Αναγνώριση Προτύπων

Επιλογή Χαρακτηριστικών

PCA - ISOMAP

Ανδρέας Λ. Συμεωνίδης

Αν. Καθηγητής

Τμήμα Ηλεκτρολόγων Μηχ/κών

&

Μηχ/κών Υπολογιστών, Α.Π.Θ.

Email: asymeon@eng.auth.gr

issel
Intelligent Systems & Software Engineering Labgroup

ARISTOTLE UNIVERSITY OF THESSALONIKI

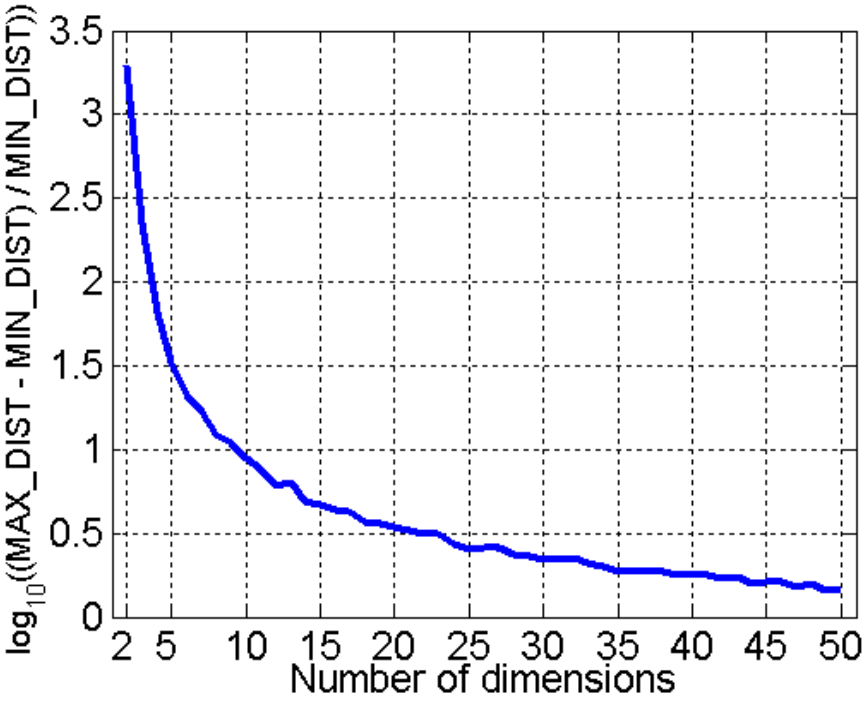
Διάρθρωση διάλεξης

- Ανάλυση κύριων συνιστωσών –
  Principal Component Analysis (PCA) και SVD
- Ισομετρική αντιστοίχιση χαρακτηριστικών –
  Isometric Feature Mapping (ISOMAP) και LLE

Η κατάρα της διαστασιμότητας

- Όταν αυξάνονται οι διαστάσεις, τα δεδομένα γίνονται αυξανόμενα αραιά

- Οι ορισμοί πυκνότητας και απόστασης ανάμεσα στα σημεία τα οποία είναι σημαντικά κριτήρια για ομαδοποίηση και ανίχνευσης outliers, χάνουν το νόημά τους



- Δημουργήστε τυχαία 500 σημεία
- Υπολογίστε την max και min απόσταση ανάμεσα σε οποιοδήποτε ζεύγος σημείων

Μείωση διαστάσεων

- Ο στόχος
  - Οι γεμάτες ουσία δομές χαμηλών διαστάσεων, κρυμμένες μέσα στις υψηλών διαστάσεων παρατηρήσεις τους.
- Κλασικές τεχνικές
  - PCA (Principle Component Analysis)
    - διατηρεί τη διακύμανση
  - MDS (MultiDimensional Scaling)
  - ISOMAP
    - διατηρεί την απόσταση μεταξύ των σημείων
  - LLE (Locally Linear Embedding)

Δεδομένα σε μορφή πίνακα

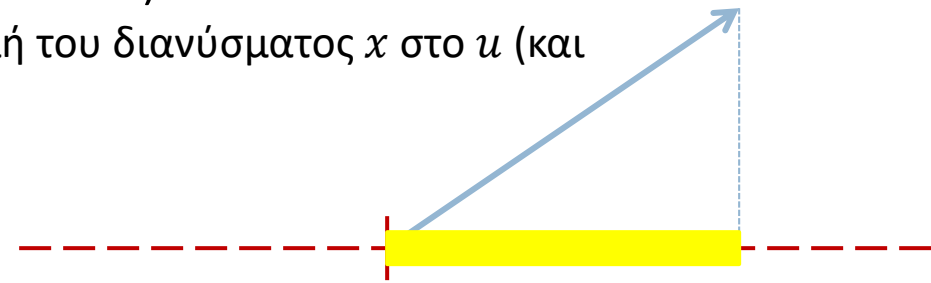
- Έστω *n* εγγραφές με *d* numerical χαρακτηριστικά. Άρα, κάθε εγγραφή περιγράφεται από *d* αριθμητικές τιμές.

- Αναπαριστούμε τα δεδομένα ως έναν $n \times d$ πίνακα *A* με πραγματικούς αριθμούς.
  - Μπορούμε να χρησιμοποιήσουμε εργαλεία γραμμικής άλγεβρας για να επεξεργαστούμε τον πίνακα

- Στόχος μας είναι να δημιουργήσουμε έναν νέο $n \times k$ πίνακα *B* τέτοιον ώστε:
  - Να διατηρεί όσο περισσότερη πληροφορία από τον αρχικό πίνακα *A*
  - Να αποκαλύπτει κάτι από τη δομή των δεδομένων του πίνακα *A*

Λίγο από γραμμική άλγεβρα…

- Υποθέτουμε ότι τα διανύσματα είναι διανύσματα στήλη.
- Ορίζουμε ως x^T για τον ανάστροφο του διανύσματος x (διάνυσμα γραμμή)
- Εσωτερικό γινόμενο: $u^T x$ $(1\times n, n\times 1 \rightarrow 1\times 1)$
  - Το εσωτερικό γινόμενο είναι η προβολή του διανύσματος x στο u (και αντίστροφα)



- $[1,2,3]\begin{bmatrix}4\\1\\2\end{bmatrix} = 12$
- $u^T x = \|x\|\|u\|\cos(u,x)$
  - Εάν $||u|| = 1$ (μοναδιαίο διάνυσμα) τότε το $u^T x$ είναι το μήκος προβολής του x στο u
- Εάν $[-1,2,3]\begin{bmatrix}4\\-1\\2\end{bmatrix} = 0$ ορθογώνια διανύσματα
  - Ορθοκανονικά διανύσματα: δυο μοναδιαία διανύσματα που είναι ορθογώνια

Ιδιοδιανύσματα

- (Δεξί) Ιδιοδιάνυσμα του πίνακα *A*: ένα διάνυσμα x τέτοιο ώστε $Ax = \lambda x$
- λ: ιδιοτιμή του ιδιοδιανύσματος x
- Τάξη του *A*: ο αριθμός των γραμμικά ανεξάρτητων διανυσμάτων γραμμή (ή στήλη)
- Ένας τετραγωνικός πίνακα *A* τάξης *r*, έχει *r* ορθοκανονικά ιδιοδιανύσματα $u_1, u_2, \dots, u_r$ με ιδιοτιμές $\lambda_1, \lambda_2, \dots, \lambda_r$.
- Τα ιδιοδιανύσματα ορίζουν μια ορθοκανονική βάση για τον χώρο στηλών του *A*

Singular Value Decomposition (SVD)

- Δεδομένου οποιουδήποτε $m \times n$ πίνακα A, ο SVD είναι ένας αλγόριθμος ο οποίος βρίσκει πίνακες U, X, και W τέτοιους ώστε:

$$A = U W X^T$$

U είναι $m \times n$ και ορθοκανονικός

W είναι $n \times n$ και διαγώνιος

X είναι $n \times n$ και ορθοκανονικός

$$\begin{pmatrix} A \end{pmatrix} = \begin{pmatrix} U \end{pmatrix} \begin{pmatrix} w_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & w_n \end{pmatrix} \begin{pmatrix} X \end{pmatrix}^T$$

SVD (cont…)

- Τα βάρη w_i ονομάζονται singular values του A
- Εάν ο A είναι singular, μερικά από τα βάρη w_i θα είναι 0
- Γενικά, η τάξη του A: *rank*(A) = k, ο αριθμός των μη-μηδενικών βαρών w_i
- *Χρησιμοποιείται συχνά στην Ανάλυση Κυρίων Συνιστωσών* (Principal Component Analysis)

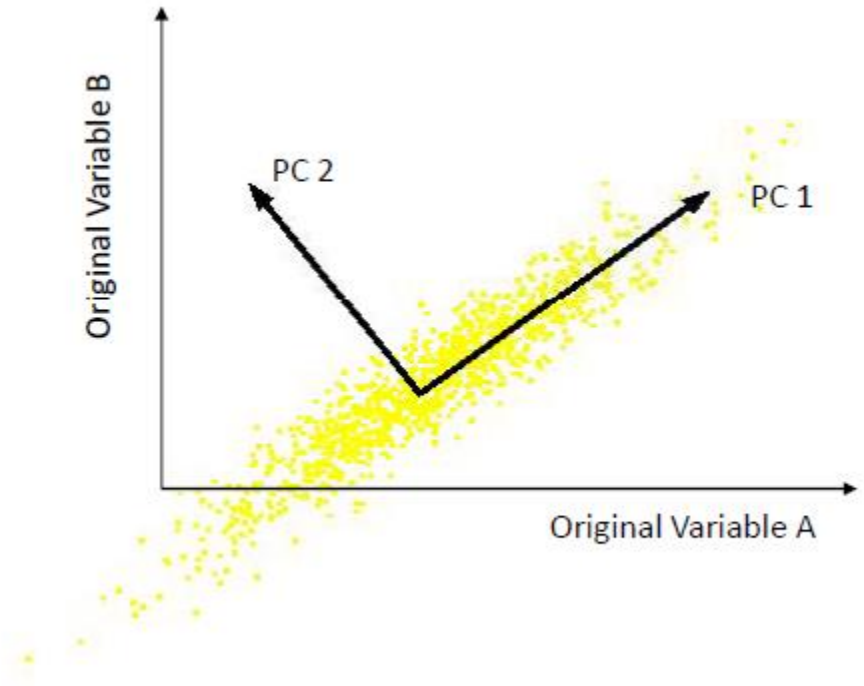
Principal Component Analysis

- Ανάλυση Κυρίων Συνιστωσών
  - Η πιο συνηθισμένη μορφή ανάλυσης
- Οι νέες μεταβλητές/διαστάσεις:
  - Είναι γραμμικός συνδυασμός των βασικών
  - Είναι ασυσχέτιστες μεταξύ τους
- Ορθογώνιες στον αρχικό χώρο διαστάσεων
  - Συλλαμβάνουν όσο το δυνατόν μεγαλύτερη από την αρχική διακύμανση στα δεδομένα
  - Ονομάζονται **Κύριες Συνιστώσες**

Principal Component Analysis (PCA)

- Γενική Αρχή
  - Μέθοδος Γραμμικής προβολής με στόχο τη μείωση του αριθμού μεταβλητών
  - Μεταφέρει ένα σετ από συσχετιζόμενες μεταβλητές σε ένα νέο σετ ασυσχέτιστων μεταβλητών
  - Αντιστοίχιση των δεδομένων σε έναν νέο χώρο χαμηλότερης διάστασης
  - Μια μορφή μη-επιβλεπόμενης μάθησης (π.χ. ομαδοποίηση)
- Ιδιότητες
  - Μπορεί να θεωρηθεί ως περιστροφή των αξόνων σε νέες θέσεις στο χώρο, οι οποίες ορίζονται από τις μεταβλητές του αρχικού χώρου
  - Οι νέοι άξονες είναι ορθογώνιοι και αναπαριστούν τις διευθύνσεις με τη μέγιστη μεταβλητότητα

Ποιοι είναι οι νέοι άξονες;



- Ορθογώνιες διευθύνσεις της μέγιστης διακύμανσης στα δεδομένα
- Οι προβολές κατά μήκος του PC1 διαχωρίζουν μέγιστα τα δεδομένα πάνω σε οποιοδήποτε άξονα.

Κύριες Συνιστώσες

- Η πρώτη κύρια συνιστώσα είναι η διεύθυνση της μεγαλύτερης διακύμανσης (συ-διακύμανσης) στα δεδομένα
- Η δεύτερη κύρια συνιστώσα είναι επόμενη ορθογώνια (ασυσχέτιστη) διεύθυνση μεγαλύτερης διακύμανσης στα δεδομένα
- Κατά συνέπεια, πρώτα αφαιρέστε όλη τη μεταβλητότητα ως προς την πρώτη συνιστώσα και στη συνέχεια βρείτε την επόμενη διεύθυνση μέγιστης μεταβλητότητας
- Και ούτω καθεξής…

Υπολογισμός των Συνιστωσών

- Διακύμανση: $\mathrm{var}(x) = \frac{1}{N}\sum_{n=1}^{N}\{u^T x - u^T \bar{x}\}^2 = u^T S u$
- Πίνακας συμμεταβλητότητας: $S = \frac{1}{N}\sum_{n=1}^{N}(x-\bar{x})(x-\bar{x})^T$
- Ο πίνακας *S* περιέχει τις συσχετίσεις (ομοιότητες) των αρχικών αξόνων, βάσει των προβολών των δεδομένων πάνω τους
- Κατά συνέπεια, θέλουμε να μεγιστοποιήσουμε το $u^T S u$, δεδομένου του ότι το *u* είναι η νέα μονάδα μέτρησης ($u^T u = 1$).

Υπολογισμός των Συνιστωσών (συν.)

- Εισαγωγή ενός Langrange πολλαπλασιαστή λ:

$$u^T S u + \lambda(1 - u^T u)$$

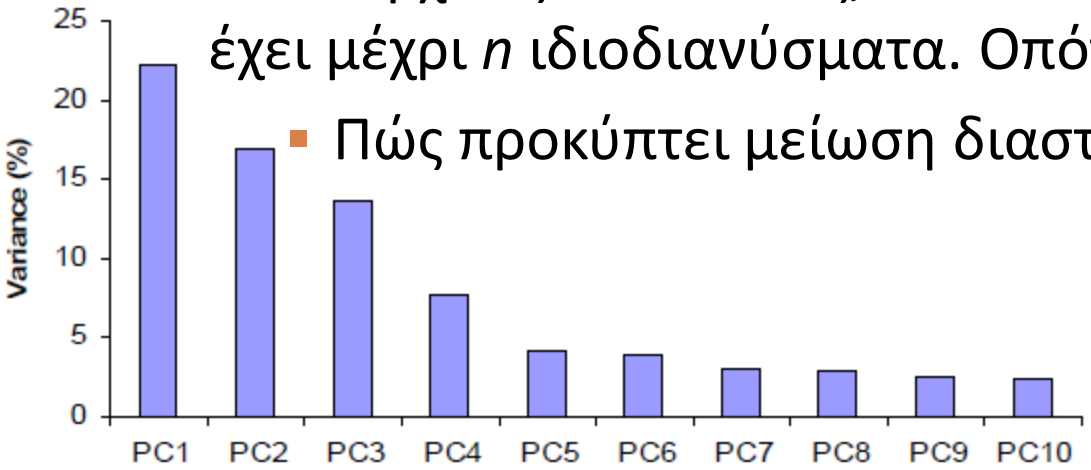
- Παραγωγίζουμε και θέτουμε το διάνυσμα των μερικών παραγώγων ίσο με μηδέν:

$$Su - \lambda u = (S - \lambda I)\, u = 0$$

- Καθώς $u \neq 0$ το u πρέπει να είναι ιδιοδιάνυσμα του xx^T με ιδιοτιμή λ
- Η ιδιοτιμή δηλώνει το βαθμό μεταβλητότητας που περικλείεται κατά μήκος της διεύθυνσης αυτής

Πόσες Κύριες Συνιστώσες;

- Για *n* αρχικές διαστάσεις, ο πίνακας συσχέτισης είναι *nxn*, και έχει μέχρι *n* ιδιοδιανύσματα. Οπότε, *n* ΚΣ.
  - Πώς προκύπτει μείωση διαστάσεων;



- Μπορούν να αγνοηθούν οι ΚΣ με μικρότερη σημασία:
  - *n* διαστάσεις στα αρχικά δεδομένα
  - υπολογισμός *n* ιδιοδιανυσμάτων και ιδιοτιμών
  - επιλέξτε μόνο τα πρώτα *p* ιδιοδιανύσματα, βάσει των ιδιοτιμών τους
  - Το τελικό σετ έχει *p* διαστάσεις

Παράδειγμα PCA – Βήμα 1

- Έστω σετ δεδομένων με 2 μεταβλητές x1, x2.
- Αφαίρεση της μέσης τιμής από κάθε μια από τις μεταβλητές. Για όλες τις *x1* τιμές αφαιρείται το $\overline{x1}$ και για όλες τις x2 τιμές αφαιρείται το $\overline{x2}$. Αυτό παράγει ένα σετ με μέσο όρο μηδέν.

  Με την αφαίρεση αυτή γίνεται πιο εύκολος ο υπολογισμός της διακύμανσης και της συμμεταβλητότητας. Διακύμανση και συμμεταβλητότητα δεν επηρεάζονται από τη μέση τιμή.

http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

Παράδειγμα PCA – Βήμα 1

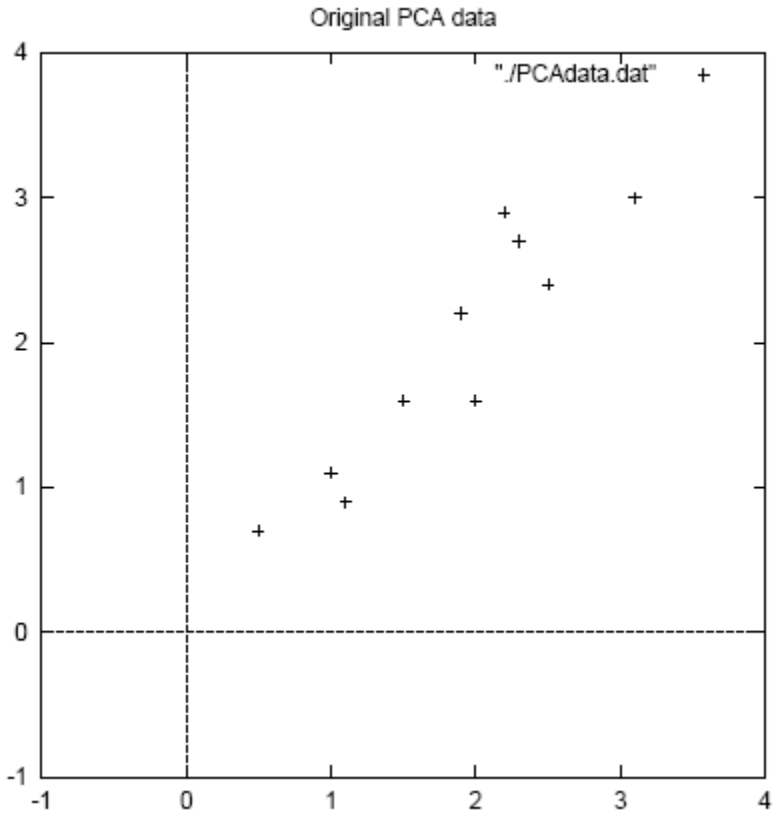
Δεδομένα:

x1	x2
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Δεδομένα με μέση τιμή μηδέν:

x1	x2
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

Παράδειγμα PCA – Βήμα 1



Παράδειγμα PCA – Βήμα 2

- Υπολογισμός του πίνακα συμμεταβλητότητας

$$\text{cov} = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

Θυμηθείτε ότι $\text{cov} = 1/n * \begin{pmatrix} E(x_1, x_1) & E(x_1, x_2) \\ E(x_2, x_1) & E(x_2, x_2) \end{pmatrix}$, $E(x, y) = xy^T$

- Καθώς στα στοιχεία του πίνακα συμμεταβλητότητας που δεν είναι στη διαγώνιο είναι θετικά, περιμένουμε ότι *x1* και *x2* αυξάνονται παράλληλα.

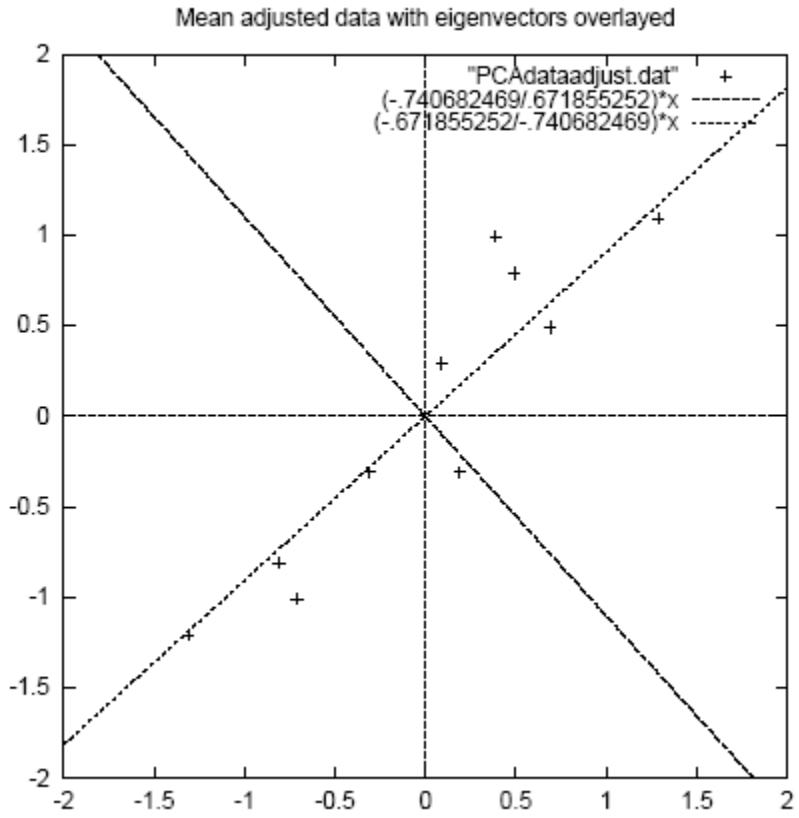
PCA Παράδειγμα PCA – Βήμα 3

- Υπολογίστε τα ιδιοδιανύσματα και τις ιδιοτιμές του πίνακα συμμεταβλητότητας

$$\text{eigenvalues} = \begin{pmatrix} .04908339 \\ 1.28402771 \end{pmatrix}$$

$$\text{eigenvectors} = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

Παράδειγμα PCA – Βήμα 3



- Τα ιδιοδιανύσματα σχεδιάζονται ως διαγώνιες διακεκομένες γραμμές στον χώρο.
- Παρατηρείστε ότι είναι κάθετες η μια στην άλλη.
- Παρατηρείστε ότι ένα από τα ιδιοδιανύσματα περνά από τη μέση των σημείων, σα να τραβά μια γραμμή βέλτιστης απεικόνισης.
- Το δεύτερο ιδιοδιάνυσμα δηλώνει το άλλο, λιγότερο σημαντικό πρότυπο στα δεδομένα, ότι όλα τα σημεία ακολουθούν την κύρια συνιστώσα, αλλά απέχουν από αυτή σε κάποιο ποσοστό.

Παράδειγμα PCA – Βήμα 4

- Μειώστε τις διαστάσεις και φτιάξτε το *διάνυσμα μεταβλητών – feature vector*

  το ιδιοδιάνυσμα με την μεγαλυτερη ιδιοτιμή είναι η κύρια συνιστώσα του σετ δεδομένων.

- Στην περίπτωση του παραδείγματος, το ιδιοδιάνυσμα με την μεγαλυτερη ιδιοτιμή είναι αυτό που περνά μέσα από τα δεδομένα.

- Ταξινομήστε τα ιδιοδιανύσματα απο αυτό με τη μεγαλύτερη ιδιοτιμή σε αυτό με τη μικρότερη ιδιοτιμή. Αυτό δηλώνει και τη σημασία των συνιστωσών.

Παράδειγμα PCA – Βήμα 4

- Τώρα μπορείτε να επιλέξετε να *αγνοήσετε* τις συνιστώσες μικρότερης σημασίας.
- Χάνετε πληροφορία, που είναι λιγότερη όσο μικρότερες είναι οι ιδιοτιμές.
- Feature Vector = (eig_1 eig_2 eig_3 … eig_n)
  Μπορείτε να επιλέξετε να φτιάξετε ένα feature vector και με τα δυο ιδιοδιανύσματα:

$$\begin{pmatrix} -.677873399 & -.735178656 \\ -.735178656 & .677873399 \end{pmatrix}$$

ή μπορείτε να αγνοήσετε το μικρότερο ιδιοδιάνυσμα:

$$\begin{pmatrix} -.677873399 \\ -.735178656 \end{pmatrix}$$

Παράδειγμα PCA – Βήμα 5

- Εξαγωγή των νέων δεδομένων

  FinalData = RowFeatureVector x RowZeroMeanData

  όπου:

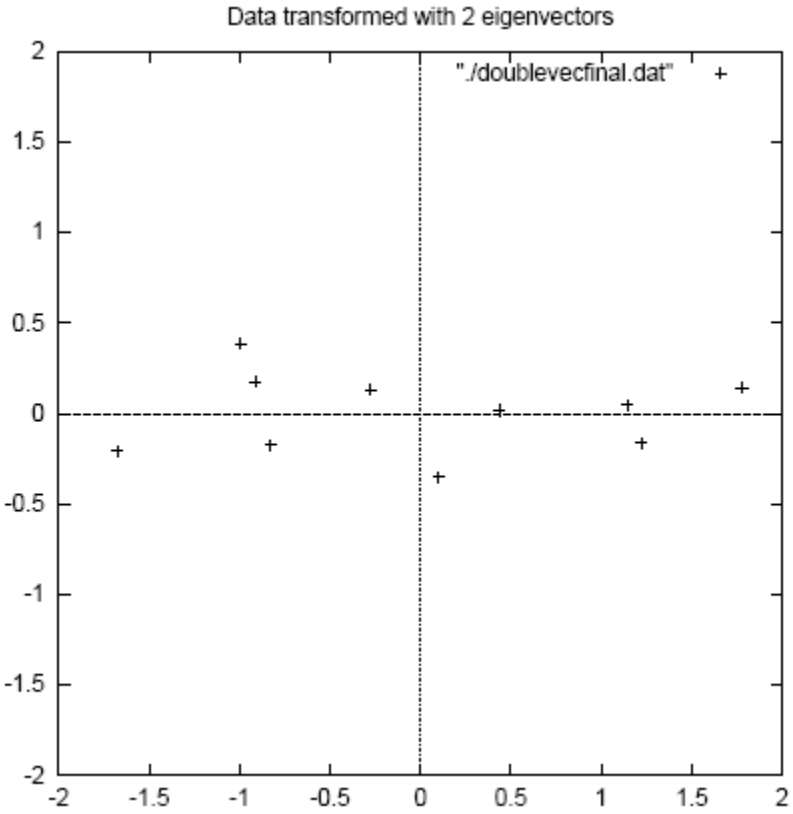
- RowFeatureVector είναι ένας πίνακας με τα ιδιοδιανύσματα στις στήλες *ανεστραμμένα*, ώστε τα ιδιοδιανύσματα να είναι σε γραμμές, με το πιο σημαντικό ιδιοδιάνυσμα στην πρώτη γραμμή
- RowZeroMeanData είναι τα κανονικοποιημένα (ως προς τη μέση τιμή) δεδομένα *ανεστραμμένα*, δηλαδή τα σημεία σε κάθε στήλη, με κάθε γραμμή να περιέχει μια διάσταση.

Παράδειγμα PCA – Βήμα 5

FinalData αναστροφή: οι διαστάσεις στις στήλες

x1	x2
-.827970186	-.175115307
1.77758033	.142857227
-.992197494	.384374989
-.274210416	.130417207
-1.67580142	-.209498461
-.912949103	.175282444
.0991094375	-.349824698
1.14457216	.0464172582
.438046137	.0177646297
1.22382056	-.162675287

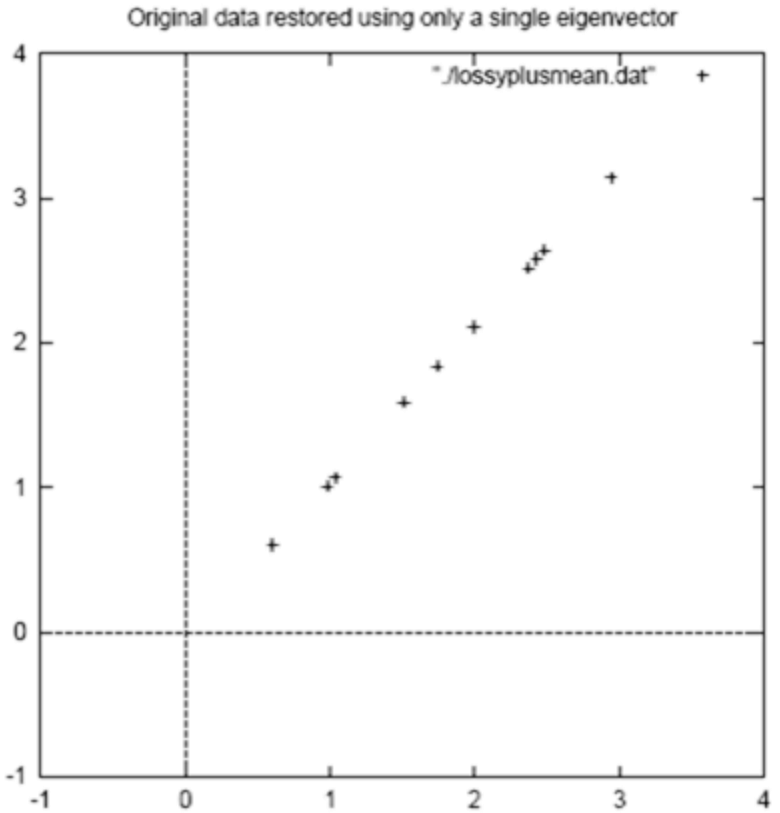
Παράδειγμα PCA – Βήμα 5



Ανακατασκευή των αρχικών δεδομένων

- Κατά την ανασκευή των δεδομένων η πληροφορία των διαστάσεων που μειώσαμε χάνεται. Έστω ότι κρατήσαμε μόνο την x1 διάσταση…

x1
-.827970186
1.77758033
-.992197494
-.274210416
-1.67580142
-.912949103
.0991094375
1.14457216
.438046137
1.22382056



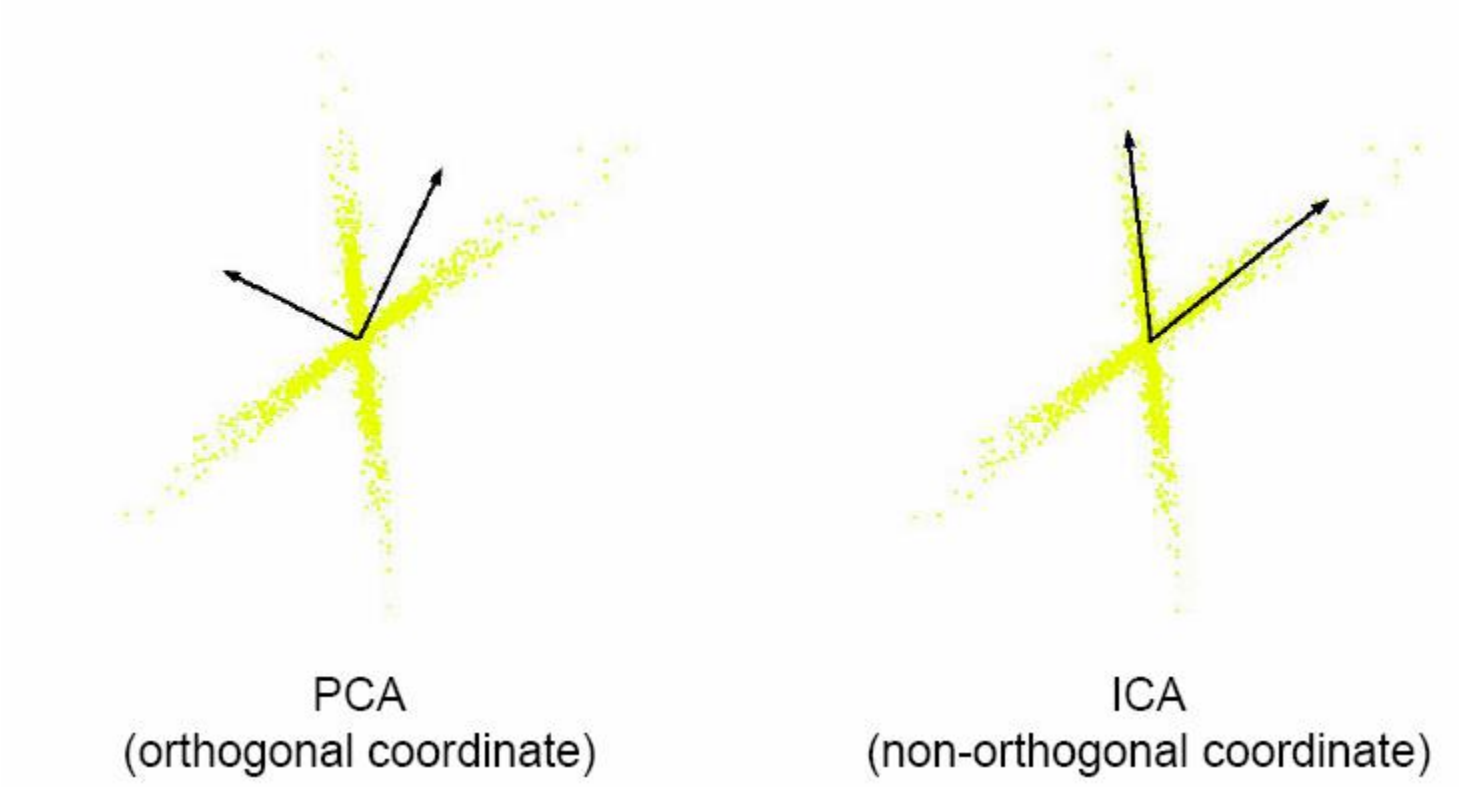
Η σημασία του PCA

- Σε δεδομένα με πολλές διαστάσεις όπου η γραφική αναπαράσταση είναι δύσκολη, η PCA είναι ένα δυνατό εργαλείο για την ανάλυση των δεδομένων και την εύρεση προτύπων μέσα σε αυτά.
- Η PCA χρησιμεύει πολύ στη συμπίεση δεδομένων
- Η πιο αποδοτική έκφραση δεδομένων είναι με τη χρήση ορθογώνιων συνιστωσών, όπως γίνεται στην PCA.

Τι είναι η ICA;

- Στην ICA, μια ανεξάρτητη συνθήκη βελτιστοποιείται (αντί μόνο για τη διακύμανση), η οποία συχνά δίνει πιο κατανοητές συνιστώσες από την PCA.
- Οι συνιστώσες αυτές ονομάζονται ανεξάρτητες (Independent Components - ICs) και αναπαριστούν μη-επικαλυπτόμενη πληροφορία.
- Για την εφαρμογή ICA θεωρούμε ότι τα δεδομένα έχουν καθοριστεί από **κάποιους βασικούς παράγοντες, οι οποίοι είναι ανεξάρτητοι μεταξύ τους**
- Η αναζήτηση συνιστωσών οι οποίες είναι όσο πιο στατιστικά ανεξάρτητες γίνεται, μπορούν να ανιχνευτούν οι παράγοντες αυτοί.
- Οι παράγοντες λέγονται Sources και το σχετικό πεδίο εφαρμογής Blind source separation - BSS

Οπτική σύγκριση PCA-ICA

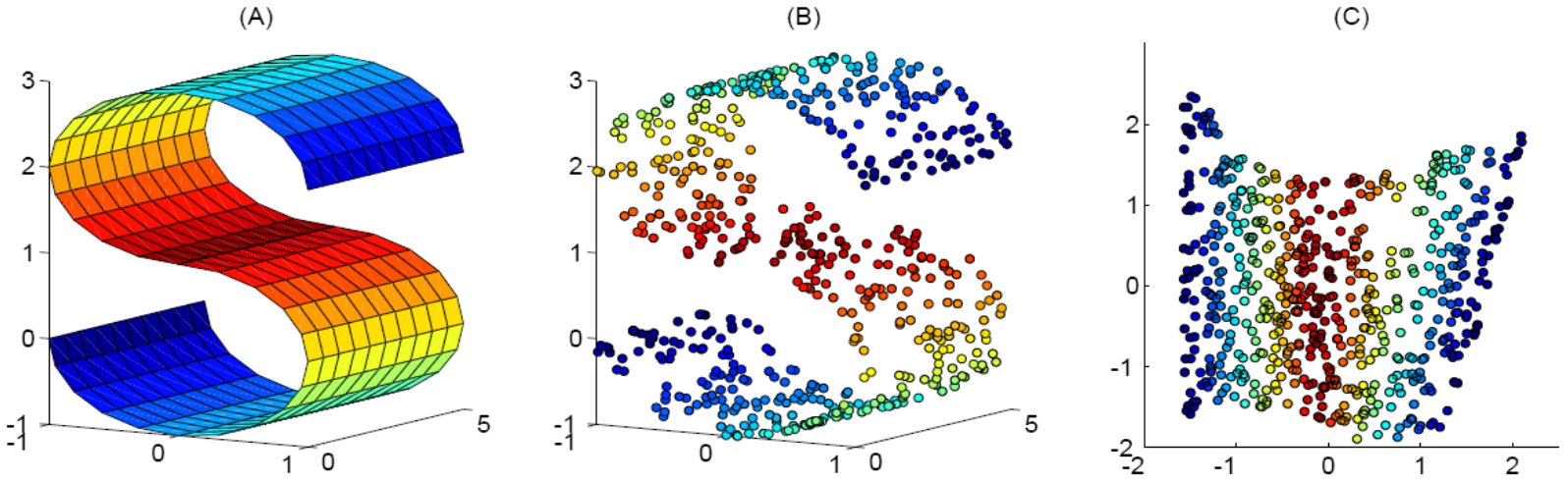


Διαφορές με την PCA

- Πρακτικά, δεν είναι τεχνική μείωσης διαστάσεων
- Δεν υπάρχει μια μοναδική λυση για τις συνιστώσες. Χρησιμοποιεί διαφορετικούς αλγορίθμους (FastICA, PearsonICA, MLICA)
- Τα ICs είναι ασυσχέτιστα, αλλά και όσο πιο ανεξάρτητα γίνεται
- Μη χρήσιμη για μεταβλητές που ακολουθούν κανονική κατανομή

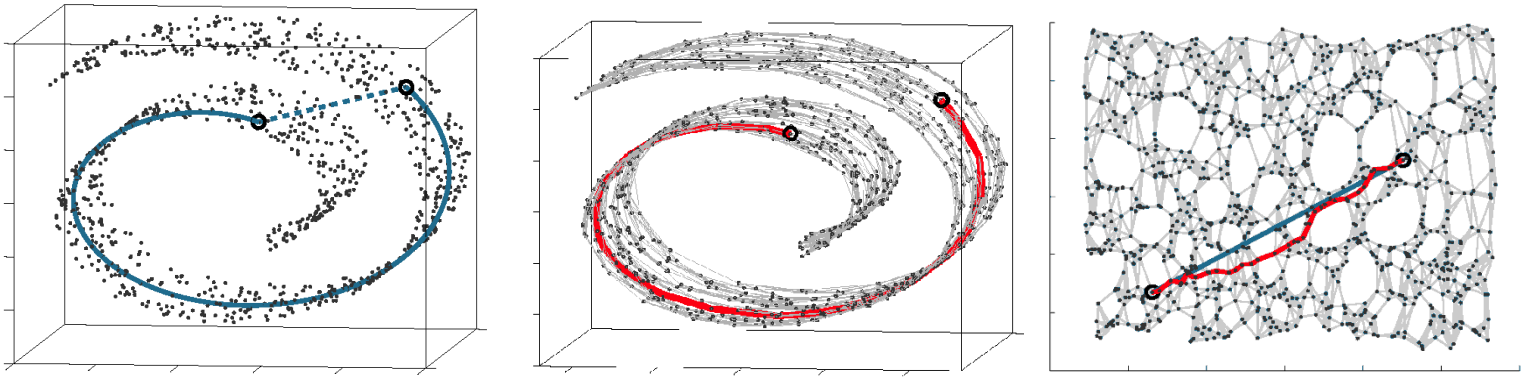
Μη γραμμικές τεχνικές μείωσης διαστάσεων

- Πολλά σετ δεδομένων περιέχουν μη γραμμικές δομές, οι οποίες δε μπορούν να ανιχνευτούν με PCA και MDS
- Χρήση μη γραμμικών τεχνικών μείωσης διαστάσεων



ISOMAP

- Παράδειγμα μη γραμμικής δομής (Swiss roll)
  - Μόνο οι γεοδεσικές αποστάσεις αναπαριστούν τη γεωμετρία του αντικειμένου.
- ISOMAP (Isometric feature Mapping)
  - Αναγνωρίζει και διατηρεί τη γεωμετρία των δεδομένων.
  - Χρησιμοποιεί γεοδεσικές αποστάσεις ανάμεσα σε όλα τα σημεία.

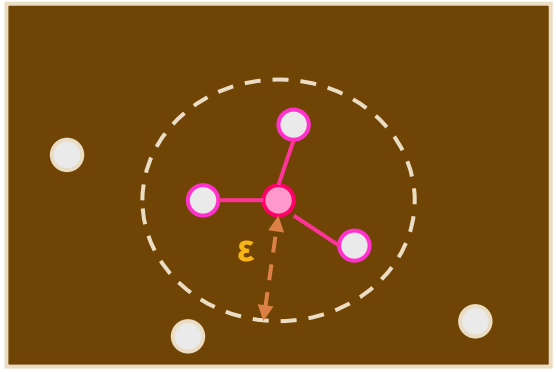


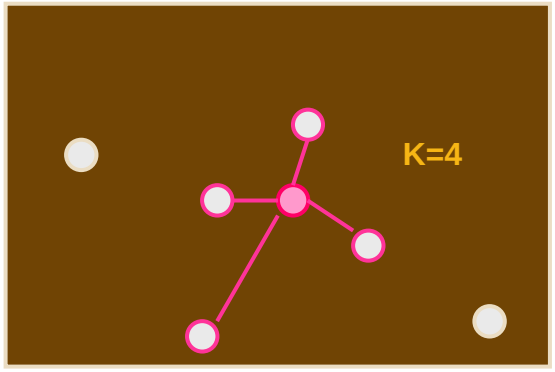
ISOMAP (Περιγραφή του αλγορίθμου)

- Βήμα 1
  - Καθορισμός των γειτονικών σημείων μέσα σε προκαθορισμένη ακτίνα βάσει των αποστάσεων στον αρχικό χώρο $d_X(i, j)$.
  - Αυτές οι σχέσεις γειτνίασης αναπαριστώνται ως ένας γράφος με βάρη G για τα δεδομένα.
- Βήμα 2
  - Υπολογισμός των γεοδεσικών αποστάσεων $d_G(i, j)$ ανάμεσα σε όλα τα ζεύγη σημείων στη δομή με τον υπολογισμό των αποστάσεων μικρότερης διαδρομής μέσα στον γράφο G.
- Βήμα 3
  - Κατασκευή μιας δομής στον d-διάστατο Ευκλείδειο χώρο Y ο οποίος αναπαριστά βέλτιστα τη γεωμετρία της αρχικής δομής.

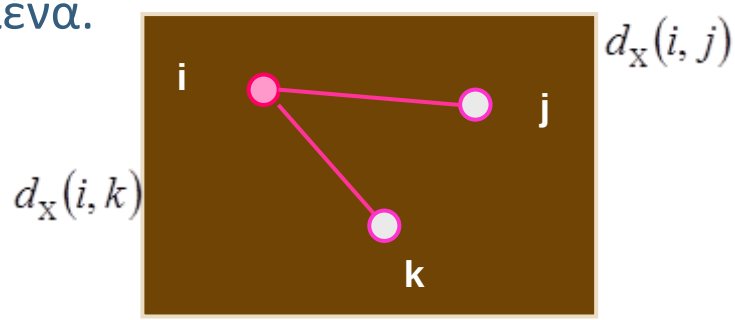
ISOMAP (Περιγραφή του αλγορίθμου)

- Βήμα 1
  - Καθορισμός των γειτονικών σημείων μέσα σε προκαθορισμένη ακτίνα βάσει των αποστάσεων στον αρχικό χώρο $d_X(i, j)$.





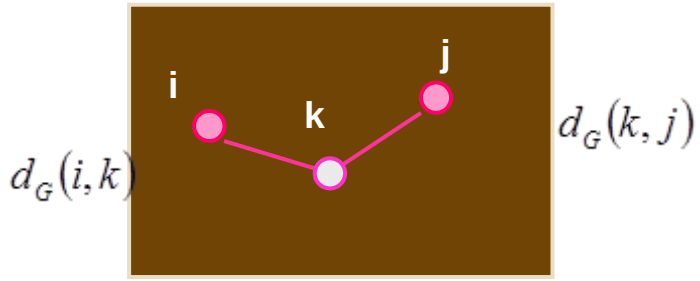
  - Αυτές οι σχέσεις γειτνίασης αναπαριστώνται ως ένας γράφος με βάρη G για τα δεδομένα.



ISOMAP (Περιγραφή του αλγορίθμου)

- Βήμα 2
  - Υπολογισμός των γεοδεσικών αποστάσεων $d_G(i,j)$ ανάμεσα σε όλα τα ζεύγη σημείων στη δομή με τον υπολογισμό των αποστάσεων μικρότερης διαδρομής μέσα στον γράφο G.
  - Μπορεί να γίνει με την εφαρμογή του αλγορίθμου του Dijkstra.

$$d_G(i,j) = d(i,j) \quad \text{neighboring } i,j$$
$$d_G(i,j) = \infty \qquad \text{othewise}$$



$$for\ \mathrm{k} = 1,2,\ldots,\mathrm{N}$$
$$d_G(i,j) = \min\{d_G(i,j),\ d_G(i,k) + d_G(k,j)\}$$

ISOMAP (Περιγραφή του αλγορίθμου)

- Βήμα 3
  - Κατασκευή μιας δομή στον d-διάστατο Ευκλείδειο χώρο Y ο οποίος αναπαριστά βέλτιστα τη γεωμετρία της αρχικής δομής.
  - Ελαχιστοποίηση της συνάρτησης κόστους.

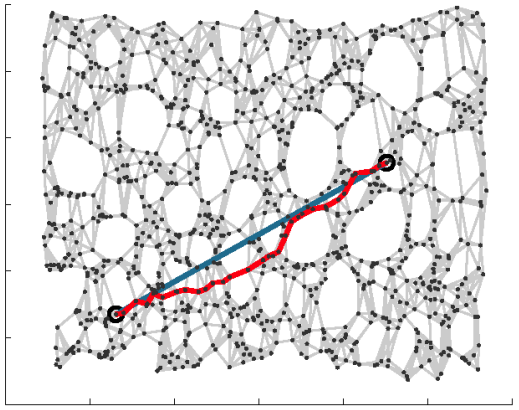
$$E = \left\| \tau(D_G) - \tau(D_Y) \right\|_{L^2}$$

$$where\ D_Y(i,j) = \left\| y_i - y_j \right\|$$

$$D_G(i,j) = d_G(i,j)$$

$$and$$

$$\tau(D) = \frac{-1}{2}(I - \frac{1}{N})D^{.2}(I - \frac{1}{N})$$



Λύση: πάρε τα πρώτα d ιδιοδιανύσματα του πίνακα $\tau(D_G)$

LLE (Local Linear Embedding)

- Διατηρεί τα πρότυπα γειτνίασης
- Αντιστοιχίζει σε ένα γενικό σύστημα συντεταγμένων λίγων διαστάσεων
- Ανακτά ολικές μη-γραμμικές δομές από τοπικά γραμμικές αντιστοιχίσεις
- Κάθε σημείο και οι γείτονές του αναμένεται να βρίσκονται πάνω ή δίπλα σε ένα τοπικά γραμμικό τμήμα.
- Κάθε σημείο κατασκευάζεται από τους γείτονές του:

$$\vec{\hat{X}}_i = \sum_j W_{ij} \vec{X}_j$$

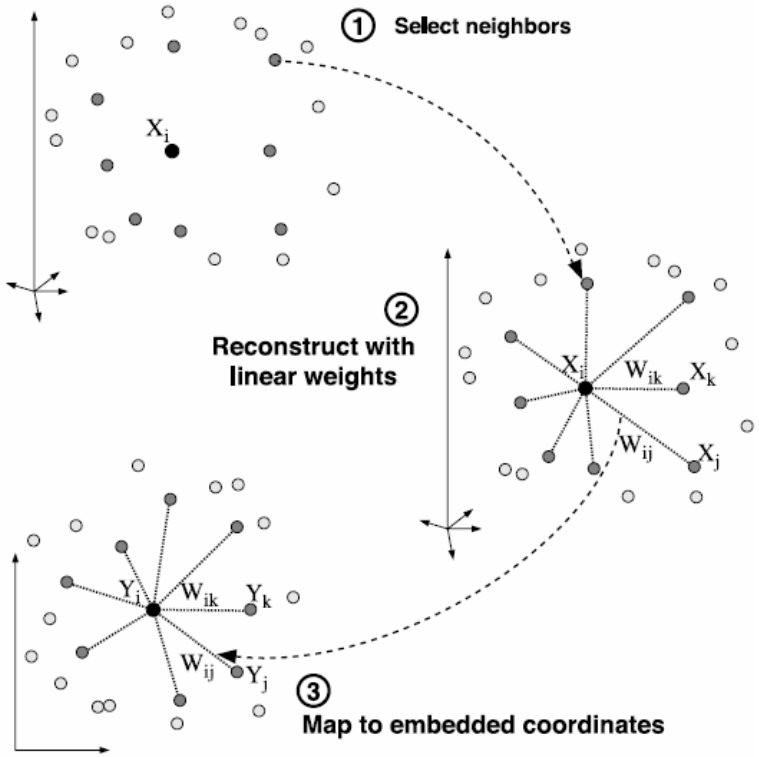
$$W_{ij} = 0 \text{ εάν } \vec{X}_j \text{ δεν είναι γείτονας στο } \vec{X}_i$$

- Όπου τα W_{ij} συνοψίζουν τη συμμετοχή του j-th σημείου στην αναπαράσταση του i-th σημείου
- Κάθε σημείο ανακατασκευάζεται μόνο από τους γείτονές του

LLE (σύνοψη αλγορίθμου)

- Βήμα 1
  - Υπολογίστε τους γείτονες για κάθε σημείο, X_i
- Βήμα 2
  - Υπολογίστε τα βάρη W_{ij} που ανακατασκευάζουν βέλτιστα κάθε σημείο X_i από τους γείτονές του, μειώνοντας το κόστος στη συν (1) με τη χρήση περιορισμένων γραμμικών τμημάτων.

$$\text{1} \quad \varepsilon(W) = \sum_i \left| \vec{X}_i - \sum_j W_{ij} \vec{X}_j \right|^2$$



- Step 3
  - Υπολογίστε τα διανύσματα Y_i τα οποία ανακατασκευάζονται βέλτιστα από τα βάρη W_{ij}, ελαχιστοποιώντας την τετραγωνική μορφή της συν (2) με τα κάτω μη-μηδενικά ιδιοδιανύσματα.

$$\text{2} \quad \phi(Y) = \sum_i \left| \vec{Y}_i - \sum_j W_{ij} \vec{Y}_j \right|^2$$

ISOMAP – σύνοψη…

- Ο Isomap διαχειρίζεται μη γραμμικές δομές.
- Ο Isomap έχει τα πλεονεκτήματα των PCA και MDS.
  - Μη επαναληπτική διαδικασία
  - Πολυωνυμική διαδικασία
  - Εγγυημένη σύγκλιση
- Ο Isomap αναπαριστά τη γενική δομή των δεδομένων σε ένα μοναδικό σύστημα συντεταγμένων.

ISOMAP και LLE: σύνοψη

- ISOMAP και LLE
  - Διαχειρίζονται μη γραμμικές δομές.
- ISOMAP
  - Χρησιμοποιεί τις γεωδαισικές αποστάστεις ανάμεσα σε όλα τα ζευγάρια.
- LLE
  - Ανακτά ολικές μη-γραμμικές δομές από τοπικά γραμμικές αντιστοιχίσεις
- ISOMAP vs LLE
  - Ο LLE χρειάζεται μεγάλα σετ δεδομένων εισόδου και πρέπει να βάρη αντίστοιχων διαστάσεων
  - Ο Isomap είναι πολύ γρήγορος λόγω του αλγορίθμου του dijkstra
  - Ο Isomap είναι πιο πρακτικός από τον LLE

Επιλογή χαρακτηριστικών - Σύνοψη

- Κατά τη διαδικασία καθορισμού και προ-επεξεργασίας των δεδομένων χρειάζεται πολλές φορές και η μείωση του αριθμού των μεταβλητών στο σετ.
- Οι τεχνικές χωρίζονται (κλασικά) σε γραμμικές και μη-γραμμικές.
- Γνωστότερη γραμμική μέθοδος: PCA.
- Γνωστότερη μη-γραμμική μέθοδος: ISOMAP
- Πέρα από αυτές, υπάρχουν και μέθοδοι που εφαρμόζουν παρόμοια λογική, αλλά ακολουθούν άλλες προσεγγίσεις (ICA).