Analyse factorielle exploratoire

Mireille AGBLA, Achille GAUSSERES, Nel HERVÉ

Sommaire

1. Introduction : analyse factorielle, analyse factorielle exploratoire et ACP

2. Modèle mathématique et estimation des paramètres

3. Exemple d'application sur R

Analyse factorielle

Analyse factorielle :

Une famille de techniques statistiques visant à **réduire** un ensemble important de variables mesurées en un plus petit nombre de variables.

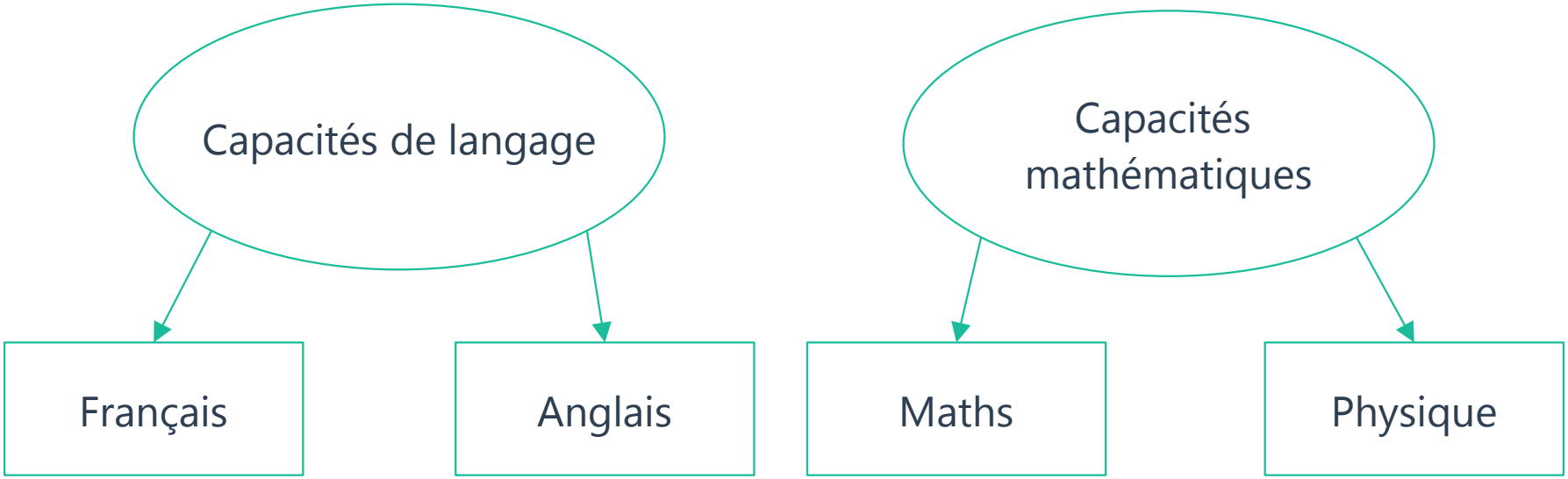
⇒ Objectif principal : **réduction** ou **synthèse** de données

Analyse factorielle

- **Analyse factorielle exploratoire (EFA)**
  - Découvrir la structure sous-jacente (variables latentes) d'un ensemble de variables manifestes
  - Ou simplement synthétiser le jeu de données

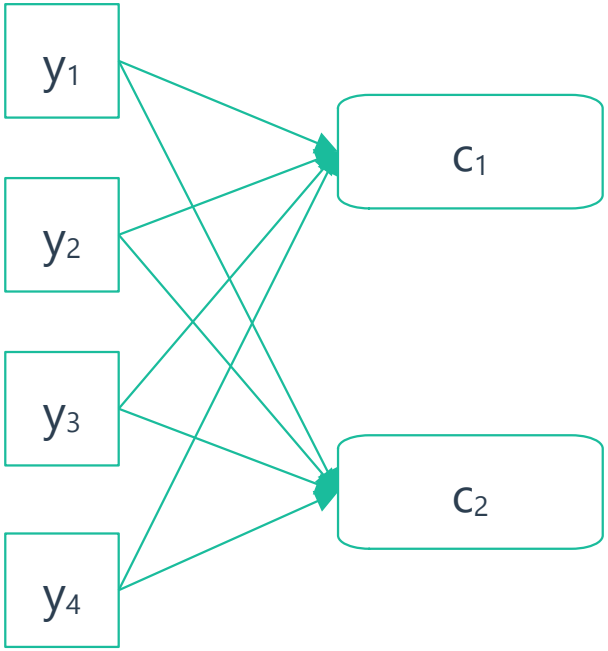
- **Analyse factorielle confirmatoire (CFA)**
  - Déterminer si le nombre de facteurs et leurs saturations correspondent à ce qui est attendu à partir de la théorie
  - Pas abordée dans ce cours

Analyse factorielle exploratoire



→ Quels facteurs peuvent expliquer nos observations ?

Comparaison : principe de l'ACP



→ c_1 et c_2 sont les **composantes principales**

→ c_1 et c_2 sont obtenues par **combinaison linéaire** des y_i

→ maximisent la **variance** des données projetées

→ forme :

$$c_1 = \lambda_{11} y_1 + \lambda_{12} y_2 + \lambda_{13} y_3 + \lambda_{14} y_4$$
$$c_2 = \lambda_{21} y_1 + \lambda_{22} y_2 + \lambda_{23} y_3 + \lambda_{24} y_4$$

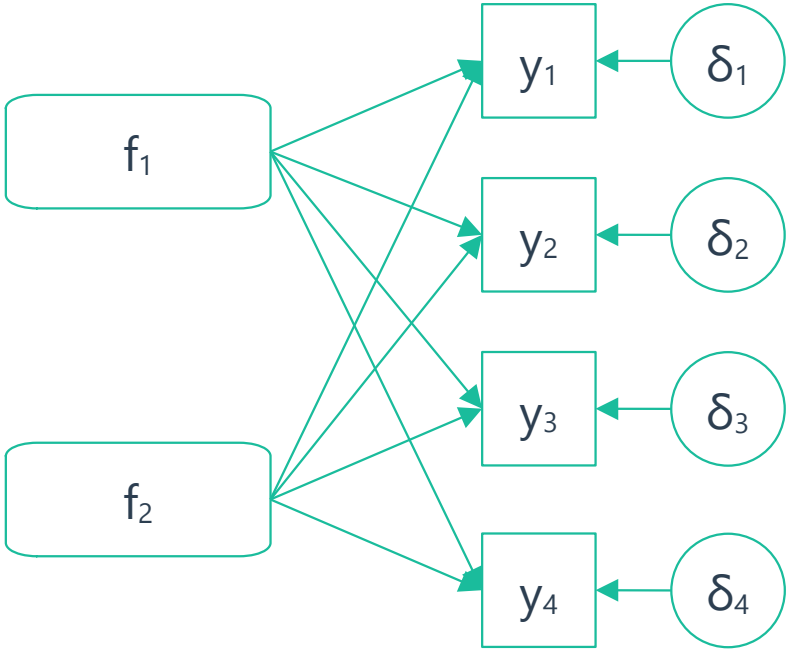
Comparaison : principe de l'EFA

→ f_1 et f_2 sont les **facteurs latents** ; les δ_i sont les **variances spécifiques** des y_i

→ les y_i sont obtenus par **combinaison linéaire** de f_1 et f_2

→ forme :

$$y_1 = \lambda_{11} f_1 + \lambda_{12} f_2 + \delta_1$$
$$y_2 = \lambda_{21} f_1 + \lambda_{21} f_2 + \delta_2$$
$$y_3 = \lambda_{31} f_1 + \lambda_{32} f_2 + \delta_3$$
$$y_4 = \lambda_{41} f_1 + \lambda_{42} f_2 + \delta_4$$



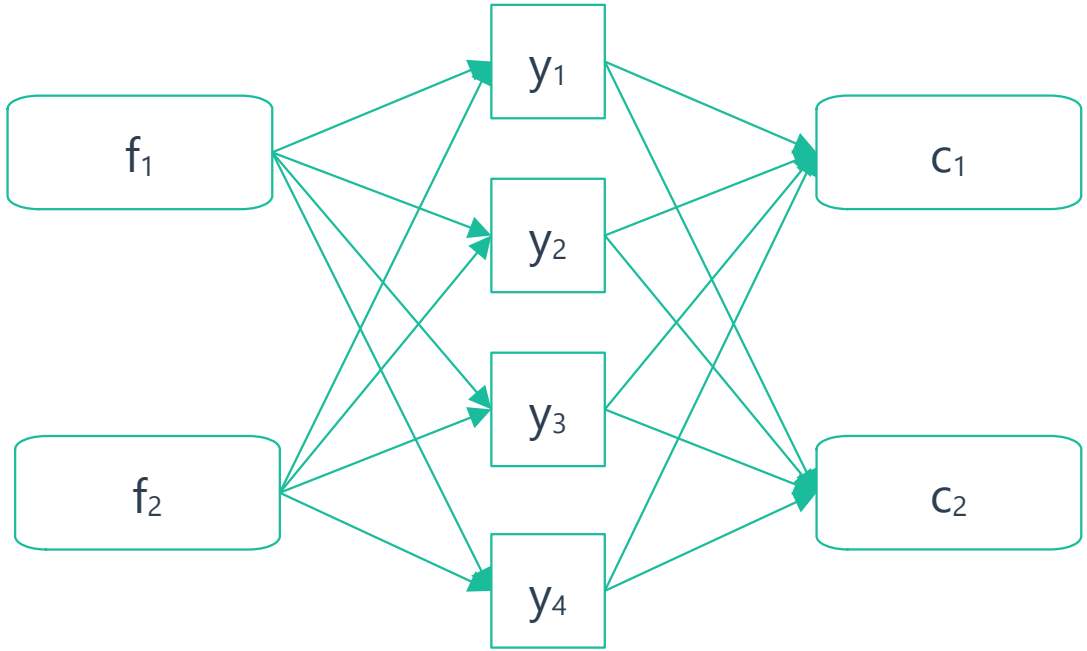
Comparaison : EFA vs ACP

EFA :

→ modèle à variables latentes

→ explique les corrélations

→ infinité de solutions pour un jeu de données



ACP :

→ procédure de calcul

→ explique la variance

→ une seule solution par jeu de données

Modèle de base de l'EFA

→ **Forme matricielle :**

$$
\left.
\begin{aligned}
Y_1 &= \lambda_{11} F_1 + \lambda_{12} F_2 + \ldots + \lambda_{1m} F_m + \delta_1 \\
Y_2 &= \lambda_{21} F_1 + \lambda_{22} F_2 + \ldots + \lambda_{2m} F_m + \delta_2 \\
&\ldots \\
Y_n &= \lambda_{n1} F_1 + \lambda_{n2} F_2 + \ldots + \lambda_{nm} F_m + \delta_n
\end{aligned}
\right\} \quad Y = \Lambda F + \delta
$$

Modèle de base de l'EFA

$$Y = \Lambda F + \delta$$

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \cdots & \cdots & \lambda_{1m} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \lambda_{n1} & \cdots & \cdots & \lambda_{nm} \end{bmatrix}_{n \times m} \times \begin{bmatrix} F_1 \\ \vdots \\ F_m \end{bmatrix}_{m \times 1} + \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_n \end{bmatrix}_{n \times 1}$$

Modèle de base de l'EFA

$$Y = \Lambda F + \delta$$

sous les hypothèses suivantes :

- Y est **centrée** : $E(Y) = 0$
- Les facteurs F_i sont **centrés-réduits** et **orthogonaux** ; pour tout $i \neq j$:
  - $E(F_i) = 0$ et $V(F_i) = I_m$
  - $cov(F_i, F_j) = 0$
- Les erreurs sont **centrées** et **décorrélées** : $E(\delta) = 0$ et $V(\delta) = diag(\psi)$
- Les facteurs F et les erreurs δ sont **indépendants**

Modèle de base de l'EFA

→ Proposition :

$$\forall i \in \{1,\dots,n\},\ V(Y_i) = \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_i^2 \qquad \text{avec} \qquad \psi_i^2 = V(\delta_i)$$

et $\forall i, j \in \{1,\dots,n\},\ cov(Y_i, Y_j) = \sum_{l=1}^{k} \lambda_{il}\lambda_{jl}$

→ Proposition :

La matrice Σ des variances-covariances de Y_i est donnée par :

$$\Sigma = \Lambda \Lambda' + \Psi \qquad \text{avec} \quad \Psi = diag(\psi_i)$$

Cette forme nous permettra d'estimer les saturations λ_{ij}.

Estimation des saturations

→ Méthodes possibles pour estimer Λ :

- **Composantes principales**
  - Avec la matrice des variances-covariances observée : $S \approx \hat{\Lambda}\,\hat{\Lambda}'$
  - Une SVD sur S donne $S = CDC' = (CD^{1/2})(CD^{1/2})'$
  - On définit $\hat{\Lambda}$ par les *m* premières colonnes de $CD^{1/2}$ (rangées dans l'ordre décroissant des valeurs propres)

- **Facteurs principaux (itérés)**
  - Estimer $S^* = S - \hat{\Psi}$, puis $\hat{\Lambda}$ avec une SVD
  - Réestimer $\hat{\Psi} = S - \hat{\Lambda}\,\hat{\Lambda}'$ et itérer jusqu'à convergence

- **Maximum de vraisemblance**
  - Hypothèse : les facteurs suivent une loi normale centrée-réduite
  - Estimation de $\hat{\Lambda}$ et $\hat{\Psi}$ par maximum de vraisemblance (méthodes itératives)

Estimation des facteurs

→ Méthode : on reprend le modèle $Y = \Lambda F + \delta$ qui devient $Y = \hat{\Lambda} F + \delta$.
On peut alors estimer F par :

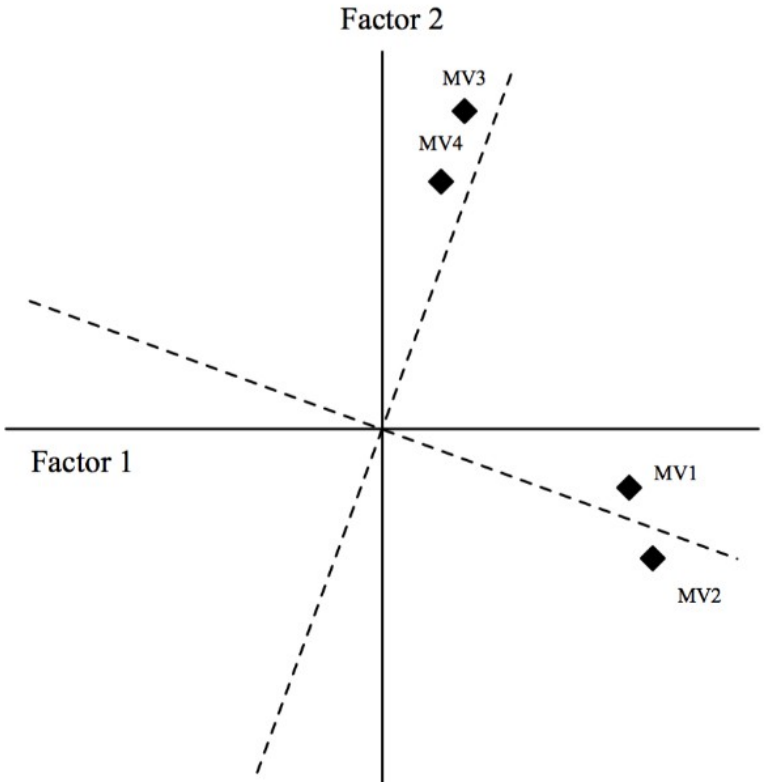
- Les MCO : $\hat{F} = (\hat{\Lambda}' \hat{\Lambda})^{-1} \hat{\Lambda}' Y$
- Maximum de vraisemblance
- …

→ Remarque : en EFA, il n'est généralement pas utile d'estimer les facteurs. Tout repose sur la valeur des saturations.

Rotation des facteurs

→ Remarque : les facteurs et saturations associées ne sont pas uniques.

On peut ainsi **pivoter les facteurs** pour faciliter l'interprétation des résultats.



Rotation des facteurs

→ Méthode la plus courante : *varimax*

→ Objectifs :
- Quelques saturations très élevées
- Autant de saturations quasi-nulles que possible

→ Méthode : optimisation itérative d'une fonction quadratique des saturations…

→ Autres options :
- Autres rotations orthogonales : *quartimax*, *equamax*…
- Rotations obliques : *oblimin*, *promax*…

Exemple

Cas concret : traits de personnalité

→ Contexte :

1 015 341 personnes ont répondu à un questionnaire qui évalue leurs traits de personnalité :

- Questions : "Je ne parle pas beaucoup", "J'aime les enfants"…
- Réponses : échelle de 1 ("Pas d'accord") à 5 ("D'accord").

Exemple

Cas concret : traits de personnalité

→ Contexte :

1 015 341 personnes ont répondu à un questionnaire qui évalue leurs traits de personnalité :

- Questions : "Je ne parle pas beaucoup", "J'aime les enfants"…
- Réponses : échelle de 1 ("Pas d'accord") à 5 ("D'accord").

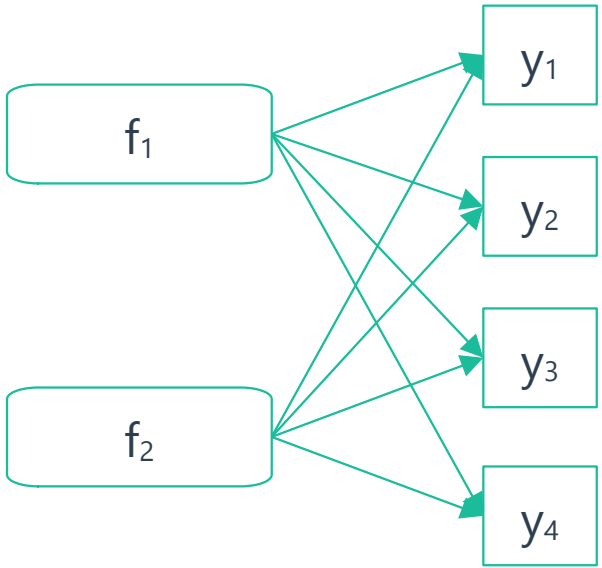
⇒ Existe-t-il des **types de personnalités sous-jacents** qui permettent d'expliquer les réponses des participants quant à leurs traits de personnalité?

Exemple

Cas concret : traits de personnalité

Existe-t-il des **types de personnalités sous-jacents** qui permettent d'expliquer les réponses des participants quant à leurs traits de personnalité?

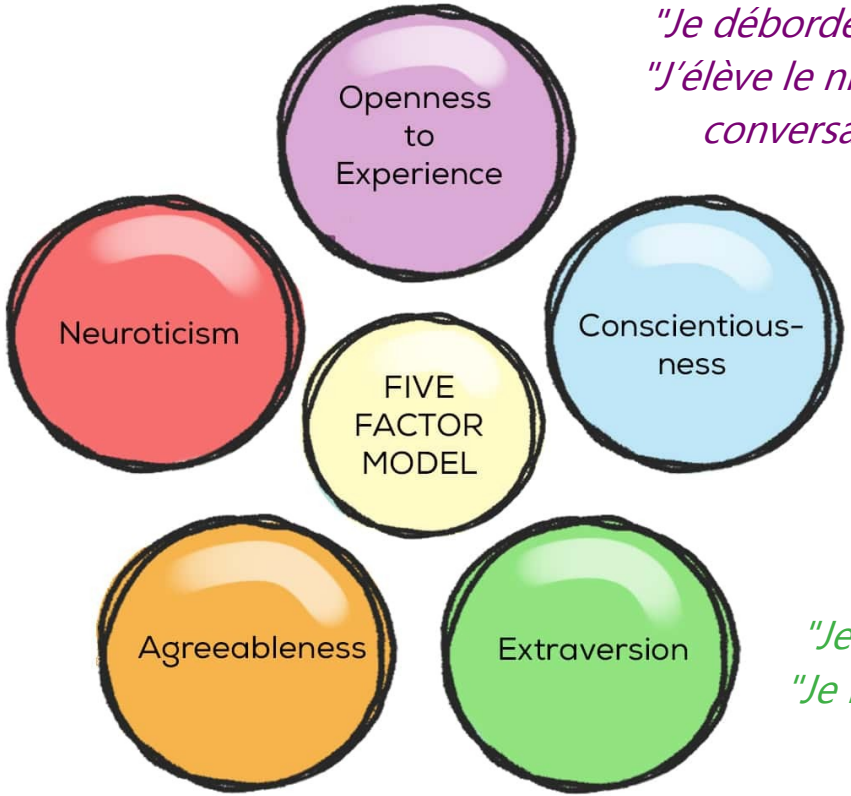
Types de personnalité (?)



Traits de personnalité (50)

Exemple

Cas concret : le modèle à 5 facteurs



"Je déborde d'idées"
"J'élève le niveau des conversations"

"Je suis irritable"
"Je me sens souvent déprimé"

"Je ne fais les choses qu'à moitié"
"Je suis perfectionniste"

"Je m'enquiers du bien-être des autres"
"J'aime les enfants"

"Je sais captiver les gens"
"Je ne parle pas beaucoup"