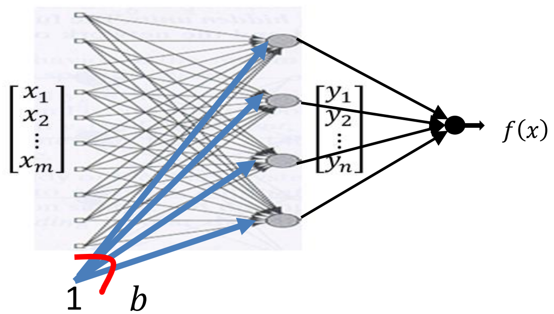
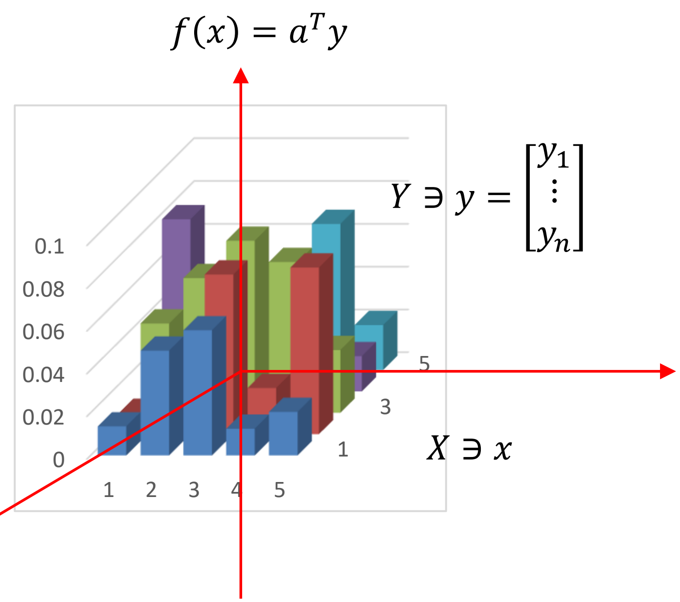
Nonlinear Mapping

$$T: X \to Y, \qquad y = \sigma(Wx + b), \qquad f(x) = a^T y$$





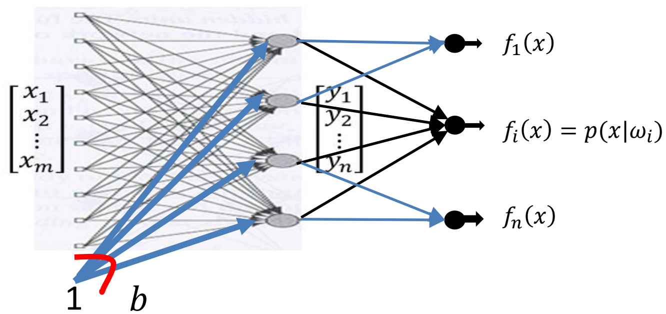
Universal Approximation Theorem Let ξ be a non-constant, bounded, and monotonically-increasing continuous activation function, $f : [0,1]^d \to \mathbb{R}$ continuous function, and $\epsilon > 0$. Then, $\exists n$ and parameters $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, $\mathbf{W} \in \mathbb{R}^{n \times d}$ s.t.

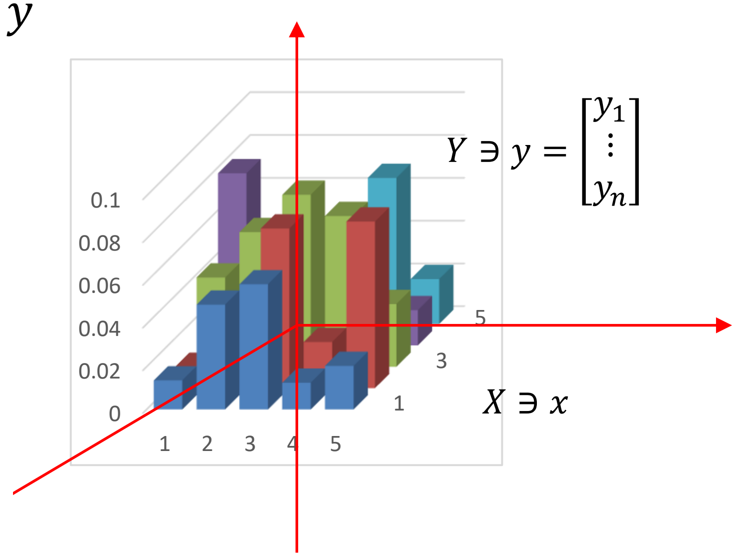
$$\left| \sum_{i=1}^{n} a_i \xi(\mathbf{w}_i^\top \mathbf{x} + b_i) - f(\mathbf{x}) \right| < \epsilon \qquad \forall \mathbf{x} \in [0,1]^d$$

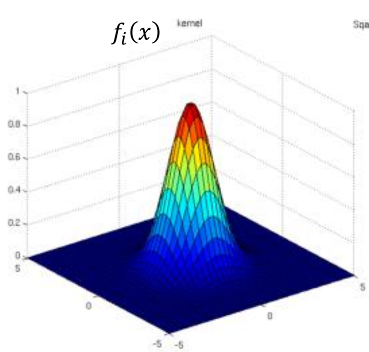
Geometric Deep Learning on graph and manifolds, Michael Bronstein, SIAM 2018, Imperial College London

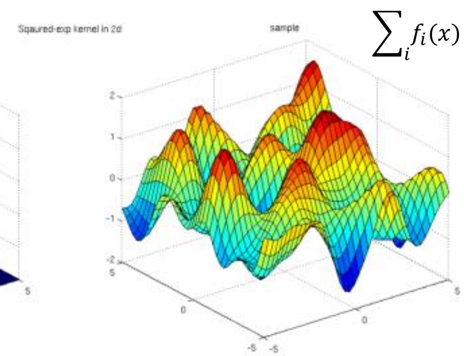
Nonlinear Mapping

$T: X \to Y, \qquad y = \sigma(Wx + b), \qquad f_i(x) = a_i^T y$



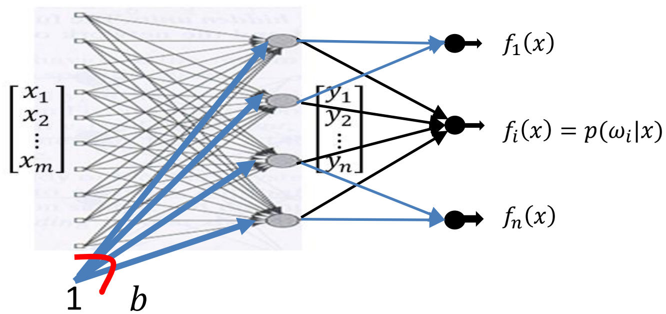


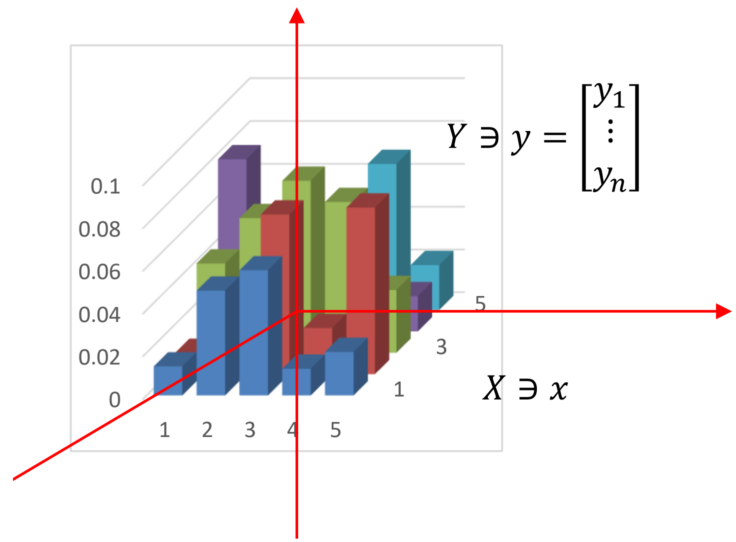


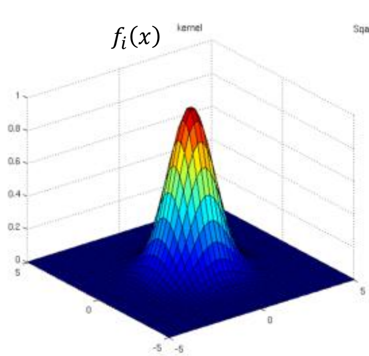


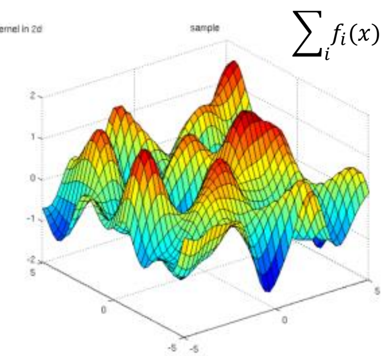
Nonlinear Mapping

$$T: X \to Y,\ \ y = \sigma(Wx + b),\ \ f_i(x) = {a_i^T y}\Big/{\sum_j a_j^T y}\ \text{(softmax)}$$









Feature Dimension Reduction: PCA & LDA (I)

Jin Young Choi

Seoul National University

Outline

Feature Extraction

Introduction of PCA & LDA

Principal Component Analysis (PCA)

Linear Discriminant Analysis (FLDA)

Multiple Discriminant Analysis (MDA)

Simple Enhancement of PCA/LDA

Feature Extraction

- Features

  Weight, Height, Width, Volume, Head size, …

  Edge, Shape, Geometric Relations …

  RGB Color for each pixel

  SIFT, SURF, HOG, …

- Feature Extraction from Raw Data

  Pixel Valued Vector is raw data vector

  Raw data vector is redundant

  The dimension should be reduced

Component Analysis and Discriminants

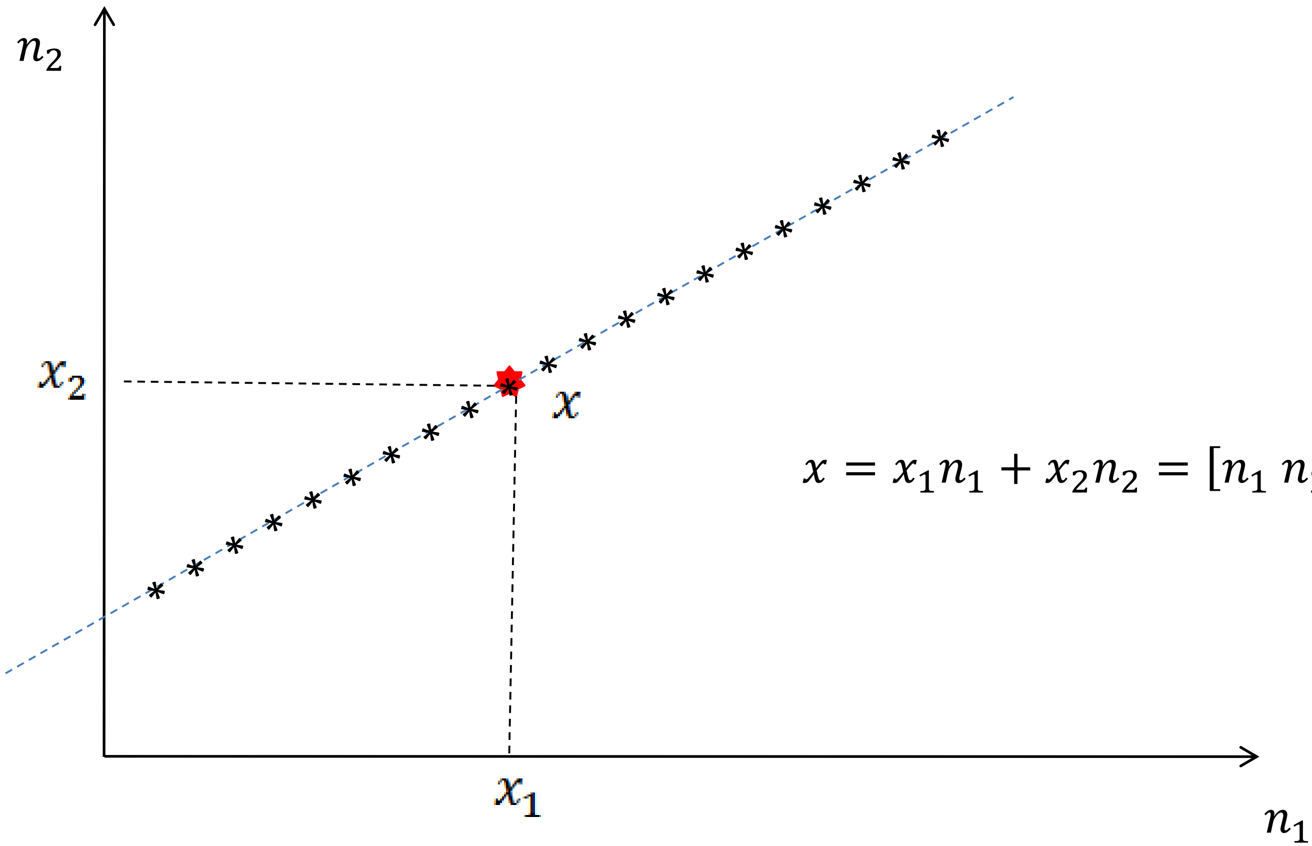
- How to reduce excessive dimensionality?
  - Answer: Combine features highly dependent to each other.

- Linear methods project high-dimensional data onto lower dimensional space.

- Principal Components Analysis (PCA)
  - seeks the projection which best represents the data in a least-square error sense.

- Linear Discriminant Analysis (LDA) or Fisher Linear Discriminant
  - seeks the projection that best separates the data in a least-square discrimination error sense.

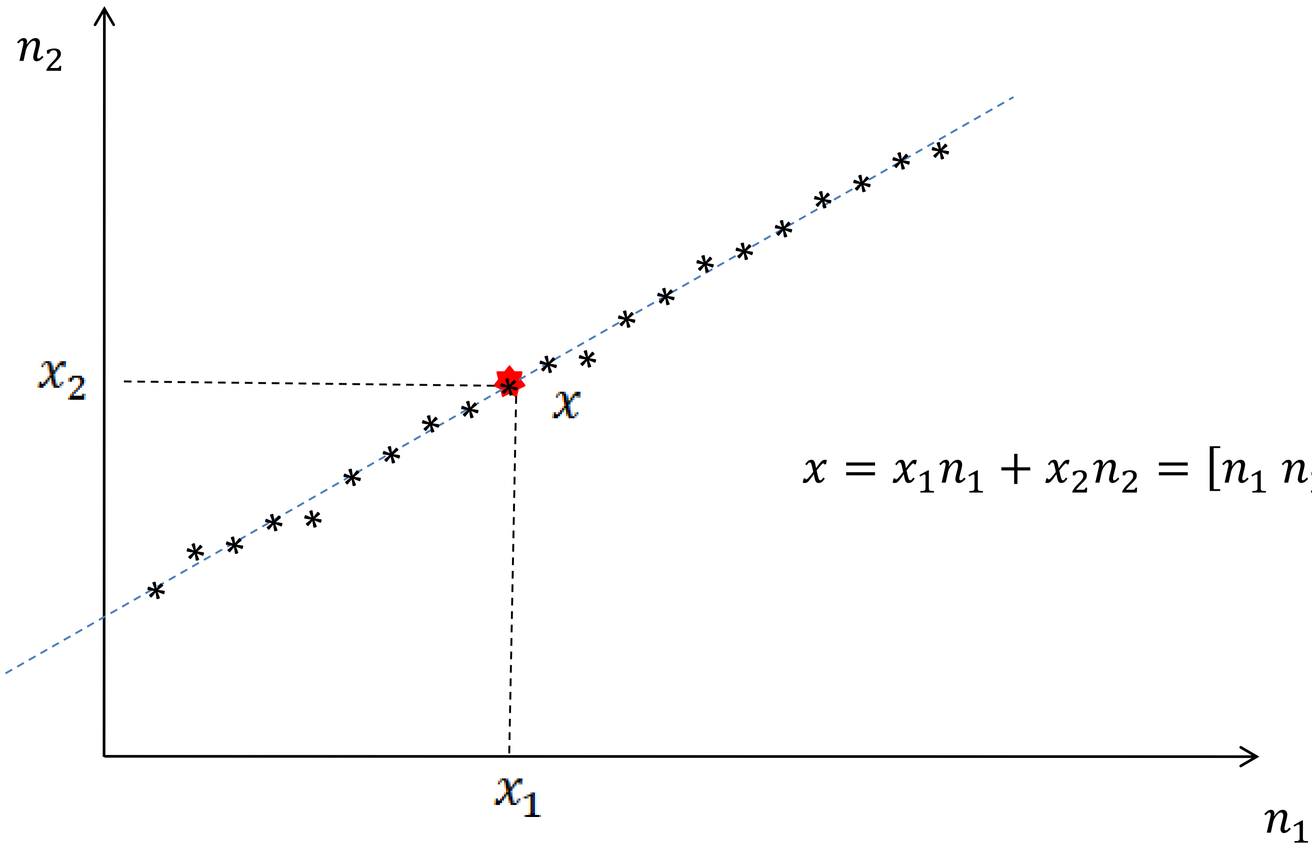
Principal Component Analysis



$$x = x_1 n_1 + x_2 n_2 = [n_1\ n_2]x = Nx$$

*

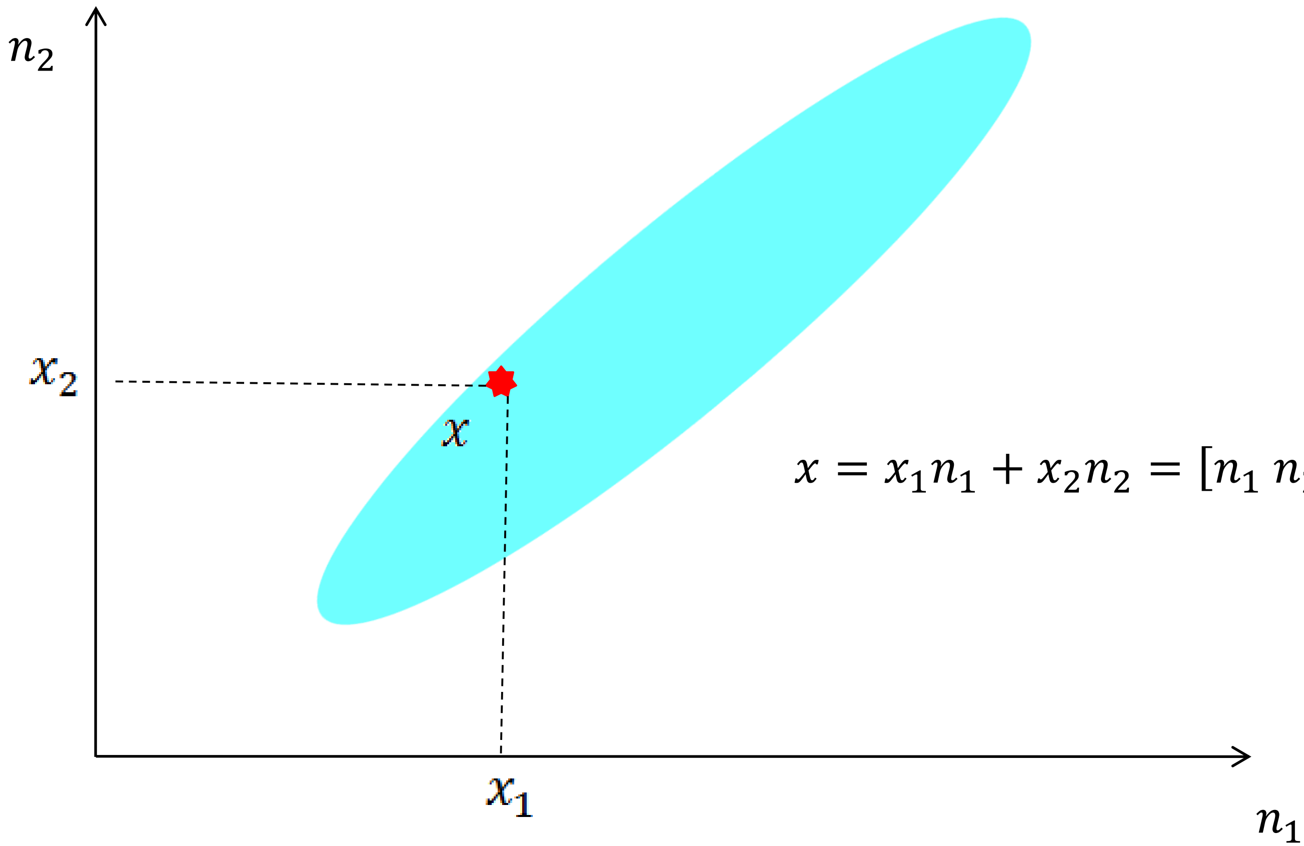
Principal Component Analysis



$$x = x_1 n_1 + x_2 n_2 = [n_1\ n_2]x = Nx$$

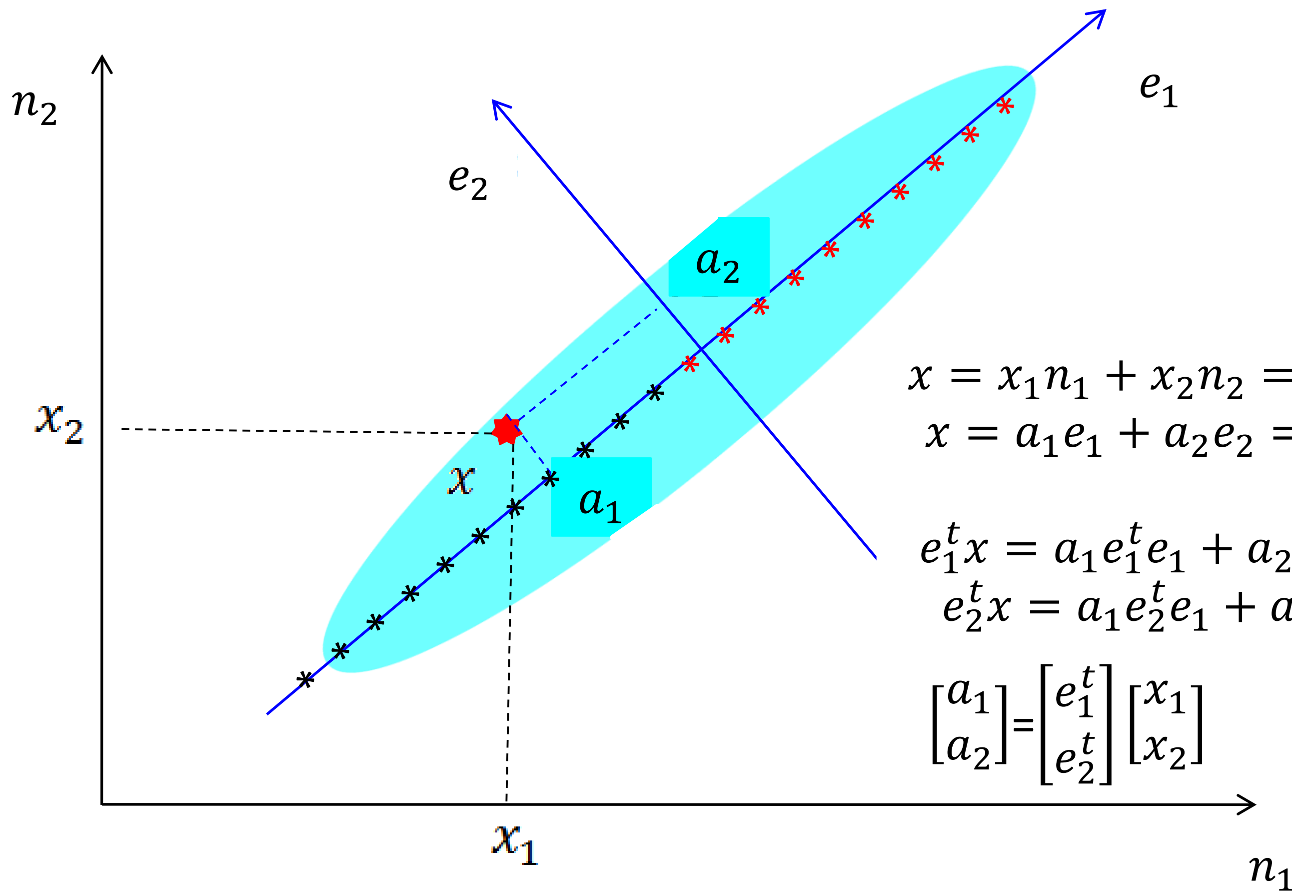
*

Principal Component Analysis



$$x = x_1 n_1 + x_2 n_2 = [n_1\ n_2]x = Nx$$

Principal Component Analysis

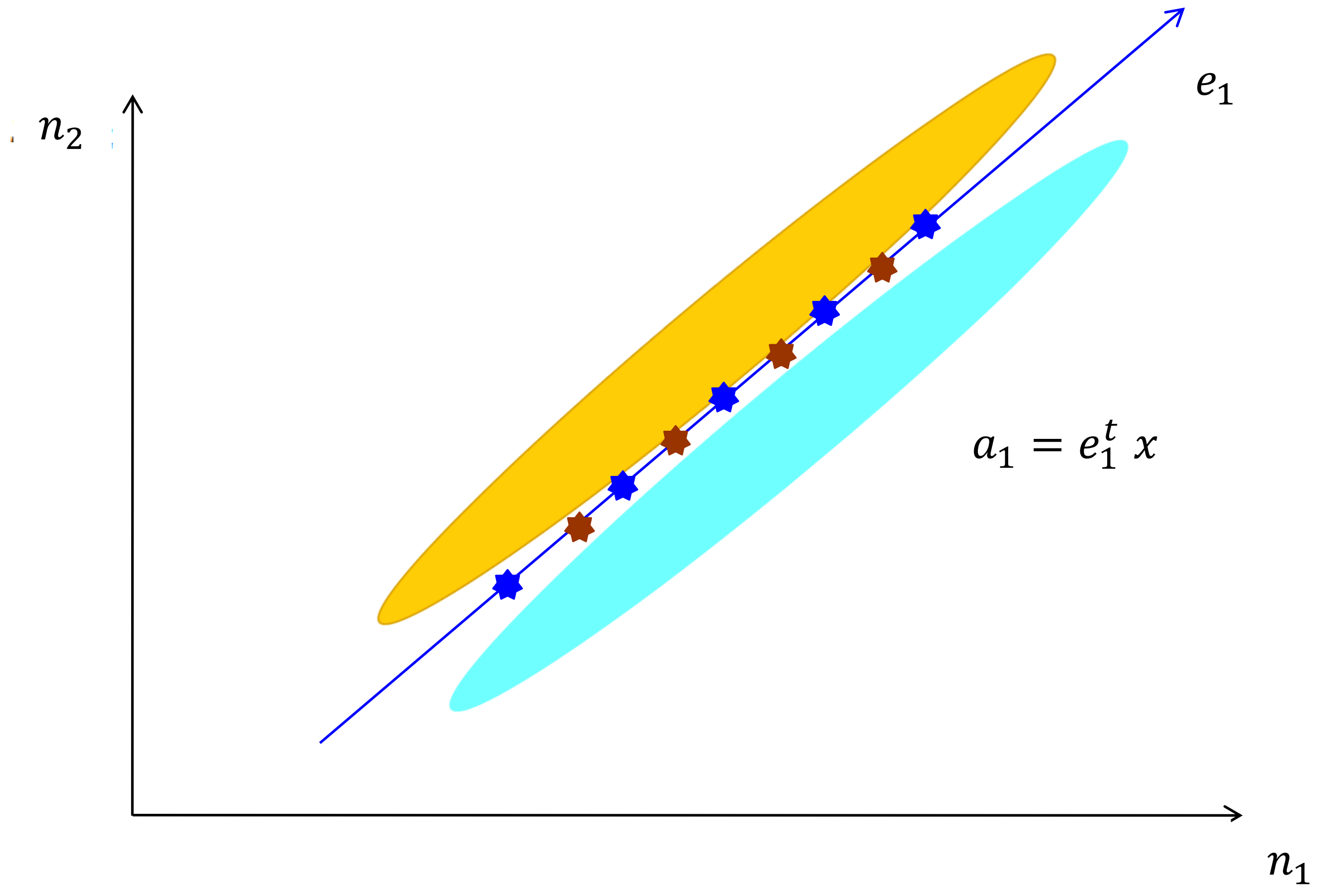


$$x = x_1 n_1 + x_2 n_2 = [n_1\ n_2]x = Nx$$
$$x = a_1 e_1 + a_2 e_2 = [e_1\ e_2]a = Ea$$

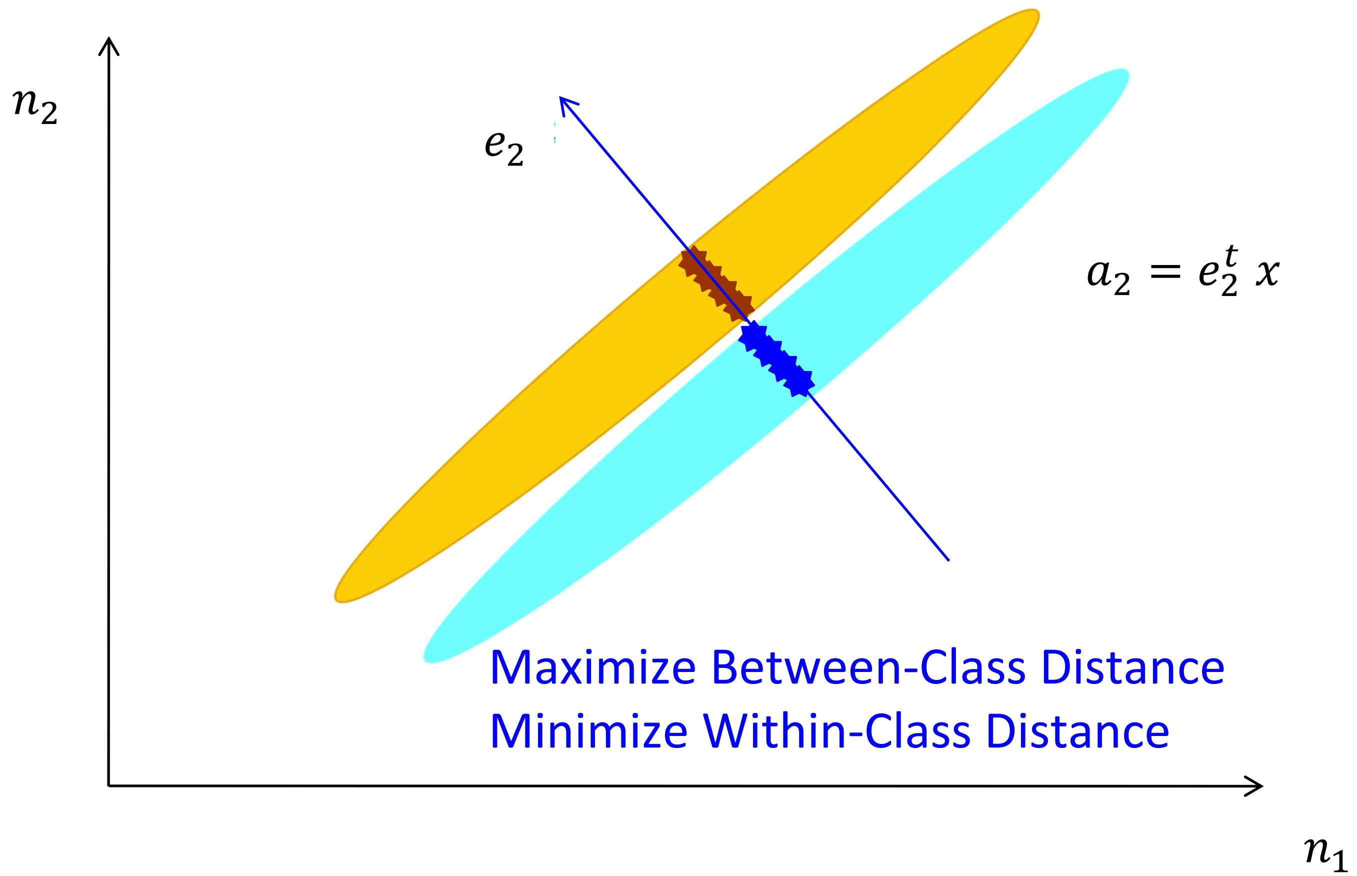
$$e_1^t x = a_1 e_1^t e_1 + a_2 e_1^t e_2 = a_1$$
$$e_2^t x = a_1 e_2^t e_1 + a_2 e_2^t e_2 = a_2$$

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} e_1^t \\ e_2^t \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad a = E^t x$$

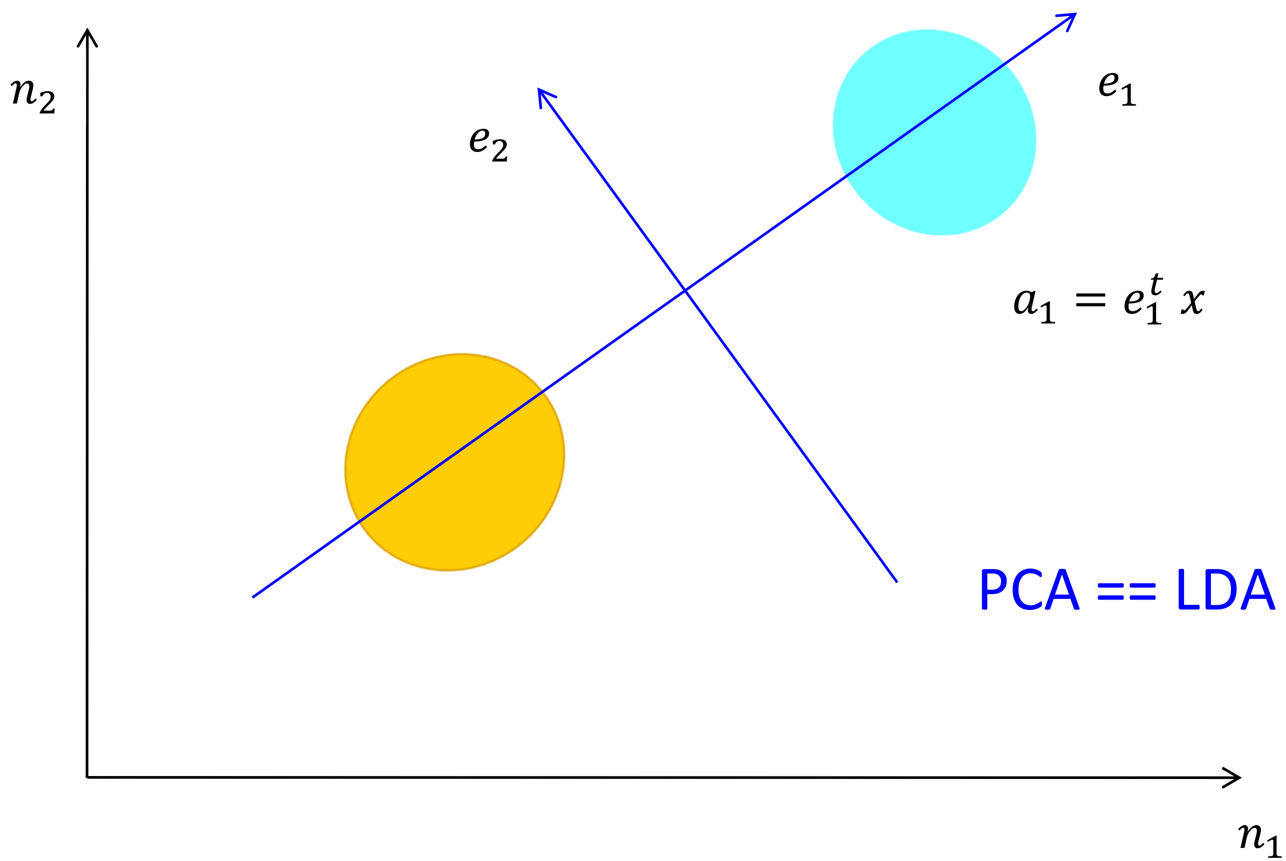
Linear Discriminant Analysis



Linear Discriminant Analysis



PCA & LDA



Principal Components Analysis (PCA)

- How to represent n d-dimensional vector samples $\{\mathbf{x}_1, .., \mathbf{x}_n\}$ by a single vector $\mathbf{x}_0$?
  - Find $\mathbf{x}_0$ that minimizes squared error correction function

$$J_0(\mathbf{x}_0) = \sum_{k=1}^{n} \| \mathbf{x}_0 - \mathbf{x}_k \|^2 .$$

Principal Components Analysis (PCA)

- How to represent *n* *d*-dimensional vector samples $\{\mathbf{x}_1, .., \mathbf{x}_n\}$ by a single vector $\mathbf{x}_0$?
  - Find $\mathbf{x}_0$ that minimizes squared error correction function

$$J_0(\mathbf{x}_0) = \sum_{k=1}^{n} \| \mathbf{x}_0 - \mathbf{x}_k \|^2 .$$

- The solution is sample mean

$$\mathrm{x}_0 = \mathrm{m} = \frac{1}{\mathrm{n}} \sum_{\mathrm{k}=1}^{\mathrm{n}} \mathrm{x}_\mathrm{k}$$

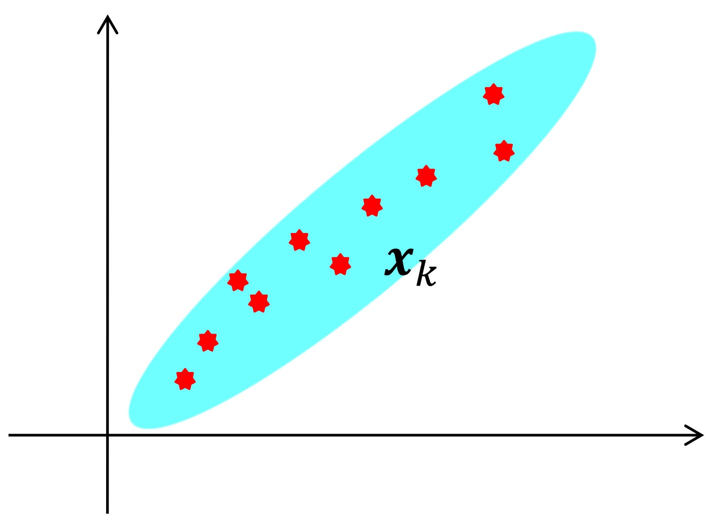
- This is zero-dimensional representation of the data set.
- One-dimensional representation by projecting the data onto a line through the sample mean reveals variability in the data.

Principal Components Analysis (PCA)

- This is zero-dimensional representation of the data set.

$$x_0 = m = \frac{1}{n}\sum_{k=1}^{n} x_k$$

- One-dimensional representation by projecting the data onto a line through the sample mean reveals variability in the data.



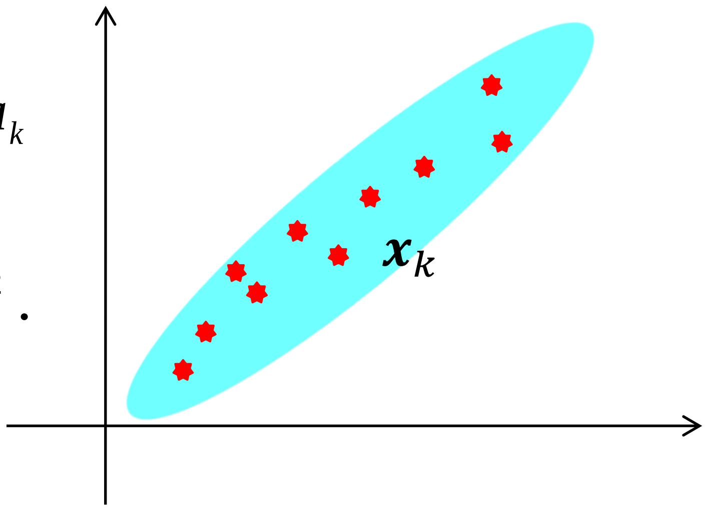
PCA ; Projection

- Let **e** be a unit vector in a direction of the line. The equation of the line

$$\mathbf{x} = \mathbf{m} + a\ \mathbf{e}$$

- Representing $\mathbf{x}_k$ by $\mathbf{m} + a_k\mathbf{e}$ find "optimal" a_k set minimizing criterion function :

$$J_1(a_1,\ldots,a_n,\mathbf{e}) = \sum_{k=1}^{n} \| \mathbf{m} + a_k\mathbf{e} - \mathbf{x}_k \|^2 .$$



PCA ; Projection

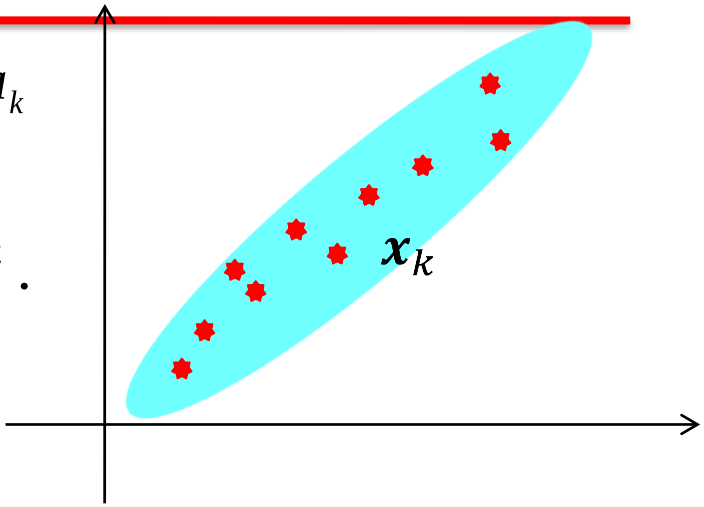
- Representing $\mathbf{x}_k$ by $\mathbf{m}+a_k\mathbf{e}$ find "optimal" a_k

  set minimizing criterion function :

$$J_1(a_1,\ldots,a_n,\mathbf{e})=\sum_{k=1}^{n}\|\mathbf{m}+a_k\mathbf{e}-\mathbf{x}_k\|^2 .$$

  from $\quad \partial J_1/\partial a_k=0$

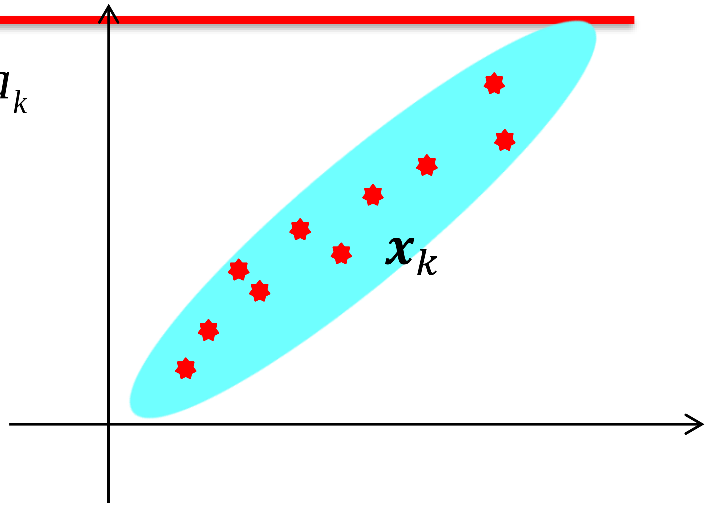
  we find $\quad a_k=\mathbf{e}^t(\mathbf{x}_k-\mathbf{m})$



PCA ; Projection

- Representing $\mathbf{x}_k$ by $\mathbf{m} + a_k\mathbf{e}$ find "optimal" a_k

$$a_k = \mathbf{e}^t(\mathbf{x}_k - \mathbf{m})$$



- How to find the *best* direction for **e** ?

- The least square solution: project the vector $\mathbf{x}_k$ onto the line in the direction of **e**, passing through the sample mean.

$$J_1(a_1, \ldots, a_n, \mathbf{e}) = \sum_{k=1}^{n} \| \mathbf{m} + a_k\mathbf{e} - \mathbf{x}_k \|^2 . \qquad a_k = \mathbf{e}^t(\mathbf{x}_k - \mathbf{m})$$

- Minimize *J* w.r.t **e.**

PCA ; Scatter matrix

- Substituting a_k into $J_1(a, \mathbf{e})$ we find

$$J_1(a,\mathbf{e}) = \sum_{k=1}^{n} a_k{}^2 \|\mathbf{e}\|^2 - 2\sum_{k=1}^{n} a_k \mathbf{e}^t(\mathbf{x}_k - \mathbf{m}) + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2$$

$$= \sum_{k=1}^{n} a_k{}^2 - 2\sum_{k=1}^{n} a_k{}^2 + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2 = -\sum_{k=1}^{n} [\mathbf{e}^t(\mathbf{x}_k - \mathbf{m})]^2 + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2$$

$$= -\sum_{k=1}^{n} \mathbf{e}^t(\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^t \mathbf{e} + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2$$

$$= -\mathbf{e}^t \mathbf{S} \mathbf{e} + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2$$

- where a *scatter matrix* **S** which is $(n-1)$ times of sample covariance matrix

$$\mathbf{S} = \sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^t.$$

PCA ; Scatter matrix

$$J_1(a,\mathbf{e}) = -\mathbf{e}^t\mathbf{S}\mathbf{e} + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2$$

- Vector **e** that minimizes J_1 also maximizes $\mathbf{e}^t\mathbf{S}\mathbf{e}$.
- So we find **e,** which maximize $\mathbf{e}^t\mathbf{S}\mathbf{e}$

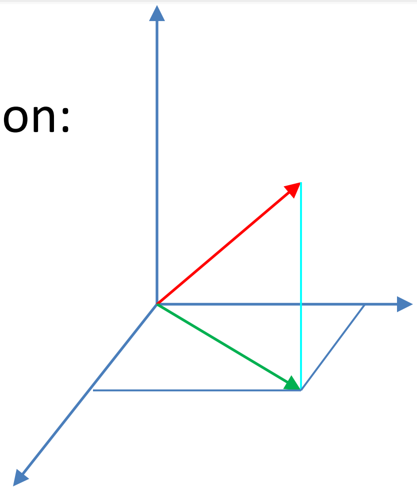
  subject to constraint $\|\boldsymbol{e}\|=1$

- Let λ be Lagrange multiplier. $L = \mathbf{e}^t\mathbf{S}\mathbf{e} - \lambda(\mathbf{e}^t\mathbf{e} - 1)$
- Differentiating L with respect to **e:** $\partial L / \partial \mathbf{e} = 2\mathbf{S}\mathbf{e} - 2\lambda\mathbf{e}$
- By setting to zero we see that **e** is an eigenvector of S:

$$\mathbf{S}\mathbf{e} = \lambda\mathbf{e} \quad \mathbf{e}^t\mathbf{S}\mathbf{e} = \lambda$$

- So to maximize $\mathbf{e}^t\mathbf{S}\mathbf{e}$ takes maximal λ

PCA ; Scatter matrix



- The result is easily extended to d' dimensional projection:

$$\mathbf{x}'_k = \mathbf{m} + \sum_{i=1}^{d'} a_k^i \mathbf{e}_i \qquad \text{where} \qquad d' \le d$$

- The criterion function

$$J_{d'} = \sum_{k=1}^{n} \left\| \left(\mathbf{m} + \sum_{i=1}^{d'} a_k^i \mathbf{e}_i \right) - \mathbf{x}_k \right\|^2$$

  is minimized when vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{d'}$ are the eigenvectors having the largest eigenvalues.

- The coefficients $a_k^i = \mathbf{e}_i^t(\mathbf{x}_k - \mathbf{m})$ are *principal components.*

Error function

- If d' < d error which is made by dropping the last terms is

$$\mathbf{x}'_k = \mathbf{m}_k + \sum_{i=1}^{d'} a_k^i \mathbf{e}_i$$

$$a_k^i = \mathbf{e}_i^t (\mathbf{x}_k - \mathbf{m})$$

$$J_{d'} = \sum_{k=1}^{n} \left\| \sum_{i=d'+1}^{d} a_k^i \mathbf{e}_i \right\|^2$$

$$= \sum_{i=d'+1}^{d} \mathbf{e}_i^t \sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^t \mathbf{e}_i$$

$$= \sum_{i=d'+1}^{d} \mathbf{e}_i^t \mathbf{S} \mathbf{e}_i = \sum_{i=d'+1}^{d} \lambda_i$$

- This is a sum of lowest eigenvalues.

PCA – the algorithm

- Input: $X^{(n)} = \{\mathbf{x}_1, .., \mathbf{x}_n\}, \quad \mathbf{x}_k = \left\langle x_1^k, ..., x_d^k \right\rangle$
- Take $d^{'} < d$
- Output: $A^{(n)} = \{\mathbf{a}_1, .., \mathbf{a}_n\} \quad \mathbf{a}_k = \{a_1^k, .., a_{d^{'}}^k\}$
- Algorithm:
  - Compute the mean of the training set $\mathbf{m} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k.$
  - Compute the scatter matrix ***S***.
  - Find eigenvectors of ***S*** and corresponding eigenvalues:

$$S\{\mathbf{e}_i, \lambda_i\}_{i=1}^{d} \;,\; \forall i : \mathbf{S}\mathbf{e}_i = \lambda \mathbf{e}_i \;, \lambda_1 \geq \lambda_2 \geq ... \lambda_d$$

  - Choose $d^{'}$ eigenvectors, and for each sample $\mathbf{x}_k$ point compute

$$\mathbf{a}_k = \{\mathbf{e}_i{}^t (\mathbf{x}_k - \mathbf{m})\}_{i=1}^{d^{'}}$$

Interim Summary

- Principal Component Analysis
  - ✓Feature Extraction
  - ✓Dimension Reduction

$$J_{d'} = \sum_{k=1}^{n} \left\| \left(\mathbf{m} + \sum_{i=1}^{d'} a_k^i \mathbf{e}_i \right) - \mathbf{x}_k \right\|^2$$

$$\mathbf{S} = \sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^t.$$

$$\mathbf{Se} = \lambda \mathbf{e}$$

$$\mathbf{e}^t \mathbf{Se} = \lambda$$

$$a_k{}^i = \mathbf{e}_i{}^t (\mathbf{x}_k - \mathbf{m}),\ i = 1, \ldots, d'$$

$$\begin{bmatrix} a_k^1 \\ a_k^2 \\ \cdots \\ a_k^{d'} \end{bmatrix} = \begin{bmatrix} \mathbf{e}_1{}^t \\ \mathbf{e}_2{}^t \\ \cdots \\ \mathbf{e}_{d'}{}^t \end{bmatrix} (\mathbf{x}_k - \mathbf{m})$$

$$\mathbf{a}_k = \mathrm{E}^t (\mathbf{x}_k - \mathbf{m})$$

$$cf)\ \ \mathrm{y}_k = \mathrm{W}^t (\mathbf{x}_k - \mathbf{m})$$

Feature Dimension Reduction: PCA & LDA (II)

Jin Young Choi

Seoul National University

Outline

Feature Extraction

Introduction of PCA & LDA

Principal Component Analysis (PCA)

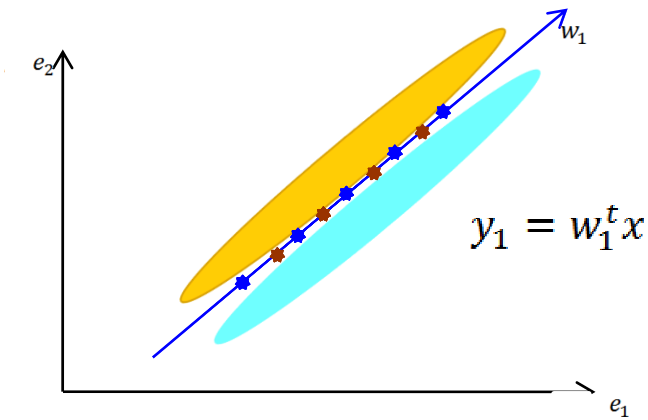
Linear Discriminant Analysis (FLDA)

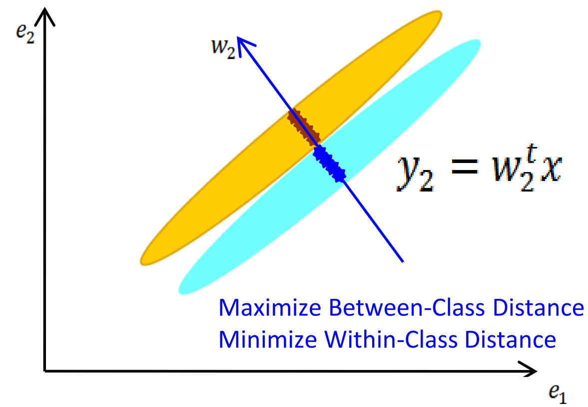
Multiple Discriminant Analysis (MDA)

Simple Enhancement of PCA/LDA

Linear Discriminant Analysis: LDA

- We have n d-dimensional samples $\mathbf{x}_1,.., \mathbf{x}_n$, n_1 in a subset D_1, labeled w_1 and n_2 in a subset D_2, labeled w_2 .
- Find direction of line **w** , that maximally separate the data.





- Let a difference between sample means be a measure of separation of projected points

Fisher Linear Discriminant cont.

- Project samples $\mathbf{x}_k$ onto **w.**

$$y_k = \mathbf{w}^t \mathbf{x}_k$$

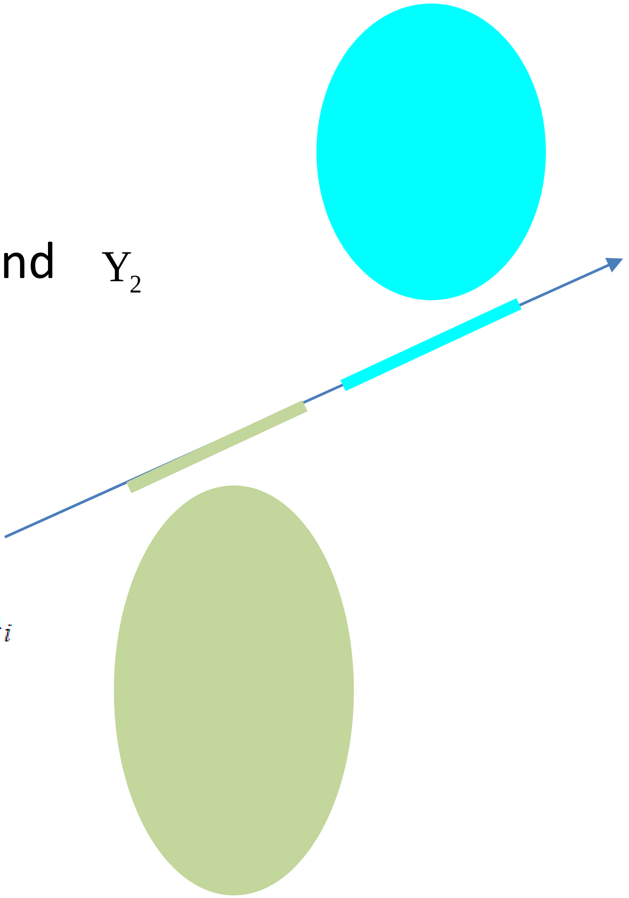
- n samples y_k are divided into the subsets Y_1 and Y_2
- Let $\mathbf{m}_i$ be the sample mean $\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{x}$

- The sample mean for projected points

$$\tilde{m}_i = \frac{1}{n_i} \sum_{y \in Y_i} y = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{w}^t \mathbf{x} = \mathbf{w}^t \mathbf{m}_i$$

- Distance between the projected means is

$$| \tilde{m}_1 - \tilde{m}_2 | = | \mathbf{w}^t (\mathbf{m}_1 - \mathbf{m}_2) |$$



Fisher Linear Discriminant cont.

- A scatter for projected samples labeled ω_i

$$\tilde{s}_i^2 = \sum_{y \in Y_i} (y - \tilde{m}_i)^2$$

$(1/n)(\tilde{s}_1^2 + \tilde{s}_2^2)$ is an estimate of the variance of the pooled data.
$\tilde{s}_1^2 + \tilde{s}_2^2$ is called total within-class scatter of the projected samples.

- The Fisher discriminant employs $\mathbf{w}^t\mathbf{x}$ for which criterion

$$J(\mathbf{w}) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

is maximum

Fisher Linear Discriminant cont.

- Define scatter matrices $\mathbf{S}_i$ and $\mathbf{S}_w$ by

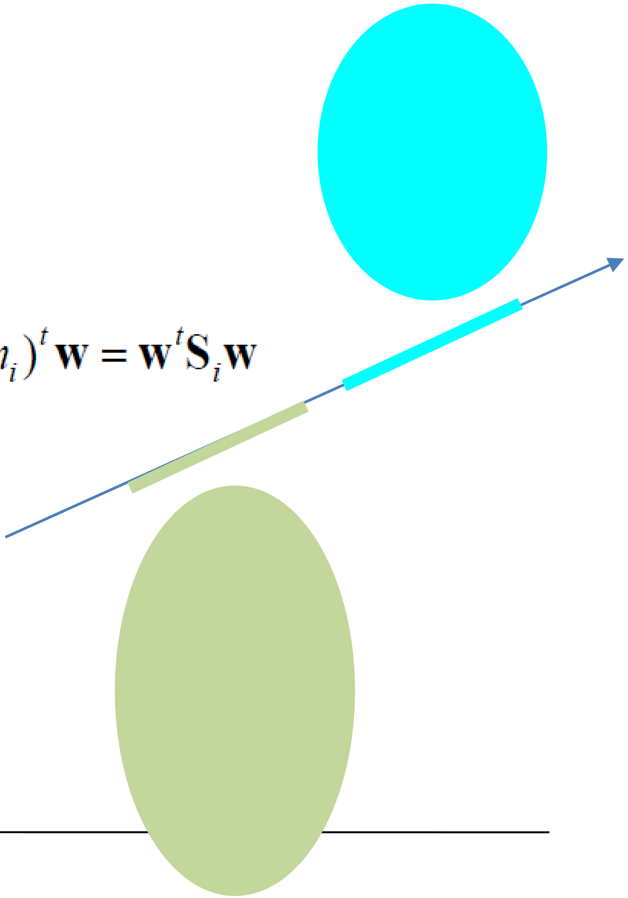
$$S_i = \sum_{x \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t$$

and

$$S_w = S_1 + S_2$$

- Then

$$\tilde{s}_i^2 = \sum_{\mathbf{x} \in D_i} (\mathbf{w}^t \mathbf{x} - \mathbf{w}^t m_i)^2 = \sum_{\mathbf{x} \in D_i} \mathbf{w}^t (\mathbf{x} - m_i)(\mathbf{x} - m_i)^t \mathbf{w} = \mathbf{w}^t \mathbf{S}_i \mathbf{w}$$



- Thus $\tilde{s}_1^2 + \tilde{s}_2^2 = \mathbf{w}^t \mathbf{S}_w \mathbf{w}$

Fisher Linear Discriminant cont.

- Similarly,

$$(\tilde{m}_1 - \tilde{m}_2)^2 = (\mathbf{w}^t\mathbf{m}_1 - \mathbf{w}^t\mathbf{m}_2)^2 = \mathbf{w}^t(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t\mathbf{w} = \mathbf{w}^t\mathbf{S}_B\mathbf{w}$$

  $\mathbf{S}_w$ is called within-class scatter matrix (proportional to sample covariance matrix)

  $\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$ is called between-class scatter matrix.

- This gives the equivalent expression for Fisher's discriminant

$$J(\mathbf{w}) = \frac{\mathbf{w}^t\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}}$$

- Which vector **w** maximizes it?

$$\nabla_{\mathbf{w}} J(\mathbf{w}) = \frac{2\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}} - \frac{\mathbf{w}^t\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}}\frac{2\mathbf{S}_W\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}} = 0$$

Fisher Linear Discriminant cont.

- Hence one gets

$$\mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_W\mathbf{w}, \quad \lambda = \frac{\mathbf{w}^t\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}},$$

  or equivalently

$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{w} = \lambda\mathbf{w},$$

- Since for any **w**, $\mathbf{S}_B\mathbf{w}$ is always in the direction of $\mathbf{m}_1$-$\mathbf{m}_2$:

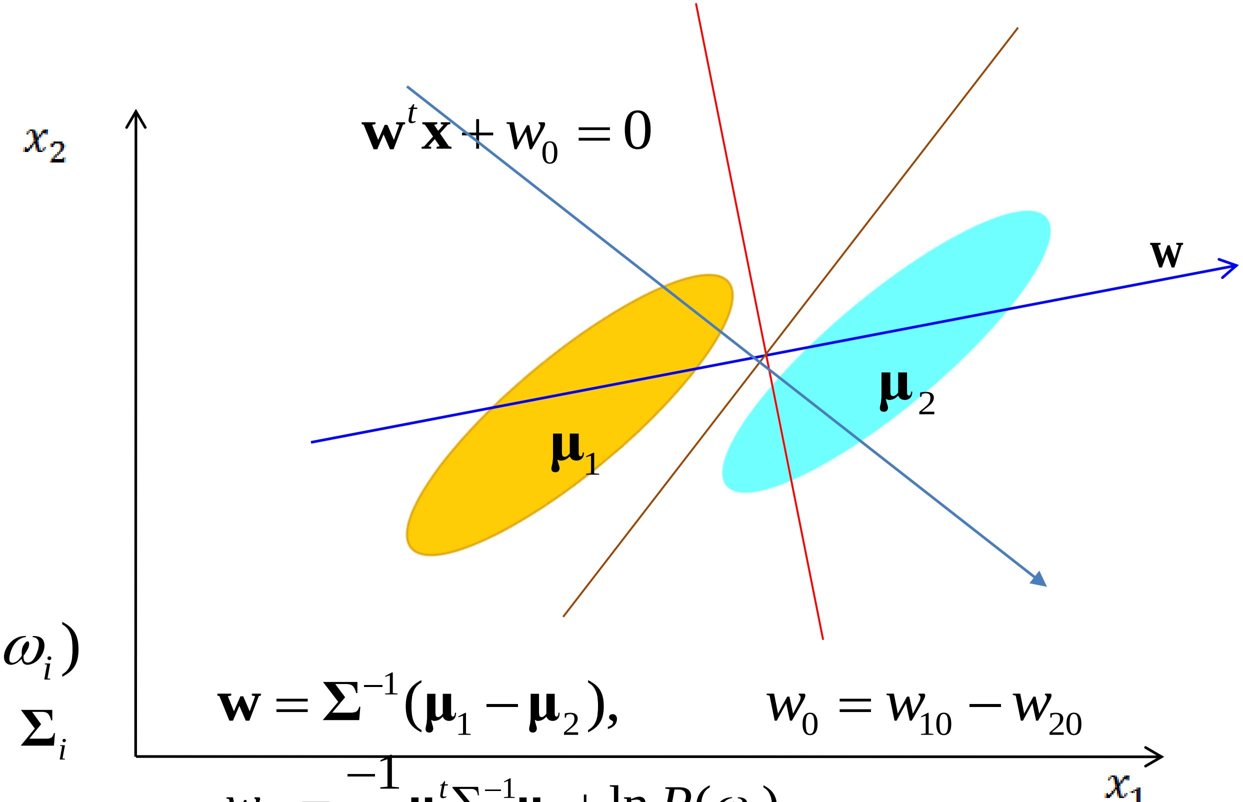
$$\mathbf{S}_B\mathbf{w} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t\mathbf{w} = \alpha(\mathbf{m}_1 - \mathbf{m}_2)$$

- It is not necessary to determine the eigenvalues of $\mathbf{S}_W^{-1}\mathbf{S}_B$.
- One simply gets

$$\mathbf{w} \propto S_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

- Scale factor for **w** is unimportant (why?).
- FLDA is one-dimensional projection

Fisher Linear Discriminant cont.



$p(\mathbf{x} \mid \omega_i)$
with $\boldsymbol{\Sigma}_i$

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \qquad w_0 = w_{10} - w_{20}$$

$$w_{i0} = \frac{-1}{2}\boldsymbol{\mu}_i^t \Sigma^{-1} \boldsymbol{\mu}_i + \ln P(\omega_i)$$

Matrix Norm

- Induced Norm

$$\|A\|_p = \sup_{\|x\|_p=1} \|Ax\|_p \qquad \|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{m} |a_{ij}| \qquad \|A\|_\infty = \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}|$$

- Schatten norm

$$\|A\|_p = \left(\sum_{i=1}^{\min\{m,n\}} \sigma_i^p(A) \right)^{1/p}$$

- nuclear norm

$$\|A\|_* = \operatorname{trace}\left(\sqrt{A^* A}\right) = \sum_{i=1}^{\min\{m,n\}} \sigma_i(A)$$

- Frobenius Norm

$$\|A\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2} = \sqrt{\operatorname{trace}\left(A^{\mathsf{T}} A\right)} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2(A)}$$

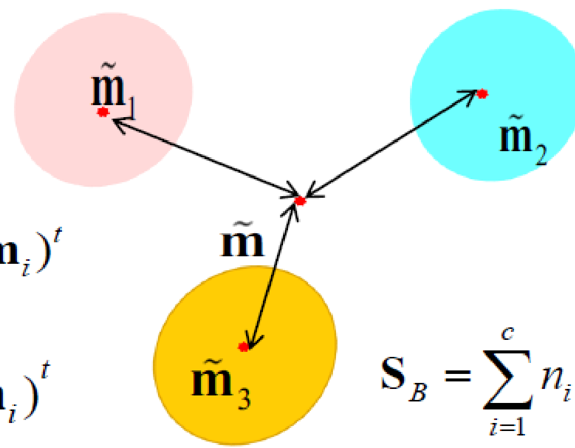
- Spectral (maximum singular value) norm

$$f(A) = \|A\|_2 = \sigma_{\max}(A) = (\lambda_{\max}(A^T A))^{1/2}$$

$$\begin{aligned}
\|A\|_2 &= \sup_{\|x\|_2=1} \|Ax\|_2 \\
&= \sup_{\|x\|_2=1} (x^T A^T A x)^{1/2} \\
&= \sup_{\|x\|_2=1} (x^T U^T \Lambda U x)^{1/2} \\
&= \sup_{\|y\|_2=1} (y^T \Lambda y)^{1/2} \Leftarrow y^T y = x^T U^T U x = 1 \\
&= \sup_{\|y\|_2=1} \left(\sum y_i^2 \lambda_i\right)^{1/2} \\
&= (\lambda_{\max}(X^T X))^{1/2}
\end{aligned}$$

Multiple Discriminant Analysis

$$d \geq c$$



$$\mathbf{S}_w = \sum_{i=1}^{c} \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t$$

$$\tilde{\mathbf{S}}_w = \sum_{i=1}^{c} \sum_{\mathbf{y} \in Y_i} (\mathbf{y} - \tilde{\mathbf{m}}_i)(\mathbf{y} - \tilde{\mathbf{m}}_i)^t$$

$$\mathbf{S}_B = \sum_{i=1}^{c} n_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^t$$

$$\tilde{\mathbf{S}}_B = \sum_{i=1}^{c} n_i (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}})(\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}})^t$$

$$cf) \quad \mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$$

Multiple Discriminant Analysis

- For the *c* -class problem we have c-1 discriminant functions.
- The projection from a *d*-dimensional space to a (*c*-1) dimension is accomplished by (*c*-1) discriminant functions (we assume $d \geq c$).
- Within-class scatter matrix is: $\mathbf{S}_w = \sum_{i=1}^{c} \mathbf{S}_i$

  where $\mathbf{S}_i = \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t$

  and $\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{x}$
- Define a total mean vector

$$\mathbf{m} = \frac{1}{n} \sum_{\mathbf{x}} \mathbf{x} = \frac{1}{n} \sum_{i=1}^{c} n_i \mathbf{m}_i$$

Multiple Discriminant Analysis

- And total scatter matrix $$\mathbf{S}_T = \sum_{\mathbf{x}} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^t$$
- It can be transformed to

$$\mathbf{S}_T = \sum_{i=1}^{c} \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mathbf{m}_i + \mathbf{m}_i - \mathbf{m})(\mathbf{x} - \mathbf{m}_i + \mathbf{m}_i - \mathbf{m})^t$$

$$= \sum_{i=1}^{c} \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t + \sum_{i=1}^{c} \sum_{\mathbf{x} \in D_i} (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^t$$

$$= \mathbf{S}_W + \sum_{i=1}^{c} n_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^t = \mathbf{S}_W + \mathbf{S}_B$$

- The between-class scatter is:

$$\mathbf{S}_B = \sum_{i=1}^{c} n_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^t$$

Multiple Discriminant Analysis

- For the c -class problem we have (c-1) discriminant functions. The projection from a d-dimensional space to a (c-1) dimensional space is accomplished by (c-1) discriminant functions:

$$y_i = \mathbf{w}_i^t \mathbf{x} \qquad i = 1, \ldots, (c-1)$$

- Taking d-by-(c-1) $\mathbf{W}$ matrix which columns are vectors $\mathbf{w}_i$, we'll get in matrix form: $\mathbf{y} = \mathbf{W}^t \mathbf{x}$
- Samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are projected to $\mathbf{y}_1, \ldots, \mathbf{y}_n$.
- We define $\tilde{\mathbf{m}}_i = \frac{1}{n_i} \sum_{\mathbf{y} \in Y_i} \mathbf{y}$, $\tilde{\mathbf{m}} = \frac{1}{n} \sum_{\mathbf{y} \in Y_i} n_i \tilde{\mathbf{m}}_i$

Multiple Discriminant Analysis cont.

$$\tilde{\mathbf{S}}_W = \sum_{i=1}^{c} \sum_{\mathbf{y} \in Y_i} (\mathbf{y} - \tilde{\mathbf{m}}_i)(\mathbf{y} - \tilde{\mathbf{m}}_i)^t$$

$$\tilde{\mathbf{S}}_B = \sum_{i=1}^{c} n_i (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}})(\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}})^t$$

- It's easy to show that $\tilde{\mathbf{S}}_W = \mathbf{W}^t \mathbf{S}_W \mathbf{W}$ *and* $\tilde{\mathbf{S}}_B = \mathbf{W}^t \mathbf{S}_B \mathbf{W}$

- The criterion function which should be maximized is:

$$J(\mathbf{W}) = \frac{|\tilde{\mathbf{S}}_B|}{|\tilde{\mathbf{S}}_W|} = \frac{|\mathbf{W}^t \mathbf{S}_B \mathbf{W}|}{|\mathbf{W}^t \mathbf{S}_W \mathbf{W}|} = \frac{tr(\mathbf{W}^t \mathbf{S}_B \mathbf{W})}{tr(\mathbf{W}^t \mathbf{S}_W \mathbf{W})} = \frac{\sum_i \mathbf{w}_i^t \mathbf{S}_B \mathbf{w}_i}{\sum_i \mathbf{w}_i^t \mathbf{S}_W \mathbf{w}_i}$$

- Every column $\mathbf{w}_i$ of **W** we should be solution of generalized eigenvalue problem

$$\mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{w}_i = \lambda_i \mathbf{w}_i$$

Multiple Discriminant Analysis cont.

- The criterion function which should be maximized is:

$$J(\mathbf{W}) = \frac{|\tilde{\mathbf{S}}_B|}{|\tilde{\mathbf{S}}_W|} = \frac{|\mathbf{W}^t \mathbf{S}_B \mathbf{W}|}{|\mathbf{W}^t \mathbf{S}_W \mathbf{W}|} = \frac{tr(\mathbf{W}^t \mathbf{S}_B \mathbf{W})}{tr(\mathbf{W}^t \mathbf{S}_W \mathbf{W})} = \frac{\sum_i \mathbf{w}_i^t \mathbf{S}_B \mathbf{w}_i}{\sum_i \mathbf{w}_i^t \mathbf{S}_W \mathbf{w}_i}$$

- Every column $\mathbf{w}_i$ of **W** we should be solution of generalized eigenvalue problem

$$\mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{w}_i = \lambda_i \mathbf{w}_i$$

$$\mathbf{S}_B \mathbf{w}_i = \lambda_i \mathbf{S}_W \mathbf{w}_i$$

$$\mathbf{S}_B \mathbf{W} = \mathbf{S}_W \mathbf{W} \Lambda$$

$$\mathbf{W}^t \mathbf{S}_B \mathbf{W} = \mathbf{W}^t \mathbf{S}_W \mathbf{W} \Lambda$$

$$tr(\mathbf{W}^t \mathbf{S}_B \mathbf{W}) = tr(\mathbf{W}^t \mathbf{S}_W \mathbf{W}) tr(\Lambda)$$

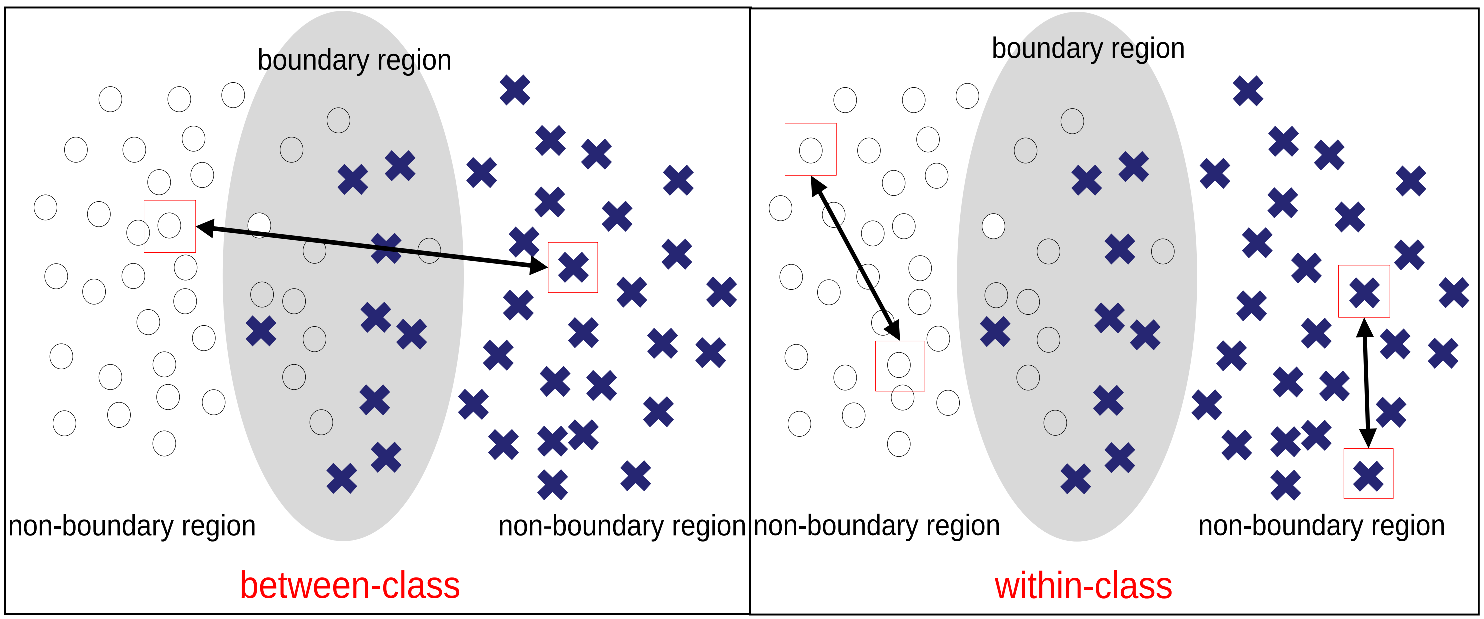
$$tr(\Lambda) = \sum_i \lambda_i = \frac{tr(\mathbf{W}^t \mathbf{S}_B \mathbf{W})}{tr(\mathbf{W}^t \mathbf{S}_W \mathbf{W})}$$

Multiple Discriminant Analysis cont.

- The MDA provides the way of reducing the dimensionality of the problem.
- The technique for finding probability density might not be feasible in the original space.
- The technique for finding probability density may work well after reducing the dimension of feature space.
- MDA may improve the separability of classes.

Simple Enhancement for PCA/LDA

- Significant pairs for between-class scatter matrix
  - Non-boundary patterns with the different class labels
- Significant pairs for within-class scatter matrix
  - Non-boundary patterns with the same class labels



Non-boundary Pattern Selection Algorithm

- Step 1. For each $\mathbf{x}_i \in \mathbf{X}$
  - Find the neighborhood defined as follows.

    $$Neighbors(\mathbf{x}_i, k) = N(\mathbf{x}_i, k) \cup \boxed{\{\mathbf{x}_i\}}$$

    where $N(\mathbf{x}_i, k)$ is the set of k nearest samples to $\mathbf{x}_i$ by L2-norm.
  - Calculate voting probabilities of $Neighbors(\mathbf{x}_i, k)$ to each class j.

    $$p_j(\mathbf{x}_i) = \frac{\sum_{\forall n \in Neighbors(\mathbf{x}_i, k)} I_j(n)}{k+1}$$

    where $I_j(n)$ is 1 if the class of neighbor n is j, otherwise 0.
  - Calculate the neighborhood entropy of $\mathbf{x}_i$.

    $$Neighbors\_Entropy(\mathbf{x}_i, k) = \sum_{j=1}^{l} p_j(\mathbf{x}_i) \log_l \frac{1}{p_j(\mathbf{x}_i)}$$

- Step 2. Obtain boundary patterns $\mathbf{X}^{(B)}$ and non-boundary patterns $\mathbf{X}^{(NB)}$

  $$\mathbf{X}^{(NB)} = \{\mathbf{x} | Neighbors\_Entropy(\mathbf{x}, k) \le \boxed{\theta(l)}, \mathbf{x} \in \mathbf{X}\}$$

  $$\mathbf{X}^{(B)} = \mathbf{X} - \mathbf{X}^{(NB)}$$

PCA using NPS (LDA in the same manner)

- Select non-boundary patterns via BNPS.
- Non-boundary patterns make up significant pairs.
- Emphasize the significant pairs.

Whole data

$$\mathbf{C}_X = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^\top$$

$$\mathbf{W}_{PCA} = \arg\max_{\mathbf{W}^\top\mathbf{W}=\mathbf{I}} \mathrm{tr}(\mathbf{W}^\top\mathbf{C}_X\mathbf{W})$$

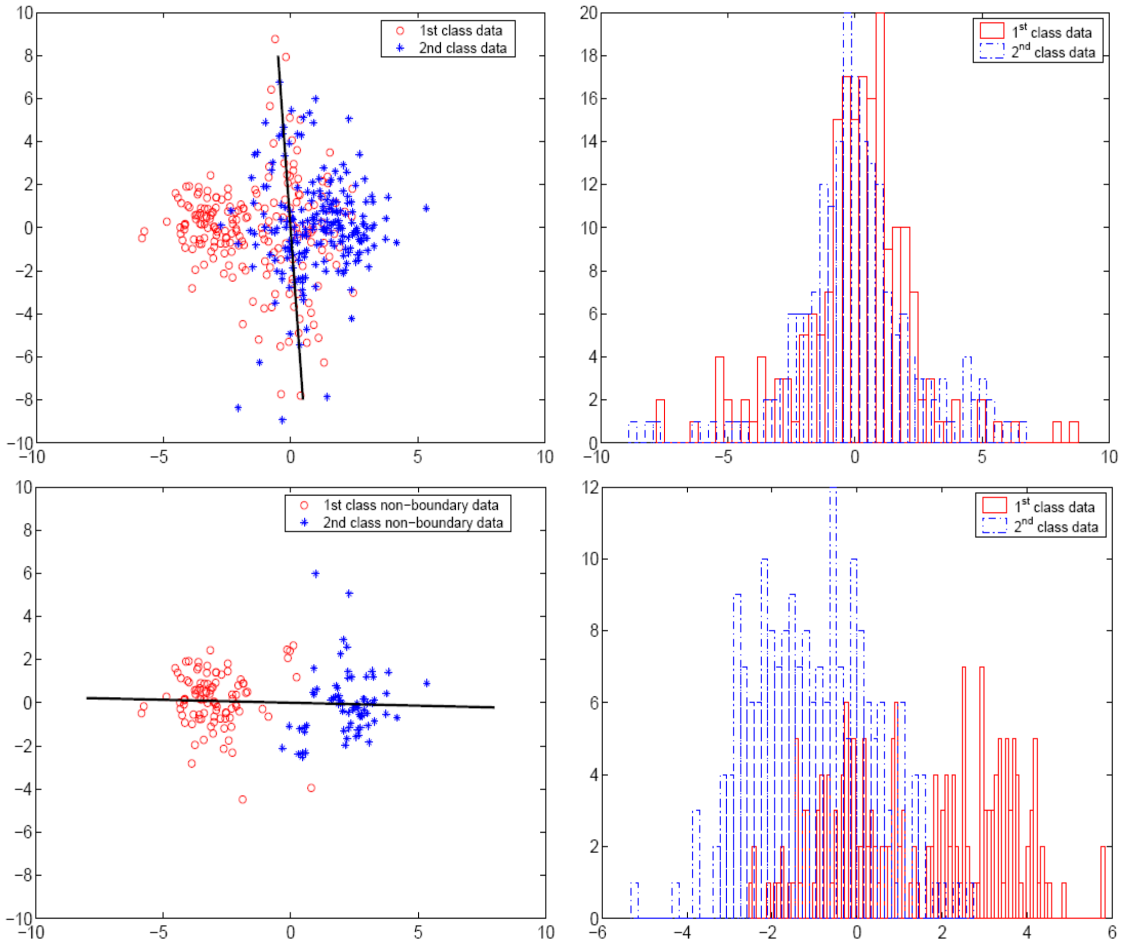
$$\Downarrow$$

Non-boundary pattern

$$\widetilde{\mathbf{C}}_X = \frac{1}{n_{NB}-1}\sum_{i=1}^{l}\sum_{j:y_j=i}(\mathbf{x}_j^{(NB)} - \widetilde{\mathbf{m}})(\mathbf{x}_j^{(NB)} - \widetilde{\mathbf{m}})^\top$$

$$\boxed{\widetilde{\mathbf{W}}_{PCA} = \arg\max_{\widetilde{\mathbf{W}}^\top\widetilde{\mathbf{W}}=\mathbf{I}} \mathrm{tr}(\widetilde{\mathbf{W}}^\top\widetilde{\mathbf{C}}_X\widetilde{\mathbf{W}})}$$

Toy Example

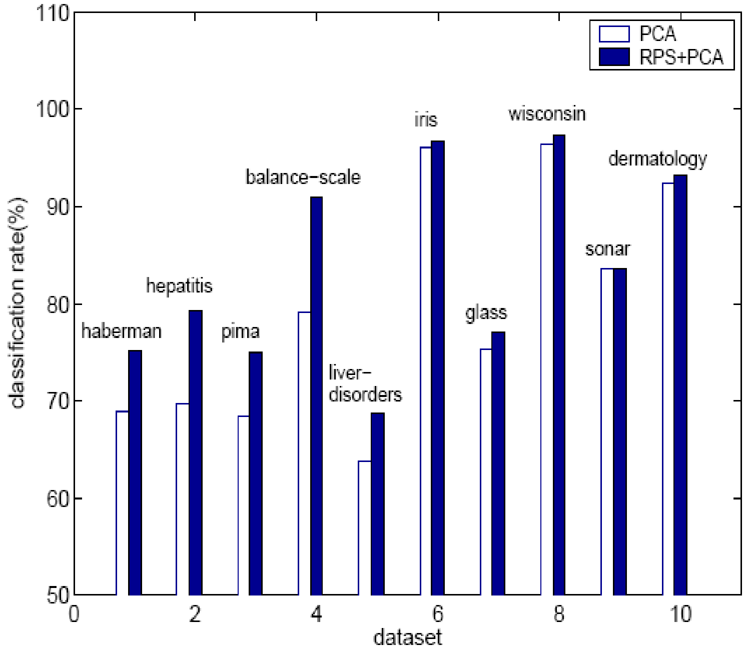


1st class data
2nd class data
1st class data
2nd class data
1st class non−boundary data
2nd class non−boundary data
1st class data
2nd class data

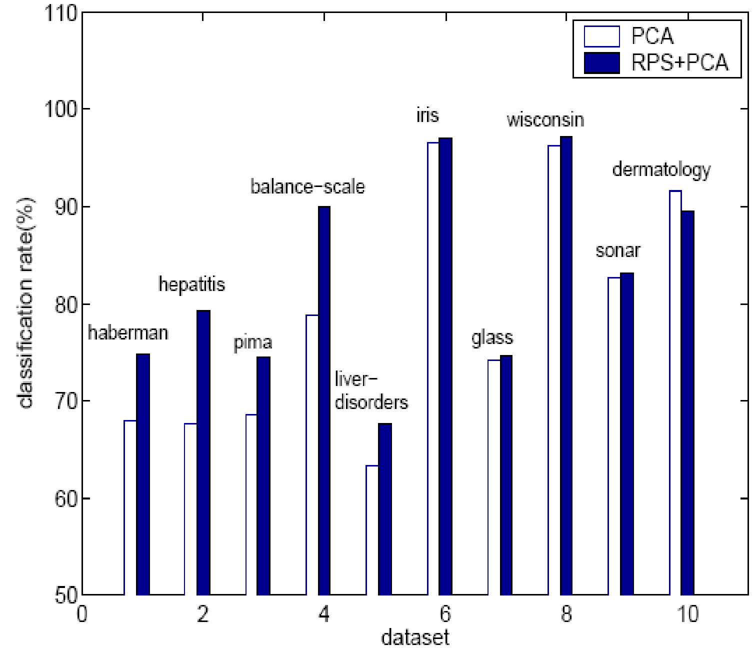
UCI Machine Learning Repository

Name	# of data	# of attributes	# of classes	Missing attributes
Haberman	306	3	2	No
Hepatitis	155	19	2	Yes
Pima	768	8	2	No
Balance-scale	625	4	3	No
Liver-disorders	345	6	2	No
Iris	150	4	3	No
Glass	214	9	6	No
Wisconsin	699	9	2	Yes
Sonar	208	60	2	No
Dermatology	366	34	6	Yes

PCA vs. NPS+PCA

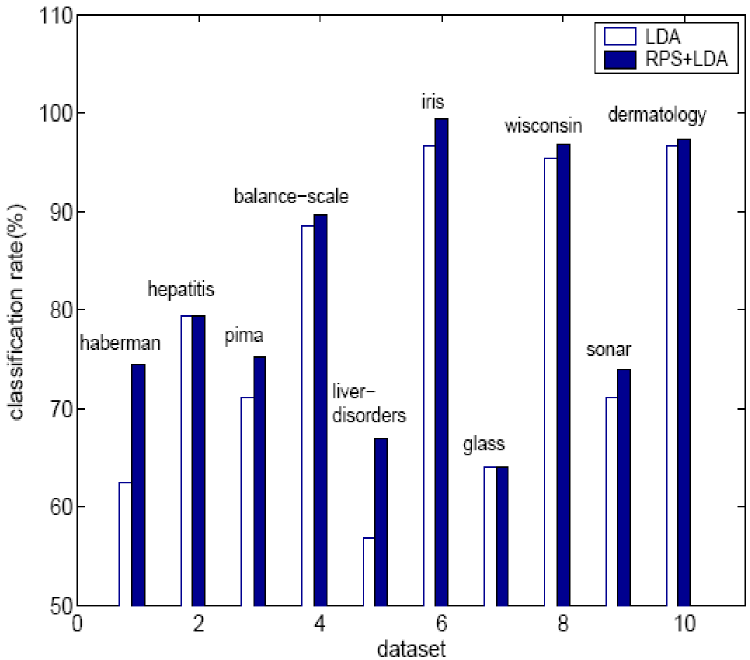


Leave-one-out
NN Classifier

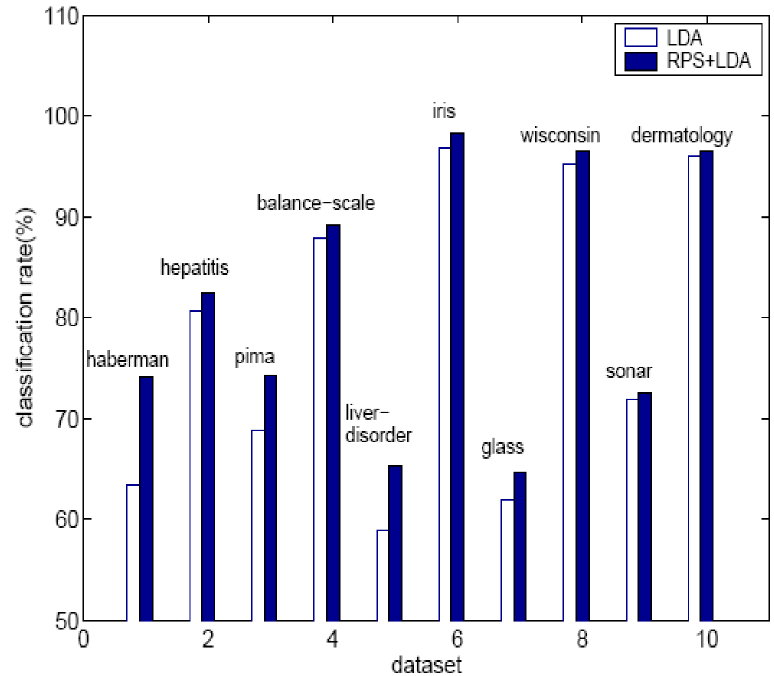


10-fold cross validation
NN Classifier

LDA vs. NPS+LDA



Leave-one-out
NN Classifier



10-fold cross validation
NN Classifier

Interim Summary

- Fisher Linear Discriminant Analysis

$$J(\mathbf{w}) = \frac{\mathbf{w}^t \mathbf{S}_B \mathbf{w}}{\mathbf{w}^t \mathbf{S}_W \mathbf{w}} \qquad \mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{w} = \lambda \mathbf{w}, \qquad \mathbf{w} \propto S_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

- Multiple Discriminant Analysis

$$\mathbf{S}_B = \sum_{i=1}^{c} n_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^t \qquad \mathbf{S}_W{}^{-1}\mathbf{S}_B\mathbf{w}_i = \lambda_i \mathbf{w}_i$$

- Simple Enhancement for PCA/LDA

$$\boxed{\widetilde{\mathbf{W}}_{PCA} = \arg \max_{\widetilde{\mathbf{W}}^\top \widetilde{\mathbf{W}} = \mathbf{I}} \operatorname{tr}(\widetilde{\mathbf{W}}^\top \widetilde{\mathbf{C}}_X \widetilde{\mathbf{W}})}$$

$$\boxed{\widetilde{\mathbf{W}}_{LDA} = \arg \max_{\widetilde{\mathbf{W}}^\top \widetilde{\mathbf{W}} = \mathbf{I}} \frac{\operatorname{tr}(\widetilde{\mathbf{W}}^\top \widetilde{\mathbf{S}}^{(b)} \widetilde{\mathbf{W}})}{\operatorname{tr}(\widetilde{\mathbf{W}}^\top \widetilde{\mathbf{S}}^{(w)} \widetilde{\mathbf{W}})}}$$