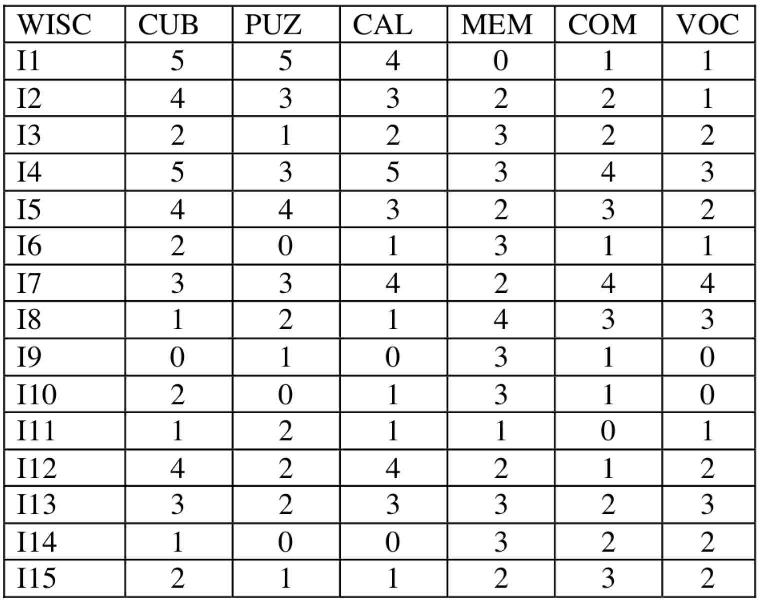
Data Mining & Big Data

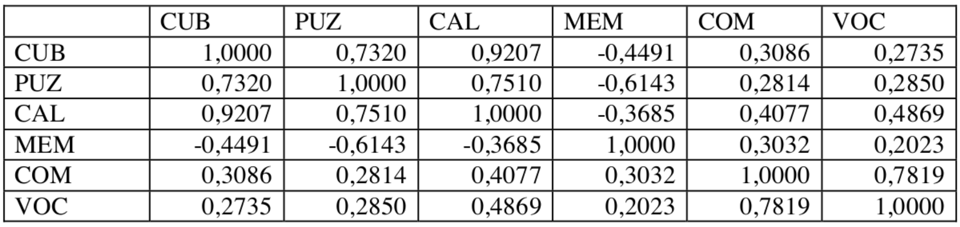
Exercise 1 PCA: the WISC dataset

The following dataset corresponds to the results of a psychological test called WISC applied on some 10 years old children [1]. Observed variable are CUB for Khos block, PUZ for object assembly, CAL for mental arithmetic, MEM for number memory, COM for sentences understanding, VOC for vocabulary.



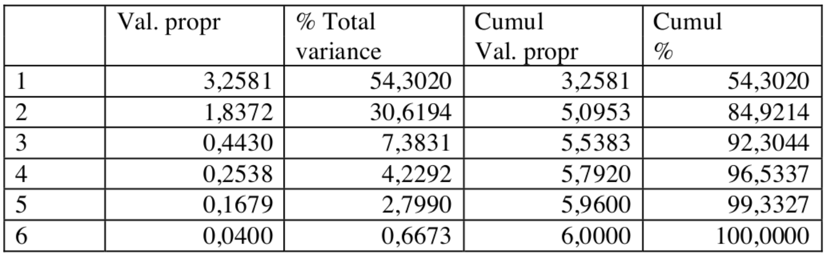
WISC	CUB	PUZ	CAL	MEM	COM	VOC
I1	5	5	4	0	1	1
I2	4	3	3	2	2	1
I3	2	1	2	3	2	2
I4	5	3	5	3	4	3
I5	4	4	3	2	3	2
I6	2	0	1	3	1	1
I7	3	3	4	2	4	4
I8	1	2	1	4	3	3
I9	0	1	0	3	1	0
I10	2	0	1	3	1	0
I11	1	2	1	1	0	1
I12	4	2	4	2	1	2
I13	3	2	3	3	2	3
I14	1	0	0	3	2	2
I15	2	1	1	2	3	2

Data are processing with a PCA. The next figures represent the results obtained.



	CUB	PUZ	CAL	MEM	COM	VOC
CUB	1,0000	0,7320	0,9207	-0,4491	0,3086	0,2735
PUZ	0,7320	1,0000	0,7510	-0,6143	0,2814	0,2850
CAL	0,9207	0,7510	1,0000	-0,3685	0,4077	0,4869
MEM	-0,4491	-0,6143	-0,3685	1,0000	0,3032	0,2023
COM	0,3086	0,2814	0,4077	0,3032	1,0000	0,7819
VOC	0,2735	0,2850	0,4869	0,2023	0,7819	1,0000

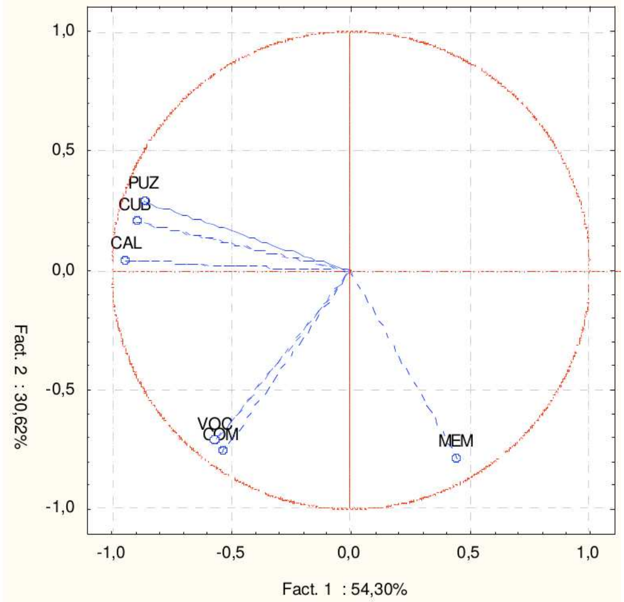
correlations associated to WISC dataset



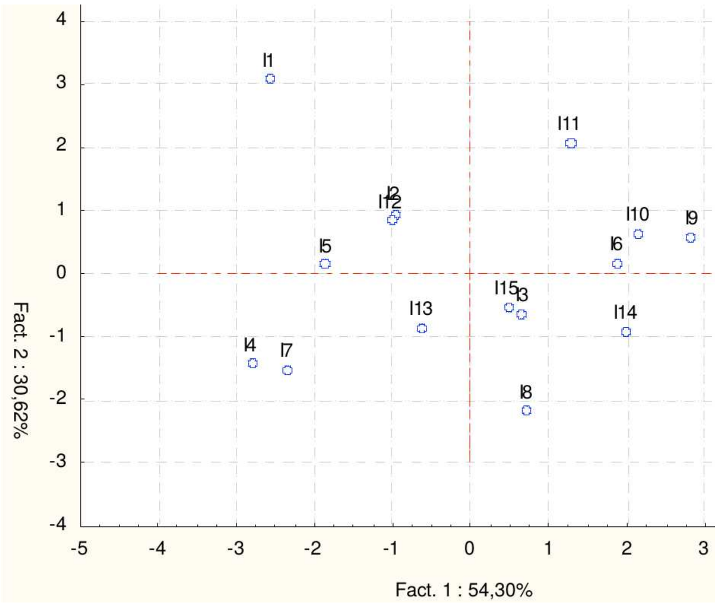
	Val. propr	% Total variance	Cumul Val. propr	Cumul %
1	3,2581	54,3020	3,2581	54,3020
2	1,8372	30,6194	5,0953	84,9214
3	0,4430	7,3831	5,5383	92,3044
4	0,2538	4,2292	5,7920	96,5337
5	0,1679	2,7990	5,9600	99,3327
6	0,0400	0,6673	6,0000	100,0000

percentage of information obtained with PCA

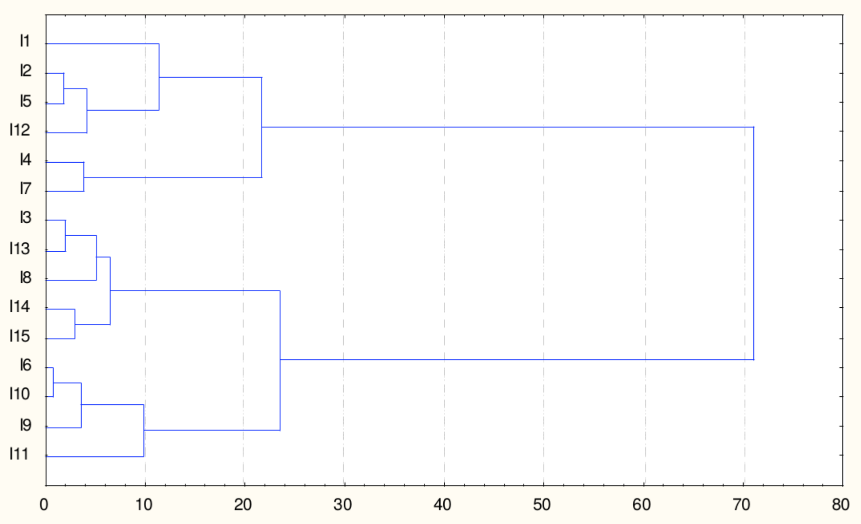
[1] The exercise is coming from http://geai.univ-brest.fr/carpentier/



correlation circle



components



1. Which variables are the most correlated ? Interpret negative and positive values.
2. Only the two first components are keeping for the rest of the study. Justify this choice.
3. The results show that all the variables are well represented. Justify it.
4. Which are the variables correlated positively in the first component ? Which are the variables correlated negatively ? What is the interpretation ?
5. Which are the variables linked with the second component ?
6. Check the results of your interpretations with the points cloud figure. Detail your reasonnig with I1, I8 and I9.
7. An individu is placed in (0,0). What would be the interpretation ?
8. The last figure is obtained after an algorithm on the projected dataset. Which method does it corresponds to ?
9. How many clusters would you keep ?
10. We finally keep 4 clusters. Describe in with classes are each individu.
11. Evaluate the position of clusters of the point cloud.