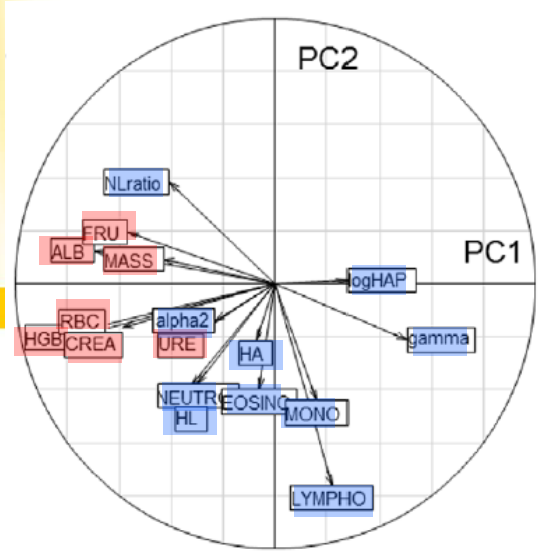
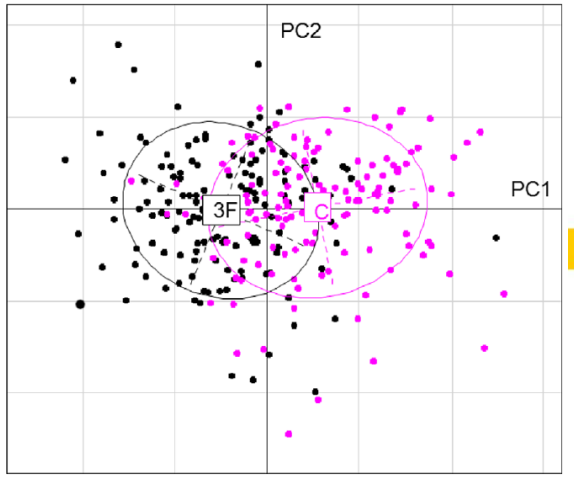
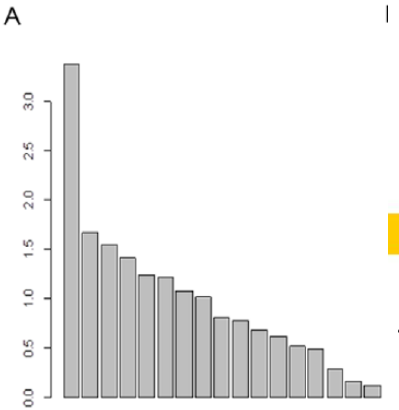
Introduction aux analyses multivariées
Analyse en Composantes Principales
ACP

Emmanuelle Gilot-Fromont

Janvier 2024







VetAgro Sup

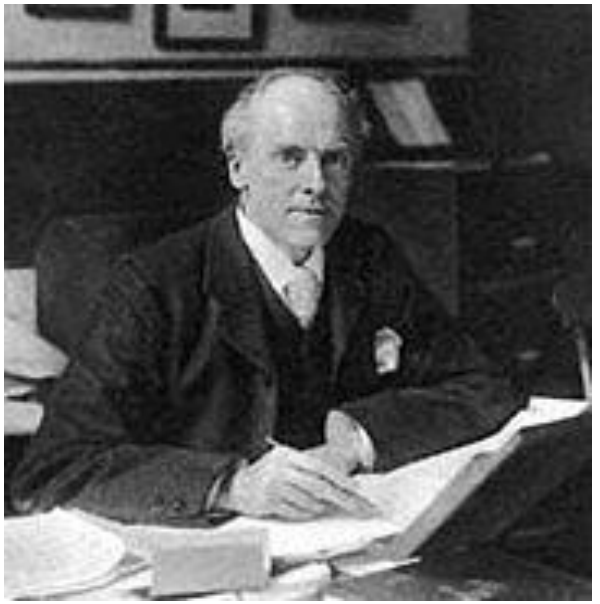
Introduction : analyses multivariées descriptives

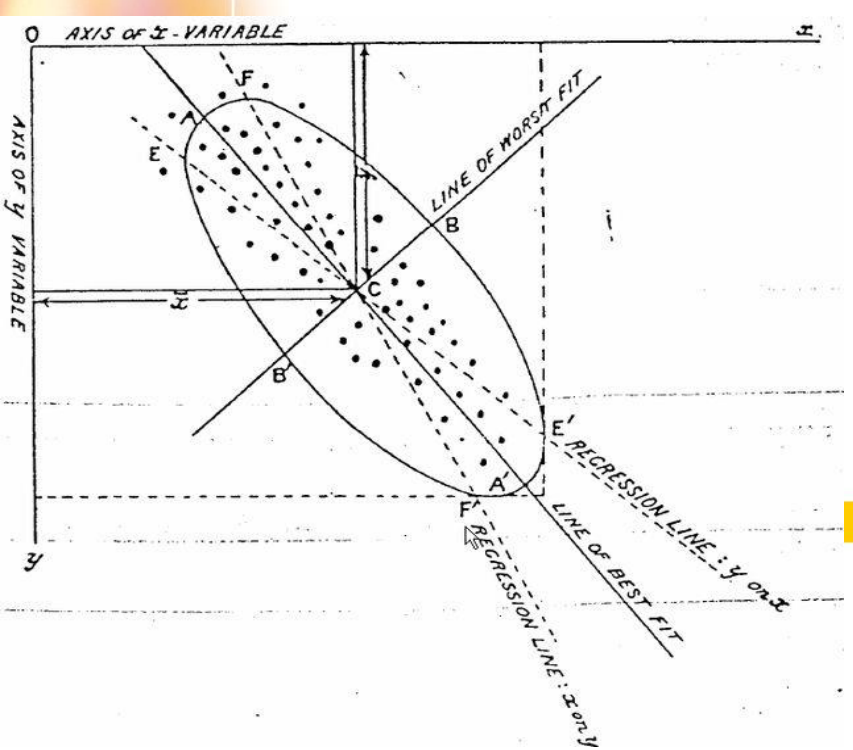
- Domaine de la biologie : données multidimensionnelles : nombreuses lignes, nombreuses colonnes, variables corrélées
- Comment étudier la structure de ces données? -> outil d'analyse
- Comment résumer cette information? -> outil de synthèse
- Analyse descriptive de données multivariées (attention, le terme « multivarié » est utilisé aussi dans le contexte inférentiel)

VetAgro Sup

Historique

Première analyse multivariée développée sur le principe proposé par K. Pearson en 1901 : utiliser les corrélations entre de nombreuses variables pour décrire et résumer l'information contenue dans un jeu de données





Développée par Harold Hotelling dans les années 30 (« transformée de Hotelling ») ainsi que l'AFC et l'analyse canonique, généralisation des analyses précédentes.

Champs d'application

- Biologie
  - biométrie, allométrie : étude des relations entre les variables mesurée chez différents individus (darwinisme, eugénisme)
  - écologie : inventaires floristiques/faunistiques : données mesurées au fil du temps ou dans l'espace
  - données moléculaires
- Recherche économique et sociale
  - traitement de données d'enquêtes : mode de vie, de consommation dans différents groupes de personnes
- Traitement d'image : données spatialisées

VetAgro Sup

Intérêt

- Facilité d'acquisition et d'échange des données : données spatialisées; réseaux, acquisition automatique
- Complexité des systèmes étudiés : nombreuses variables mesurées pour décrire un système à multiples facettes: ex: modalités d'élevage de la dinde de Bresse: 44 élevages, 44 variables de description
- Originalité de l'approche : confrontation de nombreuses informations simultanément … parfois difficile à exploiter et à expliquer : souvent complété par des approches unidimensionnelles

VetAgro Sup

Objectifs du module

- Cours : principe des analyses classiques : choisir la méthode appropriée

- TD : mise en pratique sous R et interprétation : combiner connaissance du problème et compréhension de la méthode utilisée

Prérequis: bases de biostatistiques, bases de R

VetAgro Sup

Plan du module

- Principe, types d'analyses

- Variables quantitatives : l'ACP, inter/intra

- Variables qualitatives : AFC, ACM

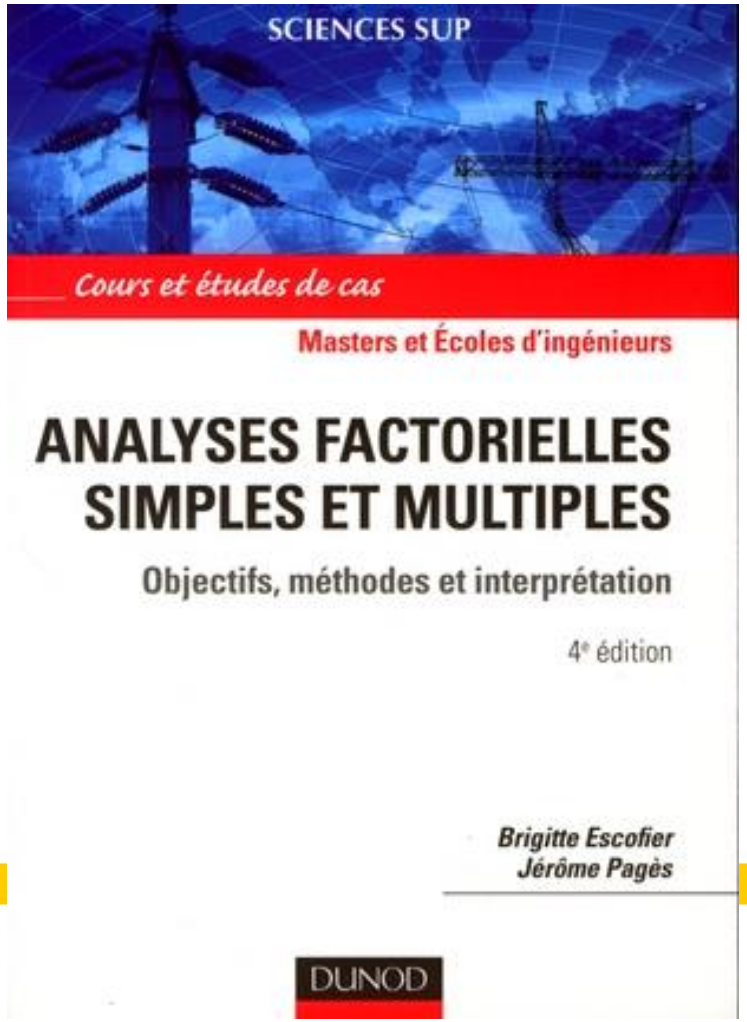
- Classification et discrimination

VetAgro Sup

Base bibliographique

- Escoffier et Pagès 4ème édition, 2008
  Disponible en ligne

- Cours et TD de statistiques en biologie
  de l'UMR 5558 :
  pbil.univ-lyon1.fr/R



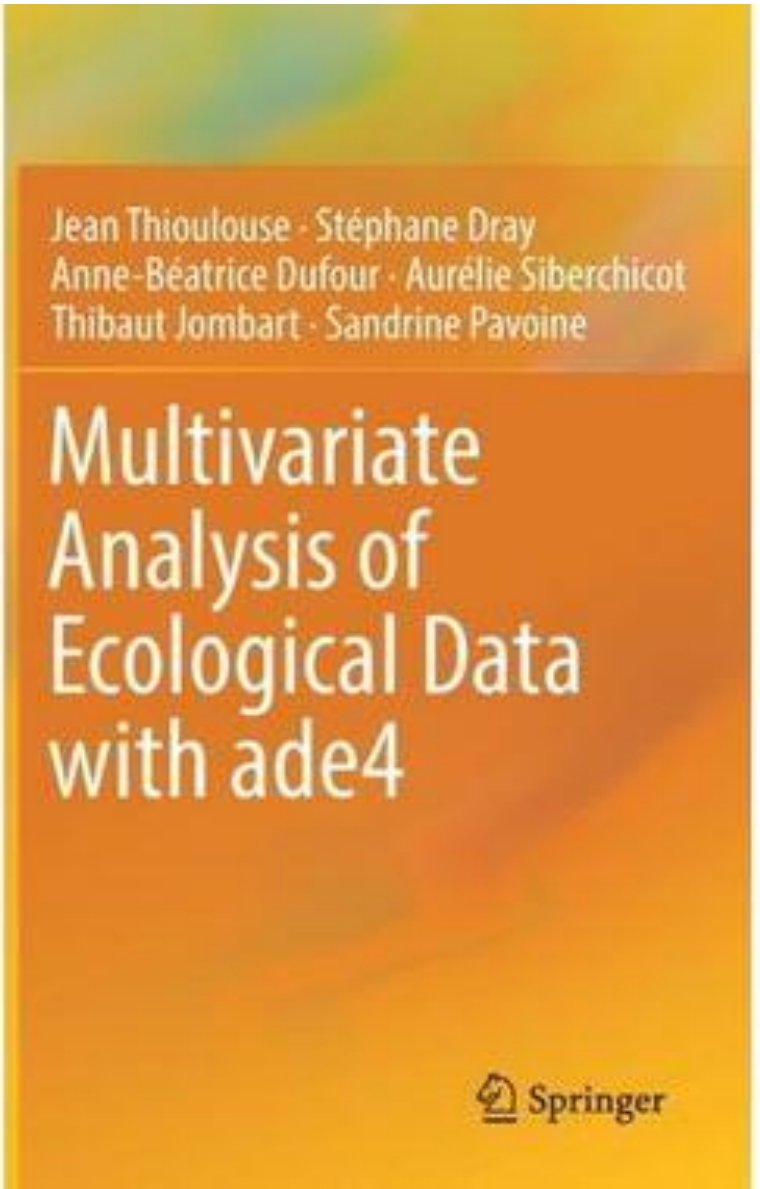
VetAgro Sup

Module ade4 de R

http://pbil.univ-lyon1.fr/ADE-4/

http://pbil.univ-lyon1.fr/R/enseignement.html

Thioulouse J, Dray S, Dufour AB, Siberchicot A, Jombart T, Pavoine S. 2018. Multivariate Analysis of Ecological Data with ade4.



VetAgro Sup

Principe, types d'analyses

1. Principe

2. Types d'analyse

VetAgro Sup

1. Principe

p variables = p mesures qualitatives (modalité) ou quantitative (nombre / mesure)

n individus = n animaux, élevages, sites, populations, espèces

1 tableau variables / individus

VetAgro Sup

Question analytique

Etudier les relations entre les variables/individus :

- Parmi les nombreux individus mesurés, y a-t-il une typologie: des ressemblances? Des groupes?
- Parmi les variables étudiées, y a-t-il une typologie : des corrélations? Des oppositions ? Des groupes de variables ?
- Dans l'association individus-variables, y a –t-il des structures fortes (association majeures)?

VetAgro Sup

Question synthétique

Sur les individus : Quelle dimension de variabilité?

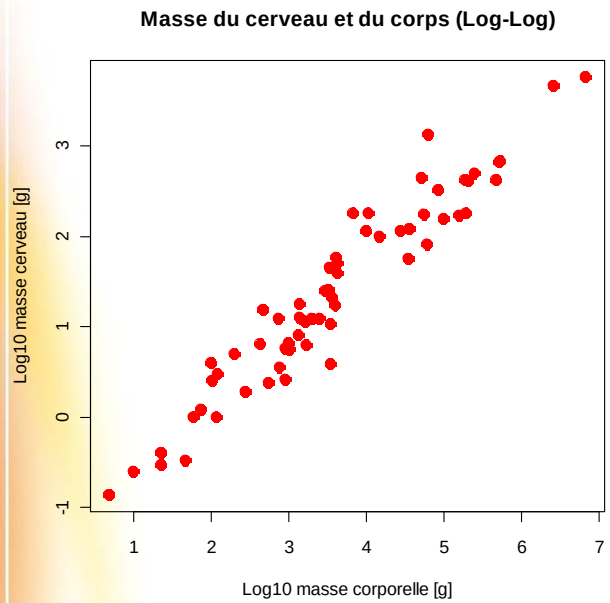
Sur les variables : peut-on résumer les informations par des variables synthétiques = peu nombreuses, non corrélées entre elles, ayant un sens biologique, en éliminant les variations qui n'ont pas de sens (« bruit »)?

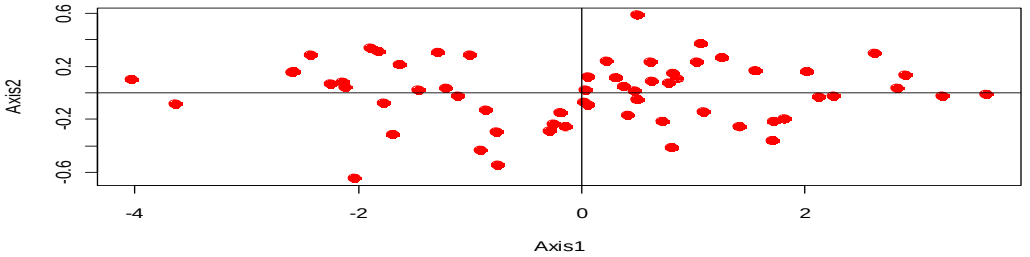
Les questions sur les individus et sur les variables ne sont pas indépendantes : idéalement on peut superposer les typologies : un groupe de variables caractérise un groupe d'individus; un groupe d'individus rassemble les individus types d'un groupe de variables

VetAgro Sup

Principe géométrique

Projection du nuage de points sur des axes :





masse corporelle/masse du cerveau chez 62 Mammifères

VetAgro Sup

Résultat 1 : décomposition de la variance

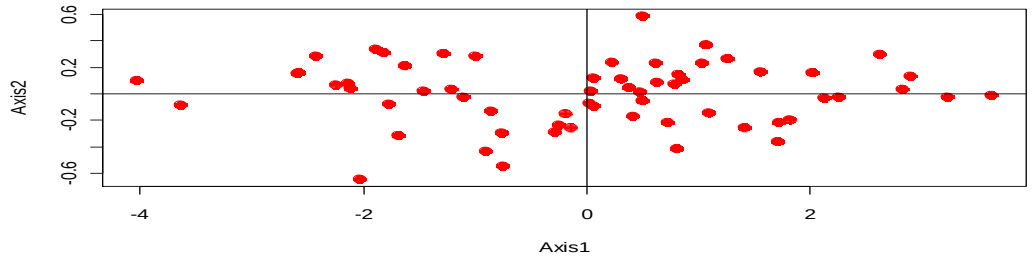
Création d'un certain nombre d'axes / composantes = nouvelles variables obtenues par combinaison linéaire des variables de départ, maximisant l'inertie projetée = résumant l'information

- Indication sur le niveau de structure du tableau: 1, 2 axes suffisent-ils à bien résumer l'information?

- Indication sur les liens variables axes

VetAgro Sup

Résultat 2: représentation graphique



- Une image permet de visualiser des tendances, des regroupements, impossibles à discerner sur un tableau de données

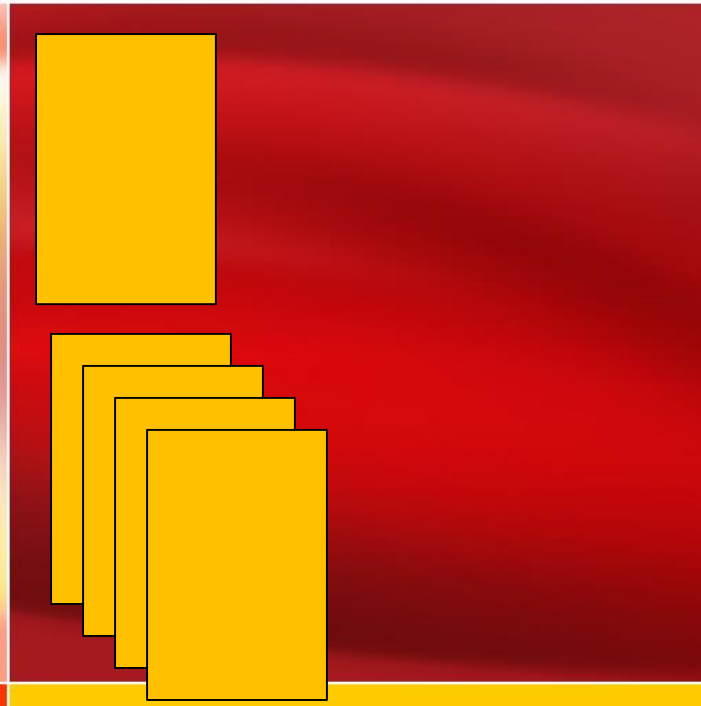
- Communication: graphes de compréhension intuitive… mais difficulté de comprendre « ce qui est derrière »

VetAgro Sup

2. Types d'analyses

1 tableau variables-individus

Plus compliqué



VetAgro Sup

Un tableau variables-individus

- Variables quantitatives mesurées sur des individus (animal, population/site, période, espèce) : Analyse en Composantes Principales ACP
- Variables qualitatives (modalités non ordonnées) mesurées sur des individus : Analyse des Correspondances Multiples ACM
- Distributions mesurées sur des individus (=table de contingence) : Analyse Factorielle des Correspondances AFC

VetAgro Sup

Analyse en composantes principales

masse/longueur/largeur/rendement

n cocons

Résultats de la mesure de chaque cocon

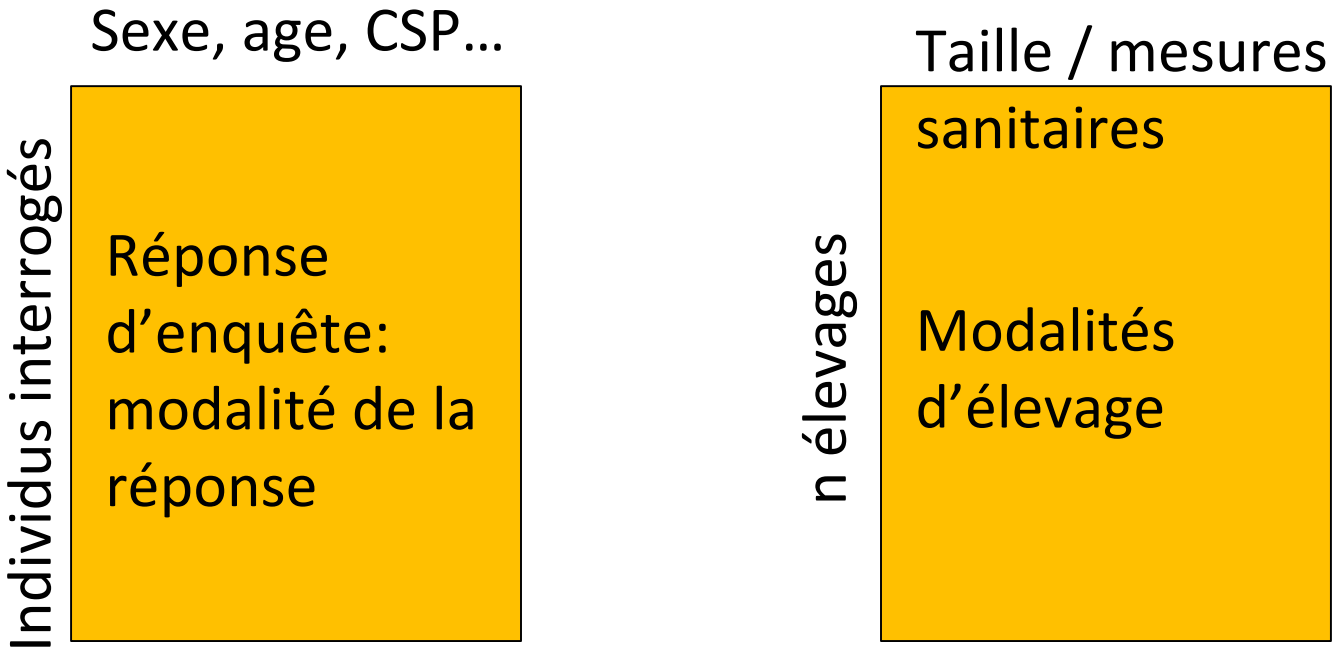
masse/durée du sommeil…

n espèces de Mammifères

Mesure moyenne

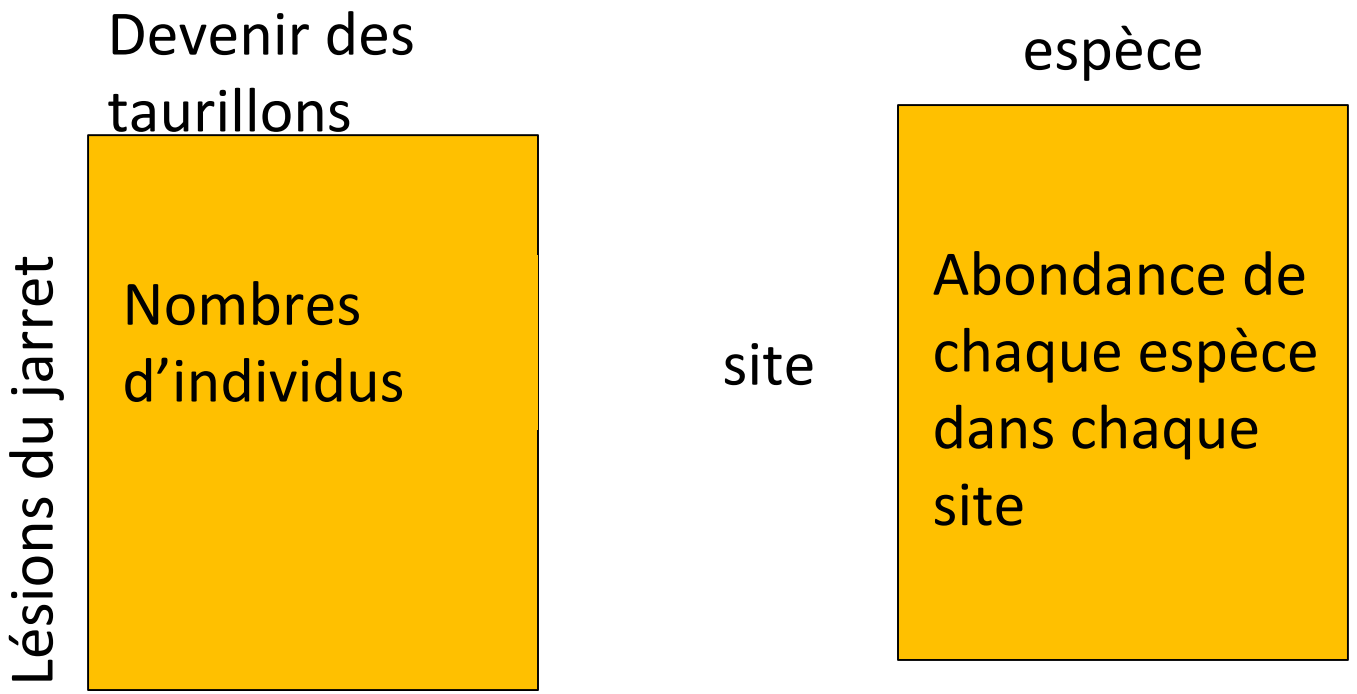
VetAgro Sup

Analyse des correspondances multiples



VetAgro Sup

Analyse factorielle des correspondances



VetAgro Sup

Analyses mixtes

Ex: Hill et Smith

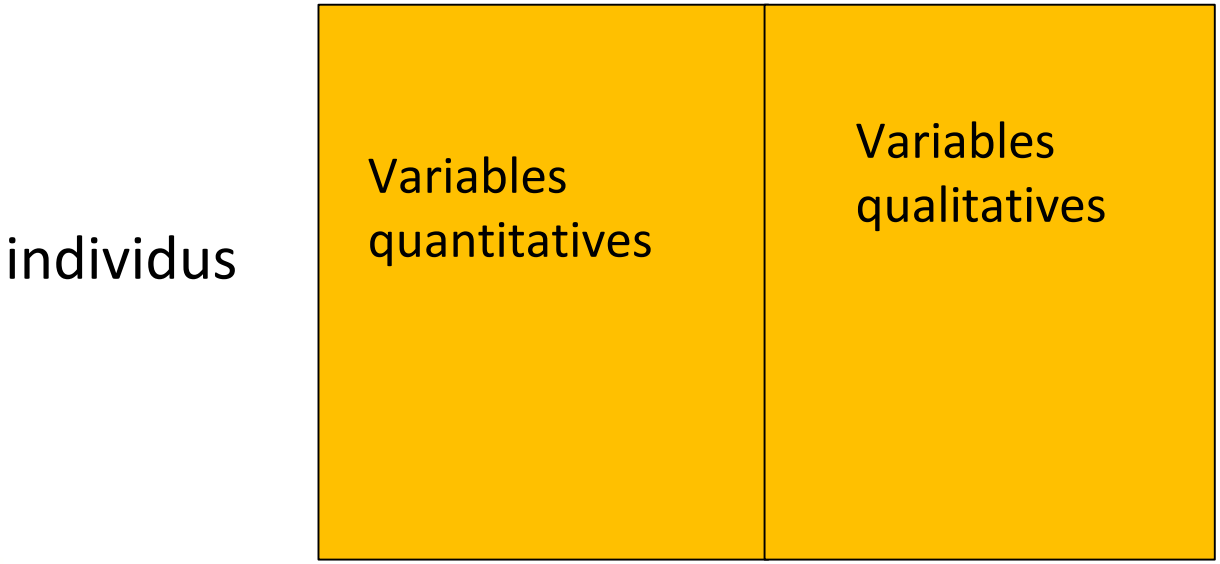
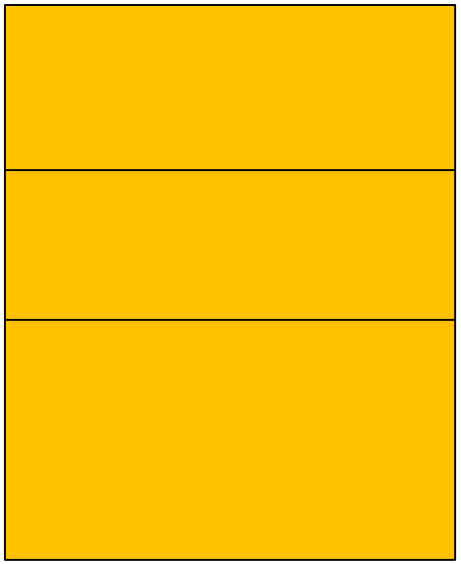


Tableau contenant une structure des lignes

- Structure connue à analyser :
  - La structure des données est-elle identique d'un groupe à l'autre : analyses inter-intra classes
  - Quelles variables sont associées à cette structuration : analyse discriminante



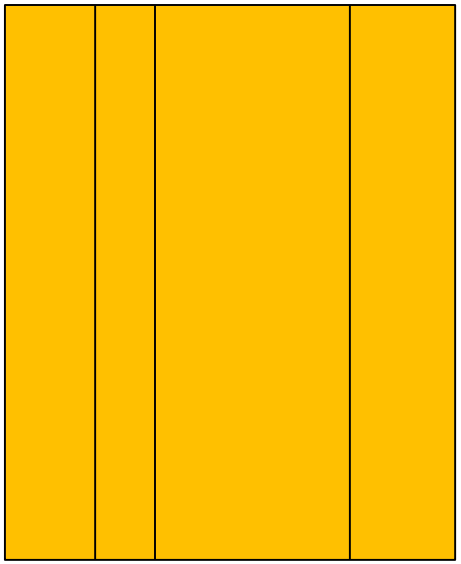
3 sites
3 dates
3 groupes

- Structure à rechercher : classification

Structure des colonnes

Plusieurs groupes de variables

- analyse factorielle multiple
  AFM



Plusieurs tableaux

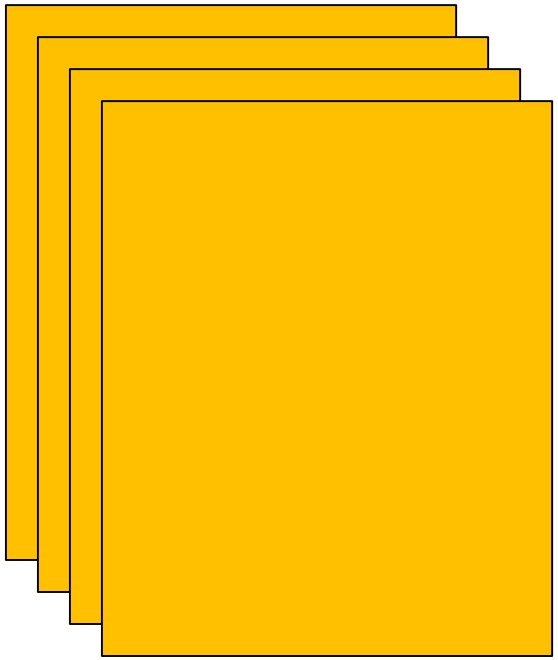
ex: variables mesurées par site et date

analyse canonique (Hotelling 1936) : combinaisons des variables de chaque tableau telles que la corrélation entre ces axes est maximale

Méthodes 2 tableaux : analyse de coinertie

Méthodes k-tableaux : coinertie multiple, analyse factorielle multiple (AFM), y compris hiérarchique (AFMH)…

Cas particulier des analyses sous contraintes spatiales



VetAgro Sup

L'ACP

1. Le nuage des individus
2. Le nuage des variables
3. L'ajustement des nuages
4. ACP centrée / normée
5. L'interprétation
6. Exemples

VetAgro Sup

1. Le nuage des individus

	1		j		p
1	x_{11}				
i			x_{ij}		
n					x_{np}

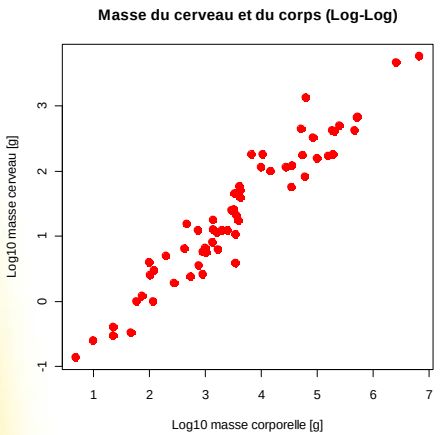
Tableau de données concernant n individus (lignes) mesurés pour p variables quantitatives (colonnes) (notations R)

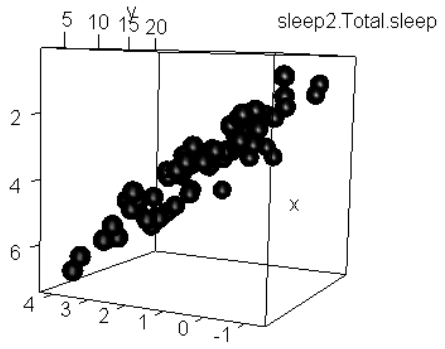
VetAgro Sup

Le nuage des individus

1 individu = 1 ligne = 1 ensemble de p nombres = valeurs prise par l'individu pour les p variables = coordonnées du point-individu sur les p axes du nuage de points

Nuage = ensemble des distances entre les n points : forme?





2 dimensions : ellipse

3 dimensions : ellipsoïde

p dimensions?

VetAgro Sup

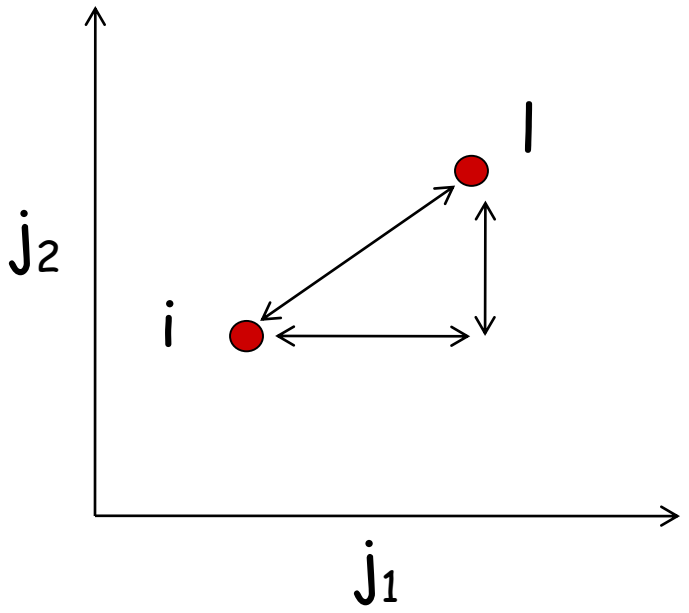
Questions sur les individus

Pour les individus, ressemblance : deux individus se ressemblent d'autant plus qu'ils possèdent des valeurs proches pour l'ensemble des variables

La distance entre deux individus est définie par :

$$d_{i,l}^2 = \sum_j (x_{ij} - x_{lj})^2$$

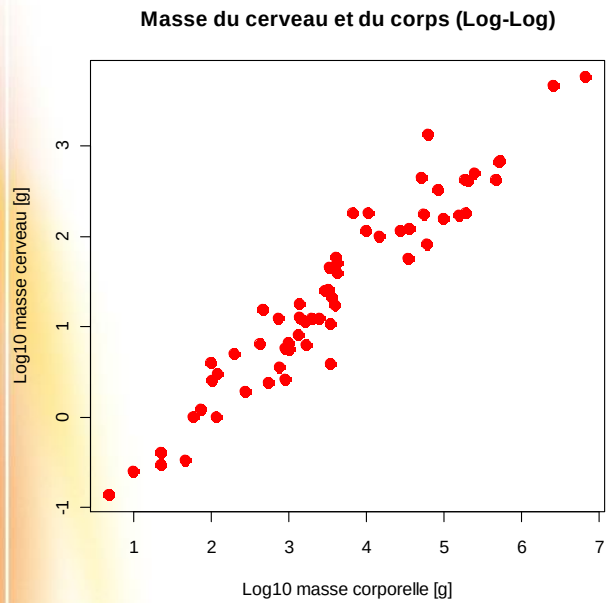
= distance euclidienne

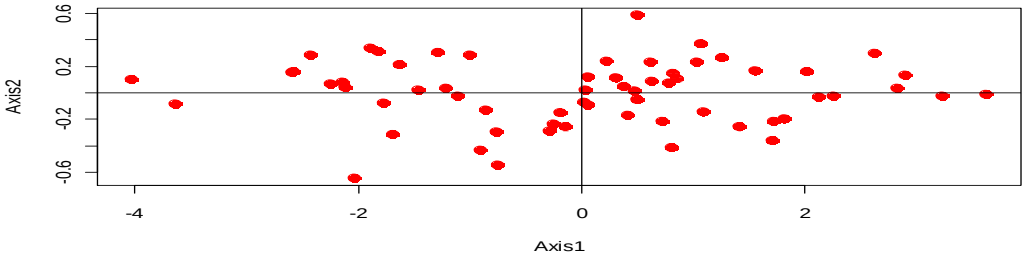


VetAgro Sup

Questions sur les individus

Y a-t-il des ressemblances, des différences, des groupes de lignes, une typologie ?



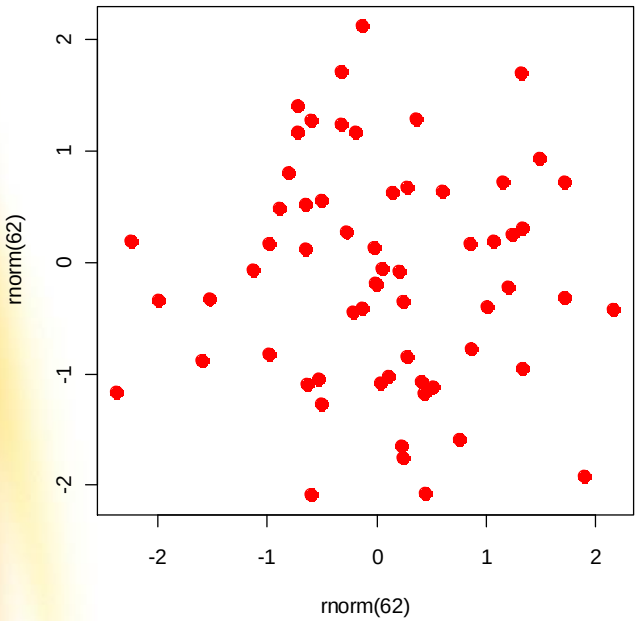


Si oui (ici: gradient), le nuage de points peut être résumé par des axes principaux : un individu est bien représenté par sa coordonnées sur l'axe 1

VetAgro Sup

Questions sur les individus

Y a-t-il des ressemblances, des différences, des groupes?



Si oui, l'information peut être résumée à quelques groupes ou 1 ou 2 gradients

Sinon, un plus grand nombre d' axes est nécessaire pour résumer l'information

On produira une ou plusieurs représentations planes du nuage à p dimensions (une représentation = 2 axes)

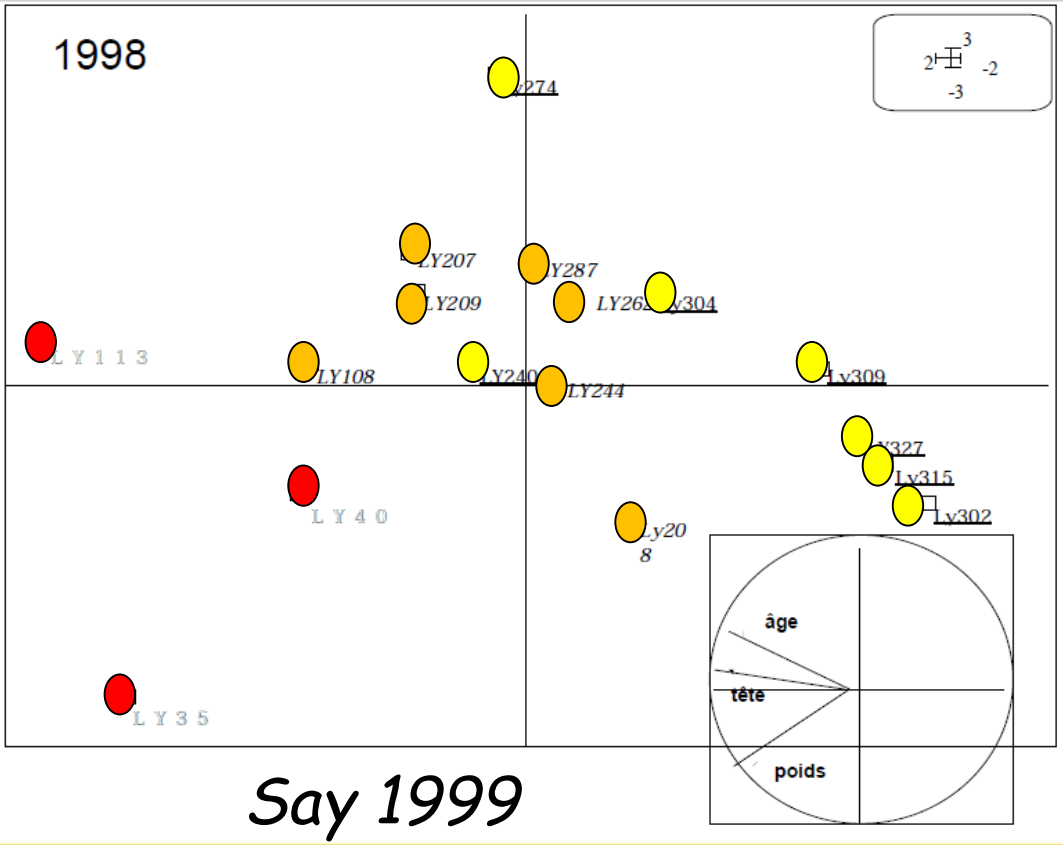
VetAgro Sup

Ex: Taille et dominance des chats

17 chats: masse, âge, volume de la tête

Effet taille : il existe une variable latente qui représente bien le gabarit d'un individu

3 groupes de succès reproducteur (LY113>108>309) bien liés à la taille



Say 1999

VetAgro Sup

2. Le nuage des variables

Pour les variables, ressemblance = corrélation : deux variables se ressemblent d'autant plus qu'elles prennent des valeurs corrélées pour l'ensemble des individus

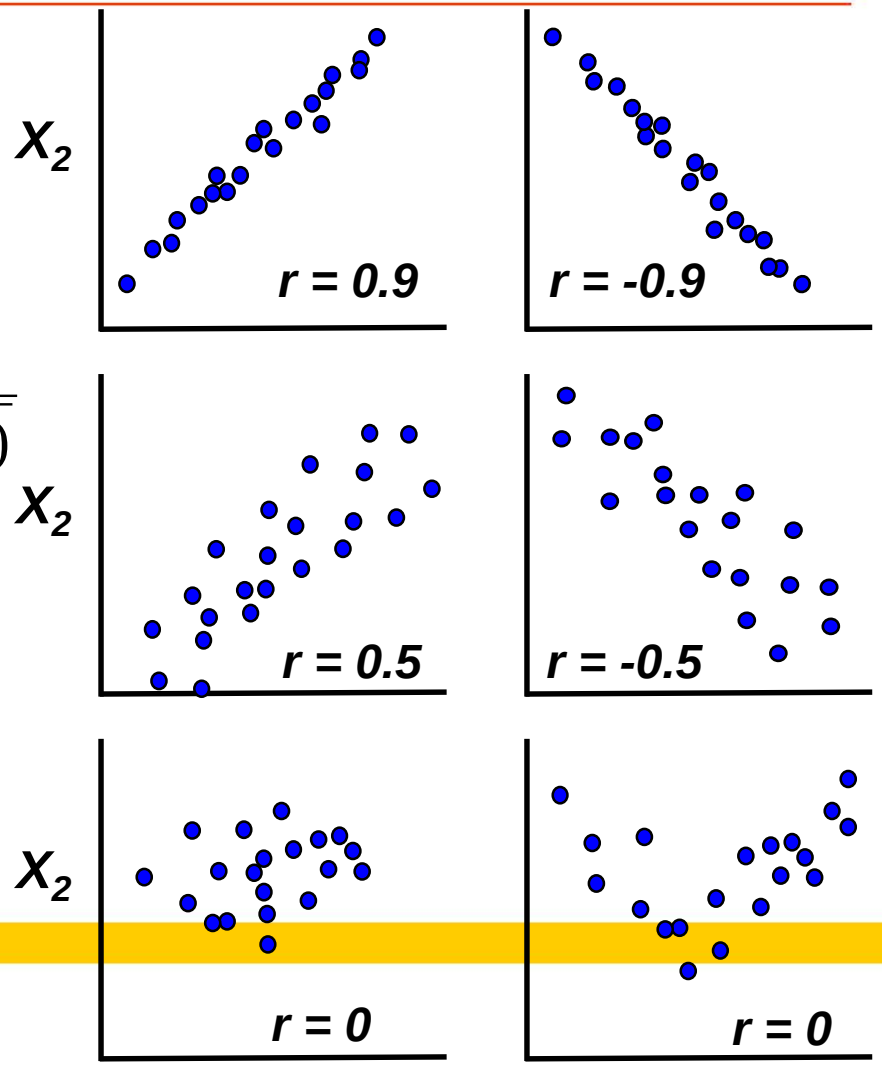
VetAgro Sup

r de Pearson

La distance entre deux variables est définie par :

$$r_{j,k} = \frac{\text{cov}\,ariance(j,k)}{\sqrt{\text{var}\,iance(j)\,\text{var}\,iance(k)}}$$

= coefficient de corrélation de Pearson, Indice de covariance absolu



$-1 \leq r \leq 1$

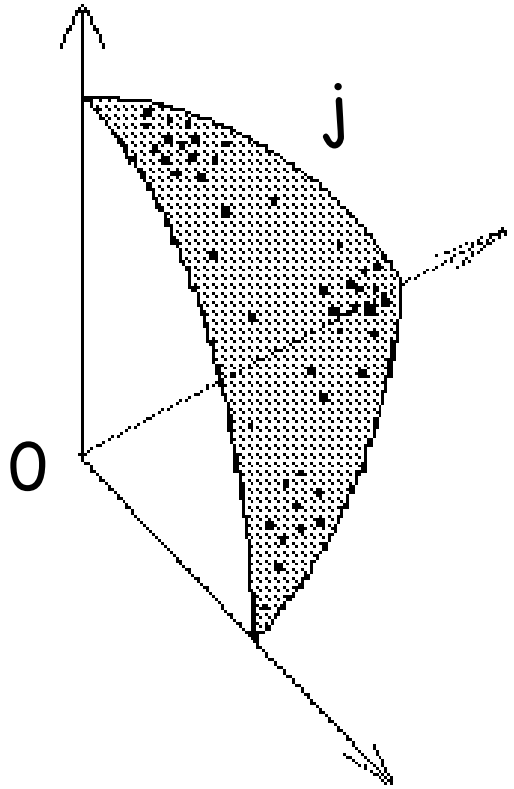
VetAgro Sup

Le nuage de points des variables

1 variable = 1 colonne = 1 ensemble de n nombres = valeurs prises par les n individus pour cette variable = coordonnées du point-variable sur les n axes du nuage de points

Si les variables sont centrées et réduites (de variance 1) => les points sont situés sur une (hyper)sphère de rayon 1 :

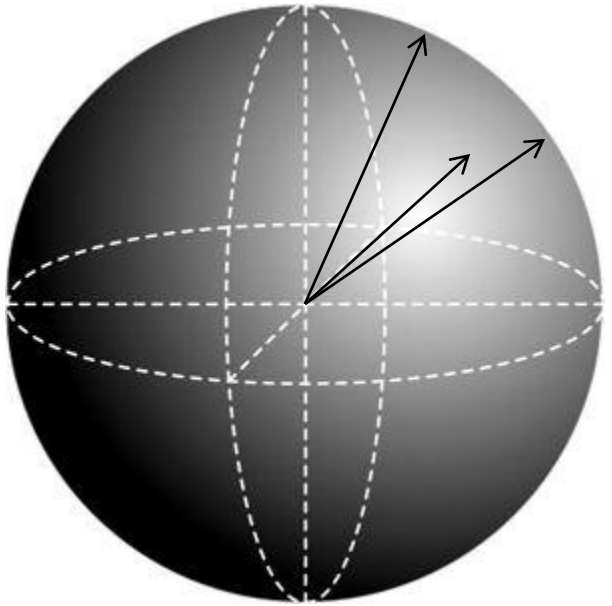
$$d(0,j)^2 = \sum_i \left(\frac{x_{ij} - \bar{x}_j}{s_j}\right)^2 = 1$$



VetAgro Sup

Le nuage de points des variables

Pour les variables, on interprète la direction des vecteurs, pas les distances entre le centre et les points

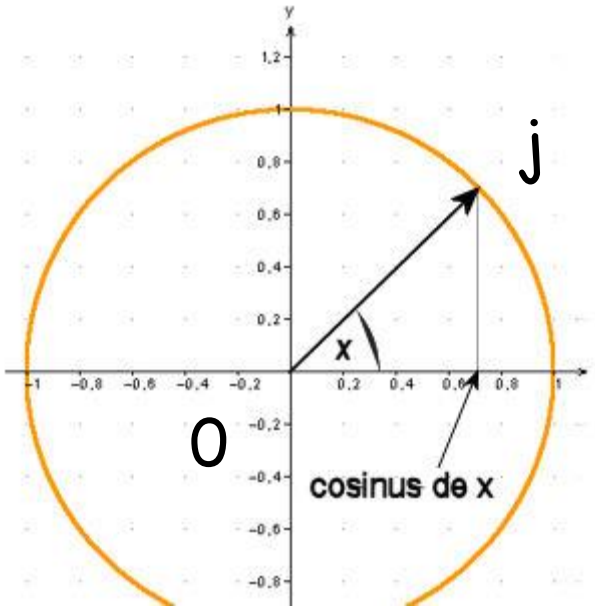


Le nuage de points des variables

De plus, si les données sont centrées, le cosinus de l'angle formé par les vecteurs représentant les 2 variables j et k est égal au r_{jk}

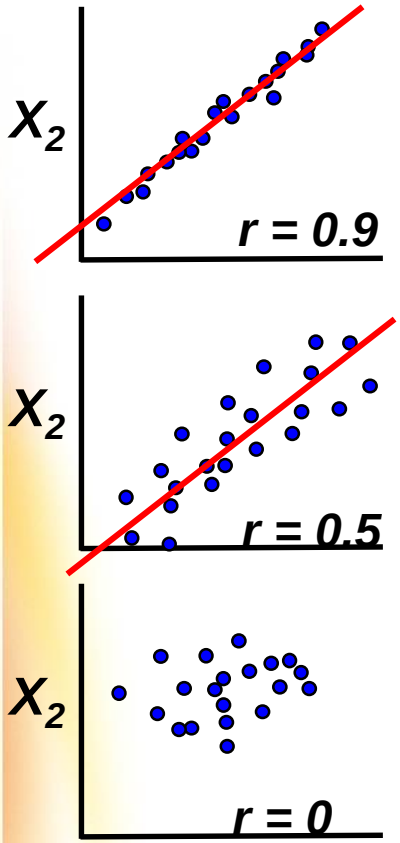
⇒ Deux vecteurs de direction proche représentent deux variables fortement corrélées

⇒ La coordonnée de la projection d'une variable sur une autre s'interprète comme un r



VetAgro Sup

Questions sur les variables



Y a-t-il des corrélations, des oppositions, des groupes?

Si oui, l'ensemble de plusieurs variables peut être représenté par une variable synthétique appelé composante principale, et les projections des variables sur la composante principale seront des corrélations

VetAgro Sup

La relation entre les nuages

La typologie des variables et celle des individus ne sont pas indépendantes car les individus sont décrits par les variables et réciproquement: on cherche :

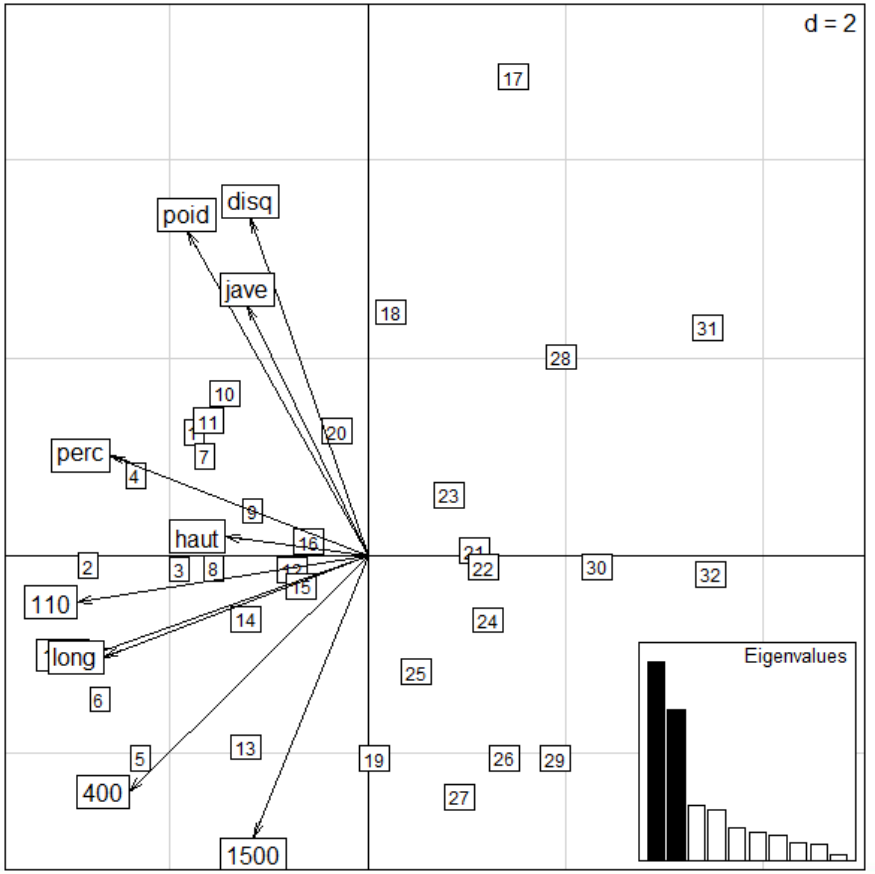
- Des axes principaux
- Des composantes principales
- Une superposition entre les deux

Ex: décathlon

Performances de 33 athlètes dans 10 disciplines

Analyse : il existe une cohérence entre les performances aux différentes disciplines

Synthèse : il y a un effet taille, les coordonnées sur l'axe 1 = variable latente « niveau »



http://pbil.univ-lyon1.fr/R/pdf/tdr51.pdf

VetAgro Sup

3. L'ajustement des nuages

On cherche :

- les axes principaux du nuage des individus
- les composantes principales du nuage des variables

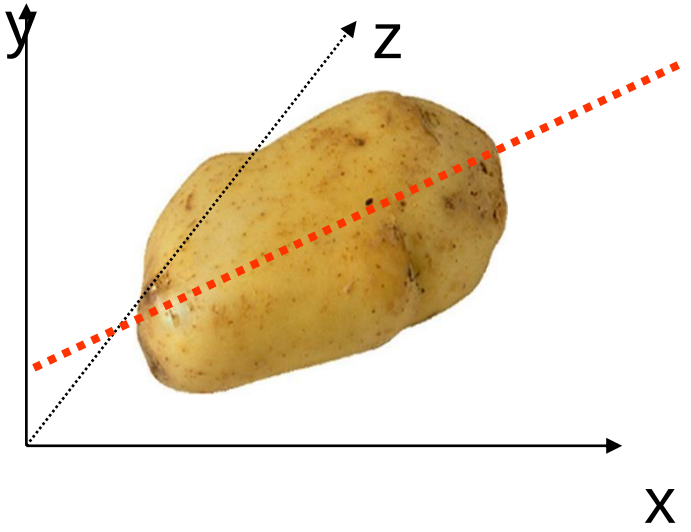
VetAgro Sup

Ajustement du nuage des individus

On cherche une suite privilégiée de directions dans le nuage de points (= axes factoriels) telle que :

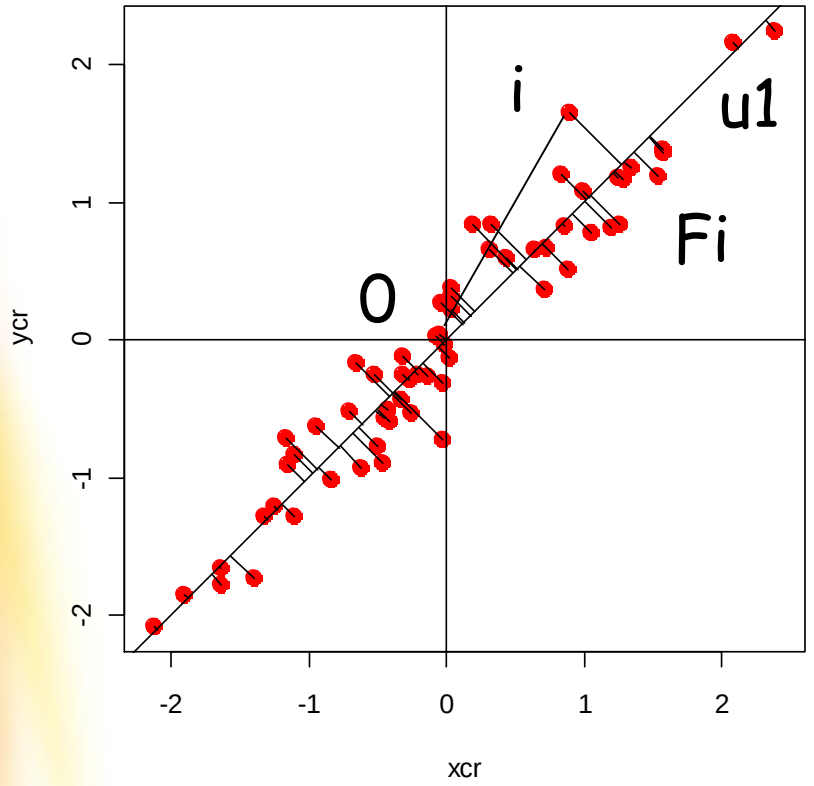
- la projection sur l'axe principal conserve au mieux la forme du nuage = maximise l'inertie projetée du nuage

- chaque axe factoriel est orthogonal au(x) précédent(s)



VetAgro Sup

Approche géométrique



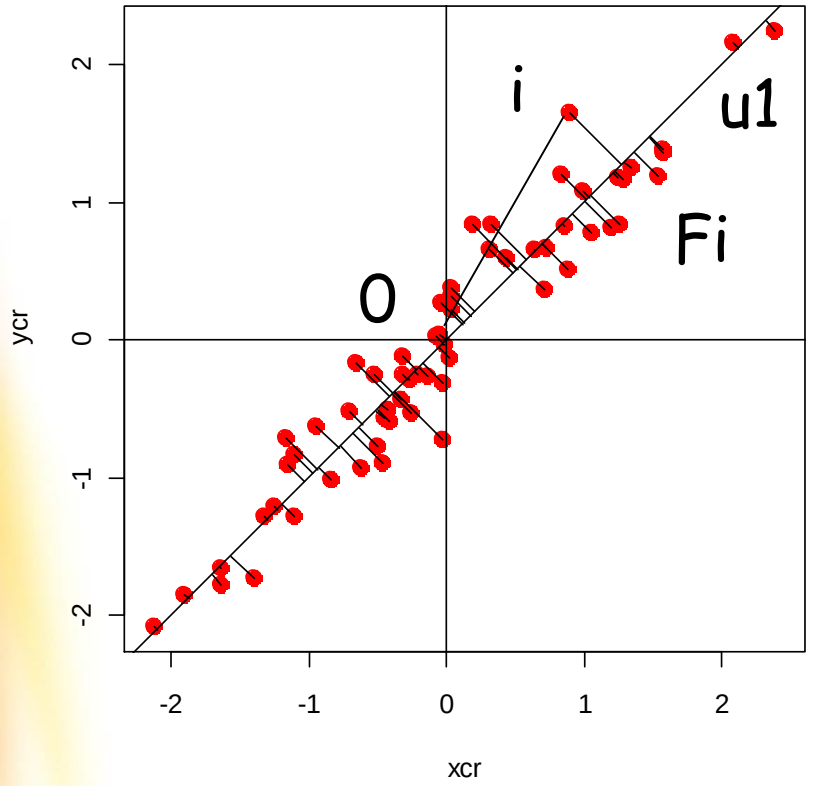
Premier axe factoriel u1 = axe qui maximise l'inertie projetée

Si Fi est le projeté orthogonal de i sur u1, critère :

$$\sum_i OF_i^2$$ maximale

VetAgro Sup

Critère de projection



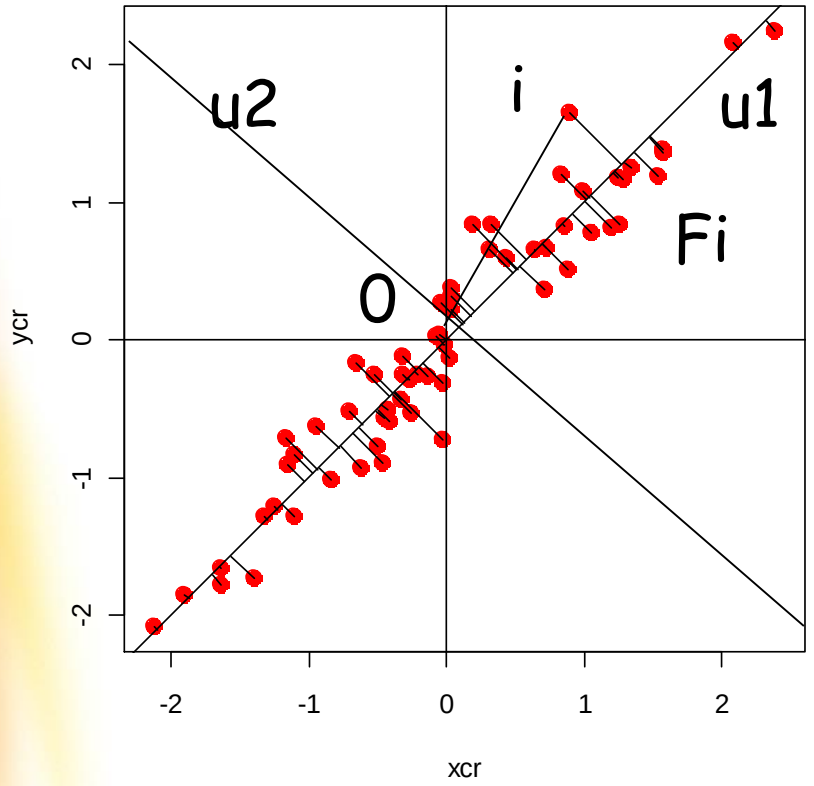
$$\sum_i OF_i^2$$ maximale =>

$$\sum_i iF_i^2$$ minimale

critère des moindres carrés avec une projection orthogonale

VetAgro Sup

Axes suivants



u2 orthogonal à u1 et maximise la variance résiduelle projetée

=> Le premier **plan factoriel** u1,u2 maximise la variance projetée sur 2 axes

VetAgro Sup

Ajustement du nuage des individus

Approche géométrique : sur le premier axe factoriel, l'inertie projetée est maximale => les distances entre points projetés sont aussi semblables que possible aux distances entre points de départ.

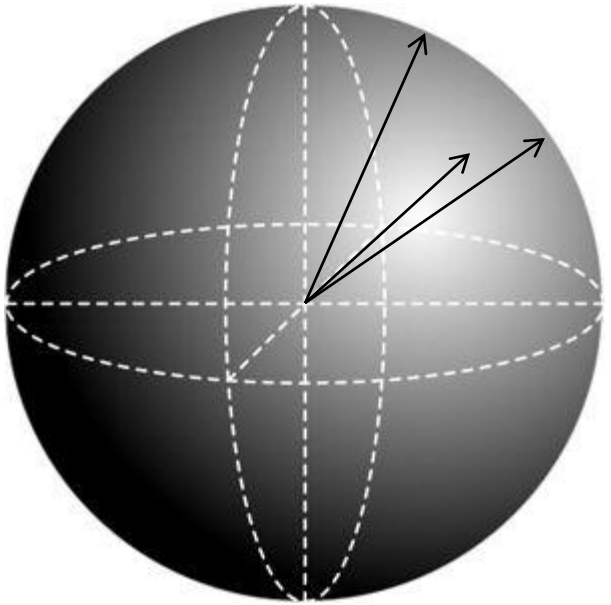
Approche statistique : le premier axe factoriel u_1 est la direction d'allongement maximale du nuage des individus = axe de variabilité maximale, axe de variance projetée maximale

Ajustement du nuage des variables

On cherche une suite privilégiée de variables synthétiques telles que :

la projection sur une composante principale conserve au mieux les relations entre variables = les angles entre vecteurs variables

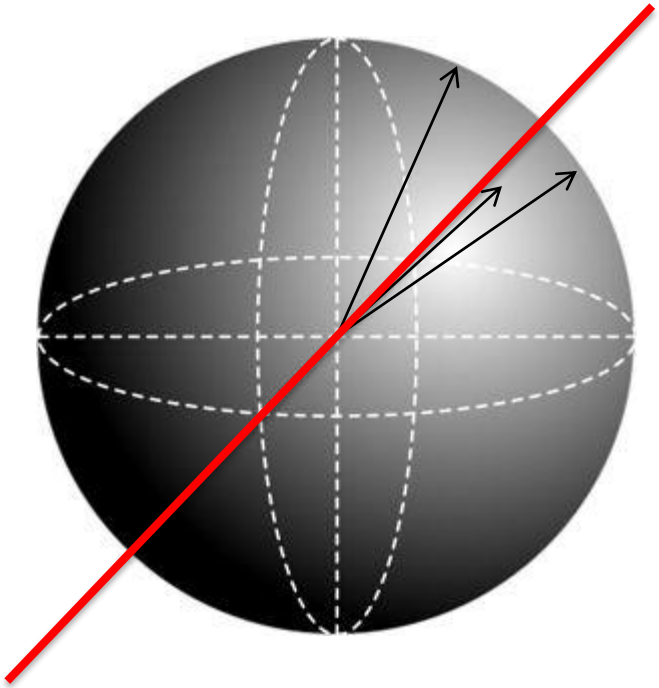
chaque axe factoriel est orthogonal au(x) précédent(s)



VetAgro Sup

Ajustement du nuage des variables

La première composante est la direction d'inertie maximale = direction qui maximise les coefficients des corrélations avec les variables de départ



Ajustement du nuage des variables

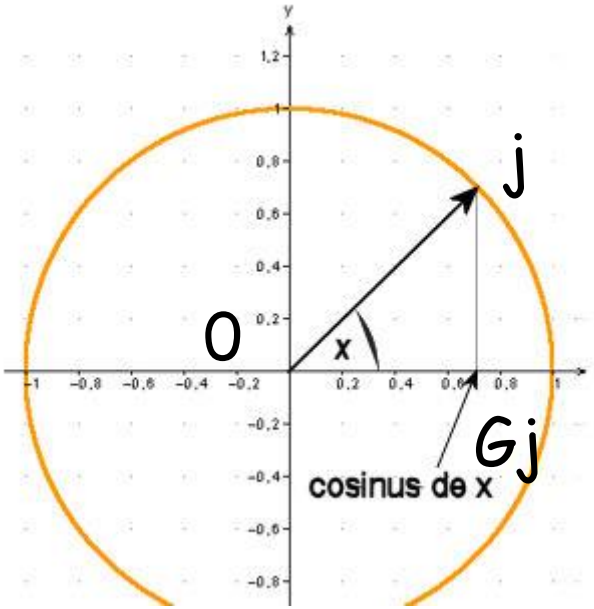
coefficients des corrélations avec les variables maximaux

= cosinus des angles de projection maximaux

= si Gj est le projeté de j sur v1:

$$= \sum_j OG_j^2 \text{ maximale}$$

Critère des moindres carrés



VetAgro Sup

La dualité

Les deux nuages sont deux représentations du même tableau :

- ils ont la même inertie totale

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

= p si les variables sont centrées et réduites

La dualité

Le premier axe principal u_1 du nuage des individus se confond, à la norme près, avec la première composante principale v1 du nuage des variables

VetAgro Sup

Relations de transition

Si l_s est l'inertie projetée sur l'axe/la composante de rang s:

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_j \left(\frac{x_{ij} - \bar{x}_j}{s_j} G_s(j) \right)$$

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_i \left(\frac{x_{ij} - \bar{x}_j}{s_j} F_s(i) \right)$$

Relations de transition

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_j \left(\frac{x_{ij} - \bar{x}_j}{s_j} G_s(j) \right)$$

La projection d'un individu $F_s(i)$ est une combinaison linéaire des projections des variables: un individu apparaitra graphiquement du côté des variables pour lesquelles il a de fortes valeurs (et à l'opposé des variables pour lesquelles il a de faibles valeurs)

La projection des variables peut être une aide à l'interprétation de la projection des individus: deux individus sont proches parce qu'ils ont de fortes valeurs pour les mêmes variables

VetAgro Sup

Relations de transition

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_i \left(\frac{x_{ij} - \bar{x}_j}{s_j} F_s(i) \right)$$

La projection d'une variable Gs(j) est une combinaison linéaire des projections des individus;

Les individus peuvent intervenir comme aide à l'interprétation des relations entre variables: si deux variables sont corrélées c'est parce que certains individus ont des valeurs fortes pour les deux variables.

VetAgro Sup

En pratique

Transformation des données :

- centrage : $$x_{ij} \rightarrow x_{ij} - \bar{x}_j$$

=> Ne modifie pas le nuage, mais le centre, pratique pour la représentation graphique

- réduction : $$x_{ij} \rightarrow \frac{x_{ij} - \bar{x}_j}{s_j}$$

=> Toutes les variables présentent la même variabilité et donc le même poids dans le calcul des distances inter-individus

VetAgro Sup

En pratique…

On cherche une projection de *M* (matrice des données) sur un vecteur *u* avec variance projetée maximale

La variance projetée de *M* sur *u* est fonction de la matrice des covariances (ou des corrélations) *C*, qui est carrée symétrique réelle => diagonalisable. Diagonaliser = obtenir:

λ : valeurs propres (racines du polynôme caractéristique det(C-λI)) : $\lambda_1 =$ inertie projetée sur le premier axe

X: vecteurs propres associé à chaque valeur de λ : base orthonormée de facteurs principaux

Changement de base: valeurs propres et vecteurs propres permettent d'obtenir les coordonnées des variables et des lignes sur les facteurs

VetAgro Sup

4. ACP centrée / normée

Lorsque

- Tous les individus ont le même poids
- les coordonnées des individus sont centrées
- les ressemblances entre variables sont exprimées par la corrélation

- ACP normée

VetAgro Sup

ACP centrée

= non normée (l'ACP normée est aussi centrée!)

Les données ne sont pas réduites, la ressemblance entre variables s'exprime par la covariance

=> une variable prend d'autant plus de poids dans l'analyse qu'elle présente des valeurs élevées => sensible au choix des unités de mesure

VetAgro Sup

Poids des individus et des variables

On peut souhaiter modifier le poids des individus:

- Individus « supplémentaires » ou illustratifs (ex: moyennes de groupes, individus particuliers ou hors de l'ensemble étudié) : poids 0

- Individus prépondérants: rare en ACP (cf AFC) : l'hypersphère n'est plus une sphère

On peut aussi ajouter des variables supplémentaires: variables extérieures au jeu de données étudié, projetées sur les axes déterminés par les variables actives: visualiser les corrélations

5. L'interprétation

- rendre clair
- donner un sens
  - Sens d'un élément au sein du tableau
  - Sens par rapport à des éléments supplémentaires (références)
  - Sens par rapport au contexte extérieur aux données analysées
- Domaine complexe qui laisse place à une approche personnelle

VetAgro Sup

Bases d'interprétation

Basée sur :

Etude de l'inertie des facteurs

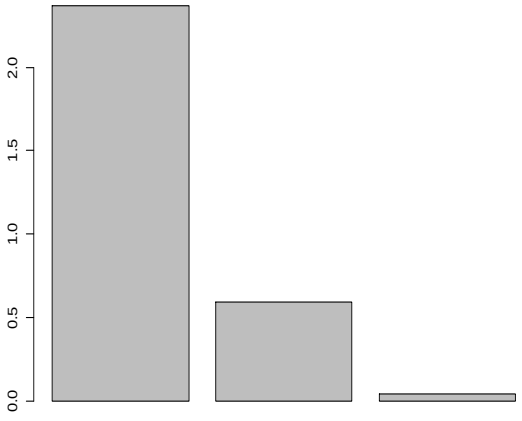
projection du nuage des individus sur les axes, projection du nuage des variables sur les composantes et superposition des 2 nuages

éléments complémentaires d'interprétation : « aides à l'interprétation »

VetAgro Sup

Etude de l'inertie des facteurs

Valeurs propres = λ_s = (eigenvalues) = inertie projetée sur chacun des facteurs, diagramme en bâtons



Plus λ_s est grande, plus elle résume de variabilité et plus le facteur s est intéressant. 1ère valeur propre comprise entre 1 (si variables indépendantes) et p (si variables corrélées)

Autant de valeurs propres que d'axes de projection = p

ACP normée: somme des λ_s = p

ACP centrée: somme des λ_s = variance du tableau

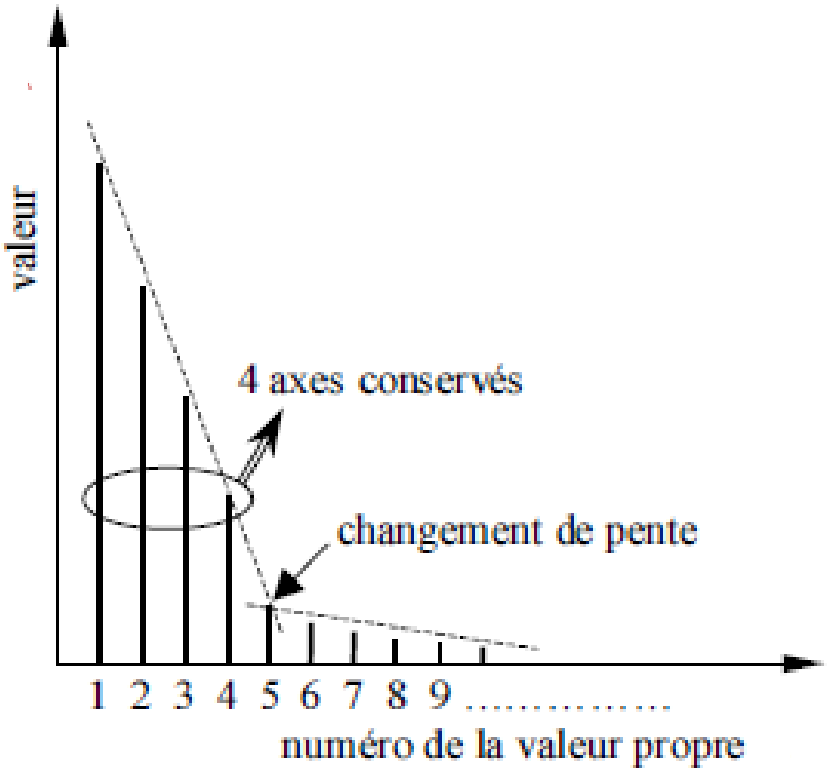
Etude de l'inertie des facteurs

Interprétation = valeur relative

(mais repère: une valeur < 1
synthétise moins d'info
qu'une variable seule)

Idée : les facteurs les plus
importants (au sens de
l'inertie) sont ceux à gauche
de la rupture de pente

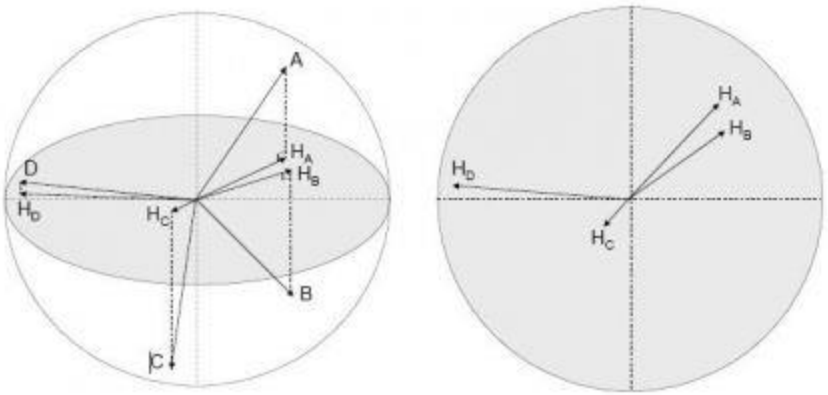
(mais ils peuvent contenir une
information triviale)



VetAgro Sup

Cercle des corrélations

Représentation liant variables et composantes principales



4 variables A, B, C, D

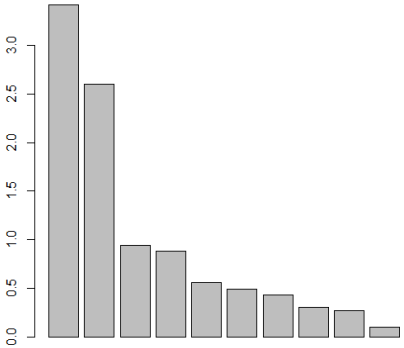
Projection de l'hypersphère de rayon 1 sur 2 composantes (en ACP normée seulement)

Le cercle permet de voir quelles variables sont bien représentées sur ce plan (proche de la circonférence, D) ou non (C)

Les vecteurs proches sont proches par la projection (A et B), les cosinus des angles entre variables et composantes représentent des corrélations

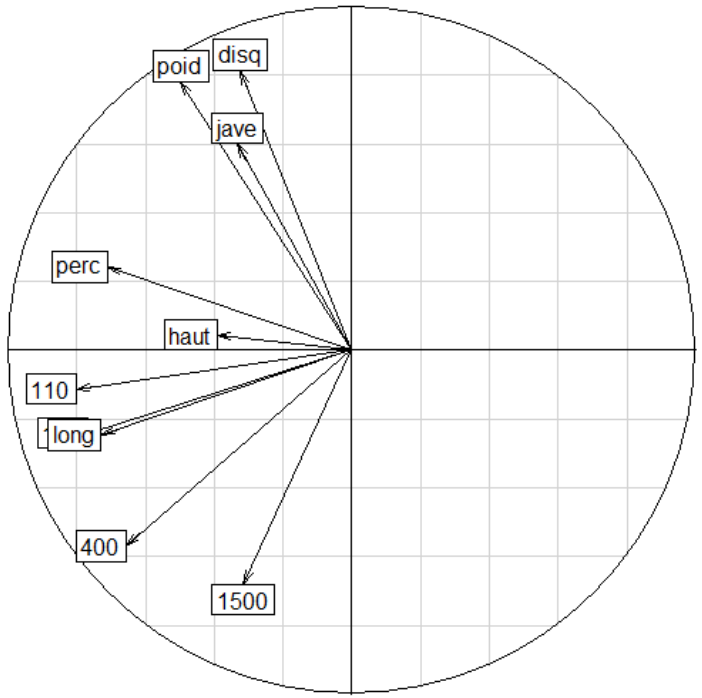
Les vecteurs proches des axes et bien représentés sont liés aux axes (D)

VetAgro Sup



Exemple à interpréter : décathlon

Le premier plan factoriel donne l'essentiel des informations



Toutes les variables sont bien représentées sur ce plan (sauf hauteur) : plan informatif

Attention, il y a 2 composantes!

Axe 1 : corrélation positive entre toutes les variables = effet taille (ici niveau); meilleure variable pour le représenter = 110m

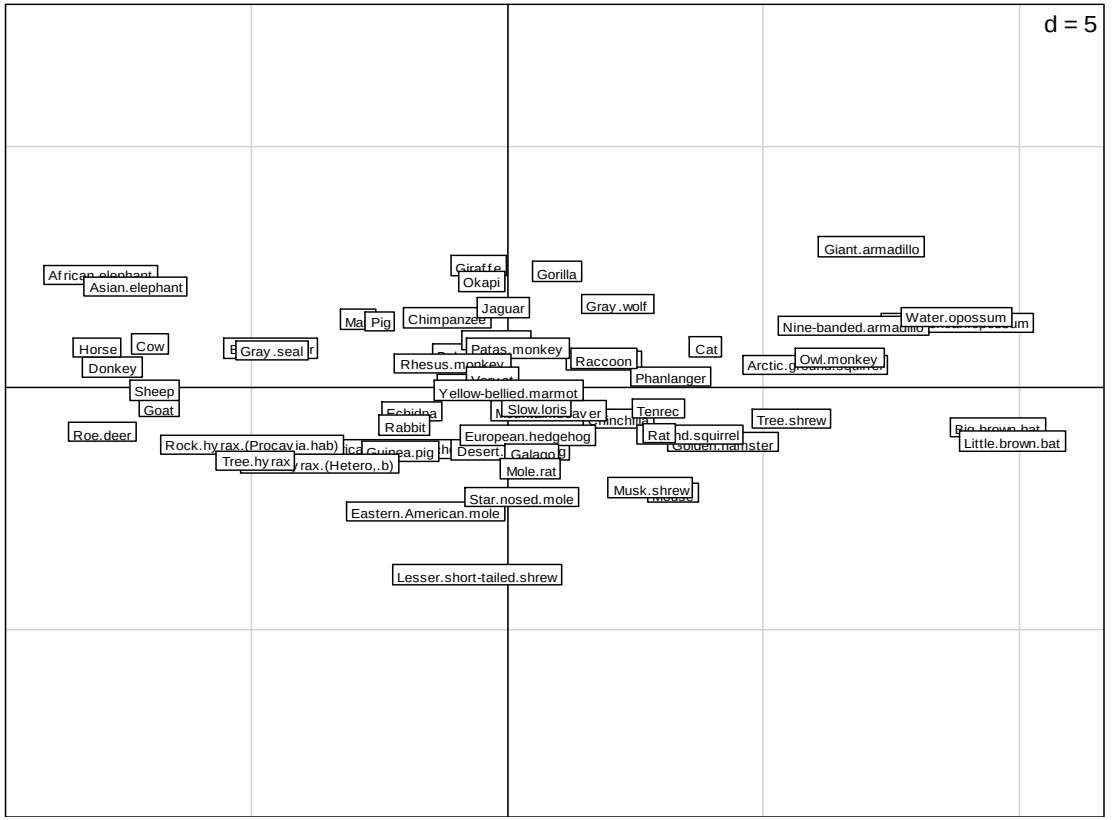
Axe 2 : corrélation négative entre les performances de lancer et de ½ fond, gradient indépendant des performances de course et saut

Saut en hauteur = compétence indépendante des autres

Carte factorielle

Représentation de la projection du nuage sur 2 axes principaux: chaque point (=ligne) est représenté par ses coordonnées sur u1 et u2

1er plan factoriel : carte sur les axes 1 et 2

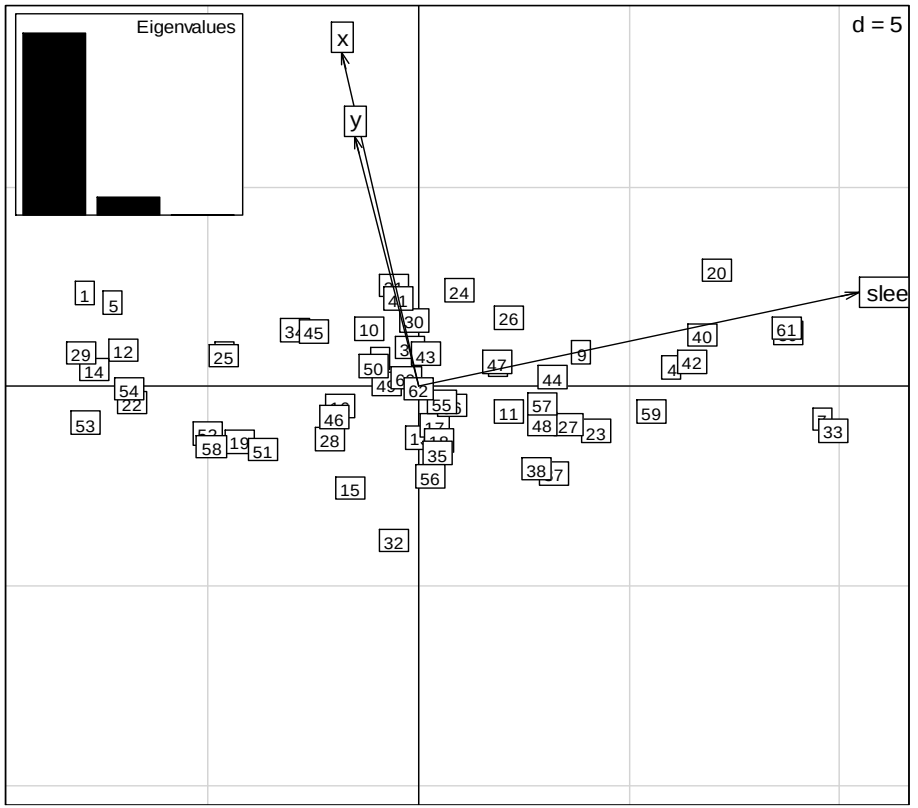


VetAgro Sup

Superposition

A droite: les espèces ayant la plus longue durée de sommeil

Mais attention: points et vecteurs sont de nature différente



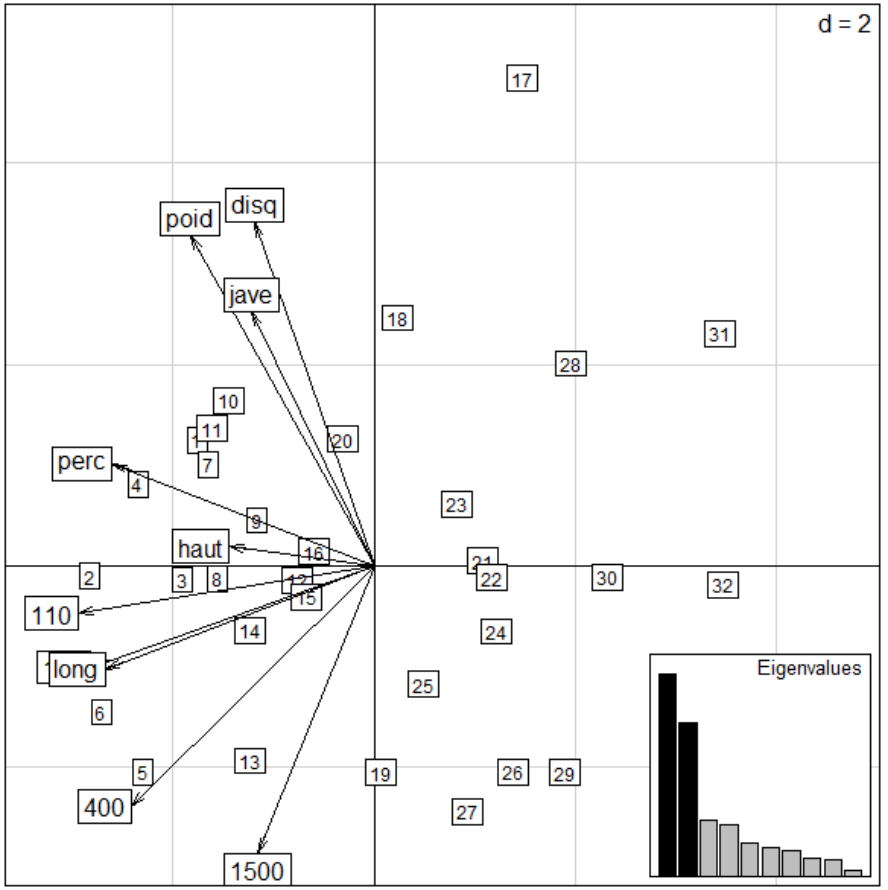
VetAgro Sup

Exemple à interpréter : décathlon

De gauche à droite :
classement au score (approximatif! Les 10 dimensions sont résumées par 2)

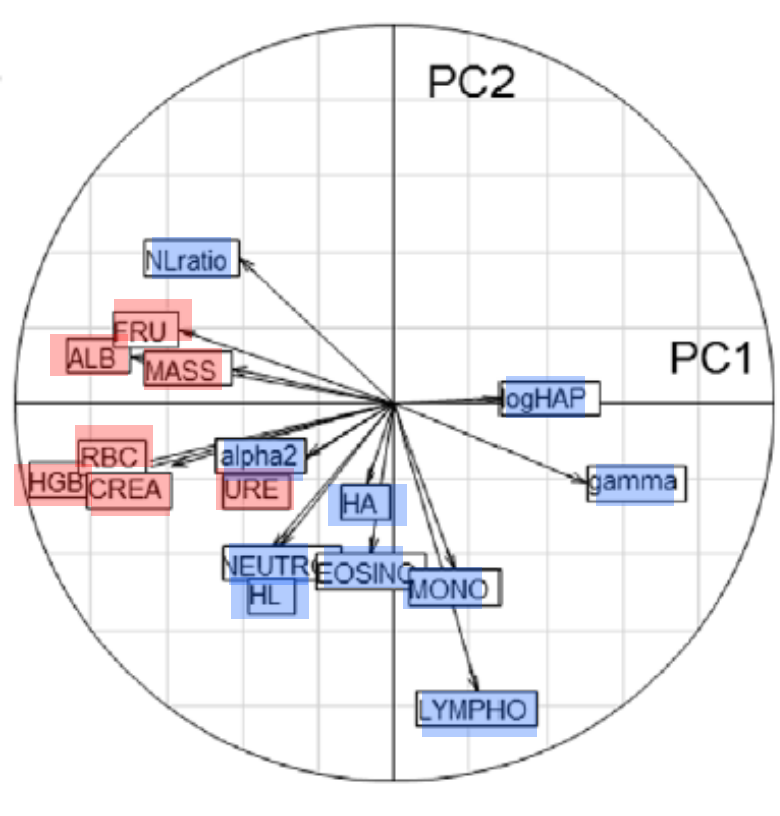
De haut en bas : par profil d'athlète

Lien entre les deux dimensions : les profils atypiques sont à droite



VetAgro Sup

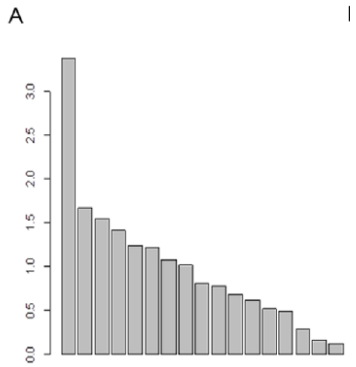
Ex : condition et immunité chez le chevreuil

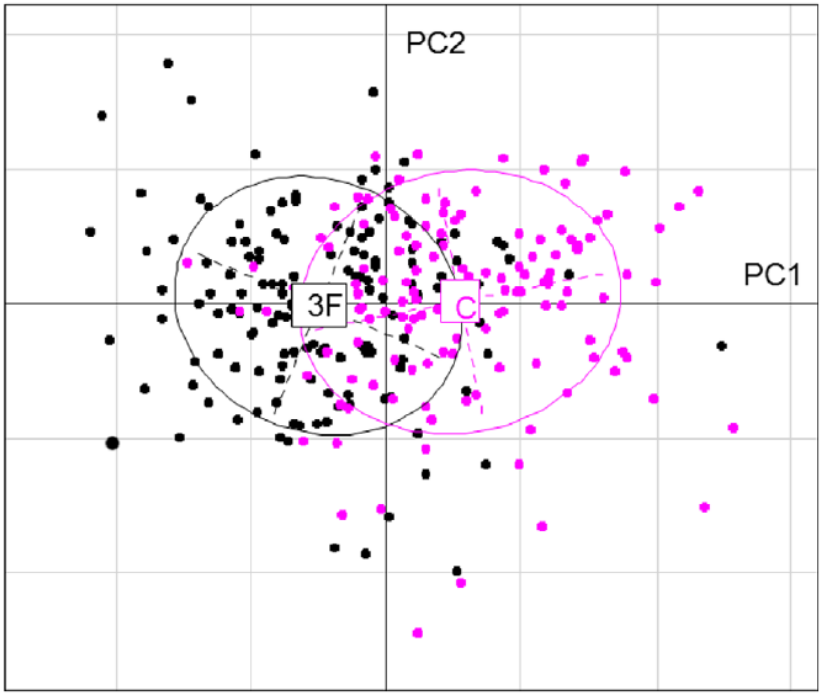


285 chevreuils issus de 2 populations (3Fontaines, Chizé)

Condition corporelle (7)

Immunité (10)





VetAgro Sup

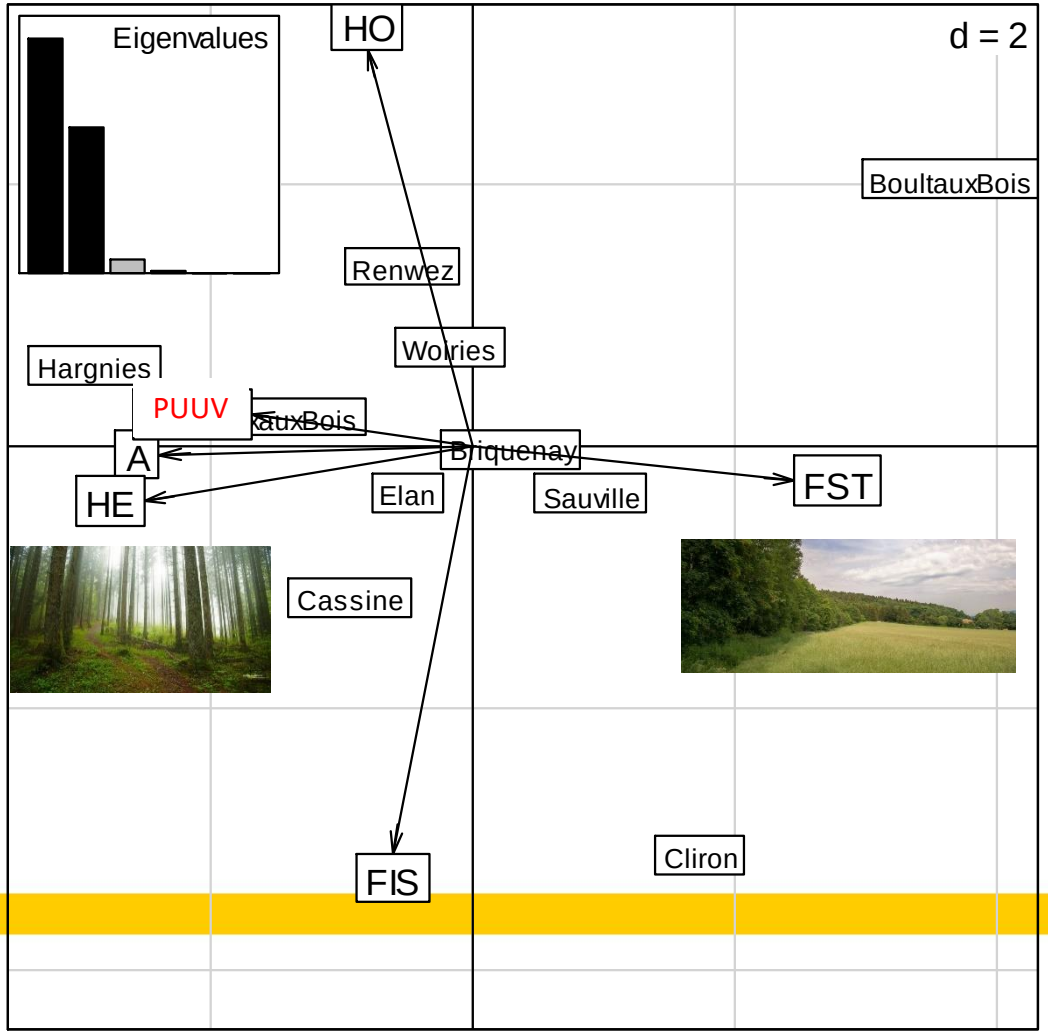
Ex: Structure génétique et épidémiologie chez le campagnol roussâtre

10 populations caractérisées par 5 variables génétiques, notamment A = diversité allélique, HE = hétérozygotie, FST = isolement génétique

Axe 1 : 58% de la variabilité

Prévalence du hantavirus Puumala = variable supplémentaire

Guivier et al. 2011



VetAgro Sup

Aides à l'interprétation

Contribution (relative) d'un axe à la représentation d'un individu ou d'une variable

$\cos^2$ (angle Oi, s)

surtout utilisé pour les variables: visible sur le cercle des corrélations

Contribution (absolue) d'un individu/d'une variable à la définition d'un axe

Inertie de la représentation de i sur s / inertie projetée sur s

Pas systématiquement utilisées en ACP

VetAgro Sup

TD

VetAgro Sup