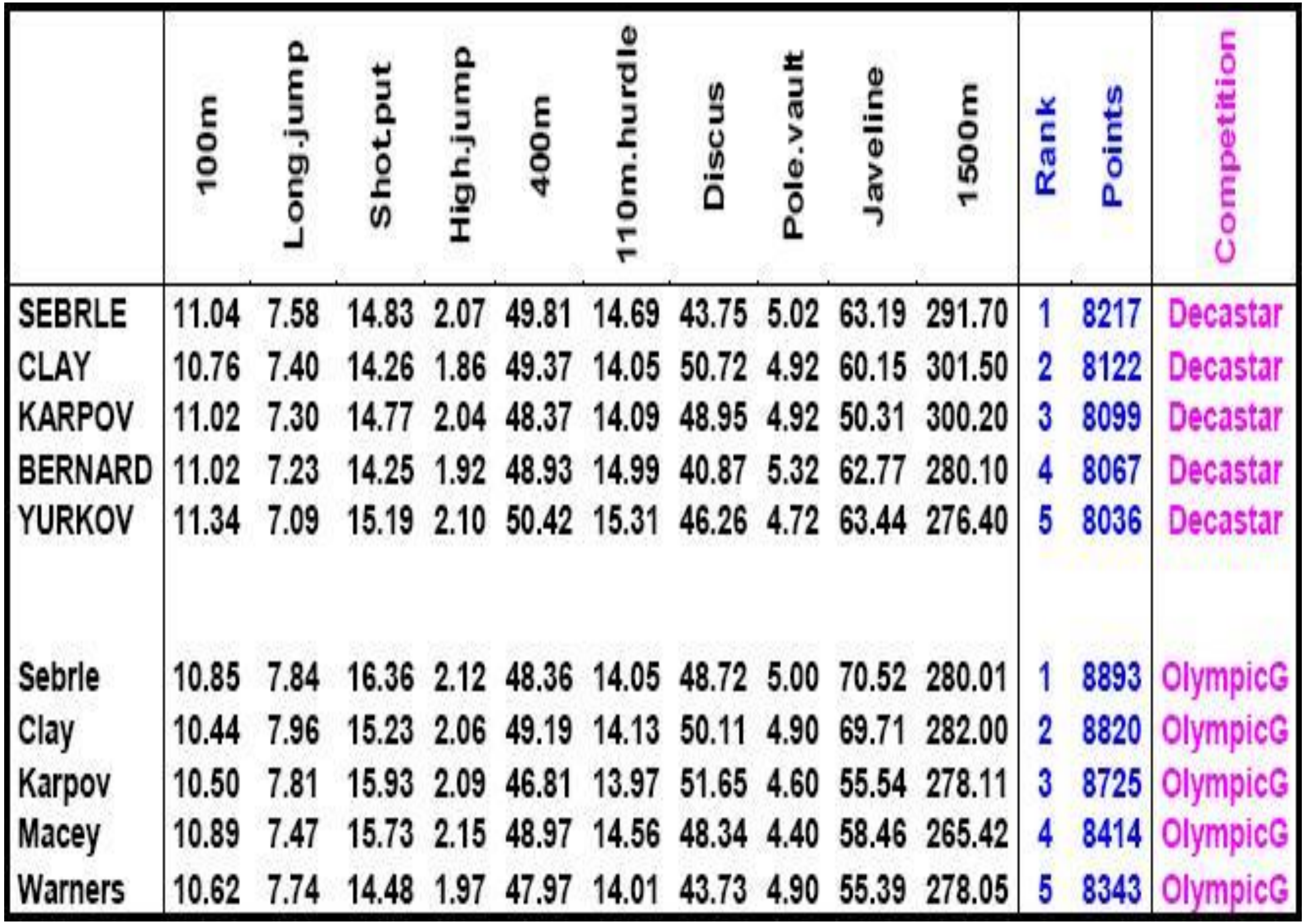
DECATHLON DATASET



	100m	Long.jump	Shot.put	High.jump	400m	110m.hurdle	Discus	Pole.vault	Javeline	1500m	Rank	Points	Competition
SEBRLE	11.04	7.58	14.83	2.07	49.81	14.69	43.75	5.02	63.19	291.70	1	8217	Decastar
CLAY	10.76	7.40	14.26	1.86	49.37	14.05	50.72	4.92	60.15	301.50	2	8122	Decastar
KARPOV	11.02	7.30	14.77	2.04	48.37	14.09	48.95	4.92	50.31	300.20	3	8099	Decastar
BERNARD	11.02	7.23	14.25	1.92	48.93	14.99	40.87	5.32	62.77	280.10	4	8067	Decastar
YURKOV	11.34	7.09	15.19	2.10	50.42	15.31	46.26	4.72	63.44	276.40	5	8036	Decastar
Sebrle	10.85	7.84	16.36	2.12	48.36	14.05	48.72	5.00	70.52	280.01	1	8893	OlympicG
Clay	10.44	7.96	15.23	2.06	49.19	14.13	50.11	4.90	69.71	282.00	2	8820	OlympicG
Karpov	10.50	7.81	15.93	2.09	46.81	13.97	51.65	4.60	55.54	278.11	3	8725	OlympicG
Macey	10.89	7.47	15.73	2.15	48.97	14.56	48.34	4.40	58.46	265.42	4	8414	OlympicG
Warners	10.62	7.74	14.48	1.97	47.97	14.01	43.73	4.90	55.39	278.05	5	8343	OlympicG

PCA → Applications

PCA to → describe a dataset, to summarize a dataset, to reduce the dimensionality.
In this example (Decathlon)

1.Individuals' study (athletes' study): two athletes will be close to each other if their results to the events are close. We want to see the variability between the individuals. Are there similarities between individuals for all the variables? Can we establish different profiles of individuals? Can we oppose a group of individuals to another one?
2.Variables' study (performances' study): We want to see if there are linear relationships between variables. The two objectives are to summarize the correlation matrix and to look for synthetic variables: can we resume the performance of an athlete by a small number of variables?
3.Link between this two studies: can we characterize groups of individuals by variables?
4.Dimensionality reduction to perform prediction or clustering tasks (IMPORTANT)

Taken together, the main purpose of principal component analysis is to:

- identify hidden pattern in a data set,
- reduce the dimensionality of the data by removing the noise and redundancy in the data,
- identify correlated variables

PCA

Athlete's profiles according to their performances only. The active variables will be only those which concern the ten events of the decathlon.

The other variables (*"Rank"*, *"Points"* and *"Competition"*) do not belong to this athletes' profiles and use an information already given by the other variables (in the case of *"Rank"* and *"Points"*) but it is interesting to confront them to the principal components as supplementary variables.

Here the variables are not measured in the same units. We must scale them in order to give the same influence for each one.

Value

Returns a list including:

eig
a matrix containing all the eigenvalues, the percentage of variance and the cumulative percentage of variance

var
a list of matrices containing all the results for the active variables (coordinates, correlation between variables and axes, square cosine, contributions)

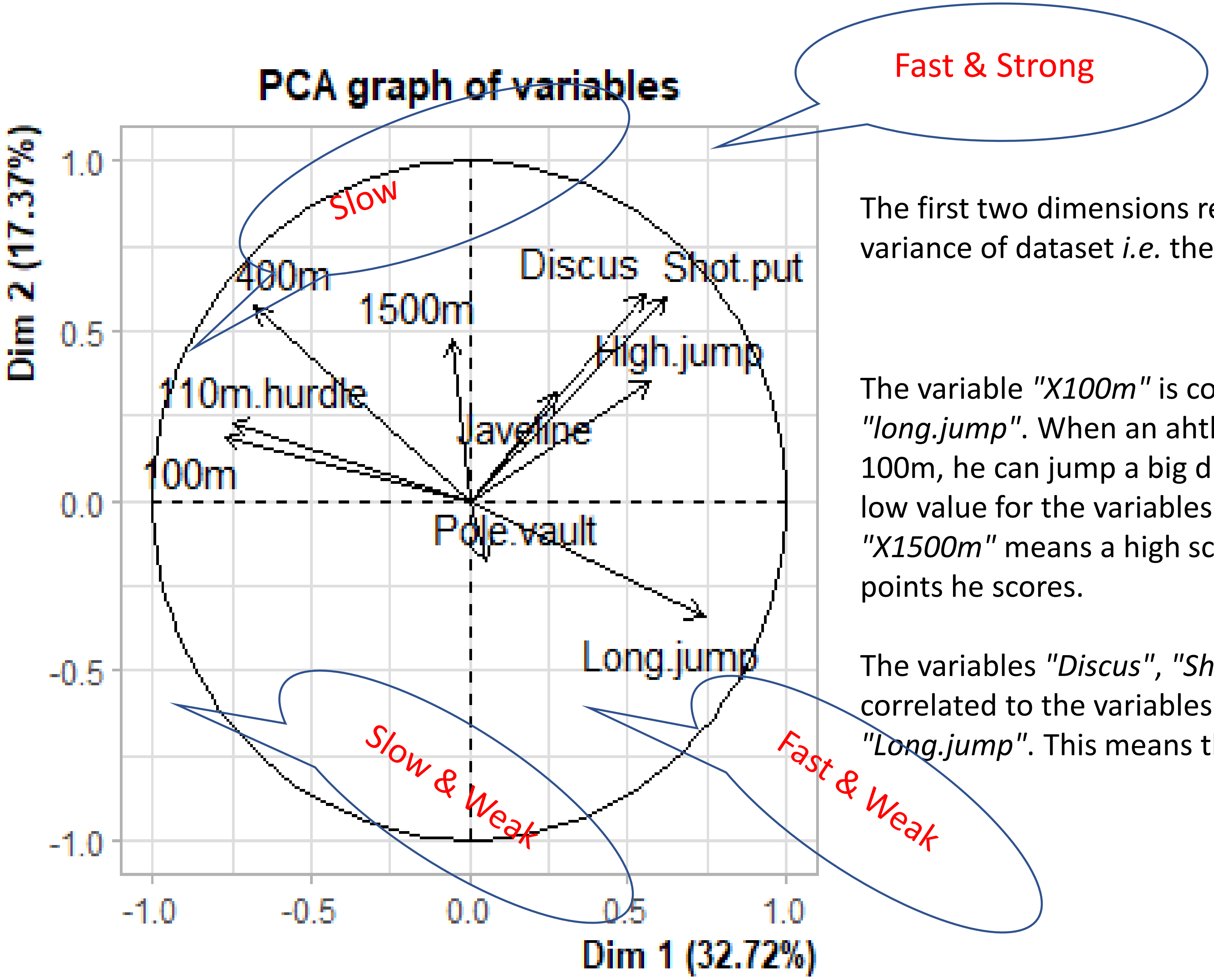
ind
a list of matrices containing all the results for the active individuals (coordinates, square cosine, contributions)

ind.sup
a list of matrices containing all the results for the supplementary individuals (coordinates, square cosine)

quanti.sup
a list of matrices containing all the results for the supplementary quantitative variables (coordinates, correlation between variables and axes)

quali.sup
a list of matrices containing all the results for the supplementary categorical variables (coordinates of each categories of each variables, v.test which is a criterion with a Normal distribution, and eta2 which is the square correlation corefficient between a qualitative variable and a dimension)

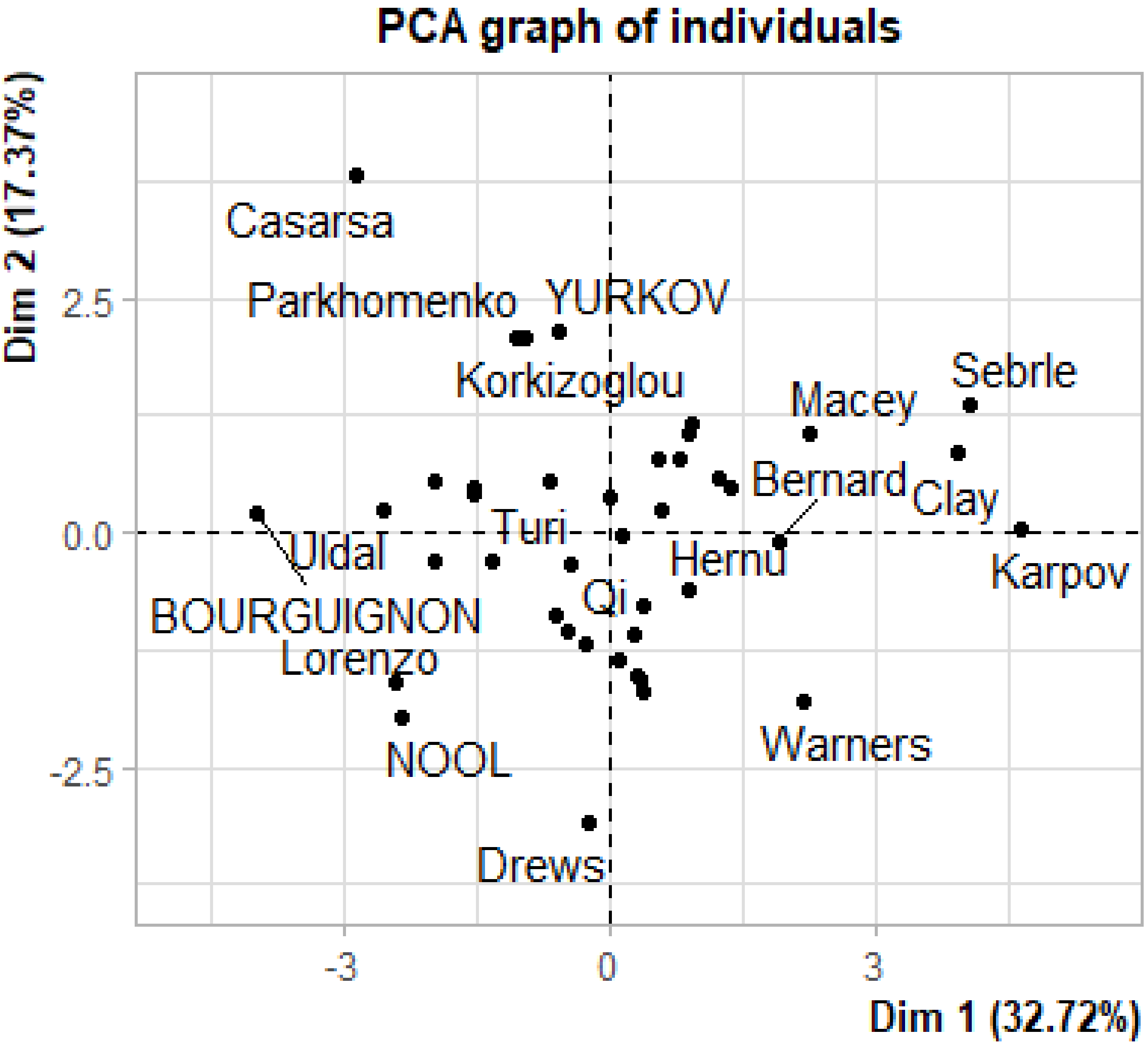
Returns the individuals factor map and the variables factor map.
The plots may be improved using the argument autolab, modifying the size of the labels or selecting some elements thanks to **the plot.PCA function.**



The first two dimensions resume 50% of the total inertia (the inertia is the total variance of dataset *i.e.* the trace of the correlation matrix).

The variable *"X100m"* is correlated negatively to the variable *"long.jump"*. When an ahtlete performs a short time when running 100m, he can jump a big distance. Here one has to be careful because a low value for the variables *"X100m"*, *"X400m"*, *"X110m.hurdle"* and *"X1500m"* means a high score: the shorter an athlete runs, the more points he scores.

The variables *"Discus"*, *"Shot.put"* and *"High.jump"* are not much correlated to the variables *"X100m"*, *"X400m"*, *"X110m.hurdle"* and *"Long.jump"*. This means that strength is not much correlated to speed.



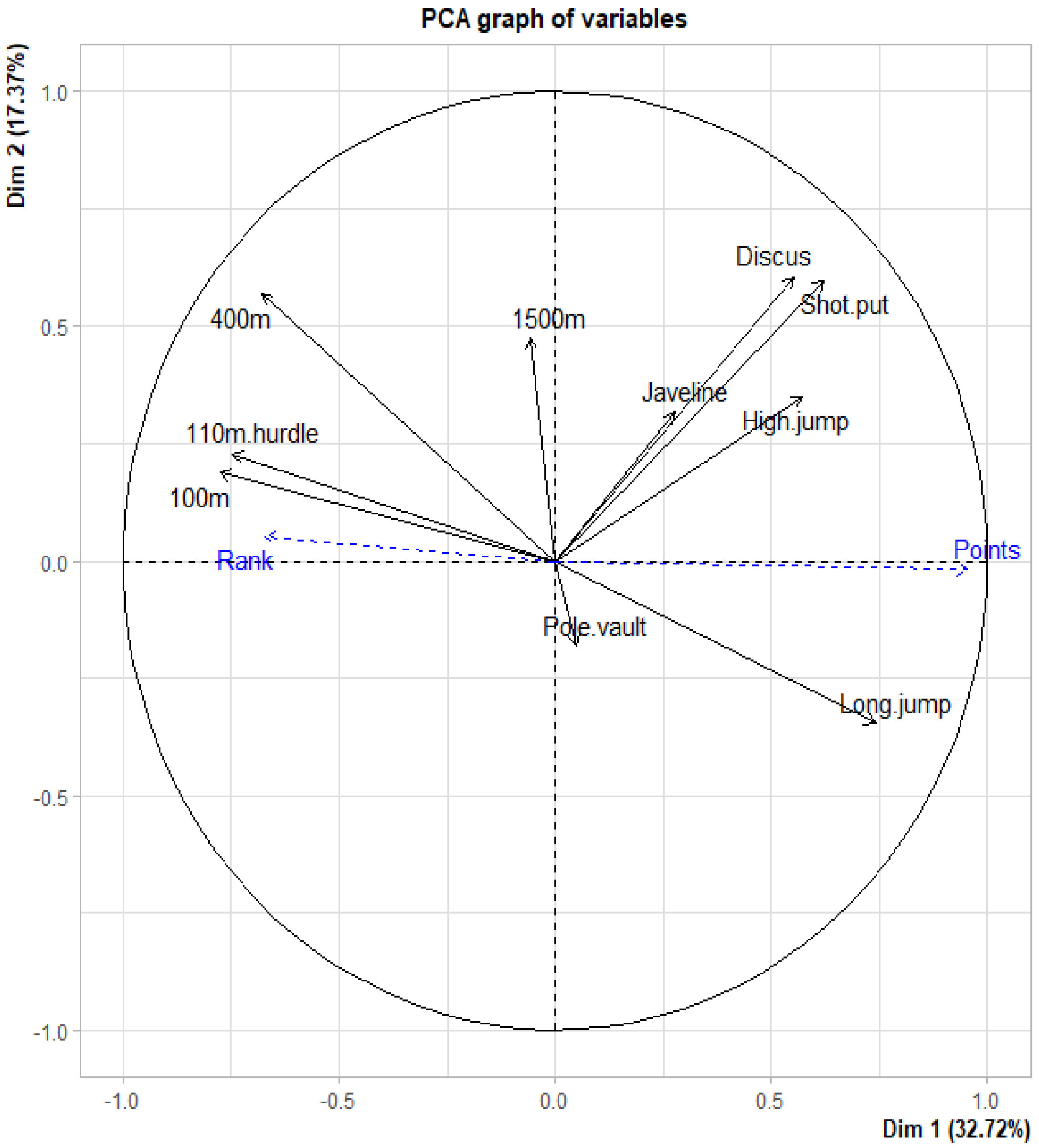
Sebrle???

Casarsa??

Nool???

Warners???

KARPOV ?? BOURGUIGNON?



The winners of the decathlon are those who scored the most (or those whose rank is low).

The variables the most linked to the number of points are the variables which refer to the speed (*"X100m"*, *"X110m.hurdle"*, *"X400m"*) and the long jump. On the contrary, *"Pole-vault"* and *"X1500m"* do not have a big influence on the number of points. Athletes who are strong for these two events are not favored.