

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Γραμμικές και Μη-γραμμικές Μέθοδοι
Αναγωγής Δεδομένων Μεγάλης Κλίμακας



ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Συγγραφέας:

Ισμήνη ΜΠΟΥΛΙΑΡΗ

Επιβλέπων Καθηγητής:

Κωνσταντίνος ΣΙΕΤΤΟΣ

Τομέας Μηχανικής

Αθήνα

Μάρτης 2017

ΕΘΝΙΚΟ ΜΕΤΕΩΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Γραμμικές και Μη-γραμμικές Μέθοδοι Αναγωγής Μεγάλης Κλίμακας Δεδομένων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Συγγραφέας:

Ισμήνη ΜΠΟΥΛΙΑΡΗ

Επιβλέπων Καθηγητής:

Κωνσταντίνος ΣΙΕΤΤΟΣ

Τριμελής Εξεταστική Επιτροπή:

- Κομίνης Ιωάννης, Επίκουρος Καθηγητής,
- Ματσόπουλος Γεώργιος, Αναπληρωτής Καθηγητής ΣΗΜΜΥ,
- Σιέττος Κωνσταντίνος, Αναπληρωτής Καθηγητής

Τομέας Μηχανικής

Αθήνα

Μάρτης 2017

Ευχαριστίες

Με την ολοκλήρωση της διπλωματικής μου εργασίας θα ήθελα να ευχαριστήσω θερμά τον καθηγητή του ΕΜΠ κ. Κωνσταντίνο Σιέττο για την άριστη συνεργασία που είχαμε και για την αμέριστη βοήθεια που μου προσέφερε σε κάθε ζήτημα που προέκυψε κατά την εκπόνηση της εργασίας. Χωρίς την καθοδήγησή του, δε θα υπήρχε αυτή η εργασία.

Επίσης θα ήθελα να ευχαριστήσω την οικογένειά μου και όλους τους φίλους μου για την συνεχή συμπαράσταση και υποστήριξή τους σε όλη τη διάρκεια των σπουδών μου.

Πρόλογος

Ο σκοπός της παρούσας διπλωματικής εργασίας είναι η παρουσίαση νέων και αποτελεσματικών αλγορίθμων εκμάθησης πολλαπλοτήτων που βρίσκουν εφαρμογή στη Βιολογία, την Ιατρική και συναφείς επιστήμες και η ανάδειξη της χρησιμότητάς τους στο πεδίο της ανάλυσης μικροσυστοιχιών γονιδίων (microarray analysis). Για την επίτευξη του σκοπού αυτού, αρχικά γίνεται μια σύντομη επισκόπηση του ορισμού της εκμάθησης πολλαπλοτήτων και παρουσίαση κάποιων από τους πιο αντιπροσωπευτικούς και ευρέως χρησιμοποιούμενους αλγορίθμους της, εμβαθύνοντας σε αποδείξεις- μαθηματικές και διαισθητικές. Στη συνέχεια, περιγράφουμε τον τρόπο χρήσης αυτών των αλγορίθμων στην εξόρυξη δεδομένων, πρώτα παραθέτοντας ένα απλό παράδειγμα από τον χώρο των μαθηματικών και κάνοντας μία σύντομη σύγκριση της αποδοτικότητας μερικών από αυτούς. Στο δεύτερο μέρος επικεντρωνόμαστε στην τεχνολογία και την ανάλυση των μικροσυστοιχιών γονιδιακής έκφρασης. Δίνεται μία σύντομη περιγραφή των βιολογικών διαδικασιών στις οποίες βασίζεται η τεχνολογία μικροσυστοιχιών. Οι πληροφορίες αυτές είναι απαραίτητες σε έναν αναγνώστη, μη εξοικειωμένο με τον χώρο της Βιολογίας, ώστε να μπορεί να κατανοήσει σε μεγαλύτερο βαθμό τη συνέχεια της εργασίας. Επιπλέον, γίνεται- για λόγους πληρότητας- μια λεπτομερής αναφορά τόσο στην τεχνολογία και στον τρόπο κατασκευής μιας μικροσυστοιχίας γονιδιακής έκφρασης (η οποία είναι πολύ εξειδικευμένη και μπορεί να παραληφθεί), όσο και στην μαθηματική επεξεργασία που πρέπει να εφαρμοστεί στα δεδομένα ώστε να είναι ικανά να παρέχουν κατά το δυνατόν ακριβέστερα αποτελέσματα, όταν τα αναλύσουμε με κάποιον αλγόριθμο αναγωγής μεγάλης κλίμακας. Επικεντρωνόμαστε στις μικροσυστοιχίες της Affymetrix, της οποίας χρησιμοποιούμε το πρωτόκολλο στο πειραματικό μέρος της εργασίας. Το τελευταίο κεφάλαιο βασίζεται στη δημοσίευση των Dawson, Rodriguez και Malyj (Dawson et al. 2005). Με την αναπαραγωγή ενός μέρους των αποτελεσμάτων των τελευταίων, αποδεικνύεται πως ο αλγόριθμος Isomap όπως αυτός εφαρμόζεται σε μία τέτοια μικροσυστοιχία- μπορεί να ανταποκριθεί άψογα στην πρόκληση της ανακάλυψης υποκείμενων δομών στα βιολογικά δεδομένα. Με άλλα λόγια, γίνεται σαφές ότι ο Isomap αποτελεί έναν αποτελεσματικό αλγόριθμο αναγωγής δεδομένων μεγάλης κλίμακας, ο οποίος μπορεί

να αποδειχτεί ένα ισχυρό εργαλείο στην ανάλυση βιολογικών/ιατρικών δεδομένων. Παραθέτουμε, τέλος, παράρτημα που περιλαμβάνει κάποιες παραπάνω πληροφορίες για τους αλγόριθμους που αναφέρονται ακροθιγώς στην εργασία και τη σχετική βιβλιογραφία σε αλφαβητική σειρά.

Abstract

The purpose of this study is to present new and effective manifold learning algorithms that can be applied in Biology, Medicine and related science fields and to highlight their utility in the field of microarray analysis. To achieve this goal, in the first part, we give the definition of the manifold and we present some of the most famous and frequently used linear (PCA, MDS) and non-linear (LLE, Isomap, Spectral Clustering, Diffusion Maps) dimensionality reduction methods. In order to gain insight to the core of dimensionality reduction methods, we also give both intuitive and mathematical proof for some of them. Furthermore, we compare their efficiency by applying them on the Swiss roll benchmark problem. In the second part, we focus on the technology and the analysis of gene expression microarrays. We use Affymetrix' s GeneChip[®] microarrays and shortly describe their construction method. We give little biological information, which is essential to a reader unfamiliar with the field of Biology, in order to be able to deeply understand the rest of this study. We describe the entire process, from the manufacture of the chip until the export of the raw intensity values, which are to be analyzed. Moreover, we reproduce the results step by step of such an experiment with the Isomap algorithm, proving that it can perfectly respond to the challenge of finding interesting structures in big biological data sets and provide important and essential information about the importance of those structures in different biological procedures. In this way, the Isomap algorithm is proved to be one of the most effective algorithms for the analysis of large data sets, such as gene expression microarrays. In the end of this paper, one can find the relevant bibliography.

Περιεχόμενα

Εισαγωγή	11
I Τεχνικές Αναγωγής Δεδομένων Μεγάλης Κλίμακας	19
1 Ανάλυση Κύριων Συνιστωσών(PCA: Principal Components Analysis)	23
1.1 Η ιδέα της PCA	24
1.2 Υποθέσεις για την PCA	25
1.3 Η λύση της PCA- Γιατί δουλεύει η PCA	26
1.4 Ο αλγόριθμος PCA	27
2 Πολυδιάστατη Κλιμάκωση (MDS: MultiDimensional Scaling)	29
2.1 Η ιδέα και τα μαθηματικά πίσω από την MDS	29
2.2 Ο αλγόριθμος MDS	30
3 Απεικόνιση Ισομετρικών Συνιστωσών (Isomap: ISometric feature MAPping)	35
3.1 Η ιδέα του αλγόριθμου Isomap	35
3.2 Ο αλγόριθμος Isomap	36
4 Τοπικά Γραμμική Εμφύτευση (LLE: Locally Linear Embedding)	39
4.1 Ο αλγόριθμος LLE	39
4.2 Η βασική ιδέα του LLE	41

5	Φασματική Ομαδοποίηση (Spectral Clustering)	45
5.1	Εισαγωγή	45
5.2	Βασικά μαθηματικά εργαλεία των αλγορίθμων φασματικής ομαδοποίησης	46
5.2.1	Περί γράφων	46
5.2.2	Γράφος Λαπλασιανού πίνακα	48
5.3	Αλγόριθμοι Φασματικής ομαδοποίησης	50
5.4	Γιατί δουλεύουν οι αλγόριθμοι φασματικής ομαδοποίησης:	52
5.4.1	TOMES ΓΡΑΦΗΜΑΤΟΣ	52
5.4.2	ΤΥΧΑΙΟΣ ΠΕΡΙΠΙΑΤΟΣ	58
5.4.3	ΘΕΩΡΙΑ ΔΙΑΤΑΡΑΧΩΝ	60
6	Πίνακες Διάχυσης (Diffusion Maps)	65
6.1	Η Μαθηματική Προσέγγιση της συνάρτησης διάχυσης	66
6.1.1	Συνδεσιμότητα	68
6.1.2	Απόσταση Διάχυσης	68
6.1.3	Συνάρτηση Διάχυσης	69
6.2	Βασικός αλγόριθμος Diffusion Mapping	71
7	Εφαρμογές των αλγορίθμων και σύγκριση της αποτελεσματικότητάς τους	73
7.1	Αποτελεσματικότητα αλγορίθμων PCA-MDS	75
7.2	Αποτελεσματικότητα αλγορίθμου LLE	76
7.3	Αποτελεσματικότητα αλγορίθμου Isomap	77
7.4	Αποτελεσματικότητα αλγορίθμων [(Spectral Clustering(Laplacian Eigenmaps)]	78
7.5	Αποτελεσματικότητα αλγορίθμου Συναρτήσεων Διάχυσης (Diffusion Maps)	79
7.6	Σύγκριση αλγορίθμων	80

II Η τεχνολογία των μικροσυστοιχιών (microar-

rays)	83
8 Βιολογικό υπόβαθρο	85
8.1 Λίγα λόγια για το DNA	85
8.2 Γονίδια και Γονιδίωμα	86
8.3 Γονιδιακή έκφραση	86
9 Μικροσυστοιχίες γονιδιακής έκφρασης	89
9.1 Τι είναι οι μικροσυστοιχίες (microarrays;)	89
9.2 Affymetrix GeneChips	90
10 Από τα βιολογικά δείγματα στις ανεπεξέργαστες εντάσεις	95
10.1 ΠΡΟΕΤΟΙΜΑΣΙΑ ΔΕΙΓΜΑΤΟΣ ΚΑΙ ΣΗΜΑΝΣΗ	95
10.1.1 Σήμανση	95
10.1.2 Υβριδοποίηση και καθαρισμός	96
10.1.3 Σάρωση και ανάλυση εικόνας	97
10.2 ΠΡΟΕΡΓΑΣΙΑ ΤΩΝ ΠΡΩΤΩΝ ΤΙΜΩΝ ΕΝΤΑΣΗΣ	
Ένα βήμα πριν την ανάλυση	98
10.2.1 Διόρθωση υποβάθρου (background correction)	99
10.2.2 Διόρθωση υποβάθρου RMA (Robust Multi-array Average)	100
10.2.3 Μέθοδοι κανονικοποίησης	101
10.2.4 Ποσοστημοριακή Κανονικοποίηση (Quantile Normalisation)	101
10.2.5 Διόρθωση PM ανιχνευτών (μόνο για μικροσυστοιχίες Affymetrix)	103
10.2.6 Περίληψη δεδομένων	105
11 Ένα πείραμα μικροσυστοιχιών	107
11.1 ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΤΩΝ ΔΕΔΟΜΕΝΩΝ	110
11.2 ΔΙΑΔΙΚΑΣΙΑ ΑΝΑΛΥΣΗΣ	116
11.2.1 Εφαρμογή αλγορίθμου Isomap	116
11.2.2 Αποτελέσματα εφαρμογής αλγορίθμου Isomap	118
11.2.3 Οπτικοποίηση αποτελεσμάτων	120

11.3 ΑΠΟΤΕΛΕΣΜΑΤΑ	124
12 ΠΑΡΑΡΤΗΜΑ 1:Λίγα λόγια για τους αλγόριθμους εκμάθησης πολλαπλοτήτων	131
Bibliography	137

Εισαγωγή

Η εκμάθηση πολλαπλοτήτων (manifold learning) είναι μία από τις πιο ευρέως χρησιμοποιούμενες μεθόδους ανάλυσης δεδομένων που χρησιμοποιείται για την ακριβή ομαδοποίηση των δεδομένων μεγάλης κλίμακας. Αποτελεί ένα σημαντικό πρόβλημα σε μια μεγάλη ποικιλία τομέων επεξεργασίας πληροφοριών (information processing) όπως η αναγνώριση προτύπων (pattern recognition), η συμπίεση δεδομένων (data compression) και η μηχανική μάθηση (machine learning). Σε πολλές περιπτώσεις τα διανύσματα δεδομένων (μετρήσεις) είναι πολύ μεγάλων διαστάσεων, αλλά μπορεί να έχουμε λόγο να πιστεύουμε ότι τα δεδομένα βρίσκονται κοντά σε μία μικρότερης διάστασης πολλαπλότητα (manifold). Με άλλα λόγια, μπορεί να πιστεύουμε πως τα μεγάλων διαστάσεων δεδομένα είναι πολλαπλές, έμμεσες μετρήσεις μιας υποκείμενης πηγής, η οποία κατά κανόνα, δεν μπορεί να μετρηθεί απευθείας. Το να γνωρίσουμε μία μικρών διαστάσεων πολλαπλότητα από μεγάλης κλίμακας δεδομένα, είναι στην ουσία η κατανόηση της υποκείμενης πηγής. Η αναγωγή δεδομένων μεγάλης κλίμακας (dimensionality reduction) μπορεί να θεωρηθεί ως η διαδικασία άντλησης ενός συνόλου βαθμών ελευθερίας που μπορούν να χρησιμοποιηθούν για την παραγωγή του μεγαλύτερου μέρους της μεταβλητότητας ενός συνόλου δεδομένων (data set). Οι τεχνικές εκμάθησης πολλαπλοτήτων μπορούν να χρησιμοποιηθούν με διάφορους τρόπους, όπως:

- Αναγωγή δεδομένων μεγάλης κλίμακας: παραγωγή μιας συμπαγούς χαμηλής διάστασης κωδικοποίησης ενός μεγάλης διάστασης συνόλου δεδομένων.
- Οπτικοποίηση δεδομένων: «μετάφραση» ενός δεδομένου συνόλου δεδομένων ως προς τους βαθμούς ελευθερίας, συνήθως σαν υποπροϊόν της αναγωγής μεγάλων δεδομένων.
- Προεπεξεργασία για επιβλεπόμενη μάθηση (supervised learning).
- Απλοποίηση, μείωση και εκκαθάριση δεδομένων για μεταγενέστερη επιβλεπόμενη μάθηση.

Αυτή η διαδικασία επεξεργασίας και ανάλυσης των δεδομένων μπορεί να προσφέρει τη δυνατότητα σε ερευνητές άλλων επιστημονικών κλάδων (βιολόγων, ιατρών, στατιστικολόγων, φυσικών κλπ) να αποκτήσουν μία πιο ολοκληρωμένη εποπτεία της εγγενούς δομής των δεδομένων και να τους βοηθήσουν στην περαιτέρω ανάλυσή τους στο μέτρο και για τους σκοπούς που επιβάλλει ο δικός τους τομέας εξειδίκευσης.

Ένα πιο συγκεκριμένο παράδειγμα τέτοιας μορφής δεδομένων μεγάλης κλίμακας που προέρχεται από τον χώρο της Βιοπληροφορικής, αποτελούν τα σύνολα δεδομένων μικροσυστοιχιών (microarrays). Πρόκειται για μια μεγάλη πηγή γενετικών δεδομένων, η οποία, μετά από κατάλληλη ανάλυση, θα μπορούσε να ενισχύσει την κατανόηση των βιολογικών διεργασιών και να βοηθήσει στην εξέλιξη κλάδων όπως η ιατρική και η φαρμακευτική. Οι μικροσυστοιχίες γονιδιακής έκφρασης αποτελούν μία δοκιμασία που μετρά τα επίπεδα έκφρασης δεκάδων χιλιάδων γονιδίων παράλληλα σε ένα μόνο τσιπ. Οι μικροσυστοιχίες μπορούν να κατασκευαστούν από μια πολύ μικρή ποσότητα ενός βιολογικού δείγματος, επιτρέποντας έτσι, έναν πειραματικό σχεδιασμό που περιλαμβάνει πολλές ομάδες δειγμάτων, επαναλήψεις, πυκνές χρονοσειρές, και δείγματα που συλλέγονται σε υψηλή διακριτικότητα από διάφορες ανατομικές θέσεις. Σήμερα, το κόστος των μικροσυστοιχιών είναι ο κύριος παράγοντας που περιορίζει τον αριθμό των δειγμάτων που μπορούν να εξεταστούν σε ένα συγκεκριμένο πείραμα. Με την ευρεία χρήση των μικροσυστοιχιών στη βασική έρευνα και την αυξανόμενη χρήση τους στην ιατρική διάγνωση, οι ερευνητές των τομέων αυτών προβλέπουν τη μείωση του κόστους των τσιπς που θα οδηγήσει σε περισσότερες μελέτες που θα κάνουν χρήση εκατοντάδων, αν όχι χιλιάδων δειγμάτων. Αυτή η επέκταση του μεγέθους του δείγματος θα προσφέρει στους ερευνητές μία καλύτερη επίγνωση των βιολογικών διαδικασιών, όπως αυτές αντανακλώνονται σε χρονικά, χωρικά και λειτουργικά μοτίβα σε σύνολα δεδομένων μικροσυστοιχιών. Για να αποκαλυφθούν αυτά τα μοτίβα, διάφοροι τύποι των τεχνικών αναγνώρισης προτύπων και ομαδοποίησης έχουν αναπτυχθεί και προσαρμόσκει για την κάλυψη των αναγκών της ανάλυσης δεδομένων μικροσυστοιχιών.

Οι βιολογικές διαδικασίες καθορίζονται από το γενετικό προφίλ του κάθε οργανισμού και των πολλαπλών περιβαλλοντικών μεταβλητών. Ωστόσο, η αλληλεπίδραση μεταξύ αυτών των παραγόντων είναι εγγενώς μη-γραμμική (Nicholson

et al. 2004). Οι βάσεις δεδομένων μικροσυστοιχιών αποτελούν μία αναπαράσταση των μη γραμμικών αλληλεπιδράσεων ανάμεσα στα ίδια τα γονίδια, αλλά και μεταξύ γονιδίων και περιβαλλοντικών παραγόντων. Ακόμα περισσότερες μελέτες μικροσυστοιχίας χρησιμοποιούν γραμμικές μεθόδους για την ερμηνεία μη-γραμμικών δεδομένων. Μια προσέγγιση για μείωση των διαστάσεων που μπορούν να λάβουν υπόψη την ενδεχόμενη μη- γραμμικότητα των δεδομένων βασίζεται στην υπόθεση ότι τα δεδομένα (γονίδια ενδιαφέροντος) βρίσκονται σε μία εμφυτευμένη (embedded), μη γραμμική πολλαπλότητα που έχει χαμηλότερη διάσταση από τον χώρο των ακατέργαστων δεδομένων και βρίσκεται μέσα σε αυτόν. Αλγόριθμοι που βασίζονται στην εκμάθηση πολλαπλοτήτων έχουν καλή απόδοση όταν προκειται για ένα τεχνητά κατασκευασμένο σύνολο δεδομένων υψηλής διάστασης. Αν και κάθε σημείο ορίζεται από χιλιάδες μεταβλητές, μπορεί να χαρακτηριστεί με ακρίβεια λαμβάνοντας υπ'όψιν μόνο μερικές από αυτές. Σε ορισμένες μελέτες δείγματα έχουν ληφθεί από μία χαμηλών διαστάσεων πολλαπλότητα εμφυτευμένη σε ένα χώρο υψηλής διάστασης (Staunton et al. 2001).

Μια συνηθισμένη εργασία για την ανάλυση μεγάλων συνόλων δεδομένων μικροσυστοιχιών είναι η ταξινόμηση του δείγματος με βάση τα πρότυπα γονιδιακής έκφρασης. Αυτή η διαδικασία μπορεί να διαιρεθεί σε δύο βήματα: την πρόβλεψη κλάσης (class prediction) και την ανακάλυψη κλάσης (class discovery). Κατά τη διάρκεια της πρόβλεψης της κλάσης τα δείγματα ταξινομούνται σε προκαθορισμένες κατηγορίες. Αντίθετα, η ανακάλυψη κλάσεων αφορά στην εύρεση νέων κλάσεων στο δείγμα. Για παράδειγμα, όταν οι συστοιχίες γονιδιακής έκφρασης χρησιμοποιούνται για την ταξινόμηση του καρκίνου, η πρόβλεψη κλάσεων ταξινομεί τα δείγματα όγκων σε προϋπάρχουσες ομάδες κακοηθειών, ενώ η ανακάλυψη κλάσης αποκαλύπτει προηγουμένως άγνωστους υποτύπους καρκίνου (Golub et al. 1999). Οι υποτύποι του όγκου μπορεί να έχουν διαφορετικές κλινικές μορφές, να αντιδρούν διαφορετικά σε ορισμένα φάρμακα, και να απαιτούν περισσότερο ή λιγότερο επιθετική χειρουργική και ραδιολογική θεραπεία. Η ανακάλυψη κλάσεων μπορεί επίσης να αποκαλύψει άγνωστες διαδικασίες στη βιολογία του καρκίνου και να καθορίσει πιο συγκεκριμένες ενδείξεις για ορισμένα φάρμακα. Ειδικά φάρμακα που μπορούν να χρησιμοποιηθούν για να στοχεύουν συγκεκριμένους υποτύπους όγκου ανακαλύφθηκαν πρόσφατα με

τη χρήση παρόμοιων τεχνολογιών, διευκολύνοντας έτσι τον φαρμακογονιδιω-
ματικό σχεδιασμό και την ανάπτυξη φαρμάκων. Οι στόχοι αυτοί θα γίνουν
σύντομα εφικτοί με τα αποτελέσματα από τις μελέτες των μικροσυστοιχιών που
χρησιμοποιούν μεγάλα δείγματα. Η πρόβλεψη και η ανακάλυψη κλάσεων με τη
χρήση μεγάλων συνόλων δεδομένων θα απαιτήσει την αξιολόγηση, την προσαρ-
μογή και την ανάπτυξη ισχυρών μαθηματικών, στατιστικών και υπολογιστικών
εργαλείων.

Πολλά πειράματα μικροσυστοιχιών έχουν σχεδιαστεί για τη διερεύνηση των
γενετικών μηχανισμών του καρκίνου, και έχουν εφαρμοστεί αναλυτικές προσεγ-
γίσεις για την ταξινόμηση διάφορων τύπων καρκίνου ή για τη διάκριση μεταξύ
καρκινικών και μη-καρκινικών ιστών. Ωστόσο, οι μικροσυστοιχιές είναι σύνολα
δεδομένων μεγάλων διαστάσεων με υψηλά επίπεδα θορύβου, γεγονός που
προκαλεί προβλήματα κατά τη χρήση μεθόδων μηχανικής εκμάθησης (machine
learning). Μια δημοφιλής προσέγγιση σε αυτό το πρόβλημα είναι η αναζη-
τηση μιας σειράς από χαρακτηριστικά που θα απλοποιήσουν τη δομή και σε
κάποιο βαθμό θα αφαιρέσουν το θόρυβο από τα δεδομένα. Αρκετοί μαθημα-
τικοί αλγόριθμοι και υπολογιστικές μέθοδοι έχουν εφαρμοστεί στην πρόβλεψη
και ανακάλυψη κλάσεων σε μεγάλα σύνολα δεδομένων γονιδιακής έκφρασης. Οι
μέθοδοι που χρησιμοποιούνται πιο συχνά βασίζονται σε τεχνικές ομαδοποίησης
όπως:

- Ιεραρχική Συσταδοποίηση (HC Hierarchical Clustering) (Eisen et al. 1998).
Η HC χρησιμοποιήθηκε για χρονική ταξινόμηση σε συνδυασμό με την
ανάλυση Fourier για την ανίχνευση γονιδίων που συσχετίζονται με πε-
ριοδικές αλλαγές στα συγχρονισμένα κύτταρα *S. cerevisiae* (Spellman
et al. 1998). Η HC εφαρμόστηκε επίσης στην ταξινόμηση του καρκίνου
(π.χ. ταξινόμηση του καρκίνου του μαστού) (Perou et al. 2000).
- Μη δομημένη συσταδοποίηση k -μέσων (unstructured k -means cluster-
ing) (Tavazoie et al. 1999),
- Αναζήτηση συγγενών ομάδων (CAST: Cluster Affinity Search) (Ben-
Dor et al. 1999),
- Ασαφής συσταδοποίηση c -μέσων (fuzzy c -means clustering) (Wang et al.

2003),

- Αμφίδρομη ομαδοποίηση (two-way clustering)(Alon et al. 1999) που χρησιμοποιήθηκε για την ανάλυση των αλληλεπιδράσεων φαρμάκων-όγκου (Scherf et al. 2000).
- Χάρτες αυτο-οργάνωσης (SOM: Self- Organizing Maps) είναι μια άλλη τεχνική που είναι κατάλληλη για διερευνητική ανάλυση των δεδομένων. Σε αντίθεση με την HC, η SOM δεν επιβάλλει μια στέρεη δομή στα δεδομένα (Tamayo et al. 1999). Η χρησιμότητα των SOM αποδείχθηκε στην ταξινόμηση της λευχαιμίας, χρησιμοποιώντας μια σταθμισμένη διαδικασία εκλογής (weighted voting procedure)(Golub et al. 1999)].
- Η σταθμισμένη ταξινόμηση εκλογής χρησιμοποιήθηκε επίσης για την πρόβλεψη της χημειοευαισθησίας της συλλογής όγκων NCI-60 (Staunton et al. 2001) και ανθρώπινων καρκίνων του μαστού (Perou et al. 2000).

Οι μέθοδοι με επίβλεψη, όπως η γραμμική διακριτική ανάλυση του Fisher(Fisher's linear discriminant analysis), έχουν το πλεονέκτημα ότι οι κλάσεις του δείγματος ορίζονται συνήθως με βάση κάποιον άλλον «χρυσό κανόνα» (π.χ., γνώσεις που προέρχονται από της ιστολογία, την κλινική έκβαση πειραμάτων, τον χρόνο επιβίωσης, κλπ). Στην επιβλεπόμενη ταξινόμηση (supervised classification), η επιλογή των ταξινομητών βασίζεται συχνά σε άλλες θεωρήσεις, π.χ., γονίδια που παίζουν ρόλο στον παθομηχανισμό μιας ορισμένης ασθένειας ή εκφράζονται σε ένα συγκεκριμένο ιστό. Αυτές οι μέθοδοι 'εμπλουτισμού' μπορούν να βελτιώσουν την ισχύροτητα της πρόβλεψης κλάσης και να μειώσουν τον αριθμό δειγμάτων που απαιτούνται για την ανάπτυξη ενός μοντέλου πρόβλεψης. Ωστόσο εισάγουν επίσης μεροληψία (bias) που μπορεί να οδηγήσει σε υπερβολικό χειρισμό, γεγονός που εγχυμονεί τον κίνδυνο μιας ενδεχόμενης αποτυχίας ανακάλυψης μη αναμενόμενων/άγνωστων κλάσεων που μπορεί να υπάρχουν στο δείγμα και θα μπορούσαν να προσφέρουν περισσότερες πληροφορίες γι' αυτό. Κάποιες μέθοδοι μείωσης διάστασης με επίβλεψη είναι:

- Μηχανές Διανυσμάτων Υποστήριξης (SVM Support Vector Machines): έχει το πλεονέκτημα ότι δεν κάνει υποθέσεις σχετικά με την κατανομή

των δεδομένων (Brown et al. 2000). Η SVM δοκιμάστηκε σε σύνολα δεδομένων για καρκίνο των ωοθηκών, λευχαιμία, και όγκων του παχέος εντέρου (Furey et al. 2000) και αποδείχθηκε επίσης ότι είναι χρήσιμη για την ταξινόμηση καρκίνου που ανήκει σε πολλές κλάσεις (Ramaswamy et al. 2001, Su et al. 2001, Yeang et al. 2001).

- Τεχνητά νευρωνικά δίκτυα (ANN :Artificial neural networks): μια άλλη μέθοδος μηχανικής μάθησης που χρησιμοποιήθηκε για να ταξινομήσει αδιαφοροποίητα κακοήθη κύτταρα(μικροκυτταρικοί όγκοι με σφαιρικά αδιαφοροποίητα κύτταρα που βράφονται με μπλε στη χρώση αιματοξυλίνης-ηωσίνης. SRBCT: small round blue-cell tumors) (Khan et al. 2001).
- Δένδρο συγκομιδής (Harvesting tree), εφαρμόστηκε πρόσφατα για τα δεδομένα γονιδιακής έκφρασης (Hastie et al. 2001). Σε αντίθεση με αυτές τις πολύπλοκες διαδικασίες, πολύ απλούστεροι ταξινομητές μπορούν επίσης να έχουν εξίσου καλή απόδοση σε ορισμένα σύνολα δεδομένων.
- Τα πλησιέστερα συρρικνωμένα κεντροειδή (nearest shrunken centroids): Εφαρμόστηκε σε κύτταρα SRBCT και λευχαιμίες (Tibshirani et al. 2002). Οι μέθοδοι μείωσης διάστασης είναι χρήσιμες για την πρόβλεψη της υποκείμενης πραγματικής διάστασης ενός συνόλου δεδομένων μικροσυστοιχιών και μειώνουν τον αριθμό των μεταβλητών που δίνονται ως είσοδοι σε οποιαδήποτε από τις διαδικασίες ταξινόμησης.
- Η πολυδιάστατη κλιμάκωση (MDS: MultiDimensional Scaling) Χρησιμοποιήθηκε για την ταξινόμηση του κυψελιδικού ραβδομοσρκώματος (Khan et al. 1998), κακοήθων δερματικών μελανώματων (Bittner et al. 2000), και καρκίνων του μαστού (Hedenfalk et al. 2001).
- Η πιο ευρέως χρησιμοποιούμενη προσέγγιση για να χαρακτηρίσει την εξόρυξη είναι η ανάλυση κύριων συνιστωσών (PCA: Principal Components Analysis), η οποία προϋποθέτει ένα πολυμεταβλητό Γκαουσιανό μοντέλο δεδομένων, χρησιμοποιήθηκε για την οπτικοποίηση χαρτών γονιδιακής έκφρασης της ανάπτυξης του κεντρικού νευρικού συστήματος (ΚΝΣ) (Wen et al. 1998) και την ταξινόμηση όγκων του ΚΝΣ σε έμβρυα (Pomeroy et al. 2002).

- Η πιθανολογική PCA (Probabilistic PCA): αποτελεί μια μέθοδο ενσωμάτωσης βιολογικών υποθέσεων εργασίας σε μοντέλα γραμμικών συντελεστών και εφαρμόστηκε πρόσφατα σε δύο σύνολα δεδομένων μικροσυστοιχιών ζυμομηκύτων (Girolami & Breitling 2004).
- Η αποσύνθεση ιδιαζουσών τιμών (SVD: Singular Value Decomposition), εφαρμόστηκε σε όγκους μαλακών ιστών (Nielsen et al. 2002), στον κυτταρικό κύκλο του *S. Cerevisiae* (ζυμομήκυτας), και σύνολα δεδομένων και ινοβλάστες που προέρχονται από ορό αίματος (Holter et al. 2000, 2001).
- Η γενικευμένη SVD (GSVD: generalized SVD) αναπτύχθηκε για να εξαχθούν πρότυπα γονιδιακής έκφρασης δύο διαφορετικών οργανισμών που είναι κοινά και συγκρίσιμα (Alter et al. 2003). Όλες αυτές οι γραμμικές μέθοδοι είναι εγγενώς ευαίσθητες σε ακραίες και σε μη διαθέσιμες τιμές, και σε κατανομές διαφορετικές από την κανονική.
- Μια παραλλαγή της SVD, η ισχυρή SVD (rSVD: robust SVD), αναπτύχθηκε πρόσφατα για την ελαχιστοποίηση της επίδρασης αυτών των αλλοιώσεων στο σύνολο δεδομένων (Liu et al. 2003).

Η Sammon χαρτογράφηση (Sammon mapping), μια μη γραμμική μέθοδος χαρτογράφησης (Sammon 1969) ενσωματώθηκε στο πακέτο `multiv` της R καθώς και λογισμικό διερευνητικής ανάλυσης και επεξεργασίας των δεδομένων γονιδιακής έκφρασης, όπως το `ENGINE` (De La Nava et al. 2003).

Μια συχνά χρησιμοποιούμενη μέθοδος για την εξεύρεση μιας κατάλληλης πολλαπλότητας, έχει εφαρμοστεί για την αντιμετώπιση παρόμοιων ζητημάτων. Ο `Isomap` (Tenenbaum et al. 2000), μια μη γραμμική τεχνική μείωσης διάστασης που αρχικά είχε σχεδιαστεί για την επίλυση κλασικών προβλημάτων της αναγνώρισης προτύπων, όπως η οπτική αντίληψη και η αναγνώριση γραφικού χαρακτήρα, που έχει ως στόχο την προβολή των δεδομένων από ένα υψηλότερο διαστάσεων χώρο σε έναν άλλον, μικρότερης διάστασης έχει εφαρμοστεί. Ο συγκεκριμένος αλγόριθμος κατασκευάζει την πολλαπλότητα ενώνοντας κάθε σημείο μόνο με τους κοντινότερους γείτονές του. Οι αποστάσεις μεταξύ των

σημείων τότε λαμβάνονται μέσω των γεωδαισιακών αποστάσεων του γραφήματος που προκύπτει.

- Ο Isomap (Tenenbaum et al. 2000) εφαρμόστηκε πρόσφατα στην ανακάλυψη βιολογικά σχετικών δομών στις μικροσυστοιχίες cDNA (Nilsson et al. 2004). Έχει, επίσης, εφαρμοστεί στην ανάλυση δύο συνόλων δεδομένων μικροσυστοιχιών καρκίνου του μαστού και πρωτεομικών φασμάτων καρκίνου του προστάτη και αποδείχτηκε ότι ξεπέρασε σταθερά την PCA στην αποκάλυψη βιολογικά σχετιζόμενων χαμηλών διαστάσεων δομών σε σύνολα δεδομένων μεγάλης κλίμακας. Οι Nilsson et al. ανεξάρτητα απέδειξαν την χρησιμότητα του Isomap χρησιμοποιώντας ένα λέμφωμα και μια cDNA μικροσυστοιχία που αντιστοιχούσε σε ένα σύνολο δεδομένων αδενοκαρκινώματος του πνεύμονα (Nilsson et al. 2004).
- Πολλές παραλλαγές του Isomap έχουν επίσης χρησιμοποιηθεί. Για παράδειγμα οι Balasubramanian και Schwartz (Balasubramanian & Schwartz 2002) παρουσίασαν μία εκδοχή ενός συνεκτικού δέντρου η οποία διαφέρει στον τρόπο κατασκευής του γράφου γειτνίασης. Τα k -κοντινότερα σημεία βρέθηκαν κατασκευάζοντας ένα ελάχιστο γεννητικό δένδρο κάνοντας χρήση μιας υπερσφαίρας (δηλ, ενός συνόλου σημείων που βρίσκονται σε μία συγκεκριμένη απόσταση από ένα δεδομένο σημείο, το κέντρο της) ακτίνας ϵ . Ο Isomap έχει δοκιμαστεί σε δεδομένα μικροσυστοιχιών με μερικά πολύ καλά αποτελέσματα (Dawson et al. 2005, Orsenigo & Vercellis 2012), μερικά εκ των οποίων θα αναλυθούν παρακάτω (Κεφ. 11). Σε σύγκριση με την PCA, ο Isomap ήταν σε θέση να αποσπάσει περισσότερες πληροφορίες σχετικές με τη δομή των δεδομένων.

Μέρος Ι

Τεχνικές Αναγωγής Δεδομένων Μεγάλης Κλίμακας

Σε αυτό το κεφάλαιο δίνουμε έναν σύντομο ορισμό της πολλαπλότητας και παρουσιάζουμε τις ιδέες που οδήγησαν σε κάποιες από τις πιο χαρακτηριστικές γραμμικές μεθόδους αναγωγής δεδομένων μεγάλης κλίμακας. Δίνονται επίσης οι μαθηματικές ή διαισθητικές- και οι αλγόριθμοι γραμμένοι σε ψευδοκώδικα για μια πιο σύντομη και περιεκτική εποπτεία τους.

Ορισμός της πολλαπλότητας (manifold)

Στα Μαθηματικά η πολλαπλότητα ορίζεται ως ένας τοπολογικός χώρος που τοπικά προσιδιάζει στον Ευκλείδειο χώρο (π.χ. κάθε σημείο μιας n -διάστατης πολλαπλότητας έχει μία γειτονιά που είναι ομοιομορφική με τον n -διάστατο Ευκλείδειο χώρο). Ο πιο συνηθισμένος τρόπος να περιγραφεί μία πολλαπλότητα είναι ένα παράδειγμα κάποιας μαθηματικής δομής που περιέχεται μέσα σε μια άλλη δομή, όπως μια ομάδα η οποία είναι μια υποομάδα. Η διαδικασία αυτή ονομάζεται εμφύτευση(embedding).

Η γραμμική περίπτωση

Κεφάλαιο 1

Ανάλυση Κύριων Συνιστωσών(PCA: Principal Components Analysis)

Η Ανάλυση Κύριων Συνιστωσών (PCA: Principal Components Analysis) (Jolliffe 2002) είναι μια πολύ διάσημη τεχνική αναγωγής μεγάλων δεδομένων, με εφαρμογές όπως η αναγνώριση προσώπων (face recognition) και η συμπίεση εικόνας (image compression).

Δοθέντος ενός συνόλου δεδομένων διάστασης n , η PCA στοχεύει στην εύρεση του πιο ουσιαστικού γραμμικού υπόχωρου διάστασης d , μικρότερης του n , τέτοιου ώστε τα σημεία των δεδομένων να βρίσκονται κατά κύριο λόγο σε αυτόν τον γραμμικό υπόχωρο. Ένας τέτοιος μειωμένος υπόχωρος επιδιώκει να διατηρήσει στο μέγιστο δυνατό βαθμό τη μεταβλητότητα των δεδομένων. (Με άλλα λόγια, η PCA ψάχνει τον καταλληλότερο πίνακα αλλαγής βάσης, ώστε να ξαναεκφράσει τα δεδομένα χωρίς απώλεια πληροφορίας και χωρίς εξωτερικές αλλοιώσεις των μετρήσεων.) Ο γραμμικός υπόχωρος μπορεί να οριστεί από d ορθογώνια διανύσματα που σχηματίζουν ένα νέο σύστημα συντεταγμένων, και καλούνται κύριες συνιστώσες (principal components). Οι κύριες συνιστώσες είναι ορθογώνιοι, γραμμικοί μετασχηματισμοί των αρχικών σημείων δεδομένων, επομένως μπορούν να είναι το πολύ n το πλήθος.

Με αυτόν τον τρόπο λύνουμε το πρόβλημα του πλεονασμού δεδομένων/

μετρήσεων- που οφείλεται είτε σε μετρήσεις παραγόντων οι οποίοι δε συμβάλουν στην καλύτερη κατανόηση των δεδομένων ή που μπορούν να εκφραστούν συναρτήσει άλλων παραγόντων, είτε σε θόρυβο-και εργαζόμαστε μόνο με τους παράγοντες που έχουν σημαντική επίδραση στο εκάστοτε πρόβλημα.

Ο συνηθέστερος ορισμός της PCA είναι πως για δοσμένο σύνολο διανυσμάτων δεδομένων $x_i, i \in 1, \dots, m$, οι d κυρίαρχοι άξονες (principal axes) είναι εκείνοι οι ορθοκανονικοί άξονες πάνω στους οποίους η διακύμανση που διατηρείται υπό προβολή είναι η μέγιστη.

1.1 Η ιδέα της PCA

Όπως αναφέρθηκε νωρίτερα, οι κύριοι στόχοι της PCA είναι

- i Η μεγιστοποίηση του σήματος (ουσιώδη δεδομένα που λαμβάνουμε από τις μετρήσεις)
- ii Η ελαχιστοποίηση των πλεοναζουσών μεταβλητών

Για την επίτευξη του πρώτου στόχου είναι προφανές πως πρέπει να είμαστε κατά το δυνατόν σίγουροι πως τα δεδομένα δεν αλλοιώνονται απο το θόρυβο. Κοινό μέτρο για τον θόρυβο αποτελεί ο λόγος SNR (signal-to-noise ratio).

$$SNR = \frac{\sigma_{signal}^2}{\sigma_{noise}^2}$$

Μεγάλη τιμή $SNR \gg 1$ φανερώνει μεγάλη ακρίβεια στα δεδομένα, ενώ μικρή τιμή δείχνει πως τα δεδομένα δεν είναι αξιόπιστα εξαιτίας μεγάλου θορύβου. Συνεπώς, ένας παράγοντας που πρέπει να εξεταστεί είναι η διακύμανση του σήματος, δηλαδή, στις περιπτώσεις που μας αφορούν, οι διακυμάνσεις των διαφόρων μεταβλητών .

Αντίστοιχα, είναι εύκολο να αντιληφθούμε πως για την ελαχιστοποίηση των πλεοναζουσών μεταβλητών, αρκεί να βρεθεί ένα μέτρο συσχέτισης όλων των μεταβλητών μεταξύ τους, το οποίο είναι φυσικά η συνδιασπορά, η οποία πρέπει στην ιδανική περίπτωση να είναι μηδενική.

Με βάση τα παραπάνω, μπορούμε να αντιληφθούμε πως το καταλληλότερο μέγεθος που συγκεντρώνει τα μεγέθη που θέλουμε να εξετάσουμε είναι ο πίνακας συνδιασποράς. Πράγματι, αν θεωρήσουμε πως τα δεδομένα μας εισάγονται $m \times n$ σε έναν πίνακα X , όπου x_1, \dots, x_m τα διανύσματα-γραμμές του X . Θεωρούμε ακόμη πως κάθε γραμμή του X αποτελεί τις μετρήσεις ενός συγκεκριμένου τύπου μεταβλητής, ενώ κάθε στήλη του αποτελεί τις μετρήσεις των διαφόρων μεταβλητών για μία συγκεκριμένη δοκιμή. Τότε ορίζεται ο πίνακας συνδιασποράς C_X , όπου

$$C_X \equiv \frac{1}{n} X X^T$$

Ο πίνακας συνδιασποράς C_X έχει τα εξής χαρακτηριστικά:

- Είναι τετραγωνικός συμμετρικός $m \times m$ πίνακας.
- Τα διαγώνια στοιχεία του είναι οι διακυμάνσεις των m μεταβλητών.
- Τα στοιχεία εκτός της διαγωνίου είναι οι συνδιακυμάνσεις μεταξύ όλων των τύπων των μεταβλητών, που όπως αναφέρθηκε, πρέπει να είναι μηδενικές.

Άρα ο πίνακας πρέπει να είναι διαγώνιος και μάλιστα οι διακυμάνσεις πρέπει να είναι ταξινομημένες κατά φθίνουσα σειρά, ώστε να είναι εύκολο να συμπεράνουμε ποιές μεταβλητές είναι σημαντικές. Για να γίνει αυτό, αρκεί να ταξινομηθούν οι γραμμές του X κατά φθίνουσα διακύμανση.

1.2 Υποθέσεις για την PCA

Έχουμε πλέον εξάγει τις υποθέσεις στις οποίες υπόκειται η PCA

1. *Γραμμικότητα*: το πρόβλημα ανάγεται στην εύρεση της κατάλληλης αλλαγής βάσης
2. *Οι μεγάλες διακυμάνσεις έχουν σημαντικές δομές*: Αυτή η υπόθεση περιλαμβάνει επίσης την πεποίθηση πως τα δεδομένα έχουν υψηλό $SNR \gg 1$)*. Έτσι, οι κύριες συνιστώσες με της μεγαλύτερες σχετικές διακυμάνσεις αντιπροσωπεύουν ενδιαφέρουσες δομές, ενώ αυτοί με μικρότερες

αντιπροσωπεύουν θόρυβο. Αξίζει να σημειωθεί ότι αυτή είναι μία ισχυρή και αρκετές φορές λανθασμένη υπόθεση.

Και μία επιπλέον,

3. *Οι κύριες συνιστώσες είναι ορθογώνιοι:* Αυτή η υπόθεση δίνει μία διαπισθητική απλοποίηση που κάνει την PCA επιλύσιμη με τεχνικές αποσύνθεσης από τη Γραμμική Άλγεβρα.

1.3 Η λύση της PCA- Γιατί δουλεύει η PCA

Έστω ότι τα δεδομένα είναι στη μορφή ενός $m \times n$ πίνακα X , όπου m το πλήθος των μεταβλητών και n ο αριθμός των δειγμάτων. Ψάχνουμε έναν ορθοκανονικό πίνακα P με $Y = PX$ τέτοιον ώστε ο C_Y να είναι διαγώνιος. Οι γραμμές του P είναι οι κύριες συνιστώσες.

Δεδομένου ότι κάθε συμμετρικός πίνακας διαγωνοποιείται από έναν πίνακα με τα ορθοκανονικά ιδιοδιανύσματά του. Επιλέγουμε ως πίνακα P τον πίνακα του οποίου οι γραμμές είναι τα ιδιοδιανύσματα του C_X , και επειδή ο αντίστροφος ενός ορθογώνιου πίνακα είναι ο ανάστροφός του ($P^{-1} = P^T$),

$$\begin{aligned} C_Y &= \frac{1}{n} Y Y^T \\ &= \frac{1}{n} P X (P X^T) \\ &= \frac{1}{n} P X X^T P^T \\ &= \frac{1}{n} P C_X P^T \end{aligned} \tag{1.1}$$

Και επομένως δείξαμε γιατί η PCA δουλεύει.

1.4 Ο αλγόριθμος PCA

Ο παρακάτω αλγόριθμος δίνει την πρακτική εφαρμογή της PCA

Ψευδοκώδικας

Είσοδος: Οργανώνουμε τα δεδομένα σε ένα $m \times n$ πίνακα X .

Βήμα 1: Αφαίρεσε το μέσο κάθε μεταβλητής από όλες της μετρήσεις για τη συγκεκριμένη μεταβλητή.

Βήμα 2: Υπολόγισε τον $m \times m$ πίνακα συνδιασποράς $C_X = \frac{1}{N-1} X X^T$

Βήμα 3: Υπολόγισε τα ορθοκανονικά ιδιοδιανύσματα του C_X και ταξινόμησέ τα κατά φθίνουσα σειρά ιδιοτιμής.

Βήμα 4: Επίλεξε m ιδιοδιανύσματα που αντιστοιχούν στις m μεγαλύτερες ιδιοτιμές να αποτελούν τη νέα βάση.

Έξοδος: Προβολή του αρχικού συνόλου δεδομένων σε μικρότερη διάσταση

Κεφάλαιο 2

Πολυδιάστατη Κλιμάκωση (MDS: MultiDimensional Scaling)

Μια διαφορετική άποψη της αναγωγής μεγάλων δεδομένων αποτελεί η Multi-dimensional Scaling (MDS) (Jolliffe 2002). Η MDS είναι άλλη μία κλασσική προσέγγιση που αντιστοιχίζει τον αρχικό μεγάλων διαστάσεων χώρο σε έναν χώρο μικρότερων διαστάσεων προσπαθώντας να διατηρήσει τις αποστάσεις ανά ζεύγη δεδομένων. Αυτό σημαίνει πως η MDS αντιμετωπίζει το πρόβλημα της κατασκευής μιας δομής από σημεία του Ευκλείδειου χώρου χρησιμοποιώντας πληροφορίες σχετικά με τις αποστάσεις ανάμεσα στα πρότυπα. Παρόλο που η μαθηματική προσέγγιση της MDS διαφέρει πολύ από αυτήν της PCA, οι δύο τεχνικές είναι τελικά πολύ στενά συνδεδεμένες και η MDS παράγει μια γραμμική ενσωμάτωση (embedding).

2.1 Η ιδέα και τα μαθηματικά πίσω από την MDS

Ένας $t \times t$ πίνακας D καλείται πίνακας απόστασης ή πίνακας συνάφειας αν είναι συμμετρικός, $d_{ii} = 0$ και $d_{ij} \neq 0, i \neq j$. Δεδομένου ενός πίνακα απόστασης D , η MDS προσπαθεί να βρει t σημεία δεδομένων y_1, \dots, y_t σε d διαστάσεις, τέτοια ώστε, αν το d_{ij} υποδηλώνει την Ευκλείδεια απόσταση μεταξύ των y_i και y_j , τότε ο \hat{D} να είναι παρόμοιος με τον D . Συγκεκριμένα, θεωρούμε τη μετρική

ή κλασική MDS, η οποία ελαχιστοποιεί το

$$\min_Y \sum_{i=1}^t \sum_{j=1}^t (d_{ij}^{(X)} - d_{ij}^{(Y)})^2 \quad (2.1)$$

όπου $d_{ij}^{(X)} = \|x_i - x_j\|^2$ και $d_{ij}^{(Y)} = \|y_i - y_j\|^2$. Ο πίνακας απόστασης $D^{(X)}$ μπορεί να μετατραπεί σε έναν πυρήνα πίνακα εσωτερικών γινομένων $X^T X$ ως εξής:

$$X^T X = \frac{1}{2} H D^{(X)} H, \quad (2.2)$$

όπου $H = I - \frac{1}{t} e e^T$ και e είναι ένα διάνυσμα-στήλη με όλα τα στοιχεία του ίσα με 1. Τώρα η 2.1 μπορεί να συμπτυχθεί στην

$$\min_Y \sum_{i=1}^t \sum_{j=1}^t (x_i^T x_j - y_i^T y_j)^2 \quad (2.3)$$

Αποδεικνύεται ότι η λύση είναι $Y = \Lambda^{\frac{1}{2}} V^T$, όπου V είναι τα ιδιοδιανύσματα του $X^T X$ που αντιστοιχούν στις πρώτες d ιδιοτιμές, και Λ είναι οι πρώτες d ιδιοτιμές του $X^T X$. Προφανώς η λύση της MDS είναι ίδια με αυτήν της PCA και όσο αφορά στις Ευκλείδειες αποστάσεις, οι δύο μέθοδοι παράγουν τα ίδια αποτελέσματα.

2.2 Ο αλγόριθμος MDS

Ψευδοκώδικας

Βήμα 1: Κατασκεύασε τον πίνακα των τετραγώνων των αποστάσεων

Βήμα 2: Υπολόγισε τον

$$X^T X = \frac{1}{2} H D^{(X)} H$$

χρησιμοποιώντας τον πίνακα $H = I - \frac{1}{t} e e^T$, όπου t το πλήθος των δεδομένων σημείων.

Βήμα 3: Βρες τις m μεγαλύτερες θετικές ιδιοτιμές $\lambda_1, \dots, \lambda_m$ και τα m αντίστοιχα ιδιοδιανύσματα e_1, \dots, e_m .

Βήμα 4: Μία m -διάστατη χωρική διάταξη των t σημείων των δεδομένων, προκύπτει από τον πίνακα συντεταγμένων $Y = \Lambda_m^{\frac{1}{2}} V_m^T$, όπου $\Lambda_m^{\frac{1}{2}}$ ο διαγώνιος πίνακας των m πρώτων ιδιοτιμών και V_m ο πίνακας των αντίστοιχων m ιδιοδιανυσμάτων.

Η μη-γραμμική περίπτωση

Κεφάλαιο 3

Απεικόνιση Ισομετρικών Συνιστωσών (Isomap: ISOmetric feature MAPping)

Ο αλγόριθμος Isomap (Tenenbaum et al. 2000) βασίζεται στην MDS και πιο συγκεκριμένα αποτελεί μία μη γραμμική γενίκευσή της. Χρησιμοποιείται για την επεξεργασία μεγάλων όγκων δεδομένων μεγάλης κλίμακας όπως τα παγκόσμια κλιματικά πρότυπα, τα αστρικά φάσματα ή οι διανομές ανθρώπινων γονιδίων. Αντίθετα με τις PCA και MDS, η προσέγγιση του Isomap είναι ικανή να βρίσκει μη γραμμικούς βαθμούς ελευθερίας που διέπουν τις πολύπλοκες φυσικές παρατηρήσεις.

3.1 Η ιδέα του αλγόριθμου Isomap

Η βασική ιδέα είναι να εφαρμόσουμε την MDS, όχι στον πολλών διαστάσεων χώρο εισόδου (input space), αλλά στον γεωδαισιακό χώρο της μη γραμμικής πολλαπλότητας των δεδομένων (data manifold). Οι γεωδαισιακές αποστάσεις αναπαριστούν τα συντομότερα μονοπάτια (paths) κατά μήκος της καμπύλης της επιφάνειας της πολλαπλότητας, μετρημένων ωςάν η επιφάνεια να ήταν επίπεδη. Αυτό μπορεί να προσεγγιστεί με μία ακολουθία μικρών βημάτων ανάμεσα σε γειτονικά σημεία του δείγματος. Ο Isomap τότε εφαρμόζει την MDS στις γεω-

δαισιακές και όχι στις Ευκλείδειες αποστάσεις για να βρει έναν μικρής διάστασης μετασχηματισμό που να διατηρεί τις αποστάσεις αυτές.

3.2 Ο αλγόριθμος Isomap

Ο αλγόριθμος Isomap δέχεται ως παραμέτρους τις αποστάσεις $d_X(i, j)$ μεταξύ όλων των ζευγαριών i, j από τα N δεδομένα σημεία στον υψηλών διαστάσεων χώρο εισόδου X , μετρημένων είτε στην κλασική Ευκλείδεια μετρική, είτε σε κάποια (domain-specific) μετρική. Ο αλγόριθμος εξάγει διανύσματα συντεταγμένων y_i σε έναν d -διάστατο Ευκλείδειο χώρο Y που αναπαριστούν καλύτερα την εγγενή γεωμετρία των δεδομένων και δουλεύει σε τρία βήματα.

Βήμα 1: Καθορίζει ποιά σημεία είναι γειτονικά με βάση τις αποστάσεις $d_X(i, j)$ των ζευγαριών σημείων i, j του χώρου των δεδομένων X με έναν από τους παρακάτω τρόπους:

- Συνδέουμε κάθε σημείο με όλα τα σημεία εντός μίας καθορισμένης ακτίνας ε (ε -Isomap), ή
- Συνδέουμε κάθε σημείο i με τους k πλησιέστερους γείτονές του, j (k -Isomap)

Αυτές οι σχέσεις γειτνίασης αναπαριστώνται σε έναν σταθμισμένο γράφο Γ των δεδομένων σημείων με κορυφές βάρους (αρχικοποίηση)
 $d_X(i, j)$ αν τα i, j συνδέονται με μία ακμή
 $d_G(i, j) = \infty$, αλλιώς

Βήμα 2: Εκτιμούμε τις γεωδαισιακές αποστάσεις $d_M(i, j)$ μεταξύ όλων των ζευγαριών σημείων M , υπολογίζοντας τα συντομότερα μονοπάτια στον γράφο G (Floyd's algorithm). Στην πράξη, για κάθε μία από τις τιμές $k = 1, \dots, N$ με τη σειρά, αντικαθιστούμε τις καταχωρήσεις $d_G(i, j)$, με το

$$\min\{d_G(i, j), d_G(i, k) + d_G(k, j)\}. \quad (3.1)$$

Ο πίνακας των τελικών τιμών $D_G = \{d_G(i, j)\}$ θα περιέχει τα συντομότερα μονοπάτια μεταξύ όλων των ζευγαριών i, j των σημείων του G .

Βήμα 3: Εφαρμόζουμε κλασική MDS στον πίνακα των αποστάσεων του γραφήματος $D_G = \{d_G(i, j)\}$, κατασκευάζοντας μία εμφύτευση των δεδομένων σε έναν d -διάστατο Ευκλείδειο χώρο Y , που διατηρεί καλύτερα την εκτιμώμενη εγγενή γεωμετρία του δείγματος. Τα διανύσματα συντεταγμένων y_i επιλέγονται ώστε να ελαχιστοποιείται η συνάρτηση κόστους

$$E = \left\| \left(\tau(D_G) - \tau(D_Y) \right) \right\|_{L^2}, \quad (3.2)$$

όπου

- D_Y : ο πίνακας ευκλειδίων αποστάσεων $\{d_y(i, j) = \|y_i - y_j\|\}$
- $\|A\|_{L^2}$: η L^2 νόρμα πίνακα $\sqrt{\sum_{i,j} A_{ij}^2}$
- τ : τελεστής που μετατρέπει αποστάσεις σε εσωτερικά γινόμενα που χαρακτηρίζουν κατά μοναδικό τρόπο τη γεωμετρία των δεδομένων, σε μορφή που υποστηρίζει την αποτελεσματική βελτιστοποίηση.

Το ολικό ελάχιστο της 3.2 επιτυγχάνεται ορίζοντας τις συντεταγμένες y_i στα d πρώτα ιδιοδιανύσματα του πίνακα $\tau(D_G)$. Δηλαδή, αν λ_p η p -οστή ιδιοτιμή σε φθίνουσα σειρά του πίνακα $\tau(D_G)$ και ν_p^i η i -οστή συνιστώσα του p -οστού ιδιοδιανύσματος, ο αλγόριθμος θέτει την p -οστή συνιστώσα του d -διάστατου διανύσματος συντεταγμένων y_i ίση με $\sqrt{\lambda_p} \nu_p^i$.

Ψευδοκώδικας

Είσοδος: Σημεία δεδομένων $x_1, \dots, x_n \in \mathbb{R}^m$.

Βήμα 1: Κατασκεύασε τον γράφο γειννίας $G = (V, E)$.

Βήμα 2: Υπολόγισε $\{d_y(i, j) = \|y_i - y_j\|\}$ για $(i, j) \in E$.

Βήμα 3: Υπολόγισε $\tilde{d}_{ij}, (i, j) \notin G$, από το συντομότερο μονοπάτι $\tilde{D}_{ij} = \tilde{d}_{ij}^2$.

Βήμα 4: Εφάρμοσε κλασική MDS.

Έξοδος: Μικρής διάστασης συντεταγμένες $y_1, \dots, y_n \in \mathbb{R}^d$.

Κεφάλαιο 4

Τοπικά Γραμμική Εμφύτευση (LLE: Locally Linear Embedding)

Ο αλγόριθμος LLE (Roweis & Saul 2000) βασίζεται στην ίδια ιδέα με τον Isomap, δηλαδή στον υπολογισμό χαμηλών διαστάσεων εμφυτεύσεων που να διατηρούν τις γειτονιές των σημείων των αρχικών μεγάλων διαστάσεων χώρων. Η διαφορά του LLE είναι πως δεν προσπαθεί να διατηρήσει τις αποστάσεις ανά ζεύγη δεδομένων, αλλά ανακτά την μη γραμμική δομή της πολλαπλότητας από τοπικά συνεχείς περιοχές της. Με αυτόν τον τρόπο δεν είναι πια αναγκαίος ο υπολογισμός αποστάσεων ανά ζεύγη πολύ απομακρυσμένων δεδομένων, γεγονός που καθιστά τον LLE υπολογιστικά «φθηνότερο» του Isomap.

Για την καλύτερη κατανόηση της ερμηνείας του LLE αλγόριθμου, παρουσιάζουμε πρώτα τα βήματά του.

4.1 Ο αλγόριθμος LLE

Βήμα 1: Καθορίζει τους γείτονες κάθε σημείου X_i , για παράδειγμα χρησιμοποιώντας τους k -πλησιέστερους γείτονες.

Βήμα 2: Υπολογίζει τα βάρη W_{ij} που ανακατασκευάζουν γραμμικά με τον καλύτερο δυνατό τρόπο το X_i , από τους γείτονές του. Ελαχιστοποιούμε την

συνάρτηση κόστους

$$\varepsilon(W) = \sum_i |\vec{X}_i - \sum_j W_{ij} \vec{X}_j|^2 \quad (4.1)$$

υπό τους ακόλουθους περιορισμούς:

- Κάθε σημείο X_i ανακατασκευάζεται μόνο από τους γείτονές του, οπότε αναγκάζουμε το $W_{ij} = 0$ αν το σημείο X_j δεν ανήκει στο σύνολο των γειτόνων του X_i .
- Το άθροισμα των βαρών κάθε γραμμής του πίνακα των βαρών είναι ίσο με 1, δηλ. $\sum_j W_{ij} = 1$.
Έτσι, τα βέλτιστα βάρη βρίσκονται με την επίλυση ενός προβλήματος ελαχίστων τετραγώνων.

Βήμα 3: Υπολογίζει τα μικρών διαστάσεων διανύσματα \vec{Y}_i που αντιπροσωπεύουν την ολική εγγενή γεωμετρία της πολλαπλότητας.

Αυτό γίνεται επιλέγοντας τις d -διάστατες συντεταγμένες των \vec{Y}_i (που ανακατασκευάζονται από τα βάρη W_{ij} ώστε να ελαχιστοποιούν τη συνάρτηση κόστους

$$\Phi(Y) = \sum_i |\vec{Y}_i - \sum_j W_{ij} \vec{Y}_j|^2. \quad (4.2)$$

Και αυτή η συνάρτηση κόστους, όπως η προηγούμενη βασίζεται σε τοπικά γραμμικά σφάλματα ανακατασκευής, με τη διαφορά ότι εδώ σταθεροποιούμε τα βάρη (όπως υπολογίστηκαν παραπάνω) και βελτιστοποιούμε τα $\text{Vec}Y_i$. Το κόστος (4.2) ορίζει μία τετραγωνική μορφή στα διανύσματα $\text{Vec}Y_i$. Υπό περιορισμούς που κάνουν το πρόβλημα καλά ορισμένο, μπορεί να ελαχιστοποιηθεί λύνοντας ένα αραιό $N \times N$ πρόβλημα ιδιοτιμών, του οποίου τα τελευταία d μη μηδενικά ιδιοδιανύσματα παρέχουν ένα διατεταγμένο σύνολο ορθογώνιων συντεταγμένων με κέντρο την αρχή των αξόνων.

Ψευδοκώδικας

Είσοδος: Σημεία δεδομένων $x_1, \dots, x_n \in \mathbb{R}^m$

Βήμα 1: Εφάρμοσε τον αλγόριθμο k-μέσων και βρες τους γείτονες του \vec{X}_i για $i = 1, \dots, n$.

Βήμα 2: Υπολόγισε τα βάρη W_{ij} που ανακατασκευάζουν καλύτερα κάθε \vec{X}_i από τους γείτονές του, λύνοντας το πρόβλημα ελαχίστων τετραγώνων με περιορισμούς (4.1).

Βήμα 3: Υπολόγισε τα χαμηλής διάστασης εμφυτευμένα διανύσματα \vec{Y}_i που ανακατασκευάζονται από τα W_{ij} ελαχιστοποιώντας τη (4.2), λύνοντας ένα αραιό $N \times N$ πρόβλημα ιδιοτιμών, του οποίου τα τελευταία d μη μηδενικά ιδιοδιανύσματα παρέχουν ένα διατεταγμένο σύνολο ορθογώνιων συντεταγμένων με κέντρο την αρχή των αξόνων.

Έξοδος: Μικρής διάστασης συντεταγμένες $y_1, \dots, y_n \in \mathbb{R}^d$.

4.2 Η βασική ιδέα του LLE

Θεωρώντας πως έχουμε επαρκή δεδομένα που αποτελούνται από N διανύσματα \vec{X}_i με πραγματικές τιμές, διαστατικότητας (dimensionality) D το καθένα, που έχουν εξαχθεί από μία υποκείμενη πολλαπλότητα, περιμένουμε κάθε σημείο δεδομένων και τα γειτονικά του να βρίσκονται πάνω ή κοντά σε μία τοπικά γραμμική περιοχή της πολλαπλότητας. Χαρακτηρίζουμε την τοπική γεωμετρία των περιοχών αυτών με γραμμικούς συντελεστές οι οποίοι ανακατασκευάζουν κάθε σημείο από τα γειτονικά του. Τα σφάλματα ανακατασκευής μετρώνται από τη συνάρτηση κόστους

$$\varepsilon(W) = \sum_i \left| \vec{X}_i - \sum_j W_{ij} \vec{X}_j \right|^2, \quad (4.3)$$

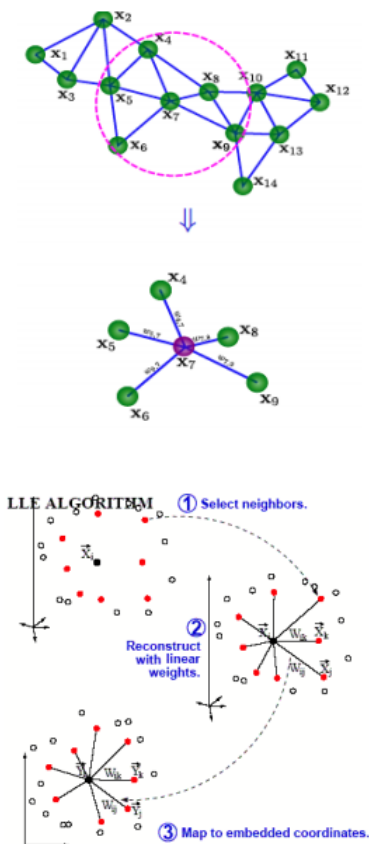
που αθροίζει τα τετράγωνα των αποστάσεων των σημείων των δεδομένων από τις αντίστοιχες ανακατασκευές τους. Τα βάρη W_{ij} συνοψίζουν τη συμβολή του j -οστού σημείου δεδομένων στην i -οστή ανακατασκευή.

Τα βάρη υπό τους περιορισμούς που αναφέρθηκαν, που ελαχιστοποιούν τα σφάλματα ανακατασκευής, παρουσιάζουν μία σημαντική συμμετρία: για κάθε ξεχωριστό σημείο των δεδομένων, είναι αναλλοίωτα σε περιστροφές, αλλαγές κλίμακας

και μεταφράσεις (translations) του συγκεκριμένου σημείου και των γειτόνων του. Λόγω συμμετρίας προκύπτει ότι τα βάρη ανακατασκευής χαρακτηρίζουν τις εγγενείς γεωμετρικές ιδιότητες κάθε γειτονιάς.

Ας υποθέσουμε ότι τα δεδομένα μίας περιοχής που βρίσκονται πάνω ή κοντά σε μία ομαλή γραμμική πολλαπλότητα χαμηλότερων διαστάσεων $d \ll D$. Τότε, σε μια καλή προσέγγιση, υπάρχει μια γραμμική απεικόνιση, που αποτελείται από μια μετάφραση, περιστροφή και αλλαγή κλίμακας που αποδίδει τις μεγάλων διαστάσεων συντεταγμένες της κάθε γειτονιάς στις πραγματικές συντεταγμένες της πολλαπλότητας. Λόγω του σχεδιασμού τους, τα βάρη αντανακλούν τις γεωμετρικές ιδιότητες των στοιχείων που είναι αναλλοίωτες σε τέτοιους μετασχηματισμούς. Βάσει αυτού περιμένουμε ότι ο χαρακτηρισμός της τοπικής γεωμετρίας του αρχικού χώρου των δεδομένων θα είναι εξίσου έγκυρος για τοπικές περιοχές της πολλαπλότητας. Ειδικότερα, τα ίδια βάρη W_{ij} που ανακατασκευάζουν το i -οστό σημείο δεδομένων στις D θα πρέπει επίσης να ανακατασκευάζει τις εμφυτευμένες συντεταγμένες της πολλαπλότητας στις d διαστάσεις. Σε αυτήν ακριβώς την ιδέα βασίζεται ο LLE.

Στην παρακάτω εικόνα (εικ. 4.1) δίνεται μία γραφική αναπαράσταση της διαδικασίας ανάλυσης με τη χρήση του αλγορίθμου LLE.



Σχήμα 4.1: Βήματα τοπικά γραμμικής εμφύτευσης: (1) Εκχώρηση γειτόνων σε κάθε δεδομένα σημείο \vec{X}_i (για παράδειγμα με χρήση των k πλησιέστερων γειτόνων). Υπολογισμός των βαρών W_{ij} που ανακατασκευάζουν γραμμικά τα \vec{X}_i από τους γείτονές τους, με επίλυση του υπό περιορισμούς προβλήματος ελαχίστων τετραγώνων της εξίσωσης 4.1. (3) Υπολογισμός των χαμηλών διαστάσεων διανυσμάτων εμφύτευσης \vec{Y}_i που ανακατασκευάζονται καλύτερα από τα βάρη W_{ij} ελαχιστοποιώντας την 4.2 με την εύρεση των μικρότερων ιδιοτιμών του αραιού συμμετρικού πίνακα της 4.1. Παρά το γεγονός ότι τα βάρη W_{ij} και τα διανύσματα \vec{Y}_i υπολογίζονται με μεθόδους γραμμικής άλγεβρας, ο περιορισμός ότι τα σημεία ανακατασκευάζονται μόνο από τους γείτονες μπορεί να οδηγήσει σε μη γραμμικές εμφυτεύσεις. (Roweis & Saul 2000)

Κεφάλαιο 5

Φασματική Ομαδοποίηση (Spectral Clustering)

5.1 Εισαγωγή

Οι αλγόριθμοι ομαδοποίησης (clustering) αποτελούν έναν γρήγορα αναπτυσσόμενο κλάδο υπολογιστικών μεθόδων στον τομέα της διερευνητικής ανάλυσης δεδομένων (exploratory data analysis) που τα τελευταία χρόνια έχουν συγκεντρώσει το ενδιαφέρον πολλών επιστημόνων στον τομέα της αναγωγής μεγάλων όγκων δεδομένων. Η ομαδοποίηση δεδομένων είναι ένα σημαντικό πρόβλημα με πολλές εφαρμογές σε τομείς όπως οι μηχανική μάθηση (machine learning), μηχανική όραση (computer vision), επιστήμη υπολογιστών (computer science) και ανάλυση σήματος (signal processing), βρίσκει όμως εφαρμογές και σε ευρύτερους τομείς όπως η Βιολογία, οι κοινωνικές επιστήμες και η Ψυχολογία.

Ο σκοπός της είναι να χωρίζει ένα σύνολο δεδομένων σε ομάδες που προκύπτουν με φυσικό τρόπο από τα ίδια τα δεδομένα ώστε τα σημεία που ανήκουν στην ίδια ομάδα (cluster) να είναι παρόμοια, ενώ αυτά που ανήκουν σε διαφορετικές ομάδες να μην παρουσιάζουν ομοιότητες.

Έχοντας ένα σύνολο δεδομένων σημείων x_1, \dots, x_n και κάποιον δείκτη ομοιότητας s_{ij} , ένας ωραίος τρόπος να εκφράσουμε τα δεδομένα είναι ο γράφος ομοιότητας $G = (V, E)$. Όπου

- κάθε κορυφή v_i αναπαριστά ένα δεδομένο σημείο x_i

- δύο κορυφές ενώνονται αν η ομοιότητα s_{ij} μεταξύ των x_i αντίστοιχων σημείων x_i και x_j είναι θετική ή μεγαλύτερη από ένα συγκεκριμένο όριο και η ακμή σταθμίζεται από το s_{ij} .

Έτσι το πρόβλημα ανάγεται στην εύρεση μιας διαμέρισης του γραφήματος, τέτοιας ώστε οι ακμές μέσα σε μία ομάδα να έχουν υψηλά βάρη (δηλαδή τα σημεία του ίδιου cluster είναι όμοια μεταξύ τους) ενώ οι ακμές σε διαφορετικές ομάδες να έχουν πολύ χαμηλά βάρη (Von Luxburg 2007).

5.2 Βασικά μαθηματικά εργαλεία των αλγορίθμων φασματικής ομαδοποίησης

5.2.1 Περί γράφων

Έστω $G = (V, E)$ ένας μη προσανατολισμένος γράφος με σύνολο κορυφών $V = \{v_1, \dots, v_n\}$. Παρακάτω υποθέτουμε ότι ο γράφος είναι σταθμισμένος, δηλ. κάθε ακμή μεταξύ δύο κορυφών έχει ένα μη αρνητικό βάρος $w_{ij} \geq 0$.

- Ο σταθμισμένος πίνακας γειτνίασης του γράφου είναι ο πίνακας $W = (w_{ij})_{i,j=1,\dots,n}$.
- Αν $w_{ij} = 0$, οι κορυφές v_i, v_j , δε συνδέονται με ακμή.
- Αφού ο G είναι μη προσανατολισμένος, απαιτούμε $w_{ij} = w_{ji}$.

Ο βαθμός κάθε κορυφής ορίζεται ως $d_i = \sum_{j=1}^n w_{ij}$.

Ο πίνακας βαθμού D ορίζεται ως ο διαγώνιος πίνακας με διαγώνια στοιχεία τους βαθμούς των κορυφών d_i, d_j . Ορίζουμε το διάνυσμα-δείκτη $\mathbf{1}_A = (f_1, \dots, f_n)' \in \mathbb{R}^n$, με

$$f_i = \begin{cases} 1 & , \text{αν } v_i \in A \\ 0 & , \text{αλλιως.} \end{cases}$$

Για δύο, όχι απαραίτητα ξένα σύνολα, $A, B \subset V$, ορίζουμε $W(A, B) := \sum_{i \in A, j \in B} w_{ij}$. Ένα υποσύνολο $A \subset V$ είναι συνδεδεμένο αν οποιεσδήποτε δύο

κορυφές στο A μπορούν να ενωθούν με ένα μονοπάτι, τέτοιο ώστε όλα τα ενδιάμεσα σημεία να βρίσκονται επίσης πάνω στο A .

Ένα υποσύνολο $A \subset V$ είναι *connected component* αν είναι συνδεδεμένο και δεν υπάρχουν ενώσεις μεταξύ κορυφών στα A και \bar{A} .

Τα μη κενά σύνολα A_1, \dots, A_k αποτελούν μία διαμέριση του γράφου αν $A_i \cap A_j = \emptyset$ και $A_1 \cup \dots \cup A_k = V$.

Διαφορετικοί γράφοι ομοιότητας

Στόχος των γράφων ομοιότητας είναι να μοντελοποιήσουν τις σχέσεις τοπικών γειτονιών των δεδομένων, χρησιμοποιώντας είτε τις ομοιότητες s_{ij} είτε τις αποστάσεις d_{ij} ανά ζεύγη δεδομένων σημείων.

- i Ο γράφος ϵ -γειτονιάς (ϵ -neighborhood graph): συνδέουμε όλα τα σημεία των οποίων οι αποστάσεις ανά ζεύγη είναι μικρότερες από ϵ . Αφού οι αποστάσεις μεταξύ όλων των συνδεδεμένων σημείων είναι περίπου ίδιας κλίμακας (το πολύ ϵ), το να σταθμίσουμε τον γράφο δε δίνει περαιτέρω πληροφορίες για τα δεδομένα. Γι'αυτό ο γράφος αυτός θεωρείται μη σταθμισμένος.
- ii Ο γράφος των k -πλησιέστερων γειτόνων (k -neighborhood graph): Εδώ συνδέουμε την κορυφή v_i με την v_j αν η δεύτερη είναι μεταξύ των k πλησιέστερων γειτόνων της v_i . Αυτός ο ορισμός οδηγεί σε προσανατολισμένο γράφο, αφού η σχέση γειτνίασης δεν είναι συμμετρική. Υπάρχουν δύο τρόποι να τον μετατρέψουμε σε μη προσανατολισμένο:
 - Απλώς αγνοούμε τις κατευθύνσεις των ακμών και συνδέουμε τα v_i και v_j με μία μη προσανατολισμένη ακμή, αν το v_i είναι μεταξύ των k πλησιέστερων γειτόνων του v_j ή αν το v_j είναι μεταξύ των k πλησιέστερων γειτόνων του v_i (*k-nearest neighbor graph*).
 - Συνδέουμε τα v_i και v_j αν και το v_i είναι μεταξύ των k πλησιέστερων γειτόνων του v_j **και** το v_j είναι μεταξύ των k πλησιέστερων γειτόνων του v_i (*mutual k-nearest neighbor graph*).

Σταθμίζουμε τις ακμές βάσει της ομοιότητας των καταληκτικών σημείων.

iii Πλήρως συνδεδεμένος γράφος (Fully connected graph): απλώς ενώνουμε όλα τα σημεία με θετική ομοιότητα μεταξύ τους και σταθμίζουμε όλες τις ακμές με το s_{ij} . Αυτή η κατασκευή είναι χρήσιμη μόνο αν η ίδια η συνάρτηση ομοιότητας μοντελοποιεί τοπικές γειτονίες (π.χ. η γκαουσιανή συνάρτηση ομοιότητας $\exp(-\|x_i - x_j\|^2 / (2\sigma^2))$).

Οι παραπάνω γράφοι ομοιότητας χρησιμοποιούνται συχνά σε αλγόριθμους φασματικής ομαδοποίησης χωρίς ωστόσο να υπάρχει θεωρητική γνώση για το πώς η επιλογή του εκάστοτε γραφήματος επηρεάζει το αποτέλεσμα.

5.2.2 Γράφος Λαπλασιανού πίνακα

Το βασικό εργαλείο των αλγορίθμων spectral clustering είναι οι γράφοι των Λαπλασιανών πινάκων. Παρακάτω, θεωρούμε πάντα ότι το G είναι ένας μη προσανατολισμένος, σταθμισμένος γράφος, με πίνακα βάρους W , όπου $w_{ij} = w_{ji} \geq 0$. Όταν χρησιμοποιούμε τα ιδιοδιανύσματα ενός πίνακα, δε θεωρούμε ότι αυτά είναι απαραίτητα κανονικοποιημένα (π.χ. το σταθερό διάνυσμα $\mathbf{1}$ και το πολλαπλάσιό του $a\mathbf{1}$ $a \neq 0$ δε θα θεωρούνται ως τα ίδια ιδιοδιανύσματα). Οι ιδιοτιμές θα διατάσσονται πάντα κατά αύξουσα σειρά. Με τον όρο « k πρώτα ιδιοδιανύσματα» θα αναφερόμαστε στα ιδιοδιανύσματα που αντιστοιχούν στις k μικρότερες ιδιοτιμές.

1. Μη κανονικοποιημένος γράφος Λαπλασιανού πίνακα (un-normalised graph Laplacian)

Πίνακας γράφου: $L = D - W$, όπου D ο πίνακας με διαγώνια στοιχεία τα αντίστοιχα d_i .

Πρόταση 1: Ιδιότητες του πίνακα L

- $\forall f \in \mathbb{R}^n, f^T L f = \frac{1}{2} \sum_{i,j=1}^n w_{ij} (f_i - f_j)^2$, (αποδ.: Άμεσα από ορισμό των d_i)
- Ο L $x^T L x \geq 0$ είναι συμμετρικός και θετικά ημι-ορισμένος (≥ 0) (αποδ.: άμεσα από προηγούμενη).
- Η μικρότερη ιδιοτιμή του L είναι 0. Το αντίστοιχο ιδιοδιάνυσμα, το σταθερό διάνυσμα $\mathbf{1}$. (αποδ.: προφανής)

- Ο L έχει n μη αρνητικές, πραγματικές ιδιοτιμές $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$. (αποδ.: από τα προηγούμενα)

Πρόταση 2: Έστω G ένας μη προσανατολισμένος γράφος με μη αρνητικά βάρη. Τότε η πολλαπλότητα k της ιδιοτιμής 0 του L ισούται με τον αριθμό των κύριων συνιστωσών A_1, \dots, A_k στον γράφο. Ο ιδιοχώρος της ιδιοτιμής 0 , παράγεται από τα διανύσματα δείκτες $\mathbf{1}_{A_1}, \dots, \mathbf{1}_{A_k}$.

Απόδειξη

Όταν $k = 1$, ένας γράφος αποτελείται μόνο από έναν *connected component* έτσι έχουμε μόνο το διάνυσμα $\mathbf{1}$ σαν ιδιοδιάνυσμα με ιδιοτιμή 0 , που προφανώς είναι το διάνυσμα-δείκτης του *connected component*.

Όταν $k > 1$, ο L μπορεί να γραφεί σαν *block* διαγώνιος πίνακας. Το φάσμα του L δίνεται από την ένωση των φασμάτων L_i , και τα αντίστοιχα ιδιοδιανύσματα του L είναι τα ιδιοδιανύσματα των L_i όπου έχουμε γεμίσει με 0 τις θέσεις των άλλων *blocks*.

2. Κανονικοποιημένος γράφος Λαπλασιανού πίνακα

Στη βιβλιογραφία υπάρχουν δύο πίνακες που καλούνται έτσι, είναι στενά συνδεδεμένοι και ορίζονται ως:

$$L_{sym} := D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$$

$$L_{rw} := D^{-1} L = I - D^{-1} W$$

Πρόταση 3: (Ιδιότητες των L_{sym} και L_{rw})

- $\forall f \in \mathbb{R}^n, f' L f = \frac{1}{2} \sum_{i,j=1}^n w_{ij} \left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right)^2$ (αποδ.: Άμεσα από ορισμό των d_i)
- Το λ είναι ιδιοτιμή του L_{rw} με ιδιοδιάνυσμα \vec{v} αν και μόνο αν το λ είναι ιδιοτιμή του L_{sym} με ιδιοδιάνυσμα $\vec{w} = D^{\frac{1}{2}} \vec{v}$.
- Το λ είναι ιδιοτιμή του L με ιδιοδιάνυσμα \vec{v} αν και μόνο αν τα λ και \vec{v} λύνουν το γενικευμένο πρόβλημα ιδιοτιμών $L \vec{v} = \lambda D \vec{v}$.
- Το 0 είναι ιδιοτιμή του L_{rw} με το σταθερό μοναδιαίο ιδιοδιάνυσμα $\mathbf{1}$. Το 0 είναι ιδιοτιμή του L_{sym} με ιδιοδιάνυσμα $D^{\frac{1}{2}} \mathbf{1}$.

- Οι L_{sym} και L_{rw} είναι θετικά ημιορισμένοι και έχουν n μη αρνητικές πραγματικές ιδιοτιμές $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$.

Πρόταση 4: Έστω G ένα μη προσανατολισμένος γράφος με μη αρνητικά βάρη. Τότε η πολλαπλότητα k της ιδιοτιμής 0 και του L_{sym} και του L_{rw} , ισούται με τον αριθμό των κύριων συνιστωσών A_1, \dots, A_k του γράφου. Για τον L_{rw} , ο ιδιοχώρος του 0 ορίζεται από τα διανύσματα-δείκτες $\mathbf{1}_{A_i}$ αυτών των components. Για τον L_{sym} ο ιδιοχώρος παράγεται από τα διανύσματα $D^{\frac{1}{2}} \mathbf{1}_{A_i}$

Απόδειξη: Ανάλογη πρότασης 2.

5.3 Αλγόριθμοι Φασματικής ομαδοποίησης

Θεωρούμε ότι τα δεδομένα μας αποτελούνται από n σημεία x_1, \dots, x_n . Μετράμε τις ανά ζεύγη ομοιότητες τους $s_{ij} = s(x_i, x_j)$, μέσω κάποιας συμμετρικής, μη αρνητικής συνάρτησης ομοιότητας και φτιάχνουμε τον αντίστοιχο πίνακα ομοιότητας.

I Μη κανονικοποιημένο Spectral clustering:

Ψευδοκώδικας

Είσοδος: Πίνακας ομοιότητας $S \in \Pi^{n \times n}$, αριθμός k των προς κατασκευή clusters.

- Κατασκεύασε έναν γράφο ομοιότητας. Έστω W ο πίνακας σταθμισμένης γειννίας του γραφήματος.
- Υπολόγισε τον μη κανονικοποιημένο Λαπλασιανό πίνακα L .
- Υπολόγισε τα πρώτα k ιδιοδιανύσματα $\vec{v}_1, \dots, \vec{v}_k$ του L .
- Έστω $U \in \Pi^{n \times k}$ ο πίνακας που περιέχει τα διανύσματα $\vec{v}_1, \dots, \vec{v}_k$ ως στήλες.
- Για $i = 1, \dots, n$, έστω $y_i \in \mathbb{R}^k$ το διάνυσμα που αντιστοιχεί στην i -οστή γραμμή του U .

- Εφάρμοσε τον αλγόριθμο k -μέσων για την ομαδοποίηση των σημείων $(y_i)_{i=1,\dots,n}$ στον \mathbb{R}^k στα clusters C_1, \dots, C_k .

Έξοδος: Clusters A_1, \dots, A_k όπου $A_i = \{j | y_j \in C_i\}$

II Κανονικοποιημένο Spectral Clustering (Shi & Malik 2000)

Ψευδοκώδικας

Είσοδος: Πίνακας ομοιότητας $S \in \Pi^{n \times n}$, αριθμός k των προς κατασκευή clusters.

- Κατασκεύασε έναν γράφο ομοιότητας. Έστω W ο πίνακας σταθμισμένης γειτνίασης του γραφήματος.
- Υπολόγισε τον μη κανονικοποιημένο Λαπλασιανό πίνακα L .
- Υπολόγισε τα πρώτα k γενικευμένα ιδιοδιανύσματα $\vec{v}_1, \dots, \vec{v}_k$ του γενικευμένου προβλήματος ιδιοτιμών (eigenproblem) $Lv = \lambda Dv$.
- Έστω $U \in \Pi^{n \times k}$ ο πίνακας που περιέχει τα διανύσματα $\vec{v}_1, \dots, \vec{v}_k$ ως στήλες.
- Για $i = 1, \dots, n$, έστω $y_i \in \mathbb{R}^k$ το διάνυσμα που αντιστοιχεί στην i -οστή γραμμή του U .
- Εφάρμοσε τον αλγόριθμο k -μέσων για την ομαδοποίηση των σημείων $(y_i)_{i=1,\dots,n}$ στον \mathbb{R}^k στα clusters C_1, \dots, C_k .

Έξοδος: Clusters A_1, \dots, A_k όπου $A_i = \{j | y_j \in C_i\}$.

III Κανονικοποιημένο Spectral Clustering (Ng et al. 2001)

Ψευδοκώδικας

Είσοδος: Πίνακας ομοιότητας $S \in \Pi^{n \times n}$, αριθμός k των προς κατασκευή clusters.

- Κατασκεύασε έναν γράφο ομοιότητας. Έστω W ο πίνακας σταθμισμένης γειτνίασης του γραφήματος.
- Υπολόγισε τον μη κανονικοποιημένο Λαπλασιανό πίνακα L_{sym} .
- Υπολόγισε τα πρώτα k γενικευμένα ιδιοδιανύσματα $\vec{v}_1, \dots, \vec{v}_k$ του L_{sym} .

- Έστω $U \in \Pi^{n \times k}$ ο πίνακας που περιέχει τα διανύσματα $\vec{v}_1, \dots, \vec{v}_k$ ως στήλες.
 - **Κατασκευάσε τον πίνακα $T \in \Pi^{n \times n}$ από τον U , με κανονικοποίηση των γραμμών, δηλ. θέσε $t_{ij} = u_{ij}(\sum_k u_{ik}^2)^{\frac{1}{2}}$.**
 - Για $i = 1, \dots, n$, έστω $y_i \in \mathbb{R}^k$ το διάνυσμα που αντιστοιχεί στην i -οστή γραμμή του T
 - Εφάρμοσε τον αλγόριθμο k -μέσων για την ομαδοποίηση των σημείων $(y_i)_{i=1, \dots, n}$ στον \mathbb{R}^k στα clusters C_1, \dots, C_k .
- Έξοδος:** Clusters A_1, \dots, A_k όπου $A_i = \{j | y_j \in C_i\}$.

5.4 Γιατί δουλεύουν οι αλγόριθμοι φασματικής ομαδοποίησης·

Παρακάτω εξετάζονται οι τρεις διαφορετικές ερμηνείες που έχουν δοθεί για την εξήγηση των αλγορίθμων φασματικής ομαδοποίησης.

5.4.1 ΤΟΜΕΣ ΓΡΑΦΗΜΑΤΟΣ

Η βασική ιδέα της ομαδοποίησης είναι να χωρίσουμε σε διαφορετικές ομάδες τα δεδομένα ανάλογα με τις ομοιότητές του. Αυτό το πρόβλημα, δεδομένων των ομοιοτήτων υπό τη μορφή γράφου ομοιότητας, μπορεί να αναχθεί στο εξής:

Βρες μία διαμέριση του γράφου ώστε τα σημεία της ίδια ομάδας να έχουν υψηλά βάρη και τα σημεία διαφορετικών ομάδων πολύ χαμηλά βάρη.

Αυτό μπορεί να γίνει επιλύοντας ένα εύκολο *mincut* πρόβλημα, δηλαδή (για δεδομένο αριθμό k υποσυνόλων) επιλέγοντας την κατάλληλη διαμέριση A_1, \dots, A_k που ελαχιστοποιεί την

$$cut(A_1, \dots, A_k) := \frac{1}{2} \sum_{i=1}^k W(A_i, \bar{A}_i)$$

. Για να αποφύγουμε το πρόβλημα τις απομόνωσης μεμονομένων κορυφών που προκύπτει συχνά, χρησιμοποιούμε κάποιες αντικειμενικές συναρτήσεις που

επιβάλλουν στα clusters να περιέχουν ισορροπημένα μεγάλο αριθμό κορυφών. Οι συνηθέστερες είναι οι παρακάτω:

- $RationCut(A_1, \dots, A_k) := \frac{1}{2} \sum_{i=1}^k \frac{W(A_i, \bar{A}_i)}{|A_i|} = \sum_{i=1}^k \frac{cut(A_i, \bar{A}_i)}{|A_i|},$

όπου $|A_i|$, ο αριθμός των κορυφών στο A_i .

- $RationCut(A_1, \dots, A_k) := \frac{1}{2} \sum_{i=1}^k \frac{W(A_i, \bar{A}_i)}{vol(A_i)} = \sum_{i=1}^k \frac{cut(A_i, \bar{A}_i)}{vol(A_i)},$

όπου $vol(A_i) = \sum_{j \in A_i} d_j$.

Αυτοί οι επιπλέον περιορισμοί όμως, κάνουν το πρόβλημα σε κλάση δυσκολίας μη ντετερμινιστικού πολυωνυμικού χρόνου (NP) Η φασματική ομαδοποίηση αποτελεί έναν τρόπο να λύνουμε χαλαρωμένες εκδοχές αυτών των προβλημάτων.

Προσεγγίζοντας το RatioCut

Για $k = 2$:

Θέλουμε να λύσουμε το πρόβλημα βελτιστοποίησης $\min_{A \subset V} RatioCut(A, \bar{A})$.

Δεδομένου ενός υποσυνόλου $A \subset V$, ορίζουμε το διάνυσμα

$f = (f_1, \dots, f_n)' \in \mathbb{R}^n$, όπου

$$f_i = \begin{cases} \sqrt{\frac{|A|}{|\bar{A}|}} & , \text{αν } v_i \in A \\ \sqrt{\frac{|\bar{A}|}{|A|}} & , \text{αν } v_i \in \bar{A} \end{cases}$$

Τώρα έχουμε:

$$\begin{aligned}
f'Lf &= \frac{1}{2} \sum_{i,j=1}^n w_{ij}(f_i - f_j) \\
&= \frac{1}{2} \sum_{i \in A, j \in \bar{A}} w_{ij} \left(\sqrt{\frac{|\bar{A}|}{|A|}} + \sqrt{\frac{|A|}{|\bar{A}|}} \right)^2 + \sum_{i \in \bar{A}, j \in A} w_{ij} \left(-\sqrt{\frac{|\bar{A}|}{|A|}} - \sqrt{\frac{|A|}{|\bar{A}|}} \right)^2 \\
&= \text{cut}(A, \bar{A}) \left(\frac{|\bar{A}|}{|A|} + \frac{|A|}{|\bar{A}|} + 2 \right) \\
&= \text{cut}(A, \bar{A}) \left(\frac{|A| + \bar{A}}{|A|} + \frac{|\bar{A}| + |A|}{|\bar{A}|} \right) \\
&= |V| \text{RatioCut}(A, \bar{A})
\end{aligned} \tag{5.1}$$

Επιπλέον,

$$\sum_{i=1}^n f_i = \sum_{i \in A} \sqrt{\frac{|\bar{A}|}{|A|}} - \sum_{i \in \bar{A}} \sqrt{\frac{|A|}{|\bar{A}|}} = |A| \sqrt{\frac{|\bar{A}|}{|A|}} + |\bar{A}| \sqrt{\frac{|A|}{|\bar{A}|}}, \text{ και}$$

$$\|f\|^2 = \sum_{i=1}^n f_i^2 = |A| \frac{|\bar{A}|}{|A|} + |\bar{A}| \frac{|A|}{|\bar{A}|} = |A| + |\bar{A}| = n.$$

Άρα το πρόβλημα τώρα γίνεται:

$$\min_{A \in V} f'Lf \text{ υπό τους περιορισμούς } f \perp \mathbf{1}, \|f\| = n, f \text{ όπως ορίστηκε παραπάνω.}$$

Μία προφανής χαλάρωση είναι να αγνοήσουμε τον περιορισμό διακριτότητας και τότε

$$\min_{A \in V} f'Lf \text{ υπό τους περιορισμούς } f \perp \mathbf{1}, \|f\| = n.$$

Τότε από το θεώρημα Rayleigh-Ritz προκύπτει ότι το f είναι το ιδιοδιάνυσμα που αντιστοιχεί στην δεύτερη μικρότερη ιδιοτιμή του L (Θυμηθείτε ότι η πρώτη μικρότερη ιδιοτιμή του L είναι το 0 που αντιστοιχεί το διάνυσμα-δείκτη $\mathbf{1}$). Χρειαζόμαστε ένα διακριτό διάνυσμα-δείκτη για να πάρουμε μία διαμέριση

του γράφου, οπότε θεωρούμε ως δείτρια-συνάρτηση την

$$f = \begin{cases} v_i \in A & , \text{αν } f_i \geq 0 \\ v_i \in \bar{A} & , \text{αν } f_i < 0. \end{cases}$$

Πιο γενική μορφή, η οποία δουλεύει για κάθε $k > 2$, είναι να θεωρηθούν οι συντεταγμένες f_i ως σημεία του \mathbb{R} και να ομαδοποιηθούν με τον αλγόριθμο k -μέσων στα clusters C, \bar{C} . Τότε επιλέγουμε

$$\begin{cases} v_i \in A & , \text{αν } f_i \geq 0 \\ v_i \in \bar{A} & , \text{αν } f_i < 0. \end{cases}$$

. Αυτός ακριβώς είναι ο μη κανονικοποιημένος αλγόριθμος spectral clustering για $k = 2$.

Για $k > 2$:

Δεδομένης μίας διαμέρισης του V σε k σύνολα A_1, A_2, \dots, A_k ,

- Ορίζουμε τα διανύσματα-δείκτες

$$h_{ij} = (h_{1,j}, \dots, h_{n,j})' := \begin{cases} \frac{1}{\sqrt{|A_j|}} & , \text{αν } v_i \in A_j \\ 0 & , \text{αλλιώς} \end{cases}$$

με $i = 1, \dots, n$ και $j = 1, \dots, k$.

- Θεωρούμε τον πίνακα $H \in \mathbb{R}^{n \times k}$ που περιέχει τα k διανύσματα-δείκτες ως στήλες (τα διανύσματα-δείκτες είναι ορθοκανονικά μεταξύ τους, δηλ. $H'H = I$) και με υπολογισμούς, όμοιους με πριν, έχουμε $h_i' L h_i = \frac{\text{cut}(A_i, \bar{A}_i)}{|A_i|}$.

Επιπλέον, εύκολα διαπιστώνεται ότι $h_i' L h_i = (H' L H)_{ii}$ και επομένως

$$\text{RationCut}(A_1, \dots, A_k) = \sum_{i=1}^k (H' L H)_{ii} = \text{Tr}(H' L H).$$

Το πρόβλημά μας ανάγεται στο

$$\min_{A_1, \dots, A_k} \text{Tr}(H' L H) \text{ υπό τους περιορισμούς } H' H = I, H \text{ όπως ορίστηκε.}$$

Το αντίστοιχο χαλαρωμένο πρόβλημα είναι το

$$\min_{A_1, \dots, A_k} \text{Tr}(H' L H) \text{ υπό τον περιορισμό } H' H = I.$$

Και πάλι (Θ. Rayleigh-Ritz) η λύση δίνεται επιλέγοντας τον H ως τον πίνακα με στήλες τα k πρώτα ιδιοδιανύσματα του L . Διακριτοποιούμε και πάλι τον πραγματικό πίνακα λύσης σε διαμέριση με τον αλγόριθμο k -μέσων. Σημειώνεται ότι ο πίνακας λύση είναι ο U που θεωρήσαμε στον μη κανονικοποιημένο αλγόριθμο (I).

Προσεγγίζοντας το Ncut

Για $k = 2$:

- Ορίζουμε το διάνυσμα-δείκτη f ως

$$h_{ij} = (h_{1,j}, \dots, h_{n,j})' := \begin{cases} \sqrt{\frac{\text{vol}(\bar{A})}{\text{vol}(A)}} & , \text{ αν } v_i \in A \\ -\sqrt{\frac{\text{vol}(A)}{\text{vol}(\bar{A})}} & , \text{ αν } v_i \in \bar{A} \end{cases}$$

όπου $i = 1, \dots, n$ και $j = 1, \dots, k$.

- Όπως και πριν, εύκολα βλέπουμε ότι $(Df)'1 = 0$, $f' Df = \text{vol}(V)$, $f' L f = \text{vol}(V) \text{Ncut}(A, \bar{A})$.

Ξαναγράφουμε ισοδύναμα το πρόβλημα ως $\min_A f' L f$ υπό τους περιορισμούς: f όπως ορίστηκε, $Df \perp 1$, $f' Df = \text{vol}(V)$.

- Χαλαρώνουμε το πρόβλημα, επιτρέποντας στο f να πάρει αυθαίρετες πραγματικές τιμές:

$\min_{f \in \mathbb{R}^n} f' L f$ υπό τους περιορισμούς: $Df \perp 1, f' D f = \text{vol}(V)$.

- Αντικαθιστούμε $g := D^{\frac{1}{2}} f$ και το πρόβλημα ανάγεται στο:
 $\min_{g \in \mathbb{R}^n} g' D^{-\frac{1}{2}} L D^{\frac{1}{2}} g$ υπό τους περιορισμούς $g \perp D^{\frac{1}{2}} 1, \|g\|^2 = \text{vol}(V)$.
(Σημειώνεται ότι $D^{-\frac{1}{2}} L D^{\frac{1}{2}} = L_{sym}, D^{\frac{1}{2}} 1$ είναι το πρώτο ιδιοδιάνυσμα του πίνακα L_{sym} , $\text{vol}(V)$ σταθερά.) Έτσι το πρόβλημα είναι στην κλασική μορφή του θεωρήματος Rayleigh-Ritz και η λύση g δίνεται από το δεύτερο ιδιοδιάνυσμα του L_{sym} .
- Αντικαθιστούμε εκ νέου $f := D^{\frac{1}{2}} g$ και βλέπουμε ότι το f είναι το δεύτερο ιδιοδιάνυσμα του L_{rw} (αντίστοιχα το γενικευμένο ιδιοδιάνυσμα του $Lv = \lambda Dv$).

Για $k > 2$:

- Ορίζουμε τα διανύσματα-δείκτες

$$h_{ij} = (h_{1,j}, \dots, h_{n,j})' := \begin{cases} \frac{1}{\sqrt{\text{vol}(A_j)}} & , \text{αν } v_i \in A_j \\ 0 & , \text{αλλιώς} \end{cases}$$

με $i = 1, \dots, n$ και $j = 1, \dots, k$.

- Θεωρούμε τον πίνακα $H \in \Pi^{n \times k}$ που περιέχει τα k διανύσματα-δείκτες ως στήλες. Παρατηρούμε ότι $H' H = I, h_i' D h_i = 1, h_i' L h_i = \frac{\text{cut}(A_i, \bar{A}_i)}{\text{vol}(A_i)}$ Το πρόβλημά μας ανάγεται στο

$\min_{A_1, \dots, A_k} \text{Tr}(H' L H)$ υπό τους περιορισμούς $H' D H = I, H$ όπως ορίστηκε.

Το αντίστοιχο πρόβλημα, χαλαρώνοντας τον περιορισμό διακριτότητας και αντικαθιστώντας $T = D^{\frac{1}{2}} H$, γίνεται

$$\min_{T \in \Pi^{n \times k}} \text{Tr}(T' D^{-\frac{1}{2}} L D^{-\frac{1}{2}} T)$$
 υπό τον περιορισμό $T' T = I$.

Και πάλι (Θ. Rayleigh-Ritz) η λύση δίνεται επιλέγοντας τον T ως τον

πίνακα με στήλες τα k πρώτα ιδιοδιανύσματα του L_{sym} . Ξαναντικαθιστούμε $H = D^{\frac{1}{2}}T$ και τότε η λύση αποτελείται από τα k πρώτα ιδιοδιανύσματα του πίνακα L_{rw} (αντίστοιχα τα k πρώτα γενικευμένα ιδιοδιανύσματα του $Lv = \lambda Dv$).

5.4.2 ΤΥΧΑΙΟΣ ΠΕΡΙΠΑΤΟΣ

Ένας τυχαίος περίπατος σε έναν γράφο είναι μία στοχαστική διαδικασία όπου τυχαία πηδάμε από κορυφή σε κορυφή. Το spectral clustering μπορεί να ερμηνευθεί ως προσπάθεια να βρούμε μια διαμέριση του γράφου τέτοια ώστε ο τυχαίος περίπατος να παραμένει επί μακρόν στο ίδιο cluster και σπάνια να μεταπηδά σε άλλα clusters. Η πιθανότητα μεταπήδησης p_{ij} από τη v_i στη v_j κορυφή σε ένα βήμα είναι ανάλογη με το βάρος w_{ij} της ακμής, $p_{ij} = w_{ij}/d_i$. Έτσι ο πίνακας μετάβασης του τυχαίου περιπάτου $P = (p_{ij})_{i,j=1,\dots,n}$ ορίζεται ως $P = D^{-1}W$. Αν ο γράφος είναι συνδεδεμένος και μη διχοτομήσιμος, ο τυχαίος περίπατος ακολουθεί πάντα μια στάσιμη κατανομή (π_1, \dots, π_n) , όπου $\pi_i = d_i/vol(V)$. Η σχέση μεταξύ L_{rw} και P είναι στενή αφού $L_{rw} = I - P$ και κατά συνέπεια, το λ είναι ιδιοτιμή του L_{rw} με ιδιοδιάνυσμα v αν και μόνο αν το $1 - \lambda$ είναι ιδιοτιμή του P με αντίστοιχο ιδιοδιάνυσμα v .

Τυχαίος περίπατος και Ncut

1η εκδοχή

Πρόταση: Έστω G συνδεδεμένος, μη διχοτομήσιμος γράφος. Υποθέτουμε ότι ξεκινάμε τον τυχαίο περίπατο $(X_t)_{t \in \mathbb{R}}$ ξεκινώντας από το X_0 στη σταθερή κατανομή π . Για τα ξένα υποσύνολα $A, B \subset V$, ορίζουμε $P(A|B) := P(X_1 \in B | X_0 \in A)$. Τότε: $Ncut(A, \bar{A}) = P(\bar{A}|A) + P(A|\bar{A})$.

Απόδειξη: Παρατηρούμε ότι

$$\begin{aligned} P(X_0 \in A, X_1 \in B) &= \sum_{i \in A, j \in B} P(X_0 = i, X_1 = j) = \\ &= \sum_{i \in A, j \in B} \pi p_{ij} \end{aligned}$$

$$= \sum_{i \in A, j \in B} \frac{d_i}{\text{vol}(V)} \frac{w_{ij}}{d_i} = \frac{1}{\text{vol}(V)} \sum_{i \in A, j \in B} w_{ij}$$

Και από τον ορισμό του $Ncut$

$$\begin{aligned} Ncut(A, \bar{A}) &= \frac{1}{2} \sum \frac{W(A, \bar{A})}{\text{vol}(A)} = \sum \frac{cut(A, \bar{A})}{\text{vol}(A)} \\ &= \frac{\sum_{i \in A, j \in \bar{A}} w_{ij}}{\text{vol}(A)} + \frac{\sum_{i \in \bar{A}, j \in A} w_{ij}}{\text{vol}(A)} \\ &= P(\bar{A}|A) + P(A|\bar{A}) \end{aligned} \quad (5.2)$$

Και επομένως, όταν ελαχιστοποιούμε το $Ncut$, στην πραγματικότητα ψάχνουμε μία τομή στον γράφο ώστε ο τυχαίος περίπατος σπάνια να μεταβαίνει από το A στο \bar{A} και αντίστροφα.

2η εκδοχή

Η απόσταση μετακίνησης (commute distance) c_{ij} είναι ο αναμενόμενος χρόνος που χρειάζεται ο τυχαίος περίπατος για να ταξιδέψει από την κορυφή u_i στην u_j και πίσω. Η απόσταση μετακίνησης, σε αντίθεση με την απόσταση του συντομότερου μονοπατιού, έχει την ιδιότητα να μειώνεται αν υπάρχουν πολλοί διαφορετικοί σύντομοι δρόμοι να βρεθεί ο τυχαίος περίπατος από τη u_i στην u_j . Έτσι η απόσταση μετακίνησης αντί να ψάχνει για το συντομότερο μονοπάτι, ψάχνει για ένα σύνολο συντομότερων μονοπατιών. Σημεία που συνδέονται με ένα σύντομο μονοπάτι στον γράφο και βρίσκονται στην ίδια περιοχή υψηλής πυκνότητας του γράφου θεωρούνται πιο κοντινά σε σχέση με σημεία που συνδέονται με ένα σύντομο μονοπάτι αλλά βρίσκονται σε διαφορετικές περιοχές υψηλής πυκνότητας. Η ιδιότητα αυτή καθιστά την απόσταση μετακίνησης κατάλληλη για σκοπούς ομαδοποίησης.

Υπολογισμός της απόσταση μετακίνησης

Η απόσταση αυτή μπορεί να υπολογιστεί με τη βοήθεια του γενικευμένου αντιστρόφου ενός Λαπλασιανου γράφου $c_{ij} = \text{vol}(V) \|z_i - z_j\|^2$ (Moore-Penrose

pseudo inverse). Συμβολίζουμε $e_i = (0, \dots, 1, \dots, 0)$ το i -οστό μοναδιαίο διάνυσμα. Υπενθυμίζουμε ότι ο L μπορεί να γραφτεί ως $L = U\Lambda U'$, όπου U ο πίνακας που περιέχει όλα τα ιδιοδιανύσματα του L ως στήλες και Λ ο διαγώνιος πίνακας με τις αντίστοιχες ιδιοτιμές $\lambda_1, \dots, \lambda_n$. Καθώς τουλάχιστον μία από τις ιδιοτιμές του L είναι 0, ο L δεν είναι αντιστρέψιμος. Γι' αυτό ορίζουμε τον γενικευμένο αντίστροφο $L^\dagger = U\Lambda^\dagger U'$, όπου ο πίνακας με διαγώνιες τιμές

$$\begin{cases} \frac{1}{\lambda_i} & , \text{αν } \lambda_i \neq 0 \\ 0 & , \text{αλλιώς.} \end{cases}$$

Οι τιμές του L^\dagger υπολογίζονται σύμφωνα με τον τύπο $\lambda_{ij}^\dagger = \sum_{k=2}^n \frac{1}{\lambda_i} u_{ik} u_{kj}$. Ο L^\dagger είναι συμμετρικός και θετικά ημιορισμένος.

Πρόταση Έστω $G = (V, E)$ ένας συνδεδεμένος, μη προσανατολισμένος γράφος. Συμβολίζουμε με c_{ij} την απόσταση μετακίνησης μεταξύ των δυο κορυφών v_i, v_j και με $L^\dagger = (l_{ij}^\dagger)_{i,j=1,\dots,n}$ τον γενικευμένο αντίστροφο του L . Τότε $c_{ij} = \text{vol}(G)(l_{ii}^\dagger - 2l_{ij}^\dagger + l_{jj}^\dagger) = \text{vol}(G)(e_i - e_j)'L^\dagger(e_i - e_j)$.

Συνέπεια Η $\sqrt{c_{ij}}$ μπορεί να θεωρηθεί ως μια συνάρτηση Ευκλείδειας απόστασης των κορυφών του γράφου. Δηλαδή, μπορούμε να κατασκευάσουμε μια εμφύτευση που να αντιστοιχίζει τις κορυφές v_i του γράφου σε σημεία $z_i \in \mathbb{R}^n$ ώστε οι Ευκλείδειες αποστάσεις μεταξύ των z_i να συμπίπτουν με τις αποστάσεις μετακίνησης στον γράφο. Καθώς ο L^\dagger είναι συμμετρικός και θετικά ημιορισμένος επάγει ένα εσωτερικό γινόμενο στον \mathbb{R}^n . Επιλέγουμε το z_i ως το σημείο του \mathbb{R}^n που αντιστοιχεί στην i -οστή γραμμή του πίνακα $U(\Lambda^\dagger)^{\frac{1}{2}}$. Τότε από την πρόταση και την κατασκευή του L^\dagger έχουμε $\langle z_i, z_j \rangle = e_i' L^\dagger e_j$ και $\|z_i - z_j\|^2 = c_{ij}$.

5.4.3 ΘΕΩΡΙΑ ΔΙΑΤΑΡΑΧΩΝ

Η θεωρία διαταραχών μελετά το πώς οι ιδιοτιμές και τα ιδιοδιανύσματα ενός πίνακα μεταβάλλονται αν προσθέσουμε μία μικρή διαταραχή H (διαταραγμένος πίνακας $A := A + H$). Τα περισσότερα θεωρήματα διαταραχών αναφέρουν πως

μια συγκεκριμένη απόσταση μεταξύ ιδιοτιμών ή ιδιοδιανυσμάτων των A και \tilde{A} φράσσεται από μια σταθερά επί μία νόρμα του H . Η σταθερά συνήθως εξαρτάται από την ιδιοτιμή την οποία εξετάζουμε και από το κατά πόσο αυτή η ιδιοτιμή είναι απομακρυσμένη από το υπόλοιπο φάσμα. Η δικαιολόγηση της φασματικής ομαδοποίησης είναι τότε η ακόλουθη:

Ας θεωρήσουμε πρώτα την ιδανική περίπτωση όπου η εντός του cluster η ομοιότητα είναι ακριβώς 0. Τότε, όπως έχουμε δει, τα k πρώτα ιδιοδιανύσματα του L ή του L_{rw} είναι τα διανύσματα-δείκτες των clusters. Σε αυτήν την περίπτωση, τα σημεία $y_i \in \mathbb{R}^k$ που κατασκευάζονται στα clusters έχουν τη μορφή $(0, \dots, 1, \dots, 0)'$ όπου η θέση του 1 υποδεικνύει τον connected component στον οποίο ανήκει το σημείο. Συγκεκριμένα όλα τα y_i που αντιστοιχούν στον ίδιο connected component συμπίπτουν. Ο αλγόριθμος k -μέσων θα βρει κατά τετριμμένο τρόπο τη σωστή διαμέριση, τοποθετώντας ένα κεντρικό σημείο σε κάθε σημείο $(0, \dots, 1, \dots, 0)' \in \mathbb{R}^k$. Σε μια σχετικά ιδανική περίπτωση όπου έχουμε ακόμα διακεκριμένα clusters αλλά η ομοιότητα εντός του cluster δεν είναι ακριβώς 0, θεωρούμε ότι οι Λαπλασιανοί πίνακες είναι διαταραγμένες εκδοχές αυτών της ιδανικής περίπτωσης. Η θεωρία διαταραχών μας λέει τότε πως τα ιδιοδιανύσματα θα είναι κοντά στα ιδανικά διανύσματα-δείκτες. Τα σημεία y_i μπορεί να μην συμπίπτουν απολύτως με τα $(0, \dots, 1, \dots, 0)'$, αλλά τα προσεγγίζουν υπό έναν μικρό όρο σφάλματος. Έτσι, εάν οι διαταραχές δεν είναι πολύ μεγάλες, ο αλγόριθμος k -μέσων θα διαχωρίσει και πάλι τις ομάδες μεταξύ τους.

Το τυπικό επιχείρημα της διαταραχής

Η τυπική βάση της προσέγγισης μέσω της θεωρίας διαταραχών για τη φασματική ομαδοποίηση είναι το θεώρημα Davis-Kahan. Αυτό το θεώρημα οριοθετεί τις διαφορές μεταξύ ιδιοχώρων συμμετρικών πινάκων υπό διαταραχές. Στη θεωρία διαταραχών, οι αποστάσεις μεταξύ υπόχωρων συνήθως μετρώνται με τη χρήση κανονικών (κύριων) γωνιών. Για να ορίσουμε τις κανονικές γωνίες, έστωσαν X_1 και X_2 δύο p -διάστατοι υπόχωροι του \mathbb{R}^d και V_1 και V_2 δύο πίνακες τέτοιοι ώστε οι στήλες τους να σχηματίζουν ορθοκανονικά συστήματα για τους X_1 , X_2 , αντίστοιχα. Τότε τα συνημίτονα $\cos \Theta_i$ των κύριων γωνιών Θ_i είναι οι ιδιάζουσες τιμές του $V_1'V_2$. Για $p = 1$, οι ως άνω ορισμένες κανονικές γωνίες

συμπίπτουν με τον κλασικό ορισμό της γωνίας. Οι κανονικές γωνίες μπορούν επίσης να οριστούν αν οι X_1, X_2 δεν έχουν την ίδια διάσταση. Ο πίνακας $\sin \Theta_i$ θα δηλώνει τον διαγώνιο πίνακα με τα ημίτονα των κανονικών γωνιών στη διαγώνιο.

Θεώρημα Davis-Kahan :

Έστωσαν $A, H \in \mathbb{R}^{n \times n}$ συμμετρικοί πίνακες και $\|\cdot\|$ η νόρμα Frobenius ή η 2-νόρμα των πινάκων αντίστοιχα. Θεωρούμε τον $\tilde{A} := A + H$ ως τη διαταραγμένη εκδοχή του A . Έστω διάστημα $S \subset \mathbb{R}$. Συμβολίζουμε $\sigma_{S_1}(A)$ το σύνολο των ιδιοτιμών του A που περιέχονται στο S_1 και με V_1 τον ιδιόχωρο που αντιστοιχεί σε όλες αυτές τις ιδιοτιμές (τυπικότερα, το V_1 είναι η εικόνα της φασματικής προβολής που επάγεται από το $\sigma_{S_1}(A)$). Συμβολίζουμε με $\sigma_{S_1}(\tilde{A})$ και \tilde{V}_1 τις αντίστοιχες ποσότητες για τον \tilde{A} . Ορίζουμε την απόσταση ως

$$\delta = \min\{|\lambda - s|; \lambda \text{ ιδιοτιμή του } A, \lambda \notin S_1, s \in S_1\}.$$

Τότε η απόσταση $d(V_1, \tilde{V}_1) := \|\sin \Theta(V_1, \tilde{V}_1)\|$ μεταξύ των δύο υποχώρων V_1 και \tilde{V}_1 φράσσεται από το $d(V_1, \tilde{V}_1) \leq \frac{\|H\|}{\delta}$. Προσπαθώντας να αποκρυπτογραφήσουμε αυτό το θεώρημα, για ευκολία θα εργαστούμε με τον Λαπλασιανό γράφο. Ο πίνακας A θα αντιστοιχεί στο γράφο L στην ιδανική περίπτωση όπου ο L έχει k κύριες συνιστώσες. Ο πίνακας \tilde{A} αντιστοιχεί σε μία διαταραγμένη περίπτωση όπου, λόγω θορύβου, οι k συνιστώσες δεν είναι πλέον εντελώς ασύνδετοι, αλλά είναι μόνο συνδεδεμένοι, μέσω μερικών ακμών με χαμηλά βάρη. Ο αντίστοιχος Λαπλασιανός γράφος είναι τότε ο \tilde{L} . Για τη φασματική ομαδοποίηση πρέπει να θεωρήσουμε τις k πρώτες ιδιοτιμές και ιδιοδιανύσματα του \tilde{L} . Συμβολίζουμε τις ιδιοτιμές του L ως $\lambda_1, \dots, \lambda_n$ και του \tilde{L} ως $\tilde{\lambda}_1, \dots, \tilde{\lambda}_n$. Το κρίσιμο σημείο είναι τώρα να επιλέξουμε το διάστημα S_1 . Θέλουμε να το επιλέξουμε έτσι ώστε τα k πρώτα ιδιοδιανύσματα και του L και του \tilde{L} να ανήκουν στο S_1 . Αυτό είναι ευκολότερο όσο μικρότερη είναι η διαταραχή $H = L - \tilde{L}$ και όσο μεγαλύτερο είναι το ιδιοκενό $|\lambda_k - \lambda_{k+1}|$. Εάν καταφέρουμε να βρούμε τέτοιο διάστημα, το θεώρημα Davis-Kahan μας λέει ότι τα ιδιοκενά που αντιστοιχούν στις k πρώτες ιδιοτιμές του ιδανικού πίνακα L και στις k πρώτες ιδιοτιμές του διαταραγμένου πίνακα \tilde{L} είναι πολύ κοντά μεταξύ τους, δηλαδή ότι η απόσταση

φράσσεται από το $\frac{\|H\|}{\delta}$. Τότε καθώς τα ιδιοδιανύσματα στην ιδανική περίπτωση είναι κατά ζεύγη σταθερά στις κύριες συνιστώσες, το ίδιο θα ισχύει προσεγγιστικά για την διαταραγμένη περίπτωση. Το πόσο καλή θα είναι η προσέγγιση εξαρτάται από την νόρμα της διαταραχής $\|H\|$ και την απόσταση δ ανάμεσα στο S_1 και στο $(k+1)$ -οστό ιδιοδιάνυσμα του L . Αν επιλέξουμε $S_1 = [0, \lambda_k]$, τότε η δ συμπίπτει με το φασματικό κενό $|\lambda_{k+1} - \lambda_k|$. Από το θεώρημα βλέπουμε πως όσο μεγαλύτερο είναι το ιδιοκενό, τόσο κοντινότερα είναι τα ιδιοδιανύσματα της ιδανικής και της διαταραγμένης περίπτωσης και κατά συνέπεια, τόσο καλύτερα δουλεύει η φασματική ομαδοποίηση.

Αν η διαταραχή H είναι πολύ μεγάλη ή το ιδιοκενό πολύ μικρό, ίσως δεν μπορούμε να βρούμε ένα σύνολο S_1 τέτοιο ώστε τόσο οι k πρώτες ιδιοτιμές τόσο του L όσο και του \tilde{L} να βρίσκονται μέσα σε αυτό. Σε αυτήν την περίπτωση, πρέπει να κάνουμε έναν συμβιβασμό επιλέγοντας το S_1 να περιέχει τις k πρώτες ιδιοτιμές του L , μα ίσως περισσότερες ή λιγότερες ιδιοτιμές του \tilde{L} . Το πόρισμα του θεωρήματος γίνεται τότε ασθενέστερο, με την έννοια ότι είτε δεν συγκρίνουμε τα ιδιοκενά που αντιστοιχούν στα k πρώτα ιδιοδιανύσματα του L και του \tilde{L} , αλλά τα ιδιοκενά που αντιστοιχούν στα k πρώτα ιδιοδιανύσματα του L και τα \tilde{k} πρώτα ιδιοδιανύσματα του \tilde{L} (όπου \tilde{k} το πλήθος των ιδιοτιμών του \tilde{L} που ανήκουν στο S_1). Είτε μπορεί η απόσταση δ να μικρύνει τόσο που το όριο της απόστασης $d(V_1, \tilde{V}_1)$ μεγαλώνει τόσο πολύ που είναι άχρηστο.

Αυτή είναι μία αδρή παρουσίαση του πώς η θεωρία διαταραχών εξηγεί τους αλγόριθμους φασματικής ομαδοποίησης που δίνεται για λόγους πληρότητας και δε θα σχολιαστεί περαιτέρω.

Κεφάλαιο 6

Πίνακες Διάχυσης (Diffusion Maps)

Οι συναρτήσεις διάχυσης (diffusion maps) (Nadler et al. 2006) είναι μια μη γραμμική τεχνική που κατορθώνει την αναγωγή δεδομένων μεγάλης κλίμακας αναδιοργανώνοντάς τα σύμφωνα με την υποκείμενη γεωμετρία τους. Η συνδεσιμότητα του συνόλου δεδομένων, μετρημένη με κάποιο μέτρο τοπικής ομοιότητας, χρησιμοποιείται για την δημιουργία μιας διαδικασίας διάχυσης που εξαρτάται από τον χρόνο. Καθώς η διαδικασία εξελίσσεται, ενσωματώνει την τοπική γεωμετρία, για να αποκαλύψει γεωμετρικές κατασκευές του συνόλου δεδομένων σε διαφορετικές κλίμακες. Ορίζοντας μια χρονικά εξαρτώμενη μετρική διάχυσης, μπορούμε τότε να μετρήσουμε την ομοιότητα μεταξύ δύο σημείων σε μία συγκεκριμένη κλίμακα (ή χρόνο), βασιζόμενοι στην γεωμετρία που έχουμε βρει. Μια συνάρτηση διάχυσης αντιστοιχεί (ή μετατρέπει) σημεία σε έναν χώρο μικρότερης διάστασης, τέτοιου ώστε η ευκλείδεια απόσταση μεταξύ σημείων να προσεγγίζει την απόσταση διάχυσης στον αρχικό χώρο. Η διάσταση του χώρου διάχυσης καθορίζεται από τη γεωμετρική κατασκευή στην οποία υπόκεινται τα δεδομένα και στην ακρίβεια με την οποία προσεγγίζεται η απόσταση διάχυσης (Nadler et al. 2005)

6.1 Η Μαθηματική Προσέγγιση της συνάρτησης διάχυσης

Έστω ένας τυχαίος περίπατος στο αρχικό σύνολο δεδομένων μας. Εύκολα αντιλαμβανόμαστε ότι ο τυχαίος περίπατος είναι πιθανότερο να μεταπηδήσει σε κοντινό σημείο δεδομένων, παρά σε κάποιο απομακρυσμένο κι έτσι μπορούμε να συσχετίσουμε την έννοια της απόστασης στον αρχικό χώρο με την έννοια της πιθανότητας. Η συνδεσιμότητα μεταξύ δύο σημείων x και y ορίζεται ως η πιθανότητα μεταπήδησης του τυχαίου περιπάτου από το x στο y σε ένα βήμα ($connectivity(x, y) = p(x, y)$) και την εκφράζουμε σαν μία μη-κανονικοποιημένη συνάρτηση πιθανότητας γνωστής ως πυρήνα διάχυσης $connectivity(x, y) \propto k(x, y)$. Ο πυρήνας καθορίζει ένα τοπικό μέτρο ομοιότητας μέσα σε μία συγκεκριμένη γειτονιά. Έξω από τη γειτονιά, η συνάρτηση γρήγορα μηδενίζεται. Για παράδειγμα, μπορούμε να θεωρήσουμε τον Γκαουσιανό πυρήνα,

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

Η γειτονιά του x μπορεί να οριστεί ως όλα εκείνα τα στοιχεία y για τα οποία $k(x, y) \geq \epsilon$ με $0 < \epsilon \ll 1$. Αυτά ορίζουν μία περιοχή μέσα στην οποία είμαστε σίγουροι ότι η μέτρηση της τοπικής μας ομοιότητας (π.χ. η Ευκλείδεια απόσταση) είναι ακριβής. Μεταβάλλοντας ελάχιστα την κλίμακα του πυρήνα (στην περίπτωση μας το α) επιλέγουμε το μέγεθος της γειτονιάς, βασιζόμενοι στην εκ των προτέρων γνώση της δομής και της πυκνότητας των δεδομένων. Για πολύπλοκες, χαμηλής διάστασης δομές, επιλέγεται μια μικρή περιοχή. Για αραιά δεδομένα, μια μεγαλύτερη γειτονιά είναι καταλληλότερη. Ο πυρήνας διάχυσης ικανοποιεί τις παρακάτω ιδιότητες:

- Ο k είναι συμμετρικός: $k(x, y) = k(y, x)$.
- Ο k είναι μη αρνητικός: $k(x, y) \geq 0$.

Η πρώτη ιδιότητα είναι αναγκαία για να εκτελέσουμε φασματική ανάλυση σε έναν πίνακα απόστασης $K_{ij} = k(x_i, x_j)$. Η δεύτερη επιτρέπει στον πυρήνα διάχυσης να ερμηνευθεί ως κλιμακωτή πιθανότητα (πάντα θετική), έτσι

ώστε $\frac{1}{d_X} \sum_{y \in X} k(x, y) = 1$. Τότε η σχέση μεταξύ του πυρήνα διάχυσης και της συνδεσιμότητας είναι

$$\text{connectivity}(x, y) = p(x, y) = \frac{1}{d_X} k(x, y)$$

όπου $\frac{1}{d_X}$, η σταθερά κανονικοποίησης.

Κατασκευάζουμε έναν πίνακα διάχυσης P , με στοιχεία $p_{ij} = p(X_i, X_j)$. Κάθε στοιχείο δίνει την συνδεσιμότητα μεταξύ οποιωνδήποτε δύο σημείων X_i και X_j των δεδομένων και συνοψίζει τις γνώσεις μας σε τοπικό επίπεδο. Αναλογικά με τον τυχαίο περίπατο, ο πίνακας αυτός δίνει τις πιθανότητες ενός βήματος από το i στο j . Παίρνοντας δυνάμεις του πίνακα διάχυσης, αυξάνουμε με τον αριθμό των βημάτων. Ο πίνακας διάχυσης μπορεί να ερμηνευθεί ως ο πίνακας μεταβασης μιας Μαρκοβιανής αλυσίδας που ορίζεται στα δεδομένα. Για παράδειγμα, έστω ένας 2×2 πίνακας διάχυσης,

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

Κάθε στοιχείο p_{ij} , είναι η πιθανότητα μεταπήδησης μεταξύ των i και j . Τότε ο P^2 , είναι

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

Παρατηρείστε ότι $P_{11} = p_{11}p_{11} + p_{12}p_{21}$, που αθροίζει δύο πιθανότητες: να παραμείνει ο τυχαίος περίπατος στο σημείο 1 και να μετακινηθεί από το 1 στο 2 και πίσω. Κάνοντας δύο μεταπηδήσεις, αυτά είναι όλα τα μονοπάτια από το i στο j . Όμοια, ο P_{ij}^t αθροίζει όλα τα μονοπάτια μήκους t από το i στο j .

6.1.1 Συνδεσιμότητα

Καθώς υπολογίζουμε τις πιθανότητες P^t για αυξανόμενη τιμή του t , παρατηρούμε το σύνολο δεδομένων σε διαφορετικές κλίμακες. Αυτή είναι η διαδικασία διάχυσης, όπου η τοπική συνδεσιμότητα παρέχει την ολική συνδεσιμότητα του συνόλου των δεδομένων. (Θεωρητικά, ο τυχαίος περίπατος είναι μία στοχαστική διαδικασία διακριτού χρόνου, ενώ η διαδικασία διάχυσης θεωρείται στοχαστική διαδικασία συνεχούς χρόνου. Εδώ εξετάζουμε μόνο τη διακριτή περίπτωση, θεωρώντας τον τυχαίο περίπατο και τη διαδικασία διάχυσης ισοδύναμα.) Για αυξανόμενες τιμές του t (προς τα εμπρός διαδικασία διάχυσης) (forward diffusion process), η πιθανότητα να ακολουθηθεί ένα μονοπάτι κατά μήκος της υποκείμενης γεωμετρίας αυξάνεται. Αυτό συμβαίνει διότι κατά μήκος της γεωμετρικής δομής τα σημεία είναι πυκνά και κατά συνέπεια πλήρως συνδεδεμένα, σχηματίζοντας σύντομα μονοπάτια και μεγάλης πιθανότητας μεταπηδήσεις. Αντίθετα, τα μονοπάτια που δεν ακολουθούν τη γεωμετρική δομή περιλαμβάνουν μία ή περισσότερες μικρής πιθανότητας μεταπηδήσεις, μειώνοντας έτσι τη συνολική πιθανότητα του μονοπατιού.

6.1.2 Απόσταση Διάχυσης

Είδαμε πώς η διαδικασία διάχυσης αποκαλύπτει την ολική γεωμετρική δομή των δεδομένων. Παρακάτω ορίζουμε μία μετρική διάχυσης βασισμένη σε αυτή τη δομή. Η μετρική υπολογίζει την ομοιότητα μεταξύ δύο σημείων στον χώρο των παρατηρήσεων ως την συνδεσιμότητα μεταξύ τους (πιθανότητα μεταπήδησης). Σχετίζεται με τον πίνακα διάχυσης και είναι

$$\begin{aligned} D_t^2(X_i, X_j) &= \sum_{u \in X} \|p_t(X_i, u) - p_t(X_j, u)\|^2 \\ &= \sum_k \|P_{ik}^t - P_{kj}^t\|^2 \end{aligned} \tag{6.1}$$

Η απόσταση διάχυσης είναι μικρή αν υπάρχουν πολλά υψηλής πιθανότητας μονοπάτια μήκους t μεταξύ δύο σημείων. Καθώς η διαδικασία διάχυσης προχωρά αποκαλύπτοντας τη γεωμετρική δομή των δεδομένων, οι κύριοι παράγοντες που συμβάλλουν στην απόσταση διάχυσης είναι μονοπάτια κατά μήκος της δο-

μής αυτής.

Ο όρος $p_t(x, u)$ στην απόσταση διάχυσης είναι η πιθανότητα μεταπήδησης από το x στο u σε t μονάδες χρόνου (βήματα), και αθροίζει τις πιθανότητες όλων των δυνατών μονοπατιών μήκους t μεταξύ των x και u . Για να παραμείνει μικρή η απόσταση διάχυσης πρέπει οι πιθανότητες των μονοπατιών μεταξύ των x και u και x και y να είναι περίπου ίσες, πράγμα που συμβαίνει όταν τα x και y είναι καλά συνδεδεμένα μέσω του u .

6.1.3 Συνάρτηση Διάχυσης

Ο υπολογισμός των αποστάσεων διάχυσης είναι υπολογιστικά ακριβός, γι'αυτό είναι πιο εύκολο να αντιστοιχίσουμε τα σημεία των δεδομένων στον ευκλείδειο χώρο, σύμφωνα με τη μετρική διάχυσης. Η απόσταση διάχυσης στον χώρο των δεδομένων γίνεται απλά η Ευκλείδεια απόσταση σ'αυτόν τον νέο χώρο διάχυσης. Ένας diffusion map, που αντιστοιχίζει τις συντεταγμένες μεταξύ του χώρου δεδομένων και του χώρου διάχυσης, στοχεύει στην αναδιοργάνωση των δεδομένων σύμφωνα με τη μετρική διάχυσης. Γι' αυτό τον εκμεταλλευόμαστε για τη μείωση της διαστατικότητας. Ο diffusion map διατηρεί την εγγενή γεωμετρία του συνόλου δεδομένων και καθώς ο μετασχηματισμός μετράει τις αποστάσεις σε μια μικρότερης διάστασης δομή, περιμένουμε να βρούμε ότι είναι απαραίτητες λιγότερες συντεταγμένες για να αναπαραστήσουμε τα δεδομένα στον καινούριο χώρο. Μένει να βρούμε ποιές διαστάσεις να αγνοήσουμε, προκειμένου να διατηρήσουμε τις αποστάσεις διάχυσης, και επομένως τη γεωμετρία, στον βέλτιστο δυνατό βαθμό. Εξετάζουμε τον μετασχηματισμό

$$Y_i := \begin{bmatrix} p_t(X_i, X_1) \\ p_t(X_i, X_2) \\ \vdots \\ p_t(X_i, X_N) \end{bmatrix} = P_{i*}^T. \quad (6.2)$$

Γι'αυτόν τον μετασχηματισμό, η Ευκλείδεια απόσταση μεταξύ δύο σημείων

X_i και X_j είναι

$$\begin{aligned} \|Y_i - Y_j\| &= \sum_{u \in X} |p_t(X_i, u) - p_t(X_j, u)|^2 \\ &= \sum |P_{ik}^t - P_{kj}^t|^2 = D_t^2(X_i, X_j) \end{aligned} \quad (6.3)$$

που είναι η απόσταση διάχυσης μεταξύ των σημείων X_i και X_j . Έχουμε βρει λοιπόν έναν τρόπο να εκφράσουμε τα δεδομένα σύμφωνα με την απόσταση διάχυσης.

Η μείωση της διαστατικότητας γίνεται αγνοώντας κάποιες διαστάσεις στον χώρο διάχυσης. Παίρνουμε τον κανονικοποιημένο πίνακα διάχυσης $P = D^{-1}K$, όπου D ο διαγώνιος πίνακας με στοιχεία τα αθροίσματα των γραμμών του K . Οι αποστάσεις διάχυσης μπορούν να εκφραστούν με τη βοήθεια των ιδιοτιμών και ιδιοδιανυσμάτων του P ως εξής:

$$Y_i' = \begin{bmatrix} \lambda_1^t \psi_1^{(i)} \\ \lambda_2^t \psi_2^{(i)} \\ \vdots \\ \lambda_n^t \psi_n^{(i)} \end{bmatrix} = P_{i*}^T. \quad (6.4)$$

Όπου το $\psi_1^{(i)}$ υποδεικνύει το i -οστό στοιχείο του πρώτου ιδιοδιανύσματος του P . Και πάλι η Ευκλείδεια απόσταση μεταξύ των σημείων Y_i' και Y_j' είναι η απόσταση διάχυσης. Το σύνολο των ορθογώνιων αριστερών ιδιοδιανυσμάτων του P σχηματίζουν μια βάση για τον χώρο διάχυσης και οι αντίστοιχες ιδιοτιμές λ_i δείχνουν τη σημαντικότητα κάθε διάστασης. Η μείωση διαστάσεων γίνεται κρατώντας τις m διαστάσεις που σχετίζονται με τα κυρίαρχα ιδιοδιανύσματα, γεγονός που εξασφαλίζει ότι το $\|Y_i' - Y_j'\|$ προσεγγίζει κατά βέλτιστο τρόπο την απόσταση διάχυσης $D_t(X_i, X_j)$. Επομένως, ο diffusion map 6.2 διατηρεί κατά βέλτιστο τρόπο την εγγενή γεωμετρία των δεδομένων.

6.2 Βασικός αλγόριθμος Diffusion Mapping

Ο βασικός αλγόριθμος για diffusion mapping σε μορφή ψευδοκώδικα είναι ο παρακάτω:

Ψευδοκώδικας

Είσοδος: Σύνολο δεδομένων υψηλής διάστασης $\{X_i\}_{i=1}^N$.

Βήμα 1: Όρισε έναν πυρήνα $k(x, y)$ και κατασκεύασε έναν πίνακα πυρήνα K , τέτοιο ώστε $K_{i,j} = k(X_i, X_j)$.

Βήμα 2: Κατασκεύασε τον πίνακα διάχυσης κανονικοποιώντας τις γραμμές του πίνακα πυρήνα.

Βήμα 3: Υπολόγισε τα ιδιοδιανύσματα του πίνακα διάχυσης.

Βήμα 4: Μετασχημάτισε στον d -διάστατο χώρο διάχυσης σε χρόνο t , χρησιμοποιώντας τα d κυρίαρχα ιδιοδιανύσματα και ιδιοτιμές, όπως φαίνεται στο (1).

Έξοδος: Μικρότερης διάστασης σύνολο δεδομένων $\{Y_i\}_{i=1}^N$

Κεφάλαιο 7

Εφαρμογές των αλγορίθμων και σύγκριση της αποτελεσματικότητάς τους

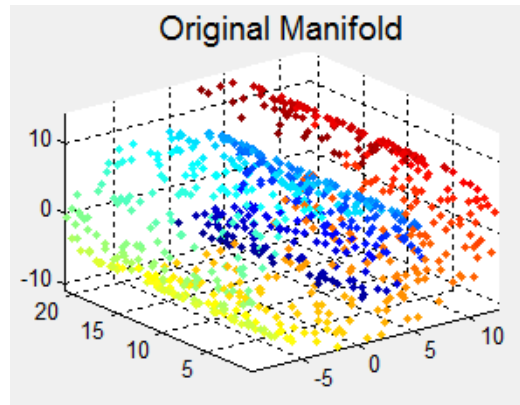
Σε αυτό το κεφάλαιο θα εφαρμόσουμε τους αλγόριθμους που αναλύθηκαν για το ίδιο πρόβλημα και θα συγκρίνουμε την αποτελεσματικότητά τους.

Για την ανάλυση χρησιμοποιήθηκε το πρόγραμμα του Todd Whitman "mani.m" που είναι διαθέσιμο στον παρακάτω σύνδεσμο:

<https://ocw.mit.edu/courses/earth-atmospheric-and-planetary-sciences/12-s990-quantifying-uncertainty-fall-2012/tools/mani.m>.

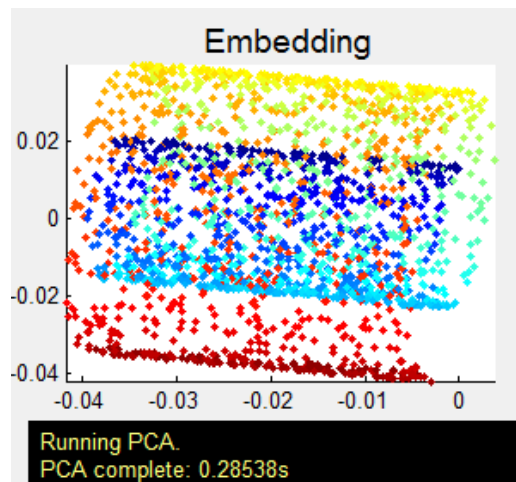
Σύνολο Δεδομένων

Έχουμε ένα τυχαίο δείγμα μεγέθους $N = 2000$ από μία σπειροειδή ζώνη ("Swiss roll").

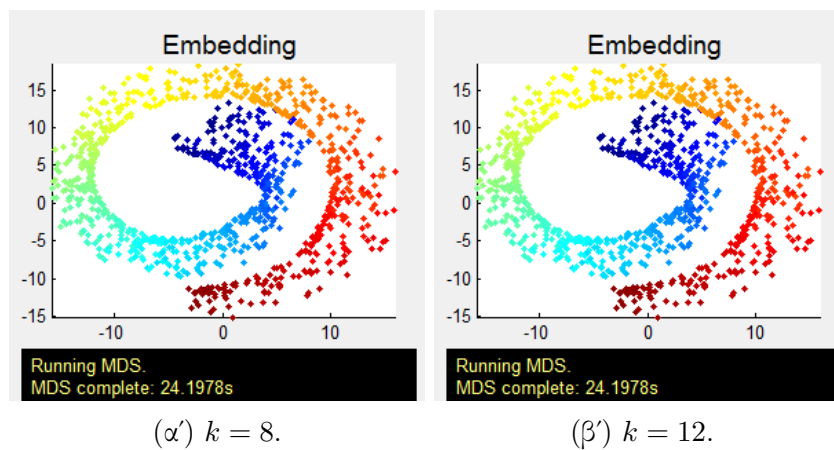


Σχήμα 7.1: Τρισδιάστατη μορφή δείγματος δεδομένων από το σύνολο 'Swiss Roll' μεγέθους $N = 2000$ σημεία.

7.1 Αποτελεσματικότητα αλγορίθμων PCA-MDS

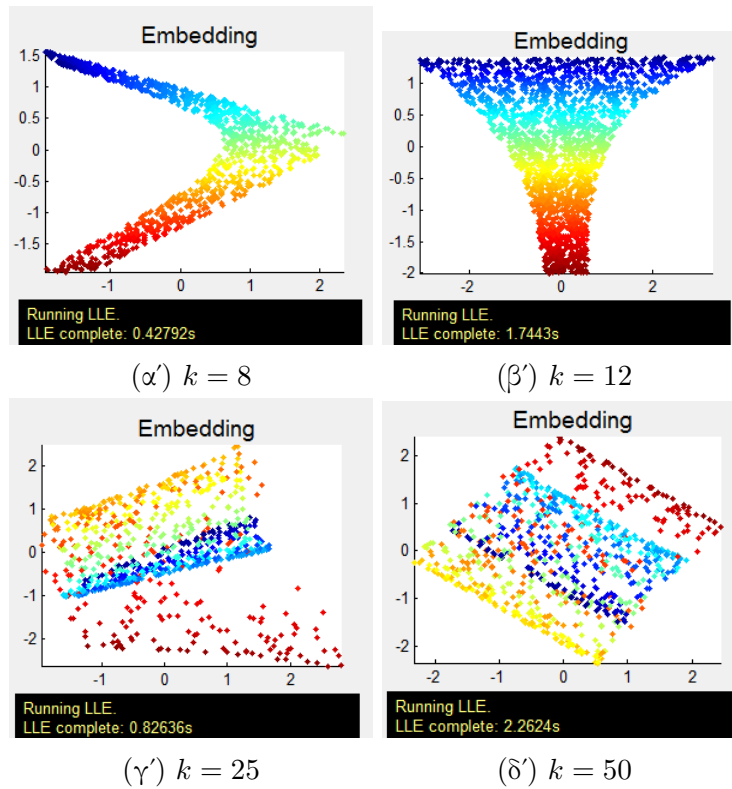


Σχήμα 7.2: Αποτελέσματα αλγορίθμου PCA μετά από εφαρμογή στο σύνολο "Swiss Roll" με $N = 2000$ σημεία.



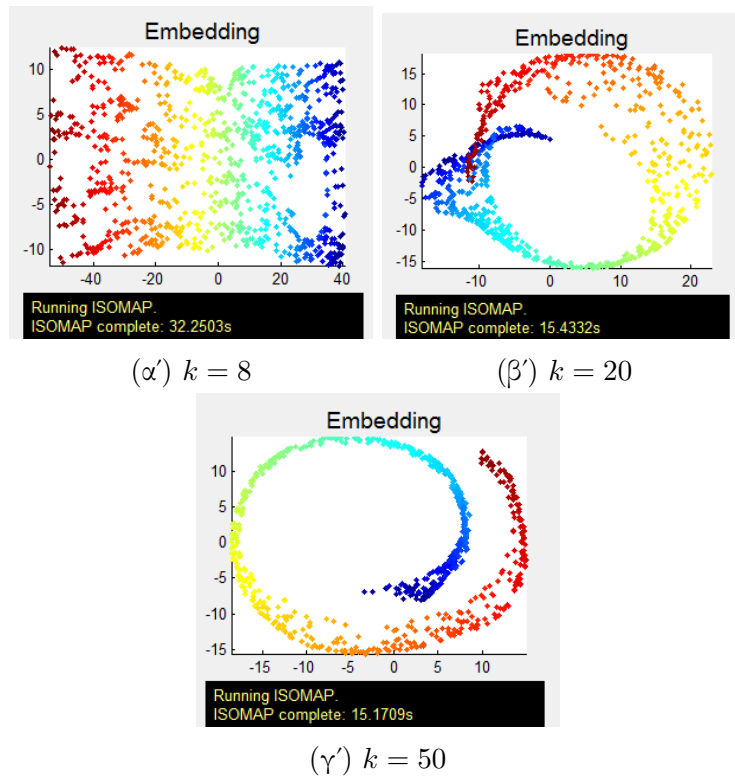
Σχήμα 7.3: Αποτελέσματα αλγορίθμου MDS μετά από εφαρμογή στο σύνολο "Swiss Roll" με $N = 2000$ (εικ. 7.1) σημεία και παραμέτρους $k = 8$ (εικ. 7.3α) και $k = 12$ (εικ. 7.3β) πλησιέστερους γείτονες

7.2 Αποτελεσματικότητα αλγορίθμου LLE



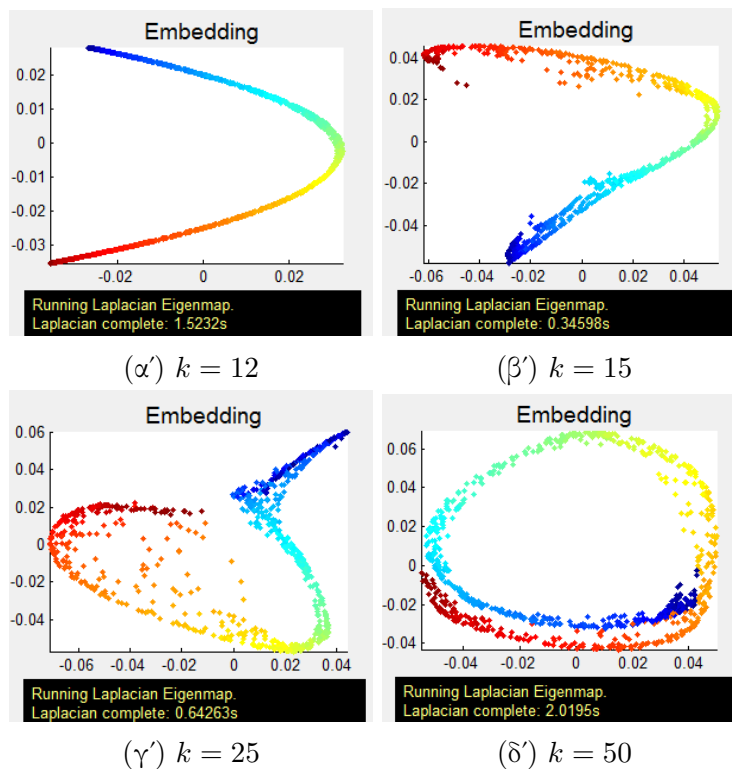
Σχήμα 7.4: Αποτελέσματα αλγορίθμου LLE μετά από εφαρμογή στο σύνολο "Swiss Roll" με $N = 2000$ σημεία και παραμέτρους $k = 8$ (εικ. 7.4α'), $k = 12$ (εικ. 7.4β'), $k = 25$ (εικ. 7.4γ') και $k = 50$ (εικ.7.4δ') πλησιέστερους γείτονες.

7.3 Αποτελεσματικότητα αλγορίθμου Isomap



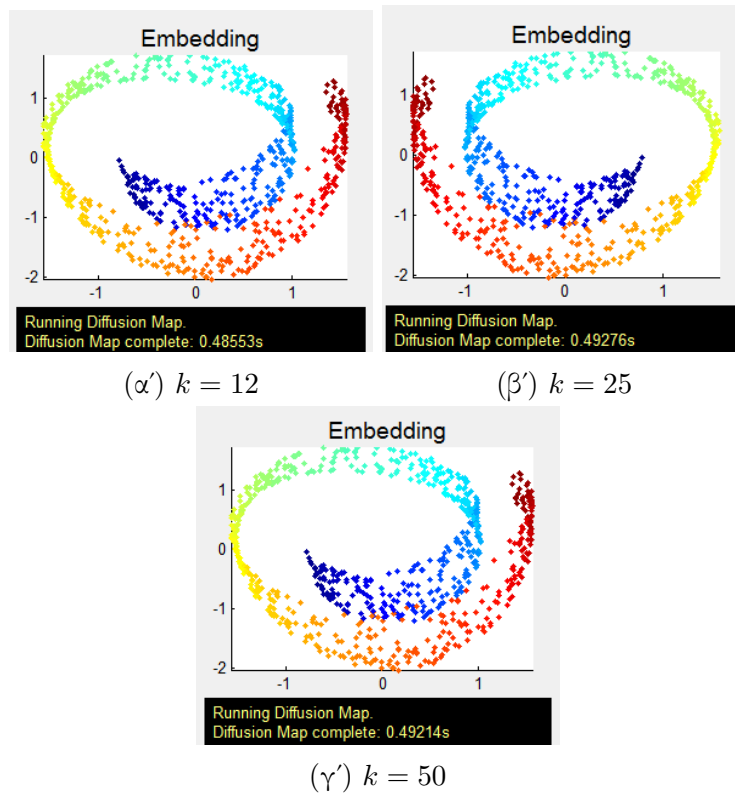
Σχήμα 7.5: Αποτελέσματα αλγορίθμου απεικόνισης Ισομετρικών Χαρακτηριστικών (Isomap) μετά από εφαρμογή στο σύνολο "Swiss Roll" με $N = 2000$ (εικ.7.1) σημεία και παραμέτρους $k = 8$ (εικ.7.5α'), $k = 20$ (εικ. 7.5β'), $k = 50$ (εικ. 7.5γ') πλησιέστερους γείτονες.

7.4 Αποτελεσματικότητα αλγορίθμων [(Spectral Clustering(Laplacian Eigenmaps))]



Σχήμα 7.6: Αποτελέσματα αλγορίθμου φασματικής ομαδοποίησης Spectral Clustering μετά από εφαρμογή στο σύνολο "Swiss Roll" με $N = 2000$ σημεία και παραμέτρους $k = 12$ (εικ. 7.6α'), $k = 15$ (εικ.7.6β'), $k = 25$ (εικ. 7.6γ') και $k = 50$ (εικ. 7.6δ') πλησιέστερους γείτονες.

7.5 Αποτελεσματικότητα αλγορίθμου Συναρτήσεων Διάχυσης (Diffusion Maps)



Σχήμα 7.7: Αποτελέσματα αλγορίθμου Συναρτήσεων Διάχυσης (Diffusion Maps) μετά από εφαρμογή στο σύνολο "Swiss Roll" με $N = 2000$ σημεία και παραμέτρους $k = 12$ (εικ. 7.7α'), $k = 25$ (εικ. 7.7β'), $k = 50$ (εικ. 7.7γ') πλησιέστερους γείτονες.

7.6 Σύγκριση αλγορίθμων

ΑΛΓΟΡΙΘΜΟΣ	ΑΠΟΔΟΤΙΚΟΤΗΤΑ		
	ΧΡΟΝΟΣ (s)	k	ΕΠΙΤΥΧΙΑ
PCA	0,28538	×	×
MDS	24,1978	8	✓
LLE		×	×
ISOMAP	15,1709	25	✓
Laplacian Eigenmaps	2,0195	50	✓
Diffusion Maps	0,48553	12	✓

Σχήμα 7.8: Συνοπτικός πίνακας σύγκρισης επιδόσεων αλγορίθμων PCA, MDS, LLE, Isomap, Spectral Clustering, Diffusion Mapping. Περιλαμβάνει τον χρόνο (στήλη 1) που χρειάστηκε ο κάθε αλγόριθμος **εφόσον αποκάλυψε με επιτυχία την εγγενή γεωμετρία των δεδομένων**(στήλη 3) για να εκτελέσει με τον καλύτερο δυνατό τρόπο την αναγωγή δεδομένων, και τον ελάχιστο δυνατό αριθμό γειτόνων στον οποίο το κατάφερε(στήλη 2).

Στην εικόνα 7.8 παρατηρούμε πως οι αλγόριθμοι που κατάφεραν να απεικονίσουν την εγγενή γεωμετρία των δεδομένων στο επίπεδο είναι η γραμμική MDS (εικ.7.3) και οι μη- γραμμικοί αλγόριθμοι Isomap (εικ. 7.5) και οι αλγόριθμοι που βασίζονται στις Λαπλασιανές ιδιοσυναρτήσεις (εικ. 7.6) και σε συναρτήσεις διάχυσης (εικ. 7.7). Η MDS κατάφερε σε 24" και χρησιμοποιώντας μόλις 8 πλησιέστερους γείτονες (εικ.7.3α') να απεικονίσει το σύνολο "swiss roll" σε μία σπείρα στο επίπεδο. Από την άλλη μεριά, ο Isomap χρειάστηκε 25 γείτονες και 10" λιγότερα για να φτάσει στο ίδιο αποτέλεσμα (εικ.7.5β'). Όσο και αν αυξήσουμε το πλήθος των γειτόνων που ζητάμε από τον αλγόριθμο να υπολογίσει, το αποτέλεσμα εξακολουθεί να είναι επιτυχημένο και ο χρόνος απόκρισης δε μεταβάλλεται σημαντικά. Ενδεικτικά, βλέπουμε πως η χρονική διαφορά μιας εφαρμογής του Isomap με αριθμό γειτόνων τέτοιο ώστε για πρώτη φορά βρίσκει τη δομή των δεδομένων ($k = 25$) (εικ. 7.5β') από μία εφαρμογή με διπλάσιο(!) αριθμό γειτόνων($k = 50$) (εικ. 7.5γ') διαφέρει μόνο κατά 0,20".

Ο αλγόριθμος φασματικής ομαδοποίησης (spectral clustering) αν και απαιτεί 50 γείτονες ώστε να αρχίσει να προσεγγίζει το ζητούμενο, εν τούτοις αποδεικνύεται πολύ αποτελεσματικός καθώς χρειάζεται μόλις 2" για να το κάνει (εικ. 7.6δ'). Ένα πρόβλημα, ωστόσο, του αλγόριθμου αυτού είναι πως αν ξεπεράσουμε ένα συγκεκριμένο εύρος για το πλήθος k , τότε ο αλγόριθμος δεν καταφέρνει πια να αναγνωρίσει τη δομή των δεδομένων και αποτυγχάνει. Ο πραγματικά αποτελεσματικότερος αλγόριθμος αποδεικνύεται αυτός που κάνει χρήση συναρτήσεων διάχυσης, καθώς καταφέρνει να προβάλλει με επιτυχία τα 3D δεδομένα στο επίπεδο σε μόλις 0,49" (!) και κάνοντας χρήση 12 πλησιέστερων γειτόνων (εικ. 7.7α').

Μέρος II

Η τεχνολογία των μικροσυστοιχιών (microarrays)

Κεφάλαιο 8

Βιολογικό υπόβαθρο

Σε αυτό το κεφάλαιο γίνεται μία σύντομη αναφορά σε κάποιες βιολογικές διαδικασίες για την ευκολότερη περαιτέρω παρακολούθηση της παρούσας εργασίας.

8.1 Λίγα λόγια για το DNA

Το DNA, όπως και το RNA, είναι ένα μακρομόριο, που αποτελείται από νουκλεοτίδια. Κάθε νουκλεοτίδιο του DNA αποτελείται από μία πεντόζη, τη δεοξυριβόζη, ενωμένη με μία φωσφορική ομάδα και μία αζωτούχο βάση. Στα νουκλεοτίδια του DNA η αζωτούχος βάση μπορεί να είναι μία από τις: αδενίνη (A), γουανίνη (G), κυτοσίνη (C) και θυμίνη (T). Οι δύο αλυσίδες ενός μορίου DNA είναι συμπληρωματικές, και αυτό υποδηλώνει ότι η αλληλουχία της μιας καθορίζει την αλληλουχία της άλλης. Οι αζωτούχες βάσεις της μιας αλυσίδας συνδέονται με δεσμούς υδρογόνου με τις αζωτούχες βάσεις της απέναντι αλυσίδας με βάση τον κανόνα της συμπληρωματικότητας. Η αδενίνη συνδέεται μόνο με θυμίνη και αντίστροφα, ενώ η κυτοσίνη μόνο με γουανίνη και αντίστροφα. Οι δεσμοί υδρογόνου που αναπτύσσονται μεταξύ των βάσεων σταθεροποιούν τη δομή του μορίου.

8.2 Γονίδια και Γονιδίωμα

Ένα γονίδιο είναι ένα τμήμα DNA με συγκεκριμένη ακολουθία που κωδικοποιεί μια συγκεκριμένη πρωτεΐνη (δηλ. μία συγκεκριμένη αλληλουχία αμινοξέων). Το μέγεθός της συνήθως μετρείται σε ζεύγη βάσεων (bp: base pairs) που αναφέρεται ως δομή διπλής έλικας του μορίου DNA. Το τυπικό μήκος του ανθρώπινου γονιδίου είναι περίπου 10.000 bp, αλλά μπορεί να φθάσει έως και 2,4 εκατομμύρια bp. Το σύνολο των γονιδίων ενός οργανισμού ονομάζεται γονιδίωμα. Το ανθρώπινο γονιδίωμα περιλαμβάνει περίπου 35.000 γονίδια ή 3 δισεκατομμύρια ζεύγη βάσεων.

8.3 Γονιδιακή έκφραση

Η γονιδιακή έκφραση είναι η διαδικασία μετατροπής των γενετικών πληροφοριών (μεταγραφής γονιδίων σε RNA αντίγραφα) σε πρωτεΐνες οι οποίες στη συνέχεια εξυπηρετούν μία βιολογική λειτουργία και αποτελείται από δύο (2) φάσεις:

- Σε ένα πρώτο στάδιο, το καλούπι DNA ενός γονιδίου μεταγράφεται στο αγγελιαφόρο- RNA (messenger- RNA: mRNA), χρησιμοποιώντας ένα ένζυμο που ονομάζεται RNA πολυμεράση. Κατά την έναρξη της μεταγραφής ενός γονιδίου η RNA προκαλεί τοπικό ξετύλιγμα της διπλής έλικας του DNA. Στη συνέχεια, τοποθετεί ριβονουκλεοτίδια απέναντι από τα δεοξυριβονουκλεοτίδια μίας αλυσίδας του DNA σύμφωνα με τον κανόνα της συμπληρωματικότητας των βάσεων (εδώ όμως απέναντι από την αδενίνη τοποθετείται το ριβονουκλεοτίδιο που περιέχει ουρακίλη).
- Το δεύτερο βήμα περιλαμβάνει τη μετάφραση της αλληλουχίας mRNA σε μια πρωτεΐνη.

Αυτή η διαδικασία δύο σταδίων είναι επίσης γνωστή ως το *θεμελιώδες δόγμα της Μοριακής Βιολογίας* (Crick 1958). Φυσικά, το επίπεδο έκφρασης ενός συγκεκριμένου γονιδίου μεταβάλλεται με την πάροδο του χρόνου.

Οι αλλαγές των συνθηκών οδηγούν σε μια αλλαγή της κατάστασης έκφρασης του γονιδίου ενός κυττάρου. Όταν ένα γονίδιο ρυθμίζεται προς τα πάνω(η έκφρασή του αυξάνεται) σε σχέση με μία προηγούμενη κατάσταση, τότε το mRNA μεταγράφεται σε ένα υψηλότερο ποσοστό. Ομοίως, μια κάτω ρύθμιση(η έκφρασή του μειώνεται) οδηγεί σε λιγότερες μεταγραφές mRNA. Έτσι, η μέτρηση του επίπεδου mRNA ενός γονιδίου υπό μεταβαλλόμενες συνθήκες είναι ένα κατάλληλο μέσο για τη διερεύνηση της λειτουργίας του εν λόγω γονιδίου. Ο σκοπός των πειραμάτων μικροσυστοιχιών είναι να καθορίσουν ταυτόχρονα τα επίπεδα έκφρασης χιλιάδων γονιδίων ή - ενδεχομένως - ολόκληρο το γονιδίωμα ενός οργανισμού.

Κεφάλαιο 9

Μικροσυστοιχίες γονιδιακής έκφρασης

Παρουσιάζεται επίσης ο τρόπος κατασκευής μια μικροσυστοιχίας Affymetrix καθώς επίσης και της διαδικασίας εξαγωγής των προς ανάλυση δεδομένων.

9.1 Τι είναι οι μικροσυστοιχίες (microarrays;)

Οι μικροσυστοιχίες (microarrays) αποτελούν μια πολύ ενδιαφέρουσα και ακόμα εξελισσόμενη τεχνολογία που έχει γίνει πλέον ένα ευρέως χρησιμοποιούμενο εργαλείο έρευνας εντός των βιολογικών επιστημών . Προσφέρουν ένα εξαιρετικό μέσο συλλογής άνευ προηγουμένου μεγάλων ποσοτήτων δεδομένων γονιδιακής έκφρασης κατά τη διάρκεια ενός μόνο πείραματος. Ωστόσο, προκαλούν επίσης τους βιολόγους, στατιστικολόγους, και επιστήμονες υπολογιστών να αναπτύξουν κατάλληλες τεχνικές και μεθόδους και να επιλύσουν τις εναπομένουσες δυσκολίες.

Ο στόχος των πειραμάτων μικροσυστοιχιών είναι η μέτρηση της αφθονίας των μεταγραφημάτων RNA ενός γονιδίου σε ένα ορισμένο είδος ιστού. Αυτό γίνεται ταυτόχρονα για ένα μεγάλο αριθμό γονιδίων ενός οργανισμού. Προκειμένου να ληφθεί μια τέτοια τιμή γονιδιακής έκφρασης για κάθε υπό έρευνα γονίδιο, ακολουθούνται πολλά διαφορετικά στάδια συμπεριλαμβανομένης της παραγωγής του τσιπ, της προετοιμασίας του δείγματος, της υβριδοποίησης, της

απόκτησης εικόνας, και προεπεξεργασίας. Η τεχνολογία μικροσυστοιχιών αναπτύχθηκε στα μέσα της δεκαετίας 1990-2000 (Schena et al. 1995) και έχει γίνει πλέον ένα ευρέως χρησιμοποιούμενο εργαλείο έρευνας. Μια μικροσυστοιχία DNA είναι κατασκευασμένη από ένα μικρό γυάλινο slide που φέρει έναν μεγάλο αριθμό μορίων DNA (ανιχνευτές) σε διατεταγμένο και συστηματικό τρόπο. Σκοπός του είναι να μετρηθεί η αφθονία των σημασμένων νουκλεϊκών οξέων(DNA, RNA) (στόχοι) που λαμβάνονται από βιολογικά δείγματα. Οι στόχοι εντοπίζουν τους ανιχνευτές που τους αντιστοιχούν και δημιουργούν σταθερό δεσμό υδρογόνου,δηλ. υβριδοποιούνται. Κατά συνέπεια, οι στόχοι μπορούν να ταυτοποιηθούν από τις ετικέτες τους σε συνδυασμό με τη γνωστή θέση των αντίστοιχων ανιχνευτών τους.

9.2 Affymetrix GeneChips

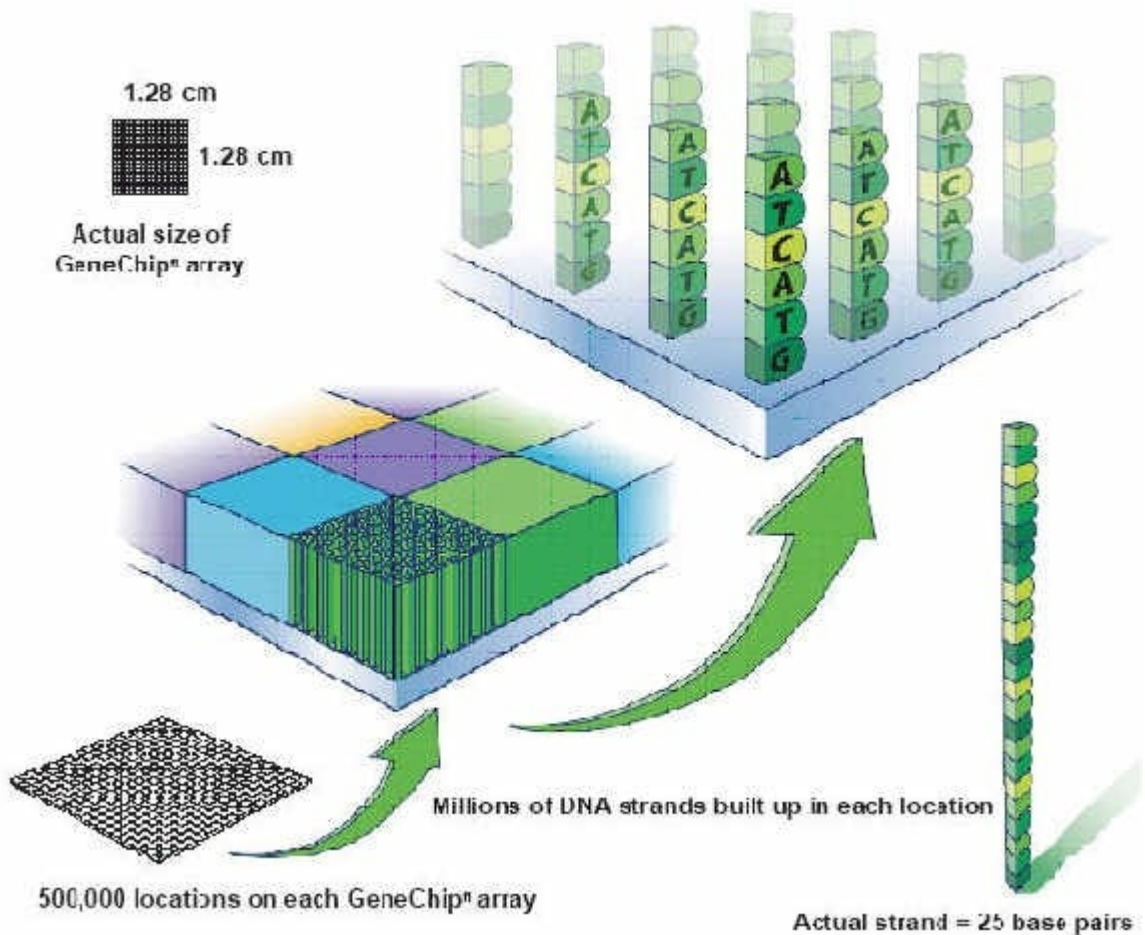
Παρουσιάζουμε εδώ κάποιες λεπτομέρειες της διαδικασίας παρασκευής ενός τσιπ της εταιρίας Affymetrix καθώς θα ασχοληθούμε παρακάτω με μία μικροσυστοιχία τέτοιου πρωτοκόλου. Η παρούσα παράγραφος παρατίθεται για λόγους πληρότητας των πληροφοριών σχετικά με την συγκεκριμένη εταιρία κατασκευής μικροσυστοιχιών και μπορεί να παραληφθεί διότι οι πληροφορίες αυτές είναι πολύ εξειδικευμένες και τεχνικές.

(Πηγή:

<https://www.vsnl.co.uk/software/genstat/htmlhelp/marray/AffymetrixChips.htm>)

Τα Affymetrix GeneChips κατασκευάζονται από τον καθορισμό των αλληλουχιών βάσεων με χρήση ενός συνδυασμού φωτολιθογραφίας που βασίζεται σε ημιαγωγούς και τεχνολογίες χημικής σύνθεσης στερεάς φάσης(χημική μέθοδος κατά την οποία τα μόρια δεσμεύονται σε ένα σφαιρίδιο και συντίθενται βήμα-βήμα σε ένα αντιδρών διάλυμα). Τα ολιγονουκλεοτίδια (oligos), συνήθως 25-μερή (μήκους 25 βάσεων), συντίθενται απευθείας πάνω σε αντικειμενοφόρο πλάκα. Κάθε συστοιχία περιέχει έως 900.000 διαφορετικά ολιγονουκλεοτίδια και κάθε ολιγονουκλεοτίδιο είναι παρόν σε εκατομμύρια αντίγραφα. Δεδομένου ότι οι ολιγονουκλεοτιδικοί ανιχνευτές συντίθενται σε γνωστές θέσεις στη συστοιχία, τα μοτίβα υβριδισμού και οι εντάσεις σήματος μπορούν να ερμηνευθούν

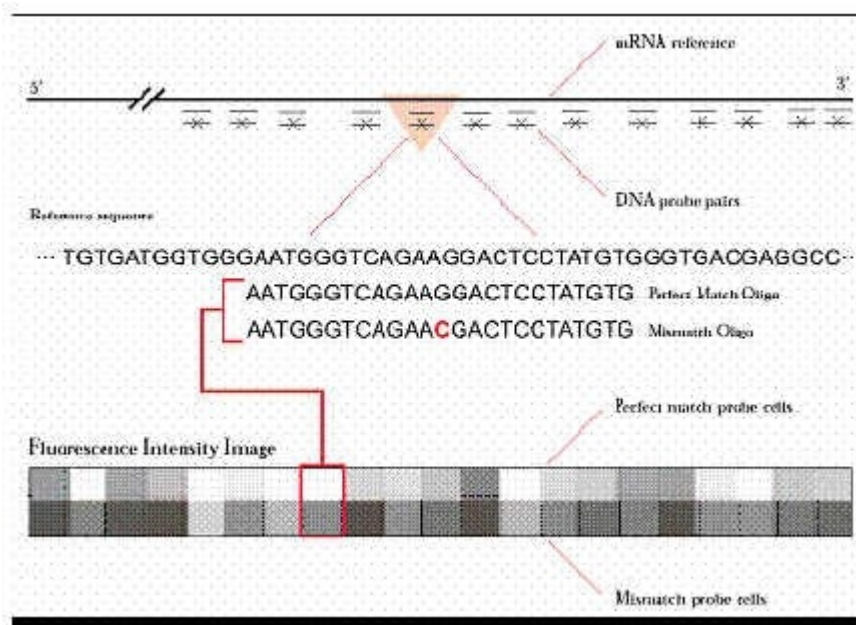
με βάση την ταυτότητα των γονιδίων και των σχετικών επιπέδων έκφρασης με το Λογισμικό Λειτουργίας της Affymetrix GeneChip. Η παρακάτω εικόνα 9.1 δίνει μια σχηματική παράσταση αυτής της διαδικασίας.



Σχήμα 9.1: Διαδικασία κατασκευής μιας μικροσυστοιχίας Affymetrix GeneChip [<https://www.vsnl.co.uk/software/genstat/htmlhelp/marray/AffymetrixChips.htm>]

Κάθε γονίδιο στη μικροσυστοιχία Affymetrix, αντιπροσωπεύεται από μια σειρά διαφορετικών ολιγονουκλεοτιδικών ανιχνευτών. Κάθε ζεύγος ανιχνευτή αποτελείται από μία τέλεια αντιστοιχισμένη ακολουθία ολιγονουκλεοτιδίων (PM perfect match) και από μία ολιγονουκλεοτιδική ανακολουθία (MM MisMatch). Ο PM ανιχνευτής έχει μία ακολουθία ακριβώς συμπληρωματική προς το συγκεκριμένο γονίδιο και έτσι μετρά την έκφραση του γονιδίου. Ο MM ανιχνευτής

διαφέρει από τον PM σε μία μόνο αντικατάσταση βάσης, στην θέση της κεντρικής βάσης(13η θέση), διαταράσσοντας τη σύνδεση του μεταγραφήματος του γονιδίου-στόχου 9.2. Αυτό βοηθά να προσδιοριστεί το υπόβαθρο(που προσθέτει θορυβο στα δεδομένα) και η μη ειδική υβριδοποίηση που συμβάλλει στο σήμα που μετράται για τον PM ανιχνευτή . Ο αλγόριθμος MAS του Λογισμικού Λειτουργίας GeneChip (GeneChip Operating Software) αφαιρεί τις εντάσεις υβριδισμού των MM ανιχνευτών από εκείνες των PM για να προσδιοριστεί η απόλυτη ή συγκεκριμένη τιμή έντασης για κάθε σετ ανιχνευτών. Όλα τα ζεύγη ανιχνευτών ενός γονιδίου συνθέτουν ένα σετ ανιχνευτών που συνήθως περιλαμβάνει 16 ή 20 τέτοια ζεύγη. Οι ανιχνευτές επιλέγονται με βάση τις τρέχουσες πληροφορίες από την Genebank και άλλες αποθήκες νουκλεοτιδίων.



Σχήμα 9.2: Διαδικασία κατασκευής PM και MM ανιχνευτών [<https://www.vsnl.co.uk/software/genstat/htmlhelp/marray/AffymetrixChips.htm>]

Το slide στην συνέχεια υβριδοποιείται με RNA από τον επιλεγμένο στόχο. Μετά την υβριδοποίηση, το τσιπ χρωματίζεται με ένα φθορίζον μόριο (στρεπταβιδίνη-φυκοερυθρίνη) που δεσμεύεται με βιοτίνη. Το πρωτόκολλο χρώσης περιλαμβάνει ένα βήμα ενίσχυσης σήματος που χρησιμοποιεί αντίσωμα κατσίκας αντί-στρεπταβιδίνης και βιοτινυλιωμένο αντίσωμα κατσίκας IgG (Η σειρά των πλύσε-

ων και τους λεκέδες με το προαναφερθέν αντιδραστήρια δεσμεύει την βιοτίνη και παρέχει μία ενισχυμένη πούδρα που εκπέμπει φως όταν το τσιπ στη συνέχεια σαρώνεται με ένα συνεστιακό λέιζερ και καταγράφεται το πρότυπο κατανομής του σήματος στη συστοιχία.

Κεφάλαιο 10

Από τα βιολογικά δείγματα στις ανεπεξέργαστες εντάσεις

10.1 ΠΡΟΕΤΟΙΜΑΣΙΑ ΔΕΙΓΜΑΤΟΣ ΚΑΙ ΣΗΜΑΝΣΗ

1. Οι ανιχνευτές αποτελούνται από βραχέα ολιγονουκλεοτίδια (μικρά τμήματα (DNA/RNA) με συνέπεια, κάθε προς εξέταση γονίδιο να αντιπροσωπεύεται από αρκετούς ανιχνευτές γνωστούς ως σετ ανιχνευτών (probe set).
2. Τρόπος κατασκευής: Τα ολιγονουκλεοτίδια συντίθενται in-situ, είναι δηλαδή κατασκευασμένα απευθείας στην επιφάνεια ενός τσιπ, εφαρμόζοντας φυσικές και χημικές τεχνικές. Ως αποτέλεσμα, οι συστοιχίες παρέχουν πολύ υψηλής πυκνότητας ανιχνευτές και ως εκ τούτου, σημαντικά μεγαλύτερος αριθμός γονιδίων μπορεί να εξεταστεί με μία και μόνο συστοιχία.

10.1.1 Σήμανση

Μετά την κατασκευή (ή την αγορά) ένα συνόλου μικροσυστοιχιών το πείραμα ξεκινά. Το πρώτο βήμα σε ένα τέτοιο πείραμα, φυσικά, είναι να συλλεχθεί το mRNA από τον ιστό που πρέπει να εξεταστεί. Πρόκειται για μια περίπλοκη διαδικασία η οποία περιλαμβάνει την εκχύλιση του RNA και τον διαχωρισμό

mRNA από tRNA και rRNA. Συνήθως, το mRNA στη συνέχεια μεταγράφεται αντίστροφα σε cDNA το οποίο είναι πιο σταθερό από το mRNA. Ως επόμενο βήμα, το ληφθέν cDNA σημαίνεται με μια φθορίζουσα χρωστική ουσία. Αυτό γίνεται με την μεταγραφή πίσω σε RNA που ενσωματώνει σημασμένα νουκλεοτίδια. Υπάρχουν διαθέσιμες διαφορετικές φθορίζουσες χρωστικές, αλλά όλες μοιράζονται την ιδιότητα να είναι σε θέση να συνδεθούν με ορισμένα νουκλεοτίδια. Επιπλέον, μπορούν να απορροφούν το φως ενός ορισμένου μήκους κύματος και στη συνέχεια εκπέμπουν φως ενός συγκεκριμένου, γνωστού μήκους κύματος εντός της ορατής περιοχής του φάσματος.

Στην περίπτωση χρησιμοποίησης ολιγονουκλεοτιδικών συστοιχιών όπως τα τσιπ που κατασκευάζονται από την Affymetrix, RNA από ένα μόνο είδος ιστού εφαρμόζεται σε κάθε τσιπ. Κατά συνέπεια, χρησιμοποιείται μόνο ένας τύπος σήμανσης. Για να συγκριθούν διαφορετικοί ιστοί, τουλάχιστον δύο τσιπ είναι απαραίτητα. Το πρωτόκολλο Affymetrix καθορίζει τον τυχαίο κατακερματισμό του σημασμένου RNA σε κομμάτια των 50 έως 100 βάσεων για τη μείωση των παρεμβολών που οφείλονται στη δευτερεύουσα δομή του στόχου και για την ελαχιστοποίηση πολλαπλών αλληλεπιδράσεων με γειτονικούς ανιχνευτές.

10.1.2 Υβριδοποίηση και καθαρισμός

Το επόμενο βήμα μετά την προετοιμασία και την σήμανση του δείγματος είναι η υβριδοποίηση των ανιχνευτών και των στόχων.

Η υβριδοποίηση είναι μια από τις βασικές έννοιες στην ανάλυση μικροσυστοιχιών. Επιτρέπει στον ερευνητή να διαχωρίζει φυσικά τις διαφορετικές αλληλουχίες RNA που ήταν πρωτύτερα αδιακρίτως μπερδεμένες. Επιπλέον, οι διαφορετικές μεταγραφές τώρα μπορούν να ποσοτικοποιηθούν δεδομένου ότι ο αριθμός των ζευγών ανιχνευτή-στόχου αναμένεται να είναι ανάλογος του συνολικού αριθμού των αντίστοιχων μεταγραφημάτων που υπάρχουν στο δείγμα. Μόλις ολοκληρωθεί η υβριδοποίηση τα τσιπς πλένονται με σκοπό την εξάλειψη περισσευμάτων RNA. Ένας άλλος σκοπός πλύσης είναι η μείωση της διασταυρούμενης υβριδοποίησης. Τα δίκλινα μόρια που δεν ταιριάζουν τέλεια είναι λιγότερο σταθερά και ως εκ τούτου, αναμένεται να σπάσουν κατά τη διάρκεια

της διαδικασίας πλύσης.

Κατά την υβριδοποίηση, μια αλληλουχία-στόχος προσδένεται στην συμπληρωματική της αλληλουχία-ανιχνευτή για την κατασκευή ενός δικλωνου μορίου. Η συστοιχία γεμίζεται με το προετοιμασμένο δείγμα και τοποθετείται σε έναν επωαστήρα για 12 έως 24 ώρες. Ο επωαστήρας εγγυάται τη σωστή θερμοκρασία και άλλες συνθήκες σε κάθε χρονική στιγμή. Η υβριδοποίηση είναι μια αντιστρεπτή διαδικασία και εξαρτάται από πολλές παραμέτρους, π.χ. θερμοκρασία, ανιχνευτής και μήκος του στόχου, το περιεχόμενο γουανίνης και κυτοσίνης (GC-content), συγκέντρωση άλατος, και συγκέντρωση φορμαμιδίου. Μια υψηλότερη θερμοκρασία διευκολύνει τον υβριδισμό όσο είναι κάτω από το σημείο τήξης των διπλών-κλώνων. Τυπικές θερμοκρασίες υβριδοποίησης είναι μεταξύ 45⁰C και 65⁰C. Μακρύτεροι στόχοι και ανιχνευτές έχουν επίσης υψηλότερη τάση υβριδοποίησης λόγω του αυξημένου αριθμού των δεσμών υδρογόνου. Από την άλλη μεριά, οι μικρότεροι ανιχνευτές μειώνουν το ποσό της διασταυρούμενης υβριδοποίησης. Το GC-περιεχόμενο είναι επίσης ένας σημαντικός παράγοντας για την αποτελεσματικότητα υβριδισμού, δεδομένου ότι τα ζευγη G-C περιέχουν τρεις δεσμούς υδρογόνου, ενώ τα ζευγη A-T έχουν μόνο δύο δεσμούς υδρογόνου. Η παρουσία χημικών ουσιών, όπως φορμαμιδίου και άλατος μειώνει τη θερμοκρασία και ελαχιστοποιεί την ηλεκτροστατική άπωση, αντίστοιχα, αυξάνοντας έτσι την αποδοτικότητα του υβριδισμού. Αξίζει να σημειωθεί πως ο υβριδισμός λαμβάνει χώρα ακόμη και αν μερικές βάσεις δεν ταιριάζουν όταν οι κλώνοι είναι αρκετά μεγάλοι. Ωστόσο, αυτές οι ατελείς διπλές έλικες είναι λιγότερο ευσταθείς.

10.1.3 Σάρωση και ανάλυση εικόνας

Τέλος, επιτυγχάνεται μια δισδιάστατη εικόνα της συστοιχίας. Για αυτόν τον σκοπό, η συστοιχία τοποθετείται σε μια ακόμη συσκευή - το σαρωτή (scanner). Ο τελευταίος περιέχει ένα ή περισσότερα λέιζερ και έναν φωτο-πολλαπλασιαστικό σωλήνα (PMT). Το λέιζερ εστιάζεται επάνω σε ένα συγκεκριμένο σημείο της συστοιχίας και διεγείρει τα φθορίζοντα σήματα που εκπέμπουν φως στο κατάλληλο μήκος κύματος. Τώρα, ο PMT ανιχνεύει την ένταση φθορισμού. Όσο υψηλότερη είναι η ένταση τόσο πιο πολύ έχουν υβριδοποιηθεί τα σημασμένα

μεταγραφήματα RNA με τους ανιχνευτές εκείνου του σημείου. Αυτή η διαδικασία επαναλαμβάνεται για κάθε σημείο στη συστοιχία δημιουργώντας μια εικόνα έντασης της συστοιχίας που αποθηκεύεται για περαιτέρω ανάλυση. Μετά τη σάρωση, η εικόνα της συστοιχίας υφίσταται περαιτέρω επεξεργασία, προκειμένου να αποκτηθεί μια ενιαία τιμή έντασης για κάθε γονίδιο ή αλληλουχία ανιχνευτή, αντίστοιχα.

Στην περίπτωση συστοιχιών Affymetrix αυτό γίνεται ως εξής: Πρώτον, ορίζεται ένα πλέγμα με βάση τα χαρακτηριστικά ευθυγράμμισης που έχουν τοποθετηθεί πάνω στο τσιπ από τον κατασκευαστή. Κάθε τμήμα πλέγματος περιέχει πολλά εικονοστοιχεία (pixels) και μπορεί εύκολα να χαρτογραφηθεί με την αντίστοιχη αλληλουχία του ανιχνευτή του, αφού η θέση του και το μέγεθός του είναι γνωστά. Στη συνέχεια, τα οριακά εικονοστοιχεία αφαιρούνται. Τέλος, το 75ο εκατοστημόριο υπολογίζεται για κάθε τμήμα του δικτύου το οποίο αναφέρεται ως πρώτη τιμή έντασης (raw intensity value) για την αντίστοιχη αλληλουχία του ανιχνευτή. Οι τιμές έντασης αποθηκεύονται σε 16 bits. Ως εκ τούτου, κυμαίνονται από 0 έως $2^{16} - 1$. Σημειώνεται ότι αυτές οι τιμές δεν είναι ίδιες με τον αριθμό των μεταγραμμένων RNA που υβριδοποιήθηκαν στο εικονοστοιχείο. Αυτοί οι αριθμοί είναι στην καλύτερη περίπτωση αναλογικοί λαμβάνοντας υπόψη ότι οι τιμές έντασης εξαρτώνται από πολλές παραμέτρους, συμπεριλαμβανομένων των ρυθμίσεων του λέιζερ και της αποτελεσματικότητας της σήμανσης.

10.2 ΠΡΟΕΡΓΑΣΙΑ ΤΩΝ ΠΡΩΤΩΝ ΤΙΜΩΝ ΕΝΤΑΣΗΣ

Ένα βήμα πριν την ανάλυση

Στο παρόν κεφάλαιο περιγράφεται το στάδιο ενός πειράματος μικροσυστοιχιών που ακολουθεί την εξαγωγή των αυθεντικών δεδομένων. Εξηγούμε γιατί είναι απαραίτητο να γίνει κάποια επεξεργασία των δεδομένων πριν αυτά να είναι έτοιμα να δοθούν ως είσοδοι σε οποιονδήποτε αλγόριθμο για την εξαγωγή συμπερασμάτων και περιγράφουμε αυτή τη διαδικασία επεξεργασίας από μαθηματική σκοπιά παρουσιάζοντας κάποιους από τους πιο διαδεδομένους αλγόριθμους που

χρησιμεύουν σε αυτήν την επεξεργασία.

Περίληψη διαδικασίας προεπεξεργασίας δεδομένων

Η προεπεξεργασία χειρίζεται ανεπεξέργαστα δεδομένα που προέκυψαν από το βήμα επεξεργασίας εικόνας και τελικά υπολογίζει μία τιμή έκφρασης για κάθε γονίδιο. Στην περίπτωση συστοιχιών ολιγονουκλεοτιδίων, αυτή υποδιαιρείται σε τέσσερα καθήκοντα:

- **Διόρθωση υποβάθρου (background correction):** προσπαθεί να εξαλείψει τον απροσδιόριστο θόρυβο του περιβάλλοντος και των τοπικών διακυμάνσεων του συνολικού επιπέδου σήματος για τα μεμονωμένα τσιπς
- **Κανονικοποίηση:** διορθώνει τις αλλοιώσεις που προκύπτουν από διαφορετικές ποσότητες RNA, αλλαγές των ρυθμίσεων του σαρωτή και άλλες πηγές που επηρεάζουν τις μετρούμενες εντάσεις σήματος
- (Ειδικά για την επεξεργασία των μικροσυστοιχιών τεχνολογίας Affymetrix) **Διόρθωση PM ανιχνευτών και περίληψη:** βήματα προεπεξεργασίας κατά τα οποία αρκετές μετρήσεις ανά γονίδιο συγχωνεύονται σε μία ενιαία (σχετική) τιμή έκφρασης.

Ένας μεγάλος αριθμός μεθόδων έχει προταθεί για κάθε ένα από τα βήματα προεπεξεργασίας και δεν είναι σαφές ποιά θα πρέπει να εφαρμόζει ο ερευνητής.

10.2.1 Διόρθωση υποβάθρου (background correction)

Τα σήματα των ανιχνευτών μετρούν την αφθονία των ειδικών σημασμένων αλληλουχιών RNA αλλά επηρεάζονται επίσης από εξωτερικούς παράγοντες, όπως ο φθορισμός της ίδιας της επιφάνειας του τσιπ. Οι μέθοδοι διόρθωσης υποβάθρου εκτιμούν την ποσότητα του σήματος του ανιχνευτή που οφείλεται στην επιφάνεια του τσιπ και την αφαιρούν αναλόγως. Δυστυχώς, αυτό δεν είναι δυνατό για τα Affymetrix τσιπς γιατί είναι τόσο πυκνά που δεν υπάρχει χώρος μεταξύ δύο ανιχνευτών. Ως εκ τούτου, το φόντο θα πρέπει να εκτιμηθεί από τα ίδια τα σήματα των ανιχνευτών.

10.2.2 Διόρθωση υποβάθρου RMA (Robust Multi-array Average)

Η RMA είναι μια διαδικασία προεπεξεργασίας μία προϋπόθεση της RMA είναι να χρησιμοποιούν αποκλειστικά PM ανιχνευτές και να αγνοήσει τους ανιχνευτές MM. Η συγκεκριμένη μέθοδος διόρθωσης υποβάθρου βασίζεται στην υπόθεση ότι το παρατηρούμενο σήμα του PM ανιχνευτή (O) αποτελείται από υπόβαθρο ή συνιστώσα θορύβου (N) που ακολουθεί την κανονική κατανομή και ένα «αληθινό» σήμα (S) που ακολουθεί την εκθετική κατανομή. Συγκεκριμένα,

$$O = N + S, N \sim N(\mu, \sigma^2), S \sim Exp(\alpha)$$

Οι παράμετροι α , μ και σ^2 θεωρούνται ίσοι για όλους τους ανιχνευτές MM σε ένα τσιπ και μπορούν συνεπώς να εκτιμηθούν από τα δεδομένα. Δυνητικά αρνητικές τιμές για το N περικόπτονται.

Δεδομένου αυτού του μοντέλου, οι παρατηρούμενες εντάσεις μπορούν να προσαρμοστούν με την αντικατάστασή τους με το αναμενόμενο πραγματικό σήμα $E(s|O = o)$

$$E(s|O = o) = a + \sigma \frac{\frac{\phi(\alpha)}{\sigma} - \frac{\phi(o-\alpha)}{\sigma}}{\frac{\Phi(\alpha)}{\sigma} - \frac{\Phi(o-\alpha)}{\sigma} - 1}$$

όπου $a = o - \mu - \sigma^2\alpha$, ϕ και Φ είναι η συνάρτηση πυκνότητας της τυποποιημένης κανονικής κατανομής και η συνάρτηση κατανομής της τυποποιημένης κανονικής κατανομής, αντίστοιχα. Η διαδικασία αυτή εξασφαλίζει ότι δεν παράγεται καμία αρνητική τιμή έντασης καθώς το S δεν μπορεί να είναι μεγαλύτερο από το O το οποίο αντικατοπτρίζεται στην αναμενόμενη τιμή $E(s|O = o)$. Οι υψηλότερες εντάσεις διορθώνονται απλά αφαιρώντας μια τιμή κοντά στο $\sim M$ ενώ οι χαμηλότερες παρατηρούμενες εντάσεις ρυθμίζονται σε μία τιμή κοντά στο μηδέν.

10.2.3 Μέθοδοι κανονικοποίησης

Γιατί κανονικοποιούμε τα δεδομένα:

Σε πολλές περιπτώσεις οι διαφορές στην επεξεργασία δύο δειγμάτων, ιδιαίτερα κατά τη διάρκεια παραγωγής cDNA, σήμανσης και ιχνηθέτησης επηρεάζουν συστηματικά τα σήματα στις συστοιχίες, στις οποίες μετρώνται οι γονιδιακές εκφράσεις των δειγμάτων. Μερικές συστηματικές διαφοροποιήσεις ανάμεσα στα τσιπς που συχνά επηρεάζουν τα δεδομένα είναι:

- Χρήση διαφορετικών ποσοτήτων RNA.
- Η μία χρωστική ενσωματώνεται καλύτερα από την άλλη (σε συστήματα 2 χρωμάτων).
- Η διαδικασία υβριδοποίησης μπορεί να μεταβεί πιο ολοκληρωμένα σε κατάσταση ισορροπίας στη μία μικροσυστοιχία σε σχέση με την άλλη.
- Οι συνθήκες υβριδοποίησης μπορεί να ποικίλλουν κατά το φυσικό εύρος μια μικροσυστοιχίας.
- Οι ρυθμίσεις των σκάνερς είναι συχνά διαφορετικές.

Με σκοπό την αναγνώριση των πραγματικών βιολογικών διαφορών ανάμεσα στα δείγματα, προσπαθούμε να αντισταθμίσουμε τα συστηματικά τεχνικά λάθη στις μετρήσεις

10.2.4 Ποσοστημοριακή Κανονικοποίηση (Quantile Normalisation)

Η μέθοδος ποσοστημοριακής κανονικοποίησης έχει περιγραφεί από τον B. Bolstad (Bolstad 2001, Bolstad et al. 2003) και είναι επίσης μέρος της διαδικασίας προεπεξεργασίας RMA. Είναι, πιθανώς, η πιο δημοφιλής μέθοδος κανονικοποίησης μεταξύ των ερευνητών, διότι είναι γρήγορη και δεν στηρίζεται σε πολλές υποθέσεις. Η μόνη παραδοχή του Bolstad είναι ότι οι εντάσεις κάθε τσιπ προέρχονται από τον ίδιο υποκείμενο τύπο κατανομής. Η έννοια της ποσοστημοριακής κανονικοποίησης είναι απλή. Βασίζεται στη λογική των QQ-plots όπου

τα ποσοστιαία σημεία (δηλ., οι ταξινομημένες μετρήσεις ή τιμές) από ένα σύνολο δεδομένων X σχεδιάζονται έναντι των ποσοστιαίων σημείων άλλου συνόλου δεδομένων U . Αν τα Q και U προέρχονται από την ίδια κατανομή τότε το QQ-plot τους δείχνει μια γραμμή περίπου κατά μήκος της διαγωνίου. Ωστόσο, στην περίπτωση του επίπεδου έντασης του ανιχνευτή τα ποσοστιαία σημεία των δύο συστοιχιών συνήθως δεν εκτείνονται κατά μήκος της διαγωνίου παρότι οι πραγματικές τιμές έκφρασης ακολουθούν την ίδια κατανομή σε επαναλαμβανόμενα δείγματα. Κάποιος θα μπορούσε να ισχυριστεί ότι οι αντίστοιχες συναρτήσεις κατανομής μεταμορφώθηκαν κατά τη διάρκεια των πειραμάτων μικροσυστοιχιών εξαιτίας τεχνικών λόγων. Για να ανακτήθει μια κοινή κατανομή, μπορούμε απλά να προβάσουμε τα ποσοστιαία σημεία πάνω στη διαγώνιο του QQ-plot. Η προβολή στη διαγώνιο είναι ισοδύναμη με την αντικατάσταση κάθε συνιστώσας ενός ποσοστημοριακού διανύσματος από τη μέση τιμή του εν λόγω διανύσματος. Η εφαρμογή αυτής της ιδέας σε υψηλότερες διαστάσεις (δηλ. περισσότερες από δύο συστοιχίες) είναι απλή.

Η κεντρική ιδέα είναι να ενοποιήσουμε τις εντάσεις όλων των ανιχνευτών ενός τσιπ συγκεντρώνοντας σε ένα σχήμα της τυποποιημένης κανονικής κατανομής, η οποία καθορίζεται συγκεντρώνοντας τις επιμέρους κατανομές εντάσεων των τσιπς. Ο αλγόριθμος αντιστοιχίζει κάθε τιμή οποιουδήποτε τσιπ στο αντίστοιχο τεταρτημόριο της κανονικής κατανομής. Για τον λόγο αυτό η μέθοδος ονομάζεται ποσοστημοριακή κανονικοποίηση. Ο αντίστοιχος τύπος είναι $x_{norm} = F_i^{-1}(F_{ref}(x))$, όπου F_i είναι η συνάρτηση κατανομής των παρατηρήσεων του τσιπ i και F_{ref} είναι η συνάρτηση κατανομής του τσιπ αναφοράς. Αν τα F_i και F_{ref} είναι αρκετά όμοια σε σχήμα, τότε πρακτικά αυτός ο μετασχηματισμός δε διαφέρει πολύ από μια ευθεία. Αυτός ο τύπος κανονικοποίησης ελαττώνει επίσης τον θόρυβο ανάμεσα σε επαναλαμβανόμενες μετρήσεις των ίδιων δειγμάτων.

Πλεονεκτήματα και Μειονεκτήματα της μεθόδου

Το κύριο πλεονέκτημα της μεθόδου ποσοστημοριακής κανονικοποίησης είναι το χαμηλό υπολογιστικό της κόστος. Άλλες μέθοδοι κανονικοποίησης μερικές φορές απαιτούν αρκετές ώρες του χρόνου της CPU, ενώ η ποσοστημοριακή

κανονικοποίηση μπορεί να γίνει μέσα σε λίγα λεπτά. Μια τόσο θετική όσο και αρνητική ιδιότητα της εκατοστημοριακής κανονικοποίησης είναι ότι δεν επιβάλλει κανένα στατιστικό μοντέλο για τα δεδομένα. Θετική επειδή τέτοια μοντέλα μπορεί να μην ταιριάζουν πάντα στα δεδομένα, αρνητική γιατί χωρίς ένα τέτοιο μοντέλο δεν φαίνεται να υπάρχει σωστός τρόπος για να εξηγηθεί ποια ακριβώς είναι η επίδραση της μεθόδου στα δεδομένα. Η υπόθεση μιας κοινής υποκείμενης κατανομής φαίνεται να δικαιολογείται στην περίπτωση των επαναλαμβανόμενων δειγμάτων, αλλά είναι προφανές ότι παραβιάζεται όταν υπάρχουν διαφορικά εκφραζόμενα γονίδια. Στην τελευταία περίπτωση δεν είναι σαφές πόσο καλά αποδίδει η συγκεκριμένη μέθοδος. Το κύριο μειονέκτημα αυτής της προσέγγισης για την κανονικοποίηση είναι η ισχυρή υπόθεση ότι οι κατανομές των εντάσεων των ανιχνευτών είναι πανομοιότυπες (ακόμη και εάν οι επιμέρους ανιχνευτές διαφέρουν ως προς τις θέσεις τους στην κατανομή). Αυτό ισχύει για γονίδια περιορισμένης επάρκειας, και σε μια αρκετά καλή προσέγγιση για τα γονίδια μέτριας επάρκειας, αλλά σίγουρα δεν είναι αλήθεια για τα λίγα γονίδια υψηλής επάρκειας, της οποίας τυπικά επίπεδα ποικίλει αισθητά από δείγμα σε δείγμα.

10.2.5 Διόρθωση PM ανιχνευτών (μόνο για μικροσυστοιχίες Affymetrix)

Κάθε γονίδιο στη μικροσυστοιχία Affymetrix, αντιπροσωπεύεται από μια σειρά διαφορετικών ολιγονουκλεοτιδικών ανιχνευτών. Κάθε ζεύγος ανιχνευτή αποτελείται από μία τέλεια αντιστοιχισμένη ακολουθία ολιγονουκλεοτιδίων (PM perfect match) και από μία ολιγονουκλεοτιδική ανακολουθία (MM MisMatch). Ο PM ανιχνευτής έχει μία ακολουθία ακριβώς συμπληρωματική προς το συγκεκριμένο γονίδιο και έτσι μετρά την έκφραση του γονιδίου. Ο MM ανιχνευτής διαφέρει από τον PM σε μία μόνο αντικατάσταση βάσης, στην θέση της κεντρικής βάσης (13η θέση), διαταράσσοντας τη σύνδεση του μεταγραφήματος του γονιδίου-στόχου. Αυτό βοηθά να προσδιοριστεί το υπόβαθρο (που προσθέτει θόρυβο στα δεδομένα) και η μη ειδική υβριδοποίηση που συμβάλλει στο σήμα που μετράται για τον PM ανιχνευτή. Ο αλγόριθμος MAS του Λογισμικού Λειτουργίας GeneChip (GeneChip Operating Software) αφαιρεί τις εντάσεις υβριδισμού των MM ανιχνευτών από εκείνες των PM για να προσδιοριστεί η

απόλυτη ή συγκεκριμένη τιμή έντασης για κάθε σετ ανιχνευτών.

Όπως προαναφέρθηκε, οι συστοιχίες Affymetrix είναι τόσο πυκνές ώστε οι συνεισφορές κάθε μη ειδικού σήματος, όπως κάποιας μη ειδικής σύνδεσης, ενδεχόμενου διασταυρούμενου υβριδισμού και αυτό-φθορισμού της επιφάνειας πρέπει να εκτιμάται από τις ίδιες τις τιμές έντασης. Για την αντιμετώπιση αυτών των προβλημάτων, η Affymetrix αποφάσισε να συμπληρώσει τους ανιχνευτές PM με τους MM ανιχνευτές. Η Affymetrix αρχικά προσαρμοσε τις εντάσεις PM αφαιρώντας τις αντίστοιχες εντάσεις MM. Δυστυχώς, οι MM ανιχνευτές δεν συμπεριφέρονται όπως αναμενόταν. Για παράδειγμα, Chudin et al. (Chudin et al. 2001) δείχνουν ότι οι ανιχνευτές MM παρέχουν επίσης ένα συγκεκριμένο σήμα. Οι Naef et al. (Naef, Lim, Patil & Magnasco 2002) επισημαίνουν ότι περίπου το 30 τοις εκατό των MM ανιχνευτών σταθερά μετρούν υψηλότερη αφθονία mRNA από τους αντίστοιχους PM ανιχνευτές. Αυτό οδηγεί σε ένα σημαντικό αριθμό αρνητικών προσαρμοσμένων τιμών ανιχνευτή. Τέλος, έχει βρεθεί ότι η ευαισθησία κάθε ανιχνευτή εξαρτάται από το περιεχόμενο γουανίνης και κυτοσίνης (GC-content), ότι οι MM ανιχνευτές είναι πιο ευαίσθητοι από τους αντίστοιχους PM αν η μεσαία βάση τους είναι μια βάση πουρίνης (αδενίνη, γουανίνη), και ότι αυτή η σχέση αντιστρέφεται για βάσεις πυριμιδίνης (κυτοσίνη, θυμίνη) (Binder et al. 2004, Naef & Magnasco 2003, Wu et al. 2004, Zhang et al. 2003). Κατά συνέπεια, μερικοί ερευνητές (π. χ. Ομάδα Speed, UC Berkeley) προτιμούν να αγνοήσουν απλά τους ανιχνευτές MM. Με το MAS 5.0, η Affymetrix άλλαξε τη μέθοδο διόρθωσης των MM. Για να αποφευχθούν αρνητικές τιμές των αισθητήρων, οι ανιχνευτές MM έχουν πλέον αντικατασταθεί από μία ιδανική εκτιμώμενη αναντιστοιχία (IM Ideal Mismatch). Η Affymetrix διακρίνει τρεις περιπτώσεις:

1. Η ένταση του MM ανιχνευτή είναι μικρότερη από την ένταση του αντίστοιχου PM. Σε αυτήν την περίπτωση, $IM = MM$.
2. Η ένταση MM είναι μεγαλύτερη από την αντίστοιχη ένταση του PM αλλά οι περισσότερες άλλες εντάσεις MM από το σετ ανιχνευτών δεν είναι. Σε αυτήν την περίπτωση, η IM εκτιμάται από τους άλλους PM και MM ανιχνευτές από αυτό το σετ ανιχνευτών. Ειδικότερα, η Affymetrix υπολογίζει έναν ισχυρό λόγο μέσων μεταξύ των PM και MM του σετ

ανιχνευτών ο οποίος στη συνέχεια χρησιμοποιείται για τη βαθμονόμηση των PM των οποίων οι MM εντάσεις είναι μεγαλύτερες, προκειμένου να ληφθεί η αντίστοιχη IM .

3. Οι περισσότερες εντάσεις των MM σε ένα σύνολο ανιχνευτών είναι μεγαλύτερες από αυτές των αντίστοιχων PM. Στην περίπτωση αυτή, η τιμή IM προσαρμόζεται σε κάποια τιμή λίγο μικρότερη από του PM. Η προσαρμοσμένη τιμή του ανιχνευτή στη συνέχεια λαμβάνεται με την αφαίρεση της IM από την PM.

10.2.6 Περίληψη δεδομένων

Προκειμένου να καθοριστεί μια ενιαία τιμή έκφρασης για κάθε γονίδιο που αξιολογείται σε ένα τσιπ Affymetrix, όλες οι εντάσεις του αντίστοιχου συνόλου του ανιχνευτή πρέπει να συνοψιστούν. Αρκετές προσεγγίσεις έχουν προταθεί (Irizarry et al. 2003, Li & Wong 2001a). Παρουσιάζουμε εδώ μια από τις πιο συνηθισμένες μεθόδους περίληψης των δεδομένων.

Καθαρισμός διαμέσου (Medianpolish)

Οι Li και Wong (Li & Wong 2001a,b) έδειξαν για πρώτη φορά ότι οι ανιχνευτές ενός σετ ανιχνευτών παρουσιάζουν συστηματικές διαφορές με τους συγγενείς ανιχνευτές τους. Προκειμένου να αιτιολογήσουν το γεγονός αυτό και να διερευνήσουν τις συγκεκριμένες συγγένειες προτείνουν ένα δείκτη έκφρασης που βασίζεται σε κάποιο μοντέλο (MBEI Model- Based Expression Index) που περιλαμβάνει όλα τα δείγματα ενός πειράματος κατά την εκτίμηση μιας τιμής έκφρασης. Το μοντέλο πίσω από τη συνοπτική μέθοδο καθαρισμού διαμέσου είναι παρόμοιο με τον MBEI, δηλ. αντιπροσωπεύει επίσης συγγένειες ειδικών ανιχνευτών, αλλά είναι λιγότερο πολύπλοκη. Η μέθοδος καθαρισμού διαμέσου είναι μέρος της διαδικασίας προεπεξεργασίας RMA. Όπως αναφέρθηκε παραπάνω, οι MM εντάσεις απλά αγνοούνται στο πλαίσιο της RMA. Συνεπώς, μια RMA τιμή έκφρασης καθορίζεται μόνο από PM εντάσεις. Βασίζεται στο ακόλουθο γραμμικό προσθετικό μοντέλο:

$$\log_2(y_{ij}) = a_i + \mu_j + \varepsilon_{ij} \quad (10.1)$$

όπου a_i είναι η επίδραση της συγγένειας ανιχνευτών για τον ανιχνευτή i , $\sum_i a_i = 0$, το μ_j αντιπροσωπεύει το επίπεδο έκφρασης της συστοιχίας j και ε_{ij} είναι ένας όρος ανεξάρτητων σφαλμάτων που προέρχονται από την ίδια κατανομή με μηδενικό μέσο (Irizarry et al. 2003). Το εκτιμώμενο μ_j είναι η ζητούμενη τιμή έκφρασης του αντιπροσωπευτικού σετ ανιχνευτών για τη συστοιχία j . Για να εκτιμηθεί το μ_j εφαρμόζεται ο αλγόριθμος καθαρισμού διαμέσου. Αυτός ο αλγόριθμος παρέχει αξιόπιστες εκτιμήσεις για δύο λόγους. Πρώτον, χρησιμοποιώντας διαμέσους και όχι μέσους όρους καθίσταται λιγότερο ευαίσθητος έναντι ακραίων τιμών, και, δεύτερον, οι εκτιμήσεις γίνονται με βάση το σύνολο των συστοιχιών. Με αυτόν τον τρόπο μπορούμε να 'δανειστούμε' πληροφορίες από άλλα τσιπς. Το μοντέλο που προσαρμόζει ο αλγόριθμος καθαρισμού διαμέσου είναι παρόμοιο με το μοντέλο 10.1,

$$y_{ij} = \mu + \alpha + \beta_j + \varepsilon_{ij}$$

όπου y_{ij} είναι ένας πίνακας, στην περίπτωση μας ο πίνακας ένταση του συγκεκριμένου ανιχνευτή, τέτοιος ώστε το i δείχνει τους ανιχνευτές και το j υποδεικνύει τα τσιπς. Ο πίνακας β_{ij} λαμβάνεται με εναλλάξ αφαίρεση των διαμέσων των γραμμών και στηλών από τα στοιχεία του πίνακα. Ένα διάνυσμα-γραμμή α και ένα διάνυσμα-στήλη \mathbf{b} ενημερώνονται κατά τη διάρκεια της κάθε επανάληψης προσθέτοντας τη διάμεσο των γραμμών και στηλών, αντίστοιχα. Αυτή η διαδικασία επαναλαμβάνεται μέχρις ότου ο πίνακας να αλλάξει κατά λιγότερο από ένα μικρό περιθώριο ή μέχρις ότου ολοκληρωθεί ένας προκαθορισμένος μέγιστος αριθμός επαναλήψεων. Το $\hat{\mu}$ μπορεί τώρα να προσδιοριστεί προσθέτοντας τη διάμεσο του διανύσματος α στη διάμεσο του \mathbf{b} , και οι εκτιμητές $\hat{\alpha}_i$ και $\hat{\beta}_j$ υπολογίζονται ως $\alpha_i - \hat{\mu}$ και $\beta_j - \hat{\mu}$, αντίστοιχα (Ihaka & Gentleman 1996). Οι εκτιμήσεις για τα $\hat{\mu}_j$ στην εξίσωση 10.1 υπολογίζονται προσθέτοντας $\hat{\mu} + \hat{\beta}_j$. Παρατηρήστε ότι οι εκτιμήσεις αυτές εξακολουθούν να είναι στη λογαριθμική κλίμακα.

Κεφάλαιο 11

Ένα πείραμα μικροσυστοιχιών

Εισαγωγή

Στόχος αυτού του κεφαλαίου είναι να αναδείξουμε πως ο αλγόριθμος Isomap αποτελεί ένα ισχυρό μαθηματικό εργαλείο αναγωγής δεδομένων μεγάλης κλίμακας, το οποίο καταφέρνει να ανταποκριθεί σε ένα πρόβλημα γονιδιακής έκφρασης μικροσυστοιχιών Affymetrix, αποκαλύπτοντας δομές βάσει χαρακτηριστικών που του ορίζουμε. Για τον σκοπό αυτό, αναπαρήγαμε ένα μέρος των αποτελεσμάτων των Kevin Dawson, Raymond L Rodriguez και Wasyl Malyj , «Sample phenotype clusters in high-density oligonucleotide microarray data sets are revealed using Isomap, a nonlinear algorithm» (DOI: 10.1186/1471-2105-6-195© Dawson et al; licensee BioMed Central Ltd. 2005). Αρχικά φτιάχνουμε κάποια διαγνωστικά γραφήματα ώστε να αποκτήσουμε μια γενική εικόνα της μορφής των δεδομένων, δηλ. των εντάσεων γονιδιακής έκφρασης όπως έχουν προκύψει από την επεξεργασία μιας μικροσυστοιχίας Affymetrix και στη συνέχεια ακολουθώντας την ενδεικνυόμενη διαδικασία επεξεργασίας των δεδομένων, συνεχίζουμε με γραφήματα των κανονικοποιημένων δεδομένων. Στη συνέχεια περνάμε στη διαδικασία της ανάλυσης με τον αλγόριθμο Isomap όπου περιγράφουμε αναλυτικά τα βήματα που ακολουθήθηκαν για την εξαγωγή των αποτελεσμάτων. Τέλος, οπτικοποιούμε και σχολιάζουμε τα αποτελέσματα.

Περιγραφή των δεδομένων και μεθόδων

167 υψηλής πυκνότητας μικροσυστοιχίες (microarrays) ολιγονουκλεοτιδίων ποντικών (samples) του είδους *Rattus Norvegicus* τύπου U34A με 8.799 γονίδια σε κάθε συστοιχία ήταν προσβάσιμα στο Gene Expression Omnibus (GEO) GSE 464. Το σύνολο δεδομένων περιέχει δείγματα από ποντικούς ελέγχου/ψευδο-εγχειρισμένους (sham-operated) και ποντικούς που υπέστησαν τραυματισμό του νωτιαίου μυελού. Τα δείγματα εξετάζονται ως προς 3 χαρακτηριστικά:

1. **Ως προς τη θέση**, έχουμε τρία (3) διαφορετικά σημεία τραυματισμού:

- ακριβώς στη θέση του σπονδύλου T T9(GEO Accession Number:GDS63),
- πάνω (GEO Accession Number: GDS872) και
- κάτω (GEO Accession Number: GDS896) από αυτόν.

2. **Ως προς τη σοβαρότητα του τραυματισμού**, έχουμε τέσσερις(4) διαβαθμίσεις:

- έλεγχος(καμία βλάβη),
- ήπιος,
- μέτριος,
- σοβαρός τραυματισμός

3. Η τρίτη κατηγορία ταξινόμησης είναι το **χρονικό διάστημα** που μεσολάβησε από τον τραυματισμό έως τη συλλογή του δείγματος. Αυτά τα χρονικά σημεία είναι τα εξής: 0 λεπτά, 30 λεπτά, 4 ώρες, 24 ώρες, 2 ημέρες, 3 ημέρες, 7 ημέρες, 14 ημέρες και 28 ημέρες.

Και τα 167 δείγματα υποβλήθηκαν σε ανάλυση Isomap με έναν ανεπίβλεπτο τρόπο χωρίς a priori γνώση των κλάσεων στις οποίες ανήκει το κάθε δείγμα. Ο Isomap προσαρμόζει μια μη γραμμική πολλαπλότητα στα 167 δείγματα. Αυτή η πολλαπλότητα χρησιμοποιείται για να εκφράσει τις αποστάσεις μεταξύ των

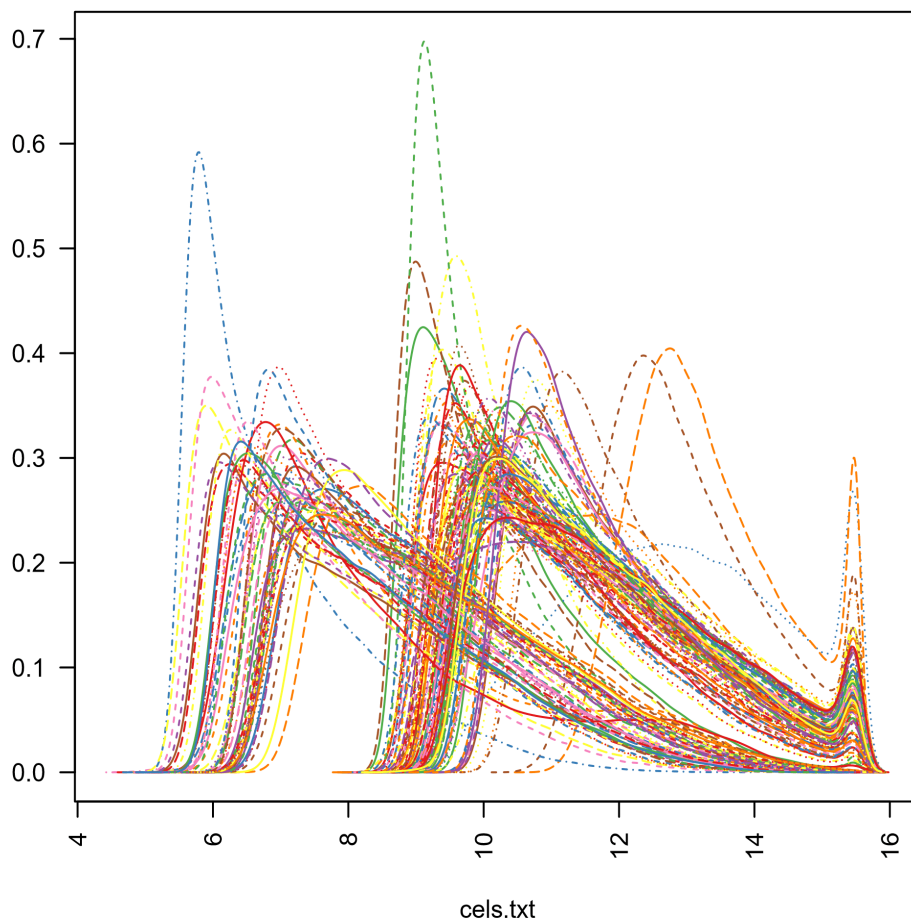
δειγμάτων ως αποστάσεις μονοπατιών (path distances) πάνω στην επιφάνεια της πολλαπλότητας και όχι τις άμεσες Ευκλείδειες αποστάσεις που δε λαμβάνουν υπόψη την ύπαρξή της. Οι αποστάσεις υπολογίστηκαν για όλα τα ζεύγη των 167 δειγμάτων και υποβλήθηκαν σε πολυδιάστατη κλιμάκωση (multi-dimensional scaling). Το αποτέλεσμα αυτής της διαδικασίας παρουσιάζεται στα παρακάτω σχήματα σε ένα τρισδιάστατο σύστημα συντεταγμένων όπως προέκυψαν από την εφαρμογή του αλγορίθμου Isomap στα δεδομένα. Κάθε σφαίρα αντιπροσωπεύει ένα δείγμα. Τα δείγματα ομαδοποιήθηκαν και χρωματίστηκαν σύμφωνα με μία εκ των τριών κύριων ιδιοτήτων: χρόνος, θέση και σοβαρότητα του τραυματισμού.

11.1 ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Διαγνωστικά γραφήματα

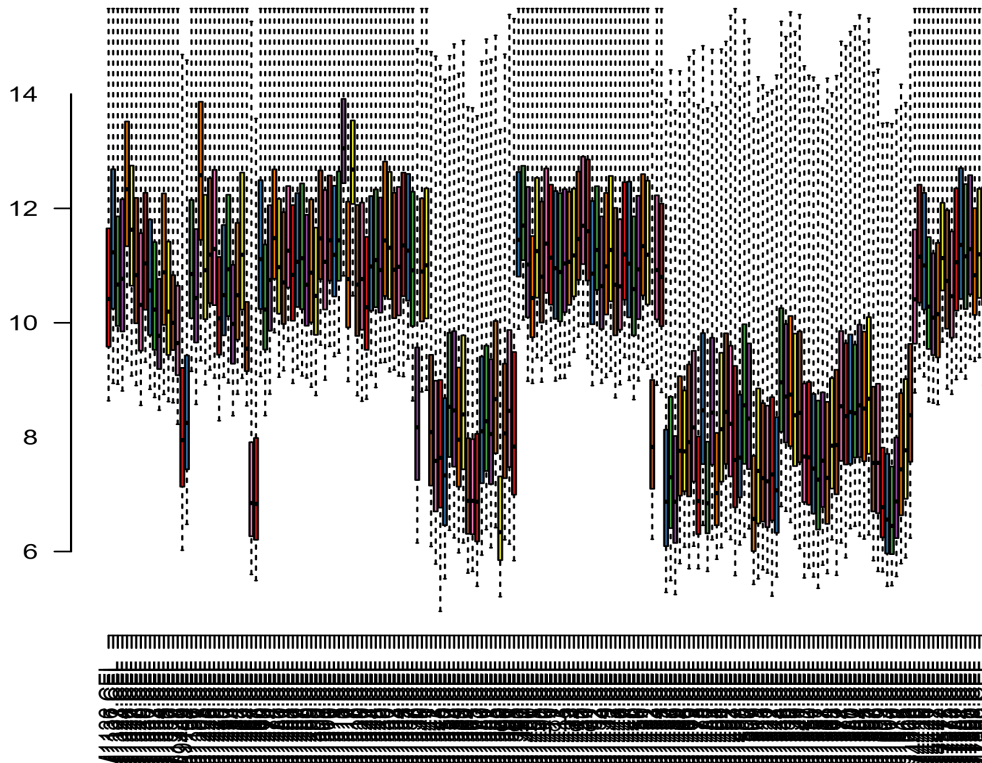
Αρχικά, φτιάχνουμε κάποια διαγνωστικά διαγράμματα ώστε να αποκτήσουμε μία γενική εικόνα για τη φύση των δεδομένων.

- Ιστόγραμμα



Σχήμα 11.1: Ιστόγραμμα πρώτων(ανεπεξεργαστων) τιμών έντασης

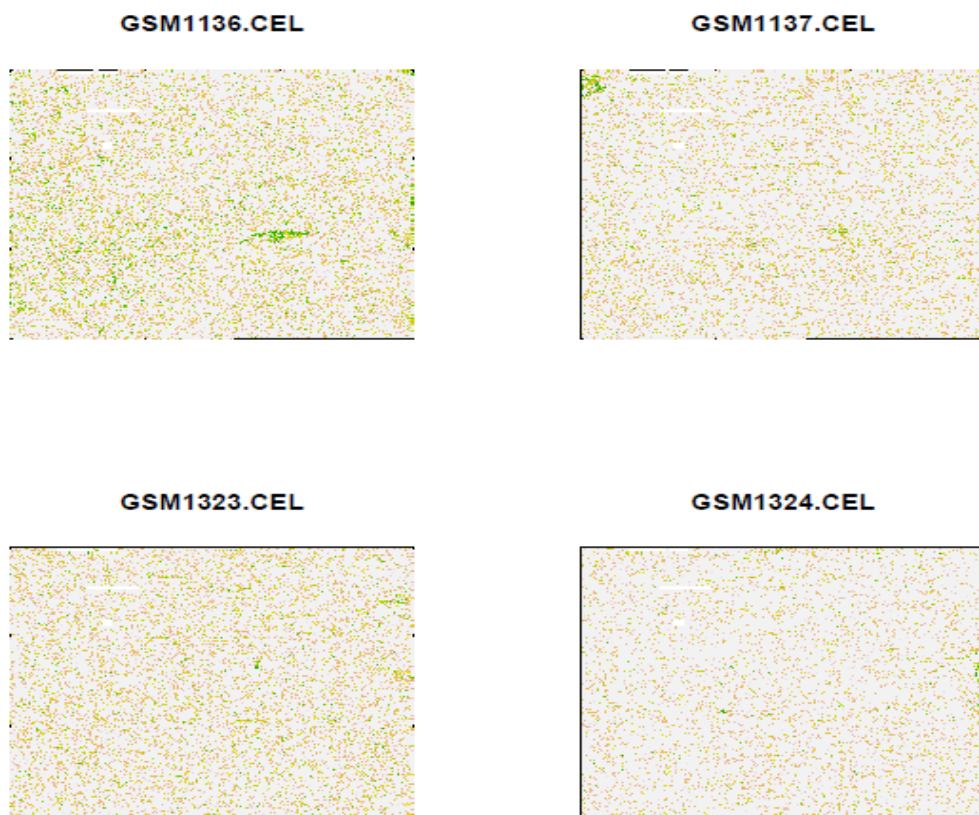
- Θηκόγραμμα (Boxplot)



Σχήμα 11.2: Θηκόγραμμα πρώτων(ανεπεξέργαστων) τιμών έντασης

Παρατηρούμε ότι εν γένει τα δείγματα ακολουθούν πανομοιότυπες κατανομές (που προσιδιάζουν στην κανονική) (εικ. 11.1), (εικ.11.2) ενώ λίγα μόνο από αυτά εμφανίζουν έκτροπες τιμές. Βρίσκοντας, μέσω του παραπάνω θηκογράμματος (εικ. 11.2) ποιά είναι εκείνα τα δείγματα που εμφανίζονται ως έκτροπες τιμές, μπορούμε να δούμε ποια εικόνα παρουσιάζουν τα αντίστοιχα τσιπς. Η αναμενόμενη εικόνα ενός τσιπ είναι η εξής:

- Εικόνες τσιπς που δεν παρουσιάζουν ανομοιομορφίες και περιμένουμε πως δεν επηρεάζονται σημαντικά από θόρυβο



Σχήμα 11.3: Αναμενόμενες εικόνες τσιπς που παρουσιάζουν ομοιομορφία, ένδειξη 'σωστής' έκβασης του πειράματος

Για τα παρακάτω δείγματα, μπορούμε να εικάσουμε ότι οι ανομοιομορφίες που εμφανίζουν μπορεί να προέρχονται από θόρυβο (όπως αναφέρθηκε παραπάνω) και επομένως αντιλαμβανόμαστε πως η διαδικασία της διόρθωσης υποβάθρου είναι ένα πολύ σημαντικό βήμα για να αποκομίσουμε τη δυνατόν περισσότερη και, φυσικά, ακριβέστερη πληροφορία.

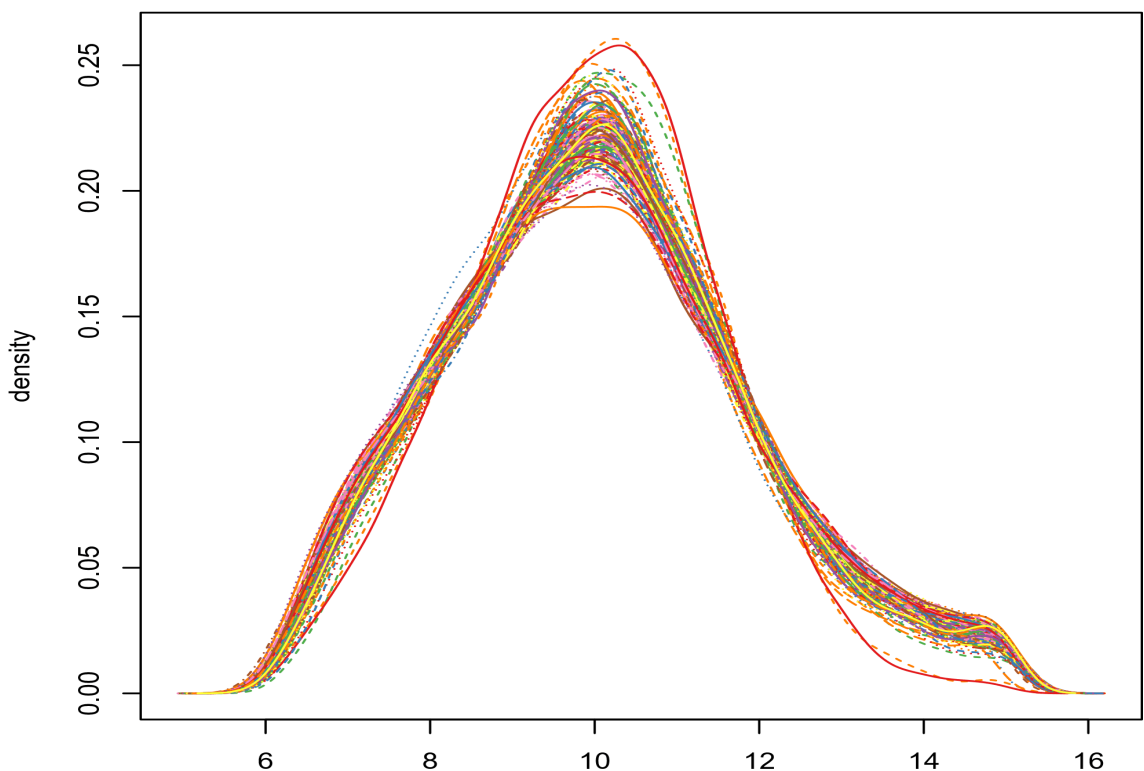


Σχήμα 11.4: Ανομοιώμορφα τσιπς που επιτρέπουν να υποθέσουμε την ύπαρξη θορύβου και άλλων αλλοιώσεων

Ποσοστημοριακή διόρθωση υποβάθρου και RMA κανονικοποίηση των δεδομένων

Με βάση τα παραπάνω, προχωράμε στην διόρθωση υποβάθρου και κανονικοποίηση των δεδομένων με χρήση του αλγορίθμου RMA. Τότε, η εικόνα των παραπάνω διαγραμμάτων έχει ως εξής:

- Ιστόγραμμα κανονικοποιημένων δεδομένων

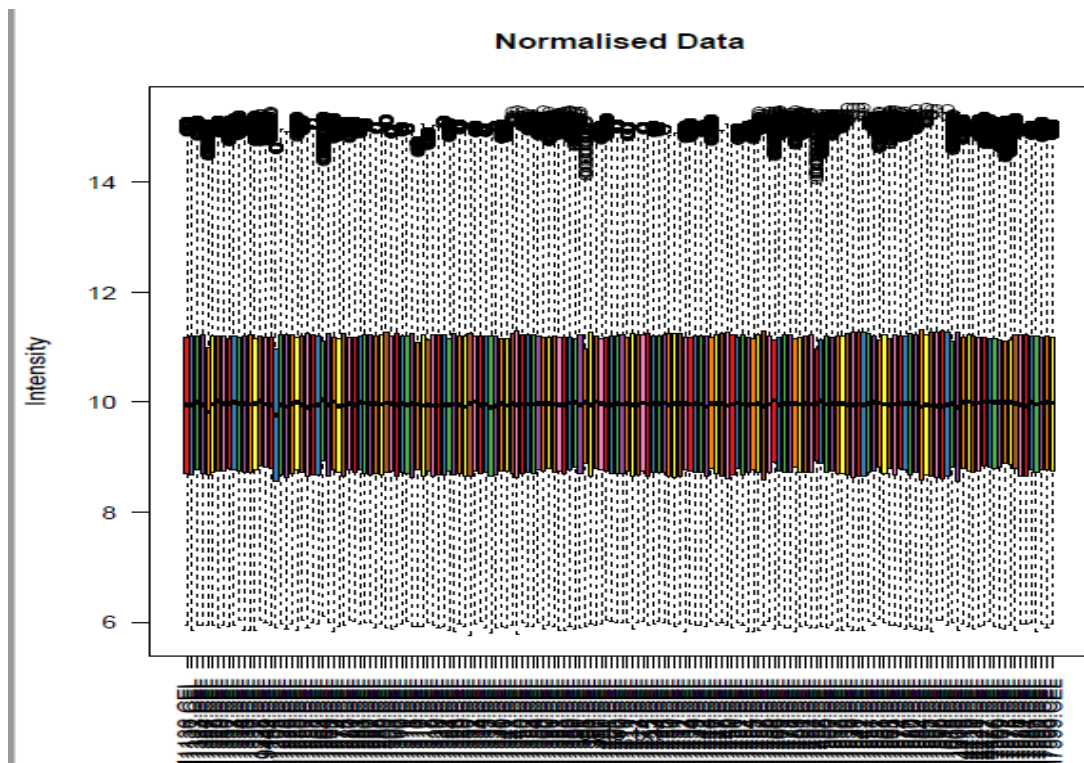


Σχήμα 11.5: Ιστόγραμμα κανονικοποιημένων πρώτων τιμών έντασης

Παρατηρούμε πως πλέον όλα τα δείγματα ακολουθούν προσεγγιστικά την τυποποιημένη κανονική κατανομή και πως ελάχιστα πλέον δείγματα (2-3) εμφανίζουν μεγάλες διαφοροποιήσεις από τα υπόλοιπα (εικ. 11.5), γεγονός που οδηγεί στην πεποίθηση πως κατά βάση τα δεδομένα μας είναι πια έτοιμα να εισαχθούν στον αλγόριθμο για περαιτέρω επεξεργασία, έχοντας απομονώσει τον θόρυβο-στον μέγιστο δυνατό βαθμό- από αυτά.

Αντίστοιχα συμπεράσματα βγάζουμε και από το παρακάτω θηκόγραμμα (εικ. 11.6).

- Θηκόγραμμα κανονικοποιημένων δεδομένων



Σχήμα 11.6: Θηκόγραμμα κανονικοποιημένων πρώτων τιμών έντασης

11.2 ΔΙΑΔΙΚΑΣΙΑ ΑΝΑΛΥΣΗΣ

11.2.1 Εφαρμογή αλγορίθμου Isomap

- Αποθηκεύουμε τις κανονικοποιημένες τιμές της γονιδιακής έκφρασης για τα δείγματα σε έναν πίνακα M με 8799 γραμμές που αντιστοιχούν στα γονίδια και 167 στήλες που αντιστοιχούν στα δείγματα.

The screenshot shows the 'Import' dialog box in Microsoft Excel. The 'Import' tab is active, and the 'VIEW' section is expanded. The 'Column delimiters' are set to 'Tab'. The 'Range' is set to 'A2:GL8800'. The 'Variable Names Row' is set to '1'. The 'Import Selection' button is highlighted. Below the dialog box, a table is visible with the following columns: A (probes), B (Symbols), C (Entrez_IDs), D (GSM1136CEL), E (GSM1137CEL), F (GSM1323CEL), G (GSM1324CEL), H (GSM1326CEL), I (GSM1327CEL), J (GSM1328CEL), and K (GSM1329CEL). The table contains 18 rows of data, with the first row being the header. The second row contains the following values: A01157cdfs_s_at, Lipf, 50682, 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506... The third row contains: A03913cdfs_s_at, Serpine2, 29366, 11.4175149..., 11.6136577..., 10.9870676..., 12.079, 2.07819506541226, Converted ToType: NUMBER, Value: 2.07819506... The fourth row contains: A04674cdfs_s_at, NA, 24867, 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506... The fifth row contains: A07543cdfs_s_at, Smr3b, 24867, 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506... The sixth row contains: A09811cdfs_s_at, Igfbp2, 25662, 12.2653573..., 12.1169493..., 12.2750666..., 12.3703840..., 11.4736896..., 11.4959274..., 12.0147748..., 12.6... The seventh row contains: A16585cdfs_s_at, Rln1, 25616, 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506... The eighth row contains: A17753cdfs_s_at, Drd3, 29238, 2.07857814..., 2.07857814..., 2.07857814..., 2.07819506..., 2.07857814..., 2.07857814..., 2.07857814..., 2.07857814..., 2.07... The ninth row contains: A30543cdfs_s_at, Cd44, 25406, 2.46810131..., 2.29973248..., 2.18499079..., 2.18499079..., 5.82709645..., 4.93153935..., 2.18499079..., 2.56... The tenth row contains: A44407cdfs_at, Ephb3, 287989, 2.13595620..., 2.13595620..., 2.35619290..., 2.13595620..., 2.13595620..., 2.13595620..., 2.13595620..., 2.68... The eleventh row contains: AA108277_at, Hsph1, 288444, 2.97333673..., 5.67455880..., 6.75809798..., 6.51868111..., 2.94902152..., 3.42945632..., 4.49341167..., 2.39... The twelfth row contains: AA108308_i_at, Mdm2, 314856, 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506..., 2.07819506... The thirteenth row contains: AA108308_s_at, Mdm2, 314856, 5.46792529..., 5.22132369..., 5.72930359..., 4.97875319..., 4.89756041..., 4.89756041..., 4.89771132..., 5.08... The fourteenth row contains: AA684537_at, Ndufb5, 294964, 11.0438745..., 10.5020049..., 10.8990815..., 10.9261772..., 11.0153274..., 11.0393065..., 10.7178659..., 11.0... The fifteenth row contains: AA684929_f_at, Ndufa4l2, 100362331, 4.22000246..., 3.63021472..., 3.63021472..., 3.63021472..., 2.48408026..., 3.59858107..., 3.63021472..., 3.63... The sixteenth row contains: AA684960_f_at, Ndufa4l2, 100362331, 5.3283866..., 5.3283866..., 5.56550924..., 5.33516849..., 4.97154720..., 5.26694116..., 5.44711967..., 5.44... The seventeenth row contains: AA684963_at, Fkbp2, 293702, 13.0755023..., 12.5183780..., 13.9958609..., 13.4732877..., 11.3246276..., 11.5682567..., 12.6427843..., 12.4... The eighteenth row contains: AA685112_at, Ndufs8, 293652, 8.11288111..., 7.43370062..., 7.74154971..., 7.97546920..., 4.13778924..., 4.13778924..., 7.95472997..., 7.81...

Σχήμα 11.7: Μορφή αρχείου δεδομένων όπως αυτά εισήχθησαν το Matlab για την ανάλυσή τους με εφαρμογή του αλγορίθμου Isomap

- Υπολογίσαμε τις Ευκλείδειες αποστάσεις μεταξύ των δειγμάτων με χρήση της συνάρτησης
 - 1 $D = L2_distance(M, M)$;
- Ορίσαμε στον αλγόριθμο Isomap να κάνει ανάλυση για διαστάσεις από 1 έως 20 αποθηκεύοντας αυτό το εύρος στην μεταβλητή dims:

```
1 options.dims=1:20;
```

– Εφαρμόζουμε τον αλγόριθμο Isomap χρησιμοποιώντας τη μέθοδο των k - πλησιέστερων γειτόνων με παραμέτρους D (πίνακας Ευκλίδειων αποστάσεων), $k = 5$ (πλήθος προς εύρεση πλησιέστερων γειτόνων), options (χρησιμοποιείται η μεταβλητή `dims=1 : 20` που ορίσαμε προτύτερα)

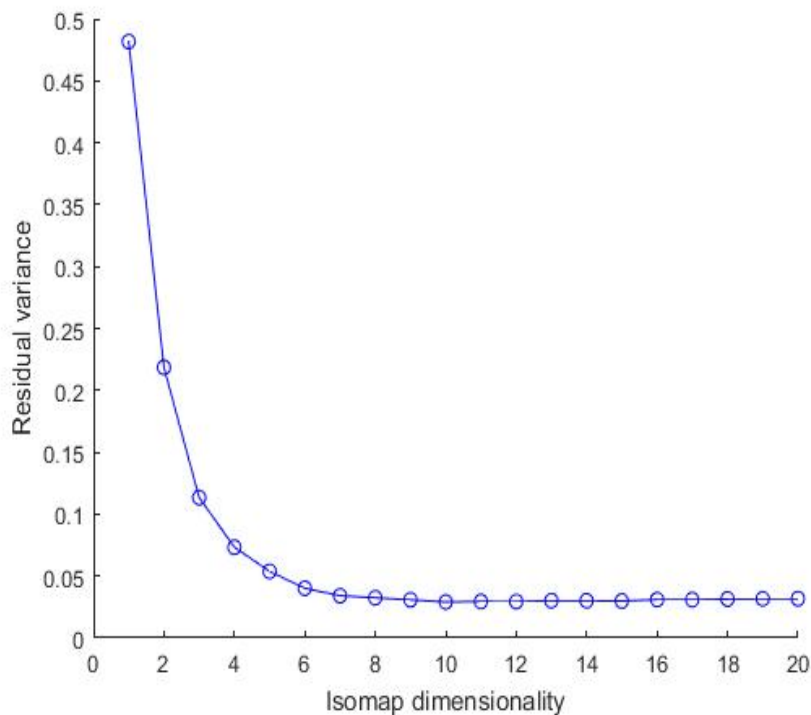
```
1 [Y, R] = Isomap(D, 'k',5, options);
```

Οι συντεταγμένες που υπολόγισε ο Isomap αποθηκεύτηκαν στο διάνυσμα `Y.coords` ενώ οι διαχυμάνσεις υπολοίπων στο διάνυσμα `R`.

11.2.2 Αποτελέσματα εφαρμογής αλγορίθμου Isomap

Παρακάτω παρουσιάζουμε γραφικά τα αποτελέσματα του αλγορίθμου Isomap αποκτώντας έτσι μια πρώτη εικόνα για την μορφή της εμφύτευσης των δεδομένων σε ένα δισδιάστατο χώρο.

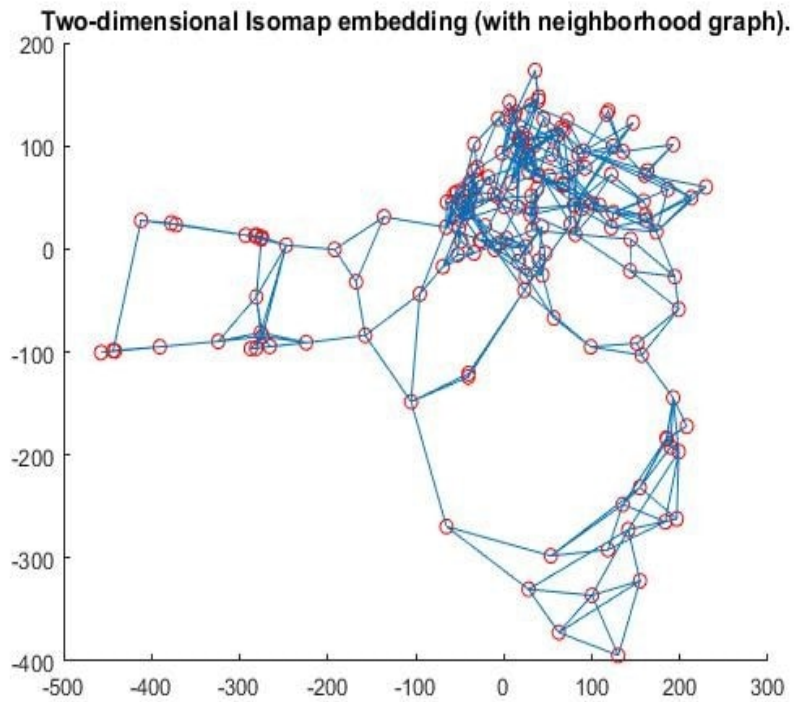
- Διάγραμμα Διαστατικότητας- Διακύμανσης υπολοίπων



Σχήμα 11.8: Διάγραμμα διαστάσεων της εμφύτευσης που παρήγαγε ο Isomap ως προς τη διακύμανση υπολοίπων όπως υπολογίστηκε από τον αλγόριθμο σε κάθε βήμα της διαδικασίας

Στο παραπάνω γράφημα (εικ. 11.8) παρατηρούμε ότι όσο αυξάνεται η διάσταση, τόσο μειώνεται η διαφορά υπολοίπων των διασπορών ανάμεσα στα μοντέλα που προκύπτουν από την εφαρμογή του αλγορίθμου. Επομένως, γίνεται σαφές ότι έχει νόημα να αναζητούμε ένα μοντέλο μέχρι διάστασης 4, εφόσον από εκεί και πέρα δεν κερδίζουμε σημαντικές πληροφορίες από την προσθήκη περαιτέρω μεταβλητών στην ανάλυσή μας.

- Γράφος γειτνίασης



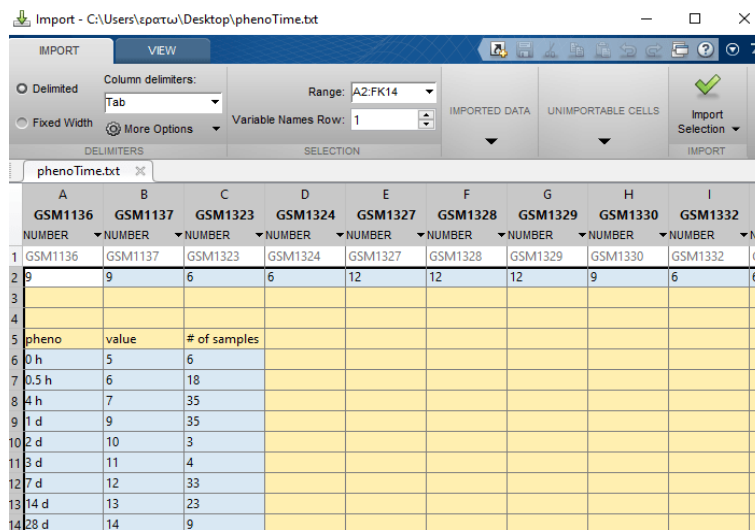
Σχήμα 11.9: Γράφος γειτνίασης όπως υπολογίστηκε από τον αλγόριθμο Isomap

Ο γράφος γειτνίασης (εικ. 11.9) που προέκυψε από την εφαρμογή του αλγορίθμου μας δίνει μία πρώτη εικόνα για τον τρόπο με τον οποίο διατάσσονται τα δείγματα των δεδομένων με βάσει κάποια κοινά τους χαρακτηριστικά.

11.2.3 Οπτικοποίηση αποτελεσμάτων

Inputs

- Αποθηκεύουμε σε ένα διάνυσμα (samples) τα ονόματα των δειγμάτων.
- Δημιουργούμε 3 διανύσματα (phenoTime, phenoSeverity, phenoLocation) που περιλαμβάνουν τους 3 διαφορετικούς φαινότυπους ως προς τους οποίους γίνεται η ανάλυση, δηλαδή χρόνος, σοβαρότητα, θέση. Οι φαινότυποι του κάθε δείγματος παρέχονται μαζί με άλλες πληροφορίες από την ιστοσελίδα <https://www.ncbi.nlm.nih.gov/geo/>.



	A	B	C	D	E	F	G	H	I	J
	GSM1136	GSM1137	GSM1323	GSM1324	GSM1327	GSM1328	GSM1329	GSM1330	GSM1332	GSM1332
1	GSM1136	GSM1137	GSM1323	GSM1324	GSM1327	GSM1328	GSM1329	GSM1330	GSM1332	GSM1332
2	9	9	6	6	12	12	12	9	6	6
3										
4										
5	pheno	value	# of samples							
6	0 h	5	6							
7	0.5 h	6	18							
8	4 h	7	35							
9	1 d	9	35							
10	2 d	10	3							
11	3 d	11	4							
12	7 d	12	33							
13	14 d	13	23							
14	28 d	14	9							

Σχήμα 11.10: Μορφή αρχείου που περιλαμβάνει μία κωδικοποίηση των φαινοτύπων με βάση το χρονικό διάστημα μετά τον τραυματισμό κατά το οποίο έγινε η μέτρηση, δηλ. 0 ώρες, 0,5 ώρες, 4 ώρες, 1 ημέρα, 2 μέρες, 3 μέρες, 7 μέρες, 14 μέρες και 28 μέρες μετά τον τραυματισμό.

	A	B	C	D	E	F	G	H	I
	GSM1136	GSM1137	GSM1323	GSM1324	GSM1327	GSM1328	GSM1329	GSM1330	GSM1332
TEXT	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER
1	GSM1136	GSM1137	GSM1323	GSM1324	GSM1327	GSM1328	GSM1329	GSM1330	GSM1332
2	2	2	2	2	2	2	2	1	1
3									
4									
5	pheno	value	# of samples						
6	control								
7	sham-oper...	1	39						
8	mild	2	60						
9	moderate	3	23						
10	severe	4	45						

Σχήμα 11.11: Μορφή αρχείου που περιλαμβάνει μία κωδικοποίηση των φαινοτύπων με βάση τη σοβαρότητα του τραυματισμού, δηλ. δείγμα ελέγχου (καθόλου τραυματισμός), ήπιος, μέτριος και σοβαρός τραυματισμός.

	A	B	C	D	E	F	G	H	I
	GSM1136	GSM1137	GSM1323	GSM1324	GSM1327	GSM1328	GSM1329	GSM1330	GSM1332
TEXT	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER
1	GSM1136	GSM1137	GSM1323	GSM1324	GSM1327	GSM1328	GSM1329	GSM1330	GSM1332
2	17	17	17	17	17	17	17	17	17
3									
4									
5									
6	pheno	value	# of samples						
7	AT	15	24						
8	BELOW	16	45						
9	ABOVE	17	98						

Σχήμα 11.12: Μορφή αρχείου που περιλαμβάνει μία κωδικοποίηση των φαινοτύπων με βάση την ανατομική θέση του τραυματισμού, δηλ. πάνω, κάτω και ακριβώς στη θέση του σπονδύλου T9. Για τα δείγματα ελέγχου (καθόλου τραυματισμός, θέση του τραυματισμού θεωρείται ακριβώς ο σπόνδυλος T9, όπως προκύπτει από τον σχολιασμό που παρέχει η Affymetrix για τα δείγματα.

Έτσι μπορούμε να δούμε σε κάθε ένα από αυτά τα διανύσματα τον αντιστοιχο φαινότυπο για το κάθε δείγμα. Από τις παρακάτω εικόνες μπορούμε να δούμε ποια είναι η μορφή των διανυσμάτων αυτών. Για παράδειγμα, το δείγμα GSM1136(1η στήλη σε κάθε αρχείο φαινοτύπων) μπορούμε να δούμε πως προήλθε από δείγμα που ελήφθη από ποντικό

- 1 ημέρα μετά τον τραυματισμό (τιμή *phenoTime* = 9)
- Αφορά σε ήπιο τραυματισμό (τιμή *phenoSeverity* = 2)
- Σε θέση κάτω από το σημείο τραυματισμού (τιμή *phenoLocation* = 17)

Στα συγκεκριμένα αρχεία γίνεται μία κωδικοποίηση, όπως φαίνεται και στο τέλος του αρχείου που συνίσταται στην αντιστοίχιση συγκεκριμένων ακεραίων σε κάθε ένα από τα χαρακτηριστικά. Έχει επίσης σημειωθεί το πλήθος των δειγμάτων που αντιστοιχεί σε κάθε φαινότυπο. Αυτό θα αποτελέσει έναν ακόμη έλεγχο της ορθότητας του αλγορίθμου Isomap, όταν θα οπτικοποιήσουμε τα δεδομένα, οπότε και θα είναι εμφανές το πλήθος των δειγμάτων κάθε κατηγορίας.

Στη συνέχεια, συνοψίζουμε τα δεδομένα σε μία δομή μονέλου (Model structure) ώστε να επιτύχουμε μια τρισδιάστατη απεικόνιση των ομαδοποιήσεων που έχει επιτύχει ο Isomap, χρησιμοποιώντας τη συνάρτηση `Makemodel`. Έτσι έχουμε τρία μοντέλα, ένα για κάθε κατηγορία φαινότυπου.

- 1 `Model1 = makemodel(M,Y, options , samples , phenoTime , 'i', 3,5);`
- 2 `Model2 = makemodel(M,Y, options , samples , phenoSeverity , 'i', 3,5);`
- 3 `Model3 = makemodel(M,Y, options , samples , phenoLocation , 'i', 3,5);`

οπου τα

- `M,Y(Y.coords)`, `options(options.dims)`, `samples`, `phenoTime`, `phenoSeverity`, `phenoLocation` όπως ορίστηκαν νωρίτερα και

- 'i' υποδεικνύει το είδος του αλγορίθμου που χρησιμοποιείται (Isomap),
- 3 υποδεικνύει ότι το μοντέλο μας είναι τρισδιάστατο και
- 5 το πλήθος των πλησιέστερων γειτόνων που ορίσαμε στον αλγόριθμο να αναζητήσει.

Χρησιμοποιούμε τη συνάρτηση `showmodel` για να οπτικοποιήσουμε τα μοντέλα που προέκυψαν από τα παραπάνω

```
1 showmodel(model, dim, p, q, maxrad),
```

```
,
```

όπου `model` τα μοντέλα `model1`, `model2`, `model3` που αναφέρονται στους φαινότυπους για τον χρόνο, τη σοβαρότητα και τη θέση αντίστοιχα, `dim` η διάσταση στην οποία θέλουμε να αναπαράστήσουμε γραφικά τα δεδομένα,

`p` οι διάμετροι των σφαιρών (προκαθορίζεται στο 0.95),

Η διάμετρος μιας σφαίρας ορίζεται (εδώ) ως το 95-εκατοστημόριο της κατανομής των αποστάσεων από τον πλησιέστερο γείτονα μέσα σε μια κλάση.

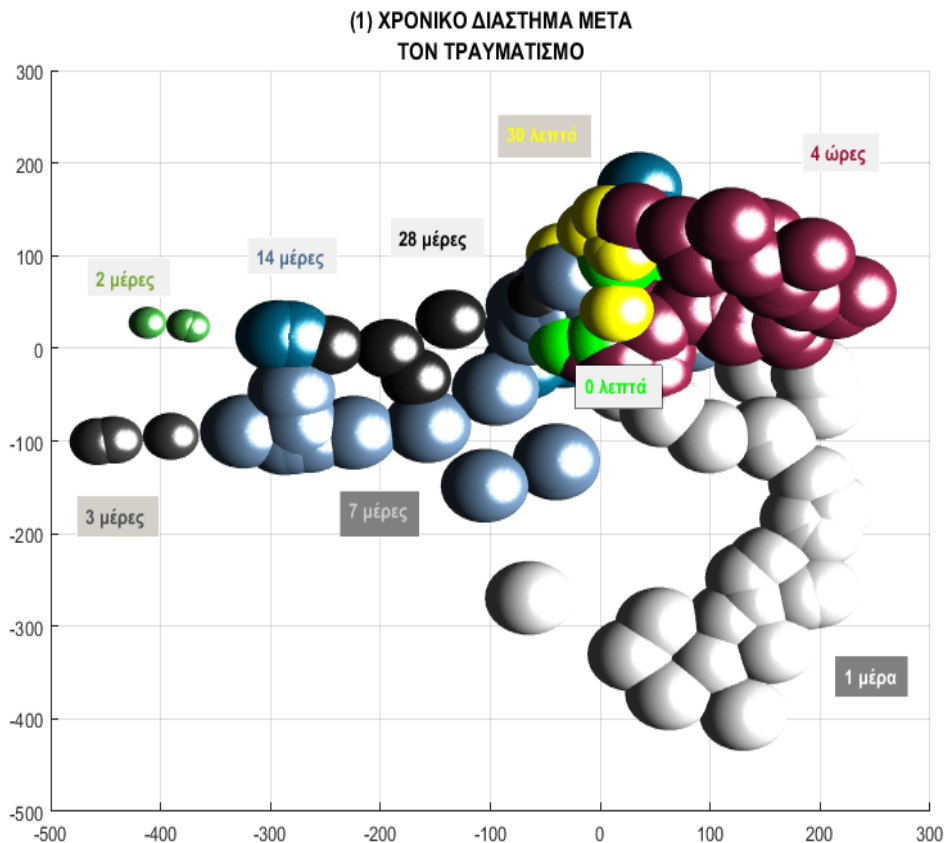
Οι διάμετροι των σφαιρών εκφράζουν τη συμπαγεια της ομάδας: **όσο πιο συμπαγής είναι η ομάδα, τόσο μικρότερη είναι η διάμετρος της σφαίρας. Αν τα μέλη μιας ομάδας εκτείνονται σε μεγαλύτερο χώρο, τότε εμφανίζονται σφαίρες μεγαλύτερης διαμέτρου.** `q`: η ποιότητα της εικόνας (προκαθορίζεται στο 50). `maxrad`: μέγιστη διάμετρος σφαίρας

Το μοντέλο που προσαρμόζεται από τον Isomap ομαδοποιεί επιτυχώς τα όμοια δείγματα σε ομάδες που βρίσκονται σε σαφώς καθορισμένες θέσεις του τρισδιάστατου μοντέλου. Τα λιγότερο επηρεασμένα ψευδοεγχειρισμένα (*sham operated*) δείγματα φαίνονται στον κεντρικό εσωτερικό πυρήνα του μοντέλου. Αντιθέτως, τα πιο επηρεασμένα δείγματα, δηλαδή στις 24 ώρες μετά από μέτριο ως σοβαρό τραυματισμό(η πιο ενεργή φάση του τραυματισμού του νωτιαίου μυελού) εμφανίζονται στα περιφερειακά σημεία του μοντέλου.

11.3 ΑΠΟΤΕΛΕΣΜΑΤΑ

Σε αυτήν την ενότητα παρουσιάζουμε γραφικά την ομαδοποίηση που έκανε ο Isomap για κάθε χαρακτηριστικό(φαινότυπο). Με άλλα λόγια, οι εντολές που χρησιμοποιούμε στο Matlab, χρησιμοποιούν-σε επίπεδο γραφικής απεικόνισης-τον γράφο γειτνίασης(ομαδοποίηση)που παράγαγε ο αλγόριθμος, αποδίδοντας σε κάθε σημείο το αντίστοιχο προς εξέταση χαρακτηριστικό, το οποίο είναι χρωματισμένο με διαφορετικό χρώμα. Έτσι βλέπουμε εν τέλει με σαφήνεια τον βαθμό στον οποίο έχει καταφέρει ο Isomap να ομαδοποιήσει σωστά τα δεδομένα αυτά. Αυτό σημαίνει πως κάθε γειτονιά του γράφου πρέπει να είναι τέτοια ώστε τα κοινά χαρακτηριστικά για κάθε έναν από τους φαινότυπους να είναι συγχροτημένα σε ομάδες ίδιου χρώματος κατά μήκος του γραφήματος. Και αυτό συμβαίνει, αποδεικνύοντας της αποτελεσματικότητα του Isomap στην ανάλυση βιολογικών δεδομένων μικροσυστοιχιών.

I Γραφική απεικόνιση της ομαδοποίησης του Isomap ως προς το **χρο- νικό διάστημα** που μεσολάβησε μεταξύ τραυματισμού και μέτρη- σης.

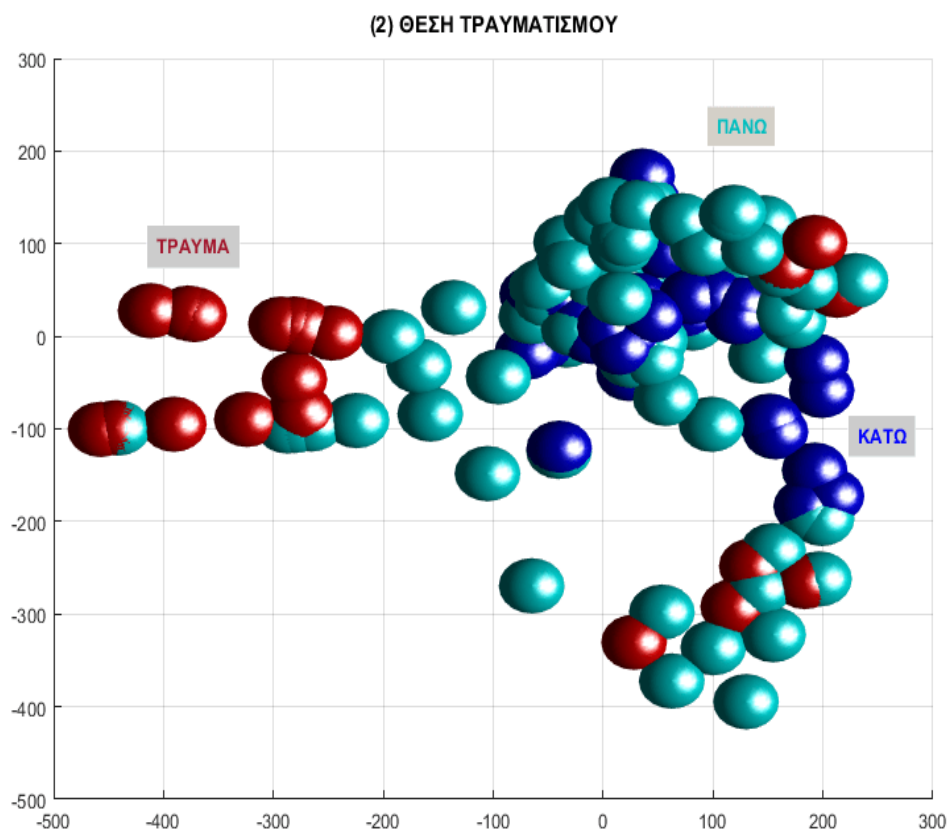


Σχήμα 11.13: Ανάλυση Isomap ως προς τον φαινότυπο του χρόνου που με- σολάβησε μεταξύ τραυματισμού και μέτρησης(χρόνος) για 167 συστοιχίες ολι- γονουκλεοτιδίων υψηλής πυκνότητας που προέκυψαν από ποντίκια όπως περι- γράφεται στην αρχή του κεφαλαίου.

Εικ. 11.13: Δείχνει έναν ξεκάθαρα χρονικά εξαρτώμενο διαχωρισμό των δειγμάτων κατά μήκος του άξονα x. Η δεξιά πλευρά του μον- τέλου δείχνει δείγματα στα 30 λεπτά, 4 ώρες, 1 μέρα μετά τον τραυ- ματισμό. Τα νωρίτερα συλλεγμένα δείγματα βρίσκονται στο κέντρο τού μοντέλου. Αντίθετα, δείγματα από μεταγενέστερες χρονικές

στιγμές εντοπίζονται σε πιο απομακρυσμένες θέσεις στη δεξιά πλευρά του μοντέλου. Ο Isomap διαχωρίζει αποτελεσματικά την πρώτερη ενεργή φάση της μετατραυματικής καταστροφής από την ύστερη φάση των 48 ωρών-28 ημερών. Δείγματα από την τελευταία φάση αναγέννησης (2, 3, 7, 14, 28 μέρες) φαίνονται στην αριστερή πλευρά του μοντέλου. Κατά τη διάρκεια της αναγέννησης, τα νωρίτερα δείγματα βρίσκονται σε πιο απομακρυσμένες θέσεις και τα πιο μεταγενέστερα σε πιο κεντρικές θέσεις του μοντέλου. Είναι αξιοπρόσεκτο το γεγονός ότι κάποια δείγματα βρίσκονται στην περιφέρεια του μοντέλου ακόμα και 28 μέρες μετά τον τραυματισμό. Αυτό μπορεί να αποδοθεί σε μη ολοκληρωμένη αναγέννηση και μόνιμη βλάβη που προκλήθηκε από έναν πιο σοβαρό τραυματισμό του νωτιαίου μυελού. Το δεξί μέρος του μοντέλου περιλαμβάνει δείγματα από την πρώτη φάση της μετατραυματικής ζημιάς (30 λεπτά-24 ώρες). Σε αυτά τα δείγματα τα κυρίαρχα γεγονότα είναι ο τραυματισμός, οι δευτερογενείς βιοχημικές μεταβολές και οι διαδικασίες αυτοκαταστροφής. Από την άλλη μεριά, το αριστερό μέρος του μοντέλου περιλαμβάνει μεταγενέστερα δείγματα (48 ώρες-28 μέρες) όπου νευροπροστατευτικά φαινόμενα ανάρρωσης αντιπαρέρχονται τα προηγούμενα.

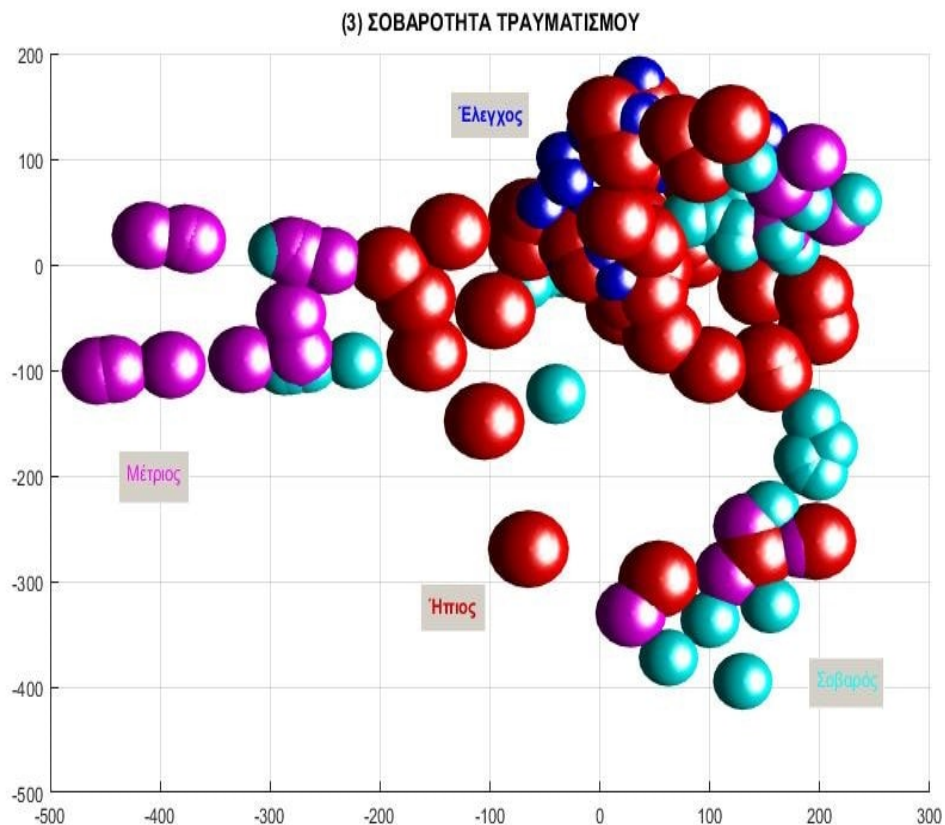
II Γραφική απεικόνιση της ομαδοποίησης του Isomap ως προς τη θέση του τραυματισμού.



Σχήμα 11.14: Ανάλυση Isomap ως προς τον φαινότυπο της θέσης του τραυματισμού για 167 συστοιχίες ολιγονουκλεοτιδίων υψηλής πυκνότητας που προέκυψαν από ποντίκια όπως περιγράφεται στην αρχή του κεφαλαίου.

Εικ. 11.14: Δείχνει πως τα δείγματα που προέρχονται από τα σημεία πάνω και κάτω από τον τραυματισμό βρίσκονται στον κεντρικό πυρήνα του μοντέλου. Παράλληλα, τα δείγματα από την ίδια τη θέση τού τραυματισμού εμφανίζονται στα περιφερειακά σημεία τού μοντέλου. Δεν υπάρχει εμφανής διαχωρισμός ανάμεσα στα μη επηρεασμένα δείγματα και αυτά που προήλθαν από τις περιοχές πάνω και κάτω από τη θέση του τραυματισμού.

III Γραφική απεικόνιση της ομαδοποίησης του Isomap ως προς τη **σοβαρότητα** του τραυματισμού.



Σχήμα 11.15: Ανάλυση Isomap ως προς τον φαινότυπο της σοβαρότητας του τραυματισμού για 167 συστοιχίες ολιγονουκλεοτιδίων υψηλής πυκνότητας που προέκυψαν από ποντίκια όπως περιγράφεται στην αρχή του κεφαλαίου.

Εικ. 11.15: Τα δείγματα εικονικά χειρουργημένων ποντικιών (sham-operated) βρίσκονται στον κεντρικό πυρήνα του μοντέλου που περιβάλλεται από ένα κέλυφος δειγμάτων από ποντικούς με ήπια κάκωση νωτιαίου μυελού. Τα πιο απομακρυσμένα δείγματα στους δύο λοβούς του μοντέλου είναι εκείνα που εκπροσωπούν μέτρια έως σοβαρή βλάβη. Στο δεξιό λοβό του μοντέλου, ο οποίος εκπροσωπεί τα χρονικά σημεία από 30 λεπτά ως 24 ώρες, δεν υπάρχει σαφής διαχωρισμός

μεταξύ τού μέτριου και τού σοβαρού τραυματισμού. Απροσδόκητα, στον αριστερό λοβό που αντιπροσωπεύει τα μεταγενέστερα χρονικά σημεία, τα δείγματα με μέτριο τραυματισμό είναι πιο απομακρυσμένα από εκείνα με σοβαρό τραυματισμό. Παρά το γεγονός ότι αυτός ο διαχωρισμός μεταξύ της μέτριας και σοβαρής ζημίας δεν ήταν αναμενόμενος, ο διαχωρισμός μεταξύ ήπιας και μέτριας έως σοβαρής ζημίας είναι πολύ σαφής. Ο διαχωρισμός αυτός απαντά επίσης σε μια ερώτηση που τέθηκε στον πίνακα χρόνου. Στον αριστερό λοβό τού μοντέλου, έξι δείγματα είναι ορατά στις 28 ημέρες μετά την κάκωση.

Κεφάλαιο 12

ΠΑΡΑΡΤΗΜΑ 1: Λίγα λόγια για τους αλγόριθμους εκμάθησης πολλαπλοτήτων

Σε αυτό το παράρτημα παρουσιάζουμε πολύ συνοπτικά τις ιδέες πάνω στις οποίες βασίστηκαν κάποιες από τις μεθόδους εκμάθησης πολλαπλοτήτων και ομαδοποίησης που έχουν χρησιμοποιηθεί κατά καιρούς για την αναγωγή δεδομένων μεγάλης κλίμακας σε σύνολα δεδομένων μικροσυστοιχιών και οι οποίες αναφέρονται στην εισαγωγή.

1. Η Ιεραρχική ομαδοποίηση (HC Hierarchical Clustering) είναι μια μέθοδος ανάλυσης κατά συστάδες που επιζητά να φτιάξει μια ιεραρχία των συστάδων. Οι στρατηγικές της HC γενικά εμπίπτουν σε δύο κατηγορίες:
 - Συσσωρευτικός (Agglomerative): Αυτή είναι μια «από κάτω-προς-τα-πάνω» προσέγγιση. Κάθε παρατήρηση ξεκινά από τη δική του ομάδα και ζεύγη ομάδων συγχωνεύονται καθώς προχωράμε παραπάνω στην ιεραρχία.
 - Διαίρετικός (Divisive): Αυτή είναι μία «από-πάνω-προς-τα-κάτω» προσέγγιση. Όλες οι παρατηρήσεις ξεκινούν σε μια ομάδα και γίνονται διαχωρισμοί αναδρομικά καθώς προχωρά κανείς προς τα κάτω στην ιεραρχία.

Γενικά οι συγχωνεύσεις και διαιρέσεις γίνονται με άπληστο (greedy) τρόπο (αναζήτηση δραστηριότητας με ελάχιστο χρόνο ολοκλήρωσης οδηγεί σε συνολικά βέλτιστη λύση) και τα αποτελέσματα του αλγόριθμου συνήθως παρουσιάζονται σε ένα δενδρόγραμμα (dendrogram).

2. Το τεχνητό νευρωνικό δίκτυο ANN: Artificial Neural Network είναι ένα δίκτυο από απλούς υπολογιστικούς κόμβους (νευρώνες, νευρώνια), διασυνδεδεμένους μεταξύ τους. Οι νευρώνες είναι τα δομικά στοιχεία του δικτύου. Κάθε τέτοιος κόμβος δέχεται ένα σύνολο αριθμητικών εισόδων από διαφορετικές πηγές (είτε από άλλους νευρώνες, είτε από το περιβάλλον), επιτελεί έναν υπολογισμό με βάση αυτές τις εισόδους και παράγει μία έξοδο. Η εν λόγω έξοδος είτε κατευθύνεται στο περιβάλλον, είτε τροφοδοτείται ως είσοδος σε άλλους νευρώνες του δικτύου. Υπάρχουν τρεις τύποι νευρώνων: οι νευρώνες εισόδου, οι νευρώνες εξόδου και οι υπολογιστικοί νευρώνες ή κρυμμένοι νευρώνες. Οι νευρώνες εισόδου δεν επιτελούν κανέναν υπολογισμό, μεσολαβούν απλώς ανάμεσα στις περιβαλλοντικές εισόδους του δικτύου και στους υπολογιστικούς νευρώνες. Οι νευρώνες εξόδου διοχετεύουν στο περιβάλλον τις τελικές αριθμητικές εξόδους του δικτύου. Οι υπολογιστικοί νευρώνες πολλαπλασιάζουν κάθε είσοδό τους με το αντίστοιχο συναπτικό βάρος και υπολογίζουν το ολικό άθροισμα των γινομένων. Το άθροισμα αυτό τροφοδοτείται ως όρισμα στη συνάρτηση ενεργοποίησης, την οποία υλοποιεί εσωτερικά κάθε κόμβος. Η τιμή που λαμβάνει η συνάρτηση για το εν λόγω όρισμα είναι και η έξοδος του νευρώνα για τις τρέχουσες εισόδους και βάρη.
3. Δένδρο συγκομιδής (Harvesting tree): μια νέα μέθοδος της επιβλεπόμενης μάθησης Αυτή η τεχνική ξεκινά με την ιεραρχική ομαδοποίηση των γονιδίων, και στη συνέχεια μοντελοποιεί την μεταβλητή που έχει προκύψει ως ένα άθροισμα των μέσων προφίλ έκφρασης επιλεγμένων ομάδων και των προϊόντων τους.
4. Τα πλησιέστερα συρρικνωμένα κεντροειδή (nearest shrunken centroids) Εν συντομία, η μέθοδος υπολογίζει ένα τυποποιημένο κεντροειδές για κάθε κλάση. Αυτή είναι η μέση τιμή γονιδιακής έκφρασης για κάθε γο-

νίδιο σε κάθε κατηγορία διαιρεμένη με την τυπική απόκλιση ολόκληρης της κλάσης για το συγκεκριμένο γονίδιο. Η ταξινόμηση πλησιέστερων κέντροειδών λαμβάνει το προφίλ γονιδιακής έκφρασης ενός νέου δείγματος, και το συγκρίνει με καθένα από αυτά τα κέντροειδή της κλάσης. Η κλάση της οποίας το κέντροειδές είναι πιο κοντά, ως τετράγωνο της απόστασης, είναι η προβλεπόμενη κλάση για το νέο δείγμα.

5. Η πιθανολογική PCA (Probabilistic PCA). Θεωρεί πως οι κύριοι άξονες ενός συνόλου διανυσμάτων παρατηρήσεων (σύνολο δεδομένων) μπορούν να προσδιορίζονται μέσω της εκτίμησης μέγιστης πιθανοφάνειας των παραμέτρων ενός μοντέλου λανθάνουσών μεταβλητών (latent variable model) στενά συνδεδεμένου με την παραγοντική ανάλυση.
6. Η αποσύνθεση ιδιαζουσών τιμών (Singular Value Decomposition SVD), Τα δεδομένα γονιδιακής έκφρασης n γονιδίων, καθενός μετρημένου σε m διακριτά χρονικά σημεία οργανώνονται σε ένα $n \times m$ πίνακα A . λαμβάνεται με εφαρμογή της SVD. Τα δεδομένα έκφρασης για κάθε γονίδιο μπορούν να θεωρηθούν ως ένα μοναδιαίο διάνυσμα σε έναν χώρο υψηλών διαστάσεων, οι άξονες καθενός εκ των οποίων αντιπροσωπεύουν το επίπεδο έκφρασης μιας μέτρησης στα χρονικά του πειράματος. Η κατασκευή μέσω της SVD εξασφαλίζει ότι οι μέσοι αντιστοιχούν σε γραμμικώς ανεξάρτητα διανύσματα βάσης, ένας γραμμικός συνδυασμός των οποίων περιγράφει ακριβώς το μοτίβο έκφρασης του κάθε γονιδίου.
7. Ο αλγόριθμος k -μέσων (k -means) είναι ένας από τους απλούστερους αλγόριθμους μάθησης χωρίς επίβλεψη που μπορεί να λύσει το γνωστό πρόβλημα της ομαδοποίησης. Η ομαδοποίηση k -μέσων είναι μια μέθοδος κβαντισμού διανυσμάτων, αρχικά από το πεδίο της επεξεργασίας σήματος (signal processing) που είναι δημοφιλής για την ανάλυση κατά συστάδες στην εξόρυξη δεδομένων. Η συσταδοποίηση k -μέσων είναι μια μέθοδος που χρησιμοποιείται για να ταξινομήσει ημι δομημένα ή μη-δομημένα σύνολα δεδομένων. Αποτελεί μία από τις συνηθέστερες και πιο αποτελεσματικές μεθόδους για την ταξινόμηση δεδομένων λόγω της απλότητάς της και της ικανότητας να χειριστεί τα ογκώδη σύνολα δεδομένων. Δέχε-

ται ως παραμέτρους τον αριθμό των συστάδων και το αρχικό σύνολο των κεντροειδών. Υπολογίζεται η απόσταση του κάθε στοιχείου στο σύνολο δεδομένων από κάθε ένα από τα κεντροειδή της αντίστοιχης συστάδας. Το αντικείμενο τότε τοποθετείται στη συστάδα από την οποία απέχει μικρότερη απόσταση. Το κεντροειδές της συστάδας στην οποία το στοιχείο εντάχθηκε υπολογίζεται εκ νέου.

8. CAST: Η τεχνική CAST είναι ένας αλγόριθμος που προτείνουν οι [Bendor et al. 1999] για την ομαδοποίηση δεδομένων γονιδιακής έκφρασης. Η είσοδος του αλγόριθμου περιλαμβάνει τα ζεύγη ομοιότητας των γονιδίων, και μια παράμετρο αποκοπής (threshold) (Η οποία είναι ένας πραγματικός αριθμός μεταξύ 0 και 1 και μπορεί να θεωρηθεί ως το αντίστροφο της απόστασης μετρική μεταξύ δύο γονιδίων). Οι ομάδες κατασκευάζονται μία-μία. Ο αλγόριθμος λειτουργεί και με προσθήκη και με αφαίρεση γονιδίων από μια ομάδα, προσαρμόζοντας κάθε φορά τις συγγένειες των γονιδίων στην τρέχουσα ομάδα, και συνεχίζοντας αυτή τη διαδικασία μέχρις ότου δεν μπορούν να γίνουν περαιτέρω αλλαγές στην τρέχουσα ομάδα. Η συγγένεια (affinity) ενός γονιδίου ορίζεται ως το άθροισμα των τιμών ομοιότητας (similarity values) και όλων των γονιδίων στην υπό κατασκευή συστάδα.
9. Ασαφής συσταδοποίηση *c*-μέσων (fuzzy *c*-means clustering): είναι μια τεχνική ομαδοποίησης δεδομένων στην οποία ένα σύνολο δεδομένων ομαδοποιούνται σε n συστάδες με κάθε σημείο του συνόλου δεδομένων να ανήκει σε κάθε συστάδα σε έναν ορισμένο βαθμό. Για παράδειγμα, ένα ορισμένο σημείο δεδομένων που βρίσκεται κοντά στο κέντρο της συστάδας θα ανήκει σε μεγάλο βαθμό στη συγκεκριμένη ομάδα και κάποιο άλλο σημείο που βρίσκεται μακριά από το κέντρο θα έχει μικρότερη συμμετοχή στη συγκεκριμένη συστάδα. Ο FCMC ομαδοποιεί τα δεδομένα σε ασαφείς συστάδες ελαχιστοποιώντας το άθροισμα της αντικειμενικής συνάρτησης τετραγωνικού σφάλματος μέσα σε μια ομάδα.
10. Αμφίδρομη Ομαδοποίηση Two-way clustering μία τεχνική εξόρυξης δεδομένων που επιτρέπει την ταυτόχρονη ομαδοποίηση των γραμμών και

των στηλών ενός πίνακα. Δοθέντος ενός $m \times n$ πίνακα δεδομένων, οι αλγόριθμοι biclustering παράγουν biclusters - ένα υποσύνολο γραμμών που εμφανίζουν παρόμοια συμπεριφορά σε ένα υποσύνολο στηλών, ή το αντίστροφο.

11. Αυτοοργανωνόμενοι Χάρτες (SOM: Self- Organizing Maps) είναι ένα από τα πιο δημοφιλή μοντέλα νευρωνικών δικτύων. Ο αλγόριθμος SOM βασίζεται σε μη επιβλεπόμενη, ανταγωνιστική (competitive) μάθηση. Παρέχει μια χαρτογράφηση διατηρεί τις τοπολογικές ιδιότητες από τον χώρο υψηλών διαστάσεων σε μονάδες χάρτη (map units)(μονάδα μετρησης απόστασης γονιδίων σε έναν χρωμόσωμα) . Οι μονάδες χάρτη, ή νευρώνες, συνήθως σχηματίζουν ένα διδιάστατο πλέγμα και έτσι η χαρτογράφηση είναι μια χαρτογράφηση από τον χώρο υψηλής διάστασης πάνω σε ένα επίπεδο. Η ιδιότητα να διατηρεί την τοπολογία σημαίνει ότι η χαρτογράφηση διατηρεί τη σχετική απόσταση μεταξύ των σημείων. Τα σημεία που βρίσκονται κοντά μεταξύ τους στο χώρο εισόδου αντιστοιχίζονται σε κοντινές μονάδες χάρτη στον SOM. Ο SOM μπορεί να χρησιμεύσει ως εργαλείο για την ανάλυση διασποράς των υψηλών διαστάσεων των δεδομένων. Επίσης, ο SOM έχει την ικανότητα να γενικεύει. Αυτό σημαίνει ότι το δίκτυο μπορεί να αναγνωρίσει ή να χαρακτηρίσει εισροές που δεν έχει αντιμετωπίσει ποτέ πριν. Μια νέα είσοδος εξομοιώνεται με τη μονάδα χάρτη στην οποία έχει αντιστοιχιστεί.
12. Μηχανές Διανυσμάτων Υποστήριξης (SVM Support Vector Machines): Αποτελούν έναν τρόπο εκπαίδευσης των νευρωνικών δικτύων. Πιο συγκεκριμένα, μια μηχανή διανυσμάτων υποστήριξης κατασκευάζει ένα υπερεπίπεδο ή ένα σύνολο από υπερεπίπεδα σε έναν υψηλής ή άπειρης διαστάσης χώρο, που μπορεί να χρησιμοποιηθεί για ταξινόμηση, παλινδρόμηση, ή άλλες εργασίες. Διαισθητικά, ένας καλός διαχωρισμός επιτυγχάνεται από το υπερεπίπεδο που έχει τη μεγαλύτερη απόσταση από το πλησιέστερο σημείο των εκπαιδευμένων-δεδομένων οποιασδήποτε κλάσης το λεγόμενο λειτουργικό περιθώριο (functional margin), δεδομένου ότι σε γενικές γραμμές όσο μεγαλύτερο είναι το περιθώριο τόσο χαμηλότερη είναι η λάθος γενίκευση του ταξινομητή.

Bibliography

- Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. & Levine, A. J. (1999), ‘Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays’, *Proceedings of the National Academy of Sciences* **96**(12), 6745–6750.
- Alter, O., Brown, P. O. & Botstein, D. (2003), ‘Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms’, *Proceedings of the National Academy of Sciences* **100**(6), 3351–3356.
- Bah, B. (2008), Diffusion maps: analysis and applications, PhD thesis, University of Oxford.
- Balasubramanian, M. & Schwartz, E. L. (2002), ‘The isomap algorithm and topological stability’, *Science* **295**(5552), 7–7.
- Belkin, M. & Niyogi, P. (2001), Laplacian eigenmaps and spectral techniques for embedding and clustering, *in* ‘NIPS’, Vol. 14, pp. 585–591.
- Belkin, M. & Niyogi, P. (2003), ‘Laplacian eigenmaps for dimensionality reduction and data representation’, *Neural computation* **15**(6), 1373–1396.
- Belkin, M. & Niyogi, P. (2008), ‘Towards a theoretical foundation for laplacian-based manifold methods’, *Journal of Computer and System Sciences* **74**(8), 1289–1308.
- Ben-Dor, A., Shamir, R. & Yakhini, Z. (1999), ‘Clustering gene expression patterns’, *Journal of computational biology* **6**(3-4), 281–297.

- Bengio, Y., Paiement, J.-F., Vincent, P., Delalleau, O., Le Roux, N. & Ouimet, M. (2003), ‘Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering’, *Mij* **1**, 2.
- Bengio, Y., Vincent, P., Paiement, J.-F., Delalleau, O., Ouimet, M. & Le Roux, N. (2003), *Spectral clustering and kernel PCA are learning eigenfunctions*, Vol. 1239, Citeseer.
- Binder, H., Kirsten, T., Loeffler, M. & Stadler, P. F. (2004), ‘Sensitivity of microarray oligonucleotide probes: variability and effect of base composition’, *The Journal of Physical Chemistry B* **108**(46), 18003–18014.
- Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A. et al. (2000), ‘Molecular classification of cutaneous malignant melanoma by gene expression profiling’, *Nature* **406**(6795), 536–540.
- Bolstad, B. (2001), ‘Probe level quantile normalization of high density oligonucleotide array data’, *Unpublished manuscript*.
- Bolstad, B. M., Irizarry, R. A., Åstrand, M. & Speed, T. P. (2003), ‘A comparison of normalization methods for high density oligonucleotide array data based on variance and bias’, *Bioinformatics* **19**(2), 185–193.
- Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M. & Haussler, D. (2000), ‘Knowledge-based analysis of microarray gene expression data by using support vector machines’, *Proceedings of the National Academy of Sciences* **97**(1), 262–267.
- Burden, C. J. (2008), ‘Understanding the physics of oligonucleotide microarrays: the affymetrix spike-in data reanalysed’, *Physical Biology* **5**(1), 016004.
- Chudin, E., Walker, R., Kosaka, A., Wu, S. X., Rabert, D., Chang, T. K. & Kreder, D. E. (2001), ‘Assessment of the relationship between signal intensities and transcript concentration for affymetrix genechip® arrays’, *Genome biology* **3**(1), research0005–1.

- Crick, F. H. (1958), ‘On protein synthesis’, *Symposia of the Society for Experimental Biology* **12**(138-63), 8.
- Cung, B., Jin, T., Ramirez, J., Thompson, A., Boutsidis, C. & Needell, D. (2012), ‘Spectral clustering: An empirical study of approximation algorithms and its application to the attrition problem’, *arXiv preprint arXiv:1211.3444* .
- Dawson, K., Rodriguez, R. L. & Malyj, W. (2005), ‘Sample phenotype clusters in high-density oligonucleotide microarray data sets are revealed using isomap, a nonlinear algorithm’, *BMC Bioinformatics* **6**(1), 195.
- De La Nava, J. G., Santaella, D. F., Alba, J. C., Carazo, J. M., Trelles, O. & Pascual-Montano, A. (2003), ‘Engene: the processing and exploratory analysis of gene expression data’, *Bioinformatics* **19**(5), 657–658.
- Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998), ‘Cluster analysis and display of genome-wide expression patterns’, *Proceedings of the National Academy of Sciences* **95**(25), 14863–14868.
- Forsyth, D. A. & Ponce, J. (2003), ‘A modern approach’, *Computer vision: a modern approach* pp. 88–101.
- Freudenberg, J. M. (2005), ‘Comparison of background correction and normalization procedures for high-density oligonucleotide microarrays’, *Institut für Informatik* **120**(11).
- Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M. & Haussler, D. (2000), ‘Support vector machine classification and validation of cancer tissue samples using microarray expression data’, *Bioinformatics* **16**(10), 906–914.
- GENECLUSTER, B. (n.d.), ‘Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation’.
- Girolami, M. & Breitling, R. (2004), ‘Biologically valid linear factor models of gene expression’, *Bioinformatics* **20**(17), 3021–3033.

- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A. et al. (1999), ‘Molecular classification of cancer: class discovery and class prediction by gene expression monitoring’, *Science* **286**(5439), 531–537.
- Hastie, T., Tibshirani, R., Botstein, D. & Brown, P. (2001), ‘Supervised harvesting of expression trees’, *Genome Biology* **2**(1), research0003–1.
- Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M. et al. (2001), ‘Gene-expression profiles in hereditary breast cancer’, *New England Journal of Medicine* **344**(8), 539–548.
- Holter, N. S., Maritan, A., Cieplak, M., Fedoroff, N. V. & Banavar, J. R. (2001), ‘Dynamic modeling of gene expression data’, *Proceedings of the National Academy of Sciences* **98**(4), 1693–1698.
- Holter, N. S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J. R. & Fedoroff, N. V. (2000), ‘Fundamental patterns underlying gene expression profiles: simplicity from complexity’, *Proceedings of the National Academy of Sciences* **97**(15), 8409–8414.
- Huang, L., Yan, D., Taft, N. & Jordan, M. I. (2009), Spectral clustering with perturbed data, in ‘Advances in Neural Information Processing Systems’, pp. 705–712.
- Ihaka, R. & Gentleman, R. (1996), ‘R: a language for data analysis and graphics’, *Journal of computational and graphical statistics* **5**(3), 299–314.
- Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. & Speed, T. P. (2003), ‘Exploration, normalization, and summaries of high density oligonucleotide array probe level data’, *Biostatistics* **4**(2), 249–264.
- Jain, A. K. (2010), ‘Data clustering: 50 years beyond k-means’, *Pattern recognition letters* **31**(8), 651–666.

- Jolliffe, I. (2002), *Principal component analysis*, Wiley Online Library.
- Khan, J., Simon, R., Bittner, M., Chen, Y., Leighton, S. B., Pohida, T., Smith, P. D., Jiang, Y., Gooden, G. C., Trent, J. M. et al. (1998), ‘Gene expression profiling of alveolar rhabdomyosarcoma with cdna microarrays’, *Cancer Research* **58**(22), 5009–5013.
- Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C. et al. (2001), ‘Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks’, *Nature medicine* **7**(6), 673–679.
- Li, C. & Wong, W. H. (2001*a*), ‘Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection’, *Proceedings of the National Academy of Sciences* **98**(1), 31–36.
- Li, C. & Wong, W. H. (2001*b*), ‘Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application’, *Genome biology* **2**(8), research0032–1.
- Liu, L., Hawkins, D. M., Ghosh, S. & Young, S. S. (2003), ‘Robust singular value decomposition analysis of microarray data’, *Proceedings of the National Academy of Sciences* **100**(23), 13167–13172.
- Nadler, B., Lafon, S., Coifman, R. & Kevrekidis, I. (2005), ‘Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators’, pp. 955–962.
- Nadler, B., Lafon, S., Coifman, R. R. & Kevrekidis, I. G. (2006), ‘Diffusion maps, spectral clustering and reaction coordinates of dynamical systems’, *Applied and Computational Harmonic Analysis* **21**(1), 113–127.
- Naef, F., Lim, D. A., Patil, N. & Magnasco, M. (2002), ‘Dna hybridization to mismatched templates: a chip study’, *Physical Review E* **65**(4), 040902.
- Naef, F. & Magnasco, M. O. (2003), ‘Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays’, *Physical Review E* **68**(1), 011906.

- Naef, F., Magnasco, M. & Socci, N. (2002), Extracting more signal at high intensities in oligonucleotide arrays, Technical report.
- Ng, A. Y., Jordan, M. I., Weiss, Y. et al. (2001), On spectral clustering: Analysis and an algorithm, *in* ‘NIPS’, Vol. 14, pp. 849–856.
- Nicholson, J. K., Holmes, E., Lindon, J. C. & Wilson, I. D. (2004), ‘The challenges of modeling mammalian biocomplexity’, *Nature biotechnology* **22**(10), 1268–1274.
- Nielsen, T. O., West, R. B., Linn, S. C., Alter, O., Knowling, M. A., O’Connell, J. X., Zhu, S., Fero, M., Sherlock, G., Pollack, J. R. et al. (2002), ‘Molecular characterisation of soft tissue tumours: a gene expression study’, *The Lancet* **359**(9314), 1301–1307.
- Nilsson, J., Fioretos, T., Höglund, M. & Fontes, M. (2004), ‘Approximate geodesic distances reveal biologically relevant structures in microarray data’, *Bioinformatics* **20**(6), 874–880.
- Orsenigo, C. & Vercellis, C. (2012), ‘An effective double-bounded tree-connected isomap algorithm for microarray data classification’, *Pattern Recognition Letters* **33**(1), 9–16.
- Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslén, L. A. et al. (2000), ‘Molecular portraits of human breast tumours’, *Nature* **406**(6797), 747–752.
- Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y., Goumnerova, L. C., Black, P. M., Lau, C. et al. (2002), ‘Prediction of central nervous system embryonal tumour outcome based on gene expression’, *Nature* **415**(6870), 436–442.
- Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P. et al. (2001), ‘Multiclass cancer diagnosis using tumor gene expression signatures’, *Proceedings of the National Academy of Sciences* **98**(26), 15149–15154.

- Roweis, S. T. & Saul, L. K. (2000), ‘Nonlinear dimensionality reduction by locally linear embedding’, *science* **290**(5500), 2323–2326.
- Sammon, J. W. (1969), ‘A nonlinear mapping for data structure analysis’, *IEEE Transactions on computers* **100**(5), 401–409.
- Sarwar, B., Karypis, G., Konstan, J. & Riedl, J. (2000), Application of dimensionality reduction in recommender system—a case study, Technical report, DTIC Document.
- Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995), ‘Quantitative monitoring of gene expression patterns with a complementary dna microarray’, *Science* **270**(5235), 467.
- Scherf, U., Ross, D. T., Waltham, M., Smith, L. H., Lee, J. K., Tanabe, L., Kohn, K. W., Reinhold, W. C., Myers, T. G., Andrews, D. T. et al. (2000), ‘A gene expression database for the molecular pharmacology of cancer’, *Nature genetics* **24**(3), 236–244.
- Shi, J. & Malik, J. (2000), ‘Normalized cuts and image segmentation’, *IEEE Transactions on pattern analysis and machine intelligence* **22**(8), 888–905.
- Sonka, M., Hlavac, V. & Boyle, R. (2014), *Image processing, analysis, and machine vision*, Cengage Learning.
- Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B. (1998), ‘Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization’, *Molecular biology of the cell* **9**(12), 3273–3297.
- Staunton, J. E., Slonim, D. K., Collier, H. A., Tamayo, P., Angelo, M. J., Park, J., Scherf, U., Lee, J. K., Reinhold, W. O., Weinstein, J. N. et al. (2001), ‘Chemosensitivity prediction by transcriptional profiling’, *Proceedings of the National Academy of Sciences* **98**(19), 10787–10792.

- Su, A. I., Welsh, J. B., Sapinoso, L. M., Kern, S. G., Dimitrov, P., Lapp, H., Schultz, P. G., Powell, S. M., Moskaluk, C. A., Frierson, H. F. et al. (2001), ‘Molecular classification of human carcinomas by use of gene expression signatures’, *Cancer research* **61**(20), 7388–7393.
- Szlam, A., Kluger, Y. & Tygert, M. (2014), ‘An implementation of a randomized algorithm for principal component analysis’, *arXiv preprint arXiv:1412.3510* .
- Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. & Golub, T. R. (1999), ‘Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation’, *Proceedings of the National Academy of Sciences* **96**(6), 2907–2912.
- Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. (1999), ‘Systematic determination of genetic network architecture’, *Nature genetics* **22**(3), 281–285.
- Tenenbaum, J. B., De Silva, V. & Langford, J. C. (2000), ‘A global geometric framework for nonlinear dimensionality reduction’, *Science* **290**(5500), 2319–2323.
- Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. (2002), ‘Diagnosis of multiple cancer types by shrunken centroids of gene expression’, *Proceedings of the National Academy of Sciences* **99**(10), 6567–6572.
- Trefethen, L. N. (2000), *Spectral methods in MATLAB*, SIAM.
- Von Luxburg, U. (2007), ‘A tutorial on spectral clustering’, *Statistics and computing* **17**(4), 395–416.
- Wang, J., Bø, T. H., Jonassen, I., Myklebost, O. & Hovig, E. (2003), ‘Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data’, *BMC bioinformatics* **4**(1), 60.

- Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Barker, J. L. & Somogyi, R. (1998), ‘Large-scale temporal gene expression mapping of central nervous system development’, *Proceedings of the National Academy of Sciences* **95**(1), 334–339.
- Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F. & Spencer, F. (2004), ‘A model-based background adjustment for oligonucleotide expression arrays’, *Journal of the American statistical Association* **99**(468), 909–917.
- Yeang, C.-H., Ramaswamy, S., Tamayo, P., Mukherjee, S., Rifkin, R. M., Angelo, M., Reich, M., Lander, E., Mesirov, J. & Golub, T. (2001), ‘Molecular classification of multiple tumor types’, *Bioinformatics* **17**(suppl 1), S316–S322.
- Zhang, L., Miles, M. F. & Aldape, K. D. (2003), ‘A model of molecular interactions on short oligonucleotide microarrays’, *Nature biotechnology* **21**(7), 818–821.