# Analiza wariancji i metody klasyfikacyjne
# Analysis of variance and classification methods

## Analiza Składowych Głównych
## Principal Component Analysis PCA

## lecture 5

*18 November 2019*

Ilona Anna Urbaniak (PK)

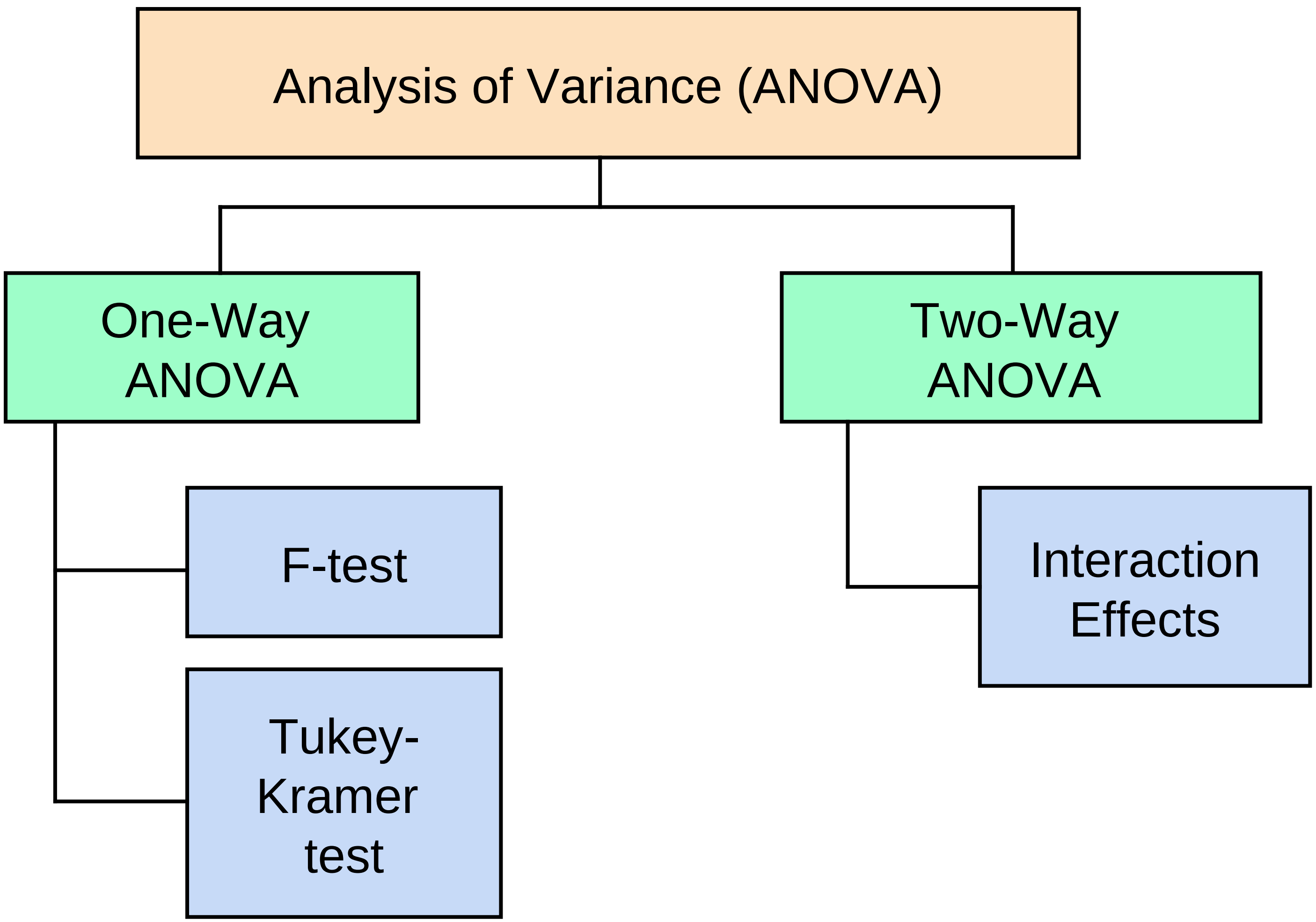
Marcin Wolter (IFJ PAN)

*e-mail: marcin.wolter@ifj.edu.pl, phone: 12 662 8024*

Slides: https://indico.ifj.edu.pl/event/271/

# Summary of ANOVA 1 & 2 way

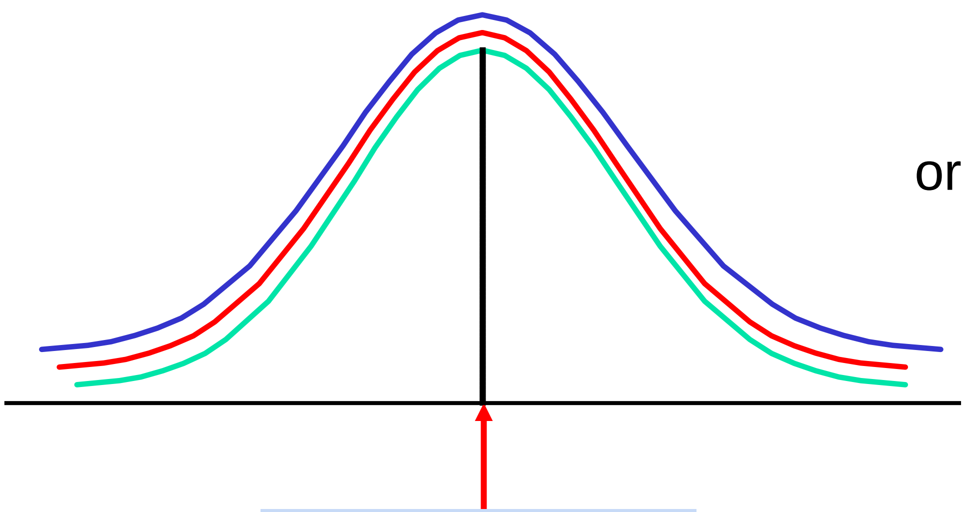
# What we have learned?



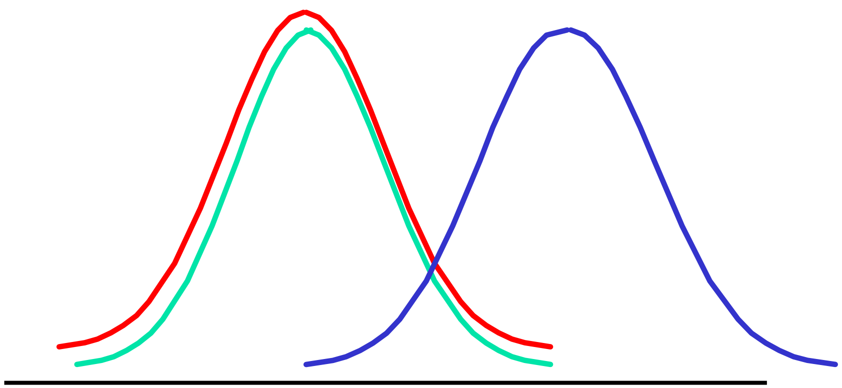
# One-Factor ANOVA

$$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_c$$

$$H_1: \text{Not all } \mu_i \text{ are the same}$$



or



$$\mu_1 = \mu_2 = \mu_3$$

$$\mu_1 = \mu_2 \neq \mu_3$$

# Two-Way ANOVA
# Sources of Variation

SST = SSA + SSB + SSAB + SSE

Degrees of Freedom:

SST
Total Variation

n - 1

SSA
Factor A Variation

r – 1

SSB
Factor B Variation

c – 1

SSAB
Variation due to interaction between A and B

(r – 1)(c – 1)

SSE
Random variation (Error)

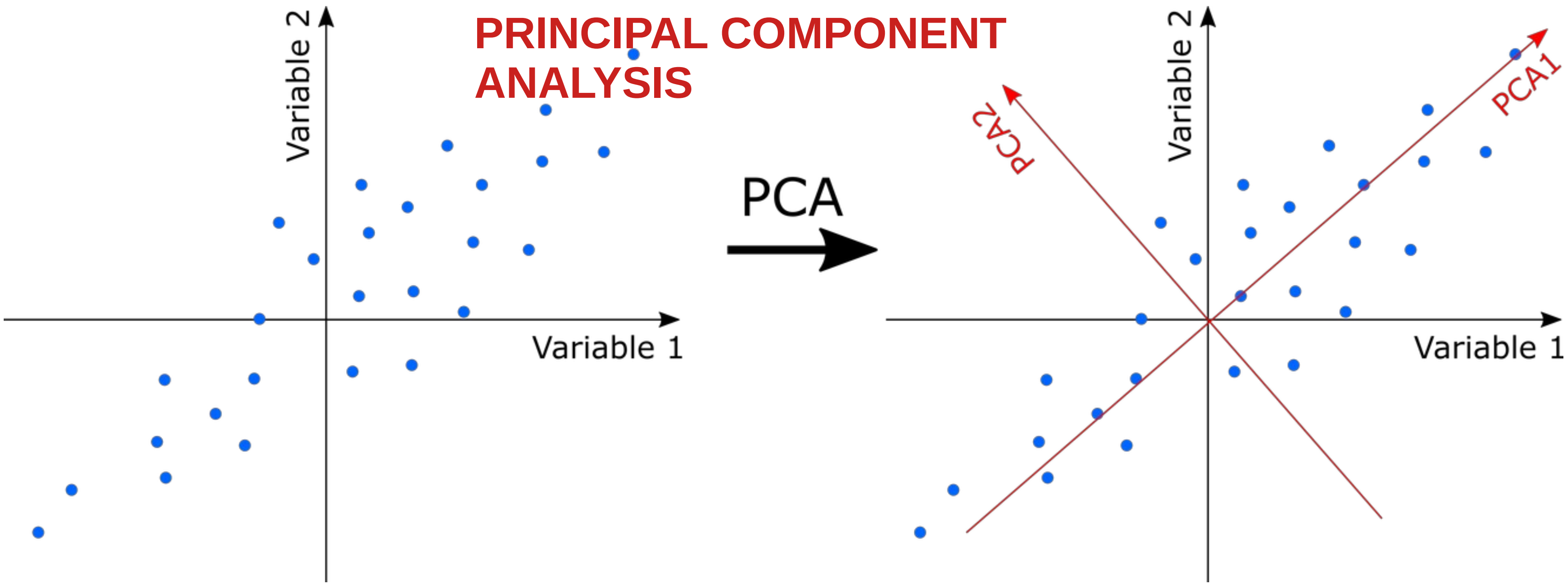
rc(n’ – 1)

# Principal Component Analysis PCA

# Problems

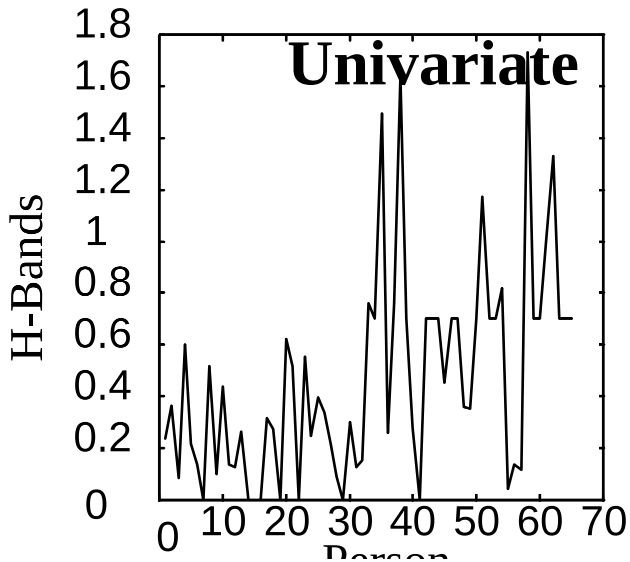
- Which variables are responsible for the highest variance?
- Can we build by linear transformation new variables and rank then according to the variance they create?
- If we have multidimensional data, can we visualize them in 2D using most discriminating variables out of a set of new variables?

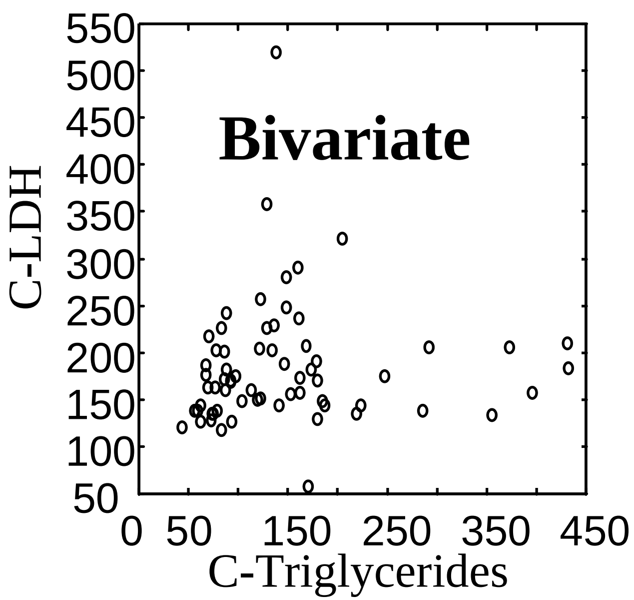
**PRINCIPAL COMPONENT ANALYSIS**

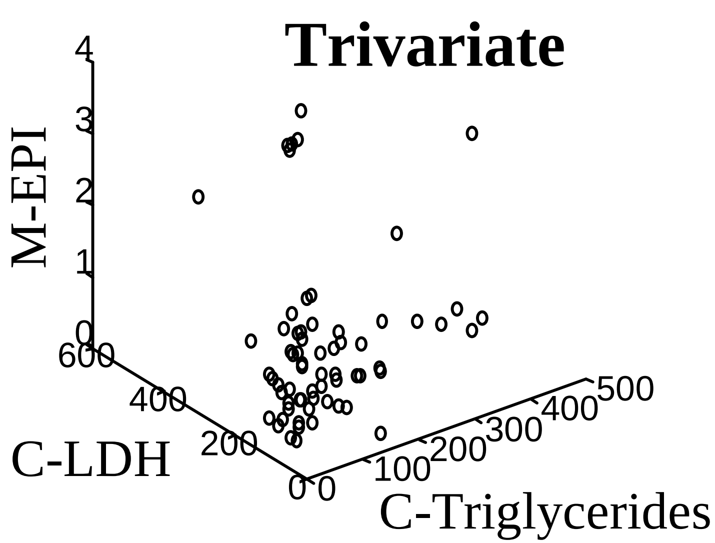



- PCA is sensitive to the scaling of the variables.

# Data Presentation




How to find the 'best' low dimension space that conveys maximum useful information? One answer: **Find "Principal Components"**

# The Goal

We wish to explain/summarize the underlying variance-covariance structure of a large set of variables through **a few** linear combinations of these variables.

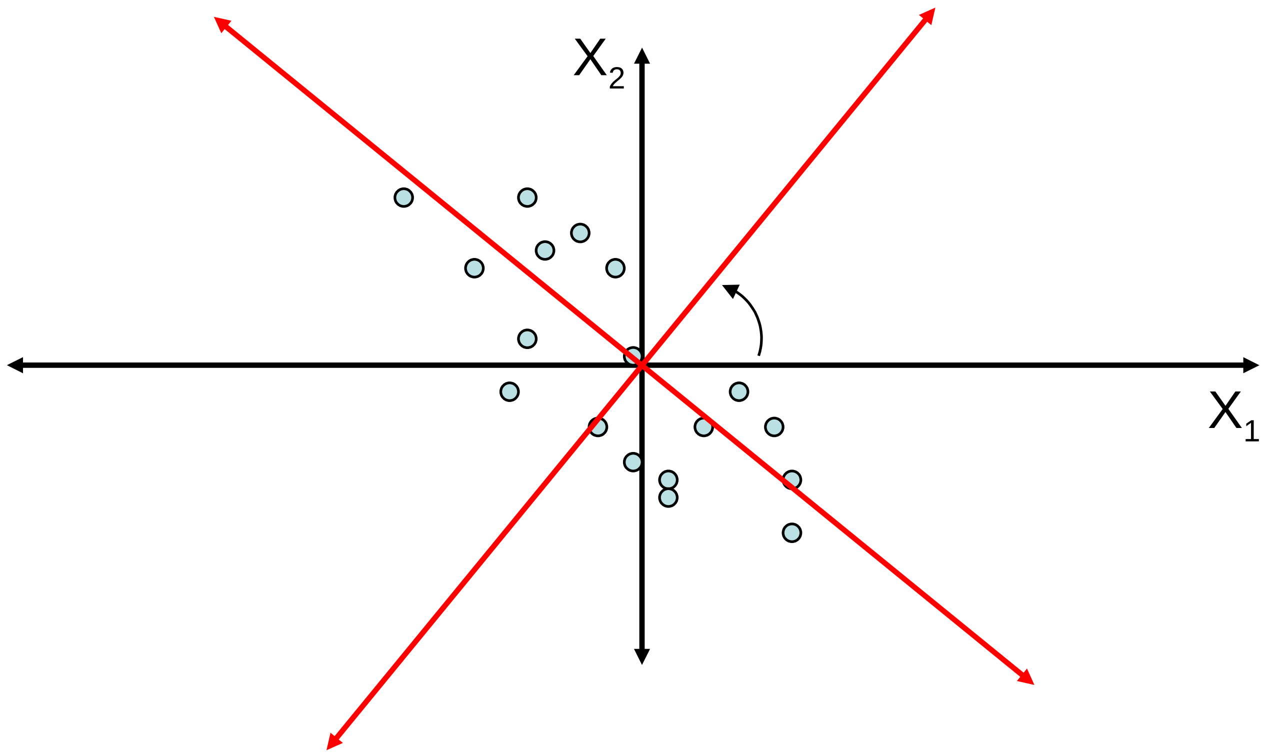
# Applications

- Uses:
  - Data Visualization
  - Data Reduction
  - Data Classification
  - Noise Reduction

- Examples:
  - How many unique “sub-sets” are in the sample?
  - How are they similar / different?
  - Which measurements are needed to differentiate?
  - How to best present what is “interesting”?
  - Which “sub-set” does this new sample rightfully belong?

# Trick: Rotate Coordinate Axes

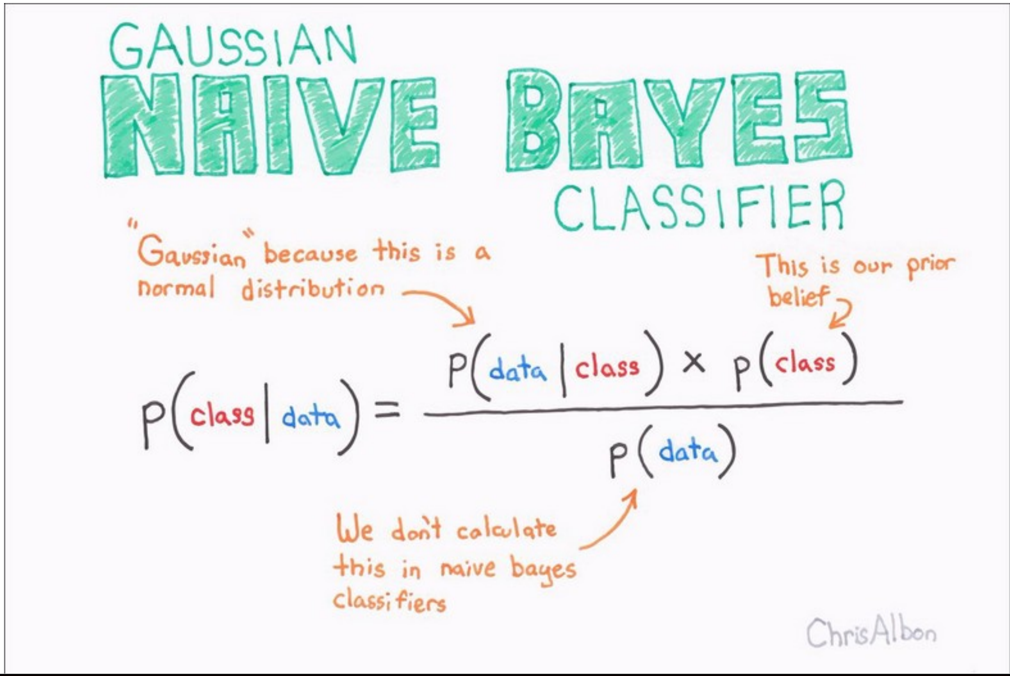
Suppose we have a population measured on p random variables $X_1,\ldots,X_p$. Note that these random variables represent the p-axes of the Cartesian coordinate system in which the population resides. Our goal is to develop a new set of p axes (linear combinations of the original p axes) in the directions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




This is accomplished by rotating the axes.

# Two examples

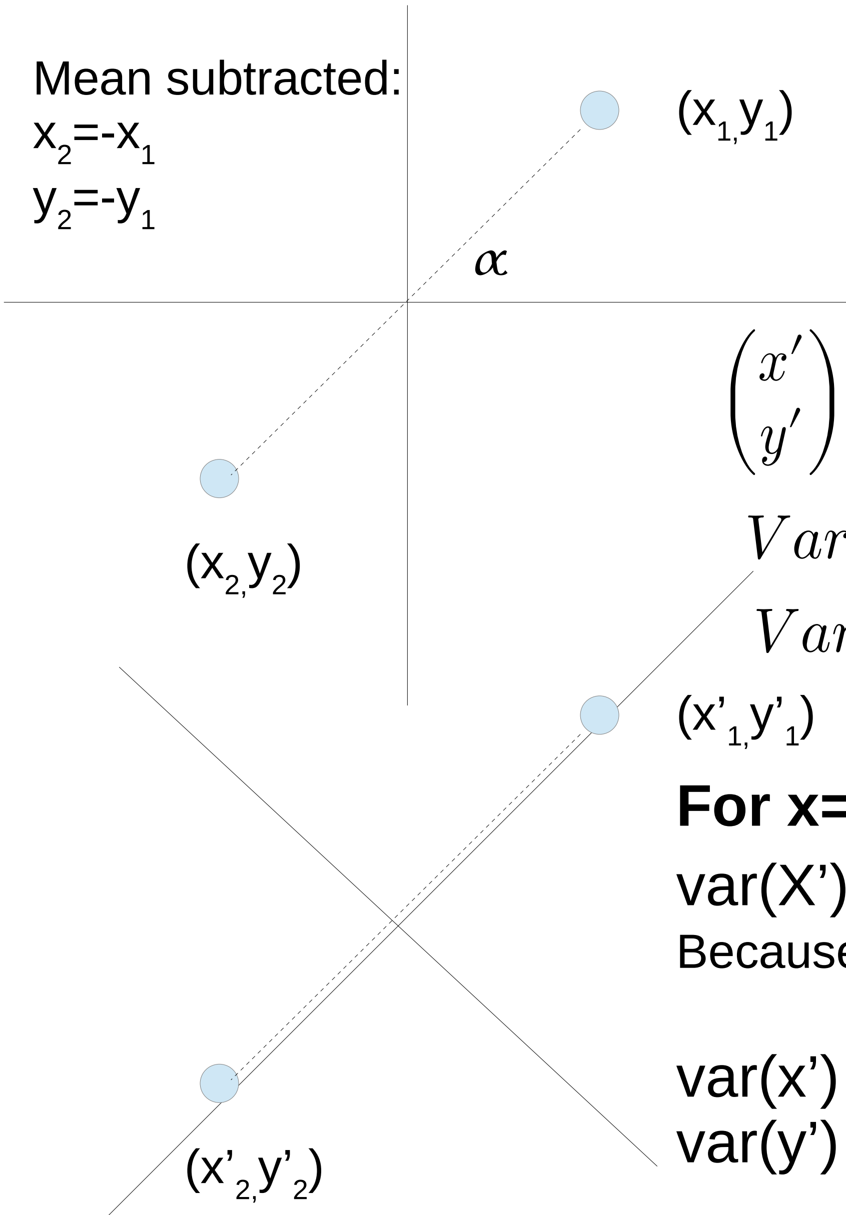
- principal_component_analysis.ipynb
  - Principal component analysis on famous IRIS dataset
  - PCA is done once manually and once using sklearn package
  - Sklearn is a machine learning package
- plot_digits_simple_classif.ipynb
  - Analize hand-written digits - 8x8 pixel maps
  - PCA performed on 64 input variables
  - Naive Bayes method used for classification on n first principal components
  - Digits visualized on 2D space



# Just two points

$$mean\ subtracted: x_1 = -x_2 = x$$
$$y_1 = -y_2 = y$$

Mean subtracted:
$x_2 = -x_1$
$y_2 = -y_1$




$$var(X) = 2x^2$$
$$var(Y) = 2y^2$$

After rotation by an angle $\alpha$:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x\cos(\alpha) + y\sin(\alpha) \\ x\sin(\alpha) - y\cos(\alpha) \end{pmatrix}$$

$$Var(X') = 2x'^2 = 2(x\cos(\alpha) + y\sin(\alpha))^2$$
$$Var(Y') = 2y'^2 = 2(-x\sin(\alpha) + y\cos(\alpha))^2$$

**For x=y    maximum var(X') at $\alpha=45^0$**

var(X') = $2*x^2(\cos(\alpha)+\sin(\alpha))^2 = 2*x^2(1+\sin(2\alpha))$

Because:

$$\sin(2\theta) = 2\sin\theta\cos\theta = \frac{2\tan\theta}{1+\tan^2\theta}$$

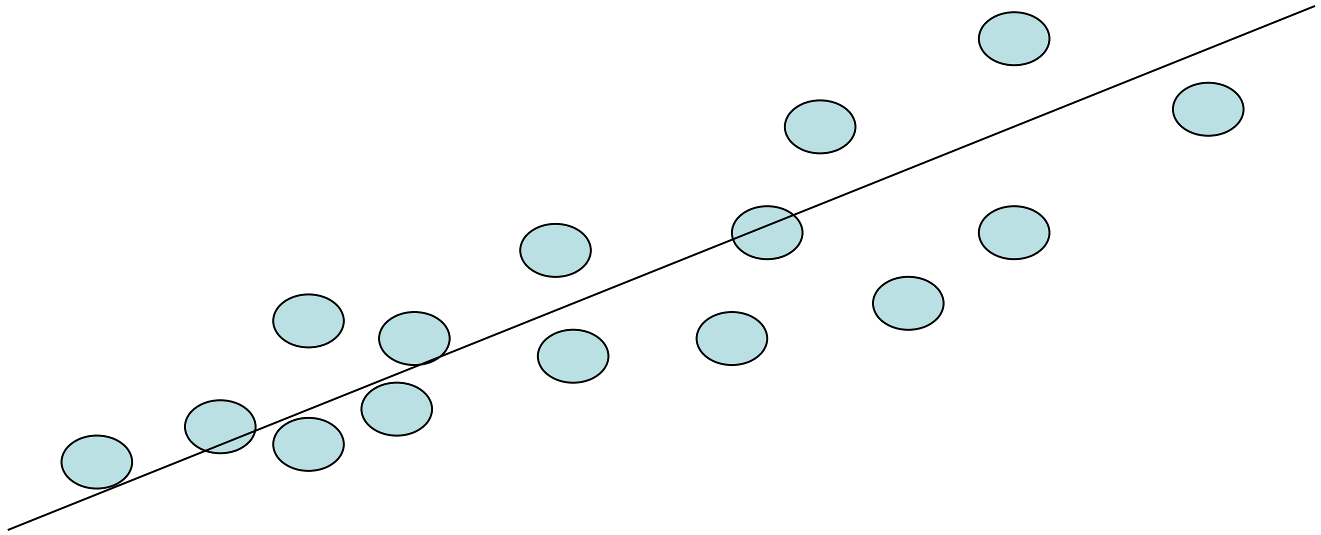
var(x') = 2 var(x)
var(y') = 0

# Algebraic Interpretation

- Given m points in a n dimensional space, for large n, how does one project on to a low dimensional space while preserving broad trends in the data and allowing it to be visualized?

# Algebraic Interpretation

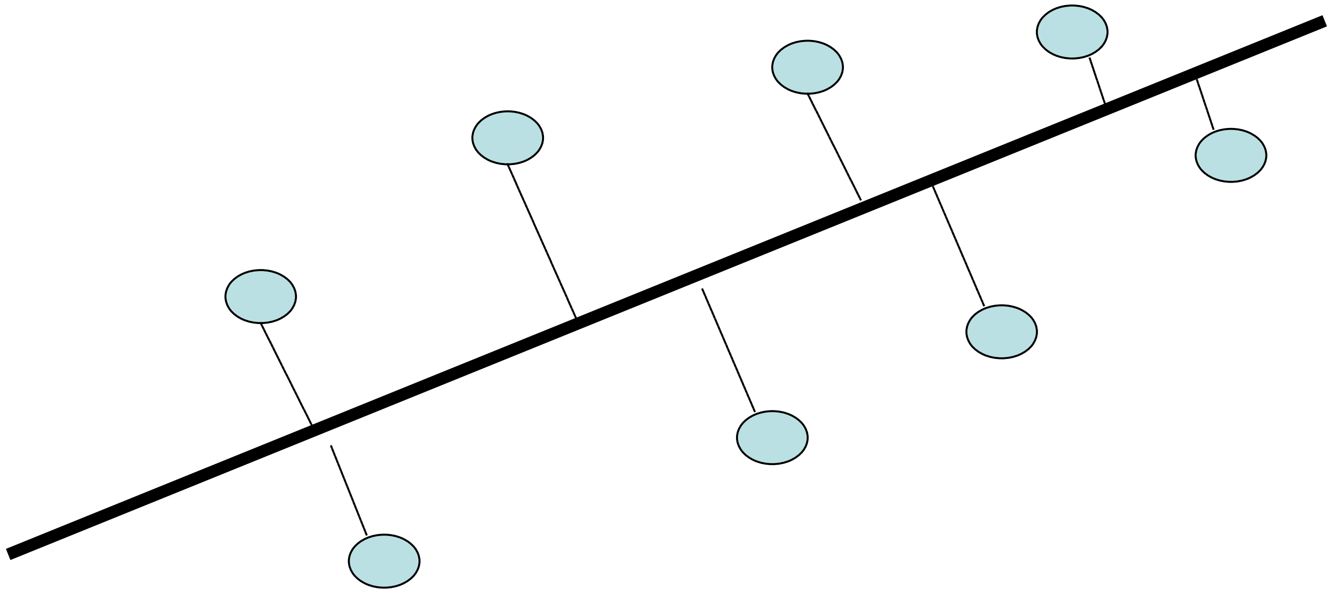
- Given m points in a n dimensional space, for large n, how does one project on to a 1 dimensional space?



- Choose a line that fits the data so the points are spread out well along the line

# Algebraic Interpretation – 2D

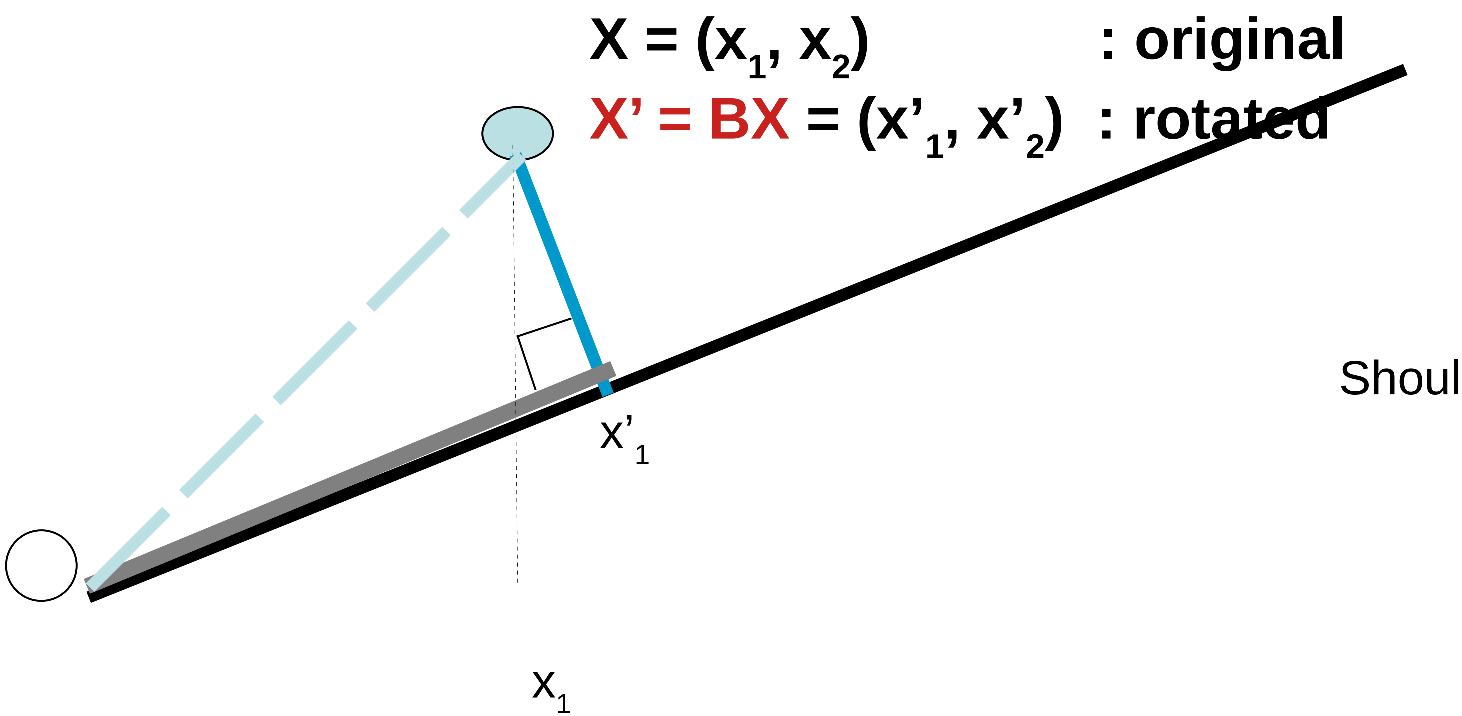
- Formally, minimize sum of squares of distances to the line.



- Why sum of squares? Because it allows fast minimization, assuming the line passes through 0

# Algebraic Interpretation

- Minimizing sum of squares of distances to the line is the same as maximizing the sum of squares of the projections on that line, thanks to Pythagoras.

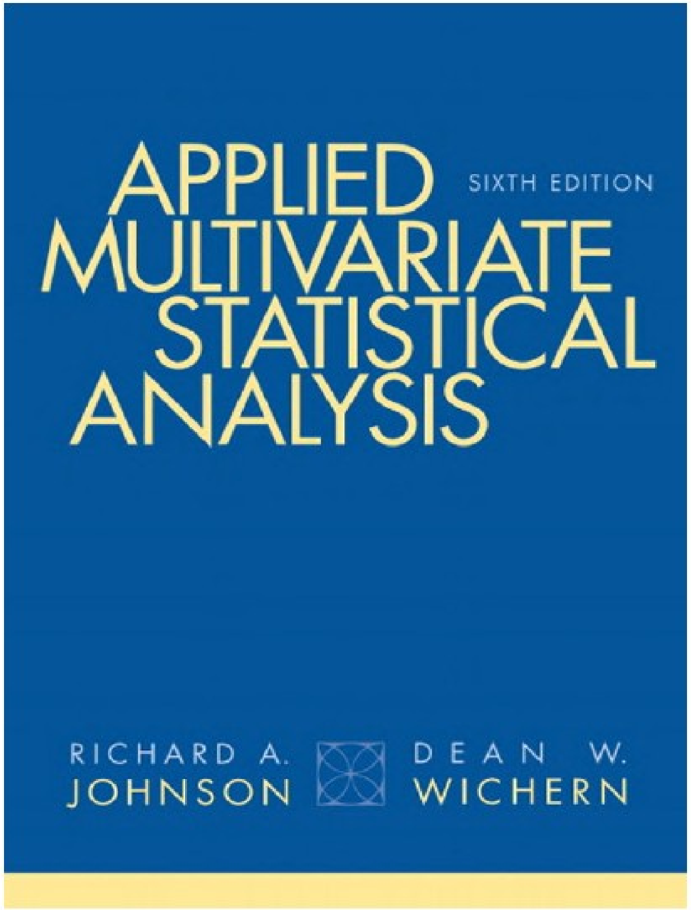


# Algebraic Interpretation

- How is the sum of squares of projection lengths expressed in algebraic terms?

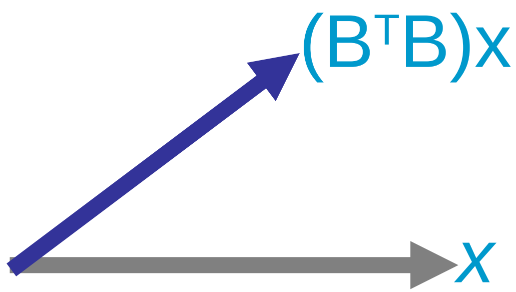
Nicely explained in:

http://docshare04.docshare.tips/files/12598/125983744.pdf

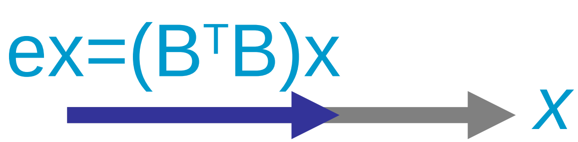


# Algebraic Interpretation

- $(B^T B)x$ points in some other direction in general




$x$ is an eigenvector and $e$ an eigenvalue if



# Algebraic Interpretation

- How many eigenvectors are there?
- For Real Symmetric Matrices

  – except in degenerate cases when eigenvalues repeat, there are n eigenvectors

    *$x_1 \ldots x_n$ are the eigenvectors*

    *$e_1 \ldots e_n$ are the eigenvalues*

  – all eigenvectors are mutually orthogonal and therefore form a new basis
    - Eigenvectors for distinct eigenvalues are mutually orthogonal
    - Eigenvectors corresponding to the same eigenvalue have the property that any linear combination is also an eigenvector with the same eigenvalue; one can then find as many orthogonal eigenvectors as the number of repeats of the eigenvalue.

# Algebraic Interpretation

- For matrices of the form $B^T B$
  - All eigenvalues are non-negative (try to show this?)

# Some mathematics

**Maximization of Quadratic Forms.** Let **B** (pxp) be a positive definite matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ and associated normalized eigenvectors are $\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_p}$ . Then:

$$max_{x\neq 0}\frac{\mathbf{x^T B x}}{\mathbf{x^T x}} = \lambda_1 \quad (attained\ when\ \mathbf{x} = \mathbf{e_1})$$
$$min_{x\neq 0}\frac{\mathbf{x^T B x}}{\mathbf{x^T x}} = \lambda_p \quad (attained\ when\ \mathbf{x} = \mathbf{e_p})$$

**Proof:** Let **P** (pxp) be the orthogonal matrix whose columns are the eigenvectors $\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_p}$ and $\mathbf{\Lambda}$ be the diagonal matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ along the main diagonal. Let $\mathbf{B^{1/2} = P\Lambda^{1/2}P^T}$ and $\mathbf{y=P^Tx}$ (sizes: $\mathbf{y}$(px1), $\mathbf{x}$(px1), $\mathbf{P^T}$(pxp)).
Consequently, $\mathbf{x\neq 0}$ implies $\mathbf{y\neq 0}$**.** Thus,

$$\frac{\mathbf{x^T B x}}{\mathbf{x^T x}} = \frac{\mathbf{x^T B^{1/2} B^{1/2} x}}{\mathbf{x^T P P^T x}}$$

$$= \frac{\mathbf{x^T P \Lambda^{1/2} P^T P \Lambda^{1/2} P^T x}}{\mathbf{y^T y}} = \frac{\mathbf{y^T \Lambda y}}{\mathbf{y^T y}}$$

$$= \frac{\sum_{i=1}^{p} \lambda_i y_i^2}{\sum_{i=1}^{p} y_i^2} \leq \lambda_1 \frac{\sum_{i=1}^{p} y_i^2}{\sum_{i=1}^{p} y_i^2} = \lambda_1$$

# So some calculations

- **Σ** – covariance matrix of a data **X**
- **Σ** has the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ and associated eigenvectors are $\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_p}$ .
- **X' = BX** – transformation of X to the new coordinate system
- thus covariance $Cov(x'_1) = Cov(B_{11}x_1+\ldots+B_{1p}x_p) = B_1^T\boldsymbol{\Sigma}B_1$, where $B_1=(B_{11},B_{12},\ldots,B_{1p})$

We know that $$max_{x\neq 0}\frac{\mathbf{x^TBx}}{\mathbf{x^Tx}} = \lambda_1 \quad (attained\ when\ \mathbf{x}=\mathbf{e_1})$$

So: $$max_{B_1\neq 0}\frac{\mathbf{B_1^T}\Sigma\mathbf{B_1}}{\mathbf{B_1^TB_1}} = \lambda_1 \quad B_1\ eigenvector\ of\ \Sigma,\ \lambda_1\ eigenvalue)$$

$$\lambda_1 = \frac{e_1^T\Sigma e_1}{e_1^Te_1} = e_1^T\Sigma e_1 = Var(X_1')$$

Def. of λ

$e_1^Te_1=1$

**What we wanted to show!**

**Max. variance of $X_1' = \lambda_1$ - 1st eigenvalue of covariance matrix Σ,**

**The 1st PCA axis is the eigenvector $e_1$ of covariance matrix Σ**

# Just two points again

- Try to do it using matrix calculations

$$Data: B = \begin{vmatrix} x & y \\ -x & -y \end{vmatrix}$$

$$Covariance: Cov(B) = \frac{1}{N} B^T B = \frac{1}{2} \begin{vmatrix} 2x^2 & 2xy \\ 2xy & 2y^2 \end{vmatrix} = \begin{vmatrix} x^2 & xy \\ y^2 & xy \end{vmatrix}$$

$$Eigenvalues: Det(Cov(B) - \lambda I) = Det\left(\begin{vmatrix} x^2 - \lambda & xy \\ xy & y^2 - \lambda \end{vmatrix}\right) = 0$$

$$(x^2 - \lambda)(y^2 - \lambda) - x^2 y^2 = 0 \Longrightarrow \lambda_1 = x^2 + y^2, \ \lambda_2 = 0$$

Eigenvalues are: $\lambda_1 = x_1^2 + y_1^2$, $\lambda_2 = 0$

These eigenvalues are our two variances!

Corresponding eigenvectors:
(modulo normalization)

$$e_1 = \begin{pmatrix} 1/y \\ 1/x \end{pmatrix} \qquad e_2 = \begin{pmatrix} 1/x \\ -1/y \end{pmatrix}$$

# PCA: *General*

From *k* original variables: $x_1, x_2, \ldots, x_k$:

Produce *k* new variables: $y_1, y_2, \ldots, y_k$:

$$y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1k}x_k$$

$$y_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2k}x_k$$

...

$$y_k = a_{k1}x_1 + a_{k2}x_2 + \ldots + a_{kk}x_k$$

# PCA: *General*

From $k$ original variables: $x_1, x_2, \ldots, x_k$:

Produce $k$ new variables: $y_1, y_2, \ldots, y_k$:

$y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1k}x_k$

$y_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2k}x_k$

...

$y_k = a_{k1}x_1 + a_{k2}x_2 + \ldots + a_{kk}x_k$

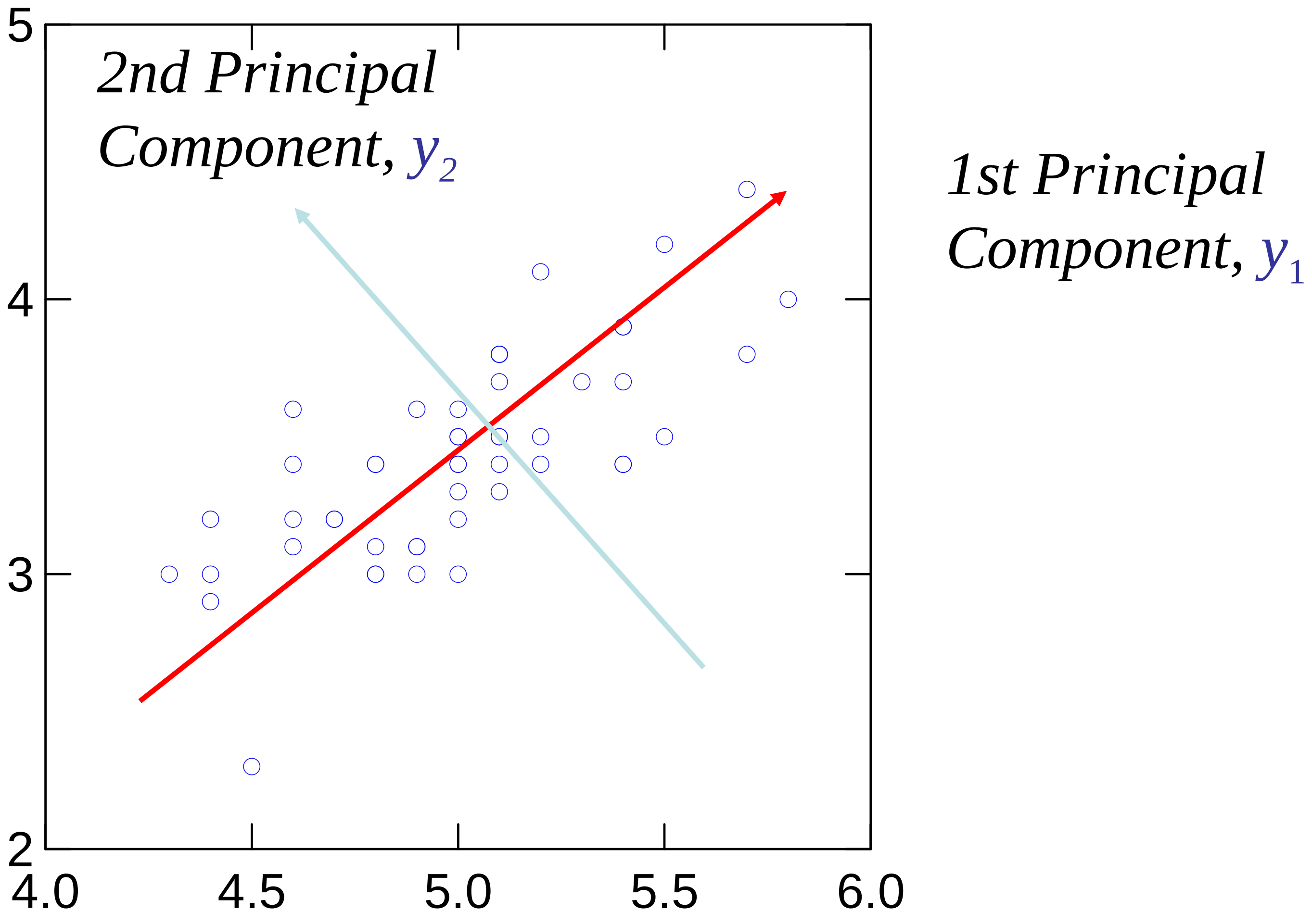
*such that:*

$y_k$'s are uncorrelated (orthogonal)

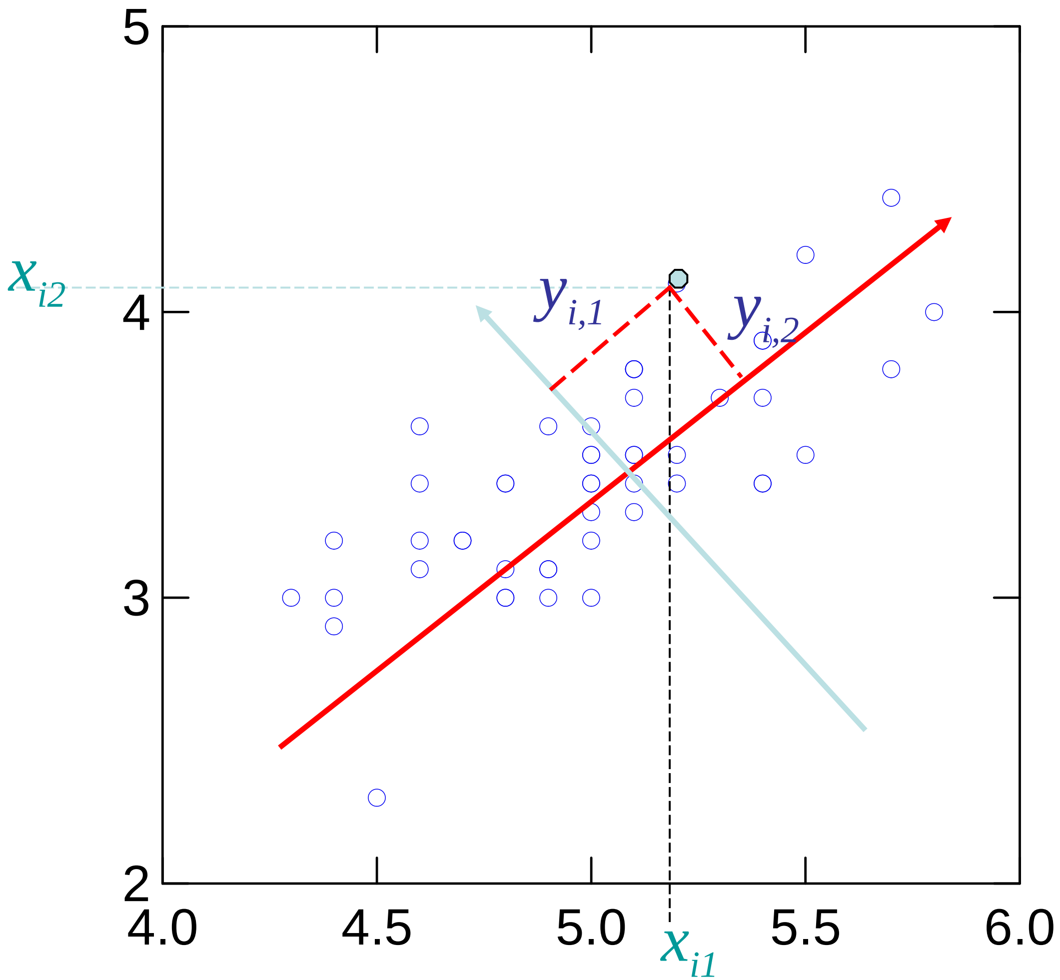
$y_1$ explains as much as possible of original variance in data set

$y_2$ explains as much as possible of remaining variance

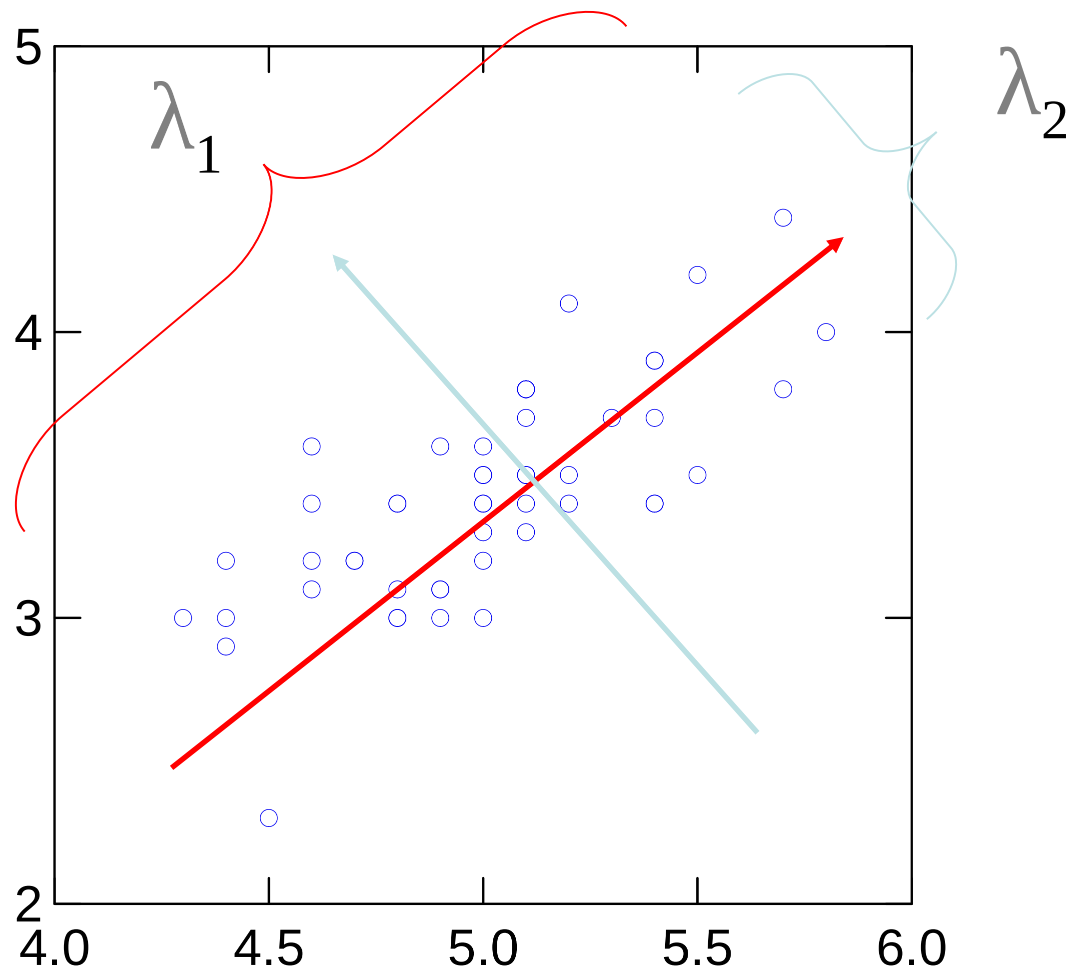
etc.


2nd Principal Component, $y_2$
1st Principal Component, $y_1$
5
4
3
2
4.0
4.5
5.0
5.5
6.0

# PCA Scores

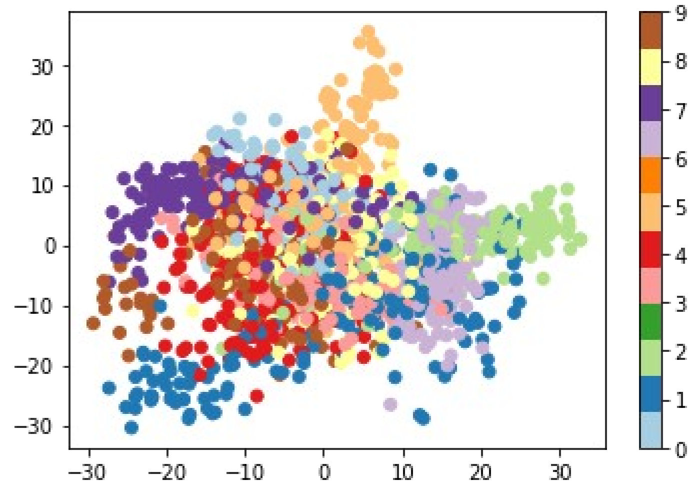


# PCA Eigenvalues

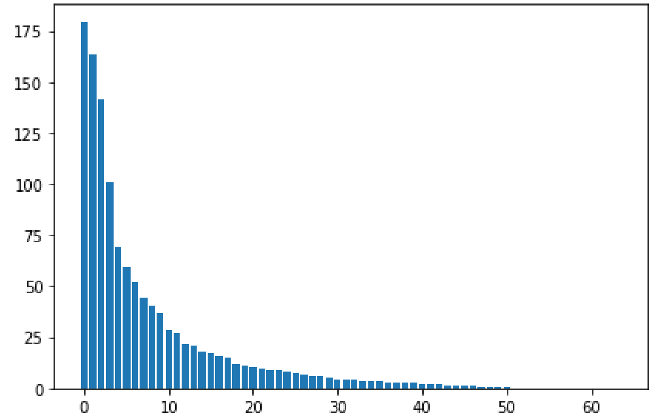


# Summary

- PCA helps to visualize the multidimensional data in 2D:




- Few components can explain most of the variance:




- Further analysis / classification might be much easier.