Correlations,
Principal Component Analysis

Correlation is "normalized covariance"

- Also called:
  Pearson correlation coefficient

$\rho_{XY}=\sigma_{XY}/\sigma_X\sigma_Y$
is the covariance normalized to be $-1 \leq \rho_{XY} \leq 1$

Karl Pearson (1852– 1936)
English mathematician and biostatistician

Spearman rank correlation

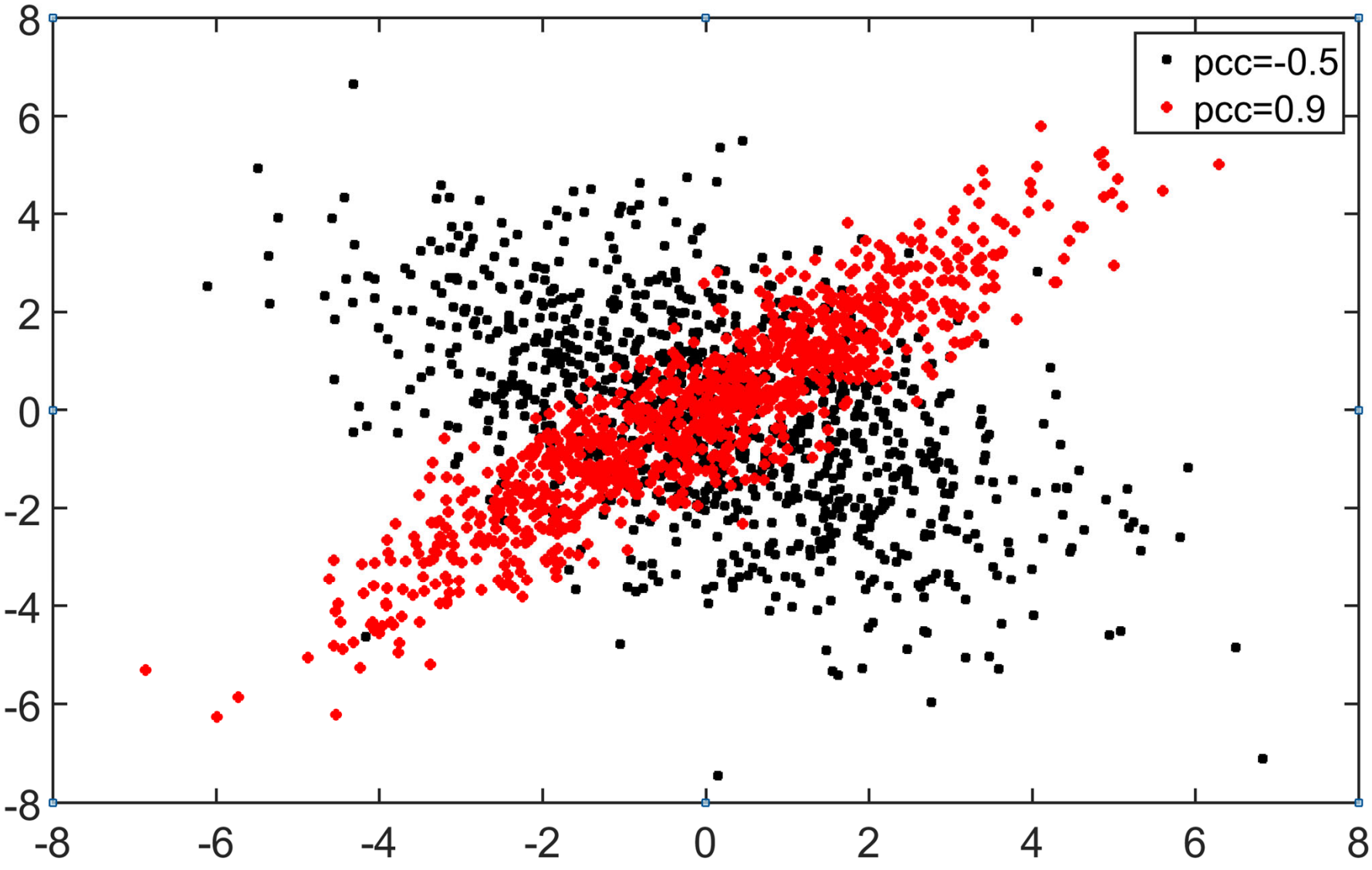
- Pearson correlation tests for linear relationship between X and Y
- Unlikely for variables with broad distributions → non-linear effects dominate
- Spearman correlation tests for any monotonic relationship between X and Y
- Calculate ranks (1 to n), $r_X(i)$ and $r_Y(i)$ of variables in both samples. Calculate Pearson correlation between ranks: Spearman(X,Y) = Pearson(r_X, r_Y)
- Ties: convert to fractions, e.g. tie for 6s and 7s place both get 6.5. This can lead to artefacts.
- If lots of ties: use Kendall rank correlation (Kendall tau)

Matlab exercise: Correlation/Covariation

- Generate a sample with Stats=100,000 of two Gaussian random variables r1 and r2 which have mean 0 and standard deviation 2 and are:
  - Uncorrelated
  - Correlated with correlation coefficient 0.9
  - Correlated with correlation coefficient -0.5
  - Trick: first make uncorrelated r1 and r2. Then make anew variable: r1mix=mix.*r2+(1-mix.^2)^0.5.*r1; where mix= corr. coeff.
- For each value of mix calculate covariance and correlation coefficient between r1mix and r2
- In each case make a scatter plot: plot(r1mix,r2,'k.');

Matlab exercise: Correlation/Covariation

```
1. Stats=100000;
2. r1=2.*randn(Stats,1);
3. r2=2.*randn(Stats,1);
4. disp('Covariance matrix='); disp(cov(r1,r2));
5. disp('Correlation=');disp(corr(r1,r2));
6. figure; plot(r1,r2,'k.');
7. mix=0.9; %Mixes r2 to r1 but keeps same variance
8. r1mix=mix.*r2+sqrt(1-mix.^2).*r1;
9. disp('Covariance matrix='); disp(cov(r1mix,r2));
10. disp('Correlation=');disp(corr(r1mix,r2));
11. figure; plot(r1mix,r2,'k.');
12. mix=-0.5; %REDO LINES 8-11
```


pcc=-0.5
pcc=0.9
8
6
4
2
0
-2
-4
-6
-8
-8
-6
-4
-2
0
2
4
6
8

Linear Functions of Random Variables

- A function of multiple random variables is itself a random variable.
- A function of random variables can be formed by either linear or nonlinear relationships. We will only work with linear functions.
- Given random variables $X_1, X_2, \ldots, X_p$ and constants $c_1, c_2, \ldots, c_p$

  $Y = c_1X_1 + c_2X_2 + \ldots + c_pX_p$ (5-24)

  is a linear combination of $X_1, X_2, \ldots, X_p$.

Mean & Variance of a Linear Function

$Y = c_1X_1 + c_2X_2 + \ldots + c_pX_p$

$$E(Y) = c_1E(X_1) + c_2E(X_2) + \ldots + c_pE(X_p) \qquad (5\text{-}25)$$

$$\mathrm{V}(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \ldots + c_p^2V(X_p) + 2\sum_{i<j}\sum c_ic_j \operatorname{cov}(X_iX_j) \qquad (5\text{-}26)$$

If $X_1, X_2, \ldots, X_p$ are independent, then $\operatorname{cov}(X_iX_j) = 0$,

$$V(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \ldots + c_p^2V(X_p) \qquad (5\text{-}27)$$

Example 5-31: Error Propagation

A semiconductor product consists of three layers. The variances of the thickness of each layer is 25, 40 and 30 nm^2. What is the variance of the finished product?

Answer:

IMPORTANT

$$X = X_1 + X_2 + X_3$$

Add variations not standard deviations

$$V(X) = \sum_{i=1}^{3} V(X_i) = 25 + 40 + 30 = 95 \text{ nm}^2$$

$$SD(X) = \sqrt{95} = 9.7 \text{ nm}$$

If adding SDs one would get $\sqrt{25}$ nm + $\sqrt{40}$ nm + $\sqrt{30}$ nm = 16.08 nm

IMPORTANT:

p independent identically distributed (iid) variables

Average $\bar{X} = \dfrac{X_1 + X_2 + X_3 + \ldots + X_p}{p}$

$$E(\bar{X}) = \frac{p \cdot E(X)}{p} = \frac{p \cdot \mu}{p} = \mu$$

$$V(\bar{X}) = \frac{p \cdot V(X)}{p^2} = \frac{p \cdot \sigma^2}{p^2} = \frac{\sigma^2}{p}$$

Standard deviation $(\bar{X}) = \sqrt{V(\bar{X})} = \dfrac{\sigma}{\sqrt{p}}$

Mean & Variance of an Average

If $\bar{X} = \dfrac{\left(X_1 + X_2 + \ldots + X_p\right)}{p}$ and $E\left(X_i\right) = \mu$

Then $E\left(\bar{X}\right) = \dfrac{p \cdot \mu}{p} = \mu$ (5-28a)

If the X_i are independent with $V\left(X_i\right) = \sigma^2$

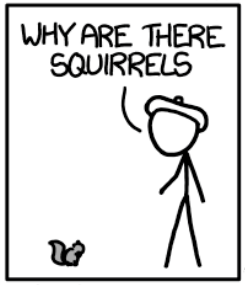
Then $V\left(\bar{X}\right) = \dfrac{p \cdot \sigma^2}{p^2} = \dfrac{\sigma^2}{p}$ (5-28b)

Credit: XKCD comics

QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT



WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

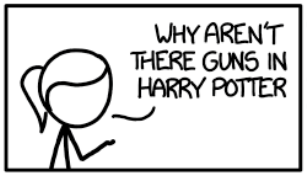
WHY ARE THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD

WHY IS GPS FREE

WHY ARE THERE FEMALE MR MIMES
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE



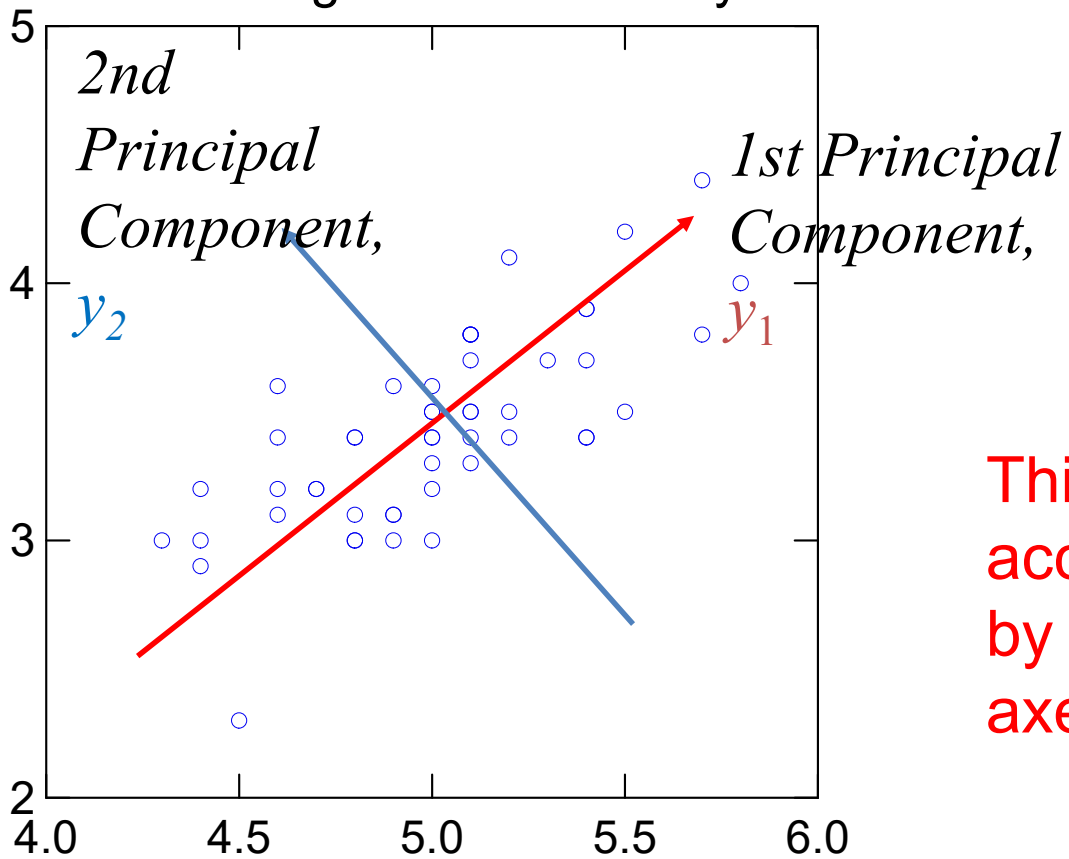
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS

Principal Component Analysis (PCA)

Trick: Rotate Coordinate Axes

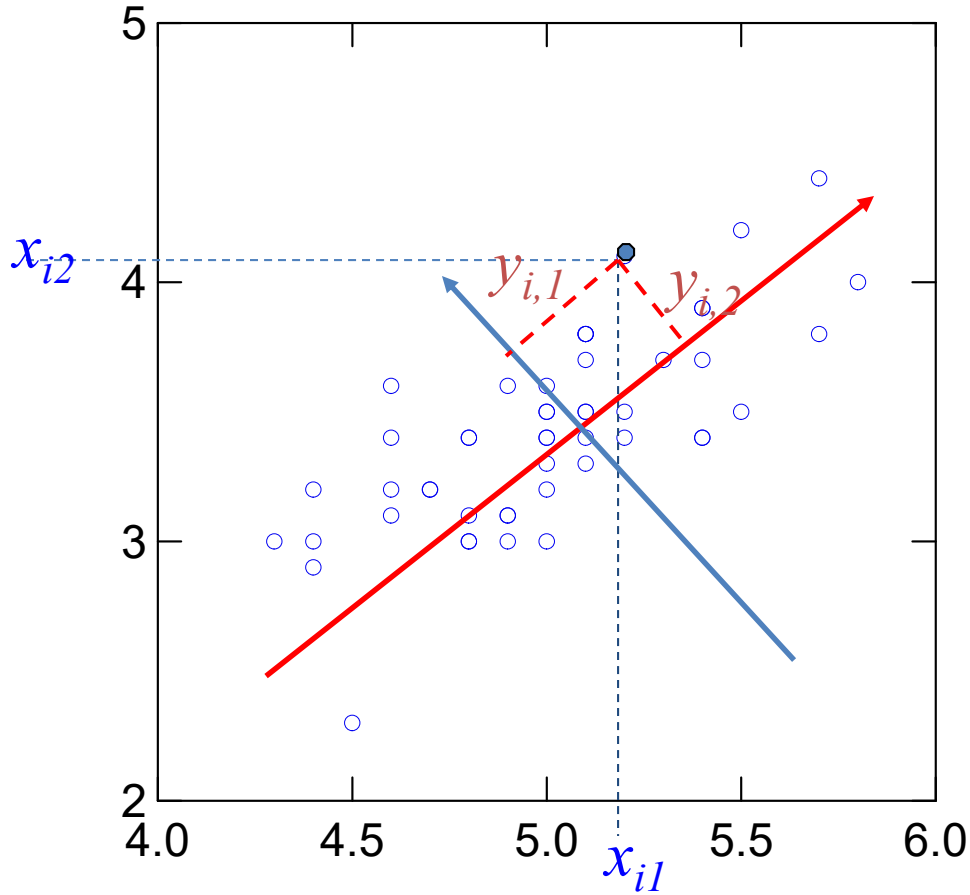
Suppose we have a population measured on p random variables $X_1,\ldots,X_p$. Note that these random variables represent the p-axes of the Cartesian coordinate system in which the population resides. Our goal is to develop a new set of p axes (linear combinations of the original p axes) in the directions of greatest variability:



This is accomplished by rotating the axes.

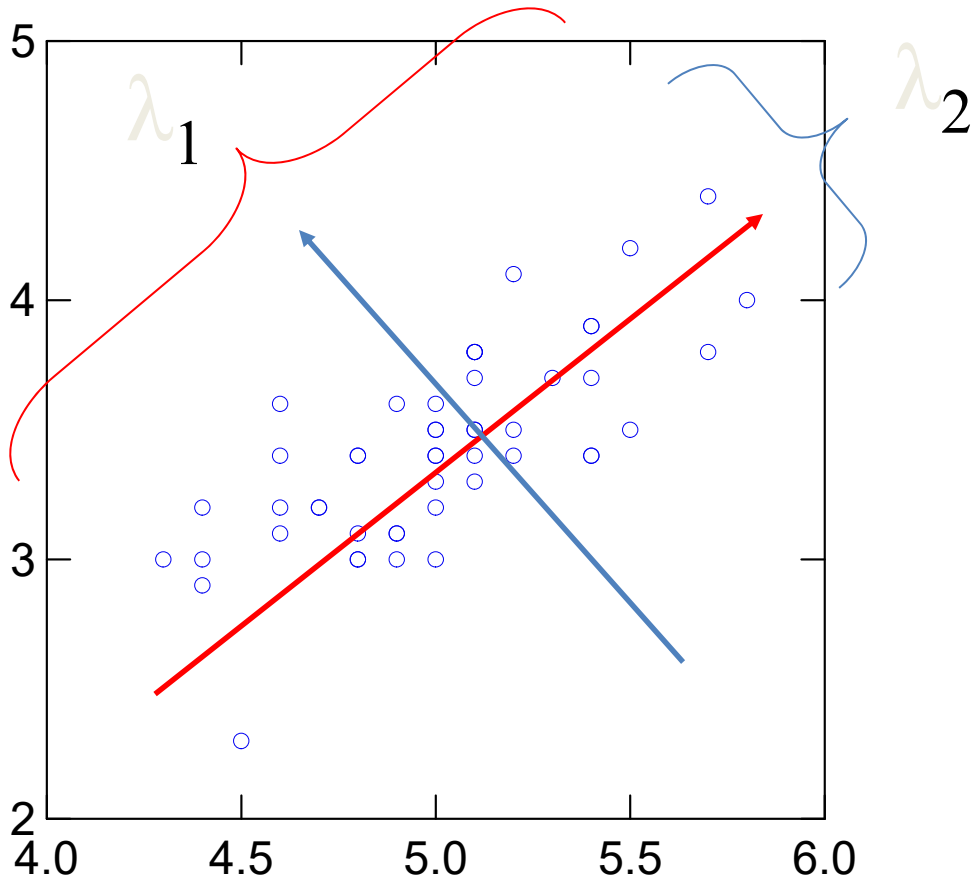
Adapted from slides by Prof. S. Narasimhan, "Computer Vision" course at CMU

PCA Scores



Adapted from slides by Prof. S. Narasimhan, “Computer Vision” course at CMU

PCA Eigenvalues and Eigenevectors



Adapted from slides by Prof. S. Narasimhan, “Computer Vision” course at CMU

PCA: *General*

From p original variables: $x_1, x_2, \ldots, x_p$:
I need to produce p new variables:
$y_1, y_2, \ldots, y_p$:

$$y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p$$
$$y_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2p}x_p$$
$$\ldots$$
$$y_p = a_{p1}x_1 + a_{p2}x_2 + \ldots + a_{pp}x_p$$

Answer: PCA diagonalize the p x p symmetric matrix of corr. coefficients

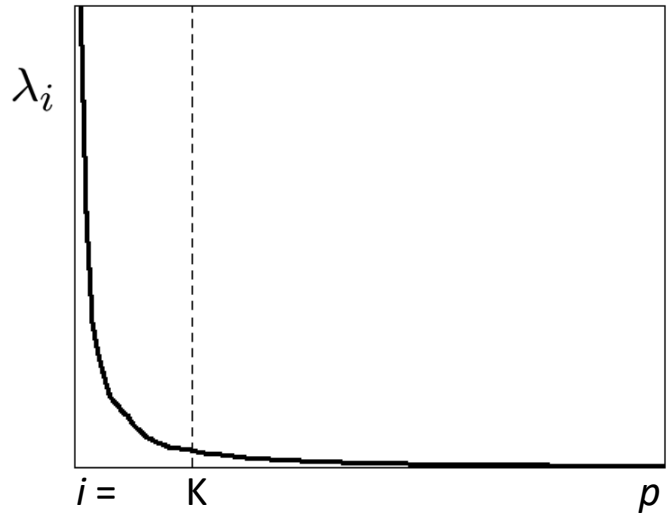
$$r_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

such that:
y_k's are uncorrelated (orthogonal)
y_1 explains as much as possible of original variance in data set
y_2 explains as much as possible of remaining variance
etc.

Adapted from slides by Prof. S. Narasimhan, "Computer Vision" course at CMU

Choosing the Dimension K

eigenvalues

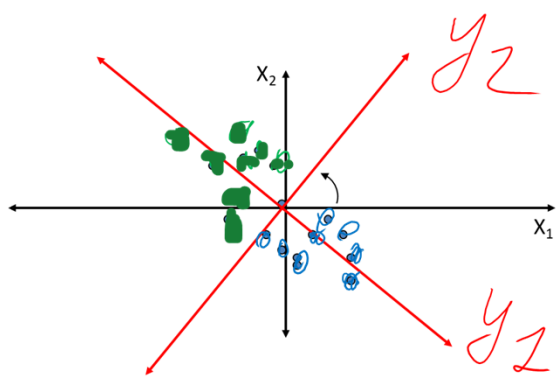


- How many eigenvectors to use?
- Look at the decay of the eigenvalues
  - the eigenvalue tells you the amount of variance "in the direction" of that eigenvector
  - ignore eigenvectors with low variance

Adapted from slides by Prof. S. Narasimhan, "Computer Vision" course at CMU

Applications of PCA

- Uses:
  - Data Visualization
  - Dimensionality Reduction
  - Data Classification



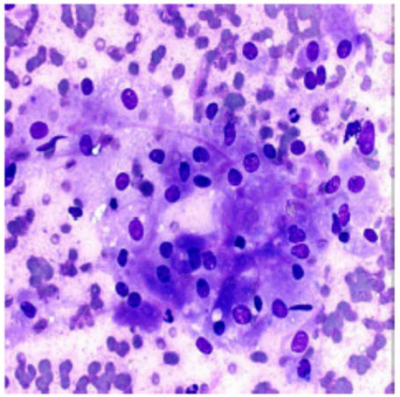
Examples:

- How to best present what is "interesting"?
- How many unique subsets (clusters, modules) are there in the sample?
- How are they similar / different
- What are the underlying factors that most influence the samples?
- Which measurements are best to differentiate between samples?
- Which subset does this new sample rightfully belong?

Adapted from slides by Prof. S. Narasimhan, "Computer Vision" course at CMU

Let's work with real cancer data!

- Data from Wolberg, Street, and Mangasarian (1994)
- Fine-needle aspirates = biopsy for breast cancer
- Black dots – cell nuclei. Irregular shapes/sizes may mean cancer
- 212 cancer patients and 357 healthy individuals (column 1)
- 30 other properties (see table)



Variable			
Radius (average distance from the center)	Col 2	Col 12	Col 22
Texture (standard deviation of gray-scale values)	Col 3	Col 13	Col 23
Perimeter	Col 4	Col 14	Col 24
Area	Col 5	Col 15	Col 25
Smoothness (local variation in radius lengths)	Col 6	Col 16	Col 26
Compactness (perimeter^2 / area - 1.0)	Col 7	Col 17	Col 27
Concavity (severity of concave portions of the contour)	Col 8	Col 18	Col 28
Concave points (number of concave portions of the contour)	Col 9	Col 19	Col 29
Symmetry	Col 10	Col 20	Col 30
Fractal dimension ("coastline approximation" - 1)	Col 11	Col 21	Col 31

Matlab exercise 1

- Download cancer data in cancer_wdbc.mat
- Data in the table X (569x30). First 357 patients are healthy. The remaining 569-357=212 patients have cancer.
- Calculate the correlation matrix of all-against-all variables: 30*29/2=435 correlations.
  Hint: look at the help page for corr
- Visualize 30x30 table of correlations using pcolor
- Plot the histogram of these 435 correlation coefficients