Introduction aux statistiques multivariées

Parfois aussi nommées méthodes *d'analyse de données* (ce qui constitue un abus de langage et n'a pas beaucoup de sens...). On les inclut aussi souvent dans l'ensemble du "*machine learning*" = apprentissage machine (ce qui n'a guère de sens non plus). Toutes ces méthodes appartiennent en réalité à l'ensemble des statistiques.

METHODES DESCRIPTIVES :

- L'analyse en composantes principales (ACP) cherche à représenter dans un espace de dimension faible (<< *p*) un nuage de points représentant *n* individus, ou objets, décrits par *p variables quantitatives* (donc de dimension *p*) en utilisant les corrélations existant entre ces variables.
- L'analyse des correspondances (AFC ou ACM) étudie les proximités entre individus décrits par deux ou plusieurs *variables qualitatives* ainsi que les proximités entre les modalités de ces variables.
- Les méthodes de classification (*clustering*) ou de typologie procèdent par regroupement des individus en classes homogènes (classifications hiérarchiques, arbres phylogénétiques, moyennes mobiles (*K-means*), ...).

METHODES EXPLICATIVES ET PREDICTIVES :

- L'analyse discriminante (AFD) étudie la prévision d'une variable qualitative par des variables numériques. C'est une méthode géométrique en espace réduit.
- Les arbres de décision et régressions (glm) étudient la prévision d'une variable qualitative ou quantitative dépendante par une combinaison linéaire de variables explicatives (modèles de régression)

METHODES PUREMENT PREDICTIVES :

- Les réseaux neuronaux visent à établir par apprentissage un modèle susceptible d'affecter à un jeu d'essai (différent du jeu d'apprentissage) une qualité (reconnaissance d'image), une valeur (estimation numérique), appartenance à un groupe de malades à partir d'un profil d'expression génétique (micro-arrays ou "puce à ADN")...

l'ACP : de quoi s'agit-il ?

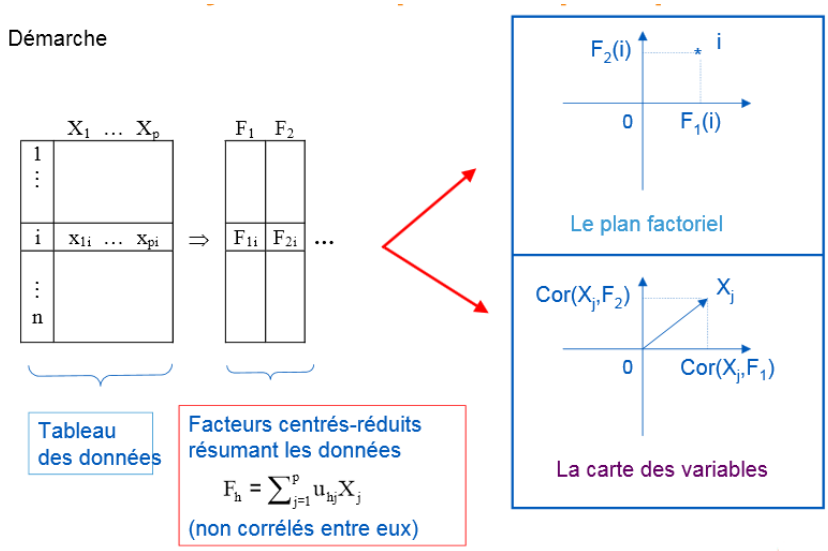
Données types. Un tableau ayant en colonnes *p* variables quantitatives (métriques) et en lignes *n* individus

	1.....***j***.....***p***				
1					
.					
.					
i			X_{ij}		
.					
.					
n					

X_{ij} = valeur de la variable *j* prise par l'individu *i*

Objectif : résumer l'information de dimension *p* contenue dans ce tableau dans un espace réduit. On prend le risque de perdre une partie de l'information totale pour en extraire la plus importante (la plus informative).

Démarche (très résumée) :



Principe

On substitue aux variables initiales (*X*) des "indices synthétiques" (*F*) qui sont des combinaisons linéaires de ces variables. Ces indices sont appelés axes ou *composantes principales*.

- Le premier axe (F1) sera tel que la variance des coordonnées des individus projetés sur cet axe sera maximale. Il explique donc un % de la variance totale du tableau.
- Le second (F2) sera orthogonal au premier (corrélation nulle avec F1) et aura la variance la plus élevée possible (mais inférieure à la première)
- Et ainsi de suite pour tous les autres (il y a exactement *p* composantes).

L'ACP permet :

Représenter les variables en fonction de leurs corrélations:

- quelles sont les variables *corrélées*, *anticorrélées* et *non corrélées* entre elles.

Représenter les individus en fonction de leurs proximités :

- quels sont les individus *ressemblants* et les individus *dissemblants* (éloignés),
- comment les individus se situent sur les composantes qui sont des *axes synthétiques ayant une signification biologique*, écologique ou autre, dépendant de la nature des variables mesurées et de la nature des individus.

NB :

Il est possible ensuite d'opérer une classification des individus sur la base de leur coordonnées factorielles au moyen d'un algorithme de clustering et d'obtenir ainsi des groupes d'individus sur lesquels on pourra calculer moyennes et variances de chacune des variables qui les caractérisent.

PETIT VADE-MECUM D'INTERPRETATION DE L'ACP

AVANT L'ACP

Quels types de tableaux ?

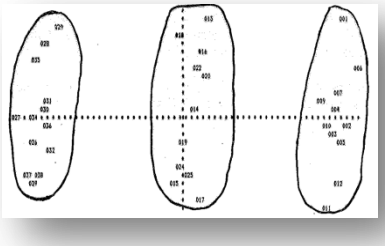
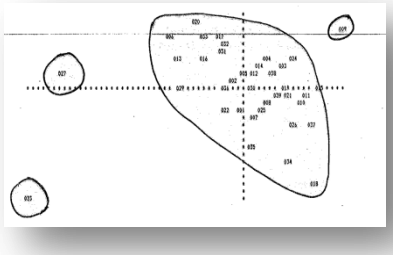
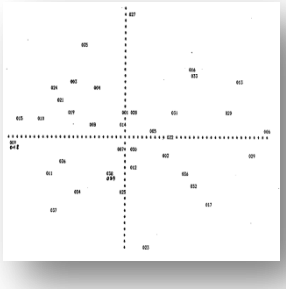
1. Tableaux de **mesures**. (dosages, densités optiques (absorbance), comptages (nombres d'individus, d'événements, etc…). Les variables (descripteurs) sont continues ou entières, *quantitatives.*
2. Tableaux de **notes**. Evaluation d'intensité de maladie, de qualité, etc. L'œil (l'odorat, le toucher…) remplace l'instrument de mesure. L'intervalle de notation doit être suffisamment grand (8 à 10 classes de notes au moins). Il s'agit de *variables qualitatives ordinales*[1].
3. Tableaux de **rangs**. Les *variables sont des rangs*, les n observations sont classées de 1 à n (du plus faible au plus fort, du plus rapide au plus lent, etc.)

Remarque : l'ACP est fortement influencée par l'ordre de grandeur des variables. En général, les variables ayant les plus grandes variances engendrent les premières composantes. En cas de fortes hétérogénéité des dimensions (mesures dans des unités différentes) on recommande d'effectuer l'ACP sur des données centrées-réduites (matrice des corrélations). Sinon, effectuer l'analyse sur les données centrées seulement (matrice de dispersion). Cette remarque est sans objet pour les tableaux de rangs (variances identiques). Attention : une variable à faible variance (donc de peu d'intérêt et devant être éliminée avant analyse) se retrouvera avec un poids équivalent aux autres variables après normalisation par centrage et réduction, ce qui n'est pas souhaitable.

Examiner les données

Réaliser des histogrammes pour chaque variable. L'utilisation du plot pairs(données), permet de prendre connaissance des données et détecter des erreurs de mesure et autres. Examiner les variances et écart-types de chaque variable.
Notez que si des variables sont très fortement corrélées, elles sont redondantes. Deux variables fortement corrélées disent la même chose ! Il convient peut-être d'en éliminer une.
Décider des variables et observations supplémentaires. Dans l'exemple du blé tendre, 10 variables ont été mesurées. La 11ème variable (Rendement= Rdt) peu représenter une variable à expliquer par les 10 précédentes. Dans ce cas, on optera pour la mise de rendement en « variable supplémentaire », les 10 autres variables étant qualifiées d' « actives ». Rdt ne participera pas à l'analyse mais sa projection aura du sens. De même, imaginons que 3 nouvelles variétés soient mises sur le marché. Comment se positionnent-elles par rapport aux 39 disponibles ? Là encore, il est possible d'opter pour leur mise en « observations supplémentaires ».

Une première analyse

Cas 1 : plusieurs sous populations. Il faut réaliser une ACP pour chacun des groupes.
Cas 2 : Des données erronées ou bien des observations sont atypiques. Les éliminer de l'analyse.
Cas 3 : OK, on poursuit par l'interprétation.

INTERPRETATION

1. Examiner les statistiques élémentaires (moyennes, écart-types et corrélations)
2. Examiner les valeurs propres ou mieux les pourcentages de variation expliqués par chaque composante principale (histogramme). Ceci détermine les plans à examiner.
3. Sur les plans retenus :
   - La structure des variables (cercle des corrélations)
   - La répartition des individus à partir de leurs coordonnées sur les axes principaux et les cos² d'angles indiquant la **qualité de leur représentation** dans le plan considéré.

[1] Ce type de données peut être aisément traité par une méthode alternative : l'Analyse Factorielle de Correspondances (AFC), y compris les tableaux de présence-absence (0 ou 1).

Examen des **valeurs propres** (extrait de l'ACP sur temperatures.txt)

```
                            PC1    PC2     PC3 …
Standard deviation        3.0954 1.5088 0.26460 …
Proportion of Variance    0.7985 0.1897 0.00583 …
Cumulative Proportion     0.7985 0.9882 0.99402 …
```

La troisième ligne montre que le plan 1-2 des composantes principales rassemble 98.8 % (=79.9 + 18.9) de la variance totale de la matrice initiale. L'analyse de ce plan est très suffisante.

Examen des **vecteurs propres** (x = acp$rotation de la fonction prcomp). Extrait de l'acp sur « ble tendre.txt »

```
      PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8    PC9    PC10
HEP  0.386  0.0317  0.325 -0.286  0.062 -0.306  0.710 -0.233 -0.064 -0.000
DEP -0.387 -0.2079 -0.331 -0.088 -0.087  0.124  0.131 -0.804 -0.022 -0.013
QTE -0.078 -0.4821 -0.023  0.590  0.575 -0.170  0.199  0.050 -0.061  0.074
H20 -0.358 -0.2137 -0.365 -0.277 -0.192 -0.124  0.338  0.479 -0.459 -0.080
PMS  0.412  0.0084 -0.021  0.454 -0.371  0.293  0.028 -0.133 -0.611 -0.072
EM2 -0.025  0.5548 -0.358  0.110  0.227  0.276  0.320  0.064  0.014  0.561
CTE -0.257  0.4403 -0.006  0.340 -0.204 -0.736 -0.092 -0.127 -0.119 -0.001
G.E -0.298 -0.2608  0.491  0.108 -0.435  0.078  0.091  0.059  0.080  0.612
GM2 -0.374  0.2536  0.257  0.297 -0.087  0.335  0.401  0.105  0.237 -0.540
PMG  0.316 -0.2084 -0.462  0.220 -0.437 -0.145  0.201  0.110  0.573 -0.011
```

Ce tableau contient les valeurs propres c.à.d. les coefficients à affecter aux variables initiales permettant le calcul des composantes principales :

0.386*HEP -0.387*DEP + ….+ 0.316*PMG

Les variables ayant les coefficients les plus forts (en valeur absolues) sont celles qui contribuent le plus à la composante… Ici PMS est celle ayant le plus contribué à la PC1 et QTE à PC2. En examinant les variables ayant le plus contribué à une composante, on en déduit le **sens biologique de la composante**

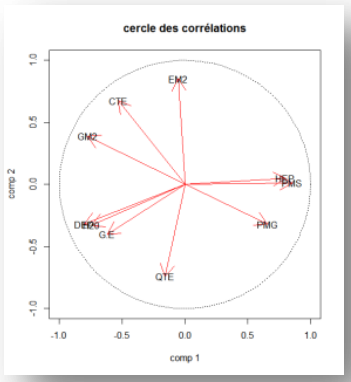
La représentation du cercle des corrélations s'obtient ainsi : on récupère les deux premières colonnes (plan I-2) du tableau des vecteurs propres que l'on multiplie par la racine carré des values propres correspondantes (= écart-types).

Avec les résultats de la fonction prcmp( ) :

```
# On trace un cercle
a=seq(0,2*pi,length=100)
plot(cos(a),sin(a),type='l',lty=3,xlab='comp 1', ylab='comp 2', main="cercle des corrélations")

#récupération des vecteurs propres
r = acp$r ; r          # matrice des vecteurs propres
v = t(acp$r)           # on transpose
v = v[1:2,] ; v        # on récupère les deux premières lignes

# acp$sdev contient les racines carrées des valeurs propres = écart-types
arrows(0,0,acp$sdev[1]*v[1,],acp$sdev[2]*v[2,],col='red')
text(acp$sdev[1]*v[1,],acp$sdev[2]*v[2,],labels=colnames(v))
```

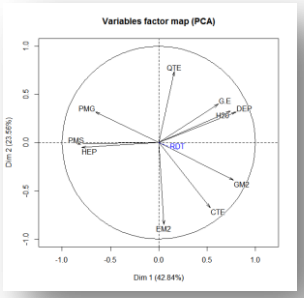


L'interprétation se fait à partir des directions des variables. Ainsi, PMS n'est pas corrélée à EM2 (produit scalaire nul), très bien corrélée (positivement) à HEP mais négativement à DEP et H2O…

Il s'agit aussi d'estimer la bonne qualité de représentation d'autant meilleure qu'elle s'approche du cercle. Sa quantification de cette représentation est donnée par le carré de sa corrélation à l'axe considéré (on somme sur les 2 axes d'un plan).

La fonction PCA(ex. ble tendre.txt) du package FactoMineR trace automatiquement ce graphique. Cette fonction retourne les éléments suivants :

$eig	valeurs propres
$var	résultats pour les variables
varcoord	coordonnées des variables
varcor	corrélations variables - dimensions
varcos2	cos2 des variables
varcontrib	contributions des variables
$ind	résultats pour les individus
indcoord	coordonnées des individus
indcos2	cos2 des individus
indcontrib	contributions des individus
$call	résumé des statistiques
$call$centre	moyennes des variables
$call$ecart.type	écart-type des variables
$call$row.w	poids des individus
$call$col.w	poids des variables



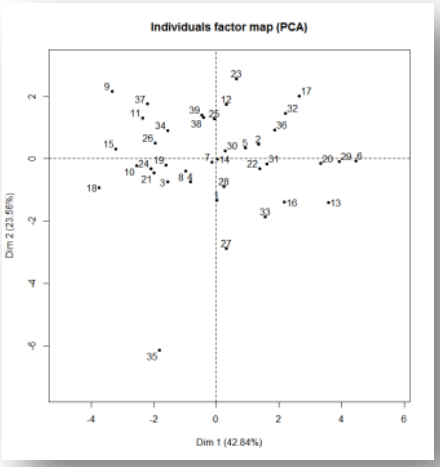
Une ACP avec la fonction PCA.
RDT est en variable supplémentaire.

<u>La structure des variables</u> résulte de ces examens. On en déduit la signification écologique/biologique des axes (composantes).

<u>Examen des individus.</u>

Les coordonnées des individus sont calculées en multipliant la matrice des données (centrées réduites au besoin) par la matrice des vecteurs propres (**$\mathbf{F} = [y - \overline{y}]\mathbf{U}$).**
La démarche très similaire : examen des corrélations et de leurs contributions aux composantes.

L'examen d'un biplot permet de rapprocher les variables des individus



Projection des 39 individus dans le plan 1-2.

Fonction PCA du package FactoMineR

Description

Performs Principal Component Analysis (PCA) with supplementary individuals, supplementary quantitative variables and supplementary categorical variables. Missing values are replaced by the column mean.

Usage

```
PCA(X, scale.unit = TRUE, ncp = 5, ind.sup = NULL,  quanti.sup = NULL, quali.sup = NULL, row.w = NULL, col.w = NULL,
      graph = TRUE, axes = c(1,2))
```

Arguments

X	a data frame with *n* rows (individuals) and *p* columns (numeric variables)
ncp	number of dimensions kept in the results (by default 5)
scale.unit	a boolean, if TRUE (value set by default) then data are scaled to unit variance
ind.sup	a vector indicating the indexes of the supplementary individuals
quanti.sup	a vector indicating the indexes of the quantitative supplementary variables
quali.sup	a vector indicating the indexes of the categorical supplementary variables
row.w	an optional row weights (by default, a vector of 1 for uniform row weights)
col.w	an optional column weights (by default, uniform column weights)
graph	boolean, if TRUE a graph is displayed
axes	a length 2 vector specifying the components to plot

Value

Returns a list including:

eig	a matrix containing all the eigenvalues, the percentage of variance and the cumulative percentage of variance
var	a list of matrices containing all the results for the active variables (coordinates, correlation between variables and axes, square cosine, contributions)
ind	a list of matrices containing all the results for the active individuals (coordinates, square cosine, contributions)
ind.sup	a list of matrices containing all the results for the supplementary individuals (coordinates, square cosine)
quanti.sup	a list of matrices containing all the results for the supplementary quantitative variables (coordinates, correlation between variables and axes)
quali.sup	a list of matrices containing all the results for the supplementary categorical variables (coordinates of each categories of each variables, v.test which is a criterion with a Normal distribution, and eta2 which is the square correlation coefficient between a qualitative variable and a dimension)

Exemple :

```
acp <- PCA(temperatures, quanti.sup= c(13,14,15,16))
```