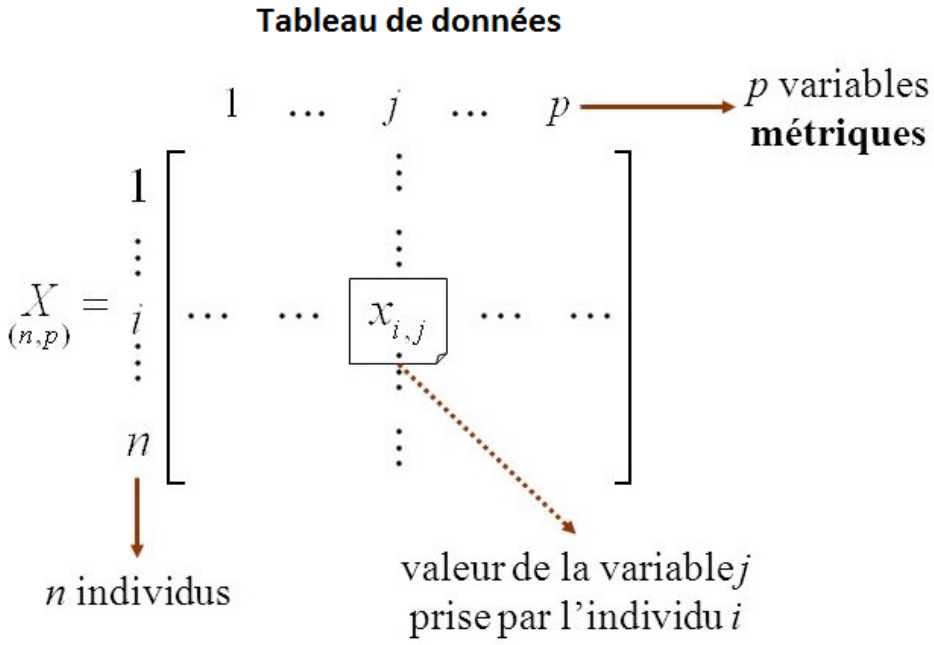
1 La Théorie

1.1 Tableau de données

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

L'image ci-dessous illustre un tableau de donnees a *n-individus* et a *p-variables*. Un *individu* est noté $i = (x_{i1},x_{i2},...,x_{ip}) \in \mathbb{R}^p$ et une *variable* est notée $x_j = (x_{1j},x_{2j},...,x_{nj}) \in \mathbb{R}^n$

NB: Il est important de souligner que les *variables* sont **quantitatives** en ACP, c'est à dire que $x_{ij} \in \mathbb{R}$, $\forall (i,j) \in [n] \times [p]$.



1.2 Moyenne et Variance

Pour une *variable* x_j, on note sa moyenne $\bar{x}_j$ et son écart-type s_j telles que:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \qquad s_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$$

1.3 Distance et Point moyen

En ACP, il est important de savoir quels *individus* sont proches les un des autres, pour pouvoir, si possible, creer des groupes d'*individus* selon leur proximité. Il est alors necessaire de definir une distance entre deux *individus* i et i', soit d cette distance:

$$d^2(i,i') = \sum_{j=1}^{p}(x_{ij} - x_{i'j})^2$$

De plus, à partir de la moyenne de chaque *variables*, on construit un point noté G et appelé **point moyen** du nuage des *individus*, ce point peut etre interpreté comme le centre de gravité du nuage des *individus*:

$$G = (\bar{x}_1,\bar{x}_2,...,\bar{x}_p)$$

1.4 Centrage et Réduction

Ayant le point moyen, qui peut être vue comme le centre de gravité du nuage, il est conseillé de translater notre nuage de telle façon que G soit confondue avec le centre du repère, cela ce traduit par: "Replacer x_j par $x_{ij} - \bar{x}_j$ dans le tableau de données";

Cette processus s'appelle **centrage** et elle améliore grandement l'affichage et l'interprétation des graphiques de nos données.

$$x_{ij} \leftarrow x_{ij} - \bar{x}_j \quad \forall i \in [n], \forall j \in [p],$$

Parfois, les *variables* ne sont du même unité, ceci peut entrainer une ambiguïté dans l'interprétation des données, en effet, une variable a une "quantité" plus petite, dans le tableau, lorsqu'elle est exprimé en mettre, que lorsqu'elle est exprimé en centimètre. Ce qui peut entrainer, respectivement aux unités, une faible ou une forte importance par rapport aux autres variables qui faussera alors l'analyse.

Un moyen efficace d'eviter ce probleme est la **reduction**, elle consiste a diviser les *variables* par leur écart-type.

Comme on a convenu plutot qu'il est preferable de centrer les *variables*, alors le **centrage** et la **reduction** se traduit par: "Replacer x_j par $(x_{ij} - \bar{x}_j)/s_j$ dans le tableau de données";

$$x_{ij} \leftarrow \frac{x_{ij} - \bar{x}_j}{s_j} \quad \forall i \in [n], \forall j \in [p]$$

NB: Apres le centrage et la reduction, nos nouvelles *variables* sont alors "asymptotiquement" de moyenne nulle et de variance egale a un. c'est a dire:

$$\bar{x}_j \approx 0 \quad s_j^2 \approx 1 \quad \forall j \in [p]$$

1.5 Inertie

Par definition, l'inertie I des données est:

$$I = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(\frac{x_{ij} - \bar{x}_j}{s_j} \right)^2$$

C'est donc, au coefficient 1/n pres, la somme des carrés de toutes les cellules du tableau de données apres centrage et reduction, mais elle peut egalement etre interpretée par rapport aux *individus* et aux *variables*.

Inertie et Individus

Soit l'individu $i \in [n]$, la quantité $\sum_{j=1}^{p} \left(\frac{x_{ij}-\bar{x}_j}{s_j}\right)^2$ de l'inertie represente la distance au carrée entre cet individue et le point moyen G. Par conséquent, l'inertie peut etre vue comme la somme des carres des distances au centre de gravite pour tous les individus.

Ainsi, l'inertie renseigne sur la "forme" du nuage des individus. En effet, plus la distance entre les individus sont grande (resp. petite), plus l'inertie est grande (resp. petite).

Inertie et Variables

Dans la definition de l'inertie, les deux somme $\sum$ peuvent etre interverti, ainsi, on a une autre expression:

$$I = \frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n} \left(\frac{x_{ij} - \bar{x}_j}{s_j}\right)^2$$

Ici, on peut remarquer que la quantité $\sum_{i=1}^{n} \left(\frac{x_{ij}-\bar{x}_j}{s_j}\right)^2$ correspond au carré de la norme de la variable centrée reduite x_j ou $j \in [p]$. Or cette quantité est egale à n, ainsi, l'inertie est toujour egale au nombre de variables, en effet:

$$\begin{aligned}
I &= \frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n} \left(\frac{x_{ij} - \bar{x}_j}{s_j}\right)^2 \\
&= \frac{1}{n} \sum_{j=1}^{p} \left(\frac{1}{s_j^2}\, n s_j^2\right) \\
&= \frac{1}{n} \sum_{j=1}^{p} \left(\frac{1}{s_j^2}\, n s_j^2\right) \\
&= \frac{1}{n} \sum_{j=1}^{p} n \\
&= \frac{1}{n}\, pn \\
&= p
\end{aligned}$$

1.6 Représentation Simplifiée

On a vu que les *variables* sont des vecteurs de $\mathbb{R}^p$, alors, quand $p > 3$, il n'est plus possible de representer les variables. L'ACP vise a fournir une image simplifiée du nuage des *individus* la plus fidele possible, cest a dire, trouver une sous-espace de dimension plus petite qui resume au mieux les données.

Comment retrouver la meilleur sous-espace?

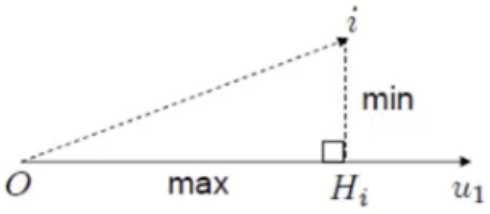
Pensons pour cela a l'image d'un chameau, la figure ci-dessous propose deux representations simplifiees de cette image: des representations en dimension 2, la vue de face et la vue de profil.

Il est evident de dire que la meilleure representation simplifiee est la vue de profil. La raison est que l'image projetee du chameau dans ce plan est plus proche de l'image initiale dans le sens ou la variabilite des points qui la representent est plus grande et donc restitue mieux la variabilitee des points d'origine en dimension 3.

On a alors les etapes suivantes pour retouver analytiquement la meilleure representation simplifiee du nuage des *individus*:

Etape 1: **Trouver l'axe qui deforme le moins possible le nuage**

On cherche un axe dans $\mathbb{R}^p$ de sorte que les distances entre les points initiaux i (*individu*) soient les plus proches possibles de leurs projetes orthogonaux sur cette axe et cela en tenant compte de tous les autres points. Notons $\vec{u}_1$ la direction de cet axe, et H_i la projetée orthogonale de i.



Le but est de minimiser la distance iH_i, ce qui revient a maximiser la distance OH_i

Plus formellement, on cherche la direction $\vec{u}_1$ de $\mathbb{R}^p$ telle que $\sum_{i=1}^{n} OH_i^2$ soit maximum.

On dira qu'on cherche $\vec{u}_1$ telle que l'inertie projetee I_p est maximum.

Etape 2: **Trouver le meilleur plan**

Cette fois ci, on cherche le meilleur plan $\mathcal{P}$ ou l'inertie projetee I_p est maximal sur ce plan, pour ce faire, on cherche une direction $\vec{u}_1$ qui maximise I_p (fait à l'etape 1), puis on cherche une autre direction $\vec{u}_2$ orthogonale a $\vec{u}_1$ et qui, elle aussi, maximise I_p.

Le plan meilleur plan $\mathcal{P}$ est formé par les deux meilleurs axes de directions respectives $\vec{u}_1$ et $\vec{u}_2$, orthogonals l'un a l'autre.

Etape 3: **Trouver un troisieme meilleur axe**

On peut chercher un troisieme axe et sequentielement, chercher les axes les un apres les autres et a chaque fois, un axe doit etre orthogonal aux axes precedents, et maximise l'inertie projetee I_p.

1.7 Correlation

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

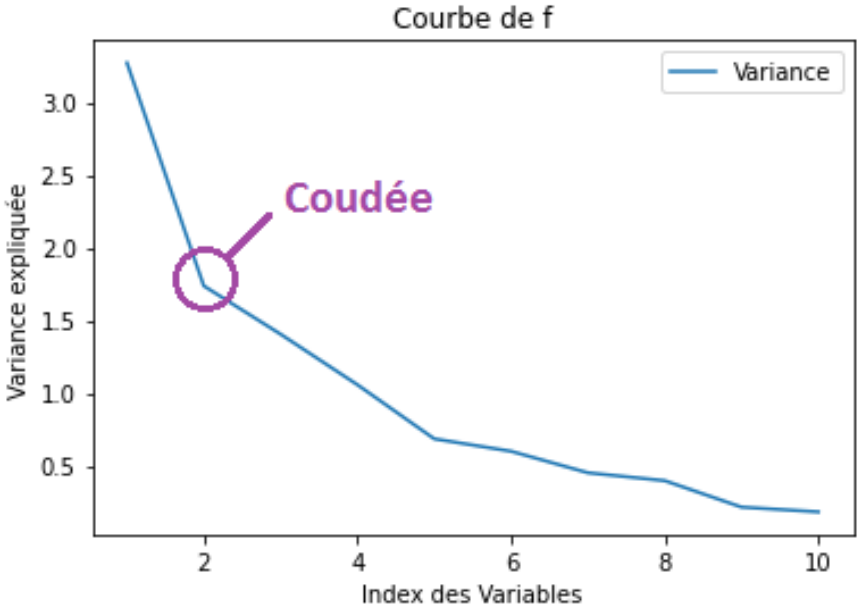
Parfois, il y a des *Variables* dont l'importance est néglisable par rapport aux autres, il est important de les reconnaitre ou même de les éliminer pour avoir une meilleur analyse des données. Les *Variables* restantes issue de cette élimination seront appelées *Variables Explicatives*.

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*

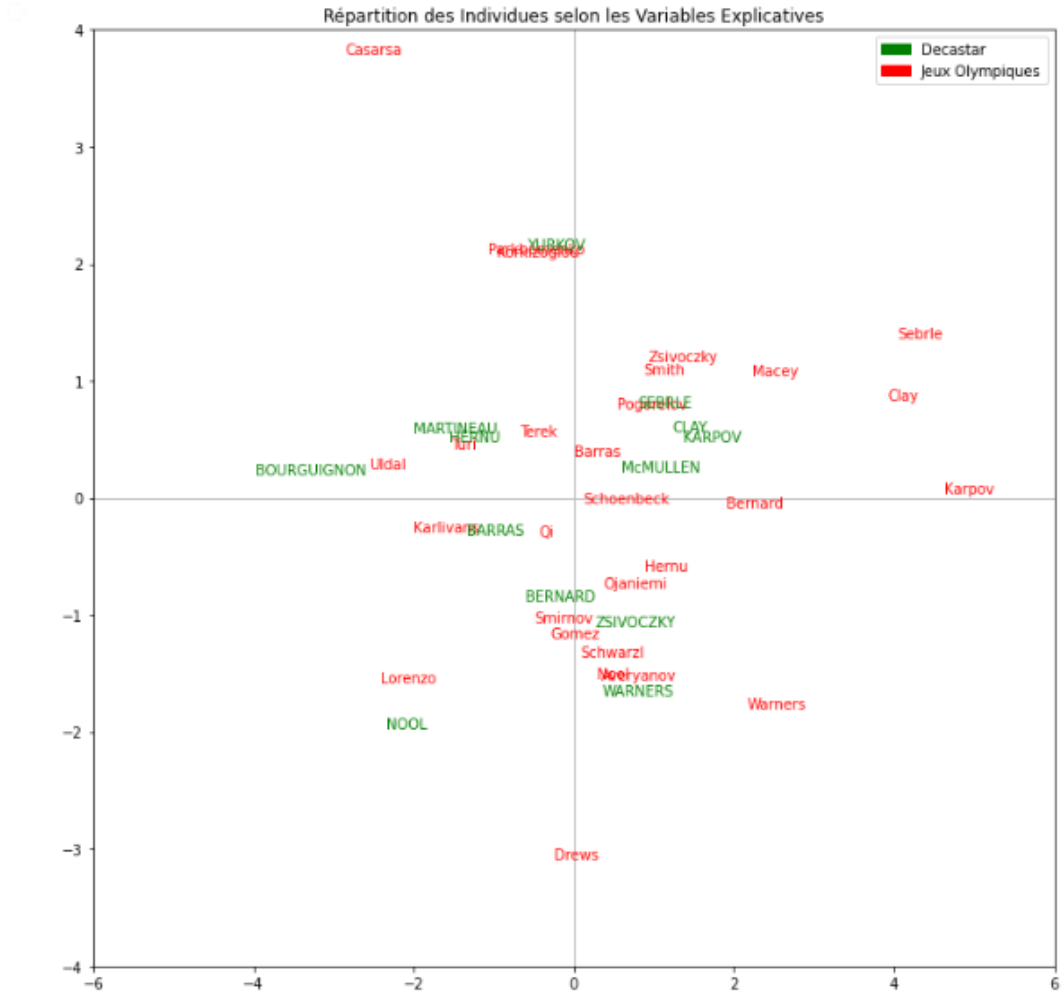
Pour ce faire, on a donc besoin des variation expliquée de chaque *Variable* de notre données normalisées et de trace la courbe de la fonction f définie par:

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



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:

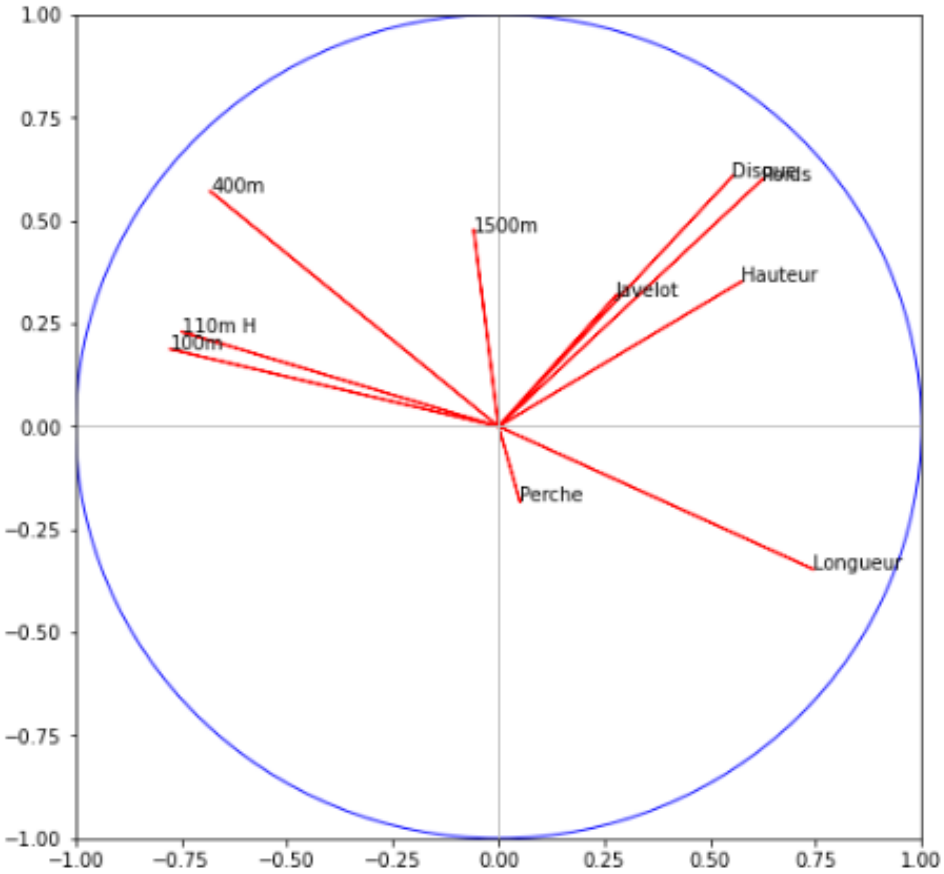


Interpretation du premier axe

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.

Interpretation du 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