L’analyse factorielle exploratoire.

Objectifs de la section

Au terme de cette section vous serez en mesure :

- De saisir la différence entre l’analyse factorielle et l’analyse en composantes principales
- D’identifier les questions de recherche qui appellent l’utilisation d’une analyse factorielle exploratoire
- De vous assurer que vos données rencontrent ses conditions d’application
- De prendre les décisions appropriées quant au nombre de facteurs à extraire
- D’effectuer une analyse factorielle exploratoire sur SPSS
- D’interpréter adéquatement les résultats obtenus

Introduction

L’analyse factorielle exploratoire (AFE) est une technique qui permet de mettre en évidence la **structure latente** d’une masse de données. On entend par structure latente, la présence d’un certain nombre de facteurs (ou de dimensions sous-jacentes) permettant d’expliquer pourquoi certaines de nos variables sont intercorrélées, alors que d’autres variables ne le sont pas. Les variables latentes (communément appelées facteurs) ne sont pas directement observables, mais elles sont inférées en tenant compte du patron de corrélation observé entre nos variables. L’AFE ressemble beaucoup à l’analyse en composantes principales (ACP) et on peut faire de nombreux parallèles entre les deux techniques: dans les deux cas, les analyses portent sur des matrices de corrélation pouvant comporter un nombre élevé de variables; de plus, dans les deux cas, il s’agit d’expliquer une portion importante de variance présente dans les données en ayant recours à un nombre limité de dimensions.

La différence fondamentale entre les deux techniques concerne justement la portion de variance que l’on cherche à expliquer. Alors que l’ACP s’intéresse à la **variance totale** présente dans la matrice de corrélation, l’analyse factorielle exploratoire quant à elle se concentre strictement sur la portion de **variance commune** partagée par certaines variables. Nous aurons l’occasion de revenir sur cet aspect, puisqu’il constitue l’élément clé pour comprendre la distinction entre les deux techniques. Si vous avez déjà consulté le document portant sur l’analyse en composantes principales, vous développerez rapidement un sentiment de familiarité avec l’analyse factorielle exploratoire, puisque nous utiliserons la même procédure FACTOR de SPSS et que nous l’appliquerons au même exemple fictif portant sur les motivations des consommateurs de bière, exemple que nous avons tiré de Wuensch (2001).

Quelques exemples réels d'application

Depuis quelques années plusieurs auteurs se sont penchés sur le rôle que peut jouer l'estime de soi des parents sur la qualité des interactions parents-enfants. L'échelle du sentiment parental de compétence [Parenting Sense of Competence (PSOC)] développée par Johnston et Mash (1989) est l'une des mesures les plus populaires de l'estime des soi des parents. Dès l'origine le questionnaire de 17 items a été construit de manière a évaluer deux aspects distincts de l'estime de soi parentale: la satisfaction d'être parent et le sentiment d'efficacité dans le rôle parental. L'existence de ces deux dimensions a d'ailleurs été mise en évidence par Johnston et Mash (1989) dans leur analyse en composantes principales. Plus récemment, Ohan, Leung et Johnston (2000) se sont donné comme objectif de confirmer la structure factorielle du PSOC en la vérifiant auprès d'un échantillon de parents d'enfants n'éprouvant pas de problèmes particuliers.

Le PSOC fut donc administré à 110 pères et 110 mères ayant au moins un enfant entre 5 et 12 ans. Les auteurs optèrent pour une **analyse factorielle exploratoire** parce que différentes contraintes les empêchaient d'utiliser une analyse de type confirmatoire (nombre de variables trop élevé sur chaque facteur, taille de l'échantillon plutôt petite et pourcentage de variance expliquée plutôt faible). Des analyses factorielles de type « Principal axis factoring » ont été calculées séparément avec les donnés des mères et des pères. Dans les deux cas le critère de Kaiser indiquait l'existence de trois facteurs pour lesquels les valeurs propres (eigenvalues) étaient > 1.0. Cependant, les graphiques d'accumulation de variance de Cattell (« scree test ») indiquaient que des solutions en deux dimensions devaient plutôt être considérées. Ce critère fut donc privilégié, d'autant plus que les troisièmes facteurs auraient été sous définis puisqu'ils auraient expliqué moins de 10% de variance, avec pas plus de deux variables atteignant une pondération de .40. Des solutions à deux facteurs ont donc été calculées. La première dimension expliquait 31.80% de la variance chez les mères et 26.79% chez les pères. Les pourcentages correspondants pour le deuxième facteur étaient 11.44% et 14.09%. Comme Ohan et al. (2000) s'appuyaient sur une théorie voulant que les deux dimensions constituant l'estime de soi parentale soient corrélées, ils exécutèrent une **rotation oblique**. Dans le cas des mères, la corrélation obtenue entre les deux facteurs a été de .48, alors qu'elle a été de .27 dans l'échantillon des pères. Finalement, l'examen des pondérations des variables sur les deux facteurs a ensuite permis de constater que la structure factorielle du PSOC correspondait à ce qui avait déjà été observé par Johnston et Mash (1989) dans leur analyse en composantes principales. Ces résultats appuient donc une interprétation bidimensionnelle de l'estime de soi parentale, impliquant d'une part la satisfaction d'être parent et d'autre part le sentiment d'efficacité dans le rôle parental.

On voit, dans ce premier exemple, que l'analyse factorielle s'insère dans une perspective théorique plus importante que ne le fait l'analyse en composantes principales. En effet, si Ohan et al. (2000) ont choisi l'AFE plutôt que l'ACP c'est que leur travail s'appuyait sur une base théorique solide et bien délimitée qui leur permettait de prédire que deux dimensions corrélées (satisfaction et efficacité) seraient en mesure d'expliquer des regroupements d'items au PSOC. Leur objectif de recherche n'était pas simplement de réduire la complexité des 17 items du PSOC, mais surtout d'en vérifier la **structure latente**.

Voici un deuxième exemple quelque peu différent d'utilisation de l'analyse factorielle. Cooney, Kadden et Litt (1990) se sont intéressés à la mesure de la sociopathie chez les alcooliques. En effet, plusieurs auteurs ont suggéré que la sociopathie ou les troubles de la personnalité antisociale constituaient un facteur de risque important dans le développement de l'alcoolisme. Cependant, plusieurs mesures différentes de la sociopathie ont été utilisées chez les alcooliques, certaines étant des mesures catégorielles, d'autres des mesures continues. Devant ce constat, Cooney et al. (1990) ont décidé de confronter quatre mesures différentes de la sociopathie couramment utilisées et de **vérifier si elles mesuraient bien un construit unitaire**. Ils ont évalué 118 alcooliques hospitalisés à qui ils ont administré les quatre mesures suivantes : 1) l'échelle Pd de psychopathie du MMPI-168, 2) la liste des 22 symptômes de psychopathie de Hare, 3) l'échelle de socialisation de l'Inventaire de personnalité de Californie, et 4) l'interview diagnostic de la personnalité antisociale du NIMH. Les quatre mesures furent soumises à une analyse factorielle (sans rotation) qui démontra clairement la présence d'un seul facteur général sur lequel les quatre mesures obtenaient des pondérations significatives, bien que ce soit l'échelle de l'inventaire de Californie qui obtenait la pondération la plus importante (.90). Ce facteur général de psychopathie expliquait à lui seul 57.3% de la variance des quatre instruments.

On constate par ce deuxième exemple que l'objectif principal de l'analyse factorielle **n'est pas** de procéder à une réduction des données comme c'était le cas avec l'ACP. En effet, avec quatre mesures, on ne peut certainement pas prétendre être confronté à un nombre particulièrement élevé de mesures. Par contre, il n'y a pas de doute que les quatre instruments utilisés possèdent des formats d'administration vraiment différents et qu'ils ont été développés selon des perspectives théoriques spécifiques; étant donné cette grande diversité conceptuelle, on pourrait fort bien être en présence de dimensions multiples et différentes de la sociopathie. La question de savoir si ces quatre instruments mesurent un construit hypothétique commun peut être avantageusement examinée à l'aide de la technique de l'analyse factorielle.

Malgré leur diversité évidente, les deux exemples précédents collent parfaitement à la définition de l'AFE telle qu'on la retrouve dans le « *Dictionary of statistics and methodology* » (Vogt, 1993) :

L'analyse factorielle exploratoire: Analyse factorielle conduite dans le but de découvrir quelles sont les variables latentes (facteurs) sous-jacentes à un ensemble de variables ou de mesures. Cette technique est généralement mise en contraste avec l'analyse factorielle confirmatoire qui permet de tester formellement des théories et de mettre à l'épreuve des hypothèses concernant les facteurs anticipés. (Vogt, 1993, page 87).

Description sommaire de la technique

Le postulat fondamental à la base de l'analyse factorielle est le suivant: si des variables sont corrélées les unes avec les autres dans nos données, c'est parce qu'elles subissent l'influence de certains facteurs qui leur sont communs. L'analyse a pour objectif de mettre en évidence ces facteurs communs, qui ne sont malheureusement pas directement observables, mais qui pourront être estimés. Le modèle mathématique à la base de l'analyse factorielle s'exprime dans un ensemble d'équations s'apparentant à des équations de régression multiple et ayant la forme suivante :

$$
\begin{aligned}
var_1 &= \hat{a}_1 F_1 + \hat{a}_2 F_2 + \hat{a}_3 F_3 \ldots + \hat{a}_k F_k + U_1 \\
var_2 &= \hat{a}_1 F_1 + \hat{a}_2 F_2 + \hat{a}_3 F_3 \ldots + \hat{a}_k F_k + U_2 \\
var_3 &= \hat{a}_1 F_1 + \hat{a}_2 F_2 + \hat{a}_3 F_3 \ldots + \hat{a}_k F_k + U_3 \\
&\vdots
\end{aligned}
$$

Quelques observations importantes doivent être faites à l'égard de cet ensemble d'équations. Tout d'abord vous remarquerez que les équations sont identiques dans leur forme pour chacune des variables soumises à l'analyse (var_1, var_2, var_3...). Cependant, il faut bien comprendre que les coefficients de régression $\hat{a}_1$, $\hat{a}_2$ et $\hat{a}_3$ correspondent à des paramètres qui seront déterminés lors de l'analyse et qui prendront dans les faits des valeurs différentes d'une variable à l'autre. Chaque variable sera donc prédite à l'aide des mêmes facteurs communs mais recevant des pondérations différentes. Une deuxième observation importante concerne les facteurs F_1, F_2, F_3 et suivants qui ne sont pas des variables prévisionnelles individuelles comme on est habitué de les voir en analyse de régression. Ce sont plutôt des **facteurs latents**, non directement observables, mais définis eux-mêmes par différents regroupements de variables. L'une des tâches du chercheur sera justement d'identifier ou de nommer ces facteurs en examinant les variables qui les composent. Enfin, vous remarquerez que chaque équation comporte également un terme U correspondant à une source de **variation unique** pour chaque variable. La source de variance U est très importante en analyse factorielle puisqu'elle permet justement de distinguer cette technique de l'analyse en composantes principales. En effet, dans le modèle de l'analyse factorielle, chaque variable possède une portion de variance unique qui est indépendante de toutes les autres sources de variance et qui n'a rien à voir avec la variance commune; cette variance unique est elle-même constituée de deux portions, à savoir 1) la portion de variance réelle de la variable, non assujettie à l'influence des facteurs communs et donc appelée **variance spécifique**, et 2) la variance d'erreur de la variable. L'analyse factorielle met de côté la variance unique des variables et **se concentre sur la portion de variance commune** pour en extraire les dimensions sous-jacentes. En contraste, l'analyse en composantes principales ne fait pas ces nuances entre variance commune et variance unique, ne fait aucune place à la variance d'erreur, assume que la variance totale est importante et l'utilise donc au complet pour en extraire des composantes. Pour ces raisons, il faut reconnaître que l'AFE et l'ACP sont fort différentes tant au niveau conceptuel qu'au niveau des modèles mathématiques qui les sous-tendent.

Une conséquence importante découle de ce qui vient d'être dit : en analyse factorielle, le chercheur doit préciser dès le début pour chacune des variables analysées, la portion de variance commune sur laquelle il souhaite se pencher. Or, paradoxalement, cette information ne sera disponible qu'au terme de l'analyse factorielle. La solution à ce dilemme consiste à fournir au logiciel d'analyse (SPSS ou autres) un estimé de la variance commune de chaque variable. L'approche la plus courante consiste à utiliser le $\mathbf{R^2}$ comme **estimé de la variance commune** (rappelez-vous que la corrélation multiple R correspond à la corrélation entre une variable et un ensemble d'autres variables). Avant l'analyse proprement dite, les estimés de départ des variances communes sont donc placés dans la diagonale principale de la matrice de corrélation à être analysée, remplaçant ainsi les valeurs 1.0 que l'on retrouve habituellement le long de cette diagonale. On parle désormais d'une **matrice de corrélation réduite**, puisque la variance totale de chacune des variables n'est plus 1.0, mais bien une valeur réduite (R^2) correspondant à la variance commune estimée.

L'analyse factorielle de la matrice réduite peut alors démarrer et les facteurs sont extraits selon le nombre requis. Cette première extraction mène à une solution dans laquelle les variances communes de chaque variable sont calculées. Or, il est bien possible que ces variances communes calculées diffèrent des variances communes estimées au départ; si c'est le cas, les anciennes valeurs (R^2) de la diagonale principale de la matrice analysée sont remplacées par les nouvelles valeurs calculées et le processus d'extraction des facteurs recommence… Ce processus itératif continue autant de fois que nécessaire, c'est-à-dire jusqu'à ce qu'il n'entraîne plus de changements importants dans le calcul des variances communes. On dit alors que la solution a convergé, c'est-à-dire qu'elle a atteint une phase stable et qu'elle peut maintenant être interprétée.

Particularités des matrices réduites

En analyse factorielle, la matrice de corrélation réduite comporte une diagonale principale correspondant à des estimés de la variance commune de chaque variable. Les coefficients R multiple au carré (R^2) sont généralement utilisés à cet effet. Vous les retracerez dans la diagonale de la matrice que l'on s'apprête à analyser et qui est présentée au tableau 2.1. Cette matrice de corrélation est identique à celle que nous avions utilisée comme exemple dans le document portant sur l'ACP (tableau 1.2, page 6), sauf pour le remplacement des valeurs de la diagonale. Rappelez-vous que la somme des valeurs de cette diagonale correspond à la variance qui sera analysée. Alors que l'ACP s'intéressait à 7 unités de variance (7 variables X 1.0), l'AFE se limitera aux 5.69 unités de variance commune (.74 + .91 + .86 + .50 + .93 + .86 + .89), soit 81.2% de la variance totale dans ce cas-ci.

Tableau 2.1 Matrice d'intercorrélation réduite entre sept variables mesurant les motivations à acheter une marque de bière particulière.

	PRIX	QUANTITE	ALCOOL	PRESTIGE	COULEUR	AROME	GOÛT
PRIX	**.74**	.83	.77	-.41	.02	-.05	-.06
QUANTITE	.83	**.91**	.90	-.39	.18	.10	.03
ALCOOL	.77	.90	**.86**	-.46	.07	.04	.01
PRESTIGE	-.41	-.39	-.46	**.50**	-.37	-.44	-.44
COULEUR	.02	.18	.07	-.37	**.93**	.91	.91
AROME	-.05	.10	.04	-.44	.91	**.86**	.87
GOÛT	-.06	.03	.01	-.44	.91	.87	**.89**

Il va sans dire que les matrices de corrélation réduites soumises à une AFE doivent posséder les mêmes qualités globales que celles que nous avions notées dans le cas de l'ACP. Ainsi, une variable qui ne serait corrélée à aucune autre devrait certainement être retirée de l'analyse, puisque nous nous intéressons à la variance commune partagée entre les variables. À cet égard, le calcul des mesures d'adéquacité de l'échantillonnage de Kaiser-Meyer-Olkin (Measure of Sampling Adequacy, MSA ») est tout indiqué.

Rappelons les valeurs repères suggérées par Kaiser : inacceptable en dessous de 0.5, médiocre entre 0.5 et 0.6, moyen entre 0.6 et 0.7, bien entre 0.7 et 0.8, très bien entre 0.8 et 0.9 et excellent au delà de 0.9. Le tableau 2.2 présente les mesures d'adéquacité de l'échantillonnage obtenues; elles sont identiques à celles rapportées au tableau 1.3 du document traitant de l'ACP et nous avons déjà mentionné que c'est probablement la nature fictive des données utilisées qui explique leur qualité discutable.

Tableau 2.2 Mesures Kaiser-Meyer-Olkin d'adéquacité de l'échantilonnage calculées pour la matrice d'intercorrélation réduite du tableau 2.1.

Variable	Indice d'adéquacité de Kaiser-Meyer-Olkin
Prix	.78512
Quantité	.55894
Alcool	.64103
Prestige	.74289
Couleur	.59006
Arôme	.79444
Goût	.67012
Matrice globale	.66646

Aussi, les situations de multicolinéarité modérée ne sont pas problématiques en AFE, mais il faut à tout prix éviter la situation de **singularité** où une variable serait entièrement prédite par une combinaison de plusieurs variables. Rappelons que cette condition peut être détectée en calculant le « déterminant » de la matrice de corrélation $|R|$ qui est égal à 0.00000 dans le cas d'une matrice singulière. Dans le cas présent, le déterminant est de 0.0004927 indiquant qu'il ne s'agit pas d'une matrice singulière. À l'opposé, il est également contre-indiqué de procéder à une AFE si nous sommes en présence d'une **matrice d'identité**, c'est-à-dire une matrice où tous les coefficients de corrélation auraient la valeur 0.0. Le **test de sphéricité de Bartlett** permettra de mettre à l'épreuve l'hypothèse nulle d'une matrice d'identité dans la population. Évidemment, le résultat au test de Bartlett est identique à ce qui avait été noté dans le document sur l'ACP : la valeur obtenue (729.82, p <.0004927) permet de rejeter l'hypothèse nulle qui voudrait que nous soyons en présence d'une matrice d'identité.

Détermination du nombre de facteurs communs à extraire

La détermination du nombre de facteurs communs à extraire dans une analyse factorielle se fait en s'appuyant sur les mêmes critères que ceux déjà présentés dans le cas de l'ACP. Essentiellement, le chercheur peut appuyer sa décision en considérant les quatre points suivants: 1) le critère de Kaiser (1960) indiquant la limite inférieure du nombre de facteurs à extraire, 2) l'inspection du graphique d'accumulation de variance de Cattell (1966), 3) l'analyse parallèle de Horn (1965) et 4) la facilité à interpréter les facteurs retenus. (Voir Baillargeon (2003) pour une présentation plus détaillée de ces critères.)

Preacher et MacCallum (2002) insistent sur le fait qu'aucun critère n'est absolu et que la décision quant au nombre de facteurs à extraire comportera toujours un élément important de subjectivité. Ils suggèrent de combiner l'analyse parallèle et l'inspection du graphique d'accumulation de variance de manière à prendre une décision éclairée. Dans le cas où les différents critères ne concordent pas, il est fortement recommandé de procéder à plusieurs analyses complètes en modifiant le nombre de facteurs extraits et en s'appuyant ultimement sur le critère plus subjectif relié à la facilité à interpréter la solution finale.

L'analyse factorielle exploratoire sur SPSS

1. La lecture de la matrice de corrélation

Puisque l'analyse factorielle s'intéresse aux facteurs latents pouvant rendre compte de la variance commune entre différentes variables, il est essentiel d'avoir accès à la **matrice de corrélation** pour procéder à ce genre d'analyse. Dans le cas du logiciel SPSS, la matrice de corrélation peut être calculée à partir des données brutes, c'est-à-dire des scores individuels sur les différentes variables. À l'occasion, il arrive qu'un chercheur n'ait pas accès à l'ensemble des données individuelles, mais qu'il souhaite quand même procéder à une analyse factorielle sur une matrice de corrélation reproduite dans une publication. Dans ce cas, il est possible d'utiliser la commande MATRIX de SPSS pour lire directement la matrice de corrélation et ensuite procéder à l'analyse factorielle proprement dite. C'est l'approche que nous avons utilisée avec la matrice de corrélation des motivations des consommateurs de bière (Wuensch, 2001). Notez que la matrice est lue avec des valeurs 1.0 dans la diagonale principale et que ce n'est qu'à l'étape suivante que ces valeurs seront remplacées par des estimés de la variance commune.

```
MATRIX DATA
   VARIABLES=PRIX QUANTITE ALCOOL PRESTIGE COULEUR AROME GOUT
   /CONTENT = CORR
   /N=100
BEGIN DATA
1.00
 .83 1.00
 .77  .90 1.00
-.41 -.39 -.46 1.00
 .02  .18  .07 -.37 1.00
-.05  .10  .04 -.44  .91 1.00
-.06  .03  .01 -.44  .91  .87 1.00
END DATA
```


2. La requête d'exécution d'une analyse factorielle exploratoire

Nous utilisons la commande FACTOR pour exécuter une analyse factorielle exploratoire. Vous remarquerez que cette commande est pratiquement identique à celle utilisée pour une analyse en composantes principales. La différence critique se situe dans la portion `/ EXTRACTION PAF` où l'on précise une extraction de type PAF « Principal Axis Factoring » plutôt que PC (« Principal Components »).

```
FACTOR MATRIX=IN(CORR=*)
       / PRINT CORRELATION KMO AIC DET INITIAL EXTRACTION ROTATION
       / EXTRACTION PAF
       / PLOT EIGEN
       / ROTATION VARIMAX
       / ROTATION OBLIMIN
```

Notez que trois autres sous-commandes très utiles ont déjà été présentées dans le document traitant de l'ACP et qu'elles s'appliquent tout aussi bien dans le cas de l'analyse factorielle.

La première `/ FORMAT = SORT BLANK (.3)` facilite l'interprétation de la solution finale en en mettant en ordre les variables en fonction de la taille de leurs pondérations respectives et en se limitant aux pondérations > .30. La sous-commande `/ CRITERIA = FACTORS (n)` quant à elle vous donne le plein contrôle sur le nombre de facteurs à extraire. Sachez que si vous n'utilisez pas cette sous-commande, SPSS appliquera par défaut le critère de Kaiser. Enfin, la sous-commande `/ SAVE = REG (ALL C)` permet de sauvegarder, pour chaque individu de notre échantillon, les nouveaux scores factoriels générés par l'AFE. Rappelons que cette option n'est disponible que si vous avez exécuté votre analyse factorielle à partir d'un fichier contenant les données individuelles sur chacune des variables analysées.

3. La rotation des facteurs initiaux

Dans le but de faciliter l'interprétation des facteurs latents extraits par l'analyse factorielle, il est fortement suggéré de procéder à une rotation de ces facteurs. Rappelons que la décision concernant le choix d'une rotation orthogonale ou oblique donne lieu à des débats assez virulents. Les tenants de la rotation orthogonale soulignent sa simplicité mathématique, alors que les défenseurs de la rotation oblique affirment que seule une rotation oblique est en mesure de bien refléter la réalité des phénomènes psychologiques étudiés. En parlant de la rotation orthogonale, Pedhazur et Pedhazur Schmelkin (1991) affirment que « … de telles solutions sont, la plupart du temps, des représentations naïves et irréalistes des phénomènes socio-comportementaux (p. 615). » et un peu plus loin, « Tout se ramène à la question suivante : Les aspects que nous postulons à propos d'un construit multidimensionnel sont-ils intercorrélés? La réponse à cette question est reléguée à un simple statut de supposition lorsque nous employons une rotation orthogonale (p. 615). » Pour cette raison, Pedhazur et Pedhazur Schmelkin (1991) recommandent vivement de procéder aux deux types de rotation; si la rotation oblique démontre une corrélation importante entre les dimensions et que cet état de fait correspond à la position théorique entretenue à l'égard du construit étudié, il faut alors privilégier cette solution plus représentative de la réalité. Si par ailleurs, la solution oblique démontre l'absence de corrélation (ou une corrélation négligeable) entre les facteurs, il est alors approprié de se rabattre sur la solution orthogonale plus simple. C'est essentiellement la même approche qui a été récemment recommandée par Preacher et MacCallum (2002).

4. L'examen des résultats produits par SPSS

SPSS présente d'abord la matrice de corrélation soumise à l'analyse. Rappelons qu'il s'agit exactement de la même matrice de corrélation que celle que nous avons déjà analysée à l'aide de la technique de l'ACP.

Correlation Matrix:

	PRIX	QUANTITE	ALCOOL	PRESTIGE	COULEUR	AROME	GOUT
PRIX	1.00000						
QUANTITE	.83000	1.00000					
ALCOOL	.77000	.90000	1.00000				
PRESTIGE	-.41000	-.39000	-.46000	1.00000			
COULEUR	.02000	.18000	.07000	-.37000	1.00000		
AROME	-.05000	.10000	.04000	-.44000	.91000	1.00000	
GOUT	-.06000	.03000	.01000	-.44000	.91000	.87000	1.00000

Puisque notre programme SPSS comportait la sous-commande suivante `/ PRINT CORRELATION KMO AIC DET INITIAL EXTRACTION ROTATION` nous retrouvons sous la matrice de corrélation les statistiques Kaiser-Meyer-Olkin, le test de sphéricité de Bartlett, de même que le déterminant de la matrice.

```
Determinant of Correlation Matrix =     .0004927

Kaiser-Meyer-Olkin Measure of Sampling Adequacy =  .66646

Bartlett Test of Sphericity = 729.82355, Significance =     .00000
```

À cette étape de l'analyse, les résultats présentés par SPSS sont en tous points identiques à ceux préalablement obtenus dans l'ACP. Ils nous permettent simplement de juger des propriétés de la matrice de corrélation et de détecter diverses anomalies qui pourraient compromettre la faisabilité de l'analyse factorielle. Dans le cas présent, nous constatons que le déterminant est > .00001, que la mesure d'adéquacité de l'échantillonnage est « moyenne » et que l'hypothèse d'une matrice d'identité est facilement rejetée.

Nous retrouvons ensuite dans la diagonale de la matrice « Anti-image Correlation AIC » les mesures Kaiser-Meyer-Olkin d'adéquacité de l'échantillonnage pour chaque variable.

Anti-image Correlation Matrix:

	PRIX	QUANTITE	ALCOOL	PRESTIGE	COULEUR	AROME	GOUT
PRIX	**.78512**						
QUANTITE	-.52716	**.55894**					
ALCOOL	.07219	-.79326	**.64103**				
PRESTIGE	.27057	-.10526	.21809	**.74289**			
COULEUR	.06286	-.47688	.36276	-.25614	**.59006**		
AROME	.17294	.04027	-.05012	.30806	-.57590	**.79444**	
GOUT	-.09184	.43552	-.30274	.27034	-.71534	-.05696	**.67012**

La section suivante intitulée « Initial Statistics » correspond à l'essence même de l'analyse factorielle. Dans la section gauche de ce tableau, sous la colonne « Communality » vous constaterez que les valeurs ne sont plus des 1.0 comme en ACP, mais qu'elles sont toutes inférieures à 1.0. C'est ici que l'analyse factorielle se démarque de l'ACP puisque désormais c'est la variance commune qui est analysée et non plus la variance totale de chaque variable. Les valeurs .73829, .90663 … .88809 correspondent aux R^2 de chaque variable, utilisés ici comme estimés de départ de leur variance commune. Autrement dit, nous assumons que .73829 unités de variance de la variable Prix correspond à de la variance importante partagée par les autres variables, alors que .26171 unités, c'est-à-dire 1.0 - .73829, serait de la variance négligeable, puisque correspondant soit à de la variance spécifique de cette variable, soit à son erreur de mesure. J'attire votre attention sur la variable Prestige qui semble avoir une proportion plutôt importante de variance unique. Dans le cas où une variable se démarquerait vraiment beaucoup de l'ensemble des autres variables, il serait certainement opportun de s'interroger sur la pertinence de la conserver dans l'analyse.

Si vous faites la somme des variances communes, vous arriverez à un total de 5.68157 correspondant à 81% de la variance totale (5.68 / 7). C'est cette proportion de **variance commune** qui sera éventuellement répartie de façon décroissante entre les différents facteurs retenus. La portion de droite du tableau, quant à elle, présente la proportion décroissante de variance **TOTALE** attribuée à chaque facteur. Il semble que ce soit simplement pour des raisons de facilité que SPSS présente ici ces pourcentages en termes de variance totale (plutôt que commune).

```
Initial Statistics:


Variable     Communality  *  Factor   Eigenvalue   Pct of Var   Cum Pct
                          *
PRIX              .73829  *     1       3.31217       47.3         47.3
QUANTITE          .90663  *     2       2.61662       37.4         84.7
ALCOOL            .85932  *     3        .57780        8.3         93.0
PRESTIGE          .50439  *     4        .23840        3.4         96.4
COULEUR           .92641  *     5        .13526        1.9         98.3
AROME             .85844  *     6        .08297        1.2         99.5
GOUT              .88809  *     7        .03678         .5        100.0
```

La courbe d'accumulation de variance (« scree test ») de Cattell est également reproduite puisque la sous-commande `/ PLOT  = EIGEN` a été utilisée. Dans le cas présent, c'est encore une fois le critère de Kaiser qui a provoqué par défaut l'extraction de deux facteurs, tel qu'indiqué au bas du graphique. Vous devrez utiliser la sous-commande `/ CRITERIA  = FACTORS (n)` et relancer la tâche SPSS si vous jugez qu'un nombre différent de facteurs aurait dû être retenu.

```
Eigenvalue (Scree) Plot

      3.312 +   *
            |
            |
 E          |
 I    2.617 +       *
 G          |
 E          |
 N          |
 V          |
 A          |
 L          |
 U          |
 E          |
 S          |
            |
       .578 +           *
       .238 +               *
       .000 +---+---+---+---+---*---*---*
                1   2   3   4   5   6   7


                      FACTORS

PAF   extracted   2 factors.    7 iterations required.
```

La matrice suivante est identifiée par SPSS comme étant la « Factor Matrix » et elle correspond à la matrice des poids factoriels (« factor loadings ») **avant** rotation. Comme nous l'avons déjà mentionné dans le cas de l'ACP, il n'y a généralement pas beaucoup d'intérêt à examiner cette matrice puisqu'elle ne donne pas une vision très nette des facteurs extraits. Vous pouvez constater que les pondérations se répartissent de manière importante sur l'ensemble des variables rendant l'interprétation plus difficile. Il y a donc avantage à procéder à des rotations (orthogonale et oblique) des facteurs dans le but de mettre en évidence une structure simple.

```
Factor Matrix:


              Factor  1     Factor  2

PRIX            .49125        .71371
QUANTITE        .63817        .70582
ALCOOL          .58748        .72106
PRESTIGE       -.61091       -.09428
COULEUR         .79174       -.52303
AROME           .76088       -.56031
GOUT            .74227       -.59524
```

La première rotation demandée était de type Varimax, donc orthogonale. Les statistiques finales présentent les résultats de cette rotation des deux facteurs retenus. Vous constaterez que la variance commune a maintenant été redistribuée différemment.

```
Final Statistics:


Variable     Communality  *  Factor   Eigenvalue   Pct of Var   Cum Pct
                          *
PRIX             .75072   *     1       3.12370       44.6         44.6
QUANTITE         .90545   *     2       2.47821       35.4         80.0
ALCOOL           .86507   *
PRESTIGE         .38211   *
COULEUR          .90041   *
AROME            .89289   *
GOUT             .90527   *



VARIMAX   rotation   1 for extraction   1 in analysis  1 - Kaiser
Normalization.
```

Après rotation, la matrice des poids factoriels est beaucoup plus facile à analyser. Une structure simple est maintenant visible puisque le facteur 1 obtient des pondérations importantes uniquement sur les variables Couleur, Arôme et Goût, alors que le facteur 2 est défini par les variables Quantité, Alcool et Prix.

Rotated Factor Matrix:

	Factor 1	Factor 2
PRIX	-.04607	**.86521**
QUANTITE	.07515	**.94858**
ALCOOL	.02569	**.92974**
PRESTIGE	-.42659	-.44736
COULEUR	**.94643**	.06850
AROME	**.94471**	.02013
GOUT	**.95127**	-.01890

Dans le cas d'une rotation orthogonale les poids factoriels sont équivalents à des coefficients de corrélation entre les variables et les facteurs. Le facteur 1 est donc fortement corrélé avec des scores élevés attribués aux motivations touchant la couleur, l'arôme et le goût de la bière; on retrouve ici la dimension qui avait été interprétée comme étant propre au Dégustateur de bière. La deuxième dimension correspondrait quant à elle aux motivations du gros Buveur, ce facteur étant fortement corrélé avec le prix, la quantité et le taux d'alcool de la bière consommée. Curieusement, la motivation liée au prestige est en corrélation négative avec ces deux facteurs. Une interprétation possible de ce résultat consisterait à dire qu'une dimension liée au prestige pourrait être pertinente, mais qu'elle était probablement sous-représentée parmi les variables soumises à l'analyse, ce qui a pu empêcher l'émergence d'un troisième facteur; l'ajout de quelques variables touchant au prestige pourrait peut-être venir renforcer cette dimension. Somme toute, la rotation orthogonale de cette analyse factorielle reproduit essentiellement les mêmes résultats que ceux que nous avions obtenus par l'ACP.

Comme plusieurs auteurs (Pedhazur & Pedhazur Schmelkin, 1991; Preacher & MacCallum, 2002) l'ont mentionné, la rotation orthogonale assume que les facteurs sont indépendants sans vraiment vérifier cette prétention. Il est donc fortement recommandé d'examiner aussi les résultats d'une rotation oblique, même si cette solution est quelques fois plus difficile à interpréter en raison même de la corrélation entre les facteurs.

Pattern Matrix:

	Factor 1	Factor 2
PRIX	-.10942	.87535
QUANTITE	.00669	.95057
ALCOOL	-.04179	.93517
PRESTIGE	-.39737	-.41988
COULEUR	.94886	.00028
AROME	.95066	-.04835
GOUT	.96010	-.08816

La section identifiée par SPSS sous le nom de « Pattern Matrix » présente la matrice des pondérations factorielles. Ces valeurs correspondent aux coefficients de régression lorsque l'on tente d'expliquer les variables à l'aide des différents facteurs utilisés comme variables prévisionnelles. Ainsi, on peut affirmer que le facteur 1 a un poids de .94886 pour expliquer la motivation liée à la couleur de la bière, alors que le facteur 2 n'a qu'un poids négligeable de .00028 pour expliquer cette même motivation. Cependant, comme il peut exister une corrélation entre les facteurs 1 et 2, il n'est pas possible d'interpréter les pondérations factorielles comme étant des corrélations. Il faut inspecter la matrice identifiée « Structure Matrix » pour retrouver les coefficients de corrélation entre les facteurs et les variables. Les corrélations reproduites dans l'encadré ci-bas permettent encore une fois de proposer une interprétation en deux facteurs, le premier correspondant aux motivations du dégustateur de bière, le deuxième aux motivations du gros buveur.

Structure Matrix:

	Factor 1	Factor 2
PRIX	.01625	**.85965**
QUANTITE	.14316	**.95153**
ALCOOL	.09247	**.92917**
PRESTIGE	-.45765	-.47693
COULEUR	**.94890**	.13650
AROME	**.94372**	.08813
GOUT	**.94745**	.04967

Finalement, la solution oblique nous permet de vérifier l'existence ou non d'une corrélation entre les deux dimensions. À toutes fins pratiques, dans le cas présent, cette corrélation est négligeable. Tenant compte de ce résultat, il serait tout à fait indiqué de revenir à la solution orthogonale plus simple et de mettre de côté la solution oblique.

Factor Correlation Matrix:

	Factor 1	Factor 2
Factor 1	1.00000	
Factor 2	**.14356**	1.00000

Remarques finales

L’analyse factorielle que nous venons d’examiner à l’aide de données fictives portant sur les motivations des buveurs de bière a permis de dresser un tableau pratiquement identique à celui auquel nous étions arrivés précédemment avec une analyse en composantes principales. Bien sûr les résultats chiffrés ne sont pas exactement les mêmes, mais la solution globale semble révéler la présence des deux mêmes dimensions (facteurs ou composantes principales).

On est donc en droit de se demander si les deux techniques sont à ce point équivalentes que l’on pourrait choisir l’une ou l’autre indifféremment. Il est vrai que dans bien des situations concrètes de recherche l’analyse en composantes principales et l’analyse factorielle donneront des solutions fort comparables. Stevens (1992) souligne que des solutions semblables seront effectivement obtenues lorsque le nombre de variables dépasse 30 et que la variance commune d’aucune de ces variables n’est inférieure à .4. Cependant, cette concordance fréquente entre les résultats des deux types d’analyse ne devrait pas nous faire oublier que ces techniques reposent sur des postulats et des modèles mathématiques bien distincts. Pour cette raison, le chercheur qui souhaite réduire une masse de données en conservant le maximum de variance totale de ses variables devrait privilégier l’analyse en composantes principales. En contrepartie, le chercheur qui prétend pouvoir expliquer la variance commune entre certaines variables à l’aide de variables latentes devrait certainement opter pour une analyse factorielle.

Références

Baillargeon, J. (2003). *L'analyse en composantes principales*. UQTR. [Document inédit disponible en ligne]. http://www.uqtr.ca/cours/srp-6020/acp/acp.pdf

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245-276.

Cooney, N. L., Kadden, R. M., & Litt, M. D. (1990). A comparison of methods for assessing sociopathy in male and female alcoholics. *Journal of Studies on Alcohol*, *51*, 42-48.

Johnston, C., & Mash, E. J. (1989). A measure of parenting satisfaction and efficacy. *Journal of Clinical Child Psychology*, *18*, 167-175.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141-151.

Ohan, J. L., Leung, D. W., & Johnston, C. (2000). The parenting sense of competence scale: Evidence of a stable factor structure and validity. *Canadian Journal of Behavioural Science*, *32*, 251-261.

Pedhazur, E. J., & Pedhazur Schmelkin, L. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: LEA.

Preacher, K. J. & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. [document disponible en ligne]. http://quantrm2.psy.ohio-state.edu/maccallum/tomswift/paper.htm

Stevens, J. (1992). *Applied multivariate statistics for the social sciences* (2e édition). Hillsdale, NJ: LEA.

Vogt, W. P. (1993). *Dictionary of statistics and methodology : A nontechnical guide for the social sciences.* Newbury Park, CA : Sage.

Wuensch, K. L. (2001). Principal component analysis. [document disponible en ligne]. http://core.ecu.edu/psyc/wuenschk/MV/FA/PCA.doc

