L'analyse en composantes principales.

Objectifs de la section

Au terme de cette section vous serez en mesure :

- D'identifier les situations propices à l'utilisation de l'analyse en composantes principales
- De vous assurer que vos données rencontrent ses conditions d'application
- De prendre les décisions appropriées quant au nombre de composantes à extraire
- D'effectuer une analyse en composantes principales sur SPSS
- D'interpréter adéquatement les résultats obtenus

Introduction

L'analyse en composantes principales (ACP) est souvent confondue avec l'analyse factorielle (AF) que nous examinerons dans une prochaine section. Plusieurs chercheurs ne font d'ailleurs pas de nuances entre les deux techniques et les considèrent comme étant deux variations d'une même méthode générale qu'ils identifient alors sous le vocable général d'analyse factorielle. Nous verrons que malgré l'apparence de grande similitude entre l'analyse en composantes principales et l'analyse factorielle proprement dite, il existe des différences fondamentales entre ces deux techniques et il importe donc de les traiter séparément. La première raison qui explique pourquoi il existe tant de confusion entre l'ACP et l'AF est sans doute le fait que ces techniques s'intéressent toutes deux à l'analyse de matrices de corrélation. Aussi, nous verrons que les deux techniques produisent des résultats qui se présentent sous des formes très similaires. Enfin, les deux techniques sont souvent offertes à l'intérieur des logiciels comme SPSS et SAS. À vrai dire, dans le cas du logiciel SPSS, nous verrons que les deux types d'analyse sont invoqués à l'aide de la même commande FACTOR, ce qui ne contribue en rien à diminuer la confusion.

Quelques exemples réels d'application

La fonction première de l'analyse en composantes principales en est une de **réduction des données**. On peut imaginer de nombreuses situations où les chercheurs sont aux prises avec un nombre très élevé de variables à considérer simultanément. Par exemple, la mise au point d'un instrument de mesure portant sur un construit théorique le moindrement complexe nécessite généralement une cueillette de données réparties sur une bonne centaine d'items. La recherche de Saintonge et Lachance (1995) illustre bien cette situation. Leur étude porte sur la validation d'une adaptation canadienne-française du SITA, un

test de séparation-individuation à l'adolescence développé initialement par Levine, Green et Millon (1986). Les auteurs ont administré le SITA comportant 103 items à un groupe de 424 jeunes âgé(e)s entre 12 et 22 ans. L'analyse en composantes principales a permis d'épurer l'instrument et de le réduire à une version comportant 52 items regroupés en 9 échelles. Saintonge et Lachance (1995) ont ensuite poursuivi leur analyse de cette version épurée du SITA en démontrant qu'il existait des différences significatives entre gars et filles (et entre adolescents plus âgés et plus jeunes) sur un certain nombre des neuf échelles identifiées par l'analyse en composantes principales.

Voici un deuxième exemple d'étude mettant à profit l'analyse en composantes principales. Carey, Faulstich, Gresham, Ruggiero et Enyart (1987) ont administré le « Children's Depression Inventory CDI » développé par Kovacs (1985) à un groupe de 153 jeunes patients hospitalisés pour soins psychiatriques de même qu'à un groupe témoin de 153 enfants provenant de la population générale. Le CDI comporte 27 items, mais l'ACP a permis de mettre en évidence deux ou trois composantes principales représentant de façon parcimonieuse une portion significative de la variance du test initial. La réduction des 27 items en un nombre limité de composantes a alors permis aux chercheurs de procéder à une analyse de fonctions discriminantes où il a été démontré que les composantes du CDI permettent de discriminer de façon adéquate les enfants appartenant au groupe clinique de ceux constituant le groupe témoin.

Terminons avec ce troisième exemple de recherche où l'on retrouve une application représentative de l'ACP. Ossenkopp et Mazmanian (1985) voulaient prédire la réponse physiologique au froid chez des rats soumis à une exposition de quatre heures en froid extrême. L'ensemble de leurs variables prévisionnelles était composé de 19 variables comportementales et physiologiques. Plutôt que d'utiliser ces 19 variables dans une analyse de régression multiple, ils ont d'abord procédé à une analyse en composantes principales qui leur a permis d'extraire 5 composantes représentant une proportion importante de la variance des mesures initiales. L'analyse de régression multiple a ensuite été menée en utilisant comme variables prévisionnelles les scores obtenus sur les cinq composantes extraites.

Les trois exemples que nous venons de citer ont en commun d'utiliser l'analyse en composantes principales dans une perspective de réduction des données et dans les trois cas on peut observer que les données ainsi réduites (les composantes) sont soumises à des analyses ultérieures. Ainsi, Saintonge et Lachance (1995) ont utilisé les composantes extraites en tant que variables dépendantes dans des analyses de variance et de tests *t*. Carey et al. (1987) quant à eux ont procédé à une analyse de fonctions discriminantes en se servant des composantes extraites pour définir leurs fonctions. Enfin, Ossenkopp et Mazmanian (1985) ont créé leur équation de régression multiple en utilisant comme variables prévisionnelles les scores obtenus préalablement par l'analyse en composantes principales.

Les trois exemples précédents illustrent parfaitement bien la définition de l'ACP telle qu'on la retrouve dans le « *Dictionary of statistics and methodology* » (Vogt, 1993) :

> L'analyse en composantes principales : Ensemble de méthodes permettant de procéder à des transformations linéaires d'un grand nombre de variables intercorrélées de manière à obtenir un nombre relativement limité de composantes non corrélées. Cette approche facilite l'analyse en regroupant les données en des ensembles plus petits et en permettant d'éliminer les problèmes de multicolinéarité entre les variables. L'analyse en composantes principales s'apparente à l'analyse factorielle, mais c'est une technique indépendante qui est souvent utilisée comme première étape à une analyse factorielle (Vogt, 1993, page 177).

Stevens (1992) met lui aussi en relief l'intérêt qu'offre l'analyse en composantes principales comme technique de réduction des données. Il énumère trois cas particuliers que l'on mettra facilement en rapport avec les trois exemples de recherches examinés plus tôt : 1) lors de la mise au point d'un instrument de mesure, un chercheur peut vouloir déterminer combien de dimensions indépendantes sont nécessaires pour rendre compte de la majorité de la variance présente dans l'ensemble des items de l'instrument; 2) dans le contexte de l'analyse de régression multiple, un chercheur peut être en présence d'un nombre trop élevé de variables prévisionnelles compte tenu de la taille limitée de l'échantillon disponible. En utilisant un nombre limité de composantes principales, un ratio « N / variables prévisionnelles » plus adéquat peut être atteint. 3) Enfin, dans le contexte des analyses de variance multivariées (MANOVA), Stevens (1992) rappelle qu'il est généralement préférable de limiter le nombre de variables dépendantes soumises à une même analyse.

Description sommaire de la technique

L'idée à la base de l'analyse en composantes principales est de pouvoir expliquer ou rendre compte de la variance observée dans la masse de données initiales en se limitant à un nombre réduit de composantes, définies comme étant des transformations mathématiques pures et simples des variables initiales. L'algorithme utilisé pour la détermination de ces composantes obéit à deux contraintes importantes. Tout d'abord, la première composante extraite doit correspondre à un score composite qui **maximise** la proportion de variance expliquée dans les variables initiales. Pour comprendre cette idée il est avantageux de faire une analogie avec la technique de régression multiple. Comme vous le savez, dans une analyse de régression multiple nous cherchons à expliquer le maximum de variance possible dans une variable critère (variable dépendante) en déterminant mathématiquement les pondérations optimales des différentes variables prévisionnelles (variables indépendantes). Dans le cas de

l'analyse de régression nous avons en main non seulement les variables prévisionnelles, mais aussi la variable critère, puisqu'elle a été directement mesurée par le chercheur. Par analogie, l'analyse en composantes principales serait un peu comme une analyse de régression pour laquelle nous ne connaîtrions pas la variable critère à expliquer. Il s'agirait de la découvrir. Heureusement, l'algorithme utilisé dans l'ACP assure que la composante C_1, la première extraite, correspondra à la plus grande proportion possible de variance présente dans les variables initiales. Ainsi, l'analyse en composantes principales nous mettra en présence d'une équation très apparentée à l'équation de régression classique ayant la forme suivante :

$$C_1 = \hat{a}_1 \, var_1 + \hat{a}_2 \, var_2 + \hat{a}_3 \, var_3 \ldots + \hat{a}_k \, var_k$$

Idéalement, nous aimerions que cette première composante C_1 corresponde à une proportion très importante de la variance présente dans nos données initiales; ainsi, 80% ou 70% de variance expliquée à l'aide d'une première composante serait certainement un résultat très apprécié du chercheur. Cependant la réalité est souvent moins gratifiante et il est fréquent de n'expliquer que 40%, 30%, ou même 20% lors de l'extraction d'une première composante. La variance restante, inexpliquée par C_1, n'est pas laissée de côté dans l'analyse des composantes principales; au contraire, elle est soumise à son tour au même processus d'extraction des composantes. Mais ici, l'algorithme à la base de l'ACP obéit à une deuxième contrainte importante : il cherche à extraire une deuxième composante, **indépendante de la première**, qui expliquerait à son tour la plus grande proportion de variance possible parmi la variance laissée inexpliquée par la composante C_1. La composante C_2 sera donc représentée à son tour par une nouvelle équation où les coefficients $\hat{a}_1$, $\hat{a}_2$, $\hat{a}_3$ … $\hat{a}_k$ correspondront à autant de nouvelles pondérations des différentes variables initiales en tant que variables prévisionnelles de C_2.

$$C_2 = \hat{a}_1 \, var_1 + \hat{a}_2 \, var_2 + \hat{a}_3 \, var_3 \ldots + \hat{a}_k \, var_k$$

On peut déceler différentes conséquences de cette approche analytique. Tout d'abord, il devrait être évident que les composantes extraites expliqueront chacune une proportion de variance de moins en moins importante. C_1 explique plus de variance que C_2, C_2 plus que C_3, C_3 plus que C_4, etc. Par ailleurs, la proportion de variance totale cumulée à travers les différentes composantes pourra éventuellement atteindre 100% si le processus d'extraction est mené à terme, c'est à dire si le nombre de composantes extraites équivaut au nombre de variables initialement soumises à l'analyse. Rappelons cependant que l'objectif premier de cette technique d'analyse en est précisément un de réduction de la masse de données. Il serait donc paradoxal de vouloir réduire, disons, la complexité d'une centaine de variables en extrayant une centaine de composantes! En d'autres termes, le chercheur devra prendre une décision importante quant au nombre de composantes principales qu'il juge opportun

d'extraire. Nous reviendrons plus loin sur cette question difficile et examinerons quelques critères servant à prendre une décision judicieuse.

Quelques particularités des matrices de corrélation

On aura compris dans les pages précédentes que le chercheur qui utilise une approche en composantes principales ne s'intéresse pas particulièrement aux variables individuelles de son étude, souvent très nombreuses, mais qu'il mise plutôt sur la présence d'**intercorrélation** entre ces variables pour pouvoir en extraire des dimensions plus globales. En fait, comme nous le verrons dans les exercices pratiques, il est possible de générer une ACP à partir d'une matrice de corrélation, sans même avoir accès aux scores bruts correspondant aux données individuelles. Vous pourriez donc utiliser une matrice de corrélation apparaissant dans une recherche publiée et reproduire, vérifier ou même modifier l'analyse en composantes principales faites par d'autres chercheurs.

Il existe une particularité vraiment remarquable des matrices de corrélation : elles grossissent à vue d'œil en fonction du nombre de variables individuelles étudiées. Le tableau 1.1 illustre à quel point la masse de données peut grossir à un point tel, qu'il devient très difficile d'examiner chacun des éléments de la matrice.

Tableau 1.1 Progression du nombre de coefficients d'intercorrélation en fonction de l'augmentation du nombre de variables individuelles (*k*).

Nombre de variables individuelles *k*	Nombre de coefficients de corrélation distincts	Nombre de variables individuelles *k*	Nombre de coefficients de corrélation distincts
1	0	20	190
2	1	30	435
3	3	40	780
4	6	50	1225
5	10	60	1770
6	15	70	2415
7	21	80	3160
8	28	90	4005
9	36	100	4950
10	45	200	19900
11	55	300	44850
12	66	400	79800
13	78	500	124750
14	91	600	179700
15	105	700	244650

On voit que la progression augmente rapidement en fonction de l'équation suivante :

$$\frac{k(k-1)}{2}$$

où
k = nombre de variables individuelles

Équation 1.1

Comme les matrices de corrélation constituent les données de base soumises à l'analyse en composantes principales, il importe d'examiner leurs caractéristiques d'un peu plus près. Pour ce faire, j'ai emprunté un exemple d'étude fictive proposé par Wuensch (2001) dans laquelle une centaine de consommateurs auraient indiqué, sur des échelles de mesure appropriées, quelle est l'importance de sept critères spécifiques dans leur motivation à acheter une marque de bière plutôt qu'une autre. Le tableau 1.2 présente la matrice d'intercorrélation entre les sept variables. Comme toutes les matrices de corrélation, il s'agit d'une matrice de forme carrée, c'est-à-dire comportant un nombre égal de rangées et de colonnes correspondant au nombre de variables. Notez que la taille de la matrice de corrélation n'a aucun rapport avec le nombre de sujets dans l'échantillon. Elle demeurerait 7 x 7 avec 200, 300 ou 1000 participants. Dans le cas présent, elle est donc constituée de 49 cellules. On observe évidemment la présence de la diagonale principale où s'alignent les valeurs 1.00 correspondant à la corrélation parfaite de chaque variable avec elle-même. La diagonale principale divise la matrice en deux portions triangulaires symétriques où l'on retrouve les mêmes coefficients de corrélation, puisque, par exemple, la corrélation entre le prix et la quantité (.83) est égale à la corrélation entre la quantité et le prix (.83). Vous verrez dans les exercices pratiques sur SPSS qu'il est suffisant de fournir la portion triangulaire inférieure de la matrice de corrélation pour pouvoir procéder à une ACP.

Tableau 1.2 Matrice d'intercorrélation entre sept variables mesurant les motivations à acheter une marque de bière particulière.

	PRIX	QUANTITE	ALCOOL	PRESTIGE	COULEUR	AROME	GOÛT
PRIX	**1.00**	.83	.77	-.41	.02	-.05	-.06
QUANTITE	.83	**1.00**	.90	-.39	.18	.10	.03
ALCOOL	.77	.90	**1.00**	-.46	.07	.04	.01
PRESTIGE	-.41	-.39	-.46	**1.00**	-.37	-.44	-.44
COULEUR	.02	.18	.07	-.37	**1.00**	.91	.91
AROME	-.05	.10	.04	-.44	.91	**1.00**	.87
GOÛT	-.06	.03	.01	-.44	.91	.87	**1.00**

Que pouvons-nous dire de la taille des coefficients de corrélation apparaissant au tableau 1.2? Certains de ces coefficients sont particulièrement petits, par exemple entre la couleur de la bière et son prix (.02) ou entre le taux

d'alcool et le goût (.01). Vous comprendrez que si tous les coefficients de corrélation étaient aussi faibles que ceux-là, il n'y aurait absolument aucun intérêt à procéder à une analyse en composantes principales de ces données. En effet, pour pouvoir extraire une composante correspondant à une fonction linéaire des variables initiales, il faut nécessairement que ces variables soient intercorrélées. Heureusement la matrice du tableau 1.2 comporte un certain nombre de coefficients de tailles intéressantes (-.41, .77, -.44, etc.) et même quelques coefficients particulièrement élevés (.87, .91, .91). L'analyse en composantes principales s'accommode assez bien des situations où un certain niveau de multicolinéarité existe entre les données. Cependant, il faut absolument se méfier de la condition dite de « **singularité** » où une variable serait parfaitement corrélée avec une autre variable ou avec une combinaison de plusieurs variables. Cette condition peut être détectée en calculant le « déterminant » de la matrice de corrélation | R |. Le déterminant est une valeur numérique unique associée à une matrice carrée et qui peut prendre n'importe quelle valeur entre 0.0 et 1.0. Cependant ces deux valeurs extrêmes sont problématiques. En effet, un déterminant de 0.0 indique que la matrice est singulière c'est-à-dire qu'il existe au moins un cas de dépendance linéaire dans la matrice ou, en d'autres mots, qu'une variable peut être entièrement expliquée ou prédite par une combinaison linéaire d'autres variables. Vous seriez confrontés à cette situation problématique si votre matrice de corrélation comportait par exemple des variables comme FRANÇAIS, ANGLAIS, HISTOIRE, MATH et TOTAL.

Comme le mentionne Field (2000), on ne devrait jamais procéder à une ACP sur une matrice de corrélation dont le déterminant est plus petit que 0.00001. À l'inverse, un déterminant égal à 1.0 correspond lui aussi une condition impropre à l'ACP; il indique que la matrice de corrélation est une **matrice d'identité**, c'est-à-dire une matrice ne contenant que des valeurs 0.0 , sauf pour la présence des valeurs 1.0 dans la diagonale. Il existe un test statistique qui permet de mettre à l'épreuve l'hypothèse nulle selon laquelle la matrice de corrélation observée dans notre échantillon proviendrait d'une population où la matrice serait une matrice d'identité. C'est le test de **sphéricité de Bartlett**. Évidemment, nous souhaitons vivement que ce test soit significatif pour nous autoriser à rejeter l'hypothèse nulle d'identité indiquant l'absence de corrélation significative entre nos variables. Il faut dire que le test de Bartlett est sensible à la taille de l'échantillon et que lorsque le N est assez grand, les chances de rejeter l'hypothèse nulle sont très élevées. En ce sens, le rejet de l'hypothèse nulle ne garantit pas nécessairement que l'ACP donnera de bons résultats; à l'inverse, si le test de Bartlett ne nous permet pas de rejeter l'hypothèse nulle, nous sommes en présence d'une situation vraiment extrême où l'ACP n'est pas justifiable. Qu'en est-il de la matrice du tableau 1.2? Son déterminant est de 0.0004927, indiquant qu'il ne s'agit pas d'une matrice singulière; de plus, le test de sphéricité de Bartlett nous donne une valeur de 729.82, $p < .00000$ nous permettant évidemment de rejeter l'hypothèse nulle et d'affirmer qu'il ne s'agit pas non plus

d'une matrice d'identité. Il serait donc légitime de procéder à une ACP des données du tableau 1.2.

Nous venons de voir deux indices (le déterminant et le test de sphéricité de Bartlett) qui nous aident à vérifier si dans l'ensemble une matrice de corrélation possède les propriétés souhaitées pour l'analyse en composantes principales. Il est également important d'examiner chacune des variables de façon individuelle pour nous assurer que chacune d'elles est en relation avec l'ensemble des autres variables. Par exemple, en inspectant une à une chaque rangée de la matrice du tableau 1.2 vous constaterez que toutes les variables démontrent au moins une corrélation substantielle avec une autre variable. Lorsque nous sommes en présence d'une variable qui n'est en corrélation avec aucune autre dans la matrice, il est recommandé de retrancher cette variable avant de procéder à une ACP.

L'examen des variables individuelles est grandement facilité par le calcul des mesures d'adéquacité de l'échantillonnage de Kaiser-Meyer-Olkin (Measure of Sampling Adequacy, MSA »). Ces indices se calculent pour chacune des variables de même que pour la matrice globale et peuvent prendre elles aussi des valeurs entre 0.0 et 1.0 Pour être conservée dans une ACP, une variable doit obtenir une mesure K-M-O dépassant 0.5. Kaiser (1974) a suggéré une gradation intéressante utilisant les points de référence suivants : 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.

Tableau 1.3 Mesures Kaiser-Meyer-Olkin d'adéquacité de l'échantilonnage calculées pour la matrice d'intercorrélation du tableau 1.1.

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

Les données présentées au tableau 1.3 ne sont pas particulièrement encourageantes, mais elles reflètent probablement la nature fictive des données mesurant les motivations des consommateurs de bière. Tenant compte de cette condition particulière, nous procéderons maintenant à l'extraction des composantes principales de ces données.

Extraction des composantes principales

Le nombre maximum de composantes principales qu'il est possible d'extraire d'une matrice de corrélation est égal au nombre de variables dans la matrice. Dans l'exemple qui nous intéresse nous pourrions donc extraire jusqu'à sept composantes. Toutefois, comme nous l'avons mentionné plus haut, le pourcentage de variance expliqué par chaque composante décroît systématiquement à mesure que l'on progresse dans le processus d'extraction et peut devenir tout à fait négligeable une fois que les composantes les plus importantes auront été extraites. Ceci nous amène à considérer différents critères qui nous aideront à déterminer combien de composantes il vaut la peine d'extraire.

1. Utilisation du critère de Kaiser (1960)

Pour comprendre ce critère il faut aborder brièvement la notion de variance présente dans les données. Dans le cas d'une matrice de corrélation comme celle présentée au tableau 1.2, les valeurs apparaissant dans la diagonale principale correspondent à la variance de chaque variable. Si vous avez de la difficulté à réconcilier cette affirmation avec l'observation que ce sont toutes des valeurs 1.0 qui apparaissent dans cette diagonale c'est simplement que vous n'avez pas réalisé que le calcul d'un coefficient de corrélation entraîne toujours une standardisation des variables. Par exemple, lorsque nous calculons la corrélation entre la taille d'individus mesurée en centimètres et leurs poids corporels mesurés en kilogrammes, nous perdons la métrique de ces deux mesures (cm et kg) parce que le calcul entraîne une standardisation sur de nouvelles échelles possédant chacune une moyenne de 0.0 et un écart-type de 1.0. Sur les échelles standardisées, chacune des mesures apparaissant dans une matrice de corrélation a donc bel et bien une variance de 1.0. La variance totale dans la matrice quant à elle, correspond à la somme des variances de chaque variable. Dans le cas qui nous intéresse la variance totale présente dans les données est donc de 7.0 puisqu'il y a sept variables dans la matrice de corrélation.

Comment cette variance totale (7.0) sera-t-elle répartie entre les différentes composantes que nous voulons extraire? La réponse s'obtient en calculant ce que l'on nomme la **valeur propre** ou « **eigenvalue** » de chaque composante. Le tableau 1.4 présente ces valeurs pour les données fictives simulant les motivations à acheter une marque de bière. On constate que la valeur propre (eigenvalue) de la première composante est de 3.31217 ce qui correspond à 47.3 % de la variance totale de 7.0. Comme nous l'avons mentionné précédemment, l'algorithme utilisé en ACP fait en sorte de maximiser la variance expliquée par la première composante. Toujours selon ce même algorithme, la deuxième composante extraite viendra expliquer une portion additionnelle de variance, indépendante de la première, et correspondant à une proportion plus

faible que la précédente. L'examen du tableau 1.4 permet de constater que la composante C_2 explique 2.61662 unités de variance (sur 7.0), ce qui correspond à 37.4 % de la variance totale. Nous pouvons donc dire qu'après avoir extrait deux composantes principales le chercheur serait en mesure de rendre compte de 84.7% de la variance des motivations animant le consommateur de bière. N'est-ce pas là précisément l'objectif de l'analyse en composantes principales? Réduire les données de 7 variables à 2 composantes tout en réussissant à rendre compte de 84.7% de la variance initiale… On pourrait même se demander si cela vaut vraiment la peine de continuer à extraire d'autres composantes au-delà de la dimension C_2. Le critère de Kaiser nous dit justement qu'il ne vaut pas la peine de poursuivre l'extraction puisque la composante C_3 n'expliquerait que .57780 unités de variance, ce qui correspond à moins de variance que celle associée à une variable initiale de la matrice de corrélation. Rappelez-vous que chaque variable possède 1.0 unité de variance. Selon Kaiser (1960), l'extraction des composantes doit donc s'arrêter dès qu'une valeur propre devient inférieure à 1.0.

Tableau 1.4 Répartition des valeurs propres (eigenvalues) et des pourcentages de variance associés à chacune des composantes principales.

Composante	Valeur propre « Eigenvalue »	Pourcentage de variance	Pourcentage de variance cumulée
C_1	3.31217	47.3	47.3
C_2	2.61662	37.4	84.7
C_3	.57780	8.3	93.0
C_4	.23840	3.4	96.4
C_5	.13526	1.9	98.3
C_6	.08297	1.2	99.5
C_7	.03678	.5	100.0
Total :	7.00000	100.00	

2. Utilisation du test d'accumulation de variance « scree test » de Cattell (1966)

En 1966, Cattell a proposé une méthode graphique pour décider du nombre de composantes à extraire. Le test d'accumulation de variance communément appelé « **scree test** » demande que l'on trace un graphique illustrant la taille des valeurs propres « eigenvalues » des différentes composantes en fonction de leur ordre d'extraction. Le terme « scree » fait référence à un phénomène géomécanique où l'on observe une accumulation de dépôts rocheux au pied d'une montagne, créant ainsi un petit promontoire à l'endroit où le dénivelé de la montagne se transforme brusquement en une pente plus douce. Le critère proposé par Cattell nous amène à arrêter l'extraction des composantes à l'endroit où se manifeste le changement de pente dans le graphique.

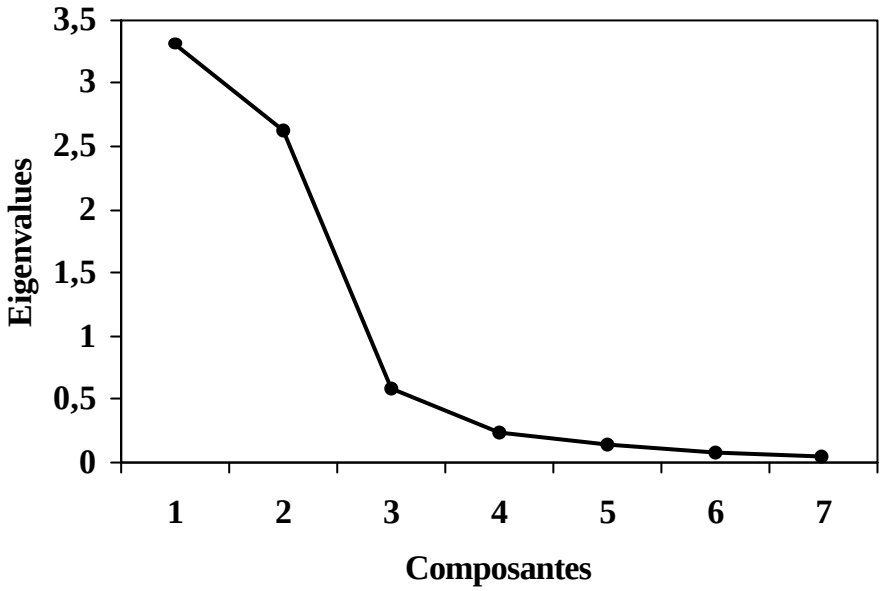

Figure 1.1 Illustration de l'accumulation de variance « scree test » de Cattell (1966).

La figue 1.1 correspond au test d'accumulation de variance pour nos données fictives. On y constate que la pente change radicalement avec la composante C_3. La représentation graphique des variances nous aide à voir que le point C_3 appartient beaucoup plus au segment C_3 à C_7 qu'au segment C_1 à C_3. Selon le critère de Cattell on devrait donc se limiter à l'extraction des deux premières composantes.

3. Utilisation de l'analyse parallèle de Horn (1965)

L'approche suggérée par Horn (1965) pour déterminer le nombre de composantes à retenir pour extraction s'appuie sur un raisonnement très différent des deux précédents. Horn indique qu'il est possible de découvrir par chance une composante pouvant expliquer une certaine proportion de variance, même en partant de données générées complètement au hasard et pour lesquelles aucune dimension réelle n'existe. Cette proportion de variance, expliquée par pure chance, pourrait donc servir comme point de comparaison afin de nous aider à décider si la variance que nous obtenons dans notre analyse est significativement plus importante que celle observable dans une matrice de données générées de façon aléatoire. L'analyse parallèle consiste donc à mener une ACP sur une matrice de corrélation générée au hasard mais comportant le même nombre de variables (et de participants) que notre étude. La série décroissante des valeurs propres (eigenvalues) calculées sur ces données aléatoires sera alors comparée aux valeurs propres calculées sur les données réelles. Si une composante existe vraiment dans nos données de recherche, sa

valeur propre correspondante devrait être significativement plus grande que celle obtenue sur les données aléatoires. Ainsi, Horn recommande de ne conserver pour extraction que les composantes dont les variances sont significativement supérieures à celles obtenues par pure chance. La prise de décision est relativement facilitée si l'on trace un graphique représentant les deux séries de valeurs propres. L'inspection de la figure 1.2 permet de constater que cette méthode indiquerait deux composantes à extraire de la matrice de corrélation portant sur les motivations de nos consommateurs de bière.

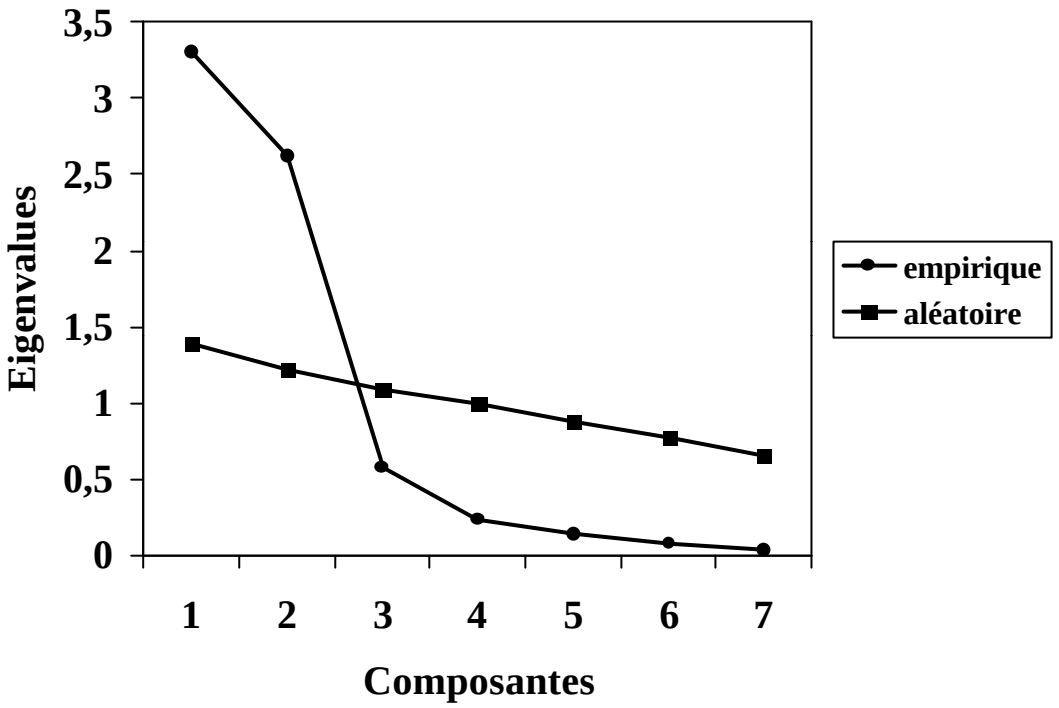

Figure 1.2 Illustration de l'analyse parallèle de Horn (1965)

Récemment, Kaufman et Dunlap (2000) ont grandement facilité l'utilisation de l'analyse parallèle en publiant un petit programme informatique qui calcule rapidement les valeurs propres que l'on obtiendrait par chance en partant de données aléatoires. L'utilisateur doit simplement indiquer le nombre de variables et le nombre de participants de son étude et le programme retourne les valeurs propres que l'on s'attendrait d'obtenir par pure chance. Ce petit programme qui fonctionne dans l'environnement Windows est disponible à l'adresse suivante : http://www.tulane.edu/~dunlap/psylib/pa.exe Vous pouvez aussi en obtenir une copie sur mon compte UNIX, tout en sachant que vous devrez le recopier dans votre environnement Windows : cp /u/b/a/baillarg/pub pa.exe pa.exe

4. Décision basée sur l'interprétation des composantes extraites

Ultimement, la décision concernant le nombre de composantes à extraire doit aussi tenir compte de la capacité des chercheurs à interpréter les dimensions extraites. Il ne sert à rien d'extraire une composante en s'appuyant sur un critère aussi rigoureux soit-il, si par ailleurs cette composante défie toute compréhension. Par ailleurs, Wood, Tataryn et Gorsuch (1966) ont démontré qu'une surestimation du nombre de composantes était généralement moins dommageable qu'une sous-estimation. Comme vous le voyiez la décision quant au nombre de composantes à extraire est difficile à prendre et comporte une part importante de subjectivité. Il est suggéré de confronter les différents critères plutôt que d'appliquer bêtement l'unique règle du eigenvalue > 1.0 de Kaiser.

L'analyse en composantes principales sur SPSS

1. La lecture de la matrice de corrélation

Il est facile d'obtenir une ACP sur SPSS à l'aide de la procédure FACTOR. Comme nous l'avons mentionné précédemment, ce type d'analyse utilise une **matrice de corrélation** comme données de départ. Si vous travaillez avec un fichier de données brutes comportant les scores individuels sur une série de variables, SPSS calculera automatiquement la matrice de corrélation nécessaire et la rendra accessible dès que vous invoquerez la commande FACTOR. Par ailleurs, vous pouvez fournir vous-même la partie triangulaire inférieure d'une matrice de corrélation, sans avoir accès aux données brutes individuelles. Dans un tel cas, votre programme SPSS comportera une section de lecture de cette matrice plutôt que la section habituelle du DATA LIST. L'encadré suivant décrit le jeu de commandes nécessaires pour lire la matrice de corrélation des motivations des consommateurs de bière.

```
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'éxécution d'une analyse en composantes principales

La commande FACTOR permet d'obtenir une ACP et de préciser un certain nombre d'options d'exécution comme le nombre de composantes à extraire, le type de rotation à effectuer, les statistiques à afficher, etc. L'encadré suivant présente la commande pour obtenir les résultats les plus utiles.

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

Trois autres sous-commandes méritent d'être commentées ici. La première prend la forme `/ FORMAT = SORT BLANK (.3)` et son utilisation facilite grandement l'interprétation des composantes extraites, surtout lorsqu'elles proviennent de données réelles complexes. L'effet de cette sous-commande est double : d'abord elle permet de mettre en ordre décroissant les pondérations des différentes variables sur chacune des composantes extraites, ensuite elle masque toutes les pondérations inférieures à .3 permettant ainsi de ne pas se laisser influencer par les variables moins significatives lors de l'interprétation des composantes extraites.

La sous-commande `/ CRITERIA = FACTORS (n)` est aussi très utile lorsque le chercheur désire fixer lui-même le nombre de composantes à extraire. En son absence, SPSS applique par défaut le critère de Kaiser et extrait automatiquement toutes les composantes dont les valeurs propres sont inférieures à 1.0. Après avoir considéré d'autres critères de décision (par ex., graphique d'accumulation de variance de Cattell ou résultat d'une analyse parallèle de Horn, etc.) le chercheur peut contrecarrer le choix imposé par SPSS et décider d'augmenter ou de réduire le nombre de composantes à extraire. Le nombre de composantes est alors indiqué dans la parenthèse.

Enfin, la sous-commande `/ SAVE = REG (ALL C)` permet de sauvegarder, pour chaque individu de notre échantillon, les nouveaux scores générés par l'ACP sur chacune des composantes extraites. Il est alors possible d'utiliser les scores de composantes (souvent nommés scores factoriels) à l'intérieur d'autres analyses comme la régression multiple, l'analyse discriminante, etc. Notez que cette portion de la commande FACTOR n'est utilisable que lorsque l'ACP est démarrée à partir d'un fichier SPSS contenant des données individuelles; elle est inapplicable lorsque qu'une matrice de corrélation est utilisée comme input.

3. La recherche d'une structure simple des composantes

L'une des étapes importantes dans l'ACP consiste à identifier et à nommer les composantes extraites. Pour ce faire, il est courant d'examiner chacune des composantes une à une et de déterminer avec lesquelles des variables initiales elles sont le plus en corrélation. Par exemple un chercheur pourrait découvrir que la composante C1 est fortement corrélée avec diverses mesures d'habileté verbale, alors que la composante C2 pourrait être plus fortement corrélée avec des mesures d'habileté visuo-spatiale. Constatant ce patron de corrélation, le chercheur serait justifié d'identifier les deux composantes extraites comme correspondant aux dimensions de l'intelligence verbale et de l'intelligence non verbale. Malheureusement, les résultats initiaux de l'ACP ne favorisent pas cette identification car la technique a tendance à produire une première composante générale sur laquelle plusieurs variables obtiennent des pondérations importantes. L'algorithme utilisé maximise la variance expliquée, mais au prix d'une interprétation souvent difficile des composantes extraites. La dernière étape de l'ACP consiste donc à transformer à nouveau la solution obtenue en faisant une rotation des axes servant à définir les différentes composantes. Cette transformation mathématique des vecteurs correspondant aux composantes **préserve la variance expliquée de chaque variable**, mais la réassigne à des composantes transformées.

On emploie le terme « rotation » parce que la détermination des nouvelles pondérations se fait en faisant pivoter les axes de référence (les composantes) de manière à simplifier la structure obtenue. Deux grands types de rotation peuvent être distingués : orthogonale et oblique. Dans le cas d'une rotation **orthogonale**, les axes de références seront déplacés en maintenant l'angle de 90 degrés qui les sépare, préservant ainsi l'indépendance des composantes. À l'opposé, une rotation **oblique** pourra déplacer les axes de références en augmentant ou en diminuant l'angle qu'ils forment entre eux. Cette section de l'ACP est particulièrement controversée : certains auteurs décrient l'approche oblique, invoquant qu'elle ajoute une transformation artificielle des données, alors que l'approche orthogonale est mathématiquement beaucoup pus simple. À l'inverse, les tenants de l'approche oblique affirment qu'elle respecte et colle beaucoup plus à la réalité des phénomènes étudiés en psychologie, puisque les construits psychologiques sont pratiquement toujours corrélés entre eux. Nous sommes effectivement de cet avis et acceptons l'idée exprimée par Preacher et MacCallum (2002) qui soutiennent qu'il est indéfendable sur un plan théorique d'imposer une structure d'indépendance à des dimensions qui sont effectivement corrélées.

Preacher et MacCallum (2002) sont catégoriques à cet égard : si un chercheur ne sait pas clairement comment des dimensions sont reliées entre elles, il n'est pas légitime d'assumer qu'elles sont indépendantes. Il est toujours préférable d'examiner la solution oblique et de vérifier s'il y a une corrélation entre les dimensions extraites, quitte à revenir ensuite à une solution orthogonale s'il n'y a vraiment pas de corrélation entre les dimensions.

4. L'examen des résultats produits par SPSS

La sortie imprimée des résultats SPSS présente d'abord la matrice de corrélation soumise à l'ACP. Rappelons que vous devriez déjà avoir examiné cette matrice avant de procéder à une ACP pour vous assurer que les variables utilisées seront pertinentes, suffisamment corrélées entre elles et qu'elles ne comporteront aucun cas de singularité où une variable serait entièrement définie par une ou plusieurs autres variables combinées.

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

Pour faciliter le diagnostic de conditions problématiques dans la matrice de corrélation, vous avez demandé d'imprimer certaines statistiques importantes comme le déterminant de la matrice, la mesure globale d'adéquacité d'échantillonnage de Kaiser-Meyer-Olkin, de même que le test de sphéricité de Bartlett. Vous trouverez les résultats de ces tests directement sous la matrice de corrélation.

Determinant of Correlation Matrix = **.0004927**

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

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

Ces statistiques nous encouragent à poursuivre l'ACP. En effet, le déterminant est > .00001 et donc ne s'approche pas trop de 0.0, la mesure d'adéquacité de l'échantillonnage peut être qualifiée de « moyenne » et le test de Bartlett nous permet de rejeter l'hypothèse nulle selon laquelle nos données proviendraient d'une population où la matrice de corrélation serait une matrice d'identité.

SPSS reproduit ensuite l'anti-image des matrices de covariance et des matrices de corrélation. La portion importante à considérer ici est la diagonale apparaissant dans la section « Anti-image Correlation Matrix. » Les valeurs présentées dans cette diagonale (.78512, .55894, … .67012) correspondent aux mesures Kaiser-Meyer-Olkin d'adéquacité de l'échantillonnage calculées pour chaque variable (nous les avions présentées au tableau 1.3). Ces valeurs sont dans la zone « médiocre – moyen – bien » et reflètent probablement la nature fictive des données analysées ici.

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 nous donne les statistiques initiales sous une forme qui peut prêter à confusion. En effet, il faut bien comprendre qu'il s'agit ici de **deux tableaux différents** placés côte à côte; ils sont à peine séparés par une ligne verticale constituée d'une série d'astérisques (*). Le premier, dans la portion de gauche, énumère la proportion de variance attribuée à chacune des sept variables de notre analyse. Dans le cas présent, les valeurs 1.0 correspondent au fait que chaque variable a une variance de 1.0 et que la totalité de cette variance sera utilisée pour déterminer les composantes principales. La portion de droite du tableau, quant à elle, nous donne les valeurs propres décroissantes et les pourcentages correspondants de chaque composante « Factor » pouvant être extraites. Prenez la peine d'aller vérifier que ces valeurs sont bien celles dont nous avions fait état précédemment dans le tableau 1.4.

Initial Statistics:

Variable	Communality	*	Factor	Eigenvalue	Pct of Var	Cum Pct
		*				
PRIX	1.00000	*	1	3.31217	47.3	47.3
QUANTITE	1.00000	*	2	2.61662	37.4	84.7
ALCOOL	1.00000	*	3	.57780	8.3	93.0
PRESTIGE	1.00000	*	4	.23840	3.4	96.4
COULEUR	1.00000	*	5	.13526	1.9	98.3
AROME	1.00000	*	6	.08297	1.2	99.5
GOUT	1.00000	*	7	.03678	.5	100.0

L'utilisation de la sous-commande `/ PLOT = EIGEN` provoque évidemment l'impression d'un graphique correspondant à la courbe d'accumulation de variance (« scree test ») de Cattell. Vous voudrez probablement inspecter ce graphique pour décider si le nombre de composantes extraites en fonction du critère de Cattell correspond au même nombre qu'indiqué par le critère de Kaiser (eigenvalue > 1.0). Aussi, vous vous êtes peut-être donné la peine de procéder à l'analyse parallèle de Horn; dans ce cas, vous pourriez superposer sur ce graphique la courbe des valeurs propres auxquelles vous vous attendez si vos données ne comportent aucune composante réelle (données générées de façon aléatoire). SPSS ne permet pas de calculer ces valeurs, mais j'ai déjà mentionné qu'il est facile de télécharger le petit programme développé à cet effet par Kaufman et Dunlap (2000).

J'ai déjà indiqué que SPSS applique par défaut le critère de Kaiser concernant le nombre de composantes à extraire; c'est ce qui explique que vous trouverez sous la courbe d'accumulation de variance une mention à l'effet que deux composantes seront extraites : « PC extracted 2 factors ». Il n'en tient qu'à vous de modifier le nombre de composantes à extraire si vous jugez que le critère de Kaiser ne devrait pas être appliqué. Utilisez à cet effet la sous-commande `/ CRITERIA = FACTORS (n)` et relancez la tâche SPSS.

```
            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

PC    extracted   2 factors.
```

La matrice suivante générée par SPSS est la matrice des poids factoriels (« factor loadings »). Elle contient les coefficients permettant d'exprimer chacune des variables en fonction des composantes extraites. Ainsi, la motivation basée sur le prix payé pour une bière peut être représentée par l'équation suivante :

$$\text{Prix} = .54984\ C_1 + .73611\ C_2$$

Alors que la motivation impliquant le goût d'une bière correspond à :

$$\text{Goût} = .71439\ C_1 - .6432\ C_2$$

Plusieurs observations intéressantes découlent de ces équations. Ainsi, on peut dire qu'un changement d'une unité de C_1 entraînera un changement correspondant de .54984 unités dans le prix, alors qu'un changement d'une unité de C_2 provoquera .73611 unités de changement dans le prix. Il s'agit là de l'interprétation classique d'une équation de régression.

Par ailleurs, puisque les diverses composantes extraites sont orthogonales, c'est-à-dire indépendantes les unes des autres, les poids factoriels sont aussi interprétables comme étant des coefficients de corrélation entre les variables et les composantes. On peut donc dire qu'il y a une corrélation de .54984 entre la composante C_1 et la variable Prix, ou encore qu'il y a 30.23% de variance commune ($.54984^2$) entre ces deux scores. De la même manière on en arrive à dire qu'il existe 54.18% de variance partagée ($.73611^2$) entre le Prix et la composante C_2. Les corrélations (et les variances communes) entre les variables et les composantes sont utiles pour nous aider à définir ou à cerner les composantes extraites. Nous y reviendrons dans un instant.

Les statistiques finales. -- L'encadré suivant présente les statistiques finales tenant compte du nombre de composantes extraites. Ici, c'est le critère de Kaiser qui a déterminé que deux composantes seraient conservées.

Variable	Communality	*	Factor	Eigenvalue	Pct of Var	Cum Pct
		*				
PRIX	.84417	*	1	3.31217	47.3	47.3
QUANTITE	.89739	*	2	2.61662	37.4	84.7
ALCOOL	.88793	*				
PRESTIGE	.54289	*				
COULEUR	.91418	*				
AROME	.91818	*				
GOUT	.92406	*				

Comme nous l'avons mentionné précédemment (dans la section des statistiques initiales), il faut bien observer que nous sommes en présence de deux tableaux présentés côte à côte par SPSS. La portion de droite nous informe que la première composante C_1 expliquera 47.3% de la variance totale des variables, alors que la composante C_2 ajoutera un autre 37.4%. Au total, notre ACP permettra donc d'expliquer 84.7% de la variance présente dans nos données à l'aide de deux composantes indépendantes.

La portion de gauche du tableau des statistiques finales nous donne une information très utile sur chacune des variables participant à l'analyse. On y retrouve la **proportion de variance commune** entre chaque variable et l'ensemble des composantes retenues dans la solution finale. Ainsi, on peut constater que 84.4% de la variance de la variable Prix est explicable à l'aide des deux composantes extraites. Notez que la variance commune (.84417) correspond à la somme des carrés des poids factoriels examinés précédemment : $(.54984^2) + (.73611^2)$. L'inspection de ces valeurs est importante car elle peut nous indiquer assez facilement les variables qui ont une variance unique, non partagée par l'ensemble des autres variables. Par exemple, on voit ici que la motivation liée au Prestige (.54289) se démarque de l'ensemble des autres motivations dans la détermination du comportement du consommateur de bière; on peut même affirmer que 45.7% de la variance du Prestige (1.0 - .54289) est de la variance unique, non expliquée par les deux composantes extraites.

L'étape de rotation. -- Comme nous l'avons mentionné précédemment, la décision à prendre sur le type de rotation à effectuer n'est pas facile. Nous examinerons d'abord les résultats d'une rotation **orthogonale** de type **VARIMAX**. Ce type de rotation a pour effet de diminuer la généralité de la première composante principale. Elle simplifie la structure de la solution en maximisant la variance des composantes. Généralement cela entraîne une redistribution des pondérations de façon telle que certaines variables seront fortement corrélées avec une dimension, alors que d'autres variables obtiendront des pondérations négligeables. L'encadré suivant présente les pondérations après rotation Varimax.

Rotated Factor Matrix:

	Factor 1	Factor 2
PRIX	-.06023	**.91681**
QUANTITE	.06814	**.94485**
ALCOOL	.01963	**.94210**
PRESTIGE	-.50777	-.53391
COULEUR	**.95429**	.05917
AROME	**.95813**	.01273
GOUT	**.96096**	-.02487

L'examen de la matrice des poids factoriels après rotation permet de constater facilement que la première composante est définie par les motivations reliées à la couleur, l'arôme et le goût de la bière consommée. La deuxième composante, quant à elle, se définit en termes de prix, de quantité et de taux d'alcool. Je vous laisse le soin de déterminer si ces deux composantes correspondent à votre perception des motivations reliées à la consommation du houblon. Quant à moi, il me semble qu'elles mettent en évidence une première dimension reliée au comportement du « dégustateur », alors que la deuxième serait plutôt reliée au comportement du « buveur ».

Cette première solution assume que les composantes « dégustateur » et « buveur » sont indépendantes l'une de l'autre, mais nous ne sommes pas en mesure de vérifier directement la validité de ce postulat. Une rotation oblique permettrait de voir plus clair sur cette question, puisqu'elle n'assumerait pas l'indépendance des deux dimensions.

Les résultats de la rotation oblique sont plus complexes parce que, précisément, les composantes peuvent être intercorrélées rendant alors leur interprétation plus difficile. Trois matrices devront être examinées. L'encadré suivant présente la matrice de pondérations après rotation de type oblique. Ces valeurs correspondent aux coefficients de régression lorsque l'on tente d'expliquer les variables à l'aide des différentes composantes comme variables prévisionnelles. Par exemple, on peut dire que la pondération de la composante C_1 est de .9688 pour expliquer la motivation reliée au goût d'une bière, quant les autres composantes sont contrôlées (ici la composante C_2). On voit bien que cette solution oblique reproduit le même patron que ce que nous avions obtenu dans la solution orthogonale : la composante C_1 est associée au goût, à l'arôme et à la couleur de la bière, alors que la composante C2 est associée au prix, au volume et au taux d'alcool. Cependant, comme les deux composantes sont peut-être corrélées entre elles, les pondérations apparaissant dans la « pattern matrix » **ne peuvent pas** être interprétés comme des coefficients de corrélation.

Pattern Matrix:

	Factor 1	Factor 2
PRIX	-.12152	.92680
QUANTITE	.00582	.94652
ALCOOL	-.04282	.94696
PRESTIGE	-.47562	-.50380
COULEUR	.95659	-.00361
AROME	.96354	-.05061
GOUT	.96888	-.08864

C'est la matrice identifiée dans SPSS comme une « structure matrix » qui présente les coefficients de corrélation entre les variables et les composantes. Dans le cas présent, cette matrice met en évidence la même structure des composantes, mais il peut arriver que les matrices « pattern » et « structure » diffèrent l'une de l'autre.

```
Structure Matrix:

              Factor  1     Factor  2

PRIX            .00007        .91086
QUANTITE        .12999        .94729
ALCOOL          .08141        .94134
PRESTIGE       -.54171       -.56619
COULEUR         .95612        .12188
AROME           .95690        .07579
GOUT            .95725        .03846
```

Finalement, SPSS produit la matrice de corrélation entre les composantes extraites. L'information disponible dans ce tableau nous permet de constater que la corrélation entre la composante C_1 (dégustateur) et la composante C_2 (buveur) est très faible (.13119). Tenant compte de cette information, il serait justifié de revenir à une solution orthogonale et de présenter uniquement les résultats de la rotation Varimax.

```
Factor Correlation Matrix:

              Factor  1    Factor  2

Factor  1      1.00000
Factor  2        .13119      1.00000
```

Remarques finales

L'analyse en composantes principales est une technique parsemée d'embûches. Nous avons vu qu'elle comporte une série de décisions critiques portant sur les propriétés des variables soumises à l'analyse, les propriétés de la matrice d'intercorrélation, le nombre de composantes à extraire, le type de rotation à utiliser, etc. Si vous décidez de pousser plus à fond l'utilisation de cet outil d'analyse vous devrez sans doute consulter des sources documentaires plus complètes; elles sont heureusement nombreuses. Toutefois, en portant

attention aux quelques points suivants, vous devriez être en mesure d'exploiter efficacement cette technique idéale de réduction des données.

- Assurez-vous d'avoir un nombre suffisant de participants à votre analyse. Il est généralement risqué de procéder à une ACP avec un n < 100. Plusieurs autres auteurs (p. ex., Grimm & Yarnold, 1995) mentionnent d'envisager un ratio de 5 à 10 fois plus de participants que de variables analysées. Par contre, les opinions plus récentes à cet effet soulignent qu'il n'y a pas de règle absolue puisque le nombre de participants doit se déterminer en fonction de la communalité des variables utilisées et de la détermination des composantes obtenues. Wuensch (2001) mentionne qu'une solution pourrait être tout à fait acceptable avec un n beaucoup inférieur à 100, si les variables ont des communalités élevées (>.6) et que les composantes possèdent plusieurs pondérations élevées (>.8).

- Assurez-vous que la matrice de corrélation analysée est adéquate, qu'elle n'est pas singulière et qu'elle n'est pas une matrice d'identité.

- Assurez-vous de prendre les bonnes décisions concernant le nombre de composantes à extraire. Ne vous contentez pas de l'option automatique programmée dans SPSS.

- N'acceptez pas sans réflexion et sans jugement critique la solution orthogonale proposée par SPSS.

Références

Carey, M. P., Faulstich, M. E., Gresham, F. M., Ruggiero, L., & Enyart, P. (1987). Children's depression inventory : Construct and discriminant validity across clinical and nonreferred (control) populations. *Journal of Consulting and Clinical Psychology*, *55*, 755-761.

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

Grimm, L. G., & Yarnold, P. R. (1995). *Reading and understanding multivariate statistics*. Washington, DC: APA.

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

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*, 31-36.

Kaufman, J. D., & Dunlap, W. P. (2000). Determining the number of factors to retain: A program for parallel analysis. *Behavior Research Methods, Instruments, & Computers*, *32*, 389-385.

Levine, J. B., Green, C., & Million, T. (1986). *Separation Individuation Test of Adolescence (SITA)*. Princeton, NJ : Educational Testing Service, TC019234, Set W.

Ossenkopp, K.-P., & Mazmanian, D. S. (1985). Some behavioral factors related to the effects of cold-restraint stress in rats: A factor analytic-multiple regression approach. *Physiology and Behavior*, *34*, 935-941.

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

Saintonge, S. & Lachance, L. (1995), Validation d'une adaptation canadienne-française du test de séparation-individuation à l'adolescence. *Revue québécoise de psychologie*, *16*, 199-218.

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

Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under- and over-extraction on principal axis factor analysis with varimax rotation. *Psychological Methods*, *1*, 354-365.

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