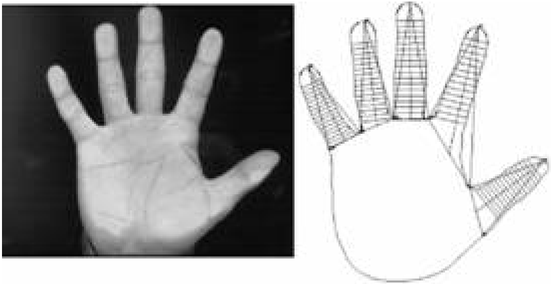
Principal Component Analysis

COMP4211

THE DEPARTMENT OF
COMPUTER SCIENCE & ENGINEERING
計算機科學及工程學系

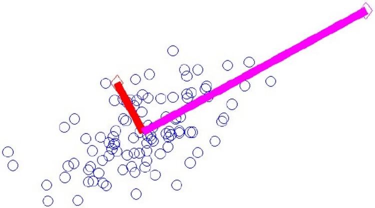
Feature Extraction



Principal Component Analysis (PCA)

- Given: n d-dimensional points $\mathbf{x}_1, \ldots, \mathbf{x}_n$

Goal: find the "right" features from the data



Principal component analysis (PCA)

- aka Karhunen-Loéve (K-L) transformation, Hotelling transformation

Zero-D Representation

How to find $\mathbf{x}_0$ that represents $\mathbf{x}_1, \ldots, \mathbf{x}_n$?

Criterion: find $\mathbf{x}_0$ such that the sum of the squared distances between $\mathbf{x}_0$ and the various $\mathbf{x}_k$ is as small as possible

- define $\mathbf{m} = \sum_{k=1}^{n} \mathbf{x}_k / n$

$$
\begin{aligned}
J_0(\mathbf{x}_0) &= \sum_{k=1}^{n} \|\mathbf{x}_0 - \mathbf{x}_k\|^2 \\
&= \sum_{k=1}^{n} \|(\mathbf{x}_0 - \mathbf{m}) - (\mathbf{x}_k - \mathbf{m})\|^2 \\
&= \sum_{k=1}^{n} \|\mathbf{x}_0 - \mathbf{m}\|^2 - 2(\mathbf{x}_0 - \mathbf{m})^t \sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{m}) + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2 \\
&= \sum_{k=1}^{n} \|\mathbf{x}_0 - \mathbf{m}\|^2 + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2
\end{aligned}
$$

Zero-D Representation...

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

- $\sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2$
  - independent of $\mathbf{x}_0$

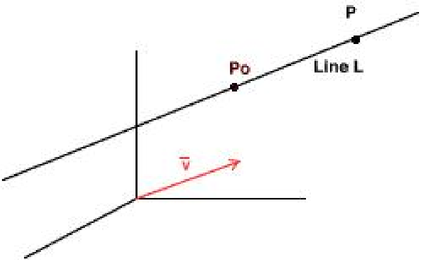
$$\tilde{J}_0(\mathbf{x}_0) = \sum_{k=1}^{n} \|\mathbf{x}_0 - \mathbf{m}\|^2$$

- minimize $\tilde{J}_0(\mathbf{x}_0) \rightarrow \mathbf{x}_0 = \mathbf{m}$

The "best" zero-dimensional representation of the data set is the sample mean

One-D Representation

How to represent the set of points by a line through $\mathbf{m}$?

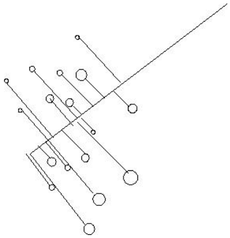


$\mathbf{x} = \mathbf{m} + a\mathbf{e}$, $\mathbf{e}$: unit vector along the line

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

One-D Representation...

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



Note that $\|\mathbf{e}\| = 1$, then $\frac{\partial}{\partial a_k} J_1(a_1, \ldots, a_n, \mathbf{e}) = 0$ gives

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

Project $\mathbf{x}_k$ onto the line in the direction of $\mathbf{e}$ that passes through the sample mean

One-D Representation...

What is the **best** direction?

Substitute $a_k = \mathbf{e}^t(\mathbf{x}_k - \mathbf{m})$ into $J_1(a_1, \ldots, a_n, \mathbf{e})$

$$
\begin{aligned}
J_1(\mathbf{e}) &= \sum_{k=1}^{n} a_k^2 \|\mathbf{e}\|^2 - 2\sum_{k=1}^{n} a_k \mathbf{e}^t(\mathbf{x}_k - \mathbf{m}) + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2 \\
&= \sum_{k=1}^{n} a_k^2 - 2\sum_{k=1}^{n} a_k^2 + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2 \\
&= -\sum_{k=1}^{n} \left(\mathbf{e}^t(\mathbf{x}_k - \mathbf{m})\right)^2 + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2 \\
&= -\sum_{k=1}^{n} \mathbf{e}^t(\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^t\mathbf{e} + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2 \\
&= -\mathbf{e}^t\left(\sum_{k=1}^{n}(\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^t\right)\mathbf{e} + \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{m}\|^2
\end{aligned}
$$

What is the Best Direction?

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

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

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

- $\mathbf{e}$ that minimizes J_1 also maximizes $\mathbf{e}^t\mathbf{S}\mathbf{e}$
- maximize $\mathbf{e}^t\mathbf{S}\mathbf{e}$ subject to $\|\mathbf{e}\| = 1$
  - constrained optimization
- method of Lagrange multipliers

(*) Constrained Optimization

Maximize $\mathbf{e}^t\mathbf{S}\mathbf{e}$ subject to $\|\mathbf{e}\| = 1$

$$L = \mathbf{e}^t\mathbf{S}\mathbf{e} - \lambda(\mathbf{e}^t\mathbf{e} - 1)$$

$$\frac{\partial L}{\partial \mathbf{e}} = 2\mathbf{S}\mathbf{e} - 2\lambda\mathbf{e} = \mathbf{0} \Rightarrow \mathbf{S}\mathbf{e} = \lambda\mathbf{e}$$

- $\mathbf{e}$ must be an eigenvector of $\mathbf{S}$

What is the Best Direction?...

Maximize $\mathbf{e}^t\mathbf{S}\mathbf{e}$ subject to $\|\mathbf{e}\| = 1$

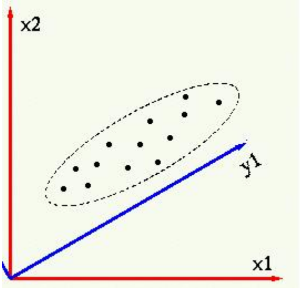
- $\mathbf{e}$ must be an eigenvector of $\mathbf{S}$

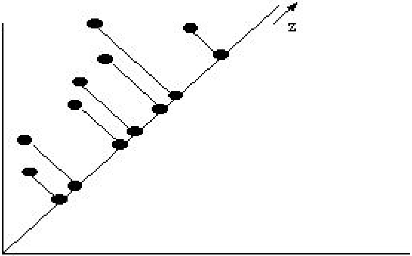
To find the one-dimensional projection of the data that is best in the least sum-of-squared-error sense

- $\rightarrow$ maximize $\mathbf{e}^t\mathbf{S}\mathbf{e} = \lambda\mathbf{e}^t\mathbf{e} = \lambda$
- $\rightarrow$ select the eigenvector $\mathbf{e}$ corresponding to the largest eigenvalue of $\mathbf{S}$

Dimensionality Reduction

Can be used to simplify a dataset by choosing a new coordinate system





- if we only keep y_1 but ignore y_2, a 50% compression rate can be achieved without losing much information in the signal

Second Best Direction?

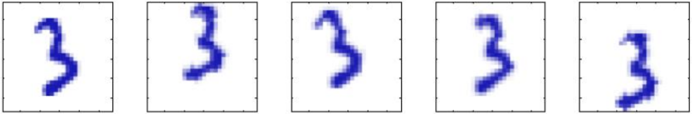
What is the second best direction?

- the second best direction should be orthogonal to the first best direction

$$\begin{aligned} \max_{\mathbf{e}} \quad & \mathbf{e}^t \mathbf{S} \mathbf{e} \\ \text{s.t.} \quad & \mathbf{e}^t \mathbf{e} = 1 \text{ and } \mathbf{e}^t \mathbf{e}_1 = 0 \end{aligned}$$

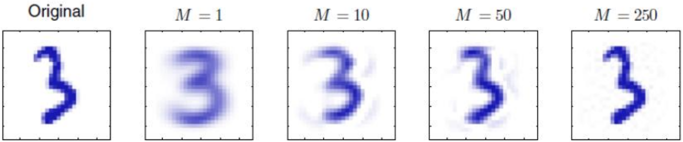
- Select eigenvector $\mathbf{e}_2$ corresponding to 2nd largest eigenvalue of $\mathbf{S}$
- Similarly, the nth best direction is the eigenvector $\mathbf{e}_n$ corresponding to the nth largest eigenvalue of $\mathbf{S}$

Example

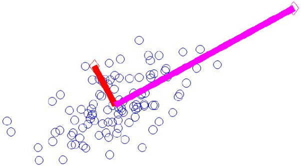


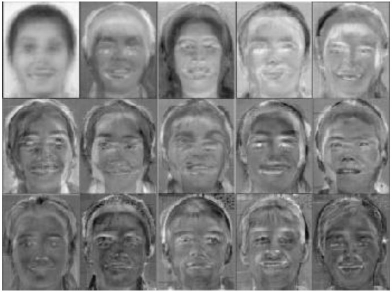
- a collection of 100×100 images created from one image by introducing random displacement and rotation

eigenvectors



EigenFace





Another Derivation

Find the projection $\mathbf{e}$ s.t. $\mathrm{var}(\mathbf{e}^t\mathbf{x})$ is maximized

$$\begin{aligned}
\mathrm{var}(\mathbf{e}^t\mathbf{x}) &= E[(\mathbf{e}^t\mathbf{x}-\mathbf{e}^t\mathbf{m})^2] \\
&= E[(\mathbf{e}^t\mathbf{x}-\mathbf{e}^t\mathbf{m})(\mathbf{e}^t\mathbf{x}-\mathbf{e}^t\mathbf{m})] \\
&= E[\mathbf{e}^t(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^t\mathbf{e}] \\
&= \mathbf{e}^t E[(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^t]\mathbf{e} \\
&= \mathbf{e}^t\boldsymbol{\Sigma}\mathbf{e}
\end{aligned}$$

- $E[(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^t] = \boldsymbol{\Sigma}$

Maximization of $\mathrm{var}(\mathbf{e}^t\mathbf{x})$ subject to $\|\mathbf{e}\| = 1$

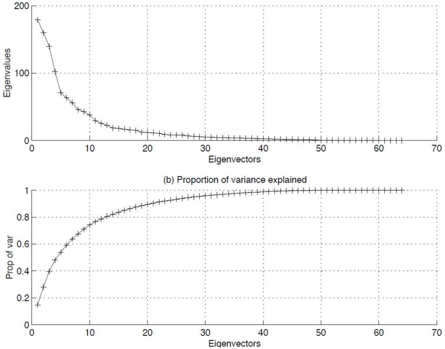
- choose the eigenvector with the largest eigenvalue for the variance to be maximum

How to Choose k?

Proportion of variance explained:

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_d}$$

- λ_i are sorted in descending order

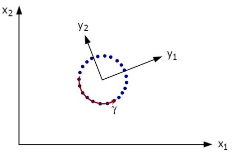


- e.g., stop at proportion of variance > 0.9

Limitations of PCA

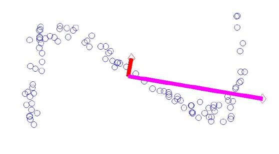
PCA is linear

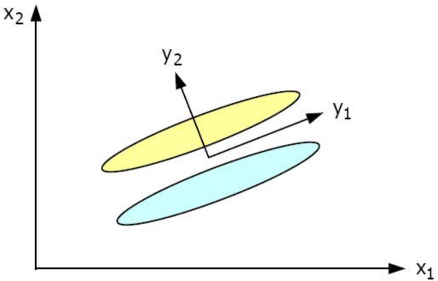
Example



- The dimensionality of the feature space is 2, but the intrinsic dimensionality of the data distribution is 1
- Each point $\mathbf{x}$ in the data set can be specified (parametrically) by a single parameter (γ), instead of two variables x_1 and x_2

Example

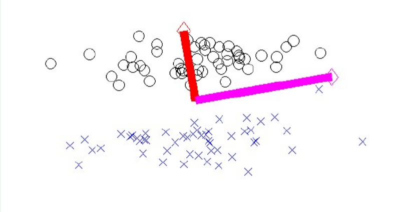




- Since the data variance is largest along the y_1 direction, PCA transforms to one dimension will remove all the ability to discriminate the two classes
- For PCA to be effective in extracting useful features for classification, large variance in the data should correspond to large variance between classes rather than large variance within each class

Limitations of PCA...

Example



Example