Statistics and learning

Multivariate statistics 1

Emmanuel Rachelson and Matthieu Vignes

ISAE SupAero

Wednesday 25th September 2013

Motivating examples (1)

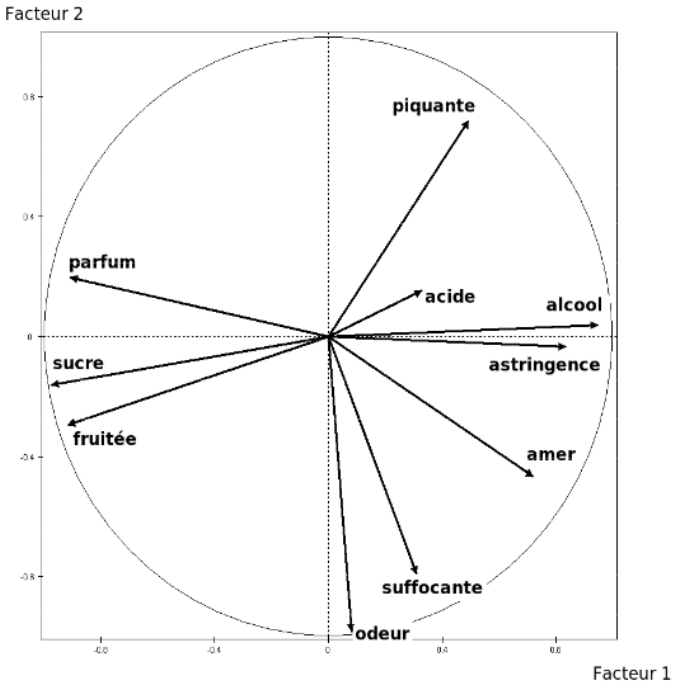
Cider get different measures gathered in

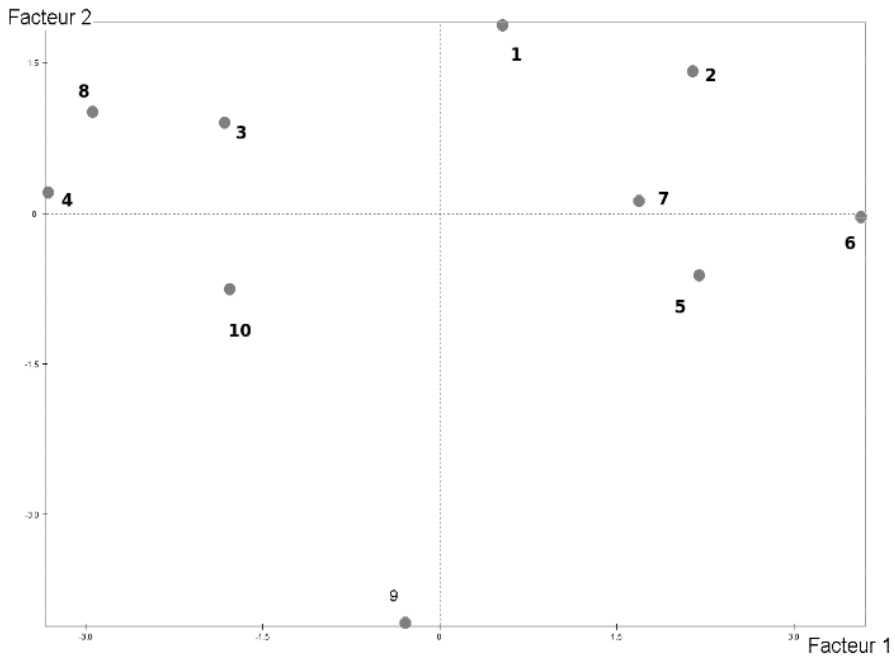
cidre	odeur	sucre	acide	amer	astringence	suffocante	piquante	alcool	parfum	fruitée
1	2,14	1,86	3,29	2,29	2	0,14	2,29	1,86	1,29	1,29
2	2,43	0,79	2,71	2,57	2	0,43	2,57	2,86	0,43	0,14
3	2,71	3,14	2,57	2,57	1,43	0,14	2,14	0,86	2,29	1,71
4	3	3,71	2,14	2,07	1,57	0	1,29	1	3,14	3,14
5	3,43	1,29	2,86	3,14	2,17	1	1,86	2,86	1,14	0,29
6	3,14	0,86	2,86	3,79	2,57	0,14	1,71	3,29	0,14	0
7	3,14	1,14	2,86	2,86	2	0,43	1,71	1,86	0,14	0
8	2,43	3,71	3,21	1,57	1,71	0	1	0,57	2,57	2,86
9	5,1	2,86	2,86	3,07	1,79	1,71	0,43	1,43	0,57	2,71
10	3,07	3,14	2,57	3	2	0	0,43	1,29	2,57	3,07

TAB. 1 – Notes obtenues par 10 marques de cidres sur 10 critères lors d'un concours agricole.

Motivating examples (1)

I claim that





represents 75% of the variance in the data !

Motivating examples (2)

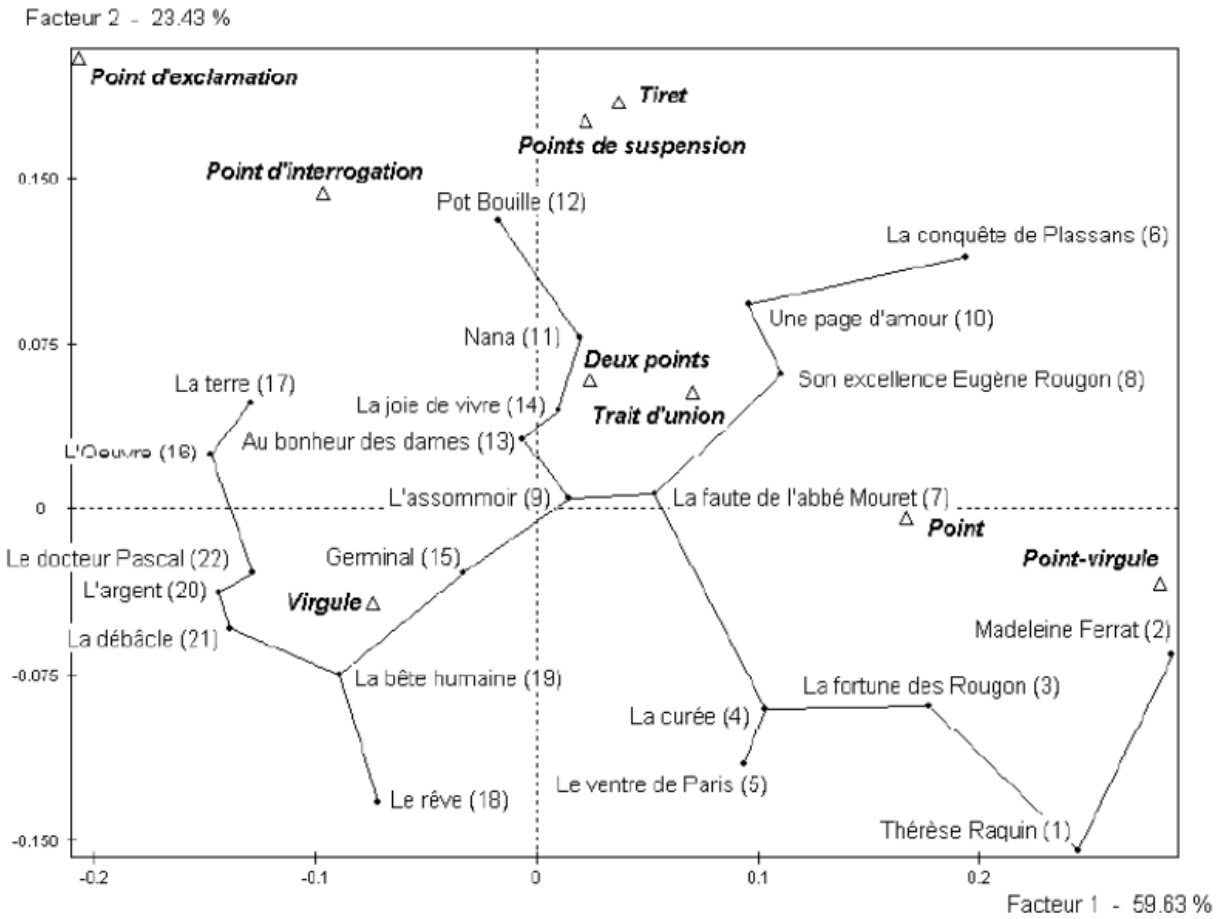
A nice representation of

Roman	.	...	!	?	,	;	:	—	-
1. Thérèse Raquin	3468	236	138	76	6195	691	168	285	543
2. Madeleine Ferrat	5131	362	236	245	8012	922	291	518	1115
3. La fortune des Rougon	6157	238	534	229	11346	936	362	711	1301
4. La curée	4958	443	357	232	11164	738	364	679	1200
5. Le ventre de Paris	5538	534	426	232	13234	1015	318	734	1201
6. La conquête de Plassans	6292	943	756	512	11585	1285	402	1432	1916
7. La faute de l'abbé Mouret	6364	679	859	462	13948	634	377	1067	1564
8. Son excellence Eugène Rougon	7258	728	1002	496	14295	889	543	1469	1907
9. L'assommoir	7820	769	1929	443	19244	1399	436	995	2272
10 Une page d'amour	6206	843	918	492	11953	647	347	1235	1409
11. Nana	7821	1007	1796	611	17881	1087	509	1523	1797
12. Pot Bouille	6875	1045	1873	651	17044	912	675	1669	1935
13. Au bonheur des dames	6916	808	1313	651	18402	972	642	1531	2114
14. La joie de vivre	5803	710	972	623	13917	602	420	1142	1590
15. Germinal	7944	606	1463	729	21388	908	621	1362	2083
16. L'Œuvre	5000	774	1692	668	18292	811	566	1107	1489
17. La terre	6979	957	2307	796	23417	947	657	1681	2113
18. Le rêve	3052	292	385	237	9551	345	230	416	650
19. La bête humaine	5484	601	929	557	18264	673	467	957	1721
20. L'argent	5022	850	1235	569	19267	684	399	1049	1677
21. La débâcle	7440	860	1833	690	26482	832	564	1398	2197
22. Le docteur Pascal	4586	621	1072	464	15598	462	315	955	1218

??

Motivating examples (2)

Information can be summarised in a sense to be precised in



Take-home message

'Simple', descriptive data analysis. And interpretations !

- INPUT: An array of data (can be more than 2D).

Take-home message

'Simple', descriptive data analysis. And interpretations !

- INPUT: An array of data (can be more than 2D).
- Identify statistical units of the population/sample and variables under study.

Take-home message

'Simple', descriptive data analysis. And interpretations !

- INPUT: An array of data (can be more than 2D).
- Identify statistical units of the population/sample and variables under study.
- Describe the variables $\rightarrow$ type, univariate description before you move on to...

Take-home message

'Simple', descriptive data analysis. And interpretations !

- INPUT: An array of data (can be more than 2D).
- Identify statistical units of the population/sample and variables under study.
- Describe the variables $\rightarrow$ type, univariate description before you move on to...
- ...bivariate (*e.g.* simple regression) and multivariate data analysis.

Take-home message

'Simple', descriptive data analysis. And interpretations !

- INPUT: An array of data (can be more than 2D).
- Identify statistical units of the population/sample and variables under study.
- Describe the variables $\rightarrow$ type, univariate description before you move on to...
- ...bivariate (*e.g.* simple regression) and multivariate data analysis.
- The goals are to describe the data and to summarise its informational content: highlight *patterns* in the data, represent in low-dimensions most of its variations.

Take-home message

'Simple', descriptive data analysis. And interpretations !

- INPUT: An array of data (can be more than 2D).
- Identify statistical units of the population/sample and variables under study.
- Describe the variables $\rightarrow$ type, univariate description before you move on to...
- ...bivariate (*e.g.* simple regression) and multivariate data analysis.
- The goals are to describe the data and to summarise its informational content: highlight *patterns* in the data, represent in low-dimensions most of its variations.
- Important point: do not forget to interpret the analysis you produce !

Take-home message

'Simple', descriptive data analysis. And interpretations !

- INPUT: An array of data (can be more than 2D).
- Identify statistical units of the population/sample and variables under study.
- Describe the variables $\rightarrow$ type, univariate description before you move on to...
- ...bivariate (*e.g.* simple regression) and multivariate data analysis.
- The goals are to describe the data and to summarise its informational content: highlight *patterns* in the data, represent in low-dimensions most of its variations.
- Important point: do not forget to interpret the analysis you produce !
- OUTPUT: a nice (set of) representations of the data with key points to explain what's in it !

First: univariate statistics

- Any data set to be 'analysed' need to be explored first !

First: univariate statistics

- Any data set to be 'analysed' need to be explored first !
- Tools might look simplistic but **robust** in interpretations.

First: univariate statistics

- Any data set to be 'analysed' need to be explored first !
- Tools might look simplistic but **robust** in interpretations.
- Way to get familiar with data set at hand: missing obs., erroneous/atypic points (outliers), (exp.) bias, rare modalities, variable distribution. . .

First: univariate statistics

- Any data set to be 'analysed' need to be explored first !
- Tools might look simplistic but **robust** in interpretations.
- Way to get familiar with data set at hand: missing obs., erroneous/atypic points (outliers), (exp.) bias, rare modalities, variable distribution...
- Allow analyst to pre-process the data: transformation(s), class recoding...

First: univariate statistics

- Any data set to be 'analysed' need to be explored first !
- Tools might look simplistic but **robust** in interpretations.
- Way to get familiar with data set at hand: missing obs., erroneous/atypic points (outliers), (exp.) bias, rare modalities, variable distribution...
- Allow analyst to pre-process the data: transformation(s), class recoding...

Quantitative variables

- From collected data to statistical table (frequency table).
- a prelude to graphical representation: 'stem-and-leaf' presentation.
- Bar and cumulative diagrams; histograms & (Kernel) density est.
- Quantiles and box(-and-whisker) plot.
- Numerical features (centrality, dispersion...).
- Minor differences for continuous and discrete quantitative variables.

Univariate statistics (con'd)

Qualitative variable

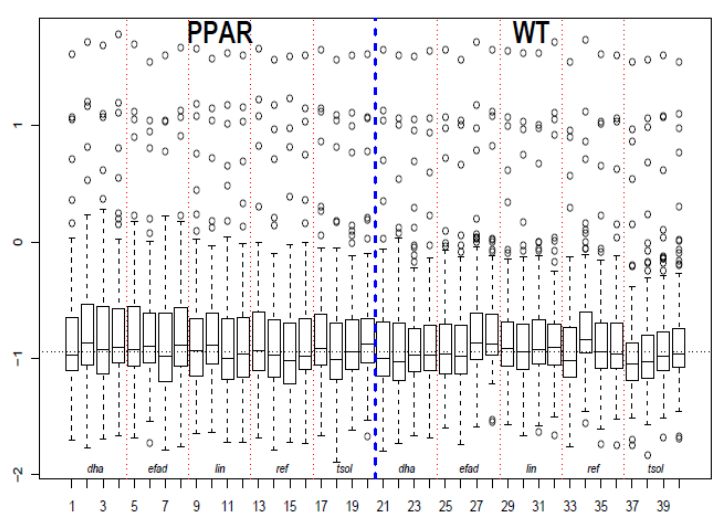
- Nominal vs. ordinal variables.
- No numerical summary from data itself $\rightarrow$ **tables** (frequency or percentages) and **graphics** (bar or pie charts).

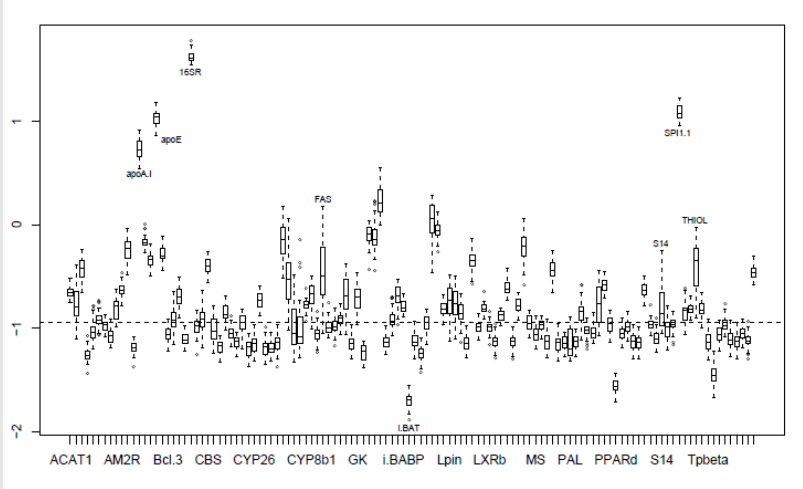
Univariate statistics (con'd)

Qualitative variable

- Nominal vs. ordinal variables.
- No numerical summary from data itself $\rightarrow$ **tables** (frequency or percentages) and **graphics** (bar or pie charts).

Genomic data





Descriptive bivariate statistics

before it's difficult to represent it

We now consider the simultaneous study of 2 variables X and Y.

The main objective is to highlight a **relationship** between these variables.

Sometimes it can be interpreted as a cause.

Descriptive bivariate statistics

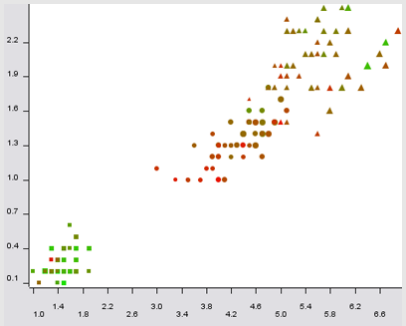
before it's difficult to represent it

We now consider the simultaneous study of 2 variables X and Y.

The main objective is to highlight a **relationship** between these variables. Sometimes it can be interpreted as a cause.

Two quantitative variables

- Scatter plot (may need to scale variables).



- Give a relationship index. *E.g.* covariance and correlation: $\mathrm{cov}(X,Y) = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})$ and $\mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$. And interpret.

Descriptive bivariate statistics (cont'd)

A quantitative variable X and a qualitative variable Y

- Parallel boxplots.
- Partial mean and sd on subpop. for all level of Y. $\rightarrow$ decomposition $\sigma_X^2 = \sigma_E^2 + \sigma_R^2$, where σ_E^2: variance explained by the partition of Y and σ_R^2: residual (between groups) variance. The ratio σ_E^2/σ_X^2 is an link index between X and Y.

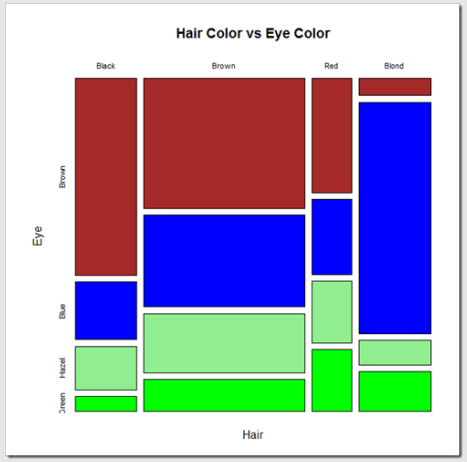
Descriptive bivariate statistics (cont'd)

A quantitative variable X and a qualitative variable Y

- Parallel boxplots.
- Partial mean and sd on subpop. for all level of Y. $\rightarrow$ decomposition $\sigma_X^2 = \sigma_E^2 + \sigma_R^2$, where σ_E^2: variance explained by the partition of Y and σ_R^2: residual (between groups) variance. The ratio σ_E^2/σ_X^2 is an link index between X and Y.

Two qualitative variables

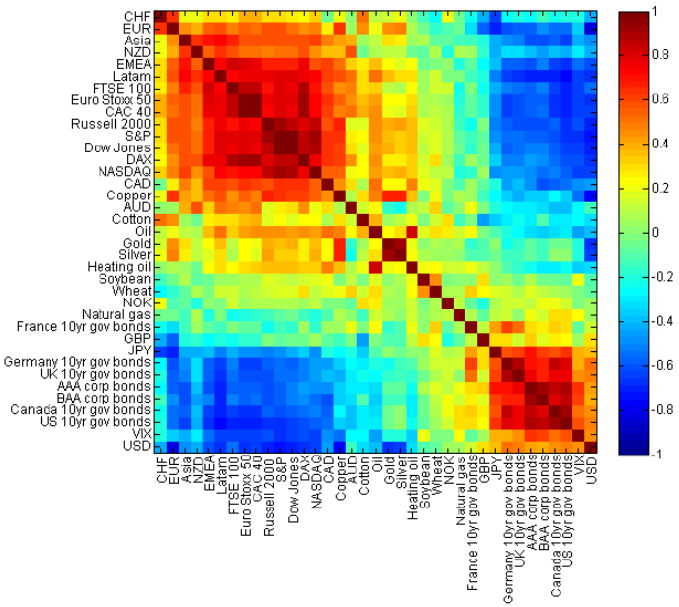
- Contingency table
- Mosaic plots with areas $\propto$ frequencies.
- Relationship index: $\chi^2 = \sum\sum \frac{(n_{kl} - s_{kl})^2}{s_{kl}}$

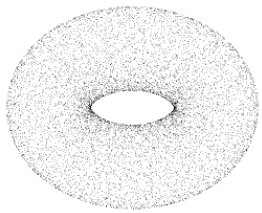


Towards multidimensional statistics

Adapting/generalising what's been seen previously:

- Matrix of correlations (symetric, positive-definite)
- Point of clouds (3D) / scatter plot matrix





Principal Component Analysis (PCA)

an introduction

- The bivariate study raised the obvious question of representing $p > 2$ variable data sets.
- Mathematically speaking, it's only a change of basis (from canonical to factor-driven). It is optimal in some sense.

Toy example

	Math.	Phys.	Engl.	Fren.
Mike	32	31	25	26
Helen	41	38	39	42
Alan	30	36	55	49
Dona	74	73	79	74
Peter	71	71	59	62
Brigit	54	51	28	35
John	26	34	70	58
William	65	62	43	47
Pam	46	48	62	61

Toy (mark) example

Toy example: data description

Elementary univariate statistics

Variable	mean	stand. dev.	min.	max
Math.	48.8	18.2	26	74
Phys.	49.3	16.1	31	73
Engl.	51.1	18.6	25	79
Fren.	50.4	14.9	26	74

Toy (mark) example

Toy example: data description

Elementary univariate statistics

Variable	mean	stand. dev.	min.	max
Math.	48.8	18.2	26	74
Phys.	49.3	16.1	31	73
Engl.	51.1	18.6	25	79
Fren.	50.4	14.9	26	74

Correlation matrix

	Math.	Phys.	Engl.	Fren.
Math.	1	0.9796	0.2316	0.4687
Phys.	0.9796	1	0.3972	0.6104
Engl.	0.2316	0.3972	1	0.9596
Fren.	0.4687	0.6104	0.9596	1

Toy (mark) example

Spectral decomposition of the covariance matrix

(Variance-)covariance matrix

	Math.	Phys.	Engl.	Fren.
Math.	330.19	286.46	78.15	126.99
Phys.	286.46	259.00	118.71	146.46
Engl.	78.15	118.71	344.86	265.69
Fren.	126.99	146.46	265.69	222.28

Toy (mark) example

Spectral decomposition of the covariance matrix

(Variance-)covariance matrix

	Math.	Phys.	Engl.	Fren.
Math.	330.19	286.46	78.15	126.99
Phys.	286.46	259.00	118.71	146.46
Engl.	78.15	118.71	344.86	265.69
Fren.	126.99	146.46	265.69	222.28

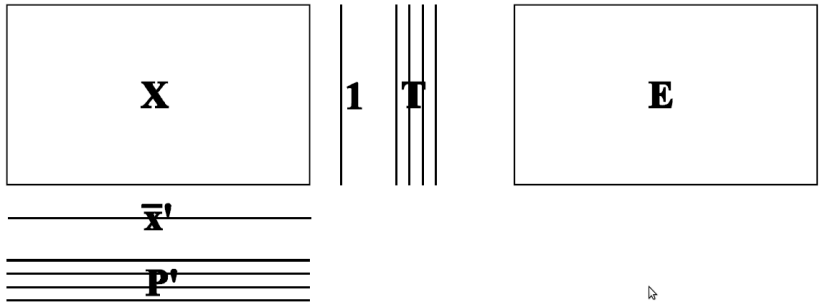
Eigen values of the covariance matrix

Factor	Eig. values	Variance percentage
F1	801.1	69.3 %
F2	351.4	30.4 %
F3	2.6	0.2 %
F4	1.2	0.1 %

PCA

- Statistical interpretation: PCA = iterative search for orthogonal linear combinations of initial variables with greatest variance.
- Geometrical interpretation: PCA = search for the best projection subspace which provides the most faithful individual/variable representation.

PCA model:
$X = {}^{\top}\bar{x} + T\,{}^{\top}P + E$

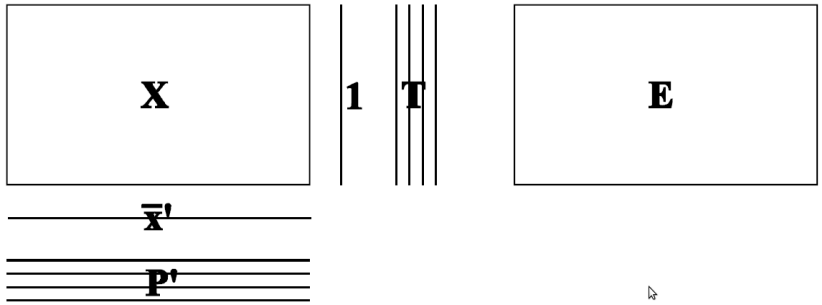


-

PCA

- Statistical interpretation: PCA = iterative search for orthogonal linear combinations of initial variables with greatest variance.
- Geometrical interpretation: PCA = search for the best projection subspace which provides the most faithful individual/variable representation.

PCA model:
$X = {}^{\top}\bar{x} + T\,{}^{\top}P + E$

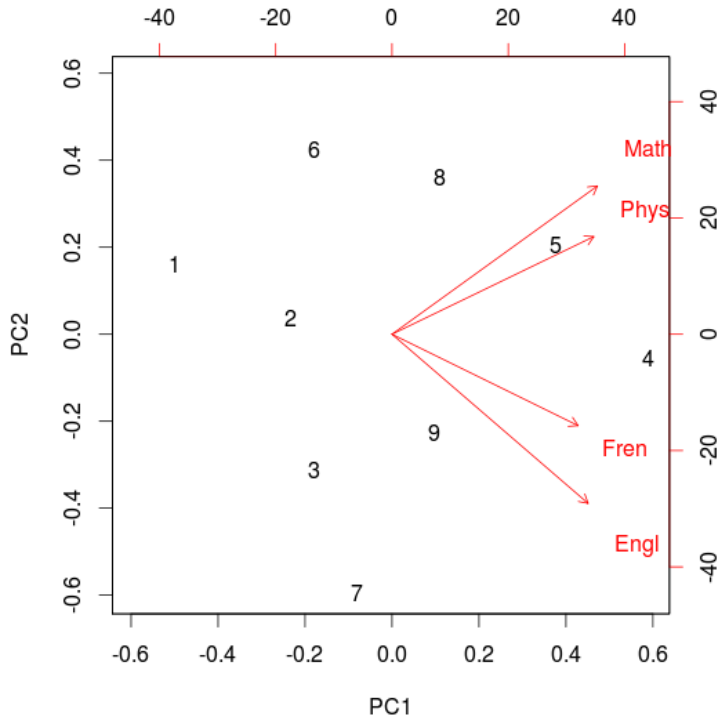


-

At the end of the day, PCA is used to (see next slide):

- Reduce the dimension of a data set
- Exhibits patterns/dependencies in high-dimensional data sets
- Represent high-dimensional data
- Bonus: detect outliers.

Studying variables and/or individuals



Note: We could have done the analysis by interpreting linear combinations of individuals who would have had contributions to the axes to represent the variables; this is equivalent !

What's next ?

Practical session !