

Analyse en Composantes Principales

Paul Chapron & Yann Ménéroux

2021-2022

IGN-ENSG-UGE

ENSG
Géomatique

ÉCOLE NATIONALE
DES SCIENCES
GÉOGRAPHIQUES

Introduction

Techniques pour quantifier la liaison entre **deux** variables (quali ou quanti).

- corrélation
- régression linéaire
- χ^2 (test d'indépendance)
- visualisation adéquate

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
- parfois des proportions (somme à 1 ou 100%)

L'analyse **factorielle** cherche à réduire la **colinéarité** et le **nombre de dimensions** (=variables) qui décrivent une population ...

... en proposant de nouvelles variables **composites décorrélées**.



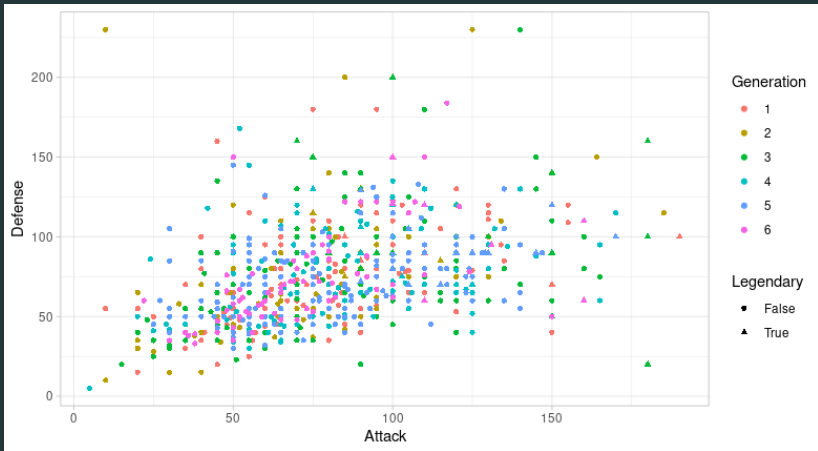
- Nom e.g. "Pikachu"
- Type 1 $\in \{Grass, Fire, Water, Bug, \dots\}$
- Type 2 idem
- HP : numérique
- Attack : numérique
- Defense : numérique
- Speed : numérique
- Special Attack :numérique
- Special Defense : numérique
- Generation : facteur $\in \{1, 2, 3, 4, 5, 6\}$
- Legendary : booléen

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

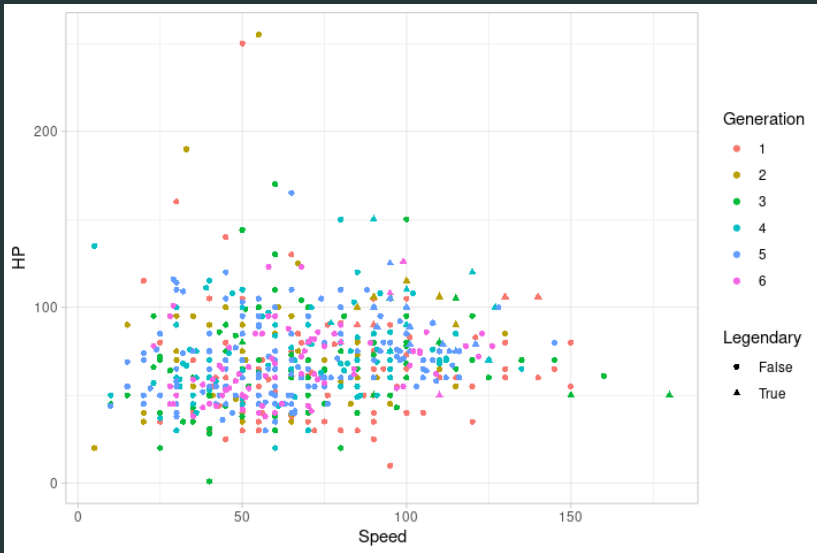
Comment les constituer ?

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

Attack vs. Defense



Speed vs. HP



L'inertie

L'inertie est l'équivalent **multi-dimensionnel** de la **variance** d'une variable.

C'est une notion centrale de l'ACP.

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

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 quantifie la **dispersion** du nuage de points

C'est 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}

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

Soient v variables , notées $X^{(k)}, k \in \{1, \dots, v\}$ 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

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

Espaces, vecteurs, axes, variables

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 ne sont pas toujours orthogonaux !

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

On trouve ces axes en combinant (linéairement), les variables de la population P : e.g.

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

L'algèbre linéaire nous donne la composition de ces combinaisons (les valeurs de α, β, γ) pour chaque des nouveaux axes

TODO SCHEMA

espace de départ \rightarrow ACP \rightarrow espace d'arrivée

Les **axes** sont les **vecteurs propres** de la matrice de corrélation de P . On peut les calculer ! (ouf)

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**.

En général on choisit $\#axes < \#dimensions$ pour **réduire la dimensionnalité**

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

Gavish & Donoho (2014) present a long overdue result on this problem and their answer is surprisingly simple and concrete. Essentially, the optimal procedure boils down to estimating the noise in the dataset, σ , and then throwing away all components whose singular values are below a specified threshold. For a square $n \times n$ matrix, this threshold is:

$$\lambda = \frac{4\sigma\sqrt{n}}{\sqrt{3}}$$

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:

- regroupements d'individus
- corrélations de variables
- contribution / représentation des variables
- contribution / représentation des individus

Contributions/ représentation des variables

Contribution / représentation des individus

Avantages

- Réduit la dimensionnalité
- Regroupe les variables et les individus

Limites

Composantes difficiles à interpréter en elles-mêmes

- This is important

- This is important
- Now this

- This is important
- Now this
- And now this

- This is really important
- Now this
- And now this

Mono message sur une diapo

$$A = \sum_{i=1}^n \left(1 + \frac{1}{x_i}\right)^{\alpha}$$