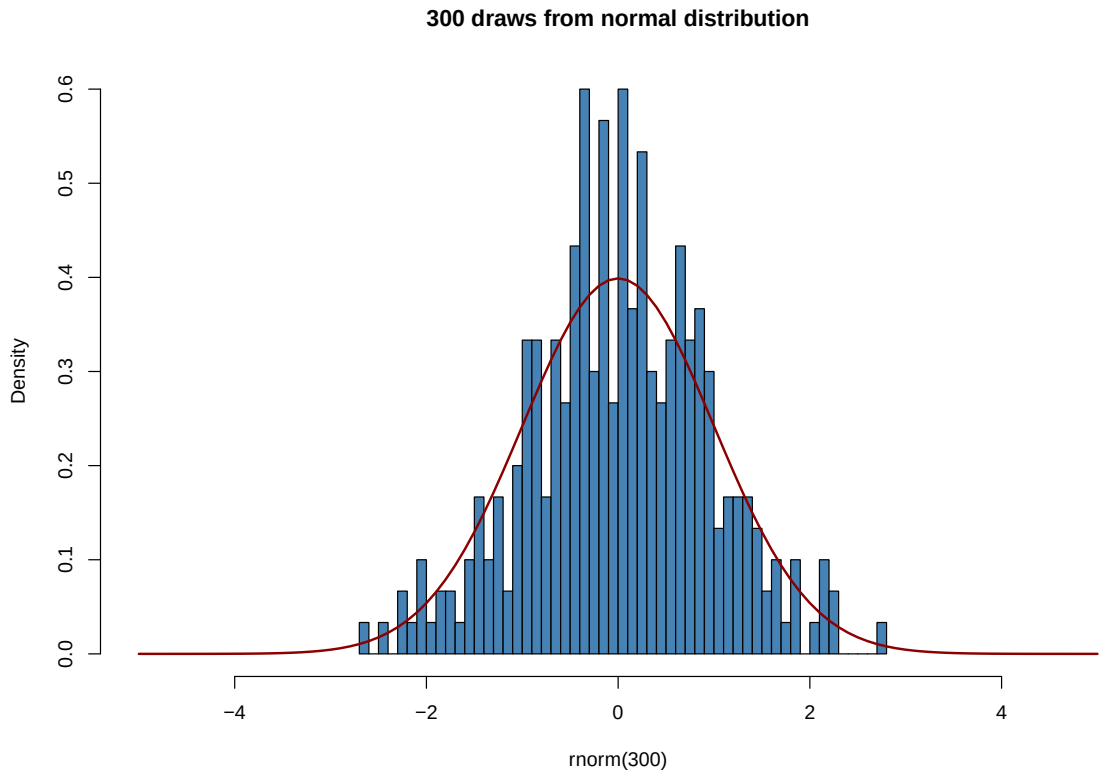
Data Exploration - PCA

Pietro Franceschi
pietro.franceschi@fmach.it

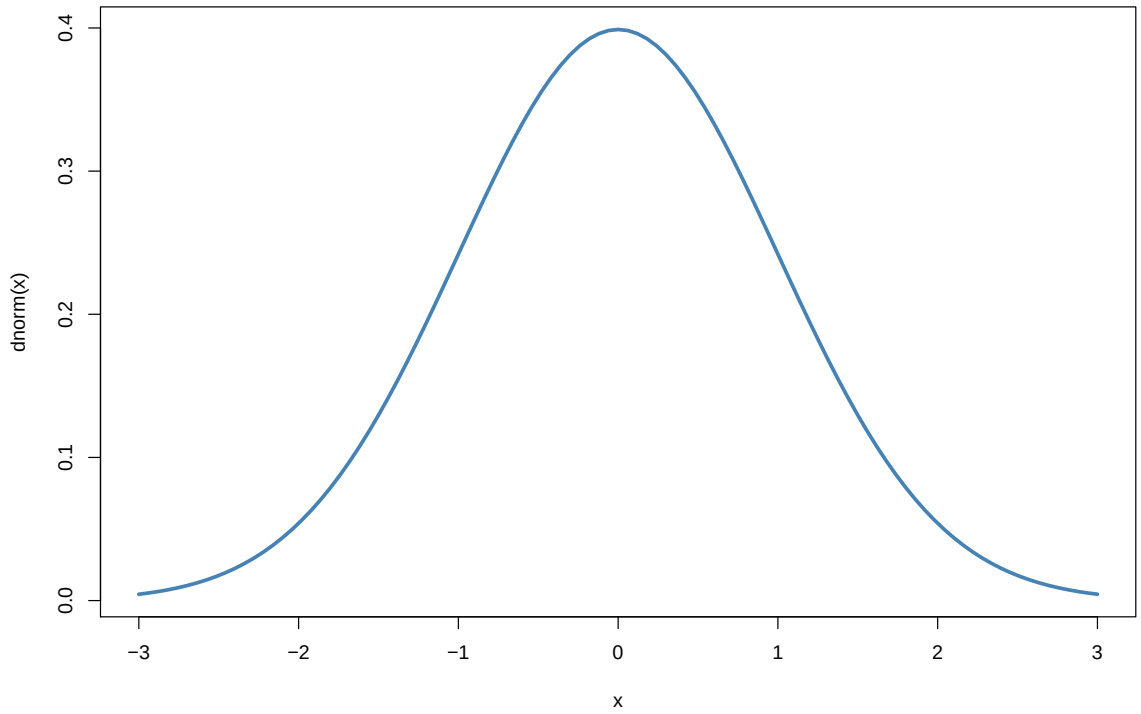
FEM - UBC

Variability and distributions



- Variability generates distributions
- Distribution is linked to probability
- Empirical distribution can be used to estimate "general" probability

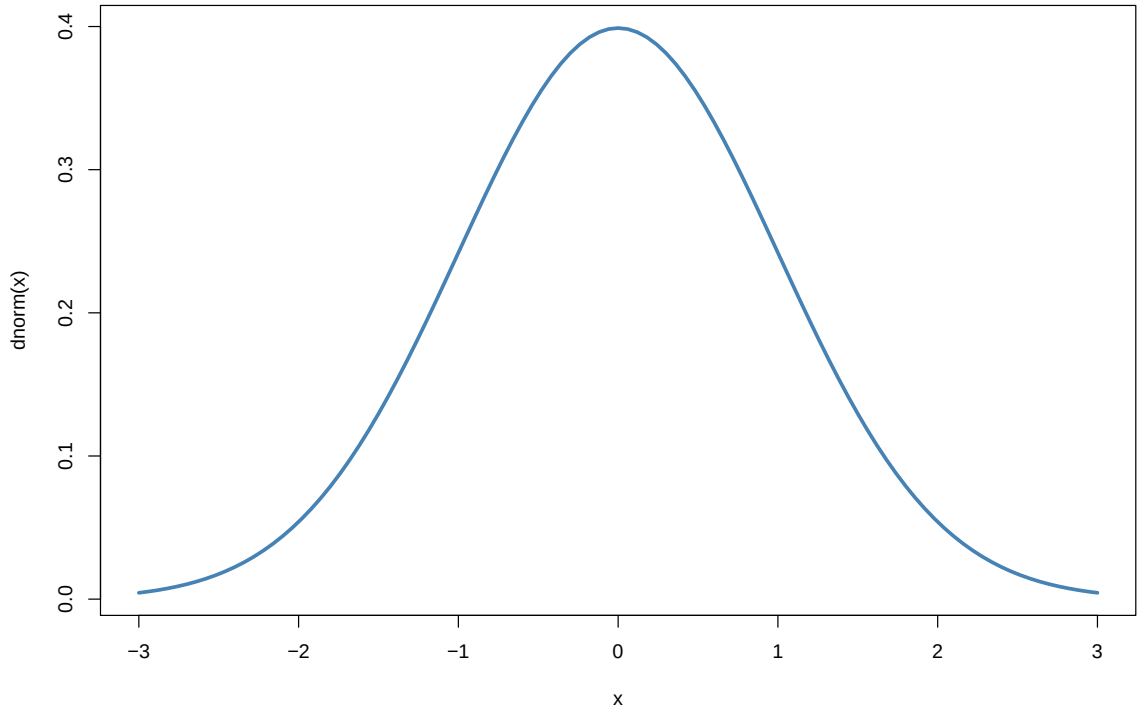
Normal Distribution



Being Normal

- has a central role in statistics
- is often a prerequisite
- is often an exception (e.g. subpopulations in my data)

Normal Distribution



Being Normal

- mean and median are the same
- the mean is the most probable value
- it makes sense to focus on the mean
- we restore normality by transforming the data

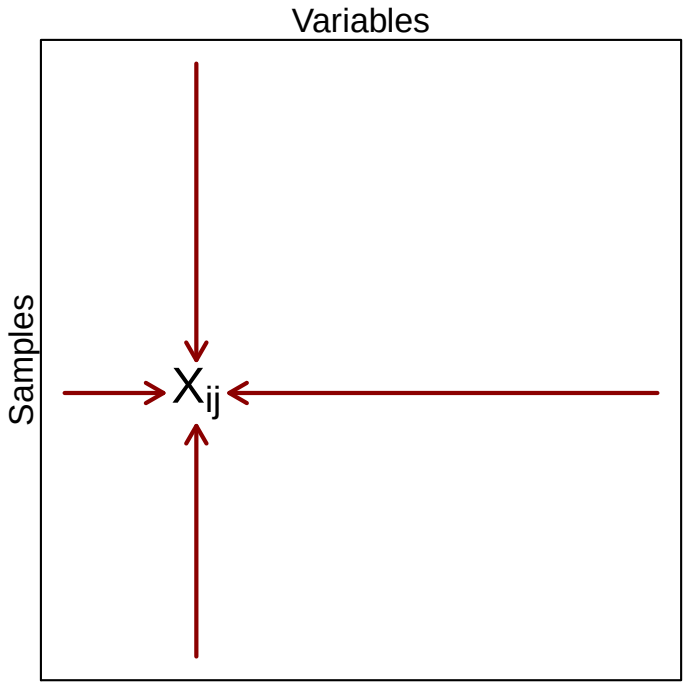
Conservation of misery

The data analyst dilemma

$$\text{Data} + \text{Knowledge} = \text{Constant}$$

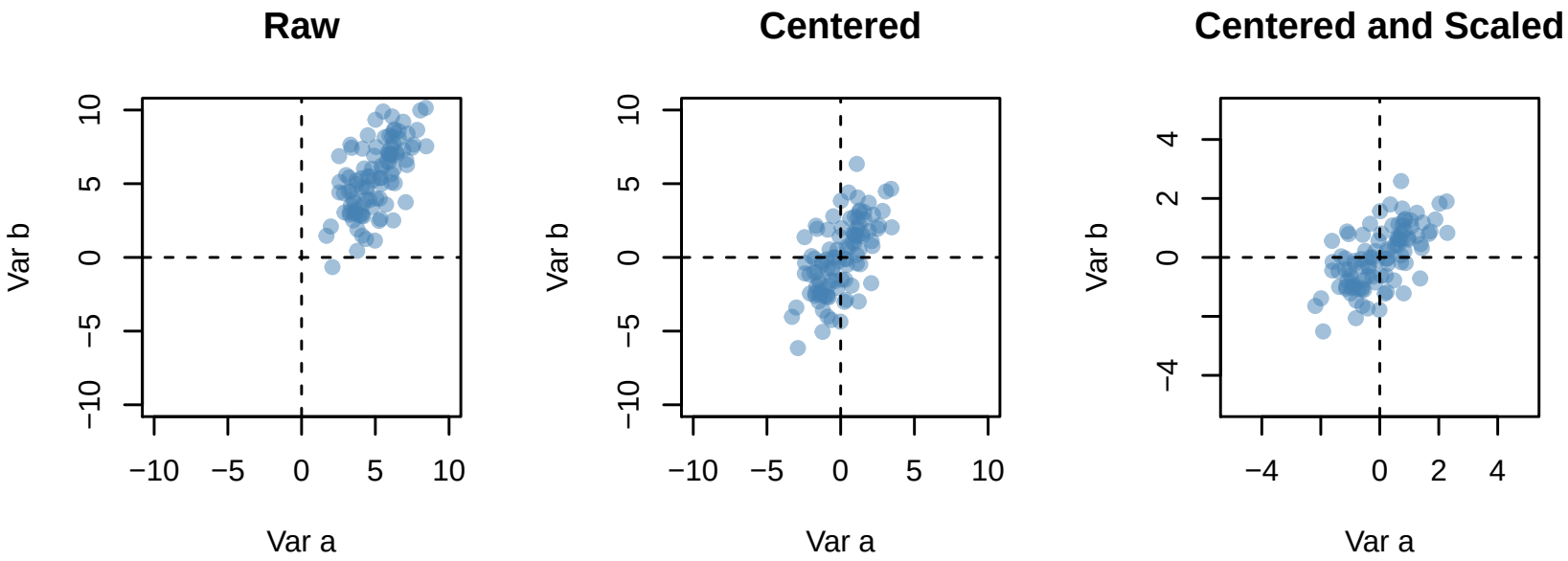
- parametric vs non parametric tests
- variable association from knowledge or data
- ...

Multivariate data - the Data Matrix



- The element X_{ij} contains the value of variable `j` in the sample `i`

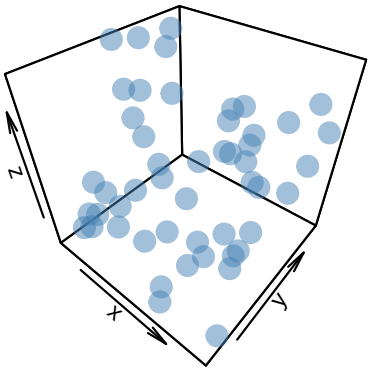
Centering and Scaling



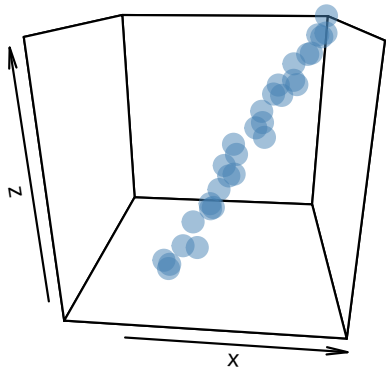
Intrinsic Dimensionality

In presence of correlation among the variables, the samples actually occupy only a "fraction" of the potential multidimensional space

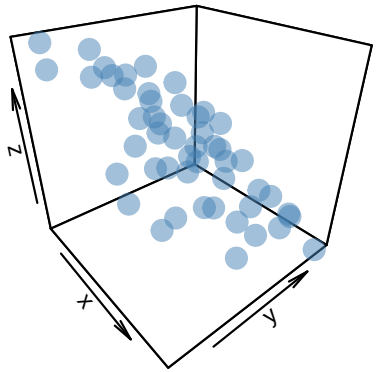
Intrinsic dim = 3



Intrinsic dim = 1



Intrinsic dim = 2



Latent Variable

Latent Variable

(Mathematical) combination of several variables. Looking to the data along **specific latent variables**, we highlight some desired property of the dataset

- Separation of sample classes (e.g. LDA)
- Prediction of sample properties (e.g. PLS)
- Good representation of the multidimensional data structure (e.g PCA, PCoA)

Loadings and Scores

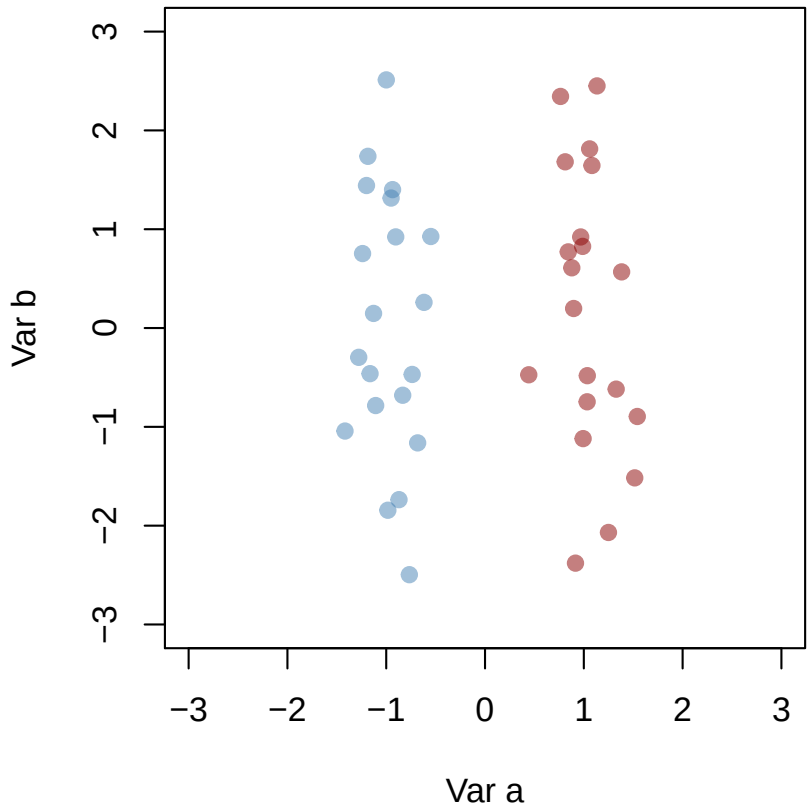
LVs and Projections

A set of latent variable can be used to reconstruct an informative representation of the dataset which captures some relevant multidimensional aspects of the data. This representation is constructed "projecting" the samples on the LVs

Loading and Scores

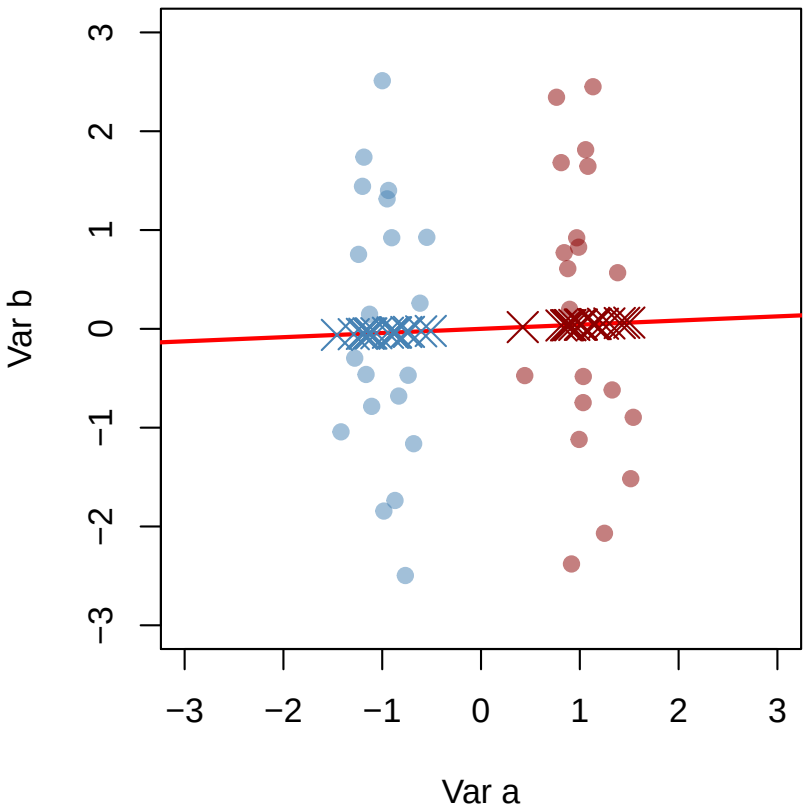
- **Scores**: the representation of the samples in the LV space
- **Loadings**: the "weight" of the original variables on the single LVs

Dummy Dataset



What LV will separate the two groups?

LV for class discrimination



- The red line represents the direction of maximal separation between the two classes
- The crosses are the **scores** along this direction

Loadings for class discrimination

The loadings represent the weight of the initial variables along the discriminating direction

- `Var a`: 0.9991168
- `Var b`: 0.0420196

Principal Component Analysis (PCA)

The aim of PCA is dimension reduction and PCA is the most frequently applied method for computing linear latent variables (components).

PCA

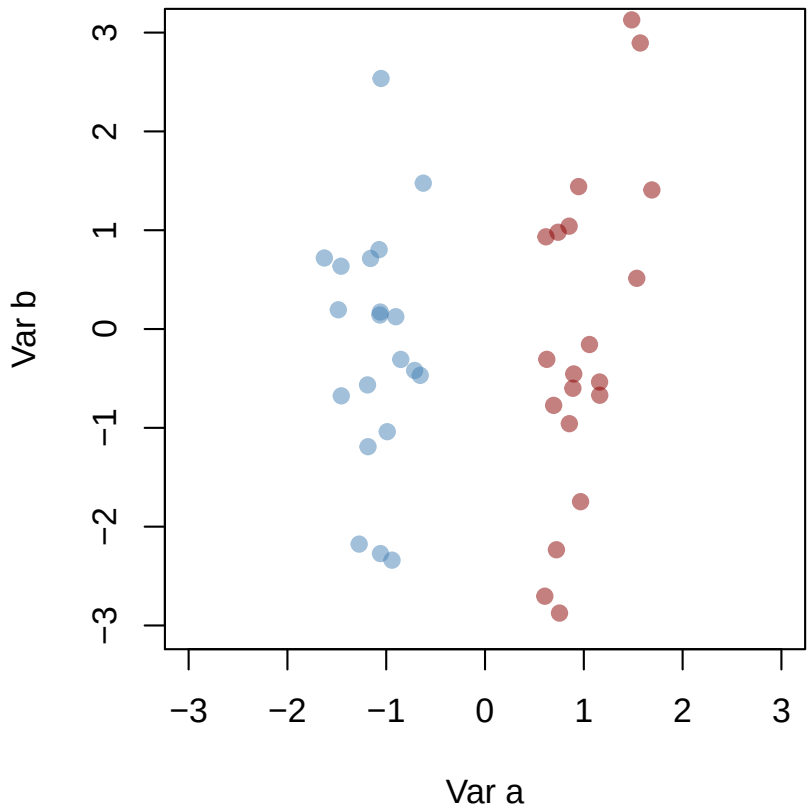
The transformation is defined in such a way way that the first principal component has the **largest possible variance** (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is **orthogonal** to the preceding components.

- In PCA the "objective" of the projection is to maximize variance
- PCA "view" will enhance the spread of the data
- The key idea is that **variability means information**

Animation!

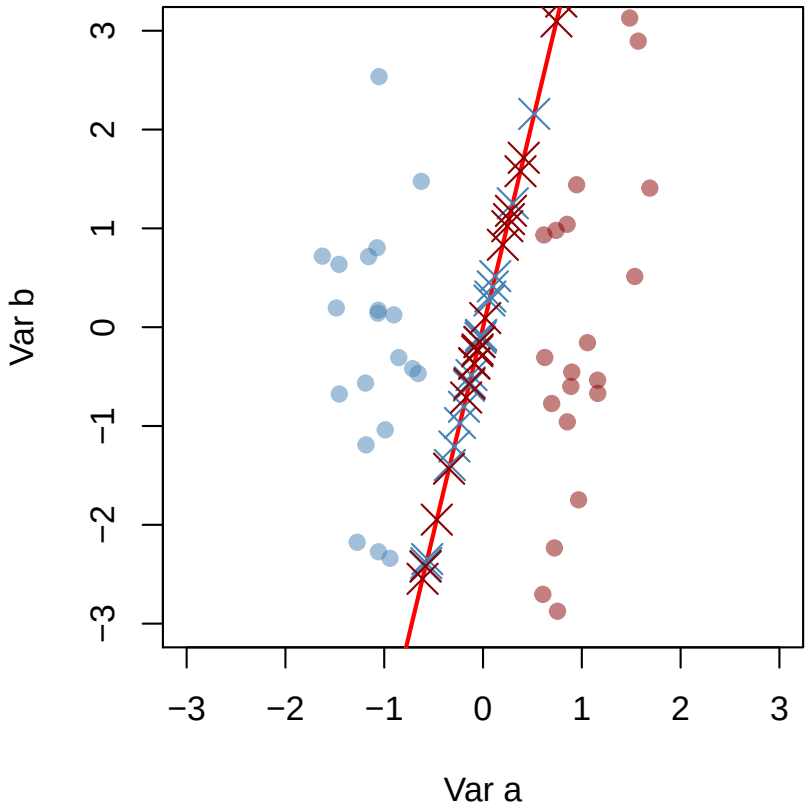
PCA Animation

Dummy Dataset



What LV will highlight the direction of maximal variance?

PCA of the dummy data



- The red line represents the direction of maximal variance (bad separation!)
- The crosses are the **scores** along this direction

Loadings for PC1

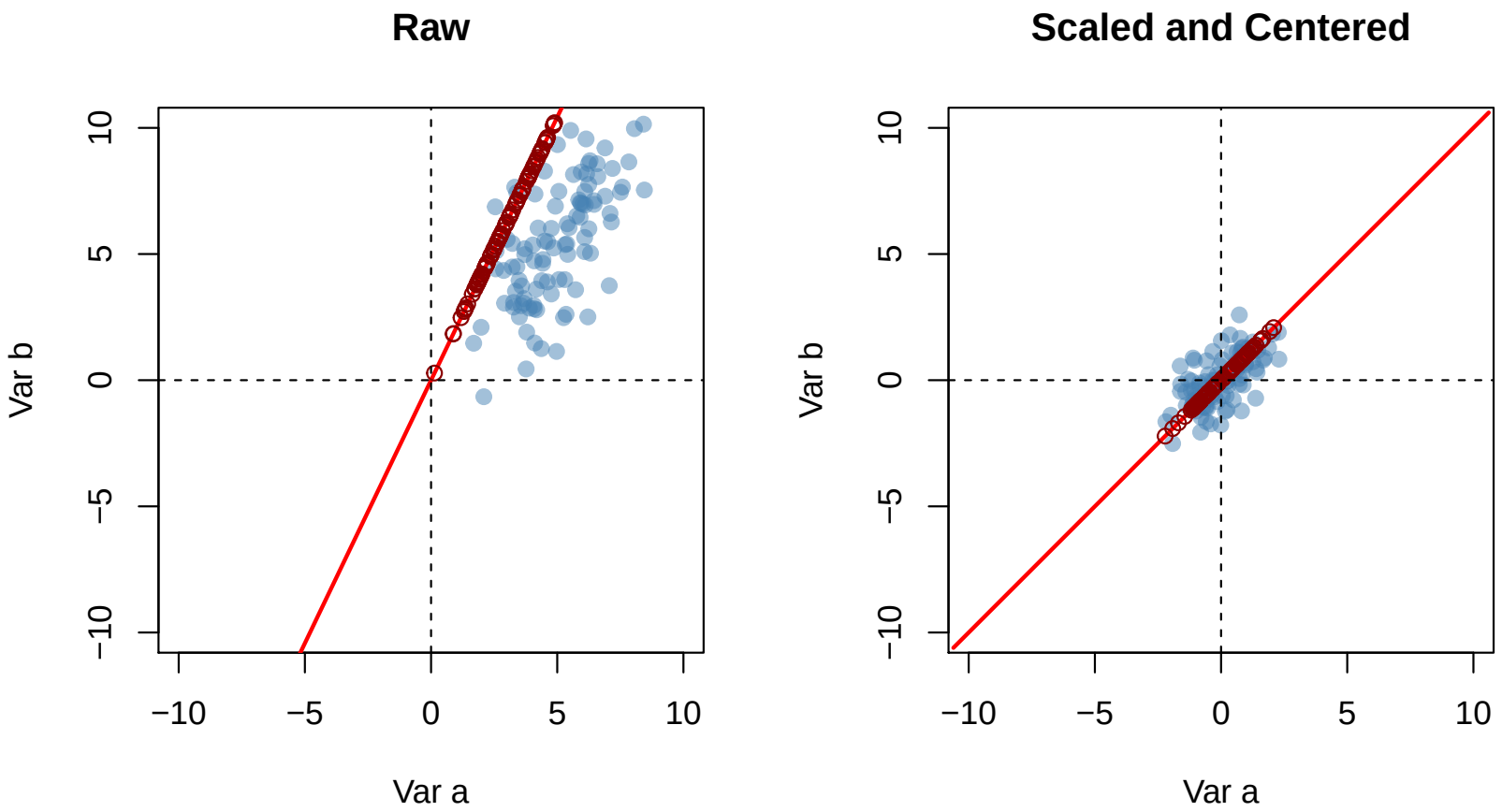
The loadings represent the weight of the initial variables along PC1

- `Var a`: 0.2336744
- `Var b`: 0.9723149

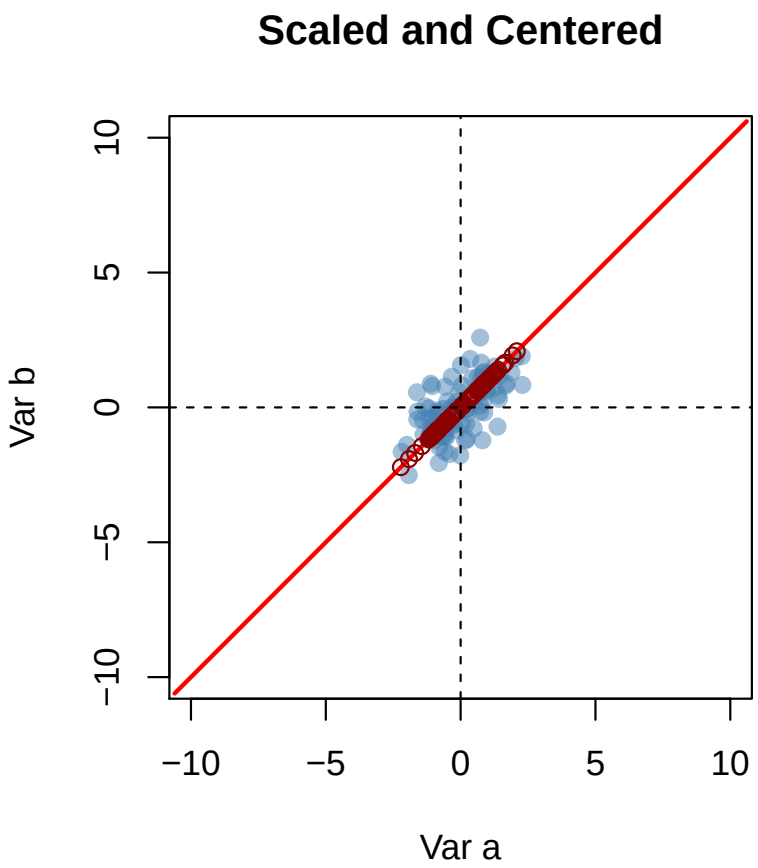
PCA uses

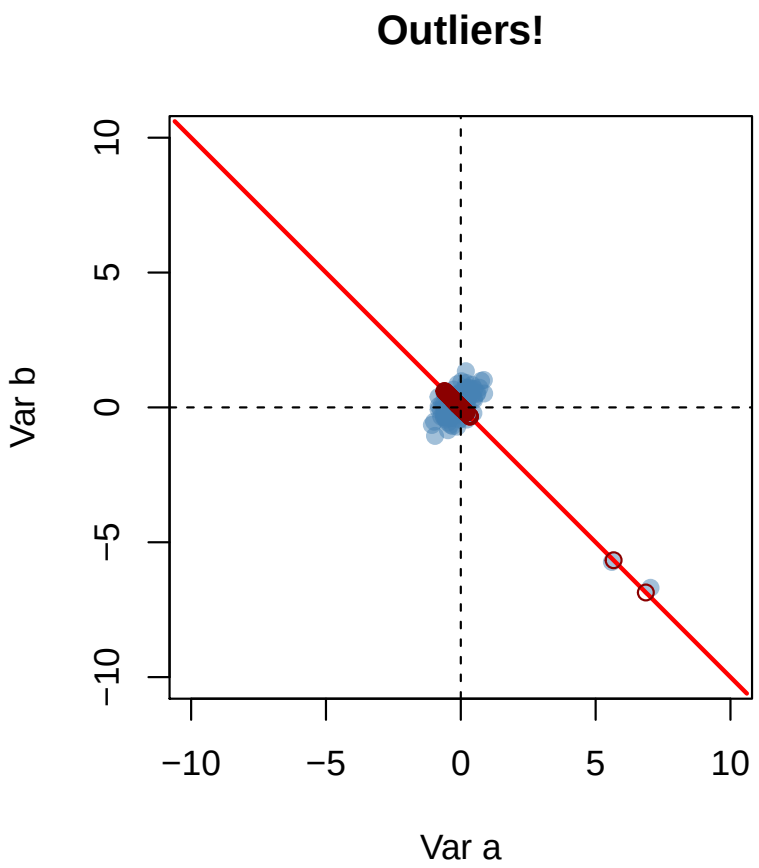
- Visualization of multivariate data by scatter plots
- Transformation of highly correlating x-variables into a smaller set of uncorrelated latent variables that can be used by other methods
- Separation of relevant information (described by a few latent variables) from noise
- Combination of several variables that characterize a chemical-technological-biological process into a single or a few "characteristic" variables
- Make the "latent properties" actually measurable

PCA is sensitive to scaling and centering



PCA is sensitive to outliers





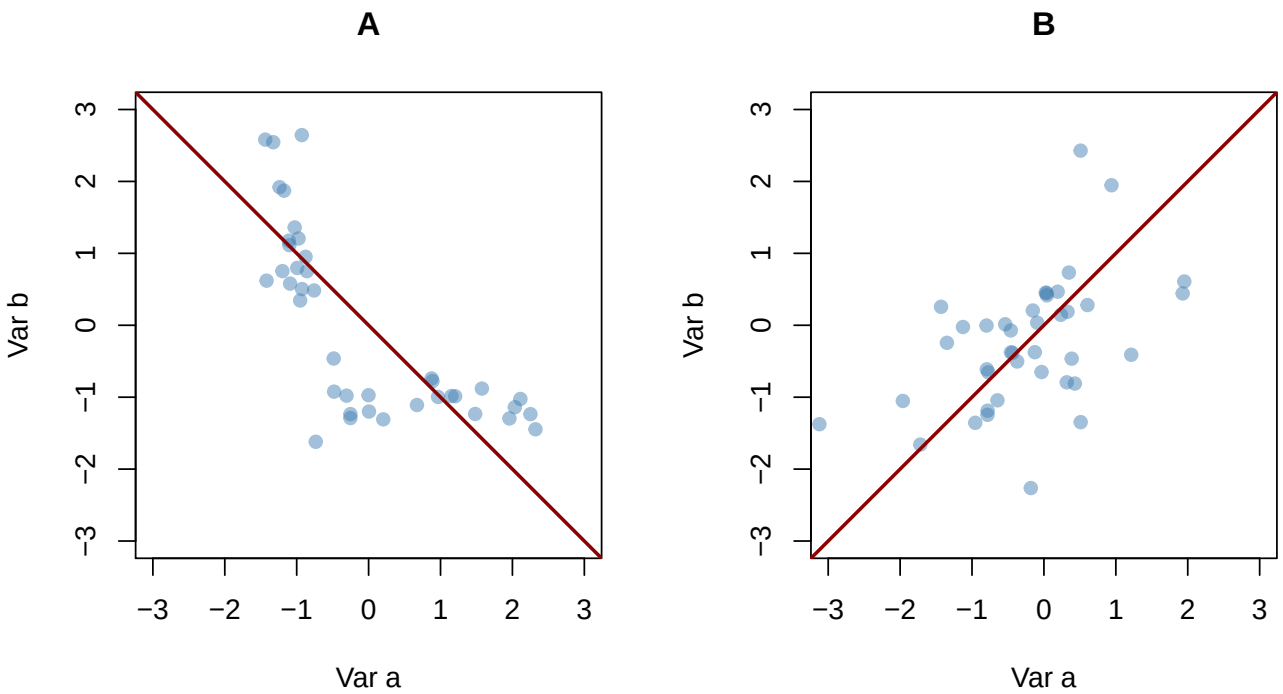
Notes

- sensitivity to outliers is useful if PCA is used to spot them ;-)
- *robust* version of PCA are available to keep all data in
- PCA show the big "structure" of my data and this can help in interpretation
- PCA will change if you add points !!!
- The loadings are not always easy to interpret

PCA as a data model

Data Model

PCA projects multidimensional data on a low dimensional subspace spanned by the LVs. This projection "models" the data. We could ask ourselves how good is this simplified representation.



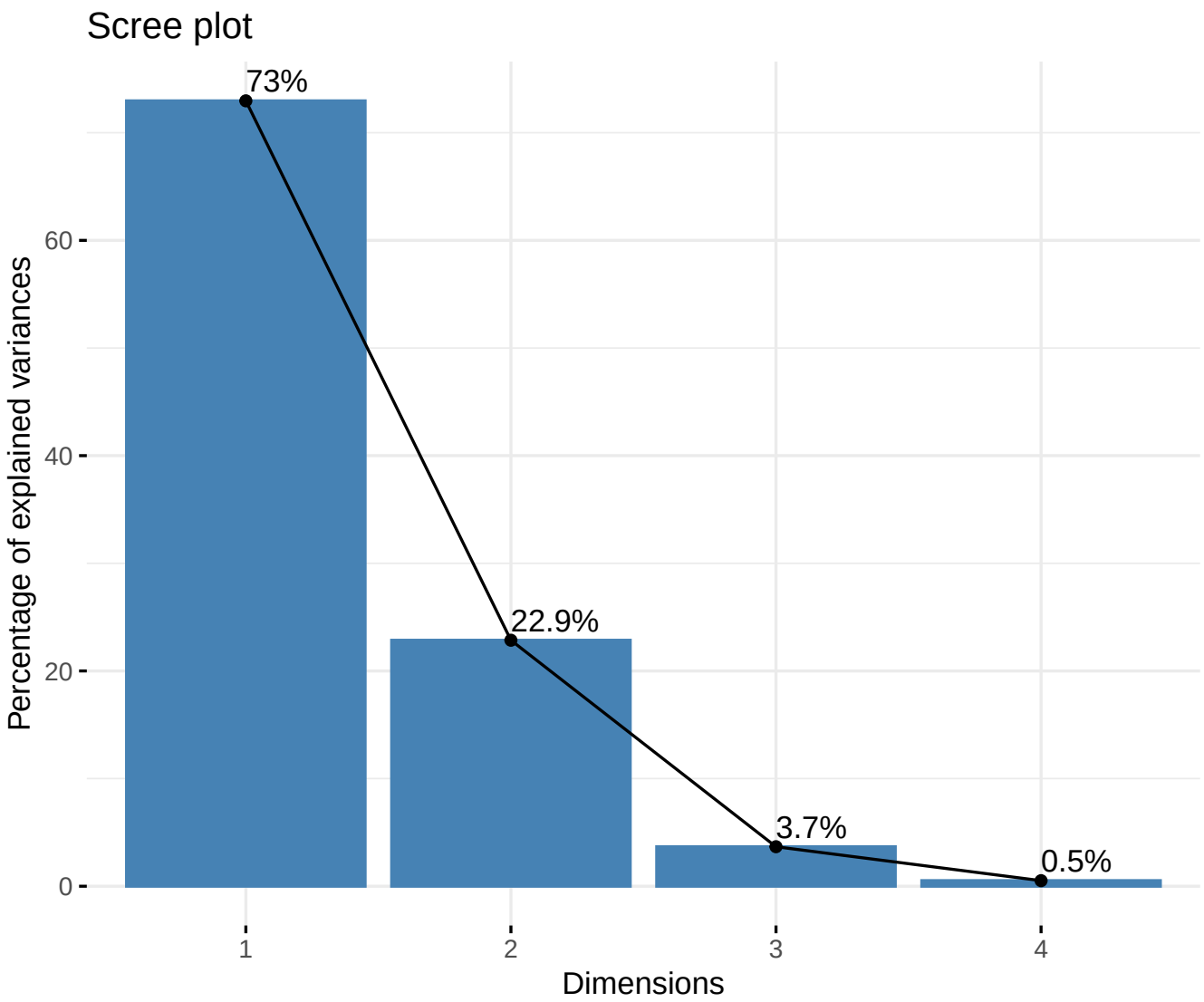
Number of Components

Model and components

- a PCA with more components will represent better the data (is a better model)
- a PCA with less components will be a more concise summary (is a model more parsimonious)
- we have to strike a balance between complexity and goodness of fit

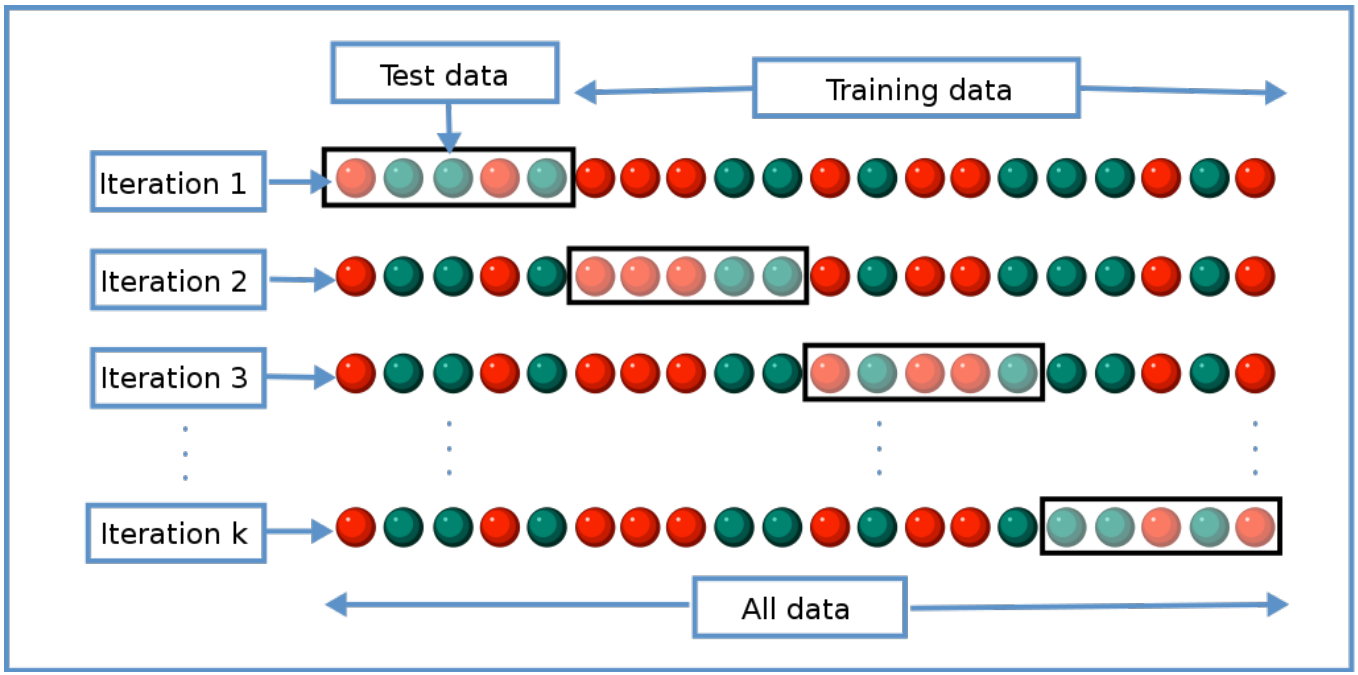
- Graphical methods (e.g. scree plot): we want to cover a big part of the variance
- Cross Validation

Scree plot


Scree plot
73%
22.9%
3.7%
0.5%
60
40
20
0
Percentage of explained variances
1
2
3
4
Dimensions

Cross Validation

- A good data model should be able to represent well new data
- I can simulate this by constructing the model on a subset of my samples and test its performance on the "leftover" samples
- To assess the variability in the performance I do that several times




LIVE
I am
here
now
being
becoming
sensing