Réduction de la dimensionnalité : Analyse en Composantes Principales

Juste Raimbault [1] (adapté du cours de Paul Chapron [1])

2024-2025

[1]IGN-ENSG-UGE

Introduction

Motivation

La plupart des phénomènes intéressants (sociaux, spatiaux) sont multi-factoriels. Les données disponibles pour les décrire sont :

- partiellement redondantes : e.g. revenu et profession
- intrinsèquement corrélées : e.g. revenu et taille du logement
- répétées : e.g. données mensuelles ou hebdomadaires

Quoi et Comment

La réduction de la dimensionnalité cherche à réduire la colinéarité et le nombre de dimensions (=variables) qui décrivent une population ...

... L'Analyse en Composantes Principales traite des variables numériques...

... en proposant de nouvelles variables composites décorrélées.

À quoi ça sert ?

- à identifier les ressemblances entre individus, à les regrouper en fonction de cette ressemblance
- à identifier les ressemblances entre variables (liaisons)

$\Longrightarrow$ résumé de l'information contenue dans les données, pour la restituer fidèlement (i.e. sans trop les déformer).

Pokemons

Une population

Plusieurs dimensions

- Nom e.g. "Pikachu"
- Type 1 $\in \{Grass,\ Fire,\ Water,\ Bug, \dots\}$
- Type 2 idem
- HP : numérique
- Attack : numérique
- Defense : numérique
- Speed : numérique
- Special Attack :numérique
- Special Defense : numérique
- Generation : facteur $\in \{1, 2, 3, 4, 5, 6\}$
- Legendary : booléen

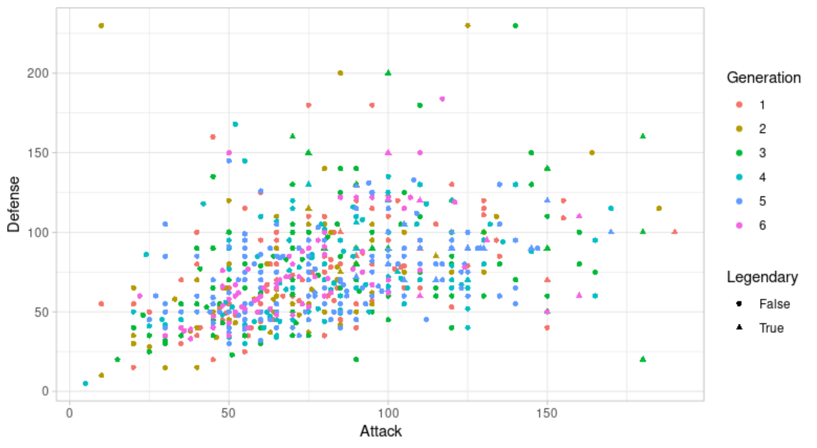
Dimensions "composites"

Existe-t-il des combinaisons qui résument bien les caractéristiques des pokemons ? (moins de six!)

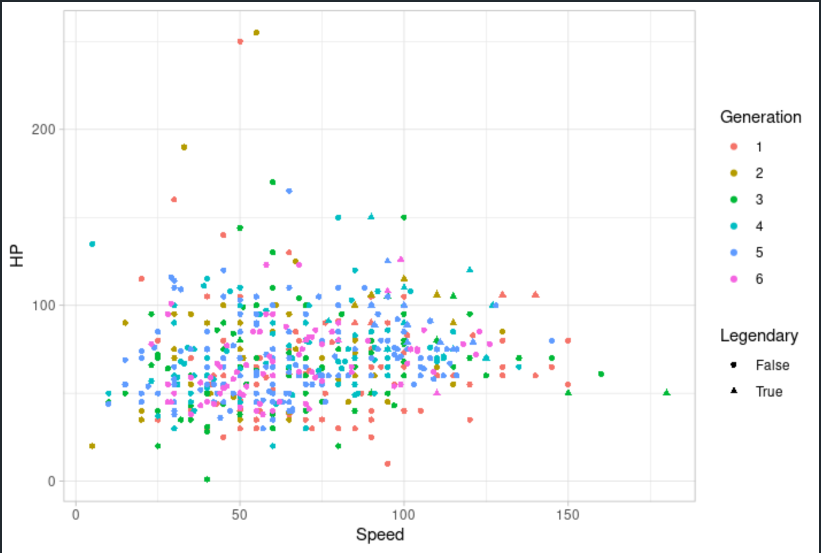
Comment les constituer ?

i.e. comment combiner les six variables numériques pour bien expliquer leur variation au sein de la population ?

Attack vs. Defense



Speed vs. HP



HP
200
100
0
0
50
100
150
Speed
Generation
1
2
3
4
5
6
Legendary
False
True

L’inertie

L'inertie

L'inertie est l'équivalent multi-dimensionnel de la variance d'une variable. C'est la dispersion des données.

C'est une notion centrale de l'ACP.

$$I = \frac{1}{n} \sum_{i=1}^{n} d^2(x_i, g)$$

Avec

- n la taille de la population
- x_i la valeur de la variable de l'individu i
- g le point moyen
- $d(x, y)$ une distance, souvent euclidienne : $(x_i - g_i)^2$

L’inertie

L’inertie quantifie la dispersion du nuage de points

L’inertie est la "moyenne du carré des distances", ou encore la somme des variances des variables

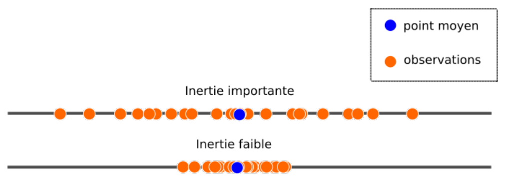
Inertie faible $\implies$ peu de variété dans les variables, individus semblables, faible quantité d’information

Soit une population P de n individus décrits par une variable X

l'inertie de la population est la variance de X :

$$I = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Le point moyen a pour "coordonnées" $\bar{x}$



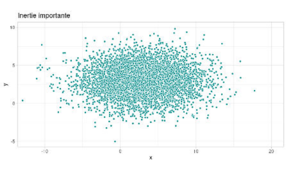
L'inertie en 2D

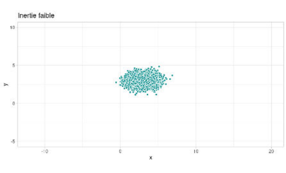
Soient X et Y deux variables qui décrivent des individus p_i de la population P, et $g = (x_g, y_g)$ le point moyen de cette population, de coordonnées $x_g = \bar{x}$ et $y_g = \bar{y}$.

L'inertie de P est :

$$I = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_g)^2 + (y_i - y_g)^2$$

On reconnaît une somme de variances : $I = var(X) + var(Y)$





Soient v variables , notées $X^{(k)}, k \in \{1, \ldots, v\}$ qui décrivent les individus d'une population P, le point moyen de P est noté g.

L'inertie de P est :

$$I = \frac{1}{n} \sum_{k=1}^{v} \sum_{i=1}^{n} (x_i^{(k)} - x_g^{(k)})^2$$

on reconnaît $I = \sum_{k=1}^{v} var(X^{(k)})$

Espaces, vecteurs, axes, variables

Individus dans l'espace d'origine

L'ACP considère une population statistique décrite par plusieurs variables (continues).

Ces variables définissent un espace vectoriel , qu'on va appeler l'espace d'origine:

- un individu i est un vecteur
- la valeur de ses variables sont les coordonnées du vecteur dans cet espace.
- chaque variable est une dimension de cet espace. elle définit un axe de l'espace. (cf. axe des x dans un repère orthonormé)

Les variables étant potentiellement corrélées, les axes de l'espace de départ ne sont pas toujours (presque jamais) orthogonaux !

Explicitation de l'ACP

L'ACP consiste à trouver de nouveaux axes orthogonaux entre eux, qui capturent le plus d'inertie possible de la population P.

Ces axes définiront un nouvel espace : l'espace d'arrivée

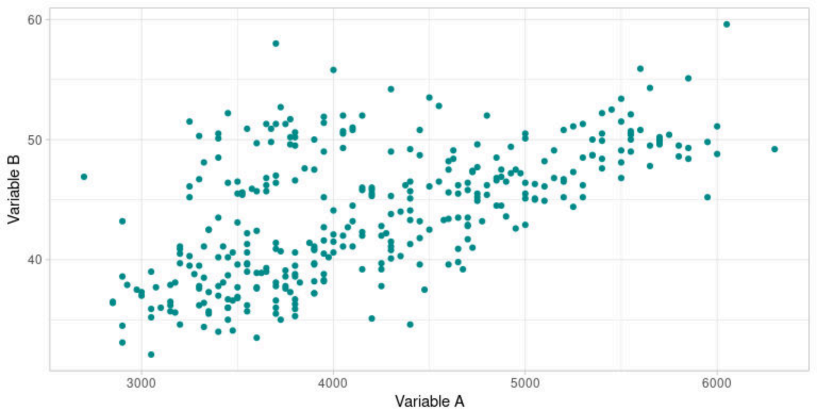
On trouve ces axes en combinant (linéairement), les variables de la population P, par exemple :

$$axe_1 = \alpha X + \beta Y + \gamma Z$$

La composition de ces combinaisons (les valeurs de α, β, γ) pour chaque axe est donnée en résolvant un système d'équations algébriques

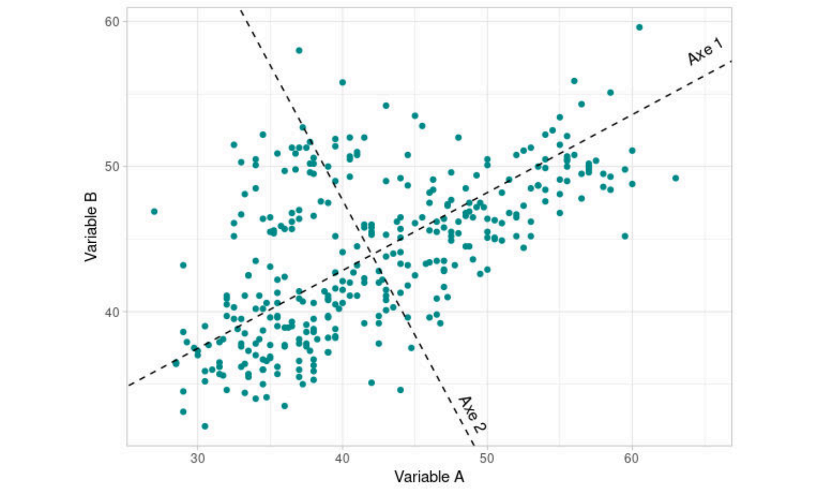
Exemple en 2D

Espace de départ



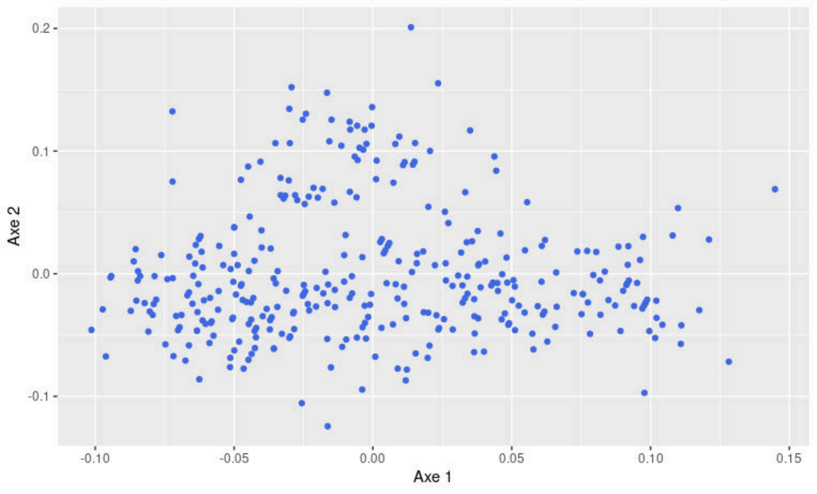
Exemple en 2D

Espace de départ + Les axes de l'espace d'arrivée

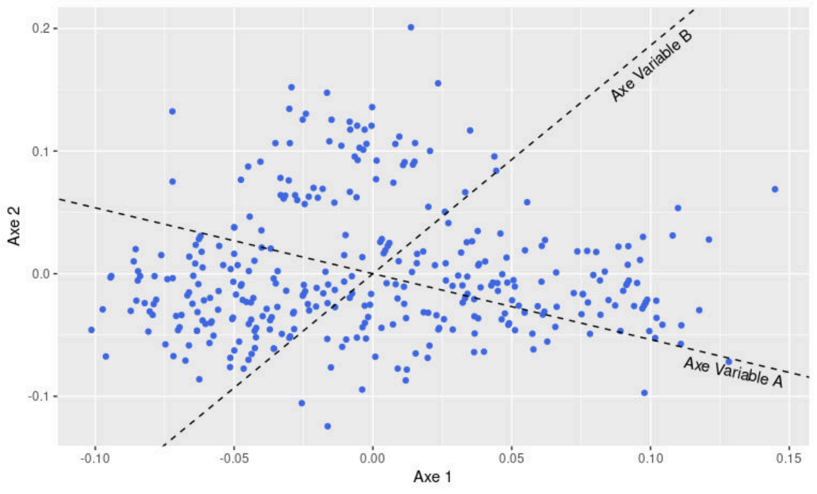


Exemple en 2D

Espace d'arrivée



Espace d'arrivée + les axes de l'espace de départ

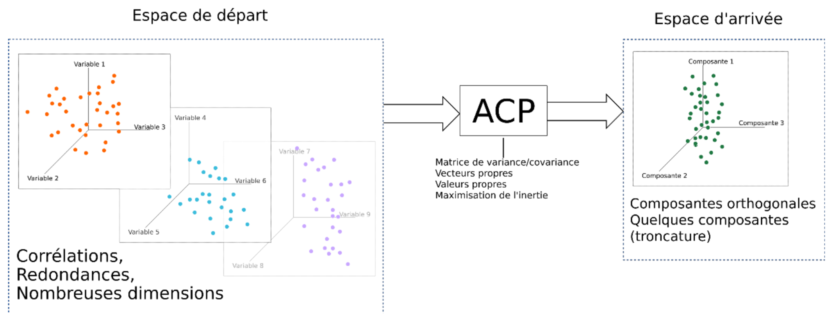


Calcul des axes

Les axes sont les vecteurs propres de la matrice de corrélation de P. On peut les calculer

l'ACP est le calcul d'une transformation linéaire qui re-projette des vecteurs-individus dans un nouvel espace – l'espace d'arrivée– constitué par les nouveaux axes.

On appelle ces axes composantes, elles sont linéairement indépendantes et forment une base de l'espace d'arrivée.


Espace de départ
Variable 1
Variable 3
Variable 2
Variable 4
Variable 6
Variable 5
Variable 7
Variable 9
Variable 8
Corrélations,
Redondances,
Nombreuses dimensions
ACP
Matrice de variance/covariance
Vecteurs propres
Valeurs propres
Maximisation de l'inertie
Espace d'arrivée
Composante 1
Composante 3
Composante 2
Composantes orthogonales
Quelques composantes
(troncature)

Centrer et réduire les variables ?

Une pratique courante de l'ACP consiste à centrer et réduire les variables du jeu de données avant de réaliser l'ACP

Nombres de composantes et inertie

L'ACP capture l'inertie de P en créant des composantes (les vecteurs propres de la matrice de variance/covariance de P).

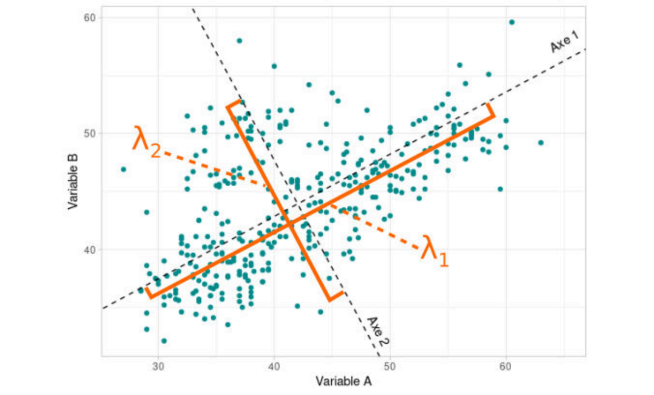
Il y a autant de composantes possibles que de dimensions de l'espace de départ.

L'intérêt de l'ACP est de pouvoir se limiter à quelques composantes :

- pour capturer suffisamment l'inertie ($\approx$ l'information) de P
- pour réduire la dimensionnalité ($\approx$ complexité) de P

Les vecteurs propres définissent la direction des axes.

Une valeur propre associée à un vecteur propre quantifie la dispersion des points le long de l'axe orienté par le vecteur propre.



Nombres de composantes et inertie

L'inertie capturée par une composante k est sa valeur propre , λ_k

On ordonne les composantes par valeur propre décroissantes:

- La 1$^{\text{ère}}$ composante correspond au vecteur propre de plus grande valeur propre, elle capture la plus grande proportion d'inertie
- La 2$^{\text{nde}}$ composante correspond au vecteur propre de la seconde plus grande valeur propre , elle capture la seconde plus grande proportion d'inertie
- etc.

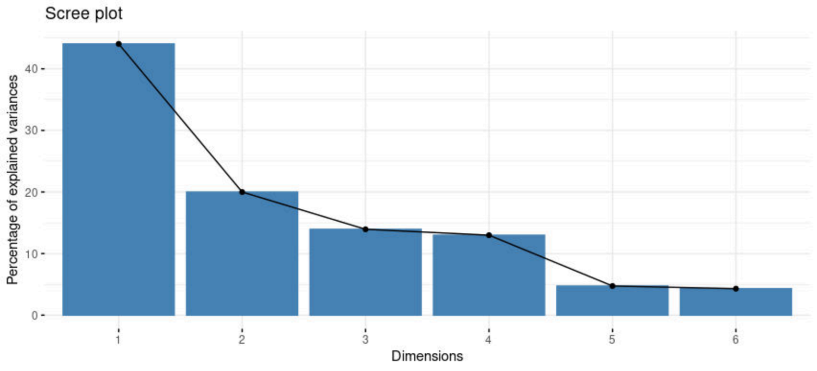
Si les variables sont centrées et réduites, leur somme vaut $Dim(P)$

Interpréter les résultats d'une ACP

Les objets à explorer dans les résultats d'une ACP

- Dimensionnalité : L'essentiel de l'inertie est-elle exprimée en peu de dimensions dans l'espace d'arrivée ?
- Colinéarité des variables : Comment les variables de l'espace de départ sont-elles corrélées entre elles et aux axes de l'espace d'arrivée ?
- Contribution : À quel point Individus et Variables contribuent aux axes de l'espace d'arrivée ?
- Représentation : Les Individus et Variables sont ils·elles bien représenté·e·s par les axes de l'espace d'arrivée ?

Le scree plot montre la proportion d'inertie capturée par les différentes composantes. La valeur propre associée aux vecteurs propres (axes) est proportionnelle à l'inertie capturée.



Dimensionnalité : Nombres de composantes et inertie

Idéalement, les premières (2 ou 3) composantes capturent une partie significative (e.g. $\gtrapprox 50\%$) de l'inertie de P.

Cela signifie que les composantes résument bien l'information contenue dans les variables de P, en peu de dimensions.

Dimensionnalité : Nombres de composantes

Pour profiter du "résumé" de l'ACP, il faut se limiter à un certain nombre de composantes pour définir l'espace d'arrivée.

Heuristiques du choix du nombre:

- On garde les q axes que l'on sait interpréter : 2 ou 3 !
- "coude" dans le scree-plot.
- ne conserver que les $\lambda > 1$ ou $\lambda > 2$
- Karlis-Saporta-Spinaki : conserver les λ t.q. $\lambda > 1 + 2\sqrt{\frac{p-1}{n-1}}$
- Gavish & Donoho (2014) : $\lambda = \frac{4\sigma\sqrt{n}}{\sqrt{3}}$ avec σ le bruit estimé dans les données.

Avec λ, les valeurs propres associées aux axes, p le nombre de variable de P, et n la taille de P

Nombres de composantes et visualisation

En pratique , si on sélectionne q composantes, il faudra projeter les individus et les variables dans C_q^2 plans pour les visualiser.

Si $q = 3$, il faut 3 graphiques $\{(q_1, q2), (q_2, q_3), (q_1, q_3)\}$.

Si $q = 4$, il en faut 6 !

L’espace d’arrivée

On sait passer de l’espace de départ à l’espace d’arrivée : On peut projeter les variables et les individus dans l’espace d’arrivée

De cette projection on tire beaucoup d’information utiles:

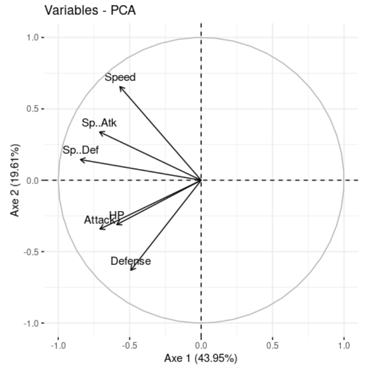
- corrélations de variables (si elle sont bien représentées !)
- contribution / représentation des variables
- contribution / représentation des individus
- regroupements d’individus, individus extrêmes

Projection des variables : colinéarité, contribution, qualité de la représentation

Colinéarité des variables

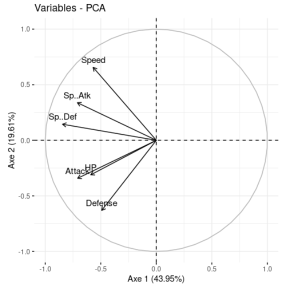
Rappel : les variables sont des vecteurs dans l'espace des individus.

On peut projeter les variables dans l'espace d'arrivée :



Si les variables sont centrées et réduites lors de l'ACP, on peut les représenter dans un cercle de corrélation et évaluer visuellement leur corrélation

Colinéarité des variables



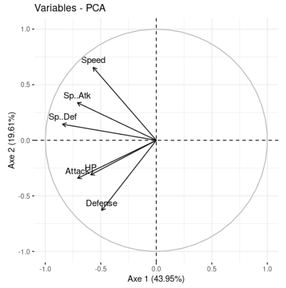
Variable $\leftrightarrow$ Flèche

Coordonnées de la variable $\leftrightarrow$ corrélation linéaire avec les composantes

Proximité au cercle $\leftrightarrow$ qualité de représentation de la variable

Angle des variables $\leftrightarrow$ corrélation des variables entre elles

Colinéarité des variables



- la corrélation de `Defense` avec l'Axe 1 est de -0.5
- `Attack` et `HP` sont très corrélées
- `Speed` et `Defense` sont (linéairement) indépendantes

Ici : regroupement de variables ? Oui !

Contribution des variables

La contribution d'une variable v à l'inertie de l'axe k est la coordonnée carrée de v sur l'axe k divisée par son inertie.

$$Contrib_{vk} = \frac{c_{vk}^2}{\lambda_k}$$

Plus la contribution d'une variable est élevée , plus elle est importante pour expliquer la variabilité de P

La qualité de représentation d'une variable v par l'axe k est la coordonnée carrée de v sur l'axe k :

$$Qlt_{vk} = c_{vk}^2$$

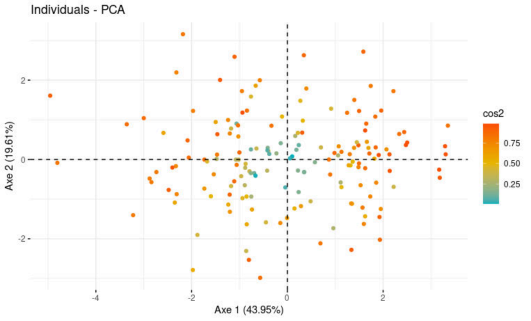
(On peut vouloir vérifier qu'une variable d'intérêt soit bien représentée dans les premières composantes.)

Projection des individus: contribution , qualité de la représentation

Nuage de points des individus dans le plan

Rappel : les individus sont des vecteurs dans l'espace des variables de P.

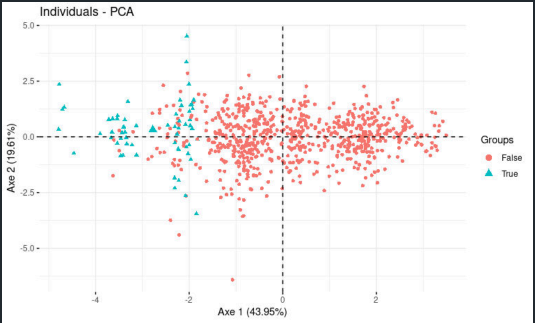
On peut projeter les individus dans l'espace d'arrivée :



une fois projetés, les individus similaires sont proches. Parfois cela fait apparaître des regroupements (ici, pas vraiment) et des individus extrêmes.

Nuage de points des individus dans le plan

Il est parfois pertinent de colorer les individus projetés par une variable tierce (i.e. non inclus dans P_i lors du calcul des composantes)



Ici : PCA sur toutes les générations de Pokemons, individus projetés sur (Axe_1, Axe_2), colorés selon le facteur `Legendary`

Contribution des individus

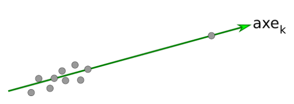
La contribution de l'individu i à l'axe k s'écrit :

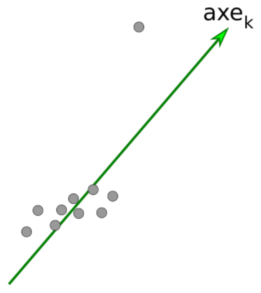
$$Contrib_{ik} = \frac{p_i c_{ik}^2}{\lambda_k}$$

Avec :

- c_{ik} la coordonnée de i selon k
- p_i le poids de l'individu, à poids constants $\forall i, p_i = \frac{1}{n}$
- λ_k la valeur propre associée à l'axe k

- Plus la valeur $Contrib_{ik}$ est extrême, plus elle influe sur la direction de l'axe k
- la coordonnée doit être rapportée à l'étirement du nuage de points donné par λ_k
- filtrer des individus extrêmes *peut* améliorer l'ACP !

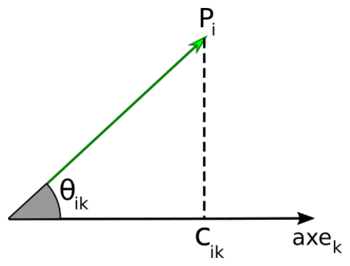




Qualité de représentation des individus

La qualité de représentation de l'individu i à l'axe k s'écrit :

$$Qlt_{ik} = cos^2(\theta_{ik}) = \frac{c_{ik}^2}{\|P_i\|^2}$$



Avec :

- c_{ik} la coordonnée de i selon k
- θ_{ik} l'angle entre le vecteur P_i et l'axe k
- $\|P_i\|$ la norme du vecteur l'individu i

Nouvelles variables, Nouveaux individus

On peut intégrer de nouvelles variables et individus:

- soit dans le calcul de l'ACP, ce qui modifie l'espace d'arrivée,
- soit a posteriori.

Bilan

Bilan de l'ACP

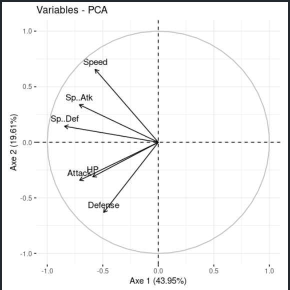
Avantages

- Réduit la dimensionnalité
- Regroupe les variables et les individus
- montre l'effet conjoint des variables

Limites

- Composantes difficiles à interpréter en elles-mêmes
- hypothèses fortes : la variance est un mélange "linéaire", et la variance est de l'information, pas du bruit ($\approx$RSB fort)
- Que faire si p est grand et si les premières composantes capturent peu d'inertie ?

Pokémonologie



- L'Axe 1 "prend tout" : c'est la puissance générale des pokémon, une sorte de score global
- L'Axe 2 sépare les variables en deux groupes : celle du combat "standard" (`Attack`, `Defense`,`HP`) et celles du combat "spécial/rapide" (`Sp..Atk`, `SP..Def`, `Speed`)
- On pourrait être tenté de diviser les pokemons en "Costauds classiques " vs. "Ninjas spéciaux" .

Merci à Anh Le , `https://anhqle.github.io/gotta-plot-them-all/`

Pour plus tard

La notion d'inertie est très utile en classification : observer la chute d'inertie intra-classe indique souvent le nombre optimal de classes ! (cf CAH)

La malédiction de la dimensionnalité (curse of dimensionality) peut nuire dans beaucoup de traitements numériques . Elle peut (parfois) être contournée, en appliquant une ACP !