UNIVERSITE D'ANTANANARIVO

FACULTE DES SCIENCES

DEPARTEMENT MATHEMATIQUES

ET INFORMATIQUE

Analyse de Données Multidimensionnelles
Théorie et Pratique

Élaboré par:
ANDRIAMANANA Hajaniaina Rivo Hery
RAMANANTSOA Zo Samoela Reno

Mathématiques Appliquées
Spécialité: Combinatoire et Optimisation

Année universitaire: 2018-2019

Table des matières

1 Introduction

Les méthodes d'analyse de données ont commencées a être développées dans les années 50, poussées par le développement de l'informatique et du stockage des données qui depuis n'a cesse de croitre. L'analyse de données a surtout été développée en France par J.P. Benzecri [Ben80a], [Ben80b] qui a su par l'analyse des correspondances représenter les données de manière simple et interprétante.

Aujourd'hui les méthodes d'analyse de données sont employées dans un grand nombre de domaines qu'il est impossible d'énumérer, a l'instar du marketing pour l'étude des marches ou de l'étude des données météorologiques.

Les analyses factorielles constituent la plupart des analyses de données. Elles sont fondées sur un principe unique, c'est pour cela que nous pouvons parler de l'analyse factorielle [EP90]. Ce principe repose sur le fait que les deux nuages de points représentant respectivement les lignes et les colonnes du tableau étudié (tableau 1.1) sont construits et représentés sur des graphiques. Ces représentations des lignes et des colonnes fortement liées entre elles permettent une analyse plus aisée pour l'opérateur. Parmi les divers méthodes d'analyse factorielle, on peut citer l'analyse en composantes principales, l'analyse factorielle des correspondances, l'analyse des correspondances multiples et l'analyse factorielle discriminante.

Dans ce pressent projet, nous nous intéresserons surtout a l'analyse en composantes principales que nous notons par la suite ACP. C'est une des premières analyses factorielles, et certainement aujourd'hui l'une des plus employées. Dans [LMP95], nous trouvons l'historique de cette méthode qui fut conçue par Karl Pearson en 1901. Elle est sans doute a la base de la compréhension actuelle des analyses factorielles.

Son utilisation a cependant été plus tardive avec l'essor des capacités de calculs. Les principales variantes de l'ACP viennent des différences de transformations du tableau de données. Ainsi, le nuage de points peut être centre ou non, réduit ou non. Le cas le plus étudié, et que nous présentons ici, est lorsque le nuage de point est centre et réduit ; dans ce cas nous parlons d'ACP normée. D'autres variantes existent telle que l'analyse en composante curviligne [DH97] pour remédier au fait que les projections sont linéaires, ou encore l'analyse en composantes indépendantes pour la séparation de sources [Pha96].

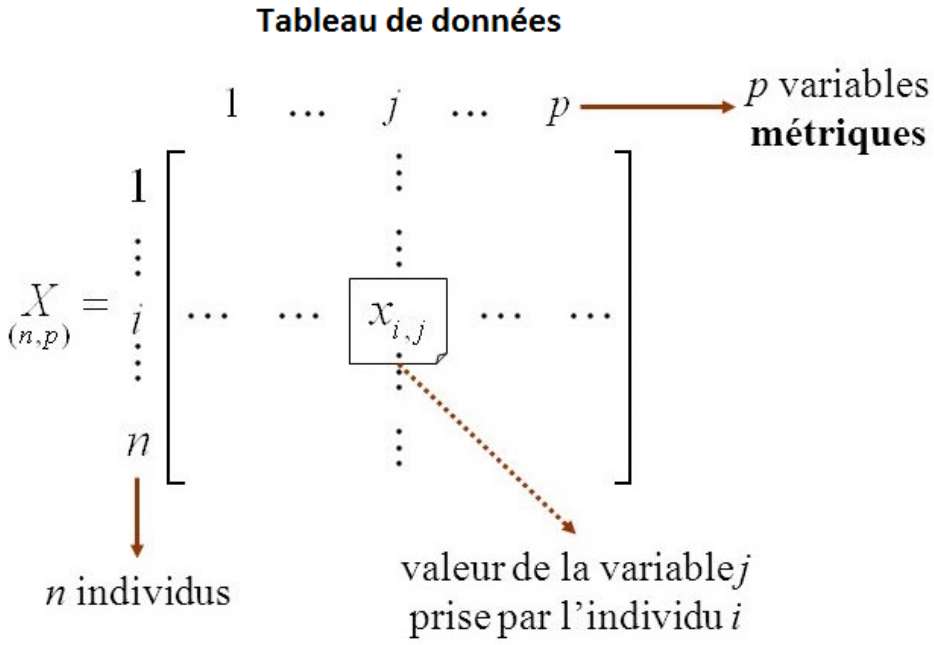
2 Théorie

2.1 Tableau de données

L'**A**nalyse en **C**omposante **P**ricipale s'intéresse à l'étude 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 données 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 $x_{ij} \in \mathbb{R}$, $\forall (i,j) \in [n] \times [p]$.



2.2 Moyenne et Variance

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

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

2.3 Distance et Point moyen

En ACP, il est important de savoir quels *individus* sont proches les un des autres, pour pouvoir, si possible, créer des groupes d'*individus* selon leur proximité. Il est alors nécessaire de définir 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 être interprété comme le centre de gravité du nuage des *individus*:

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

2.4 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'éviter ce problème est la **réduction**, elle consiste a diviser les *variables* par leur écart-type.

Comme on a convenu plutôt qu'il est préférable de centrer les *variables*, alors le **centrage** et la **réduction** 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: Après le centrage et la réduction, nos nouvelles *variables* sont alors "asymptotiquement" de moyenne nulle et de variance égale a un. c'est a dire:

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

2.5 Inertie

Par définition, 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 près, la somme des carrés de toutes les cellules du tableau de données après centrage et réduction, mais elle peut également être interprétée 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 individu et le point moyen G. Par conséquent, l'inertie peut être 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 définition de l'inertie, les deux somme $\sum$ peuvent être 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 réduite x_j ou $j \in [p]$. Or cette quantité est égale à n, ainsi, l'inertie est toujours égale au nombre de variables, en effet:

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

2.6 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 représenter les variables. L'ACP vise a fournir une image simplifiée du nuage des *individus* la plus fidèle possible, c'est a dire, trouver une sous-espace de dimension plus petite qui résume au mieux les données.

Comment retrouver la meilleur sous-espace?

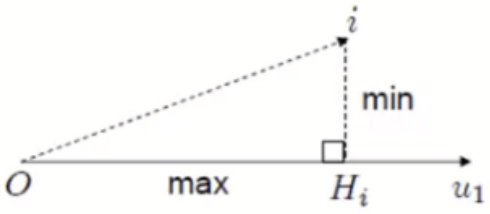
Pensons pour cela a l'image d'un chameau, la figure ci-dessous propose deux représentations simplifiées de cette image: des représentations en dimension 2, la vue de face et la vue de profil.

Il est évident de dire que la meilleure représentation simplifiée est la vue de profil. La raison est que l'image projetée du chameau dans ce plan est plus proche de l'image initiale dans le sens ou la variabilité des points qui la représentent est plus grande et donc restitue mieux la variabilité des points d'origine en dimension 3.

On a alors les étapes suivantes pour retrouver analytiquement la meilleure représentation simplifiée du nuage des *individus*:

Étape 1: **Trouver l'axe qui déforme 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 projetés 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 projetée I_p est maximum.

Étape 2: **Trouver le meilleur plan**

Cette fois ci, on cherche le meilleur plan $\mathcal{P}$ ou l'inertie projetée I_p est maximale sur ce plan, pour ce faire, on cherche une direction $\vec{u}_1$ qui maximise I_p (fait à l'étape 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 $\vec{u}_2$, orthogonales l'un a l'autre.

Étape 3: **Trouver un troisième meilleur axe**

On peut chercher un troisième axe et essentiellement, chercher les axes les un après les autres et a chaque fois, un axe doit être orthogonal aux axes précédents, et maximise l'inertie projetée I_p.

NB: A l'issu de ces étapes, on appel les axes de direction $\vec{u}_j$,$j \in [p]$, **Composantes Principales**

3 Pratique de l'ACP

3.1 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

3.2 Centrage et Reduction

On voit bien que les *variables* ne sont pas de même unité, alors en plus du centrage, on va aussi réduire les variables. Après avoir fait cette étape dans un logiciel statistique. On veut confirmer que les *variables* sont bien centrées et réduites, c'est a dire qu'elles sont asymptotiquement de moyenne nulle et de variance égale a un. Ainsi on a la figure ci-dessous:

```
#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.

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

3.3 Les axes a retenir

Un moyen de trouver les **composantes principales** ou **meilleurs axes** est de calculer les *valeurs propres* de la matrice de covariances des *variables*. En ordonnant ces *valeurs propres* par ordre décroissant, les *vecteurs propres* associes deviennent les directions $\vec{u}_j, j \in [p]$ des **composantes principales**.

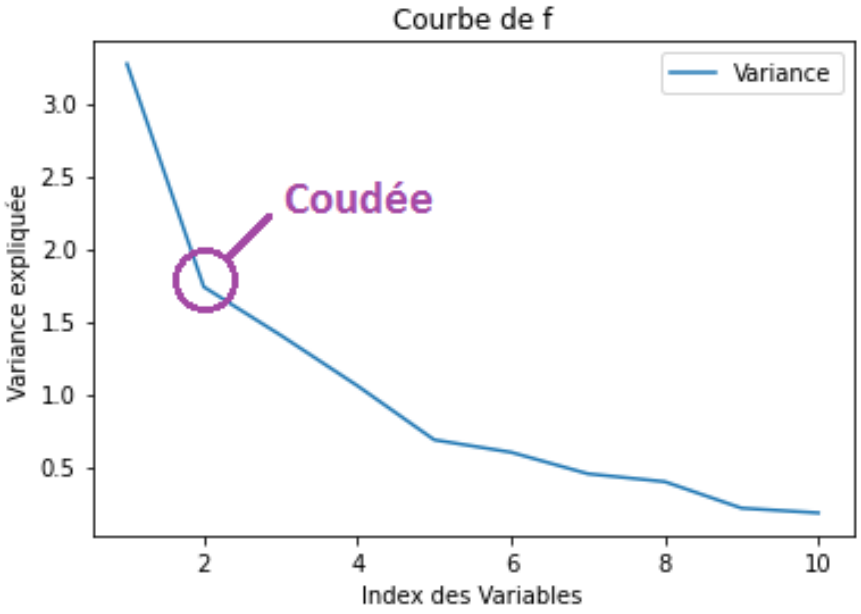
La quantité d'informations portée par un axe est alors proportionnelle a la *valeur propre* associée, ainsi, un nombre $k < p \in \mathbb{N}^*$ d'axes suffit pour représenter les individus. En effet, il y a des axes dans leur *valeur propre* est neglisable par rapport a celles des autres.

Il y a plusieurs méthodes pour retrouver les axes 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 (axes 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*

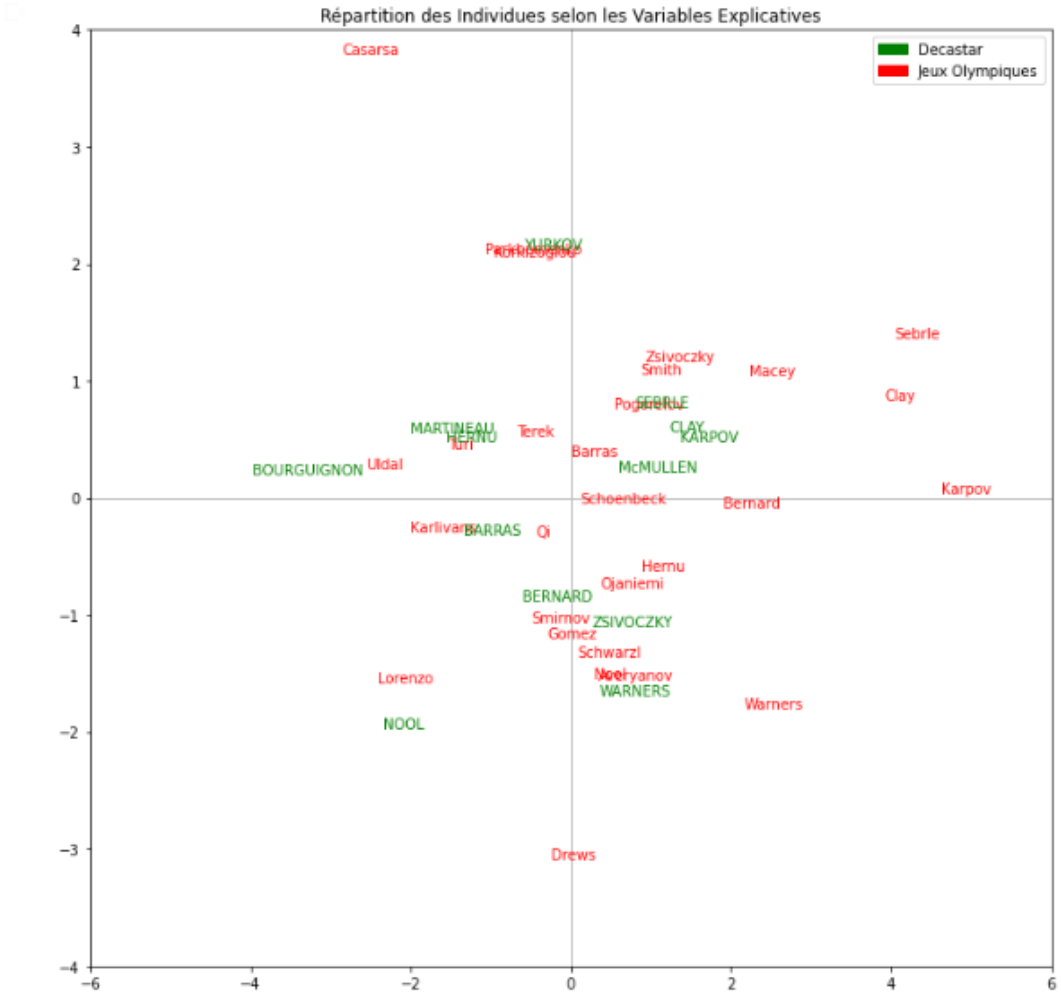
Pour ce faire, on a donc besoin des variation expliquée, qui sont les *valeurs propres* associées à nos axes et de trace la courbe de la fonction f définie par:

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



D'après la graphe de la fonction - variance expliquée - on ne va donc retenir que les deux premières **composantes principales** suivant l'ordre décroissante des *valeurs propres*.

On peut ainsi positionner les individus dans le plan formé par ces deux **composantes principales**. On a alors la figure suivante:



3.4 Interprétation du graphe des individus

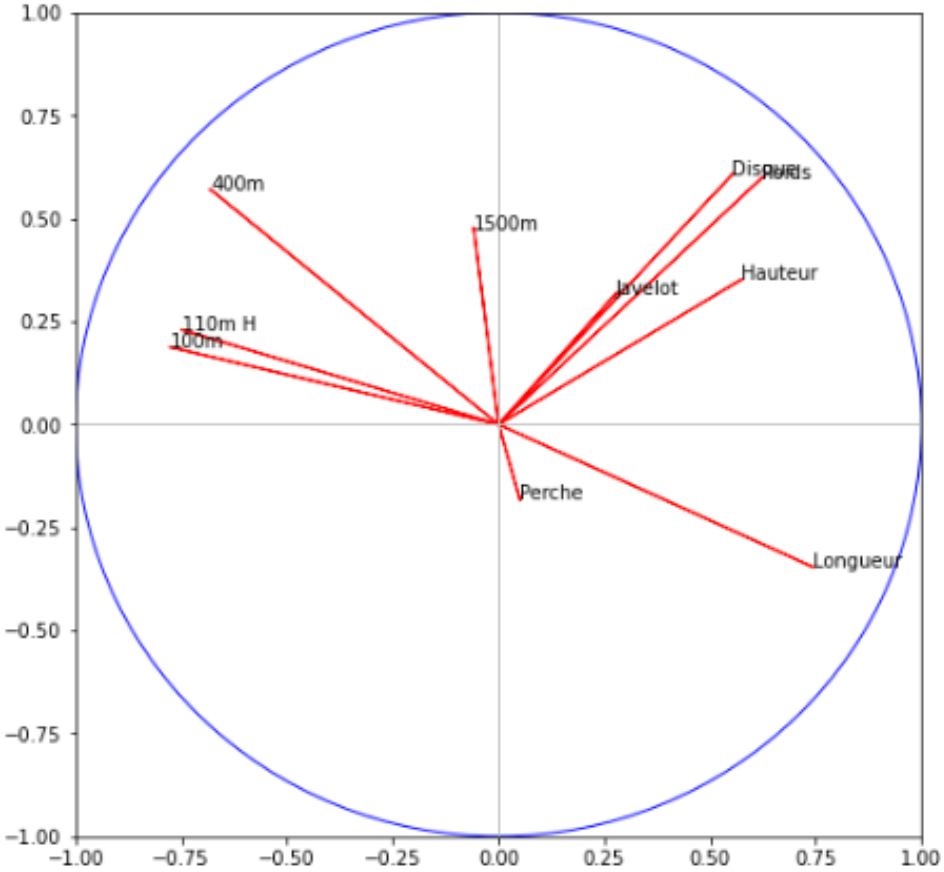
Interprétation du premier axe (Horizontale)

On peut remarquer du premier axe qu'il départage les joueurs ayant un bon classement a ceux qui ont un mauvais classement pendant les deux compétitions. En effet, les joueurs **Karpov** et **Clay** qui sont le deuxieme et le troisième 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.

Interprétation du deuxième axe (Verticale)

On peut remarquer du deuxième axe qu'il départage les joueurs forts dans les épreuves de vitesse et d'endurance a ceux qui sont fort dans les épreuves de lancés (poids, perche, javelot). En effet, les joueurs **Carcasa** et **Yurkov** qui sont meilleurs dans les épreuves de lancées respectivement aux JO et Decastar sont placées en haut du deuxième axe, alors que les joueurs **Drews** et **Bernard** forts aux épreuves respectives d'endurances pendant les JO et de vitesse pendant le Decastar sont placés en bas du deuxième axe.

Le cercle de corrélation ci-dessous peut nous donner encore plus d'informations sur nos données par l'interprétation des variables.



3.5 Interprétation du cercle de corrélation

Interprétation des corrélations positives

- On remarque que les épreuves **100m** et **110m H** sont très corrélées positivement, c'est a dire que les joueurs forts sur l'une sont aussi forts sur l'autre. Ce sont les joueurs fort en vitesse qu'on a vu dans l'interprétation des **composantes principales**.
- De même, les épreuves **Poids**, **Disque** et **Javelot** sont corrélées positivement, cette corrélation regroupe les épreuves de lancés. Ceci aussi renforce notre interprétation des **composantes principales**.

Interprétation des corrélations négatives

- On remarque que les épreuves **400m** et **Longueur** sont très corrélées négativement, c'est a dire que les joueurs meilleurs sur l'un sont mauvais sur l'autre. On peut remarquer dans ce cas le joueur **Martineau** pendant le Decastar, ou il est très fort en **400m** mais très mauvais en **Longueur**, ou le joueur **Karpov** pendant les JO, qui est très faible en **400m** mais très fort en **Longueur**.

4 Pratique de l'AFC

4.1 Recherche des axes a retenir et des coordonnées des points lignes et des points colonnes

Nous pouvons voir sur le tableau ci-dessous une partie des données de la répartition des nombres de médaillés sur les épreuves des Jeux Olympiques par les différents pays participants:

	alg	aus	bah	bar	bdi	blr	bra
10000m	0	0	0	0	0	0	0
100m	0	0	0	1	0	0	0
110mH	0	0	0	0	0	0	0
1500m	1	0	0	0	0	0	0
200m	0	0	0	0	0	0	0
20km	0	2	0	0	0	0	0
3000mSteep	0	0	0	0	0	0	0
400m	0	0	0	0	0	0	0
400mH	0	0	0	0	0	0	0
4x100m	0	0	0	0	0	0	2
4x400m	0	1	1	0	0	0	0
5000m	1	0	0	0	1	0	0
50km	0	1	0	0	0	0	0
800m	1	0	0	0	0	0	0
Decathlon	0	0	0	0	0	1	0
Disque	0	0	0	0	0	2	0
Hauteur	1	0	0	0	0	0	0
Javelot	0	0	0	0	0	0	0
Longueur	0	1	0	0	0	0	0
Marathon	0	0	0	0	0	0	1
Marteau	0	0	0	0	0	2	0
Perche	0	1	0	0	0	0	0
Poids	0	0	0	0	0	1	0
Triple saut	0	0	2	0	0	0	0

FIG. 1 – *Extrait de la repartition des nombres de medailles*

Après avoir centre puis réduit le tableau de données, nous calculons les valeurs propres de la matrice d'inertie associée.

```
print(acf.eig_)
```

```
[[ 0.81669755  0.62070983  0.54424401]
 [13.85386679 10.52927281  9.23216187]
 [13.85386679 24.3831396  33.61530146]]
```

FIG. 2 – *Valeurs Propres de la matrice associée*

L'étude des valeurs propres permet ensuite de déterminer les axes a retenir.

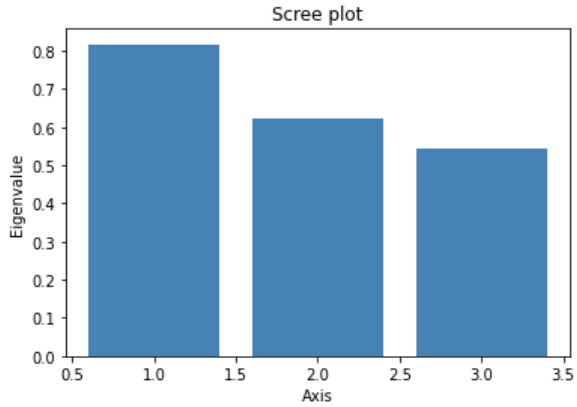


FIG. 3 – *Représentation des Valeurs Propres en pourcentage de la variance*

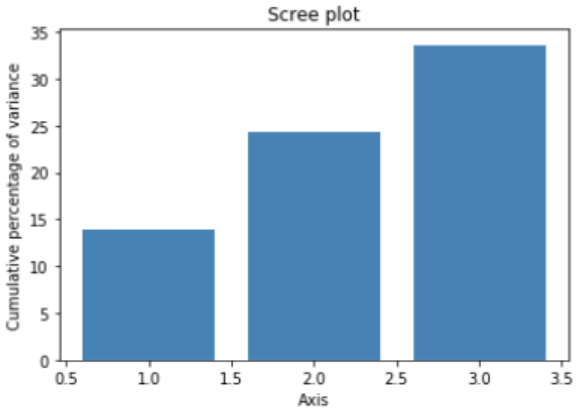


FIG. 4 – *Représentation des Valeur Propres en pourcentage cumule de la variance*

Nous retiendrons alors les 3 axes associes a ces valeurs propres pour l'AFC: en effet ces 3 axes permettent d'expliquer environ 35% de la variance totale des données. Nous déterminons ensuite les coordonnées, les contributions et les $\cos^2$ de chacun des points lignes et points colonnes pour tous les axes factoriels.

On procède au mapping simultané des points lignes et des points colonnes: nous représentons dans le même repère des axes retenus les points lignes et les points colonnes afin d'en dégager l'interprétation des données.

Dans notre cas, nous effectuons le mapping sur l'axe 1 et 2:

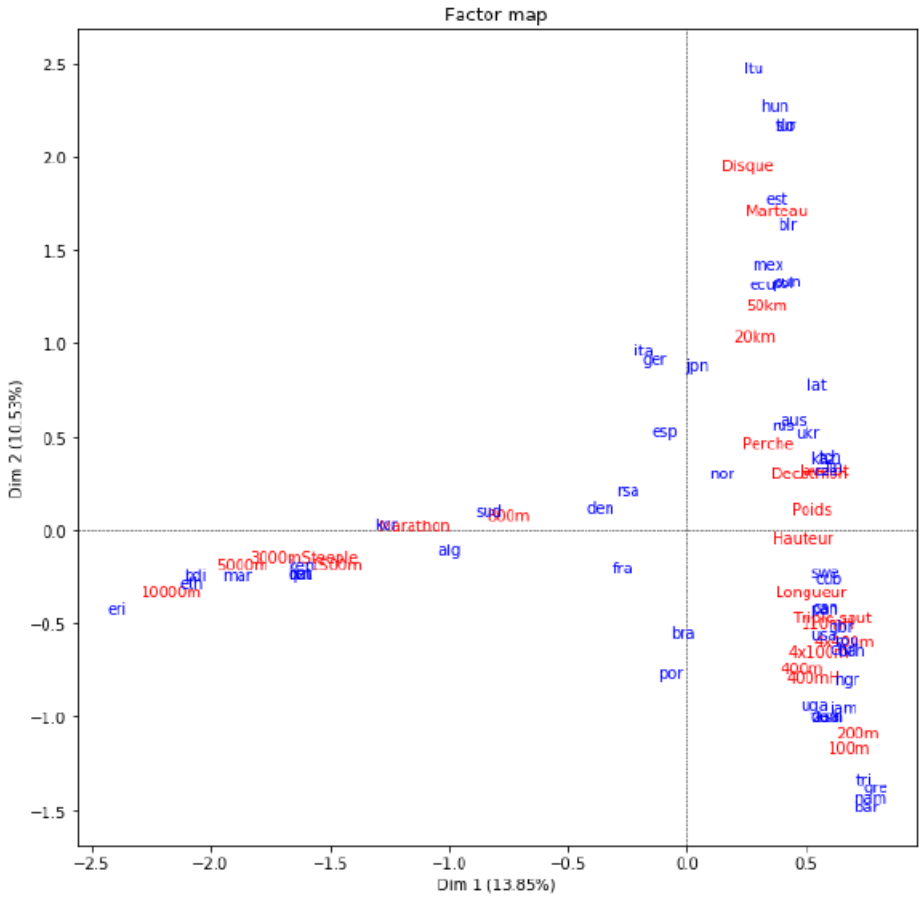


FIG. 5 – *Mapping sur l'axe 1 et 2 des points lignes et des points colonnes*

4.2 Interpretation

Dans notre interprétation, nous allons nous baser sur le mapping effectue sur les axes 1 et 2 précédent.

En reliant chaque point ligne et chaque point colonne a l'origine, on peut associer a chaque point un vecteur relatif a ce point. Ainsi, en étudiant les produits scalaires entre chaque vecteur nous distinguons essentiellement 3 cas: soit le produit scalaire est positif, soit il est négatif, soit il est proche de zero.

Dans le cas d'un produit scalaire positif, nous avons une conjonction entre les vecteurs: on a une forte correspondance entre les vecteurs.

Si le produit scalaire est négatif, nous avons une opposition entre les vecteurs: on a une forte incompatibilité entre les vecteurs.

Enfin lorsque le produit scalaire est nul, nous avons une quadrature entre les vecteurs: il n'y a pas de corrélation entre les vecteurs.

Ainsi, nous remarquons que le plan est divise en 3 régions de conjonctions.

Premièrement, nous remarquons une région ou se concentrent principalement les épreuves de forces tels que le lance de poids, le lance de disque, le lance de marteau,... ou dominent principalement domines par les Européens (Biélorussie, Russie,...).

Deuxièmement, nous observons une région ou l'on retrouve surtout les épreuves de course d'endurance telles que le 10000m, le 5000m, le marathon,... ou dominent principalement les Africains (Kenya, Éthiopie,...).

Puis finalement, nous avons une dernière région ou se concentrent surtout les épreuves de courses de vitesse telles que le 100m, le 200m, le 4*100m,... ou dominent principalement les Américains (USA, Jamaïque,...).

Ensuite, nous remarquons que la région ou se concentrent les épreuves de forces est en opposition avec celle ou se concentrent les épreuves de vitesses. Nous en déduisons que les athlètes doués en épreuves de forces sont généralement moins bons pour les épreuves de vitesses et inversement. Par conséquent, les pays ayant reçus plus de médailles dans les épreuves de forces ont généralement eu moins de médailles dans les épreuves de vitesses et réciproquement. Ainsi les Européens sont généralement bons en épreuve de force tel que le lance de marteau ou le lance de disque mais sont mauvais au 100m ou 200m. Tandis que les Américains sont plutôt bon au 100m et 200m, mais mauvais aux épreuves de lances(disques, marteau,...).

Enfin, nous remarquons que la région ou se concentrent les épreuves d'endurance est en quadrature avec ces deux régions. Donc les athlètes qui sont bons au épreuves d'endurance sont généralement moyens aux épreuves de forces et aux de vitesses. Par conséquent, les pays ayant remportes plusieurs médailles dans ces disciplines ne sont ni avantages ni désavantages dans les deux autres sortes d'épreuves. Ainsi les Africains qui sont généralement bons au marathon ne sont ni excellent ni mauvais ni au 100m ni au lance de marteau.