Recueil des données quantitatives

L’analyse factorielle

vincent.berthet@univ-lorraine.fr

1. Définition

- L'analyse factorielle (AF) est une **méthode statistique** inventée par Spearman permettant d'analyser un ensemble de **corrélations.**

  Le facteur d'intelligence général :
  - ✓ est **commun** aux différentes épreuves d'intelligence
  - ✓ permet de **résumer** les corrélations observées
  - ✓ est **hypothétique**

- L'AF est une **méthode de réduction des données**
  - ✓ nombre de facteurs < nombre de variables observées
  - ✓ facteurs = dimensions fondamentales

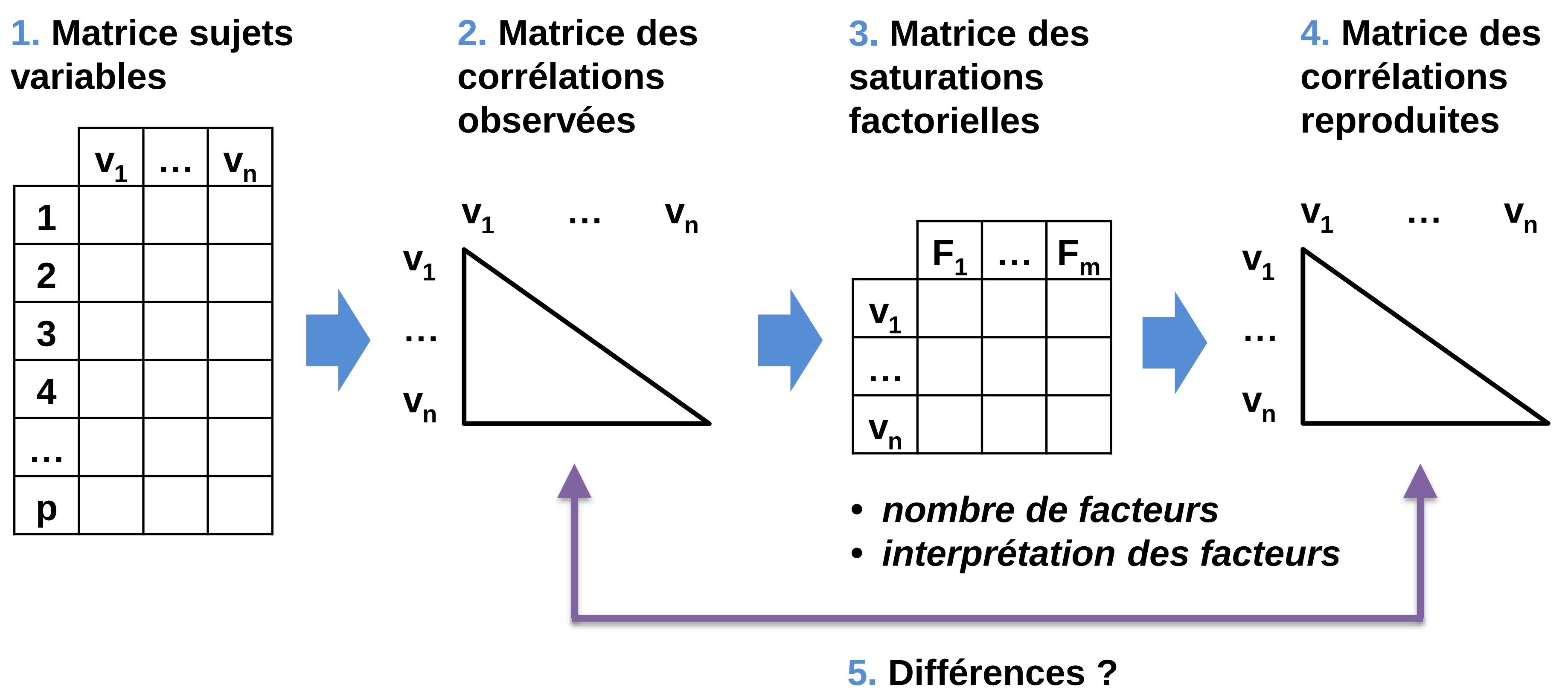
  En psychologie :
  - ✓ variables observées = comportements
  - ✓ facteurs = dimensions fondamentales des comportements

2. Fonctionnement de l’analyse factorielle

- On distingue **cinq étapes :**

1. on part d’une matrice sujets-variables
2. on calcule la matrice des **corrélations observées**
3. on estime la matrice des saturations factorielles
4. on calcule la matrice des **corrélations reproduites**
5. on compare les corrélations reproduites aux corrélations observées

2. Fonctionnement de l’analyse factorielle



3. L'hypothèse fondamentale

Les corrélations observées sont dues à la présence de facteurs latents communs aux variables

- Un facteur commun est :
- ✓ **une variable latente** : un facteur est une entité hypothétique
- ✓ **une variable causale** : les covariations des variables sont dues aux variations d'un facteur commun

- L'hypothèse fondamentale de l'analyse factorielle se décline sous deux aspects :
- ✓ le principe d'**indépendance locale**
- ✓ la notion de **corrélation reproduite**

3. L’hypothèse fondamentale

Le principe d’indépendance locale

- Chez des enfants de 5 à 12 ans, on constate une corrélation élevée entre la richesse du vocabulaire et la longueur du pouce…

  En fait, cette corrélation est artificielle car elle due à la présence d’une **covariable** : l’âge

- La corrélation entre deux variables lorsqu’une covariable est maintenue constante (ou contrôlée) est appelée **corrélation partielle.**

  Chez des enfants du **même âge**, il n’y a plus de corrélation entre la richesse du vocabulaire et la longueur du pouce

- **Principe d’indépendance locale** : si l’on maintient constant un facteur, les corrélations entre les variables observées deviennent nulles

3. L'hypothèse fondamentale

La notion de corrélation reproduite

- La **saturation** (λ) d'une variable observée dans un facteur est la **corrélation entre le facteur et la variable**

✓ $-1 \leq \lambda \leq +1$

✓ λ^2 = % de variance de la variable expliquée par le facteur

- Une saturation correspond donc au degré auquel le facteur **influence** la variable

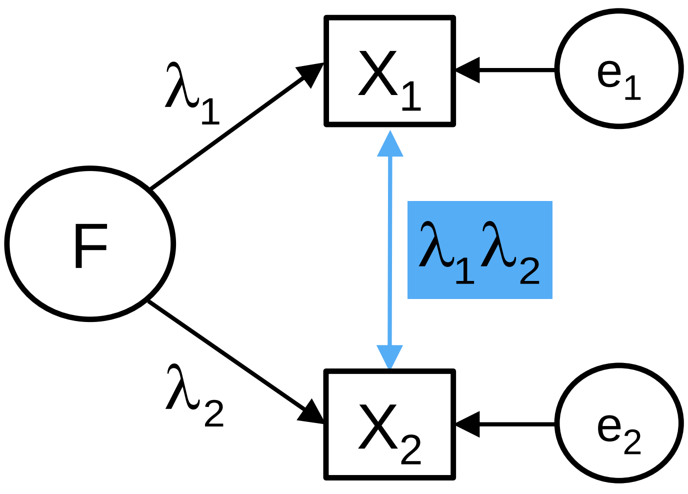
3. L'hypothèse fondamentale

- Si l'on connait les saturations des variables dans les facteurs, **on peut reproduire les corrélations entre les variables.**

  Ce sont les corrélations attendues sur la base des facteurs communs

- Si deux variables observées sont influencées par un facteur commun : corrélation reproduite = **produit de leurs saturations respectives dans ce facteur**

$$\hat{r}(X_1, X_2) = \lambda_1 \times \lambda_2$$



4. La matrice des saturations factorielles

- Cette matrice indique les **saturations des variables dans les différents facteurs.**

  Il s’agit de la matrice centrale dans l’AF car elle montre comment les variables sont reliées aux facteurs

	Facteur 1	**Facteur 2**	**…**	**Facteur *m***
variable 1				
variable 2				
…				
variable *n*				

- Une matrice de saturation est définie par deux caractéristiques :

✓ le **nombre** de facteurs

✓ la **signification** des facteurs

4. La matrice des saturations factorielles

- Pour une même matrice de corrélations, l'AF peut donner lieu à une **infinité** de matrices de saturations possibles : problème de l'**indétermination de la solution**

- Il faut donc des critères pour décider du **nombre de facteurs**
  - ✓ la matrice des corrélations résiduelles
  - ✓ le pourcentage de variance expliquée
  - ✓ les valeurs propres

- et des critères pour décider de la **signification des facteurs**
  - ✓ la cohérence théorique

4. La matrice des saturations factorielles

Le nombre de facteurs

- Le nombre de facteurs retenus définit une **solution factorielle**

solution à 1 facteur :

	Facteur 1
variable 1	
variable 2	
…	
variable *n*	

solution à 2 facteurs :

	Facteur 1	**Facteur 2**
variable 1		
variable 2		
…		
variable *n*		

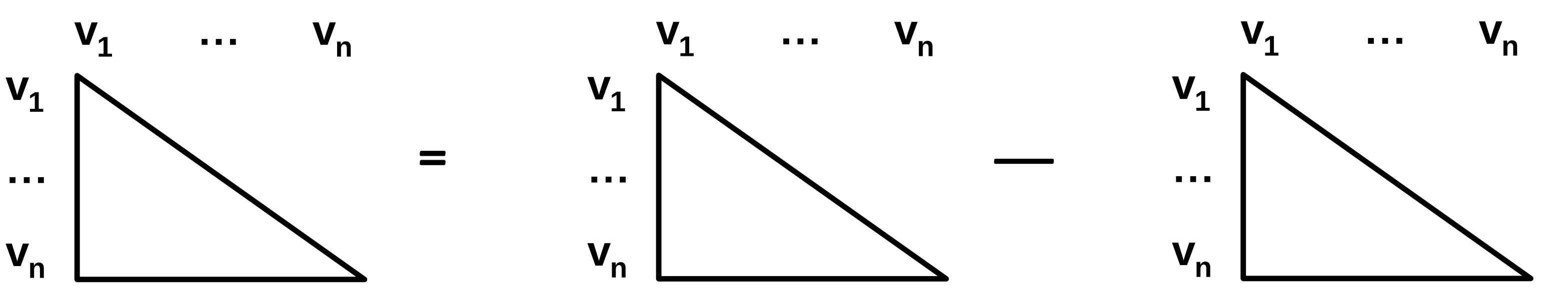
4. La matrice des saturations factorielles

Critère 1 : les corrélations résiduelles

- La matrice des saturations permet de calculer les corrélations attendues entre les variables sur la base des facteurs communs : la **matrice des corrélations reproduites**

- La **différence** entre les corrélations observées et les corrélations reproduites correspond aux **corrélations résiduelles**

Matrice des corrélations résiduelles | **Matrice des corrélations observées** | **Matrice des corrélations reproduites**



4. La matrice des saturations factorielles

Critère 1 : les corrélations résiduelles

- Les corrélations résiduelles correspondent à la part des corrélations observées qui n’est pas expliquée par les facteurs
- Le nombre de facteurs doit être tel que **les corrélations résiduelles soient faibles**
  - ✓ corrélations résiduelles faibles : bonne solution factorielle
  - ✓ corrélations résiduelles importantes : mauvaise solution factorielle

4. La matrice des saturations factorielles

Critère 2 : le pourcentage de variance expliquée

	Facteur 1	Facteur 2
item 1	0.76	-0.42
item 2	0.71	-0.36
item 3	0.72	-0.47
item 4	0.51	0.62
item 5	0.65	0.70
item 6	0.45	0.30

4. La matrice des saturations factorielles

Critère 2 : le pourcentage de variance expliquée

- La **communauté** (h^2) d'une variable exprime le % de sa variance qui est expliqué par la solution factorielle.

  Elle est égale à la somme des carrés des saturations de la variable (si les facteurs sont indépendants) :

$$h^2 = \lambda_1^2 + \lambda_2^2$$

	Facteur 1	**Facteur 2**	**h²**
item 1	0.76	-0.42	0.75
item 2	0.71	-0.36	0.63
item 3	0.72	-0.47	0.75
item 4	0.51	0.62	0.65
item 5	0.65	0.70	0.92
item 6	0.45	0.30	0.29

4. La matrice des saturations factorielles

Critère 2 : le pourcentage de variance expliquée

- La **communauté** (h^2) d'une variable exprime le % de sa variance qui est expliqué par la solution factorielle.

  Elle est égale à la somme des carrés des saturations de la variable (si les facteurs sont indépendants) :

$$h^2 = \lambda_1^2 + \lambda_2^2$$

	Facteur 1	Facteur 2	h²
item 1	0.76	-0.42	0.75
item 2	0.71	-0.36	0.63
item 3	0.72	-0.47	0.75
item 4	0.51	0.62	0.65
item 5	0.65	0.70	0.92
item 6	0.45	0.30	0.29
			66.5%

La moyenne des communautés indique le **% de variance totale expliquée par la solution factorielle**

4. La matrice des saturations factorielles

Critère 2 : le pourcentage de variance expliquée

- Le nombre de facteurs doit être tel que **le pourcentage de variance totale expliquée est élevé.**

- Plus il y a de facteurs, plus le % de variance expliquée est élevé

- Il faut trouver le juste milieu entre :
  - ✓ avoir un % de variance expliquée important
  - ✓ avoir un petit nombre de facteurs (principe de parcimonie)

4. La matrice des saturations factorielles

Critère 3 : les valeurs propres

- La **valeur propre** d'un facteur est égale à la somme des carrés des saturations des variables

	Facteur 1	**Facteur 2**	**h²**
item 1	0.76	-0.42	0.75
item 2	0.71	-0.36	0.63
item 3	0.72	-0.47	0.75
item 4	0.51	0.62	0.65
item 5	0.65	0.70	0.92
item 6	0.45	0.30	0.29
vp	**2.48**	**1.50**	

4. La matrice des saturations factorielles

Critère 3 : les valeurs propres

$$\% \text{ variance expliquée par un facteur} = \left(\frac{\text{valeur propre du facteur}}{\text{nombre variables}} \right) \times 100$$

	Facteur 1	**Facteur 2**	**h²**
item 1	0.76	-0.42	0.75
item 2	0.71	-0.36	0.63
item 3	0.72	-0.47	0.75
item 4	0.51	0.62	0.65
item 5	0.65	0.70	0.92
item 6	0.45	0.30	0.29
vp	**2.48**	**1.50**	
% var	**41.5**	**25**	**66.5**

4. La matrice des saturations factorielles

Critère 3 : les valeurs propres

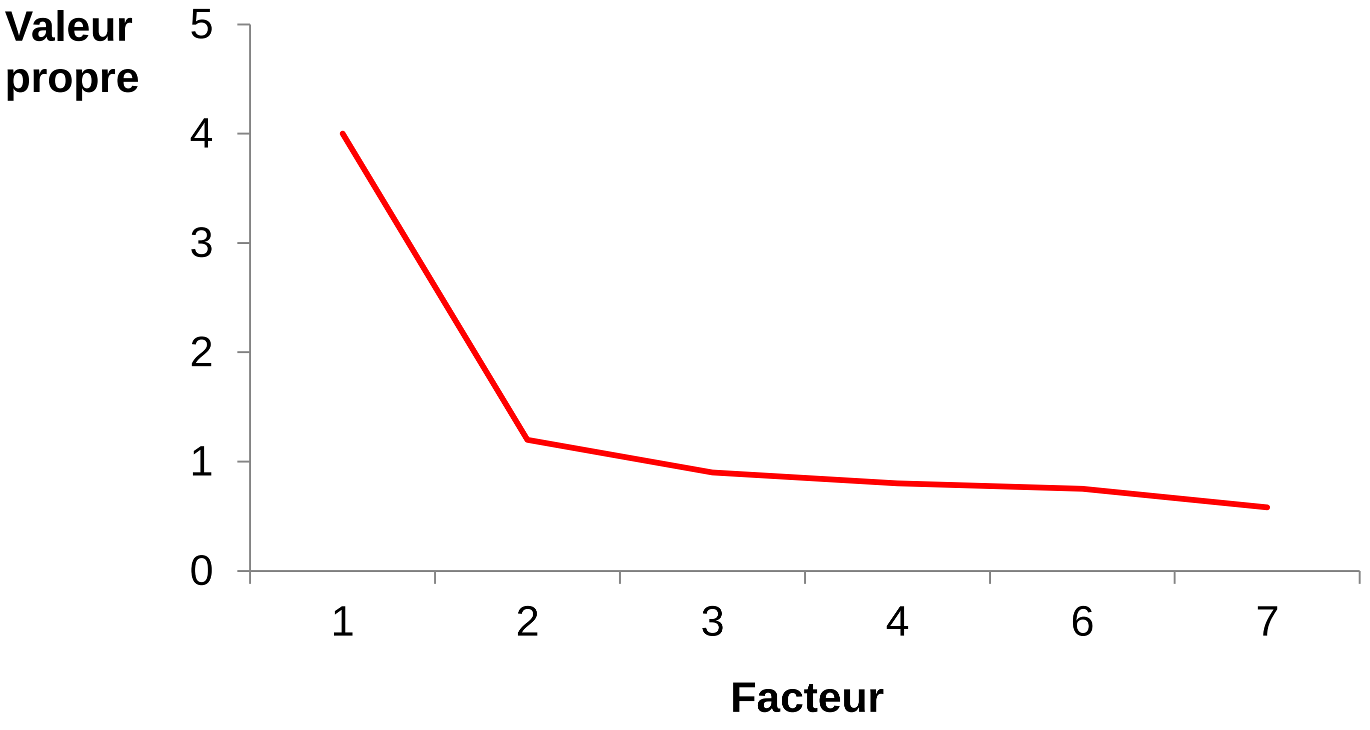
- La notion de valeur propre d’un facteur donne lieu à deux critères pour décider du nombre de facteurs à retenir.

  Le **critère de Kaiser** : ne retenir que les facteurs dont la valeur propre est > 1 (facteurs qui expliquent une part importante de la variance)

4. La matrice des saturations factorielles

Critère 3 : les valeurs propres

La **chute des valeurs propres** (Cattell) : ne pas retenir les facteurs dont les valeurs propres se situent sur une droite



4. La matrice des saturations factorielles

L’interprétation des facteurs

- Après avoir décidé du nombre de facteurs retenus, il faut les interpréter : leur attribuer une **signification psychologique.**

  Un facteur est une entité statistique qui n’a pas de signification en soi

- Chaque facteur doit être interprété en fonction du **contenu des variables qui saturent fortement dans celui-ci** (en valeur absolue)

4. La matrice des saturations factorielles

- Pour que les facteurs puissent être facilement interprétés, la matrice des saturations doit présenter une **structure simple :** chaque facteur doit avoir des **saturations proches de 0 ou de 1**.

  C’est le **principe de structure simple** (Thurstone, 1947)

Structure simple

	Facteur 1	Facteur 2
v1	-0.08	0.81
v2	0.10	0.92
v3	0.09	-0.86
v4	0.92	-0.11
v5	-0.89	0.19
v6	0.94	-0.21

Pas de structure simple

	Facteur 1	Facteur 2
v1	0.54	0.71
v2	0.69	0.74
v3	0.81	0.67
v4	-0.52	0.33
v5	0.36	-0.43
v6	0.75	0.67

4. La matrice des saturations factorielles

- La matrice des saturations peut être représentée par un graphique dans lequel :
  - ✓ facteurs = axes
  - ✓ variable = point
  - ✓ saturations = coordonnées

- La méthode standard pour trouver la structure simple est la **rotation orthogonale** (Kaiser, 1958).

  Cette méthode consiste à **modifier le pattern des saturations** des variables dans les facteurs afin de faire apparaître une structure simple.

  Graphiquement, on impose aux axes (= facteurs) une rotation

4. La matrice des saturations factorielles

- Exemple : matrice des saturations avec 4 items et 2 facteurs

  Avant rotation : pas de structure simple

	Facteur 1	Facteur 2
Je me sens tendu(e), crispé(e)	0.40	0.40
Je me sens nerveux(se), irritable	0.60	0.60
Je suis une personne chaleureuse	-0.50	0.50
Je me sens calme	-0.40	-0.40

4. La matrice des saturations factorielles

- Exemple : matrice des saturations avec 4 items et 2 facteurs

  Après rotation : structure simple

	Facteur 1	Facteur 2
Je me sens tendu(e), crispé(e)	0.57	0
Je me sens nerveux(se), irritable	0.85	0
Je suis une personne chaleureuse	0	0.71
Je me sens calme	-0.57	0

4. La matrice des saturations factorielles

- Exemple : matrice des saturations avec 4 items et 2 facteurs

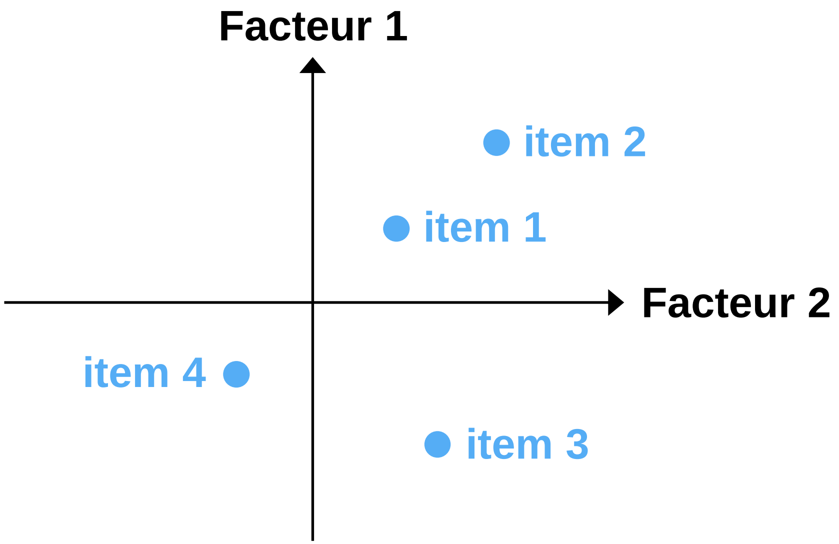
  Après rotation : structure simple

	Anxiété	Agréabilité
Je me sens tendu(e), crispé(e)	**0.57**	0
Je me sens nerveux(se), irritable	**0.85**	0
Je suis une personne chaleureuse	0	**0.71**
Je me sens calme	**-0.57**	0

4. La matrice des saturations factorielles

Avant rotation :

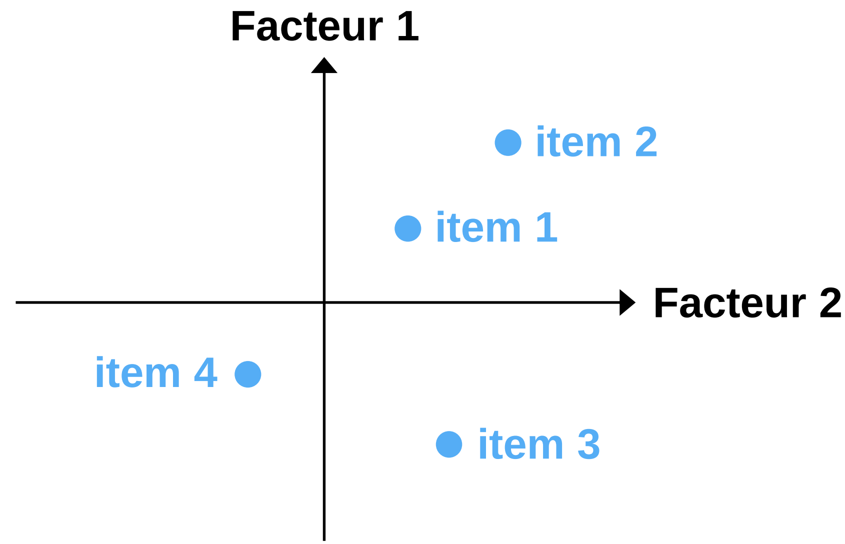
	Facteur 1	Facteur 2
item 1	0.40	0.40
item 2	0.60	0.60
item 3	-0.50	0.50
item 4	-0.40	-0.40



4. La matrice des saturations factorielles

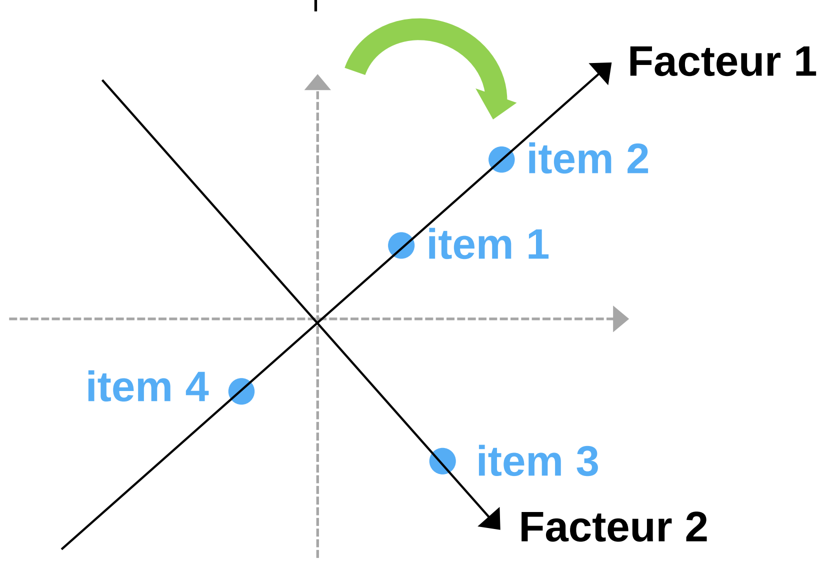
Avant rotation :

	Facteur 1	Facteur 2
item 1	0.40	0.40
item 2	0.60	0.60
item 3	-0.50	0.50
item 4	-0.40	-0.40



Après rotation :

	Facteur 1	Facteur 2
item 1	0.57	0
item 2	0.85	0
item 3	0	0.71
item 4	-0. 57	0



4. La matrice des saturations factorielles

- ATTENTION : la rotation **ne change pas le % de variance expliquée**

  La solution factorielle retenue est associée à un certain % de variance expliquée.

  Pour une solution factorielle donnée, la rotation ne change que la **répartition** de la variance expliquée entre les facteurs (principe des vases communicants)

- **Exemple :**

Avant rotation

	Facteur 1	**Facteur 2**	**h²**
item 1	0.40	0.40	0.32
item 2	0.60	0.60	0.72
item 3	-0.50	0.50	0.50
item 4	-0.40	-0.40	0.32
vp	0.93	0.93	**46.5%**

Après rotation

	Facteur 1	**Facteur 2**	**h²**
item 1	0.57	0	0.32
item 2	0.85	0	0.72
item 3	0	0.71	0.50
item 4	-0. 57	0	0.32
vp	1.37	0.50	**46.5%**

4. La matrice des saturations factorielles

- Deux types de rotations :
  - ✓ la rotation **orthogonale** (ou « varimax »)
    - les facteurs sont **indépendants**
    - graphiquement : les axes sont **orthogonaux**
  - ✓ la rotation **oblique** (ou « oblimin »)
    - les facteurs sont **corrélés**
    - graphiquement : les axes sont **obliques**

4. La matrice des saturations factorielles

- Pour une solution factorielle donnée, le principe de rotation fait qu'il y a une infinité de matrices de saturations possibles
  - ✓ toutes expliquent le même % de variance globale
  - ✓ toutes reproduisent aussi bien les corrélations observées

- La matrice de saturations qui doit être retenue est celle qui permet **l'interprétation des facteurs la plus cohérente**

- En résumé, l'objectif de l'analyse factorielle est de rendre compte d'un ensemble corrélations à l'aide d'un petit nombre de facteurs :
  - ✓ qui expliquent un **% important de la variance totale**
  - ✓ dont la **signification psychologique est claire**

4. La matrice des saturations factorielles

L'estimation des saturations

Comment les valeurs des saturations sont-elles trouvées ?

- Ces valeurs sont déterminées par un **algorithme** qui cherche à **minimiser les différences** entre les corrélations observées et les corrélations reproduites

- Cet algorithme procède par essai-erreur :
  - ✓ il démarre avec une certaine configuration de saturations
  - ✓ il calcule les corrélations résiduelles
  - ✓ il recommence avec une autre configuration de saturations
  - ✓ …
  - ✓ l'algorithme s'arrête lorsque l'ajustement ne s'améliore plus

5. Les scores factoriels

- Les facteurs latents identifiés par l'analyse factorielle sont des **nouvelles variables.**

  On peut calculer les scores des sujets à ces facteurs : il s'agit des **scores factoriels**

- **Exemple :** 3 variables et 1 facteur

  Le score d'un sujet au facteur est une **somme pondérée**

$$S_i = \lambda_1 V_1 + \lambda_2 V_2 + \lambda_3 V_3$$

sujet	v_1	v_2	v_3	F
1				
2				
…				
p				

	F
v_1	λ_1
v_2	λ_2
v_n	λ_3

5. Les scores factoriels

- **Cas particulier :** les saturations sont égales à 1

  $s_i = v_1 + v_2 + v_3$

- **Remarques :**

  Ce cas particulier caractérise l'échelonnement de Likert (on additionne les réponses aux items pour obtenir le score brut)

  Cet échelonnement est valable si les items mesurent le même facteur (**unidimensionnalité des items**)

6. Les deux logiques d'utilisation de l'AF

- Logique **exploratoire** :
  - ✓ **le modèle factoriel est défini *a posteriori*** (après les données)

  On cherche à faire émerger un modèle factoriel à partir des données qui permet d'expliquer au mieux des corrélations observées

- Logique **confirmatoire** :
  - ✓ **le modèle factoriel est défini *a priori*** (avant les données)

  On définit un ensemble d'hypothèses relatives aux facteurs et on vérifie si le modèle rend compte des données

7. Conditions d'application de l'AF

- L'application de l'AF requiert qu'un certain nombre de conditions soient remplies :
  - ✓ variables d'intervalles (*analyse factorielle des correspondances* pour variables nominales)
  - ✓ normalité des distributions des variables
  - ✓ linéarité des relations entre les variables
  - ✓ nombre d'observations : 5 sujets pour une variable