Analyse de données exploratoire

Christophe Ambroise et Cyril Dalmasso

École doctorale *« du génome aux organismes »*
Université d'Évry

27-29 janvier 2014

`http://stat.genopole.cnrs.fr/~jchiquet/fr/initiation_R`

Analyse en composantes principales

Objectif

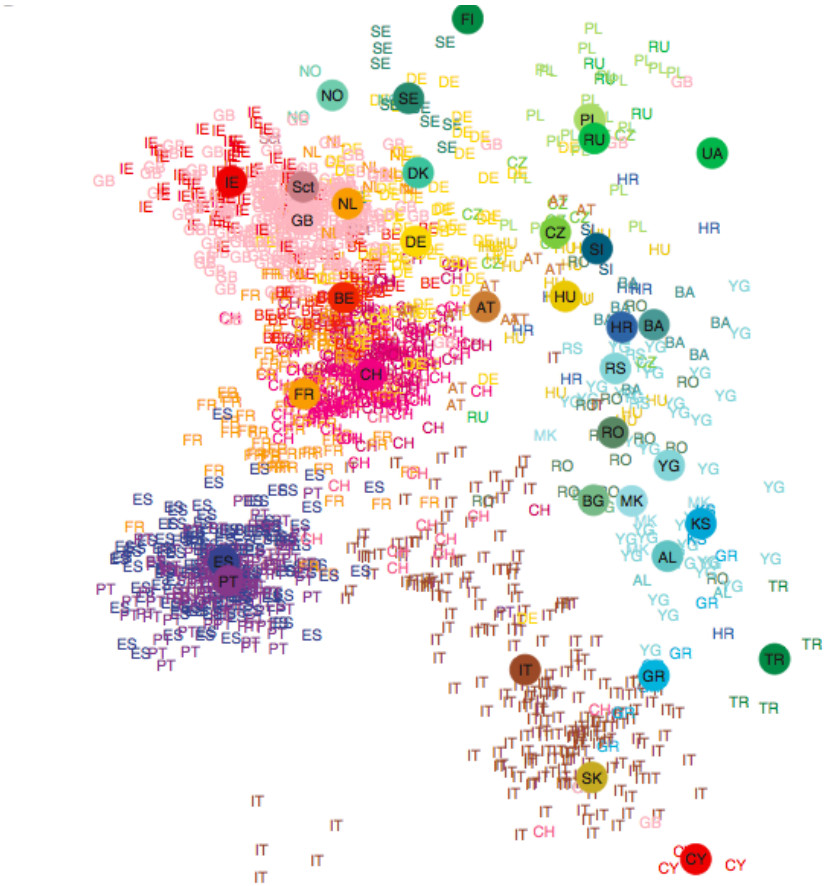
L'idée centrale de l'analyse en composante principales est de réduire la dimensionalité d'un jeu de données consitué d'un grand nombre de variables liées, tout en gardant autant d'information que possible.

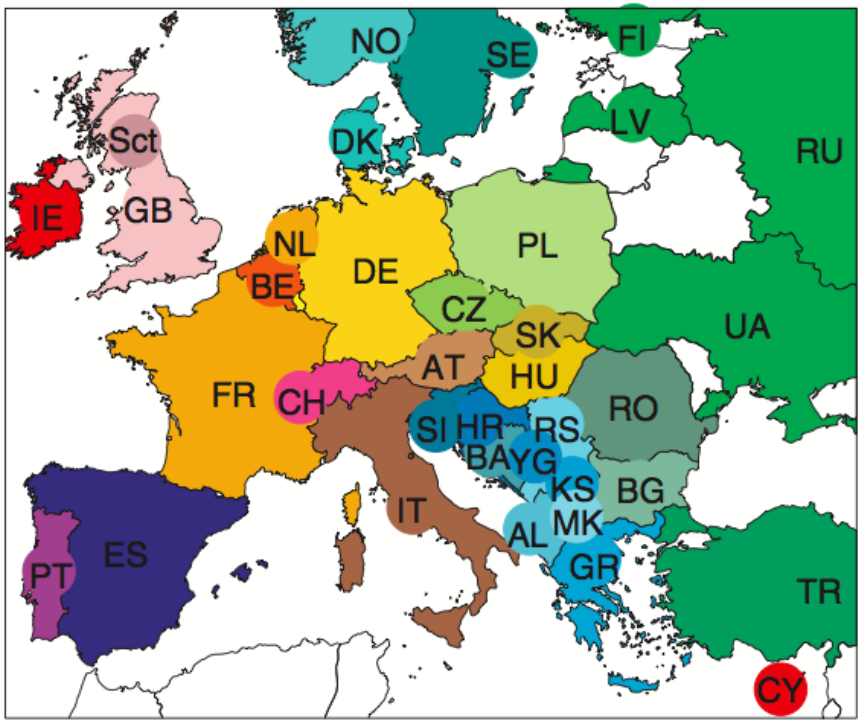
Intuition

Définir de nouvelles variables

- les plus informatives (variance élevées) possibles
- non correlées
- combinaison linéaires des variables des variables originales

Exemple




NO
SE
FI
LV
RU
Sct
DK
IE
GB
NL
DE
PL
BE
CZ
SK
UA
AT
HU
FR
CH
RO
SI
HR
RS
BAYG
KS
BG
IT
MK
AL
GR
PT
ES
TR
CY

L'ACP en un transparent

Notation

- $\mathbf{x}$, un vecteur à p variables décrivant un individu
- $\mathbf{u}_k$, un vecteur de p coefficient
- $\mathbf{u}_k^t \mathbf{x} = \sum_j u_{kj} x_j = c_k$

Procédure

- Trouver une fonction linéaire de $\mathbf{x}$, $\mathbf{u}_1^t \mathbf{x}$ de variance maximum
- Trouver une autre combinaison linéaire de $\mathbf{x}$, $\mathbf{u}_2^t \mathbf{x}$ non corrélée avec $\mathbf{u}_1^t \mathbf{x}$ de variance maximum
- Itérer ....

Solution du problème de l'ACP

Notations et hypothèses

- Σ est la matrice de variance covariance du vecteur aléatoire $\mathbf{x}$
- Lorsque Σ est inconnue, elle est remplacée par son estimateur $\mathbf{S}$

Solution en bref

- $\forall k \ \mathbf{u_k}^t \ \mathbf{u_k} = 1$ (Tous les axes principaux sont de norme unité)
- Les $\mathbf{u_k}$ sont les vecteur propres de Σ associé aux valeurs propres λ_k
- L'ordre des $\mathbf{u_k}$ correspond à l'ordre inverse des valeurs propres λ_k
- $var(c_k) = \lambda_k$ avec $\mathbf{u}_k^t \mathbf{x} = \sum_j u_{kj} x_j = c_k$

En pratique

Représentations

- individus
- variables
- individus supplémentaires
- variables supplémentaires

Indicateurs de qualité

- Représentation globale
- Contribution relative d'un axe à un individu
- Contribution relative d'un individu à un axe

En R

Fonctions

- `prcomp()`
- `princomp()`

Modules

- `FactoMiner`
- `ade4`