πιο συνηθισμένη προσέγγιση που συμβάλει στην εκτίμηση του στηρίγματος, με σεβασμό στην απόσταση Hausdorff περιορίζοντας της τιμές του $\hat{\mathbb{X}}$ στα συμπαγή σύνολα που περιέχονται στο χώρο M.

Το πεπερασμένο σύνολο $\mathbb{X}_n := \{X_1, X_2, \ldots, X_n\}$ είναι ένας φυσικός εκτιμητής του στηρίγματος $\mathbb{X}_\mu$. Θα δούμε παρακάτω ότι σε διάφορα πλαίσια συζήτησης, το $\mathbb{X}_n$ φαίνεται να έχει τη βέλτιστη αναλογία σύγκλισης στο $\mathbb{X}_\mu$ σεβόμενο τη απόσταση Hausdorff. Για κάποιες σταθερές $a, b \geq 0$, θεωρούμε ότι το μ ικανοποιεί την (α, β)-στάνταρ υπόθεση ότι αν για κάθε $x \in \mathbb{X}_\mu$ και κάθε $r \geq 0$:

$$\mu(B(x, r)) \geq min(ar^b, 1)$$

Αυτή η υπόθεση έχει χρησιμοποιηθεί ευρέως στη βιβλιογραφία των εκτιμήσεων συνόλων με την απόσταση Hausdorff. ([0], [0])

2.4.5.1 Θεώρημα [0]

Υποθέτουμε ότι το μέτρο πιθανότητας μ στο χώρο M ικανοποιεί την (a, b)-στάνταρ υπόθεση. Τότε για κάθε $\varepsilon \geq 0$:

$$\mathbb{P}(d_b(dgm(Filt(\mathbb{X}_\mu)), dgm(Filt(\mathbb{X}_n)))) \geq \varepsilon) \leq min\left(\frac{2^b}{a\varepsilon^b} exp(-na\varepsilon^b), 1\right)$$

Επιπλέον έχουμε σχεδόν βέβαια ότι:

$$lim\sup_{n\to\infty}\left(\frac{n}{\log n}\right)^{1/b} d_b(dgm(Filt(\mathbb{X}_\mu)), dgm(Filt(\mathbb{X}_n))) \leq C_1$$

Και το παρακάτω:

$$\mathbb{P}\left(d_b(dgm(Filt(\mathbb{X}_\mu)), dgm(Filt(\mathbb{X}_n))) \leq C_2\left(\frac{\log n}{n}\right)^{1/b}\right)$$

συγκλίνει στο 1 όταν το $n \to \infty$, όπου C_1 και C_2 εξαρτώνται αποκλειστικά από τα a και b.

Έστω $\mathcal{P} = \mathcal{P}(a, b, M)$ το σύνολο όλων των μέτρων πιθανότητας σε ένα μετρικό χώρο (M, ρ) που ικανοποιεί την (a, b)-στάνταρ υπόθεση στο χώρο M:

$$\mathcal{P} := \{\mu \text{ στον } \mathrm{M} | \mathbb{X}_\mu \text{ συμπαγές και } \forall x \in \mathbb{X}_\mu, \forall r \geq 0, \mu(B(x, r)) \geq min\left(1, ar^b\right)\}$$

Το ακόλουθο θεώρημα δίνει τα άνω και κάτω φράγματα για την αναλογία σύγκλισης των εμμενόντων διαγραμμάτων. Το άνω φράγμα αποτελεί συνέπεια του Θεωρήματος 2.4.5.7, ενώ το κάτω φράγμα προκύπτει από τη χρήση του Λήμματος του Le Cam.

2.4.5.2 Θεώρημα [0]

Έστω a, b δύο μη αρνητικές σταθερές. Τότε:

$$sup_{\mu \in \mathcal{P}} \mathbb{E}[d_b(dgm(Filt(\mathbb{X}_\mu)), dgm(Filt(\mathbb{X}_n)))] \leq C \left(\frac{\log n}{n}\right)^{1/b}$$

όπου η σταθερά C εξαρτάται αποκλειστικά από τα a και b, όχι στον M. Επιπλέον, υποθέτουμε ότι υπάρχει ένα μη απομονωμένο σημείο $x \in M$ και θεωρώντας οποιαδήποτε ακολουθία $(x_n) \in M$ τέτοια ώστε $\rho(x, x_n \leq (an)^{-1/b}$. Τότε οποιοσδήποτε εκτιμητής $\hat{dgm}$ του $dgm(Filt(\mathbb{X}_\mu))$:

$$\liminf_{n \to \infty} \rho(x, x_n)^{-1} \sup_{\mu \in \mathcal{P}} \mathbb{E}\left[d_b(dgm(Filt(\mathbb{X}_\mu)), \hat{dgm}_n)\right] \geq C'$$

όπου το C' είναι μία σταθερά.

Συνεπώς, ο εκτιμητής $dgm(Filt(\mathbb{X}_n))$ είναι το βέλτιστο minimax στο χώρο $\mathcal{P}(a, b, M)$ έως κάποιο λογαριθμικό όρο, εφόσον μπορέσουμε να βρούμε ένα μη απομονωμένο σημείο στο M και μία ακολουθία $(x_n) \in M$ τέτοια ώστε $\rho(x_n, x) \sim (an)^{-1/b}$. Εμφανώς αυτή είναι η περίπτωση του Ευκλείδειου χώρου $\mathbb{R}^d$.

Θεωρούμε το μοντέλο περιελιγμού, όπου οι παρατηρήσεις ικανοποιούν τη σχέση $Y_i = X_i + \varepsilon_i$. με X_1, X_2, ..., X_n. Οι παρατηρήσεις αυτές λαμβάνονται σύμφωνα με το μέτρο μ, όπως και στη προηγούμενη ενότητα,

και με $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ συμβολίζουμε τυχαίες μεταβλητές ακολουθούν i.i.d. την τυπική κανονική. Μπορούμε να συμπεράνουμε από τα αποτελέσματα των Genovese et al. (2012) ότι οι minimax αναλογίες σύγκλισης για την εκτίμηση των εμμενόντων διαγραμμάτων, σε αυτό το πλαίσιο, είναι άνω φραγμένα από κάποια αναλογία τάξης ίσης με αυτή του $(\log n)^{-1/2}$. Από την άλλη μεριά όμως, η εύρεση ενός αυστηρού κάτω φράγματος είναι σαφώς δυσκολότερη από την εκτίμηση του στηρίγματος. [0]

Εκτίμηση της εμμένουσας ομολογίας των συναρτήσεων

Το Θεώρημα 2.4.5.5 μας δίνει τη δυνατότητα να εκτιμήσουμε την εμμένουσα ομολογία των συναρτήσεων που ορίζονται επί του $\mathbb{R}^d$, σε μία υποπολλαπλότητα του $\mathbb{R}^d$ ή γενικότερα σε ένα μετρικό χώρο. Για να ακολουθήσουμε μία πολύ σημαντική κατεύθυνση στην έρευνα, σχετικά με το συγκεκριμένο θέμα, θα χρειαστούμε διάφορες εύρωστες τεχνικές της Τοπολογικής Ανάλυσης Δεδομένων. Μία επιλογή είναι να μελετήσουμε την εμμένουσα ομολογία των άνω συνόλων επιπέδων των εκτιμητών πυκνότητας. [0] Εναλλακτικά, μια προσέγγιση που είναι πιο στενά συνδεδεμένη με την έννοια της συνάρτησης απόστασης, και ταυτόχρονα εύρωστη στο θόρυβο, συντελεί στη μελέτη της εμμένουσας ομολογίας των συνόλων υποεπιπέδων της απόστασης σε μέτρο, όπως ορίζεται σε προηγούμενη ενότητα. [0] Η εμμένουσα ομολογία των συναρτήσεων παλινδρόμησης έχει επίσης μελετηθεί από τους Bubenik et al. (2010). [0] Από την άλλη μερία η εναλλακτική προσέγγιση των Bobrwski et al. (2014) που βασίζεται στην ενθετική απεικόνιση μεταξύ των εμφωλευμένων ζευγών των εκτιμώμενων εμμενόντων ομολογιών των συναρτήσεων πυκνότητας και των συναρτησεων παλινδρόμησης. [0]

Στατιστικά αποτελέσματα για άλλες υπογραφές

Οι περιοχές σύγκλισης και της περιοχές εμπιστοσύνης μπορούν να προταθούν ως εμμένοντα τοπία δίνοντας παρόμοια αποτελέσματα για της σταθερότητα. Όμως, μία πολύπλοκη περιγραφή του minimax για αυτό το πρόβλημα θα προϋπέθετε την απόδειξη των αντίστοιχων κάτω φραγμάτων. Η συναρτησιακή σύγκλιση για τα εμμένοντα τοπία και σχήματα έχει μελετηθεί από τους Chazal et al. (2015b). [0]

Περιοχές εμπιστοσύνης για την εμμένουσα ομολογία

Σε κάποιες περιπτώσεις η χρήση των εμμενόντων διαστημάτων, πιο συγκεκριμένα όταν το νέφος δεδομένων δεν προέρχεται από κάποιο γεωμετρικό σχήμα, γίνεται εξαιρετικά περίπλοκη. Το πρόβλημα έγκειται στο γεγονός ότι πολλές τοπολογικές ιδιότητες παρουσιάζονται κοντά στη διαγώνιο. Αυτό σημαίνει ότι ανταποκρίνονται σε τοπολογικές ιδιότητες που, αφού εμφανιστούν στη διήθηση, "πεθαίνουν" γρήγορα και ως εκ τούτου θεωρούνται θόρυβος και όχι πληροφορία. Οι περιοχές εμπιστοσύνης των εμμενόντων διαγραμμάτων δίνουν σαφείς απαντήσεις σχετικά με τη διάκριση μεταξύ των σημάτων και του θορύβου σε αυτές τις αναπαραστάσεις.

Τα αποτελέσματα για τη σταθερότητα, όπως αυτά αποτυπώνονται σ' αυτήν την εργασία, ωθούν στη χρήση της απόστασης Bottleneck για τον ορισμό των περιοχών εμπιστοσύνης. Φυσικά μπορούν, στο πνεύμα των αποστάσεων Wasserstein, να προταθούν και εναλλακτικές αποστάσεις. Για την εκτίμηση ενός εμμένοντος διαγράμματος dgm με κάποιον εκτιμητή dgm, ψάχνουμε τυπικά για κάποια τιμή η_α τέτοια ώστε:

$$P\left(d_b(dgm, dgm) \geq \eta_\alpha\right) \leq \alpha, \forall \alpha \in (0,1)$$

Έστω B_α μία κλειστή μπάλα ακτίνας α, για την απόσταση Bottleneck, και κέντρο στο dgm στο χώρο των εμμενόντων διαγραμμάτων. Μπορούμε να οπτικοποιήσουμε τις υπογραφές των σημείων που ανήκουν σε αυτή τη μπάλα με διάφορους τρόπους. [0] Μία πρώτη επιλογή είναι να αποτελεί κέντρο ενός "κουτιού" πλευράς 2α, σε κάθε σημείο του εμμένοντος διαγράμματος dgm. Μία διαφορετική λύση είναι να οπτικοποιήσουμε το σύνολο εμπιστοσύνης, με την πρόσθεση μίας ζώνης σε απόσταση $\eta_\alpha/2$ κατακόρυφα από τη διαγώνιο. Σημειώνεται ότι η απόσταση Bottleneck έχει οριστεί ως η νόρμα l_∞. Τα σημεία που βρίσκονται έξω από αυτή τη ζώνη θεωρούνται ότι είναι σημαντικές τοπολογικές ιδιότητες. [0]

Έχουν προταθεί αρκετές μέθοδοι για την εκτίμηση η_α στον ορισμό της περιοχής εμπιστοσύνης για την εμμένουσα ομολογία του στηρίγματος του μέτρου και για τα σύνολα υποεπιπέδων μίας συνάρτησης πυκνότητας. Με εξαίρεση τη μέθοδο Bottleneck Bootstrap, που θα δούμε παρακάτω, όλες οι υπόλοιπες προτεινόμενες μέθοδοι που παρουσιάζονται στην

παρούσα εργασία βασίζονται στα αποτελέσματα για τη σταθερότητα των εμμενόντων διαγραμμάτων. Τα σύνολα εμπιστοσύνης για τα διαγράμματα προκύπτουν από τα σύνολα εμπιστοσύνης από το δειγματικό χώρο.

Υποδειγματοληπτική προσέγγιση

Αυτή η μέθοδος βασίζεται σε μία περιοχή για το στήριγμα K της κατανομής του δείγματος σε απόσταση Hausdorff. Έστω $\tilde{\mathbb{X}}_b$ ένα υπο-δείγμα μεγέθους b από το δείγμα $\tilde{\mathbb{X}}_n$, όπου $b = o(n/\log n)$. Έστω $q_p(1-a)$ ένα τεταρτημόριο της κατανομής του $Haus(\tilde{\mathbb{X}}_b, \mathbb{X}_n)$. Σε αυτό το σημείο θα θεωρήσουμε $\hat{\eta}_\alpha := 2\hat{q}_b(1-a)$, όπου $\hat{q}_b$ είναι μία εκτίμηση του $q_b(1-a)$, αξιοποιώντας την τυπική διαδικασία Monte Carlo. Υπό μία (a,b)-στάνταρ υπόθεση και για κάποιο αρκετά μεγάλο n, έχουμε: [0]

$$\left\{ \begin{array}{ll} P(d_b(dgm(Filt(K)), dgm(Filt(\mathbb{X}_n))) \geq \hat{\eta}_\alpha) & \leq P(Haus(K, \mathbb{X}_n) \geq \hat{\eta}_\alpha) \\ & \leq a + O\left(\frac{b}{n}\right)^{1/4} \end{array} \right\}$$

Bottleneck Bootstrap

Τα αποτελέσματα σχετικά με τη σταθερότητα συχνά οδηγούν σε συντηρητικά σύνολα εμπιστοσύνης. Μία εναλλακτική στρατηγική είναι εκείνη που ονομάζεται Bottleneck Bootstrap και έχει εισηχθεί από τους Chazal et al. (2016b). [0] Θεωρούμε ότι το γενικό πλαίσιο όπου ένα εμμένον διάγραμμα, έστω $\hat{dgm}$, που ορίζεται από τις παρατηρήσεις $(X_1, X_2, \ldots, X_n)$ σε ένα μετρικό χώρο. Αυτό το εμμένον διάγραμμα αντιστοιχεί στην εκτίμηση ενός υποκείμενου εμμένοντος διαγράμματος dgm, η οποία μπορεί να συνδεθεί, για παράδειγμα με το στήριγμα του μέτρου, ή τα σύνολα υποεπιπέδων μίας συνάρτησης, που σχετίζεται με την κατανομή. Ένα τέτοιο παράδειγμα είναι η συνάρτηση πυκνότητας όταν τα X_i ανήκουν στον $\mathbb{R}^d$. Έστω $(X*_1, X*_2, \ldots, X*_n)$ ένα δείγμα από το εμπειρικό μέτρο, που ορίζεται από τις παρατηρήσεις $(X_1, X_2, \ldots, X_n)$. Έστω $\hat{dgm}*$ το εμμένον διάγραμμα που προκύπτει από αυτό το δείγμα. Μπορούμε τώρα να αντί για το η_α να πάρουμε την ποσότητα $\hat{\eta}_\alpha$ που ορίζεται από την παρακάτω εξίσωση:

$$P(d_b(\hat{dgm}*, \hat{dgm}) \geq \hat{\eta}_\alpha | X_1, X_2, \ldots, X_n) = \alpha$$

Σημειώνεται ότι το η_α μπορεί με ευκολία να εκτιμηθεί με τη χρήση των διαδικασιών Monte Carlo. Έχει παρατηρηθεί ότι το Bottleneck Bootstrap είναι έγκυρο όταν χρησιμοποιείται για τον υπολογισμό των συνόλων υποεπιπέδων ενός εκτιμητή πυκνότητας. [0]

Ζώνες εμπιστοσύνης για τοπία

Ένας αλγόριθμος Bootstrap μπορεί να αξιοποιηθεί στην κατασκευή ζωνών εμπιστοσύνης για τοπία. [0] Από την άλλη μεριά, υποθέτουμε τώρα ότι τα διάφορα παρατηρούμενα τοπία, έστω $\lambda_1, \lambda_2, \ldots, \lambda_N$ επιλέγονται i.i.d. από μία τυχαία κατανομή στο χώρο των τοπίων. Έτσι, η στρατηγική του πολλαπλασιαστή Bootstrap μπορεί να χρησιμοποιηθεί στην κατασκευή μίας ζώνης εμπιστοσύνης για το $\mathbb{E}(\lambda_1)$.

Επιπλέον για θέματα που αφορούν στην κεντρική τάση της Εμμένουσας Ομολογίας ο αναγνώστης μπορεί να ανατρέξει στα έργα των Mileyko et al. (2011) [0] και Turner et al. (2014a) [0].

- Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, P., Chepushtanova, S., Hanson, E., Motta, F., and Ziegelmeier, L. (2017). Persistence images: a stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(8):1–35.

- Balakrishna, S., Rinaldo, A., Sheehy, D., Singh, A., and Wasserman, L. A. (2012). Minimax rates for homology inference. *Journal of Machine Learning* Research - Proceedings Track, 22:64–72.

- Biau, G., Chazal, F., Cohen-Steiner, D., Devroye, L., and Rodriguez, C. (2011). A weighted k-nearest neighbor density estimate for geometric inference. *Electronic Journal of Statistics*, 5:204–237.

- Bobrowski, O., Mukherjee, S., and Taylor, J. (2014). Topological consistency via kernel estimation. *arXiv preprint arXiv:1407.5272*

- Bonis, T., Ovsjanikov, M., Oudot, S., and Chazal, F. (2016). Persistence-based pooling for shape pose recognition. In *Computational Topology in Image Context - 6th International Workshop*, CTIC 2016, Marseille, France, June 15-17, 2016, *Proceedings*, pages 19–29.

- Brécheteau, C. (2017). The dtm-signature for a geometric comparison of metric-measure spaces from samples.

- Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- Bubenik, P. (2015). Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16:77–102.
- Bubenik, P., Carlsson, G., Kim, P. T., and Luo, Z.-M. (2010). Statistical topology via morse theory persistence and nonparametric estimation. Algebraic methods in statistics and probability II, 516:75–92.
- Buchet, M., Chazal, F., Dey, T. K., Fan, F., Oudot, S. Y., and Wang, Y. (2015a). Topological analysis of scalar fields with outliers. In *Proc. Sympos. on Computational Geometry*.
- Buchet, M., Chazal, F., Oudot, S., and Sheehy, D. R. (2015b). Efficient and robust persistent homology for measures. In *Proceedings of the 26th ACM-SIAM symposium on Discrete algorithms*. SIAM. SIAM.
- Cadre, B. (2006). Kernel estimation of density level sets. *Journal of multivariate analysis*, 97(4):999–1023.
- Cang, Z. and Wei, G. (2017). Topologynet: Topology based deep convolutional and multi task neural networks for biomolecular property predictions. PLoS Computational Biology, 13(7):e1005690.
- Carlsson, G. (2009). Topology and data. *AMS Bulletin*, 46(2):255–308.
- Carrière, M., Michel, B., and Oudot, S. (2017). Statistical analysis and parameter selection for mapper.
- Carrière, M. and Oudot, S. (2015). Structure and stability of the 1-dimensional mapper. *arXiv preprint arXiv 1511.05823*.
- Carriere, M. and Oudot, S. (2017). Sliced wasserstein kernel for persistence diagrams. To appear in ICML-17.
- Chazal, F. (2017). High-dimensional topological data analysis. In *Handbook of Discrete and Computational Geometry* (3rd Edition), chapter 27. CRC Press.

- Chazal, F. (2017). High-dimensional topological data analysis. In *Handbook of Discrete and Computational Geometry* (3rd Edition), chapter 27. CRC Press.

- Chazal, F., Chen, D., Guibas, L., Jiang, X., and Sommer, C. (2011a). Data-driven trajectory smoothing. In P*roc. ACM SIGSPATIAL GIS*.

- Chazal, F., Cohen-Steiner, D., Glisse, M., Guibas, L., and Oudot, S. (2009a). Proximity of persistence modules and their diagrams. In *SCG*, pages 237–246.

- Chazal, F., Cohen-Steiner, D., Guibas, L. J., M'emoli, F., and Oudot, S. Y. (2009b). Gromov hausdorff stable signatures for shapes using persistence. *Computer Graphics Forum (proc. SGP 2009)*, pages 1393–1403.

- Chazal, F., Cohen-Steiner, D., and Lieutier, A. (2009c). Normal cone approximation and offset shape isotopy. *Comp. Geom. Theor. Appl.*, 42(6-7):566–581.

- Chazal, F., Cohen-Steiner, D., and Lieutier, A. (2009d). A sampling theory for compact sets in euclidean space. *Discrete and Computational Geometry*, 41(3):461–479.

- Chazal, F., Cohen-Steiner, D., Lieutier, A., and Thibert, B. (2008). Stability of Curvature Measures. *Computer Graphics Forum (proc. SGP 2009)*, pages 1485–1496.

- Chazal, F., Cohen-Steiner, D., and Mérigot, Q. (2010). Boundary measures for geometric inference. *Found. Comp. Math.*, 10:221–240.

- Chazal, F., Cohen-Steiner, D., and Mérigot, Q. (2011b). Geometric inference for probability measures. *Foundations of Computational Mathematics*, 11(6):733–751.

- Chazal, F., de Silva, V., Glisse, M., and Oudot, S. (2016a). *The structure and stability of persistence modules*. SpringerBriefs in Mathematics. Springer.

- Chazal, F., Fasy, B. T., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. (2014a). Robust topological inference: Distance to a measure and kernel distance JMLR.

- Chazal, F., Fasy, B. T., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. (2015a). Subsampling methods for persistent homology. To appear in Proceedings of the 32 st International Conference on Machine Learning (ICML-15).

- Chazal, F., Fasy, B. T., Lecci, F., Rinaldo, A., and Wasserman, L. (2015b). Stochastic convergence of persistence landscapes and silhouettes. *Journal of Computational Geometry*, 6(2):140–161.

- Chazal, F., Glisse, M., Labruère, C., and Michel, B. (2014b). Convergence rates for persistence diagram estimation in topological data analysis. To appear in *Journal of Machine Learning Research*.

- Chazal, F., Guibas, L. J., Oudot, S. Y., and Skraba, P. (2013). Persistence-based clustering in riemannian manifolds. *Journal of the ACM (JACM)*, 60(6):41.

- Chazal, F., Huang, R., and Sun, J. (2015c). Gromov—hausdorff approximation of filamentary structures using reeb-type graphs. *Discrete Comput. Geom.*, 53(3):621–649.

- Chazal, F. and Michael, B. (2017). An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. *Cornell University Library*, Submitted: Wed, 11 Oct 2017 11:53:32 UTC

- Chazal, F. and Lieutier, A. (2008a). Smooth manifold reconstruction from noisy and non-uniform approximation with guarantees. *Comp. Geom. Theor. Appl.*, 40(2):156–170.

- Chazal, F. and Lieutier, A. (2008b). Smooth manifold reconstruction from noisy and non uniform approximation with guarantees. *Computational Geometry Theory and Applications*, 40:156–170.

- Chazal, F., Massart, P., and Michel, B. (2016b). Rates of convergence for robust geometric inference. *Electron. J. Statist*, 10:2243–2286.

- Chazal, F. and Oudot, S. Y. (2008). Towards persistence-based reconstruction in euclidean spaces. In *Proceedings of the twenty-fourth annual symposium on Computational geometry*, SCG '08, pages 232–241, New York, NY, USA. ACM.

- Chen, Y.-C., Genovese, C. R., and Wasserman, L. (2015). Density level sets: Asymptotics, inference, and visualization. *arXiv preprint arXiv:1504.05438*.

- Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2005). Stability of persistence diagrams. In *SCG*, pages 263–271.

- Cohen-Steiner, D., Edelsbrunner, H., Harer, J., and Mileyko, Y. (2010). Lipschitz functions have l p-stable persistence. *Foundations of computational mathematics*, 10(2):127–139.

- Cuevas, A. and Rodríguez-Casal, A. (2004). On boundary estimation. *Advances in Applied Probability*, pages 340–354.

- De Silva, V. and Carlsson, G. (2004). Topological estimation using witness complexes. In *Proceedings of the First Eurographics Conference on Point-Based Graphics*, SPBG'04, pages 157–166, Aire-la-Ville, Switzerland, Switzerland. Eurographics Association.

- De Silva, V. and Ghrist, R. (2007). Homological sensor networks. *Notices of the American mathematical society*, 54(1).

- Devroye, L. and Wise, G. L. (1980). Detection of abnormal behavior via nonparametric estimation of the support. *SIAM J. Appl. Math.*, 38(3):480–488

- Dey, T. K., Mémoli, F., and Wang, Y. (2016). Multiscale mapper: topological summarization via codomain covers. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 997–1013. Society for Industrial and Applied Mathematics.

- Dey, T. K., Mémoli, F., and Wang, Y. (2017). Topological analysis of nerves, reeb spaces, mappers, and multiscale mappers. In *Proc. Sympos. Comput. Geom. (SoCG)*.

- Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2002). Topological persistence and simplification. *Discrete Comput. Geom.*, 28:511–533.

- Fasy, B. T., Kim, J., Lecci, F., and Maria, C. (2014a). Introduction to the r package tda. *arXiv preprint arXiv:1411.1830*.

- Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. (2014b). Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339.

- Federer, H. (1959). Curvature measures. *Trans. Amer. Math. Soc.*, 93:418–491.

- Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2012). Manifold estimation and singular deconvolution under hausdorff loss. *Ann. Statist.*, 40:941–963

- Ghrist, R. (2017). Homological algebra and data. *preprint*.

- Grove, K. (1993). Critical point theory for distance functions. In *Proc. of Symposia in Pure Mathematics*, volume 54.

- Guibas, L., Morozov, D., and Mérigot, Q. (2013). Witnessed k-distance. *Discrete Comput. Geom.*, 49:22–45.

- Hatcher, A. (2001). *Algebraic Topology.* Cambridge Univ. Press.

- Hofer, C., Kwitt, R., Niethammer, M., and Uhl, A. (2017). Deep learning with topological signatures. *arXiv preprint arXiv:1707.04041.*

- Judd, Charles and, McCleland, Gary (1989). *Data Analysis.* Harcourt Brace Jovanovich. ISBN 0-15-516765-0.

- Koomey, G. J.(2006). *Best Practices for Understanding Quantitative Data*

- Kovacev-Nikolic, V., Bubenik, P., Nikolić, D., and Heo, G. (2016). Using persistent homology and dynamical distances to analyze protein binding. *Statistical applications in genetics and molecular biology*, 15(1):19–38.

- Kramar, M., Goullet, A., Kondic, L., and Mischaikow, K. (2013). Persistence of force networks in compressed granular media. *Physical Review E*, 87(4):042207.

- Kramár, M., Levanger, R., Tithof, J., Suri, B., Xu, M., Paul, M., Schatz, M. F., and Mischaikow, K. (2016). Analysis of kolmogorov flow and rayleigh–bénard convection using persistent homology. *Physica D: Nonlinear Phenomena*, 334:82–98.

- Kusano, G., Fukumizu, K., and Hiraoka, Y. (2017). Kernel method for persistence diagrams via kernel embedding and weight factor. *arXiv preprint arXiv:1706.03472.*

- Lee, Y., Barthel, S. D., Dłotko, P., Moosavi, S. M., Hess, K., and Smit, B. (2017). Quantifying similarity of pore-geometry in nanoporous materials. *Nature Communications*, 8.

- Li, C., Ovsjanikov, M., and Chazal, F. (2014). Persistence-based structural recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2014 IEEE Conference on, pages 2003–2010.

- Lum, P., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J., and Carlsson, G. (2013). Extracting insights from the shape of complex data using topology. *Scientific reports*, 3.

- Maria, C., Boissonnat, J.-D., Glisse, M., and Yvinec, M. (2014). The gudhi library: Simplicial complexes and persistent homology. In *International Congress on Mathematical Software*, pages 167–174. Springer.

- McKinney, W. et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. SciPy Austin, TX.

- Mileyko, Y., Mukherjee, S., and Harer, J. (2011). Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12):124007.

- Nakamura, T., Hiraoka, Y., Hirata, A., Escolar, E. G., and Nishiura, Y. (2015). Persistent homology and many-body atomic structure for medium-range order in the glass. *Nanotechnology*, 26(30):304001.

- Niyogi, P., Smale, S., and Weinberger, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete and Computational Geometry*, 39(1-3):419–441.

- Niyogi, P., Smale, S., and Weinberger, S. (2011). A topological view of unsupervised learning from noisy data. *SIAM Journal on Computing*, 40(3):646–663.

- Obayashi, I. and Hiraoka, Y. (2017). Persistence diagrams with linear machine learning models. *arXiv preprint arXiv:1706.10082*.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

- Petrunin, A. (2007). Semiconcave functions in Alexandrov's geometry. In *Surveys in differential geometry*. Vol. XI, pages 137–201. Int. Press, Somerville, MA.

- Phillips, J. M., Wang, B., and Zheng, Y. (2014). Geometric inference on kernel density estimates. *arXiv preprint 1307.7760*.

- Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters-an excess mass approach. The Annals of Statistics, pages 855–881.

- Reininghaus, J., Huber, S., Bauer, U., and Kwitt, R. (2015). A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4741–4748.

- Schutt, Rachel; O'Neil, Cathy (2013). *Doing Data Science*. O'Reilly Media. ISBN 978-1-449-35865-5.

- Seversky, L. M., Davis, S., and Berger, M. (2016). On time-series topological data analysis: new data and opportunities. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 59–67.

- Sherman, Rick,. *Business intelligence guidebook : from data integration to analytics.* Amsterdam. ISBN 978-0-12-411528-6. OCLC 894555128

- Singh, A., Scott, C., and Nowak, R. (2009). Adaptive Hausdorff estimation of density level sets. *Ann. Statist.*, 37(5B):2760–2782.

- Singh, G., Mémoli, F., and Carlsson, G. E. (2007). Topological methods for the analysis of high dimensional data sets and 3d object recognition. In *SPBG*, pages 91–100. Citeseer.

- Skraba, P., Ovsjanikov, M., Chazal, F., and Guibas, L. (2010). Persistence-based segmentation of deformable shapes. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 45–52

- Tsybakov, A. B. et al. (1997). On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3):948–969.

- Tukey, J. *The future of data analysis*, Princeton University and Bell Telephone Laboratories

- Turner, K., Mileyko, Y., Mukherjee, S., and Harer, J. (2014a). Fréchet means for distributions of persistence diagrams. *Discrete and Computational Geometry*, 52(1):44–70.

- Turner, K., Mukherjee, S., and Boyer, D. M. (2014b). Persistent homology transform for modeling shapes and surfaces. *Information and Inference: A Journal of the IMA*, 3(4):310–344.

- Umeda, Y. (2017). Time series classification via topological data analysis. *Transactions of the Japanese Society for Artificial Intelligence*, 32(3) : $D - G72_1$.

- Villani, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society.

- Walt, S. v. d., Colbert, S. C., and Varoquaux, G. (2011). The numpy array: a structure for efficient numerical computation. *Computing in Science and Engineering*, 13(2):22–30.

- Xia, B. S., Gong, P. (2015). *Review of business intelligence through data analysis. Benchmarking*, 21(2), 300-311. doi:10.1108/BIJ-08-2012-0050

- Yao, Y., Sun, J., Huang, X., Bowman, G. R., Singh, G., Lesnick, M., Guibas, L. J., Pande, V. S., and Carlsson, G. (2009). Topological methods for exploring low-density states in biomolecular folding pathways. *The Journal of chemical physics*, 130(14):144115.

- Zomorodian, A. and Carlsson, G. (2005). Computing persistent homology. *Discrete Comput. Geom.*, 33(2):249–274.

- Διαδικτυακός ιστότοπος: https://www.microsoft.com/en-us/research/project/data-cleaning/ τελευταία ενημέρωση: 05/11/2020