EE3731C – Signal Processing Methods

Qi Zhao
Assistant Professor
ECE, NUS

Expectation

- A weighted average of all possible values, e.g., discrete random variables (rvs), finite case:

$$\mathrm{E}(X) = \sum_{i=1}^{n} x_i p_i$$

- Properties
  - $E(X + c) = E(X) + c$
  - $E(X + Y) = E(X) + E(Y)$
  - $E(cX) = cE(X)$
  - $E(cX + dY) = cE(X) + dE(Y)$
  - $E(XY) = E(X)E(Y)$?

Variance

- A measure of how far a set of numbers is spread out

$$\sigma^2 = Var(X) = E[(X - E(X))^2]$$

- Properties
  - $Var(X) \geq 0$
  - $Var(X + c) = Var(X)$
  - $Var(cX) = c^2 Var(X)$
  - $Var(cX + dY) = c^2 Var(X) + d^2 Var(Y) + 2cdCov(X, Y)$

Covariance

- Intuition:
  - A measure of how much two random variables change together.
- Covariance:

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$
$$= E(XY) - E(X)E(Y)$$

Covariance Matrix

- A matrix whose element in the *i*, *j* position is the covariance between the *i* th and *j* th elements of a random vector.
  - Each element of the vector is a scalar random variable.

- Intuitively, the covariance matrix generalizes the notion of variance to multiple dimensions.

Covariance Matrix

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_D \end{bmatrix}$$

$\sum_{ij} = Cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)]$ where $\mu_i = E(X_i)$

$$\sum = \begin{bmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] & \dots E[(X_1 - \mu_1)(X_D - \mu_D)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] & \dots E[(X_2 - \mu_2)(X_D - \mu_D)] \\ \vdots & \vdots \quad \ddots & \vdots \\ E[(X_D - \mu_D)(X_1 - \mu_1)] & E[(X_D - \mu_D)(X_2 - \mu_2)] & \dots E[(X_D - \mu_D)(X_D - \mu_D)] \end{bmatrix}$$

$$\sum = E[(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X}))^T]$$

$$\sigma^2 = Var(X) = E[(X - E(X))^2]$$

Properties of Covariance Matrix

For $\sum = E[(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X}))^T]$ and $\boldsymbol{\mu} = E(\boldsymbol{X})$

Property 1: $\sum = E(\boldsymbol{X}\boldsymbol{X}^T) - \boldsymbol{\mu}\boldsymbol{\mu}^T$

Property 2: $\sum is$ positive semi-definite and symmetric

Property 3: $Cov(\boldsymbol{A}\boldsymbol{X} + c) = \boldsymbol{A}Cov(\boldsymbol{X})\ \boldsymbol{A}^T$

Recall

- Orthogonal: $\boldsymbol{v}_i \boldsymbol{v}_j = 0$, if $\mathrm{i} \neq j$
  - Geometrically perpendicular
  - Statistically uncorrelated in terms of the second-order statistics

- Orthonormal: $\boldsymbol{v}_i \boldsymbol{v}_j = 0$, if $\mathrm{i} \neq j$; $|\boldsymbol{v}_i| = 1$

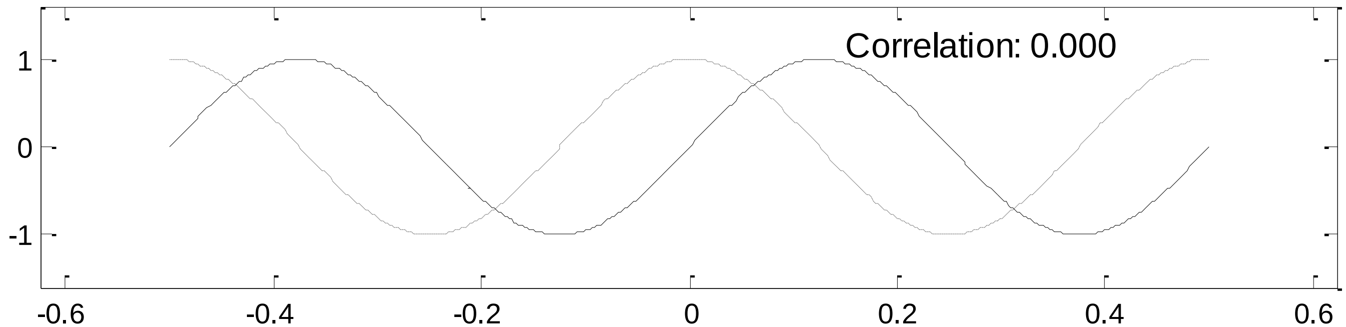
Correlation

- Intuition:
  - A measure of how much two random variables change together.
- Covariance:

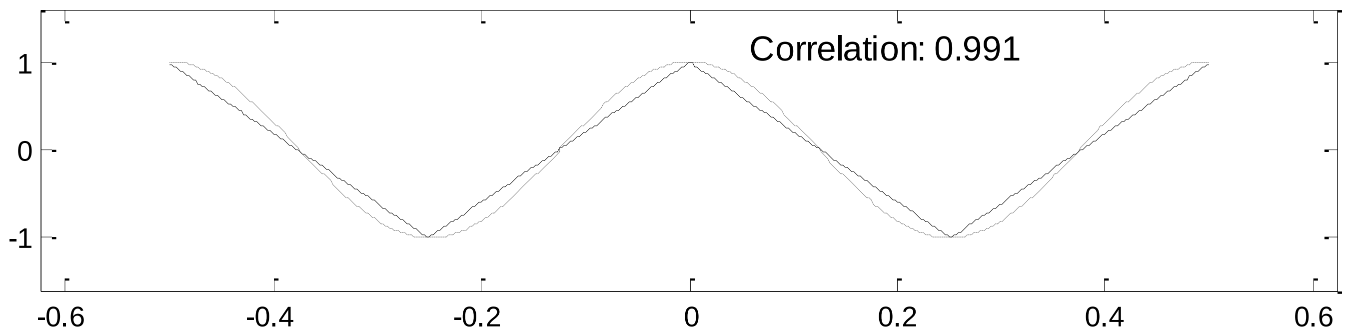
$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$
$$= E(XY)\text{-}\ E(X)E(Y)$$

  - X and Y are uncorrelated if $Cov(X,Y)$ =0
  - Correlation coefficient: $r = \frac{Cov(X,Y)}{std(X)std(Y)}$
  - If $r = 0$, it is also called that X and Y are orthogonal.

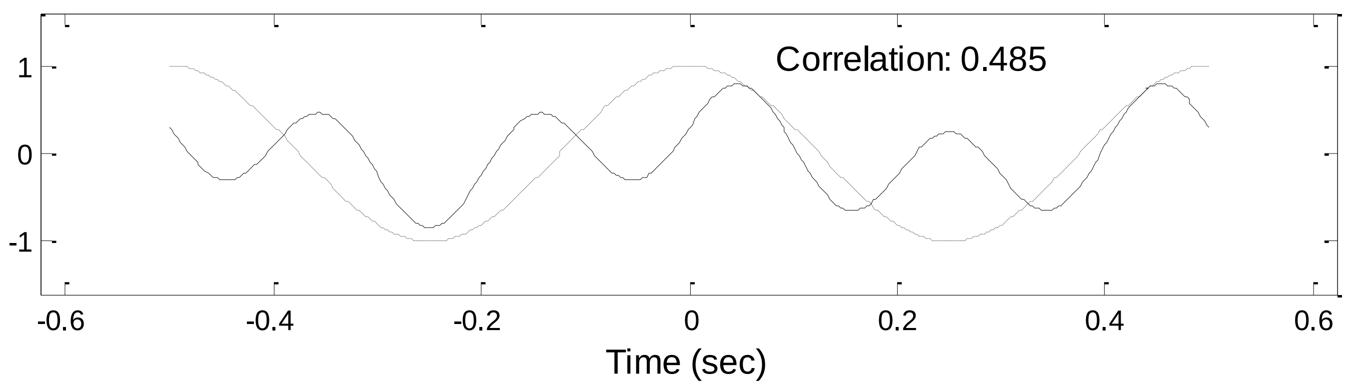
No correlation between a sine and a cosine wave.



A high correlation between a sine and a triangle.



A moderate correlation between a sine and a composite waveform.



Independence

- Intuition: X and Y are in different "worlds".
  - Two random variables are independent if the observed value of one does not affect the probability distribution of the other.
  - More strict concept.
  - Requires the signals to be probabilistically independent and uncorrelated in all the higher-order statistics.

- Statistical independence

$$P(A \cap B) = P(A)P(B)$$

  - The occurrence of one event makes it neither more nor less probable that the other occurs, e.g., rolling a die, tossing a coin, etc.

Motivation

- If two items or dimensions are highly correlated or dependent
  - They are likely to represent highly related phenomena.
  - So we want to combine related variables, and focus on uncorrelated or independent ones, especially those along which the observations have high variance.
- Suppose you have 3 variables, or 4, or 5, or 10000?
- Look for the phenomena underlying the observed covariance/co-dependence in a set of variables.

Curse of Dimensionality

- If data lie in high-dimensional space, then a large amount of data is required to learn its distributions.

- 100 dimensions, each dimension 5 levels

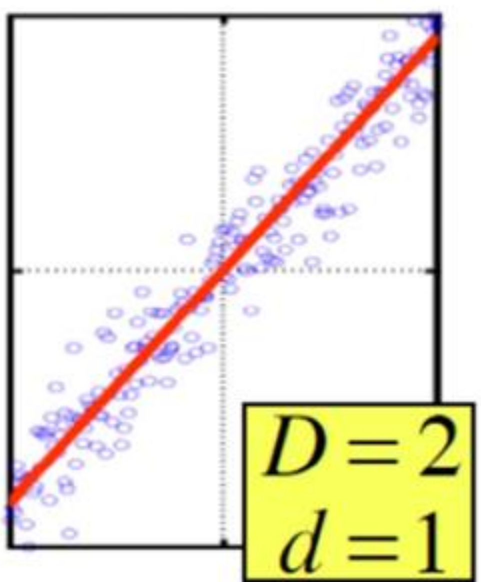
=> 5^{100}

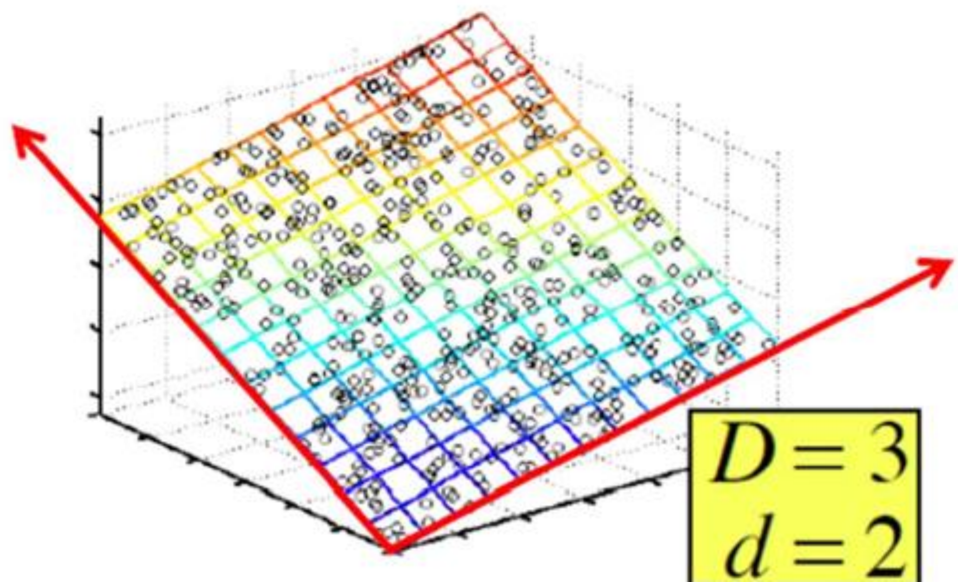
- Given data in a D-dimensional space, the hope is that data points will lie mainly in a linear subspace with d-dimensions (d<D).

What is the max value for d?

Dimensionality Reduction

- Project the data into lower-dimensional space
  - Assumption: data approximately lie in a lower-dimensional space => *Preserve structure.*
  - Less computation, easier interpretation.

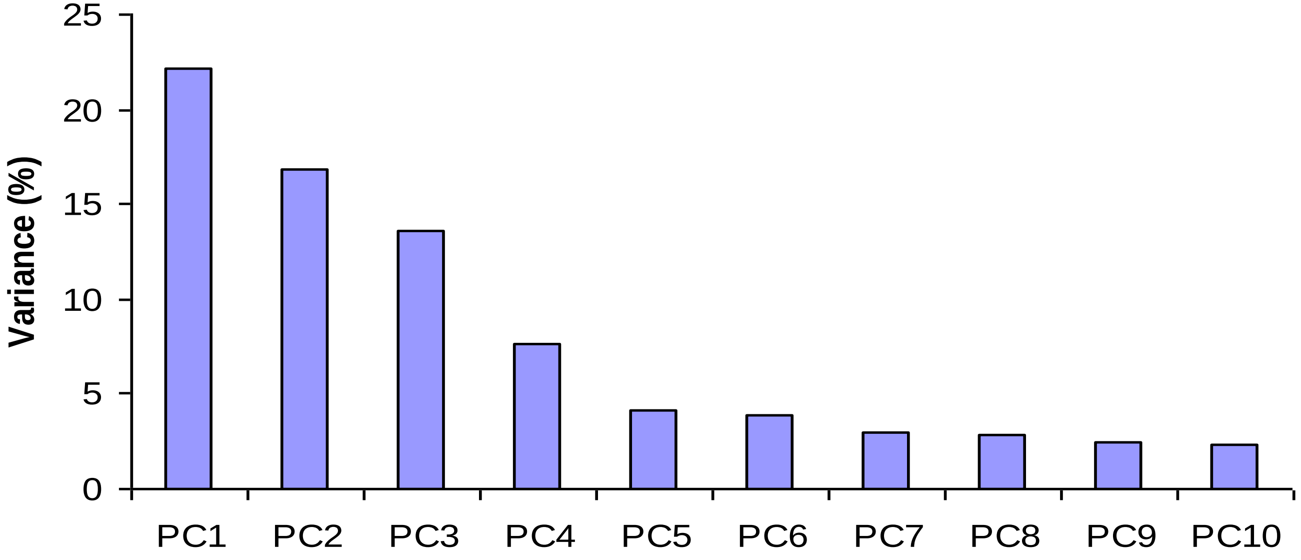




Principal Components

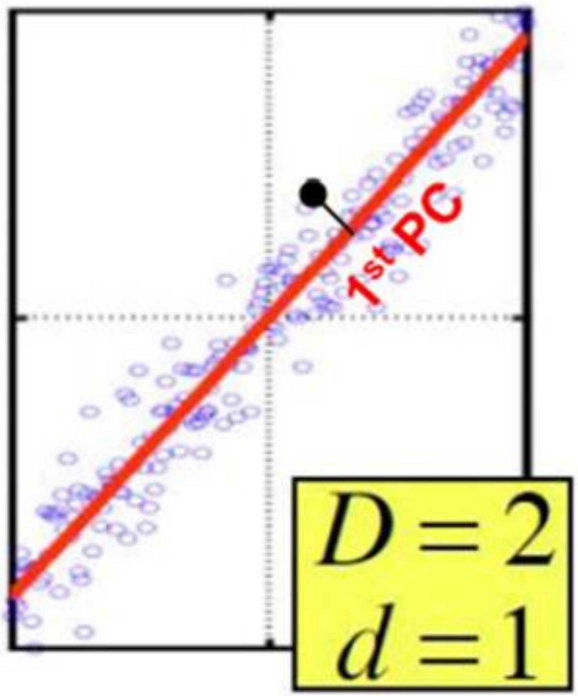
- Principal components are a new coordinate system.
- The new variables/dimensions
  - Are linear combinations of the original ones
  - Are uncorrelated with one another
    - Orthogonal in original dimension space
  - Capture as much of the original variance in the data as possible
  - Are called Principal Components

Dimensionality Reduction

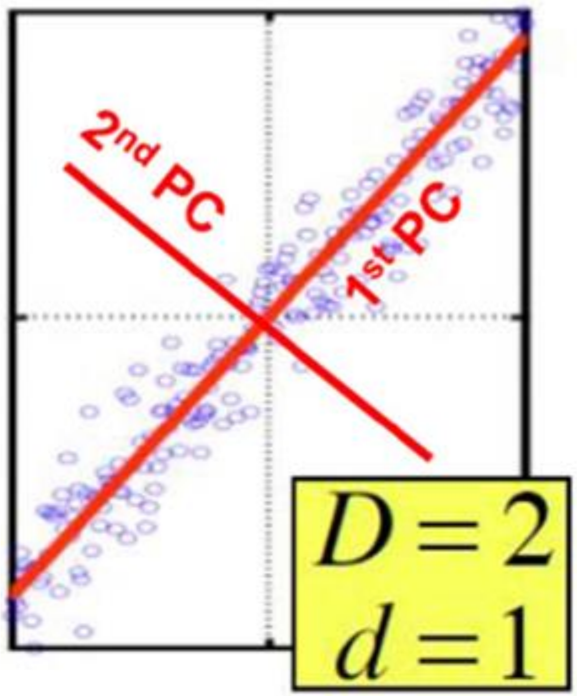


- Can ignore the components with lesser significance
- Do lose some information, but hopefully not much

Principal Component Analysis

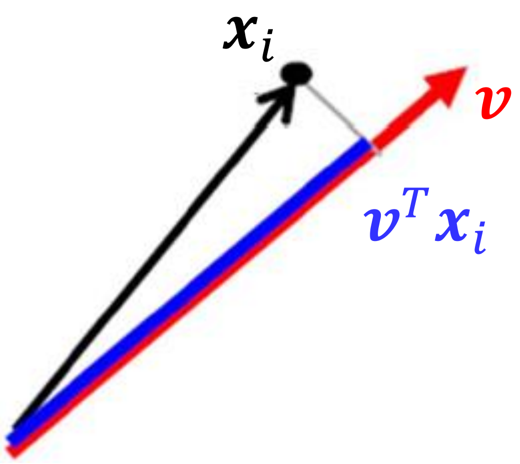


- Principal components (PCs): orthogonal directions that capture most of the variance in data
- 1st PC: direction of greatest variability in data



- 2^{nd} PC: next orthogonal direction of greatest variability

Projecting onto the PCs



- $\boldsymbol{x}_i$: a D-dimensional data point
- $\boldsymbol{v}$: 1st PC
- $\boldsymbol{v}^T\boldsymbol{x}_i$: projection of $\boldsymbol{x}_i$ onto the 1st PC

Principal Component Analysis

- Assume data are centered
- For a projection direction $\boldsymbol{v}$, variance of projected data $Var(\boldsymbol{v}^T\boldsymbol{x})$:

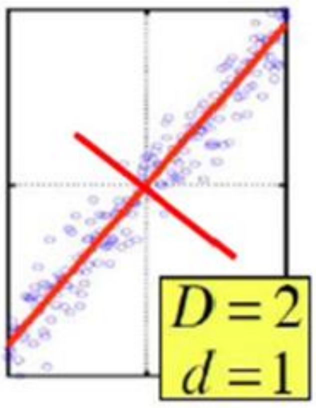
$$\frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{v}^T\boldsymbol{x}_i)^2 = \frac{1}{n-1}\boldsymbol{v}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{v}$$

  – $\boldsymbol{X}\boldsymbol{X}^T$: sample covariance

- Maximize the variance of projected data

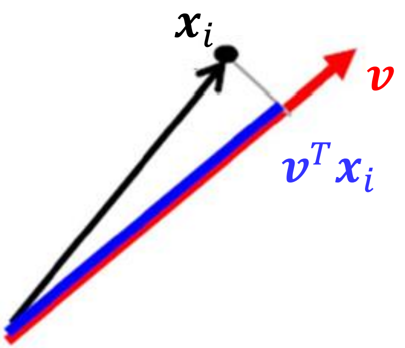
$$\max_{\boldsymbol{v}} \boldsymbol{v}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{v} \quad \text{s.t. } \boldsymbol{v}^T\boldsymbol{v} = 1$$

How to solve this?



- Maximum Variance Subspace: PCA finds vectors $\boldsymbol{v}$ such that projections onto the vectors capture maximum variance in the data

$$\max_{\boldsymbol{v}} \boldsymbol{v}^T \boldsymbol{X} \boldsymbol{X}^T \boldsymbol{v} \quad \text{s.t. } \boldsymbol{v}^T \boldsymbol{v} = 1$$



First PC

- $Var(\boldsymbol{v}^T\boldsymbol{x})$ is maximized if λ_1 is the max eigenvalue of $\boldsymbol{X}\boldsymbol{X}^T$, and the first PC is the corresponding eigenvector.

All PCs

- All the PCs are generated in this way.
  - Each is a eigenvector of $\boldsymbol{X}\boldsymbol{X}^T$ and their corresponding eigenvalues satisfy:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$$

  - So that

$$Var(\boldsymbol{v}_1{}^T\boldsymbol{x}) \geq Var(\boldsymbol{v}_2{}^T\boldsymbol{x}) \geq \cdots \geq Var(\boldsymbol{v}_D{}^T\boldsymbol{x})$$

Derivation of PCA

Assumption and More Notation

- $\mathbf{\Sigma}$ is the *known* covariance matrix for the random variable $\mathbf{x}$

Shortcut to solution

- For $k = 1, 2, \ldots, p$ the k^{th} PC $\boldsymbol{\alpha}_k$
  is an eigenvector of $\mathbf{\Sigma}$ corresponding to its k^{th} largest eigenvalue λ_k.
- If $\boldsymbol{\alpha}_k$ is chosen to have unit length (i.e. $\boldsymbol{\alpha}_k^T \boldsymbol{\alpha}_k = 1$) then
  $\mathrm{Var}(z_k) = \lambda_k \quad (z_k = \boldsymbol{\alpha}_k^T \mathbf{x})$

Derivation of PCA

First Step

- Find $\boldsymbol{\alpha}_k$ that maximizes $\text{Var}(\boldsymbol{\alpha}_k^T \mathbf{x}) = \boldsymbol{\alpha}_k^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_k$
- Without constraint we could pick a very big $\boldsymbol{\alpha}_k$.
- Choose normalization constraint, namely $\boldsymbol{\alpha}_k^T \boldsymbol{\alpha}_k = 1$ (unit length vector).

Constrained maximization - method of Lagrange multipliers

- To maximize $\boldsymbol{\alpha}_k^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_k$ subject to $\boldsymbol{\alpha}_k^T \boldsymbol{\alpha}_k = 1$ we use the technique of Lagrange multipliers. We maximize the function

$$\boldsymbol{\alpha}_k^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_k - \lambda(\boldsymbol{\alpha}_k^T \boldsymbol{\alpha}_k - 1)$$

w.r.t. to $\boldsymbol{\alpha}_k$ by differentiating w.r.t. to $\boldsymbol{\alpha}_k$.

Derivation of PCA

Constrained maximization - method of Lagrange multipliers

- This results in

$$
\begin{aligned}
\frac{d}{d\boldsymbol{\alpha}_k}\left(\boldsymbol{\alpha}_k^T\boldsymbol{\Sigma}\boldsymbol{\alpha}_k - \lambda_k(\boldsymbol{\alpha}_k^T\boldsymbol{\alpha}_k - 1)\right) &= 0 \\
\boldsymbol{\Sigma}\boldsymbol{\alpha}_k - \lambda_k\boldsymbol{\alpha}_k &= 0 \\
\boldsymbol{\Sigma}\boldsymbol{\alpha}_k &= \lambda_k\boldsymbol{\alpha}_k
\end{aligned}
$$

- This should be recognizable as an eigenvector equation where $\boldsymbol{\alpha}_k$ is an eigenvector of Σ and λ_k is the associated eigenvalue.
- Which eigenvector should we choose?

Derivation of PCA

Constrained maximization - more constraints

- The second PC, $\boldsymbol{\alpha}_2$ maximizes $\boldsymbol{\alpha}_2^T \Sigma \boldsymbol{\alpha}_2$ subject to being uncorrelated with $\boldsymbol{\alpha}_1$.
- The uncorrelation constraint can be expressed using any of these equations

$$
\begin{aligned}
\mathrm{cov}(\boldsymbol{\alpha}_1^T \mathbf{x}, \boldsymbol{\alpha}_2^T \mathbf{x}) &= \boldsymbol{\alpha}_1^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_2^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2^T \lambda_1 \boldsymbol{\alpha}_1 \\
&= \lambda_1 \boldsymbol{\alpha}_2^T \boldsymbol{\alpha}_1 = \lambda_1 \boldsymbol{\alpha}_1^T \boldsymbol{\alpha}_2 = 0
\end{aligned}
$$

- Of these, if we choose the last we can write an Langrangian to maximize $\boldsymbol{\alpha}_2$

$$
\boldsymbol{\alpha}_2^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_2 - \lambda_2(\boldsymbol{\alpha}_2^T \boldsymbol{\alpha}_2 - 1) - \phi \boldsymbol{\alpha}_2^T \boldsymbol{\alpha}_1
$$

Derivation of PCA

Constrained maximization - more constraints

- Differentiation of this quantity w.r.t. $\boldsymbol{\alpha}_2$ (and setting the result equal to zero) yields

$$
\begin{aligned}
\frac{d}{d\boldsymbol{\alpha}_2}\left(\boldsymbol{\alpha}_2^T\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda_2(\boldsymbol{\alpha}_2^T\boldsymbol{\alpha}_2 - 1) - \phi\boldsymbol{\alpha}_2^T\boldsymbol{\alpha}_1\right) &= 0 \\
\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda_2\boldsymbol{\alpha}_2 - \phi\boldsymbol{\alpha}_1 &= 0
\end{aligned}
$$

- If we left multiply $\boldsymbol{\alpha}_1$ into this expression

$$
\begin{aligned}
\boldsymbol{\alpha}_1^T\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda_2\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_2 - \phi\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_1 &= 0 \\
0 - 0 - \phi 1 &= 0
\end{aligned}
$$

then we can see that ϕ must be zero and that when this is true that we are left with

$$
\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda_2\boldsymbol{\alpha}_2 = 0
$$

Derivation of PCA

Clearly

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda_2\boldsymbol{\alpha}_2 = 0$$

is another eigenvalue equation and the same strategy of choosing α_2 to be the eigenvector associated with the second largest eigenvalue yields the second PC.

This process can be repeated for $k = 1 \dots p$ yielding up to p different eigenvectors of $\boldsymbol{\Sigma}$ along with the corresponding eigenvalues $\lambda_1, \dots \lambda_p$.

Furthermore, the variance of each of the PC's are given by

$$\text{Var}[\boldsymbol{\alpha}_k^T \mathbf{x}] = \lambda_k, \qquad k = 1, 2, \dots, p$$

Computing PCA

- Subtract off the mean
- Form the covariance matrix
- Calculate the eigenvectors and eigenvalues of the covariance matrix
- Rearrange the eigenvectors and eigenvalues
- Select a subset of eigenvectors as basis vectors

PCA Analysis on Images

- PCA analysis can be used to decompose an image into a set of orthogonal principal component images (eigen-images)
  - Image coding
  - Image denoising
  - Features for image classification
  - …

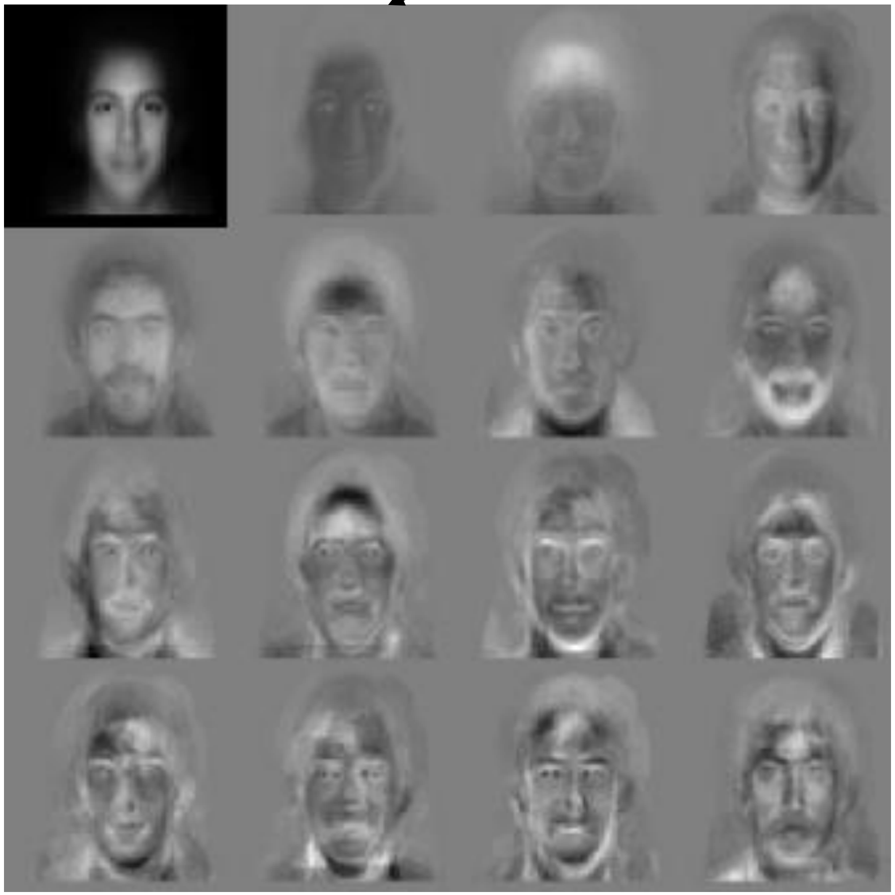
PCA on Faces: “Eigenfaces”

- Eigenfaces are a set of eigenvectors used in the computer vision problem of human face recognition.
  - These eigenvectors are derived from the covariance matrix of the probability distribution of the high-dimensional vector space of possible faces of human beings.
- Eigenfaces are the “standardized face ingredients” derived from the statistical analysis of many pictures of human faces.
- A human face may be considered to be a combination of these standard faces.

PCA on Faces: "Eigenfaces"

First principal component

Average face



Other components

For all except average,
"gray" = 0,
"white" > 0,
"black" < 0

PCA on Faces: “Eigenfaces”

- When properly weighted, eigenfaces can be summed together to create an approximate gray-scale rendering of a human face.
- Remarkably few eigenvector terms are needed to give a fair likeness of most people's faces.
- Hence eigenfaces provide a means of applying data compression to faces for identification purposes.

PCA for Relighting

- Images under different illumination

PCA for Relighting

- Most variation captured by first 5 principal components – can re-illuminate by combining only a few images

EE3731C – Signal Processing Methods

Qi Zhao
Assistant Professor
ECE, NUS

Some of the materials are from the Internet and the textbooks