INF250

Multivariate analysis

Topics for INF250

- Multivariate analysis
- PCA
- Clustering

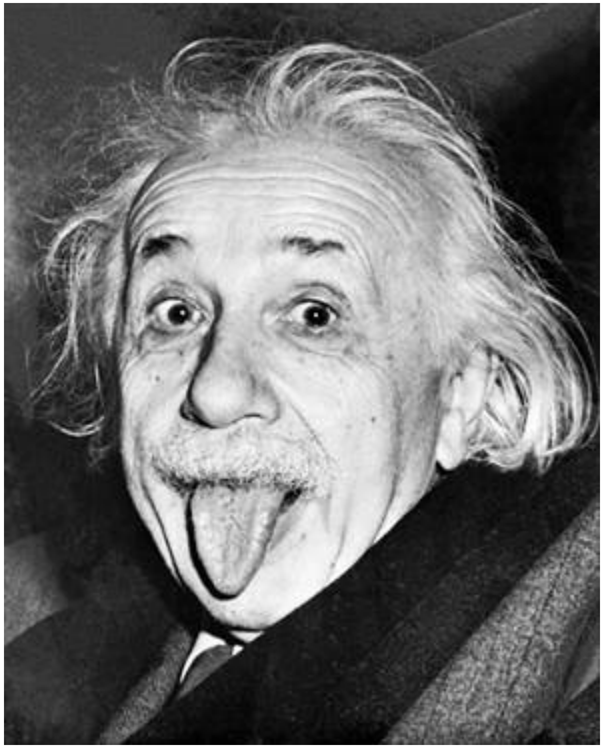
The world is multivariate

Multivariate analysis refers to any statistical technique used to analyze data that arises from more than one variable

Who is this ?

WHO IS THIS?

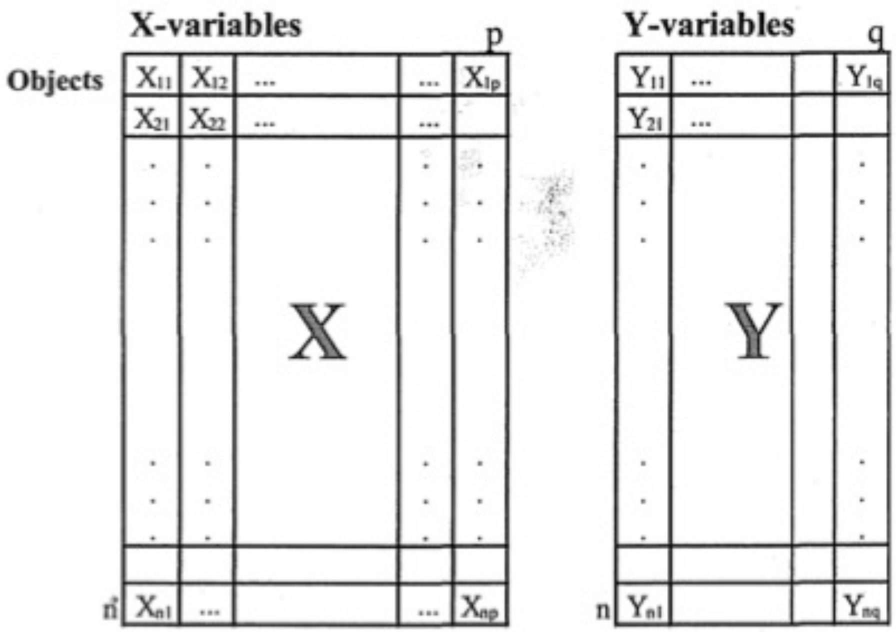
WHO IS THIS



- The brain do not collect information from single parts, but for elements in a pattern
- We need multivariate techniques!

Collecting data

- Data are often collected as sensor input (X) and actual chemical/physical/sensory measurements as (Y)



Data presentation and modelling

- One, Two, Three dimensional plots?
- Do we need a N dimension space to view data?
- How to find the optimal low dimension space that reveals maximum useful information?
- How to find the «not so useful information»
    - «NOISE»

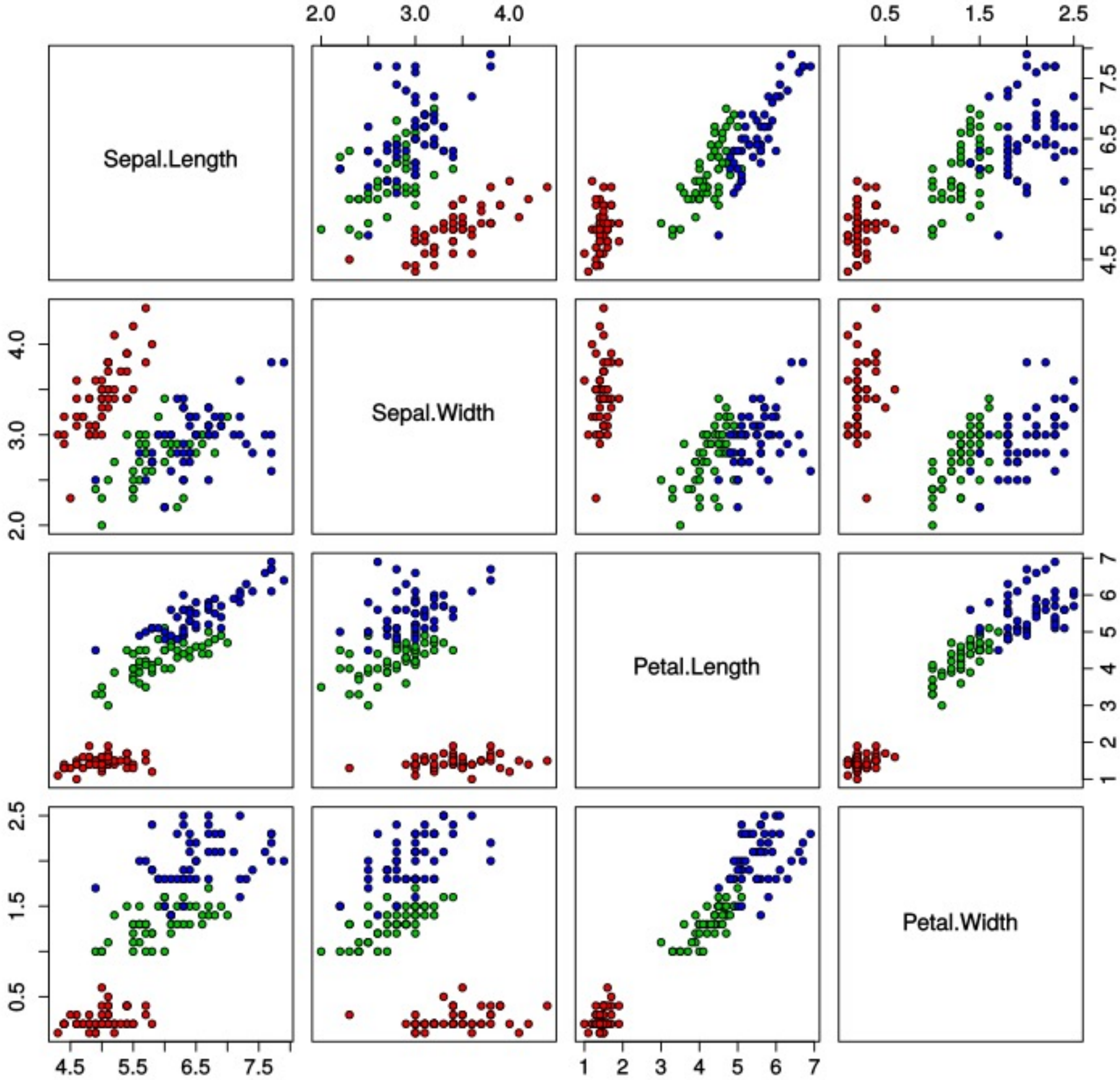
The Iris data set

- https://en.wikipedia.org/wiki/Iris_flower_data_set

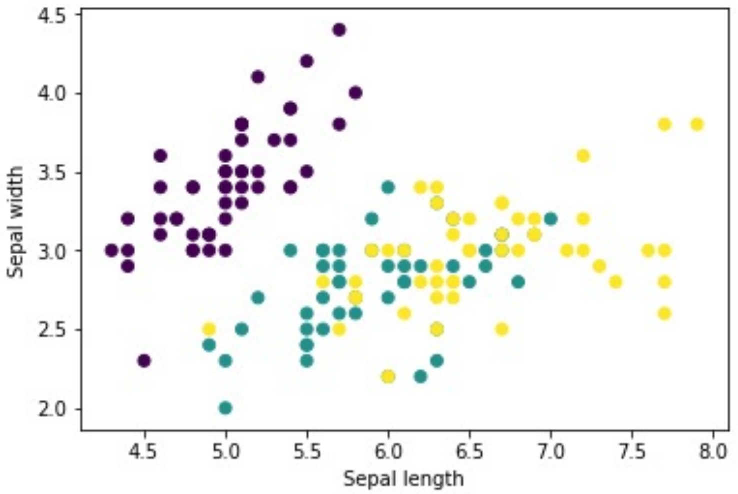
- Often used as a demo dataset for multivariate analysis
- 3 species of Iris determined from 4 features: length and width of sepal and petals

- http://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html

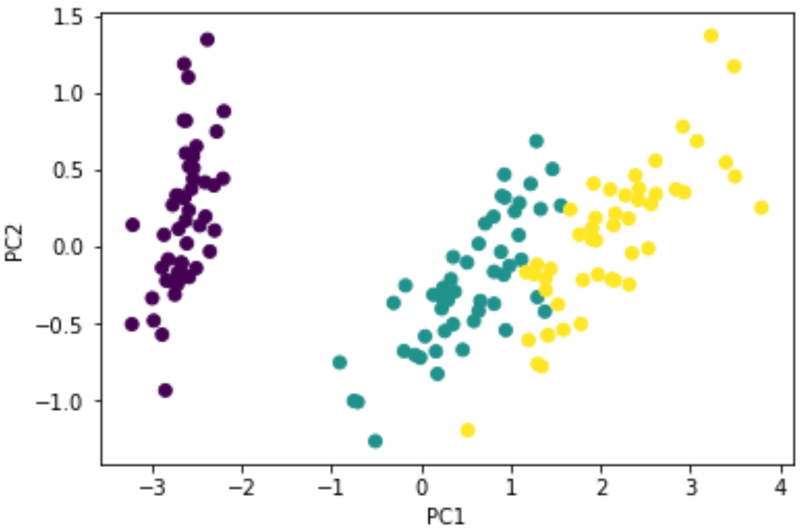
Iris Data (red=setosa,green=versicolor,blue=virginica)



Principal Component Analysis



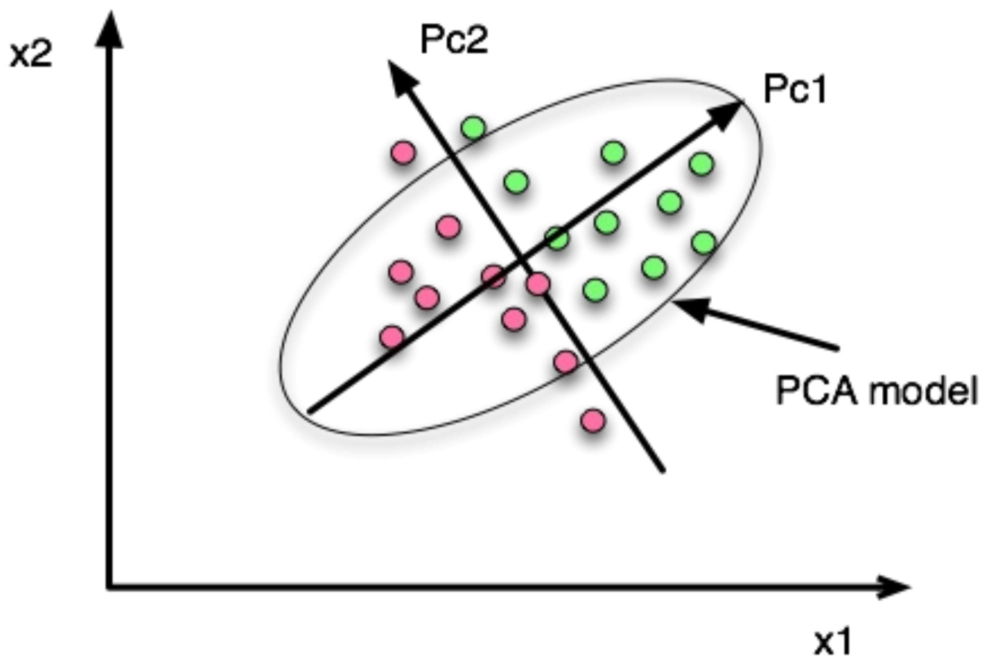

```python
import matplotlib.pyplot as plt
from sklearn import datasets
from sklearn.decomposition import PCA

# import the Iris  data
iris = datasets.load_iris()
X = iris.data
y = iris.target

plt.figure()
plt.scatter(X[:,0],X[:,1],c=y)
plt.xlabel('Sepal length')
plt.ylabel('Sepal width')
plt.show()
```




```python
#simple PCA

X_reduced = PCA(n_components=3).fit_transform(iris.data)
plt.figure()
plt.scatter(X_reduced[:,1],X_reduced[:,2],c=y)
plt.xlabel('PC1')
plt.ylabel('PC2')
plt.show()
```

GRAPHICAL INTERPRETATION OF PCA

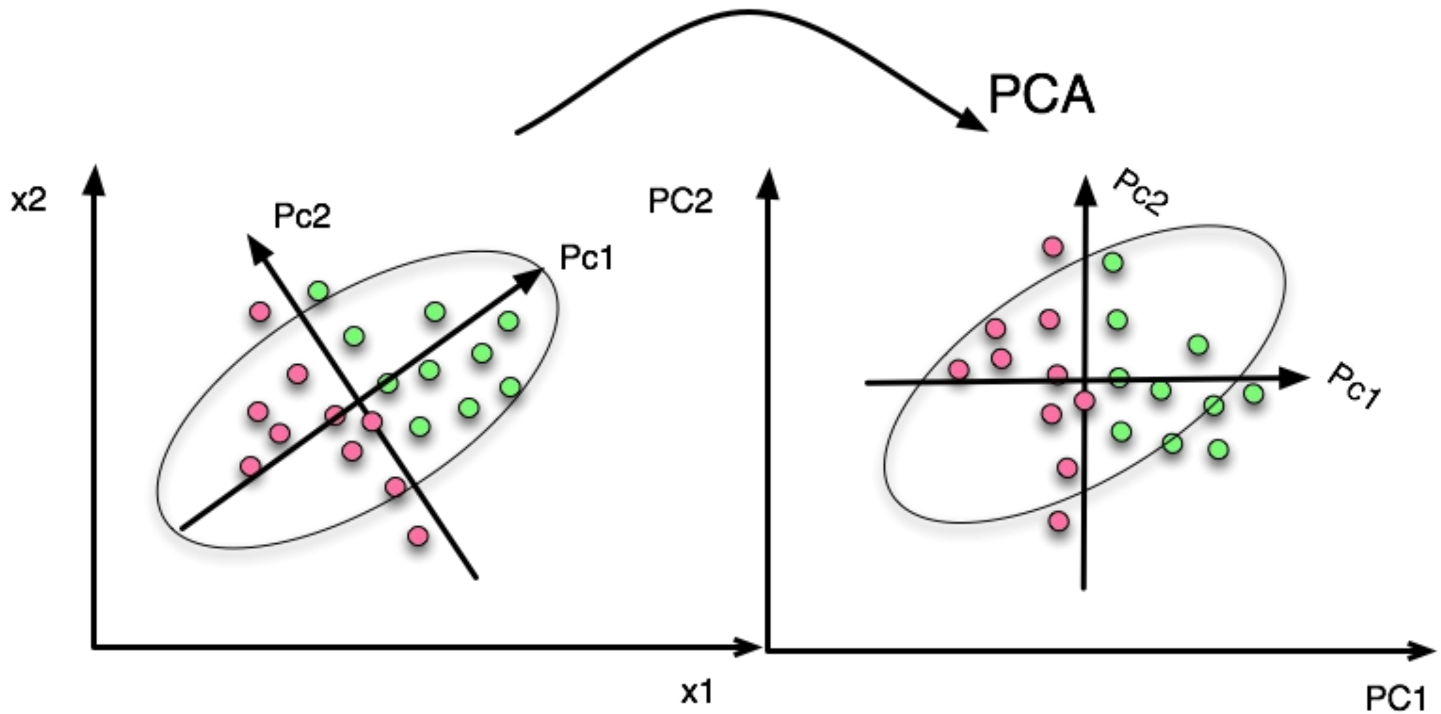
- The main idea is to form a minimum number of new variables that describes the variation of the data by a linear combination of the original values



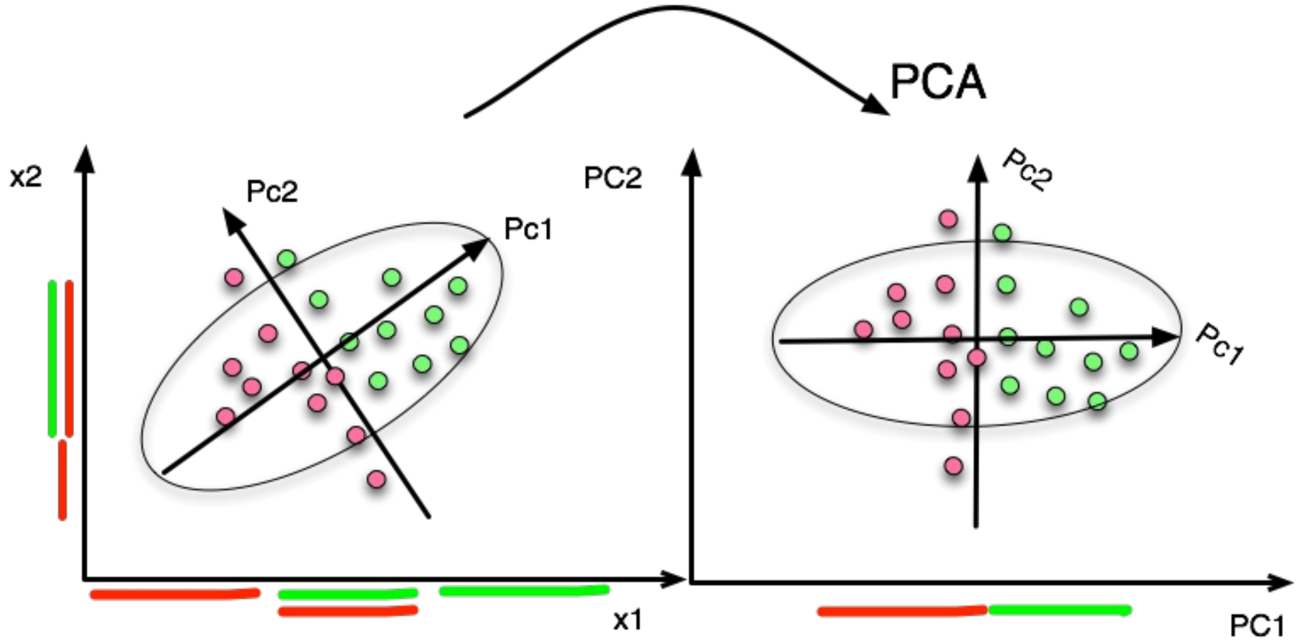
- PC1
  - Direction of the largest variation in X
- PC2
  - Direction of the next largest variation in X

The Principal Components

- The new coordinate system represents a stepwise rotation along the principal orthogonal axes



USE PCA TO SEPARATE CLASSES



Saparation of classes with a rotation

PCA basics

Data $\Rightarrow$ Systematic variation + Residuals

$$X = TP' + E$$

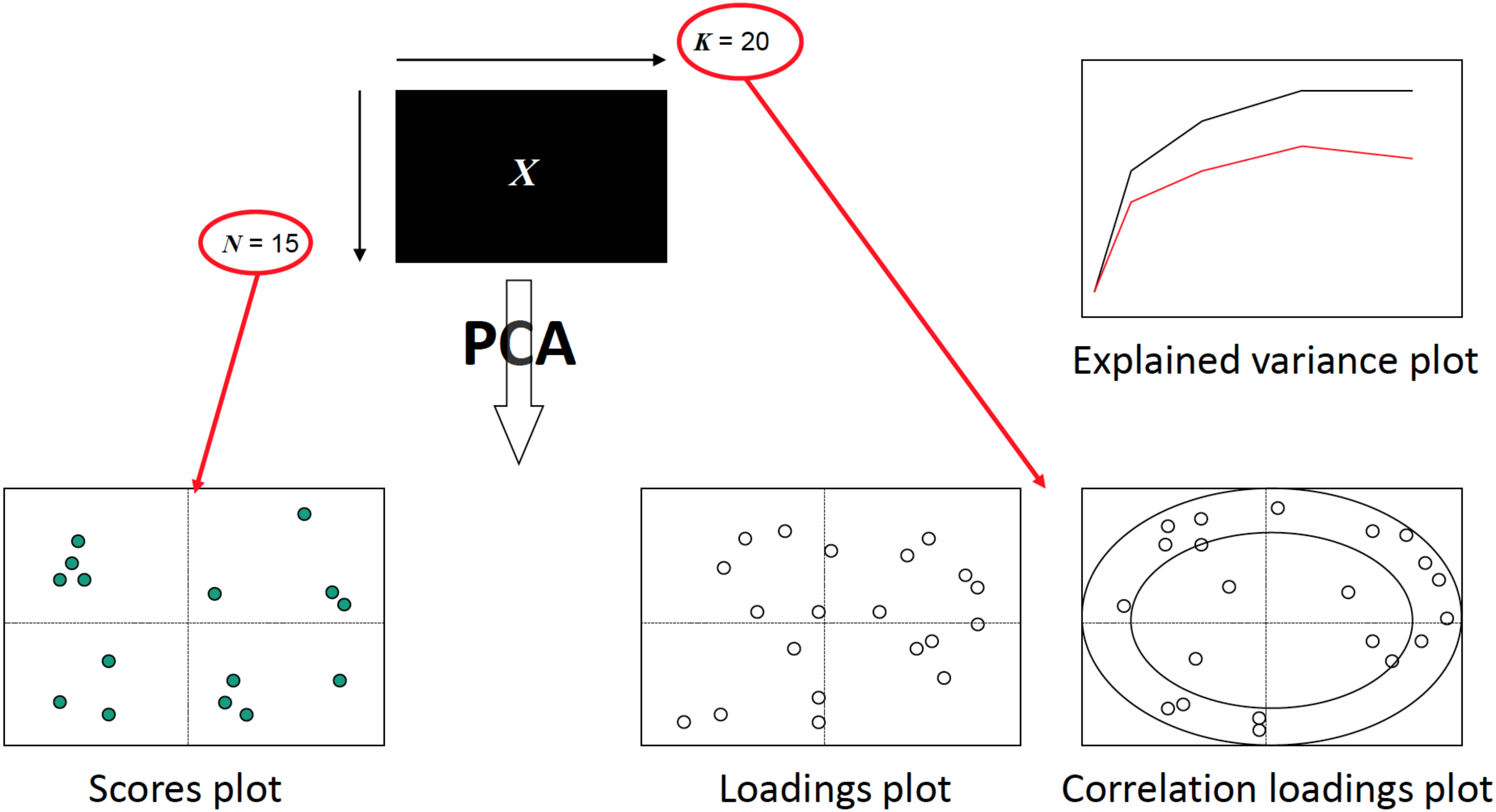
$\Downarrow$

Principal Components (PC's) describing the systematic variation in the data

X: data matrix
T: PCA scores matrix
P: PCA loadings matrix
E: residuals / noise

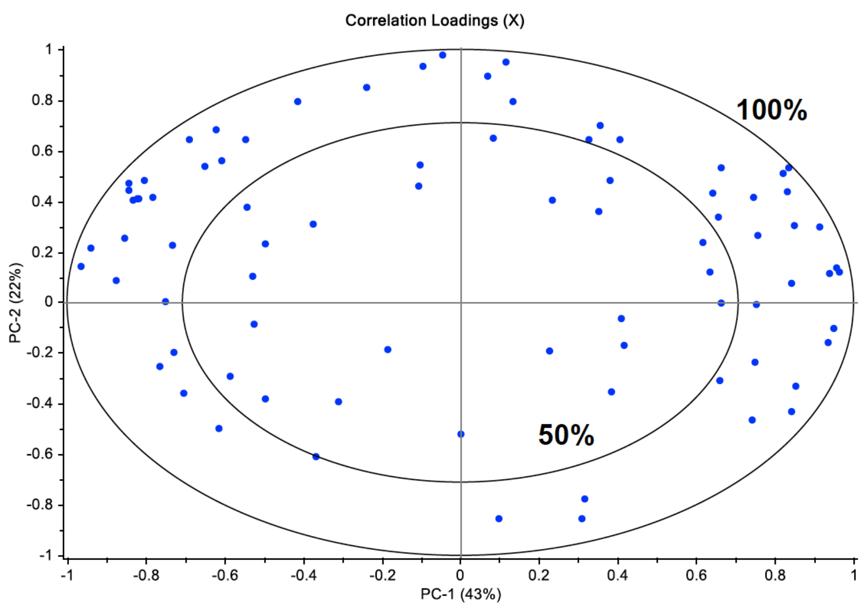
PCA basics

Data in this illustration consists of 15 observations and 20 variables ➔ data matrix X of dimension (15×20)



PCA- correlation loadings

- Circles in the plot corresponding to various degrees of explained variances
- Typically one will present a circle for **100%**explained and for **50%**explained variance by the **two** components



PCA basics – centring and standardisation

- Only purpose of PCA is to look for directions with **high variance**
- This implies: if there are variables x_k in X that have a **larger variance** than others …
  - they will be given **most** attention
  - ➔ They will **dominate** the extracted components
  - ➔ They will **dominate** the plots
- Generally one is interested in letting all variables play a role in the estimation of components (there are exceptions) ➔ standardise variables x_k in X
- Matrices in multivariate statistics are always **either** *centered* or *standardised*

PCA basics – centring and standardisation

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1K} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{NK} \end{pmatrix}$$

- Number of **objects (rows)**:
  - $n = 1 \ldots N$
- Number of **variables (columns)**:
  - $k = 1 \ldots K$
- Observed value x_{nk} for
  - n'th object
  - k'th variable

center

$$\boxed{x_{nk,cent} = x_{nk} - \bar{x}_k}$$

$$\bar{x}_k = \frac{1}{N}\sum_{n=1}^{N} x_{nk}$$

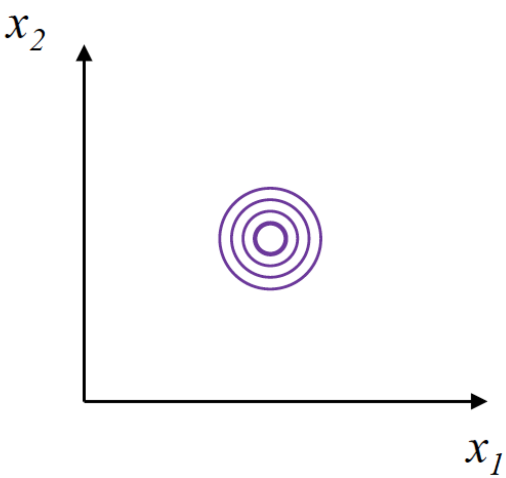
where

standardise

$$\boxed{x_{nk,stand} = \frac{x_{nk} - \bar{x}_k}{\sigma_k}}$$

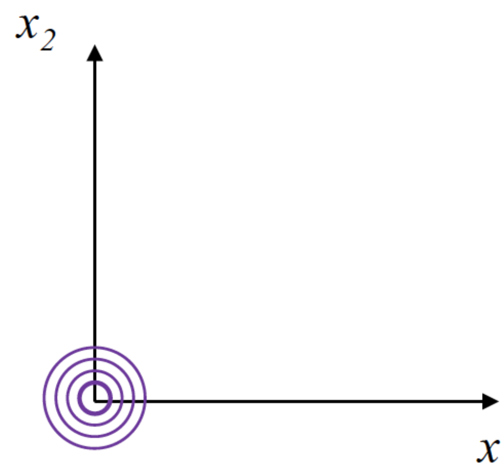
$$\sigma_k = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_{nk} - \bar{x}_k)^2}$$

PCA basics – centring and standardisation



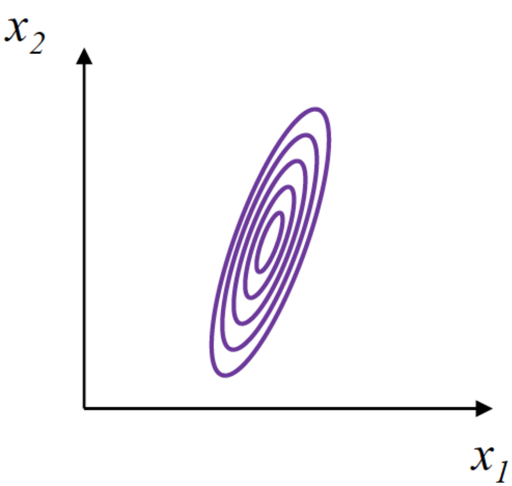
center data

$$x_{nk,cent} = x_{nk} - \bar{x}_k$$



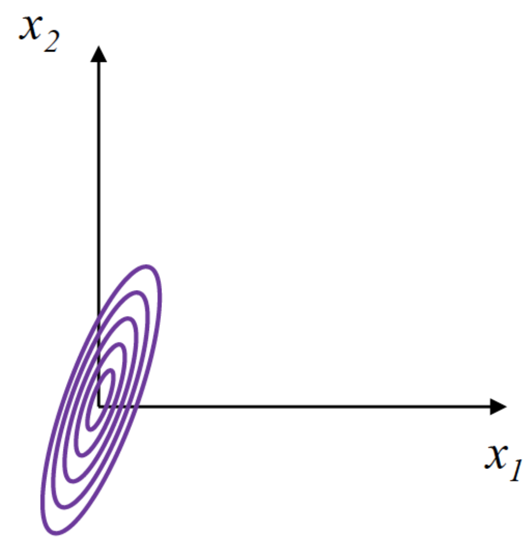
Equal variance of
x_1 and x_2

PCA basics – centring and standardisation



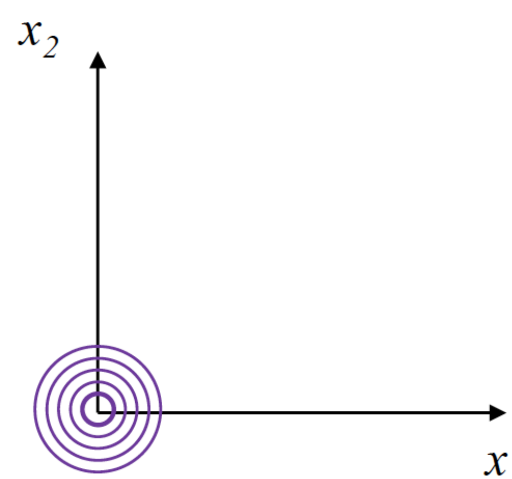
center data

$$x_{nk,cent} = x_{nk} - \bar{x}_k$$



standardise data

$$x_{nk,stand} = \frac{x_{nk} - \bar{x}_k}{\sigma_k}$$



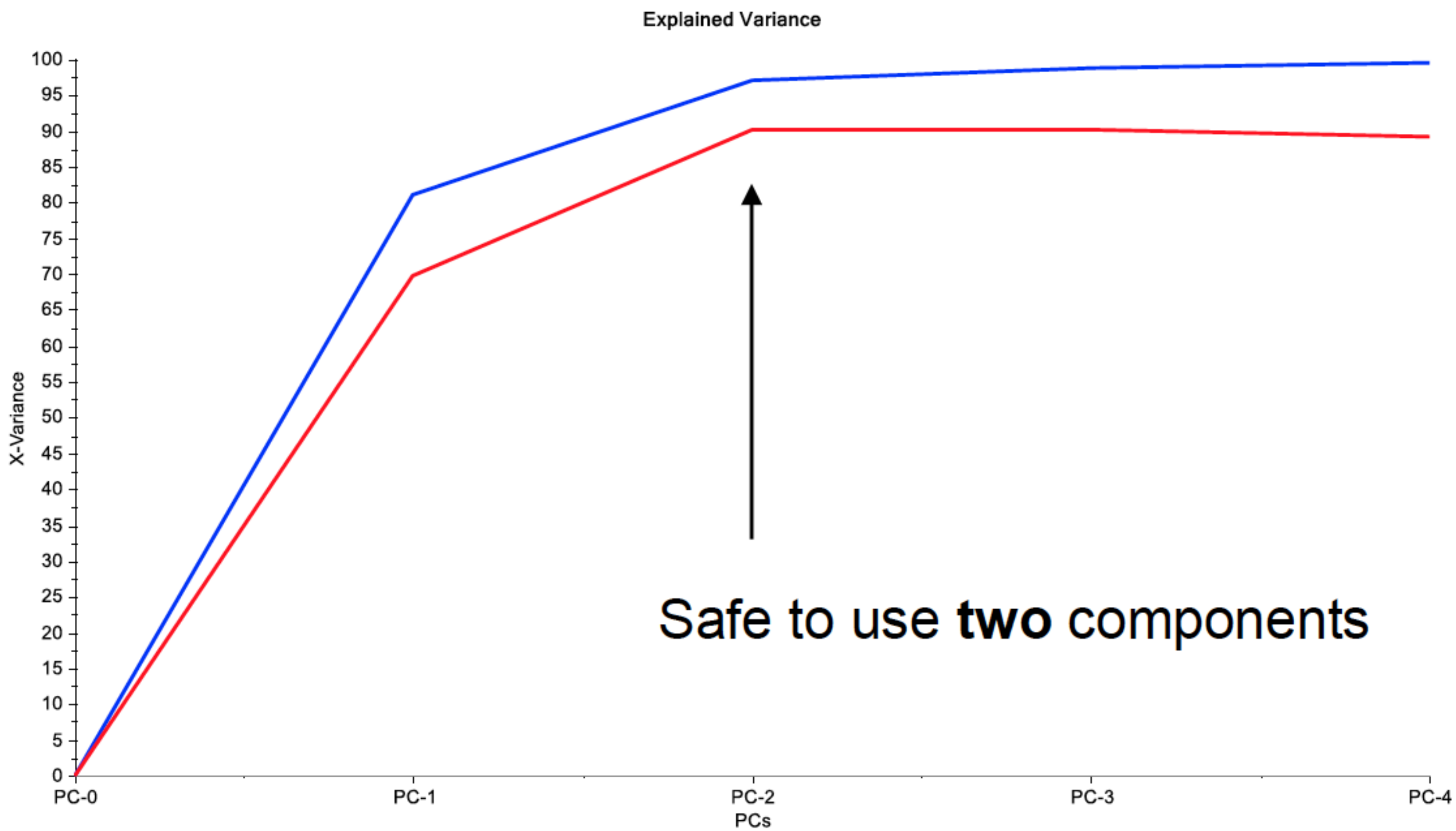
Variance of x_2 is larger than variance of x_1

PCA - validation

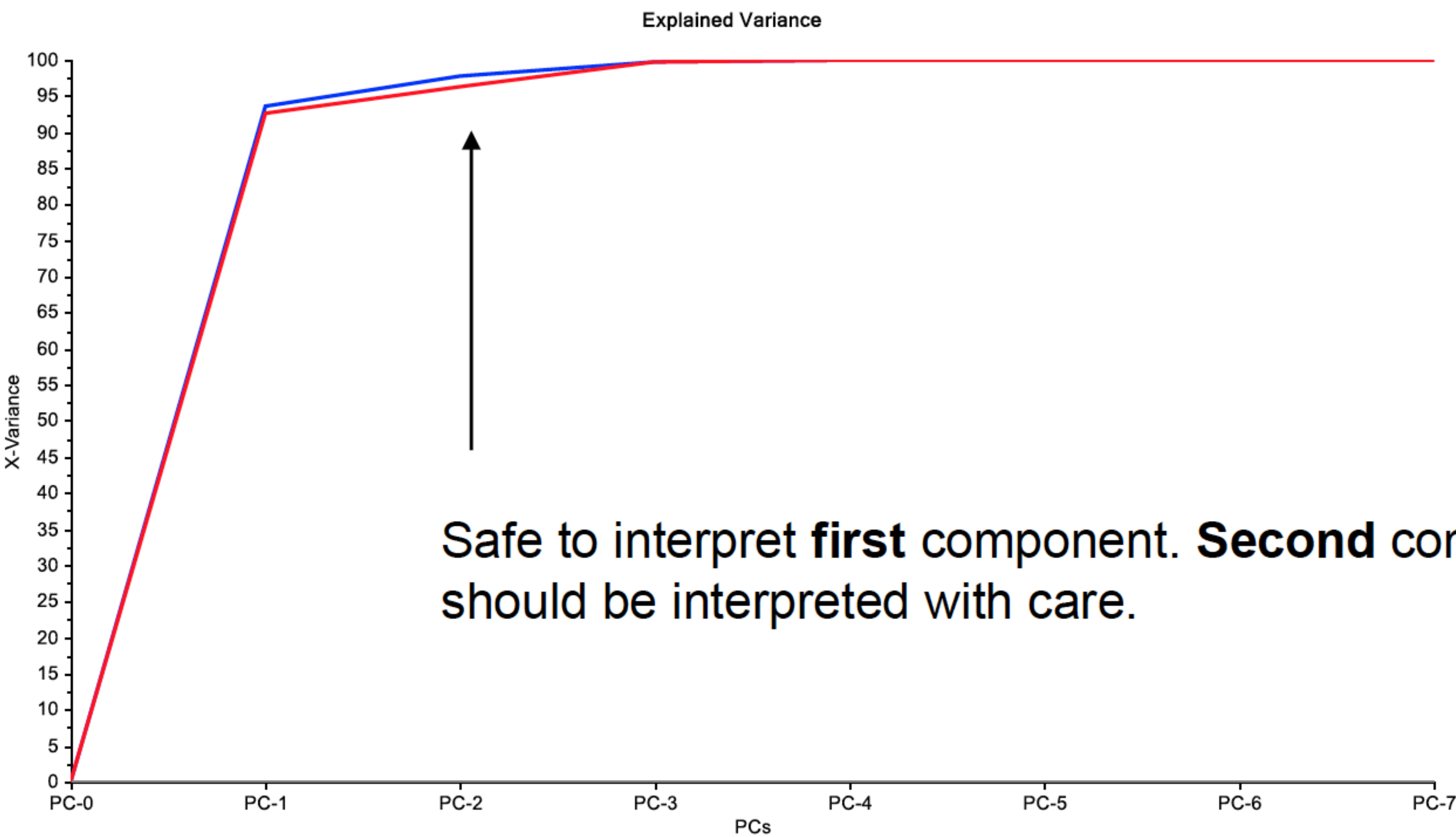
Validation is necessary to gain knowledge on **how many** components are **appropriate** for the model, i.e. how many components can be used for interpretation and further analysis.

- Use of *internal cross validation*in PCA
- **K-fold**cross validation (number of folds / splits used)
- **LOO**cross validation ("Leave-one-out")

PCA - validation

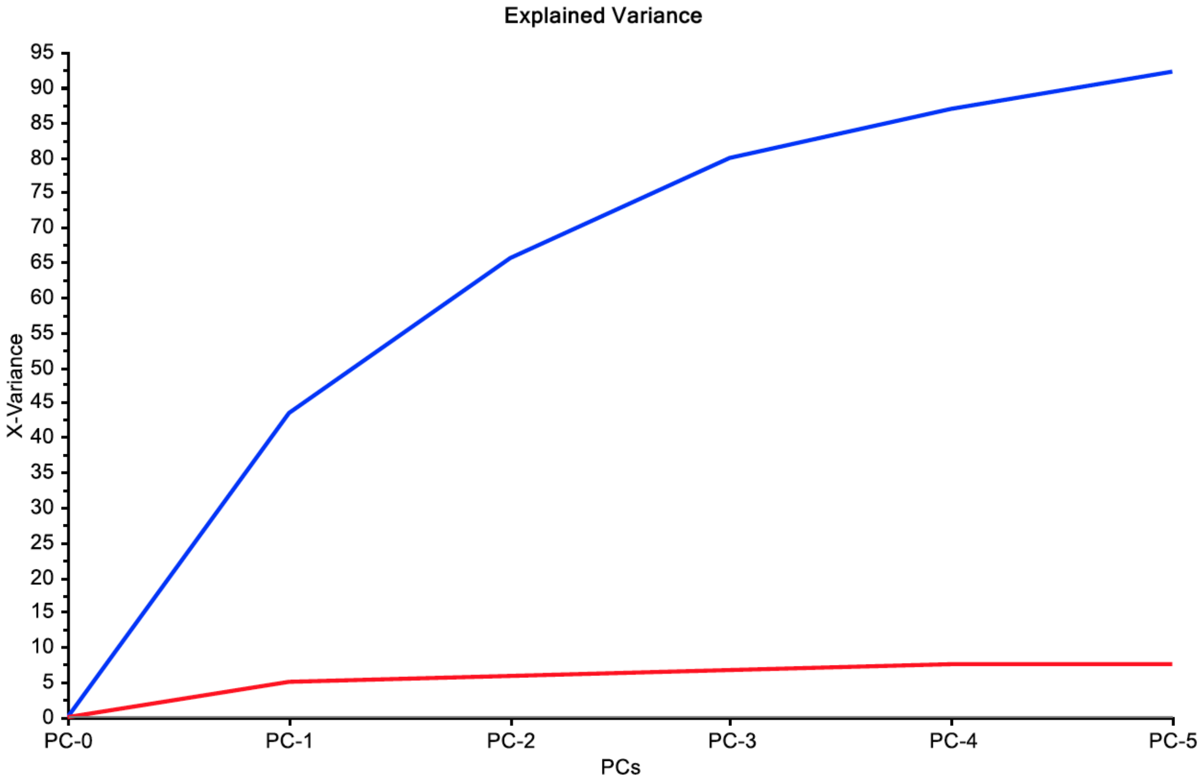


PCA - validation



Safe to interpret **first** component. **Second** component should be interpreted with care.

PCA - validation



Poor model – may be a result of few objects that are very different from each other or overfitting

Hoggorm, HoggormPlot and examples

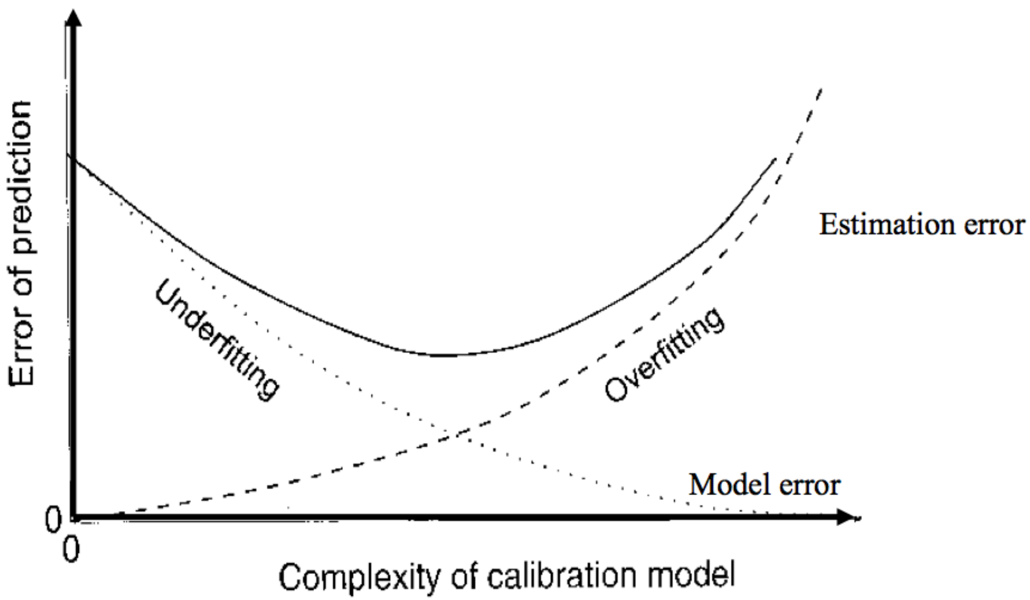
- **Hoggorm** package for multivariate statistics
  - GitHub: https://github.com/olivertomic/hoggorm
  - Read the Docs: http://hoggorm.readthedocs.io/en/latest/
- **HoggormPlot** package for convenient plotting of Hoggorm results
  - GitHub: https://github.com/olivertomic/hoggormPlot
  - Read the Docs: http://hoggormplot.readthedocs.io/en/latest/
- **Examples** of how to use Hoggorm illustrated in Jupyter notebooks
  - GitHub: https://github.com/khliland/hoggormExamples

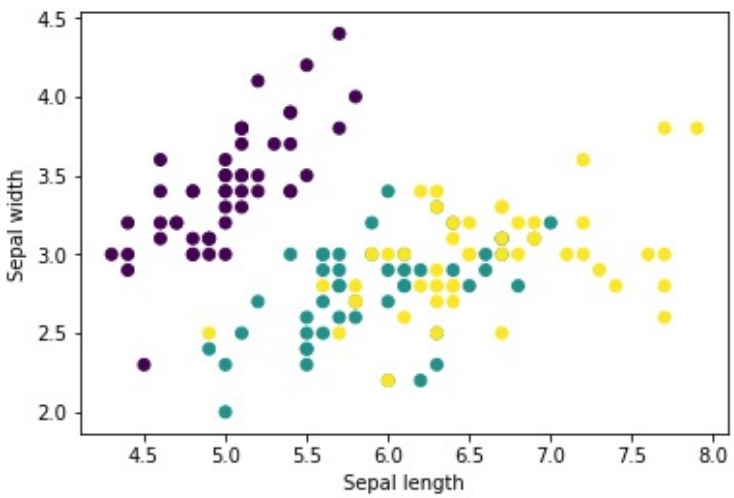
Hoggorm

```python
import hoggorm as ho
import hoggormplot as hopl
from sklearn import datasets

#import Iris data set

iris = datasets.load_iris()
X = iris.data
Y = iris.target

# Get the variabls
iris_varNames = list(iris.feature_names)

# Get the objects
iris_objNames = list(iris.target_names)

model_01 = ho.nipalsPCA(arrX=X, Xstand=True, cvType=["loo"], numComp=3)


hopl.plot(model_01, plots=[1, 2, 3, 6], line=True)
hopl.plot(model_01)
hopl.plot(model_01, plots=['scores', 'loadings', 'explainedVariance'])
```

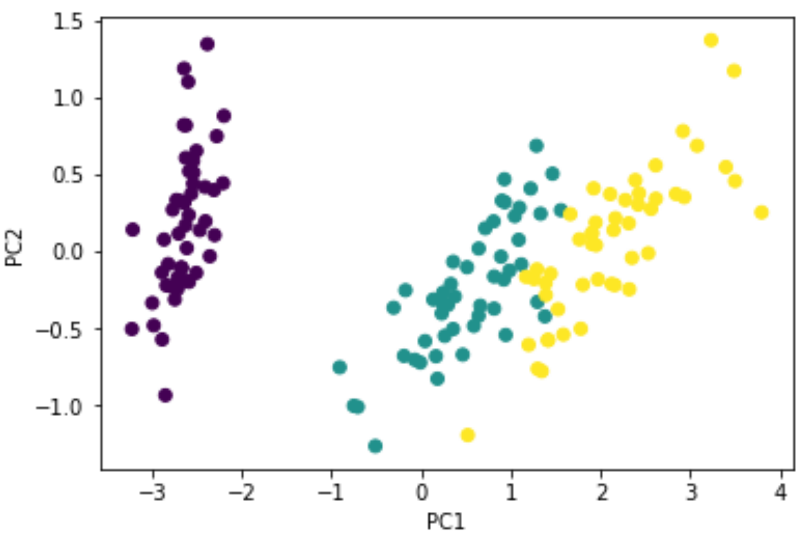
HOW MANY COMPONENTS?

- «Use the number of components that do not produce overfitting of model»
- Explained variance for PCA
- Prediction error for PCR/PLS





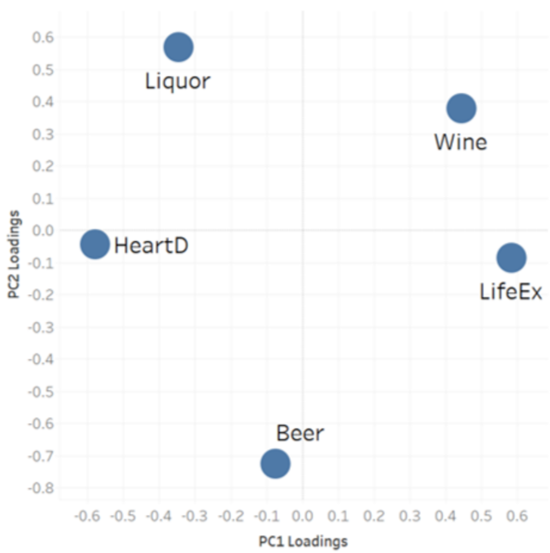

```python
import matplotlib.pyplot as plt
from sklearn import datasets
from sklearn.decomposition import PCA

# import the Iris  data
iris = datasets.load_iris()
X = iris.data
y = iris.target

plt.figure()
plt.scatter(X[:,0],X[:,1],c=y)
plt.xlabel('Sepal length')
plt.ylabel('Sepal width')
plt.show()


#simple PCA

X_reduced = PCA(n_components=3).fit_transform(iris.data)
plt.figure()
plt.scatter(X_reduced[:,1],X_reduced[:,2],c=y)
plt.xlabel('PC1')
plt.ylabel('PC2')
plt.show()
```

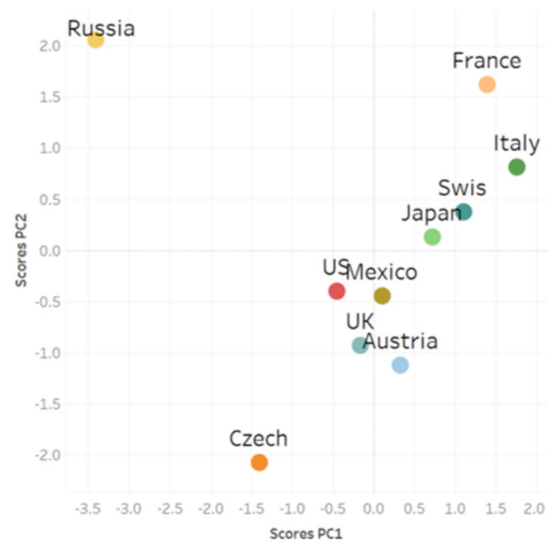


Causality ?

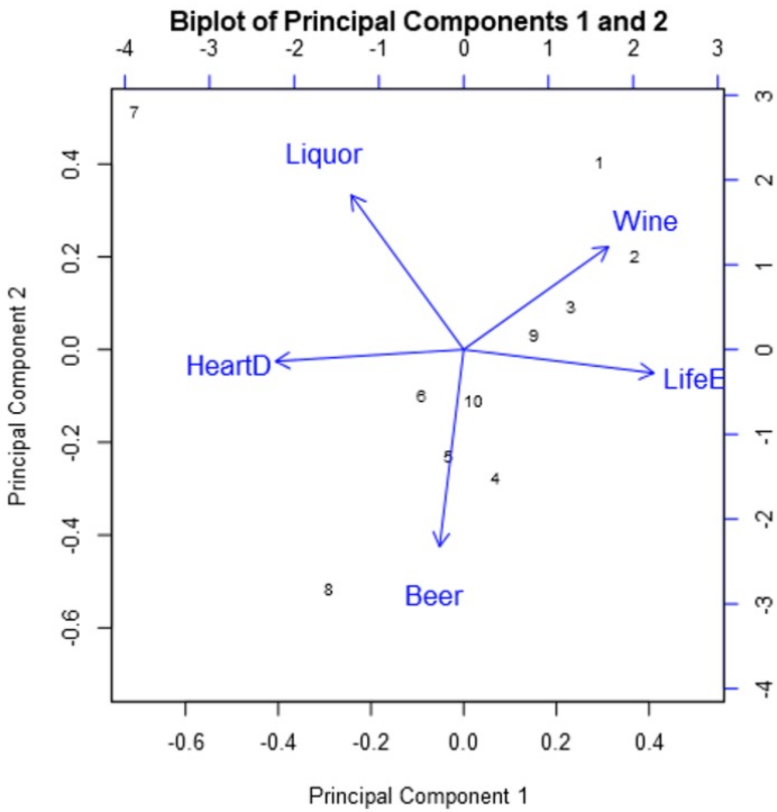
Data set from Time Magazines 1996

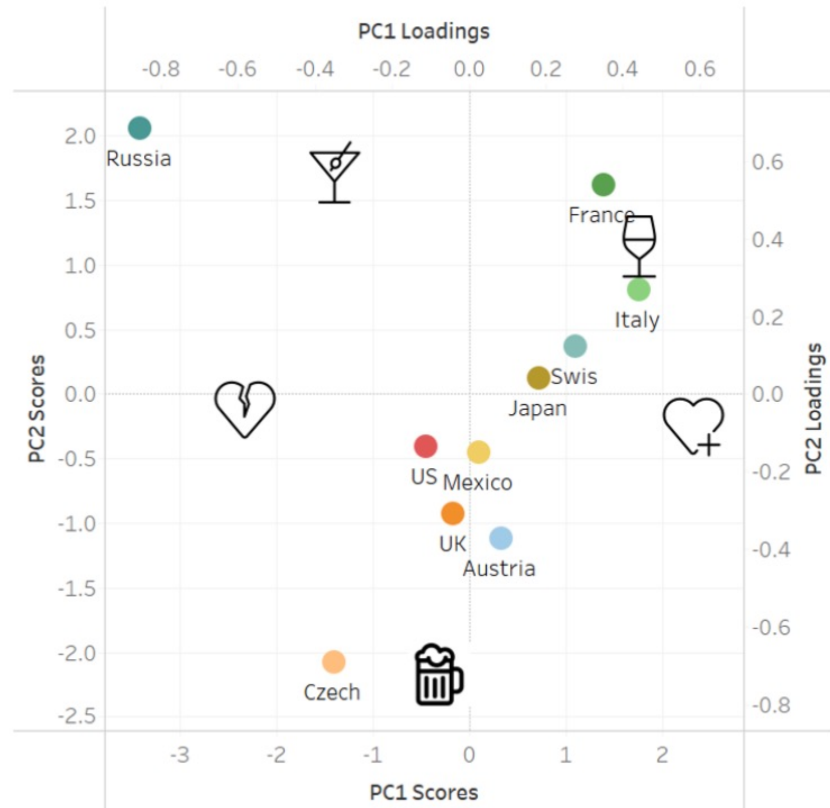
	Liquor	Wine	Beer	Life Ex	Heart D
France	2.5	63.5	40.1	78.0	61.1
Italy	0.9	58.0	25.1	78.0	94.1
Swis	1.7	46.0	65.0	78.0	106.4
Austria	1.2	15.7	102.1	78.0	173.0
UK	1.5	12.2	100.0	77.0	199.7
US	2.0	8.9	87.8	76.0	176.0
Russia	3.8	2.7	17.1	69.0	373.6
Czech	1.0	1.7	140.0	73.0	283.7
Japan	2.1	1.0	55.0	79.0	34.7
Mexico	0.8	0.2	50.4	73.0	36.4





https://www.thedataschool.co.uk/robbin-vernooij/principal-component-analysis-alteryx-example-pinguin/





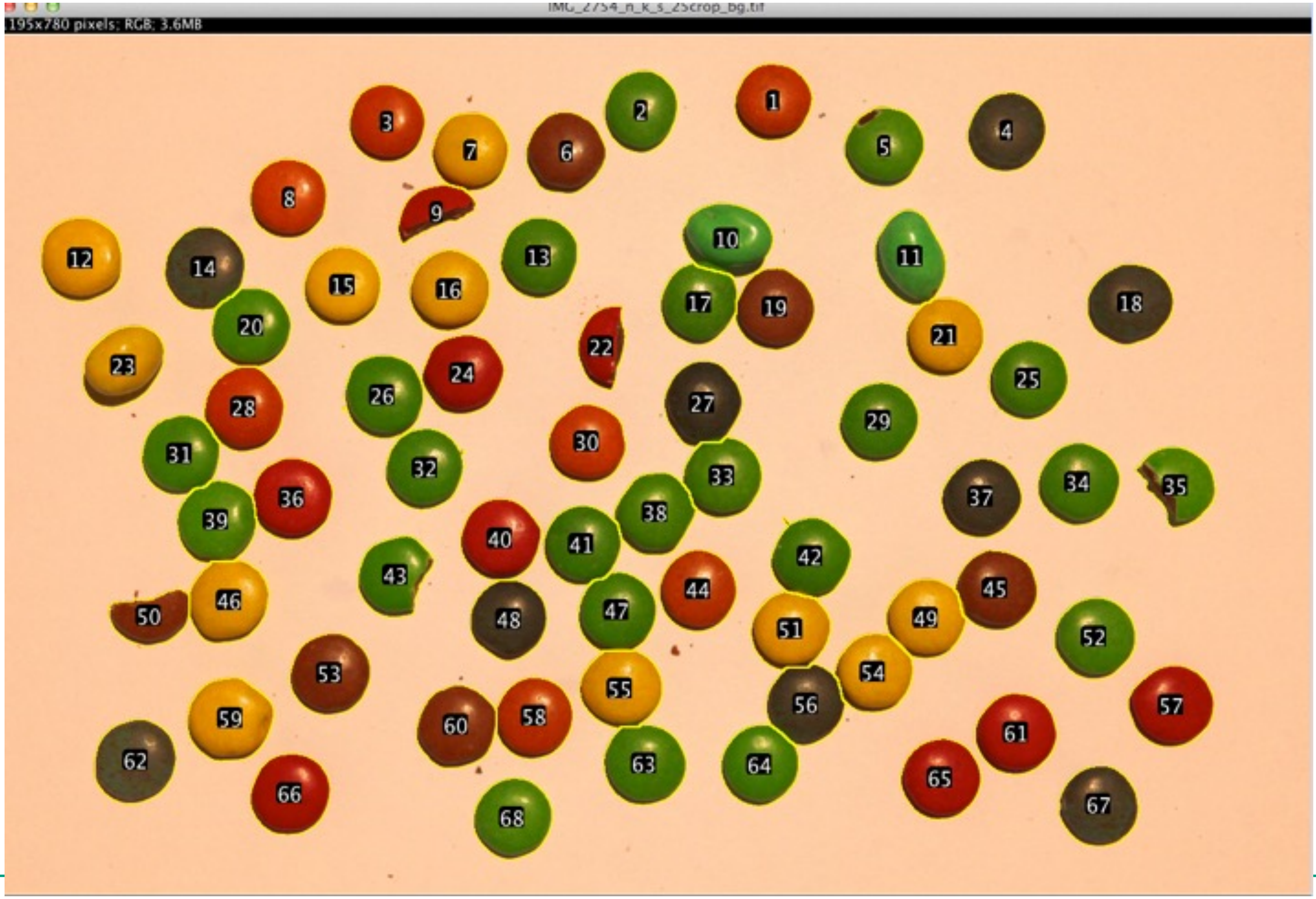
- Does this mean that you live longer if you drink wine, shorter if you drink Liquor?
- The analysis is about correlation, not causality

MULTIVARIATE ANALYSIS ON IMAGES

- Features in images are used in analysis
  - Matrix of Objects Shape, size, distribution etc
- Surface texture analysis
  - Image Surface texture

UNIVARIATE ANALYSIS ON IMAGES

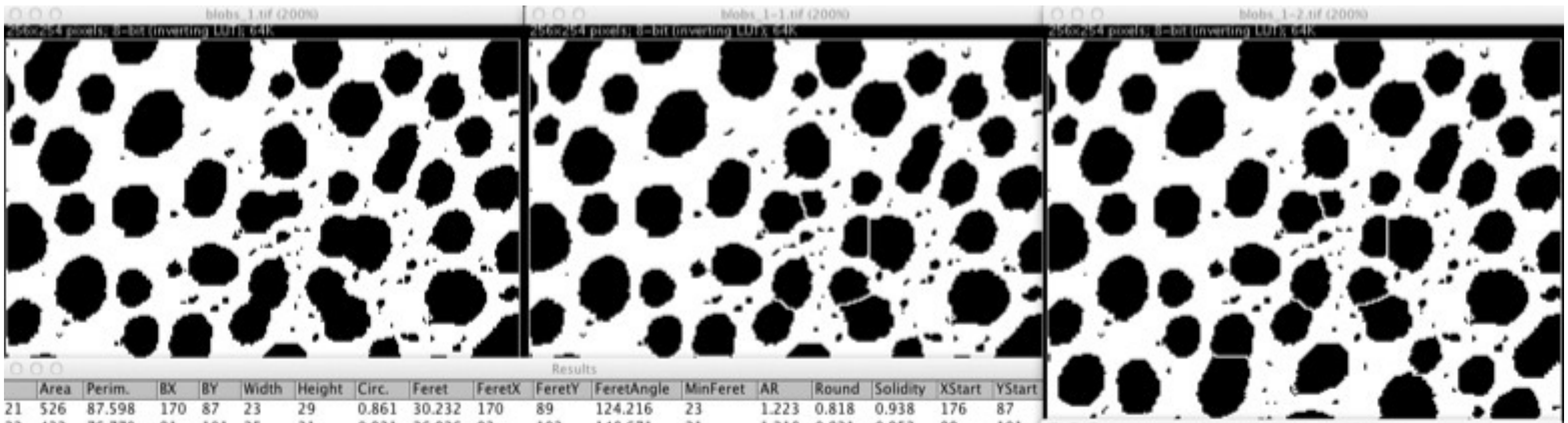
- OBJECT analysis of chocolate



UNIVARIATE ANALYSIS ON IMAGES

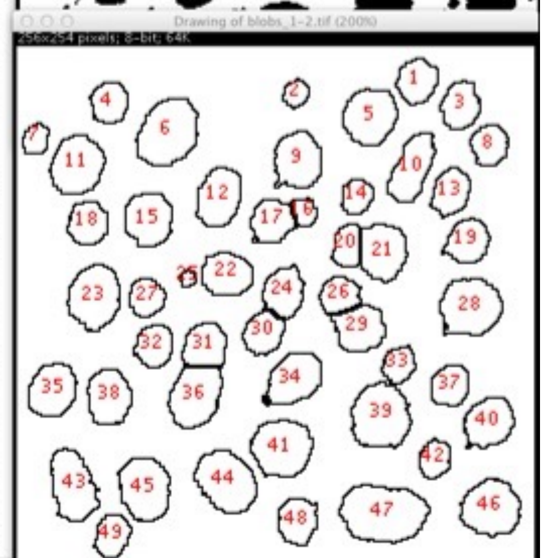
- OBJECT analysis

Figur 1. Sortert mhp «Roundness»



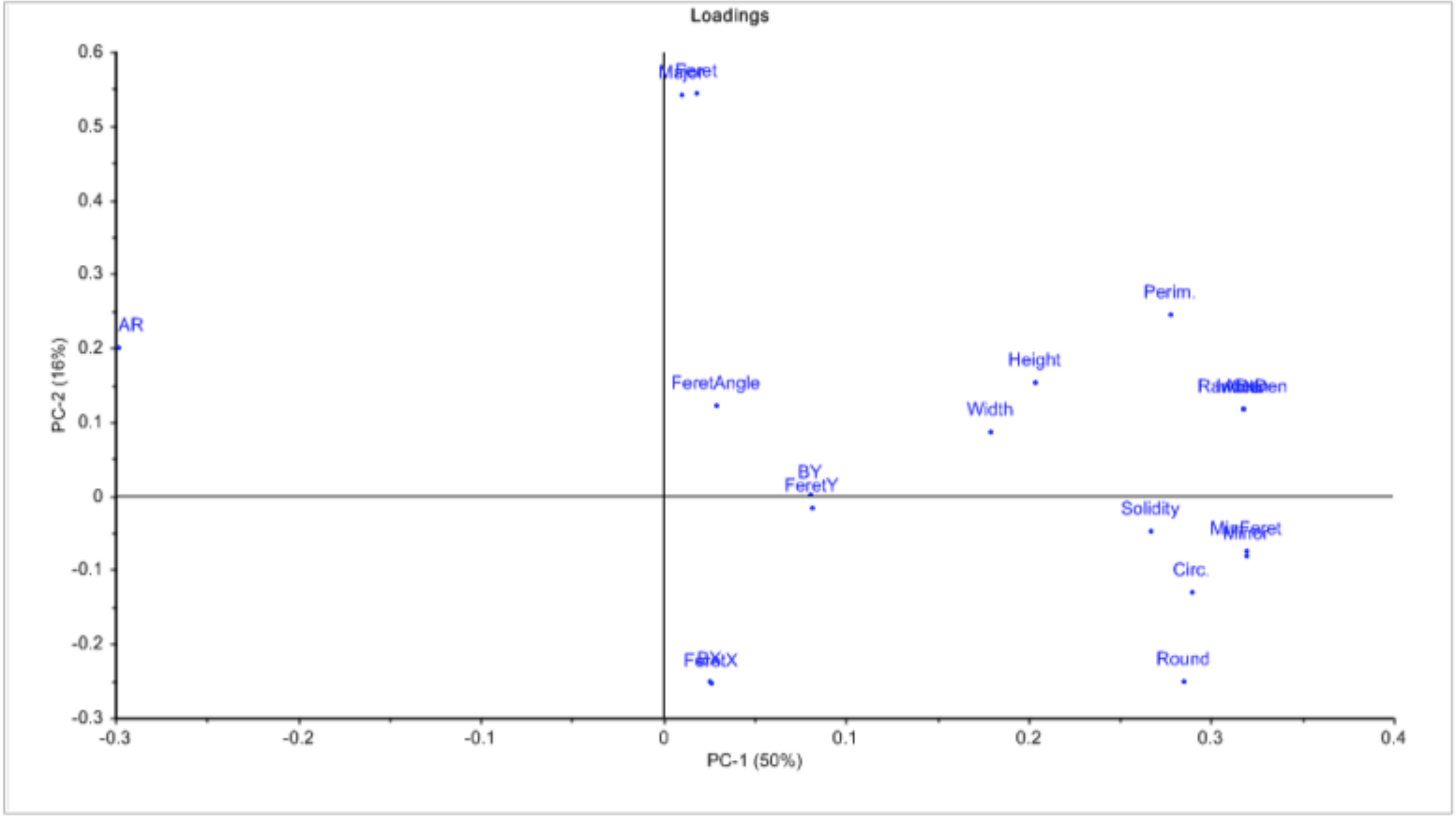
Results

	Area	Perim.	BX	BY	Width	Height	Circ.	Feret	FeretX	FeretY	FeretAngle	MinFeret	AR	Round	Solidity	XStart	YStart
21	526	87.598	170	87	23	29	0.861	30.232	170	89	124.216	23	1.223	0.818	0.938	176	87
22	432	76.770	91	101	25	21	0.921	26.926	93	103	148.671	21	1.218	0.821	0.952	99	101
23	635	97.397	24	107	26	33	0.841	33.956	34	140	76.373	26	1.269	0.788	0.936	37	107
24	385	78.912	121	107	21	27	0.777	27.459	132	134	79.509	20.478	1.263	0.792	0.902	136	107
25	48	26.385	80	109	8	9	0.866	9.849	81	118	66.038	8	1.092	0.916	0.842	84	109
26	280	66.426	149	113	21	19	0.797	22.204	152	131	35.838	18.029	1.172	0.853	0.903	156	113
27	261	59.355	55	114	18	19	0.931	19.925	60	133	72.474	18	1.072	0.933	0.927	60	114
28	688	105.983	209	114	31	28	0.770	34.828	211	142	39.174	28	1.094	0.914	0.928	220	114
29	451	84.426	155	127	27	23	0.795	29.155	155	132	157.834	22.781	1.124	0.890	0.902	170	127
30	326	70.669	111	129	21	23	0.820	23.601	113	149	36.384	19.508	1.177	0.850	0.901	121	129
31	383	75.698	81	136	22	22	0.840	25.612	82	156	51.340	21.592	1.098	0.911	0.928	91	136
32	309	67.355	57	137	19	21	0.856	22.472	64	158	69.146	18.981	1.172	0.853	0.928	66	137
33	223	60.770	180	147	17	19	0.759	21.471	184	166	62.241	16.550	1.232	0.811	0.896	193	147
34	530	101.983	121	151	29	26	0.640	35.228	121	175	34.592	22.627	1.460	0.685	0.921	134	151
35	491	82.426	5	156	24	26	0.908	29.206	7	180	51.953	23.580	1.230	0.813	0.947	17	156
36	621	99.497	74	157	27	31	0.788	35	78	186	53.130	25.048	1.371	0.729	0.930	82	157
37	305	64.284	205	157	18	20	0.927	23.022	207	176	55.620	18	1.168	0.856	0.958	212	157
38	474	83.598	34	159	22	27	0.852	28.792	39	186	69.677	22	1.294	0.773	0.940	41	159
39	740	109.296	165	165	30	33	0.778	34.132	169	196	58.173	30	1.106	0.904	0.921	179	165
40	468	89.740	221	173	25	25	0.730	31.401	222	198	52.765	23.202	1.230	0.813	0.909	232	173
41	671	100.326	115	184	31	28	0.838	32.650	122	212	49.970	28	1.066	0.938	0.936	129	184
42	228	57.355	199	193	15	20	0.871	20.396	203	213	78.690	15	1.275	0.784	0.923	206	193
43	604	104.225	16	198	24	36	0.699	37.363	22	198	105.524	23.277	1.798	0.556	0.900	22	198
44	689	101.397	87	199	31	31	0.842	35	92	201	126.870	26.163	1.325	0.755	0.937	97	199
45	626	96.569	49	203	26	31	0.844	32.894	56	203	109.537	25.608	1.301	0.769	0.940	56	203
46	630	98.225	219	213	30	29	0.821	31.321	219	233	16.699	28.986	1.065	0.939	0.918	232	213
47	971	124.711	159	216	45	29	0.785	45.177	159	231	5.080	28.056	1.563	0.640	0.939	170	216
48	353	73.598	129	223	19	23	0.819	24.839	132	244	49.899	19	1.201	0.833	0.931	134	223
49	285	64.184	37	231	19	21	0.869	22.847	42	252	66.801	17.398	1.343	0.744	0.913	43	231



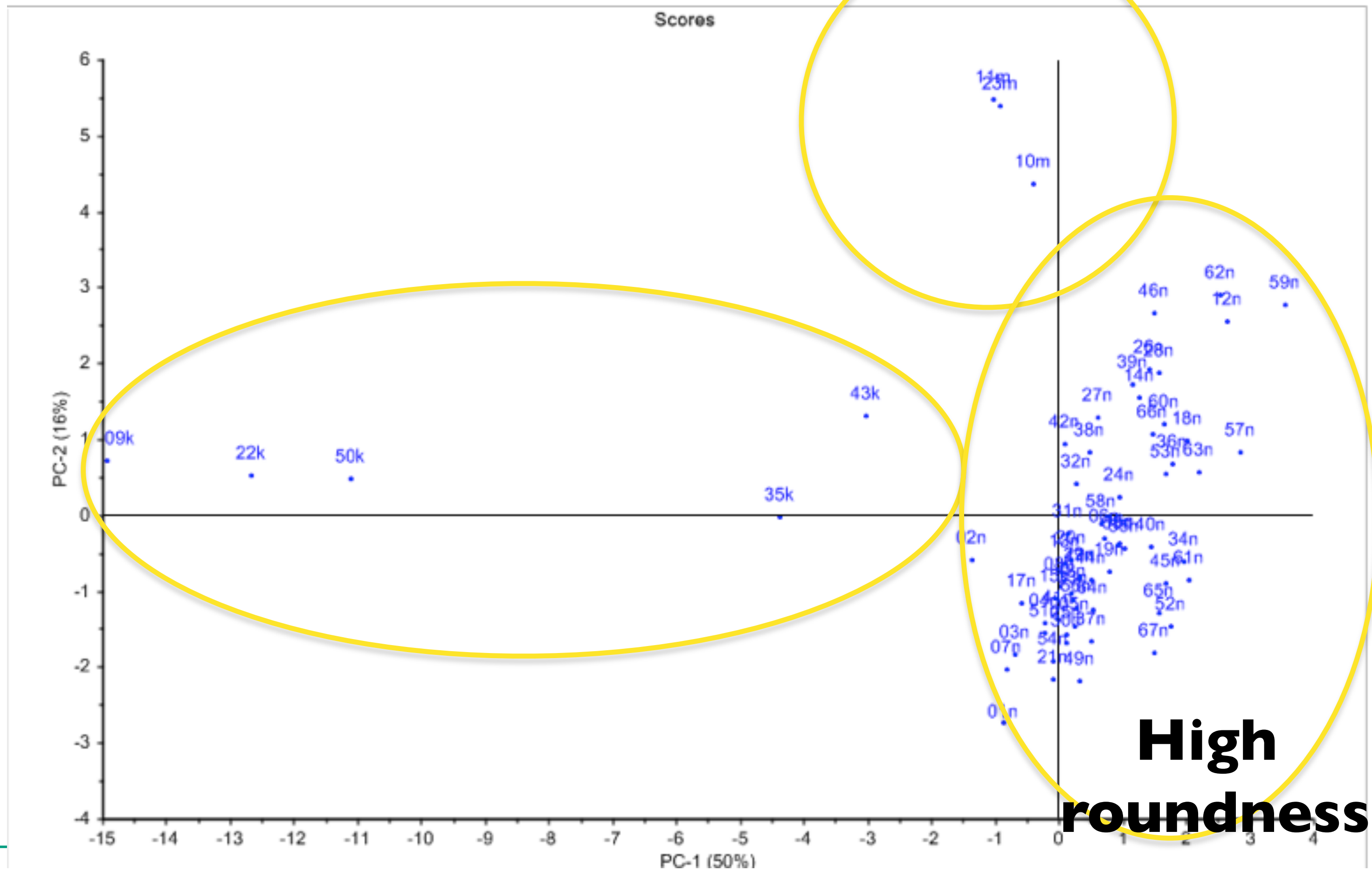
MULTIVARIATE ANALYSIS ON IMAGES

- PCA loadings on shape descriptors (area, roundness, ellipse etc)

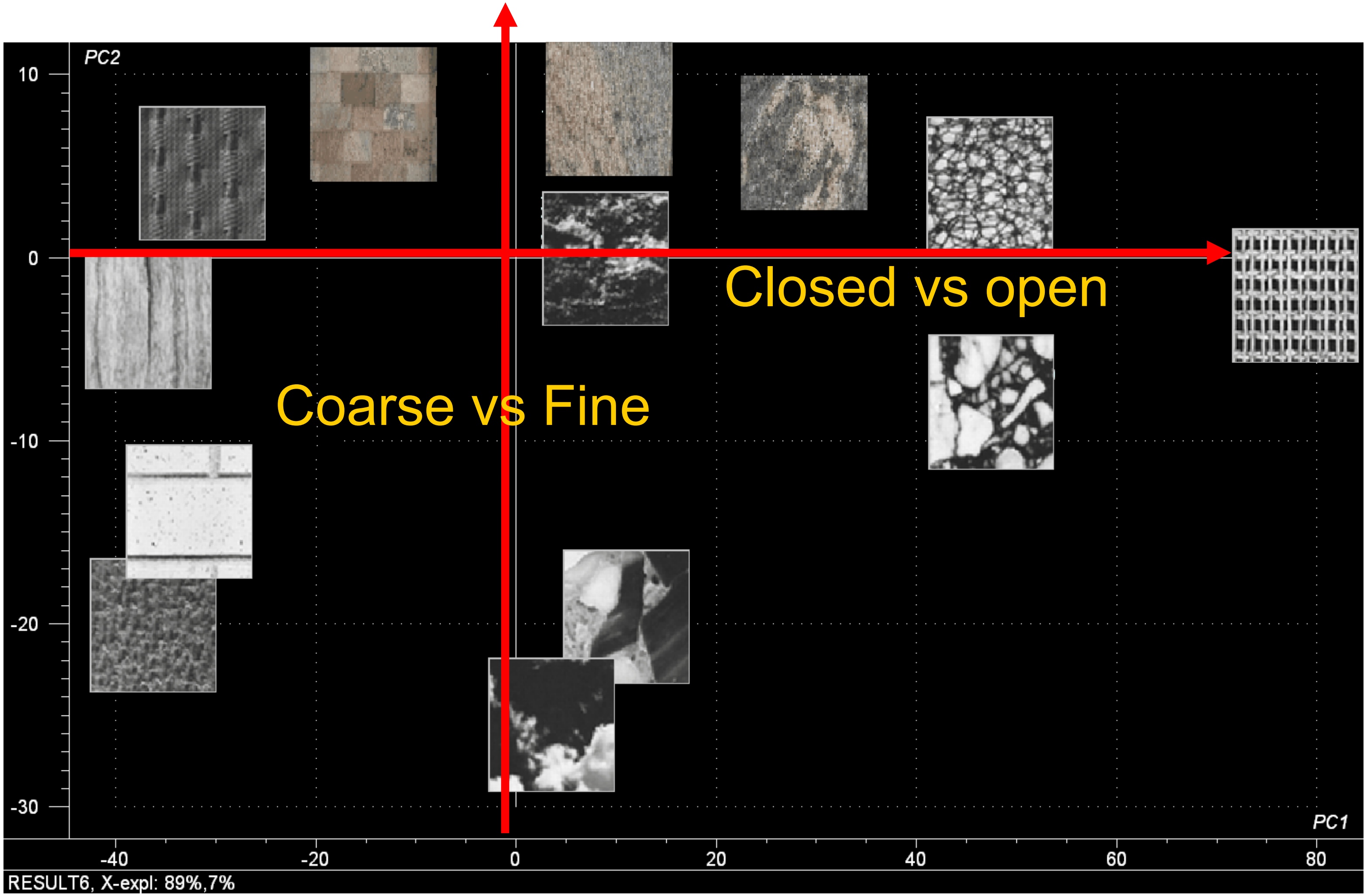


MULTIVARIATE ANALYSIS ON IMAGES

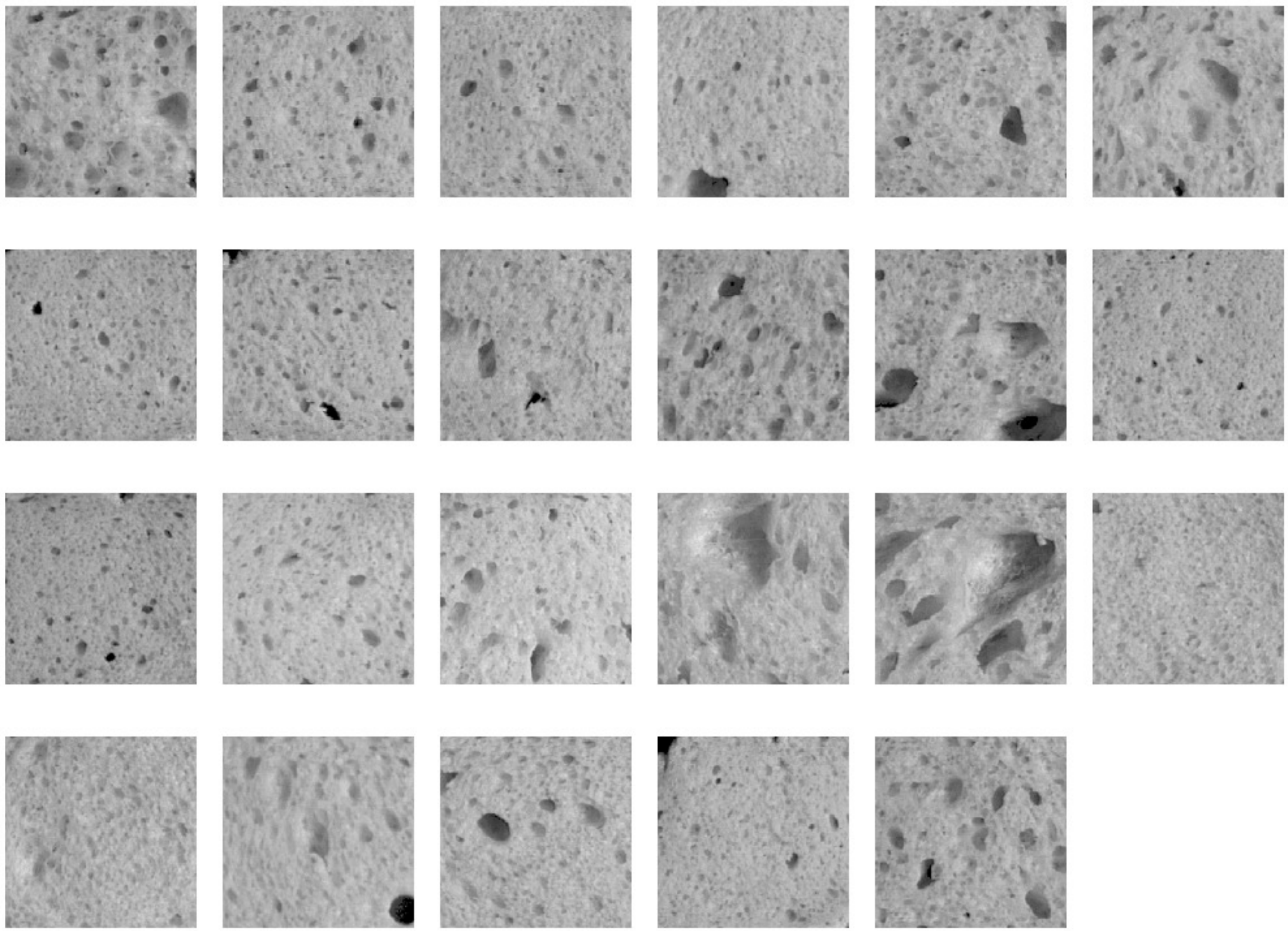
- PCA scores on shape descriptors (area, roundness etc)



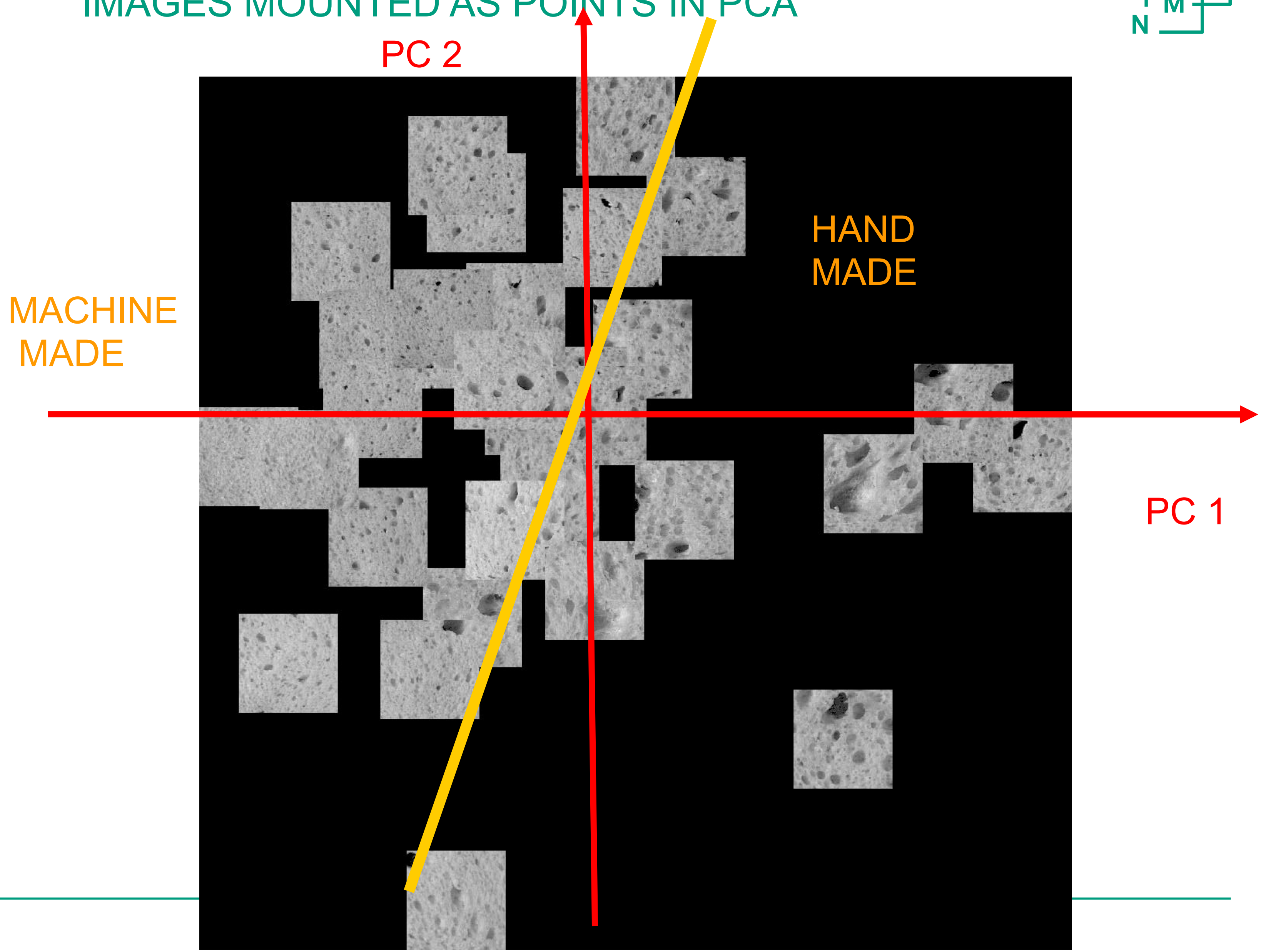
Principal Components Analysis of textures



Baguette textures and modeling of sensory porosity



IMAGES MOUNTED AS POINTS IN PCA



Frost damage detection

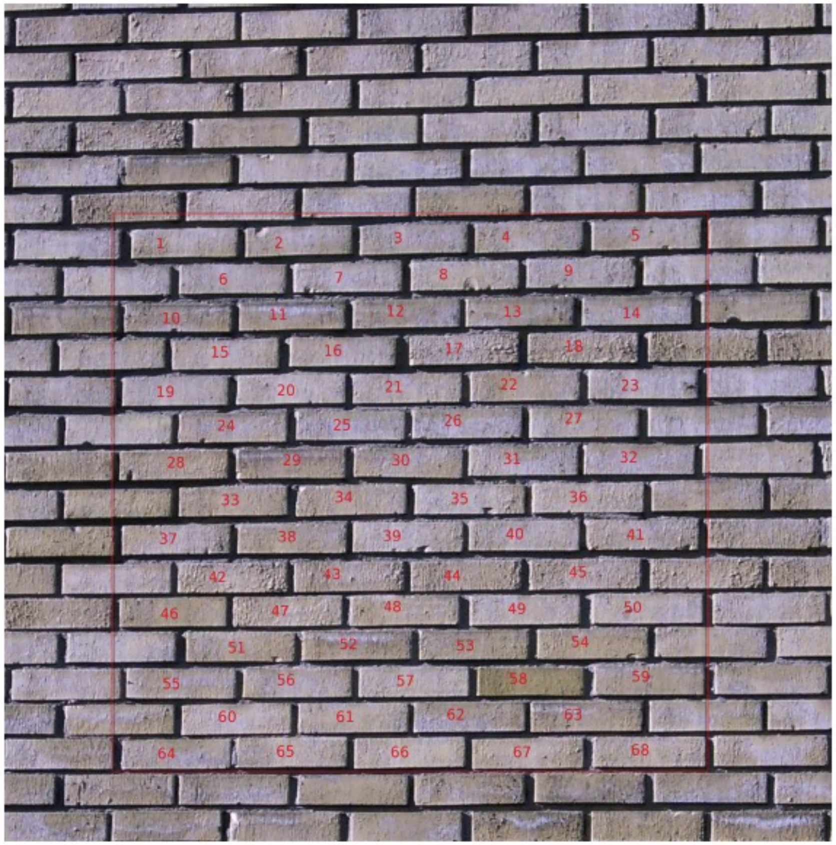


Fig 3: The original wall

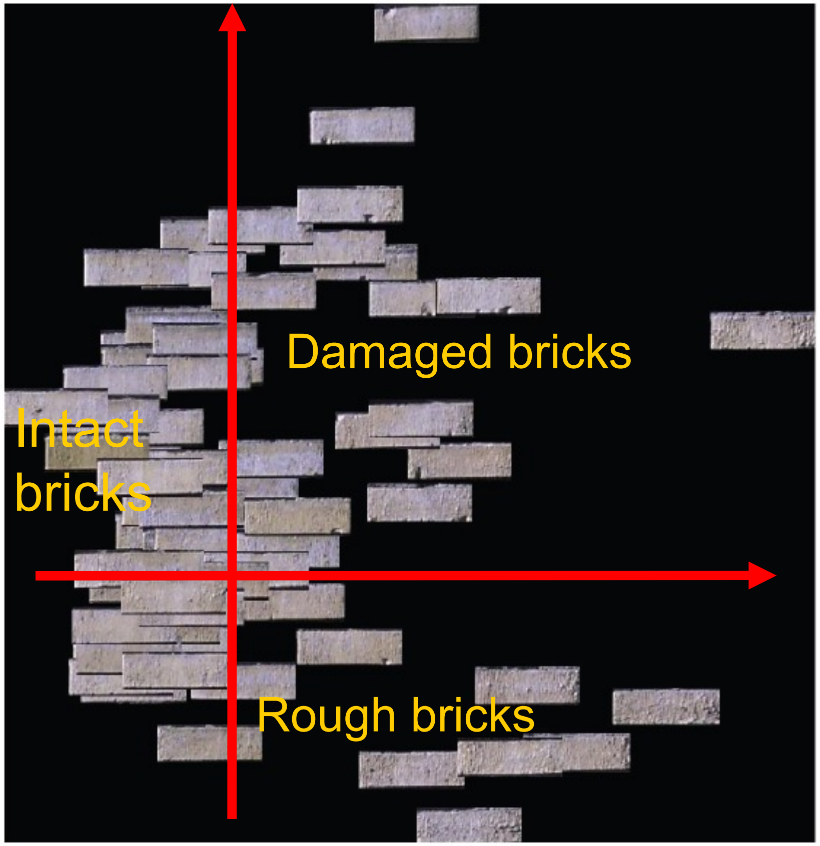
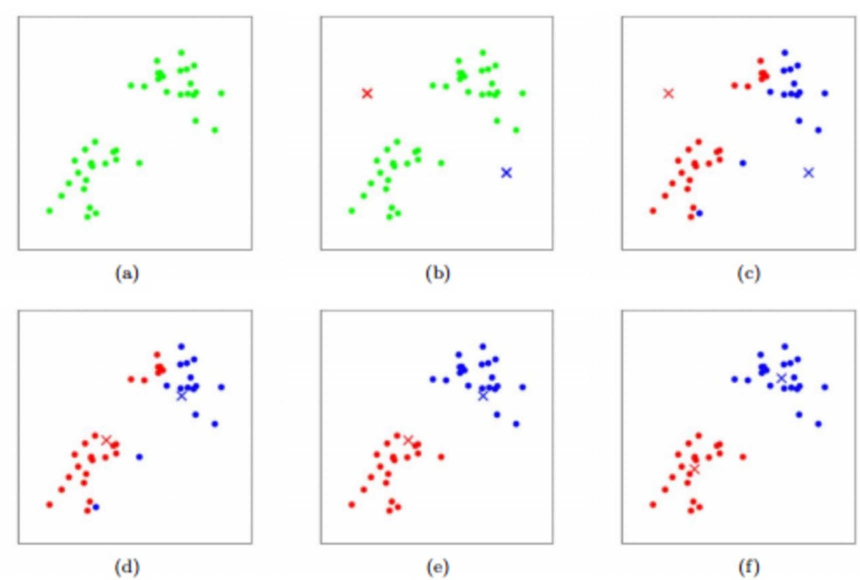


Fig 4: The bricks mounted in a PCA plot

Clustering

- K-means clustering
  - Separates samples into k groups of equal variance



  - https://www.youtube.com/watch?v=4b5d3muPQmA&ab_channel=StatQuestwithJoshStarmer

---

  - https://www.naftaliharris.com/blog/visualizing-k-means-clustering/