The University of Franche-Comté

Master CompuPhys - Statistical exploitation of measurements

Application of principal component analysis (PCA) to the analysis of sports results

Chiari Even
Legrand Maxime
M1 Compuphys 2022

Supervisor : Fabrice Devaux

December 9, 2022

Contents

1 Introduction

The main objective of this last practical work is to highlight the correlation between trial scores of players during the Tokyo 2021 decathlon by applying the Karhunen-Loève Transform (KLT) on a set of data which will be introduced below. In a first part we'll process the data to filter out its main three components, before analyzing it by plotting the truncated data.

2 Theory

The truncated matrix is obtained by only considering the highest components of the 10-dimensional space formed by the 10 decathlon event scores (N variables) and the 13 contender rank in each event (K measures). In our case, we want to reduce the dimensionality from N = 10 to L = 3.

In order to do so, one can proceed by either using the covariance matrix or the correlation one. In our case, going through the covariance matrix is enough as the data isn't highly spread out and that all events have the same weight. It can be done by projecting the Covariance matrix onto a orthogonal space and then by building the truncated matrix back. The said covariance matrix is defined by :

$$C = \frac{M_c M_c^{\mathbf{T}}}{K-1} \tag{1}$$

With :

- M_C = $M - \overline{M}$ the centered matrix of data

The rotation matrix between these two spaces is given by taking the transpose of the Eigenvector matrix $\mathbf{\Phi}$ associated to $\mathbf{C}$, which allows us to write the KLT $\mathbf{Y}$ as :

$$Y = M_C \Phi^{\mathbf{T}} \tag{2}$$

By isolating the $L = 3$ highest components in this space, one can construct the truncated matrix with its minimum error with the following relation :

$$M_{trunc} = \overline{M} + Z\Psi \tag{3}$$

With :

- Z the **K**x**L** columns of Y
- Ψ the **L**x**N** rows of Φ

3 Analysis of data

Now we can proceed to the comparison of M_{trunc} with our initial matrix the following relation :

$$M_{comp} = \frac{M - M_{trunc}}{M} \tag{4}$$

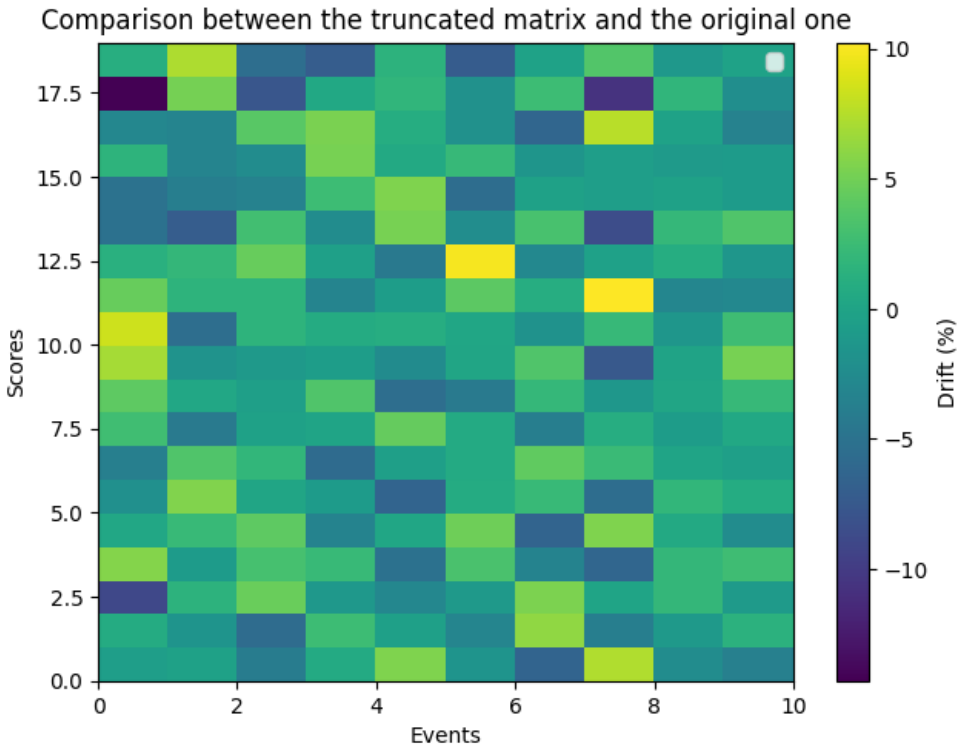

Figure 1: Comparison between Normal and data density distributions

One can notice that the drift between them is weak (ie. most scores are really close to the original ones).

We can compute the ratio of information lost with this relation :

$$Q = \frac{\sum_{n=L+1}^{N} \lambda_n}{\sum_{n=1}^{N} \lambda_n} \tag{5}$$

In our case we have :

$$Q \approx 0.25\% \tag{6}$$

We can accept this fraction of information loss due to its weakness. Now we want study the correlation coefficient matrix. In order to to this, we have plotted two graphs :

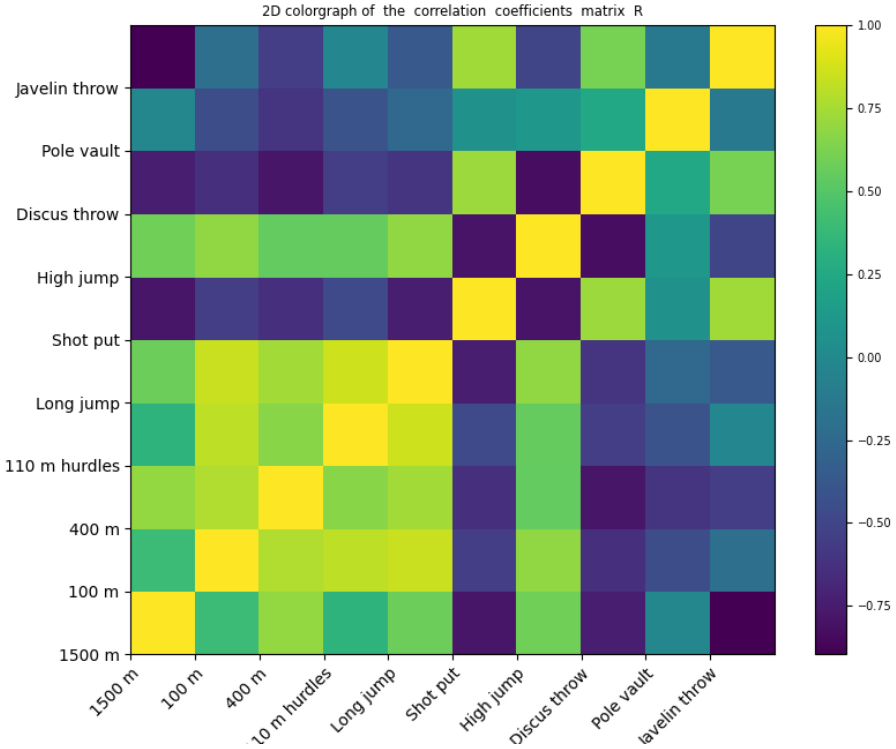

Figure 2: Color map of the correlation coefficient matrix

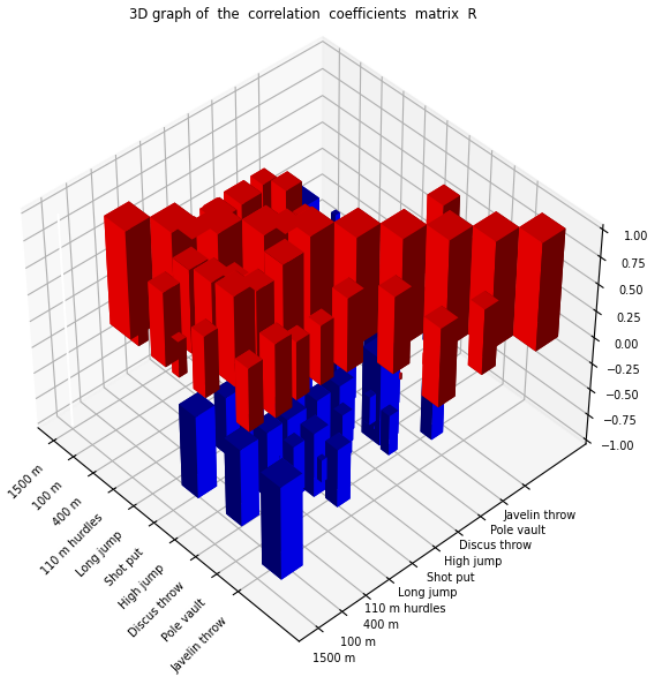

Figure 3: 3D histogram of the correlation coefficient matrix

On this two graphs we can observe the prominence of the diagonal and this is logical because all the diagonal coefficients represent the correlation of a sport with itself.

We can deduce many things from these two graphs. For example, the "shot put" sport is not very correlated with the others, it even is opposed with certain sports. Indeed, on the 3D graph, we can see the other bars are under 0, it means they are negatively correlated (if we you are good in this sport you are bad in the other). Another example could be the "1500m", it looks like correlated with many other sports of the same nature (running).

For the 3D graph we have scaled the width of the bar with the value of the correlated coefficients. It means the more the bar is large, the more it is correlated or uncorrelated with other sports.

We can run the python file to observe the 3D graph from many points of view to see correlations between the sports.

We also studied two other representations of our truncated data, we have computed the plot of two 3D graphs: one with the components of Z and the other with the components of Ψ:

For the graph of Z :

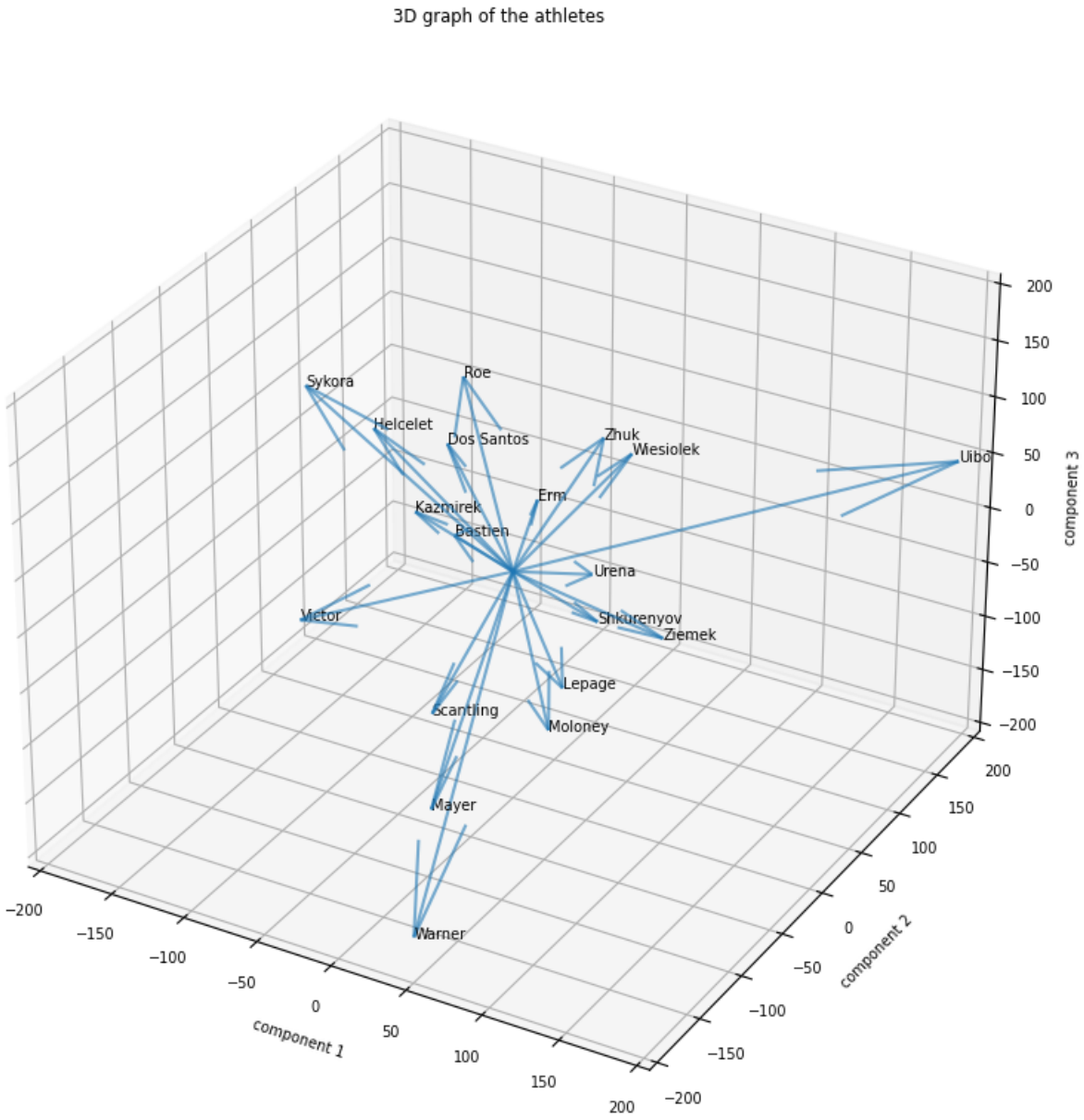

Figure 4: 3D histogram of the Z matrix

We can see on the graph the name of each athlete and their main components. We can see some athletes are opposed with others, meaning they are opposed in terms of results in the different sports. More generally, it shows the trend of an athlete to be good or not in a contest.

We can run the python file to observe the 3D graph from many points of view to see which component is the most important for each athlete.

For the graph Ψ

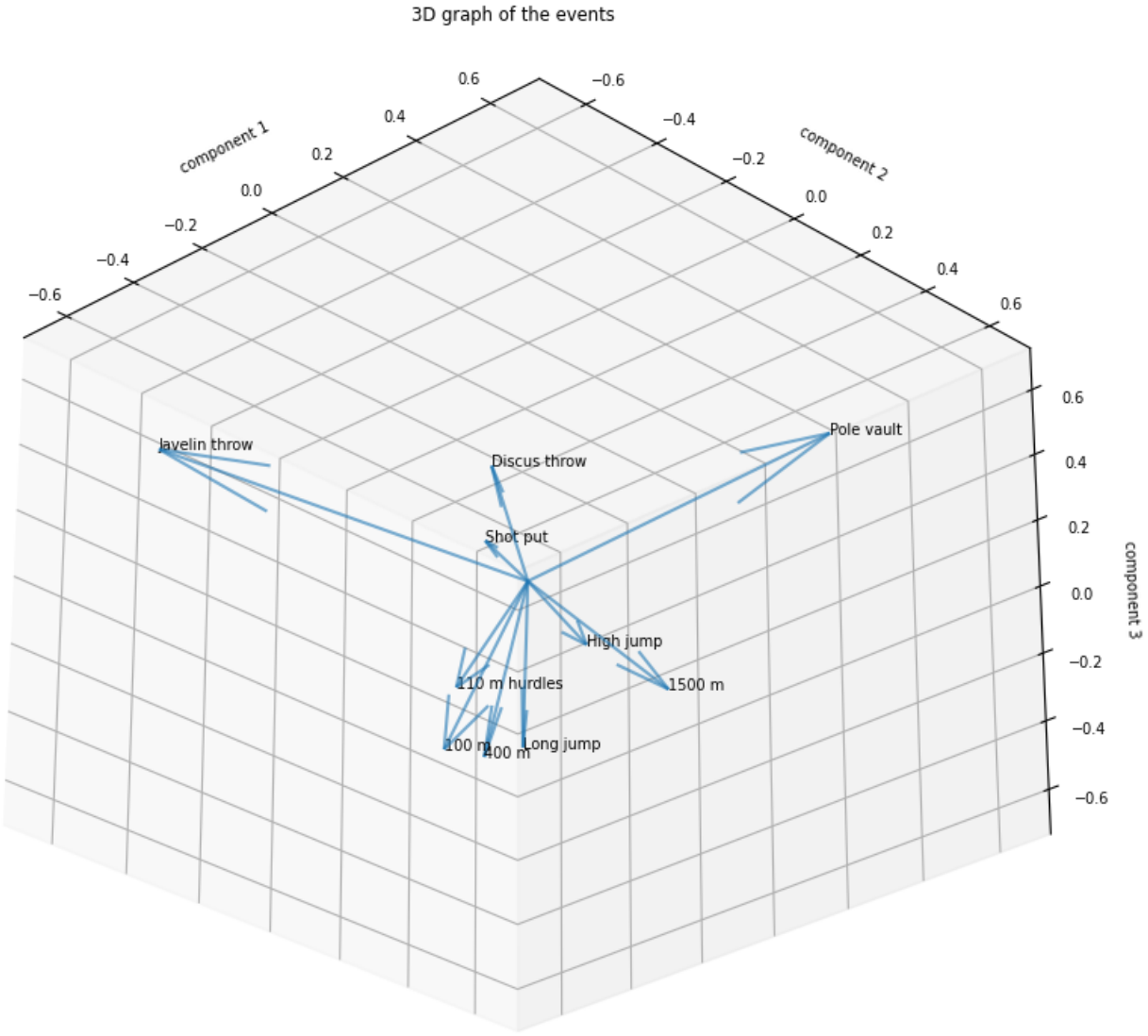

Figure 5: 3D histogram of the Ψ matrix

We can see on the graph each sport in the OG decathlon and their main components. We can see some very isolated sports in term of components compared to others. For example, the "shot put" instead of running sports that seems to have close value for their components.

We can run the python file to observe the 3D graph from many point of views to see which component is the most important for each sport and phenomena of group running sport and the isolated "shot put". The more the arrows are close to each other on an axis, the more they share a common behavior.

4 Conclusion

Mixing up the Principal Component Analysis with the Karhunen-Loeve transform gives powerful results when it comes to analyzing a set of data. Indeed, we saw that applying the PCA method on the data makes it interpretable with a minimum of information loss. Coupling it with the KLT method to get the highest correlations thus allowed us to identify the similarities between sports as well as between contenders.

We could expect that the contenders level at running sports makes all these sports strongly correlated, and we had also interesting characteristics such as a contender being good at Shot Put tends to be bad at running sports.