Data Mining
(Mining Knowledge from Data)

Statistic Methods for Data Mining

Marcel Jiřina, Pavel Kordík

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

FIT

Lecture

1) The mean value: mean/median
2) Extremes
3) Correlation
4) Principal Component Analysis (PCA)

Arithmetic mean

$$\bar{x} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

arithmetic mean

number of samples

class/value of sample

Median

- Sort (order) samples in increasing/decreasing order:

$$Median = \left(\frac{N+1}{2}\right) sample$$

- =number of samples

Demonstration – original image

Demonstration – noise added

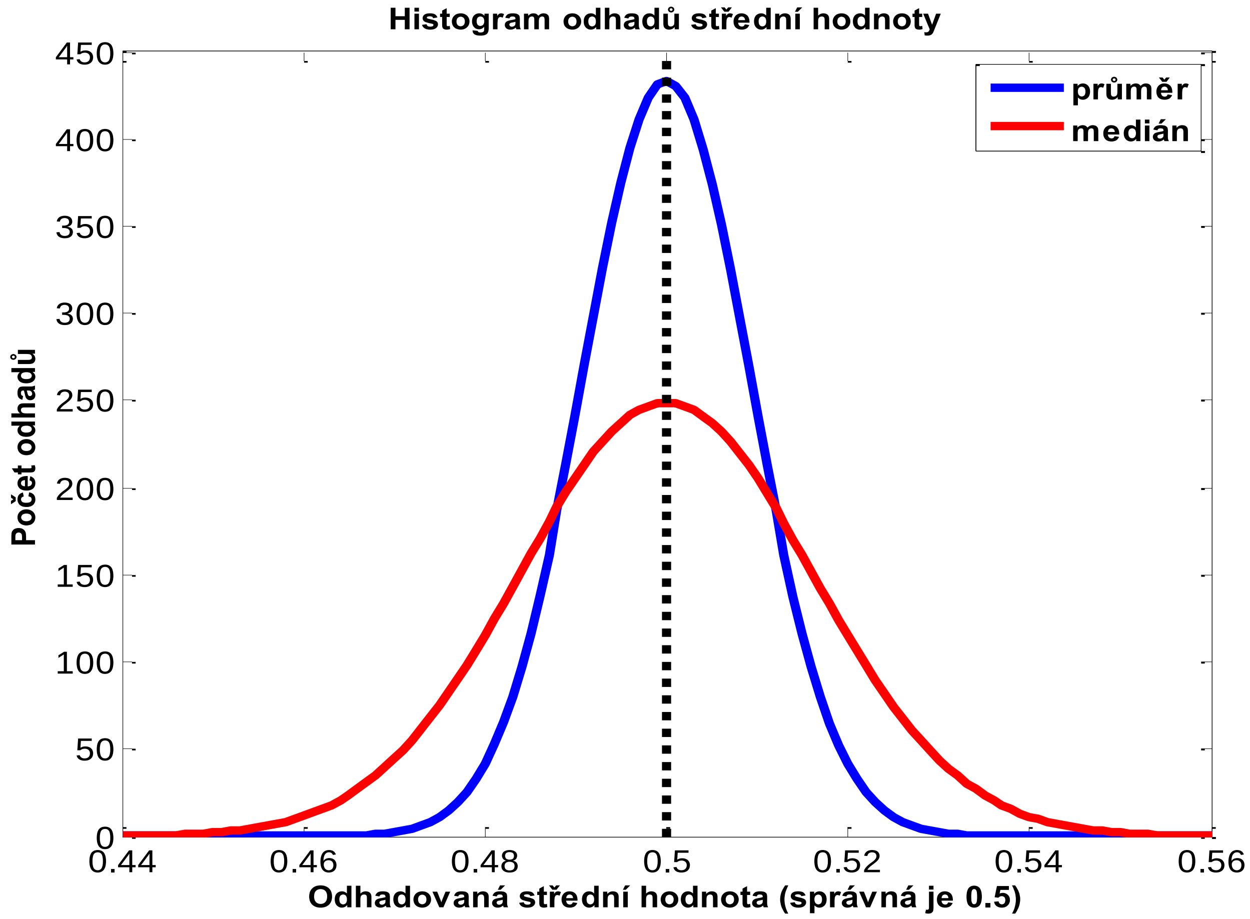
Demonstration – mean

Demonstration – median

Demonstration – original image

Symmetric noise

Let’s have 1000 randomly generated samples from range 0 to 1



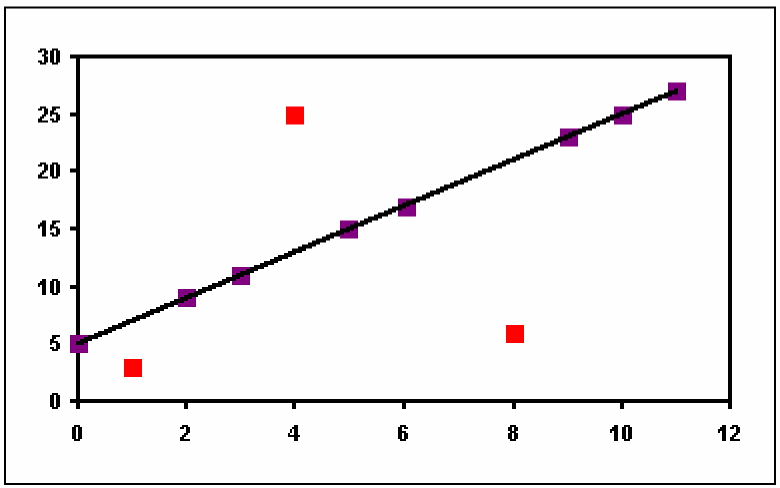
In this case, the arithmetic mean is more accurate than median.

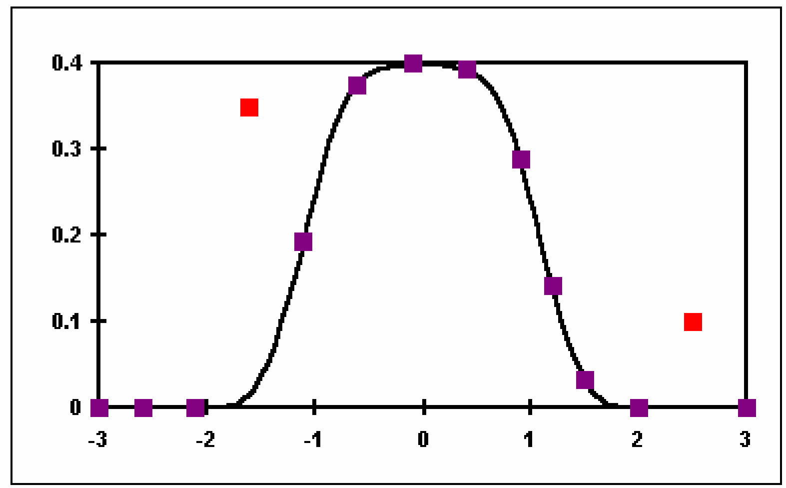
Median vs. arithmetic mean

- **Arithmetic mean**
  - Takes into account all samples, but it is prone to (asymmetric) extremes
    -> excellent on symmetric distributions
- **Median**
  - **resistant** to extreme deviations
    -> it is used usually by asymmetric distributions
  - mathematic notation is **lengthy**
  - Calculation on a computer is **lengthy**

What is an outlier?

- Outlier is a sample that differs from other samples so much that raises suspicions that it was created by a different mechanism.

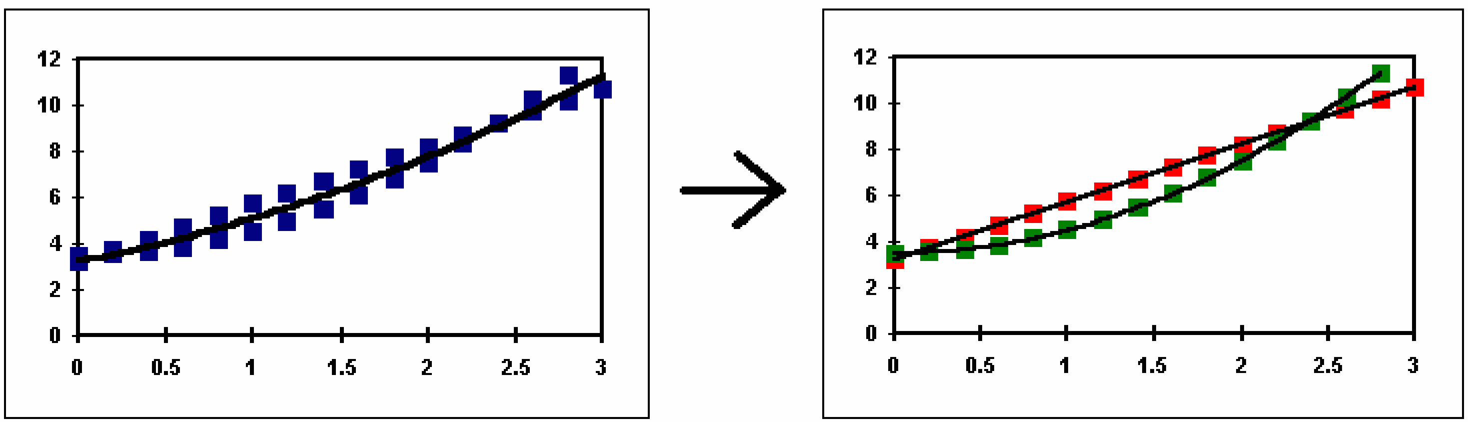




Examples of outliers (red)

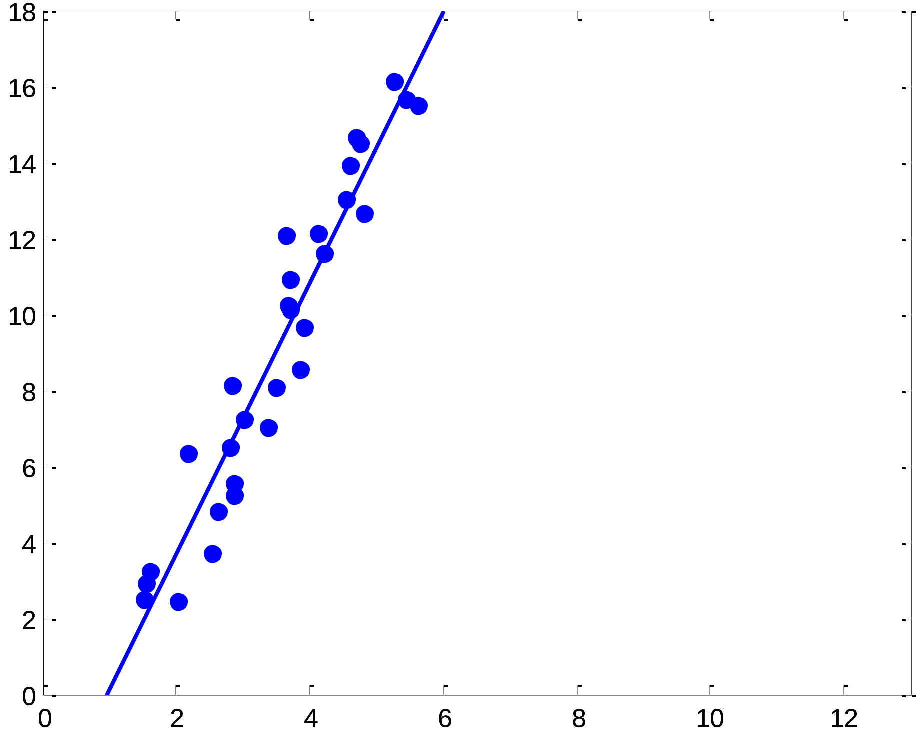
Possible causes of outliers

- Measurement Error
- Wrong assumptions (other distribution)
- Unknown data structure (multiple dist.)



- New phenomenon

Outlier harms output...



Linear regression
(Least Squares Method)

…let’s delete it!

- **Story**: In 1985 the British Antarctic expedition recorded that the concentration of ozone is about 10% lower than typical. The question was, why similarly lower value recorded a satellite. Finally, it was found that the satellite considered these values as outliers and thus deleted them? And it has been done since 1976...
- **The lesson**: Do not delete automatically outliers, because they just might be the most valuable samples in the entire dataset.

What to do with outliers?

- For normally distributed values it is expected that an outlier will appear from time to time. In this case, the outlier is kept and a robust method that can handle the outliers is applied.
- If we do not have a robust method, the outlier can be removed. But it is necessary to keep it in mind and explain why it was removed.

z-score test

- For **z-score test** mean and standard deviation of the entire dataset is calculated. Then for each sample computes? Z-score is computed:

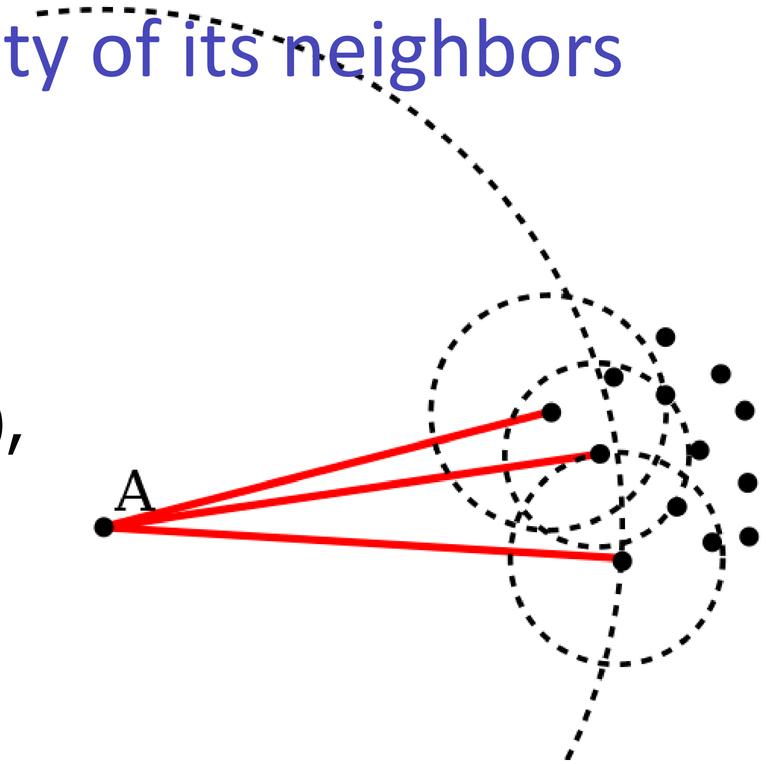
$$z = \frac{x - \mu}{\sigma}$$

- Samples with z-score greater than 3 are identified as outliers.
- This is not the most reliable method, since both μ and σ are influenced by outliers.

Local Outlier Factor

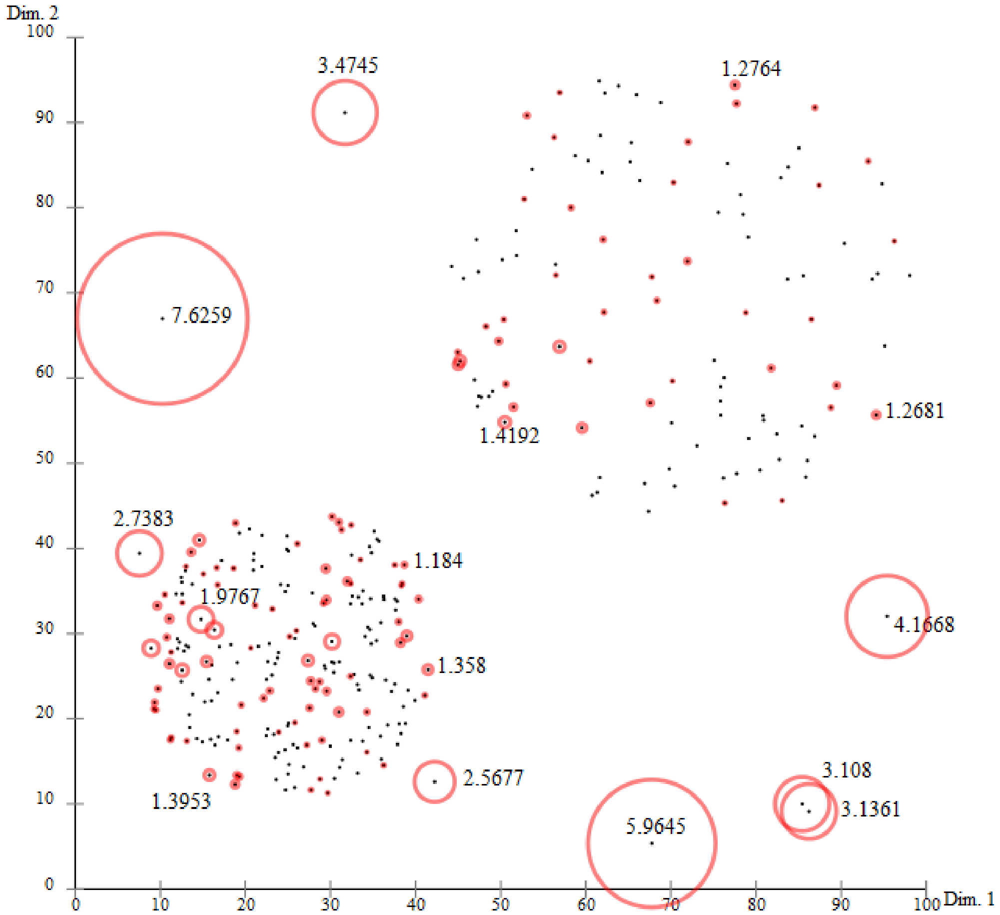
- The idea behind the **Local Outlier Factor** (LOF) is in the comparison of the local density of the sample with the local density of its neighbors

The three nearest neighbors of point A are quite far (large circle), in comparison with circles belonging to these neighbors.



http://wikipedia.com/Local_outlier_factor

Local Outlier Factor



While the top right cluster has a similar density as outliers near the lower left corner, the outliers were detected correctly.

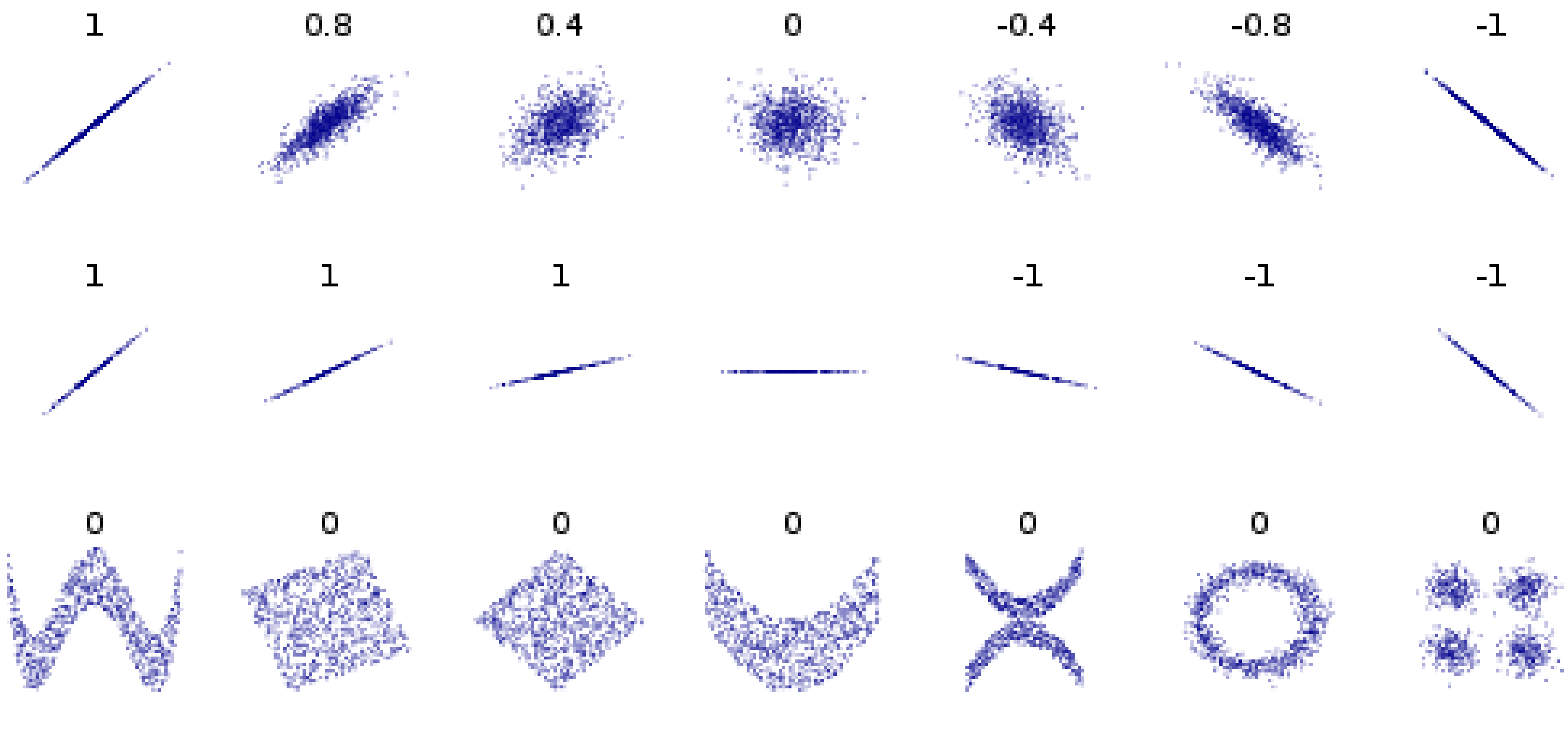
Variance

$$\sigma^2 = \frac{\sum(X_i - \mu)^2}{N}$$

= variance of population = sample

N = number of samples = mean value

Correlation



$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y},$$

Correlation matrix

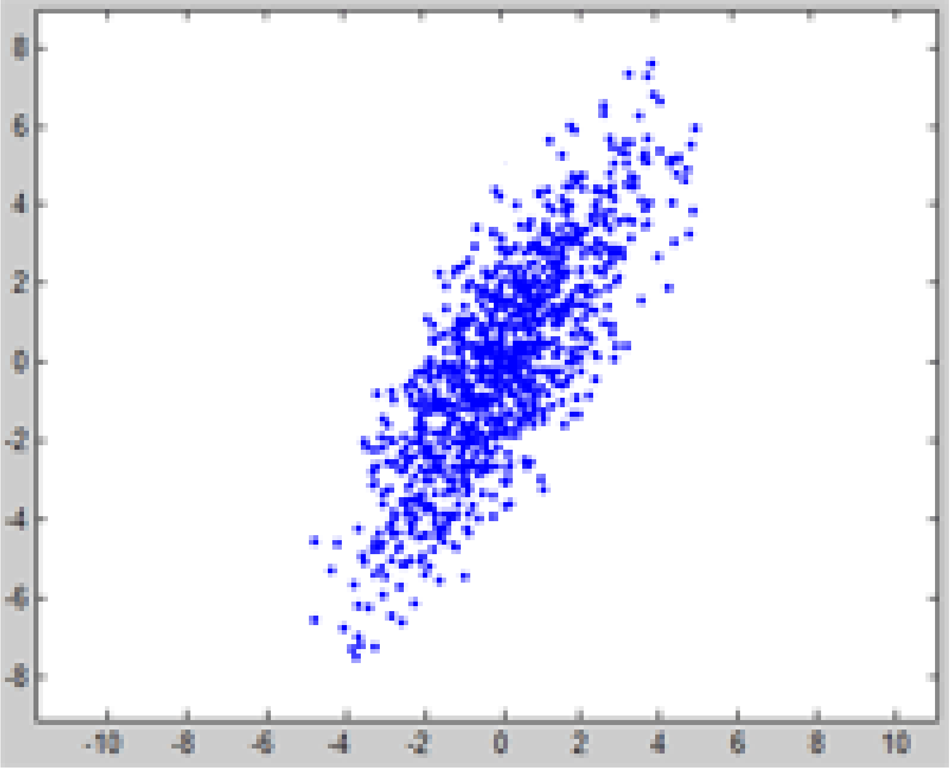
(example of preferences of Czech political parties)

	Občané.cz	VV	KSČM	ČSSD	Moravané	SPOZ	TOP 09	KDU-ČSL	Pravý Blok	Str. zelených	Suverenita	Piráti	Dělníci	SSO	ODS	Ztracenci	Sabotéři	Nechodiči	Velikost
Občané.cz	100%	3%	1%	0%	-2%	0%	0%	1%	4%	1%	1%	3%	0%	2%	0%	3%	1%	-7%	-5%
VV	3%	100%	-9%	-8%	-3%	1%	7%	-3%	2%	0%	7%	4%	1%	3%	7%	0%	0%	-29%	-7%
KSČM	1%	-9%	100%	21%	6%	4%	-43%	-2%	3%	-28%	11%	2%	5%	-7%	-38%	-2%	-1%	-4%	-31%
ČSSD	0%	-8%	21%	100%	9%	8%	-40%	5%	-4%	-24%	5%	-5%	0%	-11%	-32%	-3%	-1%	-17%	-16%
Moravané	-2%	-3%	6%	9%	100%	1%	-16%	21%	-4%	-6%	-10%	-2%	-2%	-9%	-12%	-2%	11%	0%	3%
SPOZ	0%	1%	4%	8%	1%	100%	-7%	10%	1%	-8%	1%	1%	0%	2%	-7%	-5%	-1%	-20%	-10%
TOP 09	0%	7%	-43%	-40%	-16%	-7%	100%	-10%	-1%	46%	-16%	1%	-16%	15%	55%	5%	2%	-49%	11%
KDU-ČSL	1%	-3%	-2%	5%	21%	10%	-10%	100%	0%	-11%	-11%	-2%	-9%	-5%	-13%	-7%	-2%	-27%	-16%
Pravý Blok	4%	2%	3%	-4%	-4%	1%	-1%	0%	100%	-2%	4%	4%	5%	3%	-2%	2%	-1%	-4%	-9%
Str. zelených	1%	0%	-28%	-24%	-6%	-8%	46%	-11%	-2%	100%	-15%	-1%	-9%	9%	29%	3%	5%	-25%	10%
Suverenita	1%	7%	11%	5%	-10%	1%	-16%	-11%	4%	-15%	100%	2%	8%	2%	-8%	5%	-1%	-8%	-20%
Piráti	3%	4%	2%	-5%	-2%	1%	1%	-2%	4%	-1%	2%	100%	2%	3%	0%	5%	-1%	-7%	-4%
Dělníci	0%	1%	5%	0%	-2%	0%	-16%	-9%	5%	-9%	8%	2%	100%	0%	-14%	0%	1%	9%	-4%
SSO	2%	3%	-7%	-11%	-9%	2%	15%	-5%	3%	9%	2%	3%	0%	100%	10%	4%	0%	-13%	-3%
ODS	0%	7%	-38%	-32%	-12%	-7%	55%	-13%	-2%	29%	-8%	0%	-14%	10%	100%	4%	2%	-52%	9%
Ztracenci	3%	0%	-2%	-3%	-2%	-5%	5%	-7%	2%	3%	5%	5%	0%	4%	4%	100%	0%	-5%	-4%
Sabotéři	1%	0%	-1%	-1%	11%	-1%	2%	-2%	-1%	5%	-1%	-1%	1%	0%	2%	0%	100%	-7%	2%
Nechodiči	-7%	-29%	-4%	-17%	0%	-20%	-49%	-27%	-4%	-25%	-8%	-7%	9%	-13%	-52%	-5%	-7%	100%	25%
Velikost	-5%	-7%	-31%	-16%	3%	-10%	11%	-16%	-9%	10%	-20%	-4%	-4%	-3%	9%	-4%	2%	25%	100%

Principal Component Analysis (PCA)

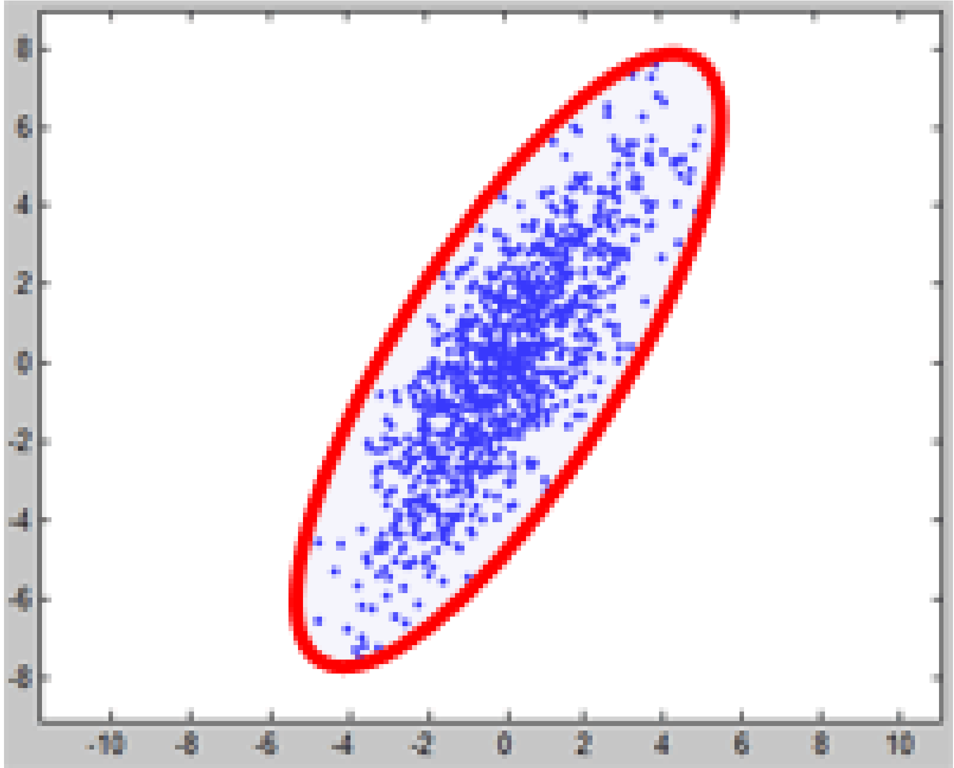
- PCA is used to reduce the number of attributes
- PCA does not select attributes, but transforms them
- PCA maximizes variances

PCA – example on 2D data



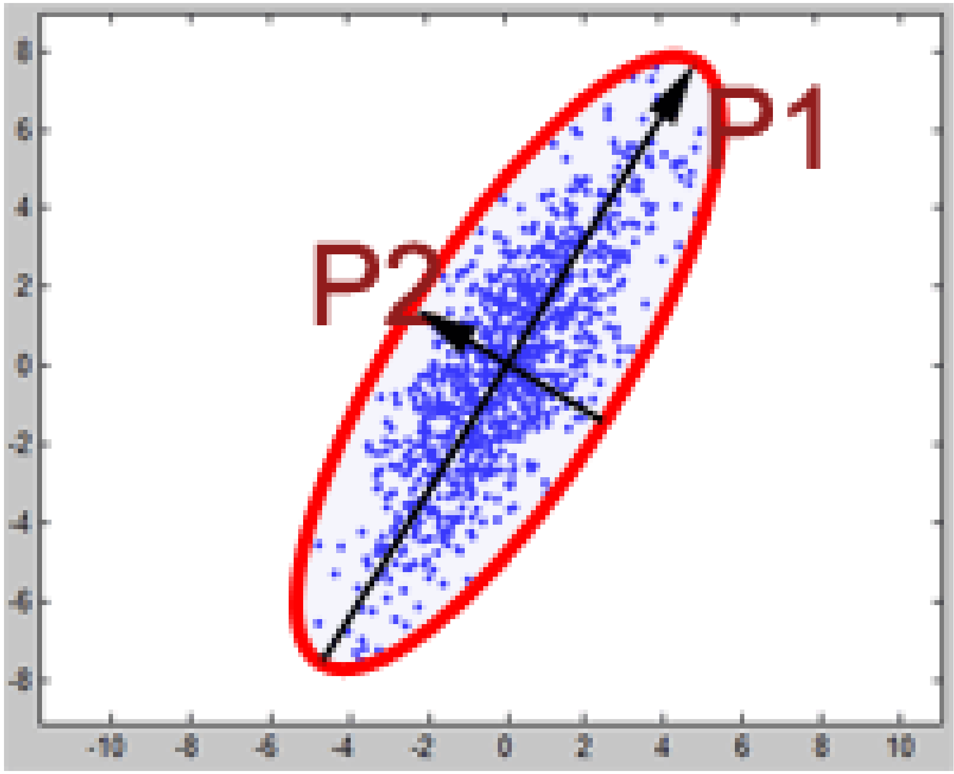
PCA works for any number of dimensions, but for clarity we use two dimensions only.

PCA – circumscribed ellipse



To see how the data are spread, we circumscribed the data by an ellipse and describe the axes.

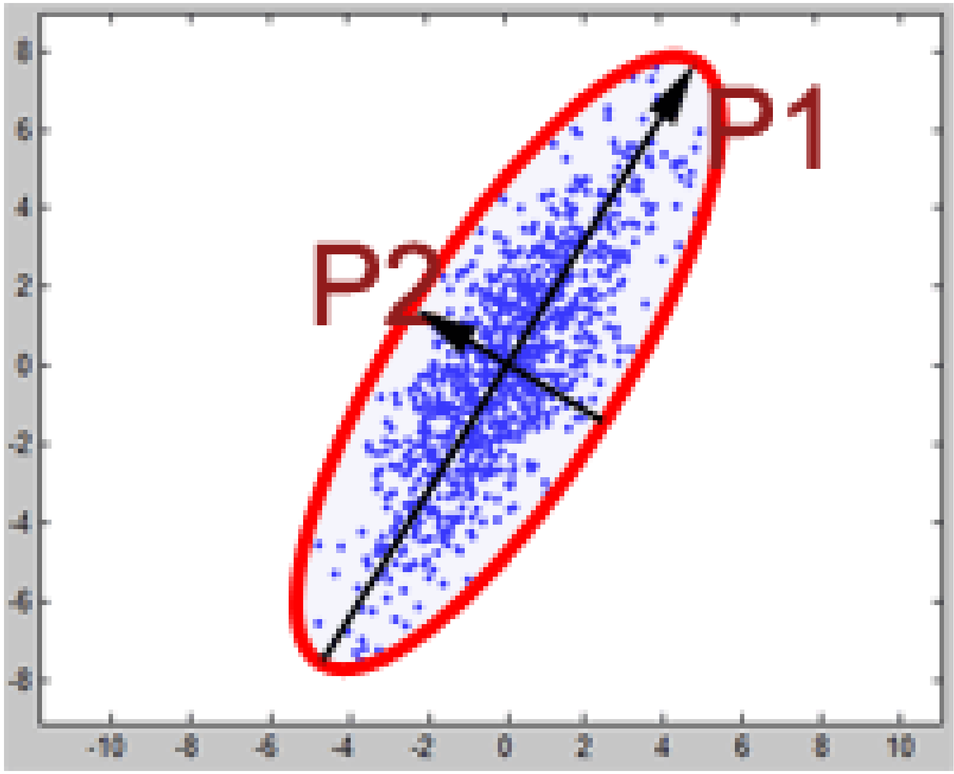
PCA – Principal Components



The first principal component maximizes the variance.

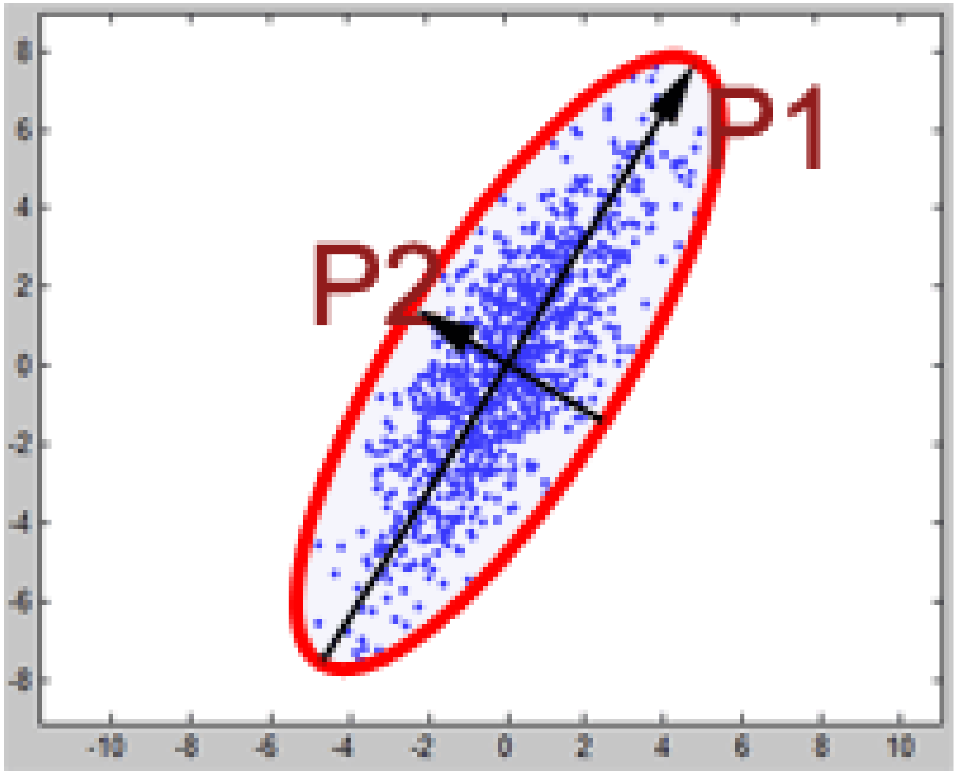
Another principal component maximizes the remaining variance.

PCA – Principal Components



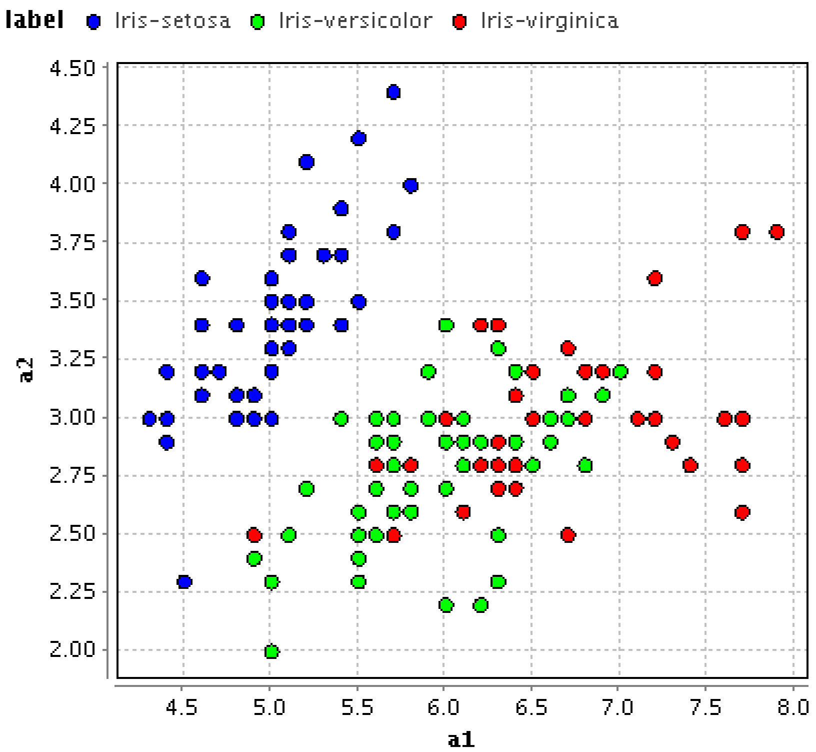
Question: What is the angle between P1 and P2?

PCA – Principal Components



Answer: Principal Components always enclose the right angle. PCA only rotates Cartesian coordinates, but not change them.

PCA – use

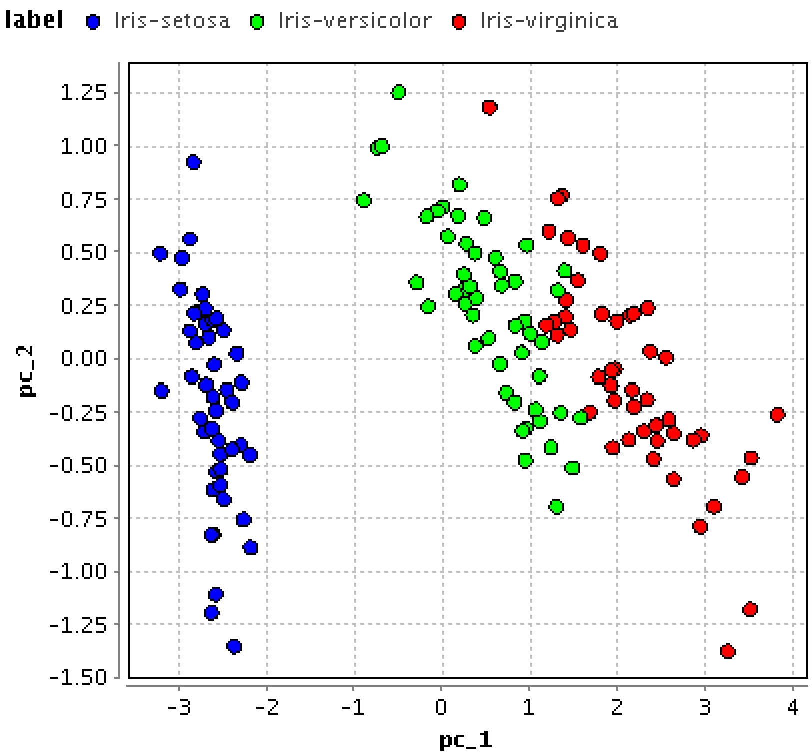


Let us have the Iris dataset, which has 4 attributes.

Let us have a classifier that accepts only two attributes.

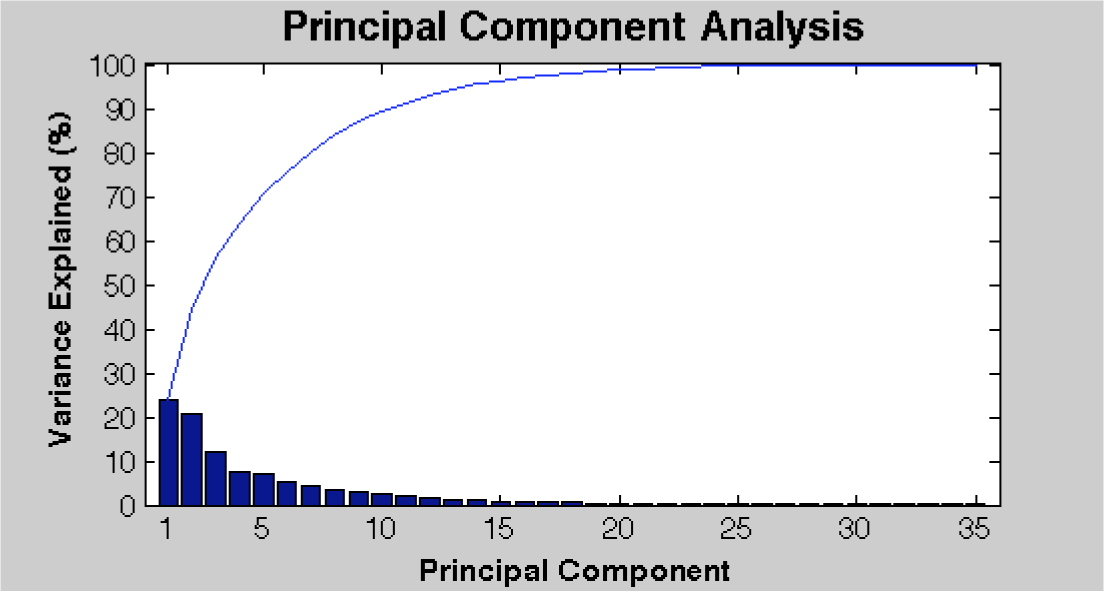
Which attributes to choose?

PCA – use



- We use PCA and then we use the first two principal components!

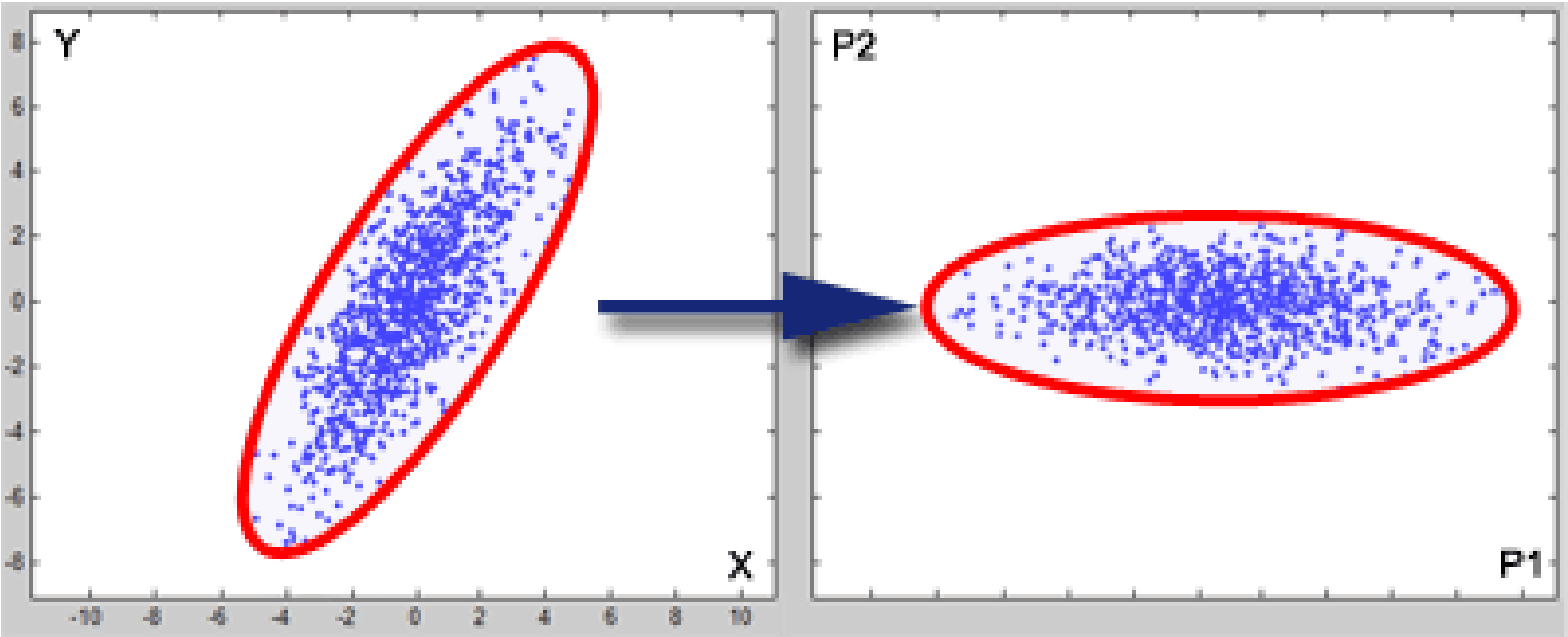
PCA - example



This is quite common dataset with 35 attributes.

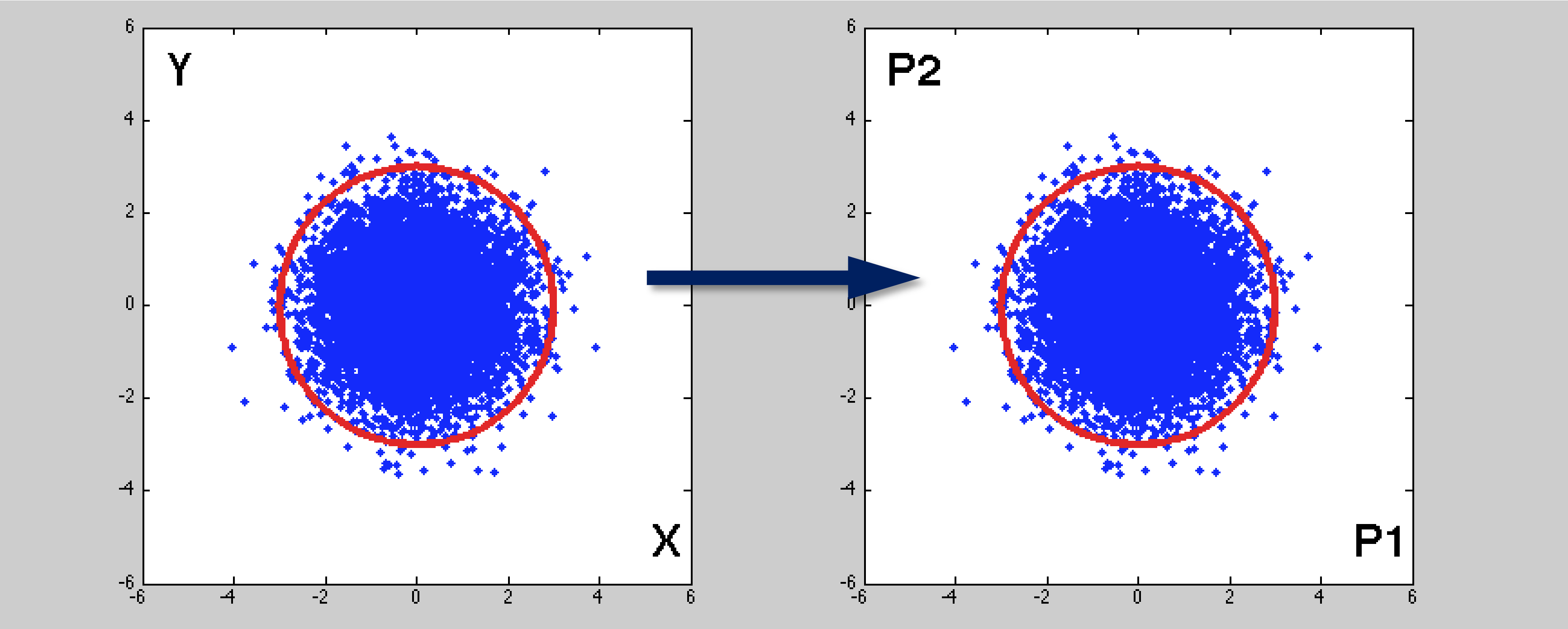
- The first 10 PCs explain 90 % of the variance.
- Another 10 PCs explain 9 % of the variance.
- Last 15 PCs explain 1 % of the variance.

PCA - limits



PCA works well when the data are distributed dominantly in one direction than in another.
Question: When PCA fails?

PCA - limits



Answer: When there is the same variance in all directions. In this case the PCA does not change anything.

PCA - factors

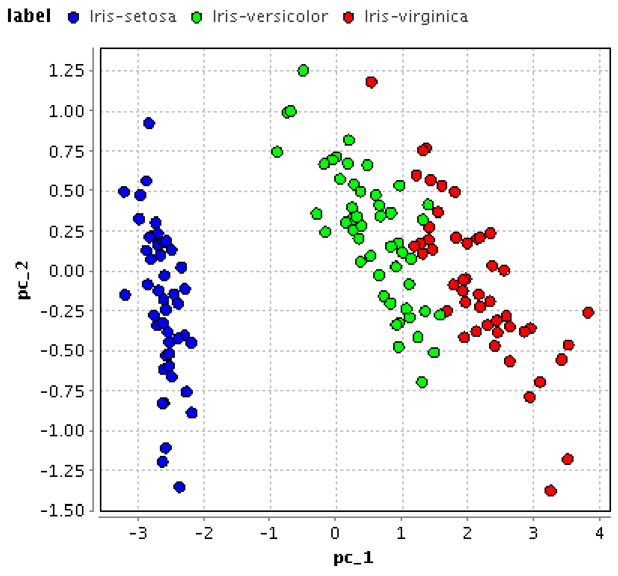
- New axes (factors) are calculated by linear combinations of the original attributes

$$F_i = W_{i1}X_1 + W_{i2}X_2 + \ldots + W_{ip}X_p$$

- PC1 corresponds to factor F_1
- How it is expressed?
- F1 = PC1 =

  w_{11}.petal_length +

  w_{12}.petal_width + …



PCA - factors

- How do we the inverse transform?
- petal_length = ?

$$X_j = A_{1j}F_1 + A_{2j}F_2 + \ldots + A_{mj}F_m + U_j$$

- What is the meaning of U?
- petal_length =

  $= a_{11}.F_1 + a_{21}.F_2 + \ldots$ remaining_varinace

Utilization of PCA

- MI-PDD, MI-ROZ
- Use – during exercises in Rapidminer
- In "R": function *princomp*