UNDERSTANDING STATISTICS & EXPERIMENTAL DESIGN

Content

1. Basic Probability Theory
2. Signal Detection Theory (SDT)
3. SDT and Statistics I and II
4. Statistics in a nutshell
5. Multiple Testing
6. ANOVA
7. Experimental Design & Statistics
8. Correlations & PCA
9. Meta-Statistics: Basics
10. Meta-Statistics: Too good to be true
11. Meta-Statistics: How big a problem is publication bias?
12. Meta-Statistics: What do we do now?

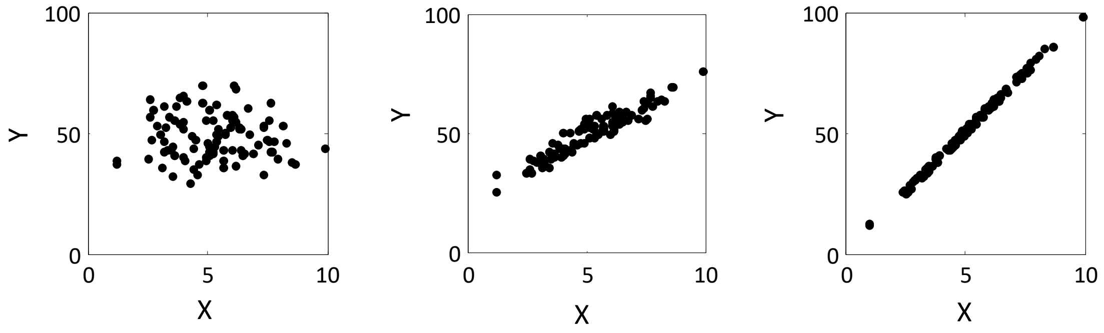
Principal component analysis (PCA)

Covariance & Correlations

No relationship

Weak positive relationship

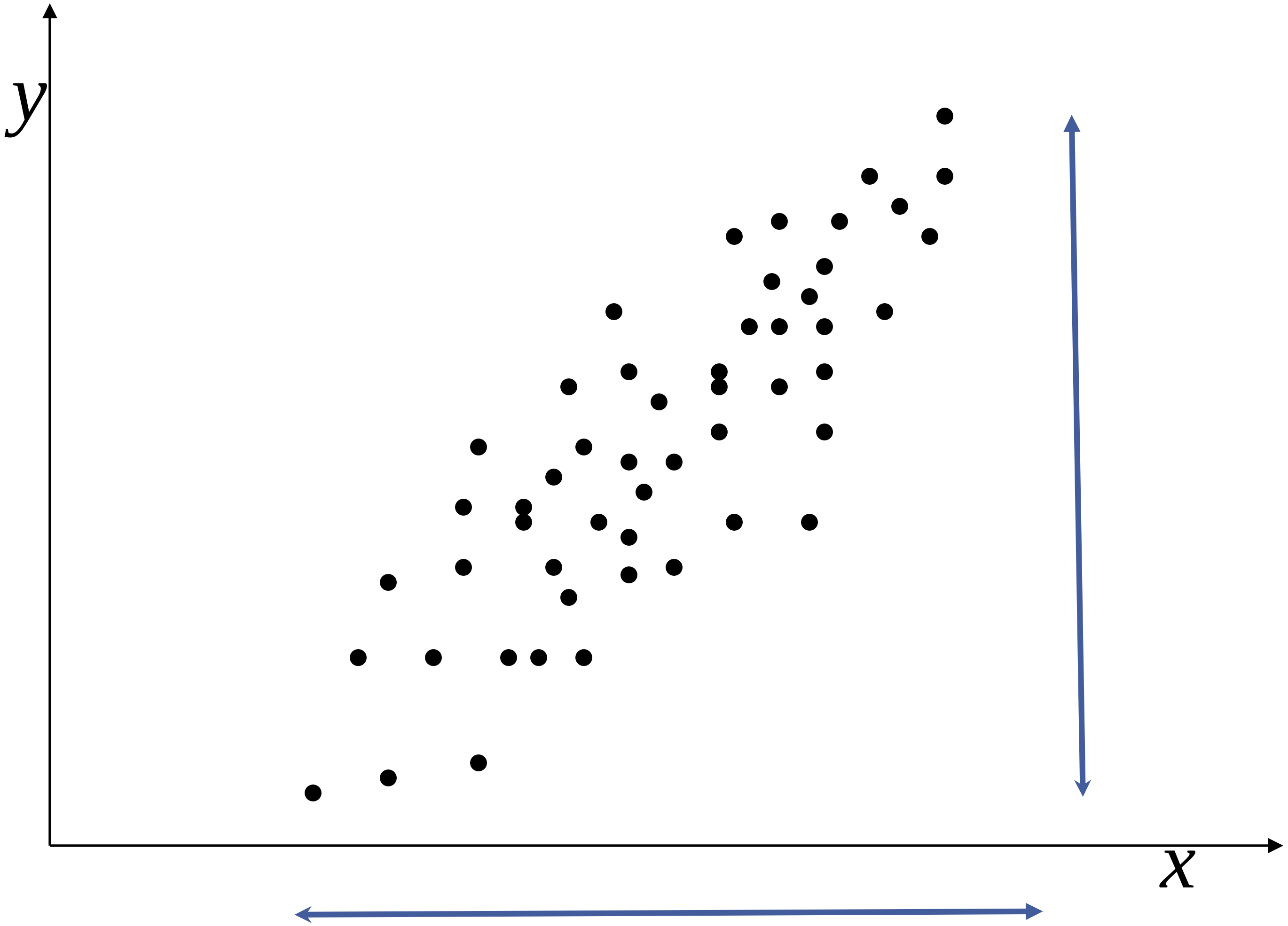
Strong positive relationship

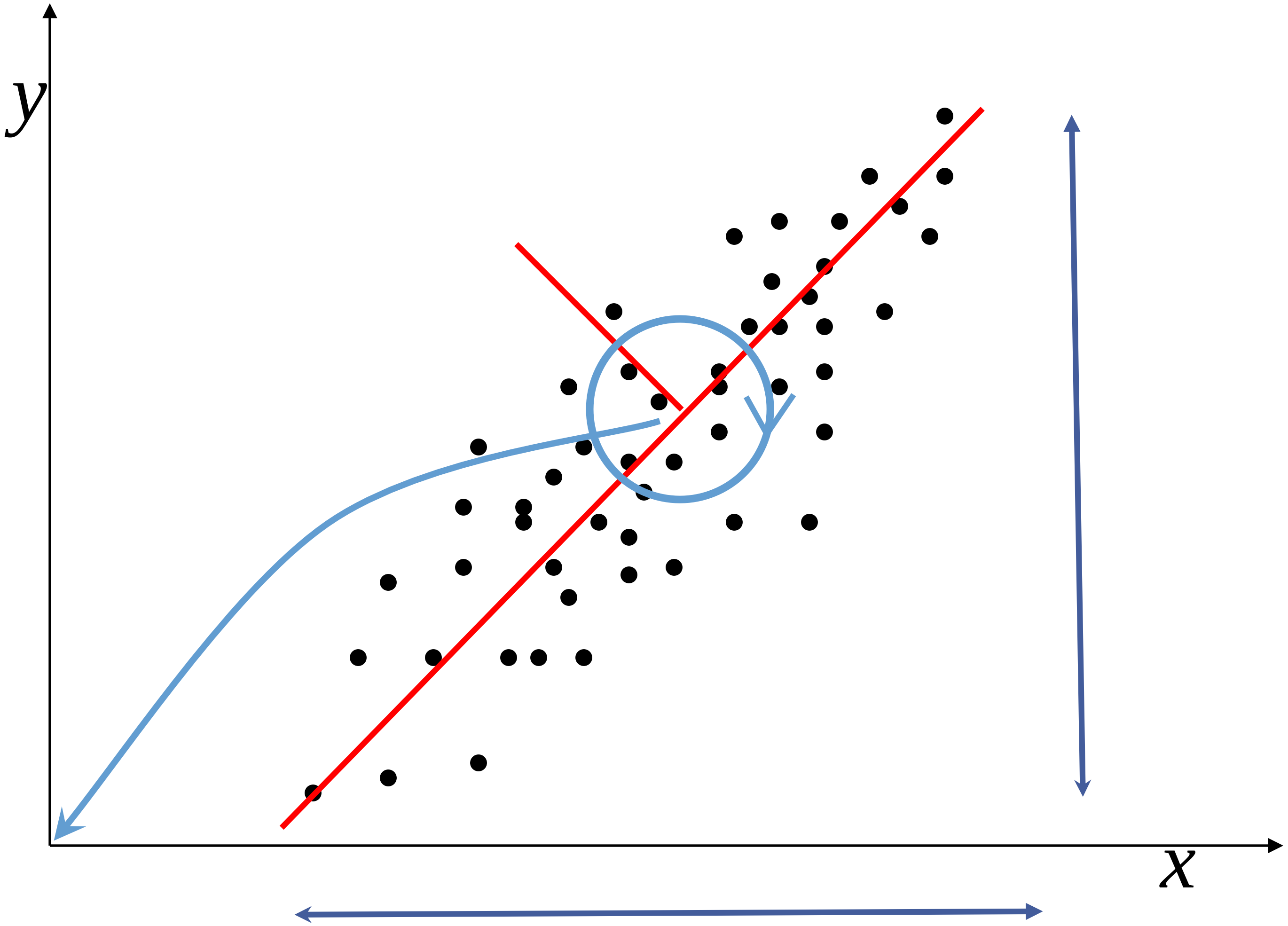


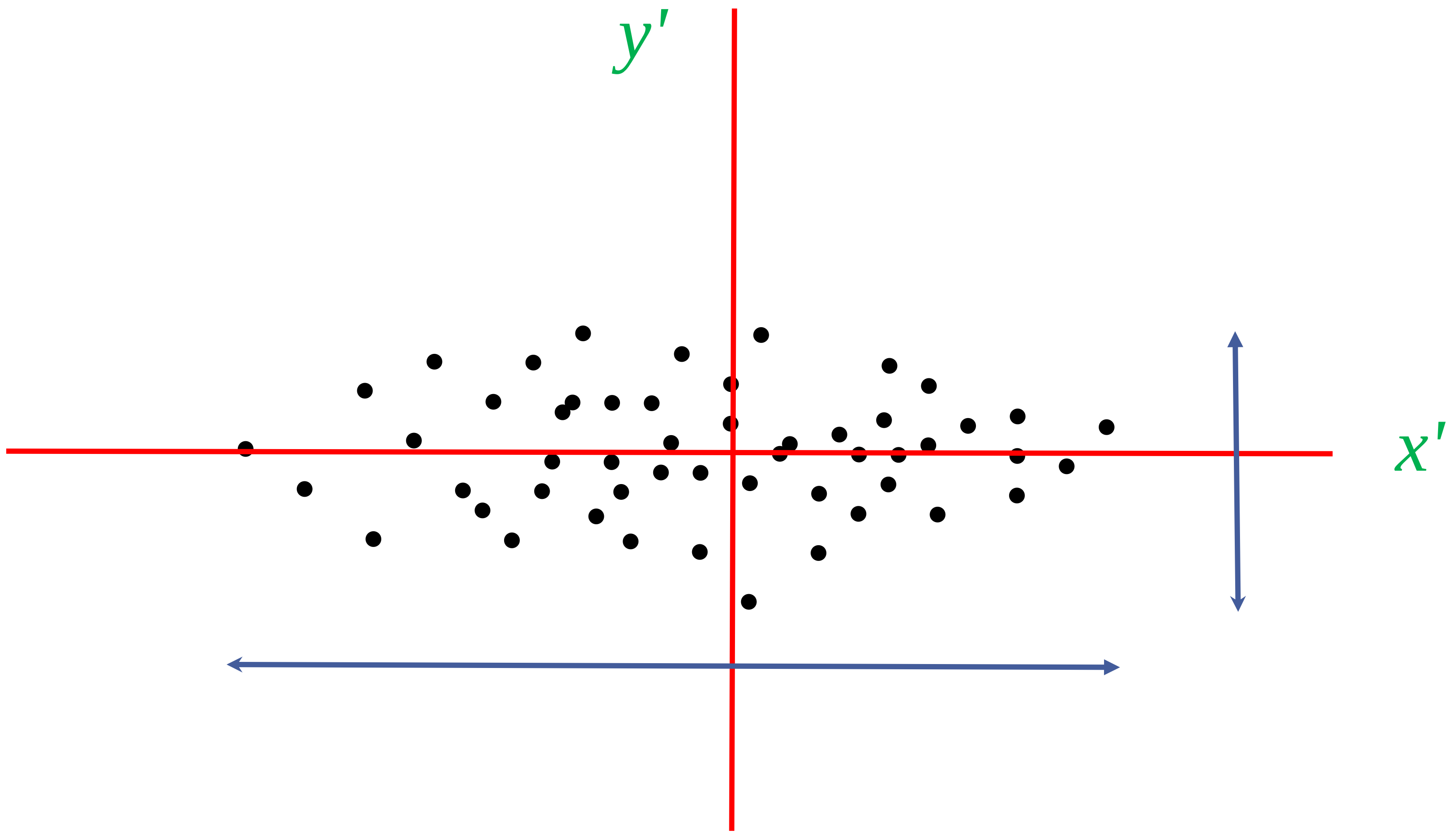
$$\mathrm{cov}(X,Y) = \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

$$r = corr(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_x \sigma_y}$$

r^2 variance explained


y
x

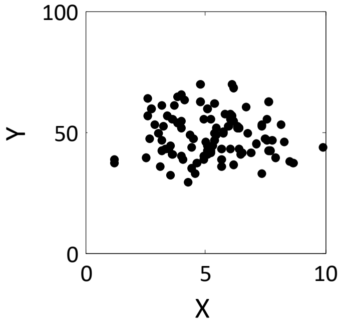

y
x



y'
x'

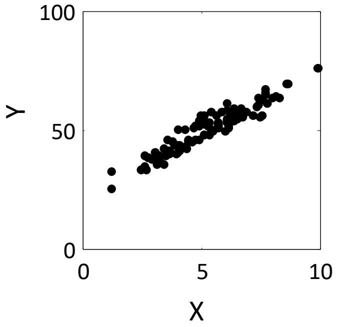
Covariance Matrix

No relationship



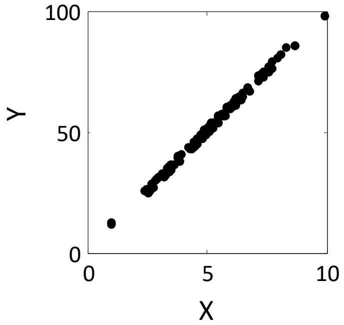
$$\begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$$

Weak positive relationship



$$\begin{pmatrix} 3 & 0.5 \\ 0.5 & 3 \end{pmatrix}$$

Strong positive relationship

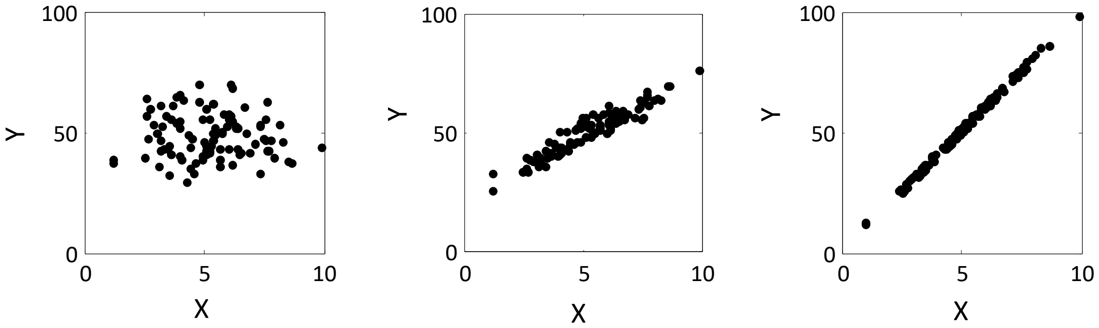


$$\begin{pmatrix} 3 & 3 \\ 3 & 3 \end{pmatrix}$$

$$\mathrm{cov}(x, y) = \begin{bmatrix} s_x^2 & s_{xy} \\ s_{yx} & s_y^2 \end{bmatrix}$$

Covariance Matrix

No relationship — Weak positive relationship — Strong positive relationship

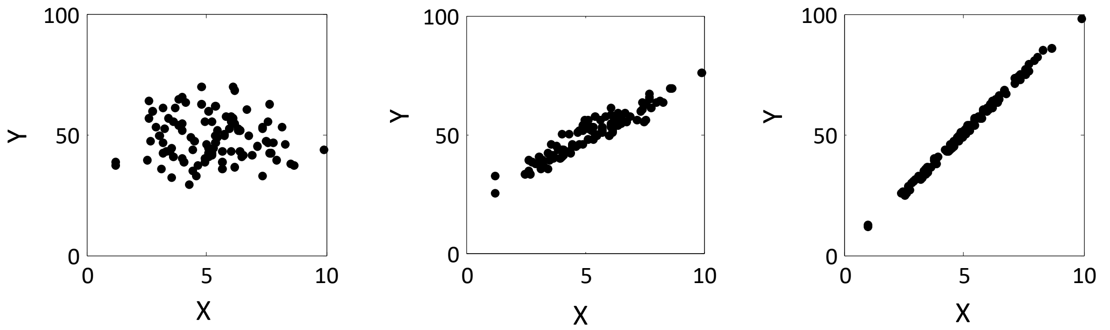


$$\begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} \longrightarrow \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} \qquad \begin{pmatrix} 3 & 0.1 \\ 0.1 & 0.2 \end{pmatrix} \longrightarrow \begin{pmatrix} 3.19 & 0 \\ 0 & 0.01 \end{pmatrix} \qquad \begin{pmatrix} 3 & 3 \\ 3 & 3 \end{pmatrix} \longrightarrow \begin{pmatrix} 6 & 0 \\ 0 & 0 \end{pmatrix}$$

$$\mathrm{cov}(x, y) = \begin{bmatrix} s_x^2 & s_{xy} \\ s_{yx} & s_y^2 \end{bmatrix}$$

Covariance Matrix

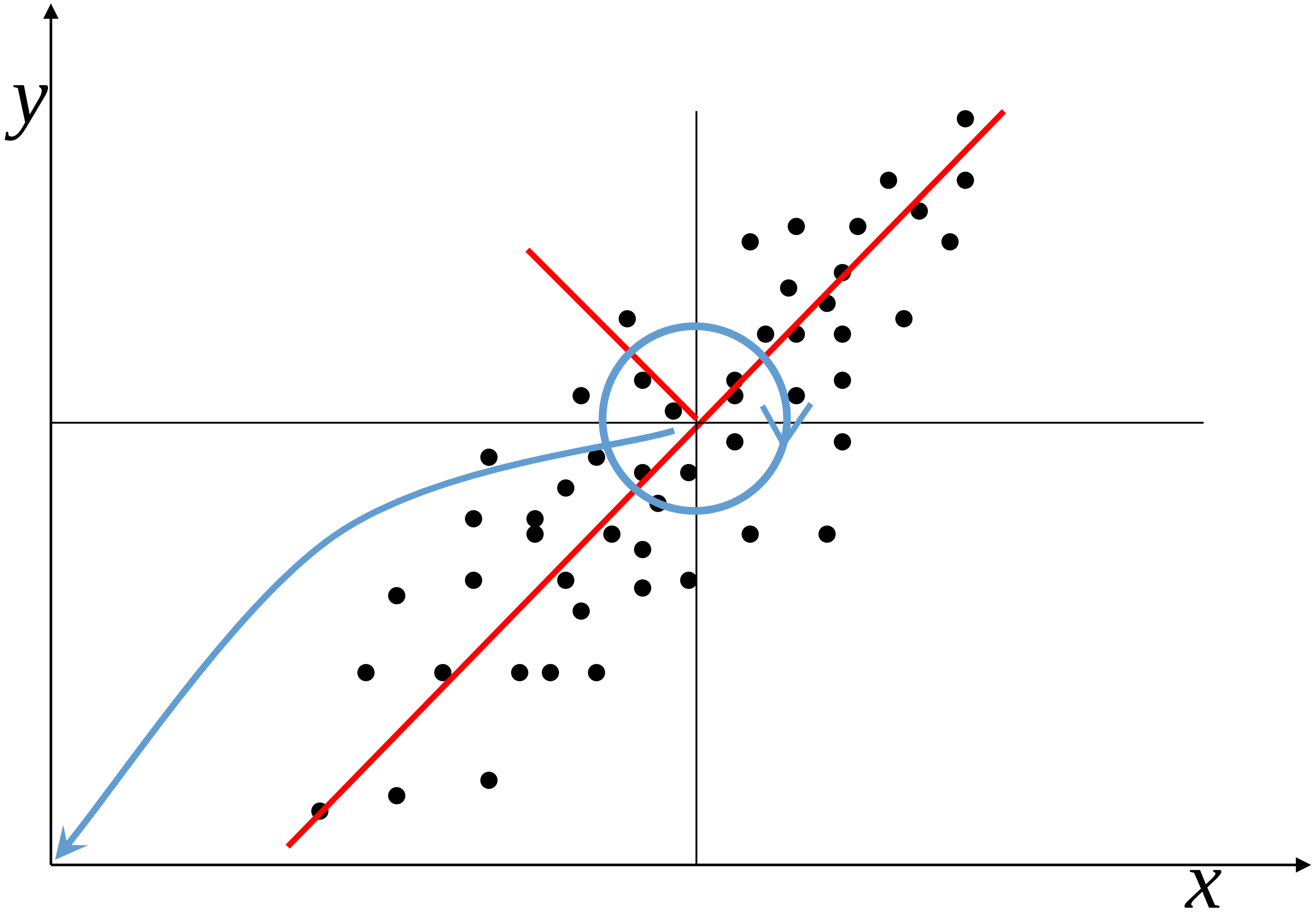
No relationship | Weak positive relationship | Strong positive relationship



$$\begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} \longrightarrow \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} \qquad \begin{pmatrix} 3 & 0.1 \\ 0.1 & 0.2 \end{pmatrix} \longrightarrow \begin{pmatrix} 3.19 & 0 \\ 0 & 0.01 \end{pmatrix} \qquad \begin{pmatrix} 3 & 3 \\ 3 & 3 \end{pmatrix} \longrightarrow \begin{pmatrix} 6 & 0 \\ 0 & 0 \end{pmatrix}$$

Observation: Total variance $s^2_x + s^2_y$ (trace of matrix) does not change but covariance and single variances may
What we want: de-correlated data, new representation

What about more dimensions?



Can we rotate and translate the Cov-Matrix for n > 2?

$$\begin{pmatrix} \lambda_1 & & & & 0 \\ & \lambda_2 & & & \\ & & \ddots & & \\ & & & \lambda_{n-1} & \\ 0 & & & & \lambda_n \end{pmatrix} = D$$

Can we rotate and translate the Cov-Matrix for n > 2?

$$\begin{pmatrix} \lambda_1 & & & & 0 \\ & \lambda_2 & & & \\ & & \ddots & & \\ 0 & & & \lambda_{n-1} & \\ & & & & \lambda_n \end{pmatrix} = D$$

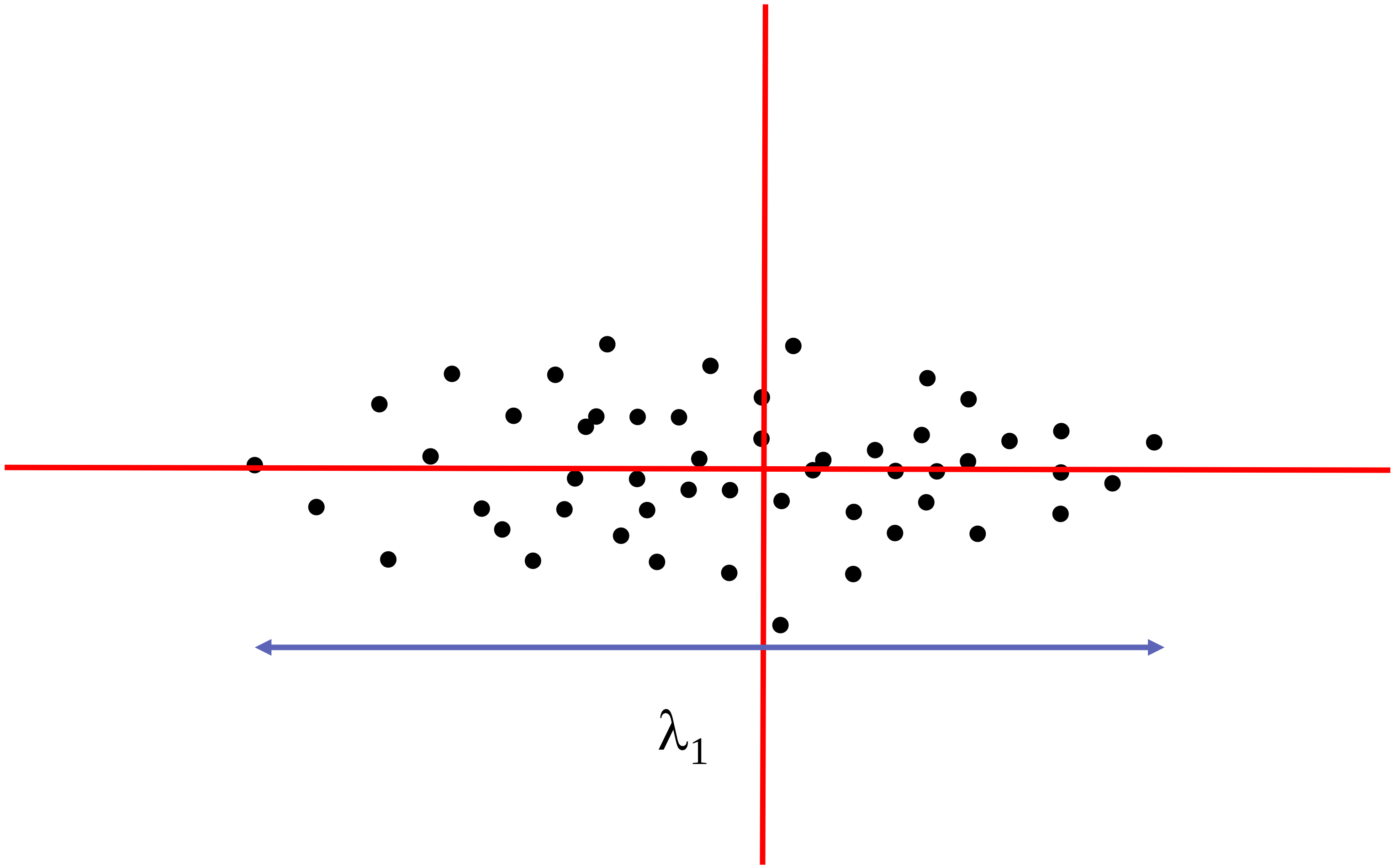
Yes, we can! Because Cov(X,Y) is symmetric

$$D = \mathbf{X}' Cov \mathbf{X}$$

We can rotate and translate the Cov-Matrix to

$$\begin{pmatrix} \lambda_1 & & & & 0 \\ & \lambda_2 & & & \\ & & \ddots & & \\ & & & \lambda_{n-1} & \\ 0 & & & & \lambda_n \end{pmatrix} = D$$

- For each Cov(X,Y), there are principal components, i.e., other than the original basis vectors, which allow for a more simple representation
- The λ_i are the eigenvalues of Cov(X,Y) and the principal components are the corresponding eigenvectors e_i
- Each λ_i indicates the variance e_i explains
- The sum of the λ_i is equal to the total variance, i.e., Trace, of Cov(X,Y)


λ_1

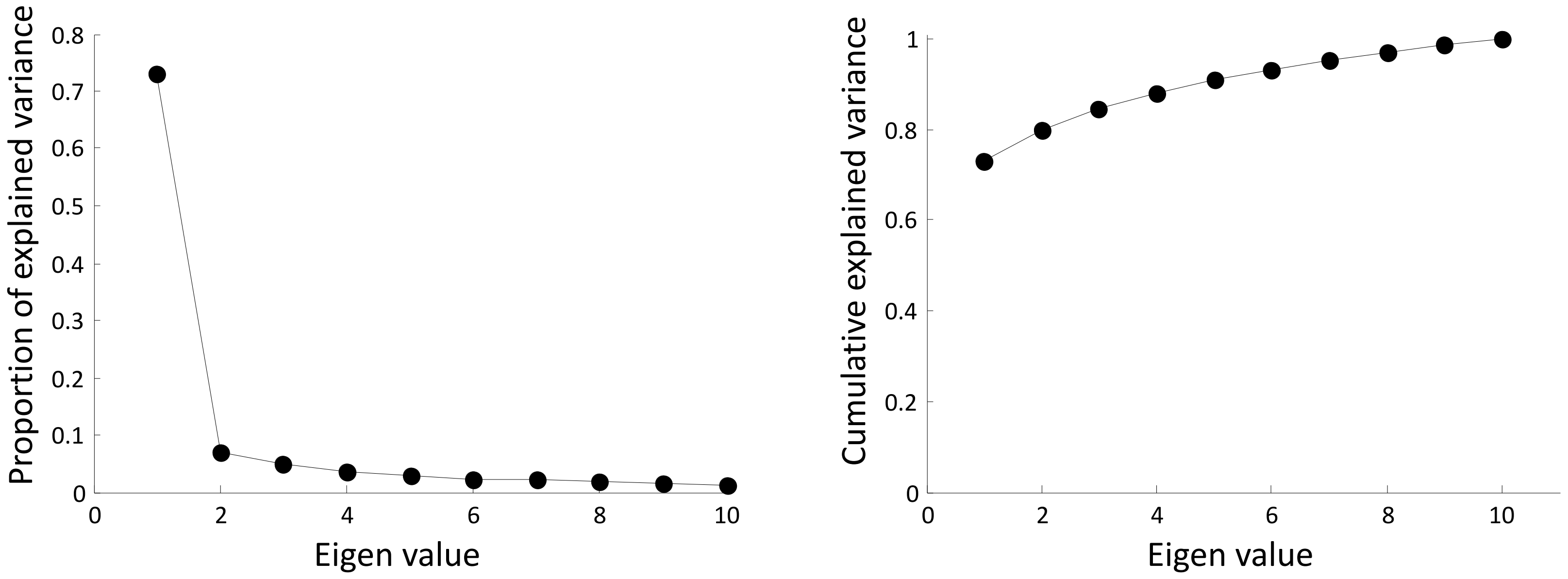
We can rotate and translate the Cov-Matrix to

$$\begin{pmatrix} \lambda_1 & & & & 0 \\ & \lambda_2 & & & \\ & & \ddots & & \\ & & & \lambda_{n-1} & \\ 0 & & & & \lambda_n \end{pmatrix} = D$$

- For each Cov(X,Y), there are principal components, i.e., other than the original basis vectors, which allow for a more simple representation
- The λ_i are the eigenvalues of Cov(X,Y) and the principal components are the corresponding eigenvectors e_i
- Each λ_i indicates the variance e_i explains
- The sum of the λ_i is equal to the total variance, i.e., Trace, of Cov(X,Y)

PCA Applications: Data Compression

Convention: λ_i with small values are "uninformative": remove
If an eigenvalue < 1, it explains less variance than one of the original variables



Scree Plot

Principal Components Analysis (PCA)

- Step 1: $x\text{-}m_x$, $y\text{-}m_y$.
- Step 2: compute the covariance matrix for *x* and *y*:

$$\text{cov}(x, y) = \begin{bmatrix} s_x^2 & s_{xy} \\ s_{yx} & s_y^2 \end{bmatrix}$$

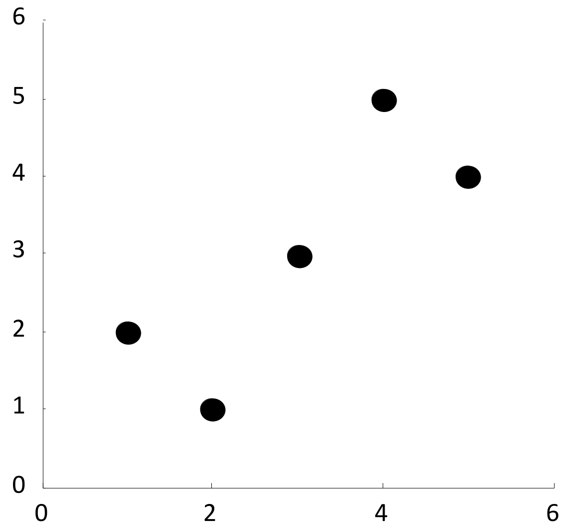
- Step 3: Solve the characteristic equation:

$$det(Cov(x, y) - \lambda I) = 0$$

$$det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$$

PCA Example

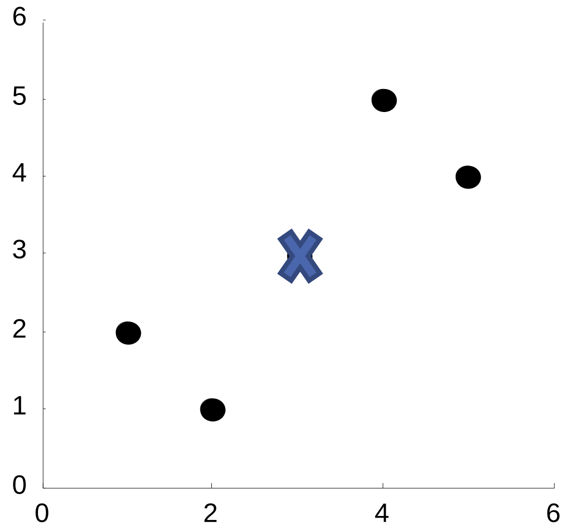
X	Y
1	2
2	1
3	3
4	5
5	4



PCA Example

X	Y
1	2
2	1
3	3
4	5
5	4

$m_x = 3$ $m_y = 3$



PCA Example

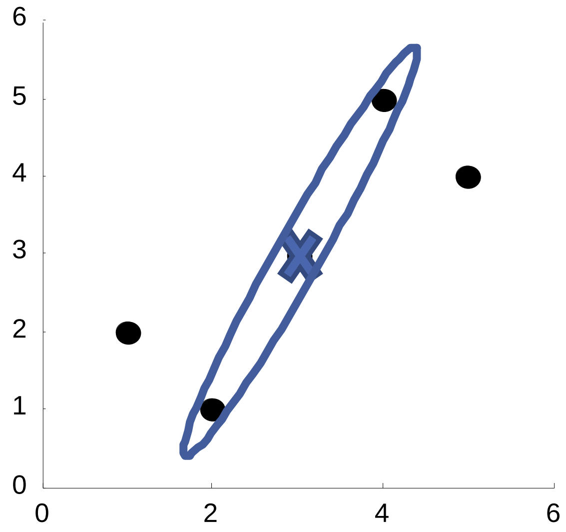
X	Y
1	2
2	1
3	3
4	5
5	4

$m_x = 3$ $m_y = 3$

$s^2_x = 2.5$ $s^2_y = 2.5$

$s_{xy} = 2$

$$\Sigma = \begin{bmatrix} s_x^2 & s_{xy} \\ s_{yx} & s_y^2 \end{bmatrix} = \begin{bmatrix} 2.5 & 2 \\ 2 & 2.5 \end{bmatrix}$$



PCA Example

$$det(\Sigma - \lambda I) = 0$$

$$det\left(\begin{bmatrix} 2.5 & 2 \\ 2 & 2.5 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right) = 0$$

$$det\left(\begin{bmatrix} 2.5 - \lambda & 2 \\ 2 & 2.5 - \lambda \end{bmatrix}\right) = 0$$

$$(2.5 - \lambda)(2.5 - \lambda) - (2)(2) = 0$$

$$\lambda^2 - 5\lambda + 2.25 = 0$$

$$(4.5 - \lambda)(0.5 - \lambda) = 0$$

$\lambda = 4.5$, or $\lambda = 0.5$

Eigenvalues

Eigenvectors

$$E = \begin{bmatrix} c \\ c \end{bmatrix}$$

$$(\Sigma - \lambda I)E = 0$$

$$\left(\begin{bmatrix} 2.5 & 2 \\ 2 & 2.5 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right) E = 0$$

$$\left(\begin{bmatrix} 2.5 & 2 \\ 2 & 2.5 \end{bmatrix} - \begin{bmatrix} 4.5 & 0 \\ 0 & 4.5 \end{bmatrix}\right) E = 0$$

$$\begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix} E = 0$$

$$\begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} E_x \\ E_y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$-2E_x + 2E_y = 0$$

$$2E_x - 2E_y = 0$$

$$-E_x + E_y = 0$$

$$E_x - E_y = 0$$

$$-E_x = -E_y$$

$$E_x = E_y$$

$\lambda = 0.5$

$$E = \begin{bmatrix} c \\ -c \end{bmatrix}$$

PCA Example

Choose "c" so that norm(E) = 1

$$E = \begin{bmatrix} c \\ c \end{bmatrix}$$

$$norm(E) = \sqrt{c^2 + c^2}$$

$$1 = \sqrt{c^2 + c^2}$$

$$1 = c^2 + c^2$$

$$1 = 2c^2$$

$$\frac{1}{2} = c^2$$

$$\sqrt{\frac{1}{2}} = c$$

$$c = 0.7071$$

$$E = \begin{bmatrix} c \\ -c \end{bmatrix}$$

$$norm(E) = \sqrt{c^2 + (-c)^2}$$

$$1 = \sqrt{c^2 + c^2}$$

$$1 = c^2 + c^2$$

$$1 = 2c^2$$

$$\frac{1}{2} = c^2$$

$$\sqrt{\frac{1}{2}} = c$$

$$c = 0.7071$$

$$E = \begin{bmatrix} c & c \\ c & -c \end{bmatrix}$$

$$E = \begin{bmatrix} 0.7071 & 0.7071 \\ 0.7071 & -0.7071 \end{bmatrix}$$

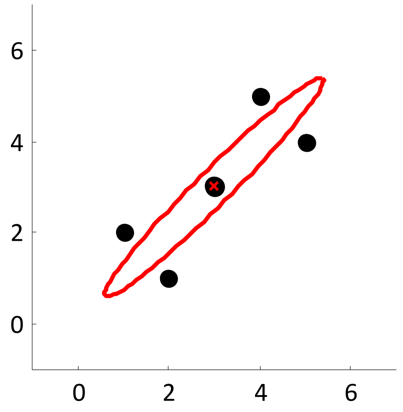
PCA Example

x	y
1	2
2	1
3	3
4	5
5	4

$m_x = 3 \quad m_y = 3$

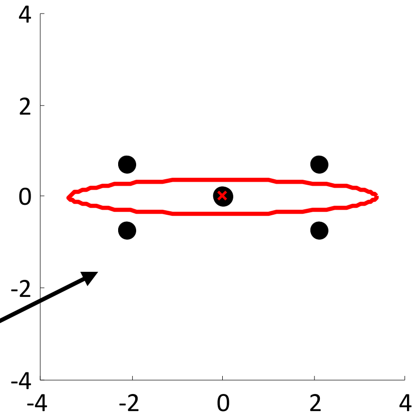
Subtract the mean

x	y
-2	-1
-1	-2
0	0
1	2
2	1



$$X * E = \begin{bmatrix} -2 & -1 \\ -1 & -2 \\ 0 & 0 \\ 1 & 2 \\ 2 & 1 \end{bmatrix} * \begin{bmatrix} 0.7071 & 0.7071 \\ 0.7071 & -0.7071 \end{bmatrix}$$

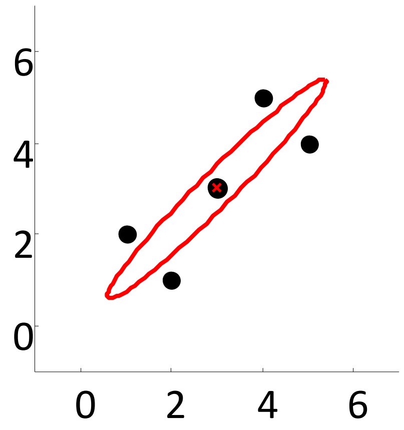
$$X * E = \begin{bmatrix} -2.1213 & -0.7071 \\ -2.1213 & 0.7071 \\ 0 & 0 \\ 2.1213 & -0.7071 \\ 2.1213 & 0.7071 \end{bmatrix}$$

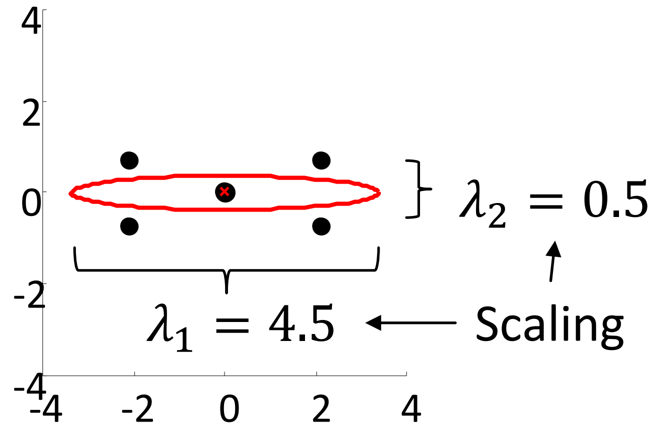


PCA Example

Rotation:

$$E = \begin{bmatrix} 0.7071 & 0.7071 \\ 0.7071 & -0.7071 \end{bmatrix}$$





PCA Example

- PCA gives you the scaling, rotating and translating matrices necessary to decorrelate your data.

PCA Example

- As correlations, PCA is linear
- Same assumptions as for Corr
- PCA on Cov(X,Y) or Corr(X,Y)?
- Scaling (Ants and Elephants)

Example of Factor Analysis

- To determine benefits from toothpaste
- Responses were obtained on 6 variables:

  V1: It is imp to buy toothpaste to prevent cavities

  V2: I like a toothpaste that gives shiny teeth

  V3: A toothpaste should strengthen your gums

  V4: I prefer a toothpaste that freshens breath

  V5: Prevention of tooth decay is not imp

  V6: The most imp consideration is attractive teeth
- Responses on a 7-pt scale (1=strongly disagree; 7=strongly agree)

Example Of Factor Analysis

RESPONDENT NUMBER	V1	V2	V3	V4	V5	V6
1	7.00	3.00	6.00	4.00	2.00	4.00
2	1.00	3.00	2.00	4.00	5.00	4.00
3	6.00	2.00	7.00	4.00	1.00	3.00
4	4.00	5.00	4.00	6.00	2.00	5.00
5	1.00	2.00	2.00	3.00	6.00	2.00
6	6.00	3.00	6.00	4.00	2.00	4.00
7	5.00	3.00	6.00	3.00	4.00	3.00
8	6.00	4.00	7.00	4.00	1.00	4.00
9	3.00	4.00	2.00	3.00	6.00	3.00
10	2.00	6.00	2.00	6.00	7.00	6.00
11	6.00	4.00	7.00	3.00	2.00	3.00
12	2.00	3.00	1.00	4.00	5.00	4.00
13	7.00	2.00	6.00	4.00	1.00	3.00
14	4.00	6.00	4.00	5.00	3.00	6.00
15	1.00	3.00	2.00	2.00	6.00	4.00
16	6.00	4.00	6.00	3.00	3.00	4.00
17	5.00	3.00	6.00	3.00	3.00	4.00
18	7.00	3.00	7.00	4.00	1.00	4.00
19	2.00	4.00	3.00	3.00	6.00	3.00
20	3.00	5.00	3.00	6.00	4.00	6.00
21	1.00	3.00	2.00	3.00	5.00	3.00
22	5.00	4.00	5.00	4.00	2.00	4.00
23	2.00	2.00	1.00	5.00	4.00	4.00
24	4.00	6.00	4.00	6.00	4.00	7.00
25	6.00	5.00	4.00	2.00	1.00	4.00
26	3.00	5.00	4.00	6.00	4.00	7.00
27	4.00	4.00	7.00	2.00	2.00	5.00
28	3.00	7.00	2.00	6.00	4.00	3.00
29	4.00	6.00	3.00	7.00	2.00	7.00
30	2.00	3.00	2.00	4.00	7.00	2.00

Table 19.1

Correlation Matrix

Variables	V1	V2	V3	V4	V5	V6
V1	1.000					
V2	-0.530	1.000				
V3	0.873	-0.155	1.000			
V4	-0.086	0.572	-0.248	1.000		
V5	-0.858	0.020	-0.778	-0.007	1.000	
V6	0.004	0.640	-0.018	0.640	-0.136	1.000

Results Of Principal Components Analysis

Bartlett's Test
Apprx. chi-square=111.3, df=15, significance=0.00
Kaiser-Meyer-Olkin msa=0.660

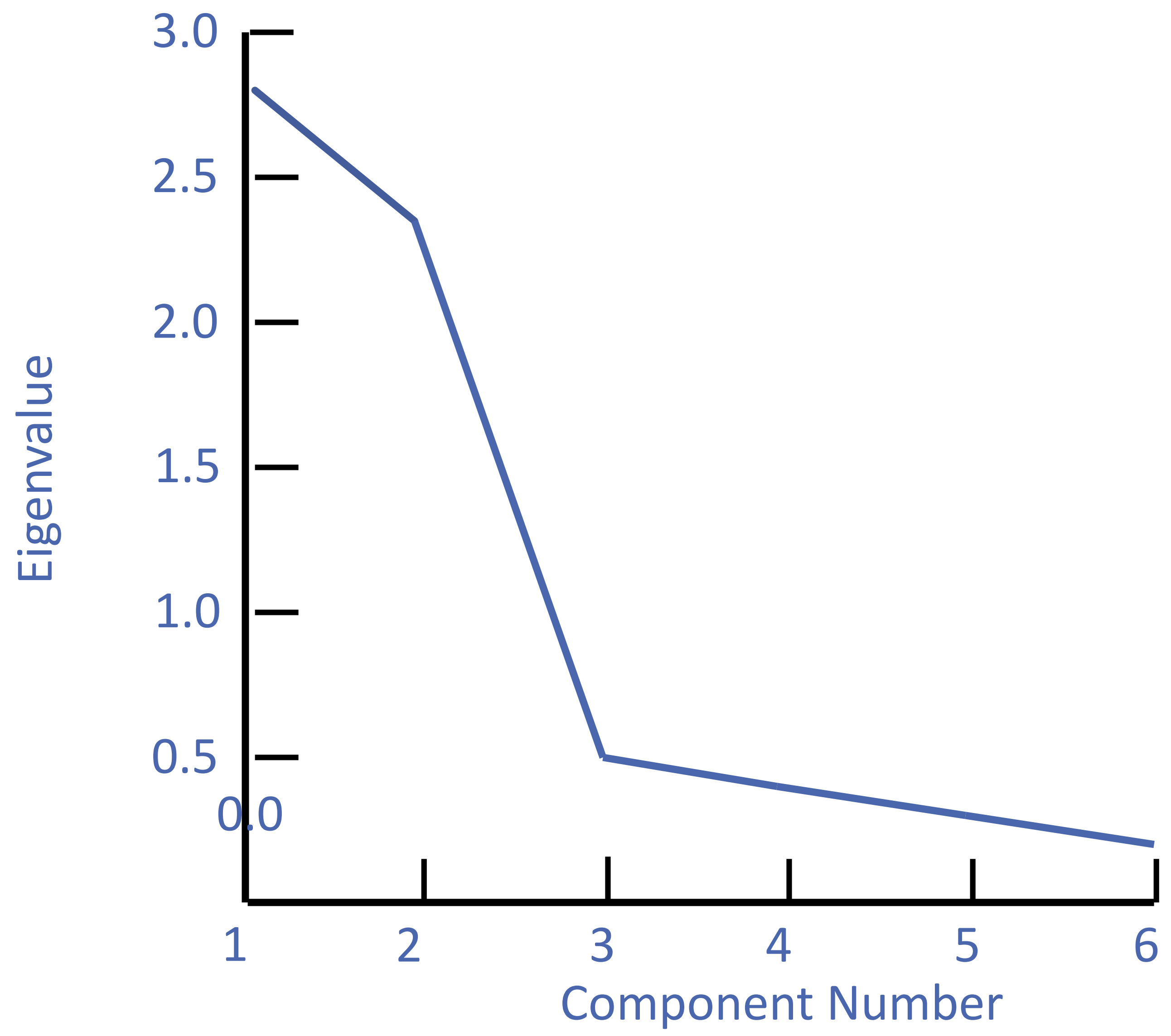
Communalities

Variables	Initial	Extraction
V1	1.000	0.926
V2	1.000	0.723
V3	1.000	0.894
V4	1.000	0.739
V5	1.000	0.878
V6	1.000	0.790

Initial Eigenvalues

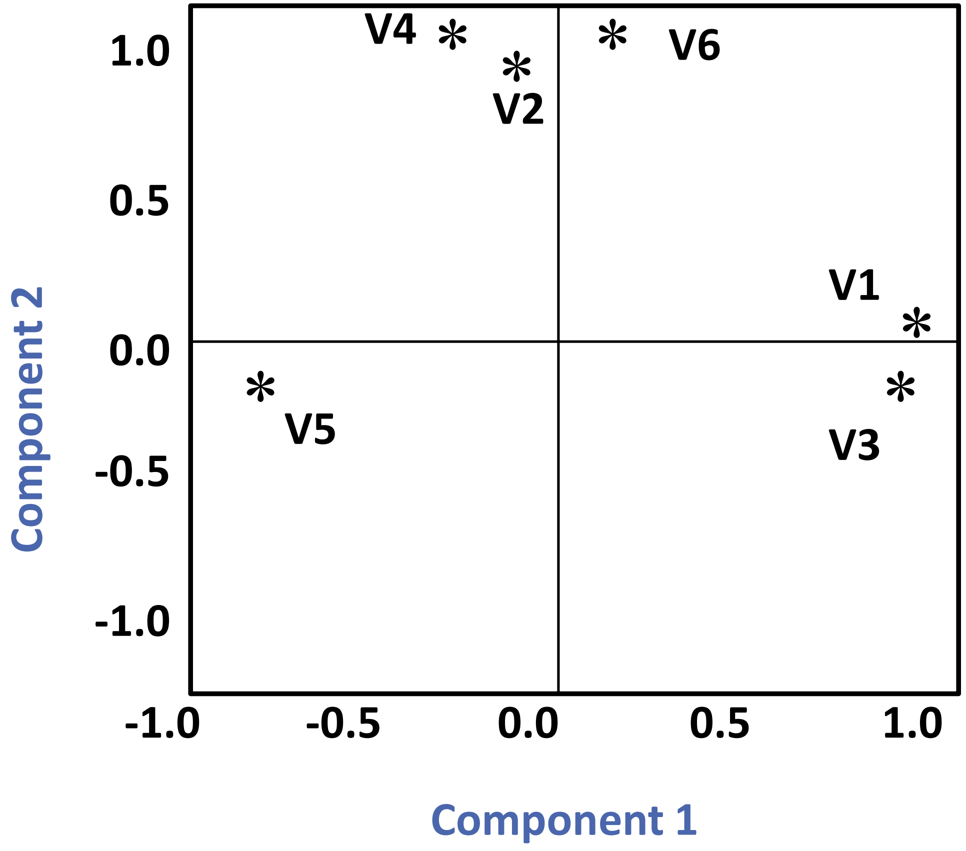
Component	Eigenvalue	% of variance	Cumulat. %
1	2.731	45.520	45.520
2	2.218	36.969	82.488
3	0.442	7.360	89.848
4	0.341	5.688	95.536
5	0.183	3.044	98.580
6	0.085	1.420	100.000

Screen Plot



Factor Loading Plot

Component Plot in Rotated Space



Rotated Component Matrix

	Component	
	1	2
V1	0.962	-2.66E-02
V2	-5.72E-02	0.848
V3	0.934	-0.146
V4	-9.83E-02	0.854
V5	-0.933	-8.40E-02
V6	8.337E-02	0.885

Another Example Of Factor Analysis

- To determine benefits from toothpaste
- Responses were obtained on 6 variables:

  V1: It is imp to buy toothpaste to prevent cavities

  V2: I like a toothpaste that gives shiny teeth

  V3: A toothpaste should strengthen your gums

  V4: I prefer a toothpaste that freshens breath

  V5: Prevention of tooth decay is not imp

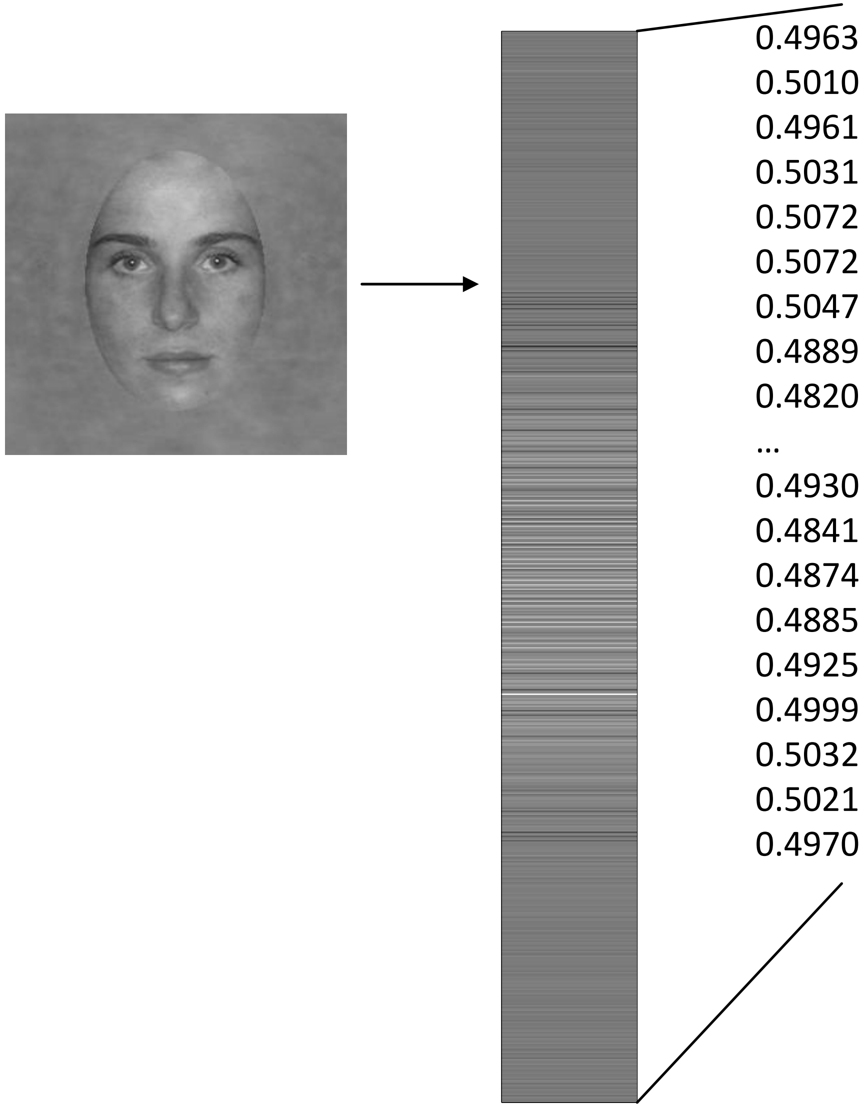
  V6: The most imp consideration is attractive teeth
- Responses on a 7-pt scale (1=strongly disagree; 7=strongly agree)

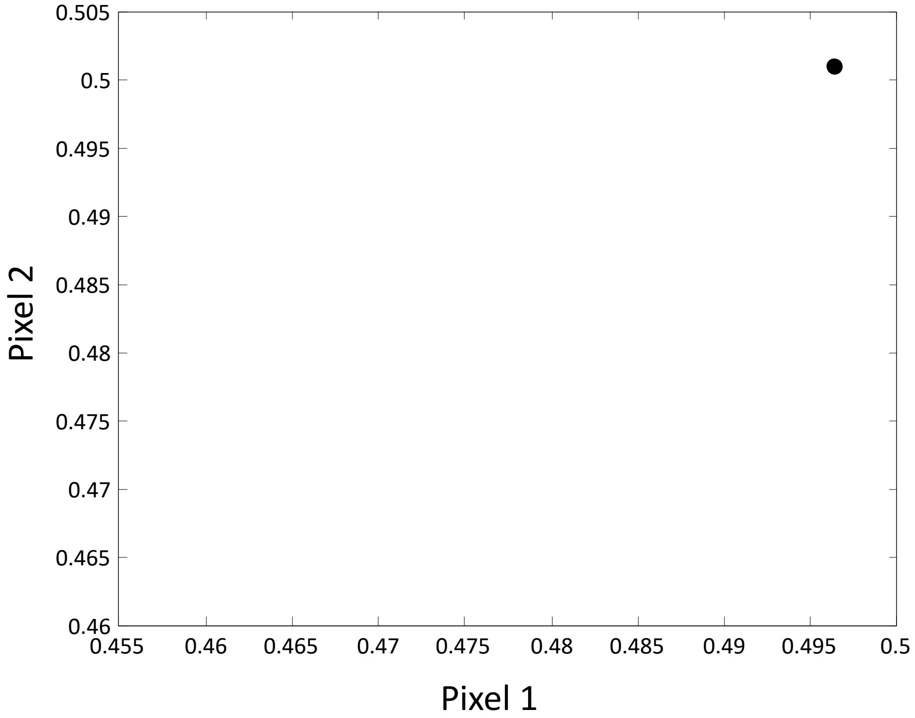
PCA Example

- PCA is not a statistical test
- The PCA “space” is determined by the arbitrary choice of variables, which may not be the best for your research question

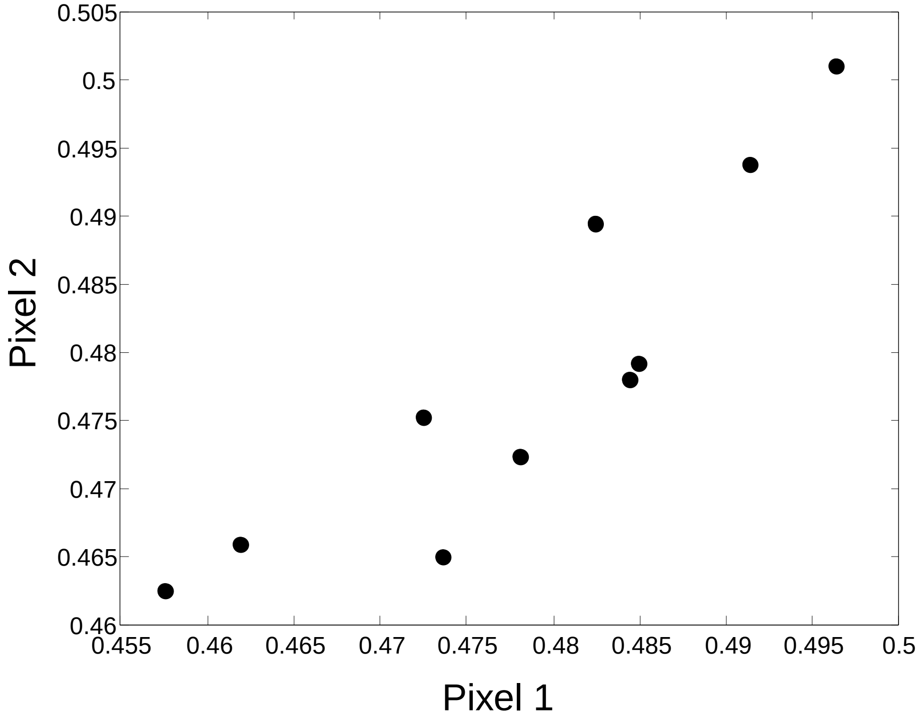
PCA Applications: Data Compression

PCA Applications: Data Compression





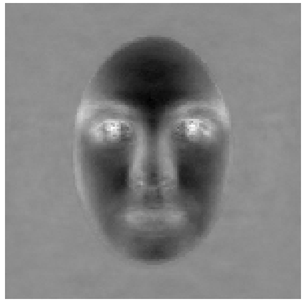
PCA Applications: Data Compression



PCA Applications: Data Compression

Eigenfaces

$\lambda_1 = 72.9242$



$\lambda_2 = 6.84$

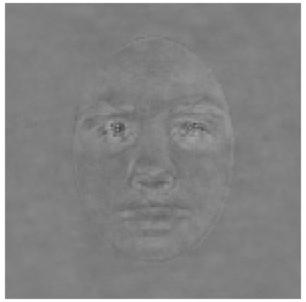
$\lambda_3 = 4.7808$

$\lambda_4 = 3.4802$

$\lambda_5 = 2.9461$

$\lambda_6 = 2.1323$

$\lambda_7 = 2.0361$



$\lambda_8 = 1.8905$

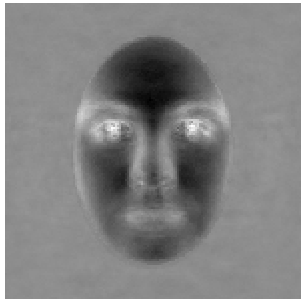
$\lambda_9 = 1.6516$

$\lambda_{10} = 1.318$

PCA Applications: Data Compression

Eigenfaces

% Explained Variance = 72.92



% Explained Variance = 6.84

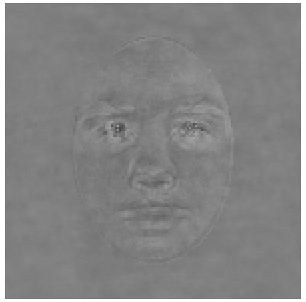
% Explained Variance = 4.78

% Explained Variance = 3.48

% Explained Variance = 2.95

% Explained Variance = 2.13

% Explained Variance = 2.04



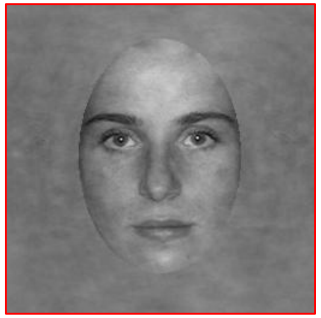
% Explained Variance = 1.89

% Explained Variance = 1.65

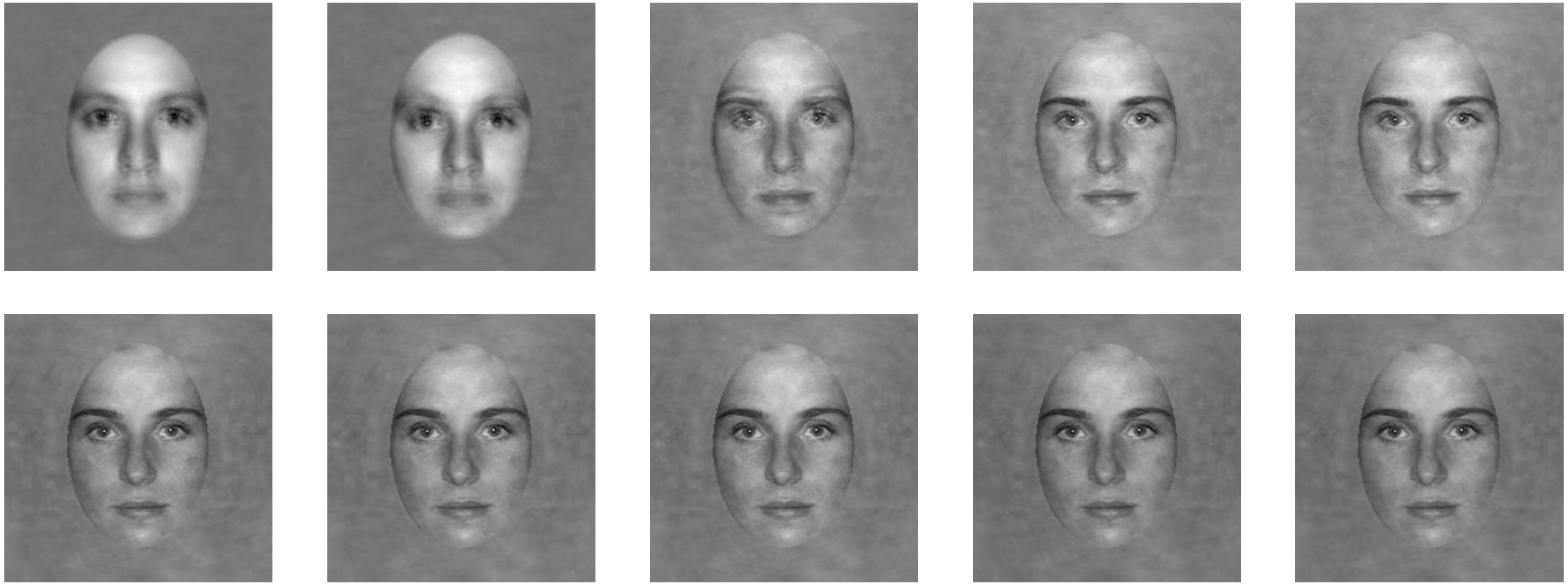
% Explained Variance = 1.32

PCA Applications: Data Compression

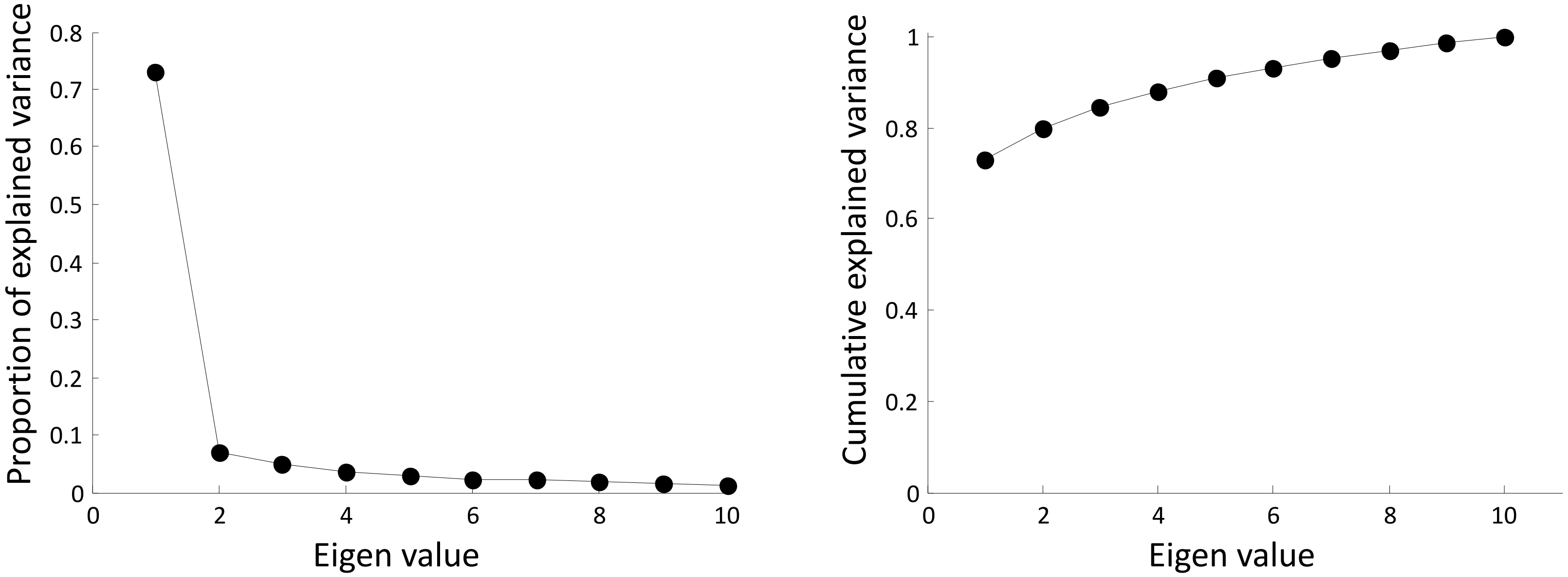
Reconstructing a single face using the eigenfaces:



$$Y = X * E$$
$$X = Y * E^{-1}$$
$$X' = Y(:,1:n) * E^{-1}(1:n,:)$$



PCA Applications: Data Compression



88% of the variability in the data is captured just by the first 4 eigenfaces!

PCA Applications: Data Compression

- Using PCA, 10 faces may be re-generated using only 4 eigenfaces!

Stats Tests Contn‘d

Ancoca
Manova
Model Based

Permutation
Bootstrap

Chi^2
Kolomogoroph-Smirnoff

ICA

FDR

Outlier detection

END Class 8a