EISTI

Co-funded by the Erasmus+ Programme of the European Union

TowardanOpenResourcesUsingServices

Multivariate analysis – R examples

Astrid JOURDAN – Mathematics department - EISTI

Computer Architecture and Environmental Science Application

Ferrara 6-10 june 2016

Co-funded by the
Erasmus+ Programme
of the European Union

Reminder of some datamining concepts

The curse of dimensionality

Principal analysis components

Practice with R

Process for a data mining study

Co-funded by the Erasmus+ Programme of the European Union

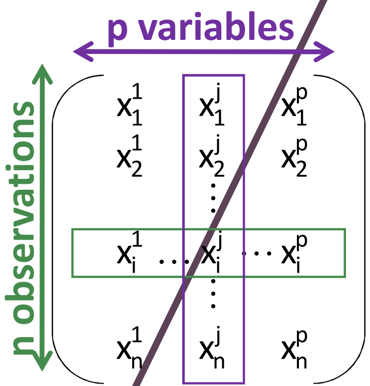
Definition of the study
what issues do you want to answer?
Define the ***statistical unit*** *and the variables*

→ **Collect the data**
select the data in accordance with your issues

→ **Pre-processing the data**
outliers, standardisation, missing data, …

→ **Apply methods**
Choice of the method depends on

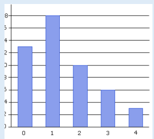
- *The nature of the variables*
- *The size of the data*
- *The objective*

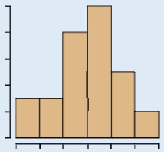


DATA is a table with

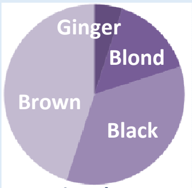
- **p columns :** *realizations of the random variables*
- **n rows :** *observations of the statistical unit*

A Quantitative variable
is numerical and represents a measurable quantity (continuous or discrete)

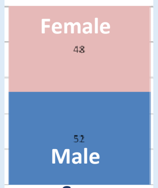

Family


Size

A Categorical variable
takes on values that are names or labels



Hair colors



Sex

Basic data mining objective

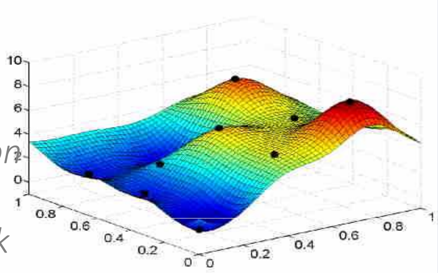
Co-funded by the Erasmus+ Programme of the European Union

Supervised learning /Prediction

Explain a target variable with other variables

- Quantitative target

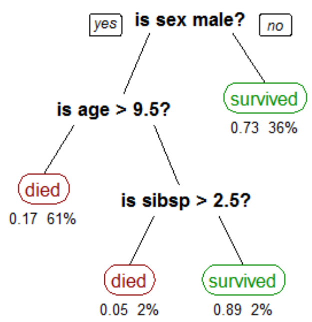
Regression



✓ *linear regression*
✓ *SVR*
✓ *neural network*
✓ ...

- Categorial target

Classification



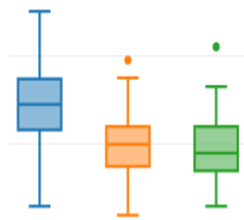
✓ *logistic regression*
✓ *decision tree*
✓ *neural network*
✓ ...

Unsupervised learning /Description

Describe hidden structure from unlabeled data

- Description

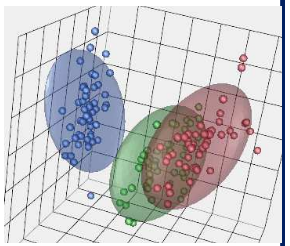
Summarize the variables, detect link between Variables, detect outliers, …



✓ *Graphical and numerical description*
✓ *PCA, CA ,,…*
✓ *…*

- Clustering

Organizing objects into groups (clusters) such that the members in each Group are similar



✓ *K-means*
✓ *Hierarchical clustering*
✓ *…*

Overfitting and generalization

The goal of supervised learning is to find a model *f* such as $Y=f(X_1,\ldots,X_p)+\varepsilon$, where ε is an error

Error measurement

- Regression : sum of residuals

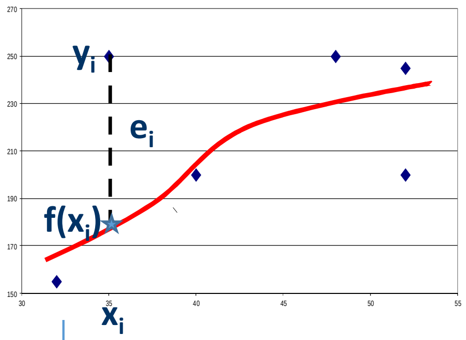
$$\sum_{i=1}^{n} L(y_i - f(x_i))$$

where L is a cost function (e.g. $L(u)=|u|$, $L(u)=u^2$)

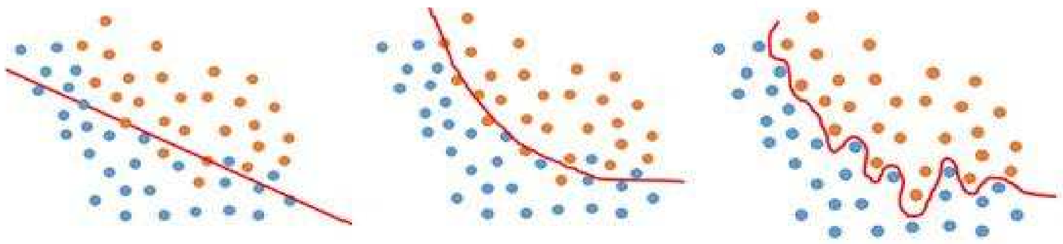
- Classification : confusion matrix

		Predicted cluster			
		C1	C2	…	Ck
True cluster	C1				
	C2				
	…				
	Ck				

diagonal = good prediction



underfitting *good fitting* *overfitting*

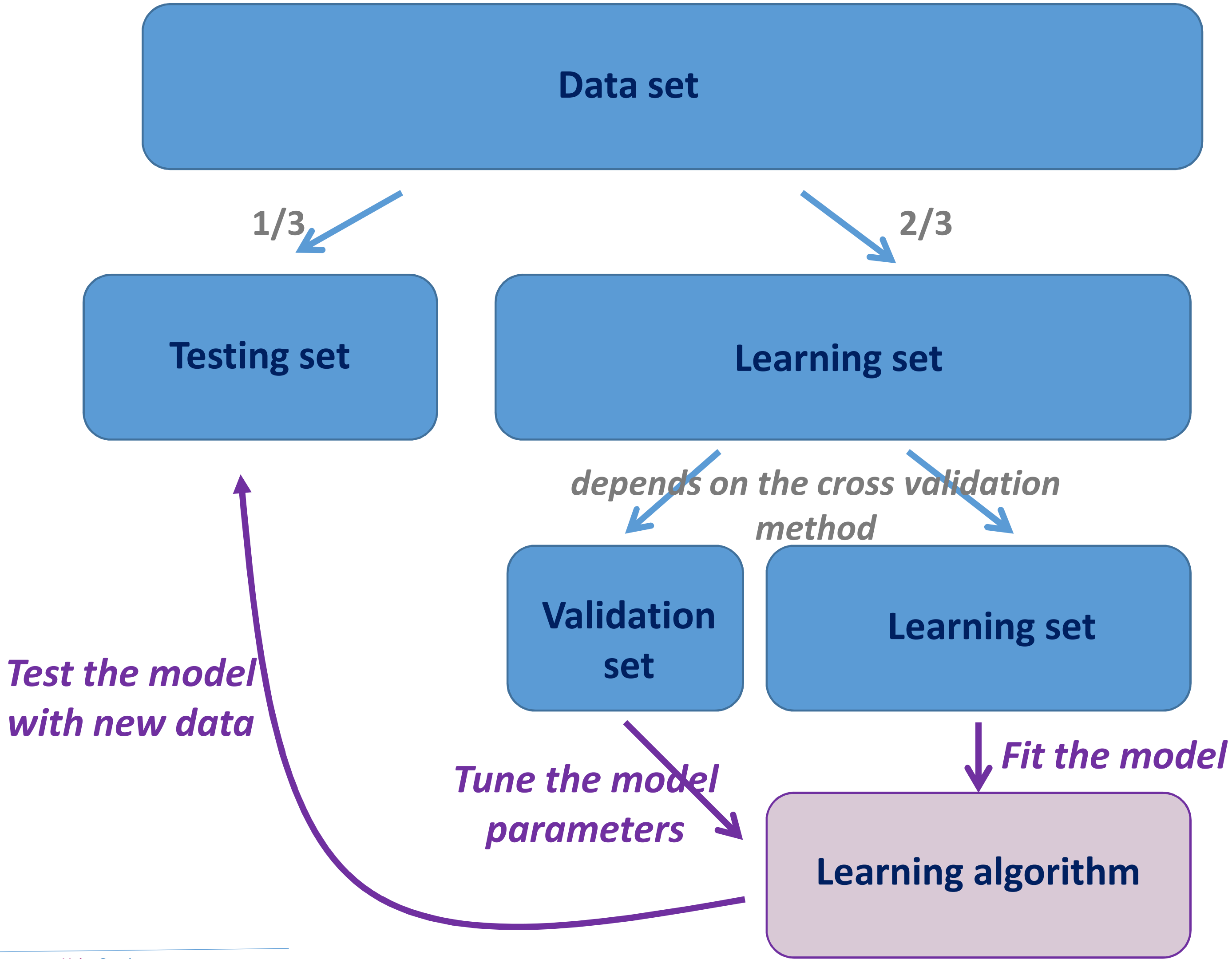


Two kinds of errors :

- **Fitting error** : computed on the training dataset . A small error means that the model reproduced the training dataset well
- **Prediction error** : computed on a new dataset. A small error means that the model is able to predict new values

Process for learning

Co-funded by the Erasmus+ Programme of the European Union



Predictive methods

variables → target ↓	1 quantitative	*n* quantitative	1 categorial	*n* categorial	Mixed
1 quantitative	Linear regression, decision tree, SVR	Multiple linear regression, PLS regression, decision tree, neural networks, SVR	ANOVA, decision tree, SVR	ANOVA, decision tree, neural networks, SVR	ANCOAV, decision tree, neural networks, SVR
1 categorial	Discriminant analysis, Logistic regression, decision tree, neural networks SVM	Multiple discriminant analysis, logistic regression, PLS logistic regression, decision tree, neural networks SVM	Discriminant analysis, logistic regression, decision tree, neural networks SVM	Multiple discriminant analysis, logistic regression, decision tree, neural networks SVM	Logistic regression, decision tree, neural networks SVM

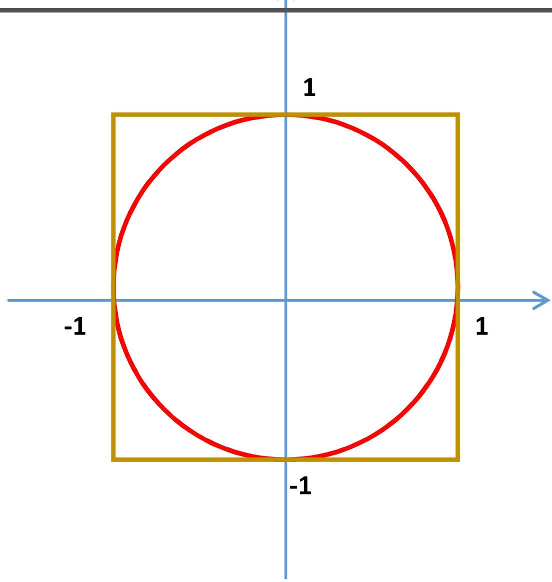
The curse of dimensonality

Co-funded by the Erasmus+ Programme of the European Union

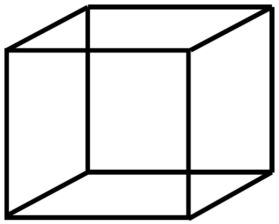
Dimension = number of variables

When the dimensionality increases, the volume of the space increases so fast that the available data become sparse.

- Impossible to analyse and organize sparse data
- An enormous amount of training data are required to ensure a good exploration of a high dimensional space.



	Covering		
d	Vol. hypercube	Vol. sphere	%
2	4	3,1	78,5%
4	16	4,9	30,8%
6	64	5,2	8,1%
8	256	4,1	1,6%
10	1024	2,6	0,2%



Grid with 2 levels = 2^d points (e.g. d=20 $\Rightarrow$ 33554432 points)
k levels = k^d points (e.g. k=5 and d=10 $\Rightarrow$ 9765625 points)

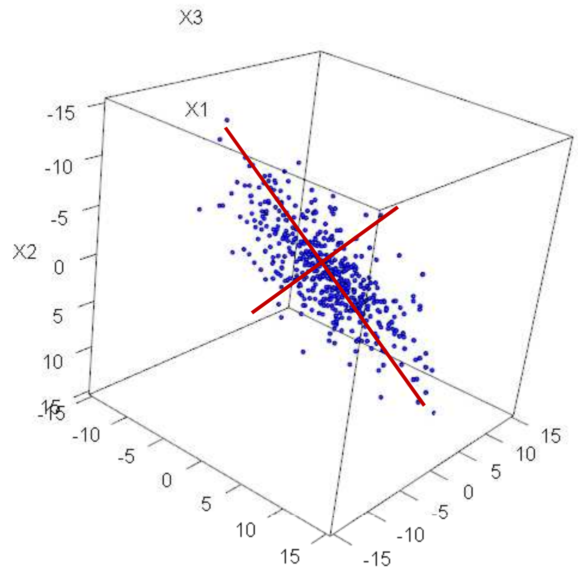
Dimensionality reduction

Dimensionality reduction

Co-funded by the Erasmus+ Programme of the European Union

A large number of variables (features):

- increases overfitting
- contains redundant variables
- increases the training time and requires a large amount of memory and computation power



Feature selection

Selection of a subset of relevant features for a model

Feature selection algorithms remove :

✓ redundant variables (correlated, mutual information,...)

✓ irrelevant variables (measure of accuracy, AIC, BIC, MSE...) ***(risk of overfitting)***

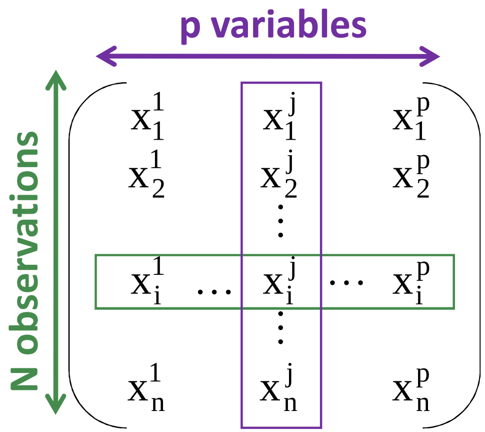
Feature extraction

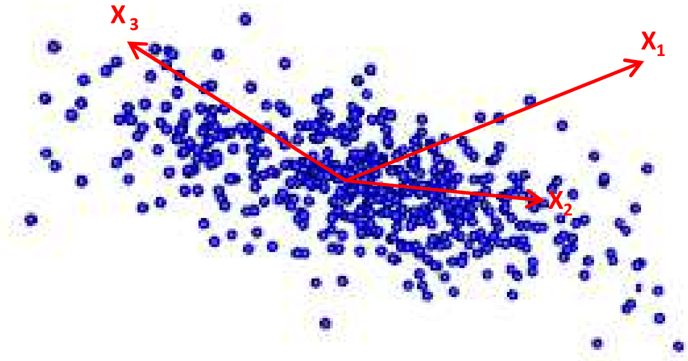
Creation of new variables from the original variables

✓Linear (e.g. PCA) or nonlinear transformation (e.g. kernel PCA)

✓ Criteria to measure the « loss of information » (variance, pairwise distances ,...)

Principal component analysis (PCA)





p-dimensional coordinate system defined by the variables

Transformation of the original coordinate system to an orthogonal coordinate system

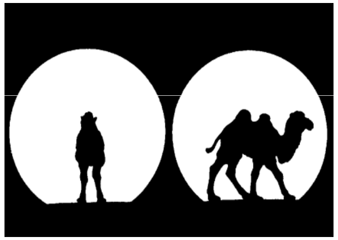
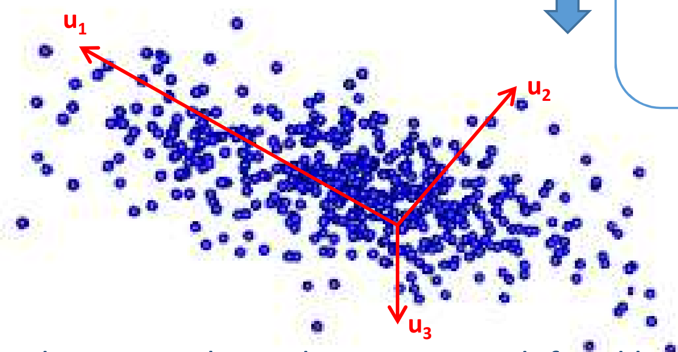


Figure J.P. Fenelon

Dimensionality reduction
=
Projection on vect{$u_1,\ldots,u_K$}
where **k<<p**
$\Rightarrow$
loss of information



p-dimensional coordinate system defined by the principal axis

EISTI

Principal component analysis

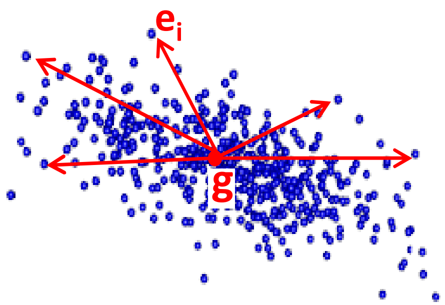
Co-funded by the Erasmus+ Programme of the European Union

Information

The information is the inertia of the data set

$$I = \frac{1}{n}\sum_{i=1}^{n} \| e_i - g \|^2$$

where g is the gravity center of the data (mean). Note that I=tr(V) where V is the covariance matrix (Inertia = sum of variances).



	Pop. (T)	Life exp.	Nb. child
Argentina	41050	75,87	2,19
Armenia	3099	74,44	1,77
...			

Distance between Argentina and Armenia
= $(41050-3099)^2+(75,87-74,44)^2+(2,19-1,77)^2=1440278405$
$\cong(41050-3099)^2$

Reduced and centered variables : $x_i^k \leftarrow \frac{x_i^k - \bar{x}^k}{s_k}$

Orthogonal random variables

Scalar prod. : $<X^k,X^h>=E[X^kX^h]$ → *Centered variables* $<X^k,X^h>=\mathrm{cov}(X^k,X^h)$

Norm : $\|X\|^2=E[X^2]$ → $\|X\|^2=\mathrm{var}(X)$

$$\cos(\widehat{X^k,X^h}) = \frac{<X^k,X^h>}{\|X^k\|\|X^h\|}$$

Cosine between two variables
Linear correlation coef.

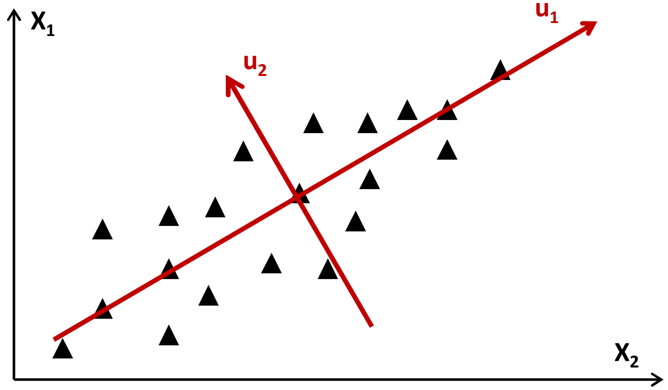
$$r(X^k,X^h) = \frac{\mathrm{cov}(X^k,X^h)}{\sqrt{\mathrm{var}(X^k)\mathrm{var}(X^h)}}$$

Orthogonal variables
=
uncorrelated variables

Principal component analyse

Construction of the principal axes

- The 1st principal axis (u_1) catches a maximum variance (inertia).
- The 2nd principal axis (u_2) catches the maximum of the remaining variance and is orthogonal to the 1st axis
- Each succeeding component is built in the same way until the last axis (u_p)



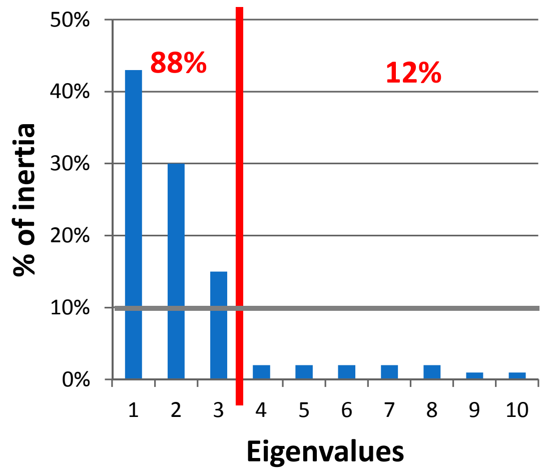
Solution

The principal axes are the eigenvectors associated to the eigenvalues $\lambda_1,\dots,\lambda_p$ of the covariance matrix V such that $\lambda_1>\dots>\lambda_p$.

- Inertia of the data projected on u_k is λ_k
- Inertia of the data projected on $<u_1,\dots,u_k>$ is $\lambda_1+\dots+\lambda_k$
- Total inertia is $I=\lambda_1+\dots+\lambda_p$

The principal components are the components of the data projected on the principal axis.

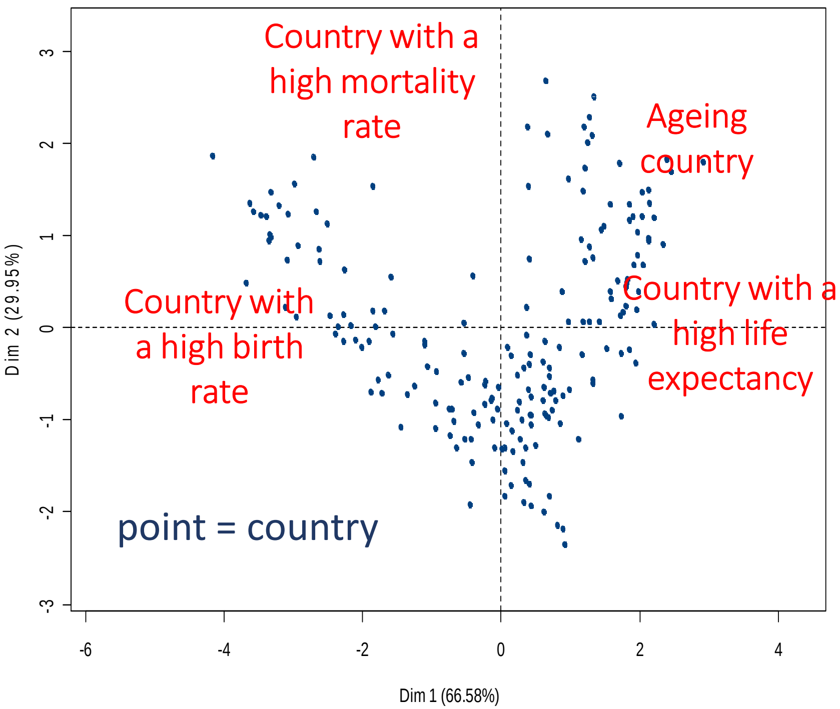
!!! PCA is sensitive to outliers !!!



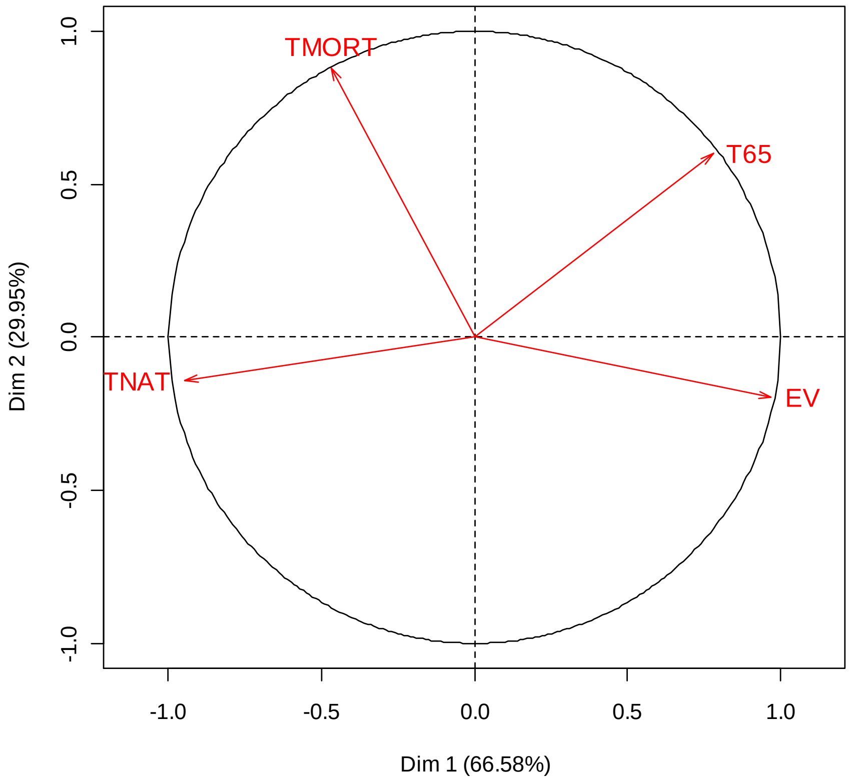
Results of PCA

- PCA is a method to reduce the dimension
- PCA is also a method to describe and understand the data

Individual graph



Variable graph

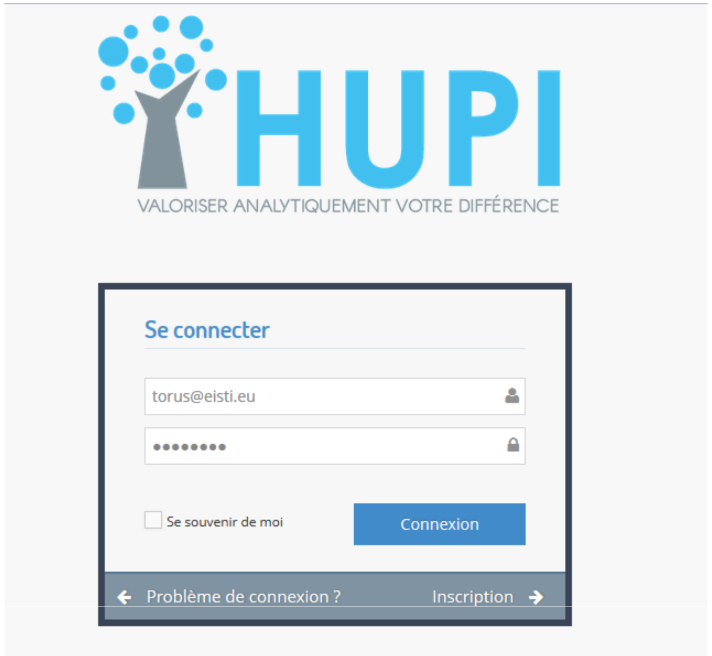


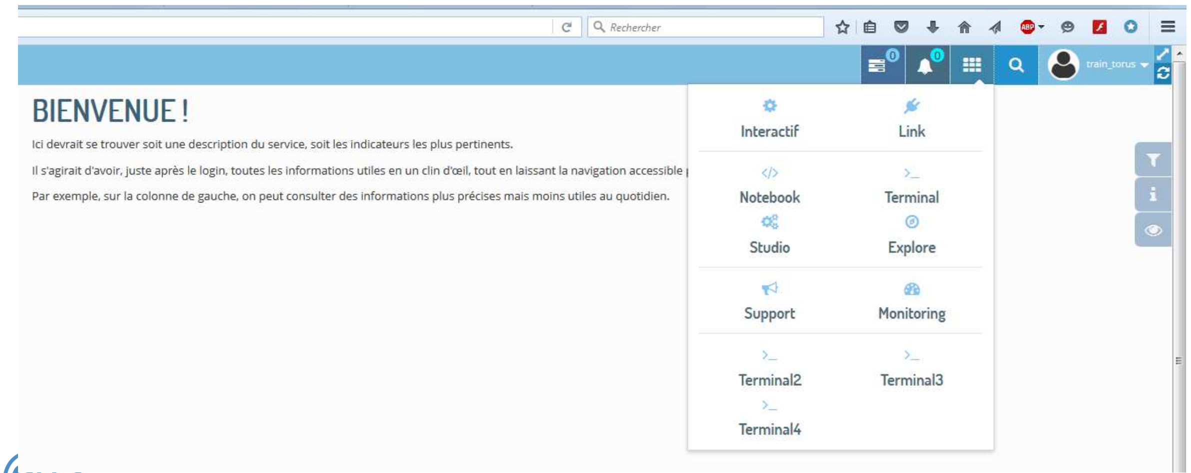
Now practice with R!

Use Mozilla Firefox to connect to HUPI

http://ecoles.hupi.io/

user : torus@eisti.eu
Password : Lz4eA8b7





Select a Terminal

login : train_torus
Password : Lz4eA8b7