Analyse multidimensionnelle des données
Master 2ème année - Psychologie Sociale des Représentations

Réf. (polycopié et fichiers de données utilisés) :
http://geai.univ-brest.fr/~carpenti/

1 Présentation

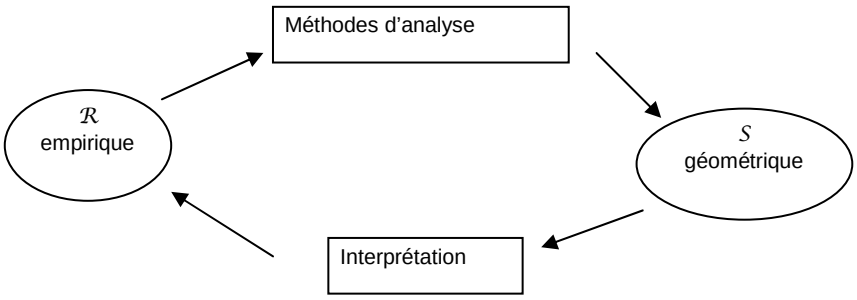
1.1 Introduction

Comment peut-on définir l'analyse multidimensionnelle des données ?

L'analyse statistique élémentaire s'applique à des situations dans lesquelles une ou deux variables ont été observées sur un ensemble d'individus statistiques (populations ou échantillons). L'extension de ces méthodes aux cas où le nombre de variables devient plus élevé est souvent appelé analyse *multivariée*. Cependant les conclusions ou résultats obtenus par ces méthodes restent de même nature, *unidimensionnelle*. Par exemple, la MANOVA (analyse de variance multivariée) permet d'étudier l'effet de facteurs de variation sur un "vecteur" de variables dépendantes, mais apporte une conclusion analogue à celle de l'ANOVA : les facteurs ont (ou n'ont pas) un effet sur le vecteur des VD.

L'analyse multidimensionnelle (ou plutôt, les méthodes qui en relèvent) étudie également des situations où un ensemble de variables doit être étudié simultanément sur un ensemble d'objets statistiques. Par nature, ces données se modélisent dans un espace à plusieurs dimensions. Mais, à la différence des méthodes précédentes, l'analyse multidimensionnelle des données s'attache à fournir des résultats en réduisant le nombre de dimensions, mais en ne se limitant pas à une seule. La plupart des méthodes d'analyse multidimensionnelle utilisent un modèle géométrique (une géométrie dans un espace de dimension supérieure à 3) et ses possibilités de projection sur des sous-espaces de dimension plus réduite, notamment sur des plans bien choisis. Les "écarts" entre objets y sont alors traduits par les distances habituelles.

G. Drouet d'Aubigny schématise ce traitement d'un tableau de données complexes, ou système relationnel empirique de la façon suivante :



Le plus souvent, les méthodes d'analyse multidimensionnelle s'appliquent à des tableaux de l'un des types suivants :

- Tableau protocole individus x variables numériques. Exemple :

On dispose des consommations annuelles de 8 types de denrées alimentaires pour 8 catégories socio-professionnelles (en 1972).

	PAO	PAA	VIO	VIA	POT	LEC	RAI	PLP
AGRI	167	1	163	23	41	8	6	6
SAAG	162	2	141	12	40	12	4	15
PRIN	119	6	69	56	39	5	13	41
CSUP	87	11	63	111	27	3	18	39
CMOY	103	5	68	77	32	4	11	30
EMPL	111	4	72	66	34	6	10	28
OUVR	130	3	76	52	43	7	7	16
INAC	138	7	117	74	53	8	12	20

Légende :

Variables :	Observations :
PAO Pain ordinaire	AGRI Exploitants agricoles
PAA Autre pain	SAAG Salariés agricoles
VIO Vin ordinaire	PRIN Professions indépendantes
VIA Autre vin	CSUP Cadres supérieurs
POT Pommes de terre	CMOY Cadres moyens
LEC Légumes secs	EMPL Employés
RAI Raisin de table	OUVR Ouvriers
PLP Plats préparés	INAC Inactifs

- Tableau de contingence. Exemple :

Répartition des étudiants selon la catégorie socio-professionnelle des parents et le type d'études suivi en 1975-1976 (simplifié) :

	Droit	Sciences	Médecine	IUT
Exp. agri.	80	99	65	58
Patron	168	137	208	62
Cadre sup.	470	400	876	79
Employé	145	133	135	54
Ouvrier	166	193	127	129

- Tableau protocole pour des variables nominales

	Sexe	Revenu	Preference
s1	F	M	A
s2	F	M	A
s3	F	E	B
s4	F	E	C

s5	F	E	C
s6	H	E	C
s7	H	E	B
s8	H	M	B
s9	H	M	B
s10	H	M	A

- Tableau individus x variables comportant des variables numériques et une variable dichotomique

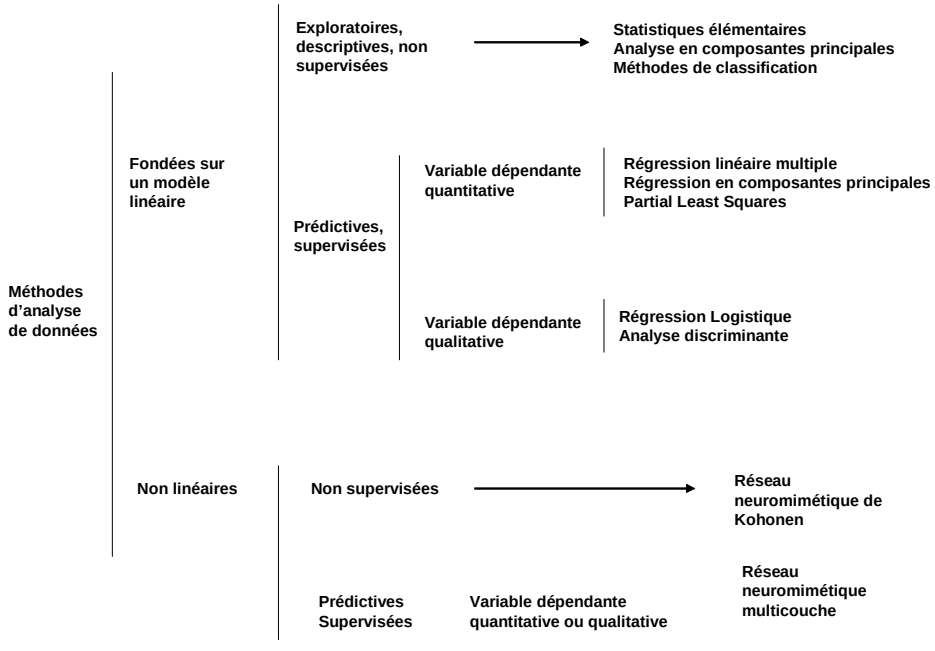
	Age	Etat-Civil	Feministe	Frequence	Agressivite	Harcelement
1	13	1	102	2	4	**0**
2	45	2	101	3	6	**0**
3	19	2	102	2	7	**1**
4	42	2	102	1	2	**1**
5	27	1	77	1	1	**0**
6	19	1	98	0	6	**1**
7	37	1	96	1	6	**0**

On cherche à analyser les résultats contenus dans ces tableaux, en explicitant plusieurs dimensions, si possible indépendantes l'une de l'autre.

1.2 Quelques méthodes utilisées

De nombreuses méthodes ont été proposées. Ces méthodes peuvent être regroupées d'une part selon les outils mathématiques utilisés (méthodes linéaires ou non linéaires), d'autre part selon la nature du résultat recherché (méthodes descriptives ou prédictives).

Méthodes descriptives : toutes les variables jouent des rôles analogues.
Méthodes prédictives : on cherche à "expliquer" ou "prévoir" une ou plusieurs variables (variables dépendantes ou VD) à l'aide des autres variables (variables indépendantes ou VI).



1.3 Concepts fondamentaux

Selon [Doise], toute distribution de réponses sur plusieurs variables peut être statistiquement décomposée en trois éléments : le niveau (la moyenne des réponses des individus), la dispersion (le degré d'éparpillement des réponses individuelles autour de la moyenne), et la corrélation (le lien entre les réponses individuelles pour deux variables). Ces composantes sont autant de points de vue sur les données.

Un tableau de données carré ou rectangulaire est appelé *matrice*. L'élément générique du tableau est désigné par une notation à double indice, par exemple x_{ij} . En général, le premier indice désigne le numéro de ligne, et le second indice le numéro de colonne. Un tableau comportant n lignes et p colonnes est dit *de dimension (n, p).*

Lorsque l'on traite un tableau Individus x Variables de dimension (n, p), les individus peuvent être représentés comme des points d'un espace à p dimensions, les variables comme des points d'un espace à n dimensions. L'ensemble des points représentant les individus est appelé *nuage des individus*.

La distance entre deux individus M_i, M_j est calculée par :

$$M_i M_j^2 = d^2(M_i, M_j) = (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{ip} - x_{jp})^2 = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$$

L'inertie du nuage de points par rapport à un point donné O de l'espace est la somme des carrés des distances des points Mi à O.

$$I = \sum_{i=1}^{n} OM_i^2$$

L'inertie du nuage de points par rapport au point moyen du nuage est encore appelée somme des carrés ou variation totale.

Le "lien" entre deux variables X_k et X_l peut être mesuré par leur coefficient de corrélation $r(X_k,X_l)$. Lorsque les variables sont centrées et réduites, ce coefficient de corrélation est, à une division par n près, le produit scalaire des vecteurs représentant ces variables. C'est aussi le cosinus de l'angle entre ces deux vecteurs. Pour des variables centrées réduites :

$$r(X_k, X_l) = \frac{1}{n} \sum_{i=1}^{n} x_{ik} x_{il} = \cos\left(\overrightarrow{X_k}, \overrightarrow{X_l}\right)$$

2 Méthodes exploratoires, descriptives

2.1 Analyse en composantes principales ou ACP

2.1.1 Introduction

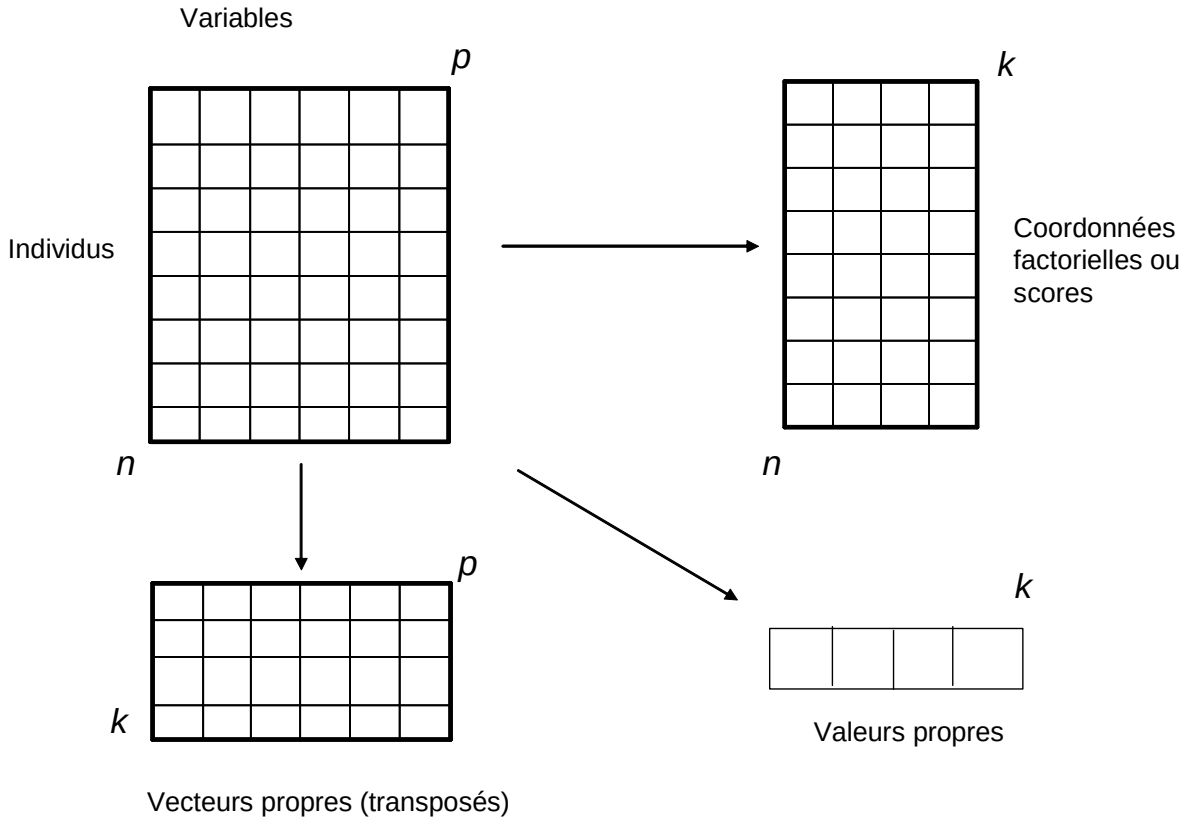
On a observé *p* variables sur *n* individus. On dit qu'il s'agit d'un protocole multivarié. Les données à traiter forment une matrice :

	X_1	X_2	...	X_p
i_1	x_{11}	x_{21}	...	x_{1p}
i_2	x_{21}	x_{22}	...	x_{2p}
...	...	...	...	...
i_n	x_{n1}	x_{n2}	...	x_{np}

On cherche à remplacer ces p variables par q nouvelles variables (composantes principales ou facteurs) résumant au mieux le protocole, avec $q \leq p$ et si possible q=2.
L'une des solutions à ce problème est l'ACP, méthode qui a l'avantage de résumer un ensemble de variables corrélées en un nombre réduit de facteurs non corrélés. Les principaux résultats d'une ACP sont donnés par :

- Les coordonnées des individus sur les composantes principales ou scores des individus ;
- Les coordonnées des variables sur les composantes principales, ou saturations des variables ; dans le cas d'une ACP normée, les saturations sont aussi les coefficients de corrélation entre les variables initiales et les composantes principales ;
- Les valeurs propres associées à chacune des composantes principales, qui représentent l'inertie du nuage prise en compte par la composante.

Principaux résultats d'une ACP



Principe de la méthode :

- Pour éliminer les effets dus aux choix d'unités des différentes variables, on fait un centrage-réduction des différentes variables.

- Les distances entre les individus sont mesurées par la distance euclidienne dans un espace de dimension p. Par exemple, pour les points représentant les individus 1 et 2 :

$$d^2(M_1, M_2) = (x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + ... + (x_{1p} - x_{2p})^2$$

- On recherche alors la direction dans laquelle le nuage de points est le plus dispersé : cette direction est le premier axe principal, et l'inertie (dispersion) le long de cet axe est la valeur propre associée à cet axe.

- On projette alors les points dans le sous-espace orthogonal au premier axe principal, et on cherche de nouveau la direction de plus grande dispersion du nuage projeté. On obtient ainsi le deuxième axe principal, et la seconde valeur propre.

- On poursuit la méthode, jusqu'à ce que l'essentiel de l'inertie du nuage de points ait été prise en compte.

2.1.2 Exemple

On reprend l'exemple donné en introduction : consommations annuelles de 8 types de denrées alimentaires pour 8 catégories socio-professionnelles (en 1972).

	PAO	PAA	VIO	VIA	POT	LEC	RAI	PLP
AGRI	167	1	163	23	41	8	6	6
SAAG	162	2	141	12	40	12	4	15
PRIN	119	6	69	56	39	5	13	41
CSUP	87	11	63	111	27	3	18	39
CMOY	103	5	68	77	32	4	11	30
EMPL	111	4	72	66	34	6	10	28
OUVR	130	3	76	52	43	7	7	16
INAC	138	7	117	74	53	8	12	20

Légende :

Variables :	Observations :
PAO Pain ordinaire	AGRI Exploitants agricoles
PAA Autre pain	SAAG Salariés agricoles
VIO Vin ordinaire	PRIN Professions indépendantes
VIA Autre vin	CSUP Cadres supérieurs
POT Pommes de terre	CMOY Cadres moyens
LEC Légumes secs	EMPL Employés
RAI Raisin de table	OUVR Ouvriers
PLP Plats préparés	INAC Inactifs

Données après centrage et réduction :

	PAO	PAA	VIO	VIA	POT	LEC	RAI	PLP
AGRI	1,43	-1,22	1,72	-1,15	0,30	0,49	-0,93	-1,50
SAAG	1,25	-0,90	1,16	-1,50	0,17	1,90	-1,38	-0,77
PRIN	-0,29	0,35	-0,70	-0,09	0,05	-0,58	0,65	1,36
CSUP	-1,44	1,92	-0,85	1,66	-1,48	-1,28	1,77	1,19
CMOY	-0,86	0,04	-0,73	0,58	-0,84	-0,93	0,20	0,46
EMPL	-0,58	-0,27	-0,62	0,23	-0,59	-0,22	-0,03	0,30
OUVR	0,10	-0,59	-0,52	-0,22	0,56	0,13	-0,70	-0,68
INAC	0,39	0,67	0,54	0,48	1,83	0,49	0,42	-0,36

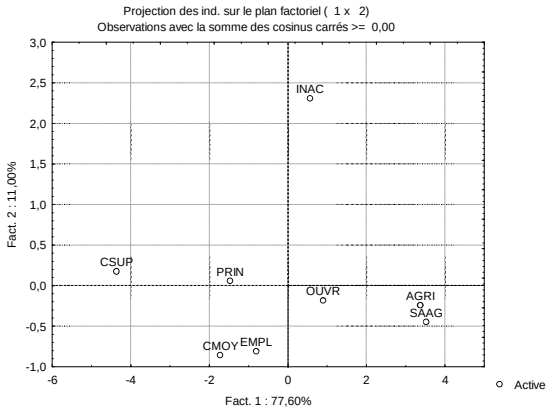
Corrélations entre variables :

	PAO	PAA	VIO	VIA	POT	LEC	RAI	PLP
PAO	1,00	-0,77	0,93	-0,91	0,66	0,89	-0,83	-0,86
PAA	-0,77	1,00	-0,60	0,90	-0,33	-0,67	0,96	0,77
VIO	0,93	-0,60	1,00	-0,75	0,52	0,79	-0,67	-0,83
VIA	-0,91	0,90	-0,75	1,00	-0,42	-0,84	0,92	0,72
POT	0,66	-0,33	0,52	-0,42	1,00	0,60	-0,41	-0,55
LEC	0,89	-0,67	0,79	-0,84	0,60	1,00	-0,82	-0,75
RAI	-0,83	0,96	-0,67	0,92	-0,41	-0,82	1,00	0,83
PLP	-0,86	0,77	-0,83	0,72	-0,55	-0,75	0,83	1,00

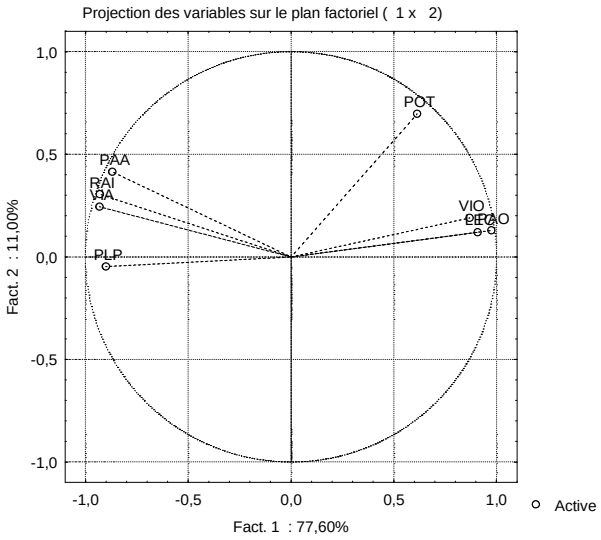
Valeurs propres de l'ACP

	Val Propre	Pourcentage	Cumul Inertie	Cumul %
1	6,2079	77,60	6,21	77,60
2	0,8797	11,00	7,09	88,60
3	0,4160	5,20	7,50	93,79
4	0,3065	3,83	7,81	97,63
5	0,1684	2,11	7,98	99,73
6	0,0181	0,23	8,00	99,96
7	0,0034	0,04	8,00	100,00

Représentation graphique des individus (plan 1-2)



Représentation graphique des variables (plan 1-2)

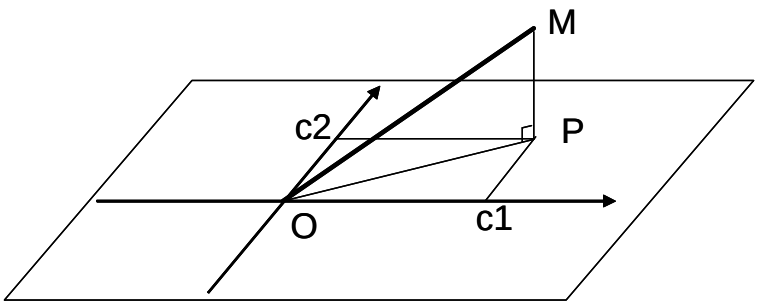


Aides à l'interprétation

Contributions ou inerties relatives des individus

	QLT	Coord. 1	Cos2	Ctr	Coord. 2	Cos2	Ctr
AGRI	0,889	1,35	0,884	22,89	-0,26	0,005	0,86
SAAG	0,913	1,41	0,898	24,97	-0,48	0,014	2,84
PRIN	0,576	-0,59	0,575	4,36	0,06	0,001	0,05
CSUP	0,943	-1,75	0,942	38,26	0,19	0,002	0,44
CMOY	0,940	-0,69	0,753	5,94	-0,91	0,187	10,43
EMPL	0,858	-0,32	0,428	1,31	-0,86	0,430	9,29
OUVR	0,376	0,36	0,361	1,63	-0,20	0,015	0,48
INAC	0,987	0,23	0,056	0,64	2,46	0,932	75,61
				100			100

Qualités de représentation



Cosinus carrés

$$Cos^2(\overrightarrow{OM}, CP_1) = \frac{Oc_1^2}{OM^2}$$

$$Cos^2(\overrightarrow{OM}, CP_2) = \frac{Oc_2^2}{OM^2}$$

$\overrightarrow{OM}$: vecteur de l'observation

$\overrightarrow{OP}$: vecteur de la projection sur le plan factoriel

$\overrightarrow{Oc_1}$: projection sur l'axe 1

$\overrightarrow{Oc_2}$: projection sur l'axe 2

Qualité

$$QUAL = Cos^2(\overrightarrow{OM}, \overrightarrow{OP}) = \frac{OP^2}{OM^2}$$

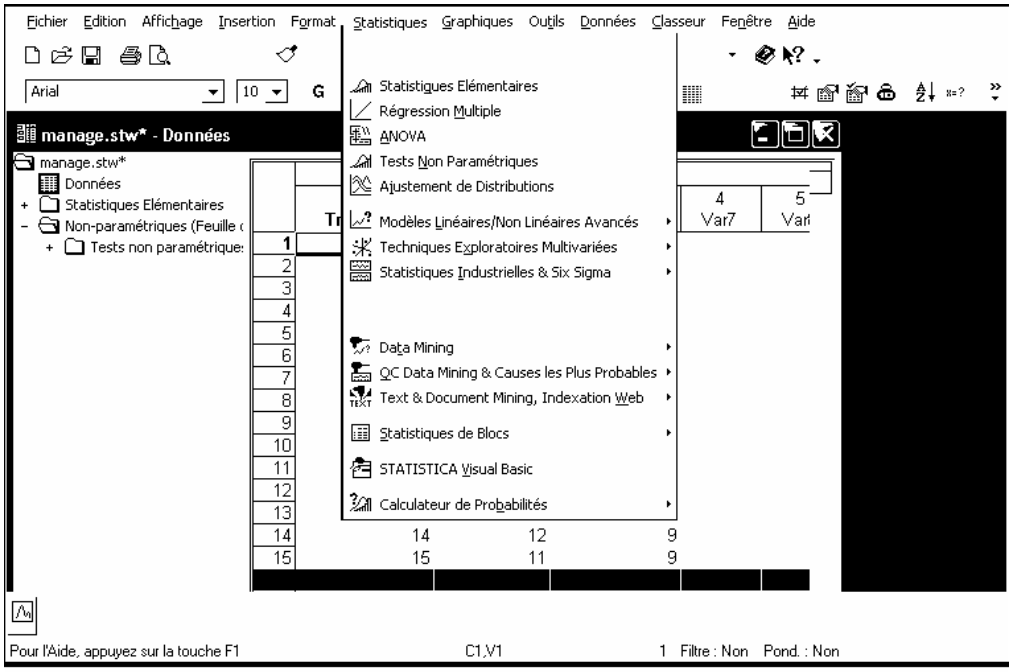
2.1.3 Analyse en composantes principales avec Statistica

2.1.3.1 Présentation de Statistica

Statistica : l'interface utilisateur

L'écran de travail

Statistica 7.1 est un logiciel dédié aux traitements statistiques. C'est également la "brique" de base des logiciels proposés par Statsoft, et ses possibilités d'interaction avec d'autres logiciels (tableurs, systèmes de gestion de bases de données, traitements de textes, ...) sont nombreuses. En revanche, l'interface utilisateur pourra sembler un peu déconcertante au premier abord.



Les objets manipulés par Statistica

La **feuille de données** est organisée en variables et observations. Les colonnes sont les variables. Chaque ligne représente un individu statistique, appelé observation.

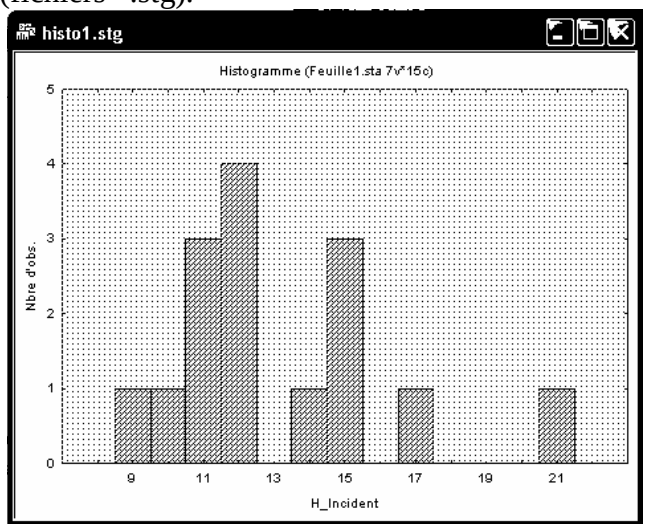
Données : Feuille1.sta (7v par 15c)

	1 Trimestre	2 H_Incident	3 D_Incident	4 Var7
1	1	11	8	
2	2	11	13	
3	3	14	12	
4	4	21	17	
5	5	12	14	
6	6	10	9	
7	7	15	10	
8	8	15	12	
9	9	17	13	
10	10	9	10	
11	11	12	8	
12	12	12	13	
13	13	15	12	
14	14	12	9	
15	15	11	9	

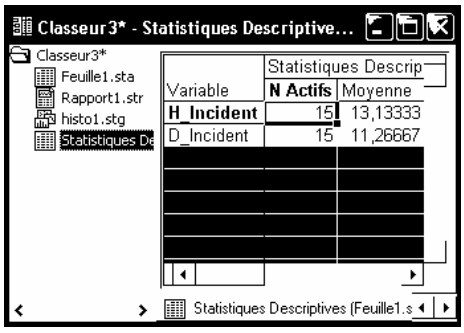
Les feuilles de données peuvent être enregistrées comme fichiers autonomes (fichiers *.sta). Elles contiennent les données d'entrée sur lesquelles s'effectuent les traitements statistiques. Les résultats de ces traitements s'affichent dans un document de sortie. Plusieurs possibilités sont offertes.

Fenêtre de rapport : C'est la méthode traditionnelle pour gérer les résultats produits par le logiciel. Un rapport se comporte plus ou moins comme un document produit par un traitement de textes. On peut insérer des commentaires, modifier la mise en forme, spécifier la mise en page, la numérotation des pages, l'en-tête et le pied de page en vue de l'impression. Les rapports peuvent être enregistrés comme fichiers autonomes (fichiers *.str).

Les résultats de sortie peuvent également être dirigés vers des fenêtres individuelles. Les résultats numériques sont alors affichés dans des fenêtres de données. Les graphiques sont affichés dans des **fenêtres de graphiques** (fichiers *.stg).

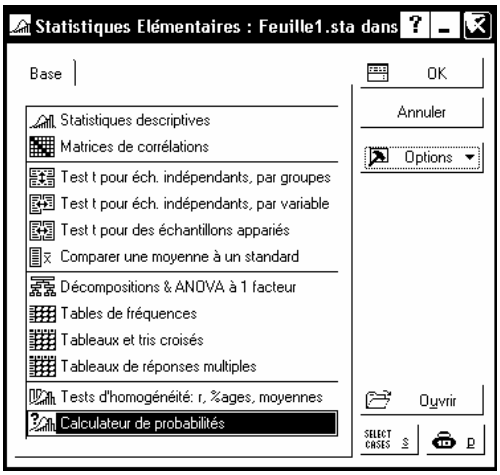


Les classeurs : les données d'entrée et de sortie peuvent également être stockées comme onglets dans un classeur. Un classeur est un "container" accueillant d'autres objets, organisés sous forme hiérarchique. Ils correspondent aux fichiers de type *.stw.



Traitements statistiques

Statistica est organisé en modules, accessibles à partir du menu Statistiques. Chaque module contient un groupe de procédures statistiques reliées entre elles. Par exemple, le module "Statistiques élémentaires" se présente comme suit :

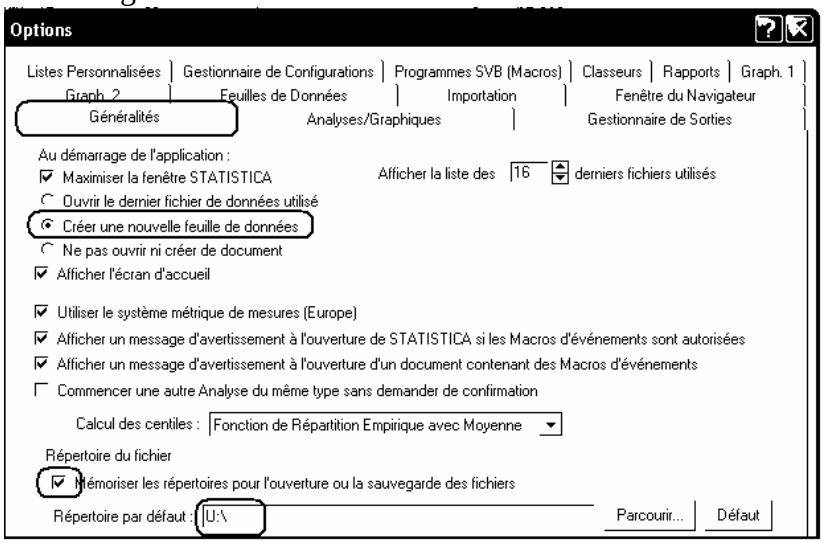


Gérer les sorties

Modifier le comportement de Statistica

Le comportement de Statistica peut être modifié en intervenant dans la fenêtre de dialogue affichée par le menu Outils - Options.
Par exemple, nous souhaitons :
- que Statistica n'ouvre plus systématiquement la dernière feuille de données utilisée lors du chargement du logiciel ;
- que Statistica nous propose par défaut le volume U: pour enregistrer nos documents, au lieu du répertoire "Mes Documents".

Exécutez le menu Outils - Options. Sous l'onglet Généralités, activez le bouton radio "Créer une nouvelle feuille de données".
Désactivez la boîte à cocher "mémoriser les répertoires pour l'ouverture ou la sauvegarde des fichiers". Complétez la zone d'édition "Répertoire par défaut" en indiquant U:\, puis réactivez la boîte à cocher (N.B. Bien que l'option soit en apparence désactivée, Statistica proposera par défaut le répertoire U:\ pour l'enregistrement de nouveaux documents.



Gérer les sorties

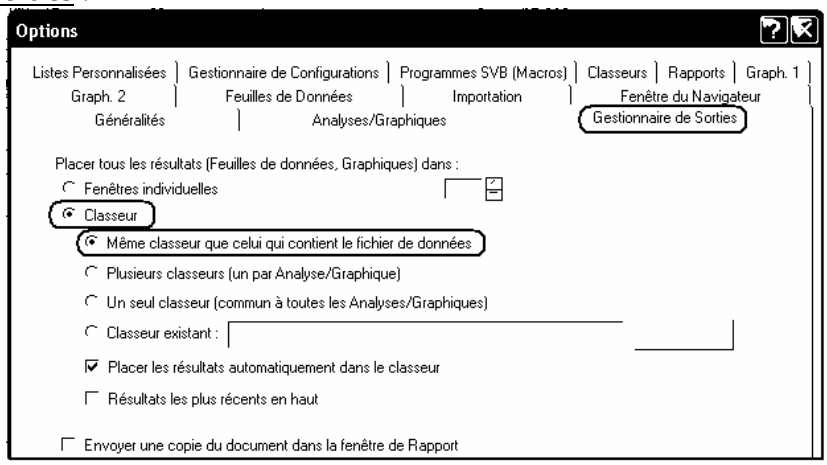
Lorsqu'on utilise Statistica sans se préoccuper des options de sortie des résultats, on se retrouve vite à la tête d'une quantité de fenêtres (classeurs, feuilles de données de résultats, fenêtres de graphiques...). Pour réaliser un travail que l'on souhaite conserver et reprendre au cours de plusieurs

séances de travail, il paraît indispensable d'organiser correctement son espace de travail et ses sauvegardes.

Enregistrer données et résultats dans un seul classeur

Cette méthode consiste à enregistrer les données, les résultats de traitements, et les commentaires éventuels comme objets d'un même classeur. Ainsi, un unique fichier du disque rassemble l'ensemble de notre travail sur un cas donné.

Ce comportement correspond aux réglages suivants dans le menu Outils - Options - Onglet Gestionnaire de Sorties :



Remarque : Le réglage ne sera actif que si la feuille de données se trouve effectivement dans un classeur. Or, ce ne sera pas le cas si la feuille de données a été ouverte à partir d'un fichier *.sta, ou importée à partir d'une feuille Excel. Dans ce cas, vous devez insérer la feuille de données dans le classeur comme il a été indiqué au paragraphe précédent.

Indiquer quelle est la feuille de données active

Lors des premières manipulations avec Statistica, nous n'avons pas eu besoin de nous préoccuper de la notion de "feuille de données active", les choix par défaut faits par Statistica nous convenant parfaitement. Cependant, cette notion permet de résoudre plusieurs problèmes :

- Ouvrir plusieurs fichiers .sta et effectuer un travail sur l'un d'eux (pas nécessairement le dernier ouvert)

- Utiliser une feuille de résultats comme feuille de données pour des traitements ultérieurs.

- Lorsque l'on travaille avec une feuille de données insérée dans un classeur, il arrive couramment que Statistica ne retrouve pas la feuille à partir de laquelle les traitements doivent être effectués. Mais on peut éviter ce comportement en spécifiant la propriété "feuille de données active" pour l'objet du classeur qui contient nos données.

Pour spécifier comme feuille de données active une feuille d'un classeur :

- Cliquez avec le bouton droit de la souris sur l'icône de la feuille de données dans le volet gauche du classeur.

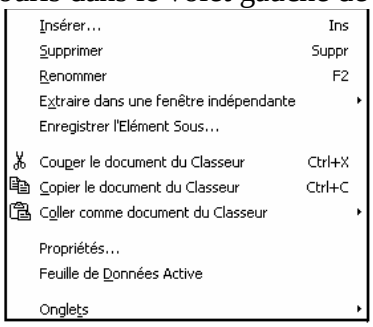
- Utilisez l'item Feuille de données active du menu local.

On peut également utiliser le menu Données - Feuille de données active.
Remarquez que le volet gauche d'un classeur indique si une feuille insérée dans le classeur est active ou non : l'icône d'un feuille active est encadrée en rouge :

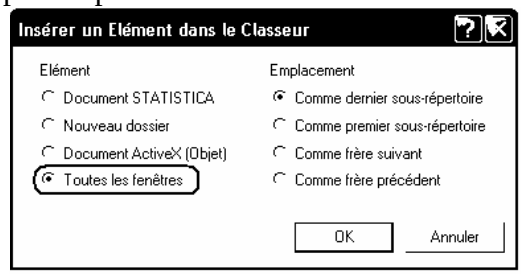
Enregistrer les données et l'ensemble des traitements réalisés dans un même classeur

Ouvrez un fichier de données (un fichier d'extension .sta) et réalisez un ou plusieurs traitements relatifs à ces données (par exemple, des statistiques descriptives et un graphique). Si vous avez gardé les options par défaut de Statistica, les résultats de tous ces traitements se trouvent dans un classeur.

Pour enregistrer données, traitements et rapport dans un seul classeur :
Affichez la fenêtre du classeur contenant les résultats.
Cliquez avec le bouton droit de la souris dans le volet gauche de la fenêtre du classeur.



Sélectionnez l'item Insérer..., puis l'option "Toutes les fenêtres" :



N'oubliez pas, ensuite, de spécifier la feuille contenant les données de base comme feuille active du classeur.

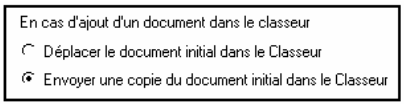
Manipuler les objets contenus dans un classeur

Copier - coller entre classeurs, entre un classeur et un objet Statistica

Pour déplacer un objet d'un classeur à un autre, il suffit de déplacer son icône depuis le volet gauche du premier classeur dans le volet gauche du second. On peut également utiliser les menus locaux Copier et Coller obtenus à l'aide d'un clic droit dans le volet gauche de chaque classeur.

Le menu local "Insérer" du volet gauche d'un classeur permet également d'insérer dans ce classeur un document contenu dans une fenêtre indépendante. Il suffit de choisir les options : Document Statistica - Créer à partir d'une fenêtre.

L'opération faite par Statistica est soit une copie (l'original de l'objet est conservé) soit un déplacement (l'original de l'objet n'est pas conservé) selon le paramétrage choisi dans le menu Outils - Options - Onglet Classeurs - Item "En cas d'ajout d'un document dans le classeur".



Supprimer un objet d'un classeur

Il est également possible de supprimer un objet d'un classeur, à l'aide d'un clic droit et de l'item de menu Supprimer. Cela permet notamment de ne garder, pour un traitement donné, que le résultat le plus abouti. Attention cependant : lorsque l'on supprime un objet qui n'est pas une feuille de la hiérarchie, on supprime en même temps tous les objets qui en dépendent.

2.1.3.2 Présentation de l'exemple

Source de l'exemple : Claude FLAMENT, Laurent MILLAND, Un effet Guttman en ACP, Mathématiques & Sciences humaines (43e année, n° 171, 2005, p. 25-49)

Cet exemple a trait à la représentation sociale de l'homosexualité. Le questionnaire, composé d'une liste de 31 traits plus ou moins sexués, a été administré à 70 hommes homosexuels et à 70 hommes hétérosexuels [Rallier, Ricou, 2000]. Tous les sujets devaient, dans un premier temps, se décrire à partir de cette liste de traits, en se positionnant à chaque fois sur une échelle allant de 1 (= négatif) à 7 (= positif). Après avoir réalisé cette auto-description, les sujets devaient répondre à ce même questionnaire « comme le feraient les X en général », la cible « X » pouvant être : les hommes, les femmes, ou les homosexuels. Nous disposons ainsi de 8 profils moyens, qui se définissent à partir de la combinaison entre les caractéristiques des répondants et les consignes données pour remplir les questionnaires. Nous travaillons ici sur un extrait des données complètes (15 traits), extrait qui respecte scrupuleusement le type de résultat obtenu sur l'ensemble des 31 traits de l'étude.
Pour faciliter le repérage des consignes, nous avons fait le choix de coder les 8 profils en repérant en premier les répondants, puis le type de consigne parmi les 4 possibles :

Ho : Soi = sujets Homosexuels répondant à la consigne d'auto-description Soi ;
Hé : Soi = sujets Hétérosexuels répondant à la consigne d'autodescription Soi ;
Ho : H = sujets Homosexuels répondant comme le feraient les Hommes ;
Hé : H = sujets Hétérosexuels répondant comme le feraient les Hommes;
Ho : F = sujets Homosexuels répondant comme le feraient les Femmes ;
Hé : F = sujets Hétérosexuels répondant comme le feraient les Femmes ;
Ho : Ho = sujets Homosexuels répondant comme le feraient les Homosexuels ;
Hé : Ho = sujets Hétérosexuels répondant comme le feraient les Homosexuels.

Nous partons ici d'un tableau de données comprenant, pour chacune des 8 conditions expérimentales, les moyennes de chaque trait calculées sur les 70 réponses obtenues dans chacune des conditions expérimentales. On retrouve, dans le tableau ci-dessous, le rang (solidarisation des variables) de chacun des 15 traits dans les 8 profils

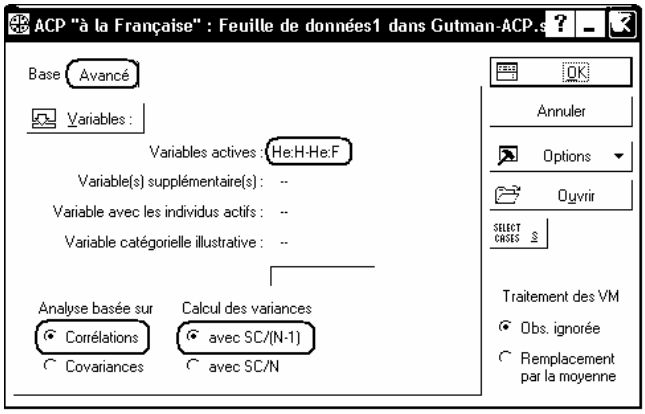
	He:H	Ho:H	He:Soi	Ho:Soi	Ho:Ho	Ho:F	He:Ho	He:F
Est meneur	5	6	12	13	15	13	14	13
Aime competition	3	3	13	14	11	14	13	14
FEMININ	15	15	15	15	13	2	4	1
A confiance en soi	4	8	6	11	14	12	12	12
Devoue	11	12	10	7	10	11	8	7
MASCULIN	1	1	1	12	12	15	15	15
Bienveillant	10	10	9	9	7	9	7	6
Attentif aux besoins des autres	12	13	11	4	9	8	5	5
Energique	8	4	5	8	6	10	11	11
Ambitieux	6	7	3	10	8	7	10	10

Sensible	14	14	14	2	1	1	1	2
Agreable	9	9	7	5	3	6	6	3
Affectueux	13	11	8	1	4	5	2	4
A du caractere	2	5	4	6	5	4	9	8
Defend ses opinions	7	2	2	3	2	3	3	9

Remarque. A l'examen du tableau précédent, on constate que les rangs ont été déterminés à l'inverse de ce qui est généralement fait en statistiques : les rangs élevés correspondent aux traits les moins typiques du stéréotype considéré, tandis que les rangs faibles correspondent aux traits les plus typiques. Cette remarque est importante pour l'interprétation des résultats de l'ACP.

Ouvrez le classeur Statistica Rep-Soc-Homo.stw.

Pour effectuer l'ACP, nous utilisons le menu Statistiques - Techniques exploratoires multivariées - ACP "à la française".



La fenêtre de dialogue permet de spécifier les variables qui participeront à l'analyse. Elle permet également d'indiquer les différentes options choisies pour le traitement.
Utilisez l'onglet "Avancé" de cette fenêtre.

- Comment seront traitées les valeurs manquantes ? Nous voyons que Statistica propose soit de neutraliser la ligne correspondante, soit de remplacer la valeur manquante par la moyenne observée sur la variable.

- L'analyse sera-t-elle basée sur les covariances ou sur les corrélations ?

- Utilise-t-on les variances et covariances non corrigées (SC/N) ou les variances et covariances corrigées (SC/(N-1)). Dans le cas d'une ACP normée, les deux méthodes fournissent des résultats presque identiques : seuls les scores des individus sont légèrement modifiés. En fait, l'ACP est une méthode descriptive et non une méthode inférentielle. Elle est effectuée dans un but exploratoire : on étudie les données pour elles-mêmes, et non en vue d'une généralisation à une population. C'est pourquoi l'utilisation des variances non corrigées est généralement justifiée.

Nous ferons ici une analyse basée sur les corrélations, en utilisant les variances et covariances corrigées (SC/(N-1)), de manière à retrouver les résultats publiés. Cliquez ensuite sur le bouton OK.

N.B. Ne fermez pas l'analyse en cours pendant la suite des manipulations. Ainsi, vous n'aurez pas à indiquer de nouveau les options ci-dessus, vos résultats seront cohérents entre eux et se rassembleront dans un même classeur.

2.1.3.3 Statistiques descriptives - Matrice des corrélations

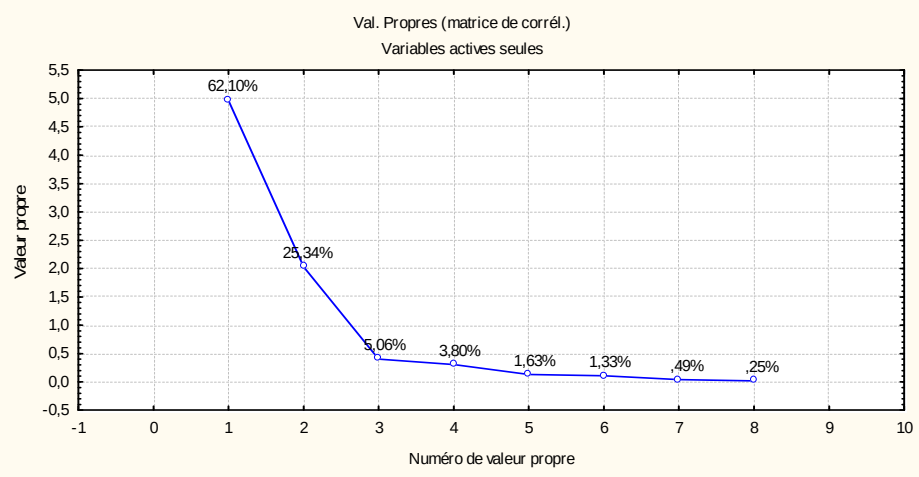
Ces résultats peuvent être obtenus à l'aide de l'onglet "Descriptives".

Variable	Corrélations (Repr-Soc-Homo dans Rep-Soc-Homo.stw)							
	He:H	Ho:H	He:Soi	Ho:Soi	Ho:Ho	Ho:F	He:Ho	He:F
He:H	1,0000	0,8679	0,5857	-0,4071	-0,3143	-0,6036	-0,8179	-0,8714
Ho:H	0,8679	1,0000	0,6786	-0,2393	-0,0607	-0,4821	-0,6429	-0,8357
He:Soi	0,5857	0,6786	1,0000	0,1679	0,2179	-0,1321	-0,2821	-0,4464
Ho:Soi	-0,4071	-0,2393	0,1679	1,0000	0,8429	0,5607	0,7143	0,5036
Ho:Ho	-0,3143	-0,0607	0,2179	0,8429	1,0000	0,6750	0,6821	0,4929
Ho:F	-0,6036	-0,4821	-0,1321	0,5607	0,6750	1,0000	0,8714	0,8071
He:Ho	-0,8179	-0,6429	-0,2821	0,7143	0,6821	0,8714	1,0000	0,8857
He:F	-0,8714	-0,8357	-0,4464	0,5036	0,4929	0,8071	0,8857	1,0000

2.1.3.4 Choix des valeurs propres

Affichez d'abord le tableau des valeurs propres et le diagramme correspondant.
Pour cela, cliquez sur les boutons "Valeurs propres" et "Tracé des valeurs propres" de l'onglet "Base".

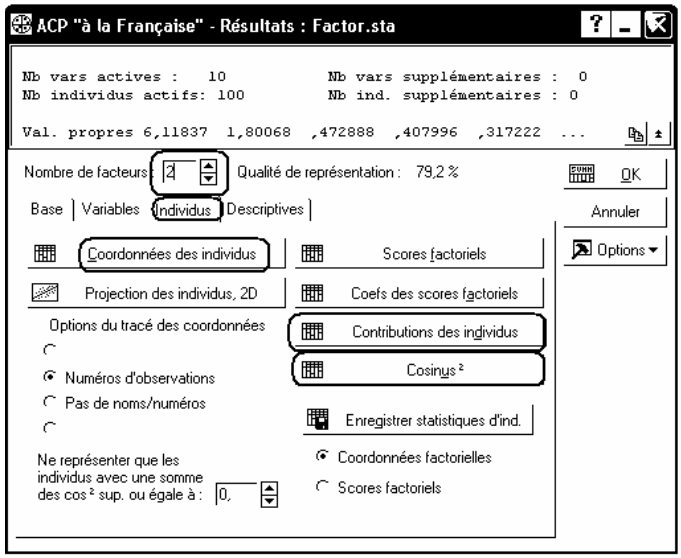
Valeur numéro	Val. Propres (matrice de corrél.) & stat. associées Variables actives seules			
	Val. propr	% Total variance	Cumul Val. propr	Cumul %
1	4,9682	62,1026	4,9682	62,10
2	2,0268	25,3355	6,9950	87,44
3	0,4045	5,0562	7,3995	92,49
4	0,3038	3,7979	7,7034	96,29
5	0,1308	1,6346	7,8341	97,93
6	0,1064	1,3301	7,9405	99,26
7	0,0391	0,4892	7,9797	99,75
8	0,0203	0,2541	8,0000	100,00



Dans notre cas, on peut choisir de retenir 2 composantes principales. Dans les manipulations qui suivent, on indiquera donc 2 dans la zone d'édition "nombre de facteurs".
Pour les résultats relatifs aux individus et aux variables, on utilisera de préférence les onglets correspondants.

2.1.3.5 Résultats relatifs aux individus

On pourra obtenir successivement les scores des individus, leurs contributions à la formation des composantes principales et leurs qualités de représentation en utilisant les boutons "Coordonnées des individus", "Contributions des individus", "Cosinus²".



Individus	Coordonnées factorielles des ind	
	Fact. 1	Fact. 2
Est meneur	-2,5273	1,6292
Aime competition	-2,7956	1,0317
FEMININ	2,2340	2,9293
A confiance en soi	-2,1315	0,6348
Devoue	0,4389	1,0207
MASCULIN	-3,9200	-0,7793
Bienveillant	0,5732	0,4503
Attentif aux besoins	1,6498	0,7581
Energique	-1,0549	-0,8752
Ambitieux	-0,8932	-0,6719
Sensible	3,9415	-0,0333
Agreable	1,4885	-0,8338
Affectueux	2,6229	-0,7360
A du caractere	-0,3598	-1,7357
Defend ses opinions	0,7335	-2,7890

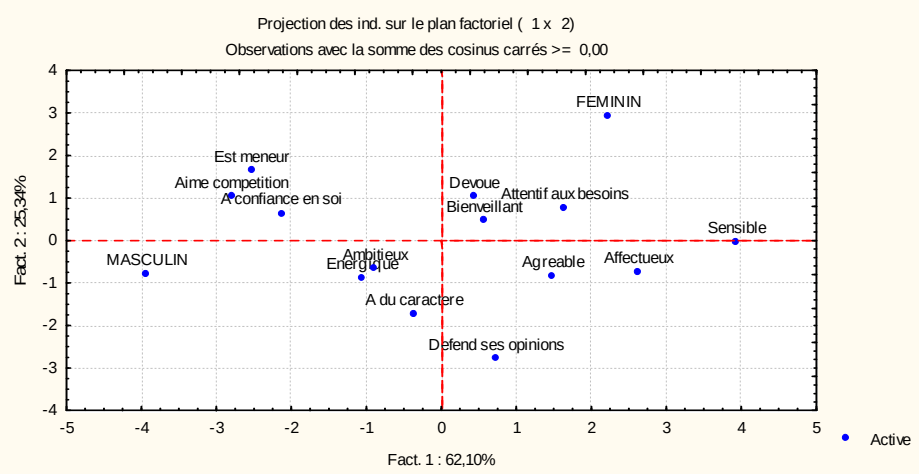
Individus	Contributions des ind	
	Fact. 1	Fact. 2
Est meneur	9,18	9,35
Aime competition	11,24	3,75
FEMININ	7,18	30,24
A confiance en soi	6,53	1,42
Devoue	0,28	3,67
MASCULIN	22,09	2,14
Bienveillant	0,47	0,71
Attentif aux besoins	3,91	2,03
Energique	1,60	2,70
Ambitieux	1,15	1,59
Sensible	22,34	0,00
Agreable	3,19	2,45
Affectueux	9,89	1,91
A du caractere	0,19	10,62
Defend ses opinions	0,77	27,41

Individus	Cosinus carrés,		
	Fact. 1	Fact. 2	Fact. 1 & 2 =v1+v2
Est meneur	0,6759	0,2809	0,9568
Aime competition	0,7203	0,0981	0,8184
FEMININ	0,3100	0,5330	0,8429
A confiance en soi	0,8041	0,0713	0,8755
Devoue	0,0875	0,4736	0,5611
MASCULIN	0,9427	0,0373	0,9800
Bienveillant	0,3866	0,2385	0,6251
Attentif aux besoins	0,6404	0,1352	0,7757
Energique	0,4364	0,3004	0,7368
Ambitieux	0,3711	0,2099	0,5810
Sensible	0,9502	0,0001	0,9503
Agreable	0,6330	0,1986	0,8317
Affectueux	0,8600	0,0677	0,9277
A du caractere	0,0284	0,6621	0,6906
Defend ses opinions	0,0582	0,8409	0,8991

Remarquez que les résultats ainsi obtenus sont présentés dans des feuilles de résultats sur lesquelles il est possible d'effectuer les mêmes transformations (tris, ajout ou suppression de colonne, etc) que sur les feuilles contenant les données de base. Ainsi, une colonne supplémentaire a été ajouté au tableau des cosinus-carrés pour indiquer la qualité de représentation des individus dans le premier plan factoriel.

On peut ensuite obtenir les projections du nuage des individus selon les premiers axes factoriels à l'aide du bouton "Projection de individus, 2D". Lorsque les individus ne sont pas anonymes (ce qui est le cas ici), il est utile d'étiqueter chaque point. Plusieurs méthodes sont possibles :

- Utiliser les identifiants d'individus figurant dans la première colonne du tableau de données
- Utiliser les numéros des observations
- Utiliser les étiquettes indiquées dans la variable "illustrative" : ces étiquettes peuvent être des identifiants des individus, mais peuvent également représenter un groupe d'appartenance, etc.



Dans certains cas, il pourra être utile de modifier les échelles sur les axes de manière à obtenir une représentation en axes orthonormés. L'importance de la part d'inertie expliquée par le premier axe principal apparaît ainsi plus clairement.

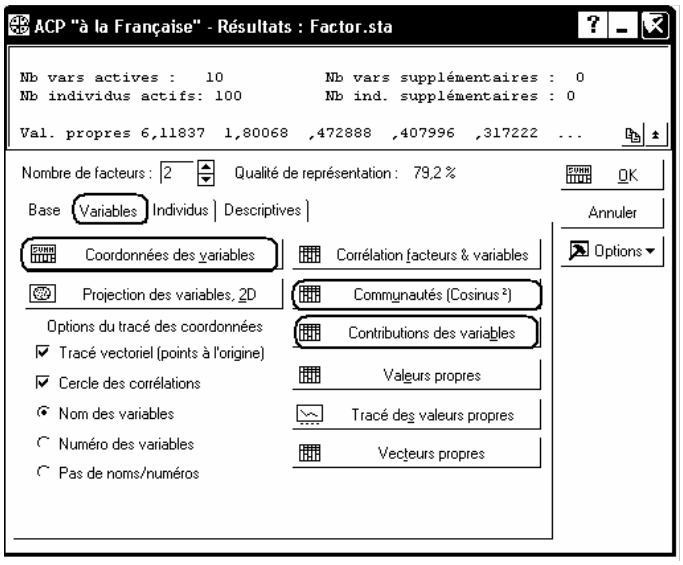
2.1.3.6 Résultats relatifs aux variables

Activons ensuite l'onglet "Variables".

On obtient les saturations des variables en cliquant sur le bouton "Coordonnées des variables" ou le bouton "Corrélation facteurs et variables" : dans le cas d'une ACP normée, ces deux traitements fournissent le même résultat.

On obtient leurs contributions à la formation des composantes principales en utilisant le bouton "Contributions des variables".

Les qualités de représentation sont calculées, de façon cumulative (qualité de la projection selon F1, puis selon le plan (F1,F2), puis selon l'espace (F1,F2,F3) en utilisant le bouton "Communautés (Cosinus²)".



Saturations des variables

Variable	Coord. factorielles des var	
	Fact. 1	Fact. 2
He:H	0,8863	0,3388
Ho:H	0,7743	0,5518
He:Soi	0,4047	0,8013
Ho:Soi	-0,6701	0,6053
Ho:Ho	-0,6317	0,7093
Ho:F	-0,8511	0,2387
He:Ho	-0,9663	0,1361
He:F	-0,9555	-0,1428

Contributions des variables

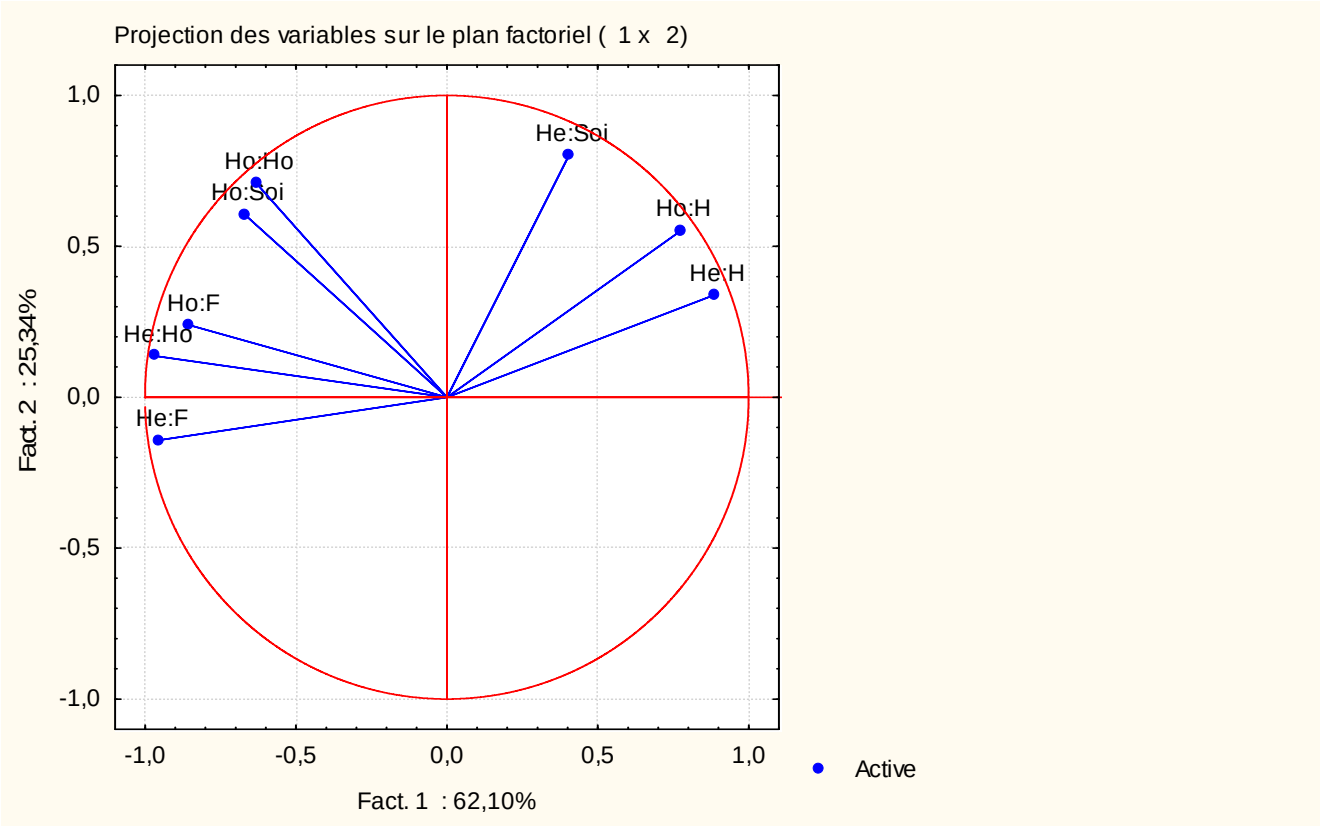
Variable	Contributions des var	
	Fact. 1	Fact. 2
He:H	0,1581	0,0566
Ho:H	0,1207	0,1502
He:Soi	0,0330	0,3168
Ho:Soi	0,0904	0,1808
Ho:Ho	0,0803	0,2482
Ho:F	0,1458	0,0281
He:Ho	0,1879	0,0091
He:F	0,1838	0,0101

Qualités des représentations des variables

Variable	Communautés,	
	Avec 1 facteur	Avec 2 facteurs
He:H	0,7856	0,9004
Ho:H	0,5996	0,9041
He:Soi	0,1638	0,8060
Ho:Soi	0,4491	0,8154
Ho:Ho	0,3991	0,9022
Ho:F	0,7243	0,7813
He:Ho	0,9337	0,9522
He:F	0,9131	0,9334

Représentation des variables

Le bouton "Projection des variables, 2D" permet d'obtenir les diagrammes représentant les projections des variables selon les plans définis par deux axes principaux.



On peut remarquer que toutes les variables se projettent dans un même demi-plan du premier plan factoriel. Autrement dit, une rotation des axes factoriels convenablement choisie permettrait de ramener toutes les variables dans le demi-plan correspondant aux valeurs positives du premier facteur.

2.1.3.7 Coefficients des variables

Les coefficients des variables (c'est-à-dire la matrice permettant de passer des variables centrées réduites aux composantes principales et vice-versa) sont obtenus à l'aide du bouton "Vecteurs propres" de l'onglet "Variables".

	Vecteurs propres de la matrice de corrélation (Repr-Soc-Homo dans Rep-Soc-Homo.stw) Variables actives seules							
Variable	Fact. 1	Fact. 2	Fact. 3	Fact. 4	Fact. 5	Fact. 6	Fact. 7	Fact. 8
He:H	0,398	0,238	0,172	-0,198	0,731	-0,039	-0,325	-0,272
Ho:H	0,347	0,388	0,135	-0,416	-0,440	-0,137	-0,329	0,466
He:Soi	0,182	0,563	0,217	0,750	-0,149	0,097	0,022	-0,092
Ho:Soi	-0,301	0,425	-0,617	0,082	0,318	-0,345	0,036	0,345
Ho:Ho	-0,283	0,498	-0,111	-0,411	-0,090	0,589	0,198	-0,309
Ho:F	-0,382	0,168	0,687	-0,142	0,196	-0,294	0,408	0,206
He:Ho	-0,434	0,096	0,065	-0,030	-0,261	-0,450	-0,496	-0,531
He:F	-0,429	-0,100	0,189	0,168	0,184	0,463	-0,577	0,401

2.1.4 Interprétation des résultats de l'ACP

2.1.4.1 Examen des valeurs propres. Choix du nombre d'axes

On examine les résultats relatifs aux valeurs propres.
Plusieurs critères peuvent nous guider :

- "méthode du coude" on examine la courbe de décroissance des valeurs propres pour déterminer les points où la pente diminue de façon brutale ; seuls les axes qui précèdent ce changement de pente seront retenus.
- si l'analyse porte sur p variables et n > p individus, la variation totale est répartie sur p axes. On peut alors choisir de conserver les axes dont la contribution relative est supérieure à $\frac{100\%}{p}$. Dans le cas d'une ACP normée, cela revient à conserver les axes correspondant aux valeurs propres supérieures à 1.

Sur le cas étudié, les différentes méthodes conduisent à ne garder que les deux premiers axes.

2.1.4.2 Interpréter les résultats relatifs aux individus

Très souvent, les individus pris en compte pour une ACP sont en nombre très élevé et sont considérés comme anonymes. Les éléments qui suivent concernent évidemment les cas où ils ne le sont pas.

Contributions des individus à la formation d'un axe

On relève, pour chaque axe, quels sont les individus qui ont la plus forte contribution à la formation de l'axe. Par exemple, on retient (pour l'analyse) les individus dont la contribution relative est supérieure à $\frac{100\%}{n}$. On note également si cette contribution intervient dans la partie positive ou dans la partie négative de l'axe.

On peut ainsi caractériser l'axe en termes d'opposition entre individus. Il peut également être intéressant d'étudier comment l'axe classe les individus.

Si un individu a une contribution très forte à la formation d'un axe, on peut choisir de recommencer l'analyse en retirant cet individu, puis de l'introduire en tant qu'individu supplémentaire.

Ainsi, pour le premier axe, on relève les traits qui ont contribué pour plus de 6,67% à sa formation et le signe de la coordonnée de chacun de ces traits. On obtient :

-	+
MASCULIN (22,09) Aime competition (11,24) Est meneur (9,18)	Sensible (22,34) Affectueux (9,89) FEMININ (7,18)

On voit que cet axe oppose le trait "masculin", et des traits qui sont souvent associés à ce sexe (meneur, aime compétition, a confiance en soi), sur la partie négative de l'axe, à des traits tels que "sensible", "affectueux", "attentif", et "féminin" sur la partie positive.

Pour le deuxième axe, la même démarche conduit au tableau suivant :

-	+
Defend ses opinions (27,41) A du caractere (10,62)	FEMININ (30,24) Est meneur (9,35)

Cet axe oppose deux traits pratiquement indépendants du premier axe (partie négative de l'axe) au trait "féminin" (partie positive de l'axe).

Projections des individus dans un plan factoriel

Même s'il s'agit du plan (F1, F2), les proximités entre individus doivent être interprétées avec prudence : deux points proches l'un de l'autre sur le graphique peuvent correspondrent à des individus éloignés l'un de l'autre. Pour interpréter ces proximités, il est nécessaire de tenir compte des qualités de représentation des individus.
Se méfier également des individus proches de l'origine : mal représentés, ou proches de la moyenne, ils ont, de toutes façons, peu contribué à la formation des axes étudiés.

2.1.4.3 Interpréter les résultats relatifs aux variables

Contributions des variables

L'examen du tableau des contributions des variables peut permettre d'identifier des variables qui ont un rôle dominant dans la formation d'un axe factoriel. Comme précédemment, on retient (par exemple) les variables dont la contribution relative est supérieure à $\frac{100\%}{p}$. On note également si cette contribution intervient dans la partie positive ou dans la partie négative de l'axe.

Ainsi, pour le premier axe, en fixant la "limite" à 12,5%, on obtient :

-	+
He:Ho (0,1879) He:F (0,1838) Ho:F (0,1458)	He:H (0,1581)

Ainsi, cet axe oppose les profils féminins et homosexuels vus par les hétérosexuels (partie négative de l'axe) au profil masculin vu par les hétérosexuels (partie positive de l'axe).

Remarque importante. L'analyse des individus (traits) avait associé la partie négative du premier axe aux traits masculins. L'analyse des variables semble a priori conduire à un résultat opposé. Mais la contradiction n'est qu'apparente : ici, le protocole des rangs accorde le rang le moins élevé au trait le plus caractéristique du profil. La variable He:H par exemple, est fortement corrélée positivement avec le facteur 1. Le trait "masculin" par exemple obtient un score faible aussi bien sur cette variable (rang 1) que sur le premier facteur (-3,92, minimum des coordonnées de points).

Pour le second axe factoriel, on obtient :

-	+
	He:Soi (0,3168) Ho:Ho (0,2482) Ho:Soi (0,1808) Ho:H (0,1502)

On remarque que les quatre variables retenues sont celles qui ne figuraient pas dans le tableau précédent. Ces quatre variables sont corrélées positivement avec le deuxième axe.

Analyse des projections des variables sur les plans factoriels

Les diagrammes représentant les projections des variables sur les axes factoriels nous fournissent plusieurs types d'informations :

- La longueur du vecteur représentant la variable est liée à la qualité de la représentation de la variable par sa projection dans ce plan factoriel

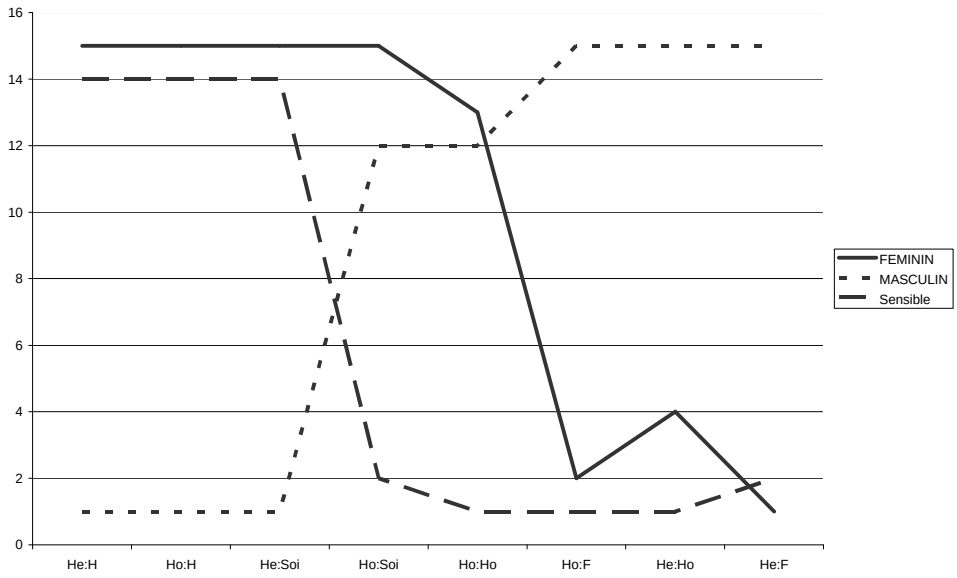
- Pour les variables bien représentées, l'angle entre deux variables est lié au coefficient de corrélation entre ces variables (si la représentation est exacte, le coefficient de corrélation est le cosinus de cet angle). Ceci permet de dégager des "groupes de variables" de significations voisines, des groupes de variables qui "s'opposent", des groupes de variables relativement indépendantes entre eux.

- De même, pour les variables bien représentées, l'angle que fait la projection de la variable avec un axe factoriel est lié au coefficient de corrélation de cette variable et de l'axe factoriel.

Ainsi, dans notre exemple, toutes les variables sont bien représentées dans le premier plan factoriel. Des variables telles que Ho:Soi et Ho:Ho par exemple, sont fortement corrélées positivement entre elles, alors que Ho:Ho et Ho:H sont pratiquement non corrélées. Les variables He:Ho et He:F par exemple, sont fortement anti-corrélées (corrélées négativement) avec le premier axe.

Synthèse des résultats obtenus

On voit que les sujets hétérosexuels ont tendance à estimer que les homosexuels se décrivent comme "féminin" plutôt que "masculin". L'étude des résultats de l'ACP pourrait nous conduire à associer la description que les homosexuels se font d'eux-mêmes à "féminin". Mais, cette conclusion est contredite par les données : les homosexuels ne se voient jamais comme "féminin", mais font appel à des items identifiés ici comme des caractéristiques féminines (sensible, affectueux, etc). Le graphique suivant, dans lequel on a représenté les scores des traits "féminin", "masculin" et "sensible" en fonction des profils convenablement ordonnés, le met en évidence :



Sur ce graphique, les profils sont ordonnés en fonction de leur ordre d'apparition sur le cercle des corrélations (graphique du paragraphe 2.1.3.6). Cet ordre peut également être schématisé de la manière suivante :

Répondants		
He	Ho	Cible
H	H	Masculine
Soi	Soi	Homosexuelle
	Ho	
	Fe	Féminine
Ho		
Fe		

2.1.5 ACP avec individus et variables supplémentaires

Lorsqu'on réalise une ACP, il est possible de déclarer certains individus "inactifs" et/ou certaines variables "supplémentaires". Les données correspondantes n'interviennent plus dans le calcul de détermination des composantes principales. En revanche, on leur applique les mêmes transformations qu'aux autres données afin de les ré-introduire dans les tableaux et graphiques de résultats.

Cette méthode peut notamment être utilisée lorsque des individus ou des variables ont une influence trop importante sur les résultats d'une ACP. On recommence alors les calculs en les déclarant comme individus inactifs ou variables supplémentaires. Elle peut également être utilisée pour introduire des variables plus synthétiques, et des moyennes par groupe d'individus, comme c'est le cas dans l'exemple ci-dessous.

Avec Statistica, il est simple de déclarer une variable comme variable supplémentaire : le premier dialogue de l'ACP prévoit une zone d'édition pour cela. Pour déclarer des individus comme "inactifs", il est nécessaire de construire une variable supplémentaire, qui ne contiendra que deux modalités, et d'utiliser les zones d'édition "Variable avec individus actifs" et "Code des individus actifs".

Ouvrez le fichier Proteines-2008.stw.

Source : Exemple fourni avec le logiciel Statistica.
Cet exemple particulier est présenté par Greenacre (1984) dans le cadre d'une comparaison entre l'analyse en composantes principales (voir l'Analyse Factorielle) et l'analyse des correspondances.

Les données du fichier d'exemple Protein.sta représentent des estimations de la consommation protéique issue de 9 sources différentes, par habitant dans 25 pays (les données ont initialement été reportées par Weber, 1973, dans un polycopié publié à l'Université de Kiel, Institut für Agrarpolitik und Marktlehre, intitulé "Agrarpolitik im Spannungsfeld der Internationalen Ernährungspolitik").

Au fichier de données initial ont été ajoutées les 5 variables suivantes :

- Consommation en protéines animales (somme des variables v1 à v5)
- Consommation en protéines végétales (somme des variables v6 à v9)
- Un code du nom du pays sur 2 ou 3 lettres
- Le groupe auquel appartient le pays (4 groupes ont été définis : NW (Europe du Nord et de l'Ouest), NE (Europe de l'Est, pays du Nord), SW (Europe de l'Ouest, pays du Sud) et SE (Europe de l'Est, pays du Sud)).
- Une variable codant pour les individus actifs (1) et inactifs (0).

Quatre individus ont été ajoutés, correspondant aux moyennes observées dans les 4 groupes de pays définis précédemment

Extrait des données :

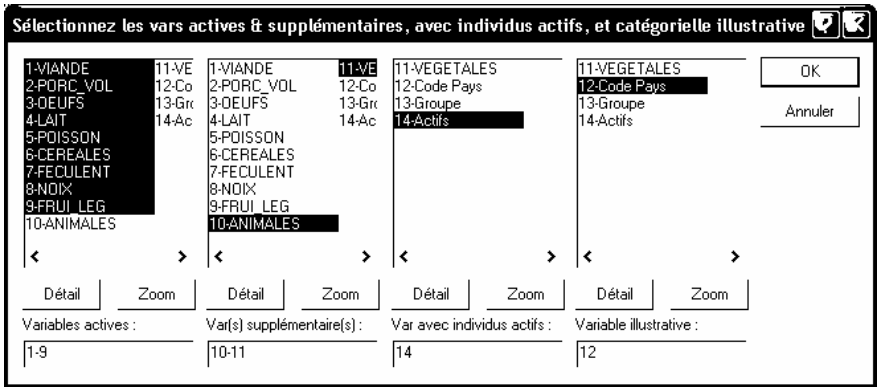
	Evaluation des consommations de protéïnes, en grammes/habitant/jour								
	1 VIANDE	2 ORC_VC	3 OEUFS	4 LAIT	5 'OISSON	6 EREALE	7 ECULEN	8 NOIX	9 RUI_LE
Belgique/Lux.	13,5	9,3	4,1	17,5	4,5	26,6	5,7	2,1	4,0
Bulgarie	7,8	6,0	1,6	8,3	1,2	56,7	1,1	3,7	4,2
Tchécoslovaquie	9,7	11,4	2,8	12,5	2,0	34,3	5,0	1,1	4,0
Danemark	10,6	10,8	3,7	25,0	9,9	21,9	4,8	0,7	2,4
R.D.A.	8,4	11,6	3,7	11,1	5,4	24,6	6,5	0,8	3,6
Finlande	9,5	4,9	2,7	33,7	5,8	26,3	5,1	1,0	1,4

Toutes les variables s'expriment ici avec la même unité (g.hab/jour). Pour réaliser une ACP, deux possibilités s'offrent à nous :

- Faire une ACP sur les valeurs non réduites. Ainsi, une information telle que "l'apport protéique des viandes, porc et volailles est, dans tous les cas, supérieur à celui des fruits et légumes" est prise en compte dans l'étude.

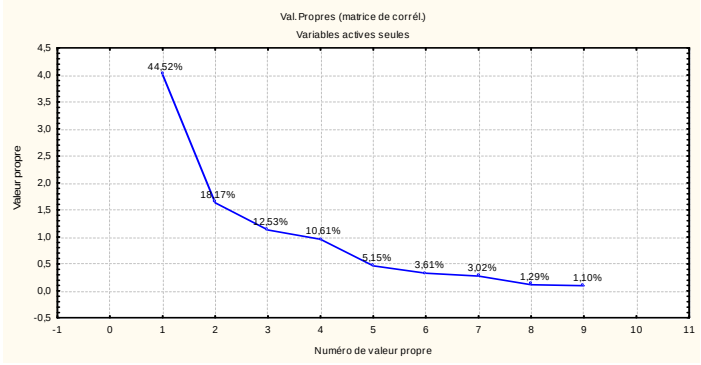
- Faire une ACP sur les valeurs réduites (ACP calculée à partir du tableau des corrélations). Dans ce cas, l'étude "gomme" les inégalités des apports protéiques des différentes sources.

Réalisons une ACP sur les corrélations en spécifiant individus actifs et variables supplémentaires comme suit :



Affichez les tableaux des covariances et des corrélations. On voit déjà apparaître une opposition entre protéines d'origine animale et protéines d'origine végétale.

Combien de valeurs propres faut-il ici retenir ? Seules 3 valeurs propres sont supérieures à 1, mais la règle du coude conduit à retenir soit 2, soit 4 axes factoriels. En fait, il faut conserver 4 axes pour mettre en évidence certaines spécificités des pays d'Europe Centrale (axe 3) ou de la France (axe 4).



Exercice : Calculez les résultats de l'ACP pour les 4 premiers axes à l'aide de Statistica, puis interprétez les résultats.

2.1.6 ACP avec rotation

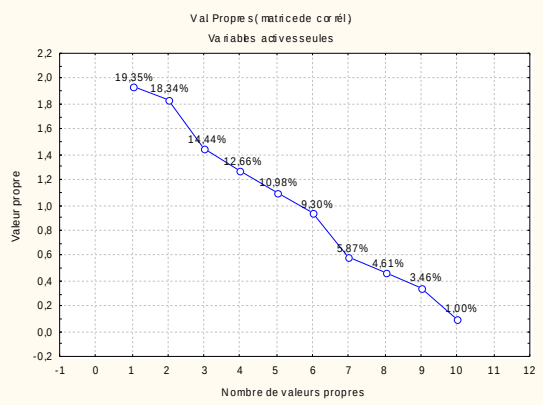
Par construction, les composantes principales sont des abstractions mathématiques et ne possèdent pas nécessairement de signification intuitive. Après avoir réalisé l'ACP, il peut parfois être intéressant de définir d'autres variables en effectuant une combinaison linéaire des composantes principales retenues, à l'aide d'une "rotation". L'objectif est généralement d'augmenter les saturations, c'est-à-dire les corrélations entre ces nouveaux "facteurs" et certaines variables de départ. Les nouveaux "facteurs" ainsi obtenus perdent les propriétés des facteurs principaux. Par exemple, le premier d'entre eux ne correspond plus à la direction de plus grande dispersion du nuage des individus. En revanche, la part de variance expliquée par les facteurs retenus reste identique. Il existe différents critères (varimax, quartimax, equamax, etc) permettant d'obtenir une rotation conduisant à des saturations proches de 1 ou -1, ou au contraire proches de 0.

Cette possibilité n'est pas disponible dans la méthode "ACP à la française" de Statistica. En revanche, on peut l'utiliser en utilisant le module "Analyse factorielle" convenablement paramétré.

2.1.7 Une ACP fournit-elle toujours des informations interprétables ?

Tout tableau de données peut être soumis à une ACP, et les méthodes d'analyse qui ont été développées permettent de "trouver des résultats". Mais ces résultats correspondent-ils à une réalité plus ou moins cachée ou ne constituent-ils qu'un artefact de la méthode ?
Pour étudier cet aspect, réalisons une ACP sur des données ... où il n'y a rien à dire (il s'agit de données produites à l'aide d'un générateur de nombres aléatoires).

Ouvrez le fichier aleatoire-20sujets.stw et réalisez une ACP normée sur ces données. La représentation graphique des valeurs propres nous indique déjà l'absence d'intérêt des données traitées :



2.2 Combiner description et prédiction : Analyse factorielle

2.2.1 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.

Comme l'ACP, l'analyse factorielle s'applique à des protocoles multivariés, c'est-à-dire des tableaux décrivant *n* sujets à l'aide de *p* variables numériques. Quelques remarques :

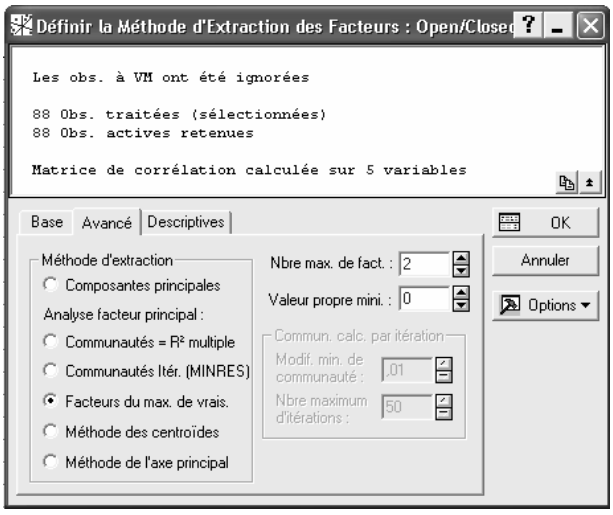
- l'intérêt porte ici sur les variables et non sur les individus statistiques ; il s'agit donc plus d'une méthode d'analyse multivariée que d'une méthode d'analyse multidimensionnelle.
- de nombreuses variantes existent : l'analyse factorielle est parfois désignée par le terme "analyse en facteurs communs 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.

2.2.2 Exemple introductif

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

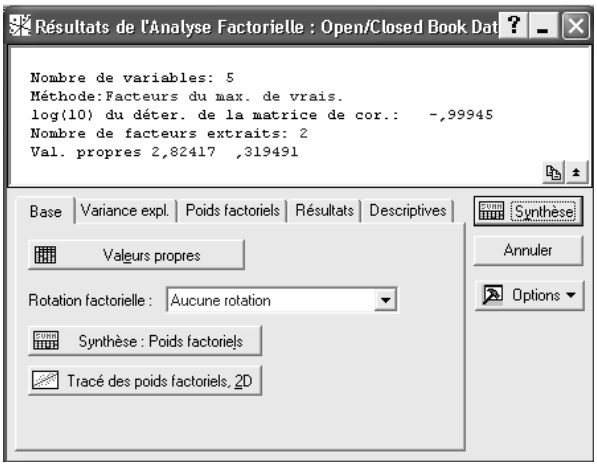
On dispose des notes obtenues par 88 sujets dans 5 matières : Mechanics(C), Vectors(C), Algebra(O), Analysis(O), Statistics(O). Pour deux matières, les étudiants n'avaient pas accès à leurs documents (closed book - C), pour les trois autres, les documents pouvaient être consultés (open book - O).

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)

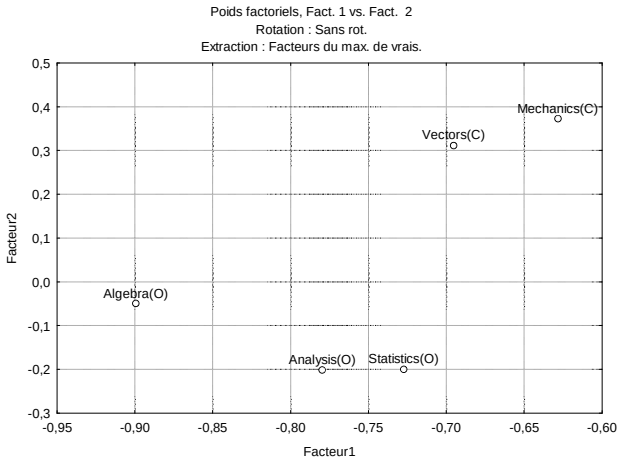
Mechanics(C)	0,47	-0,00	0,00	-0,01	0,01
Vectors(C)	-0,00	0,42	-0,00	0,01	-0,01
Algebra(O)	0,00	-0,00	0,19	-0,00	0,00
Analysis(O)	-0,01	0,01	-0,00	0,35	-0,00
Statistics(O)	0,01	-0,01	0,00	-0,00	0,43

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 ?

2.2.3 Justification conceptuelle de l'analyse factorielle exploratoire

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 :

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

ou, de façon moins formelle :

$$\text{Variable observée} = \sum \text{coeff.} \times \text{variable latente} + \text{erreur spécifique}$$

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.

2.2.4 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.

2.2.4.1 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.

2.2.4.2 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.

2.2.4.3 L'analyse factorielle du maximum de vraisemblance

Notion de 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.

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.

Méthode 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 X_i sur la population parente de l'échantillon observé. Certains auteurs expriment cette condition en termes d'asymétrie et d'aplatissement des distributions observées.

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.

2.2.5 Résultats obtenus - Scores des individus

2.2.5.1 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 :

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

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%.

2.2.5.2 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.

2.2.6 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.

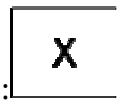
D'autres rotations ont également été proposées. Les rotations dites orthogonales produisent des facteurs non corrélés entre eux, tandis que les transformations par rotation oblique produisent de nouveaux facteurs qui peuvent être corrélés.

2.2.7 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 : 

- 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 :

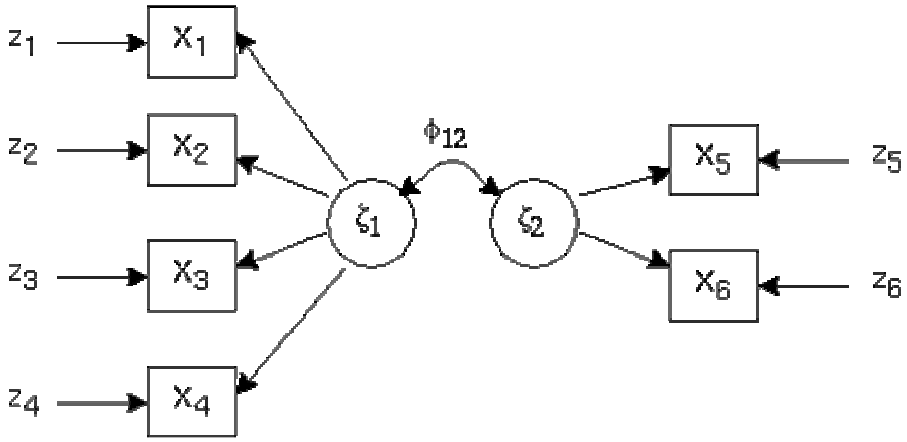
	Self	Parent	Teacher	Friend	Educ Asp	Col Plan
Self	1,00	0,73	0,70	0,58	0,46	0,56
Parent	0,73	1,00	0,68	0,61	0,43	0,52
Teacher	0,70	0,68	1,00	0,57	0,40	0,48
Friend	0,58	0,61	0,57	1,00	0,37	0,41
Educ Asp	0,46	0,43	0,40	0,37	1,00	0,72
Col Plan	0,56	0,52	0,48	0,41	0,72	1,00

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) :



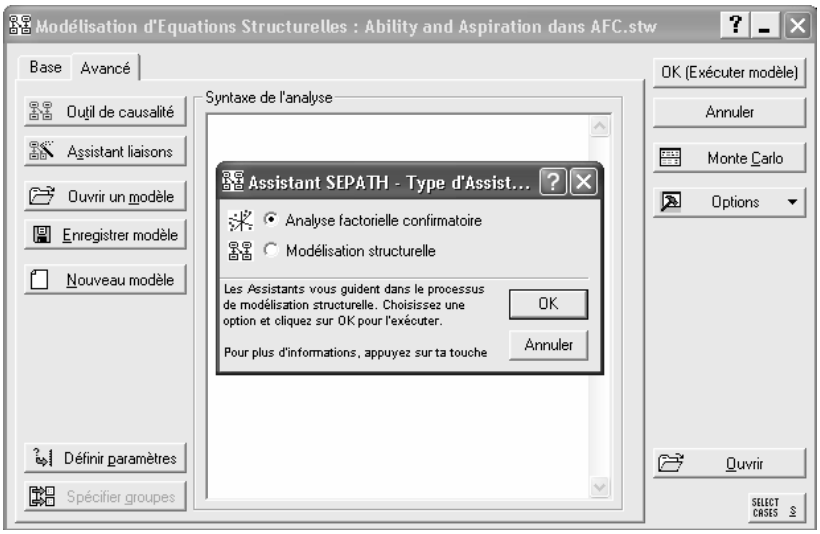
Traitement avec Statistica.

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					

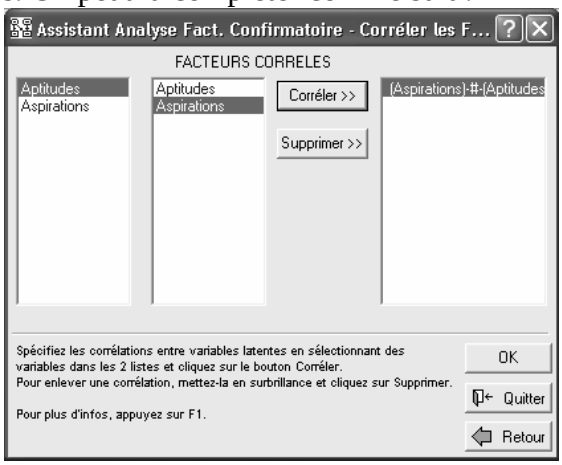
On choisit ensuite le menu Statistiques - Modèles linéaires / non linéaires avancés - Modélisation d'équations structurelles.

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 :



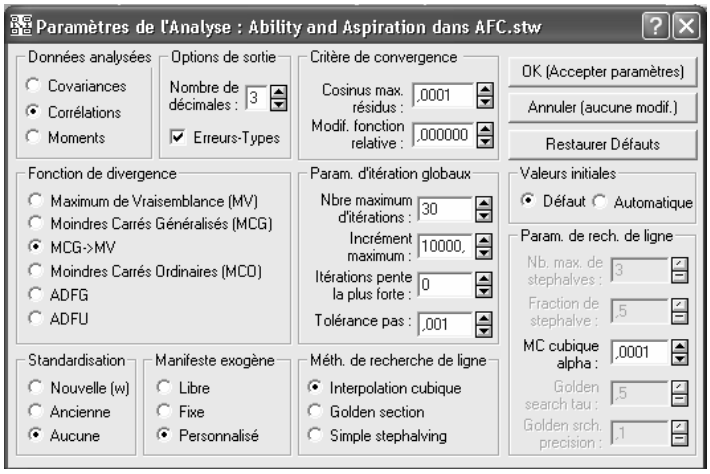
Lorsque la fenêtre suivante s'affiche, cliquer sur OK :

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		0,745	0,255
Parent	0,849		0,721	0,279
Teacher	0,805		0,648	0,352
Friend	0,695		0,483	0,517
Educ Asp		0,775	0,601	0,399
Col Plan		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)

2.2.8 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

Pages mises en ligne par Peter Tryfos
http://www.yorku.ca/ptryfos/methods.htm

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