Analyse Factorielle

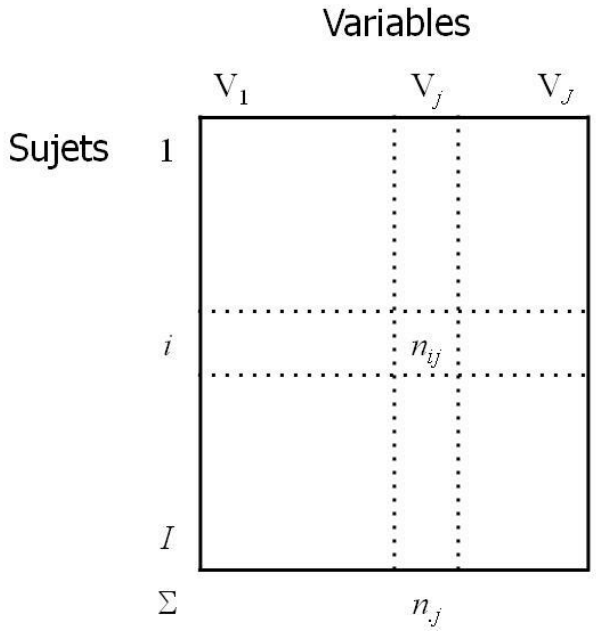
Pr Roch Giorgi

roch.giorgi@univ-amu.fr

SESSTIM, Faculté de Médecine, Aix-Marseille Université, Marseille, France
http://sesstim.univ-amu.fr/

Introduction (1)

- Étudier simultanément un nombre important de variables quantitatives
  - ✓ 2 variables quantitatives $\Rightarrow$ nuage de points (espace de dimension 2)
  - ✓ J variables $\Rightarrow$ espace de dimension J !



Introduction (2)

- Méthode pour obtenir un résumé « pertinent » des données initiales
  - ✓ Cherche à expliquer la variance commune entre les variables
  - ✓ Par de nouvelles variables – facteurs « latents »
  - ✓ Largement utilisée en sciences sociales notamment pour étudier des concepts qui ne peuvent pas être mesurés directement

Introduction (3)

- Méthode factorielle de réduction de dimension pour l'exploration statistique de données quantitatives complexes
  - ✓ Possible également avec des variables qualitatives

- Pour fournir une interprétation d'un nouvel espace créé par un nombre réduit de dimensions qui sont à la base des anciennes dimensions

Introduction (4)

- Construction du modèle statistique associé, estimation
- Représentations graphiques des individus, des variables et simultanée
- Qualité de représentation

Objectifs

- Réduction du nombre de variables
- Identification de variables latentes

$\Rightarrow$Interprétation subjective

Principes (1)

- Variables $V_1,\ldots,V_J$ liées à un petit nombre de variables (non-observées) latentes $f_1,\ldots,f_k$ avec $k<J$

- Modèle de régression

$$
\begin{aligned}
V_1 &= \lambda_{11}.f_1 + \lambda_{12}.f_2 + \ldots + \lambda_{1k}.f_k + u_1 \\
V_2 &= \lambda_{21}.f_1 + \lambda_{22}.f_2 + \ldots + \lambda_{2k}.f_k + u_2 \\
&\vdots \\
V_j &= \lambda_{j1}.f_1 + \lambda_{j2}.f_2 + \ldots + \lambda_{jk}.f_k + u_j \\
&\vdots \\
V_J &= \lambda_{J1}.f_1 + \lambda_{J2}.f_2 + \ldots + \lambda_{Jk}.f_k + u_J
\end{aligned}
$$

Principes (2)

$$V_j = \lambda_{j1}.f_1 + ... + \lambda_{jl}.f_l ... + \lambda_{jk}.f_k + u_j$$

- λ_{jl} : coefficient de régression reflétant le lien de chaque V_j sur le facteur commun
  - ✓ Valeurs élevées relient le facteur aux variables initiales à partir desquelles ont en déduit une description des facteurs
- u_j : résidu spécifique à V_j avec u_1,..., u_J non corrélés entre eux et avec f_1,...,f_k

Principes (3)

- Facteurs ne sont pas corrélés entre eux (conditionnellement aux facteurs, les variables sont indépendantes , i.e. les corrélations entre les variables observées proviennent de leurs relations avec les facteurs)

- Les coefficients de régression $\lambda_{..}$ ($\boldsymbol{\lambda}$) : facteurs extraits

Autres Hypothèses

- Facteurs sont de moyenne nulle et de variance égale à 1

- Facteurs ne sont pas corrélés entre eux

$\Rightarrow$Facteurs extraits expriment la corrélation entre les variables et les facteurs

Variance de V_j

$$\theta_j^2 = \sum_{l=1}^{k} \lambda_{jl}^2 + \psi_j$$

Variance commune partagée avec les autres variables au travers de f_l

Variance de u_j , variance spécifique à V_j non partagée avec les autres variables

Covariance entre V_j et V_m

$$\theta_{jm} = \sum_{l=1}^{k} \underbrace{\lambda_{jl}\lambda_{ml}}$$

Indépendant de V_j et V_m
La relation s'exprime au travers des facteurs communs

Remarque

- Estimation des paramètres
  - ✓ Analyse factorielle principale
  - ✓ Maximum de vraisemblance

Détermination du Nombre de Facteurs

- Règle de Guttman-Kaiser
  - ✓ Valeurs propres > 1
- Retenir les facteurs correspondant à 70-80% de la variance
- Scree plot
- Approche par Maximum de vraisemblance permet d'effectuer un test statistique (H_0 : k facteurs sont suffisant *vs* H_1 : nombre de facteurs ≥ k)

Rotation des Facteurs (1)

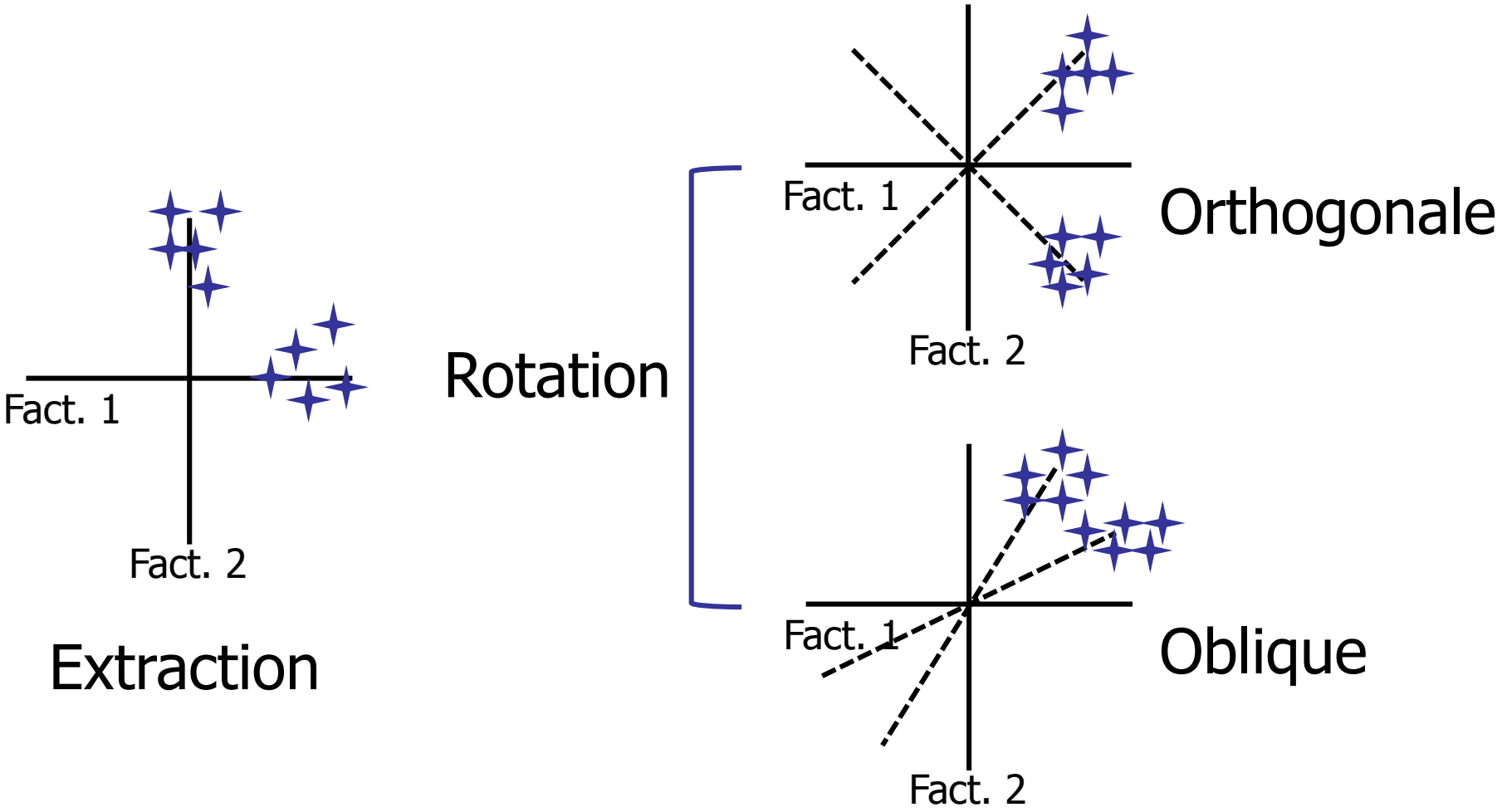
- Extraction des facteurs
  - ✓ Interpréter et nommer les facteurs

- Rotation des facteurs
  - ✓ Altère le pattern des facteurs extraits et peut améliorer l'interprétation

Facteurs Extraits et Scores

- Interprétation des facteurs extraits
  - ✓ Ceux expliquant au moins 16% de la variance (surtout vrai pour ACP)
- Les valeurs des scores peuvent être utilisées par la suite pour
  - ✓ Identifier des clusters de sujets
  - ✓ Effectuer une régression pour résoudre un problème de multicolinéarité
  - ✓ Effectuer une analyse sur les scores plutôt que sur des variables qui constitueraient en fait des mesures similaires d'une même variable sous-jacente

Rotation des Facteurs (2)



Source : *Field A. Discovering Statistics using SPSS for Windows. London-Thousand Oaks-New Delhi: Sage publications. 2000.*

AF et ACP : Similarités

- Réduction du nombre de variables
- Utilisés sur matrice de covariances ou de corrélations
- Reposent sur un modèle linéaire
- Résultats similaires que AF sans rotation

AF et ACP : Différences

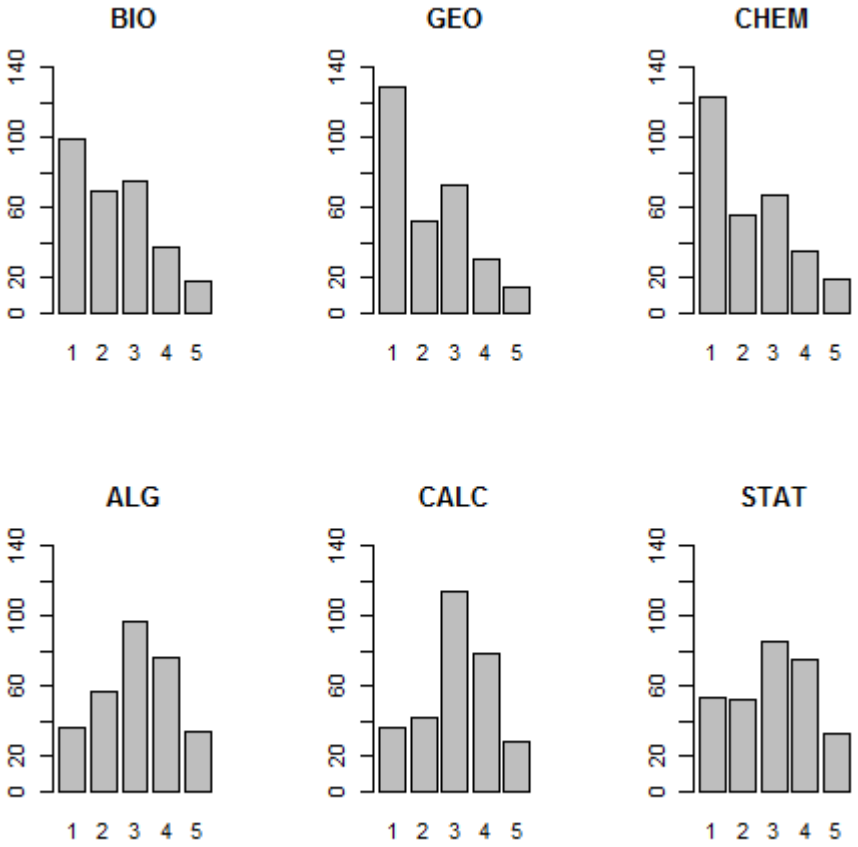
- AF repose d'emblée sur la construction d'un modèle linéaire (approche modélisation) – ACP recherche le meilleur ajustement dans un espace réduit et se fini par un modèle linéaire (approche descriptive)
- AF cherche à reproduire au mieux la matrice de covariance ou de corrélation (décomposition de la variance, commune unique et erreur) – ACP optimise la variance totale (variance totale des variables prises en compte par la moyenne des composantes ; il n'y a pas de terme d'erreur)

Exemple

- 300 étudiants, 6 variables
  - ✓ Biologie, géologie, chimie, algèbre, calcul, statistique
  - ✓ 1 → 5 ; apprécie pas du tout → apprécie fortement

Source : *http://rtutorialseries.blogspot.fr/2011/10/r-tutorial-series-exploratory-factor.html*

Descriptif



Matrice de Corrélations

	BIO	GEO	CHEM	ALG	CALC	STAT
BIO	1.0000000	0.6822208	0.7470278	0.1153204	0.2134271	0.2028315
GEO	0.6822208	1.0000000	0.6814857	0.1353557	0.2045215	0.2316288
CHEM	0.7470278	0.6814857	1.0000000	0.0838225	0.1364251	0.1659747
ALG	0.1153204	0.1353557	0.0838225	1.0000000	0.7709303	0.4094324
CALC	0.2134271	0.2045215	0.1364251	0.7709303	1.0000000	0.5073147
STAT	0.2028315	0.2316288	0.1659747	0.4094324	0.5073147	1.0000000

Résultats de l'AF (1)

- Nombre de facteurs

```
Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 345.33 on 9 degrees of freedom.
The p-value is 6.11e-69

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 2.94 on 4 degrees of freedom.
The p-value is 0.568
```

Résultats de l'AF (2)

- Sans rotation

```
Loadings:
     Factor1 Factor2
BIO   0.303   0.810
GEO   0.290   0.736
CHEM  0.230   0.835
ALG   0.781  -0.130
CALC  0.970  -0.103
STAT  0.530
```

« Science »

« Math »

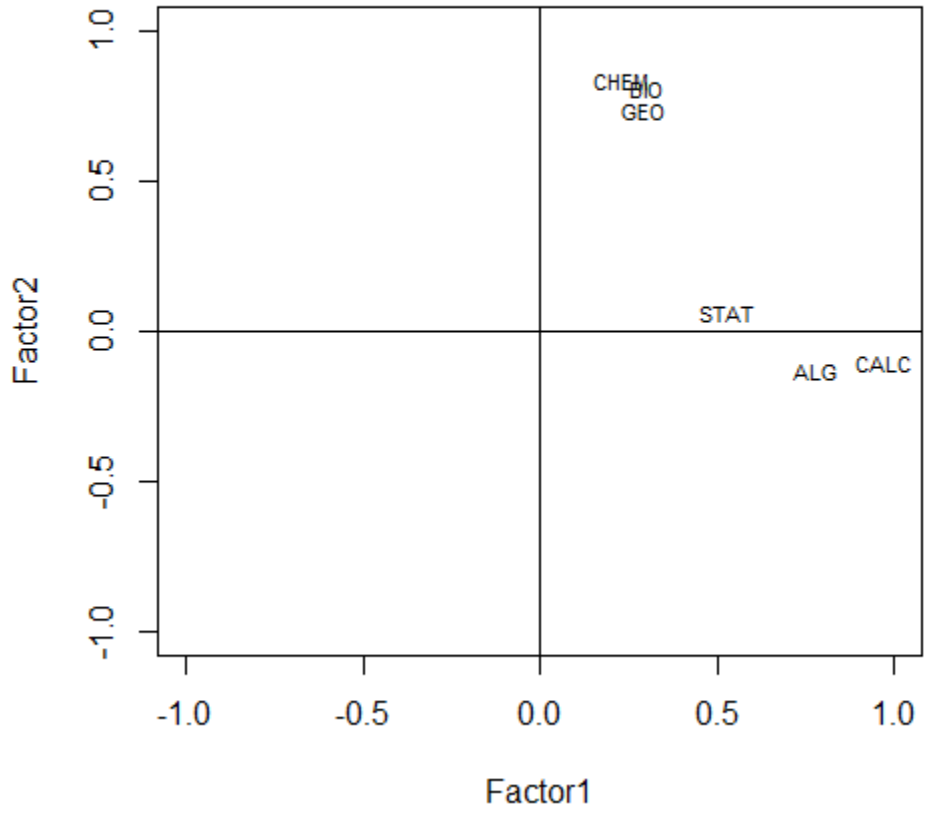
Résultats de l'AF (3)

- Sans rotation

	Factor1	Factor2
SS loadings	2.061	1.926
Proportion Var	0.344	0.321
Cumulative Var	0.344	0.665

Résultats de l'AF (4)

- Représentation graphique (sans rotation)



Résultats de l'AF (5)

- Rotation Varimax (orthogonale ; maximise la variance du nouveau facteur)

```
Loadings:
     Factor1 Factor2
BIO   0.855   0.133
GEO   0.779   0.135
CHEM  0.865
ALG           0.791
CALC          0.971
STAT  0.170   0.506
```

« Math »

« Science »

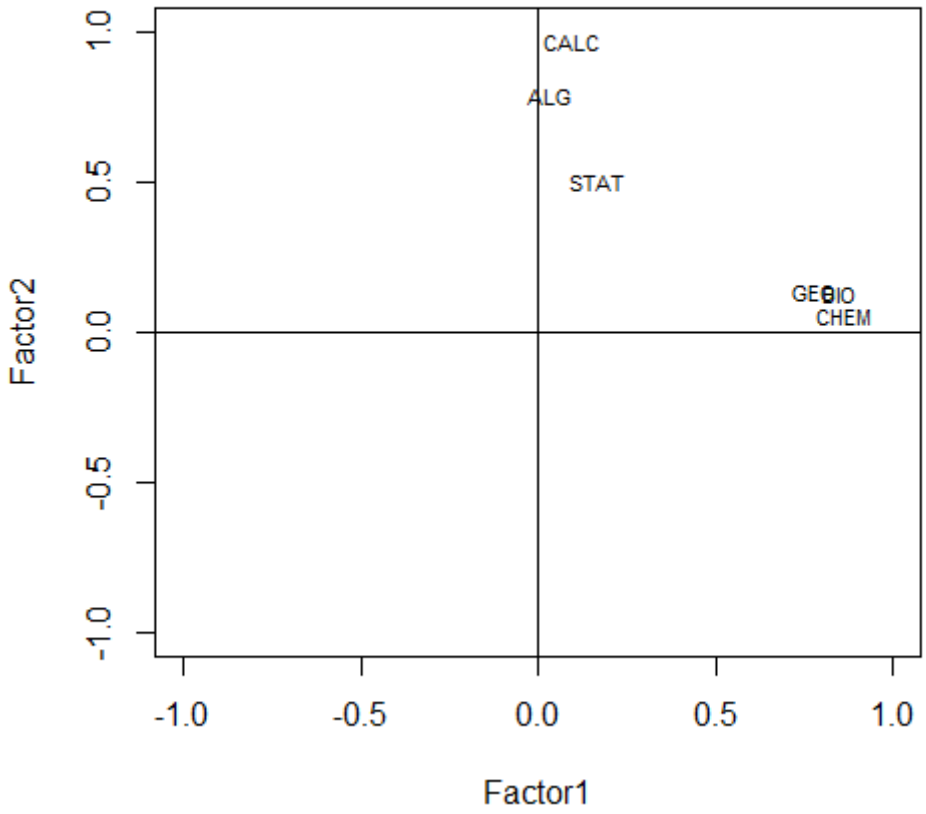
Résultats de l'AF (6)

- Rotation Varimax

	Factor1	Factor2
SS loadings	2.124	1.863
Proportion Var	0.354	0.311
Cumulative Var	0.354	0.665

Résultats de l'AF (7)

- Représentation graphique (après varimax)



Sources

- Brian S. Everitt. An R and S-PLUS Companion to Multivariate Analysis. Springer 2005.
- http://rtutorialseries.blogspot.fr/2011/10/r-tutorial-series-exploratory-factor.html