Εθνικό και Καποδιστριακό
ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Αναγνώριση Προτύπων - Μηχανική Μάθηση

Μέθοδοι μείωσης διαστάσεων

Γιάννης Παναγάκης

Workflow για την Αναγνώριση Προτύπων

1. Διατύπωση του προβλήματος ως πρόβλημα μηχανικής μάθησης
2. Συλλογή δεδομένων
3. Προεπεξεργασία δεδομένων
4. Επιλογή του στατιστικού μοντέλου που θα χρησιμοποιηθεί
5. Εκπαίδευση (learn/train/estimate/fit…) του στατιστικού μοντέλου χρησιμοποιώντας δεδομένα εκπαίδευσης
6. Αξιολόγηση της επίδοσης του μοντέλου

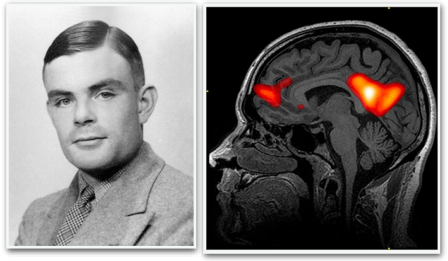
Σε αυτή τη διάλεξη:

- Curse of dimensionality
- Μέθοδοι μείωσης διαστάσεων:
  - Principal Component Analysis (PCA)
  - Non-negative Matrix Factorization NMF

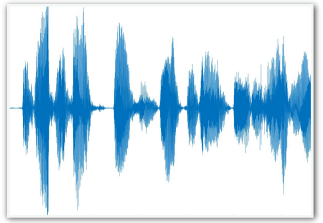
Η διαστατικότητα των δεδομένων και οι συνέπειες της

Δεδομένα εισόδου πολύ μεγάλων διαστάσεων

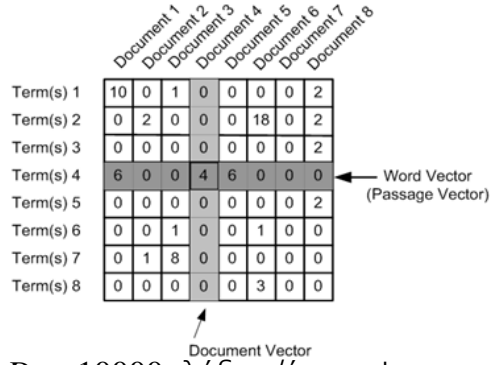
Σήματα υψηλών διαστάσεων: $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^D$



$D = 10^6$ pixels/voxels



$D = 44100$ samples/sec



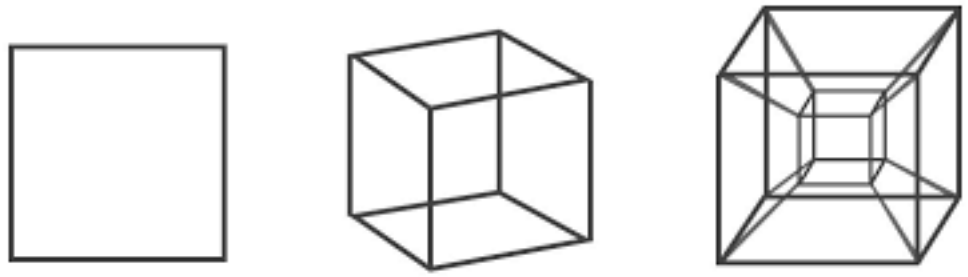
$D = 10000$ λέξεις/έγγραφο

Q: Ποιες είναι οι συνέπειες της μεγάλης διάστασης των δεδομένων στους αλγορίθμους μηχανικής μάθησης;

Γεωμετρία στις μεγάλες διαστάσεις

— Όταν σκεφτόμαστε για τη γεωμετρία χώρων πολλών διαστάσεων συνήθως βασιζόμαστε στην εμπειρία που έχουμε από τη γεωμετρία στις δύο ή τρεις διαστάσεις.

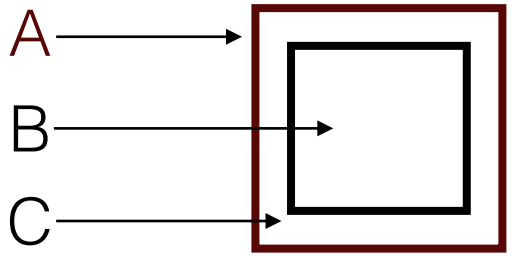
— Ωστόσο, η γεωμετρία των χώρων μεγάλων διαστάσεων είναι “περίεργη” και η διαισθητική κατανόηση της αποτυγχάνει.



Γεωμετρία στις μεγάλες διαστάσεις

Παράδειγμα:

- Έστω Α υπερκύβος D διαστάσεων με πλευρές μήκους ℓ : $[-\frac{\ell}{2}, \frac{\ell}{2}]^D$
- Έστω Β υπερκύβος D διαστάσεων με πλευρές μήκους $\alpha\ell$: $[-\frac{\alpha\ell}{2}, \frac{\alpha\ell}{2}]^D,\ 0 < a < 1$
- Ας ορίσουμε ως κέλυφος C το χώρο που μένει "άδειος" εάν ο Β είναι εγγεγραμμένος στον Α.
- Ποιο είναι το ποσοστό του όγκου του Α το οποίο καταλαμβάνει το κέλυφος C όταν το D είναι πολύ μεγάλο;



$$\frac{\ell^D - (\alpha\ell)^D}{\ell^D} = \frac{\ell^D(1-\alpha^D)}{\ell^D} = 1 - \alpha^D$$

- Αν α=0.99 και D = 1000, τότε το 99.99% του όγκου του Α καταλαμβάνεται από το κέλυφος C!

Έλλειψη τοπικότητας στις μεγάλες διαστάσεις

— Δεν υπάρχουν δεδομένα-διανύσματα μεγάλων διαστάσεων που να βρίσκονται κοντά με βάση κάποια απόσταση στις μεγάλες διαστάσεις.

Παράδειγμα:

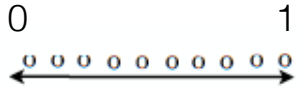
- Ας υποθέσουμε ότι θέλουμε να χρησιμοποιήσουμε τον αλγόριθμο του πλησιέστερου γείτονα για την ταξινόμηση δεδομένων και θέλουμε να συλλέξουμε ένα σύνολο εκπαίδευσης έτσι ώστε το κατώφλι απόφασης του ταξινομητή να είναι ε = 0.1. Δηλαδή για να θεωρηθούν δύο δείγματα γειτονικά θα πρέπει να έχουν απόσταση το πολύ 0.1.

  - Πόσα δεδομένα εκπαίδευσης πρέπει να συλλέξουμε για να καλύψουμε ομοιόμορφα το χώρο διάστασης d;

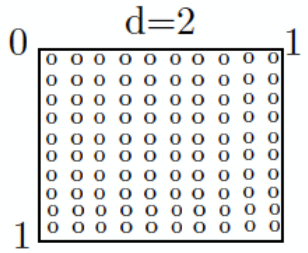
Έλλειψη τοπικότητας στις μεγάλες διαστάσεις

- Πόσα δεδομένα εκπαίδευσης πρέπει να συλλέξουμε για να καλύψουμε ομοιόμορφα το χώρο διάστασης d;

- 10 σημεία καλύπτουν το χώρο [0,1] σε απόσταση 0.1



- 100 σημεία χρειάζονται για το χώρο $[0,1]^2$



- 10^d σημεία χρειάζονται για το χώρο $[0,1]^d$

ϵ^{-d} σημεία (δεδομένα) στη γενική περίπτωση

Curse of dimensionality

— Όσο μεγαλώνει η διάσταση των δεδομένων, χρειαζόμαστε εκθετικό στη διάσταση των δεδομένων πλήθος δεδομένων έτσι ώστε να υπάρχουν εγγυήσεις γενίκευσης.

— Τα δεδομένα μεγάλης διάστασης απαιτούν αυξημένο χώρο μνήμης για την αποθήκευση τους και αυξημένο χρόνο για την εξεργασία τους.

— Το σύνολο των συνεπειών της μεγάλης διάστασης των δεδομένων αναφέρεται συχνά ως κατάρα της διαστατικότητας (curse of dimensionality).

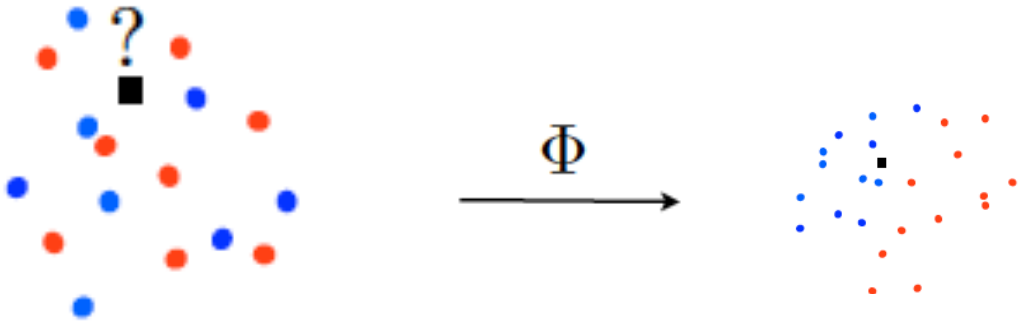
Richard E. Bellman

Q: Πώς θα προσεγγίσουμε συναρτήσεις από δεδομένα/σήματα πολλών διαστάσεων όταν έχουμε περιορισμένα σε πλήθος παραδείγματα εκπαίδευσης;

Εξαγωγή χαρακτηριστικών - Μάθηση αναπαραστάσεων

Χώρος δεδομένων

$$\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^D$$
$$D \to \infty$$



Αναπαράσταση δεδομένων

$$\Phi(x) \in \mathcal{R} \subseteq \mathbb{R}^d$$
$$d \ll D$$

$$\|\mathbf{x}_1 - \mathbf{x}_2\|$$

Δεν φέρει πληροφορία

$$\|\Phi(\mathbf{x}_1) - \Phi(\mathbf{x}_2)\|$$

Φέρει πληροφορία

Στόχος: μείωσης της διάστασης των δεδομένων, εφαρμόζοντας το μετασχηματισμό Φ(Χ)

$$\Phi : \mathcal{X} \to \mathcal{R}$$

$$\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^D \xrightarrow{\Phi} \Phi(x) \in \mathcal{R} \subseteq \mathbb{R}^d$$

Εξαγωγή χαρακτηριστικών

Χώρος δεδομένων

$$\Phi : \mathcal{X} \rightarrow \mathcal{F}$$

Αναπαράσταση σήματος

$$\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^D \xrightarrow{\Phi} \Phi(x) \in \mathcal{F} \subseteq \mathbb{R}^d$$

$$\|\mathbf{x}_1 - \mathbf{x}_2\|$$

Δεν φέρει πληροφορία

$$\|\Phi(\mathbf{x}_1) - \Phi(\mathbf{x}_2)\|$$

Φέρει πληροφορία

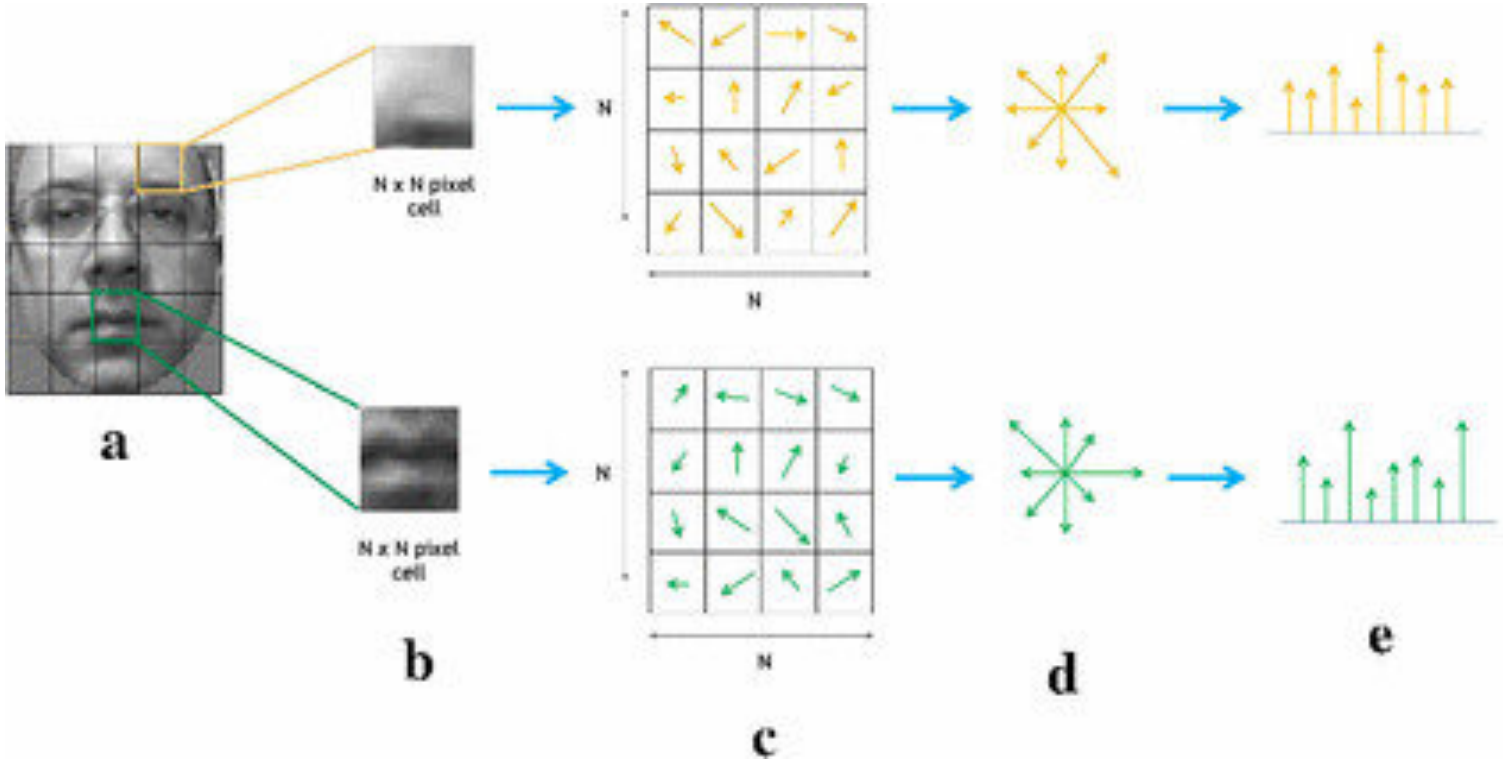
Αν ο μετασχηματισμός Φ είναι προκαθορισμένος, τότε αναφερόμαστε σε εξαγωγή χαρακτηριστικών, πχ. SIFT, HOGs, MFCCs

Εξαγωγή χαρακτηριστικών

Χώρος δεδομένων

$$\Phi : \mathcal{X} \rightarrow \mathcal{F}$$

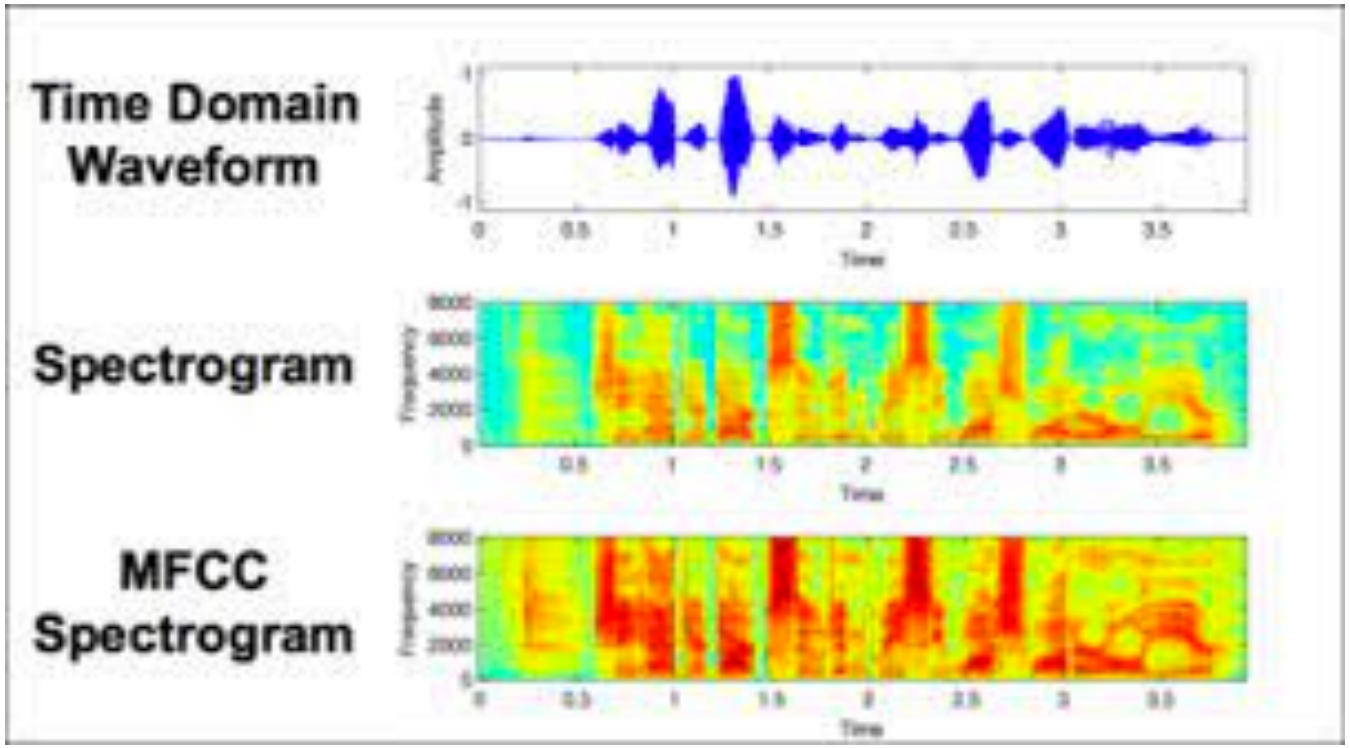
Αναπαράσταση σήματος



Histogram of oriented gradients (HoG)

Εξαγωγή χαρακτηριστικών

Χώρος δεδομένων $\Phi : \mathcal{X} \rightarrow \mathcal{F}$ Αναπαράσταση σήματος



Mel Frequency Cepstral Coefficient - MFCCs

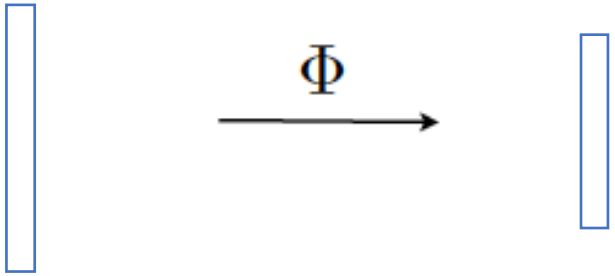
Μάθηση αναπαραστάσεων

Χώρος δεδομένων

$$\Phi : \mathcal{X} \rightarrow \mathcal{R}$$

Αναπαράσταση σήματος

$$\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^D$$



$$\Phi(x) \in \mathcal{R} \subseteq \mathbb{R}^d$$

$$\|\mathbf{x}_1 - \mathbf{x}_2\|$$

Δεν φέρει πληροφορία

$$\|\Phi(\mathbf{x}_1) - \Phi(\mathbf{x}_2)\|$$

Φέρει πληροφορία

Αν τα δεδομένα έχουν κάποια λανθάνουσα γεωμετρική δομή μπορούμε να μάθουμε το μετασχηματισμό Φ από τα δεδομένα, χρησιμοποιώντας κυρίως μεθόδους μάθησης χωρίς επίβλεψη. Στη περίπτωση αυτή αναφερόμαστε σε μάθηση αναπαραστάσεων (representation learning).

Μείωση διαστάσεων (Dimensionality reduction)

—Το πρόβλημα της μείωσης διαστάσεων μπορεί να θεωρηθεί μαθηματικά ως πρόβλημα εύρεσης μίας απεικόνισης ή μετασχηματισμού (M) από ένα χώρο πολλών διαστάσεων σε ένα χώρο (πολύ) μικρότερης διάστασης έτσι ώστε ικανοποιείται κάποιο κατάλληλο κριτήριο (π.χ. ακρίβεια ανακατασκευής των δεδομένων).

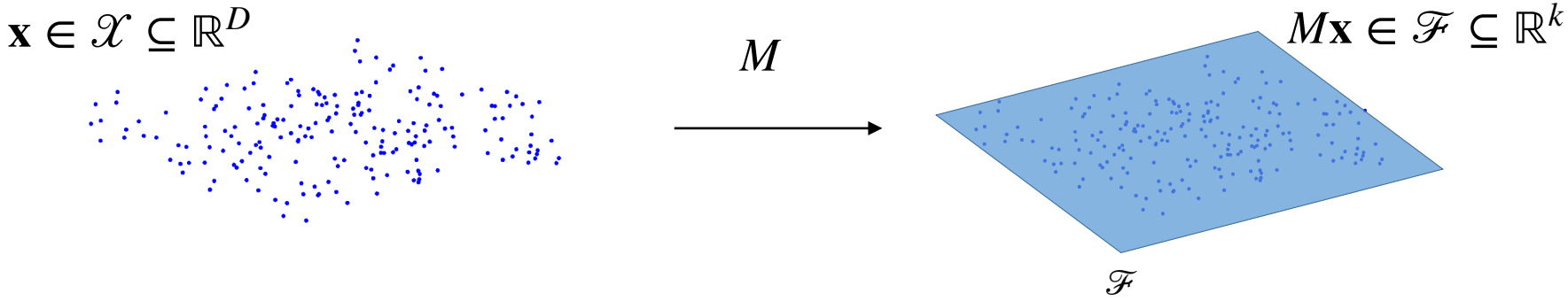
$$M : \mathbb{R}^D \rightarrow \mathbb{R}^k, \;\; k \ll D$$

Principal Component Analysis - PCA

Εξαγωγή χαρακτηριστικών - Μάθηση αναπαραστάσεων

- Η ανάλυση κύριων συνιστωσών (principal component analysis - PCA) είναι πιθανότατα η πιο διαδεδομένη μέθοδος μείωσης διαστάσεων δεδομένων.

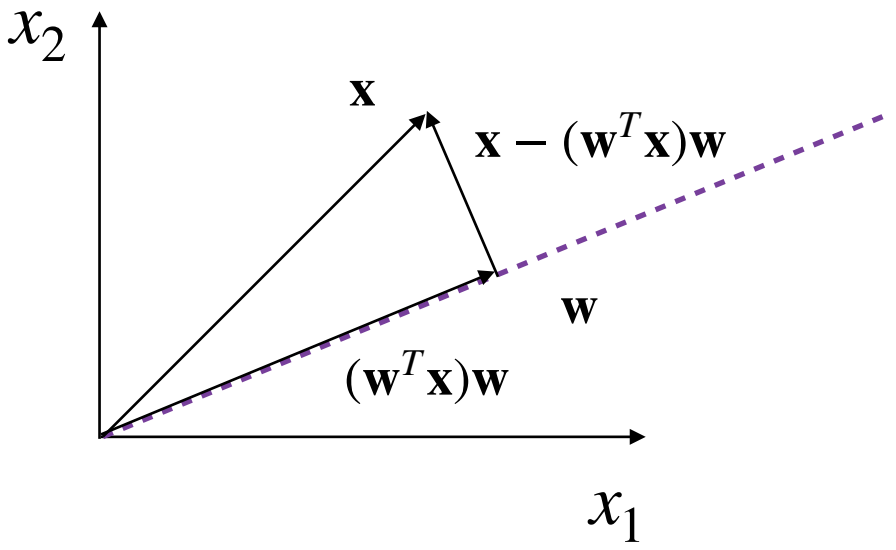
- Η PCA αποτελεί μη επιτηρούμενη μέθοδο (unsupervised) μείωσης διαστάσεων η οποία αξιοποιώντας την υποκείμενη δομή των δεδομένων εξάγει μια γραμμική απεικόνιση από τον D-διάστατο χώρο σε ένα (υπο) χώρο k διαστάσεων. Οι k διαστάσεις του χώρου μειωμένης διάστασης αποτελούν τις κύριες συνιστώσες των δεδομένων και ορίζουν ένα υπερεπίπεδο F.



Προβολή σε υποχώρους

Εάν $\mathbf{w}, \mathbf{x} \in \mathbb{R}^D$, $\|\mathbf{w}\|_2 = 1$ τότε η ορθογώνια προβολή του **x** στο **w** είναι:

$$(\mathbf{w}^T\mathbf{x})\mathbf{w} = \underbrace{\mathbf{w}\mathbf{w}^T}_{\mathbf{P}}\mathbf{x}$$



Ο τετραγωνικός πίνακας **P** καλείται πίνακας προβολής και προβάλει τα δεδομένα D-διαστάσεων στον υποχώρο μίας διάστασης που ορίζεται από το **w**.

Εύρεση κύριας συνιστώσας: ελαχιστοποίηση σφάλματος ανακατασκευής

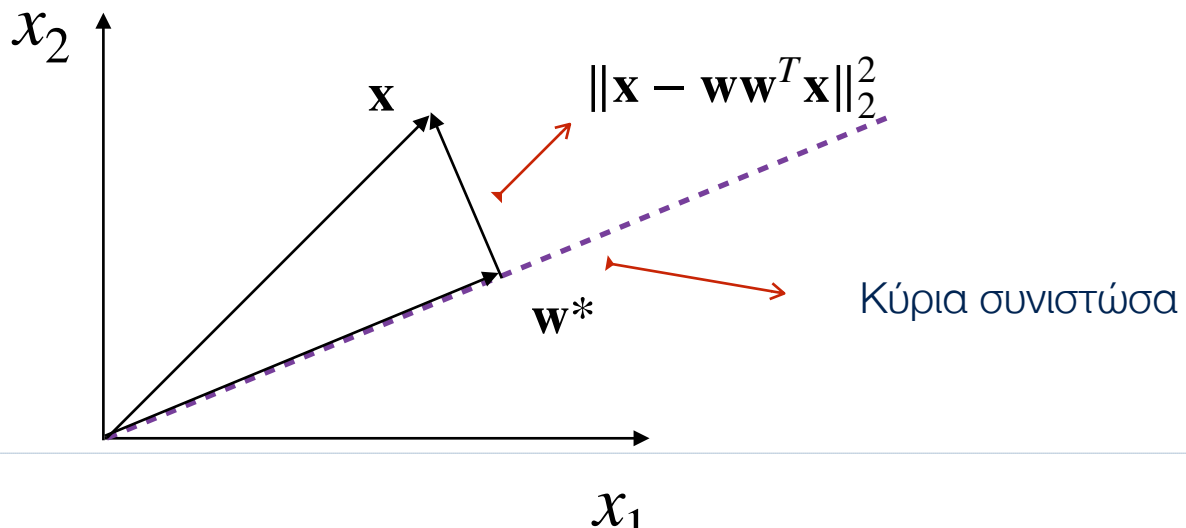
- Ας ξεκινήσουμε με το πρόβλημα εύρεσης της κύριας συνιστώσας (principal component), δηλαδή k=1, έχωντας ώς κατευθυντήρια αρχή την ανακατασκευή των δεδομένων (data reconstruction).

- Πρόβλημα: Προσδιόρισε ένα διάνυσμα **w** του οποίου η κατεύθυνση επιτρέπει τη βέλτιστη ανακατασκευή του **x**.

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \|\mathbf{x} - \mathbf{w}\mathbf{w}^T\mathbf{x}\|_2^2 \text{ subject to } \|\mathbf{w}\|_2 = 1$$

- Ο περιορισμός μοναδιαίου μήκους που επιβάλουμε στο **w** είναι απαραίτητος για να αποφύγουμε την τετριμμένη λύση.

- Το σφάλμα ανακατασκευής ποσοτικοποιεί τη πληροφορία που χάνουμε εάν προβάλουμε τα δεδομένα **x** στο **w**.



Εύρεση κύριας συνιστώσας: ελαχιστοποίηση σφάλματος ανακατασκευής

- Μπορούμε να ερμηνεύσουμε την ελαχιστοποίηση του τετραγωνικού σφάλματος υπό το πρίσμα της κωδικοποίησης - αποκωδικοποίησης.

- Κωδικοποίηση: $\mathbf{z} = \mathbf{w}^T\mathbf{x}$

- Αποκωδικοποίηση: $\tilde{\mathbf{x}} = \mathbf{w}\mathbf{z}$

- Επιθυμούμε το σφάλμα ανακατασκευής να είναι μικρό $\|\mathbf{x} - \tilde{\mathbf{x}}\|_2^2$

- Για να βρούμε τη κύρια συνιστώσα ενός συνόλου N δεδομένων, αρκεί να ελαχιστοποιούμε το αναμενόμενο σφάλμα ανακατασκευής.

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} E[\|\mathbf{x} - \mathbf{w}\mathbf{w}^T\mathbf{x}\|_2^2] \text{ subject to } \|\mathbf{w}\|_2 = 1$$

- Στη πράξη ελαχιστοποιούμε το εμπειρικό σφάλμα ανακατασκευής.

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\|_2^2 \text{ subject to } \|\mathbf{w}\|_2 = 1$$

Εύρεση κύριας συνιστώσας: μεγιστοποίηση διακύμανσης

- Μια διαφορετική προσέγγιση στην εύρεση της κύριας συνιστώσας είναι να βρούμε τη προβολή **w** η οποία συλλαμβάνει (εξηγεί) τη διακύμανση των δεδομένων.

- Ας υποθέσουμε ότι τα δεδομένα έχουν μηδενική μέση τιμή.

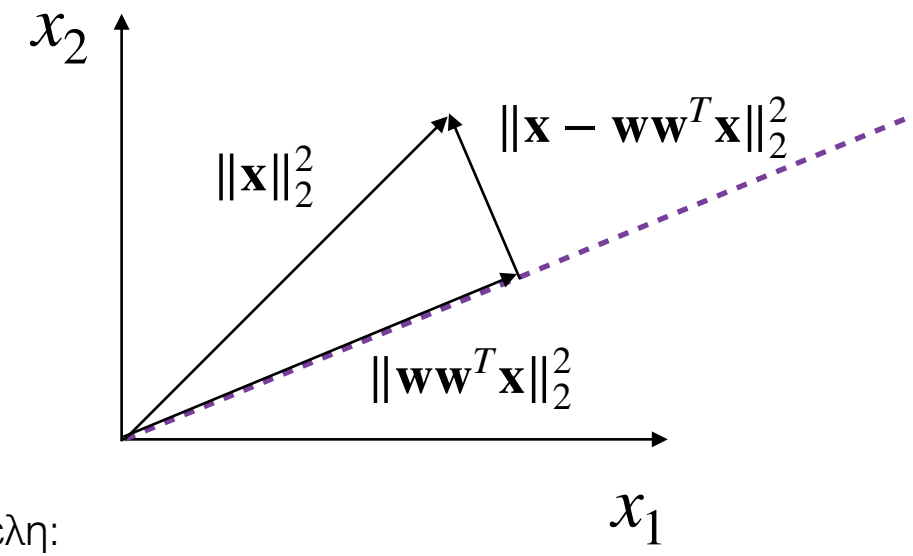
- Πυθαγόρειο θεώρημα

$$\|\mathbf{x}\|_2^2 = \|\mathbf{w}\mathbf{w}^T\mathbf{x}\|_2^2 + \|\mathbf{x} - \mathbf{w}\mathbf{w}^T\mathbf{x}\|_2^2$$

$$=$$

$$\|\mathbf{x}\|_2^2 = \|\mathbf{w}^T\mathbf{x}\|_2^2 + \|\mathbf{x} - \mathbf{w}\mathbf{w}^T\mathbf{x}\|_2^2$$

Εφόσον το **w** έχει μοναδιαίο μήκος



- Παίρνοντας την αναμενόμενη τιμή κατά μέλη:

$$E[\|\mathbf{x}\|_2^2] = E[\|\mathbf{w}^T\mathbf{x}\|_2^2] + E[\|\mathbf{x} - \mathbf{w}\mathbf{w}^T\mathbf{x}\|_2^2]$$

Διακύμανση των δεδομένων = διακύμανση προβολής + σφάλμα ανακατασκευής

σταθερή επιθυμούμε να είναι μεγάλη επιθυμούμε να είναι μικρό

Εύρεση κύριας συνιστώσας: μεγιστοποίηση διακύμανσης

- Αντικειμενική συνάρτηση: Μεγιστοποίηση της διακύμανσης των προβαλλόμενων δεδομένων.

$$\mathbf{w}^* = \arg\max_{\|\mathbf{w}\|_2=1} E[\|\mathbf{w}^T\mathbf{x}\|_2^2] = \arg\max_{\|\mathbf{w}\|_2=1} \frac{1}{N}\sum_{n=1}^{N} \|\mathbf{w}^T\mathbf{x}_n\|_2^2$$

$$= \arg\max_{\|\mathbf{w}\|_2=1} \frac{1}{N}\|\mathbf{w}^T\mathbf{X}\|_2^2$$

Πίνακας δεδομένων

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{D\times N}$$

$$= \arg\max_{\|\mathbf{w}\|_2=1} \mathbf{w}^T\left(\frac{1}{N}\mathbf{X}\mathbf{X}^T\right)\mathbf{w}$$

Πίνακας συνδιακύμανσης

$$\mathbf{C} = \frac{1}{N}\mathbf{X}\mathbf{X}^T$$

Πηλίκο Rayleigh

$$= \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T\mathbf{C}\mathbf{w}}{\mathbf{w}^T\mathbf{w}}$$

$$= \arg\max_{\mathbf{w}} \mathbf{w}^T\mathbf{C}\mathbf{w} - \lambda\mathbf{w}^T\mathbf{w}$$

Πίνακας συνδιακύμανσης

- Παράδειγμα: Δεδομένα δύο διαστάσεων με μέση τιμή μηδέν

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{2\times N} \qquad \mu = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n = \frac{1}{N}\mathbf{X}\mathbf{1}_{N\times 1} = \mathbf{0} \in \mathbb{R}^2$$

$$\mathbf{C} = \mathbf{X}\mathbf{X}^T = \begin{pmatrix} x_1^2 & x_1x_2 \\ x_1x_2 & x_2^2 \end{pmatrix} = \begin{pmatrix} \text{var}(x_1) & \text{cov}(x_1, x_2) \\ \text{cov}(x_2, x_1) & \text{var}(x_2) \end{pmatrix}$$

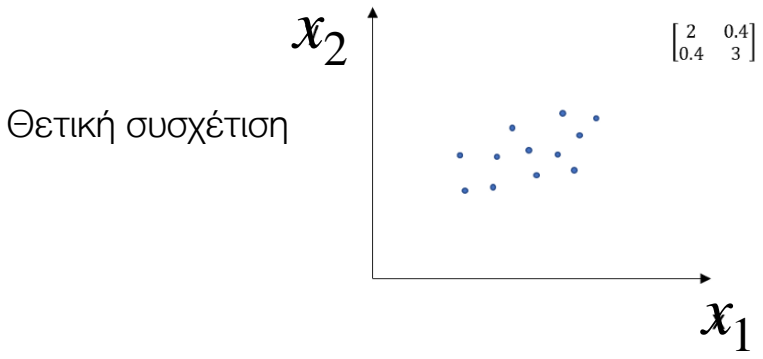
- Τα στοιχεία της κύρια διαγώνιου του πίνακα συνδιακύμανσης αποτελούν τη διακύμανση των δεδομένων.

- Τα στοιχεία εκτός της κύρια διαγώνιου του πίνακα συνδιακύμανσης αποτελούν τη συνδιακύμανση των δεδομένων.

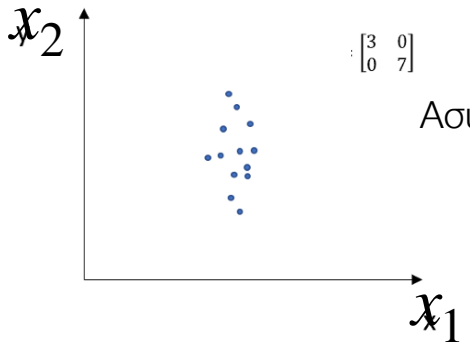
- Ο πίνακας συνδιακύμανσης είναι συμμετρικός:

$$\mathbf{C} = \mathbf{C}^T$$

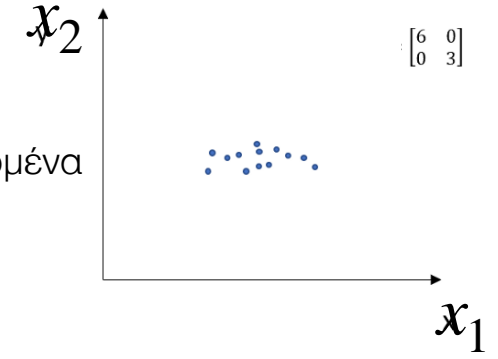
Πίνακας συνδιακύμανσης



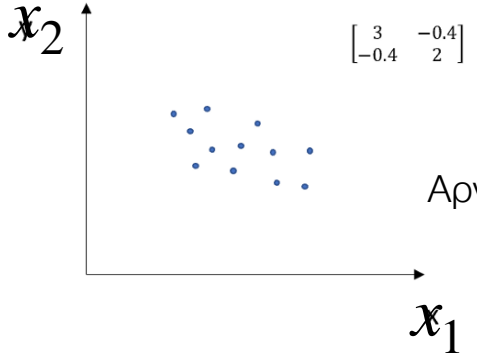
Θετική συσχέτιση



Ασυσχέτιστα δεδομένα



Ασυσχέτιστα δεδομένα



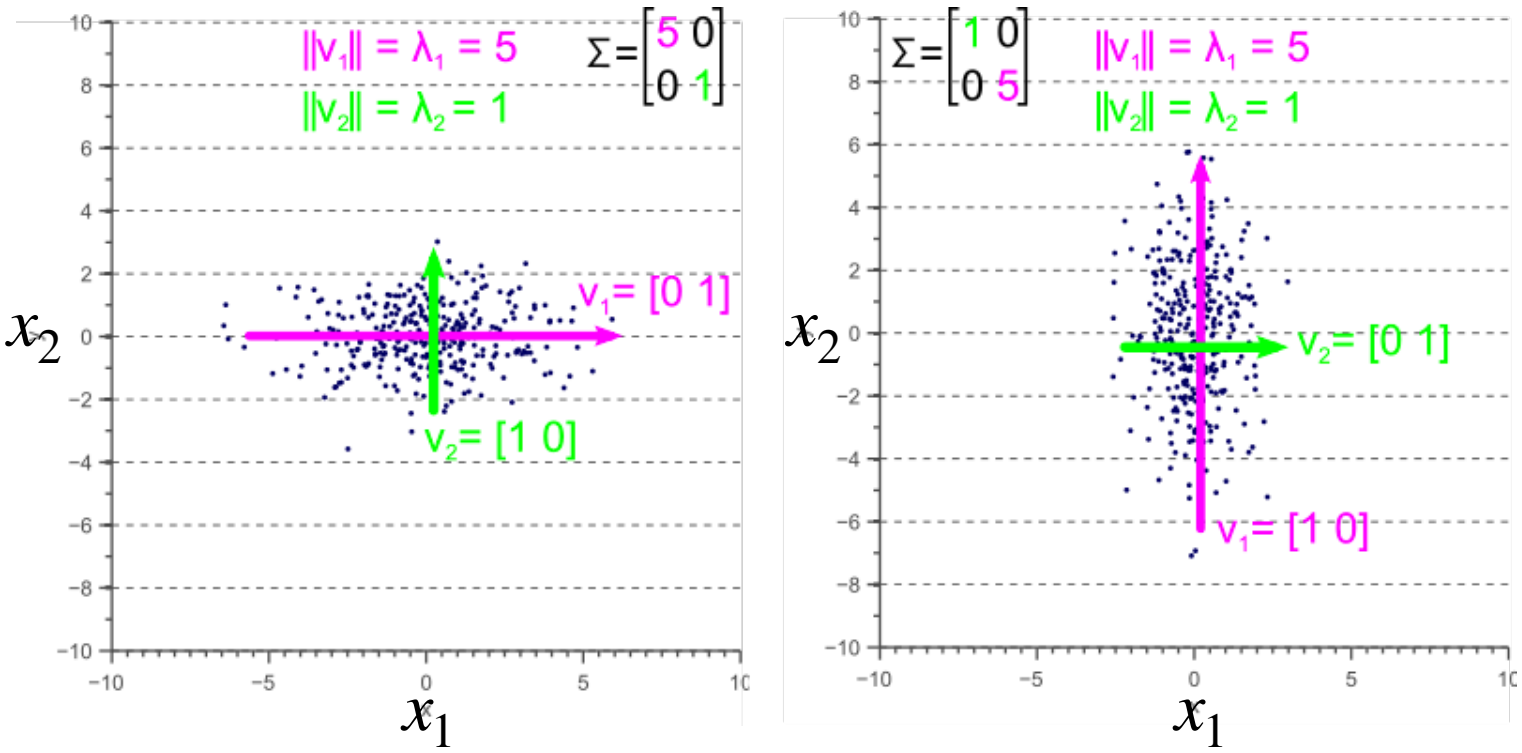
Αρνητική συσχέτιση

- Όπως μπορείτε να δείτε, ο πίνακας συνδιακύμανσης περιγράφει τόσο την διασπορά (διακύμανση) όσο και τον προσανατολισμό (συνδιακύμανση) των δεδομένων μας.

Ιδιοδιανύσματα και ιδιοτιμές του πίνακα συνδιακύμανσης

- Στο πίνακα συνδιακύμανσης (όπως και σε κάθε τετραγωνικό πίνακα), αντιστοιχούν ιδιοδιάνυσματα (eigenvectors) και ιδιοτιμές (eigenvalues) τέτοια ώστε να ισχύει η παρακάτω εξίσωση:

$$\mathbf{C}\mathbf{v}_i = \lambda_i \mathbf{v}_i$$



Εύρεση κύριας συνιστώσας: μεγιστοποίηση διακύμανσης

- Για να βρούμε την κύρια συνιστώσα των δεδομένων αρκεί να λύσουμε το παρακάτω πρόβλημα μεγιστοποίησης:

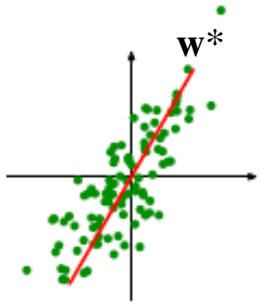
$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \ell(\mathbf{w}) = \mathbf{w}^T\mathbf{C}\mathbf{w} - \lambda\mathbf{w}^T\mathbf{w}$$

- Εφόσον η αντικειμενική συνάρτηση είναι παραγωγίσιμη για να βρούμε το μέγιστο της αρκεί να υπολογίσουμε τη παράγωγο και να τη θέσουμε ίση με το μηδέν.

$$\nabla_{\mathbf{w}}\ell(\mathbf{w}) = \mathbf{C}\mathbf{w} - \lambda\mathbf{w} = \mathbf{0}$$

- Συνεπώς προκύπτει ότι η κύρια συνιστώσα είναι το ιδιοδιάνυσμα που αντιστοιχεί στη μεγαλύτερη ιδιοτιμή του πίνακα συνδιακύμανσης.

$$\mathbf{C}\mathbf{w}^* = \lambda_{max}\mathbf{w}^*$$



- Οι ιδιοτιμές και τα ιδιοδιανύσματα πινάκων υπολογίζονται στην πράξη μέσω της ρουτίνας **eig()**.

Ισοδυναμία ελαχιστοποίησης σφάλματος - μεγιστοποίησης διακύμανσης

- Θα δείξουμε ότι η μεγιστοποίησης της διακύμανσης κατα μήκος της κύριας συνιστώσας είναι ισοδύναμη με την ελαχιστοποίηση του σφάλματος ανακατασκευής.

$$\text{Σφάλμα ανακατασκευής} = \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\|_2^2$$

$$= \sum_{n=1}^{N} (\|\mathbf{x}_n\|_2^2 - \|\mathbf{w}^T\mathbf{x}_n\|_2^2)$$

$$= \text{σταθερός όρος} - \sum_{n=1}^{N} \|\mathbf{w}^T\mathbf{x}_n\|_2^2$$

$$= \text{σταθερός όρος} - \|\mathbf{w}^T\mathbf{X}\|_2^2$$

- Συνεπώς, $\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\|_2^2$ subject to $\|\mathbf{w}\|_2 = 1 \Leftrightarrow$

$$= \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T \left(\frac{1}{N}\mathbf{X}\mathbf{X}^T\right) \mathbf{w}}{\mathbf{w}^T\mathbf{w}}$$

Εύρεση k κύριων συνιστωσών

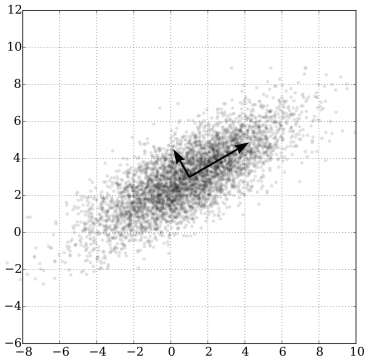
- Η παραπάνω ιδέα γενικεύεται εύκολα στη περίπτωση περισσότερων της μίας κύριων συνιστωσών (k>1).

- Για να βρούμε k ορθογώνιες μεταξύ τους κύριες συνιστώσες αρκεί να ελαχιστοποιήσουμε το σφάλμα ανακατασκευής υπό τον περιορισμό της ορθογωνιότητας.

Πίνακας συνιστωσών: $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k] \in \mathbb{R}^{D \times k}$

Ορθογωνιότητα: $\mathbf{W}^T\mathbf{W} = \mathbf{I} \in \mathbb{R}^{k \times k}$

Frobenius νορμα: $\|\mathbf{W}\|_F^2 = \sum_i \sum_j w_{ij}^2 = \sum_j \|\mathbf{w}_j\|_2^2$



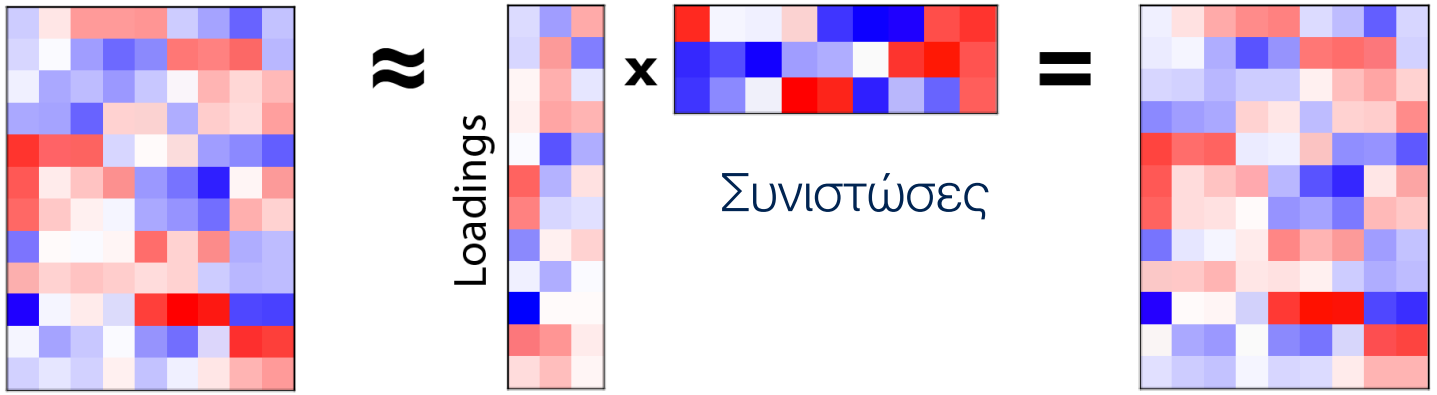
- Συνεπώς, λαμβάνοντας υπόψιν και τον ορισμό της Frobenius νάρμας, το σύνολο των k κύριών συνιστοστών προκύπτει ως λύση του προβλήματος ελαχιστοποίησης:

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\|_2^2 = \|\mathbf{X} - \mathbf{W}\mathbf{W}^T\mathbf{X}\|_F^2 \text{ subject to } \mathbf{W}^T\mathbf{W} = \mathbf{I}$$

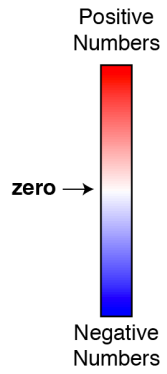
Λύση: Ο W έχει ως στήλες k ιδιοδιανύσματα που αντιστοιχούν στις k μεγαλύτερες ιδιοτιμές του πίνακα συνδιακύμανσης και αποτελούν μια ορθοκανονική βάση.

Ερμηνεία k κύριων συνιστωσών

$\mathbf{X} \in \mathbb{R}^{D \times N}$ $\mathbf{W} \in \mathbb{R}^{D \times k}$ $\mathbf{C} = \mathbf{W}^T\mathbf{X} \in \mathbb{R}^{k \times N}$ $\mathbf{W}\mathbf{W}^T\mathbf{X} \in \mathbb{R}^{D \times N}$



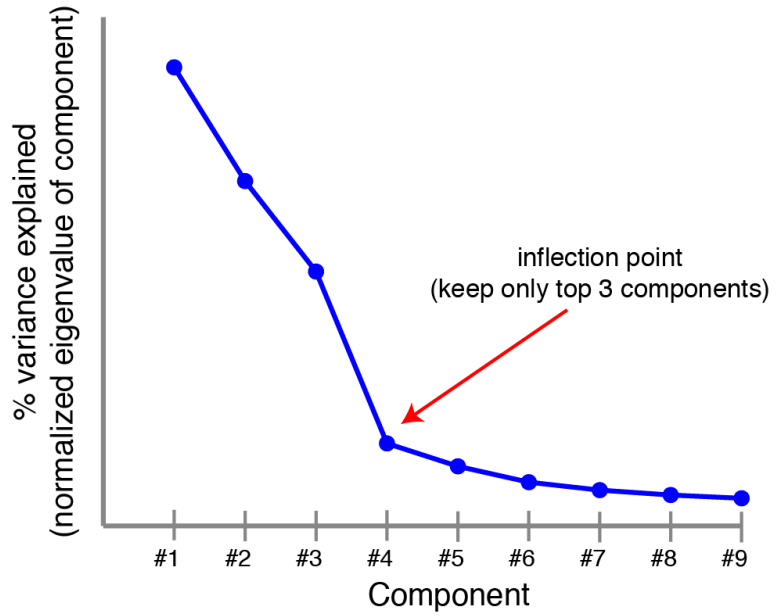
Δεδομένα Διανύσματα Βάσης Ανακατασκευή



Πώς επιλέγουμε το πλήθος των συνιστωσών;

- Οι ιδιοτιμές του πίνακα συνδιακύμανσης αντιστοιχούν στο ποσοστό της διακύμανσης των δεδομένων που περιγράφει κάθε συνιστώσα.

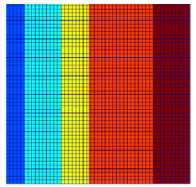
- Στη πράξη επιλέγουμε το πλήθος των συνιστωσών k τέτοιο ώστε να συλαμβανει το 85-90% της συνολικής διακύμανσης στα δεδομένα.



Πίνακες χαμηλού βαθμού και SVD

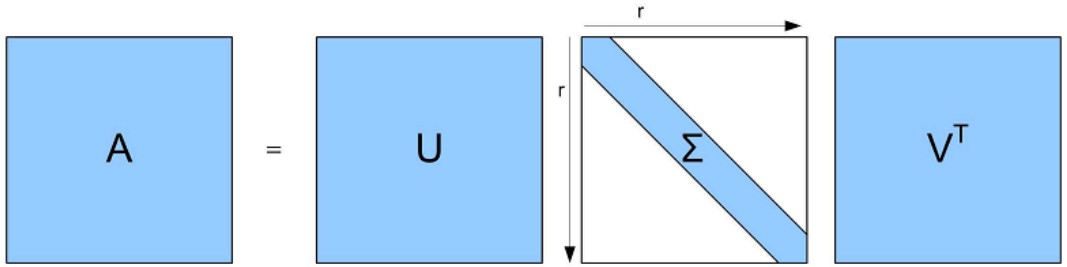
- Ορισμός: Ως βαθμός (rank) ενός πίνακα **A** ορίζεται ο μέγιστος αριθμός των γραμμικά ανεξάρτητων στηλών (γραμμών) του **A**.

Βαθμός: 5

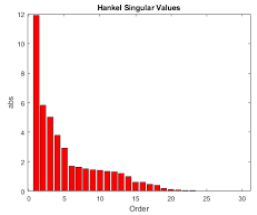


Χαμηλός βαθμός λόγω συμμετρίας του προσώπου

- SVD: Singular Value Decomposition (ανάλυση ιδιαζουσών τιμών)



$$\mathbf{U}^T\mathbf{U} = \mathbf{I} \in \mathbb{R}^{r\times r}$$



$$\mathbf{V}^T\mathbf{V} = \mathbf{I} \in \mathbb{R}^{r\times r}$$

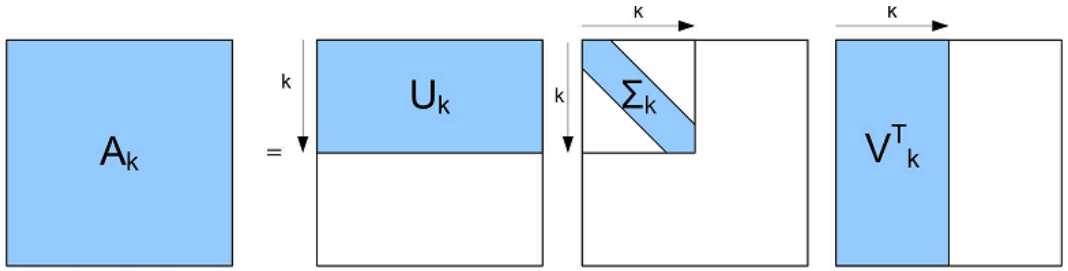
- Το πλήθος των μη-μηδενικών ιδιαζουσών τιμών ισούται με το βαθμό του πίνακα **A**.

PCA και προσέγγιση χαμηλού βαθμού

Πώς βρίσκουμε μια χαμηλού βαθμού (k) προσέγγιση (low-rank approximation) **A**k ενός πίνακα **A**;

$$\mathbf{A}_k^* = \arg\min_{\mathbf{A}_k} \|\mathbf{A} - \mathbf{A}_k\|_2^2 \text{ s.t. } \text{rank}(\mathbf{A}_k) \leq k$$

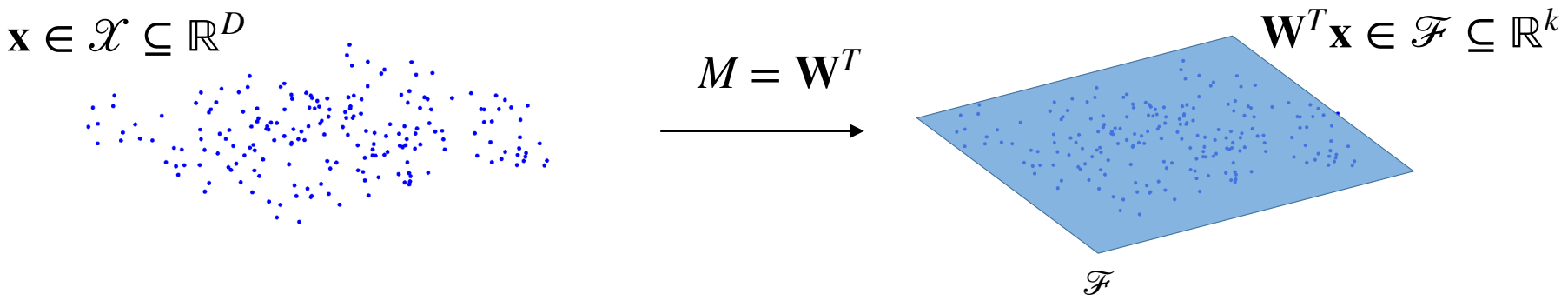
- Λύση: tSVD: Truncated Singular Value Decomposition



$$\mathbf{A}_k^* = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T = \mathbf{U}_k\mathbf{U}_k^T\mathbf{A}$$

- Στη περίπτωση της PCA ο πίνακας $\mathbf{W}\mathbf{W}^T\mathbf{X}$ έχει βαθμό μικρότερο ή ίσο του k. Κατα συνέπεια, η PCA είναι ισοδύναμη με τη χαμηλού βαθμού προσέγγιση του πίνακα δεδομένων, και μπορεί να υπολογιστεί μέσω της SVD.

Εξαγωγή χαμηλόβαθμων αναπαραστάσεων



- Για να προσδιορίσουμε την απεικόνιση M που προβάλει τα δεδομένα D διαστάσεων στο γραμμικό υποχώρο k διαστάσεων επιλύουμε:

$$\min_{\mathbf{W}} \|\mathbf{X} - \mathbf{W}\mathbf{W}^T\mathbf{X}\|_F^2 \text{ s.t. } \mathbf{W}^T\mathbf{W} = \mathbf{I}$$ Μέσω ιδιοανάλυσης

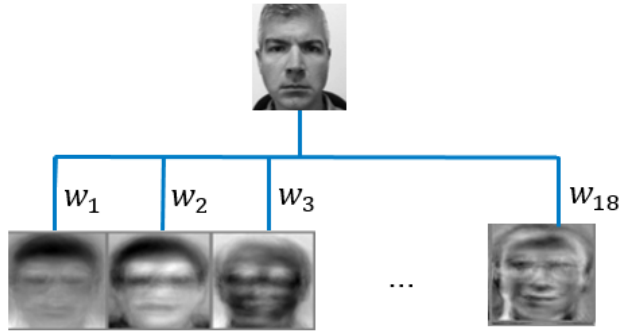
$$\min_{\mathbf{W}} \|\mathbf{X} - \mathbf{W}\|_F^2 \text{ s.t. } \text{rank}(\mathbf{W}) \leq k$$ Μέσω SVD

Χαμηλόβαθμη αναπαράσταση: $\mathbf{W}^T\mathbf{x} \in \mathcal{F}, \quad \mathbf{W} \in \mathbb{R}^k, k \ll D$

- Εφόσον εξάγουμε τις αναπαραστάσεις χαμηλών διαστάσεων, αυτές χρησιμοποιούνται στη θέση των αρχικών δεδομένων για ταξινόμηση, παλινδρόμηση, οπτικοποίηση δεδομένων κτλ.

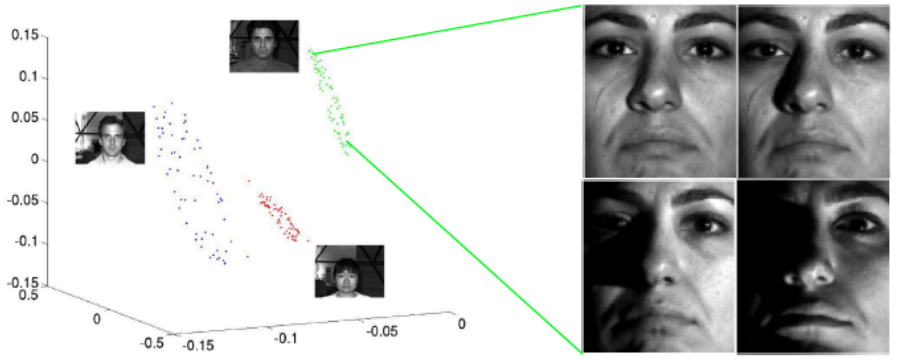
Εφαρμογές της PCA στη μοντελοποίηση προσώπου

- Ιδιοπρόσωπα - Eigenfaces [Turk and Petland, CVPR 1991]:



- Εικόνες του ιδίου προσώπου υπό διαφορετικές συνθήκες φωτισμού ζουν σε ένα γραμμικό χώρο εννέα (9) διαστάσεων

[Barsi and Jacobs, IEEE T-PAMI, 2003]



Γενικεύσεις της PCA

- Στο αρχικό μοντέλο της PCA θεωρήσαμε ότι το μοντέλο δεδομένων είναι:

$$\mathbf{X} \approx \mathbf{W}\mathbf{W}^T\mathbf{X}$$

- Ισοδύναμα, θέτοντας τη χαμηλόβαθμη αναπαράσταση ίση με $\mathbf{C} = \mathbf{W}^T\mathbf{X}$ το μοντέλο γράφεται ως εξής:

$$\mathbf{X} \approx \mathbf{W}\mathbf{C}, \mathbf{W} \in \mathbb{R}^{D\times k}, \mathbf{C} \in \mathbb{R}^{k\times N}$$

- Συνεπώς μπορούμε να σχεδιάσουμε γενικευμένα μοντέλα PCA, όπως για παράδειγμα κανονικοποιημένη PCA (regularized PCA) για να ενισχύσουμε την ικανότητα γενίκευσης:

$$\min_{W,C} \|\mathbf{X} - WC\|_F^2 + \gamma \sum_{i=1}^{D} \|\mathbf{w}_i\|_2^2 + \gamma \sum_{j=1}^{k} \|\mathbf{c}_j\|_2^2$$

- Λύση: Εναλλασσόμενη ελαχιστοποίηση

Alternating minimization:

1 choose initial starting points $W^{(0)}$ and $C^{(0)}$
2 $n \leftarrow 0$
3 **while** not converged
4 $W^{(n+1)} \leftarrow$ minimize over W while holding $C = C^{(n)}$ constant.
5 $C^{(n+1)} \leftarrow$ minimize over C while holding $W = W^{(n+1)}$ constant.
6 $n \leftarrow n + 1$
7 **end while**

Non-negative Matrix Factorization

Παραγοντοποίηση μη-αρνητικών πινάκων

- Η PCA και οι σχετικές με αυτή μέθοδοι αναπαριστούν τα δεδομένα ως γραμμικό συνδυασμό k διανυσμάτων βάσης τα οποία συλλαμβάνουν ολιστική, δύσκολα ερμηνεύσιμη πληροφορία για τα δεδομένα. Επιπλέον, δεν διατηρεί την μη-αρνητικότητα των τιμών των δεδομένων (π.χ. εικόνες, βίντεο, φασματικά δεδομένα).

Μπορούμε να σχεδιάσουμε μια μέθοδο μείωσης διαστάσεων που να διατηρεί τη μη- αρνητικότητα των δεδομένων και να εξάγει αραιές και εύκολα ερμηνεύσιμες συνιστώσες των δεδομένων;

- Απάντηση: Μη-αρνητική παραγοντοποίηση πινάκων (non-negative matrix factorization - NMF)

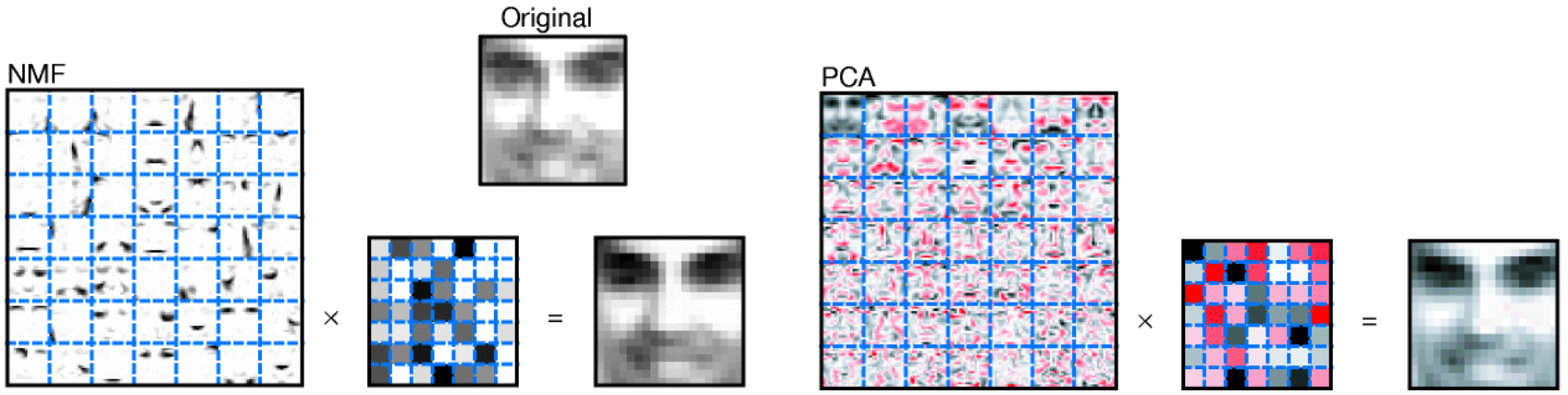
$$\mathbf{X} \approx \mathbf{WC}, \ \mathbf{X} \in \mathbb{R}_+^{D\times N}, \mathbf{W} \in \mathbb{R}_+^{D\times k}, \mathbf{C} \in \mathbb{R}_+^{k\times N}$$

NMF

$$\mathbf{X} \in \mathbb{R}_{+}^{D \times N} \qquad \mathbf{W} \in \mathbb{R}_{+}^{D \times k} \qquad \mathbf{C} = \in \mathbb{R}_{+}^{k \times N}$$

$$\begin{pmatrix} 10 & 12 & 8.5 & 15 & 0 & 31 & 20 & 9 \\ 1 & 3 & 8 & 14 & 28 & 14 & 5 & 0 \\ 2 & 6 & 3 & 2 & 4 & 15 & 10 & 0 \\ 10 & 0 & 9.5 & 29 & 8 & 7 & 0 & 15 \\ 2 & 0 & 1.5 & 5 & 0 & 1 & 0 & 3 \\ 1 & 3 & 4 & 6 & 12 & 10 & 5 & 0 \end{pmatrix} \approx \begin{pmatrix} 3 & 0 & 4 \\ 0 & 7 & 1 \\ 0 & 1 & 2 \\ 5 & 2 & 0 \\ 1 & 0 & 0 \\ 0 & 3 & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 & 1.5 & 5 & 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & 2 & 4 & 1 & 0 & 0 \\ 1 & 3 & 1 & 0 & 0 & 7 & 5 & 0 \end{pmatrix}$$

Δεδομένα Διανύσματα Βάσης Συνιστώσες



Διανύσματα Βάσης Συνιστώσες Διανύσματα Βάσης Συνιστώσες

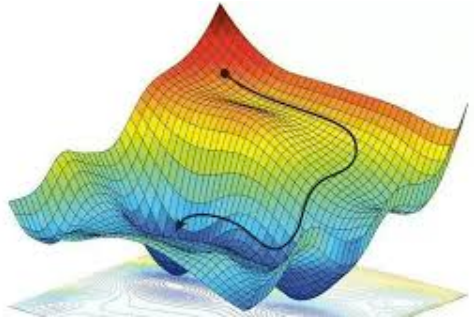
NMF: πρόβλημα βελτιστοποίησης

- Η μη-αρνητική παραγοντοποίηση πινάκων προκύπτει ως λύση του παρακάτω προβλήματος ελαχίστων τετραγώνων με περιορισμούς μη-αρνητικότητας

$$\min_{W,C} \mathcal{C}(W, C) = \|\mathbf{X} - WC\|_F^2 \quad \text{s.t.} \mathbf{W} \geq \mathbf{0}, \mathbf{C} \geq \mathbf{0}$$

- Το παραπάνω πρόβλημα βελτιστοποίησης δεν έχει λύση σε κλειστή μορφή, επομένως πρέπει να το λύσουμε επαναληπτικά.

- Σε κάθε βήμα του αλγορίθμου βελτιστοποίησης κρατάμε σταθερή τη μία μεταβλητή και τροποποιούμε την άλλη ανάλογα την κλήση της συνάρτησης (αρνητική κατεύθυνση της παραγώγου της συνάρτησης) στο τρέχον σημείο. Δηλαδή εφαρμόζουμε ένα βήμα καθόδου βασισμένο στη κλίση (gradient descent step).



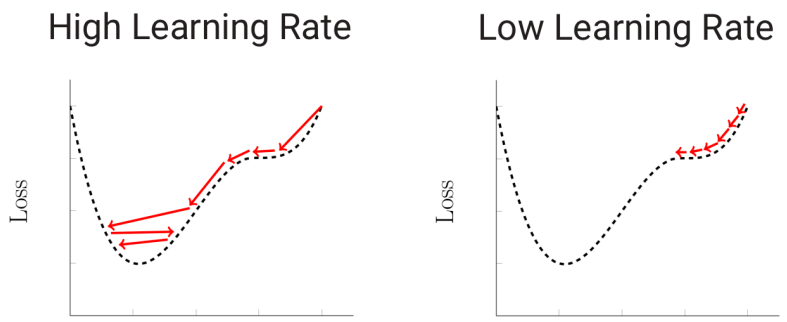
NMF: αλγόριθμος βελτιστοποίησης

- Αρχικοποιήση: Αρχικοποιούμε τις άγνωστες μεταβλητές W και C με τυχαίες μη-αρνητικές τιμές.

- Gradient descent steps

$$\mathbf{C}_{[t+1]} = \mathbf{C}_{[t]} - n_t \nabla_{C_{[t]}} \ell(W, C_{[t]})$$
$$\mathbf{W}_{[t+1]} = \mathbf{W}_{[t]} - n_t \nabla_{W_{[t]}} \ell(W_{[t]}, C)$$
$$t = t + 1$$

- Η παράμετρος n_t > 0 ονομάζεται learning rate ή step size και καθορίζει πόσο θα τροποποιηθεί η μεταβλήτή σε κάθε βήμα.



NMF: αλγόριθμος βελτιστοποίησης

Επιβάλουν αυτοί οι κανόνες επανάληψης την απαιτούμενη μη-αρνητικότητα των μεταβλητών;

- Απάντηση: Όχι, μπορούν να λάβουν και αρνητικές τιμές

- Παρατήρηση: Κάθε αριθμός μπορεί να εκφραστεί ως διαφορά μη αρνητικών αριθμών

$$\nabla_C \ell(W, C) = \nabla_C \ell(W, C)^+ - \nabla_C \ell(W, C)^- = \mathbf{W}^T\mathbf{W}\mathbf{C} - \mathbf{W}^T\mathbf{X}$$

- Εάν προσδιορίσουμε το step size κατα τρόπο που εξαρτάται από τα δεδομένα ως εξής:

$$n = \frac{\mathbf{C}}{\nabla_C \ell(W, C)^+} = \frac{\mathbf{C}}{\mathbf{W}^T\mathbf{W}\mathbf{C}}$$

- Τότε το gradient descent step για τη μεταβλητή **C** γράφεται ως εξής:

$$\mathbf{C} = \mathbf{C} - \frac{\mathbf{C}}{\mathbf{W}^T\mathbf{W}\mathbf{C}} * \nabla_C \ell(W, C) = \mathbf{C} * \frac{\mathbf{W}^T\mathbf{X}}{\mathbf{W}^T\mathbf{W}\mathbf{C}}$$

- Το * συμβολίζει το γινόμενο Hadamard (δηλ. γινόμενο στοιχείο προς στοιχείο του πίνακα)

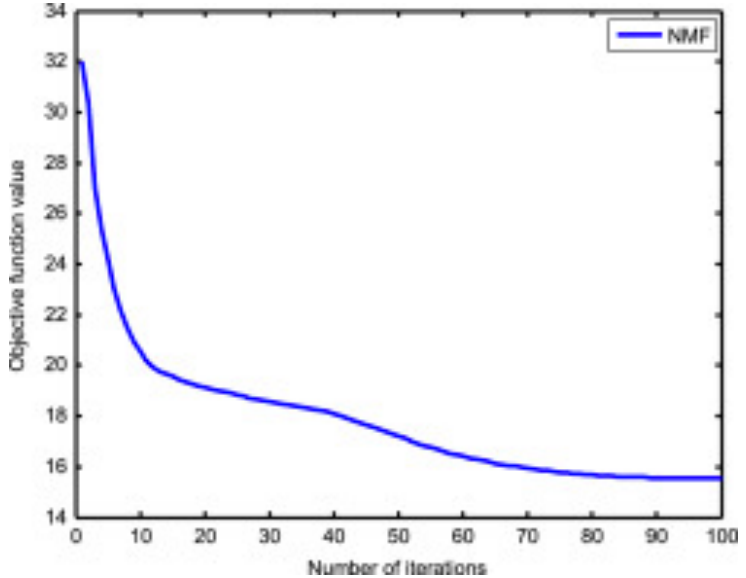
NMF: αλγόριθμος βελτιστοποίησης

- Ακολούθόντας αντίστοιχη λογική, το gradient descent step για τη μεταβλητή **W** γράφεται ως εξής:

$$\mathbf{W} = \mathbf{W} * \frac{\mathbf{X}\mathbf{C}^T}{\mathbf{W}\mathbf{C}\mathbf{C}^T}$$

- Η διάιρεση στα παραπάνω steps εκτελέιται στοιχείο προς στοιχείο του πίνακα.

- Είναι εύκολο να διαπιστώσουμε ότι ακολουθώντας αυτά τα βήματα προκύπτουν πάντα μη-αρνητικοί πίνακες.



Αναπαράσταση δεδομένων εκτός συνόλου εκπαίδευσης

- Οι μετασχηματισμοί **W** που προκύπτουν από τις μεθόδους μείωσεις διαστάσεων που συζητήσαμε (δηλ. PCA, NMF, LDA) μπορούν να χρησιμοποιηθούν για τη μείωση της διάστασης δεδομένων εκτός του συνόλου εκπαίδευσης.

- Έστω, ένα δεδομένο ελέγχου $\mathbf{x}_{test} \in \mathbf{R}^D$

- Η αναπαράσταση (ή αλλιώς τα χαρακτηριστικά) χαμηλής διάστασης προκύπτουν ως:

$$\hat{\mathbf{x}} = \mathbf{W}^T \mathbf{x}_{test} \in \mathbf{R}^k$$

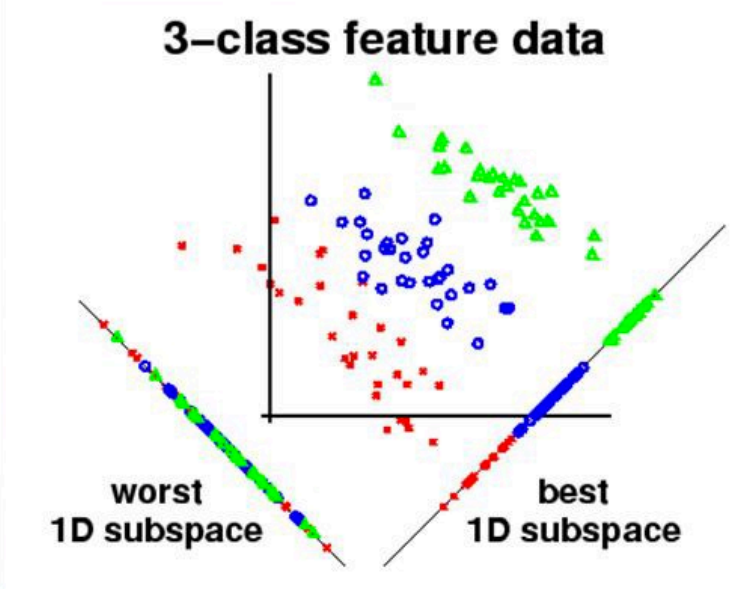
- Τα χαρακτηριστικά χρησιμοποιούνται στη συνέχεια για ταξινόμηση, παλινδρόμηση, ομαδοποίηση κτλ.

Γραμμική διακριτική ανάλυση - LDA

Διακριτικές αναπαραστάσεις

- Οι μέθοδοι μείωσης διάστασης και εξάγουν χαρακτηριστικά (αναπαραστάσεις) χωρίς να λαμβάνουν υπόψιν την πληροφορία κλάσης που πιθανόν συνοδεύουν τα δεδομένα. Συνεπώς, δεν εξάγουν χαρακτηριστικά που να είναι βέλτιστα για προβλήματα ταξινόμησης.

Μπορούμε να σχεδιάσουμε μια μέθοδο μείωσης διαστάσεων που να μεγιστοποιεί τον διαχωρισμό των δεδομένων ανάλογα με τη κλάση που ανήκουν και συνεπώς να εξάγει αναπαραστάσεις (χαρακτηριστικά) οι οποίες έχουν διακριτική ικανότητα;



Γραμμική διακριτική ανάλυση - LDA

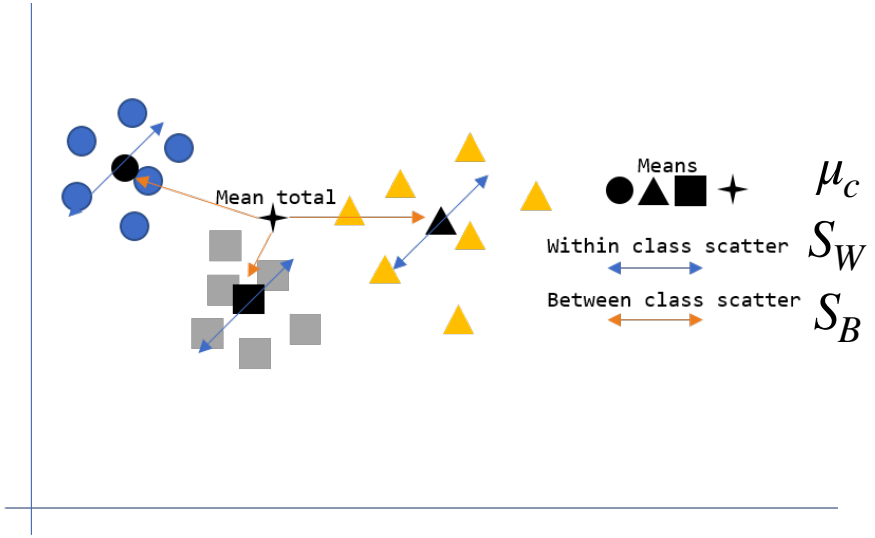
- Η γραμμική διακριτική ανάλυση (linear discriminant analysis - LDA ή Fisher discriminant analysis - FDA) στοχεύει στο να μεταχηματίσει τα δεδομένα, προβάλοντάς τα σε ένα χώρο μικρότερης διάστασης, έτσι ώστε ο διαχωρισμός των δεδομένων διαφορετικών κλάσεων να μεγιστοποιείται ενώ παράλληλα η διακύμανση των δεδομένων κάθε κλάσης ελαχιστοποιείται ώστε η πιθανότητα σύμπτωσης διαφορετικών κλάσεων να είναι μικρή.

- Πίνακας συνδιακύμανσης *μεταξύ κλάσεων:*

$$S_B = \sum_{classes\ \ c} N_c(\mu_c - \mu)(\mu_c - \mu)^T$$

- Πίνακας συνδιακύμανσης *εντός κλάσεων:*

$$S_W = \sum_{classes\ \ c} \sum_{j \in c} (x_j - \mu_c)(x_j - \mu_c)^T$$



LDA: πρόβλημα βελτιστοποίησης

- Η LDA βρίσκει το βέλτιστο μετασχηματισμό **W** διάστασης D x c-1 (το πολύ) ο οποίος να μεγιστοποιεί τη διακύμανση μεταξύ κλάσεων και να ελαχιστοποιεί την διακύμανση εντός κλάσεων των δεδομένων όταν αυτά προβάλλονται στο χώρο διάστασης c-1, λύνοντας το παρακάτω πρόβλημα βελτιστοποίησης:

$$\mathbf{W}* = \arg\max_{\mathbf{W}} \mathrm{tr}\left(\frac{\mathbf{W}^T S_B \mathbf{W}}{\mathbf{W}^T S_W \mathbf{W}}\right)$$

- Το πρόβλημα εκφράζεται ισοδύναμα ως εξής:

$$\mathbf{W}* = \arg\max_{\mathbf{W}} \ell(\mathbf{W}) = \mathrm{tr}(\mathbf{W}^T S_B \mathbf{W}) - \lambda \mathrm{tr}(\mathbf{W}^T S_W \mathbf{W})$$

- Για να βρούμε το μέγιστο, υπολογίζουμε τη παράγωγο της συνάρτησης ως προς το W και τη θέτουμε ίση με το μηδέν:

$$\nabla_W \ell(\mathbf{W}) = S_B \mathbf{W} - \lambda S_W \mathbf{W} = \mathbf{0}$$

$$S_W^{-1} S_B \mathbf{W} = \lambda \mathbf{W}$$

- Συνεπώς ο βέλτιστος μετασχηματισμός από τα k ιδιοδιανίσματα που αντιστοιχούν στις k μεγαλύτερες ιδιοτιμές του πίνακα: $S_W^{-1} S_B$