ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
DEPARTMENT OF STATISTICS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

Τμήμα Στατιστικής

Linear and nonlinear dimensionality reduction methods and application

Άγγελος —Κωνσταντίνου— Κροκίδης

Διπλωματική Εργασία
που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών στο
πλαίσιο του Προπτυχιακού Προγράμματος
Σπουδών

Αθήνα
Ιούνιος,2026

Ευχαριστίες

Θα ήθελα να εκφράσω τις θερμότερες και πιο ειλικρινείς μου ευχαριστίες στον επιβλέποντα καθηγητή μου, κ. Γιαννακόπουλο, για την πολύτιμη καθοδήγηση,την βοήθεια με όλη την βιβλιογραφία, την εμπιστοσύνη και την αμέριστη υποστήριξή του, όχι μόνο κατά την εκπόνηση της παρούσας διπλωματικής εργασίας, αλλά καθ' όλη τη διάρκεια των προπτυχιακών μου σπουδών.Η συνεργασία μας υπήρξε για εμένα πηγή έμπνευσης και ουσιαστικής μάθησης: οι συζητήσεις μας, οι εύστοχες παρατηρήσεις του και ο τρόπος με τον οποίο με ώθησε να σκέφτομαι αυστηρά και ανεξάρτητα διαμόρφωσαν καθοριστικά τόσο την παρούσα εργασία όσο και τη συνολική μου πορεία στο τμήμα Στατιστικής.Τον ευχαριστώ θερμά για τον χρόνο, την υπομονή και τη γενναιοδωρία με την οποία μοιράστηκε τις γνώσεις του. Ευχαριστώ,τέλος,την οικογένειά μου και τους ανθρώπους που με στήριξαν σε όλη αυτή τη διαδρομή.

Περιεχόμενα

Ευχαριστίες **2**

1 Εισαγωγή **6**
1.1 Κίνητρο: Δεδομένα Υψηλής Διάστασης 6
1.2 Τι είναι η Μείωση Διαστατικότητας 7
1.3 Η Διάσπαση Ιδιαζουσών Τιμών (SVD) 8
1.4 Κατηγορίες Μεθόδων μείωσης διάσπασης 11
1.5 Επίπεδα Επίβλεψης 12

2 Η μέθοδος PCA σαν πρόβλημα βελτιστοποίησης **13**
2.1 Η SVD ως βέλτιστη γραμμική συμπίεση 13
2.2 SVD και Βέλτιστη Προσέγγιση Χαμηλής Τάξης 15
2.3 PCA: Η SVD ως Μέθοδος Μείωσης Διάστασης 15
2.4 PCA: Μεγιστοποίηση Διασποράς 17
2.5 Ο Μαθηματικός Πυρήνας: Βελτιστοποίηση Ίχνους (Trace Optimization) . . 17
2.6 Το Γενικό Πρόβλημα Βελτιστοποίησης Ίχνους 20
2.7 Περιορισμός B-Ορθογωνιότητας: Το Γενικευμένο Ιδιοπρόβλημα 21
2.8 Το Πρόβλημα Λόγου Ιχνών (Trace Ratio) 22

3 Μη γραμμικές μέθοδοι (ως επεκτάσεις της PCA) **24**
3.1 Τα Όρια των Γραμμικών Μεθόδων: Πότε "Σπάει" η SVD 24
3.2 Μη Γραμμικές Μέθοδοι Βασισμένες σε Γράφους και Τοπικότητα 25
3.2.1 LLE: Τοπικά Γραμμικές Σχέσεις 25
3.2.2 Laplacian Eigenmaps και ο Λαπλασιανός Γράφος 26
3.2.3 MDS / ISOMAP: Από Αποστάσεις σε Εσωτερικά Γινόμενα 27
3.3 Γραμμικές Προβολές Διατήρησης Τοπικότητας 29
3.3.1 LPP:Locality Preserving Projections (Γραμμική Διατήρηση Τοπικότητας) 29
3.3.2 Orthogonal Neighborhood Preserving Projections (ONPP / NPP) . 30
3.3.3 LDA(Linear Discriminant Analysis): Εποπτευόμενος Διαχωρισμός Κλάσεων 30
3.4 Συνδέσεις, Ισοδυναμίες και ο Ρόλος του Λαπλασιανού Γράφου 31
3.5 Ο Πίνακας LLE ως "Σχεδόν" Λαπλασιανός 32
3.6 Ο Πίνακας Επίβλεψης H είναι Προβολεας 32
3.7 Οι Γραμμικές Μέθοδοι ως Προβολές (Rayleigh–Ritz) 33
3.8 Σύνδεση PCA – LPP 34
3.9 Σύνδεση με Φασματική Συσταδοποίηση 35

4 Πυρήνες και η Ενοποίηση Γραμμικών–Μη Γραμμικών Μεθόδων **36**
4.1 Εισαγωγή στους Πυρήνες (The Kernel Trick) 36
4.1.1 Πυρήνες και Kernel PCA 37

5 Επεξηγηματικά Παραδείγματα **40**
5.1 Σύνολα Δεδομένων και Πρωτόκολλο 40
5.2 Οπτικοποίηση σε 2D 40
5.3 Επίδραση των Πυρήνων 42

6 Συμπεράσματα **43**

Περίληψη

Στην παρούσα εργασία πραγματευόμαστε τις γραμμικές και μη γραμμικές μεθόδους μείωσης διαστατικότητας, με αφετηρια μας την Διάσπαση Ιδιαζουσών Τιμών (SVD) και την Ανάλυση Κύριων Συνιστωσών (PCA).

Αρχικα, παρουσιάζεται το κίνητρο της μελέτης, τα δεδομένα υψηλής διάστασης και οι προκλήσεις με τις οποίες αυτά συνεπάγονται και διατυπώνεται η SVD ως βέλτιστη μέθοδος γραμμικής συμπίεσης και προσέγγισης χαμηλής τάξης. Η PCA αντιμετωπίζεται ως πρόβλημα βελτιστοποίησης: μεγιστοποίησης διασποράς και βελτιστοποίησης ίχνους, ενώ εξετάζονται επεκτάσεις όπως το γενικευμένο ιδιοπρόβλημα και το πρόβλημα λόγου ιχνών.

Επίσης μελετάμε μη γραμμικές μέθοδους που αποτελούν επεκτάσεις της PCA, όπως οι LLE, Laplacian Eigenmaps, ISOMAP και MDS, γραμμικές μέθοδοι διατήρησης τοπικότητας (LPP, ONPP) και εποπτευόμενος διαχωρισμός κλάσεων μέσω LDA. Έτσι αναδεικνύεται ο ενοποιητικός ρόλος του Λαπλασιανού γράφου και οι συνδέσεις μεταξύ των μεθόδων μέσω της θεωρίας Rayleigh–Ritz.

Τέλος, εισάγεται το kernel trick (το κόλπο του πυρήνα) και η Kernel PCA ως γέφυρα μεταξύ γραμμικών και μη γραμμικών προσεγγίσεων, και παρουσιάζονται επεξηγηματικά παραδείγματα οπτικοποίησης.

Κεφάλαιο 1

Εισαγωγή

1.1 Κίνητρο: Δεδομένα Υψηλής Διάστασης

Ο όρος *data mining* (εξόρυξη δεδομένων) αναφέρεται σε ένα ευρη φάσμα,την μηχανική μάθηση, την ανάκτηση πληροφορίας την αναγνώριση προτύπων και την ανάλυση δεδομένων. Ενα κοινό χαρακτηριστικό τους είναι πως, ολο και συχνότερα, καλούμαστε να επεξεργαστούμε δεδομενα τα οποία περιγράφονται από έναν τεράστιο αριθμό χαρακτηριστικών (features), όπως τα pixels μιας εικόνας, οι συχνότητες λέξεων εντός ενος κειμένου, χαρακτηριστικών ενος προσώπου, ή οι εκφράσεις χιλιάδων γονιδιων .Τυπικα, διαθέτουμε n δείγματα, καθένα από τα οποία ζει σε έναν χώρο $\mathbb{R}^m$ με πολύ μεγάλο m. Η εργασία με τέτοια δεδομένα προσκρούει στο φαινόμενο *curse of dimensionality* (κατάρα της διαστατικότητας): όσο το m αυξάνεται, τοσο ο ογκος του χώρου μεγαλώνει εκθετικά.Οι Ευκλείδειες αποστάσεις τεινουν να ομογενοποιουνται και να χάνουν τη διακριτική τους ικανότητα. Δηλαδή οταν τα δεδομένα έχουν πάρα πολλές διαστάσεις (πολλά χαρακτηριστικά στην περίπτωση μας), οι αποστασεις μεταξύ των σημείων αρχίζουν να μοιάζουν όλες μεταξύ τους.Έτσι γίνεται δύσκολο να ξεχωρίσουμε ποια σημεία είναι πραγματικά κοντά και ποια μακριά.Το υπολογιστικό κόστος αυξάνεται πάρα πολυ, ενώ η στατιστική εκτίμηση γίνεται ασταθής λόγο του πλήθους των παραμέτρων.Έτσι γίνεται δύσκολο να ξεχωρίσουμε ποια σημεία είναι πραγματικά κοντά και ποια μακριά. Αυτό που σώζει την κατάσταση στην πράξη είναι μια εμπειρική παρατήρηση: παρότι τα δεδομένα ζουν *τυπικά* σε έναν χώρο διαστασης m, η *ουσιαστική* (intrinsic) διάστασή τους είναι συχνά πολύ μικρότερη. Τα δείγματα δεν γεμίζουν ομοιόμορφα τον $\mathbb{R}^m$, αντιθέτως, συγκεντρώνονται κοντά σε μια δομή χαμηλής διάστασης έναν υπόχωρο. Η ανίχνευση και η αξιοποίηση αυτής της δομής είναι ακριβώς το αντικείμενο της *μείωσης διαστατικότητας*.

Παρατήρηση 1.1 (Τυποποίηση της 'ουσιαστικής διάστασης'). Η διαισθητική έννοια της εγγενούς διάστασης μπορεί να οριοθετηθεί αυστηρά, αν τα δεδομένα δειγματοληπτούνται από μια *(ομαλή πολλαπλότητα* (smooth manifold) $\mathcal{M} \subset \mathbb{R}^m$, ως εγγενή διάσταση ορίζουμε την τοπολογική διάσταση $\dim \mathcal{M}$.Σε μετρικό πλαίσιο, μια χρήσιμη ποσότητα είναι η *διάσταση συσχέτισης* (correlation dimension)

$$d_{\text{corr}} = \lim_{r \to 0} \frac{\log C(r)}{\log r}, \qquad C(r) = \frac{2}{n(n-1)} \#\{(i,j) : i < j,\ \|x_i - x_j\| \le r\},$$

δηλαδή ο εκθέτης με τον οποίο το πλήθος των ζευγών εντος ακτινας r φθίνει καθώς $r \to 0$. Η *υπόθεση της πολλαπλότητας* (manifold hypothesis), ότι δηλαδή $d_{\text{corr}} \ll m$ για τα δεδομένα του πραγματικού κόσμου είναι και η συνθήκη που καθιστά τη μείωση διαστατικότητας τοσο θεμιτη όσο και αποτελεσματική.

1.2 Τι είναι η Μείωση Διαστατικότητας

Στην εξόρυξη δεδομένων συχνά έχουμε δεδομένα με χιλιάδες χαρακτηριστικά (ένα συνχό παράδειγμα ειναι τα pixels μιας εικόνας).Στόχος της μείωσης διαστατικότητας είναι να μετατρέψουμε αυτά τα δεδομένα από έναν χώρο υψηλών διαστάσεων $\mathbb{R}^m$ σε έναν χώρο χαμηλών διαστάσεων $\mathbb{R}^d$ (όπου $d \ll m$), διατηρώντας τις πιο σημαντικές πληροφορίες ή γεωμετρικές ιδιότητες των αρχικών δεδομένων. Πιο τυπικά, δίνεται ένας πίνακας δεδομένων (data matrix):

$$X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{m \times n}, \tag{1.1}$$

του οποίου κάθε στήλη $x_i \in \mathbb{R}^m$ είναι ένα δείγμα, και αναζητούμε ένα πιο χαμηλοδιάστατο ανάλογο

$$Y = [y_1, y_2, \ldots, y_n] \in \mathbb{R}^{d \times n}, \qquad d \ll m, \tag{1.2}$$

που να αποτελεί faithful representation (*πιστή αναπαράσταση* του X. Επιδιώκουμε δηλαδή μια απεικόνιση

$$\Psi : x \in \mathbb{R}^m \longrightarrow y = \Psi(x) \in \mathbb{R}^d. \tag{1.3}$$

Η απεικόνιση Ψ δεν είναι πάντοτε ρητή, η ελάχιστη απαίτηση είναι να μπορούμε να την εφαρμόσουμε στα δεδομένα μας, δηλαδή να βρούμε έναν αντιπρόσωπο $y_i \in \mathbb{R}^d$ για κάθε δείγμα x_i.

Γιατί το κάνουμε: Συχνά υπάρχει η παρανόηση ότι η μείωση διαστατικότητας γίνεται αποκλειστικά για λόγους εξοικονόμησης αποκλειστικά υπολογιστικού κόστους.Στην πραγματικότητα εξυπηρετεί τεσσερις,αλληλενδετους και εν μέρει αλληλεπικαλυπτόμενους, σκοπούς:

(i) **Οπτικοποίηση**(visualization): η προβολή σε $d = 2$ ή $d = 3$ διαστάσεις επιτρέπει την οπτική αναγνώριση δομών και συστάδων (clusters).

(ii) **Μείωση θορύβου**(denoising): οι λιγότερο σημαντικές κατευθύνσεις συχνά αντιστοιχούν σε θόρυβο και με την απόρριψή τους 'καθαρίζουμε' τα δεδομένα.

(iii) **Εξαγωγή χαρακτηριστικών** (feature extraction): τα διανύσματα y_i ονομάζονται και *χαρακτηριστικά* και χρησιμοποιούνται ως είσοδος σε επόμενες εργασίες (π.χ. την ταξινόμηση).

(iv) **Υπολογιστικό κόστος**:η εργασία σε $\mathbb{R}^d$ αντί $\mathbb{R}^m$ μειώνει δραστικά το χρόνο και την μνήμη.

Η έννοια της 'πιστότητας': Το κρίσιμο σημείο αλλα και η πηγή της ποικιλίας των μεθόδων,είναι ο ορισμός του τι εννοούμε 'πιστή αναπαράσταση'.Διαφορετικά κριτήρια οδηγούν σε διαφορετικές μεθόδους.Ενδεικτικά,μπορεί να θέλουμε να διατηρήσουμε τη *συνολική διασπορά* των δεδομένων ή τις *ζευγαρωτές αποστάσεις* $\|x_i - x_j\|$ ή τις *τοπικές γειτονιές* κάθε σημείου ή τις *γωνίες* μεταξύ γειτονικών διανυσμάτων. Όπως θα δούμε, η επιλογή του κριτηρίου είναι αυτή που τελικά καθορίζει αν μια γραμμική μέθοδος όπως η SVD/PCA θα πετύχει ή θα "σπάσει".

1.3 Η Διάσπαση Ιδιαζουσών Τιμών (SVD)

Η Διάσπαση Ιδιαζουσών Τιμών(Singular Value Decomposition) είναι το κύριο εργαλείο της αριθμητικής γραμμικής άλγεβρας πάνω στο οποίο βασίζονται οι γραμμικές μέθοδοι μείωσης διάστασης.Διαισθητικά, η SVD αποκαλύπτει τις κύριες 'κατευθύνσεις δράσης'ενός πίνακα και το πόση σημασία έχει η καθεμία.

Θεώρημα 1.2 (Διάσπαση Ιδιαζουσών Τιμών (SVD)). *[2] Για κάθε πίνακα $A \in \mathbb{R}^{m\times n}$ υπάρχουν ορθογώνιοι πίνακες $U \in \mathbb{R}^{m\times m}$ και $W \in \mathbb{R}^{n\times n}$, καθώς και ένας διαγώνιος πίνακας $\Sigma \in \mathbb{R}^{m\times n}$ με μη αρνητικά στοιχεία στην διαγώνιο*

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0, \qquad p = \min(m, n),$$

ώστε:

$$A = U\,\Sigma\,W^\top = \sum_{i=1}^{p} \sigma_i\, u_i w_i^\top. \tag{1.4}$$

Οι αριθμοί σ_i (ιδιάζουσες τιμές) είναι μονοσήμαντα ορισμένοι.Ενώ οι στήλες u_i του U και w_i του W ονομάζονται αριστερά *και* δεξιά ιδιάζοντα διανύσματα *αντίστοιχα.*

Απόδειξη: Ο πίνακας $A^\top A \in \mathbb{R}^{n\times n}$ είναι συμμετρικός και θετικά ημιορισμένος.
Από το φασματικό θεώρημα υπάρχει ορθοκανονική βάση ιδιοδιανυσμάτων $w_1, \ldots, w_n$ με ιδιοτιμές $\sigma_1^2 \geq \cdots \geq \sigma_n^2 \geq 0$, δηλαδή $A^\top A\, w_i = \sigma_i^2 w_i$.
Για κάθε i με $\sigma_i > 0$ ορίζουμε $u_i = Aw_i/\sigma_i$. Τότε

$$u_i^\top u_j = \frac{1}{\sigma_i\sigma_j} w_i^\top A^\top A\, w_j = \frac{\sigma_j^2}{\sigma_i\sigma_j} w_i^\top w_j = \delta_{ij},$$

άρα τα u_i είναι ορθοκανονικά και τα συμπληρώνουμε σε ορθοκανονική βάση του $\mathbb{R}^m$.
Εξ ορισμού $Aw_i = \sigma_i u_i$ (και $Aw_i = 0$ όταν $\sigma_i = 0$), το οποίο γραμμένο συγκεντρωτικά δίνει $AW = U\Sigma$, δηλαδή την (1.4). Η μοναδικότητα των σ_i προκύπτει από τη μοναδικότητα των ιδιοτιμών του $A^\top A$. □

Παρατήρηση 1.3 (Σύνδεση με ιδιοτιμές). Από την απόδειξη φαίνεται ότι

$$A^\top A = W\Sigma^\top\Sigma\, W^\top, \qquad AA^\top = U\Sigma\Sigma^\top U^\top,$$

δηλαδή τα σ_i^2 είναι ακριβώς οι (μη μηδενικές) ιδιοτιμές τόσο του $A^\top A$ όσο και του $AA^\top$, τα w_i τα ιδιοδιανύσματα του $A^\top A$ και τα u_i εκείνα του $AA^\top$. *Η SVD είναι,κατ' ουσίαν, ταυτόχρονη φασματική ανάλυση των δύο αυτών συμμετρικών πινάκων.*

Γεωμετρική ερμηνεία: Κάθε γραμμικός μετασχηματισμός $x \mapsto Ax$ απεικονίζει τη μοναδιαία σφαίρα του $\mathbb{R}^n$ σε ένα υπερελλειψοειδές στον $\mathbb{R}^m$. Οι ιδιάζουσες τιμές σ_i είναι τα μήκη των ημιαξόνων αυτού του υπερελλειψοειδούς, τα u_i οι διευθύνσεις τους, και τα w_i οι προεικόνες τους. Έτσι το σ_1 μετρά τη "μέγιστη επιμήκυνση" που επιφέρει ο A, και η ακολουθία $\sigma_1 \geq \sigma_2 \geq \ldots$ καταγράφει, κατά φθίνουσα σειρά σημασίας, πόσο ενεργός είναι ο πίνακας σε κάθε κατεύθυνση. Πιο απλά η μεγαλύτερη ιδιάζουσα τιμή σ_1 δείχνει την κατεύθυνση στην οποία ο πίνακας παραμορφώνει περισσότερο τα δεδομένα.Οι υπόλοιπες ιδιάζουσες τιμές δείχνουν, με σειρά σημαντικότητας, πόσο επηρεάζει ο πίνακας τις άλλες κατευθύνσεις..

Rank και βασικοί υπόχωροι. Ο αριθμός r των μη μηδενικών ιδιαζουσών τιμών ισούται με την τάξη $\mathrm{rank}(A)$.Επιπλέον, τα πρώτα r διανύσματα u_i παράγουν το $\mathrm{Range}(A)$, ενώ τα διανύσματα w_i με $\sigma_i = 0$ παράγουν τον $\mathrm{Null}(A)$, η SVD δίνει δηλαδή ορθοκανονικές βάσεις και για τους τέσσερις θεμελιώδεις υπόχωρους.Τέλος, ορίζει φυσιολογικά τον 'ψευδοαντίστροφο'(ο αντιστροφος κατα προσεγγιση πινακας) $A^\dagger = W\Sigma^\dagger U^\top$, όπου $\Sigma^\dagger$ προκύπτει αντιστρέφοντας τις μη μηδενικές σ_i.

Πλήρης έναντι οικονομικής (thin) SVD. Η μορφή του Θεωρήματος 1.2 είναι η *πλήρης* SVD, με $U \in \mathbb{R}^{m\times m}$ και $W \in \mathbb{R}^{n\times n}$ τετραγωνικούς ορθογώνιους.Στην πράξη, αν $r = \mathrm{rank}(A)$, αρκεί η *οικονομική* (thin/reduced) μορφή

$$A = U_r\,\Sigma_r\,W_r^\top, \qquad U_r \in \mathbb{R}^{m\times r},\ \Sigma_r = \mathrm{diag}(\sigma_1,\dots,\sigma_r) \succ 0,\ W_r \in \mathbb{R}^{n\times r},$$

όπου κρατάμε μόνο τις μη μηδενικές ιδιάζουσες τιμές. Οι στήλες του U_r είναι ορθοκανονική βάση του $\mathrm{Range}(A)$ και του W_r ορθοκανονική βάση του $\mathrm{Range}(A^\top) = \mathrm{Null}(A)^\perp$.

Λήμμα 1.4 (Νόρμες πινάκων μέσω ιδιαζουσών τιμών)**.** *Για κάθε $A \in \mathbb{R}^{m\times n}$ με ιδιάζουσες τιμές $\sigma_1 \ge \cdots \ge \sigma_p \ge 0$ ισχύουν*

$$\|A\|_2 = \sigma_1, \qquad \|A\|_F = \Big(\sum_{i=1}^{p}\sigma_i^2\Big)^{1/2} = \big(\,\mathrm{Tr}(A^\top A)\big)^{1/2}.$$

όπου η 1η νόρμα (Νορμα 2) κοιτάζει μόνο τη μέγιστη επιμήκυνση, αν $\sigma_1 = 10$, τότε υπάρχει κάποια κατεύθυνση στην οποία ο πίνακας μπορεί να πολλαπλασιάσει το μήκος ενός διανύσματος μέχρι και 10 φορές.Η 2η νορμα (νόρμα Frobenius) λαμβάνει υποψην όλες τις ιδιάζουσες τιμές και όχι μόνο τη μεγαλύτερη.Ενω το trace (ίχνος) που είναι ίσο με την νορμα frobenius είναι είναι το άθροισμα των στοιχείων της κύριας διαγωνίου του $A^\top A$.

Επιπλέον και οι δύο νόρμες είναι αναλλοίωτες υπό ορθογώνιους μετασχηματισμούς: $\|PAQ\| = \|A\|$ για κάθε ορθογώνιους P, Q.

Απόδειξη. Η φασματική νόρμα είναι $\|A\|_2 = \max_{\|x\|=1}\|Ax\|$.
Γράφοντας $x = \sum_i \xi_i w_i$ στη βάση των δεξιών ιδιαζόντων διανυσμάτων, $Ax = \sum_i \sigma_i\xi_i u_i$, άρα
$\|Ax\|^2 = \sum_i \sigma_i^2\xi_i^2 \le \sigma_1^2\sum_i\xi_i^2 = \sigma_1^2$, με ισότητα για $x = w_1$.
Για τη Frobenius, $\|A\|_F^2 = \mathrm{Tr}(A^\top A) = \mathrm{Tr}(W\Sigma^\top\Sigma W^\top) = \mathrm{Tr}(\Sigma^\top\Sigma) = \sum_i\sigma_i^2$, με χρήση της κυκλικότητας του ίχνους και $W^\top W = I$. Η αναλλοιότητα προκύπτει αφού ο PAQ έχει SVD $(PU)\Sigma(Q^\top W)^\top$ με τις ίδιες σ_i. □

Πρόταση 1.5 (Χαρακτηρισμός min-max των ιδιαζουσών τιμών)**.** *Έστω $A \in \mathbb{R}^{m\times n}$ με $\sigma_1 \ge \cdots \ge \sigma_p \ge 0$, $p = \min(m,n)$. Τότε για κάθε $1 \le k \le p$,*

$$\sigma_k = \max_{\substack{\mathcal{S}\subseteq\mathbb{R}^n\\ \dim\mathcal{S}=k}}\ \min_{\substack{x\in\mathcal{S}\\ x\neq 0}}\frac{\|Ax\|}{\|x\|} = \min_{\substack{\mathcal{T}\subseteq\mathbb{R}^n\\ \dim\mathcal{T}=n-k+1}}\ \max_{\substack{x\in\mathcal{T}\\ x\neq 0}}\frac{\|Ax\|}{\|x\|}.$$

Δηλαδη: Στο πρώτο σκέλος :
Διαλέγω έναν χώρο S με k διαστάσεις.
Μέσα σε αυτόν κοιτάω το χειρότερο δυνατό διάνυσμα (αυτό που 'τεντώνεται' λιγότερο), το οποίο υπολογίζω με βάση τον λόγο επιμύκησης, $\frac{\|Ax\|}{\|x\|}$.
Παίρνω τον καλύτερο δυνατό χώρο S, ώστε ακόμα και το χειρότερο διάνυσμα να έχει όσο

γίνεται μεγάλυτερη επιμήκυνση και το αποτέλεσμα ειναι το σ_k.

Ενώ στο 2ο μέρος κάνω την αντίθετη διαδικασία
Ψάχνω¨δηλαδή έναν χώρο όπου δεν υπάρχει κατεύθυνση που να δίνει μεγαλύτερη επιμήκυνση από σ_k
Παίρνω έναν μεγάλο χώρο T διαστάσεων $n-k+1$.
Βρίσκω μέσα του το διάνυσμα που τεντώνεται περισσότερο.
Επιλέγω τον χώρο που κάνει αυτό το μέγιστο όσο μικρότερο γίνεται.

Απόδειξη. Επειδή $\|Ax\|^2 = x^\top A^\top A\, x$ και οι σ_k^2 είναι οι ιδιοτιμές του συμμετρικού θετικά ημιορισμένου $A^\top A$, ο ισχυρισμός είναι το θεώρημα Courant−Fischer εφαρμοσμένο στον $A^\top A$, με λόγο Rayleigh $\dfrac{x^\top A^\top A\, x}{x^\top x} = \dfrac{\|Ax\|^2}{\|x\|^2}$· οι τετραγωνικές ρίζες δίνουν το ζητούμενο. □

Πόρισμα 1.6 (Ευστάθεια ιδιαζουσών τιμών − ανισότητα Weyl)**.** *[10] Για κάθε $A, E \in \mathbb{R}^{m\times n}$ και κάθε k,*

$$\left|\sigma_k(A+E) - \sigma_k(A)\right| \le \|E\|_2 .$$

Δηλαδή Αν προσθέσουμε στον πίνακα A ενα σφάλμα E , τότε οι ιδιάζουσες τιμές του πίνακα αλλάζουν το πολύ όσο μεγάλη είναι αυτή η αλλαγή.

Απόδειξη. Από την Πρόταση 1.5, για κάθε μοναδιαίο x ισχύει
$\|(A+E)x\| \le \|Ax\| + \|E\|_2$
παίρνοντας ελάχιστα και μεγιστοποιώντας ως προς k-διάστατους υπόχωρους,
$\sigma_k(A+E) \le \sigma_k(A) + \|E\|_2$.
Εναλλάσσοντας $A \leftrightarrow A+E$ προκύπτει η αντίστροφη ανισότητα. □

Παρατήρηση 1.7 (Δείκτης κατάστασης). Για αντιστρέψιμο τετραγωνικό A ($r = n = m$), ο *δείκτης κατάστασης* ως προς τη φασματική νόρμα είναι:
$\kappa_2(A) = \|A\|_2 \, \|A^{-1}\|_2 = \sigma_1/\sigma_n$.
Μεγάλο κ_2 σημαίνει σχεδόν μη αναστρέψιμο πίνακα − σχεδόν ιδιάζοντα πίνακα και αριθμητική αστάθεια, οπου η SVD τον εκθέτει ρητά αφού δίνει ολόκληρο το φάσμα των σ_i.
Ο δείκτης κατάστασης μας δείχνει πόσο επικίνδυνος είναι ένας πίνακας για υπολογισμούς: όσο μεγαλύτερη η διαφορά μεταξύ της μεγαλύτερης(σ_1) και μικρότερης(σ_n) ιδιάζουσας τιμής, τόσο πιο ασταθής γίνεται.

1.4 Κατηγορίες Μεθόδων μείωσης διάσπασης

Διαχωρτίζουμε τις μεθόδους σε δύο μεγάλες κατηγορίες, ανάλογα με τη φύση της απεικόνισης Ψ:

Χαρακτηριστικό	**Γραμμικές / Προβολικές**	**Μη Γραμμικές / Implicit**
Μηχανισμός	Υπολογίζουν ρητό πίνακα προβολής V· το νέο δεδομένο είναι $y = V^\top x$.	Δεν υπάρχει ρητή συνάρτηση προβολής· υπολογίζουν απευθείας τα δεδομένα Y.
Πλεονέκτημα	Προβάλλουν εύκολα νέα, άγνωστα δεδομένα (out-of-sample).	Ξεδιπλώνουν πολύπλοκες, μη γραμμικές δομές (manifolds).
Παραδείγματα	PCA, LDA, LPP, ONPP	LLE, Laplacian Eigenmaps, ISOMAP

Πίνακας 1.1: Οι δύο οικογένειες μεθόδων μείωσης διαστατικότητας.

Γραμμικές / προβολικές μέθοδοι. Εδώ η απεικόνιση Ψ είναι γραμμική: ένας πίνακας $V \in \mathbb{R}^{m\times d}$ μετασχηματίζει ρητά τα δεδομένα σε $Y = V^\top X$. Αφού 'μάθουμε' τον V, κάθε νέο δείγμα προβάλλεται απλώς με $y = V^\top x$ εξ ου και η ευκολία επέκτασης σε out-of-sample δεδομένα.Σύμφωνα με την Ενότητα 2.5, όλες αυτές οι μέθοδοι έχουν τη μορφή

$$\min_{V^\top BV=I} \operatorname{Tr}\big(V^\top X A X^\top V\big), \tag{1.5}$$

όπου ο πυρήνας A διαφέρει σε καθε μέθοδο (π.χ. για την PCA λαμβάνουμε $A = -I$ και μεγιστοποιούμε, για την LPP(Locality Preserving Projections) ένας Λαπλασιανός πίνακας γράφου,για την ONPP(Orthogonal Neighborhood Preserving Projections) ο πίνακας της LLE(Locally Linear Embedding).

Μη γραμμικές / implicit μέθοδοι. Εδώ δεν υπάρχει ρητή απεικόνιση υπολογίζουμε κατευθείαν το Y λύνοντας:

$$\min_{YBY^\top=I} \operatorname{Tr}\big(YAY^\top\big). \tag{1.6}$$

Οι μεθόδοι αυτές ξεκινούν χτίζοντας έναν *γράφο γειτνίασης* (affinity graph) $G = (V, E)$, του οποίου οι κόμβοι είναι τα δείγματα και οι ακμές συνδέουν 'κοντινούς' κόμβους (π.χ. τους k πλησιέστερους γείτονες, k-NN graph), με βάρη w_{ij}.Ο γράφος αυτός κωδικοποιεί την *τοπική* γεωμετρία, και το αντίστοιχο ιδιοπρόβλημα ξεδιπλώνει τη δομή.
Δύο αντιπροσωπευτικές περιπτώσεις:

- **LLE** (Locally Linear Embedding): κάθε δείγμα εκφράζεται ως γραμμικός συνδυασμός των γειτόνων του, $x_i \approx \sum_j w_{ij} x_j$ με $\sum_j w_{ij} = 1$, και ζητάμε το Y να σέβεται τις ίδιες σχέσεις, ελαχιστοποιώντας

  $$F_{\mathrm{LLE}}(Y) = \operatorname{Tr}\big[\, Y(I-W)^\top (I-W) Y^\top \,\big], \quad YY^\top = I.$$

  Ο πίνακας $M = (I-W)^\top(I-W)$ ονομάζεται *πίνακας LLE*.

- **Laplacian Eigenmaps**: επιβάλλει ποινή στη μεταφορά γειτονικών κόμβων σε μακρινά σημεία, ελαχιστοποιώντας
$\sum_{i,j} w_{ij} \|y_i - y_j\|^2 = 2 \operatorname{Tr}[Y(D-W)Y^\top]$ υπό $YDY^\top = I$,
όπου $D = \operatorname{diag}(\sum_j w_{ij})$ και $L = D - W$ ο *Λαπλασιανός* του γράφου.

1.5 Επίπεδα Επίβλεψης

Πέρα από τη γραμμική/μη γραμμική διάκριση, οι μέθοδοι διαφοροποιούνται ανάλογα με τη διαθεσιμότητα *ετικετών κλάσης* (class labels):

- **Μη εποπτευόμενη** (unsupervised): δεν υπάρχουν ετικέτες, ο αλγόριθμος (όπως η PCA ή το LLE) βασίζεται αποκλειστικά στη γεωμετρική δομή των δεδομένων.

- **Ημι-εποπτευόμενη** (semi-supervised): συνδυάζει έναν μικρό αριθμό δεδομένων με ετικέτες και ένα μεγάλο πλήθος δεδομένων χωρίς ετικέτες.Αποτελεί μια ιδιαίτερα ρεαλιστική προσέγγιση στις σύγχρονες εφαρμογές, όπου η χειροκίνητη σήμανση (labeling) είναι κοστοβόρα, ενώ τα μη επισημασμένα(unlabeled) δεδομένα είναι άφθονα.

- **Εποπτευόμενη** (supervised): κάθε δείγμα συνοδεύεται από μια ετικέτα που υποδεικνύει την κλάση του (π.χ. "αυτή η εικόνα είναι το ψηφίο 7"). Μέθοδοι όπως η LDA (Linear Discriminant Analysis) αξιοποιούν αυτή την πληροφορία για να μεγιστοποιήσουν τον διαχωρισμό μεταξύ των κλάσεων.

Στην εποπτευόμενη περίπτωση, η πληροφορία των κλάσεων ενσωματώνεται απευθείας στον γράφο γειτνίασης. Μια τυπική επιλογή είναι η κατασκευή ενός *εποπτευόμενου γράφου* (supervised graph), όπου δύο κόμβοι συνδέονται με ακμή *αν και μόνο αν* ανήκουν στην ίδια κλάση. Ο αντίστοιχος πίνακας γειτνίασης $H \in \mathbb{R}^{n\times n}$ λαμβάνει τότε μια μπλοκ-διαγώνια μορφή:

$$H = \operatorname{diag}[H_1, H_2, \ldots, H_c] \tag{1.7}$$

όπου c είναι το πλήθος των κλάσεων,
n_j το πλήθος των δειγμάτων της j-οστής κλάσης
και κάθε μπλοκ ορίζεται ως $H_j = \frac{1}{n_j}\mathbf{1}_{n_j}\mathbf{1}_{n_j}^\top$.

Όπως θα δούμε στη συνέχεια, ο πίνακας H αποδεικνύεται *ορθογώνιος προβολέας* ($H^2 = H$ και $H^\top = H$), γεγονός που οδηγεί σε κομψές και χρήσιμες θεωρητικές ισοδυναμίες για παράδειγμα, η κλασική LDA ταυτίζεται μαθηματικά με τις εποπτευόμενες εκδοχές των γραφο-μεθόδων (όπως το supervised LPP).

Συγκεκριμένα, η LDA επιλέγει τις κατευθύνσεις προβολής a που **μεγιστοποιούν** τον λόγο της διασποράς μεταξύ των κλάσεων (S_B) προς την διασπορά εντός των κλάσεων (S_W):

$$\max_a \frac{a^\top S_B\, a}{a^\top S_W\, a}, \tag{1.8}$$

το οποίο αποτελεί ένα γενικευμένο πρόβλημα Rayleigh quotient (λόγου ιχνών) της μορφής (2.11).

Κεφάλαιο 2

Η μέθοδος PCA σαν πρόβλημα βελτιστοποίησης

2.1 Η SVD ως βέλτιστη γραμμική συμπίεση

Η ιδιότητα που καθιστά την SVD τον ακρογωνιαίο λίθο της *γραμμικής* μείωσης διάστασης είναι το ακόλουθο θεώρημα:
Αν θέλουμε να προσεγγίσουμε τον A με έναν πίνακα μικρής τάξης, η καλύτερη δυνατή επιλογή είναι απλώς να 'κόψουμε' το άθροισμα (1.4).

Θεώρημα 2.1 (Eckart-Young-Mirsky)**.** *[3]*
Έστω $A = \sum_{i=1}^{p} \sigma_i u_i w_i^\top$ *η SVD του* A
και για $k \leq r = \operatorname{rank}(A)$ *θέτουμε:*

$$A_k = \sum_{i=1}^{k} \sigma_i u_i w_i^\top = U_k \Sigma_k W_k^\top .$$

Τότε ο A_k *ελαχιστοποιεί την απόσταση* $\|A - B\|$ *ως προς* B *τάξης το πολύ* k*, τόσο για τη νόρμα Frobenius όσο και για τη φασματική νόρμα. Επιπλέον,*

$$\|A - A_k\|_F = \Big(\sum_{i=k+1}^{p} \sigma_i^2 \Big)^{1/2}, \qquad \|A - A_k\|_2 = \sigma_{k+1}. \tag{2.1}$$

Απόδειξη. Γράφουμε την SVD ως
$A = \sum_{i=1}^{p} \sigma_i u_i w_i^\top$ με $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$.
Το υπόλοιπο της περικοπής,
$A - A_k = \sum_{i=k+1}^{p} \sigma_i u_i w_i^\top$,
είναι άθροισμα ορθοκανονικών όρων τάξης 1, οπότε αμέσως

$$\|A - A_k\|_2 = \sigma_{k+1}, \qquad \|A - A_k\|_F^2 = \sum_{i=k+1}^{p} \sigma_i^2.$$

Μένει να δείξουμε ότι καμία B με $\operatorname{rank}(B) \leq k$ δεν δίνει μικρότερο σφάλμα.
Φασματική νόρμα. Έστω B με $\operatorname{rank}(B) \leq k$, οπότε $\dim \operatorname{Null}(B) \geq n - k$. Ο υπόχωρος $\mathcal{S} = \operatorname{span}\{w_1, \ldots, w_{k+1}\}$ έχει διάσταση $k + 1$, άρα

$$\dim \operatorname{Null}(B) + \dim \mathcal{S} \;\geq\; (n - k) + (k + 1) = n + 1 > n.$$

Συνεπώς οι δύο υπόχωροι τέμνονται μη τετριμμένα: υπάρχει μοναδιαίο $z \in \mathrm{Null}(B) \cap \mathcal{S}$, $z = \sum_{i=1}^{k+1} c_i w_i$ με $\sum_i c_i^2 = 1$. Τότε $Bz = 0$, ενώ

$$\|Az\|^2 = \Big\| \sum_{i=1}^{k+1} c_i \sigma_i u_i \Big\|^2 = \sum_{i=1}^{k+1} c_i^2 \sigma_i^2 \;\geq\; \sigma_{k+1}^2 \sum_{i=1}^{k+1} c_i^2 = \sigma_{k+1}^2,$$

αφού $\sigma_i \geq \sigma_{k+1}$ για $i \leq k+1$. Επομένως

$$\|A - B\|_2 \;\geq\; \|(A-B)z\| = \|Az\| \;\geq\; \sigma_{k+1} = \|A - A_k\|_2 .$$

Νόρμα Frobenius. Έστω B με $\mathrm{rank}(B) \leq k$ και έστω $Q \in \mathbb{R}^{m \times k}$ με ορθοκανονικές στήλες, του οποίου το πεδίο τιμών περιέχει το $\mathrm{Range}(B)$. Για σταθερό $\mathrm{Range}(Q)$, η πλησιέστερη (κατά Frobenius) επιλογή με $\mathrm{Range}(B) \subseteq \mathrm{Range}(Q)$ είναι η ορθογώνια προβολή $B = QQ^\top A$: πράγματι $A - QQ^\top A \perp QQ^\top A - B$, οπότε

$$\|A - B\|_F^2 = \big\|A - QQ^\top A\big\|_F^2 + \big\|QQ^\top A - B\big\|_F^2 \;\geq\; \big\|A - QQ^\top A\big\|_F^2 .$$

Άρα, ελαχιστοποιώντας πρώτα ως προς B και έπειτα ως προς Q,

$$\min_{\mathrm{rank}(B)\leq k} \|A - B\|_F^2 = \min_{Q^\top Q = I} \big\|A - QQ^\top A\big\|_F^2 = \min_{Q^\top Q = I} \Big(\|A\|_F^2 - \mathrm{Tr}\big[Q^\top A A^\top Q\big] \Big),$$

όπου χρησιμοποιήσαμε $\big\|A - QQ^\top A\big\|_F^2 = \|A\|_F^2 - \big\|Q^\top A\big\|_F^2$ και $\big\|Q^\top A\big\|_F^2 = \mathrm{Tr}[Q^\top A A^\top Q]$. Αρκεί λοιπόν να *μεγιστοποιήσουμε* το $\mathrm{Tr}[Q^\top A A^\top Q]$ υπό $Q^\top Q = I$. Από την αρχή μεγιστοποίησης ίχνους (Θεώρημα 2.7, που αποδεικνύεται στην Ενότητα 2.5)[1] , το μέγιστο ισούται με το άθροισμα των k μεγαλύτερων ιδιοτιμών του $AA^\top$, δηλαδή $\sum_{i=1}^k \sigma_i^2$, και επιτυγχάνεται για $Q = U_k = [u_1, \ldots, u_k]$. Συνεπώς

$$\min_{\mathrm{rank}(B)\leq k} \|A - B\|_F^2 = \|A\|_F^2 - \sum_{i=1}^{k} \sigma_i^2 = \sum_{i=1}^{p} \sigma_i^2 - \sum_{i=1}^{k} \sigma_i^2 = \sum_{i=k+1}^{p} \sigma_i^2 = \|A - A_k\|_F^2 ,$$

με το ελάχιστο να επιτυγχάνεται για $B = U_k U_k^\top A = A_k$. □

Παρατήρηση 2.2 (Γενίκευση Mirsky και η αρχή πίσω από το αποτέλεσμα). Το Θεώρημα 2.1 ισχύει, στην πραγματικότητα, για *κάθε* ορθογώνια αναλλοίωτη νόρμα $\|\cdot\|$ (δηλαδή με $\|PAQ\| = \|A\|$ για ορθογώνιους P, Q) — αυτή είναι η εκδοχή *Mirsky*. Ο λόγος είναι ότι κάθε τέτοια νόρμα εξαρτάται *μόνο* από το διάνυσμα των ιδιαζουσών τιμών $\sigma(A)$ και είναι αύξουσα ως προς αυτό κατά την έννοια της κυριαρχίας (majorization). Καθώς $\sigma(A - B)$ κυριαρχείται κάτωθεν από το $(0, \ldots, 0, \sigma_{k+1}, \sigma_{k+2}, \ldots)$ όταν $\mathrm{rank}(B) \leq k$ (συνέπεια των ανισοτήτων Weyl, Πόρισμα 1.6), η ελάχιστη τιμή κάθε ορθογώνια αναλλοίωτης νόρμας επιτυγχάνεται από την περικομμένη SVD. Έτσι η βελτιστότητα της A_k δεν είναι ιδιαιτερότητα των νορμών Frobenius/φασματικής αλλά γενικό γεγονός.

Παρατήρηση 2.3 (Αριθμητική τάξη). Λόγω θορύβου και πεπερασμένης ακρίβειας, οι "μηδενικές" ιδιάζουσες τιμές είναι απλώς πολύ μικρές. Ορίζουμε την *αριθμητική τάξη* σε ανοχή ε ως $r_\varepsilon = \#\{i : \sigma_i > \varepsilon\}$. Η (2.1) δίνει τότε $\|A - A_{r_\varepsilon}\|_2 \leq \varepsilon$, συνδέοντας τη γεωμετρική ιδέα της "σχεδόν χαμηλής τάξης" με ένα μετρήσιμο αριθμητικό κριτήριο.

Με άλλα λόγια, η *περικομμένη* SVD (truncated SVD) τάξης k είναι η **βέλτιστη προσέγγιση χαμηλής τάξης**, και το σφάλμα της ελέγχεται πλήρως από τις ιδιάζουσες τιμές που απορρίπτουμε. Αυτό το γεγονός κρύβει και τον λόγο της αποτυχίας της: αν οι ιδιάζουσες τιμές φθίνουν αργά (δηλαδή τα δεδομένα δεν είναι "σχεδόν χαμηλής τάξης"), τότε κανένα μικρό k δεν δίνει μικρό σφάλμα. Θα επανέλθουμε σε αυτό στην Ενότητα 3.1.

2.2 SVD και Βέλτιστη Προσέγγιση Χαμηλής Τάξης

Τι κανει η svd; Η SVD 'σπάει' κάθε πίνακα σε άθροισμα πινάκων τάξης 1, ταξινομημένων κατά φθίνουσα σημασία $\sigma_1 \geq \sigma_2 \geq \dots$.
Απο τους πρώτους k όρους παίρνουμε τη βέλτιστη προσέγγιση τάξης k (Eckart–Young), ενώ το σφάλμα είναι η υπολειπόμενη ενέργεια των ιδιαζουσών τιμών που δεν κρατάμε.

$$A = U\Sigma W^\top = \sum_{i=1}^{p} \sigma_i u_i w_i^\top, \qquad A_k = \sum_{i=1}^{k} \sigma_i u_i w_i^\top, \qquad \|A - A_k\|_F^2 = \sum_{i>k} \sigma_i^2. \tag{2.2}$$

Έστω ο πίνακας

$$A = \begin{pmatrix} 2 & 2 \\ 1 & -1 \end{pmatrix}.$$

Τότε: $A^\top A = \left(\begin{smallmatrix} 5 & 3 \\ 3 & 5 \end{smallmatrix}\right)$ με ιδιοτιμές 8 και 2,
άρα $\sigma_1 = 2\sqrt{2},\ \sigma_2 = \sqrt{2}$ και με δεξί ιδιάζον διάνυσμα $w_1 = \frac{1}{\sqrt{2}}(1,1)^\top$.
Έτσι υπολογίζουμε $u_1 = \frac{1}{\sigma_1} A w_1 = \frac{1}{2\sqrt{2}} \cdot \frac{1}{\sqrt{2}} \left(\begin{smallmatrix} 4 \\ 0 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right)$
και τέλος η προσέγγιση rank(τάξης) 1 θα είναι:

$$A_1 = \sigma_1 u_1 w_1^\top = 2\sqrt{2} \cdot \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 0 \end{pmatrix} (1,1) = \begin{pmatrix} 2 & 2 \\ 0 & 0 \end{pmatrix}, \qquad A - A_1 = \begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix},$$

με $\|A - A_1\|_F = \sqrt{1+1} = \sqrt{2} = \sigma_2$, όπως περιμέναμε και απο την (2.2).

Το σφάλμα $\sum_{i>k} \sigma_i^2$, όπως έχουμε πει είναι η ενέργεια που χάνεται και με βάση αυτό μπορούμε να ορίσουμε το *ποσοστό ενέργειας που διατηρίεται*, ως τον λόγο: $\left(\sum_{i\leq k} \sigma_i^2\right) / \left(\sum_i \sigma_i^2\right)$, με αυτον τον τρόπο έχουμε κάποιο κριτήριο επιλογής του k. Εδώ παρατηρούμαι το πρόβλημα που είπαμε στην προηγούμενη ενότητα για το τί συμβαίνει αν τα δεδομένα δεν φθήνουν γρήγορα , όταν η δομή είναι μη γραμμική.

2.3 PCA: Η SVD ως Μέθοδος Μείωσης Διάστασης

Η πιο γνωστή γραμμική μέθοδος είναι η *Ανάλυση Κύριων Συνιστωσών* (Principal Component Analysis, PCA), η οποία δεν είναι παρά η SVD εφαρμοσμένη στα *κεντραρισμένα* δεδομένα.
Συμβολίζουμε με $\mu = \frac{1}{n}\sum_i x_i$ τον μέσο
και με

$$\bar{X} = X\left(I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^\top\right)$$

τον πίνακα των κεντραρισμένων δεδομένων ($\bar{x}_i = x_i - \mu$).

Λήμμα 2.4 (Ο τελεστής κεντραρίσματος)**.** *Ο πίνακας $J = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top \in \mathbb{R}^{n\times n}$ είναι ορθογώνιος προβολέας: $J^\top = J$ και $J^2 = J$. Έχει $J\mathbf{1} = 0$ και $\operatorname{rank}(J) = n-1$, και προβάλλει κάθε διάνυσμα στον υπόχωρο $\mathbf{1}^\perp = \{v : \mathbf{1}^\top v = 0\}$. Συνεπώς ο $\bar{X} = XJ$ έχει στήλες με μηδενικό μέσο, και $\bar{X}\mathbf{1} = 0$.*

Απόδειξη. Η συμμετρία είναι προφανής. Επειδή $\mathbf{1}^\top\mathbf{1} = n$,

$$J^2 = \left(I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^\top\right)^2 = I - \tfrac{2}{n}\mathbf{1}\mathbf{1}^\top + \tfrac{1}{n^2}\mathbf{1}(\mathbf{1}^\top\mathbf{1})\mathbf{1}^\top = I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^\top = J.$$

Άρα J είναι προβολέας· $J\mathbf{1} = \mathbf{1} - \frac{1}{n}\mathbf{1}(\mathbf{1}^\top\mathbf{1}) = 0$, και η εικόνα του είναι το $\mathbf{1}^\perp$ διάστασης $n-1$. Τέλος $\bar{X} = XJ$ αφαιρεί από κάθε στήλη τον μέσο $\mu = \frac{1}{n}X\mathbf{1}$. □

Η PCA αναζητά έναν ορθοκανονικό πίνακα $V \in \mathbb{R}^{m\times d}$ που *μεγιστοποιεί τη διακύμανση* των προβεβλημένων δειγμάτων $y_i = V^\top x_i$:

$$\max_{\substack{V\in\mathbb{R}^{m\times d}\\ V^\top V = I}} \sum_{i=1}^{n} \Big\| y_i - \tfrac{1}{n}\textstyle\sum_j y_j \Big\|_2^2 . \tag{2.3}$$

Αναπτύσσοντας, η αντικειμενική συνάρτηση γράφεται ως ίχνος,

$$F_{\mathrm{PCA}}(V) = \operatorname{Tr}\big[\, V^\top \bar{X}\bar{X}^\top V \,\big], \tag{2.4}$$

οπότε το πρόβλημα γίνεται

$$\max_{V^\top V = I} \operatorname{Tr}\big[\, V^\top \bar{X}\bar{X}^\top V \,\big]. \tag{2.5}$$

Όπως θα δούμε αμέσως παρακάτω (Θεώρημα 2.7), η λύση είναι ο πίνακας των d *αριστερών ιδιαζόντων διανυσμάτων* του $\bar{X}$ που αντιστοιχούν στις μεγαλύτερες ιδιάζουσες τιμές — ισοδύναμα, τα d κορυφαία ιδιοδιανύσματα του πίνακα συνδιακύμανσης $\bar{X}\bar{X}^\top$. Αν $\bar{X} = U\Sigma Z^\top$ είναι η SVD, τότε $V = U_d$ και

$$Y = U_d^\top \bar{X} = \Sigma_d Z_d^\top . \tag{2.6}$$

Παραγωγή μέσω πολλαπλασιαστών Lagrange. Για $d = 1$ το πρόβλημα (2.5) γίνεται $\max_{\|v\|=1} v^\top \bar{X}\bar{X}^\top v$. Με $\mathcal{L}(v,\lambda) = v^\top \bar{X}\bar{X}^\top v - \lambda(v^\top v - 1)$ και $\nabla_v \mathcal{L} = 2\bar{X}\bar{X}^\top v - 2\lambda v = 0$ προκύπτει

$$\bar{X}\bar{X}^\top v = \lambda v,$$

δηλαδή κάθε στατιστικό σημείο είναι ιδιοδιάνυσμα του πίνακα συνδιακύμανσης, με τιμή της αντικειμενικής $v^\top \bar{X}\bar{X}^\top v = \lambda$. Άρα το μέγιστο δίνεται από το κορυφαίο ιδιοδιάνυσμα u_1 ($\lambda_1 = \sigma_1^2$)· οι επόμενες συνιστώσες προκύπτουν διαδοχικά υπό ορθογωνιότητα, με το γενικό αποτέλεσμα να είναι η αρχή μεγιστοποίησης ίχνους (Θεώρημα 2.7).

Παρατήρηση 2.5 (Διατηρούμενη διασπορά). Η συνολική διασπορά των κεντραρισμένων δεδομένων ισούται με $\frac{1}{n}\operatorname{Tr}(\bar{X}\bar{X}^\top) = \frac{1}{n}\sum_i \sigma_i^2 = \frac{1}{n}\sum_i \lambda_i$. Η PCA με d συνιστώσες διατηρεί κλάσμα $\rho_d = \big(\sum_{i\le d}\lambda_i\big)/\big(\sum_i \lambda_i\big)$ της συνολικής διασποράς· η "αργή" άνοδος του ρ_d προς το 1 είναι ακριβώς το σύμπτωμα μη γραμμικής δομής.

Πρόταση 2.6 (Ισοδυναμία διασποράς – σφάλματος ανακατασκευής)**.** *Η μεγιστοποίηση της προβεβλημένης διασποράς* (2.5) *είναι ισοδύναμη με την ελαχιστοποίηση του σφάλματος ανακατασκευής:*

$$\min_{V^\top V = I} \big\| \bar{X} - VV^\top \bar{X} \big\|_F^2 .$$

Απόδειξη. Επειδή $\big\|\bar{X} - VV^\top\bar{X}\big\|_F^2 = \operatorname{Tr}[\bar{X}^\top\bar{X}] - \operatorname{Tr}[V^\top\bar{X}\bar{X}^\top V]$ και ο πρώτος όρος είναι σταθερός, η ελαχιστοποίηση του αριστερού μέλους ισοδυναμεί με τη μεγιστοποίηση του $\operatorname{Tr}[V^\top\bar{X}\bar{X}^\top V]$. □

Συνδυάζοντας με το Θεώρημα 2.1, βλέπουμε ότι **η PCA είναι ακριβώς η περικομμένη SVD των κεντραρισμένων δεδομένων**: παρέχει τον βέλτιστο d-διάστατο *γραμμικό* υπόχωρο, υπό την έννοια της ελάχιστης τετραγωνικής απόστασης ανακατασκευής. Αυτή είναι η "ιδανική" περίπτωση όπου η SVD δουλεύει τέλεια· το ζητούμενο της εργασίας είναι να κατανοήσουμε πότε και γιατί παύει να αρκεί.

2.4 PCA: Μεγιστοποίηση Διασποράς

Η PCA είναι η (2.12) με $A = \bar{X}\bar{X}^\top$ (ο μη κανονικοποιημένος πίνακας συνδιακύμανσης). Επιλέγει τους άξονες μέγιστης διακύμανσης· ισοδύναμα (Πρόταση στο Κεφ. 1), ελαχιστοποιεί το τετραγωνικό σφάλμα ανακατασκευής. Είναι, όπως είπαμε, η περικομμένη SVD των κεντραρισμένων δεδομένων.

Ιδανική περίπτωση. Έχουμε 3 σημεία στο επίπεδο: $x_1 = (-2, -1),\ x_2 = (0, 0),\ x_3 = (2, 1)$ ενώ ο μέσος είναι $(0, 0)$, άρα $\bar{X} = \left(\begin{smallmatrix} -2 & 0 & 2 \\ -1 & 0 & 1 \end{smallmatrix}\right)$ και

$$\bar{X}\bar{X}^\top = \begin{pmatrix} 8 & 4 \\ 4 & 2 \end{pmatrix}, \qquad \lambda_1 = 10,\ \lambda_2 = 0,$$

με $u_1 = \frac{1}{\sqrt{5}}(2, 1)^\top$.
Τα σημεία βρίσκονται *ακριβώς* πάνω στην ευθεία κατεύθυνσης $(2, 1)$, άρα η πρώτη συνιστώσα συγκεντρώνει το 100% της διακύμανσης ($\lambda_2 = 0$): μία διάσταση αρκεί απόλυτα.Σε αυτό το εξαιρετικά σπάνιο σενάριο η SVD λειτουργεί τέλεια.

$$\max_{V^\top V = I} \operatorname{Tr}\big[V^\top \bar{X}\bar{X}^\top V\big], \qquad Y = U_d^\top \bar{X}, \qquad \bar{X} = X\big(I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^\top\big). \tag{2.7}$$

Μη ιδανικό σενάριο Έστω τώρα σημεία ομοιόμορφα πάνω στε εναν *κύκλο* $x = \cos\theta,\ y = \sin\theta,\ \theta \in [0, 2\pi)$. Με τον μέσο να είναι $(0, 0)$ και

$$\mathrm{Var}(x) = \mathbb{E}[\cos^2\theta] = \tfrac{1}{2}, \quad \mathrm{Var}(y) = \tfrac{1}{2}, \quad \mathrm{Cov}(x, y) = \mathbb{E}[\cos\theta \sin\theta] = 0,$$

άρα ο πίνακας συνδιακύμανσης μας είναι $\frac{1}{2}I$: *ισότροπος*, με ιδιοτιμές $\frac{1}{2}, \frac{1}{2}$.

Εδώ συναντάμε τα εξής 2 προβλήματα:
(i) η PCA δεν έχει *καμία* προτιμητέα κατεύθυνση, εφόσον καμία συμπίεση δεν είναι καλύτερη, και
(ii) οποιαδήποτε προβολή σε 1 διάσταση *συγχωνεύει* αντιδιαμετρικά σημεία (π.χ. τα θ και $-\theta$ απεικονίζονται στο ίδιο x). Παρότι η εγγενής διάσταση της καμπύλης είναι 1, η γραμμική προβολή αποτυγχάνει να δώσει μια πιστή μονοδιάστατη αναπαράσταση.

Απο αυτην την αντιπαράθεση φαίνεται και ολόκληρη η ιστορία της εργασίας σε μια μικρογραφία.
Στην ευθεία, ο πίνακας συνδιακύμανσης έχει *τάξη* 1 ($\lambda_2 = 0$) και η γραμμική τάξη συμπίπτει με την εγγενή διάσταση σε αντίθεση με τον κύκλο οπου η εγγενής διάσταση είναι 1 αλλά ο πίνακας συνδιακύμανσης είναι *πλήρους τάξης* με ίσες ιδιοτιμέ και έτσι η μη γραμμικότητα χωρίζει τη δομή σε όλες τις γραμμικές κατευθύνσεις.
Το ποσοστό διατηρούμενης διακύμανσης ($\lambda_1/(\lambda_1 + \lambda_2)$) είναι 100% στην πρώτη περίπτωση και 50% στη δεύτερη.
Εδώ η SVD αποτυγχάνει πλήρως και είναι και το ποιο απλό παράδειγμα κινήτρου της χρήσης μη γραμμικών μεθόδων.

2.5 Ο Μαθηματικός Πυρήνας: Βελτιστοποίηση Ίχνους (Trace Optimization)

Το παραπάνω δεν είναι σύμπτωση ειδικά για την PCA. Όπως αναδεικνύει η εργασία των Kokiopoulou, Chen & Saad, η αναζήτηση του χαμηλοδιάστατου χώρου καταλήγει σχεδόν

πάντα στο ίδιο μαθηματικό πρόβλημα: αναζητούμε έναν πίνακα προβολής V που *μεγιστοποιεί* (ή *ελαχιστοποιεί*) το ίχνος ενός γινομένου πινάκων, υπό έναν περιορισμό ορθογωνιότητας. Η γενική μορφή είναι

$$\max_{V^\top B V = I} \operatorname{Tr}\big(V^\top A V\big), \tag{2.8}$$

όπου A είναι ένας πίνακας που *κωδικοποιεί τις σχέσεις των δεδομένων* (π.χ. συνδιακύμανση ή γειτνίαση)
B ένας πίνακας που *ορίζει τον περιορισμό* (συχνά απλώς ο μοναδιαίος I), και
V ο ζητούμενος πίνακας προβολής.
Η λύση δίνεται από τα ιδιοδιανύσματα που αντιστοιχούν στις μεγαλύτερες (ή μικρότερες) ιδιοτιμές του *γενικευμένου προβλήματος ιδιοτιμών*

$$Au = \lambda B u. \tag{2.9}$$

Η θεμελιώδης αρχή πίσω από αυτό το αποτέλεσμα είναι η ακόλουθει:

Θεώρημα 2.7 (Μεγιστοποίηση ίχνους / αρχή Ky Fan)**.** *Έστω $A \in \mathbb{R}^{n\times n}$ συμμετρικός με ιδιοτιμές $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ και αντίστοιχα ορθοκανονικά ιδιοδιανύσματα $u_1, \ldots, u_n$. Τότε, για $V \in \mathbb{R}^{n\times d}$,*

$$\max_{V^\top V = I} \operatorname{Tr}\big(V^\top A V\big) = \lambda_1 + \cdots + \lambda_d, \tag{2.10}$$

και το μέγιστο επιτυγχάνεται για $V = [u_1, \ldots, u_d]$.

Απόδειξη. Γράφουμε $A = \sum_{i=1}^n \lambda_i u_i u_i^\top$ και θέτουμε $P = VV^\top$, τον ορθογώνιο προβολέα τάξης d στον χώρο των στηλών του V (αφού $V^\top V = I$). Τότε

$$\operatorname{Tr}(V^\top A V) = \operatorname{Tr}(AP) = \sum_{i=1}^n \lambda_i \, u_i^\top P u_i = \sum_{i=1}^n \lambda_i \, c_i, \qquad c_i := u_i^\top P u_i = \|P u_i\|^2 .$$

Επειδή ο P είναι προβολέας ισχύει $0 \leq c_i \leq 1$, ενώ $\sum_i c_i = \operatorname{Tr}(P) = d$. Μεγιστοποιούμε λοιπόν τη γραμμική συνάρτηση $\sum_i \lambda_i c_i$ πάνω σε αυτό το σύνολο περιορισμών· το μέγιστο επιτυγχάνεται δίνοντας βάρος 1 στις d μεγαλύτερες ιδιοτιμές και 0 στις υπόλοιπες, δηλαδή $\sum_i \lambda_i c_i \leq \lambda_1 + \cdots + \lambda_d$. Η ισότητα ισχύει όταν ο P προβάλλει στον $\operatorname{span}\{u_1, \ldots, u_d\}$. □

Πόρισμα 2.8 (Αρχή ελαχιστοποίησης ίχνους)**.** *Με τις υποθέσεις του Θεωρήματος 2.7, $\min_{V^\top V = I} \operatorname{Tr}(V^\top A V) = \lambda_{n-d+1} + \cdots + \lambda_n$, με ελάχιστο για $V = [u_{n-d+1}, \ldots, u_n]$.*

Απόδειξη. Εφαρμόζουμε το Θεώρημα 2.7 στον $-A$: $\max_{V^\top V = I} \operatorname{Tr}(V^\top(-A)V) = -(\lambda_{n-d+1} + \cdots + \lambda_n)$. Αυτή η μορφή είναι που χρειάζονται οι LLE/Eigenmaps, οι οποίες *ελαχιστοποιούν* το ίχνος. □

Παρατήρηση 2.9 (Η οπτική της κυριαρχίας (majorization)). Το επιχείρημα της απόδειξης - μεγιστοποίηση της $\sum_i \lambda_i c_i$ με $0 \leq c_i \leq 1$, $\sum_i c_i = d$ - είναι γραμμικό πρόγραμμα σε πολύτοπο με κορυφές τα $\{0,1\}$-διανύσματα με ακριβώς d μονάδες. Το βέλτιστο πιάνεται σε κορυφή, θέτοντας $c_i = 1$ στις d μεγαλύτερες ιδιοτιμές. Ισοδύναμα, η $\sum_i \lambda_i c_i$ είναι Schur–κοίλη ως προς το (c_i) (Hardy-Littlewood-Pólya), αυτός είναι ο κοινός λόγος για τον οποίο υπερισχύουν, πάντα, οι ακραίες ιδιοτιμές.

Λήμμα 2.10 (Θεώρημα διαχωρισμού Poincaré / παρεμβολή Cauchy)**.** *Έστω $A \in \mathbb{R}^{n\times n}$ συμμετρικός με ιδιοτιμές $\lambda_1 \geq \cdots \geq \lambda_n$, και $V \in \mathbb{R}^{n\times d}$ με $V^\top V = I$. Αν $\theta_1 \geq \cdots \geq \theta_d$ είναι οι ιδιοτιμές του $V^\top A V$, τότε*

$$\lambda_i \;\geq\; \theta_i \;\geq\; \lambda_{i+n-d}, \qquad i = 1, \ldots, d.$$

Απόδειξη. Από Courant–Fischer για τον $V^\top AV$, $\theta_i = \max_{\dim \mathcal{S}=i} \min_{0\neq y\in\mathcal{S}} \frac{y^\top (V^\top AV) y}{y^\top y}$. Θέτοντας $x = Vy$ (οπότε $\|x\| = \|y\|$ αφού $V^\top V = I$), το x διατρέχει i-διάστατο υπόχωρο του $\mathrm{Range}(V)$, και η ποσότητα γίνεται λόγος Rayleigh του A σε περιορισμένη οικογένεια υπόχωρων· σύγκριση με τη μεγιστοποίηση πάνω σε *όλους* τους i-διάστατους υπόχωρους δίνει $\theta_i \le \lambda_i$. Εφαρμόζοντας το ίδιο στον $-A$ προκύπτει $\theta_i \ge \lambda_{i+n-d}$. □

Παρατήρηση 2.11. Το Λήμμα 2.10 είναι το υπόβαθρο των *μεθόδων προβολής* (Rayleigh–Ritz) του Κεφαλαίου 3.4: όταν περιορίζουμε ένα ιδιοπρόβλημα σε υπόχωρο, οι προσεγγιστικές (Ritz) ιδιοτιμές *παρεμβάλλονται* ανάμεσα στις πραγματικές, και ταυτίζονται μ' αυτές ακριβώς όταν ο υπόχωρος είναι αναλλοίωτος.

Παρατήρηση 2.12 (Σημαντικότητα Υποχωρου). Η βέλτιστη λύση *δεν* είναι μοναδική: κάθε ορθοκανονική βάση του ιδιοχώρου των d μεγαλύτερων ιδιοτιμών είναι εξίσου βέλτιστη. Αυτό που έχει σημασία είναι ο *υπόχωρος*, όχι η συγκεκριμένη βάση. Το παρατηρούμε επανειλημμένα στις μεθόδους μείωσης διάστασης: ένας περιστροφικός μετασχηματισμός $\hat{V} = VQ$ (με Q ορθογώνιο) δεν αλλάζει το αποτέλεσμα.

Περιορισμός B-ορθογωνιότητας. Όταν ο περιορισμός είναι $V^\top BV = I$ με B συμμετρικό θετικά ορισμένο, η ανάλυση ανάγεται στην προηγούμενη με αλλαγή μεταβλητής. Θέτοντας $\tilde{V} = B^{1/2}V$ και $\tilde{A} = B^{-1/2}AB^{-1/2}$, το (2.8) γίνεται $\max_{\tilde{V}^\top\tilde{V}=I} \mathrm{Tr}(\tilde{V}^\top\tilde{A}\tilde{V})$, του οποίου η λύση (Θεώρημα 2.7) δίνεται από τα κορυφαία ιδιοδιανύσματα του $\tilde{A}$· επιστρέφοντας στο V προκύπτουν ακριβώς τα ιδιοδιανύσματα του γενικευμένου προβλήματος (2.9).

Το πρόβλημα λόγου ιχνών (trace ratio). Σε ορισμένες μεθόδους (π.χ. στη φυσική διατύπωση της LDA) εμφανίζεται το δυσκολότερο πρόβλημα

$$\max_{V^\top CV = I} \frac{\mathrm{Tr}(V^\top AV)}{\mathrm{Tr}(V^\top BV)}. \tag{2.11}$$

Συνήθως αντιμετωπίζεται μέσω της βοηθητικής συνάρτησης

$$f(\theta) = \max_{V^\top V = I} \mathrm{Tr}\big[V^\top (A - \theta B) V\big],$$

η οποία (όταν $B \succ 0$) είναι *γνησίως φθίνουσα* ως προς θ, με $f(0) > 0$ και $f(\theta) < 0$ για $\theta > \lambda_{\max}(A, B)$.
Από το Θεώρημα μεγιστοποίησης ίχνους (2.7), $f(\theta)$ ισούται με το άθροισμα των d μεγαλύτερων ιδιοτιμών του πινίου $A - \theta B$. Η βέλτιστη τιμή μ του λόγου είναι και η μοναδική ρίζα της f έτσι η επίλυση του (2.11) απαιτεί έναν εξωτερικό βρόχο αναζήτησης ρίζας, και κατά συνέπεια *περισσότερα* ιδιοπροβλήματα από το απλό (2.9).
Αυτό το μοτίβο 'βελτιστοποίηση ίχνους ⇒ (γενικευμένο) ιδιοπρόβλημα' θα μας συνοδεύει σε όλη την εργασία και μας επιτρέπει να συγκρίνουμε άμεσα μεθόδους που, με την πρώτη ματιά, φαίνονται ασύνδετες.

2.6 Το Γενικό Πρόβλημα Βελτιστοποίησης Ίχνους

Στην εξίσωση που ακωλουθεί (2.12) :
Ο A είναι συμμετρικός και κωδικοποιεί τις σχέσεις των δεδομένων δηλαδή την συνδιακύμανση ή την γειτνίαση και αναζητουμε d ορθοκανονικές κατευθύνσεις (άρα τις στήλες του V) που μεγιστοποιούν το συνολικό 'μέγεθος' του A πάνω στον υπόχωρο που ορίζουν.
Το ίχνος $\mathrm{Tr}(V^\top AV) = \sum_{i=1}^{d} v_i^\top A v_i$ είναι το άθροισμα των μορφών Rayleigh κατά μήκος των d αξόνων.
και η απάντηση είναι : να πάρουμε τους άξονες των d μεγαλύτερων ιδιοτιμών.

$$\max_{\substack{V\in\mathbb{R}^{n\times d}\\ V^\top V=I}} \mathrm{Tr}\big(V^\top AV\big) \;=\; \lambda_1+\lambda_2+\cdots+\lambda_d. \tag{2.12}$$

Έστω

$$A=\begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix}, \qquad d=1.$$

Το χαρακτηριστικό πολυώνυμο είναι $(2-\lambda)^2-1=(\lambda-3)(\lambda-1)$, άρα $\lambda_1=3,\ \lambda_2=1$, με ιδιοδιανύσματα $u_1=\frac{1}{\sqrt{2}}(1,1)^\top$ και $u_2=\frac{1}{\sqrt{2}}(1,-1)^\top$.
Για $v=u_1$:

$$v^\top Av=\tfrac{1}{2}\,(1,1)\begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix}\begin{pmatrix}1\\ 1\end{pmatrix}=\tfrac{1}{2}\,(1,1)\begin{pmatrix}3\\ 3\end{pmatrix}=\tfrac{1}{2}\cdot 6=3=\lambda_1.$$

Καμία άλλη μοναδιαία διεύθυνση δεν δίνει τιμή >3. Για $d=2$ παίρνουμε $V=I$ και $\mathrm{Tr}(V^\top AV)=\mathrm{Tr}(A)=\lambda_1+\lambda_2=4$.

Γιατί ομως νικάνε τα κορυφαία ιδιοδιανύσματα;
Θέτουμε $P=VV^\top$ (ορθογώνιος προβολέας τάξης d). Με $A=\sum_i \lambda_i u_i u_i^\top$ έχουμε

$$\mathrm{Tr}(V^\top AV)=\mathrm{Tr}(AP)=\sum_{i=1}^{n}\lambda_i\,\underbrace{u_i^\top P u_i}_{c_i}, \qquad 0\le c_i\le 1, \quad \sum_i c_i=\mathrm{Tr}(P)=d.$$

Μεγιστοποιούμε δηλαδή μια γραμμική συνάρτηση $\sum_i \lambda_i c_i$ πάνω σε ένα "κουτί-σιμπλεξ": η βέλτιστη επιλογή είναι να βάλουμε όλο το βάρος ($c_i=1$) στις d μεγαλύτερες ιδιοτιμές.
Αυτό έχει 2 σημαντικές συνέπειες:

- **Μη μοναδικότητα**: σημασία έχει ο *υπόχωρος* $\mathrm{span}\{u_1,\dots,u_d\}$, όχι η συγκεκριμένη βάση (όπως είπαμε και στην παρατήρηση 2.12). Κάθε περιστροφή VQ (Q ορθογώνιος) δίνει το ίδιο ίχνος.
- **Ελαχιστοποίηση** αντί μεγιστοποίησης (που χρειάζονται οι LLE/Eigenmaps): απλώς αντιστρέφουμε το επιχείρημα και κρατάμε τις d *μικρότερες* ιδιοτιμές.

Στατιστικά σημεία και πολλαπλασιαστές Lagrange. Πέρα από το επιχείρημα του "κουτιού–σιμπλεξ", η ίδια λύση προκύπτει από τις συνθήκες πρώτης τάξης.Με συμμετρικό πίνακα πολλαπλασιαστών $\Lambda\in\mathbb{R}^{d\times d}$ για τον περιορισμό $V^\top V=I$, $\mathcal{L}(V,\Lambda)=\mathrm{Tr}(V^\top AV)-\mathrm{Tr}[\Lambda(V^\top V-I)]$, και

$$\frac{\partial\mathcal{L}}{\partial V}=2AV-2V\Lambda=0 \quad\Longrightarrow\quad AV=V\Lambda.$$

Αφού ο Λ είναι συμμετρικός, μπορεί να διαγωνιοποιηθεί και διαγωνιοποιείται $\Lambda = Q\Theta Q^\top$ θέτωντας $\hat{V} = VQ$ (επίσης ορθοκανονικό, βλ. Παρατήρηση 2.12) παίρνουμε $A\hat{V} = \hat{V}\Theta$, δηλαδή οι στήλες του $\hat{V}$ είναι ιδιοδιανύσματα του A. Η τιμή της αντικειμενικής είναι $\mathrm{Tr}(\Theta) =$ άθροισμα των επιλεγμένων ιδιοτιμών, η ολική μεγιστοποίηση επιλέγει τις d μεγαλύτερες. Έτσι τα στατιστικά σημεία αντιστοιχούν σε *όλες* τις επιλογές d ιδιοδιανυσμάτων, ενώ το ολικό μέγιστο στις κορυφαίες d.

2.7 Περιορισμός B-Ορθογωνιότητας: Το Γενικευμένο Ιδιοπρόβλημα

Πολλές μέθοδοι δεν ζητούν απλή ορθοκανονικότητα ($V^\top V = I$) αλλά *B-ορθογωνιότητα* ($V^\top BV = I$), όπου $B \succ 0$ ορίζει μια σταθμισμένη γεωμετρία (π.χ. $B = XDX^\top$ στην LPP, ή $B = S_W$ στην LDA). Η λύση δεν είναι το συνηθισμένο πλεον ιδιοπρόβλημα, αλλά το *γενικευμένο*.

$$\max_{V^\top BV = I} \mathrm{Tr}\big(V^\top AV\big) \qquad \Longleftrightarrow \qquad Au_i = \lambda_i\, Bu_i. \tag{2.13}$$

Αν είχαμε τον πίνακα :

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}, \qquad d = 1.$$

Το $Au = \lambda Bu$ δίνει $\lambda = A_{ii}/B_{ii}$ για τις διαγώνιες διευθύνσεις: $\lambda^{(1)} = 2/1 = 2$ (διεύθυνση e_1), $\lambda^{(2)} = 1/4$ (διεύθυνση e_2).// 'Αρα η βέλτιστη κατεύθυνση είναι η e_1 (μεγαλύτερη γενικευμένη ιδιοτιμή), η οποία πρέπει να κανονικοποιηθεί ώστε $v^\top Bv = 1$
Εδώ $e_1^\top Be_1 = 1$, άρα $v = e_1$. Παρατήρουμε ότι το B 'τιμωρεί' τη δεύτερη διεύθυνση (μεγάλο $B_{22} = 4$), μειώνοντας τον λόγο της.

Ο μηχανισμός είναι μια αλλαγή μεταβλητής.
Αφού $B \succ 0$, ορίζουμε $\tilde{V} = B^{1/2}V$, και οπότε ο περιορισμός γίνεται $\tilde{V}^\top\tilde{V} = I$ και η αντικειμενική $\mathrm{Tr}(\tilde{V}^\top\tilde{A}\tilde{V})$ με $\tilde{A} = B^{-1/2}AB^{-1/2}$.
Από την (2.12), βέλτιστα είναι τα κορυφαία ιδιοδιανύσματα του $\tilde{A}$, $\tilde{A}\tilde{u}_i = \lambda_i\tilde{u}_i$. Επιστρέφοντας με $u_i = B^{-1/2}\tilde{u}_i$:

$$B^{-1/2}AB^{-1/2}\big(B^{1/2}u_i\big) = \lambda_i B^{1/2}u_i \implies Au_i = \lambda_i Bu_i.$$

Πρακτικά, αντί για το (αριθμητικά) επικίνδυνο $B^{-1/2}$, χρησιμοποιούμε παραγοντοποίηση Cholesky $B = RR^\top$ και $\hat{u} = R^\top u$.
Προσοχή πρεπει να δωθεί στις ιδιομορφίες: αν ο B είναι μόνο θετικά ημιορισμένος, το πρόβλημα παραμένει καλά ορισμένο μόνο όταν $\dim\mathrm{Null}(B) < d$ αν επιπλέον $A \succeq 0$, απαιτούμε $\mathrm{Null}(A) \cap \mathrm{Null}(B) = \{0\}$.

Λήμμα 2.13 (Ταυτόχρονη διαγωνιοποίηση ζεύγους πινάκων)**.** *Έστω A συμμετρικός και $B \succ 0$ στον $\mathbb{R}^{n\times n}$. Τότε υπάρχει αντιστρέψιμος $V = [v_1, \ldots, v_n]$ με*

$$V^\top BV = I, \qquad V^\top AV = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n),$$

όπου τα λ_i είναι πραγματικά και ικανοποιούν $Av_i = \lambda_i Bv_i$, με $v_i^\top Bv_j = \delta_{ij}$.

Απόδειξη. Αφού $B \succ 0$, ορίζεται $B^{1/2} \succ 0$ με αντίστροφο $B^{-1/2}$. Ο $\tilde{A} = B^{-1/2} A B^{-1/2}$ είναι συμμετρικός, άρα (φασματικό θεώρημα) $\tilde{V}^\top \tilde{A} \tilde{V} = \Lambda$, $\tilde{V}^\top \tilde{V} = I$, με πραγματικές λ_i. Θέτουμε $V = B^{-1/2}\tilde{V}$: τότε $V^\top B V = \tilde{V}^\top \tilde{V} = I$ και $V^\top A V = \tilde{V}^\top \tilde{A} \tilde{V} = \Lambda$, ενώ από $\tilde{A}\tilde{v}_i = \lambda_i \tilde{v}_i$ προκύπτει $A v_i = \lambda_i B v_i$. □

Παρατήρηση 2.14. Το Λήμμα 2.13 εξηγεί γιατί τα γενικευμένα ιδιοπροβλήματα (2.13) είναι εξίσου "καλά" με τα συνηθισμένα: το ζεύγος (A, B) με $B \succ 0$ ανάγεται πλήρως σε συμμετρικό πρόβλημα. Αριθμητικά χρησιμοποιείται η Cholesky $B = RR^\top$, λύνοντας το συμμετρικό $R^{-1} A R^{-\top} \hat{v} = \lambda \hat{v}$ με $\hat{v} = R^\top v$.

2.8 Το Πρόβλημα Λόγου Ιχνών (Trace Ratio)

Η "φυσική" διατύπωση μεθόδων όπως η LDA δεν είναι λόγος δύο ξεχωριστών μεγιστοποιήσεων αλλά μεγιστοποίηση ενός *λόγου* ιχνών. Είναι ουσιωδώς δυσκολότερο πρόβλημα από το (2.9), και λύνεται έμμεσα μέσω της βοηθητικής συνάρτησης $f(\theta)$.

$$\max_{V^\top C V = I} \frac{\operatorname{Tr}(V^\top A V)}{\operatorname{Tr}(V^\top B V)}, \qquad f(\theta) = \max_{V^\top V = I} \operatorname{Tr}\big[V^\top (A - \theta B) V\big]. \tag{2.14}$$

Σε μια μονοδιάστατη περίπτωση με $d = 1$ και $C = I$ ο λόγος γίνεται το *γενικευμένο πηλίκο Rayleigh*

$$\rho(v) = \frac{v^\top A v}{v^\top B v}.$$

Η μέγιστη τιμή του είναι ακριβώς η μεγαλύτερη γενικευμένη ιδιοτιμή $\lambda_{\max}(A, B)$, με βέλτιστο v το αντίστοιχο γενικευμένο ιδιοδιάνυσμα. Π.χ. για $A = \left(\begin{smallmatrix} 3 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, $B = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 2 \end{smallmatrix}\right)$: $\rho(e_1) = 3$, $\rho(e_2) = 1/2$, άρα $\max \rho = 3$. Στη μονοδιάστατη περίπτωση το πρόβλημα λόγου ιχνών ταυτίζεται με το γενικευμένο ιδιοπρόβλημα· η δυσκολία εμφανίζεται μόνο για $d > 1$.

Γιατί χρειάζεται η $f(\theta)$ για $d > 1$; Έστω μ η ζητούμενη μέγιστη τιμή του λόγου. Τότε $\operatorname{Tr}[V^\top (A - \mu B) V] \le 0$ για κάθε ορθογώνιο V, με ισότητα στη βέλτιστη λύση U· δηλαδή

$$\max_{V^\top V = I} \operatorname{Tr}\left[V^\top (A - \mu B) V\right] = 0, \qquad \text{άρα} \qquad f(\mu) = 0.$$

Όταν $B \succ 0$, η f είναι *γνησίως φθίνουσα*, με $f(0) > 0$ και $f(\theta) < 0$ για $\theta > \lambda_{\max}(A, B)$· επομένως έχει μοναδική ρίζα, την μ. Από την (2.12), $f(\theta)$ ισούται με το άθροισμα των d μεγαλύτερων ιδιοτιμών του πενίου $A - \theta B$.

Κόστος: η εύρεση του μ απαιτεί εξωτερικό βρόχο αναζήτησης ρίζας, δηλαδή *πολλά* ιδιοπροβλήματα αντί για ένα. Αυτός είναι ο λόγος που, στην πράξη, η απλούστερη μορφή (2.13) (που λύνει το πρόβλημα ως ένα γενικευμένο ιδιοπρόβλημα) προτιμάται συχνά ως προσέγγιση του (2.14).

Πρόταση 2.15 (Ιδιότητες της βοηθητικής συνάρτησης f)**.** *Έστω $f(\theta) = \max_{V^\top V = I} \operatorname{Tr}[V^\top (A - \theta B) V]$ με A, B συμμετρικούς και $B \succ 0$. Τότε: (i) η f είναι κυρτή και συνεχής (κατά σημείο supremum αφινικών συναρτήσεων του θ)· (ii) είναι γνησίως φθίνουσα, με $f'(\theta) = -\operatorname{Tr}(V_\theta^\top B V_\theta) < 0$ σε σημεία διαφορισιμότητας· (iii) έχει μοναδική ρίζα μ^* ίση με τη μέγιστη τιμή του λόγου ιχνών.*

Απόδειξη. (i) Για σταθερό V, η $\theta \mapsto \operatorname{Tr}(V^\top A V) - \theta \operatorname{Tr}(V^\top B V)$ είναι αφινική· το supremum οικογένειας αφινικών είναι κυρτή και συνεχής. (ii) Από το θεώρημα φακέλου (Danskin), $f'(\theta) = -\operatorname{Tr}(V_\theta^\top B V_\theta)$, που είναι < 0 αφού $B \succ 0$. (iii) $f(0) = \sum_{i \le d} \lambda_i(A) > 0$ (μη τετριμμένη περίπτωση), ενώ για μεγάλο θ ο $A - \theta B \prec 0$ δίνει $f(\theta) < 0$· μονοτονία και συνέχεια δίνουν μοναδική ρίζα μ^*. Στη ρίζα, $\operatorname{Tr}(V^\top A V) \le \mu^* \operatorname{Tr}(V^\top B V)$ για κάθε V, με ισότητα στον βέλτιστο, δηλαδή $\mu^* = \max_V \operatorname{Tr}(V^\top A V) / \operatorname{Tr}(V^\top B V)$. □

Επαναληπτικό σχήμα τύπου Dinkelbach. Η ρίζα της f βρίσκεται με την επανάληψη

$$V_k = \arg\max_{V^\top V = I} \operatorname{Tr}[V^\top (A - \theta_k B) V], \qquad \theta_{k+1} = \frac{\operatorname{Tr}(V_k^\top A V_k)}{\operatorname{Tr}(V_k^\top B V_k)},$$

όπου κάθε βήμα είναι ένα συνηθισμένο ιδιοπρόβλημα· η θ_k συγκλίνει μονότονα στο μ^* (ισοδύναμα, μέθοδος Newton στην f). *Λόγος ιχνών έναντι ίχνους λόγου*: το *trace ratio* διαφέρει από το ευκολότερο *ratio trace* $\max_V \operatorname{Tr}[(V^\top B V)^{-1}(V^\top A V)]$, που λύνεται με ένα γενικευμένο ιδιοπρόβλημα $Av = \lambda Bv$· συμπίπτουν για $d = 1$ αλλά εν γένει διαφέρουν.

Κεφάλαιο 3

Μη γραμμικές μέθοδοι (ως επεκτάσεις της PCA)

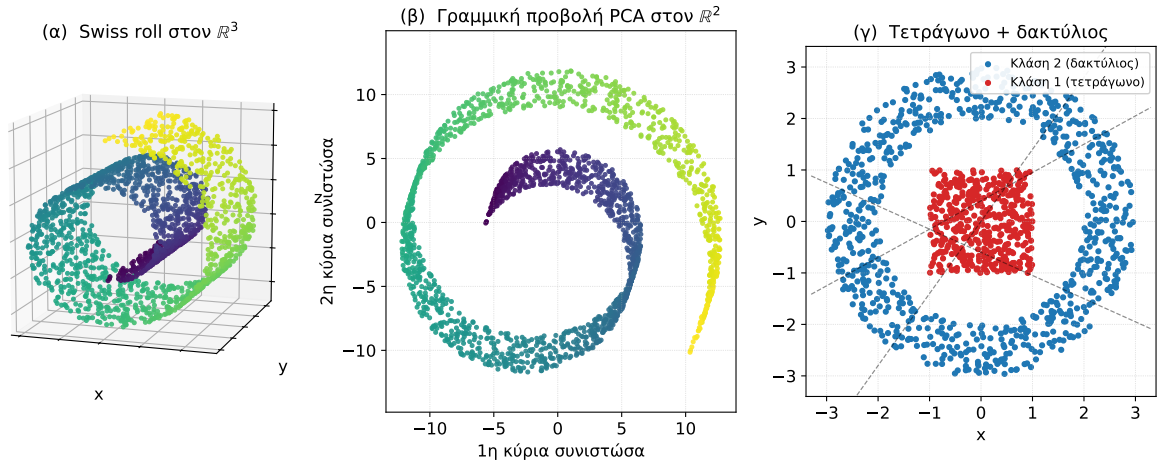
3.1 Τα Όρια των Γραμμικών Μεθόδων: Πότε "Σπάει" η SVD

Φτάνουμε τώρα στον πυρήνα του προβλήματος της παρούσας εργασίας. Είδαμε ότι η SVD/PCA είναι *βέλτιστη* — αλλά πάντοτε ως προς ένα *γραμμικό*, Ευκλείδειο κριτήριο: ελαχιστοποιεί το τετραγωνικό σφάλμα ανακατασκευής πάνω σε έναν *γραμμικό υπόχωρο* και διατηρεί τη συνολική διακύμανση. Το ερώτημα είναι: τι συμβαίνει όταν η εγγενής δομή των δεδομένων δεν είναι γραμμική;

Πρώτη όψη: αργά φθίνον φάσμα. Από τη σχέση σφάλματος (2.1), η PCA δίνει καλή d-διάστατη αναπαράσταση *μόνο* αν η "ενέργεια" $\sum_{i>d} \sigma_i^2$ είναι μικρή, δηλαδή αν τα δεδομένα είναι σχεδόν χαμηλής τάξης. Όταν όμως τα σημεία κάθονται πάνω σε μια *καμπύλη* πολλαπλότητα (π.χ. μια σπείρα ή ένα "swiss roll"), η δομή αυτή — αν και ουσιαστικά χαμηλοδιάστατη — "ξεδιπλώνεται" σε πολλές γραμμικές κατευθύνσεις του $\mathbb{R}^m$. Συνεπώς η διακύμανση σκορπίζεται σε πολλές ιδιάζουσες τιμές, το φάσμα φθίνει αργά, και κανένα μικρό d δεν επαρκεί. *Με αυτή την έννοια η SVD "σπάει": η γραμμική τάξη των δεδομένων είναι μεγάλη, ακόμη κι όταν η μη γραμμική, εγγενής διάστασή τους είναι μικρή.*

Δεύτερη όψη: λάθος κριτήριο πιστότητας. Ακόμη και όταν εφαρμοστεί, η PCA διατηρεί *Ευκλείδειες* αποστάσεις και *ολική* διακύμανση — όχι όμως τις *γεωδαισιακές* αποστάσεις πάνω στην πολλαπλότητα ούτε τις *τοπικές γειτονιές*. Δύο σημεία μπορεί να είναι Ευκλείδεια κοντά αλλά γεωδαισιακά μακριά (όπως δύο στρώματα ενός swiss roll), και η γραμμική προβολή τα "κολλάει" λανθασμένα μαζί. Η μεγιστοποίηση της διακύμανσης δεν συμπίπτει με τη διατήρηση της τοπικής γεωμετρίας ή με τον διαχωρισμό των κλάσεων.

Ένα κανονικό παράδειγμα. Θεωρούμε δύο κλάσεις στο επίπεδο: ένα μικρό τετράγωνο γύρω από την αρχή των αξόνων και έναν *δακτύλιο* (annulus) που το περιβάλλει. Οι δύο κλάσεις είναι σαφώς διακριτές, αλλά *δεν διαχωρίζονται από καμία ευθεία*. Οποιαδήποτε γραμμική μέθοδος — η PCA συμπεριλαμβανομένη — αποτυγχάνει να τις ξεχωρίσει, γιατί η ζητούμενη διαχωριστική επιφάνεια είναι εγγενώς καμπύλη. Αυτό το απλό παράδειγμα συμπυκνώνει την αδυναμία: *η γραμμικότητα της απεικόνισης Ψ, και όχι η αριθμητική της SVD, είναι ο περιορισμός.*



Σχήμα 3.1: Παραδείγματα μη γραμμικής δομής. (α) Το *swiss roll* είναι ουσιαστικά μονοδιάστατο (χρωματισμός κατά την εγγενή παράμετρο), αλλά (β) η γραμμική προβολή PCA το αφήνει "τυλιγμένο": σημεία με πολύ διαφορετική εγγενή θέση (διαφορετικό χρώμα) πέφτουν δίπλα-δίπλα — Ευκλείδεια κοντά αλλά γεωδαισιακά μακριά — ενώ η διακύμανση σκορπίζεται σχεδόν ισόποσα στους τρεις άξονες (η 2D προβολή κρατά μόλις ∼71%). (γ) Δύο σαφώς διακριτές κλάσεις (τετράγωνο και δακτύλιος) που *καμία ευθεία* δεν διαχωρίζει.

Η διαπίστωση αυτή θέτει το ερώτημα που διατρέχει την εργασία: *πώς μπορούμε να διορθώσουμε αυτή την αποτυχία, διατηρώντας ταυτόχρονα την υπολογιστική κομψότητα των ιδιοπροβλημάτων;*

3.2 Μη Γραμμικές Μέθοδοι Βασισμένες σε Γράφους και Τοπικότητα

3.2.1 LLE: Τοπικά Γραμμικές Σχέσεις

Η LLE κάνει μια τοπική υπόθεση: κάθε σημείο είναι περίπου γραμμικός συνδυασμός των γειτόνων του, με βάρη που αθροίζουν στη μονάδα.
Η (3.1) βρίσκει αυτά τα βάρη και η (3.2) βρίσκει τη χαμηλοδιάστατη ενσωμάτωση Y που *σέβεται τις ίδιες σχέσεις*.Με αυτον τον τροπο η LLE διατηρεί την τοπική γεωμετρία αντί της ολικής διακύμανσης.

$$\min_{w} \left\| x_i - \textstyle\sum_j w_{ij} x_j \right\|^2, \quad \sum_j w_{ij} = 1, \tag{3.1}$$

$$\min_{YY^\top = I} \operatorname{Tr}\big[Y\,(I-W)^\top (I-W)\,Y^\top \big], \qquad M = (I-W)^\top (I-W). \tag{3.2}$$

Υπολογισμός βαρών. Έστω $x_i = (0,0)$ με δύο γείτονες $x_1 = (1,0)$, $x_2 = (0,1)$. Ορίζουμε τον τοπικό πίνακα Gram $C_{jk} = (x_i - x_j)^\top (x_i - x_k)$:

$$x_i - x_1 = (-1,0),\ x_i - x_2 = (0,-1) \ \Rightarrow\ C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Η λύση είναι $w = \dfrac{C^{-1}\mathbf{1}}{\mathbf{1}^\top C^{-1}\mathbf{1}} = \dfrac{(1,1)^\top}{2} = \left(\tfrac{1}{2},\tfrac{1}{2}\right)$, δηλαδή η ανακατασκευή είναι το μέσο σημείο $(\tfrac{1}{2},\tfrac{1}{2})$ και το σφάλμα $\left\|(0,0)-(\tfrac{1}{2},\tfrac{1}{2})\right\|^2 = \tfrac{1}{2}$.
Το ίδιο αποτέλεσμα επίσης προκύπτει λύνοντας απευθείας το $\min(w_1^2+w_2^2)$ υπό $w_1+w_2=1$.

Ωστόσο η μέθοδος αυτή έχει 3 σημεία στα οποία αξίζει να σταθούμε με προσοχή:

- **Τοπικότητα και κόστος**: τα w_{ij} προκύπτουν από ένα μικρό σύστημα $k\times k$ (μόνο οι γείτονες), οπότε ο υπολογισμός είναι φθηνός. Μετά την ολοκλήρωσή του, $X \approx XW^\top$, δηλαδή οι στήλες του $X^\top$ είναι (κατά προσέγγιση) αριστερά μηδενοδιανύσματα του $I-W$.
- **Σταθερότητα των βαρών**: επειδή $\sum_j w_{ij}=1$, τα βάρη είναι αναλλοίωτα σε μεταφορές, περιστροφές και ομοιόθετες κλιμακώσεις της τοπικής γειτονιάς — ακριβώς οι μετασχηματισμοί κάτω από τους οποίους θέλουμε να διατηρείται η "μορφή".
- **Απόρριψη του σταθερού ιδιοδιανύσματος**: ο πίνακας $M=(I-W)^\top(I-W)$ έχει μηδενικά αθροίσματα γραμμών (αφού $(I-W)\mathbf{1}=0$), άρα το $\mathbf{1}$ είναι ιδιοδιάνυσμα με ιδιοτιμή 0. Αυτό αντιστοιχεί στη σταθερή (εκφυλισμένη) λύση και απορρίπτεται· η ενσωμάτωση είναι $Y=[u_2,\dots,u_{d+1}]^\top$, δηλαδή τα ιδιοδιανύσματα από το 2ο έως το $(d{+}1)$.

Πρόταση 3.1 (Κλειστός τύπος των βαρών LLE)**.** *Έστω x_i με γείτονες $\{\eta_1,\dots,\eta_k\}$ και $C\in\mathbb{R}^{k\times k}$ ο τοπικός πίνακας Gram $C_{ab}=(x_i-\eta_a)^\top(x_i-\eta_b)$. Το $\min_w \|x_i-\sum_a w_a\eta_a\|^2$ υπό $\sum_a w_a=1$ έχει λύση $w=\dfrac{C^{-1}\mathbf{1}}{\mathbf{1}^\top C^{-1}\mathbf{1}}$ (όταν C αντιστρέψιμος).*

Απόδειξη. Λόγω $\sum_a w_a=1$, $x_i-\sum_a w_a\eta_a=\sum_a w_a(x_i-\eta_a)$, άρα το σφάλμα είναι $w^\top Cw$. Με $w^\top Cw-2\lambda(\mathbf{1}^\top w-1)$ προκύπτει $Cw=\lambda\mathbf{1}$, δηλαδή $w=\lambda C^{-1}\mathbf{1}$· επιβάλλοντας $\mathbf{1}^\top w=1$, $\lambda=1/(\mathbf{1}^\top C^{-1}\mathbf{1})$. Στην πράξη λύνεται $Cw=\mathbf{1}$ και κανονικοποιείται. □

Κανονικοποίηση όταν $k>m$. Όταν οι γείτονες υπερβαίνουν τη διάσταση, ο C είναι (σχεδόν) ιδιάζων η συνήθης διόρθωση Tikhonov $C\leftarrow C+\frac{\delta\,\mathrm{Tr}(C)}{k}I$ ($\delta\ll 1$) ευνοεί ομαλά βάρη και αποκαθιστά την ευστάθεια. *Ο M είναι θετικά ημιορισμένος*: εξ ορισμού $z^\top Mz=\|(I-W)z\|^2\ge 0$, άρα $M\succeq 0$ και το (3.2) είναι καλά ορισμένο.

3.2.2 Laplacian Eigenmaps και ο Λαπλασιανός Γράφος

Τα Laplacian Eigenmaps βάζουν ποινή όταν *γειτονικοί* κόμβοι του γράφου απεικονίζονται *μακριά*. Η ελαχιστοποίηση της (3.3) γράφεται ως ίχνος του *Λαπλασιανού* $L=D-W$, οπότε ανάγεται πάλι σε (γενικευμένο) ιδιοπρόβλημα.

$$\min_{YDY^\top=I}\ \sum_{i,j} w_{ij}\,\|y_i-y_j\|^2 = 2\ \mathrm{Tr}\big[Y(D-W)Y^\top\big], \tag{3.3}$$

$$(D-W)\,u_i=\lambda_i D\,u_i, \qquad L=D-W,\quad D=\mathrm{diag}\Big(\textstyle\sum_j w_{ij}\Big). \tag{3.4}$$

ο Λαπλασιανός 'βλέπει' τις συστάδες. Δύο ζεύγη κόμβων με ακμές $1{-}2$ και $3{-}4$ (βάρος 1), χωρίς σύνδεση μεταξύ τους:

$$W=\begin{pmatrix}0&1&0&0\\1&0&0&0\\0&0&0&1\\0&0&1&0\end{pmatrix},\quad D=I,\quad L=D-W=\begin{pmatrix}1&-1&0&0\\-1&1&0&0\\0&0&1&-1\\0&0&-1&1\end{pmatrix}.$$

Ο L είναι μπλοκ-διαγώνιος με δύο μπλοκ $\left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$, καθένα με ιδιοτιμές 0 και 2.
Άρα το φάσμα του L είναι $\{0, 0, 2, 2\}$, και τα δύο μηδενικά ιδιοδιανύσματα είναι $(1, 1, 0, 0)^\top$ και $(0, 0, 1, 1)^\top$, σταθερά σε κάθε συνεκτική συνιστώσα. *Ο αριθμός των μηδενικών ιδιοτιμών ισούται με τον αριθμό των συνεκτικών συνιστωσών.*

Αν τώρα συνδέσουμε ασθενώς τις δύο συστάδες με ακμή $w_{23} = \varepsilon$ (μικρό), οι δύο μηδενικές ιδιοτιμές "σχίζονται": μία μένει ακριβώς 0 (με ιδιοδιάνυσμα το ολικό σταθερό $(1, 1, 1, 1)^\top$) και μία γίνεται μικρή $\lambda_2 = \mathcal{O}(\varepsilon)$, με ιδιοδιάνυσμα (*Fiedler vector*) προσήμων $(+, +, -, -)$, που διαχωρίζει τις δύο συστάδες.

- **Γιατί ο L είναι ιδιάζων**: εξ ορισμού $D_{ii} = \sum_j w_{ij}$, άρα $L\mathbf{1} = 0$· το σταθερό διάνυσμα είναι πάντα ιδιοδιάνυσμα με ιδιοτιμή 0 και απορρίπτεται (όπως στην LLE). Η πληροφορία διαχωρισμού βρίσκεται στα *επόμενα* μικρότερα ιδιοδιανύσματα.

- **Βάρη και η παράμετρος σ**: συνήθης επιλογή είναι ο πυρήνας θερμότητας $w_{ij} = \exp(-\|x_i - x_j\|^2/\sigma^2)$ (έναντι σταθερών βαρών 0/1). Το σ ρυθμίζει πόση σημασία δίνεται στους πιο κοντινούς γείτονες· είναι κρίσιμη παράμετρος και επιλέγεται συχνά ως το μισό της διαμέσου των ζευγαρωτών αποστάσεων.

- **Κανονικοποιημένος Λαπλασιανός**: με $\hat{Y} = YD^{1/2}$ και $\hat{W} = D^{-1/2}WD^{-1/2}$, η (3.3) απλοποιείται σε $\min_{\hat{Y}\hat{Y}^\top = I} \operatorname{Tr}[\hat{Y}(I - \hat{W})\hat{Y}^\top]$, όπου $\hat{L} = I - \hat{W} = D^{-1/2}LD^{-1/2}$ είναι ο *κανονικοποιημένος* Λαπλασιανός. Η σύγκριση της (3.3) με την (3.2) δείχνει πόσο κοντά είναι, στο πνεύμα, LLE και Eigenmaps: αλλάζει μόνο ο πίνακας μέσα στο ίχνος.

Παραγωγή της ταυτότητας ίχνους. Δείχνουμε την ισότητα της (3.3). Με συμμετρικό W και $D_{ii} = \sum_j w_{ij}$,

$$\sum_{i,j} w_{ij}\,\|y_i - y_j\|^2 = \sum_{i,j} w_{ij}\big(\|y_i\|^2 + \|y_j\|^2 - 2y_i^\top y_j\big) = 2\operatorname{Tr}(YDY^\top) - 2\operatorname{Tr}(YWY^\top) = 2\operatorname{Tr}(YLY^\top).$$

Πρόταση 3.2 (Φασματικές ιδιότητες του Λαπλασιανού)**.** *[13] Έστω γράφος με $w_{ij} = w_{ji} \geq 0$, $D = \operatorname{diag}(\sum_j w_{ij})$, $L = D - W$. Τότε: (i) $x^\top Lx = \frac{1}{2}\sum_{i,j} w_{ij}(x_i - x_j)^2 \geq 0$, άρα $L \succeq 0$· (ii) $L\mathbf{1} = 0$· (iii) η πολλαπλότητα της ιδιοτιμής 0 ισούται με το πλήθος των συνεκτικών συνιστωσών· (iv) ο γράφος είναι συνεκτικός ανν $\lambda_2 > 0$ (αλγεβρική συνεκτικότητα / τιμή Fiedler).*

Απόδειξη. (i) Ανάπτυξη: $\frac{1}{2}\sum_{i,j} w_{ij}(x_i^2 - 2x_ix_j + x_j^2) = x^\top Dx - x^\top Wx = x^\top Lx \geq 0$. (ii) $(L\mathbf{1})_i = D_{ii} - \sum_j w_{ij} = 0$. (iii) Αν $x^\top Lx = 0$, τότε $x_i = x_j$ όποτε $w_{ij} > 0$, άρα το x είναι σταθερό σε κάθε συνιστώσα· οι δείκτες των συνιστωσών παράγουν τον μηδενόχωρο. (iv) Άμεση συνέπεια του (iii). □

3.2.3 MDS / ISOMAP: Από Αποστάσεις σε Εσωτερικά Γινόμενα

Το (μετρικό) MDS θέλει να διατηρήσει τις ζευγαρωτές *αποστάσεις* $\|y_i - y_j\| \approx \|x_i - x_j\|$. Αντί όμως να δουλέψει με αποστάσεις, μεταφράζει το πρόβλημα σε *εσωτερικά γινόμενα*: ο πίνακας Gram G ανακατασκευάζεται από τις αποστάσεις μέσω της "διπλής κεντραρίσματος" (double centering), και η ενσωμάτωση προκύπτει από τη φασματική ανάλυση του G.

$$G = -\tfrac{1}{2}\left(I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^\top\right) S \left(I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^\top\right), \qquad \min_{Y \in \mathbb{R}^{d\times n}} \left\|G - Y^\top Y\right\|_F^2, \qquad Y = \Lambda_d^{1/2} Z_d^\top. \quad (3.5)$$

Σε ενα ορθογώνιο τρίγωνο 3-4-5: Τρία σημεία με τετραγωνικές αποστάσεις $s_{12} = 9,\ s_{13} = 16,\ s_{23} = 25$ (πίνακας S με μηδενική διαγώνιο). Με αθροίσματα γραμμών $r_1 = 25,\ r_2 = 34,\ r_3 = 41$ και ολικό άθροισμα 100, η συνιστώσα $(1,2)$ του G είναι

$$g_{12} = -\tfrac{1}{2}\Big(s_{12} - \tfrac{r_1}{3} - \tfrac{r_2}{3} + \tfrac{100}{9}\Big) = -\tfrac{1}{2}\Big(\tfrac{81-75-102+100}{9}\Big) = -\tfrac{1}{2}\cdot\tfrac{4}{9} = -\tfrac{2}{9}.$$

Αν τοποθετήσουμε τα σημεία ως $(0,0), (3,0), (0,4)$ και τα κεντράρουμε (μέσος $(1, \frac{4}{3})$), το πραγματικό εσωτερικό γινόμενο των δύο πρώτων κεντραρισμένων σημείων είναι $(-1)(2) + (-\frac{4}{3})(-\frac{4}{3}) = -2 + \frac{16}{9} = -\frac{2}{9}$ — ταυτίζεται με το g_{12} που υπολογίσαμε *μόνο* από αποστάσεις. Αυτή είναι η ουσία της (3.5).

Η λύση $Y = \Lambda_d^{1/2} Z_d^\top$ (κορυφαία d ιδιοδιανύσματα του G) δίνει *ακριβώς το ίδιο αποτέλεσμα με την PCA*: αν $\bar{X} = U\Sigma Z^\top$, τότε $Y = \Sigma_d Z_d^\top$. Η διαφορά είναι μόνο ο δρόμος — η PCA περνά από τον πίνακα συνδιακύμανσης ($m \times m$), το MDS από τον πίνακα Gram ($n \times n$) — και είναι χρήσιμη όταν έχουμε αποστάσεις αλλά όχι συντεταγμένες. Η μοναδικότητα ισχύει μόνο μέχρι ορθογώνιο μετασχηματισμό (περιστροφή/ανάκλαση δεν αλλάζει αποστάσεις). Το **ISOMAP** αντικαθιστά τις Ευκλείδειες με *γεωδαισιακές* αποστάσεις (συντομότερα μονοπάτια στον γράφο γειτνίασης), και έτσι "ξεδιπλώνει" καμπύλες πολλαπλότητες όπου η Ευκλείδεια απόσταση παραπλανά (π.χ. swiss roll).

Πρόταση 3.3 (Διπλό κεντράρισμα: από αποστάσεις σε εσωτερικά γινόμενα)**.** *Έστω σημεία $z_1, \dots, z_n$ με $S_{ij} = \|z_i - z_j\|^2$ και $Z = [z_1, \dots, z_n]$. Τότε*

$$G = -\tfrac{1}{2}\, JSJ, \qquad J = I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^\top,$$

ισούται με τον πίνακα Gram των κεντραρισμένων σημείων, $G_{ij} = \bar{z}_i^\top \bar{z}_j$· συνεπώς $G \succeq 0$ και $\operatorname{rank}(G) = \dim \operatorname{span}\{\bar{z}_1, \dots, \bar{z}_n\}$.

Απόδειξη. Γράφουμε $S = \mathbf{1}a^\top + a\mathbf{1}^\top - 2Z^\top Z$ με $a = (\|z_i\|^2)_i$. Επειδή $J\mathbf{1} = 0$, οι όροι τάξης 1 μηδενίζονται κατά τον πολλαπλασιασμό $J(\cdot)J$, και

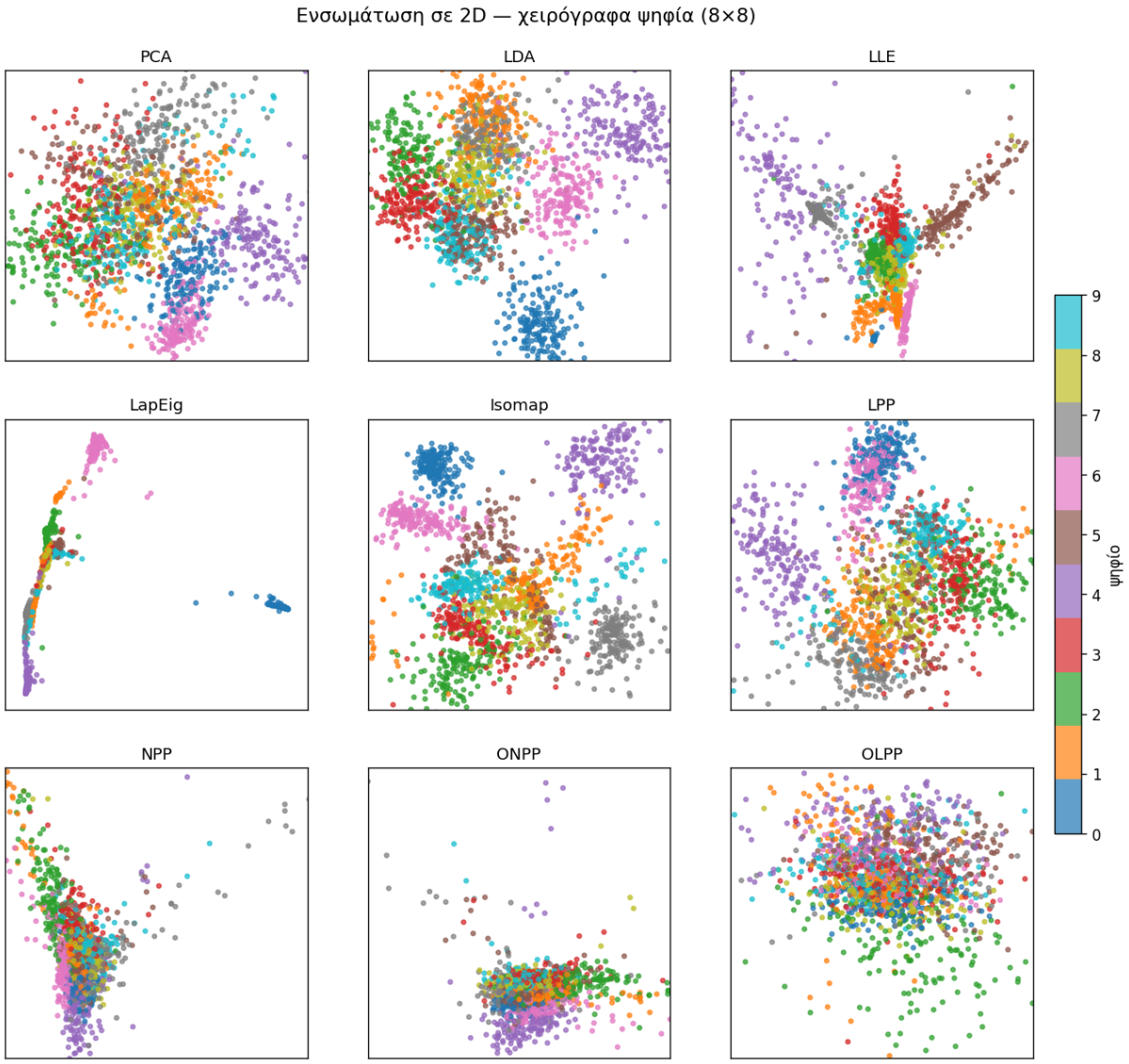
$$JSJ = -2\, JZ^\top ZJ = -2\,(ZJ)^\top (ZJ) = -2\,\bar{Z}^\top \bar{Z},$$

όπου $\bar{Z} = ZJ$ ο πίνακας των κεντραρισμένων σημείων (Λήμμα 2.4). Άρα $G = -\frac{1}{2}JSJ = \bar{Z}^\top \bar{Z} \succeq 0$, με τάξη ίση με τη διάσταση του χώρου που παράγουν τα $\bar{z}_i$. □

Πόρισμα 3.4 (Κλασικό MDS ≡ PCA)**.** *Αν τα $z_i = \bar{x}_i$ (πραγματικά Ευκλείδειες αποστάσεις), η ενσωμάτωση από τα κορυφαία d ιδιοδιανύσματα του G ταυτίζεται με τις συντεταγμένες PCA $\Sigma_d Z_d^\top$ της* (2.6) *(μέχρι ορθογώνιο μετασχηματισμό).*

Απόδειξη. Από την Πρόταση 3.3, $G = \bar{X}^\top \bar{X}$. Οι $\bar{X}^\top \bar{X}$ ($n \times n$) και $\bar{X}\bar{X}^\top$ ($m \times m$) έχουν το ίδιο μη μηδενικό φάσμα $\{\sigma_i^2\}$, και τα ιδιοδιανύσματα του πρώτου είναι τα δεξιά ιδιάζοντα διανύσματα Z_d του $\bar{X}$. Άρα $Y = \Lambda_d^{1/2} Z_d^\top = \Sigma_d Z_d^\top$, ταυτόσημο με την ενσωμάτωση PCA. □

Σε αυτό το σημείο θα ήθελα να σταθώ στο οτι το Isomap (Global Non-linear): είναι ξεκάθαρα ο νικητής των unsupervised μεθόδων στο τέστ που τρέξαμε έστω για τα unsupervised μοντέλα.



3.3 Γραμμικές Προβολές Διατήρησης Τοπικότητας

3.3.1 LPP:Locality Preserving Projections (Γραμμική Διατήρηση Τοπικότητας)

$$\min_{V^\top X D X^\top V = I} \operatorname{Tr}\big[V^\top X(D-W)X^\top V\big], \qquad X(D-W)X^\top v_i = \lambda_i\, XDX^\top v_i. \tag{3.6}$$

Η LPP θα μπορούσε να χαρακτηριστεί ως οτι είναι η γραμμική αδελφή των Eigenmaps: χρησιμοποιεί την ίδια αντικειμενική συνάρτηση (ποινή για απομάκρυνση γειτόνων), αλλά επιβάλλει $y_i = V^\top x_i$, δηλαδή ψάχνει *ρητό* πίνακα προβολής V. Έτσι αποκτά αυτό που λείπει από τα Eigenmaps: άμεση προβολή νέων (out-of-sample) δειγμάτων.

Αν ορίσουμε τον 'γράφο' ως πλήρη και ομοιόμορφο, $L = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ (ο προβολέας κεντραρίσματος), τότε $XLX^\top$ είναι ένας (κλιμακωμένος) πίνακας συνδιακύμανσης. Αγνοώντας τον περιορισμό, η LPP θα έδινε τις *χαμηλότερες* ιδιοκαταστάσεις αντί των υψηλότερων της PCA δηλαδή το "αντίστροφο" της PCA. Αυτό δείχνει πόσο στενά συνδέονται οι μέθοδοι μέσω της επιλογής του L.

Απο την σύγκριση των ιδιοπροβλημάτων (3.4) και (3.6) αποκαλύπτεται και ότι: **η LPP**

είναι μια προβολή τύπου Rayleigh-Ritz των Eigenmaps πάνω στον υπόχωρο $\mathrm{span}\{X^\top\}$.
Πράγματι, αναζητώντας λύση της μορφής $u = X^\top v$,δηλαδή στον χώρο που παράγουν οι γραμμές των δεδομένων, και επιβάλλοντας τις συνθήκες ορθογωνιότητας Galerkin, το $n \times n$ πρόβλημα των Eigenmaps μετατρέπεται στο $m \times m$ πρόβλημα της LPP. Πόρισμα: στην *υποδειγματοληπτημένη* περίπτωση ($m \geq n$ και $\mathrm{rank}(X) = n$), ο $\mathrm{span}\{X^\top\}$ είναι όλος ο χώρος και οι δύο μέθοδοι δίνουν *ταυτόσημο* αποτέλεσμα. Η κανονικοποιημένη εκδοχή ($\hat{X} = XD^{1/2}$) δίνει $\hat{X}(I - \hat{W})\hat{X}^\top v = \lambda\, \hat{X}\hat{X}^\top v$.

3.3.2 Orthogonal Neighborhood Preserving Projections (ONPP / NPP)

Οι ONPP/NPP κάνουν για το LLE ό,τι η LPP για τα Eigenmaps: χρησιμοποιούν τον πίνακα LLE M, αλλά με γραμμική προβολή $y_i = V^\top x_i$. Διαφέρουν μόνο στον περιορισμό — η ONPP ζητά *ορθογώνιο* V ($V^\top V = I$), η NPP *ορθογώνια προβεβλημένα δεδομένα* ($V^\top X X^\top V = I$, δηλαδή $YY^\top = I$ όπως στο LLE).

$$\underbrace{\min_{V^\top V = I} \mathrm{Tr}\big[V^\top X M X^\top V\big]}_{\text{ONPP}} \qquad \text{έναντι} \qquad \underbrace{\min_{V^\top X X^\top V = I} \mathrm{Tr}\big[V^\top X M X^\top V\big]}_{\text{NPP}}, \quad M = (I-W)^\top (I-W). \tag{3.7}$$

Επειδή $(I-W)\mathbf{1} = 0$, ο $\tilde{M} = XMX^\top$ κληρονομεί ιδιομορφίες: όταν $m > n$, ο $\tilde{M}$ (μεγέθους $m \times m$) έχει τάξη το πολύ n και είναι ιδιάζων. Όταν $m \leq n$, δεν είναι κατ' ανάγκη ιδιάζων, αλλά στην πράξη η απόρριψη της μικρότερης ιδιοτιμής βοηθά.

Με το ίδιο επιχείρημα προβολής όπως στην LPP: **η NPP είναι η προβολή του LLE στον** $\mathrm{span}\{X^\top\}$, και στην υποδειγματοληπτημενη περίπτωση τα δύο ταυτίζονται.
Η ONPP, αντίθετα, χρησιμοποιεί *ορθογώνια* προβολή προσπαθόντας να διατηρείσει καλύτερα τις αποστάσεις, αφού $\|V^\top(x-y)\| \leq \|x-y\|$), γι' αυτό συμπεριφέρεται πολύ καλά σε εργασίες ταξινόμησης (π.χ. αναγνώριση προσώπων), ενώ όμως δίνει φτωχότερες δισδιάστατες οπτικοποιήσεις.

3.3.3 LDA(Linear Discriminant Analysis): Εποπτευόμενος Διαχωρισμός Κλάσεων

Η LDA είναι εποπτευόμενη: αξιοποιεί τις ετικέτες για να προβάλει τα δεδομένα έτσι ώστε οι κλάσεις να *διαχωρίζονται* όσο γίνεται επίσης μεγιστοποιεί τον λόγο της διασποράς μεταξύ κλάσεων S_B προς την διασπορά εντός κλάσεων S_W ένα πρόβλημα λόγου ιχνών (§2.8).

$$S_B = \sum_{k=1}^{c} n_k(\mu^{(k)} - \mu)(\mu^{(k)} - \mu)^\top, \quad S_W = \sum_{k=1}^{c} \sum_{x_i \in X_k} (x_i - \mu^{(k)})(x_i - \mu^{(k)})^\top, \tag{3.8}$$

$$\max_a \frac{a^\top S_B\, a}{a^\top S_W\, a}, \qquad \bar{X}(I-H)\bar{X}^\top u_i = \lambda_i(\bar{X}\bar{X}^\top)u_i. \tag{3.9}$$

Παράδειγμα ,μονοδιάστατο, 2 κλάσεις. Κλάση 1: $\{0, 2\}$ (μέσος $\mu^{(1)} = 1$). Κλάση 2: $\{4, 6\}$ (μέσος $\mu^{(2)} = 5$). Ολικός μέσος $\mu = 3$. Τότε

$$S_B = 2(1-3)^2 + 2(5-3)^2 = 8+8 = 16, \qquad S_W = \big[(0-1)^2+(2-1)^2\big] + \big[(4-5)^2+(6-5)^2\big] = 2+2 = 4,$$

άρα ο λόγος είναι $S_B/S_W = 16/4 = 4$. Επαλήθευση: η ολική διασπορά είναι $S_T = \sum_i (x_i - 3)^2 = 9 + 1 + 1 + 9 = 20 = S_B + S_W$, που επιβεβαιώνει την ταυτότητα $S_T = S_B + S_W$. Όσο μεγαλύτερος ο λόγος, τόσο πιο απομακρυσμένοι οι μέσοι των κλάσεων σε σχέση με τη διασπορά εντός τους.

- **Περιορισμός τάξης**: ο S_B έχει τάξη το πολύ c (κάθε μπλοκ του H έχει τάξη 1). Άρα η LDA δίνει το πολύ c ουσιαστικές κατευθύνσεις προβολής περιοριστικό όταν θέλουμε χώρο διάστασης $> c$.
- **Μορφές ίχνους**: με απλές πράξεις, $S_B = \bar{X} H \bar{X}^\top$, $S_W = \bar{X}(I - H)\bar{X}^\top$, $S_T = \bar{X}\bar{X}^\top$, όπου H ο μπλοκ-διαγώνιος πίνακας της επίβλεψης. Αυτό φέρνει τη LDA στην ίδια οικογένεια με τις γραφο-μεθόδους: ο $I - H$ παίζει τον ρόλο Λαπλασιανού.
- **Ο H είναι προβολέας**: γράφοντας $H = \sum_k g_k g_k^\top / n_k$ (όπου g_k το χαρακτηριστικό διάνυσμα της κλάσης k) βλέπουμε ότι $H^2 = H$. Συνέπεια: όταν $W = \hat{W} = H$, ισχύει $(I - W)^\top (I - W) = I - H$, οπότε **LDA, εποπτευόμενη LPP και εποπτευόμενη NPP ταυτίζονται μαθηματικά**.

Πρόταση 3.5 (Ταυτότητα διασποράς $S_T = S_B + S_W$)**.** *Με $S_T = \sum_i (x_i - \mu)(x_i - \mu)^\top = \bar{X}\bar{X}^\top$ και S_B, S_W όπως στην* (3.8)*, ισχύει $S_T = S_B + S_W$.*

Απόδειξη. Για x_i της κλάσης k, $x_i - \mu = (x_i - \mu^{(k)}) + (\mu^{(k)} - \mu)$. Αναπτύσσοντας το εξωτερικό γινόμενο και αθροίζοντας ανά κλάση, ο σταυρωτός όρος μηδενίζεται διότι $\sum_{x_i \in X_k} (x_i - \mu^{(k)}) = 0$· απομένουν ακριβώς οι S_W και S_B. □

Πρόταση 3.6 (Λύση της LDA μέσω γενικευμένου ιδιοπροβλήματος)**.** *Για $d = 1$, $\max_a \dfrac{a^\top S_B a}{a^\top S_W a} = \lambda_{\max}(S_B, S_W)$, με βέλτιστο το γενικευμένο ιδιοδιάνυσμα $S_B a = \lambda S_W a$ (ισοδύναμα κορυφαίο ιδιοδιάνυσμα του $S_W^{-1} S_B$ όταν $S_W \succ 0$). Για $d > 1$, στη διατύπωση ίχνους λόγου οι στήλες του βέλτιστου V είναι τα d κορυφαία γενικευμένα ιδιοδιανύσματα.*

Απόδειξη. Είναι η μονοδιάστατη περίπτωση του πηλίκου Rayleigh (§2.8) για το ζεύγος (S_B, S_W) με $S_W \succ 0$· από το Λήμμα 2.13 το ζεύγος διαγωνιοποιείται ταυτόχρονα και ο λόγος μεγιστοποιείται από τη μέγιστη γενικευμένη ιδιοτιμή. Λόγω $S_T = S_B + S_W$, αυτό είναι ισοδύναμο με το ζεύγος (S_W, S_T) της (3.9) (κρατώντας τις μικρότερες ιδιοτιμές). □

Παρατήρηση 3.7 (Τάξη του S_B και το πρόβλημα μικρού δείγματος). Επειδή $\sum_k n_k (\mu^{(k)} - \mu) = 0$, οι c όροι του S_B ικανοποιούν μία γραμμική σχέση, άρα $\operatorname{rank}(S_B) \le c - 1$. Επιπλέον, όταν $m > n$ ο S_W είναι ιδιάζων (*small sample size problem*)· τυπικές αντιμετωπίσεις είναι προκαταρκτικό βήμα PCA ή η κανονικοποίηση $S_W \leftarrow S_W + \gamma I$.

3.4 Συνδέσεις, Ισοδυναμίες και ο Ρόλος του Λαπλασιανού Γράφου

Στο Κεφάλαιο **??** είδαμε ότι όλες οι μέθοδοι ανάγονται στην ίδια μορφή βελτιστοποίησης ίχνους. Εδώ κάνουμε τις συνδέσεις αυτές *αυστηρές*: αποδεικνύουμε τις βασικές ισοδυναμίες που ενοποιούν τις φαινομενικά διαφορετικές τεχνικές, ακολουθώντας το πλαίσιο των Kokiopoulou–Chen–Saad. Το συμπέρασμα είναι ότι οι γραμμικές (προβολικές) μέθοδοι είναι, ουσιαστικά, *προβολές* των μη γραμμικών (implicit) στον υπόχωρο που παράγουν τα δεδομένα.

3.5 Ο Πίνακας LLE ως "Σχεδόν" Λαπλασιανός

Συγκρίνοντας τις αντικειμενικές της LLE, $\mathrm{Tr}[Y(I-W)^\top(I-W)Y^\top]$, και των Eigenmaps, $\mathrm{Tr}[Y(D-W)Y^\top]$, η ομοιότητα είναι εμφανής: αλλάζει μόνο ο πίνακας μέσα στο ίχνος. Γεννιέται το ερώτημα: είναι ο πίνακας LLE $M=(I-W)^\top(I-W)$ ένας Λαπλασιανός; Υπενθυμίζουμε ότι ένας Λαπλασιανός L είναι συμμετρικός, με *μη θετικά* εκτός διαγωνίου στοιχεία και *μηδενικά αθροίσματα γραμμών*.

Πρόταση 3.8 (Δομή του πίνακα LLE)**.** *Έστω W ο πίνακας βαρών της LLE, με $w_{ii}=0$ και αθροίσματα γραμμών ίσα με 1 (δηλαδή $W\mathbf{1}=\mathbf{1}$). Ο συμμετρικός πίνακας $M=(I-W)^\top(I-W)$ έχει μηδενικά αθροίσματα γραμμών και στηλών. Επιπλέον, αν $w_{:j}$ συμβολίζει την j-στή στήλη του W,*

$$m_{jj}=1+\|w_{:j}\|^2\,,\qquad m_{ij}=-(w_{ij}+w_{ji})+\langle w_{:i},\,w_{:j}\rangle\,,\quad i\neq j. \tag{3.10}$$

Απόδειξη. : Αφού $W\mathbf{1}=\mathbf{1}$, ισχύει $(I-W)\mathbf{1}=0$, άρα $M\mathbf{1}=(I-W)^\top(I-W)\mathbf{1}=(I-W)^\top 0=0$: τα αθροίσματα γραμμών του M είναι μηδέν. Επειδή ο M είναι συμμετρικός, το ίδιο ισχύει και για τα αθροίσματα στηλών. Για τα στοιχεία, αναπτύσσουμε $M=I-W-W^\top+W^\top W$, οπότε

$$m_{ij}=e_i^\top M e_j=\delta_{ij}-w_{ij}-w_{ji}+(W^\top W)_{ij},\qquad (W^\top W)_{ij}=\sum_k w_{ki}w_{kj}=\langle w_{:i},\,w_{:j}\rangle\,.$$

Θέτοντας $i=j$ (και $w_{ii}=0$) προκύπτει $m_{jj}=1+\|w_{:j}\|^2\cdot$ για $i\neq j$ προκύπτει η δεύτερη σχέση. □

Παρατήρηση 3.9 (Γιατί δεν είναι γνήσιος Λαπλασιανός). Ο M ικανοποιεί τη συνθήκη των μηδενικών αθροισμάτων, αλλά *όχι* εκείνη των μη θετικών εκτός διαγωνίου: από την (3.10), όταν $w_{ij}=w_{ji}=0$ έχουμε $m_{ij}=\langle w_{:i},\,w_{:j}\rangle\geq 0$. Συνεπώς, σε ζεύγη μη γειτονικών κόμβων το στοιχείο μπορεί να είναι θετικό. Παρ' όλα αυτά, ο M μοιράζεται με τον Λαπλασιανό δύο κρίσιμες ιδιότητες — είναι συμμετρικός θετικά ημιορισμένος και σχετίζει γειτονικά δείγματα — γι' αυτό και οι δύο μέθοδοι (LLE, Eigenmaps) συμπεριφέρονται παρόμοια.

3.6 Ο Πίνακας Επίβλεψης H είναι Προβολεας

Στην εποπτευόμενη ρύθμιση (Ενότητα 1.5) ορίσαμε τον μπλοκ-διαγώνιο πίνακα $H=\mathrm{diag}[H_1,\ldots,H_c]$ με $H_k=\frac{1}{n_k}\mathbf{1}_{n_k}\mathbf{1}_{n_k}^\top$. Ισοδύναμα, αν $g_k\in\mathbb{R}^n$ είναι το *χαρακτηριστικό διάνυσμα* της κλάσης k (με i-στό στοιχείο 1 αν x_i ανήκει στην κλάση k, αλλιώς 0), τότε

$$H=\sum_{k=1}^{c}\frac{g_k g_k^\top}{n_k}. \tag{3.11}$$

Πρόταση 3.10. *Ο H είναι ορθογώνιος προβολέας (δηλαδή $H^\top=H$ και $H^2=H$) τάξης c, με $H\mathbf{1}=\mathbf{1}$.*

Απόδειξη. Η συμμετρία είναι προφανής. Τα χαρακτηριστικά διανύσματα διαφορετικών (ξένων) κλάσεων ικανοποιούν $g_k^\top g_l=n_k\,\delta_{kl}$. Άρα από την (3.11):

$$H^2=\sum_{k,l}\frac{g_k(g_k^\top g_l)g_l^\top}{n_k n_l}=\sum_k\frac{g_k\,n_k\,g_k^\top}{n_k^2}=\sum_k\frac{g_k g_k^\top}{n_k}=H.$$

Κάθε όρος $g_k g_k^\top/n_k$ έχει τάξη 1 και οι όροι έχουν ξένα πεδία τιμών, άρα $\mathrm{rank}(H)=c$. Τέλος, αφού $\mathbf{1}=\sum_k g_k$, έχουμε $H\mathbf{1}=\sum_k g_k(g_k^\top\mathbf{1})/n_k=\sum_k g_k n_k/n_k=\sum_k g_k=\mathbf{1}$. □

Πρόταση 3.11 (LDA ≡ εποπτευόμενη LPP ≡ εποπτευόμενη NPP)**.** *Όταν στις γραφομεθόδους χρησιμοποιηθεί ο εποπτευόμενος γράφος, $W = \hat{W} = H$, οι μέθοδοι LDA, (κανονικοποιημένη) εποπτευόμενη LPP και εποπτευόμενη NPP καταλήγουν στο ίδιο ιδιοπρόβλημα.*

Απόδειξη. Πρώτα, επειδή κάθε γραμμή του H αθροίζει στη μονάδα ($H\mathbf{1} = \mathbf{1}$ από την Πρόταση 3.10, και κατ' επέκταση το άθροισμα κάθε γραμμής είναι 1), ο διαγώνιος πίνακας βαθμών είναι $D = \mathrm{diag}(\text{αθροίσματα γραμμών}) = I$. Δεύτερον, αφού ο H είναι προβολέας, και ο $I - H$ είναι προβολέας, οπότε

$$(I - W)^\top (I - W) = (I - H)^\top (I - H) = (I - H)^2 = I - H, \qquad I - \hat{W} = I - H.$$

Συνεπώς:

- Εποπτευόμενη NPP: $X(I - W)^\top (I - W)X^\top u = \lambda (XX^\top)u \;\Rightarrow\; X(I - H)X^\top u = \lambda XX^\top u$.
- Κανονικοποιημένη εποπτευόμενη LPP (με $\hat{X} = XD^{1/2} = X$): $\hat{X}(I - \hat{W})\hat{X}^\top v = \lambda \hat{X}\hat{X}^\top v \;\Rightarrow\; X(I - H)X^\top v = \lambda XX^\top v$ — ταυτόσημο με το παραπάνω.

Για τη LDA, το ιδιοπρόβλημα είναι $\bar{X}(I - H)\bar{X}^\top u = \lambda(\bar{X}\bar{X}^\top)u$ με κεντραρισμένα δεδομένα. Όμως, αφού $(I - H)\mathbf{1} = 0$ (διότι $H\mathbf{1} = \mathbf{1}$), ο $I - H$ *εξουδετερώνει τη σταθερή κατεύθυνση*: με $J = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ τον τελεστή κεντραρίσματος ισχύει $J(I - H)J = I - H$, άρα $\bar{X}(I - H)\bar{X}^\top = XJ(I - H)JX^\top = X(I - H)X^\top$. Οι δύο μέθοδοι (LDA και εποπτευόμενη LPP/NPP) δίνουν επομένως τα ίδια μη τετριμμένα ιδιοδιανύσματα· διαφέρουν μόνο στον πίνακα κανονικοποίησης ($\bar{X}\bar{X}^\top$ έναντι $XX^\top$), διαφορά που αφορά αποκλειστικά την (απορριπτόμενη) σταθερή συνιστώσα. □

3.7 Οι Γραμμικές Μέθοδοι ως Προβολές (Rayleigh–Ritz)

Η πιο βαθιά σύνδεση της εργασίας είναι ότι οι προβολικές μέθοδοι (LPP, NPP) είναι *μέθοδοι προβολής* (projection methods) για τα ιδιοπροβλήματα των αντίστοιχων implicit μεθόδων (Eigenmaps, LLE). Υπενθυμίζουμε το σχετικό εργαλείο.

Ορισμός 3.12 (Ορθογώνια μέθοδος προβολής / Rayleigh–Ritz)**.** Για το ιδιοπρόβλημα $Au = \lambda u$ και έναν υπόχωρο $\mathcal{K} = \mathrm{Range}(V)$ διάστασης k, η *ορθογώνια μέθοδος προβολής* αναζητά προσεγγιστικό ζεύγος $(\tilde{u}, \tilde{\lambda})$ με

$$\tilde{u} \in \mathcal{K}, \qquad A\tilde{u} - \tilde{\lambda}\tilde{u} \;\perp\; \mathcal{K}.$$

Γράφοντας $\tilde{u} = Vy$, η συνθήκη ορθογωνιότητας (Galerkin) δίνει το συμπυκνωμένο πρόβλημα

$$V^\top A V\, y = \tilde{\lambda}\, V^\top V\, y.$$

Πρόταση 3.13 (LPP/NPP ως προβεβλημένες Eigenmaps/LLE)**.** *(i) Η LPP είναι μαθηματικά ισοδύναμη με μια μέθοδο προβολής στον υπόχωρο* $\mathrm{span}\{\hat{X}^\top\}$ *εφαρμοσμένη στον κανονικοποιημένο Λαπλασιανό $\hat{L} = I - \hat{W}$· δηλαδή είναι μια προβεβλημένη εκδοχή των Eigenmaps.*

(ii) Η NPP είναι μαθηματικά ισοδύναμη με μια μέθοδο προβολής στον $\mathrm{span}\{X^\top\}$ *εφαρμοσμένη στον πίνακα $M = (I - W)^\top (I - W)$· δηλαδή είναι μια προβεβλημένη εκδοχή του LLE.*

Απόδειξη. (ii) Το LLE λύνει $Mu = \lambda u$. Περιορίζουμε την αναζήτηση της λύσης στον υπόχωρο $\mathrm{span}\{X^\top\}$, θέτοντας $u = X^\top v$. Με βάση $V \equiv X^\top$, η συνθήκη Galerkin (Ορισμός 3.12) δίνει

$$(X)\, M\, (X^\top v) = \lambda\, (X)(X^\top v) \quad \Longleftrightarrow \quad XMX^\top v = \lambda\, XX^\top v,$$

που είναι ακριβώς το ιδιοπρόβλημα της NPP. (i) Όμοια, τα (κανονικοποιημένα) Eigenmaps λύνουν $(I - \hat{W})\hat{u} = \lambda\hat{u}$· περιορίζοντας $\hat{u} = \hat{X}^\top v$ και εφαρμόζοντας Galerkin προκύπτει $\hat{X}(I - \hat{W})\hat{X}^\top v = \lambda\, \hat{X}\hat{X}^\top v$, δηλαδή η (κανονικοποιημένη) LPP. □

Πόρισμα 3.14 (Ακριβής ισοδυναμία στην υποδειγματοληπτημένη περίπτωση)**.** *Αν η τάξη στηλών του X ισούται με n (υποδειγματοληπτημένη περίπτωση, $m \geq n$), τότε* $\mathrm{span}\{X^\top\} = \mathbb{R}^n$ *και η προβολή είναι ταυτοτική. Επομένως η LPP δίνει ακριβώς το ίδιο αποτέλεσμα με τα Eigenmaps και η NPP με το LLE.*

Απόδειξη. Όταν $\mathrm{rank}(X) = n$, ο $X^\top$ έχει τάξη n, άρα ο $\mathrm{span}\{X^\top\}$ είναι ολόκληρος ο $\mathbb{R}^n$ και είναι τετριμμένα αναλλοίωτος υπό κάθε πίνακα. Μια μέθοδος προβολής σε ολόκληρο τον χώρο επιστρέφει τα ακριβή ιδιοζεύγη, οπότε η προβεβλημένη και η αρχική μέθοδος ταυτίζονται. □

Παρατήρηση 3.15 (Ποιότητα της προβολής μέσω Poincaré). [12] Έξω από την υποδειγματοληπτημένη περίπτωση, η προβολή στον $\mathrm{span}\{X^\top\}$ είναι μόνο *προσεγγιστική*. Το Λήμμα 2.10 (διαχωρισμός Poincaré) δίνει ποσοτικό περιεχόμενο: οι τιμές Ritz της προβεβλημένης (γραμμικής) μεθόδου *παρεμβάλλονται* ανάμεσα στις ιδιοτιμές της αρχικής (implicit) μεθόδου. Έτσι η LPP/NPP προσεγγίζει τα Eigenmaps/LLE "από μέσα", με ακρίβεια ακριβώς όταν ο $\mathrm{span}\{X^\top\}$ είναι αναλλοίωτος (π.χ. $\mathrm{rank}(X) = n$) — εξήγηση, σε επίπεδο φάσματος, του γιατί οι γραμμικές μέθοδοι "χάνουν" πληροφορία όταν $m < n$.

3.8 Σύνδεση PCA – LPP

Παρότι η PCA ζητά τις *μεγαλύτερες* ιδιοτιμές και η LPP τις *μικρότερες*, μπορούμε να ανακτήσουμε την πρώτη ως ειδική περίπτωση της δεύτερης, με κατάλληλη επιλογή του γράφου.

Πρόταση 3.16. *Έστω $X \in \mathbb{R}^{m\times n}$ με $m < n$, πλήρους τάξης γραμμών και κεντραρισμένο ($X\mathbf{1} = 0$). Με $D = I$ και*

$$I - \hat{W} = X^\top (XX^\top)^{-2} X \quad (= SS^\top, \;\; S = X^\dagger),$$

η LPP γίνεται μαθηματικά ισοδύναμη με την PCA.

Απόδειξη. Αντικαθιστούμε στο (κανονικοποιημένο) ιδιοπρόβλημα της LPP, $\hat{X}(I-\hat{W})\hat{X}^\top v = \lambda\hat{X}\hat{X}^\top v$ με $\hat{X} = X$ (αφού $D = I$):

$$X\big[X^\top (XX^\top)^{-2} X\big] X^\top v = (XX^\top)(XX^\top)^{-2}(XX^\top)\, v = v.$$

Άρα το ιδιοπρόβλημα της LPP ανάγεται σε $v = \lambda\, (XX^\top)v$, δηλαδή $XX^\top v = \frac{1}{\lambda} v$. Οι *μικρότερες* λ της LPP αντιστοιχούν στις *μεγαλύτερες* ιδιοτιμές $\frac{1}{\lambda}$ του πίνακα συνδιακύμανσης $XX^\top$ — δηλαδή στην PCA. (Η ταυτότητα $X^\dagger = X^\top (XX^\top)^{-1}$ για πλήρη τάξη γραμμών δίνει $X^\dagger (X^\dagger)^\top = X^\top (XX^\top)^{-2} X$, εξηγώντας τη μορφή $SS^\top$.) □

3.9 Σύνδεση με Φασματική Συσταδοποίηση

Οι ίδιοι Λαπλασιανοί που εμφανίζονται στα Eigenmaps διέπουν και τη *φασματική συσταδοποίηση* (spectral clustering). Δοσμένου σταθμισμένου γράφου $G = (V, E)$, μια διαμέριση σε k ξένα υποσύνολα $V_1, \dots, V_k$ ζητά να ελαχιστοποιηθεί το συνολικό βάρος των ακμών που τέμνουν διαφορετικά μέρη, με ισορροπημένα μεγέθη. Μια τυπική συνάρτηση κόστους είναι

$$F(V_1, \dots, V_k) = \sum_{\ell=1}^{k} \frac{\sum_{i \in V_\ell,\, j \in V_\ell^c} w_{ij}}{\sum_{i \in V_\ell} d_i}, \qquad d_i = \sum_j w_{ij}. \tag{3.12}$$

Πρόταση 3.17 (Χαλάρωση φασματικής συσταδοποίησης)**.** *Ορίζοντας τον $n \times k$ πίνακα Z με στήλη ℓ έναν δείκτη συστάδας,*

$$Z(j, \ell) = \begin{cases} \big(\sum_{i \in V_\ell} d_i\big)^{-1/2}, & j \in V_\ell, \\ 0, & \text{αλλιώς}, \end{cases}$$

η συνάρτηση (3.12) *γράφεται ως $F = \mathrm{Tr}(Z^\top L Z)$ με τον περιορισμό $Z^\top D Z = I$, όπου $L = D - W$. Χαλαρώνοντας τον (διακριτό) περιορισμό μορφής του Z ώστε να κρατάμε μόνο τις δύο αλγεβρικές συνθήκες, η λύση δίνεται από τα k μικρότερα ιδιοδιανύσματα του γενικευμένου προβλήματος*

$$L z_i = \lambda_i D z_i \qquad (\textit{normalized cut}). \tag{3.13}$$

*Αντικαθιστώντας το "μέγεθος" $\sum_{i \in V_\ell} d_i$ με το πλήθος $|V_\ell|$, παίρνουμε τον περιορισμό $Z^\top Z = I$ και το πρόβλημα $L z_i = \lambda_i z_i$ (*ratio cut*).*

Παρατήρηση 3.18 (Η γέφυρα προς τα Eigenmaps). Το γενικευμένο πρόβλημα (3.13) είναι *ακριβώς* το ιδιοπρόβλημα των Laplacian Eigenmaps (3.4). Άρα ο πίνακας Z είναι η χαμηλοδιάστατη ενσωμάτωση Eigenmaps της πολλαπλότητας των δεδομένων: μια καλή συσταδοποίηση των γραμμών του Z αντιστοιχεί σε καλή συσταδοποίηση των αρχικών δεδομένων. Το δεύτερο μικρότερο ιδιοδιάνυσμα (*Fiedler vector*) είναι το πιο πληροφοριακό για διμερή διαχωρισμό. Όπως θα δούμε στο Κεφάλαιο 4, ο ratio cut προκύπτει επίσης ως Kernel PCA με πυρήνα $K = L^\dagger$.

Κεφάλαιο 4

Πυρήνες και η Ενοποίηση Γραμμικών–Μη Γραμμικών Μεθόδων

Στο κεφάλαιο αυτό αποδεικνύουμε ότι οι μέθοδοι πυρήνα δεν είναι παρά οι γραμμικές προβολικές μέθοδοι εφαρμοσμένες σε έναν χώρο χαρακτηριστικών (feature space), και ότι, με αυτόν τον τρόπο, οι "μη γραμμικές" μέθοδοι (LLE, Eigenmaps, φασματική συσταδοποίηση) ανακτώνται ως ειδικές περιπτώσεις. Αυτό κλείνει το επιχείρημα: η μη γραμμική διόρθωση στην αποτυχία της SVD είναι, σε επίπεδο εξισώσεων, η αντικατάσταση του εσωτερικού γινομένου με έναν πυρήνα.

4.1 Εισαγωγή στους Πυρήνες (The Kernel Trick)

Υπάρχουν οι μέθοδοι *εκμάθησης πολλαπλοτήτων* της Ενότητας 1.4 (LLE, Laplacian Eigenmaps, ISOMAP), που χτίζουν έναν γράφο τοπικής γειτνίασης και έτσι αποτυπώνουν την εγγενή γεωμετρία πριν προβάλουν, που είδαμε εως τώρα. Και οι *μέθοδοι πυρήνα* που συζητάμε σε αυτό το κεφάλαιο. Μία από τις πιο σημαντικές συνεισφορές είναι η χρήση των μεθόδων πυρήνα ως **γέφυρα** μεταξύ γραμμικών και μη γραμμικών μεθόδων. Αντί να προβάλλουμε τα δεδομένα απευθείας, τα μεταφέρουμε νοητά μέσω μιας μη γραμμικής συνάρτησης $\Phi(x)$ σε έναν χώρο πολύ υψηλότερων διαστάσεων (feature space) $\mathcal{H}$, και εφαρμόζουμε εκεί τις γραμμικές μεθόδους. Το 'τέχνασμα του πυρήνα' (kernel trick) είναι ότι δεν χρειάζεται ποτέ να υπολογίσουμε ρητά το $\Phi(x)$: αρκεί να γνωρίζουμε τα εσωτερικά γινόμενα μέσω μιας συνάρτησης πυρήνα

$$k(x_i, x_j) = \langle \Phi(x_i),\, \Phi(x_j) \rangle\,, \qquad K_{ij} = k(x_i, x_j), \tag{4.1}$$

όπου ο $K \in \mathbb{R}^{n\times n}$ είναι ο πίνακας Gram (kernel matrix). Το ότι αυτό είναι θεμιτό εξασφαλίζεται από το θεώρημα Moore–Aronszajn: *κάθε συμμετρικός θετικά (ημι)ορισμένος πυρήνας αντιστοιχεί σε ένα εσωτερικό γινόμενο πάνω σε κάποιον χώρο Hilbert* (RKHS). Έτσι ο πυρήνας ορίζει, ουσιαστικά, ένα εναλλακτικό — εξαρτημένο από τα δεδομένα — "μέτρο ομοιότητας" στον αρχικό χώρο.

Kernel PCA. Εφαρμόζοντας κλασική PCA στο σύνολο $\{\Phi(x_i)\}$, καταλήγουμε στο ιδιοπρόβλημα του (κεντραρισμένου) πίνακα Gram:

$$\bar{K} z_i = \lambda_i z_i, \qquad [z_1, \ldots, z_d] = Y^\top, \qquad \bar{K} = \big(I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^\top\big) K \big(I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^\top\big).$$

Παρατηρούμε ότι ο πίνακας δεδομένων X *εξαφανίζεται*: ο,τιδήποτε χρειαζόμαστε κρύβεται στον K. Αυτό είναι κοινό χαρακτηριστικό όλων των kernel μεθόδων — τα ιδιοπροβλήματά τους είναι $n \times n$ (όπως ακριβώς στις LLE/Eigenmaps), σε αντίθεση με τα $m \times m$ ιδιοπροβλήματα των αντίστοιχων γραμμικών μεθόδων.

Οι θεμελιώδεις ισοδυναμίες. Το εκπληκτικό αποτέλεσμα που αποδεικνύουν οι συγγραφείς είναι ότι οι "μη γραμμικές" μέθοδοι δεν είναι παρά οι γραμμικές (προβολικές) μέθοδοι εφαρμοσμένες σε έναν κατάλληλο χώρο πυρήνων. Συγκεκριμένα:

Πρόταση 4.1 (Ισοδυναμίες πυρήνων, Kokiopoulou–Chen–Saad)**.** *Ισχύουν οι εξής (μαθηματικές) ισοδυναμίες:*

(i) *Η* ***Kernel PCA*** *με πυρήνα $K = L^\dagger$ είναι μαθηματικά ισοδύναμη με το* ***spectral ratio cut*** *(στον χώρο χαρακτηριστικών).*

(ii) *Το* ***Kernel LPP*** *είναι (πρακτικά) ισοδύναμο με τα* ***Laplacian Eigenmaps****.*

(iii) *Το* ***Kernel NPP*** *είναι ισοδύναμο με το* ***LLE****.*

Ουσιαστικά, λοιπόν, οι φαινομενικά "μη γραμμικές" μέθοδοι είναι απλώς οι γραμμικές προβολικές μέθοδοι εφαρμοσμένες σε έναν κατάλληλο χώρο πυρήνων. Η "μη γραμμική διόρθωση" στην αποτυχία της SVD παίρνει έτσι μια πολύ καθαρή μορφή: *αντικαθιστούμε το (σταθερό, Ευκλείδειο) εσωτερικό γινόμενο με έναν πυρήνα που κωδικοποιεί τη σωστή, εξαρτημένη από τα δεδομένα γεωμετρία* — ή, ισοδύναμα, εργαζόμαστε με τον Λαπλασιανό ενός γράφου γειτνίασης. Και στις δύο περιπτώσεις προσγειωνόμαστε ξανά σε ένα πρόβλημα ιδιοτιμών, διατηρώντας την υπολογιστική απλότητα του γραμμικού κόσμου.

Παρατήρηση 4.2 (Πυκνοί πυρήνες έναντι αραιών Λαπλασιανών)*.* Είναι χρήσιμο να σημειωθεί ότι οι πυρήνες ορίζονται τυπικά ως ολοκληρωτικοί τελεστές και οδηγούν σε *πυκνούς* πίνακες $K\cdot$ οι Λαπλασιανοί γράφων, αντιθέτως, είναι *αραιοί* και αντιστοιχούν σε αντιστρόφους ολοκληρωτικών τελεστών. Αυτή η δυϊκότητα (συμπαγείς τελεστές $\leftrightarrow$ πυκνοί πίνακες, διαφορικοί τελεστές $\leftrightarrow$ αραιοί πίνακες) εξηγεί γιατί, παρά τις ισοδυναμίες, οι δύο προσεγγίσεις έχουν πολύ διαφορετική υπολογιστική συμπεριφορά σε μεγάλη κλίμακα.

4.1.1 Πυρήνες και Kernel PCA

Η ιδέα του πυρήνα είναι να μεταφέρουμε (νοητά) τα δεδομένα μέσω μιας μη γραμμικής Φ σε έναν χώρο υψηλών διαστάσεων και να εφαρμόσουμε εκεί γραμμική PCA. Το "τέχνασμα" είναι ότι δεν χρειαζόμαστε ποτέ το $\Phi(x)$ ρητά: αρκεί ο $n \times n$ πίνακας Gram K, μέσω της συνάρτησης πυρήνα $k(x, y)$.

$$K_{ij} = \langle \Phi(x_i),\, \Phi(x_j) \rangle\,, \qquad \bar{K} = \big(I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^\top\big) K \big(I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^\top\big), \qquad \bar{K} z_i = \lambda_i z_i. \tag{4.2}$$

Γιατί όμως εξαφανίζεται ο X). Η PCA στον χώρο χαρακτηριστικών θα απαιτούσε τα κορυφαία ιδιοδιανύσματα του $\bar{\Phi}\bar{\Phi}^\top$ ένας πίνακας $N \times N$ με N τεράστιο/άγνωστο. Πολλαπλασιάζοντας από αριστερά με $\bar{\Phi}^\top$:

$$\bar{\Phi}\bar{\Phi}^\top u = \lambda u \;\xrightarrow{\bar{\Phi}^\top \cdot}\; \underbrace{\bar{\Phi}^\top\bar{\Phi}}_{\bar{K}} (\bar{\Phi}^\top u) = \lambda\,(\bar{\Phi}^\top u) \quad \Rightarrow \quad \bar{K} z = \lambda z, \;\; z = \bar{\Phi}^\top u.$$

Τα z είναι (ανάστροφα) οι γραμμές της ενσωμάτωσης Y, άρα $Y^\top = [z_1, \dots, z_d]$ — και ο πίνακας δεδομένων δεν εμφανίζεται πουθενά. Ως κανονικό παράδειγμα της δύναμης των πυρήνων: το πρόβλημα "τετράγωνο + δακτύλιος" (§3.1 του Κεφ. 1), που *καμία* ευθεία δεν διαχωρίζει, γίνεται γραμμικά διαχωρίσιμο με Gaussian πυρήνα $k(x, y) = \exp(-\|x - y\|^2 / \sigma^2)$ για κατάλληλο σ.

- **Θεμελίωση (Moore–Aronszajn)**: κάθε συμμετρικός θετικά (ημι)ορισμένος πυρήνας αντιστοιχεί σε εσωτερικό γινόμενο πάνω σε κάποιον χώρο Hilbert (RKHS). Έτσι η επιλογή πυρήνα ισοδυναμεί με την επιλογή ενός εναλλακτικού, εξαρτημένου από τα δεδομένα "μέτρου ομοιότητας".

- **Η διάσταση των ιδιοπροβλημάτων**: όλες οι kernel μέθοδοι λύνουν $n \times n$ προβλήματα (όπως LLE/Eigenmaps), σε αντίθεση με τα $m \times m$ των γραμμικών εκδοχών. Αυτό είναι το γεωμετρικό "γεφύρωμα" γραμμικού–μη γραμμικού κόσμου.

- **Οι ισοδυναμίες**: αποδεικνύεται ότι η Kernel PCA με $K = L^\dagger$ ισοδυναμεί με το spectral ratio cut, το Kernel LPP με τα Laplacian Eigenmaps, και το Kernel NPP με το LLE. Δηλαδή οι "μη γραμμικές" μέθοδοι *είναι* οι γραμμικές προβολικές μέθοδοι εφαρμοσμένες σε κατάλληλο χώρο πυρήνων. Αυτή είναι, σε επίπεδο εξισώσεων, η τελική απάντηση στο ερώτημα "πώς διορθώνεται η αποτυχία της SVD".

Θεώρημα 4.3 (Mercer, μορφή πινάκων)**.** *Συμμετρικός $K \in \mathbb{R}^{n \times n}$ είναι έγκυρος πίνακας πυρήνα (δηλαδή $K_{ij} = \langle \Phi(x_i), \Phi(x_j) \rangle$ για κάποια Φ) αν και μόνο αν $K \succeq 0$.*

Απόδειξη. ($\Rightarrow$) $c^\top K c = \|\sum_i c_i \Phi(x_i)\|^2 \geq 0$. ($\Leftarrow$) Αν $K \succeq 0$, $K = Q \Lambda Q^\top$ με $\Lambda \succeq 0$· θέτοντας $\Phi(x_i) = \Lambda^{1/2} q_i$ ($q_i^\top$ η i-στή γραμμή του Q), $\langle \Phi(x_i), \Phi(x_j) \rangle = (Q \Lambda Q^\top)_{ij} = K_{ij}$. □

RKHS και αναπαραγωγική ιδιότητα. Ο πυρήνας k ορίζει χώρο Hilbert με αναπαραγωγικό πυρήνα $\mathcal{H}_k$, με $\Phi(x) = k(x, \cdot)$, $\langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}_k} = k(x, y)$ και την *αναπαραγωγική ιδιότητα* $\langle f, k(x, \cdot) \rangle_{\mathcal{H}_k} = f(x)$. Αυτό είναι το ακριβές νόημα του "τεχνάσματος του πυρήνα": εργαζόμαστε στον (ίσως απειροδιάστατο) $\mathcal{H}_k$ μόνο μέσω αποτιμήσεων του k.

Θεώρημα 4.4 (Θεώρημα αναπαράστασης)**.** *[11] Για $\min_{f \in \mathcal{H}_k} \Omega(\|f\|_{\mathcal{H}_k}) + \mathcal{R}(f(x_1), \dots, f(x_n))$ με Ω γνησίως αύξουσα, υπάρχει ελαχιστοποιητής της μορφής $f(\cdot) = \sum_{i=1}^n \alpha_i \, k(x_i, \cdot)$.*

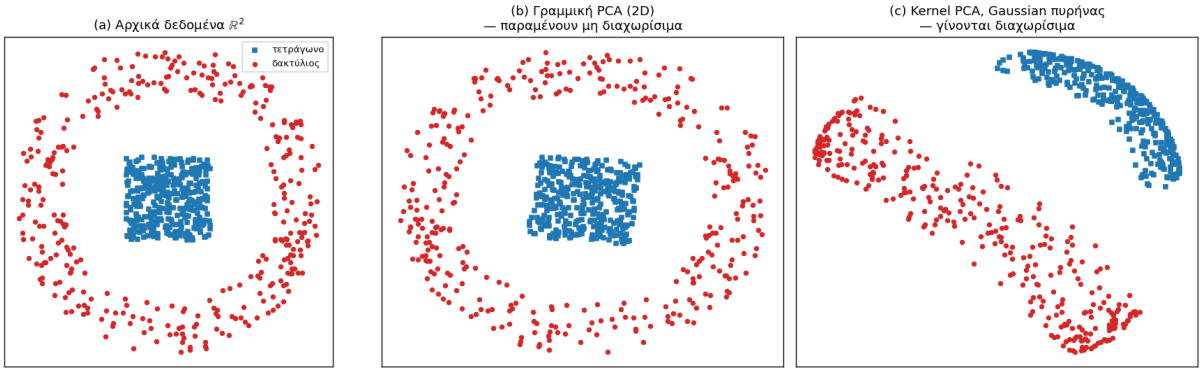
Σκαρίφημα. Αναλύουμε $f = f_\parallel + f_\perp$ με $f_\parallel \in \operatorname{span}\{k(x_i, \cdot)\}$. Από την αναπαραγωγική ιδιότητα $f(x_j) = f_\parallel(x_j)$, άρα το $\mathcal{R}$ εξαρτάται μόνο από το $f_\parallel$· και $\|f\|^2 = \|f_\parallel\|^2 + \|f_\perp\|^2$, οπότε το Ω ελαχιστοποιείται για $f_\perp = 0$. □

Παραγωγή του κεντραρισμένου πυρήνα. Με $\bar{\Phi}(x_i) = \Phi(x_i) - \frac{1}{n} \sum_\ell \Phi(x_\ell)$,

$$\big\langle \bar{\Phi}(x_i), \bar{\Phi}(x_j) \big\rangle = K_{ij} - \tfrac{1}{n} \sum_\ell K_{i\ell} - \tfrac{1}{n} \sum_\ell K_{\ell j} + \tfrac{1}{n^2} \sum_{\ell, m} K_{\ell m} = (JKJ)_{ij},$$

δηλαδή $\bar{K} = JKJ$ — το κεντράρισμα γίνεται χωρίς να υπολογιστεί ποτέ το Φ.

Συνήθεις πυρήνες. (α) γραμμικός $k(x, y) = x^\top y$ (επαναφέρει την PCA)· (β) πολυωνυμικός $(x^\top y + c)^p$, με χώρο χαρακτηριστικών πεπερασμένης διάστασης (μονώνυμα βαθμού $\leq p$)· (γ) Gaussian $\exp(-\|x - y\|^2/\sigma^2)$, με απειροδιάστατο χώρο χαρακτηριστικών — ο πυρήνας που καθιστά γραμμικά διαχωρίσιμο το "τετράγωνο + δακτύλιος". Η σ επιλέγεται συχνά ως το μισό της διαμέσου των ζευγαρωτών αποστάσεων. Εδώ βλέπουμε πως μια μέθοδος kernel pca καταφέρνει να διαχωρίσει τα δεδομένα μας ξεκάθαρα σε ενα παράδειγμα που η απλή PCA δεν τα καταφέρνει.



Κεφάλαιο 5

Επεξηγηματικά Παραδείγματα

Στο κεφάλαιο αυτό παρουσιάζονται τα αριθμητικά αποτελέσματα των μεθόδων μείωσης διαστάσεων που αναλύθηκαν θεωρητικά στα προηγούμενα κεφάλαια. Η υλοποίηση βασίστηκε σε ένα ολοκληρωμένο υπολογιστικό περιβάλλον σε γλώσσα R, χρησιμοποιώντας μια κεντρική συνάρτηση-πυρήνα `embed()` για τη διαχείριση των αλγορίθμων.

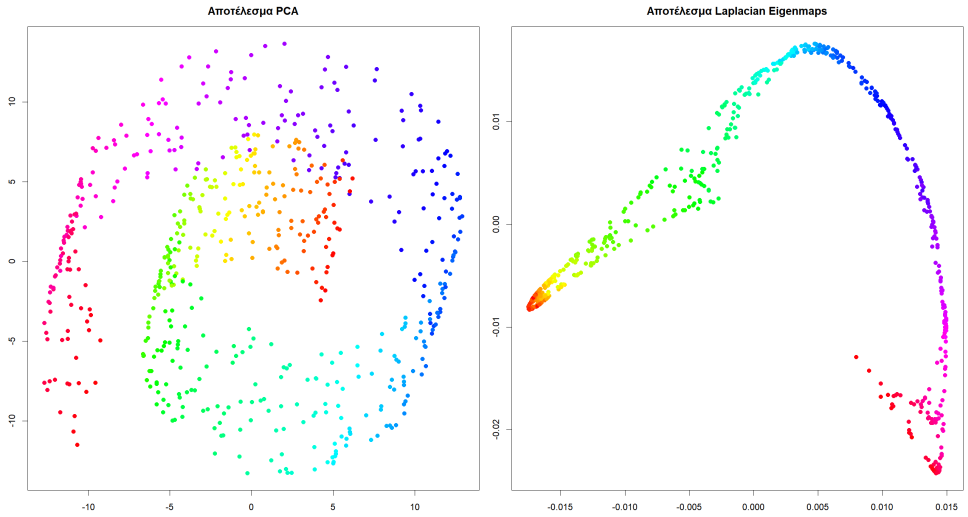
5.1 Σύνολα Δεδομένων και Πρωτόκολλο

Για την αξιολόγηση των αλγορίθμων χρησιμοποιήθηκαν συνθετικά σύνολα δεδομένων (όπως το Swiss Roll, κύκλοι και Gaussian blobs) που επιτρέπουν την επαλήθευση της ικανότητας των μεθόδων να ανακαλύπτουν κρυμμένες δομές (manifold learning). Η διαδικασία περιλαμβάνει:

- **Γεννήτριες Δεδομένων:** Υλοποιήθηκαν συναρτήσεις όπως οι `make_swiss_roll`, `make_circle` και `make_blobs` για τη δημιουργία τοπολογιών με γνωστή γεωμετρία.
- **Προεπεξεργασία:** Για τη διασφάλιση της αριθμητικής ευστάθειας, ενσωματώθηκε προαιρετικό βήμα PCA για μείωση της αρχικής διάστασης πριν από την εφαρμογή των μη γραμμικών μεθόδων.
- **Αριθμητική Επίλυση:** Χρήση παραγοντοποίησης Cholesky για την επίλυση των γενικευμένων ιδιοπροβλημάτων ($Av = \lambda Bv$) στις συναρτήσεις `seig` και `geig`.

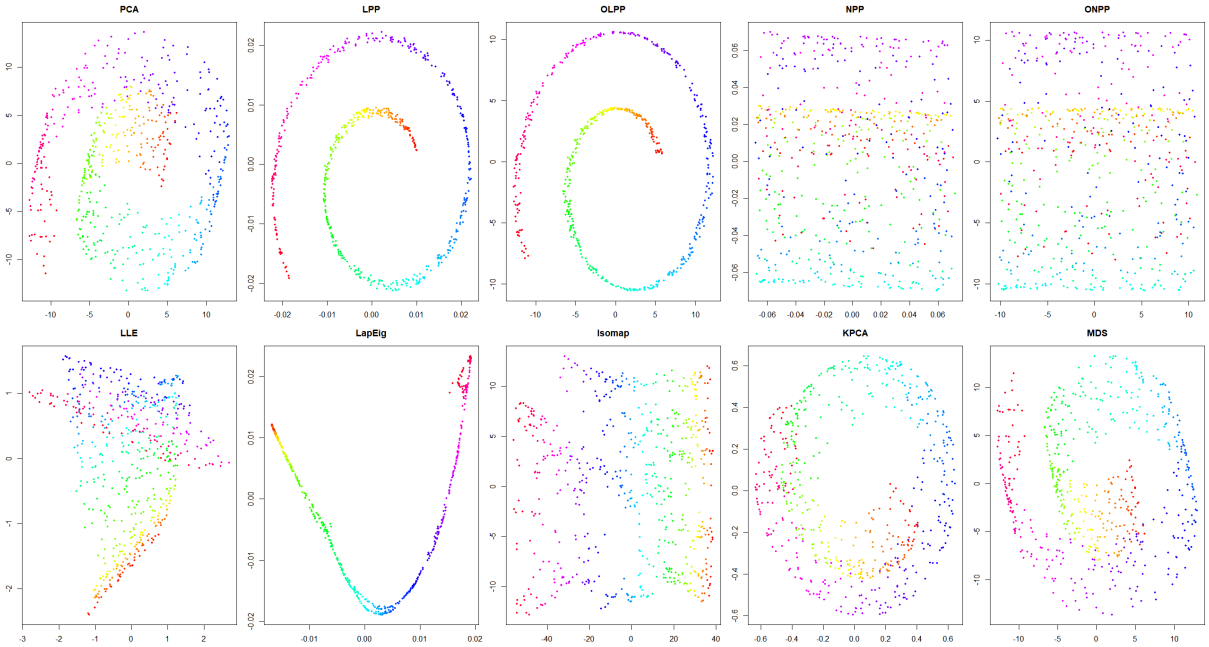
5.2 Οπτικοποίηση σε 2D

Η οπτική αντιπαραβολή των μεθόδων πραγματοποιήθηκε προβάλλοντας τα δεδομένα σε διάσταση $d = 2$. Η σύγκριση χωρίζεται σε δύο επίπεδα:



Σχήμα 5.1: Graph 1: Απευθείας σύγκριση PCA (γραμμική) και Laplacian Eigenmaps (μη γραμμική). Είναι εμφανές οτι η PCA συμπιέζει τη δομή, ενώ τα Laplacian Eigenmaps ανακτούν την εσωτερική της καμπυλότητα.

Στο Σχήμα 5.2 (Graph 2) παρουσιάζεται η συνολική συμπεριφορά των 10 αλγορίθμων, η οποία επιβεβαιώνει τη θεωρητική υπεροχή των μεθόδων που βασίζονται σε γεωδαισιακές αποστάσεις:



Σχήμα 5.2: Graph 2: Συνολικό πλέγμα αποτελεσμάτων για τις 10 μη επιβλεπόμενες μεθόδους. Παρατηρείται η σταδιακή μετάβαση από τη γραμμική συμπίεση (PCA/MDS) στο επιτυχημένο "ξετύλιγμα" της πολλαπλότητας μέσω του Isomap.

- **PCA, MDS, NPP, ONPP:** Παρατηρούμαι οτι εμφανίζουν περιορισμένη ικανότητα αναπαράστασης της μη γραμμικής δομής,λ όπως και περιμέναμε.
- **LPP, OLPP:** Αντίθετα, οι μέθοδοι αυτοι διατηρούν επιτυχώς την τοπική συνεκτικότητα (locality) των δεδομένων, ωστόσο παραμένουν περιορισμένες από τη γραμμική

φύση του μετασχηματισμού, αδυνατώντας να αντιστοιχίσουν πλήρως την καμπυλότητα της πολλαπλότητας σε ένα επίπεδο.

- **Isomap:** Επιτυγχάνει το βέλτιστο αποτέλεσμα, "ισιώνοντας" το Swiss Roll στο δισδιάστατο επίπεδο.

5.3 Επίδραση των Πυρήνων

Η μετάβαση σε kernel εκδοχές των μεθόδων επιτυγχάνεται αντικαθιστώντας τα εσωτερικά γινόμενα με έναν πίνακα Gram K. Σε αντίθεση με τις γραμμικές μεθόδους, οι kernel μέθοδοι:

1. Ανακτούν τις μη γραμμικές μεθόδους (π.χ. Kernel LPP $\equiv$ Laplacian Eigenmaps).
2. Απαιτούν την ορθή επιλογή του παραμέτρου σ του Gaussian πυρήνα, η οποία καθορίζεται συνήθως ως η μισή διάμεσος των αποστάσεων μεταξύ των σημείων.

Η παρατήρηση εμας πιβεβαιώνει τη θεωρία: η μη γραμμική διόρθωση στην αποτυχία της SVD επιτυγχάνεται μέσω του πυρήνα, μεταφέροντας το πρόβλημα σε έναν χώρο Hilbert όπου η γραμμική προβολή αποκτά μη γραμμική ισχύ.

Αναμενόμενες παρατηρήσεις που επιβεβαιώνονται από τα πειράματα:

- Οι ορθογώνιες μέθοδοι (ONPP, OLPP) τείνουν να υπερτερούν, καθώς η διατήρηση των αποστάσεων $\left\|V^\top(x-y)\right\| \leq \|x-y\|$ εξασφαλίζει ότι η γειτονιά στον μειωμένο χώρο αντιστοιχεί πιστά στη γειτονιά του αρχικού χώρου.Ακόμα και αν είναι ιδαίτερα αυστηρές.
- Η LDA περιορίζεται αυστηρά σε $c-1$ κατευθύνσεις, καθιστώντας την ιδανική για ταξινόμηση αλλά περιορισμένη για οπτικοποίηση σε υψηλές διαστάσεις.

Κεφάλαιο 6

Συμπεράσματα

Σύνοψη: Η εργασία ξεκίνησε από τη SVD ως τον γραμμικό πυρήνα της μείωσης διάστασης (μέσω PCA/Eckart–Young), ανέδειξε πότε 'σπάει' σε μη γραμμικά δεδομένα, και παρουσίασε τη μη γραμμική διόρθωση (εκμάθηση πολλαπλοτήτων και πυρήνες) μέσα από ένα ενιαίο πλαίσιο βελτιστοποίησης ίχνους και ιδιοπροβλημάτων. Το κεντρικό μήνυμα είναι : παρά την ποικιλία των μεθόδων, όλες μοιράζονται την ίδια μαθηματική δομή και η φαινομενική 'μη γραμμικότητα' κάποιων εξ αυτών δεν είναι παρά η εφαρμογή των γραμμικών εργαλείων σε έναν κατάλληλα επιλεγμένο χώρο πυρήνων.

Βασικά συμπεράσματα.

- **Ενιαία δομή βελτιστοποίησης.** Όλες οι μέθοδοι είναι εκδοχές του ίδιου προβλήματος $\min / \max \operatorname{Tr}(\cdot)$ υπό περιορισμό ορθογωνιότητας, που λύνεται με (γενικευμένο) ιδιοπρόβλημα. Η διαφορά μεταξύ PCA, LDA, LPP, LLE και Laplacian Eigenmaps έγκειται αποκλειστικά στον πίνακα A που εισέρχεται στο ίχνος: ο πίνακας συνδιακύμανσης $\bar{X}\bar{X}^\top$ για την PCA, ο Λαπλασιανός $L = D - W$ για τα Eigenmaps, ο πίνακας LLE $M = (I - W)^\top (I - W)$ για το LLE. Αυτή η ενοποίηση, που αναδεικνύεται μέσα από το πλαίσιο των Kokiopoulou–Chen–Saad [**?**], καθιστά δυνατή την άμεση σύγκριση και ταξινόμηση των μεθόδων.

- **Γραμμικές μέθοδοι ως προβολές (Rayleigh–Ritz).** Οι γραμμικές μέθοδοι (PCA, LPP, ONPP, LDA) είναι προβεβλημένες εκδοχές των αντίστοιχων μη γραμμικών (implicit) μεθόδων, με χώρο προβολής τον $\operatorname{span}\{X^\top\}$
  Οπερ σημαίνει ότι οι Ritz τιμές της γραμμικής μεθόδου παρεμβάλλονται ανάμεσα στις αληθινές ιδιοτιμές (Λήμμα 2.10), ποσοτικοποιώντας ακριβώς την πληροφορία που χάνεται λόγω της γραμμικής δέσμευσης. Στην υποδειγματοληπτημένη περίπτωση ($\operatorname{rank}(X) = n$) η προβολή καλύπτει ολόκληρο τον $\mathbb{R}^n$ και η ισοδυναμία γίνεται ακριβής (Πόρισμα 3.14): LPP $\equiv$ Eigenmaps και NPP $\equiv$ LLE.

- **Ισοδυναμίες εποπτευόμενων μεθόδων.** Στην εποπτευόμενη ρύθμιση, ο μπλοκ-διαγώνιος πίνακας επίβλεψης H είναι ορθογώνιος προβολέας (Πρόταση 3.10), όπου και αυτή η ιδιότητα αρκεί για να αποδειχθεί ότι LDA, εποπτευόμενη LPP και εποπτευόμενη NPP καταλήγουν στο *ίδιο* ιδιοπρόβλημα (Πρόταση 3.11).
  Η LDA δεν είναι ξεχωριστή μέθοδος: είναι η γραφο-μέθοδος με εποπτευόμενο γράφο $W = H$.

- **Πυρήνες ως πλήρης γέφυρα.** Οι πυρήνες γεφυρώνουν πλήρως τις δύο οικογένειες: Kernel PCA $\equiv$ spectral ratio cut, Kernel LPP $\equiv$ Laplacian Eigenmaps, Kernel NPP

$\equiv$ LLE. Το κοινό τέχνασμα — πολλαπλασιασμός από αριστερά με $\Phi^\top$ και αντικατάσταση $\Phi^\top\Phi \to K$ — μετατρέπει κάθε ιδιοπρόβλημα του χώρου χαρακτηριστικών σε $n \times n$ πρόβλημα, ανεξάρτητα από τη διάσταση του $\mathcal{H}$. Έτσι η "μη γραμμική διόρθωση" στην αποτυχία της SVD αποκτά μια καθαρή αλγεβρική ερμηνεία: αντικαθιστούμε το Ευκλείδειο εσωτερικό γινόμενο $x_i^\top x_j$ με έναν πυρήνα $k(x_i, x_j)$ που κωδικοποιεί την εγγενή γεωμετρία των δεδομένων.

- **Αριθμητικά αποτελέσματα.** Τα πειράματα στο Swiss Roll (Κεφάλαιο 5) επιβεβαίωσαν ποσοτικά τη θεωρητική ιεραρχία. Η PCA και το MDS, περιορισμένα από τη γραμμικότητά τους, αδυνατούν να "ξετυλίξουν" την πολλαπλότητα: η διακύμανση σκορπίζεται σχεδόν ισόποσα στους τρεις άξονες και η 2D προβολή κρατά μόλις $\sim$71% της συνολικής. Οι μέθοδοι LPP/OLPP διατηρούν την τοπική συνεκτικότητα αλλά παραμένουν δέσμιες της γραμμικής φύσης του μετασχηματισμού. Το Isomap, αντικαθιστώντας τις Ευκλείδειες με γεωδαισιακές αποστάσεις, επιτυγχάνει το βέλτιστο αποτέλεσμα ανάμεσα στις μη εποπτευόμενες μεθόδους, "ισιώνοντας" πλήρως το Swiss Roll στο δισδιάστατο επίπεδο.

Περιορισμοί και Μελλοντικές Κατευθύνσεις. Παρά την εννοιολογική ομορφιά του ενιαίου πλαισίου, τρία ζητήματα παραμένουν ανοιχτά και χρήζουν περαιτέρω διερεύνησης.

Πρώτον, **η επιλογή του πυρήνα και της παραμέτρου** σ είναι κρίσιμη και δεν έχει ακόμη θεωρητικά τεκμηριωμένη αυτόματη λύση. Η εμπειρική επιλογή της σ ως η μισή διάμεσος των ζευγαρωτών αποστάσεων αποδίδει καλά στην πράξη, αλλά δεν υπάρχουν εγγυήσεις βέλτιστης απόδοσης. Μέθοδοι επικύρωσης (cross-validation) ή εκμάθησης του πυρήνα από τα ίδια τα δεδομένα (kernel learning) αποτελούν φυσικές επεκτάσεις.

Δεύτερον, **η υπολογιστική κλιμάκωση** παραμένει εμπόδιο. Οι kernel μέθοδοι οδηγούν σε πυκνούς $n \times n$ πίνακες K, ενώ και η ίδια η SVD απαιτεί $O(\min(m,n)^2 \max(m,n))$ πράξεις. Για n της τάξης εκατομμυρίων, απαιτούνται τεχνικές όπως η Nyström προσέγγιση για τον K, οι τυχαιοποιημένες μέθοδοι SVD (randomized SVD), ή αραιές παραλλαγές των Λαπλασιανών γράφων.

Τρίτον, **η επιλογή της διάστασης** d εξακολουθεί να βασίζεται κυρίως σε εμπειρικά κριτήρια (ποσοστό διατηρούμενης διακύμανσης, "elbow" στο φάσμα). Αυστηρά κριτήρια για τον αυτόματο προσδιορισμό της εγγενούς διάστασης — πέρα από τη διάσταση συσχέτισης d_{corr} (Παρατήρηση 1.1) — παραμένουν ενεργό πεδίο έρευνας.

Βιβλιογραφία

[1] E. Kokiopoulou, J. Chen, and Y. Saad, *Trace optimization and eigenproblems in dimension reduction methods*, Numerical Linear Algebra with Applications, 18:565–602, 2011.

[2] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed., Johns Hopkins University Press, 2013.

[3] C. Eckart and G. Young, *The approximation of one matrix by another of lower rank*, Psychometrika, 1(3):211–218, 1936.

[4] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed., Springer, 2002.

[5] S. Roweis and L. Saul, *Nonlinear dimensionality reduction by locally linear embedding*, Science, 290:2323–2326, 2000.

[6] M. Belkin and P. Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural Computation, 15(6):1373–1396, 2003.

[7] J. B. Tenenbaum, V. de Silva, and J. C. Langford, *A global geometric framework for nonlinear dimensionality reduction*, Science, 290:2319–2323, 2000.

[8] B. Schölkopf, A. Smola, and K.-R. Müller, *Nonlinear component analysis as a kernel eigenvalue problem*, Neural Computation, 10:1299–1319, 1998.

[9] N. Aronszajn, *Theory of reproducing kernels*, Transactions of the American Mathematical Society, 68:337–404, 1950.

[10] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed., Cambridge University Press, 2013.

[11] B. Schölkopf and A. J. Smola, *Learning with Kernels*, MIT Press, 2002.

[12] U. von Luxburg, *A tutorial on spectral clustering*, Statistics and Computing, 17(4):395–416, 2007.

[13] F. R. K. Chung, *Spectral Graph Theory*, American Mathematical Society, 1997.