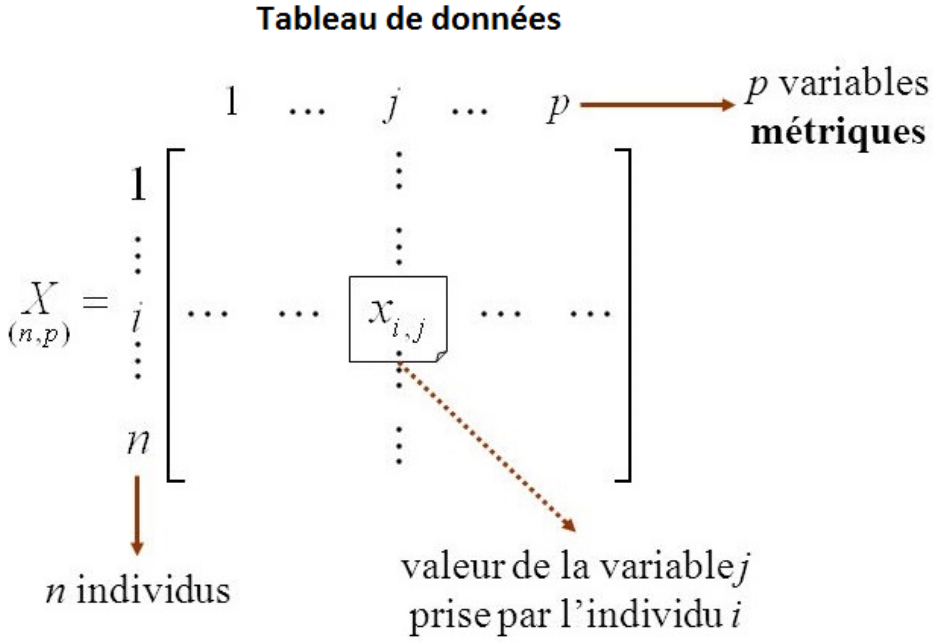
1 La Théorie

1.0.1 Tableau de données

L'Analyse en Composante Pricipale s'interesse à l'etude des tableaux de données rectangulaires dont les colonnes sont appelées *Variables* et les lignes sont appelées *Individus*.

L'image ci-dessous illustre un tableau de donnees a *p-Variables* et a *n-Individus*. Un individu est noté $X_i = (x_{i1},x_{i2},...,x_{ij},...,x_{ip})$ où $i = 1..n$.



1.0.2 Moyenne et Variance

Pour une *variable* x_k, on note sa moyenne $\bar{x_k}$ et sa variance σ_k telles que:

$$\bar{x_k} = \frac{1}{n}\sum_{i=1}^{n} x_{ik} \qquad \sigma_k = \frac{1}{n}\sum_{i=1}^{n}(x_{ik} - \bar{x_k})^2$$

1.0.3 Distance et Centre de Gravité

On define la distance d entre deux *individus* i et j par:

$$d^2(i,j) = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2$$

2 La pratique

2.1 Les données

Nous allons étudier les résultats des épreuves de Decastar et des Jeux Olympiques, dont les *Individus* sont les joueurs et les *Variables* sont les jeux eux même, soient:

- 100m, pour la course de vitesse 100m
- Longueur, pour le saut en longueur
- Poids, pour le lancé de poids
- Hauteur, pour le saut en hauteur
- 400m, pour la course de 400m
- 110m H, pour la course de vitesse 110m Homme
- Disque, pour le lancé de disque
- Perche, pour le lancé de perche
- Javelot, pour le lancé de javelot
- 1500m, pour la course d'endurance 1500m
- Classement, pour la classement finale de chaque joueur
- Points, pour le point finale obtenu par chaque joueur
- Competition, pour le type de compétition

Voici une partie des données utilisées:

	100m	Longueur	Poids	Hauteur	400m	110m H	Disque	Perche	Javelot	1500m	Classement	Points	Competition
Sebrle	10.85	7.84	16.36	2.12	48.36	14.05	48.72	5.00	70.52	280.01	1	8893	JO
Clay	10.44	7.96	15.23	2.06	49.19	14.13	50.11	4.90	69.71	282.00	2	8820	JO
Karpov	10.50	7.81	15.93	2.09	46.81	13.97	51.65	4.60	55.54	278.11	3	8725	JO
Macey	10.89	7.47	15.73	2.15	48.97	14.56	48.34	4.40	58.46	265.42	4	8414	JO
Warners	10.62	7.74	14.48	1.97	47.97	14.01	43.73	4.90	55.39	278.05	5	8343	JO
Zsivoczky	10.91	7.14	15.31	2.12	49.40	14.95	45.62	4.70	63.45	269.54	6	8287	JO
Hernu	10.97	7.19	14.65	2.03	48.73	14.25	44.72	4.80	57.76	264.35	7	8237	JO
Nool	10.80	7.53	14.26	1.88	48.81	14.80	42.05	5.40	61.33	276.33	8	8235	JO
Bernard	10.69	7.48	14.80	2.12	49.13	14.17	44.75	4.40	55.27	276.31	9	8225	JO
Schwarzl	10.98	7.49	14.01	1.94	49.76	14.25	42.43	5.10	56.32	273.56	10	8102	JO

On voit bien ici que les *Variables* ne sont pas de même unité, alors on va d'abord les normalisées. On peut confirmer que les *Variables* sont bien normalisées quand elles sont asymptotiquement de moyenne nulle et de variance égale a un.

```
#Moyenne de Z
print(Z.mean())
```

```
-9.694142507702587e-16
```

```
#Variance de Z
print(Z.var())
```

```
1.0
```

D'après ces résultats, les moyennes et les variances de nos *Variables Normalisées* tendent bien vers 0 et 1 respectivement.

A partir d'ici, on va travailler avec les **Données Normalisées**

2.2 Les variables explicatives

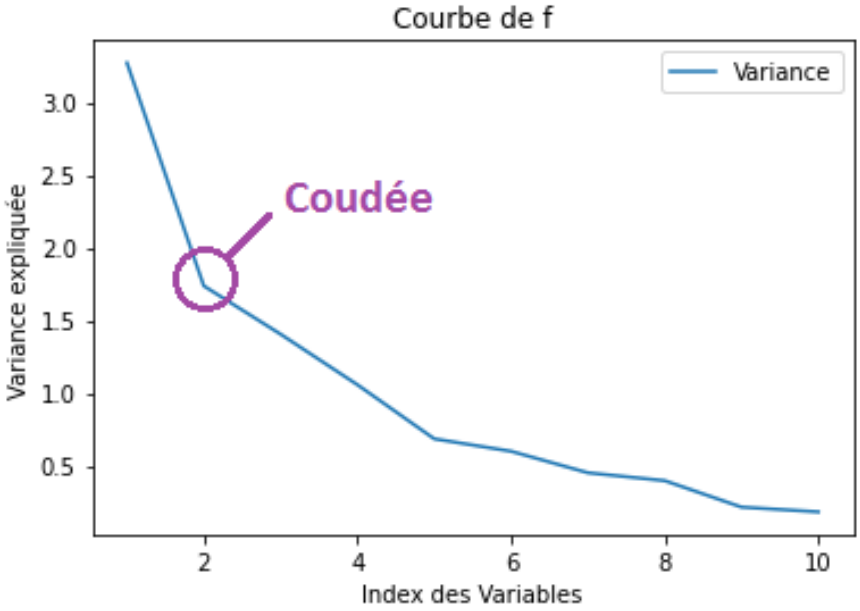
Parfois, il y a des *Variables* dont l'importance est néglisable par rapport aux autres, il est important de les reconnaitre ou même de les éliminer pour avoir une meilleur analyse des données. Les *Variables* restantes issue de cette élimination seront appelées *Variables Explicatives*.

Il y a plusieurs méthodes pour retrouver les variables à éliminer, mais pour notre problème, on va utiliser la méthode du coudé, d'après Wikipedia - Méthode du coude (clustering)

La méthode du coude est une heuristique utilisée pour déterminer le nombre de clusters (Variables dans notre cas) dans un ensemble de données . La méthode consiste à tracer la **variation expliquée** *en fonction du nombre de clusters, et à choisir le coude de la courbe comme le nombre de clusters à utiliser. La même méthode peut être utilisée pour choisir le nombre de paramètres dans d'autres modèles basés sur les données, comme le nombre de composants principaux pour décrire un ensemble de données*

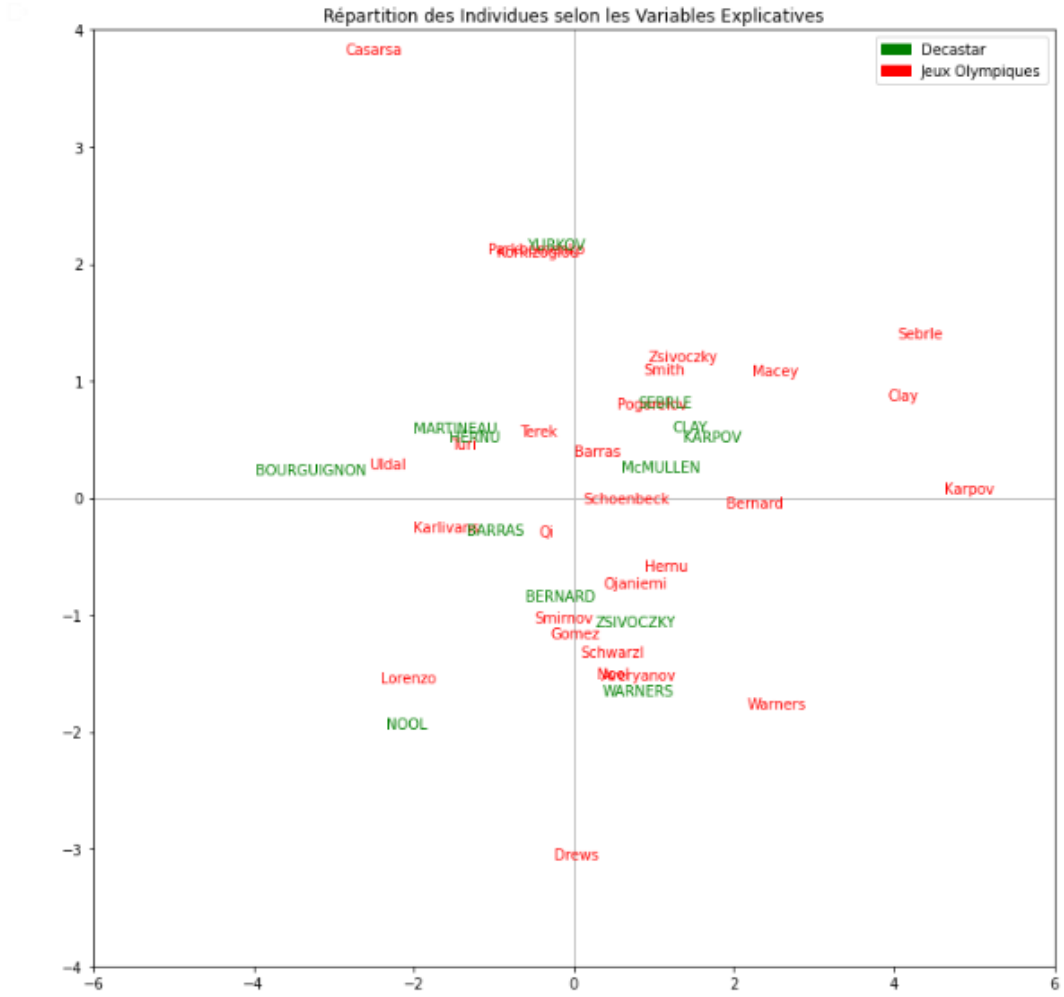
Pour ce faire, on a donc besoin des variation expliquée de chaque *Variable* de notre données normalisées et de trace la courbe de la fonction f définie par:

$$f(V_e) = \frac{(n-1)}{n} * V_e \quad , V_e \textit{ est la variance explicative.}$$



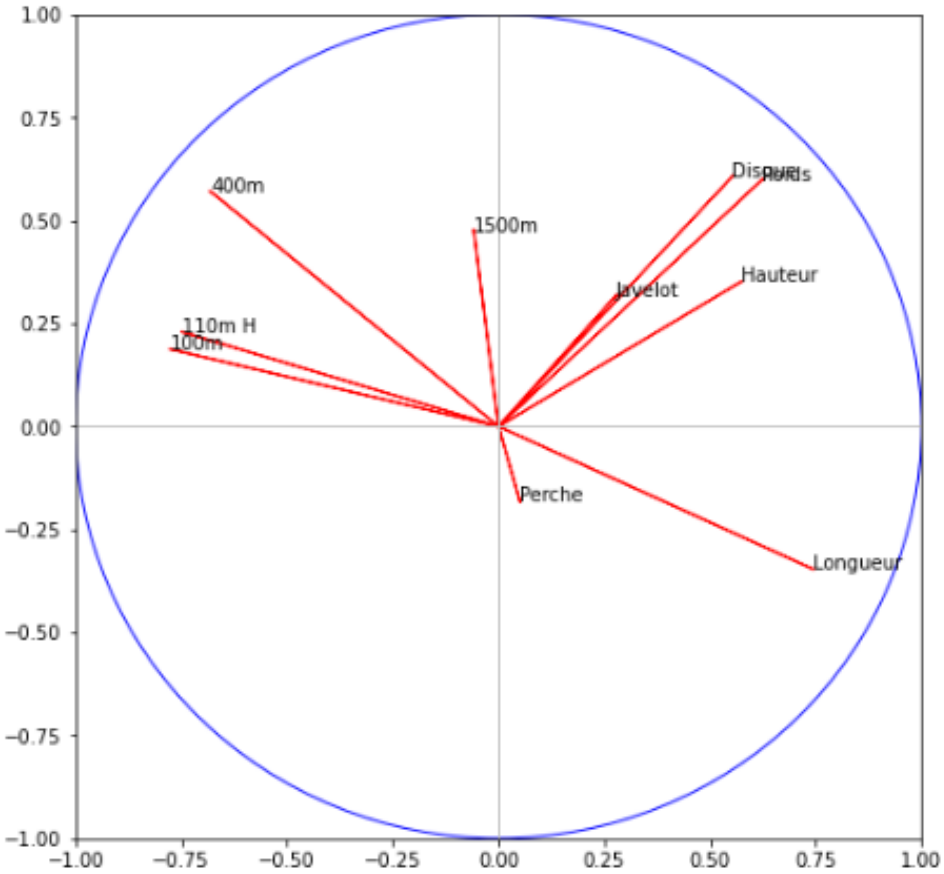
D'après la graphe de la fonction - variance expliquée - on ne va donc retenir que les deux premières *Variables* qui sont: la *Variable 100m* et la *Variables Longueur*.

A partir de ces deux *Variables*, on peut positionner les individus dans le plan. On a alors la figure suivante:



2.2.1 Interpretation

Premier Axe: On peut remarquer du premier axe qu'elle departage les joueurs ayant un bon classement a ceux qui ont un mauvais classement pendant les deux competitions. En effet, les joueurs **Karpov** et **Clay** qui sont le deuxieme et le troisieme dans le classement des Jeux Olympiques, et des Jeux de Decastar sont placés à droite dans la premier axe alors que le joueur **Uldal**, avant dernier du classement des JO et le joueur **Bourguignon**, dernier du classement de Decastar sont placés a gauche dans la premier axe.
Deuxieme Axe: On peut remarquer du deuxieme axe qu'elle departage les joueurs


1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
400m
1500m
Hauteur
Javelot
110m H
100m
Perche
Longueur