Course 495: Advanced Statistical Machine Learning/Pattern Recognition

Deterministic Component Analysis

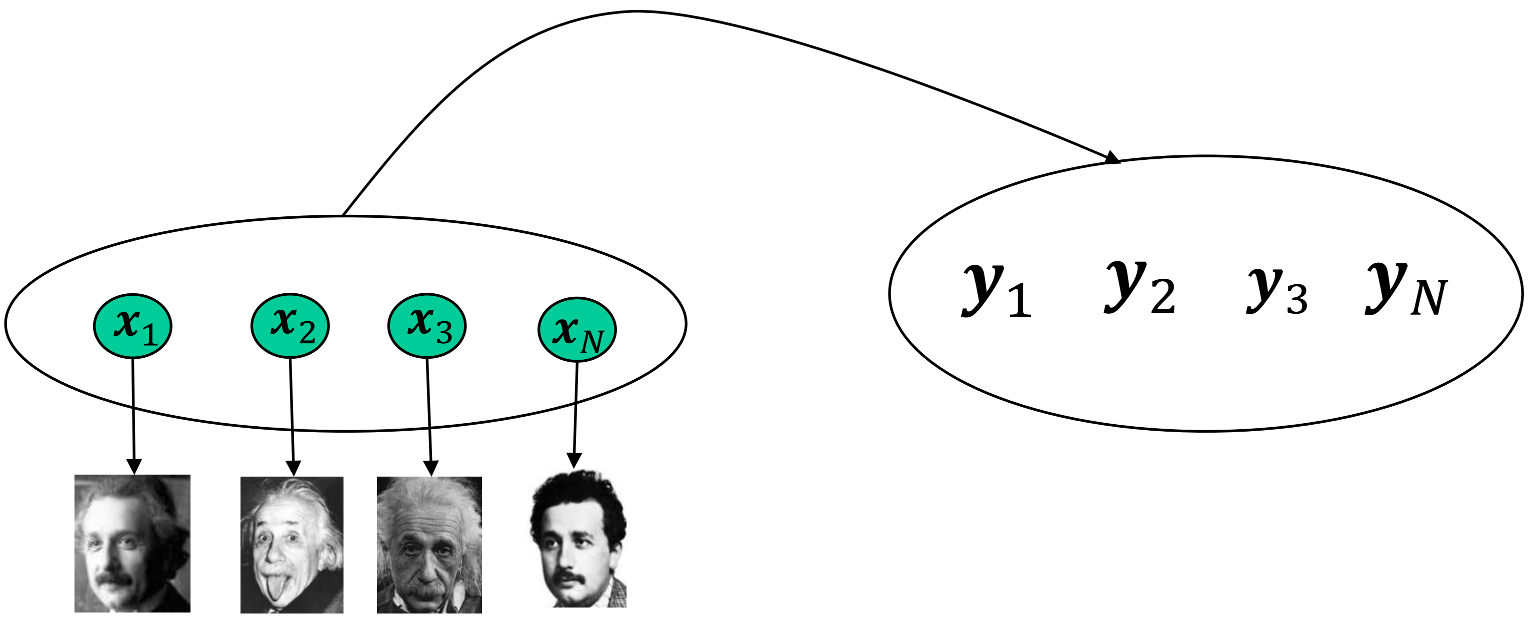
- Goal (Lecture): To present standard and modern Component Analysis (CA) techniques such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Graph/Neighbourhood based Component Analysis
- Goal (Tutorials): To provide the students the necessary mathematical tools for deeply understanding the CA techniques.

Materials

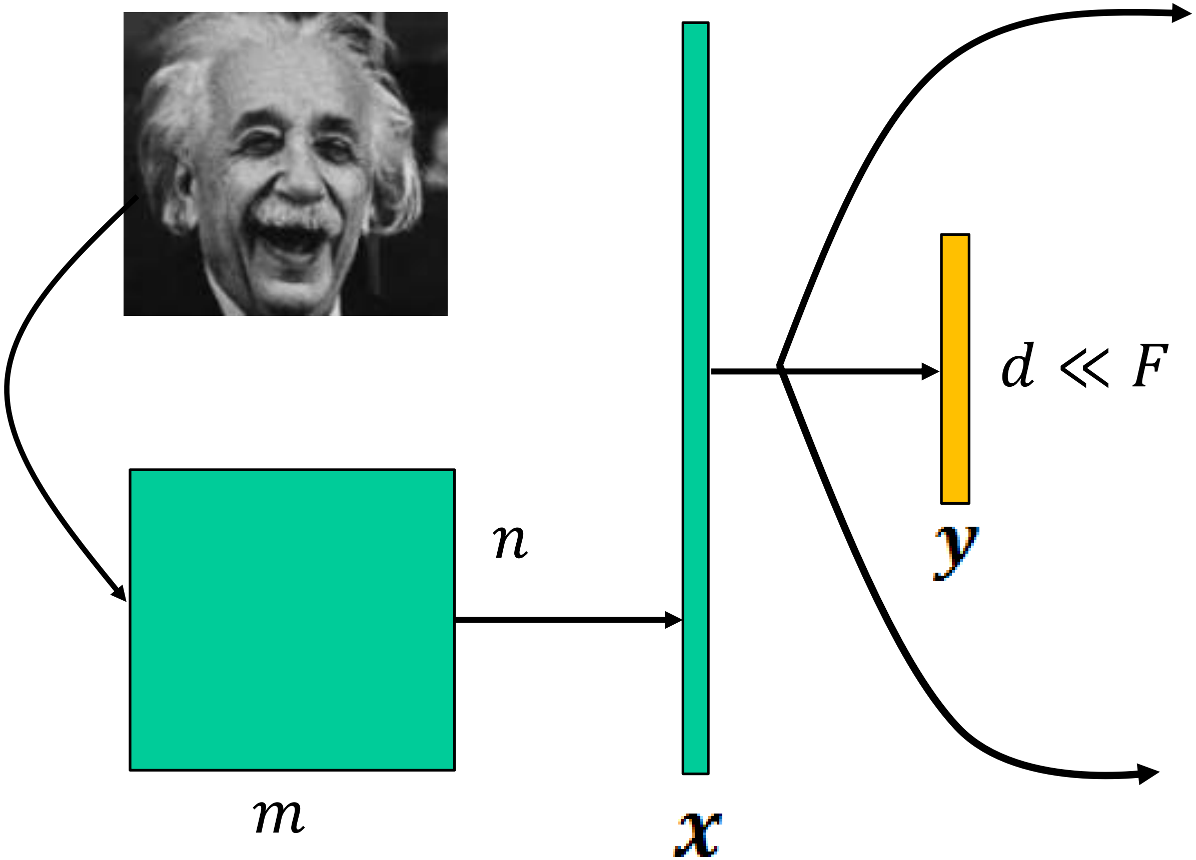
- Pattern Recognition & Machine Learning by C. Bishop Chapter 12
- Turk, Matthew, and Alex Pentland. "Eigenfaces for recognition." *Journal of cognitive neuroscience* 3.1 (1991): 71-86.
- Belhumeur, Peter N., João P. Hespanha, and David J. Kriegman. "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection." *Pattern Analysis and Machine Intelligence*, *IEEE Transactions on* 19.7 (1997): 711-720.
- He, Xiaofei, et al. "Face recognition using laplacianfaces." *Pattern Analysis and Machine Intelligence*, *IEEE Transactions on* 27.3 (2005): 328-340.
- Belkin, Mikhail, and Partha Niyogi. "Laplacian eigenmaps for dimensionality reduction and data representation." *Neural computation* 15.6 (2003): 1373-1396.

Deterministic Component Analysis

Problem: Given a population of data $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\} \in R^F$ (i.e. observations) find a latent space $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\} \in R^d$ (usually d $\ll F$) which is relevant to a task.



Latent Variable Models



Linear:

$$\boldsymbol{y} = \boldsymbol{W}^T\boldsymbol{x}$$

Parameters: $\theta = \{\boldsymbol{W}\}$

Non-linear:

$$\boldsymbol{y}$$

$$F = mn$$

What to have in mind?

- ✓ What are the properties of my latent space?
- ✓ How do I find it (linear, non-linear)?
- ✓ Which is the cost function?
- ✓ How do I solve the problem?

A first example

- Let's assume that we want to find a descriptive latent space, i.e. best describes my population as a whole.
- How do I define it mathematically?
- Idea! This is a statistically machine learning course, isn't?
- Hence, I will try to preserve global statistical properties.

A first example

- What are the data statistics that can used to describe the variability of my observations?

- One such statistic is the variance of the population (how much the data deviate around a mean).

- Attempt: I want to find a low-dimensional latent space where the “majority” of variance is preserved (or in other words maximized).

PCA (one dimension)

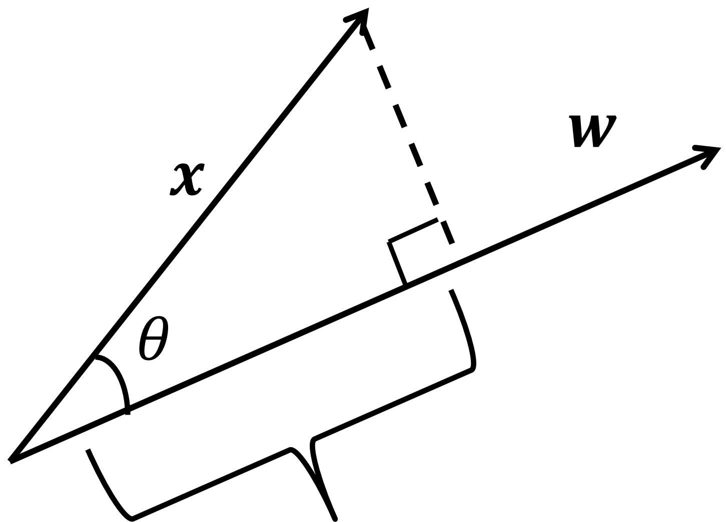
- Variance of the latent space $\{y_1, y_2, \ldots, y_N\}$

$$\sigma_y{}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_y)^2 \qquad \mu_y = \sum_{i=1}^{N} y_i$$

- We want to find $\{y_1{}^o, \ldots y_n{}^o\} = \underset{\{y_1,\ldots,y_n\}}{\mathrm{arg}max}\, \sigma_y{}^2$
- But we are missing something … The way to do it.
- Via a linear projection $\boldsymbol{w}$

i.e., $y_i = \boldsymbol{w}^T \boldsymbol{x}_i$

PCA (geometric interpretation of projection)

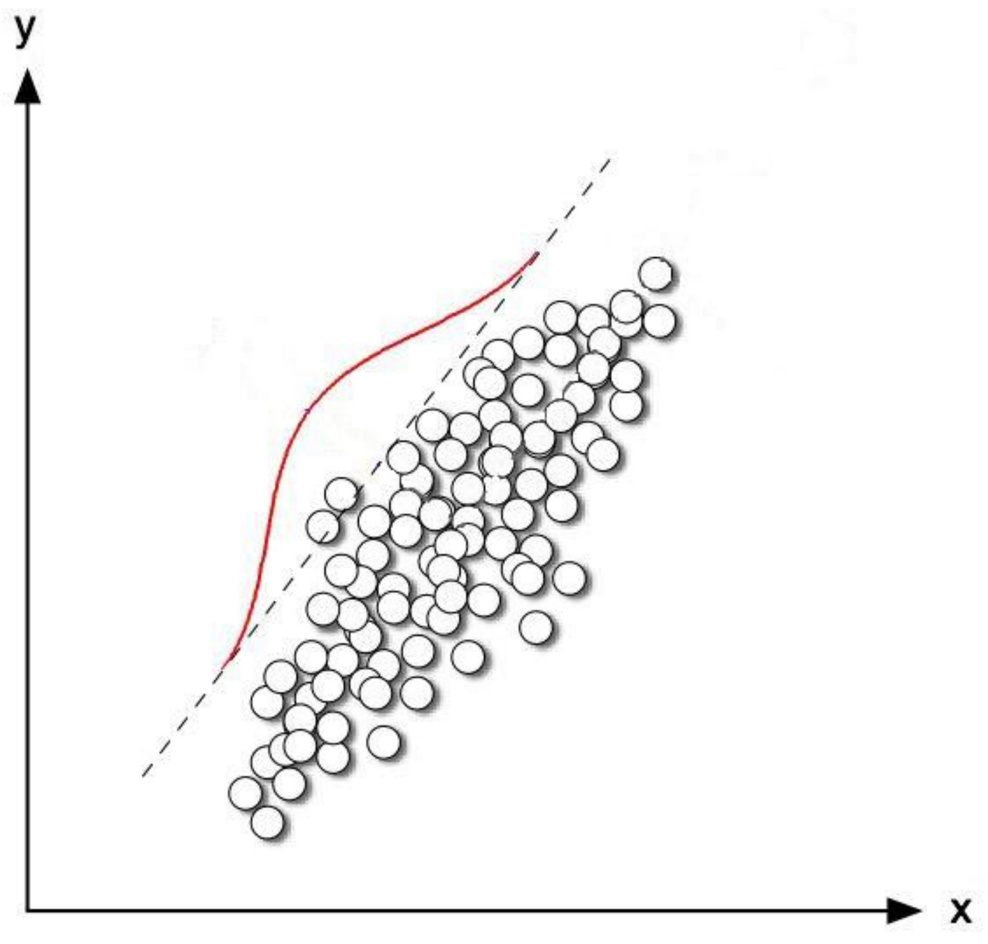


$$\cos(\theta) = \frac{\boldsymbol{x}^T\boldsymbol{w}}{|\boldsymbol{x}||\boldsymbol{w}|}$$

$$|\boldsymbol{x}| = \sqrt{\boldsymbol{x}^T\boldsymbol{x}}$$
$$|\boldsymbol{w}| = \sqrt{\boldsymbol{w}^T\boldsymbol{w}}$$

$$|\boldsymbol{x}|\cos(\theta) = \frac{\boldsymbol{x}^T\boldsymbol{w}}{|\boldsymbol{w}|}$$

PCA (one dimension)



PCA (one dimension)

Assuming $y_i = \boldsymbol{w}^T \boldsymbol{x}_i$ latent space

$$\{y_1, \ldots, y_N\} = \{\boldsymbol{w}^T \boldsymbol{x}_1, \ldots, \boldsymbol{w}^T \boldsymbol{x}_N\}$$

$$\boldsymbol{w}_0 = \underset{\boldsymbol{w}}{\operatorname{argmax}}\, \sigma^2 = \arg\max_{w} \frac{1}{N} \sum (\boldsymbol{w}^T \boldsymbol{x}_i - \boldsymbol{w}^T \boldsymbol{\mu})^2 \quad \boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i$$

$$= \underset{\boldsymbol{w}}{\operatorname{argmax}} \frac{1}{N} \sum (\boldsymbol{w}^T (\boldsymbol{x}_i - \boldsymbol{\mu}))^2$$

$$= \underset{\boldsymbol{w}}{\operatorname{argmax}} \frac{1}{N} \sum \boldsymbol{w}^T (\boldsymbol{x}_i - \boldsymbol{\mu})\, (\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{w}$$

$$= \underset{\boldsymbol{w}}{\operatorname{argmax}} \frac{1}{N} \boldsymbol{w}^T \left(\sum (\boldsymbol{x}_i - \boldsymbol{\mu})\, (\boldsymbol{x}_i - \boldsymbol{\mu})^T \right) \boldsymbol{w}$$

$$= \underset{\boldsymbol{w}}{\operatorname{argmax}}\, \boldsymbol{w}^T \mathbf{S}_t \boldsymbol{w}$$

PCA (one dimension)

$$\boldsymbol{w}_o = \underset{\boldsymbol{w}}{\operatorname{argmax}}\, \sigma^2 = \underset{\boldsymbol{w}}{\operatorname{argmax}} \frac{1}{N} \boldsymbol{w}^T \mathbf{S}_t \boldsymbol{w} \geq 0$$

$$\boldsymbol{S}_t = \frac{1}{N} \sum (\boldsymbol{x}_i - \boldsymbol{\mu})\,(\boldsymbol{x}_i - \boldsymbol{\mu})^T$$

- There is a trivial solution of $\boldsymbol{w} = \infty$
- We can avoid it by adding extra constraints (a fixed magnitude on $\boldsymbol{w}$ ($||\boldsymbol{w}||^2 = \boldsymbol{w}^T \mathbf{w}=1$)

$$\boldsymbol{w}_0 = \arg\max_{\boldsymbol{w}} \boldsymbol{w}^T \boldsymbol{S}_t \boldsymbol{w}$$

$$\text{subject to (s.t.)} \quad \boldsymbol{w}^T \boldsymbol{w} = 1$$

PCA (one dimension)

Formulate the Lagrangian

$$L(\boldsymbol{w}, \lambda) = \boldsymbol{w}^T \boldsymbol{S}_t \boldsymbol{w} - \lambda(\boldsymbol{w}^T \boldsymbol{w} - 1)$$

$$\frac{\partial \boldsymbol{w}^T \boldsymbol{S}_t \boldsymbol{w}}{\partial \boldsymbol{w}} = 2\boldsymbol{S}_t \boldsymbol{w} \qquad \frac{\partial \boldsymbol{w}^T \boldsymbol{w}}{\partial \boldsymbol{w}} = 2\boldsymbol{w}$$

$$\frac{\partial L}{\partial \boldsymbol{w}} = \mathbf{0} \quad \Rightarrow \quad \boxed{\boldsymbol{S}_t \boldsymbol{w} = \lambda \boldsymbol{w}}$$

$\boldsymbol{w}$ is the largest eigenvector of $\boldsymbol{S}_t$

PCA Properties of $\boldsymbol{S}_t$

$$\boldsymbol{S}_t = \frac{1}{N}\sum (\boldsymbol{x}_i - \boldsymbol{\mu})\ (\boldsymbol{x}_i - \boldsymbol{\mu})^T = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^T$$

$$\boldsymbol{X} = [\boldsymbol{x}_1 - \boldsymbol{\mu}, \dots, \boldsymbol{x}_N - \boldsymbol{\mu}]$$

$\boldsymbol{S}_t$ is a symmetric matrix ⇒ all eigenvalues are real

$\boldsymbol{S}_t$ is a positive semi-definite matrix, i.e.

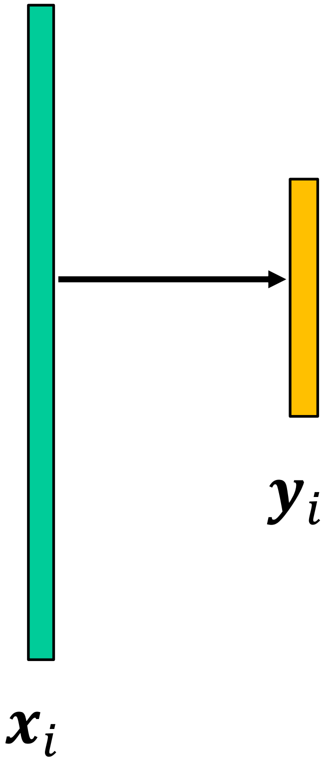
$\forall \boldsymbol{w} \neq \boldsymbol{0} \qquad \boldsymbol{w}^T\boldsymbol{S}_t\boldsymbol{w} \geq 0$ (all eigenvalues are non negative)

$$\mathrm{rank}(\boldsymbol{S}_t) = \min(N-1, F)$$

$$\boldsymbol{S}_t = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T \qquad \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}, \boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$$

PCA (more dimensions)

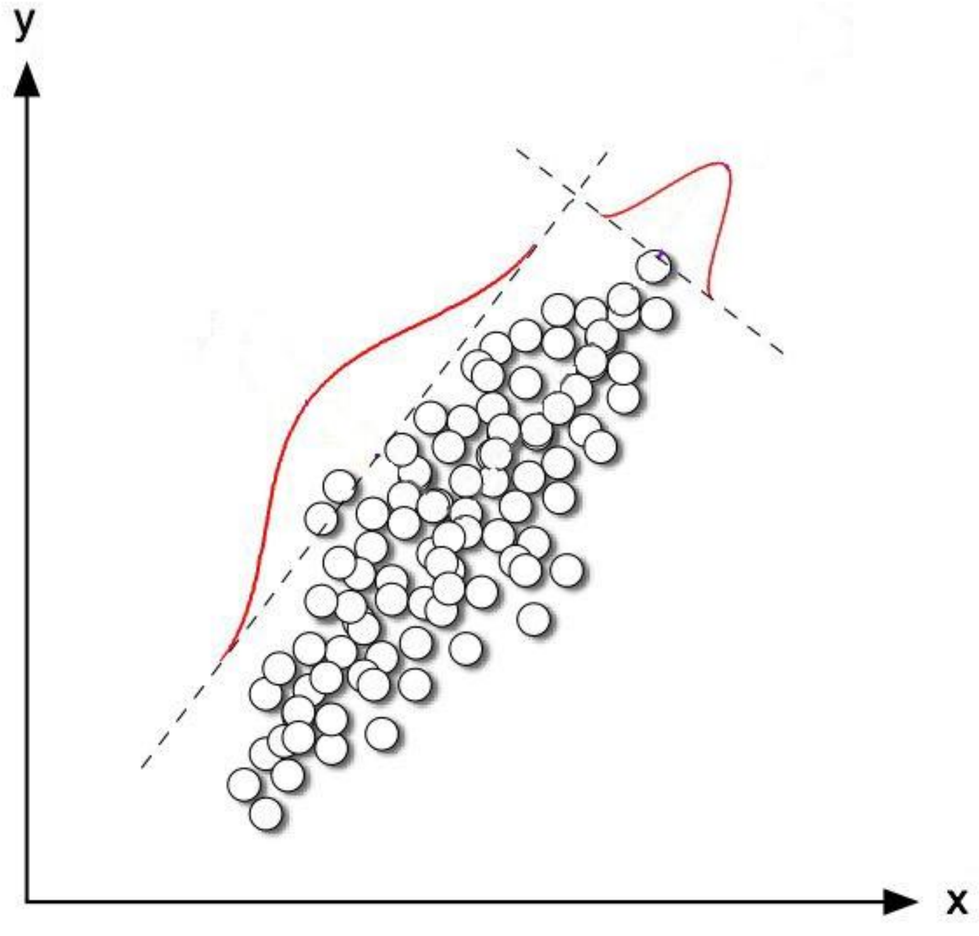
- How can we find a latent space with more than one dimensions?
- We need to find a set of projections $\{\boldsymbol{w}_1, \dots, \boldsymbol{w}_d\}$



$$\boldsymbol{y}_i = \begin{bmatrix} y_{i1} \\ \dots \\ y_{id} \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}_1{}^T \boldsymbol{x}_i \\ \dots \\ \boldsymbol{w}_d{}^T \boldsymbol{x}_i \end{bmatrix} = \boldsymbol{W}^T \boldsymbol{x}_i$$

$$\boldsymbol{W} = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_d]$$

PCA (more dimensions)



PCA (more dimensions)

Maximize the variance in each dimension

$$\boldsymbol{W}_o = \arg\max_{\boldsymbol{W}} \frac{1}{N} \sum_{k=1}^{d} \sum_{i=1}^{N} (y_{ik} - \mu_{ik})^2$$

$$= \arg\max_{\boldsymbol{W}} \frac{1}{N} \sum_{k=1}^{d} \sum_{i=1}^{N} \boldsymbol{w}_k{}^T (\boldsymbol{x}_i - \boldsymbol{\mu}_i)(\boldsymbol{x}_i - \boldsymbol{\mu}_i)^T \boldsymbol{w}_k$$

$$= \arg\max_{\boldsymbol{W}} \sum_{k=1}^{d} \boldsymbol{w}_k{}^T \boldsymbol{S}_t \boldsymbol{w}_k = \arg\max_{\boldsymbol{W}} \mathrm{tr}[\boldsymbol{W}^T \boldsymbol{S}_t \boldsymbol{W}]$$

PCA (more dimensions)

$$\boldsymbol{W}_o = \arg\max_{\boldsymbol{W}} \text{tr}[\boldsymbol{W}^T \boldsymbol{S}_t \boldsymbol{W}]$$

s.t. $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}$

Lagrangian $L(\mathbf{W}, \boldsymbol{\Lambda}) = \text{tr}[\mathbf{W}^T \boldsymbol{S}_t \mathbf{W}]\text{-tr}[\boldsymbol{\Lambda}(\mathbf{W}^T\mathbf{W} - \mathbf{I})]$

$$\frac{\partial \text{tr}[\mathbf{W}^T \boldsymbol{S}_t \mathbf{W}]}{\partial \boldsymbol{W}} = 2\boldsymbol{S}_t\boldsymbol{W} \qquad \frac{\partial \text{tr}[\boldsymbol{\Lambda}(\mathbf{W}^T\mathbf{W} - \mathbf{I})]}{\partial \boldsymbol{W}} = 2\boldsymbol{W}\boldsymbol{\Lambda}$$

$$\frac{\partial \text{L}(\boldsymbol{W}, \boldsymbol{\Lambda})}{\partial \boldsymbol{W}} = \mathbf{0} \qquad \boldsymbol{S}_t\boldsymbol{W} = \boldsymbol{W}\boldsymbol{\Lambda}$$

Does it ring a bell?

PCA (more dimensions)

- Hence, $\boldsymbol{W}$ has as columns the d eigenvectors of $\boldsymbol{S}_t$ that correspond to its d largest nonzero eigenvalues

$$\boldsymbol{W} = \boldsymbol{U_d}$$

$$\mathrm{tr}[\boldsymbol{W}^T \boldsymbol{S}_t \boldsymbol{W}] = \mathrm{tr}[\boldsymbol{W}^T \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\boldsymbol{T}} \boldsymbol{W}] = \mathrm{tr}[\boldsymbol{\Lambda}_d]$$

Example: $\boldsymbol{U}$ be 5x5 and $\boldsymbol{W}$ be a 5x3

$$\boldsymbol{W}^T\boldsymbol{U} = \begin{bmatrix} u_1.u_1 & u_1.u_2 & u_1.u_3 & u_1.u_4 & u_1.u_5 \\ u_2.u_1 & u_2.u_2 & u_2.u_3 & u_2.u_4 & u_2.u_5 \\ u_3.u_1 & u_3.u_2 & u_3.u_3 & u_3.u_4 & u_3.u_5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

PCA (more dimensions)

$$\boldsymbol{W}^T\boldsymbol{U}\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and

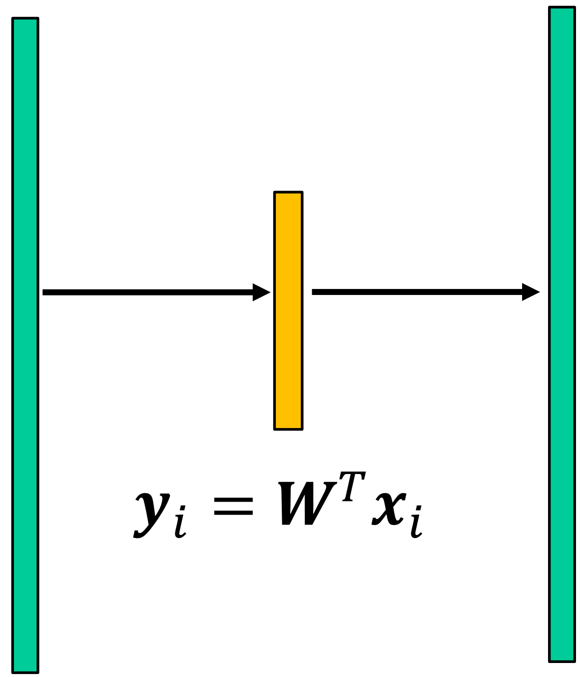
$$\boldsymbol{W}^T\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{W} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}$$

Hence the maximum is

$$\mathrm{tr}[\boldsymbol{\Lambda}_d] = \sum_{i=1}^{d} \lambda_d$$

PCA (another perspective)

- We want to find a set of bases $\boldsymbol{W}$ that best reconstructs the data after projection



PCA (another perspective)

Let us assume for simplicity centred data (zero mean)

- Reconstructed data $\tilde{\boldsymbol{X}} = \boldsymbol{W}\boldsymbol{Y} = \boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X}$

$$\boldsymbol{W}_{\boldsymbol{0}} = \arg\min_{\boldsymbol{W}} \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{F}(x_{ij} - \widetilde{x_{ij}})^2$$

$$= \arg\min_{\boldsymbol{W}} \left|\left|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{X}\right|\right|^2_F \ (1)$$

$$\text{s.t.}\, \boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{W} = \boldsymbol{I}$$

PCA (another perspective)

$$\left\| \boldsymbol{X} - \boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{X} \right\|_{\boldsymbol{F}}^{\boldsymbol{2}}$$

$$= \operatorname{tr}\left[\left(\mathbf{X} - \mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{X}\right)^{T}\left(\mathbf{X} - \mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{X}\right)\right]$$

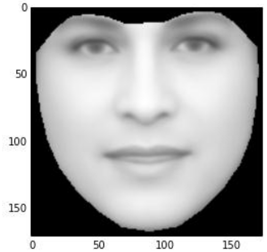
$$= \operatorname{tr}[\mathbf{X}^{\mathrm{T}}\mathbf{X} - \mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{T}\mathbf{X} - \mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{T}\mathbf{X} + \mathbf{X}^{\mathrm{T}}\mathbf{W}\underbrace{\mathbf{W}^{T}\mathbf{W}}_{\mathbf{I}}\mathbf{W}^{T}\mathbf{X}]$$

$$= \operatorname{tr}[\underbrace{\mathbf{X}^{\mathrm{T}}\mathbf{X}}_{\text{constant}}] - \operatorname{tr}[\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{T}\mathbf{X}]$$

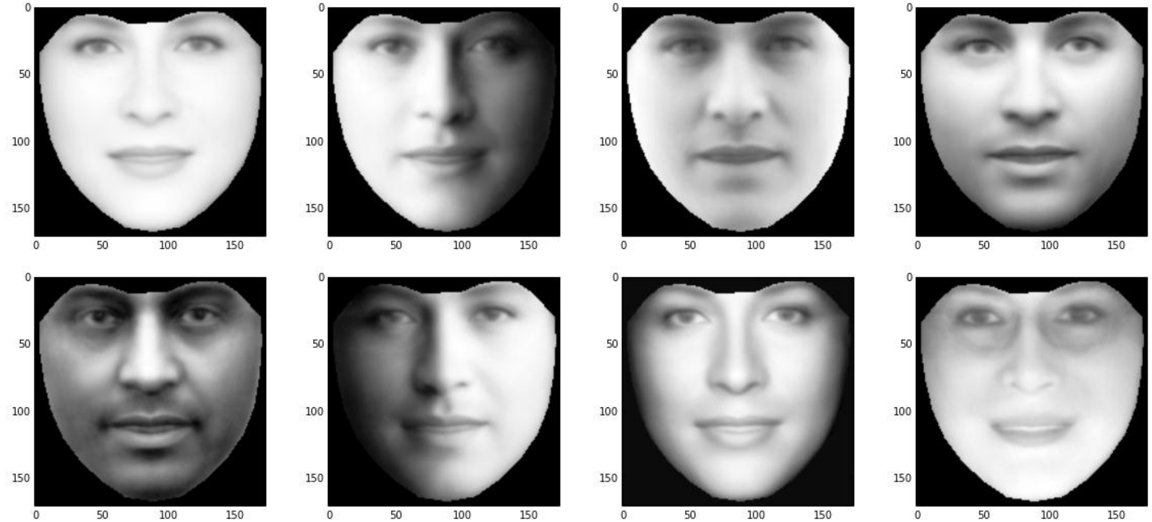
$$\min(1) \Rightarrow \quad \max_{W} \operatorname{tr}[\mathbf{W}^{T}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}]$$

PCA (applications)

- TEXTURE



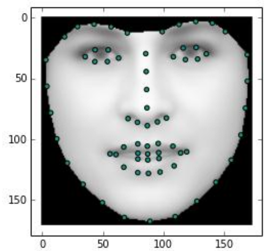
- Mean



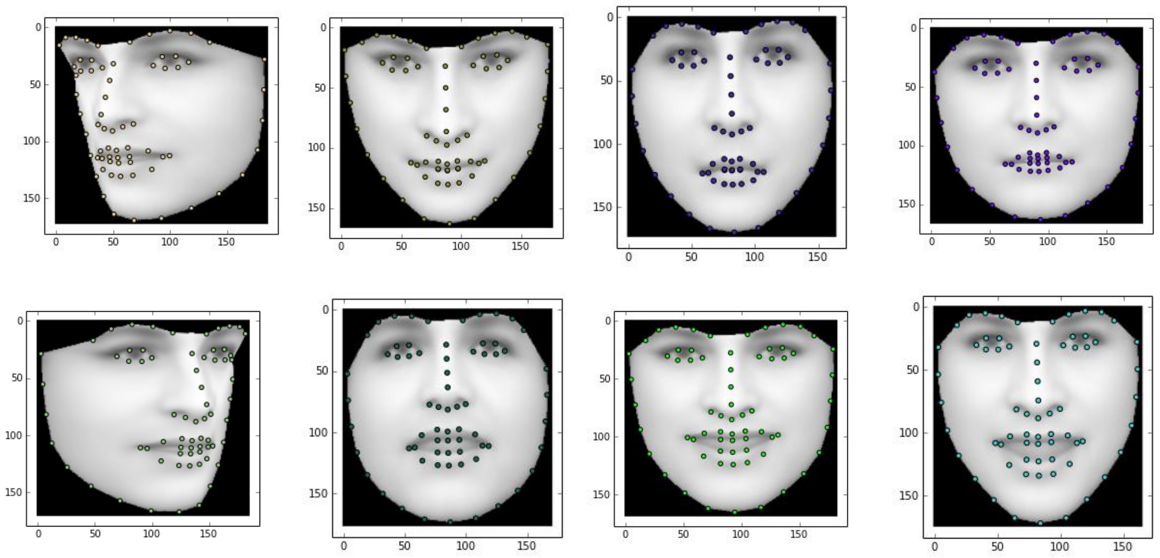
- 1st PC
- 2nd PC
- 3rd PC
- 4th PC

PCA (applications)

- SHAPE

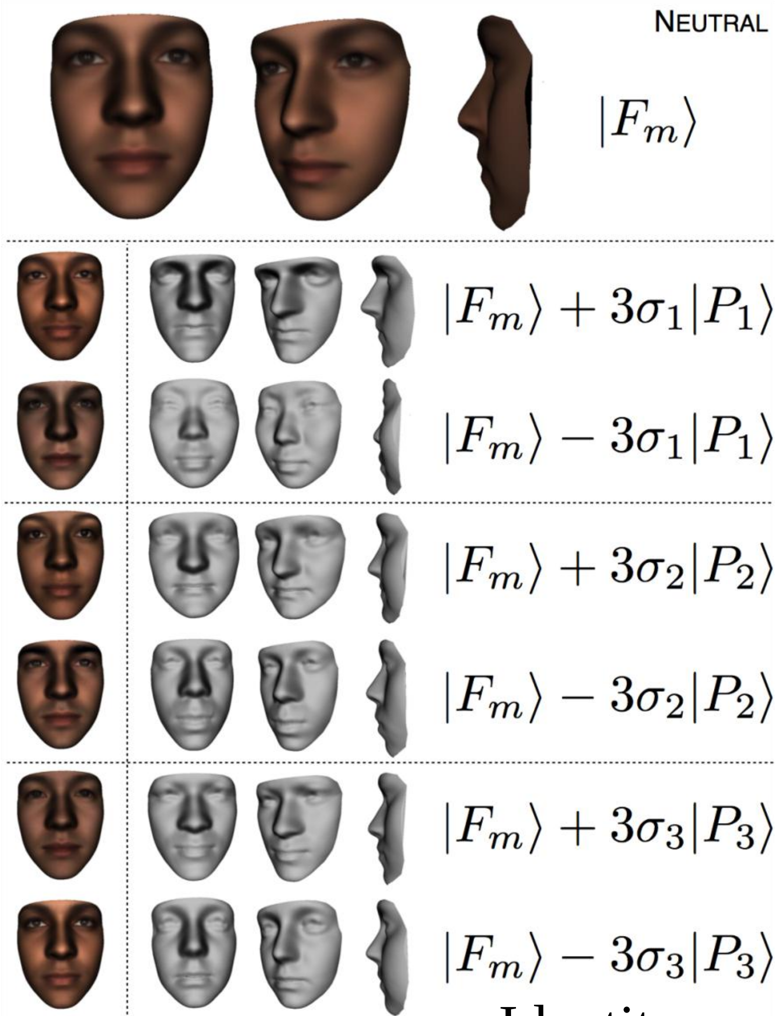


- Mean

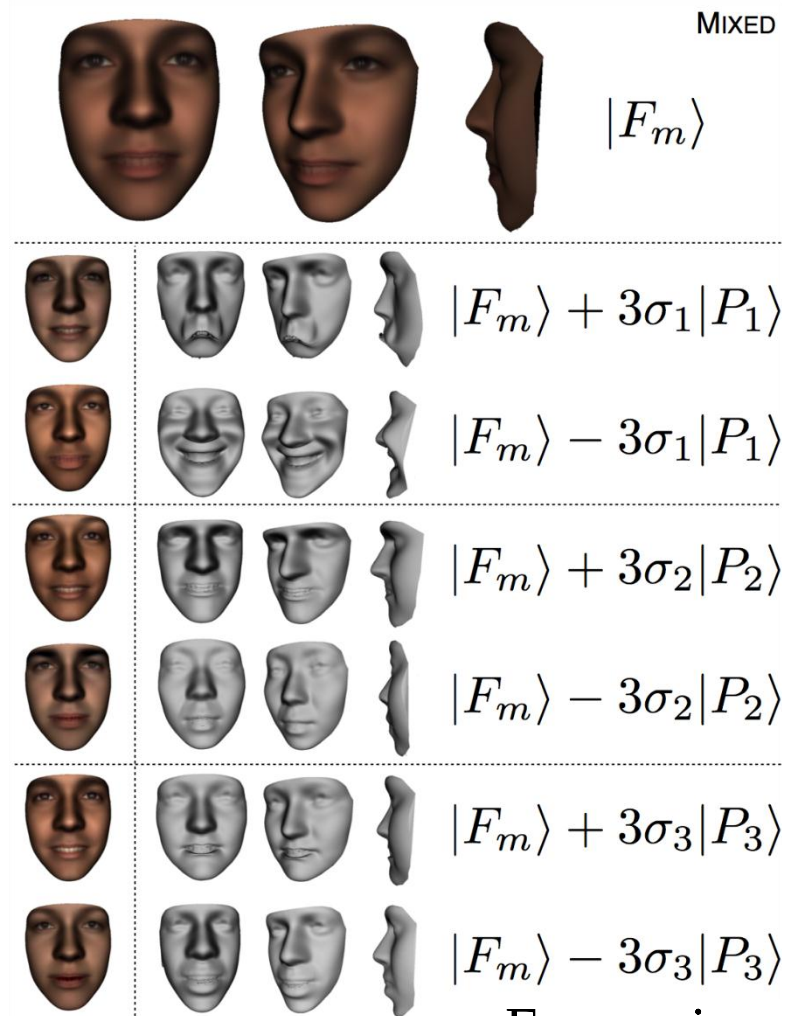


- 1st PC
- 2nd PC
- 3rd PC
- 4th PC

PCA (applications)



Identity



Expression

Linear Discriminant Analysis

- PCA: Unsupervised approach good for compression of data and data reconstruction. Good statistical prior.

- PCA: Not explicitly defined for classification problems (i.e., in case that data come with labels)

- How do we define a latent space it this case? (i.e., that helps in data classification)

Linear Discriminant Analysis

- We need to properly define statistical properties which may help us in classification.

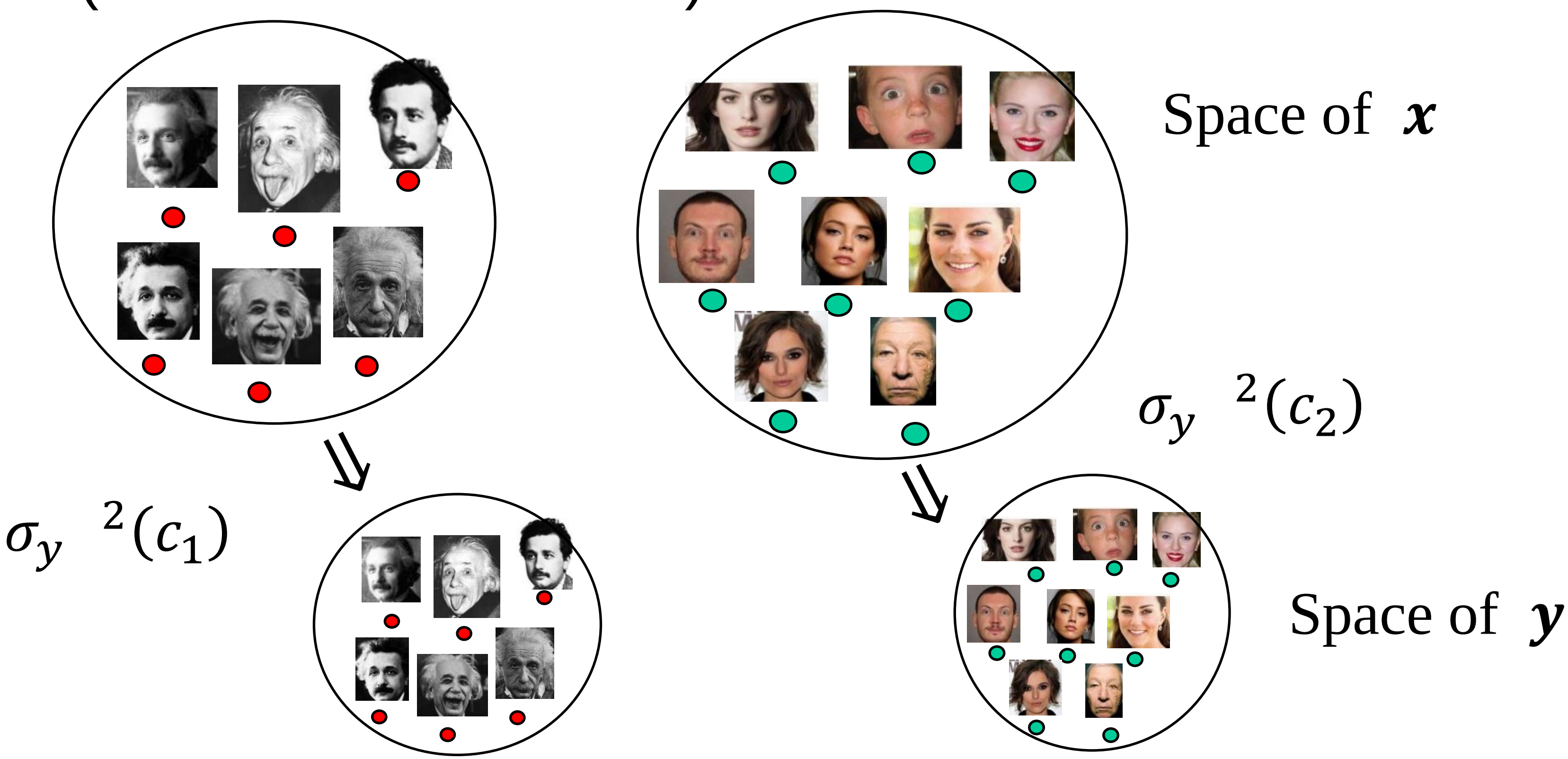
- Intuition: We want to find a space in which

(a) the data consisting each class look more like each other, while

(b) the data of separate classes look more dissimilar.

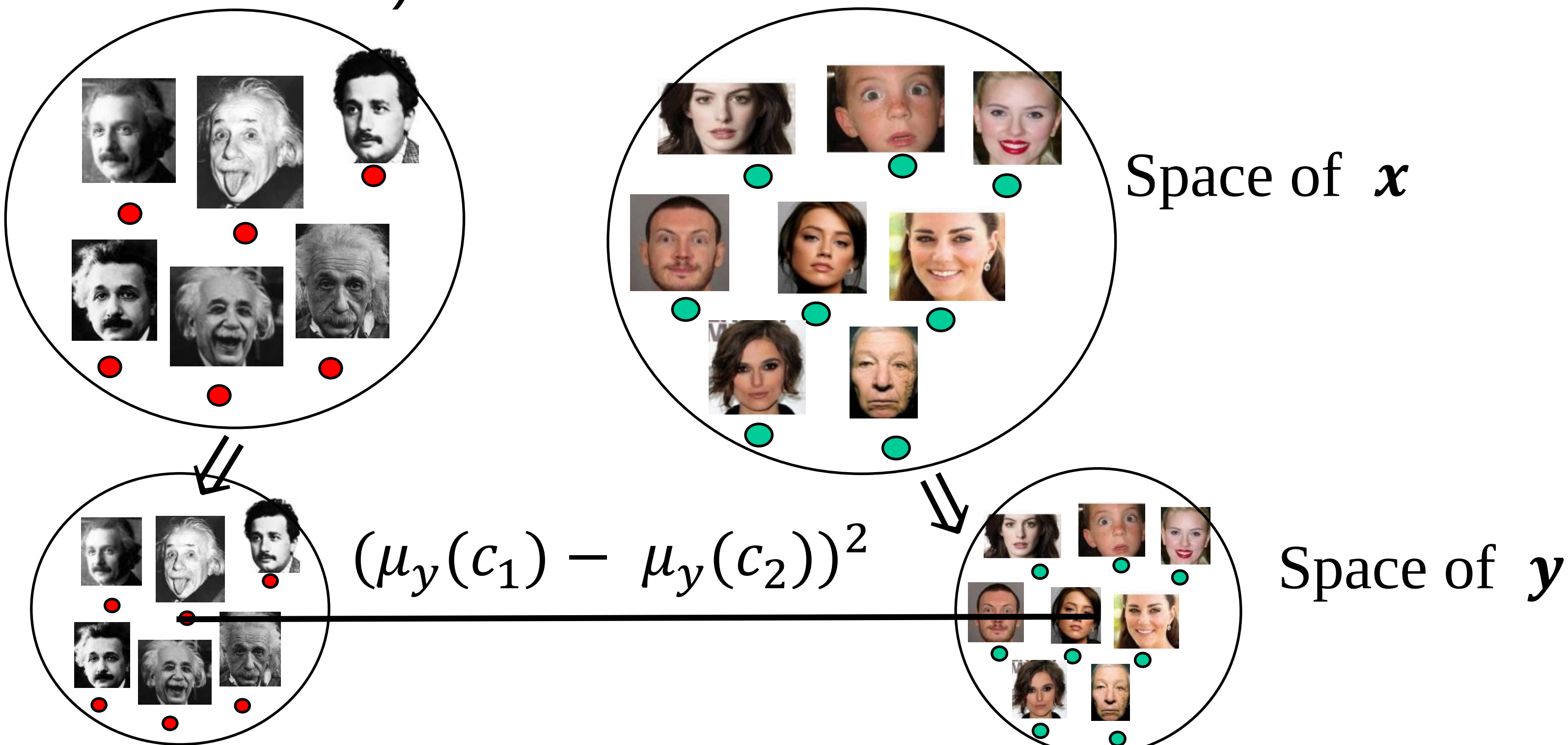
Linear Discriminant Analysis

- How do I make my data in each class look more similar? Minimize the variability in each class (minimize the variance)



Linear Discriminant Analysis

- How do I make the data between classes look dissimilar? I move the data from different classes further away from each other (increase the distance between their means).



Linear Discriminant Analysis

A bit more formally. I want a latent space y such that:

$\sigma_y{}^2(c_1) + \sigma_y{}^2(c_2)$ is minimum

$(\mu_y(c_1) - \mu_y(c_2))^2$ is maximum

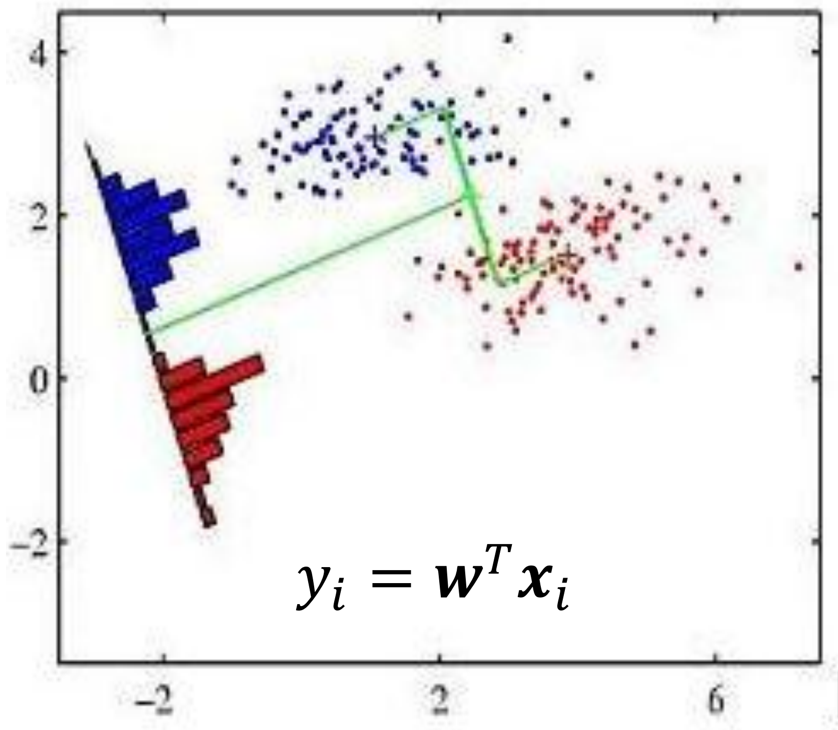
How do I combine them together?

$$\text{minimize } \frac{\sigma_y{}^2(c_1) + \sigma_y{}^2(c_2)}{(\mu_y(c_1) - \mu_y(c_2))^2}$$

$$\text{Or maximize } \frac{(\mu_y(c_1) - \mu_y(c_2))^2}{\sigma_y{}^2(c_1) + \sigma_y{}^2(c_2)}$$

Linear Discriminant Analysis

How can I find my latent space?



$$y_i = \boldsymbol{w}^T \boldsymbol{x}_i$$

Linear Discriminant Analysis

$$\sigma_y{}^2(c_1) = \frac{1}{N_{c_1}} \sum_{x_i \in c_1} (y_i - \mu_y(c_1))^2$$

$$\boldsymbol{\mu}(c_1) = \frac{1}{N_{c_1}} \sum_{x_i \in c_1} \boldsymbol{x}_i$$

$$= \frac{1}{N_{c_1}} \sum_{x_i \in c_1} (\boldsymbol{w}^T (\boldsymbol{x}_i - \boldsymbol{\mu}(c_1)))^2$$

$$= \frac{1}{N_{c_1}} \sum_{x_i \in c_1} \boldsymbol{w}^T (\boldsymbol{x}_i - \boldsymbol{\mu}(c_1))(\boldsymbol{x}_i - \boldsymbol{\mu}(c_1))^T \boldsymbol{w}$$

$$= \boldsymbol{w}^T \frac{1}{N_{c_1}} \sum_{x_i \in c_1} (\boldsymbol{x}_i - \boldsymbol{\mu}(c_1))(\boldsymbol{x}_i - \boldsymbol{\mu}(c_1))^T \boldsymbol{w}$$

$$= \boldsymbol{w}^T \boldsymbol{S}_1 \boldsymbol{w}$$

$$\sigma_y{}^2(c_2) = \boldsymbol{w}^T \boldsymbol{S}_2 \boldsymbol{w}$$

Linear Discriminant Analysis

$$\sigma_y{}^2(c_1) + \sigma_y{}^2(c_2) = \boldsymbol{w}^T(\underbrace{\boldsymbol{S}_1 + \boldsymbol{S}_2})\boldsymbol{w}$$

$\boldsymbol{S}_w$ within class scatter matrix

$$(\mu_y(c_1) - \mu_y(c_2))^2$$
$$= \boldsymbol{w}^T\underbrace{(\boldsymbol{\mu}(c_1) - \boldsymbol{\mu}(c_2))(\boldsymbol{\mu}(c_1) - \boldsymbol{\mu}(c_2))^T}\boldsymbol{w}$$

$\boldsymbol{S}_b$ between class scatter matrix

$$\frac{(\mu_y(c_1) - \mu_y(c_2))^2}{\sigma_y{}^2(c_1) + \sigma_y{}^2(c_2)} = \frac{\boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w}}$$

Linear Discriminant Analysis

$\max \boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w}$ s.t. $\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}=1$

Lagrangian: $L(\boldsymbol{w}, \lambda) = \boldsymbol{w}^T \boldsymbol{S}_b \mathbf{w} - \lambda(\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}\text{-}1)$

$$\frac{\partial \boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}}{\partial \boldsymbol{w}} = 2\boldsymbol{S}_w \boldsymbol{w} \qquad \frac{\partial \boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w}}{\partial \boldsymbol{w}} = 2\boldsymbol{S}_b \boldsymbol{w}$$

$$\frac{\partial L}{\partial \boldsymbol{w}} = \mathbf{0} \quad \Rightarrow \quad \boxed{\lambda \boldsymbol{S}_w \boldsymbol{w} = \boldsymbol{S}_b \boldsymbol{w}}$$

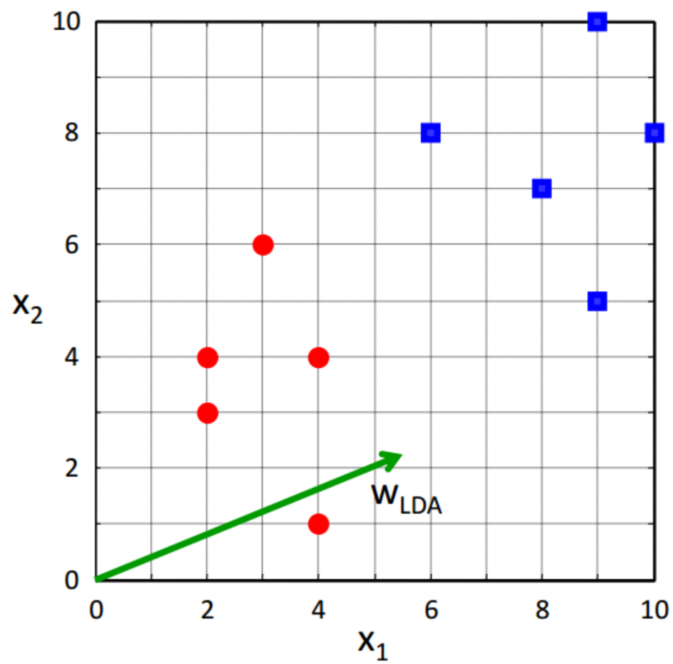
$\boldsymbol{w}$ is the largest eigenvector of $\boldsymbol{S}_w{}^{-1} \boldsymbol{S}_b$

$$\mathbf{w} \propto \boldsymbol{S}_w{}^{-1}(\boldsymbol{\mu}(c_1) - \boldsymbol{\mu}(c_2))$$

Linear Discriminant Analysis

Compute the LDA projection for the following 2D dataset

$$c_1 = \{(4,1), (2,4), (2,3), (3,6), (4,4)\}$$
$$c_2 = \{(9,10), (6,8), (9,5), (8,7), (10,8)\}$$



Linear Discriminant Analysis

Solution (by hand)

•The class statistics are

$$\boldsymbol{S}_1 = \begin{bmatrix} 0.8 & -0.4 \\ -0.4 & 2.64 \end{bmatrix} \quad \boldsymbol{S}_2 = \begin{bmatrix} 1.84 & -0.04 \\ -0.04 & 2.64 \end{bmatrix}$$

•The within and between class scatter are

$$\boldsymbol{\mu}_1 = [3.0\ 3.6]^T \qquad \boldsymbol{\mu}_2 = [8.4\ 7.6]^T$$

$$\boldsymbol{S}_b = \begin{bmatrix} 29.16 & 21.6 \\ 21.6 & 16.0 \end{bmatrix} \quad \boldsymbol{S}_w = \begin{bmatrix} 2.64 & -0.44 \\ -0.44 & 5.28 \end{bmatrix}$$

Linear Discriminant Analysis

The LDA projection is then obtained as the solution of the generalized eigenvalue problem

$$\boldsymbol{S}_w{}^{-1}\boldsymbol{S}_b\boldsymbol{w} = \lambda\boldsymbol{w} \rightarrow \left|\boldsymbol{S}_w{}^{-1}\boldsymbol{S}_b - \lambda\boldsymbol{I}\right| = 0 \rightarrow$$

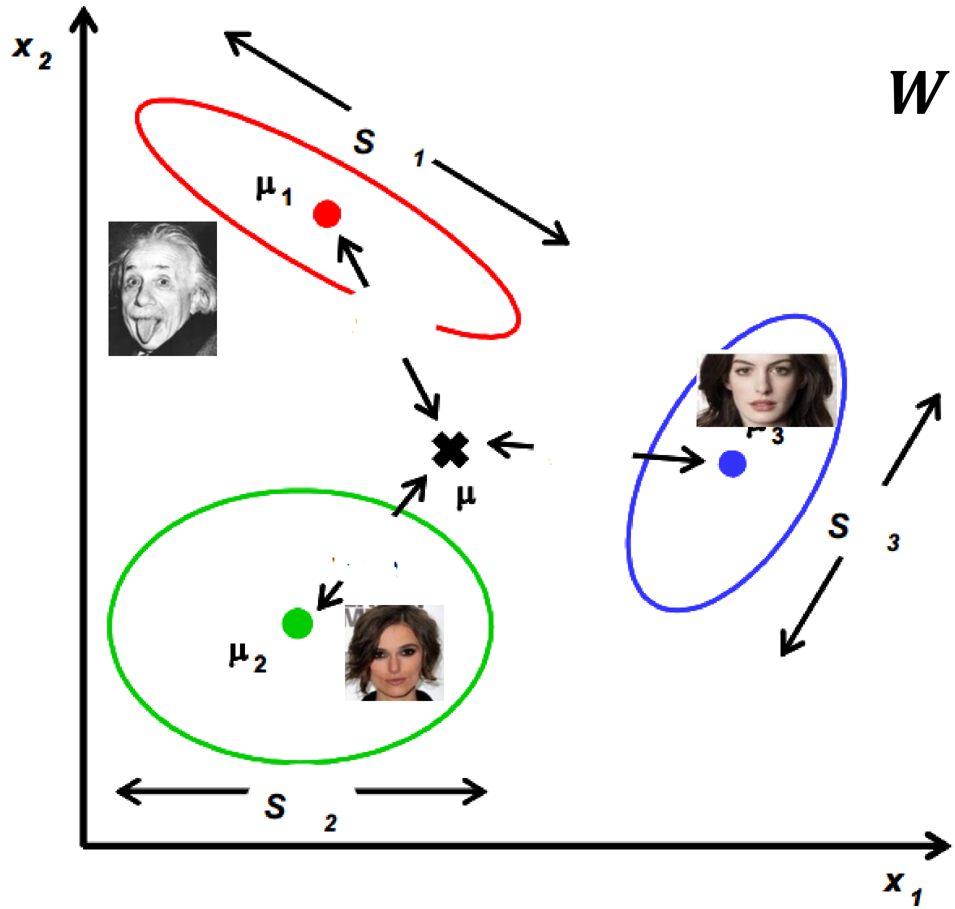
$$\begin{vmatrix} 11.89 - \lambda & 8.81 \\ 5.08 & 3.76 - \lambda \end{vmatrix} = 0 \rightarrow \quad \lambda = 15.65$$

$$\begin{bmatrix} 11.89 & 8.81 \\ 5.08 & 3.76 \end{bmatrix}\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = 15.65\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \rightarrow \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 0.91 \\ 0.39 \end{bmatrix}$$

Or directly by

$$\boldsymbol{w}^* = \boldsymbol{S}_W{}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = [-0.91 \ \ -0.39]^T$$

LDA (Multiclass & Multidimensional case)



$$\boldsymbol{W} = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_d]$$

LDA (Multiclass & Multidimensional case)

Within-class scatter matrix

$$\boldsymbol{S}_w = \sum_{j=1}^{C} \boldsymbol{S}_j = \sum_{j=1}^{C} \frac{1}{N_{c_j}} \sum_{x_i \in c_j} (\boldsymbol{x}_i - \boldsymbol{\mu}(c_j))(\boldsymbol{x}_i - \boldsymbol{\mu}(c_j))^T$$

Between-class scatter matrix

$$\boldsymbol{S}_b = \sum_{j=1}^{c} (\boldsymbol{\mu}(c_j) - \boldsymbol{\mu})\,(\boldsymbol{\mu}(c_j) - \boldsymbol{\mu})^{\mathrm{T}}$$

LDA (Multiclass & Multidimensional case)

$$\max \operatorname{tr}[\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{W}] \text{ s.t. } \boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W}=\mathbf{I}$$

Lagranging: $L(\boldsymbol{W},\boldsymbol{\Lambda}) = tr[\mathbf{w}^T\boldsymbol{S}_b\mathbf{w}] - tr[\boldsymbol{\Lambda}(\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W} - \boldsymbol{I})]$

$$\frac{\partial \operatorname{tr}[\mathbf{W}^T\boldsymbol{S}_b\mathbf{W}]}{\partial \boldsymbol{W}} = 2\boldsymbol{S}_b\boldsymbol{W} \qquad \frac{\partial \operatorname{tr}[\boldsymbol{\Lambda}(\mathbf{W}^T\boldsymbol{S}_w\mathbf{W} - \mathbf{I})]}{\partial \boldsymbol{W}} = 2\boldsymbol{S}_w\boldsymbol{W}\boldsymbol{\Lambda}$$

$$\frac{\partial \mathrm{L}(\boldsymbol{W},\boldsymbol{\Lambda})}{\partial \boldsymbol{W}} = \mathbf{0} \qquad \Rightarrow \qquad \boldsymbol{S}_b\boldsymbol{W} = \boldsymbol{S}_w\boldsymbol{W}\boldsymbol{\Lambda}$$

the eigenvectors of $\boldsymbol{S}_w{}^{-1}\boldsymbol{S}_b$ that correspond to the largest eigenvalues