

PRINCIPAL COMPONENT ANALYSIS

PCA

UNSUPERVISED LEARNING

1. Part of data **exploration**

2. Used for **visualizing**

...understand the structure in data (sub-groups)

...reveal basic trends

3. or data **pre-processing**

...variable dimensionality reduction

USE CASE

Drug design case:

We have data about 1200 drugs candidates from two different studies. For each drug we have calculated 187 physicochemical parameters. We would like to pool the data from these two studies together to have more degrees of freedom.

Before we can start doing any regression we would need to make sure that the dataset (study 1, study 2) are homogenous or at least overlapping.

VISUALIZATION

We have n datapoints

...with p features X_1, X_2, \dots, X_p

We would like to have a look on the data structure:

...we could look at the scatter plots one-by-one

$p = 10$ results in 45 plots

Most likely none of them will be very informative, as they contain only a small fraction of the information present in the dataset

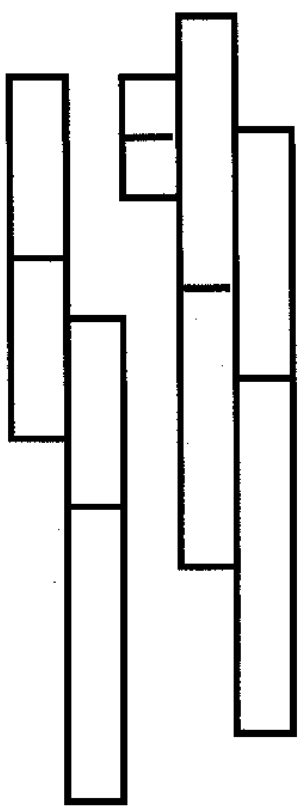

We need **low dimensionality** representation capturing **as much information as possible**

DATA PREPROCESSING

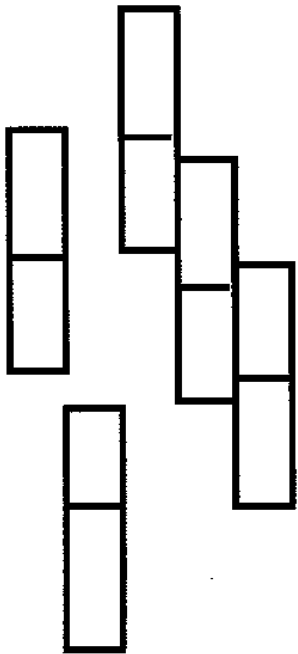

Centering – bringing the mean of the variable values to zero

Scaling – reducing the standard deviation of the variable values to 1

measured
values
&
"length"



unit variance
scaling



mean-
centering



0



FIRST PRINCIPAL COMPONENTS

The first principal component of p features is the respective normalized linear combination

$$Z_1 = \varphi_{11}X_1 + \varphi_{21}X_2 + \cdots + \varphi_{p1}X_p$$

We reduce each sample/datapoint to one value:

$$z_1 = \varphi_{11}x_1 + \varphi_{21}x_2 + \cdots + \varphi_{p1}x_p$$

FINDING THE LINEAR COMBINATION

Loadings are normalized:

$$\sum_{j=1}^p \varphi_{j1}^2 = 1$$

$$\text{maximize} \left\{ \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^p \varphi_j x_{ij} \right)^2 \right\}$$

we maximize the sample variance of the n values of z_{i1}

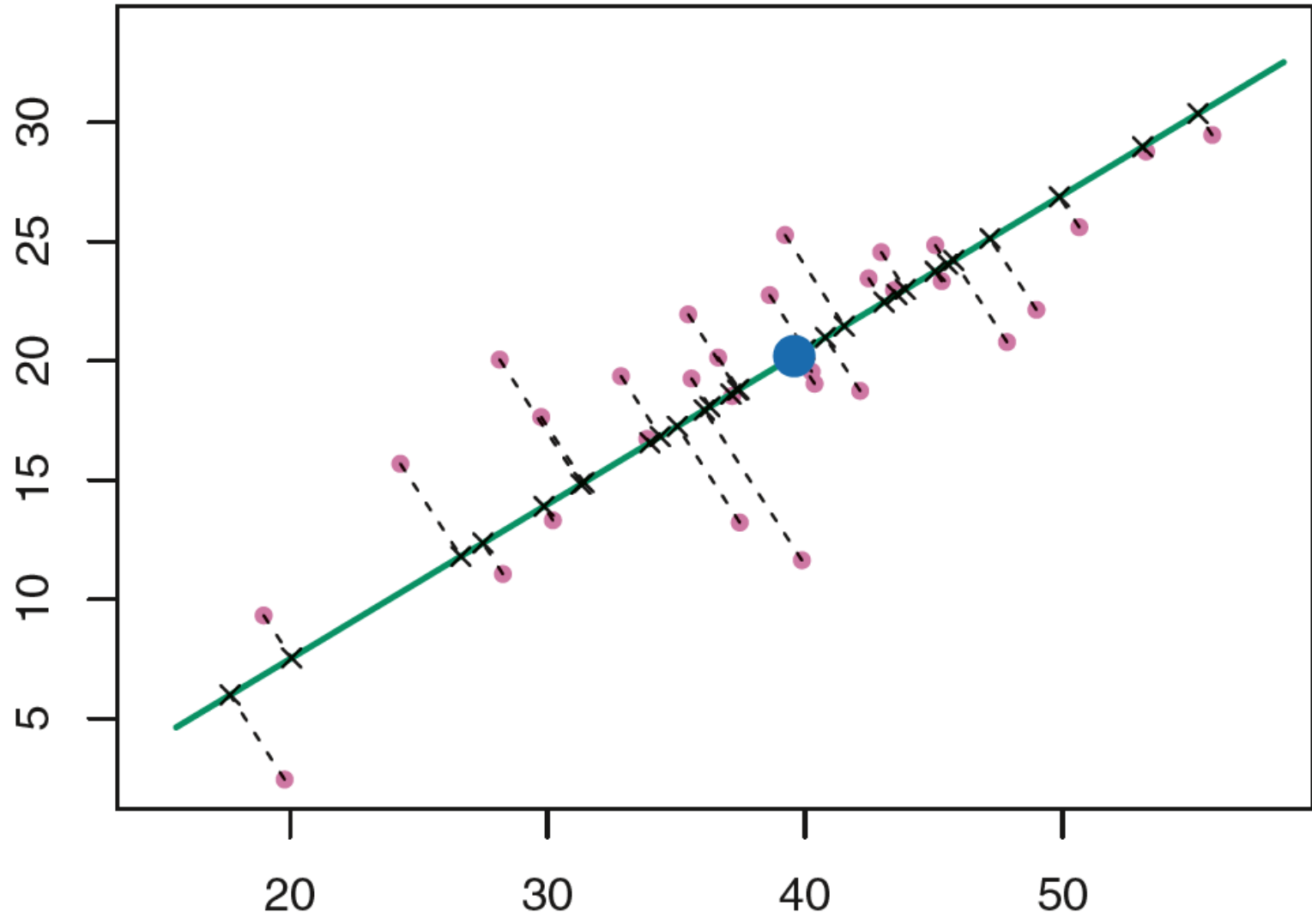
We refer to $z_{11}, z_{21}, \dots, z_{n1}$ as **scores**.

LOADINGS

Define a direction in the features space along which the data vary most

Scores are the projections of these datapoints to this direction

NB! Calculating loadings with different software (packages) at different times may flip the signs of the loadings.

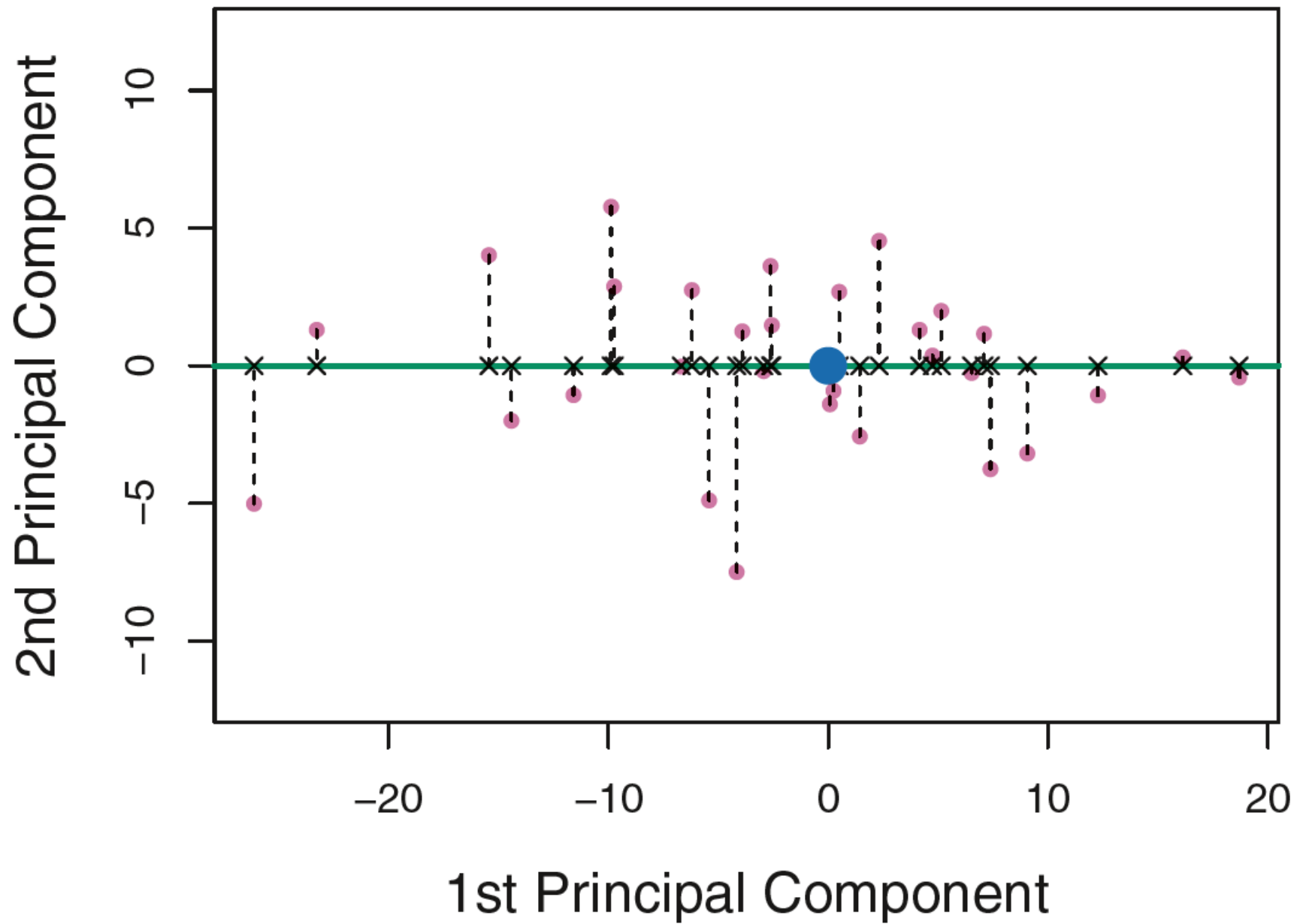


SECOND PRINCIPAL COMPONENT

Again a linear combination that has maximum variance and is **uncorrelated** to Z_1

$$Z_2 = \varphi_{12}X_1 + \varphi_{22}X_2 + \cdots + \varphi_{p2}X_p$$

e.g. the direction of second principal component is orthogonal to the first principal component

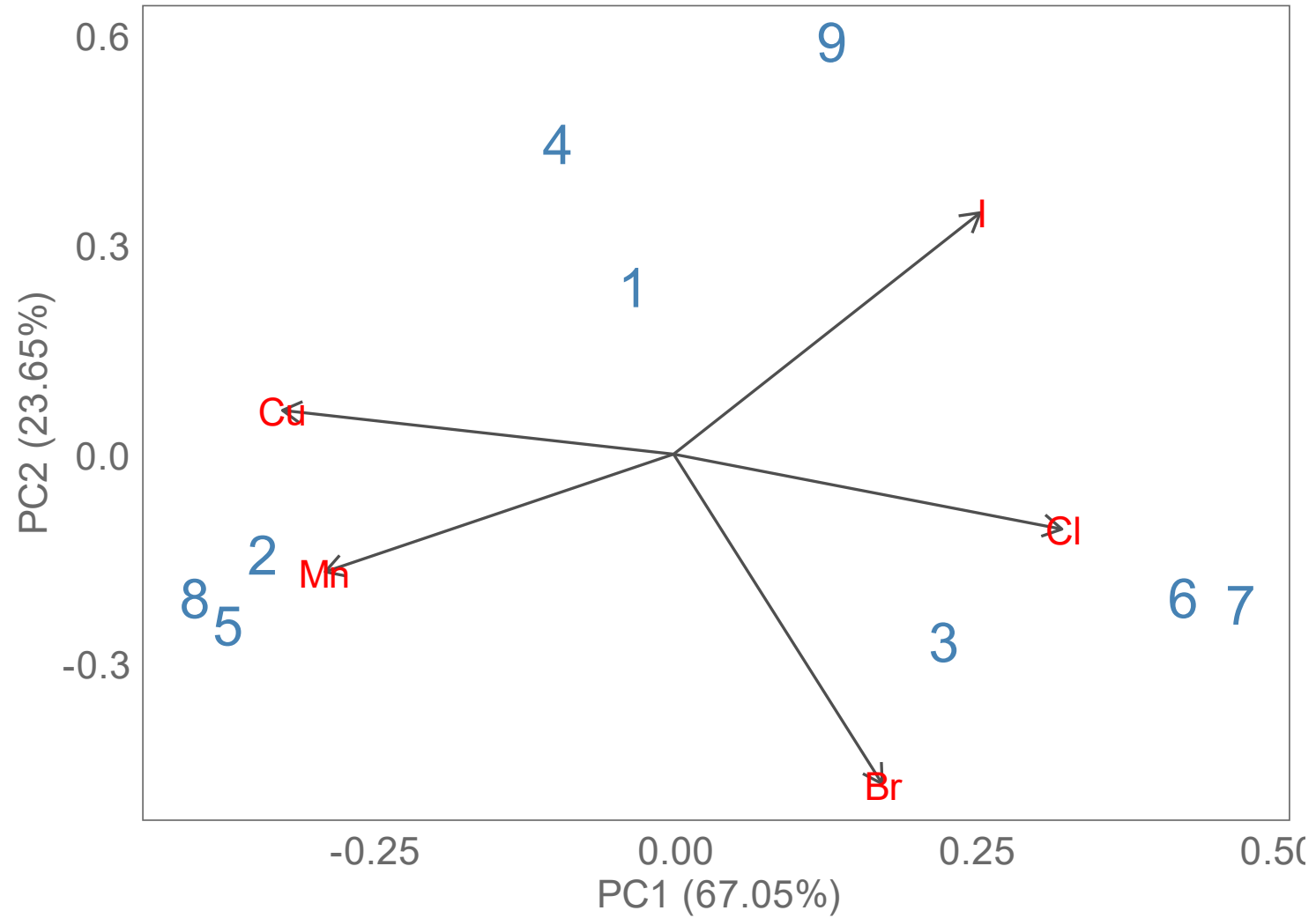


EXAMPLE

ELEMENTAL ANALYSIS OF 9 SAMPLES

Cu	Mn	Cl	Br	I
9.2	0.3	1730	12	3.6
12.4	0.39	930	50	2.3
7.2	0.32	2750	65.3	3.4
10.2	0.36	1500	3.4	5.3
10.1	0.5	1040	39.2	1.9
6.5	0.2	2490	90	4.6
5.6	0.29	2940	88	5.6
11.8	0.42	867	43.1	1.5
8.5	0.25	1620	5.2	6.2

PLOTTING THE RESULTS



EXPLAINED VARIANCE

Total variance in the data

$$\sum_{j=1}^p \text{Var}(X_j) = \sum_{j=1}^p \frac{1}{n} \sum_{i=1}^n x_{ij}^2$$

Variance explained by the m th principal component

$$\frac{1}{n} \sum_{i=1}^n z_{im}^2 = \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^p \phi_{jm} x_{ij} \right)^2$$

Proportion of variance explained

$$\frac{\sum_{i=1}^n \left(\sum_{j=1}^p \phi_{jm} x_{ij} \right)^2}{\sum_{j=1}^p \sum_{i=1}^n x_{ij}^2}$$

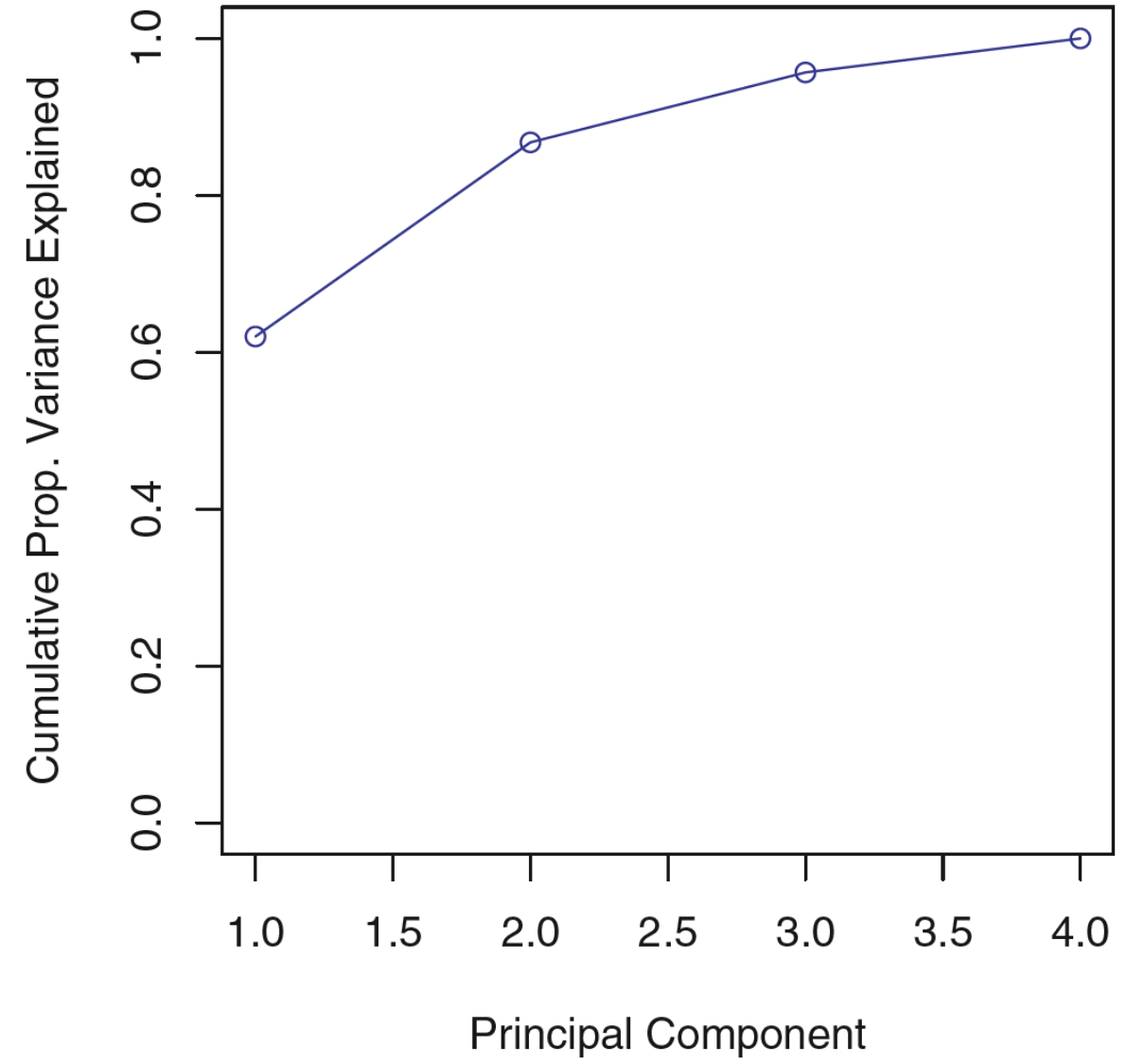
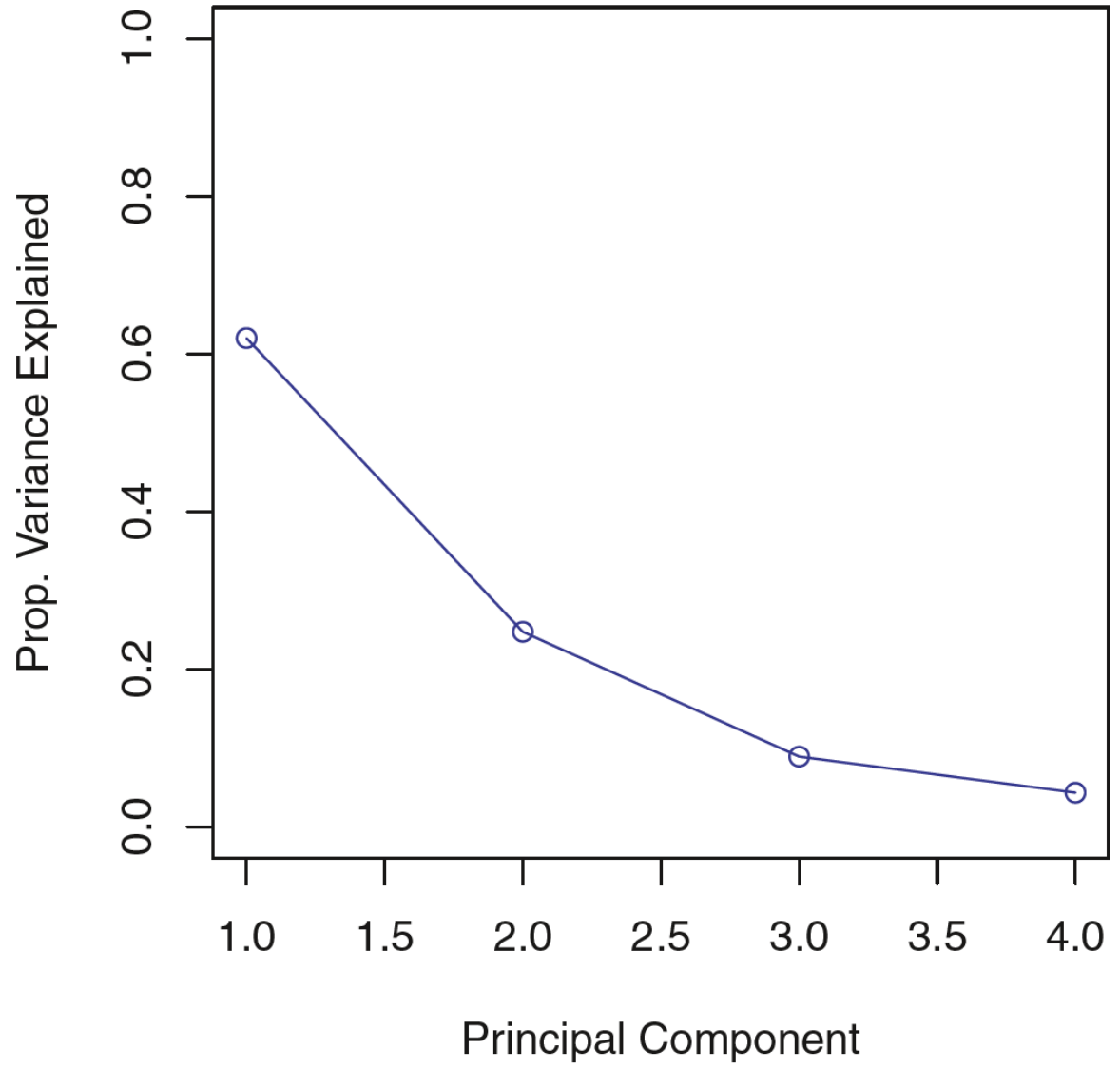
CHOOSING THE NUMBER OF PC-s?

If #PCs = #features than PCA explains 100% of the variability in the data

We need to find the # of PCs explaining the significant part of the total variance present in the data

Usually aim at 50%+

Elbow method



AFTER PCA

PCA can be used as

1. Input for regression
2. Input for classification
3. Input for clustering
4. Visualizing the homogeneity of the data
5. Visualizing the relationship between datasets

EXAMPLE OF VISUALIZING FIBER CLASSIFICATION

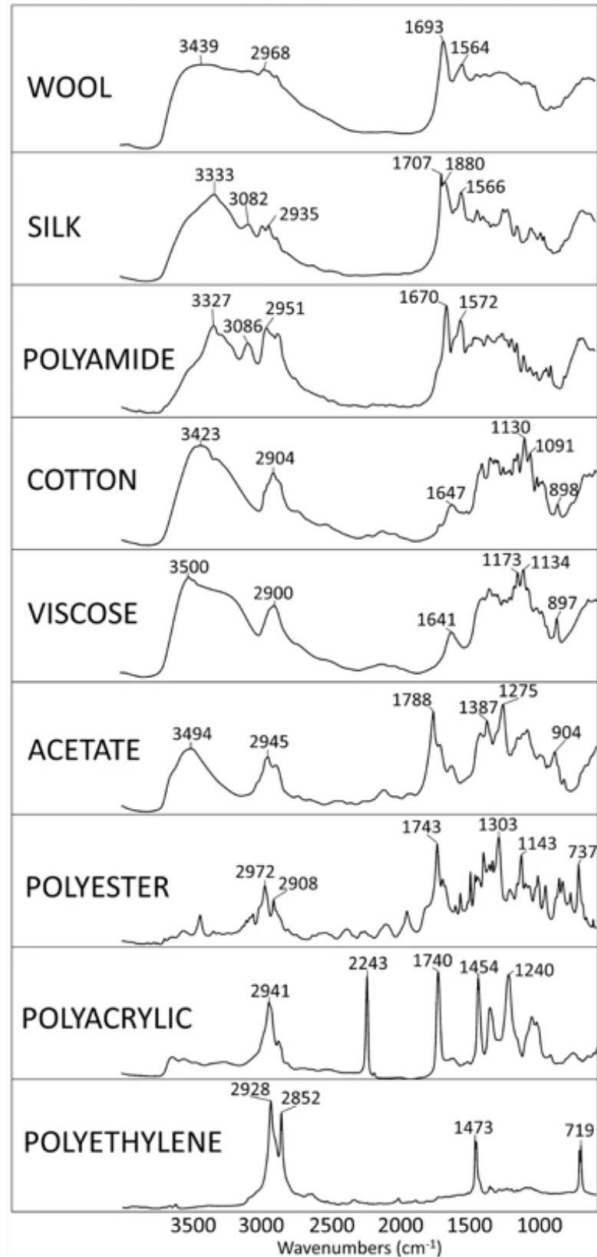
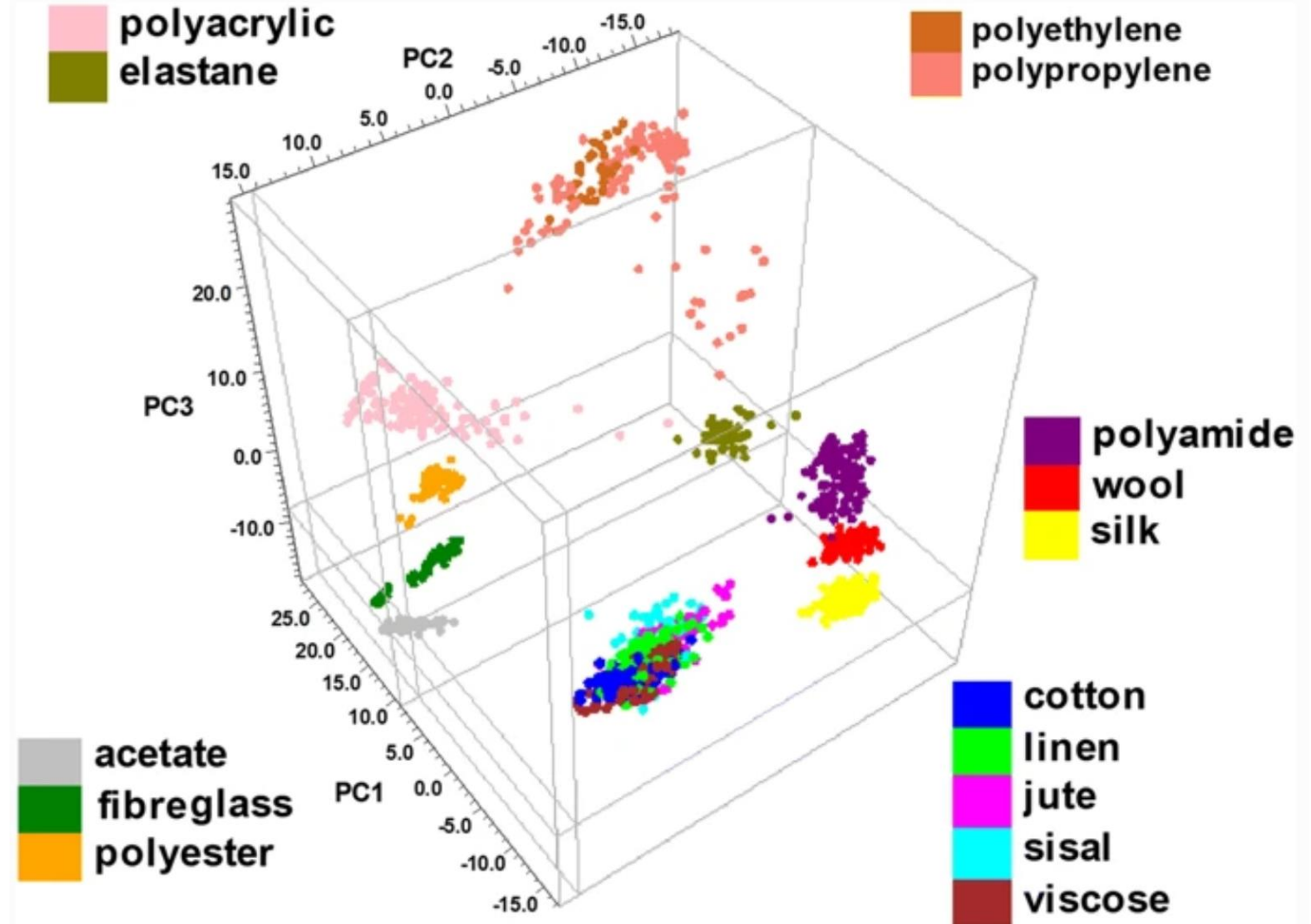


Fig. 5



NB! PCA IS NOT A CLASSIFICATION TOOL

Give me 3 reasons why not:

- 1.
- 2.
- 3.

VALIDATING THE RESULTS

Generally complicated for unsupervised learning methods

No good mechanism for cross-validation

No possibility for independent dataset based validation