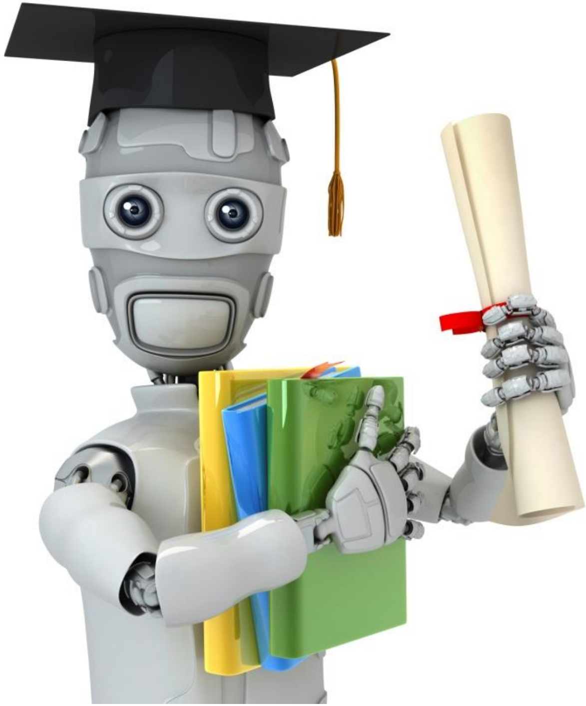
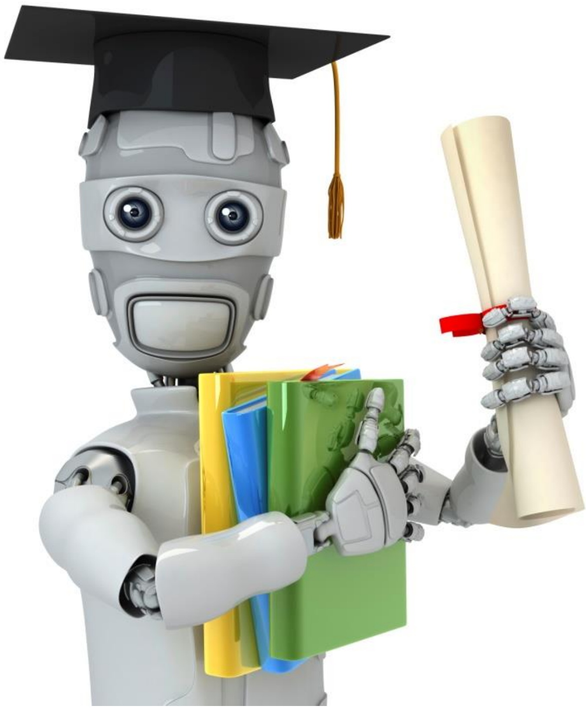


Machine Learning

Principal Component Analysis (PCA)

Fouad Hadj Selem



Introduction and first example

Machine Learning

PCA applies to data tables where rows are considered as individuals and columns as quantitative variables

Which kinds of data?

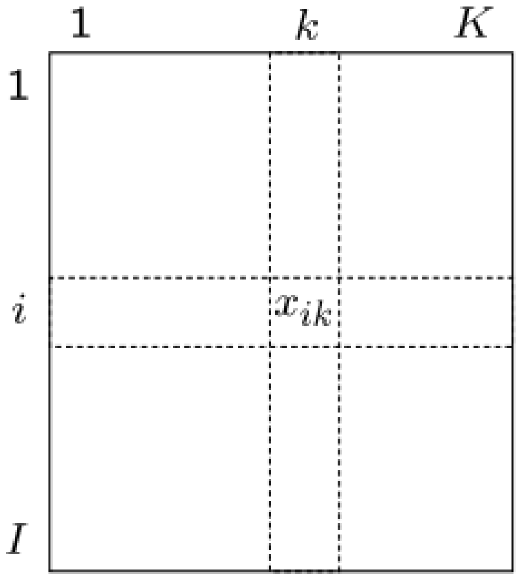


Figure: Data table in PCA

For variable k, we note:

the mean: $\bar{x}_k = \frac{1}{I}\sum_{i=1}^{I} x_{ik}$

the standard-deviation:

$$s_k = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(x_{ik} - \bar{x}_k)^2}$$

Examples :

- Sensory analysis: score for attribute k of product i
- Ecology: concentration of pollutant k in river i
- Economics: indicator value k for year i
- Genetics: expression of gene k for patient i
- Biology: measure k for animal i
- Marketing: value of measure k for brand i
- Sociology: time spent on activity k by individuals from social class i
- etc.

$\Rightarrow$ There exist many data tables like these

Issues and goals

The data table can be seen as a set of rows or a set of columns

Studying individuals

- When can we say that 2 individuals are similar (or dissimilar) with respect to all the variables?
- If there are many individuals, is it possible to categorize them?

$\Rightarrow$ groups of individuals, partitions between them

Issues and goals

Studying variables

- For individuals, we interpret similarity in terms of the variables' values
- Between variables, we talk instead of "relationships"
- Linear relationships are commonplace, and a first approximation of many links $\Rightarrow$ correlation coefficient

$\Rightarrow$ visualization of the correlation matrix
$\Rightarrow$ find a small number of synthetic variables to summarize many variables (e.g. of a prior synthetic variable: the mean. But here we search for posterior synthetic variables from the data)

Issues and goals

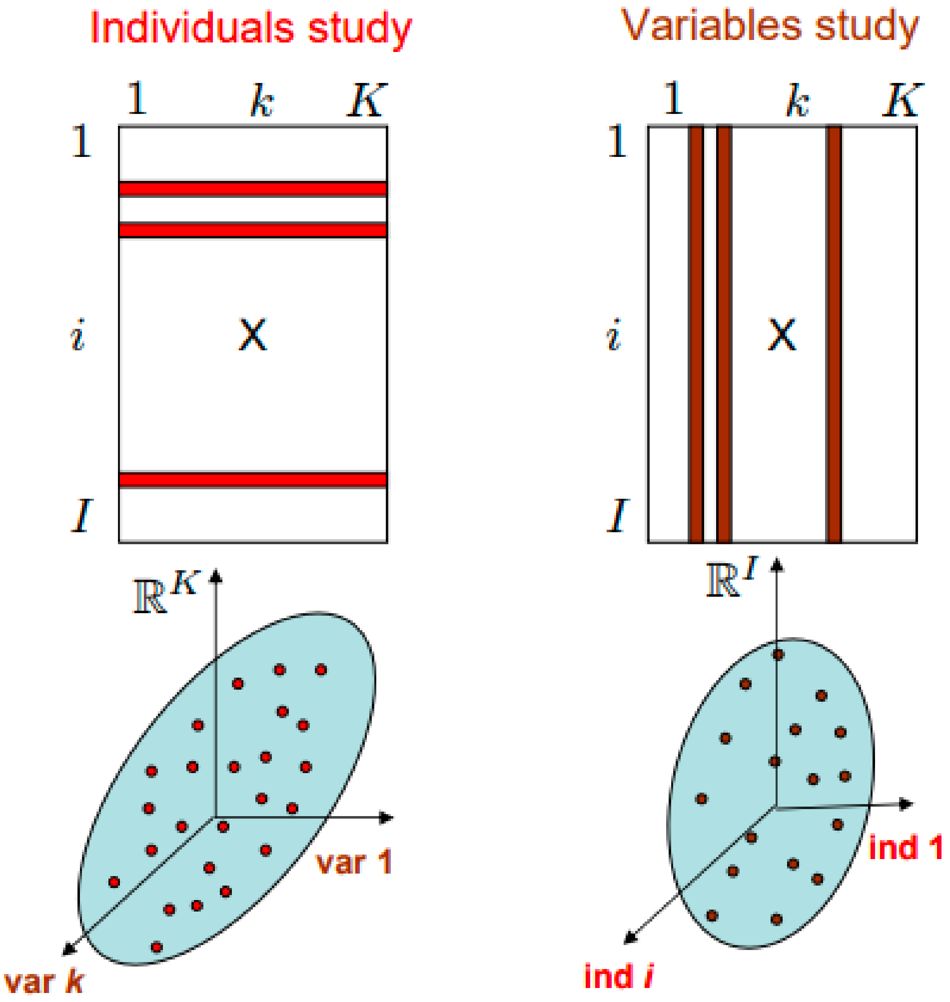
Links between the two points-of-view

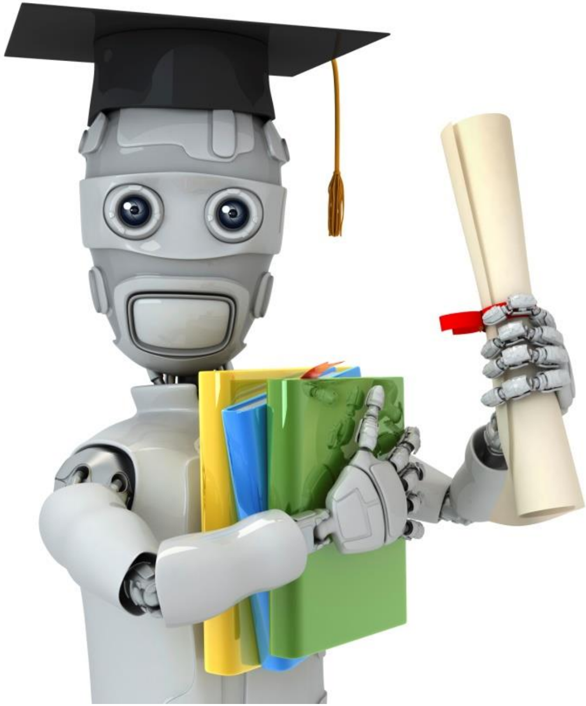
- Characterize groups of individuals using the variables
  $\Rightarrow$ need an automatic procedure
- Use specific individuals to better understand links between variables
  $\Rightarrow$ use of extreme individuals (return to individuals to understand more simply)

PCA issues:

- Descriptive method to explore data: visualization of data with simple plots
- Data compression - summarize a big data table of *individuals* $\times$ *quantitative variables*

Two-point clouds





Machine Learning

Individuals study:

The cloud of individuals NI

1 individual = 1 row of the data table $\Rightarrow$ 1 point in $\mathbb{R}^k$

- If $K = 1$: axial representation
- If $K = 2$: scatter plot
- If $K = 3$: 3D graphical representation (more difficult)
- If $K = 4$: impossible to "see" BUT the concept is easy

Notion of similarity: (squared) distance between individuals i and i':

$$d^2(i, i') = \sum_{k=1}^{K} (x_{ik} - x_{i'k})^2 \qquad \text{(thanks Mr Pythagoras)}$$

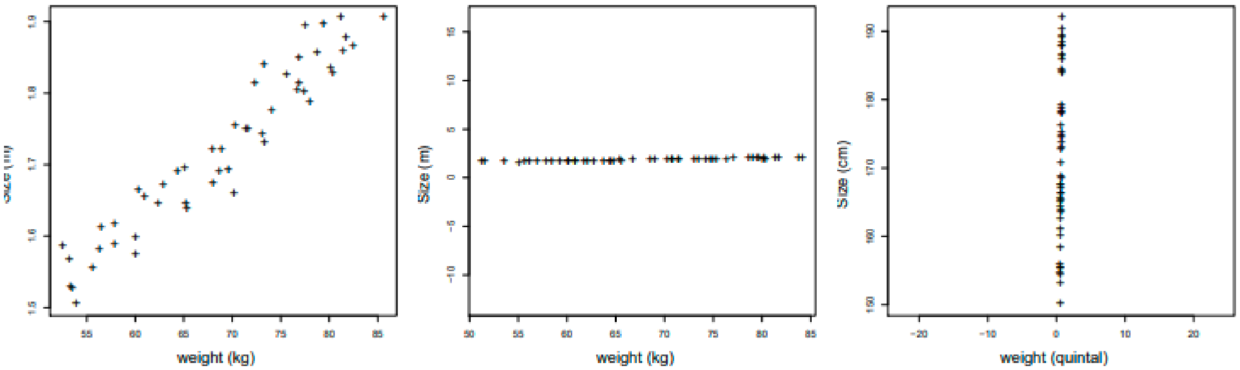
Studying the individuals $\equiv$ Studying the shape of the cloud N_I

The cloud of individuals NI

- Study the structure, *i.e.*, the shape, of the cloud of individuals
- Individuals are in $\mathbb{R}^K$

Centering – standardizing data

- Centering does not modify the shape of the cloud
  $\Rightarrow$ centering is always done



- Standardizing data is necessary if units are different between variables

$$x_{ik} \hookrightarrow \frac{x_{ik} - \bar{x}_k}{s_k}$$

Centering – standardizing data

	O.fruity	O.passion	O.citrus	...	Sweetness	Acidity	Bitterness	Astringency	Aroma.intensity	Aroma.persistency	Visual.intensity
S Michaud	-0,17	0,45	1,50	...	-0,30	0,11	0,20	-1,79	0,95	1,07	0,06
S Renaudie	0,02	1,03	1,16	...	-0,46	1,39	-0,31	-0,65	0,99	0,82	-1,08
S Trotignon	0,79	1,73	1,16	...	-0,67	0,48	0,20	-0,60	-0,44	0,07	-1,34
S Buisse Domaine	-0,17	0,45	-0,07	...	-0,02	-0,25	-2,01	0,19	-2,24	-1,66	-0,55
S Buisse Cristal	1,30	1,03	-0,12	...	-0,39	1,20	1,39	0,34	-0,44	-1,66	-0,90
V Aub Silex	-0,60	-0,97	-0,27	...	2,93	-2,07	-1,33	-0,60	-0,84	-0,92	-0,64
V Aub Marigny	-2,44	-0,97	-1,94	...	-0,30	0,84	1,39	1,45	-0,18	0,98	0,76
V Font Domaine	0,79	-1,11	-0,85	...	-0,67	-0,12	0,03	-0,44	0,29	0,41	1,03
V Font Brûlés	0,79	-0,84	0,13	...	-0,02	-0,61	0,03	0,34	0,75	0,07	1,73
V Font Coteaux	-0,29	-0,82	-0,69	...	-0,11	-0,98	0,37	1,76	1,15	0,82	0,94

PCA $\equiv$ Studying the standardized data set

Difficult to visualize the cloud $N_I \Rightarrow$ try to get an approximate view of it

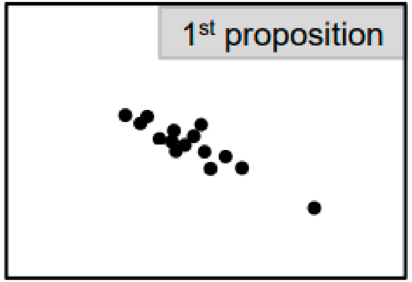
Fitting the cloud of individuals

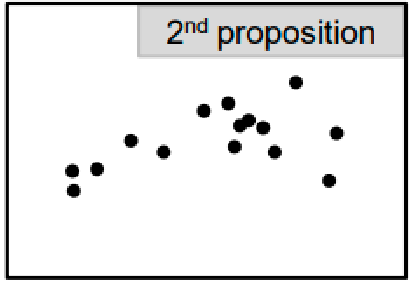
PCA searches for the best summary space for optimal visualization of N_I

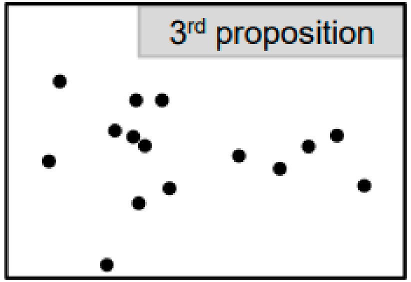
$\Longleftrightarrow$ Find a subspace that sums up the data the best

Viewpoint quality:

- faithfully reproduce the cloud's shape (*animation*)







Fitting the cloud of individuals

PCA searches for the best summary space for optimal visualization of N_I

$\Longleftrightarrow$ Find a subspace that sums up the data the best

Viewpoint quality:

- faithfully reproduce the cloud's shape (*animation*)
- best representation of diversity, variability
- doesn't distort distances between individuals

How to quantify the quality of a viewpoint?

notion of dispersion, of variability, also called **inertia**

inertia $\equiv$ variance generalized to several dimensions

How to quantify the quality of a viewpoint?

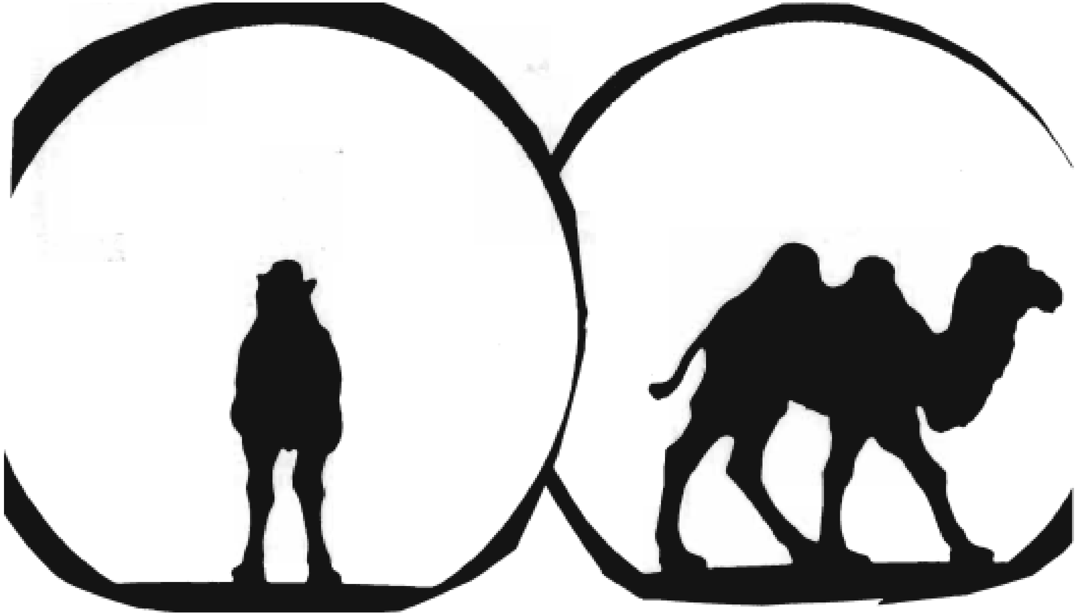
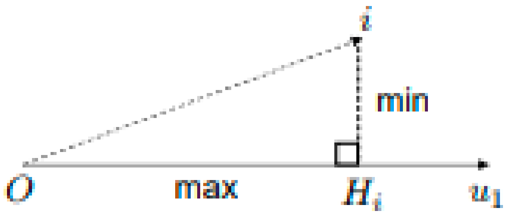


Figure: Camel or dromedary? (*illustration by J.P. Fénelon*)

Fit the individuals' cloud

How to find the best view to approximate the cloud?

1. find an axis that distorts the cloud the least



$(iH_i)^2$ small with $H_i \in$ axis $\Leftrightarrow$ $(OH_i)^2$ large (Pythagoras)
$\Rightarrow$ we want $\sum_i (OH_i)^2$ large

2. Find the best plane: maximize $\sum_i (OH_i)^2$ with $H_i \in$ plane
   The best plane contains the best axis: we search for $u_2 \perp u_1$ and maximizing $\sum_i (OH_i)^2$

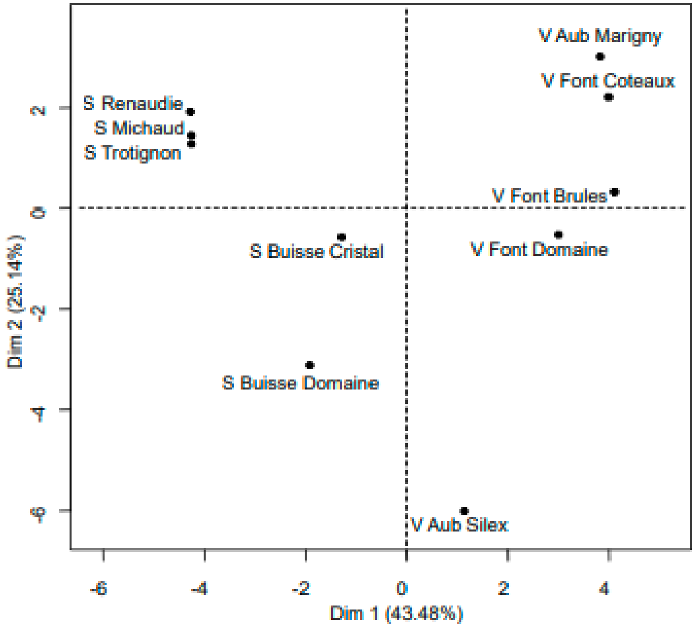
3. we can look for a third axis (etc.) with maximum inertia

- Sensory descriptors are used as active variables: only these variables are used to construct the axes
- Variables are (centered and) standardized

Example: wine data

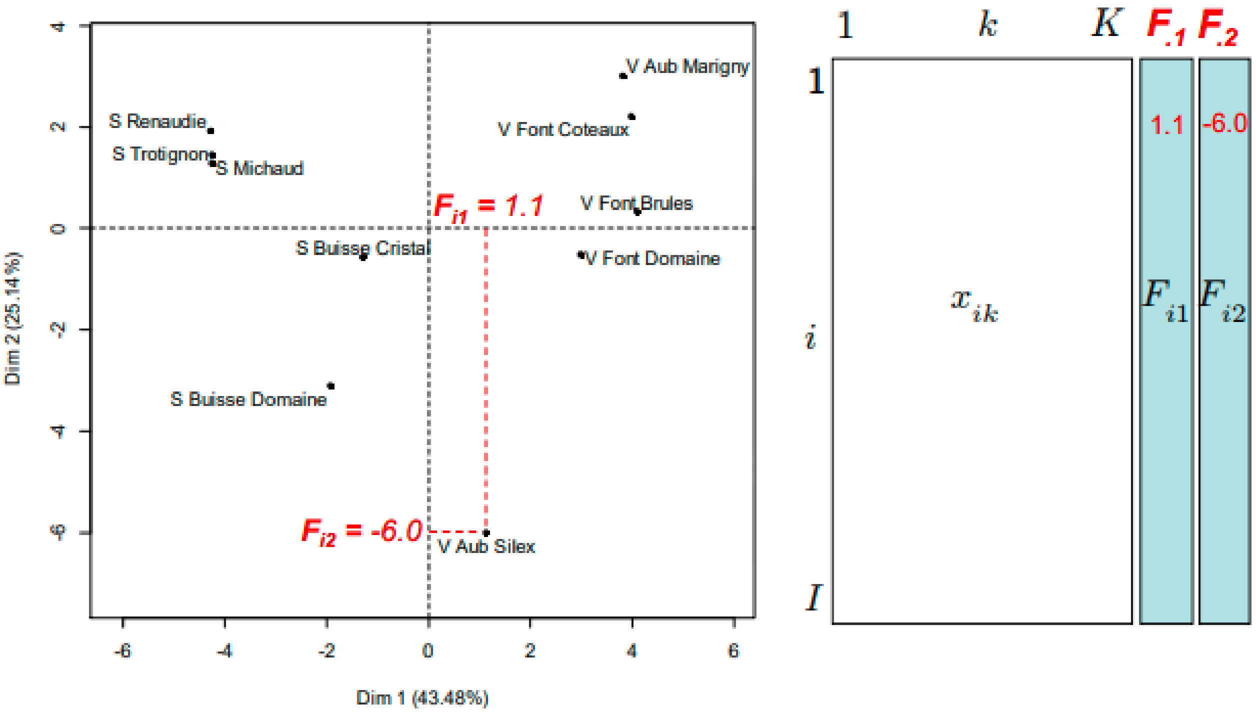
	O.fruity	O.passion	O.citrus	...	Sweetness	Acidity	Bitterness	Astringency	Aroma.intensity	Aroma.persistency	Visual.intensity	Odor.preference	Overall.preference	Label
S Michaud	4,3	2,4	5,7	...	3,5	5,9	4,1	1,4	7,1	6,7	5,0	6,0	5,0	Sauvignon
S Renaudie	4,4	3,1	5,3	...	3,3	6,8	3,8	2,3	7,2	6,6	3,4	5,4	5,5	Sauvignon
S Trotignon	5,1	4,0	5,3	...	3,0	6,1	4,1	2,4	6,1	6,1	3,0	5,0	5,5	Sauvignon
S Buisse Domaine	4,3	2,4	3,6	...	3,9	5,6	2,5	3,0	4,9	5,1	4,1	5,3	4,6	Sauvignon
S Buisse Cristal	5,6	3,1	3,5	...	3,4	6,6	5,0	3,1	6,1	5,1	3,6	6,1	5,0	Sauvignon
V Aub Silex	3,9	0,7	3,3	...	7,9	4,4	3,0	2,4	5,9	5,6	4,0	5,0	5,5	Vouvray
V Aub Marigny	2,1	0,7	1,0	...	3,5	6,4	5,0	4,0	6,3	6,7	6,0	5,1	4,1	Vouvray
V Font Domaine	5,1	0,5	2,5	...	3,0	5,7	4,0	2,5	6,7	6,3	6,4	4,4	5,1	Vouvray
V Font Brûlés	5,1	0,8	3,8	...	3,9	5,4	4,0	3,1	7,0	6,1	7,4	4,4	6,4	Vouvray
V Font Coteaux	4,1	0,9	2,7	...	3,8	5,1	4,3	4,3	7,3	6,6	6,3	6,0	5,7	Vouvray

Example: graphing the individuals



How to interpret the dimensions? Why are S. Trotignon and V. Font Brules far apart? $\Rightarrow$ Need variables to interpret the directions of variability

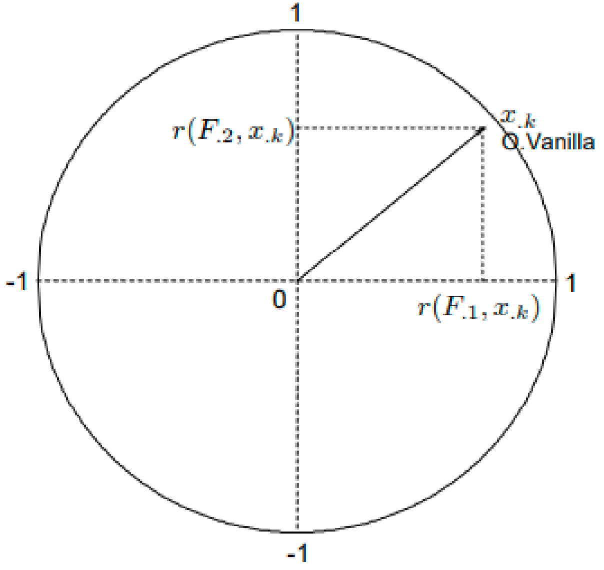
New individuals' coordinates : principal components



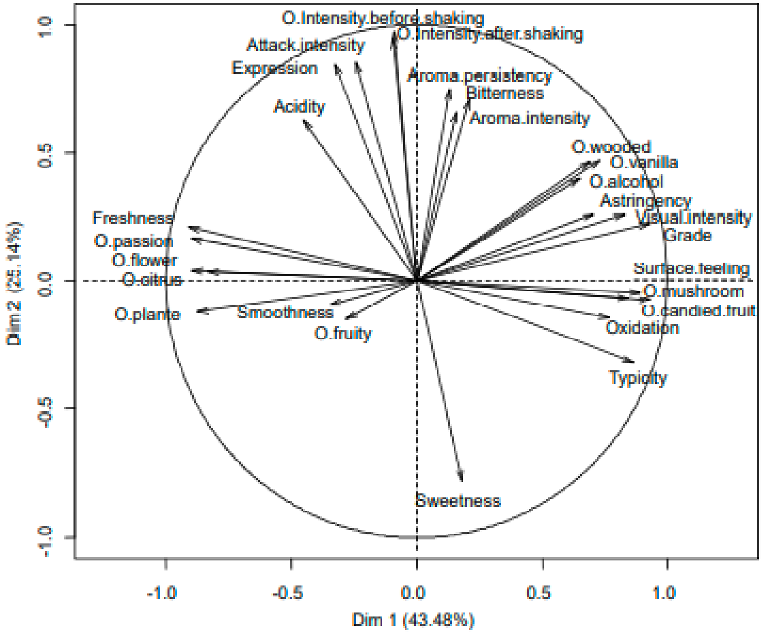
Representation of the variables as an interpretation aid for the individuals' cloud

Correlation circle explain new components using original variables

- Correlations between the variable $x_{.k}$ and $F_{.1}$ (and $F_{.2}$)



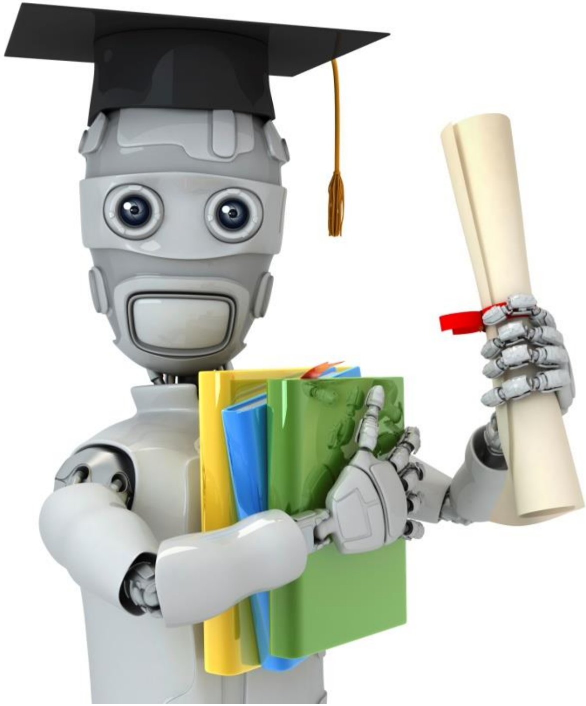
Representation of the variables as an interpretation aid for the individuals' cloud



Main directions of variability: ....

How to interpret the first dimension?

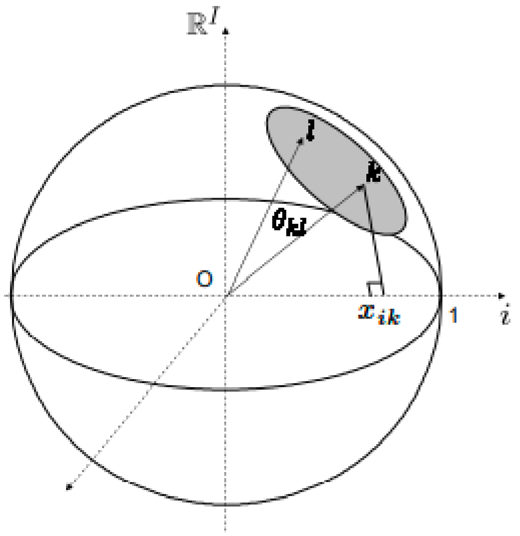
How to interpret the second dimension?



Machine Learning

Variables study:

Fitting the variables' cloud



1 variable = 1 point in an I-dimensional space

$$\cos(\theta_{kl}) = \frac{< x_{.k}, x_{.l} >}{\|x_{.k}\| \, \|x_{.l}\|} = \frac{\sum_{i=1}^{I} x_{ik} x_{il}}{\sqrt{\sum_{i=1}^{I} x_{ik}^2}\sqrt{\sum_{i=1}^{I} x_{il}^2)}}$$

Since variables are centered, $\cos(\theta_{kl}) = r(x_{.k}, x_{.l})$

If variables are standardized $\Rightarrow$ points are on an I-sphere of radius 1

Fitting the variables' cloud.

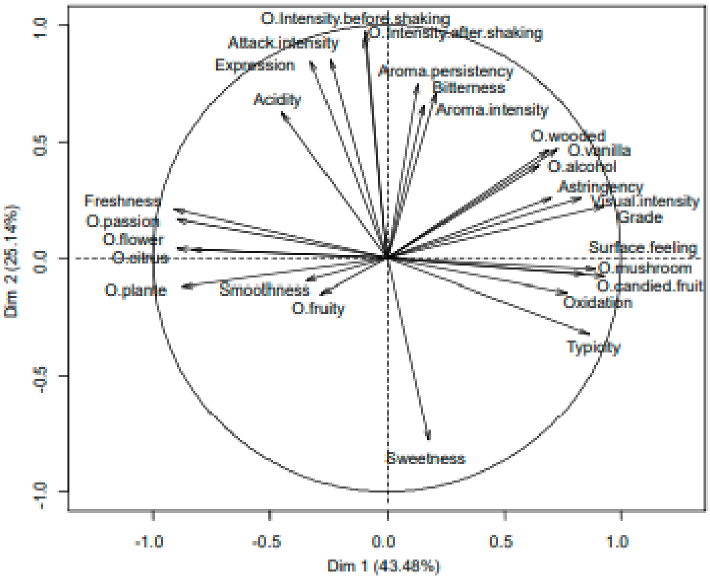
Similar strategy as for individuals: sequentially find orthogonal axes:

$$\underset{v_1 \in \mathbb{R}^I}{\arg\max} \sum_{k=1}^{K} r(v_1, x_{.k})^2$$

$\Rightarrow$ v_1 is the best synthetic variable for summarizing the variables

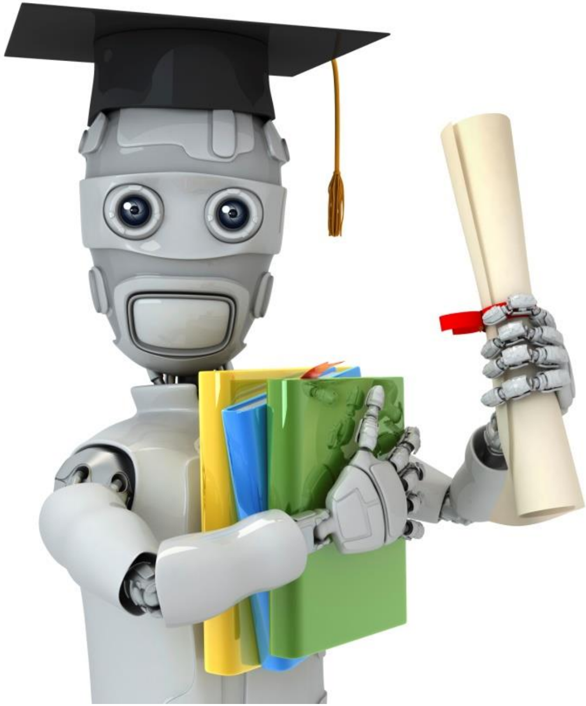
Find the 2$^{\mathrm{nd}}$ axis, then the 3$^{\mathrm{rd}}$, etc.

New presentation for our variables : same graph as before (when considering individuals)



⇒ Same graph as before!!!!

- interpretation aid for the individuals' graph
- optimal representation of the variables' cloud
- visualization of the correlation matrix

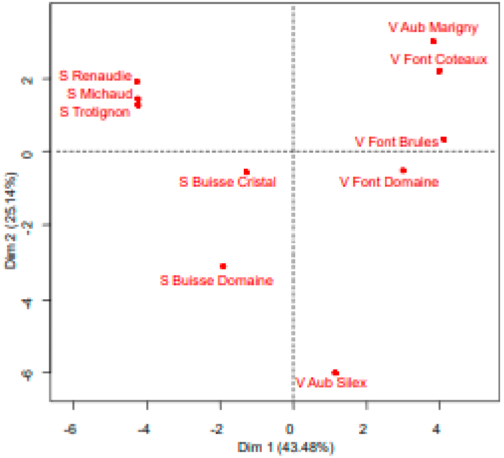


Machine Learning

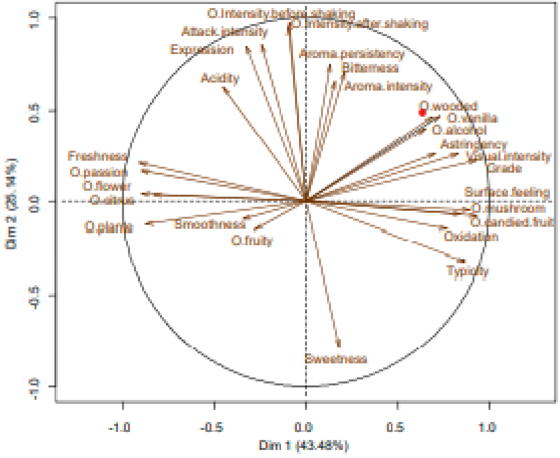
Interpretation of PCA results:

Interpretation is now possible using two graphs

$\Longrightarrow$ Individuals are on the same side as their corresponding variables with high values



	Aub Silex
O.intensity.after.shaking	-2,54
O.intensity.before.shaking	-2,37
Expression	-2,25
Acidity	-2,07
Attack.intensity	-1,36
Bitterness	-1,33
Freshness	-1,15
...	...
Typicity	1,01
Sweetness	2,93



Characterizing the axes

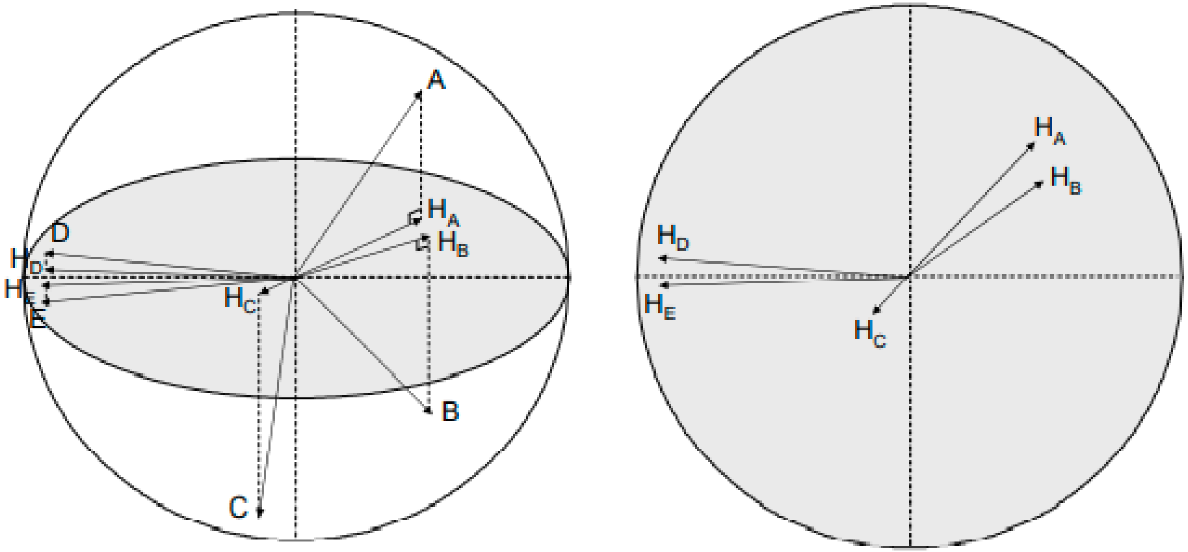
Using the continuous variables:

- correlation between each variable and the principal component of rank s is calculated
- correlation coefficients are sorted and significant ones are output

```
> dimdesc(res.pca)
                   $Dim.1$quanti                                    $Dim.2$quanti
                    corr p.value                                     corr p.value
O.candied.fruit    0.93 9.5e-05   O.intensity.before.shaking  0.97 3.1e-06
Grade              0.93 1.2e-04   O.intensity.after.shaking   0.95 3.6e-05
Surface.feeling    0.89 5.5e-04   Attack.intensity             0.85 1.7e-03
Typicity           0.86 1.4e-03   Expression                   0.84 2.2e-03
O.mushroom         0.84 2.3e-03   Aroma.persistency            0.75 1.3e-02
Visual.intensity   0.83 3.1e-03   Bitterness                   0.71 2.3e-02
   ...              ...     ...   Aroma.intensity              0.66 4.0e-02
O.plante          -0.87 1.0e-03
O.flower          -0.89 4.9e-04
O.passion         -0.90 4.5e-04
Freshness         -0.91 2.9e-04   Sweetness                   -0.78 8.0e-03
```

Quality of representation

$r(A, B) = cos(\theta_{A,B})$
$cos(\theta_{A,B}) \approx cos(\theta_{H_A,H_B})$ if the variables are well-projected



Only well-projected variables can be interpreted!

Quality of the representation

- $\cos^2(\theta_{iH_i})$ for the individuals: distance between individuals can only be interpreted for well-projected individuals

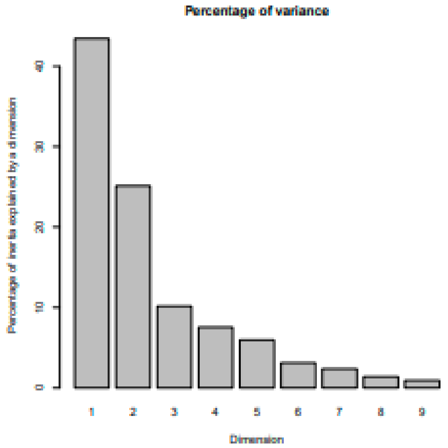
```
> round(res.pca$ind$cos2,2)
                  Dim.1 Dim.2
S Michaud          0.62  0.07
S Renaudie         0.73  0.15
S Trotignon        0.78  0.07
```

- $\cos^2(\theta_{kH_k})$ for the variables: only well-projected variables (high $\cos^2$) can be interpreted!

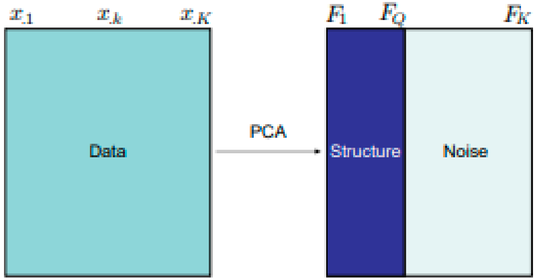
```
> round(res.pca$var$cos2,2)
                  Dim.1 Dim.2
O.fruity           0.08  0.02
O.passion          0.80  0.03
O.citrus           0.69  0.00
```

Choosing the number of dimensions

Bar chart of eigenvalues,
tests,
confidence intervals,
cross-validation (`estim_ncp` function),
Kaiser Rule

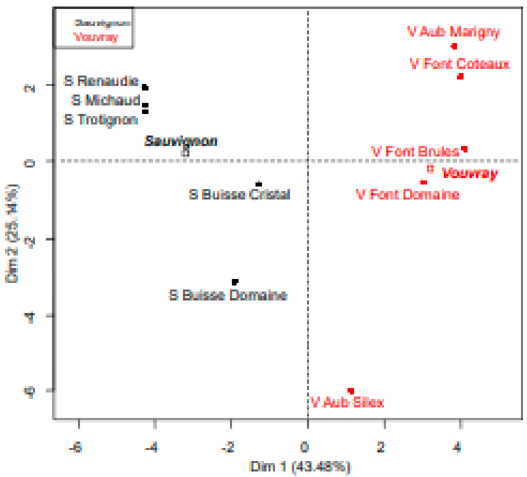


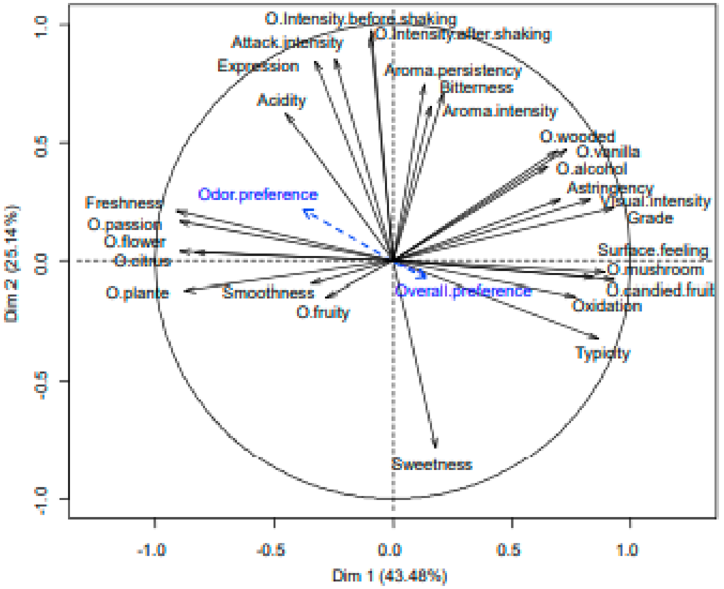
Two goals:
⇒ Interpretation
⇒ Separate structure from noise



Additional variables

- For the quantitative variables: project supplementary variables onto the axes
- For categorical variables: project the barycenter of individuals in each category





⇒ Supplementary information not used to build the axes

Contributions

⇒ Contributions to components:

- for an individual: $Ctr_s(i) = \frac{F_{is}^2}{\sum_{i=1}^{I} F_{is}^2} = \frac{F_{is}^2}{\lambda_s}$

  ⇒ Individuals with a large coordinate value contribute most

```
> round(res.pca$ind$contrib,2)
              Dim.1 Dim.2
S Michaud     15.49  3.10
S Renaudie    15.56  5.56
S Trotignon   15.46  2.43
```

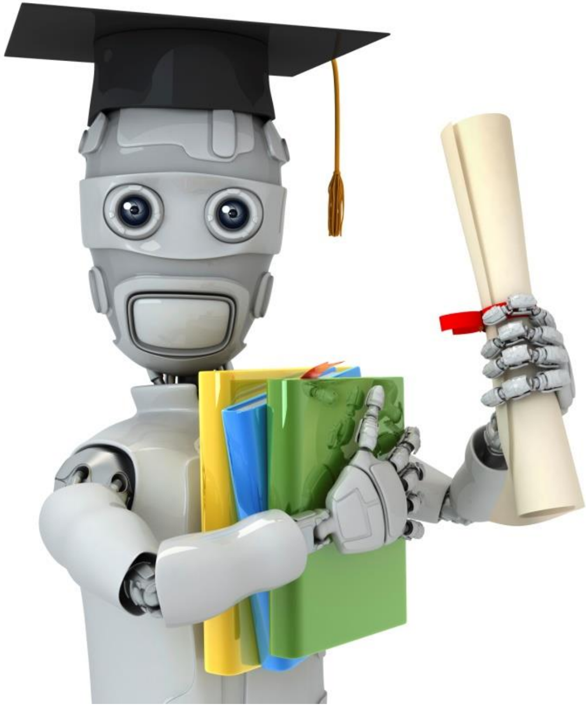
- for a variable: $Ctr_s(k) = \frac{r(x_{.k}, v_s)^2}{\sum_{k=1}^{K} r(x_{.k}, v_s)^2} = \frac{r(x_{.k}, v_s)^2}{\lambda_s}$

  ⇒ Variables highly correlated with the principal component contribute the most

```
> round(res.pca$var$contrib,2)
              Dim.1 Dim.2
O.fruity       0.67  0.34
O.passion      6.84  0.40
O.citrus       5.89  0.02
```

Main steps of PCA in practice

1. Choose active variables
2. Rescale (or not) the variables
3. Perform PCA
4. Choose the number of dimensions to interpret
5. Joint analysis of the cloud of individuals and the cloud of variables
6. Use indicators to enrich interpretation
7. Go back to raw data for interpretation



Machine Learning

See you next chapter

Thanks :)