Analyse en Composantes Principales

Paul Chapron [1] & Yann Méneroux [1]

2021-2022

[1]IGN-ENSG-UGE

ENSG Géomatique

ÉCOLE NATIONALE DES SCIENCES GÉOGRAPHIQUES

Introduction

Dans les cours précédents ...

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

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

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

Quoi et Comment

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.

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

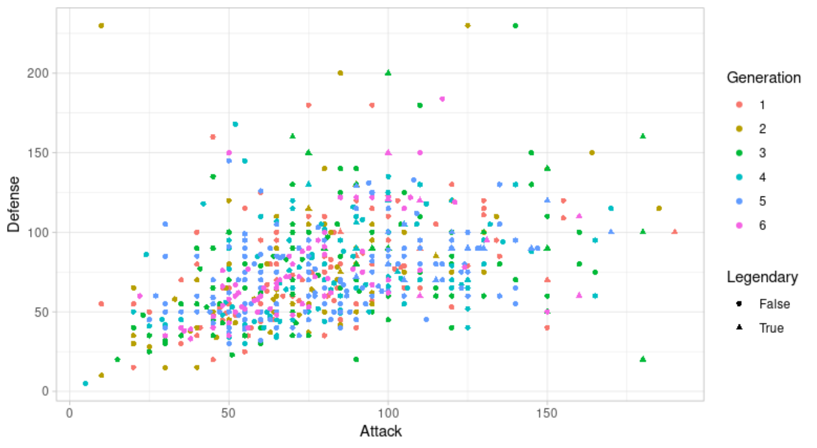
Dimensions ”composites”

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

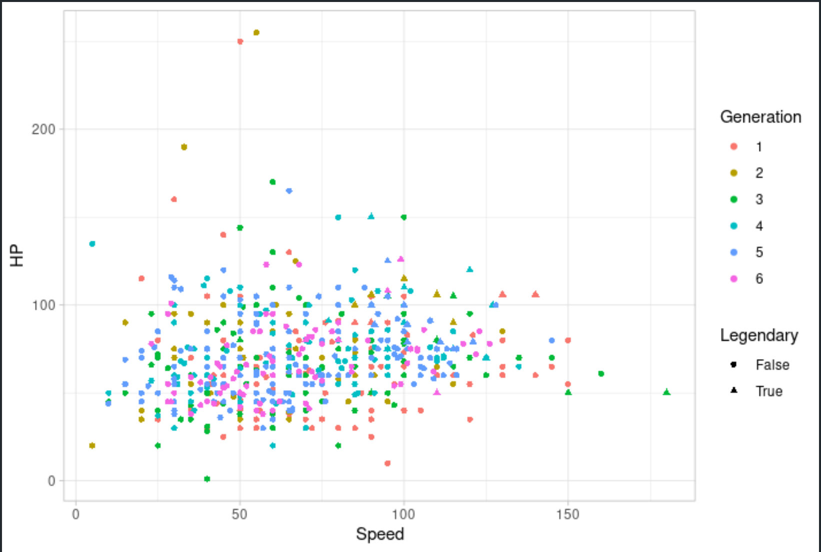
Comment les constituer ?

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
Speed
0
50
100
150
200
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 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

L'inertie en 1D

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 : $I = var(X) + var(Y)$

L'inertie en nD

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 !

Espaces, vecteurs, axes, variables

Individus dans l'espace d'origine

Les individus sont des vecteurs dans l'espace des variables

mais également

Les variables sont des vecteurs dans l'espace des individus

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

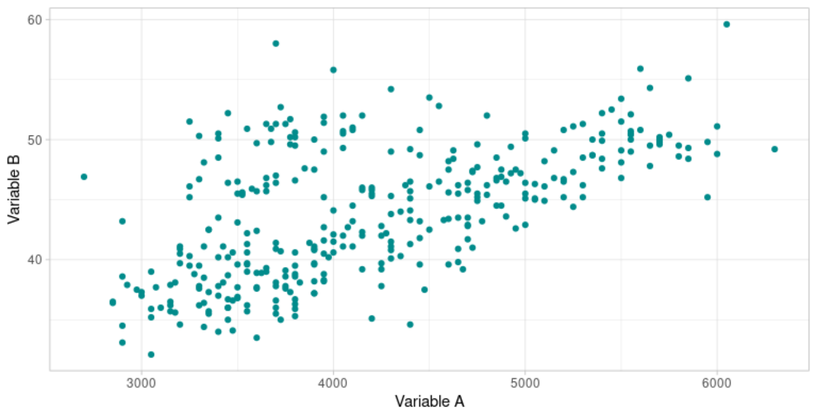
On trouve ces axes en combinant (linéairement), les variables de la population P : par exemple

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

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

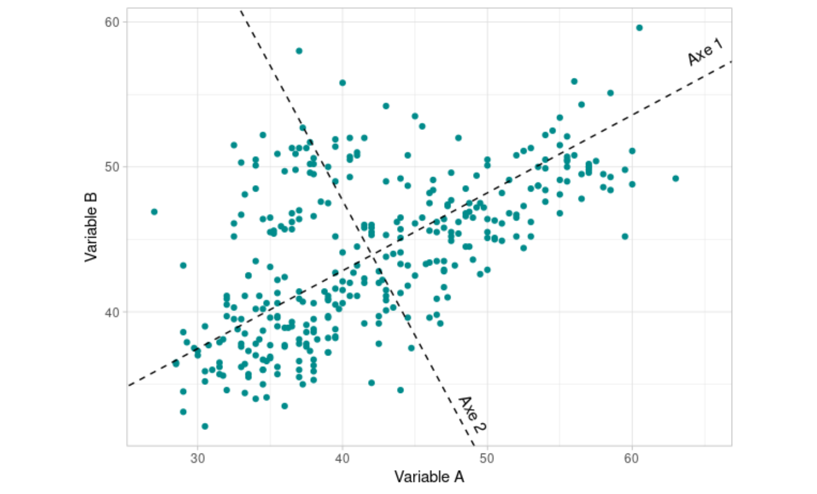
Exemple en 2D

Espace de départ



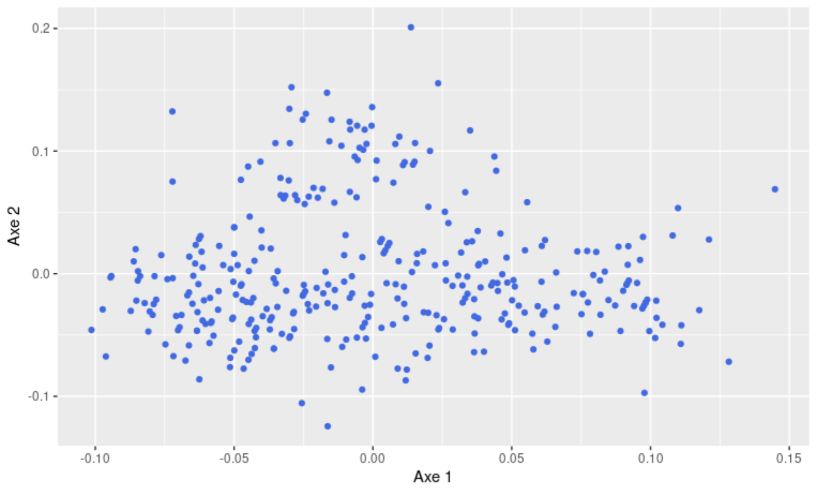
Exemple en 2D

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



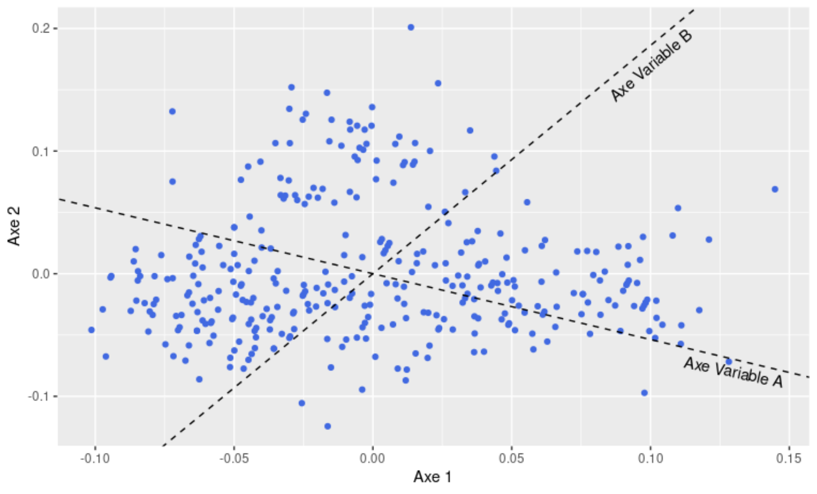
Exemple en 2D

Espace d'arrivée



Exemple en 2D

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



TODO SCHEMA

espace de départ -¿ ACP -¿ espace d'arrivée

Calcul des axes

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.

Nombres de composantes et inertie

Nombres de composantes et inertie

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:

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

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:

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

Interpréter les résultats d'une ACP

L’espace d’arrivée

Bilan de l'ACP

Avantages	**Limites**
• Réduit la dimensionnalité • Regroupe les variables et les individus	Composantes difficiles à interpréter en elles-mêmes

Animation

- This is important

Animation

- This is important
- Now this

Animation

- This is important
- Now this
- And now this

Animation

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

Mono message sur une diapo

Formules

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