2. Principal Component Analysis

In the last lecture

- Visualizing multivariate data
- Geometric interpretation of PCA
- Mathematical interpretation
- Example(s)

What is a latent variable?

- All variables are not independent.
  - They are redundant images of few ***"latent"*** variables
  - Example: your health.

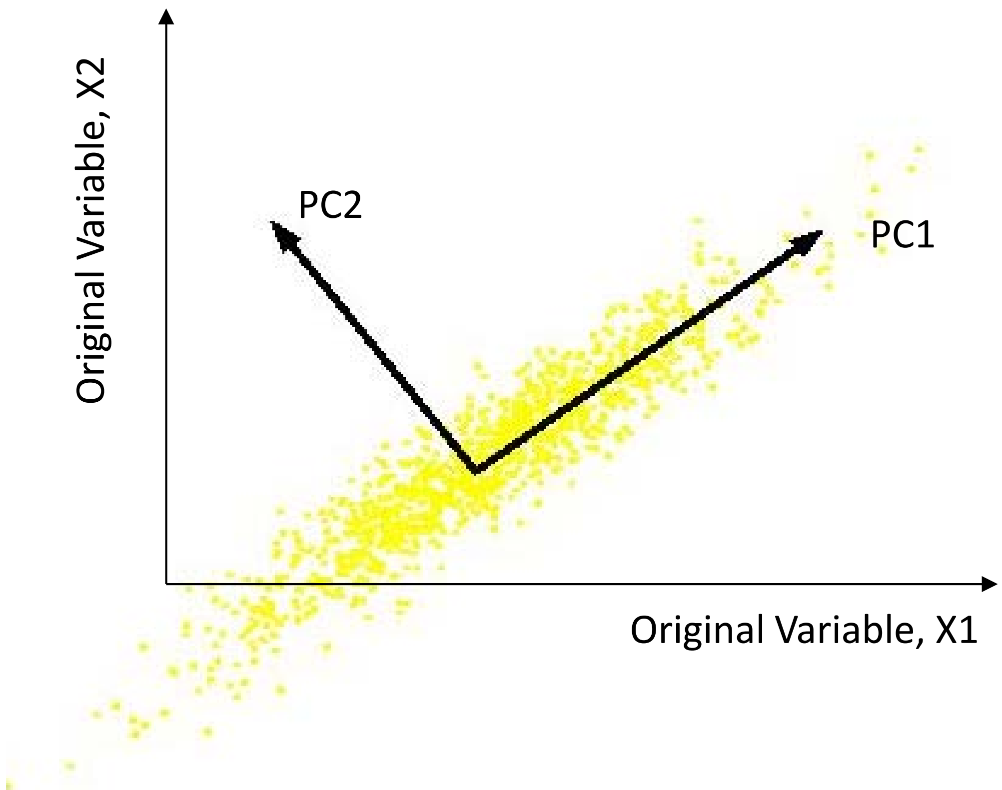
Your health

- No single measurement of "health"
  - blood pressure
  - cholesterol
  - weight
  - waist, hip (waist:hip ratio)
  - blood sugar
  - temperature, *etc*
- Combine these in some way? Trained doctor does this mentally.

Health is a latent (hidden) variable

Geometric Interpretation

- In summary,
  - PCA finds a **few** orthogonal axes of greatest variance in data. (K>>A)



Geometric Interpretation

- New latent variables are linear combinations of the original variables.

$$PC1 = a_1\ X1 + a_2\ X2 + a_3\ X3$$

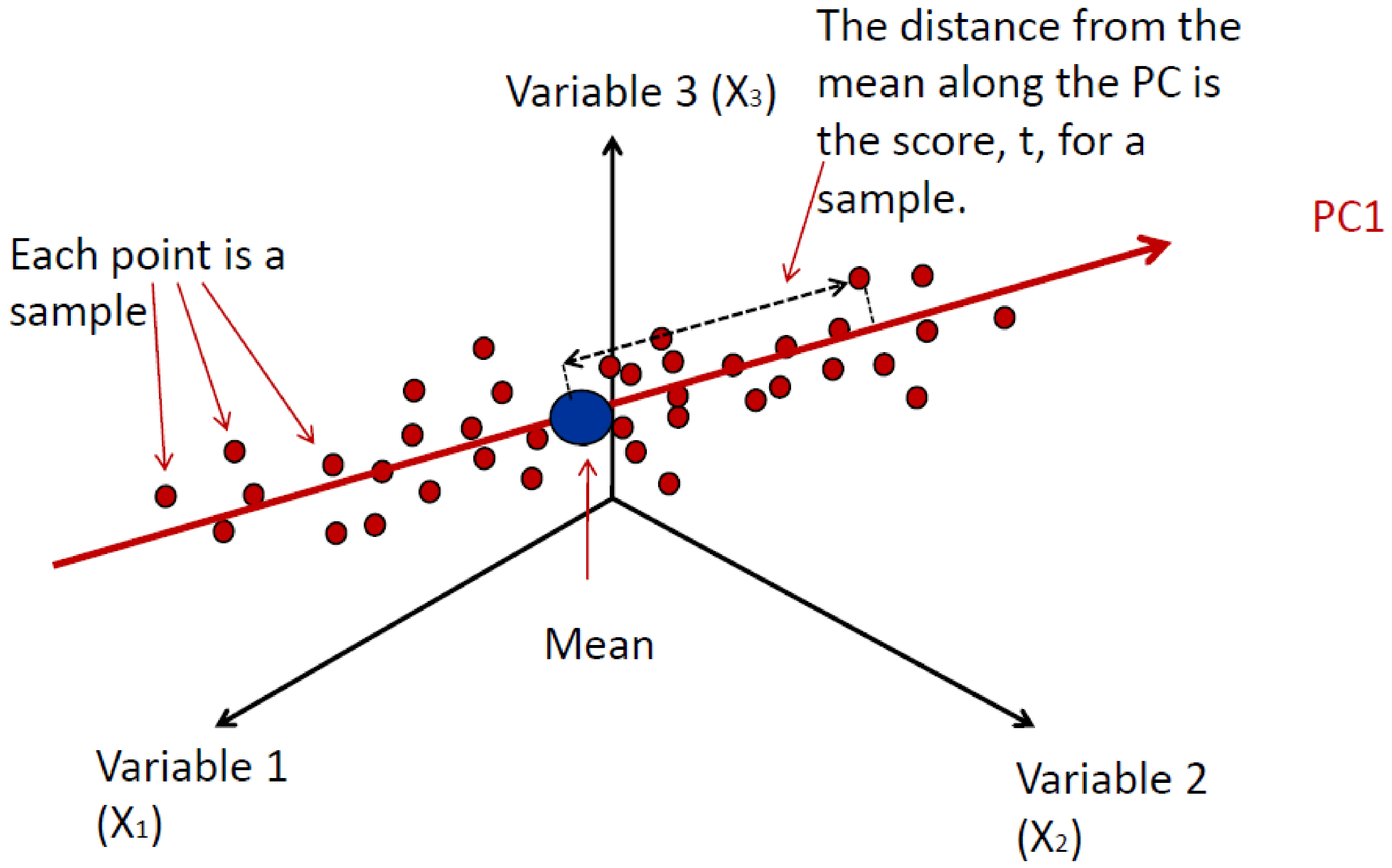
$$X = Mean + b_1\ PC1 + b_2\ PC2 + Error$$

Constraints :

- Maximise the dispersion of samples along the latent variables (the variance)
- Orthogonality

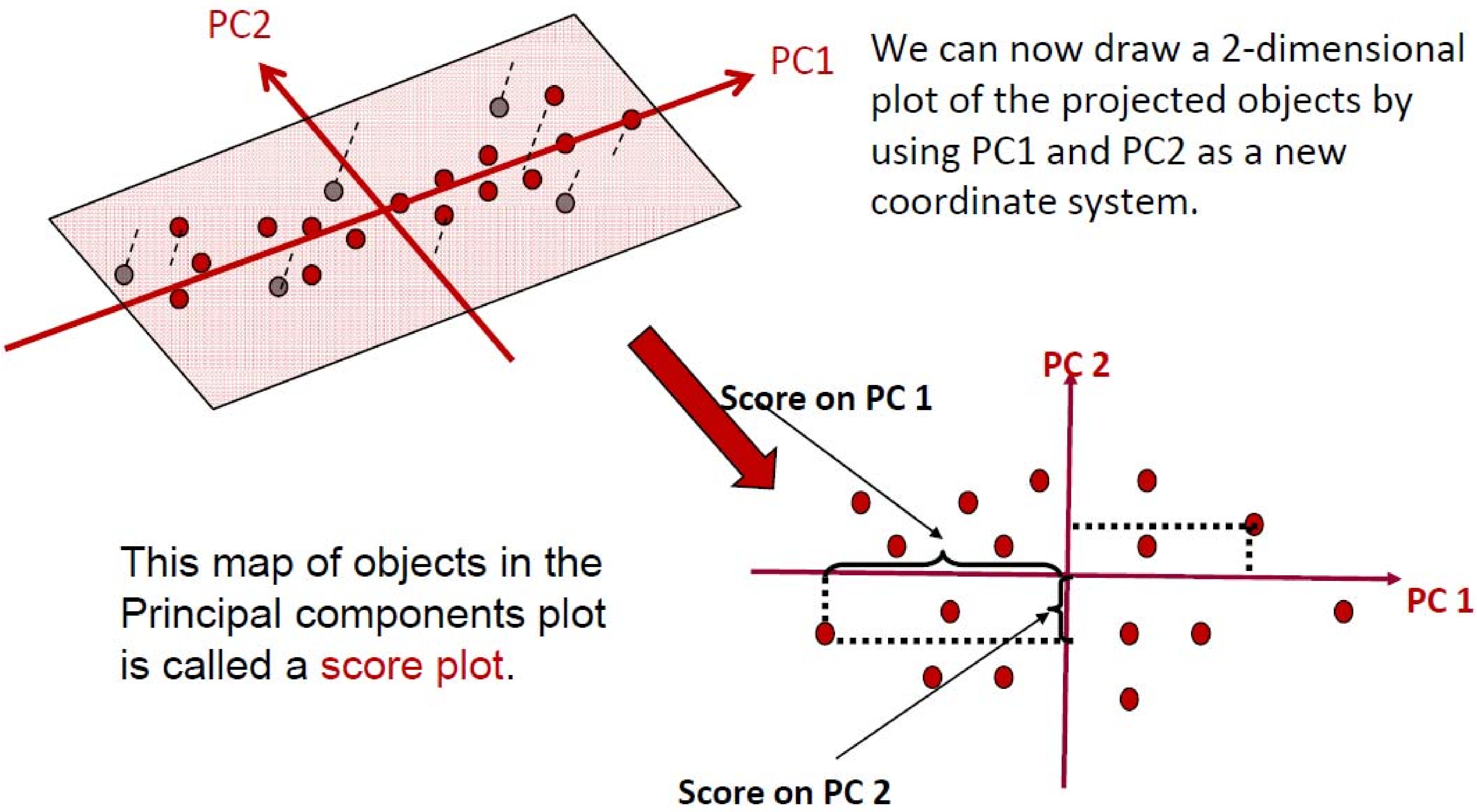
Review of PCA

- What is score?



Review of PCA

- Score plot – low dimensional summary of samples



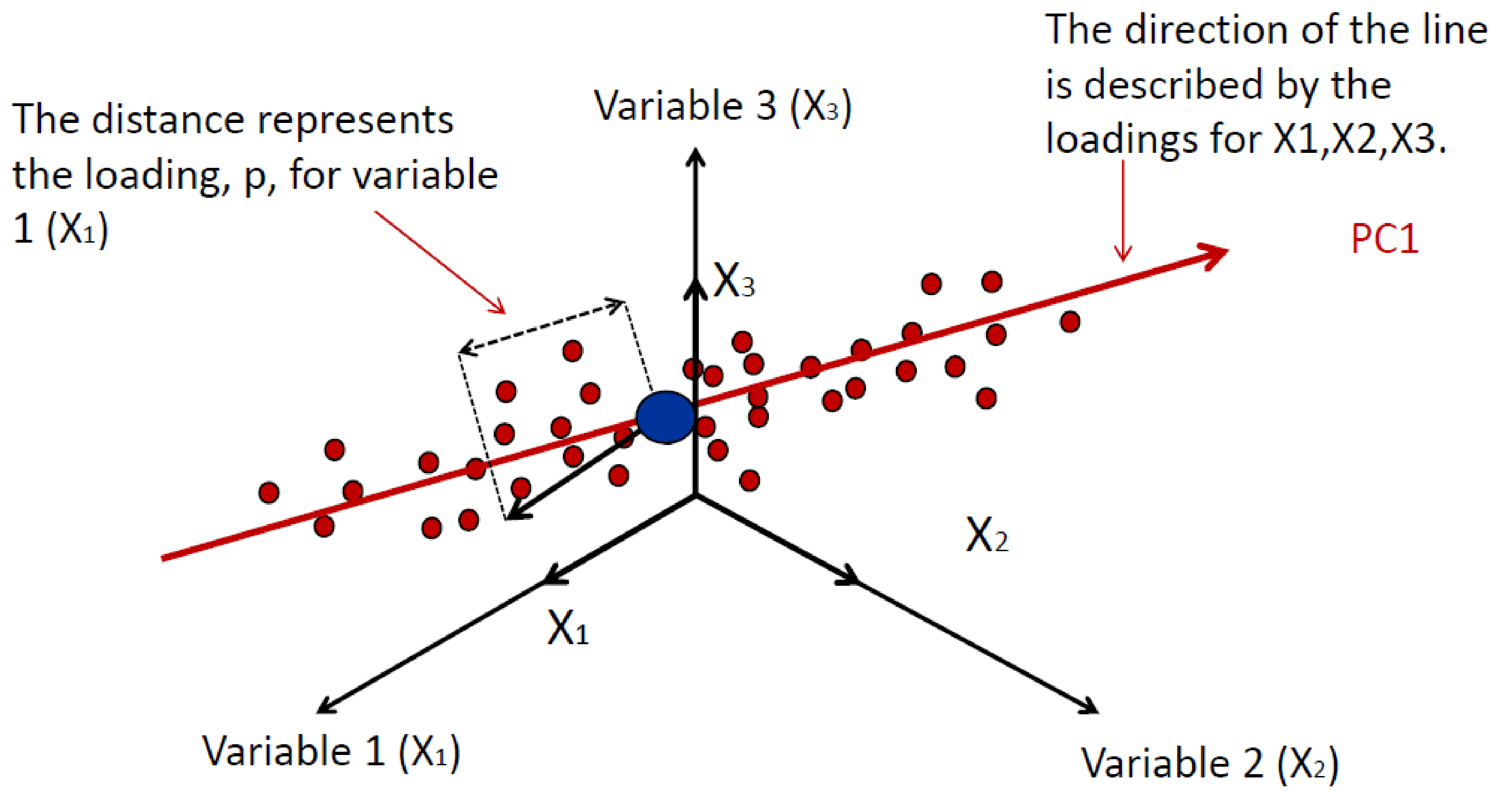
We can now draw a 2-dimensional plot of the projected objects by using PC1 and PC2 as a new coordinate system.

This map of objects in the Principal components plot is called a score plot.

Review of PCA

- What is loading?

Coefficients in the linear combination $PC1 = a_1\ X1 + a_2\ X2 + a_3\ X3$



In this lecture

- Tutorials & a bit more on PCA
- NIPALS algorithm
- Assignment #1

2. Principal Component Analysis

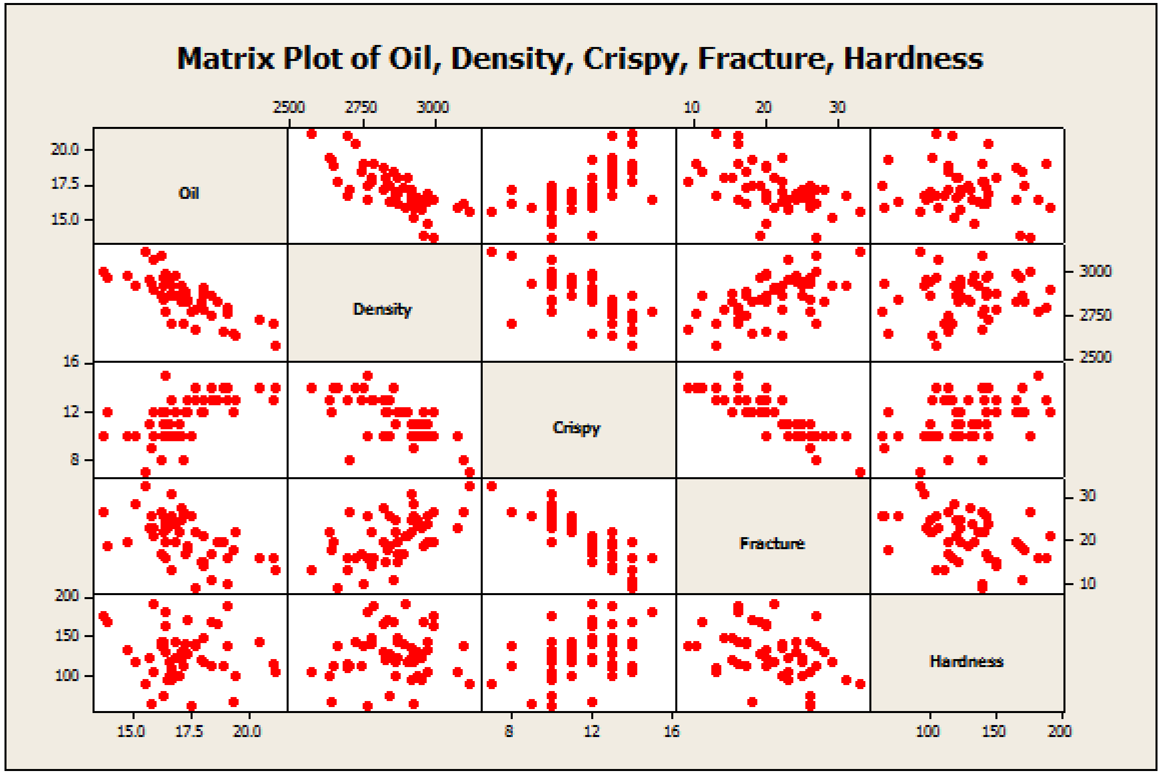
- Tutorial 1: Food texture example (food-texture.csv)

5 quality attributes are measured from pastries:

1. Percentage oil
2. Density
3. Crispiness measurement: from 7 (soft) to 15 (crispy)
4. Fracture angle
5. Hardness: force required before it breaks

Tutorial 1

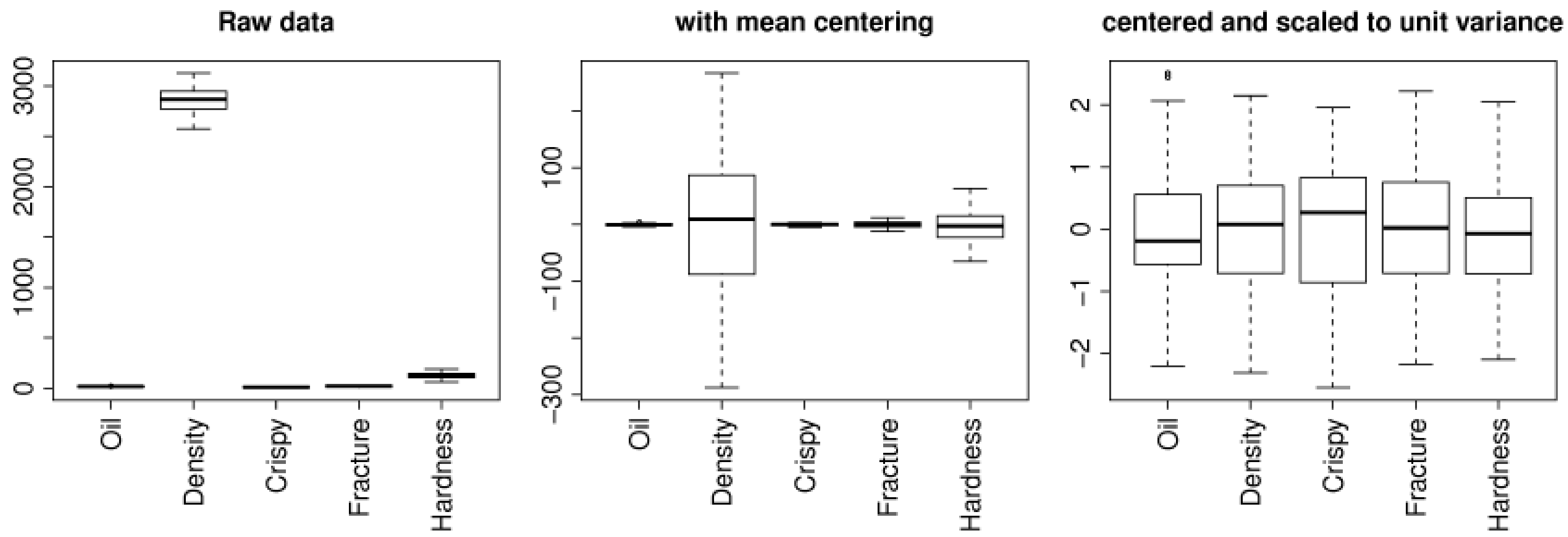
- Let’s see in a univariate fashion



- This data set has only five variables.

Tutorial 1: preprocessing

- Mean-centering & unit variance scaling (why?)

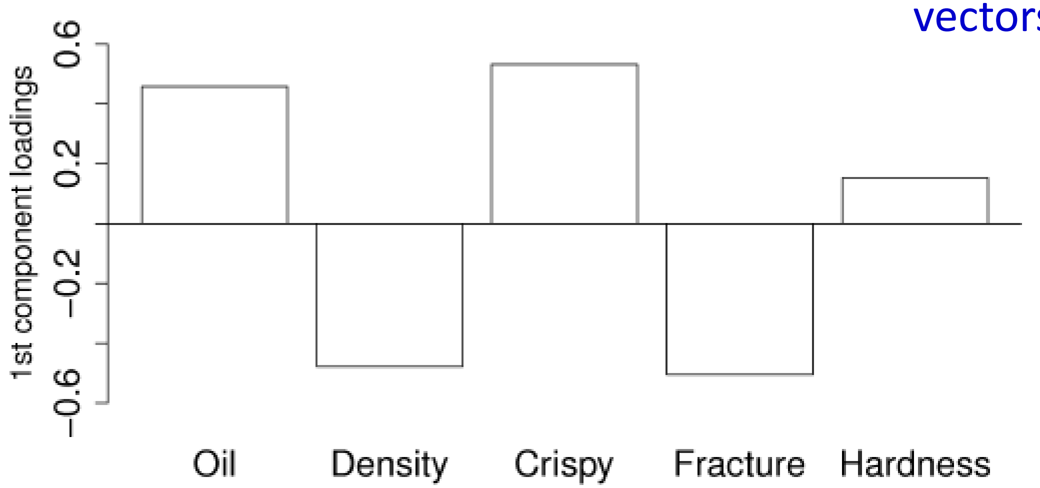


- Centering: $\mathbf{x}_{k,\text{center}} = \mathbf{x}_{k,\text{raw}} - \text{mean}\left(\mathbf{x}_{k,\text{raw}}\right)$
- Scaling: $\mathbf{x}_k = \dfrac{\mathbf{x}_{k,\text{center}}}{\text{standard deviation}\left(\mathbf{x}_{k,\text{center}}\right)}$
- Does not change relationships between variables.

Tutorial 1: loadings & scores

Loadings = direction vector

※We will se how to calculate the vectors later.



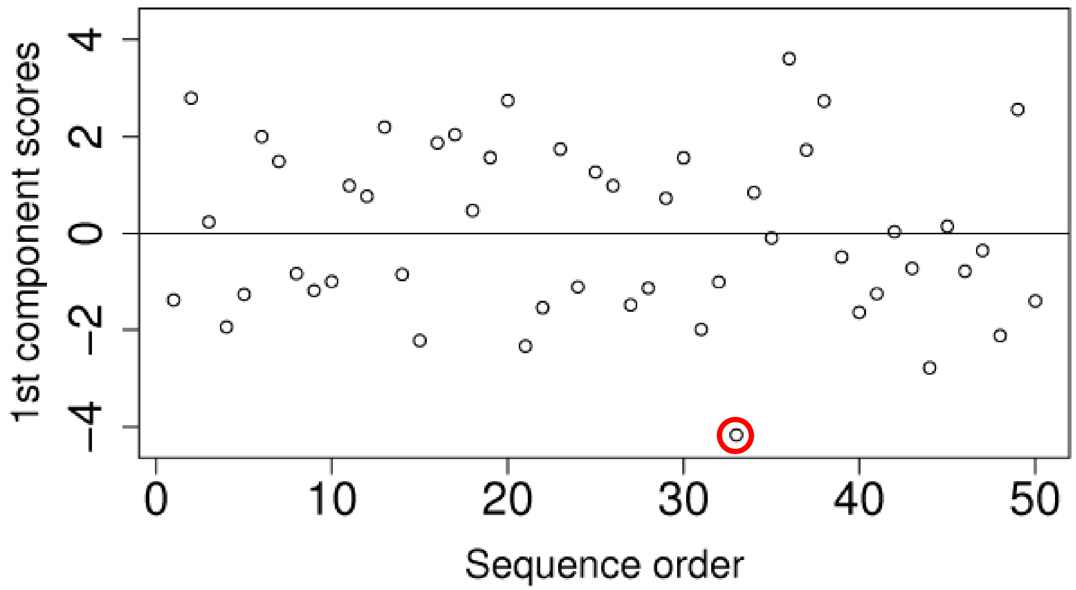
$$\mathbf{p}_1^T = [0.46 \quad -0.47 \quad 0.53 \quad -0.50 \quad 0.15]$$

$$t_{1,i} = 0.46x_{\text{oil}} - 0.47x_{\text{density}} + 0.53x_{\text{crispy}} - 0.50x_{\text{fract}} + 0.15x_{\text{hard}}$$

- $x_{\text{oil}} = \dfrac{x_{\text{oil, raw}} - \text{mean}(x_{\text{oil, raw}})}{\text{standard deviation}(x_{\text{oil,raw}})}$
- same for the other variables

Tutorial 1: loadings & scores

- Sample 33:
  - Oil = 15.5%
  - Density = 3125
  - Crisp = 7
  - Fracture = 33
  - Hardness = 92
- Mark these points on the scatterplot matrix



Tutorial 1: loadings & scores

- Sample 33: [Oil=15.5, Density=3125, Crispy=7, Fract=33, Hard=92]
- Sample 33: $t_1 = -4.2$
- $t_1 = 0.46x_{\text{oil}} - 0.47x_{\text{density}} + 0.53x_{\text{crispy}} - 0.50x_{\text{fract}} + 0.15x_{\text{hard}}$
  - $x_{\text{oil}} = (15.5 - 17.2)/1.59 = -1.07$
  - $x_{\text{density}} = (3125 - 2857)/124.5 = 2.15$
  - $x_{\text{crisp}} = (7 - 11.52)/1.78 = -2.53$
  - $x_{\text{fracture}} = (33 - 20.9)/5.47 = 2.2$
  - $x_{\text{hardness}} = (92 - 128)/31.1 = -1.16$

$t_1 = 0.46(-1.07) - 0.47(2.15) + 0.53(-2.53) - 0.50(2.2) + 0.15(-1.16) = -4.2$ $t_1 = -0.50 - 1.01 - 1.35 - 1.1 - 0.17 = -4.2$
Each measurement *contributes* to the t_1 value.

Tutorial 1: loadings & scores

- Examine sample 36: $t_1 = 3.6$

  Sample 36: 21.1% (Oil), 2570 (Density), 14 (Crispy), 13 (Fracture), 105 (Hardness)

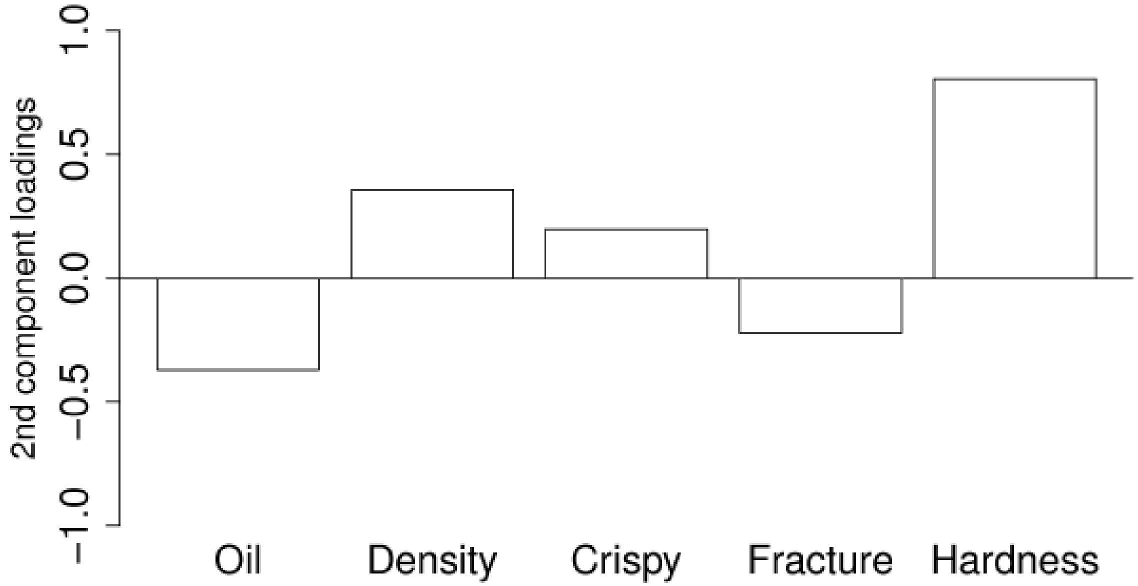
  - Characteristics of a high t_1 sample?

Pastries with high t_1 values:

This is only a **correlation** - we can only guess what the true **cause** is.
First component: explains 61% of the variability in the data.

Tutorial 1: loadings & scores

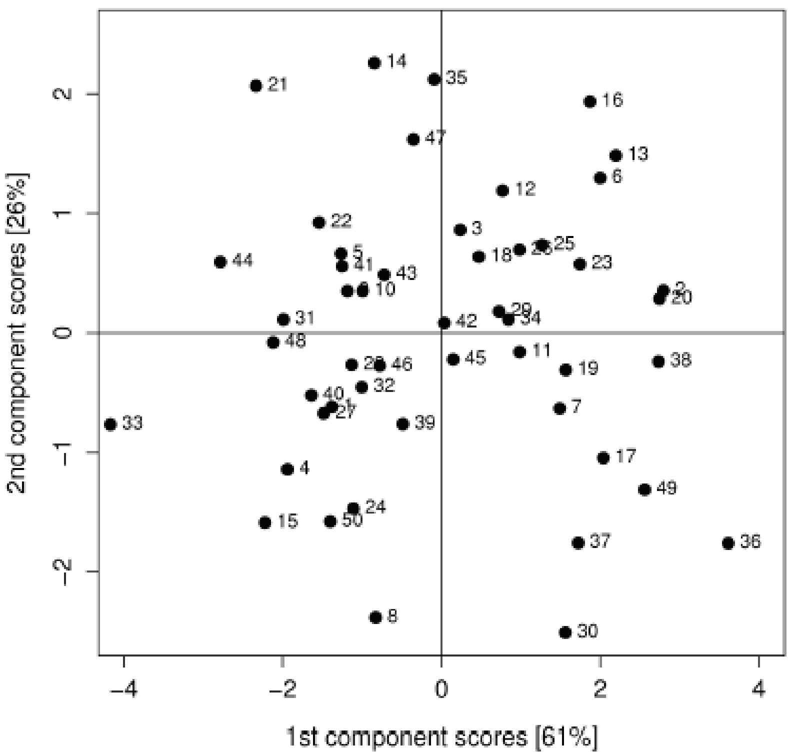
The second loading vector:

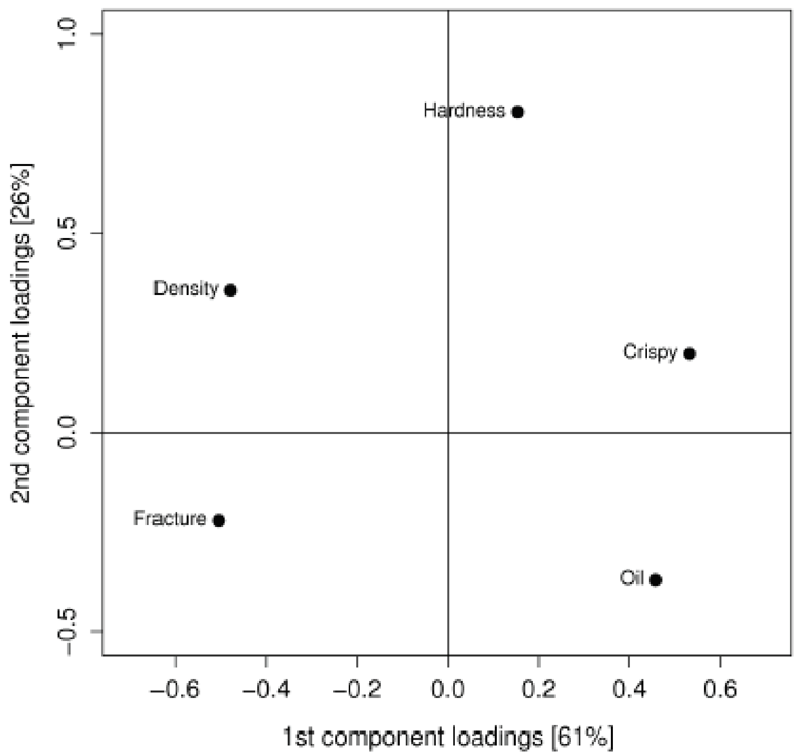


- Interpretation?
- Explains 26% of additional variability
- Is orthogonal (independent) to p_1. This means...
  - can adjust process conditions for hardness without affecting other pastry properties

Tutorial 1: loadings & scores

- In 2-D plot





Interpretation of scores & loadings

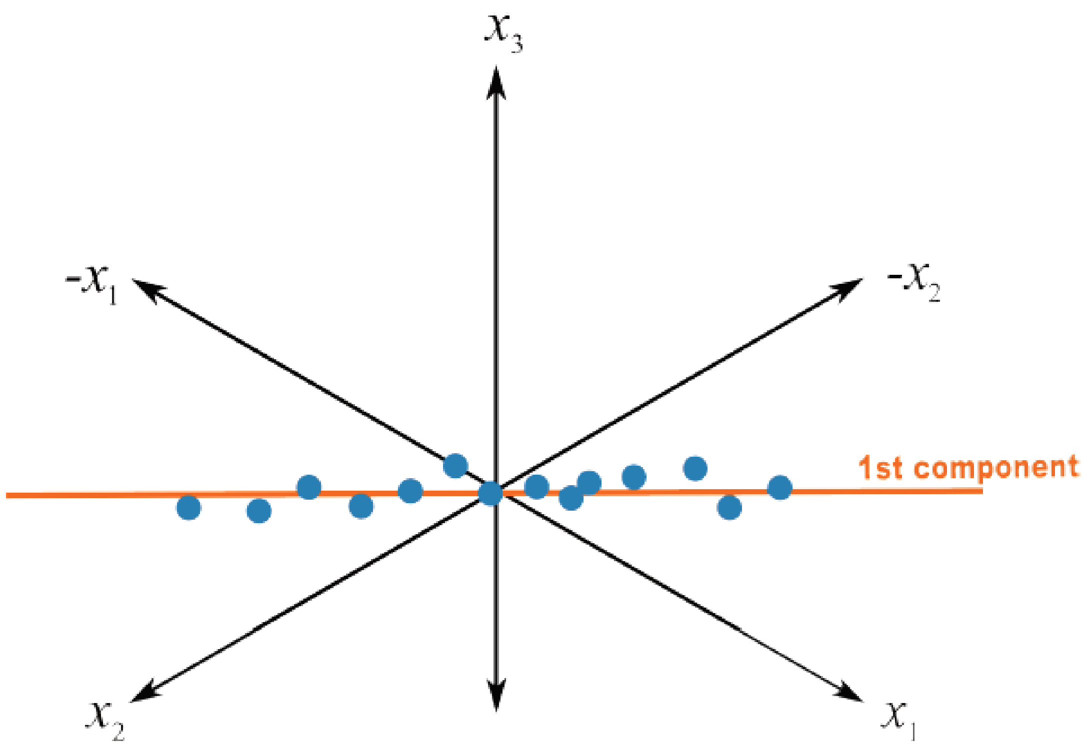
- Interpretation

  Key equation:

  $t_{i,a} = x_{i,1}p_{1,a} + x_{i,2}p_{2,a} + \ldots + x_{i,k}p_{k,a} + \ldots + x_{i,K}p_{K,a}$

  - Time-series plots of the scores
    - patterns in the data
  - Scatter plots: t_i vs t_j
    - clustering
    - outliers

Interpretation of scores & loadings



- Two variables important: $p_1 = [+1, -1, 0]$
- Or as a unit vector: $p_1 = [+0.707, -0.707, 0]$
- Unimportant variables: close to zero
- Important variables for a component:

Outliers

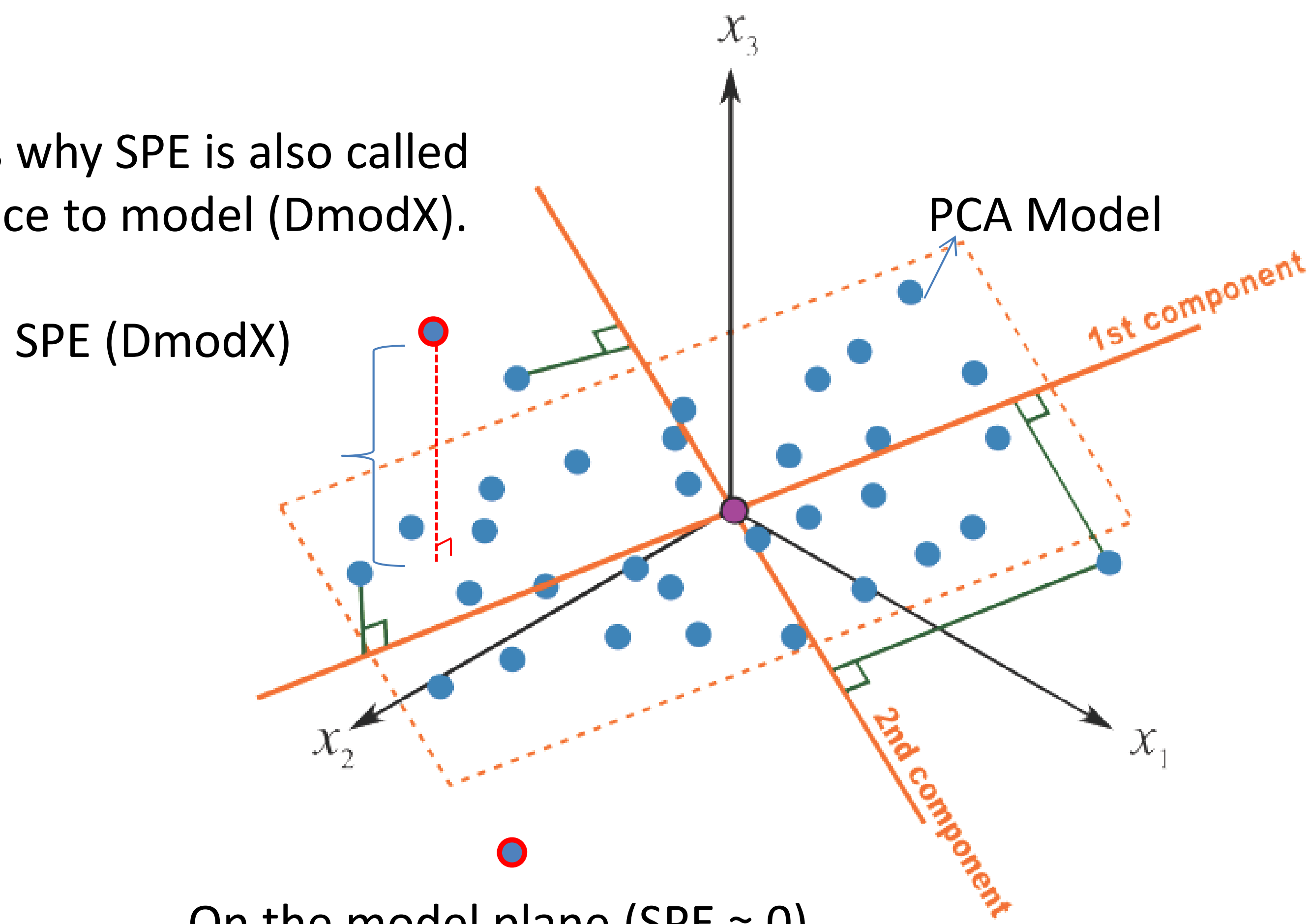
- Outliers
  - Observations poorly explained by the model
    - something new (or unusual) in this observation
  - Detected by using SPE or Hotelling's T^2.
  - SPE or Hotelling's T^2 : Complementary to each other.

- SPE (from last lecture)

$$
\begin{aligned}
\text{SPE}_i &= \sqrt{\mathbf{e}_{i,A}^T \mathbf{e}_{i,A}} \\
(1 \times 1) &= (1 \times K)(K \times 1) \\
\mathbf{e}_{i,A}^T &= \mathbf{x}_i^T - \hat{\mathbf{x}}_{i,A}^T \\
(1 \times K) &= (1 \times K) - (1 \times K)
\end{aligned}
$$

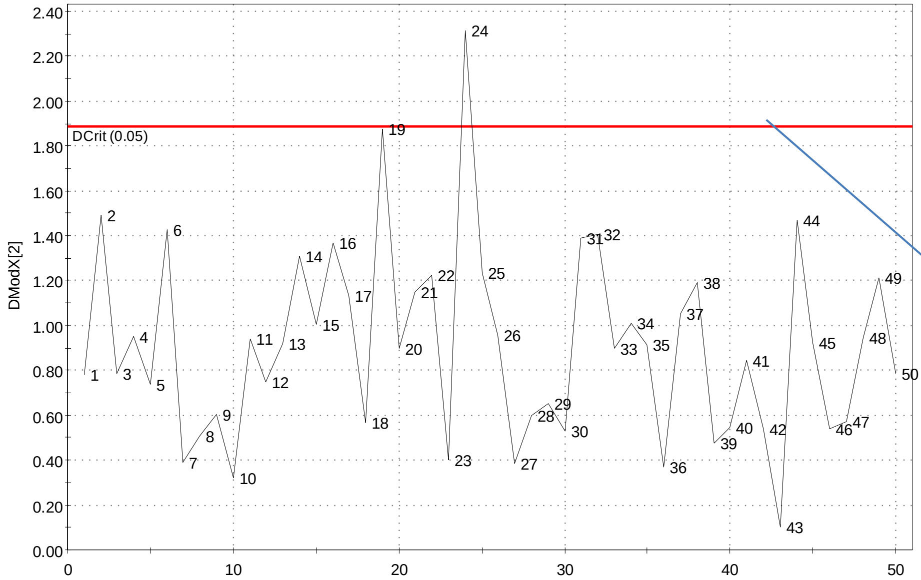
Outliers

This is why SPE is also called distance to model (DmodX).



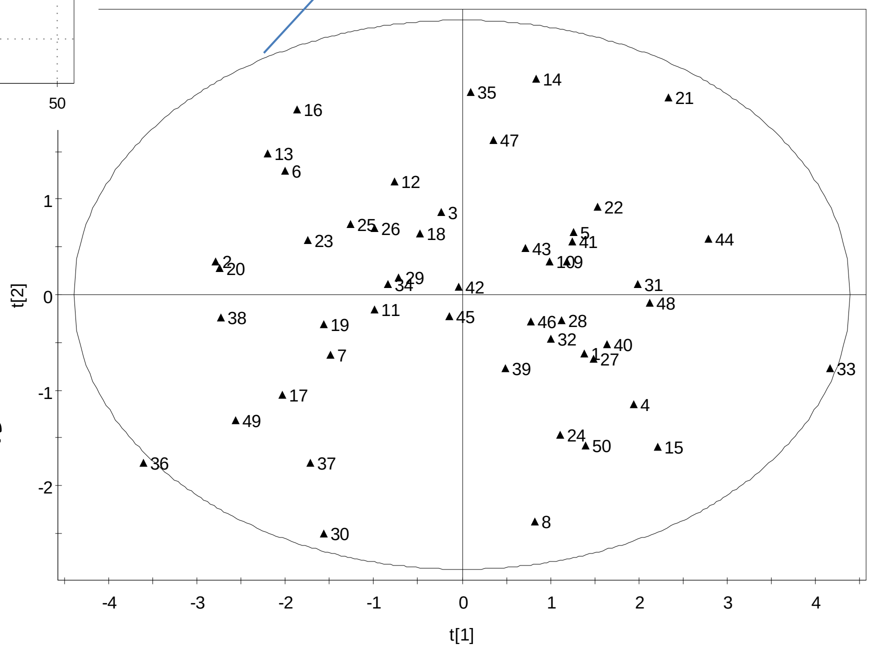
On the model plane (SPE ≈ 0)
outside of the usual range

Tutorial 1: outliers



Back to food texture example

95% C.I



Hotelling's T^2: next example

Contribution plot

- Tells why an observation differs from the others in
  - X score (**t**)
  - SPE(DModX)
  - DModY, or in the predicted Y. (PLS)
- For scores
  - Weights * ($\mathbf{x}_{outlier}$ - $\mathbf{x}_{average}$)
  - Weights: **p**(loading), variable R^2, …
- For SPE
  - Weights * $e_{outlier,k}$
  - Weights: **p**(loading), variable R^2, …

Tutorial 1: contribution plot