# Chapter 8 Principal Component Analysis & Linear Discriminant Analysis

## Curse of Dimensionality (维数灾难)

The curse of dimensionality refers to the phenomena that occur when classifying, organizing, and analyzing high dimensional data that **does not occur in low dimensional spaces**

- Computational Complexity
- Overfitting
  - $\#paramet \gg \#examples$ which leads to un reliable parameter estimation

## Principal Component Analysis (PCA 主成分分析)

### Definition

- Goal: Find linear projections with good representation ability
- Input: A set of $n$ d-dimensional samples $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ $\left(\mathbf{x}_k \in \mathbf{R}^d\right)$
- Output: Orthonormal (标准正交) projection bases $\{\mathrm{e}_1, \mathrm{e}_2, \ldots, \mathrm{e}_{d'}\}$ $(d' \leq d)$ which can best represent the (centered) samples

$$
\begin{array}{ll}
\min_{\mathbf{e}_1,\ldots,\mathbf{e}_{d'}} & J_{d'} \\
\text{s.t.} & \mathbf{e}_i^t \mathbf{e}_i = 1 \, (1 \leq i \leq d') \\
 & \mathbf{e}_i^t \mathbf{e}_j = 0 (i \neq j)
\end{array}
$$

where

- sample mean

$$
\mathbf{m} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k
$$

- projection of centered $\mathbf{x}_k$ on $\mathbf{e}_i$

$$
a_{ki} = \mathbf{e}_i^t \left(\mathbf{x}_k - \mathbf{m}\right)
$$

- PCA criterion function

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

so we have

$$
\begin{aligned}
J_{d'} &= \sum_{k=1}^{n} \left\| \left( \sum_{i=1}^{d'} a_{ki} \mathbf{e}_i \right) - (\mathbf{x}_k - \mathbf{m}) \right\|^2 \\
&= \sum_{k=1}^{n} \left[ \left( \sum_{i=1}^{d'} a_{ki} \mathbf{e}_i \right)^t \left( \sum_{i=1}^{d'} a_{ki} \mathbf{e}_i \right) - 2 \left( \sum_{i=1}^{d'} a_{ki} \mathbf{e}_i \right)^t (\mathbf{x}_k - \mathbf{m}) + \|\mathbf{x}_k - \mathbf{m}\|^2 \right] \\
&= \sum_{k=1}^{n} \left[ \sum_{i=1}^{d'} a_{ki}^2 \|\mathbf{e}_i\|^2 - 2 \sum_{i=1}^{d'} a_{ki} \mathbf{e}_i^t (\mathbf{x}_k - \mathbf{m}) + \|\mathbf{x}_k - \mathbf{m}\|^2 \right] \\
&= \sum_{k=1}^{n} \left[ - \sum_{i=1}^{d'} a_{ki}^2 + \|\mathbf{x}_k - \mathbf{m}\|^2 \right] \\
&= - \sum_{i=1}^{d'} \sum_{k=1}^{n} \mathbf{e}_i^t (\mathbf{x}_k - \mathbf{m}) (\mathbf{x}_k - \mathbf{m})^t \mathbf{e}_i + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2 \\
&= - \sum_{i=1}^{d'} \mathbf{e}_i^t \mathbf{S} \mathbf{e}_i + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2
\end{aligned}
$$

where

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

is a constant which can be ignored.

and where $\mathbf{S}$ is **symmetric** and **positive semi-definite**

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

so that

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

Using Lagrange function we have

$$
\mathbf{S} \mathbf{e}_i = \lambda_i \mathbf{e}_i
$$

where

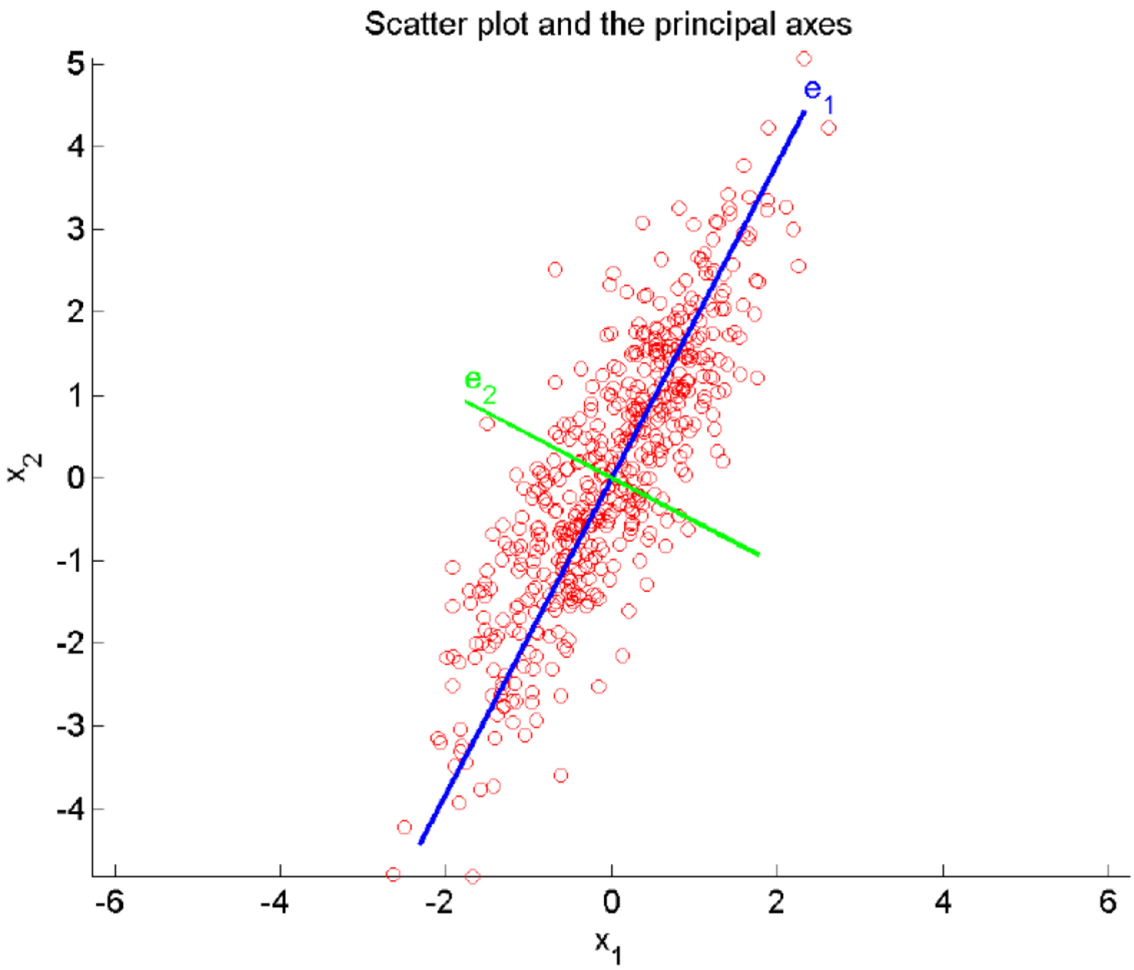
- $\lambda_i \geq 0$ : eigenvalue of $\mathbf{S}$
- $\mathbf{e}_i$ : unit-norm eigenvector of $\mathbf{S}$ w.r.t. $\lambda_i$

Finally

$$
J_{d'} = - \sum_{i=1}^{d'} \mathbf{e}_i^t \mathbf{S} \mathbf{e}_i = - \sum_{i=1}^{d'} \lambda_i \|\mathbf{e}_i\|^2 = - \sum_{i=1}^{d'} \lambda_i
$$

**Algorithm**

1. Set $\mathbf{m} = \frac{1}{n} \sum_{k=1}^{n} \mathrm{x}_k$ and $\mathbf{S} = \sum_{k=1}^{n} (\mathrm{x}_k - \mathrm{m}) (\mathrm{x}_k - \mathrm{m})^t$
2. Identify top $d'$ eigenvalues $\{\lambda_1, \ldots, \lambda_{d'}\}$ of $\mathbf{S}$ and their unit-norm eigenvectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_{d'}\}$
3. Form the $d \times d'$ linear projection matrix $\mathbf{W}$ by aligning the unit-norm eigenvectors in column, i.e. $\mathbf{W} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{d'}]$
4. Set $\tilde{\mathbf{x}}_k = \mathbf{W}^t (\mathbf{x}_k - \mathbf{m}) \, (1 \leq k \leq n)$




# Linear Discriminant Analysis (LDA 线性判别分析)

## Definition

- a.k.a Fisher Discriminant Analysis (FDA)
- Goal: Find linear projections with **good discriminant ability**
- Input:
  - The set of $c$ class labels $\Omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$
  - The set of $n$ d-dimensional training examples $\mathcal{D} = \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_c$ where $\mathcal{D}_i$ consists of $n_i$ training examples ( $n = \sum_{i=1}^{c} n_i$) with label $\omega_i$.
- Output:
  - A linear projection direction $\mathrm{w} \in \mathrm{R}^d$ where the projected training examples with different labels are well separated

## Two heuristic(启发式的) Principles

### Principle 1: within-class variance should be small

- Global sample mean

$$\mathbf{m} = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{D}} \mathbf{x}$$

- Sample mean for $w_i$

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in \mathcal{D}_i} \mathbf{x}$$

- Within-class variance after projection

$$
\begin{aligned}
J_W(\mathbf{w}) & =\sum_{i=1}^c\left(\sum_{\mathbf{x} \in \mathcal{D}_i}\left(\mathbf{w}^t\left(\mathbf{x}-\mathbf{m}_i\right)\right)^2\right) \\
& =\sum_{i=1}^c\left(\sum_{\mathbf{x} \in \mathcal{D}_i} \mathbf{w}^t\left(\mathbf{x}-\mathbf{m}_i\right)\left(\mathbf{x}-\mathbf{m}_i\right)^t \mathbf{w}\right) \\
& =\sum_{i=1}^c\left(\mathbf{w}^t\left(\sum_{\mathbf{x} \in \mathcal{D}_i}\left(\mathrm{x}-\mathbf{m}_i\right)\left(\mathbf{x}-\mathbf{m}_i\right)^t\right) \mathbf{w}\right) \\
& =\sum_{i=1}^c \mathbf{w}^t \mathbf{S}_i \mathbf{w} \\
& =\mathbf{w}^t \mathbf{S}_W \mathbf{w}
\end{aligned}
$$

where $\mathbf{S}_W$ is within-class scatter matrix(类内散度矩阵)

$$
\begin{aligned}
\mathbf{S}_W & =\sum_{i=1}^c \mathbf{S}_i \\
& =\sum_{i=1}^c\left(\sum_{\mathbf{x} \in \mathcal{D}_i}\left(\mathbf{x}-\mathbf{m}_i\right)\left(\mathbf{x}-\mathbf{m}_i\right)^t\right)
\end{aligned}
$$

**Principle 2: between-class variance should be large**

- Global sample mean

$$
\mathbf{m}=\frac{1}{n} \sum_{\mathbf{x} \in \mathcal{D}} \mathbf{x}
$$

- Sample mean for $w_i$

$$
\mathbf{m}_i=\frac{1}{n_i} \sum_{\mathbf{x} \in \mathcal{D}_i} \mathbf{x}
$$

- between-class variance after projection

$$
\begin{aligned}
J_B(\mathbf{w})= & \sum_{i=1}^c\left(n_i\left(\mathbf{w}^t\left(\mathbf{m}_i-\mathbf{m}\right)\right)^2\right) \\
& =\sum_{i=1}^c\left(n_i \mathbf{w}^t\left(\mathbf{m}_i-\mathbf{m}\right)\left(\mathbf{m}_i-\mathbf{m}\right)^t \mathbf{w}\right) \\
& =\mathbf{w}^t\left(\sum_{i=1}^c n_i\left(\mathbf{m}_i-\mathbf{m}\right)\left(\mathbf{m}_i-\mathbf{m}\right)^t\right) \mathbf{w} \\
& =\mathbf{w}^t \mathbf{S}_B \mathbf{w}
\end{aligned}
$$

where $\mathbf{S}_B$ is between-class scatter matrix(类间散度矩阵)

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

## Solution

We want to maximize LDA criterion function which is

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

where

$$
\begin{aligned}
\mathbf{S}_W & =\sum_{i=1}^c\left(\sum_{\mathbf{x} \in \mathcal{D}_i}\left(\mathbf{x}-\mathbf{m}_i\right)\left(\mathbf{x}-\mathbf{m}_i\right)^t\right) \\
\mathbf{S}_B & =\sum_{i=1}^c n_i\left(\mathbf{m}_i-\mathbf{m}\right)\left(\mathbf{m}_i-\mathbf{m}\right)^t
\end{aligned}
$$

which is equal to

$$\begin{aligned}\max_{\mathbf{w}} \ & \mathbf{w}^t \mathbf{S}_B \mathbf{w} \\ \text{s.t.}: \ & \mathbf{w}^t \mathbf{S}_W \mathbf{w} = 1\end{aligned}$$

Using Lagrange function:

$$L(\mathbf{w}, \lambda) = \mathbf{w}^t \mathbf{S}_B \mathbf{w} + \lambda \left(1 - \mathbf{w}^t \mathbf{S}_W \mathbf{w}\right)$$

let

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{0}$$

we h ave a generalized eigenvalue problem

$$\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w}$$

if $\mathbf{S}_W$ is nonsingular

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

so

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

## Notes

#### Total scatter matrix (总体散度矩阵)

$$\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B = \sum_{\mathbf{x} \in \mathcal{D}} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^t$$

#### Block matrix multiplication w.r.t $\mathbf{S}_B$

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

where $\mathbf{A}$ and $\mathbf{B}$ is

$$\begin{aligned}\mathbf{A} &= \left[n_1 \left(\mathbf{m}_1 - \mathbf{m}\right), \ldots, n_c \left(\mathbf{m}_c - \mathbf{m}\right)\right] \in \mathbf{R}^{d \times c} \\ \mathbf{B} &= \left[\left(\mathbf{m}_1 - \mathbf{m}\right), \ldots, \left(\mathbf{m}_c - \mathbf{m}\right)\right]^t \in \mathbf{R}^{c \times d}\end{aligned}$$

#### About Rank

Because

$$\sum_{i=1}^{c} n_i \left(\mathbf{m}_i - \mathbf{m}\right) = \mathbf{0}$$

so the c columns of $\mathbf{A}$ are linearly dependent

$$\begin{aligned}&\operatorname{rank}(\mathbf{A}) \leq c - 1 \\ &\operatorname{rank}(\mathbf{S}_B) \leq c - 1 \\ &\operatorname{rank}\left(\mathbf{S}_W^{-1} \mathbf{S}_B\right) \leq c - 1\end{aligned}$$

$c - 1$ non-zero eigenvalues $\lambda_j$ for $\mathrm{S}_W^{-1} \mathbf{S}_B$ along with their orthogonal eigenvectors $\mathbf{w}_i$

## Algorithm

$$\mathcal{D} = \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_c \left(\mathbf{x} \in \mathcal{D} \subset \mathbf{R}^d\right) \longrightarrow \tilde{\mathcal{D}} = \tilde{\mathcal{D}}_1 \cup \cdots \cup \tilde{\mathcal{D}}_c \left(\tilde{\mathbf{x}} \in \tilde{\mathcal{D}} \subset \mathbf{R}^{c-1}\right)$$

1. Set

$$\mathbf{m} = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{D}} \mathbf{x}$$

   and

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in \mathcal{D}_i} \mathbf{x} (1 \leq i \leq c)$$

2. Set

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

   and

$$\mathbf{S}_W = \sum_{i=1}^{c} \left( \sum_{\mathbf{x} \in \mathcal{D}_i} \left(\mathbf{x} - \mathbf{m}_i\right) \left(\mathbf{x} - \mathbf{m}_i\right)^t \right)$$

3. Choose the $c$ - 1 non-zero eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_{c-1}\}$ for $\mathbf{S}_W^{-1} \mathbf{S}_B$ and identify their orthogonal eigenvectors $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{c-1}\}$ with $\mathbf{w}_i^t \mathbf{S}_W \mathbf{w}_i = 1 (1 \leq i \leq c-1)$
4. Form the $d \times (c-1)$ linear projection matrix $\mathbf{W}$ by aligning the orthogonal eigenvectors in column, i.e. $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{c-1}]$ Linear Discriminant
5. Set $\tilde{\mathrm{x}} = \mathbf{W}^t \mathrm{x}\ (\forall \mathrm{x} \in \mathcal{D})$