Analyse Factorielle

1 Analyse factorielle exploratoire

1.1 Introduction

Le terme *analyse factorielle* (factor analysis ou FA) désigne un ensemble de techniques dont les origines peuvent être situées dans les travaux de Pearson (1901). Elle a été tout d'abord développée par des psychologues, sans que les justifications théoriques, au niveau statistique ne soient clairement établies et a donné lieu à diverses controverses entre psychologues. C'est pourquoi on a pu parler à son sujet de "mouton noir des statistiques". Ce n'est que plus tard, vers 1940 que les fondements théoriques, au niveau statistique, ont été établis pour certaines des variantes de l'analyse factorielle.
Quelques noms associés à ces méthodes : Spearman, Thomson, Thurstone, Burt, etc.

Quelques remarques :

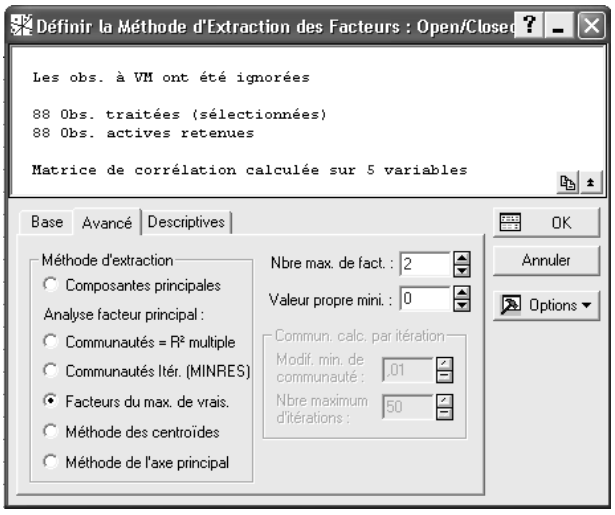
- l'intérêt porte ici sur les variables et non sur les individus statistiques ; il s'agit donc plus d'une méthode d'analyse multivariée que d'une méthode d'analyse multidimensionnelle.
- de nombreuses variantes existent : l'analyse factorielle est parfois désignée par le terme "analyse en facteurs communs et spécifiques", selon les variantes on parlera d'"analyse factorielle exploratoire" (exploratory factor analysis ou EFA) ou d'analyse factorielle confirmatoire (confirmatory factor analysis ou CFA). L'analyse en facteurs principaux (principal factor analysis ou PFA) est l'une des variantes de l'analyse factorielle.

1.2 Exemple introductif

Source : Mardia, K.V., Kent, J.T., Bibby, J.M., Multivariate Analysis, Academic Press, London 1979.

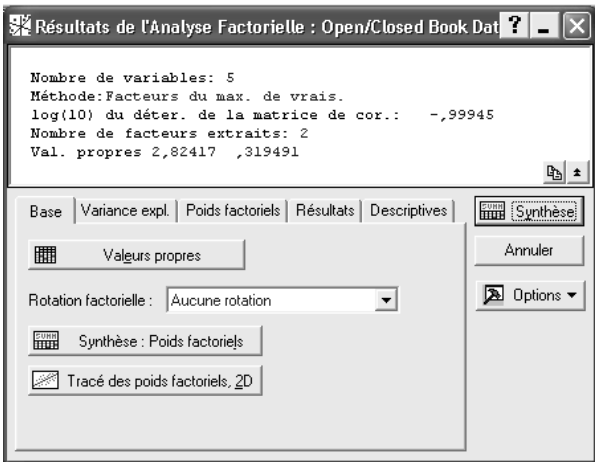
On dispose des notes obtenues par 88 sujets dans 5 matières : Mechanics(C), Vectors(C), Algebra(O), Analysis(O), Statistics(O). Pour deux matières, les étudiants n'avaient pas accès à leurs documents (closed book - C), pour les trois autres, les documents pouvaient être consultés (open book - O).

On utilise le menu Statistiques - Statistiques exploratoires multivariées - Analyse Factorielle de Statistica. Sous l'onglet "Avancé", on obtient le dialogue suivant :



Nous voyons que Statistica nous demande de fixer a priori le nombre de facteurs à extraire et nous propose plusieurs méthodes d'extraction des facteurs. Choisissons d'extraire deux facteurs par la méthode du maximum de vraisemblance.

Statistica fournit alors les résultats sous plusieurs onglets :



Sous l'onglet "Variance expliquée", on obtient notamment les 4 tableaux de résultats suivants :

- un tableau de "valeurs propres" :

Val. Propres (Open/Closed Book Data)				
Extraction : Facteurs du max. de vrais.				
	Val Propre	% Total variance	Cumul Val propre	Cumul %
1	2,824170	56,48341	2,824170	56,48341
2	0,319491	6,38983	3,143662	62,87323

- un tableau des "communautés" :

Communautés (Open/Closed Book) Rotation : Sans rot.			
	Pour 1 Facteur	Pour 2 Facteurs	R-deux Multiple
Mechanics(C)	0,394878	0,534103	0,376414
Vectors(C)	0,483548	0,580944	0,445122
Algebra(O)	0,808935	0,811431	0,671358
Analysis(O)	0,607779	0,648207	0,540864

Statistics(O)	0,529029	0,568977	0,479319

- un test d'adéquation du modèle aux données, utilisant une statistique du khi-2

Qualité d'ajust.,2 (Open/Closed Book Data) (Test de la nullité des éléments en dehors de la diagonale dans la matrice de corr.)				
	% expl.	Chi²	dl	p
Résultat	62,87323	0,074710	1	0,784601

- un tableau dit "de corrélation des résidus" :

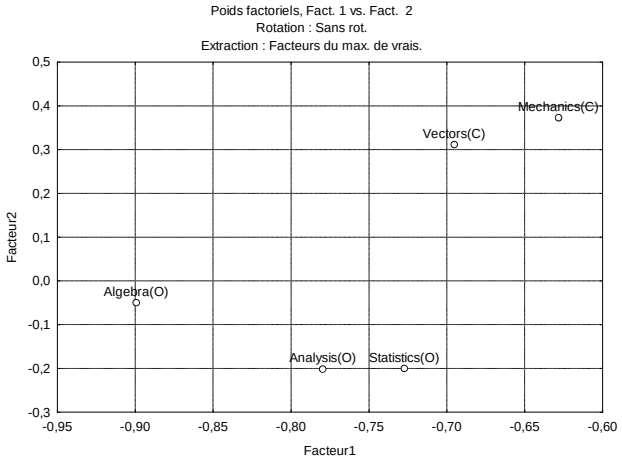
Corrélations des Résidus (Open/Closed Book Data) (Résidus marqués sont > ,100000)					
	Mechanics(C)	Vectors(C)	Algebra(O)	Analysis(O)	Statistics(O)
Mechanics(C)	0,47	-0,00	0,00	-0,01	0,01
Vectors(C)	-0,00	0,42	-0,00	0,01	-0,01
Algebra(O)	0,00	-0,00	0,19	-0,00	0,00
Analysis(O)	-0,01	0,01	-0,00	0,35	-0,00
Statistics(O)	0,01	-0,01	0,00	-0,00	0,43

L'onglet "Poids factoriels" nous offre la possibilité de transformer les facteurs par rotation. Il nous donne également les résultats suivants :

- les poids factoriels des variables selon chacun des facteurs :

Poids Factoriels(Sans rot.) (Open/Closed Book Data) (Poids marqués >,700000)		
	Facteur 1	Facteur 2
Mechanics(C)	-0,628393	0,373128
Vectors(C)	-0,695376	0,312083
Algebra(O)	-0,899408	-0,049958
Analysis(O)	-0,779602	-0,201066
Statistics(O)	-0,727344	-0,199869
Var. Expl.	2,824170	0,319491
Prp.Tot	0,564834	0,063898

- Le graphique correspondant :



Enfin, l'onglet "Résultats" nous fournit :

- les coefficients des scores factoriels :

Coefficients des Scores Factoriels (Open/Closed Book Data) Extraction : Facteurs du max. de vrais.		
	Facteur 1	Facteur 2
Mechanics(C)	-0,131635	0,457102
Vectors(C)	-0,161949	0,425053
Algebra(O)	-0,465496	-0,151209
Analysis(O)	-0,216280	-0,326209
Statistics(O)	-0,164691	-0,264662

- les scores factoriels des individus :

Scores Factoriels (Open/Closed Book Data) Extraction : Facteurs du max. de vrais.		
	Facteur 1	Facteur 2
1	-2,05705	0,73671
2	-2,51565	-0,00951
3	-2,09181	0,35850
4	-1,51263	0,02871
....	...	....

Comme on peut le voir, l'analyse factorielle, par certains aspects, semble ressembler à l'analyse en composantes principales. Mais qu'en est-il véritablement ?

1.3 Justification conceptuelle de l'analyse factorielle

L'analyse en composantes principales est une méthode qui, à partir d'un ensemble X_1, X_2, ... X_p de variables observées corrélées entre elles permet d'obtenir un nouvel ensemble Y_1, Y_2, ... , Y_p de variables non corrélées tout en conservant la dispersion observée entre les individus. La méthode travaille sur les variances dans la mesure ou Y_1 est la combinaison linéaire des X_i ayant la plus grande variance, Y_2 satisfait à la même condition tout en étant non corrélée avec Y_1, etc. L'analyse en composantes principales est essentiellement une transformation des données. C'est une méthode descriptive qui ne fait aucune hypothèse a priori sur les variables à traiter.

L'analyse factorielle est une méthode inférentielle qui vise à expliquer la matrice des covariances par un minimum, ou un petit nombre de variables hypothétiques (non observables) : les facteurs.

Par exemple, Spearman fait passer trois tests d'aptitude à un échantillon de sujets et les scores observés aux trois tests produisent la matrice de corrélation suivante :

$$\begin{bmatrix} 1 & 0,83 & 0,78 \\ 0,83 & 1 & 0,67 \\ 0,78 & 0,67 & 1 \end{bmatrix}$$

On souhaiterait étudier l'hypothèse suivante :
Les valeurs observées sont la somme de deux éléments :
- Une quantité proportionnelle à une variable ou facteur (non observable) mesurant l'intelligence du sujet
- Une quantité spécifique au test, à laquelle s'ajoute une erreur aléatoire.

Autrement dit :
- On a observé un ensemble X_1, X_2, ..., X_p de variables sur un échantillon
- On fait l'hypothèse que ces variables dépendent (linéairement) en partie de k variables non observables, ou variables latentes ou facteurs F_1, F_2, ..., F_k.

On cherche donc à décomposer les variables observées X_i (supposées centrées) de la façon suivante :

$$X_i = \sum_{r=1}^{k} l_{ir} F_r + E_i$$

avec les conditions suivantes :
- Le nombre k de facteurs est fixé à l'avance.
- Les facteurs F_r sont centrés réduits, non corrélés entre eux
- Les termes d'erreur E_i sont non corrélés avec les facteurs
- Les termes d'erreur E_i sont non corrélés entre eux.

Remarque. Dans la formulation ci-dessus, on a choisi pour simplifier, de ne pas distinguer les paramètres observés sur l'échantillon des paramètres théoriques sur la population. Comme nous n'envisageons de développements théoriques à partir de ces équations, ce choix n'a guère d'importance.

Afin d'exploiter les conditions indiquées ci-dessus, le traitement mathématique porte sur les matrices de covariance (si les données ne sont pas réduites) ou de corrélation (si elles le sont). Notons c_{ij} la covariance des variables X_i et X_j et v_i la variance de la variable E_i.

On a les égalités :

$$c_{ii} = \sum_{r=1}^{k} l_{ir}^2 + v_i$$

$$c_{ij} = \sum_{r=1}^{k} l_{ir} l_{jr} \quad \text{si} \quad i \neq j$$

c'est-à-dire, matriciellement :

$$C = LL' + V .$$

Ce problème n'admet en général pas une solution unique. On ajoute alors une condition supplémentaire telle que :

$$J = L'V^{-1}L \quad \text{est diagonale}$$

Mais, toute rotation des facteurs ainsi déterminés fournit également aussi une solution.

Vocabulaire : les coefficients l_{ir} sont appelés *poids factoriels* (loadings) des variables sur les facteurs. La quantité $h_i^2 = \sum_{r=1}^{k} l_{ir}^2$ qui représente la partie de la variance de X_i due aux facteurs et dont "partagée" avec les autres variables est appelée communauté (*communality*).

Remarque. L'analyse factorielle n'exige pas que les données de départ soient centrées et réduites. Pour certaines méthodes insensibles aux échelles (scale free) les résultats ne dépendent pas d'une éventuelle réduction des données. Il importe par ailleurs de remarquer que, lorsque les données sont centrées réduites, les poids factoriels sont les coefficients de corrélation entre les facteurs et les variables, et la communauté d'une variable représente le carré du coefficient de corrélation multiple de cette variable par rapport aux facteurs.

1.4 Méthodes d'extraction des facteurs

Comme nous le montre Statistica, plusieurs méthodes d'extraction des facteurs ont été proposées et fournissent des résultats analogues, mais pas identiques.

1.4.1 Analyse en composantes principales

Une première méthode (souvent appelée PCA, *principal component analysis* dans les ouvrages anglo-saxons) utilise les valeurs propres et la diagonalisation des matrices. Les résultats sont alors identiques à ceux obtenus par ACP normée, se limitant à k axes. La différence la plus importante par rapport à l'ACP est la possibilité d'effectuer une rotation des facteurs.

1.4.2 Méthode de l'axe principal

La méthode de l'axe principal (PFA, principal factor analysis ou PAF, principal axis factoring) est une méthode itérative cherchant à maximiser les communautés. Les estimations initiales des communautés sont les coefficients de corrélation multiple de chaque variable par rapport à toutes les autres.

1.4.3 Un aperçu sur la notion d'estimation du maximum de la vraisemblance

1.4.3.1 Vraisemblance d'une valeur d'un paramètre :

On cherche à répondre à des questions du type : "Etant donné des résultats observés sur un échantillon, est-il vraisemblable qu'un paramètre donné de la population ait telle valeur ?".

Exemple 1 : (variable discrète) Lors d'un référendum, on interroge trois personnes. Deux déclarent voter "oui", la troisième déclare voter "non".
Au vu de ces observations, laquelle de ces deux hypothèses est la plus vraisemblable :

- Le résultat du référendum sera 40% de "oui"
- Le résultat du référendum sera 60% de "oui".

Solution. Si le résultat du référendum est de 40% de "oui", la probabilité d'observer trois personnes votant respectivement "oui", "oui" et "non" est : P1 = 0,4x0,4x0,6 = 0,096. Si le résultat du référendum est de 60% de oui, la même probabilité est : P2 = 0,6x0,6x0,4 = 0,144. La seconde hypothèse est donc plus vraisemblable que la première.

Exemple 2 (variable continue) Lors d'un test effectué sur un échantillon de 5 sujets, on a observé les scores suivants :

90, 98, 103, 107, 112.

Deux modèles sont proposés pour représenter la distribution des scores dans la population parente :

- La loi normale de moyenne 100 et d'écart type 15
- La loi normale de moyenne 102 et d'écart type 10.

Quel est le modèle le plus vraisemblable ?

Dans le cas d'une variable continue, on utilise la valeur de la distribution de la loi théorique au lieu de la probabilité de la valeur observée. La vraisemblance associée à chaque hypothèse, calculée à l'aide d'Excel, est donc :

Obs	Modèle 1	Modèle 2

90	0,02130	0,01942
98	0,02636	0,03683
103	0,02607	0,03970
107	0,02385	0,03521
112	0,01931	0,02420
Vraisemblance	6,74E-09	2,42E-08

On voit que le modèle 2, dont la vraisemblance est de 2,42 10^{-8} est plus vraisemblable que le modèle 1.

1.4.4 Estimation du maximum de vraisemblance

L'estimation du maximum de vraisemblance (EMV, maximum likelihood estimation ou MLE dans les ouvrages anglo-saxons) est la valeur du paramètre pour laquelle la vraisemblance est maximum. Reprenons l'exemple du référendum.

Si le pourcentage de "oui" est p, la probabilité d'observer trois personnes votant respectivement "oui", "oui" et "non" est : $P = p^2(1-p)$. La dérivée de cette fonction est $P' = p(2 - 3p)$. Cette dérivée s'annule pour p=2/3=0,67, et cette valeur correspond à un maximum de P. Ainsi, au vu des observations, le résultat le plus vraisemblable est : 67% de "oui" ... ce qui n'est guère surprenant.

On notera que les calculs de vraisemblance sont souvent multiplicatifs et conduisent à des nombres très proches de 0. C'est pourquoi on utilise généralement la fonction L, opposée du logarithme de la vraisemblance. Dans le cas précédent on aurait ainsi :

$$L = - \ln P = - 2 \ln p - \ln(1 - p).$$

La recherche de l'estimation du maximum de vraisemblance revient alors à chercher le minimum de cette fonction.

1.4.5 L'analyse factorielle du maximum de vraisemblance

La méthode du maximum de vraisemblance est la seule qui permette de calculer un test statistique d'adéquation du modèle.

Dans cette méthode, on fixe a priori un nombre k de facteurs à extraire. Les poids factoriels des variables sur les différents facteurs sont alors déterminés de manière à optimiser une fonction de vraisemblance.

Cette méthode utilise des concepts de statistique inférentielle classiques. Mais elle suppose que les données vérifient des propriétés de régularité convenables. La condition d'application est la multinormalité des variables Xi sur la population parente de l'échantillon observé.

Un test statistique permet d'évaluer la validité du résultat. Selon Lawley et Maxwell, les hypothèses H_0 et H_1 du test sont :

H_0 : Il y a exactement *k* facteurs communs.

H_1 : Plus de *k* facteurs sont nécessaires.

La statistique utilisée dépend évidemment des covariances des X_i et des poids factoriels obtenus. Elle dépend également de la taille de l'échantillon tiré. Elle suit approximativement une loi du khi-2 avec $\frac{1}{2}\left[(p-k)^2 - (p+k)\right]$ degrés de liberté (p : nombre de variables, k : nombre de facteurs extraits).

Selon Lawley et Maxwell, si le khi-2 trouvé excède la valeur critique correspondant au niveau de significativité choisi, H0 est rejetée, et il faut considérer au moins k+1 facteurs dans le modèle.

Remarques.

1. On doit avoir $(p+k)<(p-k)^2$ ce qui limite le nombre de facteurs.

2. Certains auteurs énoncent une règle en termes de taille des échantillons pour utiliser cette statistique. Par exemple, Mardia et Kent indiquent : $n \geq p+50$.

3. Cette statistique peut être utilisée pour déterminer le nombre de facteurs à extraire. On calcule alors la statistique pour k=1, k=2, ... L'extraction d'un facteur supplémentaire se traduit par une diminution de la valeur de la statistique, mais également par une diminution du nombre de degrés de liberté. La p-value correspondante n'est donc pas nécessairement améliorée par l'augmentation du nombre de facteurs. On choisit ensuite le nombre de facteurs qui conduit à la meilleure p-value (celle qui est la plus proche de 1).

4. Cette statistique est malheureusement très sensible à la taille de l'échantillon.

1.5 Résultats obtenus - Scores des individus

1.5.1 Poids factoriels et communautés

Les résultats obtenus sont essentiellement constitués des poids factoriels des variables sur les différents facteurs et des communautés des différentes variables. Sur l'exemple donné en introduction, les poids factoriels sont donnés par :

Poids Factoriels(Sans rot.) (Open/Closed Book Data) (Poids marqués >,700000)		
	Facteur 1	Facteur 2
Mechanics(C)	-0,628393	0,373128
Vectors(C)	-0,695376	0,312083
Algebra(O)	-0,899408	-0,049958
Analysis(O)	-0,779602	-0,201066
Statistics(O)	-0,727344	-0,199869
Var. Expl.	2,824170	0,319491
Prp.Tot	0,564834	0,063898

On cherche alors à attribuer une signification à chacun des facteurs. Sur notre exemple, toutes les variables sont fortement corrélées (négativement) avec le premier facteur, qui peut ainsi apparaître comme une mesure "globale" relative à l'individu. Quant au deuxième facteur, il oppose les matières évaluées à livre fermé (poids factoriels positifs) à celles évaluées à livre ouvert (poids factoriels négatifs). On pourra parler de facteur *unipolaire* dans le premier cas, de facteur *bipolaire* dans le second.

Comme nous l'avons souligné plus haut, les facteurs ne sont pas déterminés de manière unique, et notamment, toute transformation des facteurs par rotation orthogonale conduit à une autre solution. Il peut intéressant d'effectuer une telle rotation pour obtenir des facteurs plus faciles à interpréter. C'est ce que nous ferons un peu plus loin.

Dans l'exemple traité en introduction les communautés sont les suivantes :

Communautés (Open/Closed Book) Rotation : Sans rot.			
	Pour 1 Facteur	Pour 2 Facteurs	R-deux Multiple
Mechanics(C)	0,394878	0,534103	0,376414
Vectors(C)	0,483548	0,580944	0,445122
Algebra(O)	0,808935	0,811431	0,671358
Analysis(O)	0,607779	0,648207	0,540864
Statistics(O)	0,529029	0,568977	0,479319

Ces quantités se calculent facilement à partir du tableau des poids factoriels. Par exemple, pour la variable Mechanics(O), la communauté se calcule de la manière suivante :

$$h_1^2 = (\text{-}0{,}628393)^2 + (0{,}373128)^2 = 0{,}534103$$

Pour une ACP, ces quantités sont interprétées en termes de qualité de représentation, ou de déformation due à la projection. Dans le cadre de l'analyse factorielle, elles nous indiquent quelle est la part de variabilité de chacune des variables observées qui participe à la variance "commune" et, par différence, quelle est la part qui est spécifique à chaque variable, et donc non prise en compte dans le modèle factoriel. Par exemple, pour la variable Algebra(O), la part "commune" est de 81% et la part spécifique, non prise en compte par les facteurs est de 19%.

1.5.2 Scores des individus

Les valeurs prises par les différents facteurs (qui sont des variables statistiques, même si elles ne sont pas observables directement) sur les individus statistiques composant l'échantillon sont appelées *scores des individus*. Contrairement à l'ACP, l'exploitation des résultats d'une analyse factorielle n'utilise généralement pas ces scores. En effet, les facteurs ne prennent pas en compte la totalité de la variation observée sur les données et celles-ci comportent une part de variation aléatoire due aux fluctuations d'échantillonnage. Les scores des individus ne peuvent donc pas être calculés de manière exacte mais seulement estimés à partir des autres résultats. Plusieurs méthodes ont été proposées, par exemple une méthode basée sur le maximum de vraisemblance a été proposée par Bartlett : le *Bartlett factor score*. La justification de ces méthodes approchées est particulièrement délicate lorsqu'on travaille sur les corrélations et non sur les covariances.

Dans l'exemple donné en introduction, Statistica nous donne d'une part l'expression des facteurs en fonction des variables :

Coefficients des Scores Factoriels (Open/Closed Book Data) Extraction : Facteurs du max. de vrais.		
	Facteur 1	Facteur 2
Mechanics(C)	-0,131635	0,457102
Vectors(C)	-0,161949	0,425053
Algebra(O)	-0,465496	-0,151209
Analysis(O)	-0,216280	-0,326209
Statistics(O)	-0,164691	-0,264662

Ainsi, par exemple :

$$\text{Facteur}\,1 = -0{,}132 \times \text{Mechanics} - 0{,}162 \times \text{Vectors} - 0{,}465 \times \text{Algebra} - 0{,}216 \times \text{Analysis} - 0{,}165 \times \text{Statistics}$$

D'autre part, il donne également les valeurs des facteurs sur les différentes observations, telles qu'elles peuvent être calculées à partir des formules précédentes et des valeurs centrées réduites associées aux valeurs observées. Par exemple pour le premier sujet, le logiciel indique :

	Facteur 1	Facteur 2
1	-2,05705	0,73671

Les valeurs centrées réduites des 5 variables sont :

	Mechanics(C)	Vectors(C)	Algebra(O)	Analysis(O)	Statistics(O)
1	2,17573873	2,38907869	1,54334732	1,36866891	2,24235647

Et on vérifie que :

$$\text{Facteur}\,1_{\text{Sujet}\,1} = -0{,}132 \times 2{,}176 - 0{,}162 \times 2{,}390 - 0{,}465 \times 1{,}543 - 0{,}216 \times 1{,}369 - 0{,}165 \times 2{,}242 = -2{,}057$$

Remarque. A l'exception des scores factoriels des individus, l'ensemble des résultats d'une analyse factorielle peut être obtenu à partir de la matrice des corrélations (ou des covariances) des variables, et de la taille de l'échantillon. C'est pourquoi Statistica propose de deux formats pour les données d'entrée : données brutes ou matrice de corrélations.

1.6 Rotation des facteurs : rotations orthogonales, rotations obliques

Les facteurs extraits par l'une ou l'autre des méthodes précédentes ne sont pas déterminés de manière unique et c'est généralement une condition arbitraire qui permet de choisir une solution dans l'ensemble des solutions possibles.
Il en résulte que les facteurs ainsi produits ne sont pas toujours simples à interpréter. Mais toute rotation sur les facteurs produit une autre solution et on peut être tenté de rechercher une solution qui "fasse sens", c'est-à-dire qui produise des facteurs plus simples à interpréter.

Il importe de noter que la transformation par rotation n'affecte pas l'adéquation du modèle aux données. Les communautés, notamment, restent les mêmes. Mais les solutions avant ou après rotation peuvent être interprétés de façon notablement différente.

Ainsi, sur notre exemple :

	Poids Factoriels (sans rotation)		Poids Factoriels (après rotation varimax normalisé)	
	Facteur 1	Facteur 2	Facteur 1	Facteur 2
Mechanics(C)	-0,628393	0,373128	0,270028	0,679108
Vectors(C)	-0,695376	0,312083	0,360346	0,671636
Algebra(O)	-0,899408	-0,049958	0,742939	0,509384
Analysis(O)	-0,779602	-0,201066	0,740267	0,316563
Statistics(O)	-0,727344	-0,199869	0,698141	0,285615
Var. Expl.	2,824170	0,319491	1,790119	1,353543
Prp.Tot	0,564834	0,063898	0,358024	0,270709

On examine les poids factoriels après rotation varimax. Les trois matières évaluées à livre ouvert sont alors fortement corrélées avec le premier facteur, alors que le second facteur correspond aux deux matières évaluées à livre fermé et dans une moindre mesure à l'algèbre.

La rotation la plus fréquemment utilisée est la rotation varimax (Kaiser 1958). L'effet produit par une telle rotation est généralement le suivant : pour chaque facteur, les poids factoriels élevés concernent un nombre réduit de variables et les autres poids factoriels sont proches de 0.

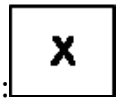
D'autres rotations ont également été proposées. Les rotations dites orthogonales produisent des facteurs non corrélés entre eux, tandis que les transformations par rotation oblique produisent de nouveaux facteurs qui peuvent être corrélés.

2 Analyse factorielle confirmatoire

L'analyse factorielle confirmatoire est apparentée à l'analyse factorielle exploratoire. Mais c'est aussi un cas particulier de modélisation d'équations structurelles (SEM : structural equation modelling). Différents algorithmes ont été développés dans ce cadre (par exemple : LISREL).

En analyse factorielle confirmatoire, le point de vue est différent de celui de l'analyse factorielle exploratoire : on se fixe a priori un modèle :

- nombre de facteurs
- corrélations éventuelles entre ces facteurs
- termes d'erreur attachés à chaque variable observée et corrélations éventuelles entre eux
- pour chaque facteur, variables avec lesquelles il sera significativement corrélé.

- Une variable observée est représentée dans un rectangle : 

- Une variable latente (un facteur) est représentée dans un ovale : 

- Un terme d'erreur, ou perturbation du modèle, est représenté par une variable sans cadre : E1

- Une flèche entre deux variables signifie que les variations de la seconde sont dues, au moins en partie, aux variations de la première.

Exemple :

Source : pages en ligne de Michael Friendly à l'adresse :
http://www.psych.yorku.ca/lab/psy6140/fa/facfoils.htm

Calsyn et Kenny (1971) ont étudié la relation entre les aptitudes perçues et les aspirations scolaires de 556 élèves du 8è grade. Les variables observées étaient les suivantes :

Self : auto-évaluation des aptitudes
Parent : évaluation par les parents
Teacher : évaluation par l'enseignant
Friend : évaluation par les amis
Educ Asp : aspirations scolaires
Col Plan : projets d'études supérieures

Sur l'échantillon étudié, les corrélations observées entre ces six variables sont les suivantes :

	Self	Parent	Teacher	Friend	Educ Asp	Col Plan
Self	1,00	0,73	0,70	0,58	0,46	0,56
Parent	0,73	1,00	0,68	0,61	0,43	0,52
Teacher	0,70	0,68	1,00	0,57	0,40	0,48

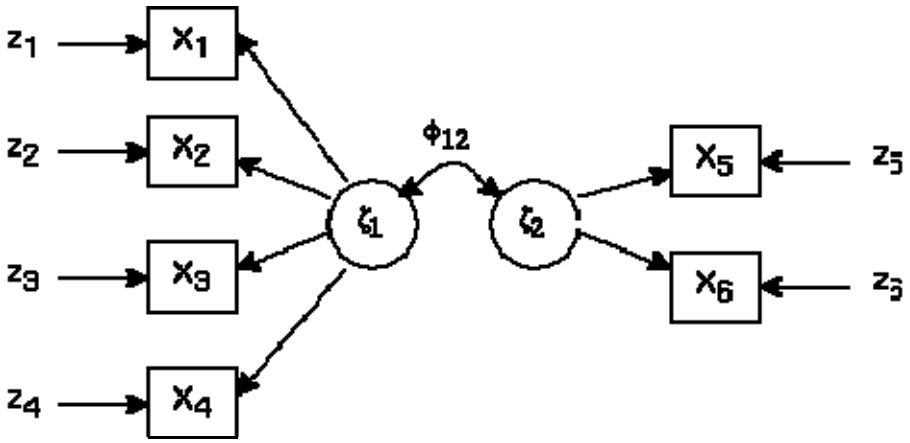
Friend	0,58	0,61	0,57	1,00	0,37	0,41
Educ Asp	0,46	0,43	0,40	0,37	1,00	0,72
Col Plan	0,56	0,52	0,48	0,41	0,72	1,00

Le modèle à tester fait les hypothèses suivantes :

- Les 4 premières variables mesurent la variable latente "aptitudes"
- Les deux dernières mesurent la variable latente "aspirations".

Ce modèle est-il valide ? Et, s'il en est bien ainsi, les deux variables latentes sont elles corrélées.

Le schéma correspondant à ce modèle peut être représenté ainsi (les variables sont renommées X_1 à X_6 et les facteurs sont désignés par la lettre grecque ζ dans ce schéma emprunté à Michael Friendly) :



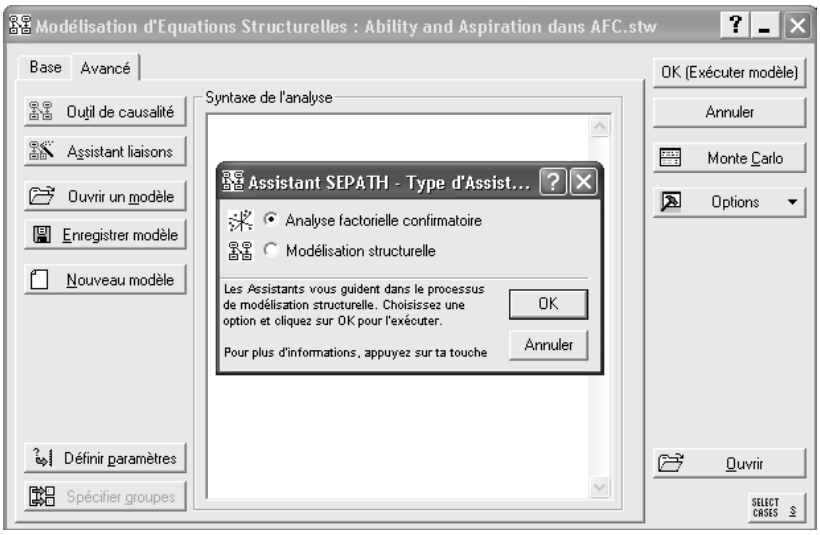
Traitement avec Statistica.

La matrice de corrélations précédente est saisie comme objet de type "matrice" de Statistica :

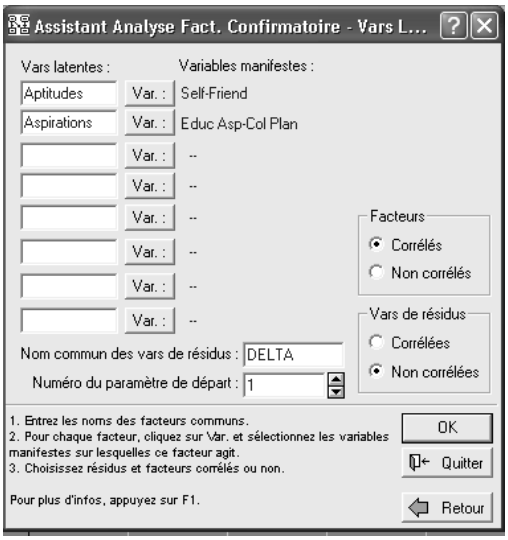
	Feuille de données3					
	1 Self	2 Parent	3 Teacher	4 Friend	5 Educ Asp	6 Col Plan
Self	1,00	0,73	0,70	0,58	0,46	0,56
Parent	0,73	1,00	0,68	0,61	0,43	0,52
Teacher	0,70	0,68	1,00	0,57	0,40	0,48
Friend	0,58	0,61	0,57	1,00	0,37	0,41
Educ Asp	0,46	0,43	0,40	0,37	1,00	0,72
Col Plan	0,56	0,52	0,48	0,41	0,72	1,00
Moyennes	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
Ec-Types	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000
Nb Obs.	556,00000					
Matrice	1,00000					

On choisit ensuite le menu Statistiques - Modèles linéaires / non linéaires avancés - Modélisation d'équations structurelles.

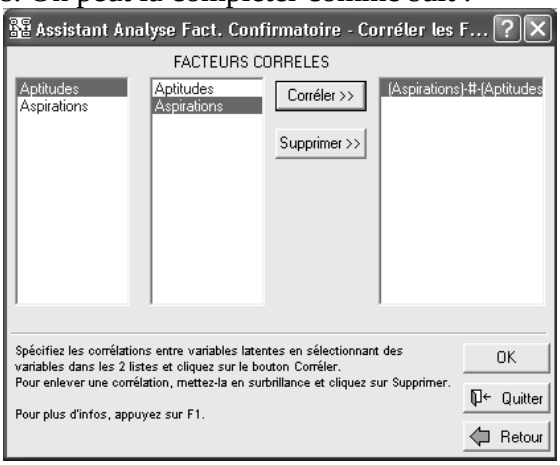
Sous l'onglet "Avancé", on clique sur le bouton "Assistant liaisons" et on choisit l'option "Analyse factorielle confirmatoire" :



On peut alors saisir le modèle sous la forme suivante :



Lorsqu'on clique sur le bouton OK, Statistica affiche une fenêtre permettant d'indiquer les corrélations entre les facteurs. On peut la compléter comme suit :

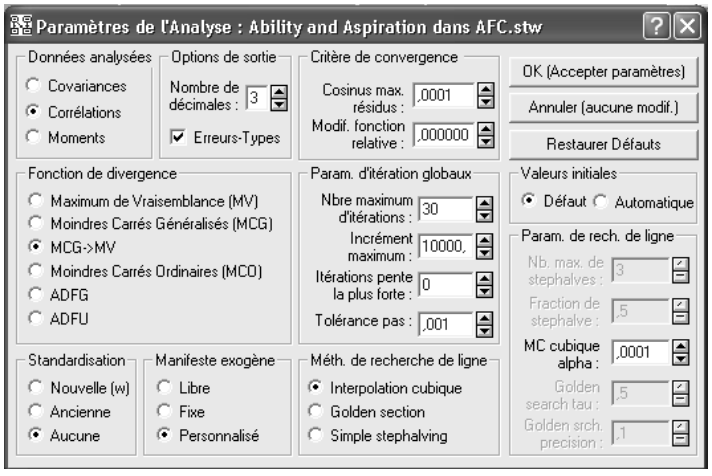


Lorsque la fenêtre suivante s'affiche, cliquer sur OK :

Le modèle spécifié est alors traduit en "langage" PATH1 sous la forme suivante :

```
(Aptitudes)-1->[Self]
(Aptitudes)-2->[Parent]
(Aptitudes)-3->[Teacher]
(Aptitudes)-4->[Friend]

(Aspirations)-5->[Educ Asp]
(Aspirations)-6->[Col Plan]

(DELTA1)-->[Self]
(DELTA2)-->[Parent]
(DELTA3)-->[Teacher]
(DELTA4)-->[Friend]
(DELTA5)-->[Educ Asp]
(DELTA6)-->[Col Plan]

(DELTA1)-7-(DELTA1)
(DELTA2)-8-(DELTA2)
(DELTA3)-9-(DELTA3)
(DELTA4)-10-(DELTA4)
(DELTA5)-11-(DELTA5)
(DELTA6)-12-(DELTA6)

(Aspirations)-13-(Aptitudes)
```

Ce "programme" peut éventuellement être enregistré dans un fichier autonome.
Cliquez ensuite sur le bouton "Paramètres de l'analyse". Le dialogue qui s'affiche est particulièrement abscons, mais nous nous contenterons d'y indiquer que les données analysées sont de type "corrélations", en laissant les autres paramètres à leurs valeurs par défaut :



Cliquez ensuite sur OK (Exécuter modèle), puis sur le bouton OK de la fenêtre suivante.

Le bouton "Synthèse du modèle" permet d'obtenir la feuille de résultats suivante :

Modèle Estimé (Ability and Aspiration dans AFC.stw)				
	Estimation Paramètre	Erreur Type	Stat. T	Niveau Proba

(Aptitudes)-1->[Self]	0,863	0,015	57,973	0,000
(Aptitudes)-2->[Parent]	0,849	0,016	54,296	0,000
(Aptitudes)-3->[Teacher]	0,805	0,018	44,287	0,000
(Aptitudes)-4->[Friend]	0,695	0,025	28,217	0,000
(Aspirations)-5->[Educ Asp]	0,775	0,026	30,279	0,000
(Aspirations)-6->[Col Plan]	0,929	0,024	39,165	0,000
(DELTA1)-->[Self]				
(DELTA2)-->[Parent]				
(DELTA3)-->[Teacher]				
(DELTA4)-->[Friend]				
(DELTA5)-->[Educ Asp]				
(DELTA6)-->[Col Plan]				
(DELTA1)-7-(DELTA1)	0,255	0,026	9,915	0,000
(DELTA2)-8-(DELTA2)	0,279	0,027	10,487	0,000
(DELTA3)-9-(DELTA3)	0,352	0,029	12,020	0,000
(DELTA4)-10-(DELTA4)	0,517	0,034	15,078	0,000
(DELTA5)-11-(DELTA5)	0,399	0,040	10,061	0,000
(DELTA6)-12-(DELTA6)	0,137	0,044	3,111	0,002
(Aspirations)-13-(Aptitudes)	0,666	0,031	21,528	0,000

On retrouve dans ce tableau le poids factoriel de chacune des variables sur le facteur spécifié par le modèle (sur une seule colonne - ce qui ne facilite pas la lecture du tableau). On y trouve également les variances des termes d'erreur DELTA1 à DELTA6 et enfin l'estimation de la corrélation entre les facteurs Aspirations et Aptitudes : 0,666.

Ces résultats seraient plus lisibles disposés de la façon (plus classique) suivante :

Modèle Estimé (Ability and Aspiration dans AFC.stw)				
	Aptitudes	Aspirations	Communauté	Spécificité
Self	0,863	0	0,745	0,255
Parent	0,849	0	0,721	0,279
Teacher	0,805	0	0,648	0,352
Friend	0,695	0	0,483	0,517
Educ Asp	0	0,775	0,601	0,399
Col Plan	0	0,929	0,863	0,137

Dans ce tableau, les communautés sont simplement les carrés des poids factoriels et les spécificités sont les compléments à 1 des communautés.

Le logiciel donne ensuite de nombreux indices évaluant la qualité du modèle.

En particulier, le bouton "Statistiques de synthèse" nous fournit la valeur d'une statistique du khi-2 du maximum de vraisemblance :

Statistiques de Synthèse (Ability and Aspiration dans AFC.stw)	
	Valeur
Chi-Deux MV	9,256
Degrés de Liberté	8,000
Niveau p	0,321

La valeur trouvée ici (p-value = 0,32) montre une bonne adéquation du modèle aux données.
D'autres indices de qualités

D'autres indices sont aussi couramment utilisés :

- AIC (Akaike Information Criterion ou Critère d'information de Akaike)
- BIC (Bayesian Information Criterion ou Critère Bayésien de Schwarz)
- TLI (Tucker-Lewis Index) : les modèles "acceptables" doivent vérifier TLI>0,90, les "bons" modèles, TLI>0,95
- RMSEA (root mean square error of approximation). les modèles "acceptables" doivent vérifier RMSEA<=0,08, les "bons" modèles, RMSEA <= 0,05
- CFI (Comparative Fit Index)

3 Bibliographie :

Ouvrages :

Lawley, D.N., Maxwell, A.E., Factor Analysis as a Statistical Method, Butterworths Mathematical Texts, England, 1963.
Mardia, K.V., Kent, J.T., Bibby, J.M., Multivariate Analysis, Academic Press, London 1979.

Articles :

Sites internet :
http://faculty.chass.ncsu.edu/garson/PA765/factor.html

Documents mis en ligne par Michael Friendly et notamment :
http://www.psych.yorku.ca/lab/psy6140/lectures/

Une discussion intéressante sur l'utilisation pratique de l'analyse factorielle :
http://core.ecu.edu/psyc/wuenschk/stathelp/EFA.htm

Site pour télécharger ce polycopié et les fichiers d'exemples :
http://geai.univ-brest.fr/~carpenti/

4 Table des matières