Analyse en Facteurs : présentation et comparaison des logiciels SAS, SPAD et SPSS

Marie Chavent, Vanessa Kuentz, Jérôme Saracco

Analyse en Facteurs : présentation et comparaison des logiciels SAS, SPAD et SPSS

Marie Chavent[1], Vanessa Kuentz[1], Jérôme Saracco[1,2]

[1] Universités Bordeaux 1 et 2,
IMB, UMR CNRS 5251,
351 Cours de la Libération, 33405 Talence Cedex, France
vanessa.kuentz,marie.chavent@math.u-bordeaux1.fr

[2] Université Montesquieu - Bordeaux 4,
GREThA, UMR CNRS 5113,
Avenue Léon Duguit, 33608 Pessac Cedex, France
jerome.saracco@u-bordeaux4.fr

Abstract In data analysis, factorial methods are essential. These techniques can be used as an end in themselves, seeking to highlight underlying common factors in a group of variables. They can also be used as input to another analysis. Then, they consist in data dimension reduction and operate by replacing the original variables, sometimes highly correlated, by a smaller number of linearly independent variables. Factor Analysis (F.A.) is one possible method for quantitative data. This article aims at presenting in a synthetic way the F.A. model, rarely described in French books, but frequent in the Anglo-Saxon literature, and often available in softwares. The presentation of the estimation techniques for the F.A. model enables to establish the existing connection between Principal Component Analysis (P.C.A.) and F.A. The usefulness of rotation techniques, which can facilitate the interpretation of the results, will also be shown. An application on crime data of American cities will be carried out and will allow to describe the results provided by three of the most used statistical softwares : SAS, SPAD and SPSS. Then it will help to clarify the vocabulary, sometimes confused for the user.

Keywords : Factor Analysis, Principal Component Analysis, Singular Value Decomposition, Rotation.

Résumé En analyse des données, les méthodes factorielles sont fondamentales. Ces techniques peuvent être utilisées comme but en soi, il s'agit alors de faire ressortir des facteurs sous-jacents communs à un groupe de variables. Elles peuvent également constituer une étape préalable à d'autres études. Elles consistent alors à réduire la dimension des données en remplaçant les variables d'origine, qui peuvent être corrélées, par un plus petit nombre de variables linéairement indépendantes. Lorsque les données sont quantitatives, l'Analyse en Facteurs (A.F.) est une des méthodes possibles. L'objectif de cet article est de dresser une présentation synthétique du modèle d'A.F., peu développé dans les manuels francophones, mais fréquent dans la littérature anglo-saxonne, et souvent présent dans les logiciels statistiques. La

présentation des techniques d'estimation du modèle d'A.F. permet d'établir le lien existant entre l'Analyse en Composantes Principales (A.C.P.) et l'A.F. Il s'agit également de montrer l'utilité des techniques de rotation, qui peuvent faciliter l'interprétation des résultats. Un exemple d'application sur des données de criminalité de villes américaines permet enfin de décrire les résultats fournis par trois des logiciels statistiques les plus utilisés : SAS, SPAD et SPSS, et ainsi de clarifier le vocabulaire, parfois confus pour l'utilisateur.

Mots-clés : Analyse en Facteurs, Analyse en Composantes Principales, Décomposition en Valeurs Singulières, Rotation.

1 Introduction

L'A.F. trouve son origine en psychométrie lorsqu'en 1904, Spearman développe une théorie psychologique selon laquelle l'esprit humain s'explique par un facteur commun à tous les individus et par plusieurs facteurs spécifiques à chacun. Ce modèle est généralisé pour plusieurs facteurs communs par Garnett en 1919. De nombreuses applications sont alors réalisées pour déterminer un nombre relativement faible de tests qui permettraient de décrire l'esprit humain de façon aussi complète que possible.

Ainsi, l'A.F. vise à écrire chaque variable aléatoire du problème en fonction de facteurs sous-jacents communs à toutes les variables, et d'un facteur spécifique ou unique à la variable aléatoire considérée. Il repose sur différentes hypothèses dont principalement la non corrélation des facteurs communs. Différentes méthodes d'estimation existent, les plus courantes sont l'estimation via les composantes principales, la méthode du facteur principal et le maximum de vraisemblance. L'estimation du modèle d'A.F. via l'A.C.P. ne garantit pas que les hypothèses du modèle soient vérifiées. Cependant cette technique est la plus utilisée car elle fournit souvent une approximation convenable.

Cet article met également en lumière un point essentiel de l'A.F. : le choix du nombre q de facteurs communs. Différents critères empiriques et théoriques existent pour le choisir. Nous insisterons sur le fait que ces règles sont une aide partielle qui ne doit pas se substituer à une interprétation rigoureuse des résultats. Notons que l'enjeu de ce choix est majeur car la qualité des résultats en dépend.

Suite à l'estimation du modèle d'A.F., la lecture des résultats peut s'avérer délicate. Les facteurs obtenus peuvent être difficiles à interpréter, sembler ne pas avoir d'intérêt pour l'étude, ou ne pas exliquer le phénomène considéré, etc. Des résultats sont pourtant parfois présents, mais leur lecture n'est pas directe et intuitive. L'utilisateur peut alors passer à côté de résultats importants. Une rotation orthogonale des facteurs peut aider dans cette phase. La justification de la possibilité d'effectuer une rotation provient de la non-unicité de la solution du modèle d'A.F. Nous verrons de plus que la rotation est possible en A.C.P. à condition d'effectuer convenablement la transformation. Bien que les techniques de rotation peuvent faciliter de façon significative la lecture des résultats, elles sont peu présentées dans les ouvrages francophones, contrairement à leurs voisins anglo-saxons. L'utilité de la rotation des facteurs sera mise en exergue sur une application concernant la criminalité de seize villes américaines (données issues de *U.S. Statistical Abstract*, 1970).

Enfin, l'estimation du modèle peut se faire à l'aide de logiciels statistiques, comme SAS, SPAD et SPSS. Le vocabulaire employé diffère d'un logiciel à l'autre et peut rapidement devenir source de confusion. L'exemple d'application précise ce vocabulaire et pourra ainsi aider les utilisateurs dans la lecture des sorties numériques des logiciels.

Le présent article s'articule autour de cinq parties. Le modèle d'A.F. est présenté à la section 2. L'estimation des paramètres du modèle est ensuite décrite à la section 3. Les techniques de rotation des facteurs, facilitant la détection de groupes de variables corrélées, sont présentées à la section 4. Enfin, à la section 5, une application de ce modèle d'A.F. est réalisée sur des données de criminalité dans différentes villes des Etats-Unis et permet de comparer les résultats fournis par les trois logiciels statistiques SAS, SPAD et SPSS.

2 Le modèle d'A.F.

Soit $\mathbf{x} = (x^1, x^2, \ldots, x^p)'$ un vecteur aléatoire de $\mathbb{R}^p$ d'espérance $\mu \in \mathbb{R}^p$. On note $\tilde{\mathbf{x}} = \mathbf{x} - \mu$ la version centrée de $\mathbf{x}$.

Le modèle d'A.F. s'écrit :

$$\underset{(p\times 1)}{\tilde{\mathbf{x}}} = \underset{(p\times q)(q\times 1)}{A_q\mathbf{f}} + \underset{(p\times 1)}{\mathbf{e}} \tag{1}$$

où :

- A_q est une matrice $(p\times q)$ de coefficients a_j^α, $j = 1, \ldots, p$, $\alpha = 1, \ldots, q$ ("loadings" en anglais). Elle est appelée matrice de saturation ("factor loadings matrix" ou "factor pattern matrix").
- $\mathbf{f} = (f^1, \ldots, f^q)'$ est un vecteur aléatoire de $\mathbb{R}^q$, composé des q facteurs communs ("common factors") aux p variables aléatoires $\tilde{x}^1, \tilde{x}^2, \ldots, \tilde{x}^p$.
- $\mathbf{e} = (e^1, \ldots, e^p)'$ est un vecteur aléatoire centré de $\mathbb{R}^p$, composé des p facteurs spécifiques (ou uniques) ("unique factors") à chaque variable $\tilde{x}^j, j = 1, \ldots, p$.

Il découle de (1) et de $\mathbb{E}(\mathbf{e}) = 0$ la propriété suivante :

$$\mathbb{E}(\mathbf{f}) = 0. \tag{2}$$

Pour tout $j = 1, \ldots, p$, on a :

$$\tilde{x}^j = \sum_{\alpha=1}^{q} a_j^\alpha f^\alpha + e^j. \tag{3}$$

Chaque variable $\tilde{x}^j$ s'écrit comme la somme d'une combinaison linéaire de facteurs $f^1, \ldots, f^q$ communs à toutes les variables $\tilde{x}^1, \ldots, \tilde{x}^p$ et d'un facteur e^j spécifique à la variable considérée $\tilde{x}^j$.

On insiste sur le fait que les facteurs communs $f^1, \ldots, f^q$ sont aléatoires. Ainsi, le modèle d'A.F. est souvent désigné comme un modèle à effets aléatoires ou modèle structurel (Baccini et Besse, 2005).

Le modèle (1) repose sur plusieurs hypothèses.

(H_1) : $\mathbb{E}(\mathbf{f}\mathbf{f}') = I_q$, où I_q est la matrice identité $(q \times q)$.

(H_2) : $\mathbb{E}(\mathbf{e}\mathbf{e}') = \Xi$, où $\Xi = diag(\xi^j, j = 1, \ldots, p)$.
(H_3) : $\mathbb{E}(\mathbf{e}\mathbf{f}') = 0$.

L'hypothèse (H_1) signifie que les facteurs communs $f^\alpha, \alpha = 1, \ldots, q$, sont non corrélés et de variance 1. Cette hypothèse de non corrélation des facteurs s'explique par le fait que l'on souhaite exprimer les variables aléatoires $\tilde{x}^j$ en fonction du plus petit nombre de facteurs possible, et donc éviter des redondances.

L'hypothèse (H_2) signifie que les facteurs uniques $e^j, j = 1, \ldots, p$, ne sont pas corrélés. Ils expriment pour chaque variable la part non expliquée par les facteurs communs. Ils ont chacun une variance spécifique ξ^j.

L'hypothèse (H_3) traduit le fait que chaque variable $e^j, j = 1, \ldots, p$, traduit la part spécifique à la variable $\tilde{x}^j$ qui n'a pu être exprimée par les facteurs communs $f^\alpha, \alpha = 1, \ldots, q$, donc les variables e^j et f^α, ne sont pas corrélées.

On note Σ la matrice de variance covariance de $\mathbf{x}$. On déduit du modèle (1) que :

$$\begin{aligned} \mathbb{E}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}') &= A_q \mathbb{E}(\mathbf{f}\mathbf{f}') A_q' + \mathbb{E}(\mathbf{e}\mathbf{e}') \text{ et donc} \\ \Sigma &= A_q A_q' + \Xi. \end{aligned} \tag{4}$$

L'équation (4) est appelée modèle de structure de covariance.
D'après (1) ou (3), on peut écrire pour tout $j = 1, \ldots, p$:

$$\begin{aligned} \mathbb{V}(x^j) &= (a_j^1)^2 + (a_j^2)^2 + \ldots + (a_j^q)^2 + \xi^j \\ &= \sum_{\alpha=1}^{q} (a_j^\alpha)^2 + \xi^j \\ &= h_j^2 + \xi^j. \end{aligned} \tag{5}$$

De même, pour $j \neq k$:

$$cov(x^j, x^k) = \sum_{\alpha=1}^{q} a_j^\alpha a_k^\alpha + 0. \tag{6}$$

On voit ainsi que les covariances des variables aléatoires $x^j, j = 1, \ldots, p$, sont complètement reconstituées par la matrice de saturation A_q tandis que les variances se décomposent en une part due aux facteurs communs, appelée communalité ou variance commune, et une part due aux facteurs spécifiques, appelée variance spécifique ou résiduelle.

On remarque également que A_q est la matrice des covariances entre les variables aléatoires $x^j, j = 1, \ldots, p$, et les facteurs communs $f^\alpha, \alpha = 1, \ldots, q$. En effet :

$$\begin{aligned} cov(\mathbf{x}, \mathbf{f}) = \mathbb{E}(\mathbf{x}\mathbf{f}') &= \mathbb{E}((A_q\mathbf{f} + \mathbf{e} + \mu)\mathbf{f}') \\ &= A_q \mathbb{E}(\mathbf{f}\mathbf{f}') + \mathbb{E}(\mathbf{e}\mathbf{f}') + \mu \mathbb{E}(\mathbf{f}') \\ &= A_q. \end{aligned} \tag{7}$$

On travaille maintenant sur les variables standardisées, c'est-à-dire que $\tilde{\mathbf{x}}$ correspond au vecteur $\mathbf{x}$ centré réduit : $\tilde{\mathbf{x}} = \Sigma^{-1/2}(\mathbf{x} - \mu)$.

Dans ce cas, la matrice A_q devient la matrice des corrélations linéaires entre les variables x^j et les facteurs f^α, et l'équation (4) s'écrit :

$$\Upsilon = A_q A_q' + \Xi \tag{8}$$

où Υ est la matrice de corrélation linéaire de $\mathbf{x}$.

De façon analogue à (5), on a :

$$1 = h_j^2 + \xi^j. \tag{9}$$

Dans la suite de cet article, nous considérons que le vecteur $\tilde{\mathbf{x}}$ correspond au vecteur $\mathbf{x}$ centré réduit.

3 Estimation du modèle

On veut estimer A_q et $\mathbf{f}$ dans le modèle (1). Rigoureusement, on ne devrait pas parler d'estimation pour $\mathbf{f}$ car il s'agit d'un vecteur aléatoire, on va donc obtenir une réalisation et non une estimation de $\mathbf{f}$. Nous nous conformerons cependant à cet abus de langage, fréquent dans la littérature.

Pour cela, on dispose d'un échantillon $\{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ de n réalisations indépendantes et identiquement distribuées du vecteur aléatoire $\mathbf{x}$ de $\mathbb{R}^p$.

D'après (1), on peut écrire pour tout $i = 1, \ldots, n$:

$$\tilde{\mathbf{x}}_\mathbf{i} = A_q \mathbf{f_i} + \mathbf{e_i}. \tag{10}$$

On note :
- $\tilde{X}$ la matrice $(n \times p)$ des données centrées réduites.
- F_q la matrice $(n \times q)$ correspondant aux n réalisations des q facteurs communs. Elle est appelée matrice des scores des facteurs communs ("factor scores matrix").
- E_q la matrice $(n \times p)$ des erreurs spécifiques.

Le modèle d'A.F. sur échantillon s'écrit alors :

$$\underset{(n \times p)}{\tilde{X}} = \underset{(n \times q)(q \times p)}{F_q A_q'} + \underset{(n \times p)}{E_q.} \tag{11}$$

Nous présentons ici trois méthodes d'estimation de A_q et F_q. Pour toute méthode d'estimation, il faut ensuite choisir le nombre q (avec $q \leq p$) de facteurs communs que l'on retient. Quelques critères pour le choisir sont discutés dans la section 3.4.

3.1 Estimation du modèle via les composantes principales

Cette technique utilise l'A.C.P. comme méthode d'estimation du modèle d'A.F. Nous rappelons donc dans un premier temps le principe de l'A.C.P., puis nous expliquons comment cette méthode est utilisée pour estimer le modèle d'A.F.

3.1.1 Présentation de l'A.C.P.

L'A.C.P est proposée pour la première fois par Pearson en 1901, elle est ensuite intégrée à la statistique mathématique par Hotelling en 1933. L'A.C.P. peut être considérée selon différents points de vue. La présentation la plus fréquente dans la littérature francophone est géométrique. L'A.C.P est alors vue comme une technique visant à représenter de façon optimale des données, selon certains critères géométriques et algébriques. Le lecteur pourra se reporter à l'ouvrage de Lebart et al. (1997). L'A.C.P peut être considérée sur un plan probabiliste, elle est alors un cas particulier du modèle d'A.F. où les variances spécifiques sont nulles ou égales, voir par exemple Tipping et Bishop (1999). Dans cet article, nous adopterons une présentation de l'A.C.P. qui nous permettra de faire le lien avec l'A.F.

L'A.C.P présente deux variantes, elle peut être réalisée à partir des données centrées ou des données centrées réduites. Dans le premier cas, on parle d'A.C.P. non normée ou A.C.P sur matrice des covariances. Dans le second cas, on parle d'A.C.P normée ou A.C.P. sur matrice des corrélations. Nous présentons dans cet article l'A.C.P. normée, variante la plus utilisée.

L'objectif de l'analyse du nuage des individus $\{x_1, \ldots, x_n\}$ en A.C.P. est de déterminer q nouvelles variables $\psi^1, \ldots, \psi^q$ avec $q \leq p$, permettant de résumer "au mieux" les p variables $\tilde{x}^1, \ldots, \tilde{x}^p$. Ces q nouvelles variables sont appelées les composantes principales des individus. Elles sont définies comme des combinaisons linéaires des p variables $\tilde{x}^1, \ldots, \tilde{x}^p$. On a donc, pour $\alpha = 1, \ldots, q$:

$$\psi^\alpha = v_1^\alpha \tilde{x}^1 + \ldots + v_p^\alpha \tilde{x}^p = \tilde{X} v^\alpha. \tag{12}$$

On suppose que l'espace $\mathbb{R}^n$ est muni de la métrique M, matrice de dimension $(n \times n)$, avec $M = diag(1/\sqrt{m}, \ldots, 1/\sqrt{m})$, où $m = n$ ou $m = n - 1$ selon l'estimateur de la variance choisi.

On veut que ces composantes soient de variance maximale et deux à deux orthogonales. Par construction, les colonnes de $\tilde{X}$ sont centrées et donc les composantes principales le sont aussi. On a donc :

$$\mathbb{V}(\psi^\alpha) = (\psi^\alpha)' M \psi^\alpha = (v^\alpha)' R v^\alpha \tag{13}$$

où $R = \tilde{X}' M \tilde{X}$ est la matrice des corrélations empiriques entre les variables initiales $x^1, \ldots, x^p$.

En ajoutant la contrainte $(v^\alpha)'(v^\alpha) = 1$, on démontre, voir par exemple Lebart et al. (1997), que pour $\alpha = 1, \ldots, q$, v^α est le vecteur propre associé à la $\alpha^{\text{ème}}$ plus grande valeur propre de la matrice des corrélations R.

On construit ainsi la matrice Ψ_q dont les colonnes sont les composantes principales des individus $\psi^\alpha, \alpha = 1, \ldots, q$:

$$\Psi_q = \tilde{X} V_q \tag{14}$$

où V_q est la matrice $(p \times q)$ dont les colonnes sont les vecteurs propres $v^\alpha, \alpha = 1, \ldots, q$, associés aux q plus grandes valeurs propres de la matrice R.

3.1.2 Estimation du modèle d'A.F.

L'A.C.P. peut être utilisée comme méthode d'estimation du modèle d'A.F. Le lien entre l'A.C.P. et l'A.F. s'obtient facilement à partir de la décomposition en valeurs singulières (D.V.S.) de la matrice $Z = M\tilde{X}$.

On note r (avec $r \leq p < n$) le rang de la matrice Z et on écrit sa D.V.S. :

$$Z = U\Lambda V' \tag{15}$$

où :
- $\Lambda = diag(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$ des valeurs singulières des matrices ZZ' et $Z'Z$ rangées par ordre décroissant ($\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \ldots \geq \sqrt{\lambda_r} > 0$).
- U est la matrice orthonormée $(n \times r)$ dont les colonnes sont les vecteurs propres de ZZ' associés aux r valeurs propres.
- V est la matrice orthonormée $(p \times r)$ dont les colonnes sont les vecteurs propres de $Z'Z$ associés aux r valeurs propres.

On a donc :

$$\tilde{X} = M^{-1}Z = M^{-1}U\Lambda V'. \tag{16}$$

On note U_q, Λ_q et V_q les matrices contenant respectivement les q premières colonnes de U, Λ et V.

• **Avec q = r.**
Pour se ramener au modèle d'A.F. (11), on pose :

$$\hat{F}_q = M^{-1}U_q \tag{17}$$
$$\hat{A}_q = V_q\Lambda_q. \tag{18}$$

On note que dans ce cas $\hat{E}_q = 0$.

Comme $U_q'U_q = I_r$, on montre que :

$$\hat{A}_q = Z'U_q = \tilde{X}'M^{-1}U_q. \tag{19}$$

Cette écriture est utilisée pour démontrer que les éléments de la matrice $\hat{A}_q$, notés $\hat{a}_j^\alpha$, sont les corrélations empiriques entre les variables x^j et les facteurs f^α (détails en annexe 7.1).

Comme $V_qV_q' = I_p$, on montre également que :

$$\hat{F}_q = \tilde{X}\underbrace{V_q\Lambda_q^{-1}}_{V_q^*}. \tag{20}$$

Cette écriture de $\hat{F}_q$ en fonction de $\tilde{X}$ fait ainsi apparaitre la matrice V_q^* des coefficients des scores des facteurs communs, calculée par certains logiciels statistiques.

• **Avec q < r.**
En ne retenant que les vecteurs propres associés aux q plus grandes valeurs propres, on a l'approximation de $\tilde{X}$ suivante :

$$\tilde{X} = \hat{F}_q\hat{A}_q + \hat{E}_q \tag{21}$$

où :

– $\hat{F}_q$ contient les q premières colonnes de $\hat{F}_q$ définie dans (17).
– $\hat{A}_q$ contient les q premières colonnes de $\hat{A}_q$ définie dans (18).
– $\hat{E}_q$ est la matrice des erreurs associée à cette approximation.

Avec cette méthode d'estimation, on montre facilement (voir les détails en annexe 7.1) que les facteurs communs estimés possèdent les bonnes propriétés mais que les hypothèses du modèle ne sont pas nécessairement toutes vérifiées.

3.1.3 Lien avec l'A.C.P.

Les facteurs communs estimés par cette méthode correspondent aux composantes principales des individus (trouvées en A.C.P.) standardisées. En effet, d'après les égalités (14) et (20), on voit que :

$$\hat{F}_q = \Psi_q \Lambda_q^{-1} \tag{22}$$

De plus, la matrice de saturations $\hat{A}_q$ est égale à la matrice des composantes principales des variables. En effet, si on présente l'A.C.P. d'un point de vue géométrique, on réalise généralement non seulement l'analyse des points-individus, comme présenté ici, mais également celle des points-variables. On montre que ces composantes correspondent aux corrélations entre les variables x^j et les facteurs f^α, et donc aux saturations (voir l'ouvrage de Lebart et al., 1997).

3.1.4 Quelques éléments de vocabulaire

On peut introduire, à partir de ces premiers résultats, le vocabulaire utilisé en A.C.P. et en A.F. (tableau 1).

TAB. 1 – Quelques éléments de vocabulaire en A.F. et A.C.P.

	Matrices	Français	Anglais
ACP	Ψ_q	Composantes principales	Principal component scores
	V_q	Coefficient des composantes principales	Principal component scoring coefficients
AF	F_q	Facteurs communs	Factor scores ou Standardized principal component scores
	V_q^*	Coefficients des facteurs communs	Factor scoring coefficients ou Standardized principal component scoring coefficients

3.2 Méthode du facteur principal

A partir de l'échantillon $\{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$, on calcule la matrice des corrélations empiriques définie par :

$$R = \tilde{X}' M \tilde{X} \tag{23}$$

L'équation (8) du modèle de structure de covariance sur échantillon s'écrit alors :

$$R = \hat{A}_q \hat{A}_q' + \hat{\Xi} \tag{24}$$

Il faut donc déterminer $\hat{A}_q$ et $\hat{\Xi}$.

Pour cela, la méthode du facteur principal estime Ξ (en fait $\Upsilon - \Xi$) et factorise $R - \hat{\Xi}$ pour obtenir $\hat{A}_q \hat{A}'_q$ en utilisant les valeurs propres et vecteurs propres de $R - \hat{\Xi}$.

• **Estimation de Ξ.**
D'après l'équation (9), un estimateur de $\Upsilon - \Xi$ est donné par :

$$R - \hat{\Xi} = \begin{pmatrix} \hat{h}_1^2 & r_{12} & \dots & r_{1p} \\ r_{21} & \hat{h}_2^2 & \dots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \dots & \hat{h}_p^2 \end{pmatrix}$$

où $\hat{h}_j^2$ est l'estimation de la $j^{\text{ème}}$ communalité définie par : $\hat{h}_j^2 = 1 - \hat{\xi}^j$.

D'après (5), la communalité h_j^2 traduit la part commune entre x^j et les $p-1$ variables restantes. Ainsi, une estimation courante pour la communalité h_j^2 est R_j^2, le coefficient de corrélation multiple entre x^j et les $p-1$ variables restantes.

Ainsi,

$$\hat{h}_j^2 = R_j^2 = 1 - \frac{1}{r^{jj}} \tag{25}$$

où r^{jj} est le $j^{\text{ème}}$ élément diagonal de R^{-1}.

Pour effectuer ces estimations, R doit être régulière. Si R est singulière, on utilise pour $\hat{h}_j^2$ la valeur absolue (ou le carré) de la plus grande corrélation de x^j avec les $p-1$ autres variables. Un autre moyen de remédier à cette singularité est de remarquer que, puisque R est singulière, cela signifie qu'il existe des combinaisons linéaires des variables. On peut donc supprimer les redondances et rendre ainsi R inversible.

Notons t le rang de la matrice $R - \hat{\Xi}$.

• **Avec $\mathbf{q = t}$.**

- Estimation de $\mathbf{A_q}$.
On écrit la décomposition spectrale de la matrice $R - \hat{\Xi}$:

$$R - \hat{\Xi} = CDC' = (CD^{1/2})(CD^{1/2})' = \hat{A}_q \hat{A}'_q \tag{26}$$

où :
- $D = diag(\theta_1, \theta_2, \dots, \theta_t)$ des valeurs propres non nulles de $R - \hat{\Xi}$.
- C est la matrice orthonormale dont les colonnes sont les vecteurs propres normés de $R - \hat{\Xi}$ associés aux t valeurs propres non nulles.

- Estimation de $\mathbf{F_q}$.
Après avoir estimé A_q, il faut "estimer" F_q. Une méthode possible est de choisir l'estimateur linéaire $\hat{\mathbf{f}} = L\tilde{\mathbf{x}}$ qui minimise l'erreur quadratique moyenne :

$$E[\|\hat{\mathbf{f}} - \mathbf{f}\|^2] = E[\|L\tilde{\mathbf{x}} - \mathbf{f}\|^2] = E[\|LA_q\mathbf{f} + L\mathbf{e} - \mathbf{f}\|^2]. \tag{27}$$

Seber (1984) montre que (27) est égale à :

$$\text{trace}(L'L\Sigma) - 2\text{trace}(LA_q) + q. \tag{28}$$

En différenciant par rapport à A_q, on obtient :

$$\begin{aligned} 2L\Sigma - 2A'_q &= 0 \\ L &= A'_q\Sigma^{-1} \end{aligned} \tag{29}$$

et donc :

$$\hat{\mathbf{f}} = A'_q\Sigma^{-1}\tilde{\mathbf{x}}. \tag{30}$$

En remplaçant A_q par son estimateur (26), et Σ par la matrice de variance covariance empirique S, on obtient :

$$\begin{aligned} \hat{F}_q &= \tilde{X}S^{-1}\hat{A}_q \\ &= \tilde{X}R^{-1}\hat{A}_q \text{ car on travaille avec la matrice } \tilde{X} \text{ centrée réduite.} \end{aligned} \tag{31}$$

Remarque. En développant le calcul de $\Sigma^{-1} = (A_qA'_q + \Xi)^{-1}$ dans (30), on montre (voir l'ouvrage de Seber, 1984) que $\hat{\mathbf{f}}$ est l'estimateur "ridge" de $\mathbf{f}$:

$$\hat{\mathbf{f}} = (I_q + A'_q\Xi^{-1}A_q)^{-1}A'_q\Xi^{-1}\tilde{\mathbf{x}}. \tag{32}$$

• Avec q < t.
Afin de retenir seulement q facteurs communs dans le modèle d'A.F., on ne conserve que les q, premières colonnes des matrices $\hat{A}_q$ et $\hat{F}_q$ définies respectivement dans (26) et (31).

• Itération de la méthode.
Cette méthode du facteur principal peut facilement être itérée afin d'améliorer l'estimation de A_q. Après avoir estimé A_q à partir de (26), nous pouvons obtenir une nouvelle estimation de la communalité en utilisant (5) :

$$\hat{h}_j^2 = \sum_{\alpha=1}^{q} (\hat{a}_j^\alpha)^2.$$

Ces valeurs sont alors insérées dans la diagonale de $R - \hat{\Xi}$, ce qui nous permet d'obtenir une nouvelle estimation de A_q à partir de la décomposition spectrale de la nouvelle matrice $R - \hat{\Xi}$, comme dans l'équation (26). Ce processus est alors itéré jusqu'à ce que les estimations de la communalité se stabilisent.

Cependant, un défaut majeur de la méthode itérée est qu'elle ne converge pas toujours. De plus, lors de ces itérations, $\hat{h}_j^2$ peut devenir supérieur à 1, ce qui implique $\hat{\xi}^j < 0$. Or, ceci est impossible car on ne peut pas avoir une variance spécifique estimée négative. Ce problème est connu sous le nom de *Heywood case* (Heywood, 1931).

3.3 Maximum de vraisemblance

On suppose que l'échantillon $\{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ est issu d'une loi multi-normale $N_p(\mu, \Sigma)$. Alors $(n-1)S$ suit la loi de Wishart $W_p(n-1, \Sigma)$ et la log-vraisemblance de l'échantillon est donnée par :

$$\log L(A_q, \Xi) = c - \frac{1}{2}(n-1)(\ln|\Sigma| + \text{trace}(\Sigma^{-1}S)) \tag{33}$$

où c est une constante et $|M|$ désigne le déterminant de M.

Les paramètres A_q et Ξ vont alors pouvoir être estimés en maximisant $\log L(A_q, \Xi)$, sous la contrainte $\Sigma = A_q A_q' + \Xi$ avec Ξ matrice diagonale. La condition suivante : $A_q' \Xi^{-1} A_q$ diagonale, est souvent rajoutée afin d'avoir une solution unique.

L'équation du maximum de vraisemblance n'a pas de solution analytique, la résolution se fait donc numériquement par itérations successives. Cependant, cette méthode ne converge pas toujours. De plus, des cas de variances spécifiques estimées négatives peuvent là encore se produire (*Heywood case*).

Notons qu'avec cette méthode, les facteurs communs obtenus $\hat{f}^\alpha$ ne sont pas forcément ordonnés par variance expliquée décroissante comme avec la méthode des composantes principales et la méthode du facteur principal.

Après avoir estimé A_q et Ξ, il faut estimer la matrice F_q des scores des facteurs communs. Pour cela, on utilise souvent la méthode des moindres carrés généralisés :

$$\hat{F}_q = (\hat{A}_q' \hat{\Xi}^{-1} \hat{A}_q)^{-1} \hat{A}_q' \hat{\Xi}^{-1} \tilde{X}. \tag{34}$$

On ne retient ensuite que les q, avec $q \leq p$, premières colonnes des matrices $\hat{F}_q$ et $\hat{A}_q$.
Pour de plus amples détails sur cette méthode d'estimation, le lecteur pourra se référer à l'ouvrage de Seber (1984).

3.4 Choix du nombre de facteurs

Comme dans toute méthode factorielle, une étape importante de l'A.F. est le choix du nombre q de facteurs communs. La qualité des estimations du modèle dépend de q. En effet, si q est trop grand, certains facteurs spécifiques vont être mélangés aux facteurs communs. A l'inverse si q est trop petit, des facteurs communs importants risquent d'être oubliés. Différents critères théoriques et empiriques peuvent être utilisés pour choisir q.

Voici deux critères théoriques reposant sur la normalité de l'échantillon :

• Critère 1.
Ce critère consiste à déterminer si les $(p-k)$ dernières valeurs propres de la matrice de covariance Σ sont significativement différentes entre elles. On fait pour cela l'hypothèse que les n observations sont les réalisations d'un vecteur aléatoire gaussien dont les $(p-k)$ dernières valeurs propres $\lambda_{k+1}, \ldots, \lambda_p$ de la matrice Σ sont égales. Sous cette hypothèse,

la moyenne arithmétique m_a des $(p-k)$ dernières valeurs propres doit être peu différente de la moyenne géométrique m_g. On définit :

$$T_1 = \left(n - \frac{2p+11}{6}\right)(p-k)\log\left(\frac{m_a}{m_g}\right). \tag{35}$$

Sous H_0, on peut montrer que T_1 suit une loi du Khi-deux à $v_1 = \frac{(p-k+2)(p-k-1)}{2}$ degrés de liberté.

On rejette donc H_0 au seuil de signification α si l'inégalité suivante est vérifiée :

$$T_1 > \chi^2_{v_1,1-\alpha} \tag{36}$$

où $\chi^2_{v_1,1-\alpha}$ est le fractile d'ordre $(1-\alpha)$ de la loi du Khi-deux à v_1 degrés de liberté.

Certains auteurs (voir par exemple Bouveyron, 2006) soulignent le fait que ce critère surestime très souvent le nombre de facteurs à retenir.

• Critère 2.
Ce critère est utilisé lorsque la méthode d'estimation du modèle d'A.F. est le maximum de vraisemblance.
On désire tester l'hypothèse que q_0 est le bon nombre de facteurs communs. Les hypothèses sont donc : $H_0 : \Sigma = A_{q_0}(A_{q_0})' + \Xi$, où A_{q_0} est de dimension $(p \times q_0)$ contre $H_1 : \Sigma = A_q A_q' + \Xi$, où A_q est de dimension $(p \times q)$ avec $q > q_0$.
La statistique de test est :

$$T_2 = \left(n - \frac{2p+4q_0+11}{6}\right)\log\left(\frac{|\hat{A}_{q_0}\hat{A}'_{q_0} + \hat{\Xi}|}{|S|}\right) \tag{37}$$

où $\hat{A}_{q_0}$ et $\hat{\Xi}$ sont les estimateurs du maximum de vraisemblance de A_q et Ξ, obtenus avec q_0 facteurs communs.

Sous H_0, on peut montrer que T_2 suit une loi du Khi-deux à $v_2 = \frac{1}{2}[(p-q_0)^2 - p - q_0]$ degrés de liberté.

On rejette donc H_0 au seuil de signification α si la condition (36) est vérifiée (où l'on aura préalablement substitué v_2 à v_1). Si H_0 est rejetée, cela signifie que le nombre q_0 de facteurs communs est trop petit.

En pratique, on commence souvent avec $q_0 = 1$, et on ajoute des facteurs jusqu'à trouver la valeur q pour laquelle H_0 est vérifiée. Ainsi, le risque associé à la procédure pour trouver le bon nombre de facteurs q_0 est supérieur à α, du fait de la multiplicité des tests.

Il faut noter que cette technique est souvent utilisée pour fixer la borne supérieure du nombre de facteurs. En effet, on peut trouver dans la littérature, voir par exemple Rencher (2002), que lorsque le nombre d'observations est grand, cette méthode a tendance à surestimer le nombre de facteurs communs.

Différentes règles empiriques peuvent également être utilisées pour choisir q. Dans la définition des critères ci-dessous, $\lambda_i, i = 1, \ldots, p$, fait référence aux valeurs propres de R

(ou S) ou bien $R-\hat{\Xi}$ (ou $S-\hat{\Xi}$), selon que la méthode d'estimation utilisée est la méthode des composantes principales ou du facteur principal. Voici quelques exemples de critères empiriques.

• Pourcentage de variance expliquée.
On choisit q tel que le pourcentage de variance expliquée par les q facteurs soit supérieur ou égal à un seuil fixé par l'utilisateur. L'appréciation de ce pourcentage doit tenir compte du nombre de variables p et du nombre d'observations n. En effet, un pourcentage de 10% peut être considéré comme élevé pour $p = 100$ variables et au contraire, faible pour $p = 10$. Notre expérience nous a montré que ce critère a souvent tendance à surestimer le nombre de facteurs q.

• Le test du coude.
On utilise le test du coude de Cattell, ou *scree-test*, basé sur l'analyse des différences consécutives entre les valeurs propres. On calcule les différences premières :

$$\epsilon_i = \lambda_i - \lambda_{i+1}$$

puis les différences secondes :

$$\delta_i = \epsilon_i - \epsilon_{i+1}.$$

On retient alors les valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_k, \lambda_{k+1}$ tels que $\delta_1, \ldots, \delta_k$ soient tous positifs.
Visuellement, ce critère revient à détecter un coude dans le graphe de l'éboulis des valeurs propres. En pratique, la détection graphique de ce coude peut se révéler difficile.

• La règle de Kaiser.
Sachant que la variance expliquée par un facteur $\hat{f}^\alpha$ est λ_α, il s'agit de retenir les facteurs dont la variance expliquée dépasse la moyenne de la variance totale expliquée : $\overline{\lambda} = \frac{\sum_{i=1}^{p} \lambda_i}{p}$. Pour la matrice de corrélation R, on a : $\overline{\lambda} = 1$. Cette valeur 1 peut aussi être vue comme la variance de chaque variable x^j et on retient donc un facteur s'il explique au moins autant de variance qu'une variable toute seule.

On insiste sur le fait que ces critères ne doivent pas se substituer à une analyse approfondie de l'interprétation des facteurs. Il est indispensable d'examiner l'information apportée par un facteur, et ainsi juger de sa pertinence et de son intérêt quant aux objectifs de l'étude. L'utilisateur retiendra, par exemple, un facteur dont la part de variance expliquée est faible, mais dont l'intérêt est significatif pour la problématique traitée. Au contraire, il pourra rejeter un facteur qui possède une part de variance expliquée élevée, mais qui n'aide pas à comprendre le phénomène étudié. Ainsi, on peut utiliser ces critères comme valeur initiale du nombre q_0 de facteurs, puis au vu de l'interprétation des résultats et des objectifs de l'étude, on peut augmenter ou diminuer ce nombre q_0 afin de trouver une interprétation des résultats satisfaisante.

3.5 Choix de la méthode d'estimation

On trouve dans la littérature (voir par exemple Rencher, 2002) que les solutions obtenues avec la méthode des composantes principales et la méthode du facteur principal (itérée ou non) sont très proches lorsque l'une des deux conditions suivantes est vérifiée :

– Les corrélations entre les variables $x^j, j = 1, \ldots, p$, sont élevées.
– Le nombre de variables p est grand.

Cependant, il est important de noter que la méthode d'estimation la plus utilisée est celle des composantes principales. C'est une technique qui fournit une approximation convenable de la solution et qui est facile à mettre en oeuvre. Ainsi, c'est la méthode utilisée par défaut lorsqu'on estime un modèle d'A.F. sous les logiciels SAS et SPSS. Enfin, contrairement aux deux autres techniques, elle ne présente pas le problème de *Heywood case*.

4 La rotation des facteurs

Dans cette section, nous allons présenter les motivations de la rotation orthogonale des facteurs estimés, puis nous montrerons que cette rotation est justifiée car elle conserve les propriétés des facteurs. Quelques critères permettant la mise en place d'une rotation optimale seront ensuite discutés. Enfin, nous montrerons brièvement que la rotation est possible en A.C.P., comme en A.F., à condition d'effectuer convenablement la transformation.

4.1 Motivations

Après avoir estimé le modèle d'A.F., on peut vouloir interpréter les facteurs communs obtenus en détectant des groupes de variables corrélées aux différents facteurs. La matrice de saturation A_q, représentant les corrélations entre les variables x^j et les facteurs communs f^α, il s'agit de faire apparaître des variables fortement corrélées à un même facteur.

Il est donc souhaitable que pour chaque colonne de A_q les valeurs soient proches soit de 0, soit de 1 et qu'il n'y ait ainsi pas de valeur intermédiaire. Cela permet alors d'associer clairement des variables à un facteur. Il faut également s'assurer que sur chaque ligne de A_q, il n'y aura qu'une seule valeur proche de 1. En effet, si la corrélation entre une variable x^j et un facteur f^α est proche de 1, les corrélations de cette variable avec les facteurs restants doivent être proches de 0, car les facteurs sont orthogonaux entre eux. Cela se traduit par la condition d'orthonormalité de la matrice de transformation T. Ainsi, chaque variable x^j ne pourra être parfaitement associée qu'à un seul facteur f^α.

Cependant, l'estimation $\hat{A}_q$ trouvée ne présente pas toujours une telle structure. Afin de se rapprocher de cette situation, il est possible de réaliser une rotation des facteurs. La justification de la possibilité de faire une rotation provient de la non-unicité de la solution du modèle d'A.F. Ainsi, il s'agit de choisir la solution optimale du point de vue de l'interprétation des résultats.

Remarque. Il existe des rotations qui ne conservent pas la propriété d'orthogonalité des facteurs communs. Ce type de rotation dites obliques ne sera pas abordé dans cet article. Pour plus de détails, le lecteur pourra se reporter à l'ouvrage de Rencher (2002).

4.2 Justifications de la rotation

La solution du modèle d'A.F. n'est pas unique. En effet, soit T une matrice orthonormale de dimension $(q \times q)$. On peut écrire :

$$\begin{aligned}\tilde{X} &= \hat{F}_q \hat{A}'_q + \hat{E}_q \\ &= \hat{F}_q T T' \hat{A}'_q + \hat{E}_q \\ &= \hat{G}_q \hat{B}'_q + \hat{E}_q \text{ avec } \hat{G}_q = \hat{F}_q T \text{ et } \hat{B}_q = \hat{A}_q T. \end{aligned} \tag{38}$$

Ainsi $\hat{G}_q$ est l'estimation de la matrice des facteurs après la rotation et $\hat{B}_q$ est l'estimation de la matrice de saturation après la rotation.
La transformation orthogonale entraîne une rotation "rigide" des q axes définis par les facteurs communs, c'est-à-dire que les q nouveaux axes restent perpendiculaires après la rotation.

Propriété :
Les facteurs communs g^α et les saturations $\hat{b}_j^\alpha$ vérifient toujours les propriétés et hypothèses du modèle d'A.F. après la rotation.

La démonstration est disponible en annexe 7.2.

On montre en particulier que les saturations après rotation, $\hat{b}_j^\alpha$, sont toujours les corrélations des variables d'origine x^j aux facteurs après rotation, g^α.

Cependant, même si suite à la rotation, les communalités estimées sont inchangées et que l'on a :

$$\hat{h}_j^2 = \sum_{\alpha=1}^{q} (\hat{b}_j^\alpha)^2 = \sum_{\alpha=1}^{q} (\hat{a}_j^\alpha)^2, \tag{39}$$

la variance expliquée par chaque facteur f^α change lors de la rotation. En effet :

$$\begin{aligned}\sum_{j=1}^{p} (\hat{b}_j^\alpha)^2 &= \sum_{j=1}^{p} (\hat{a}_j t^\alpha)^2 \text{ où } t^\alpha \text{ est la } \alpha^{\text{ème}} \text{ colonne de } T \\ &= \sum_{j=1}^{p} (\sum_{k=1}^{q} \hat{a}_j^k t_k^\alpha)^2 \\ &\neq \sum_{j=1}^{p} (\hat{a}_j^\alpha)^2. \end{aligned} \tag{40}$$

Après la rotation, les facteurs ne sont donc plus forcément rangés par ordre de variance expliquée décroissante.

4.3 Comment faire la rotation ?

Afin d'effectuer la rotation, il faut déterminer la matrice T qui fournit la "meilleure" interprétation des résultats, c'est-à-dire telle que les éléments $\hat{b}_j^\alpha$ de la matrice $\hat{B}_q = \hat{A}_q T$ soient proches de 0 ou de 1, on parle de "structure simple" de $\hat{B}_q$. Différents critères existent, le plus utilisé est Varimax.

4.3.1 Varimax

Comme toutes les variables n'ont pas la même commmunalité, le critère Varimax est souvent appliqué sur les valeurs standardisées de $\hat{B}_q$, obtenues en divisant chaque ligne de la matrice $\hat{B}_q$ par $\hat{h}_j$. On travaille avec le carré de ces éléments, $(\hat{b}_j^\alpha/\hat{h}_j)^2$, afin de se ramener à des valeurs comprises entre 0 et 1. On note $\hat{B}_q^*$ cette matrice. On veut que ses éléments soient aussi proches que possible de 0 ou de 1. Pour cela, il faut maximiser la variance empirique de chaque colonne de $\hat{B}_q^*$ afin de donner plus de poids aux valeurs extrêmes 0 et 1.

Ceci équivaut à maximiser la somme sur l'ensemble des q facteurs des variances empiriques de chaque colonne de $\hat{B}_q^*$, c'est-à-dire la quantité :

$$\sum_{\alpha=1}^{q}\left\{\frac{\sum_{j=1}^{p}((\hat{b}_j^\alpha)^2/\hat{h}_j^2)^2}{p}-\left(\frac{\sum_{j=1}^{p}((\hat{b}_j^\alpha)^2/\hat{h}_j^2)}{p}\right)^2\right\} \tag{41}$$

avec $\hat{b}_j^\alpha = \hat{a}_j t^\alpha$, $\hat{a}_j$ est la $j^{\text{ème}}$ ligne de la matrice $\hat{A}_q$, et t^α est la $\alpha^{\text{ème}}$ colonne de la matrice T.

La maximisation de la quantité (41) se fait donc de façon itérative, par rapport à $t^\alpha, \alpha = 1,\ldots,q$, sous la contrainte $t^\alpha(t^\alpha)' = 1$ et $t^k(t^l)' = 0$ pour $k \neq l$.

4.3.2 Quartimax

Le critère Quartimax maximise la somme des variances des éléments $(\hat{b}_j^\alpha)^2$ sur toute la matrice $\hat{B}_q$, c'est-à-dire la quantité :

$$\frac{\sum_{\alpha=1}^{q}\sum_{j=1}^{p}(\hat{b}_j^\alpha)^4}{pq}-\left(\frac{\sum_{\alpha=1}^{q}\sum_{j=1}^{p}(\hat{b}_j^\alpha)^2}{pq}\right)^2. \tag{42}$$

On trouve dans la littérature (Jobson, 1992) que ceci équivaut à maximiser la quantité $\sum_{\alpha=1}^{q}\sum_{j=1}^{p}(\hat{b}_j^\alpha)^4$. Il est de plus mentionné que cette méthode a tendance à produire un facteur commun général car elle maximise la variance des $(\hat{b}_j^\alpha)^2$ sur la totalité de la matrice de saturation $\hat{B}_q^*$ et non sur chaque colonne, comme le critère Varimax.

4.3.3 Orthomax

Le critère Orthomax est une généralisation des critères de rotation orthogonale. Il s'agit de maximiser la quantité :

$$\sum_{\alpha=1}^{q}\left\{\sum_{j=1}^{p}(\hat{b}_j^\alpha)^4-\frac{\delta}{p}\left(\sum_{j=1}^{p}(\hat{b}_j^\alpha)^2\right)^2\right\} \tag{43}$$

Pour $\delta = 0$ et $\delta = 1$, on retrouve respectivement le critère Quartimax et la version non standardisée de Varimax. Pour $\delta = 0.5$, le critère s'appelle Biquartimax et pour $\delta = \frac{q}{2}$, ce critère est connu sous le nom de Equamax.

4.4 Remarques sur la rotation et l'A.C.P.

La rotation en A.C.P est possible, comme en A.F., mais il faut être prudent et appliquer la transformation T aux bonnes matrices.

Si on applique la rotation directement sur les composantes principales Ψ_q, les nouvelles composantes après rotation $\Psi_q T$ ne sont pas nécessairement non corrélées. En effet :

$$(\Psi_q T)' M \Psi_q T = T' \Psi_q' M \Psi_q T = T' \Lambda^2 T. \tag{44}$$

On en déduit que $(\Psi_q T)' M \Psi_q T$ n'est pas forcément la matrice identité.

Afin d'effectuer convenablement une rotation en A.C.P. il faut donc introduire la matrice T au bon endroit dans l'écriture de $\tilde{X}$.

Il ne faut pas écrire :

$$\tilde{X} = \overbrace{M^{-1} U \Lambda}^{\Psi_q} \mathbf{TT'} V' \tag{45}$$

mais :

$$\tilde{X} = \overbrace{M^{-1} U}^{\hat{F}_q} \mathbf{TT'} \Lambda V'. \tag{46}$$

Ainsi les composantes obtenues après rotation correspondent en fait aux facteurs communs obtenus après rotation $\hat{G}_q = M^{-1} U T$. Nous avons déjà vérifié en annexe 7.2 que ces facteurs ne sont pas corrélés.

On comprend ainsi la raison pour laquelle les logiciels qui calculent les composantes principales ne proposent pas de rotation des composantes, celles-ci deviendraient en effet corrélées. A l'inverse, les logiciels qui construisent les facteurs communs proposent une rotation.

5 Un exemple d'application sur des données de criminalité

5.1 Problématique

• Données.
Dans cette exemple, nous étudions la criminalité de seize villes américaines, données étudiées par de nombreux auteurs dont Rencher (2002) (extraites de *U.S. Statistical Abstract*, 1970). Pour cela, sept types d'effractions sont relevés et un taux pour 100 000 habitants est calculé (tableau 2). L'objectif est de résumer la criminalité de ces villes à l'aide de facteurs communs. Nous allons étudier les résultats fournis par les logiciels

SAS, SPAD et SPSS. Nous choisissons d'estimer le modèle d'A.F. via les composantes principales.

Il faut noter que le logiciel SPAD utilise la définition de l'estimateur biaisé de l'écart-type ($m = n$), les logiciels SAS et SPSS utilise au contraire la définition de l'estimateur non biaisé de l'écart-type ($m = n - 1$). Il est cependant possible de préciser au logiciel SAS d'utiliser l'estimation biaisée de l'écart-type avec l'option "*vardef=n*".

TAB. 2 – Criminalité

Ville	Meurtre	Viol	Vol	Agression	Cambriolage	Vol avec effraction	Vol de voiture
Atlanta	16.5	24.8	106	147	1112	905	494
Boston	4.2	13.3	122	90	982	669	954
Chicago	11.6	24.7	340	242	808	609	645
Dallas	18.1	34.2	184	293	1668	901	602
Denver	6.9	41.5	173	191	1534	1368	780
Detroit	13	35.7	477	220	1566	1183	788
Hartford	2.5	8.8	68	103	1017	724	468
Honolulu	3.6	12.7	42	28	1457	1102	637
Houston	16.8	26.6	289	186	1509	787	697
Kansas City	10.8	43.2	255	226	1494	955	765
Los Angeles	9.7	51.8	286	355	1902	1386	862
New Orleans	10.3	39.7	266	283	1056	1036	776
New York	9.4	19.4	522	267	1674	1392	848
Portland	5	23	157	144	1530	1281	488
Tucson	5.1	22.9	85	148	1206	756	483
Washington	12.5	27.6	524	217	1496	1003	793

On note $X = (x_i^j), i = 1, \ldots, n, j = 1, \ldots, p$, la matrice des données présentées dans le tableau 2 avec $n = 16$ observations et $p = 7$ variables.

• **Choix de q.**
Les différents auteurs, qui ont étudiés ces données, ont conservé $q = 3$ facteurs communs (voir par exemple Rencher, 2002). Nous vérifions, par une étude approfondie, que la valeur 3 permet une bonne interprétation des résultats. Pour cela, nous comparons les valeurs de q proposées par les critères empiriques discutés dans la section 3.4 et examinons attentivement les valeurs propres de la matrice des corrélations R (figure 1). Dans cette étape, nous privilégions l'interprétation des facteurs et leur intérêt pour la problématique étudiée. Nous vérifions ainsi que la valeur 3 permet des interprétations intéressantes pour l'étude menée. Dans la suite de cet article, la valeur q est donc choisi égale à 3, pour l'A.C.P. comme pour l'A.F.

FIG. 1 – Valeurs propres de la matrice R

Eigenvalues of the Correlation Matrix

	Valeur propre	Différence	Proportion	Cumulée
1	3.46216658	2.13250001	0.4946	0.4946
2	1.32966657	0.37144677	0.1900	0.6845
3	0.95821980	0.34730456	0.1369	0.8214
4	0.61091525	0.25013602	0.0873	0.9087
5	0.36077923	0.19073695	0.0515	0.9602
6	0.17004228	0.06183198	0.0243	0.9845
7	0.10821030		0.0155	1.0000

5.2 Logiciel SAS

La première partie présente la procédure PRINCOMP qui réalise une A.C.P. sur les données. Dans un second temps, nous décrivons les résultats de la procédure FACTOR avec pour méthode d'estimation l'A.C.P. et ainsi nous comparons les résultats des deux procédures.

5.2.1 Procédure PRINCOMP

Le code SAS de la procédure PRINCOMP est présenté dans la figure 2. L'option "$n = 3$" permet de réduire l'affichage des résultats à 3 composantes principales.

FIG. 2 – Code SAS de la procédure PRINCOMP

```
proc PRINCOMP data=doncrim.crimrates n=3 outstat=loadACP out=comp;
var Meurtre Viol Vol Agression Cambriolage Vol_avec_effraction Vol_de_voiture;
run;
```

La procédure propose comme résultats la matrice des corrélations empiriques $R = \tilde{X}'M\tilde{X}$, ses valeurs propres λ_α et les vecteurs propres associés, c'est-à-dire la matrice V_q (figure 3). Cette matrice peut également être obtenue dans une table avec l'option "*outstat=loadACP*". Il s'agit en fait de la matrice des coefficients des composantes principales.

FIG. 3 – Sorties numériques de la procédure PRINCOMP

Correlation Matrix

	Meurtre	Viol	Vol	Agression	Cambriolage	Vol_avec_ effraction	Vol_de_ voiture
Meurtre	1.0000	0.4349	0.4374	0.5558	0.2318	-.0681	0.0630
Viol	0.4349	1.0000	0.3150	0.7722	0.4973	0.4574	0.3725
Vol	0.4374	0.3150	1.0000	0.6065	0.3402	0.3201	0.5433
Agression	0.5558	0.7722	0.6065	1.0000	0.4200	0.3425	0.3790
Cambriolage	0.2318	0.4973	0.3402	0.4200	1.0000	0.7592	0.2751
Vol_avec_effraction	-.0681	0.4574	0.3201	0.3425	0.7592	1.0000	0.3088
Vol_de_voiture	0.0630	0.3725	0.5433	0.3790	0.2751	0.3088	1.0000

Eigenvalues of the Correlation Matrix

	Valeur propre	Différence	Proportion	Cumulée
1	3.46216658	2.13250001	0.4946	0.4946
2	1.32966657	0.37144677	0.1900	0.6845
3	0.95821980		0.1369	0.8214

Eigenvectors

	Prin1	Prin2	Prin3
Meurtre	0.284077	0.601420	-.290666
Viol	0.434117	0.057951	-.280953
Vol	0.387340	0.193196	0.454238
Agression	0.458700	0.265038	-.090328
Cambriolage	0.388251	-.403606	-.293904
Vol_avec_effraction	0.346301	-.596641	-.125837
Vol_de_voiture	0.315820	-.092125	0.721022

FIG. 4 – Matrice V_q des coefficients des composantes principales

SCORE	Prin1	0.2840770255	0.4341165199	0.3873395476	0.4587002129	0.3882506994	0.346300896	0.3158201451
SCORE	Prin2	0.6014200757	0.0579510469	0.1931960648	0.2650381889	-0.403606485	-0.596640949	-0.092124862
SCORE	Prin3	-0.290666091	-0.280952874	0.4542381853	-0.090328373	-0.293904228	-0.125837414	0.7210217837

L'option "*out=comp*" permet d'obtenir dans une table appelée "*comp*", la matrice Ψ_q des composantes principales (figure 5).

FIG. 5 – Matrice Ψ_q des composantes principales

Prin1	Prin2	Prin3
-1.211983352	1.1808854874	-1.303102471
-2.133959083	-0.107709329	2.2119885684
-0.872935491	2.1520641374	0.7123636412
1.130005234	1.184421741	-1.571231827
0.9796205812	-1.48142512	-0.255612807
1.8426143897	0.077312182	0.4616195197
-3.344991804	-0.241562028	-0.108412417
-2.202267721	-1.901723853	-0.066729639
0.3019595472	1.1949108748	-0.242587142
1.0299148373	0.2190691351	-0.160284845
3.3434931322	-1.012810751	-0.486160635
0.7842181022	0.7215288423	0.356669141
1.9519465049	-0.895021258	1.213102309
-0.815483767	-1.590279827	-1.04171094
-2.086689049	-0.025286754	-0.716503761
1.3045379405	0.5256265198	0.996593304

5.2.2 Procédure FACTOR

La procédure SAS permettant d'estimer un modèle d'A.F. est FACTOR. Cette procédure nous propose différentes méthodes d'estimation dont la méthode des composantes principales que nous spécifions avec l'option "*method=prin*" (figure 6). L'option "*nfactors=3*" permet de fixer q.

FIG. 6 – Code SAS de la procédure FACTOR

```
proc FACTOR data=doncrim.crimrates method=prin nfactors=3 outstat=loadAF out=fact;
var Meurtre Viol Vol Agression Cambriolage Vol_avec_effraction Vol_de_voiture;
run;
```

La procédure FACTOR fournit comme la procédure PRINCOMP les valeurs propres λ_α de la matrice des corrélations empiriques $R = \tilde{X}'M\tilde{X}$.

En précisant l'option "*out=fact*", le logiciel calcule la réalisation des facteurs communs $f^\alpha, \alpha = 1, \ldots, q$, sur les n observations, c'est-à-dire la matrice $\hat{F}_q$ (figure 7).

FIG. 7 – Matrice $\hat{F}_q$ des facteurs communs

Factor1	Factor2	Factor3
-0.651362384	1.024085956	-1.331208242
-1.146864496	-0.093407542	2.2596975134
-0.469146166	1.8663102248	0.72772815
0.6073044668	1.0271526611	-1.605120706
0.5264824771	-1.28471954	-0.261125953
0.9902856339	0.0670465686	0.4715758913
-1.797715978	-0.209487103	-0.110750694
-1.183575955	-1.649210454	-0.068168887
0.162283657	1.0362490339	-0.247819347
0.5535123752	0.1899808466	-0.163741925
1.7969105388	-0.878328404	-0.496646318
0.4214663279	0.6257232916	0.3643619061
1.0490445493	-0.776179155	1.2392669249
-0.438269594	-1.379120376	-1.064178926
-1.121459921	-0.021929146	-0.731957565
0.701104468	0.4558331377	1.0180881778

Le logiciel calcule la matrice V_q^* des coefficients des facteurs communs. Dans les sorties du logiciel, elle est appelée "*standardized scoring coefficents*" (figure 8). Avec l'option "*outstat=loadAF*", on peut également obtenir cette matrice dans la table "*loadAF*".

FIG. 8 – Matrice V_q^* des coefficients des scores des facteurs communs

Standardized Scoring Coefficients

	Factor1	Factor2	Factor3
Meurtre	0.15267	0.52156	-0.29694
Viol	0.23331	0.05026	-0.28701
Vol	0.20817	0.16754	0.46404
Agression	0.24652	0.22985	-0.09228
Cambriolage	0.20866	-0.35002	-0.30024
Vol_avec_effraction	0.18611	-0.51742	-0.12855
Vol_de_voiture	0.16973	-0.07989	0.73657

SCORE	Factor1	0.1526729623	0.2333094518	0.2081698654	0.2465215911	0.2086595503	0.1861142539	0.1697328287
SCORE	Factor2	0.5215627254	0.0502562305	0.1675432367	0.2298460689	-0.350015084	-0.517418178	-0.079892402
SCORE	Factor3	-0.296935279	-0.287012564	0.4640353537	-0.092276607	-0.300243257	-0.12855152	0.7365730343

La matrice de saturation estimée $\hat{A}_q$ est présentée sous le nom de "factor pattern" (figure 9). Ses coefficients sont les corrélations des variables d'origine x^j aux facteurs communs f^α.

FIG. 9 – Matrice $\hat{A}_q$ de saturation

Factor Pattern

	Factor1	Factor2	Factor3
Meurtre	0.52858	0.69350	-0.28453
Viol	0.80776	0.06682	-0.27502
Vol	0.72072	0.22278	0.44465
Agression	0.85350	0.30562	-0.08842
Cambriolage	0.72241	-0.46540	-0.28770
Vol_avec_effraction	0.64436	-0.68799	-0.12318
Vol_de_voiture	0.58764	-0.10623	0.70580

En rajoutant l'option "*rotate=varimax*" dans le code présenté dans la figure 6, nous demandons au logiciel d'effectuer une rotation orthogonale avec le critère Varimax. La matrice T de transformation orthogonale, estimée selon ce critère, est présentée dans la figure 10.

FIG. 10 – Matrice T de transformation orthogonale

Orthogonal Transformation Matrix

	1	2	3
1	0.61059	0.60511	0.51090
2	0.69393	-0.71967	0.02304
3	-0.38163	-0.34046	0.85933

En rajoutant l'option "*out=factrotation*", on obtient la matrice $\hat{G}_q = \hat{F}_q T$ des facteurs communs après la rotation (figure 11).

La matrice de saturation après rotation, $\hat{B}_q = \hat{A}_q T$, est calculée et nommée "rotated factor pattern" (figure 12). Les valeurs de cette matrice sont les corrélations des variables d'origine x^j aux facteurs communs g^α après la rotation, donnés en figure 11.

On voit très clairement sur cet exemple que la rotation des "loadings" facilite la lecture des résultats. Les valeurs de $\hat{A}_q$ (figure 9) sont très dispersées et rendent difficile la détection de groupes de variables corrélées à un même facteur. Au contraire, la matrice $\hat{B}_q$ (figure 12) possède beaucoup plus de valeurs soit proches de 1, soit proches de 0. On peut ainsi associer clairement chaque variable à un facteur.

FIG. 11 – Matrice $\hat{G}_q$ des facteurs après la rotation

Factor1	Factor2	Factor3
0.8209573792	-0.677921774	-1.453131596
-1.627438853	-1.396103804	1.3537329652
0.7309213041	-1.874778112	0.4286739055
1.6961441603	0.1747618885	-1.045383689
-0.470394333	1.3320583558	0.0149855921
0.4712167452	0.390426783	0.9127247156
-1.200768377	-0.899349739	-1.018458959
-1.84110518	0.4939011466	-0.701275648
0.9127510435	-0.563183957	-0.10616891
0.5322902033	0.2539612555	0.146461291
0.6772028014	1.8885272871	0.4710274804
0.5525029634	-0.31933239	0.5428539416
-0.371019377	0.7714553414	1.5830136881
-0.818502468	1.0896243119	-1.17017204
-0.420633311	-0.413620834	-1.202456046
0.3558752995	-0.25042576	1.2435733079

FIG. 12 – Matrice $\hat{B}_q$ de saturation après la rotation

Rotated Factor Pattern

	Factor1	Factor2	Factor3
Meurtre	0.91257	-0.08237	0.04153
Viol	0.64453	0.53433	0.17789
Vol	0.42497	0.12440	0.75545
Agression	0.76696	0.32662	0.36712
Cambriolage	0.22793	0.87003	0.11113
Vol_avec_effraction	-0.03698	0.92698	0.20750
Vol_de_voiture	0.01574	0.19174	0.90429

On peut alors décrire les trois facteurs communs de la criminalité dans ces seize villes américaines. Le premier facteur fait référence aux crimes violents contre une personne : meurtre, viol et agression. Le second facteur se rapporte aux crimes en rapport avec la maison : cambriolage et vol avec effraction. Enfin le troisième facteur fait référence aux vols commis à l'extérieur : vol et vol de voiture.

5.3 Logiciel SPSS

Avec le logiciel SPSS, dans le menu "*Analyse* → *Factorisation* → *Analyse factorielle*", on peut choisir différentes méthodes d'extraction des facteurs en cliquant sur le bouton "*Extraction*" : Maximum de vraisemblance, Composantes principales, Factorisation en axes principaux, etc (figure 13).

En cliquant sur le bouton "*Facteur*"(figure 13), on peut demander au logiciel d'afficher l'estimation de la matrice $\hat{F}_q$ des scores des facteurs communs, et l'estimation de la matrice des coefficients des scores des facteurs communs V_q^*.

L'estimation de ces deux matrices $\hat{F}_q$ et V_q^* est présentée à la figure 14. Nous retrouvons les résultats de la procédure FACTOR de SAS, présentés aux figures 7 et 8.

Le logiciel propose comme résultats la matrice de saturation estimée, $\hat{A}_q$, appelée "matrice des composantes" (figure 15). On voit là l'erreur commise par le logiciel car le terme "matrice des composantes" est réservé à Ψ_q. Ce problème de vocabulaire provient peut-être d'une mauvaise traduction française de ce logiciel anglais.

Le logiciel nous propose également différentes rotations. En choisissant le critère Varimax,

FIG. 13 – Estimation du modèle d'A.F. avec SPSS

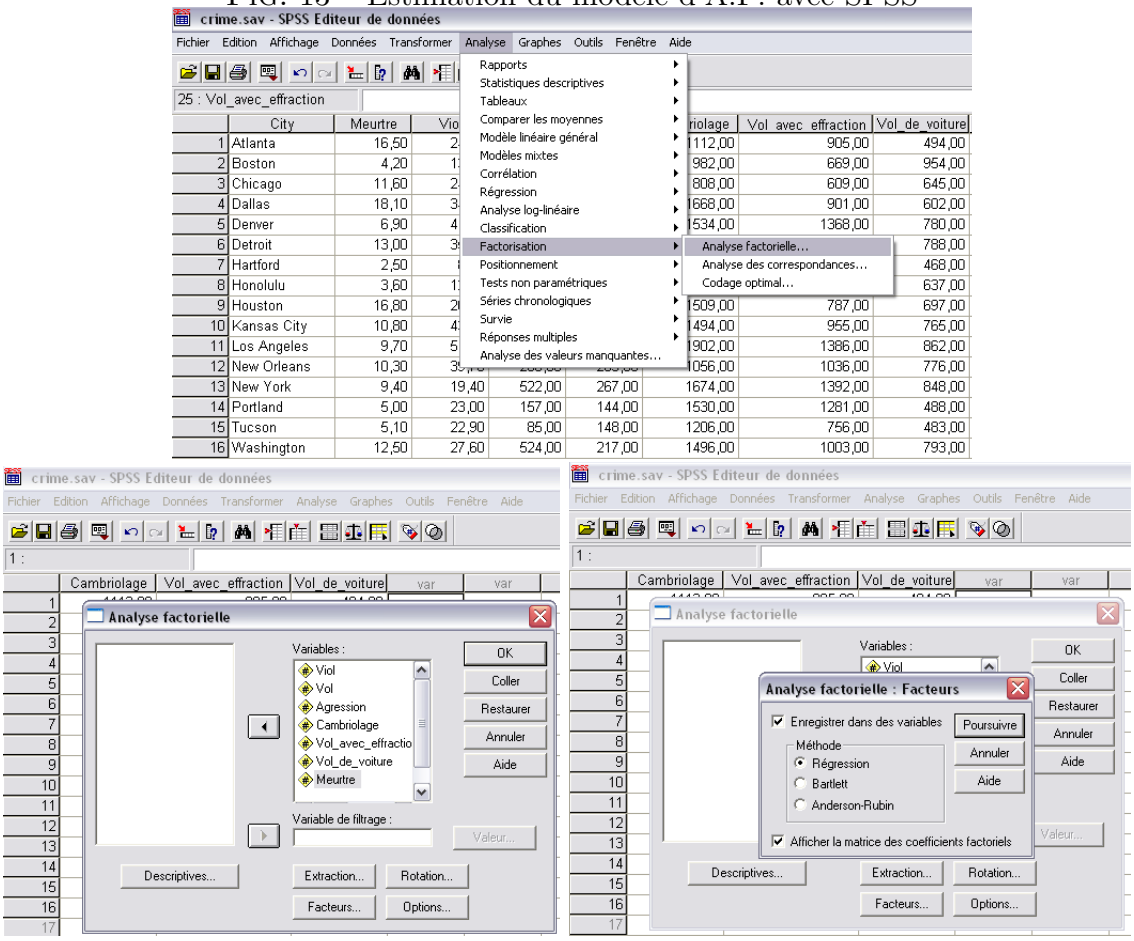

FIG. 14 – Matrice $\hat{F}_q$ et V_q^*

FAC1_1	FAC2_1	FAC3_1
-,65136	1,02409	-1,33121
-1,14686	-,09341	2,25970
-,46915	1,86631	,72773
,60730	1,02715	-1,60512
,52648	-1,28472	-,26113
,99029	,06705	,47158
-1,79772	-,20949	-,11075
-1,18358	-1,64921	-,06817
,16228	1,03625	-,24782
,55351	,18998	-,16374
1,79691	-,87833	-,49665
,42147	,62572	,36436
1,04904	-,77618	1,23927
-,43827	-1,37912	-1,06418
-1,12146	-,02193	-,73196
,70110	,45583	1,01809

Matrice des coefficients des coordonnées des composantes

	Composante		
	1	2	3
Viol	,233	,050	-,287
Vol	,208	,168	,464
Agression	,247	,230	-,092
Cambriolage	,209	-,350	-,300
Vol_avec_effraction	,186	-,517	-,129
Vol_de_voiture	,170	-,080	,737
Meurtre	,153	,522	-,297

Méthode d'extraction : Analyse en composantes principales.

Scores composante.

nous retrouvons les matrices $\hat{G}_q$ et $\hat{B}_q$ (figure 16) de la procédure FACTOR du logiciel SAS (figures 11 et 12).

FIG. 15 – Matrice $\hat{A}_q$ de saturation

Matrice des composantes[a]

	Composante		
	1	2	3
Meurtre	,529	,694	-,285
Viol	,808	,067	-,275
Vol	,721	,223	,445
Agression	,853	,306	-,088
Cambriolage	,722	-,465	-,288
Vol_avec_effraction	,644	-,688	-,123
Vol_de_voiture	,588	-,106	,706

Méthode d'extraction : Analyse en composantes principales.

a. 3 composantes extraites.

FIG. 16 – Matrice $\hat{G}_q$ et $\hat{B}_q$ après la rotation

FAC1_2	FAC2_2	FAC3_2
,82096	-,67793	-1,45313
-1,62745	-1,39610	1,35373
,73091	-1,87478	,42868
1,69615	,17476	-1,04538
-,47039	1,33206	,01498
,47122	,39043	,91273
-1,20077	-,89935	-1,01846
-1,84110	,49391	-,70128
,91275	-,56319	-,10617
,53229	,25396	,14646
,67721	1,88853	,47103
,55250	-,31933	,54286
-,37102	,77146	1,58301
-,81850	1,08963	-1,17017
-,42063	-,41362	-1,20246
,35587	-,25043	1,24357

Matrice des composantes après rotation[a]

	Composante		
	1	2	3
Viol	,645	,534	,178
Vol	,425	,124	,755
Agression	,767	,327	,367
Cambriolage	,228	,870	,111
Vol_avec_effraction	-,037	,927	,207
Vol_de_voiture	,016	,192	,904
Meurtre	,913	-,082	,042

Méthode d'extraction : Analyse en composantes principales.
Méthode de rotation : Varimax avec normalisation de Kaiser.

a. La rotation a convergé en 5 itérations.

5.4 Logiciel SPAD

Parmi les méthodes d'analyse factorielle du logiciel SPAD, l'A.C.P. est proposée mais l'A.F. n'est pas disponible. Afin d'effectuer une A.C.P., on insère la méthode des composantes principales appelée "COPRI" (figure 17).

Le logiciel réalise alors l'analyse des points-variables (voir l'ouvrage de Lebart et al., 1997) et propose comme résultats les coordonnées des variables sur les composantes calculées. Cette matrice est égale à la matrice des corrélations variable-facteur, il s'agit de l'estimation de la matrice de saturation, notée $\hat{A}_q$ (figure 18). Sur cette figure, on retrouve également la matrice des "anciens axes unitaires" qui correspond en fait à la matrice V_q des vecteurs propres de R.

Pour obtenir la matrice des composantes principales Ψ_q (figure 19) , il faut demander à SPAD, lors du paramétrage de la filière, d'afficher les résultats pour les individus.

FIG. 17 – Filière SPAD

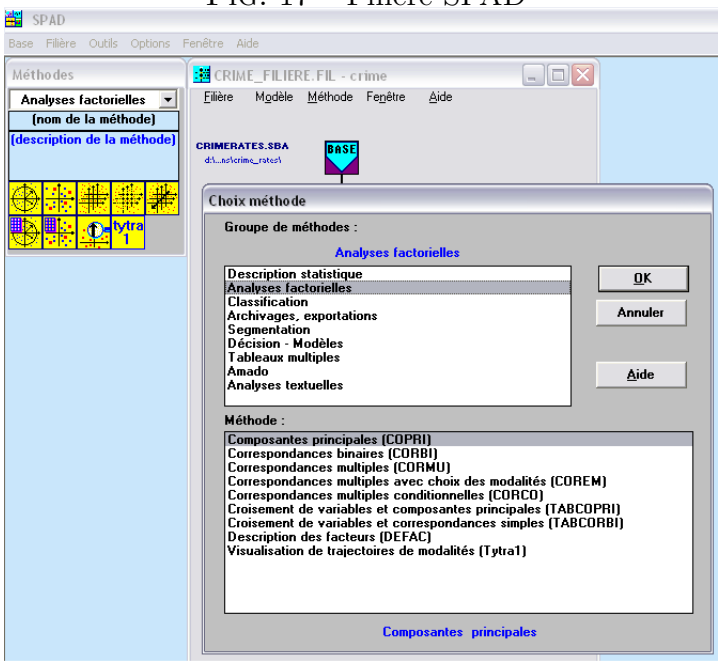

FIG. 18 – Matrice $\hat{A}_q$ de saturation

COORDONNEES DES VARIABLES SUR LES AXES 1 A 3
VARIABLES ACTIVES

VARIABLES	COORDONNEES					CORRELATIONS VARIABLE-FACTEUR					ANCIENS AXES UNITAIRES				
IDEN - LIBELLE COURT	1	2	3	0	0	1	2	3	0	0	1	2	3	0	0
C2 - Meurtre	0.53	-0.69	-0.28	0.00	0.00	0.53	-0.69	-0.28	0.00	0.00	0.28	-0.60	-0.29	0.00	0.00
C3 - Viol	0.81	-0.07	-0.28	0.00	0.00	0.81	-0.07	-0.28	0.00	0.00	0.43	-0.06	-0.28	0.00	0.00
C4 - Vol	0.72	-0.22	0.44	0.00	0.00	0.72	-0.22	0.44	0.00	0.00	0.39	-0.19	0.45	0.00	0.00
C5 - Agression	0.85	-0.31	-0.09	0.00	0.00	0.85	-0.31	-0.09	0.00	0.00	0.46	-0.27	-0.09	0.00	0.00
C6 - Cambriolage	0.72	0.47	-0.29	0.00	0.00	0.72	0.47	-0.29	0.00	0.00	0.39	0.40	-0.29	0.00	0.00
C7 - Vol_avec_effraction	0.64	0.69	-0.12	0.00	0.00	0.64	0.69	-0.12	0.00	0.00	0.35	0.60	-0.13	0.00	0.00
C8 - Vol_de_voiture	0.59	0.11	0.71	0.00	0.00	0.59	0.11	0.71	0.00	0.00	0.32	0.09	0.72	0.00	0.00

FIG. 19 – Matrice Ψ_q des composantes principales

INDIVIDUS			COORDONNEES				
IDENTIFICATEUR	P.REL	DIST0	1	2	3	0	0
01	6.25	6.03	-1.25	-1.22	-1.35	0.00	0.00
02	6.25	11.95	-2.20	0.11	2.28	0.00	0.00
03	6.25	7.18	-0.90	-2.22	0.74	0.00	0.00
04	6.25	6.43	1.17	-1.22	-1.62	0.00	0.00
05	6.25	4.60	1.01	1.53	-0.26	0.00	0.00
06	6.25	4.66	1.90	-0.08	0.48	0.00	0.00
07	6.25	12.69	-3.45	0.25	-0.11	0.00	0.00
08	6.25	9.77	-2.27	1.96	-0.07	0.00	0.00
09	6.25	3.26	0.31	-1.23	-0.25	0.00	0.00
10	6.25	2.29	1.06	-0.23	-0.17	0.00	0.00
11	6.25	14.73	3.45	1.05	-0.50	0.00	0.00
12	6.25	3.65	0.81	-0.75	0.37	0.00	0.00
13	6.25	9.14	2.02	0.92	1.25	0.00	0.00
14	6.25	5.31	-0.84	1.64	-1.08	0.00	0.00
15	6.25	5.88	-2.16	0.03	-0.74	0.00	0.00
16	6.25	4.42	1.35	-0.54	1.03	0.00	0.00

5.5 Tableau récapitulatif des différents résultats

Le tableau ci-dessous récapitule un certain nombre de résultats et permet de faire le lien entre les facteurs ou composantes, obtenus avec les trois logiciels. Pour faciliter la

lecture, les matrices ne sont pas indicées par q.

TAB. 3 – Tableau comparatif des différents résultats

	SAS **Proc** PRINCOMP	**SAS** **Proc** FACTOR **(estimation par A.C.P.)**	**SPAD**	**SPSS**
Méthode factorielle	A.C.P.	A.F.	A.C.P.	A.F.
Estimateur de la variance	non biaisé	non biaisé	biaisé	non biaisé
Facteurs/composantes	$\Psi_{sas} = \tilde{X}V$	$F_{sas} = \tilde{X}V\Lambda^{-1}$	$\Psi_{spad} = \tilde{X}V$	$F_{spss} = \tilde{X}V\Lambda^{-1}$
Norme au carré des facteurs	$(n-1)\lambda_\alpha$	$n-1$	$n\lambda_\alpha$	$n-1$
Variance des facteurs	λ_α	1	λ_α	1
Matrice de saturation ? Nom donné par le logiciel	non fournie	$\hat{A} = V\Lambda$ "factor pattern"	$\hat{A} = V\Lambda$ "corrélations variables facteurs"	$\hat{A} = V\Lambda$ "matrice des composantes"
Rotation possible ?	non	oui	non	oui
Lien	ψ^α_{sas}	$f^\alpha_{sas} = \frac{\psi^\alpha_{sas}}{\sqrt{\lambda_\alpha}}$	$\psi^\alpha_{spad} = \sqrt{\frac{n}{n-1}}\psi^\alpha_{sas}$	$f^\alpha_{spss} = f^\alpha_{sas}$

Pour mettre en place un modèle d'A.F., on peut donc utiliser la procédure FACTOR de SAS ou le logiciel SPSS.

Si on utilise la procédure PRINCOMP de SAS ou le logiciel SPAD, la méthode réalisée est une A.C.P. et il faut standardiser les composantes principales pour obtenir l'estimation des facteurs communs. Ceci peut se faire facilement sous le logiciel SAS, en spécifiant l'option "*standard*" dans le code de la procédure PRINCOMP (figure 2). On obtient alors la matrice des composantes principales standardisées présentée dans la figure (20), qui correspond bien à la matrice $\hat{F}_q$ de la procédure FACTOR (figure 7).

FIG. 20 – Matrice Ψ_q des composantes principales standardisées

Prin1	Prin2	Prin3
-0.651362384	1.024085956	-1.331208242
-1.146864496	-0.093407542	2.2596975134
-0.469146166	1.8663102248	0.72772815
0.6073044668	1.0271526611	-1.605120706
0.5264824771	-1.28471954	-0.261125953
0.9902856339	0.0670465686	0.4715758913
-1.797715978	-0.209487103	-0.110750694
-1.183575955	-1.649210454	-0.068168887
0.162283657	1.0362490339	-0.247819347
0.5535123752	0.1899808466	-0.163741925
1.7969105388	-0.878328404	-0.496646318
0.4214663279	0.6257232916	0.3643619061
1.0490445493	-0.776179155	1.2392669249
-0.438269594	-1.379120376	-1.064178926
-1.121459921	-0.021929146	-0.731957565
0.701104468	0.4558331377	1.0180881778

De plus, la procédure PRINCOMP de SAS ne fournit pas l'estimation de la matrice de saturation. Il faut la calculer : $\hat{A}_q = V_q\Lambda_q$. Cependant, si elle contient des valeurs très dispersées entre 0 et 1, il sera difficile d'associer des variables entre elles. Ce problème peut également se rencontrer avec le logiciel SPAD qui ne propose de rotation.

On peut cependant souligner deux avantages du logiciel SPAD par rapport à SAS et SPSS : l'interactivité et la possibilité de réaliser facilement des graphiques.

Ainsi, on voit l'avantage d'utiliser la procédure FACTOR de SAS ou le logiciel SPSS pour estimer le modèle d'A.F., car ils fournissent l'estimation de la matrice A_q de saturation et proposent une rotation des facteurs.

6 Conclusion

Le modèle d'A.F. est une méthode factorielle linéaire. Cette technique écrit un ensemble de p variables aléatoires comme une combinaison linéaire de q facteurs non corrélés, communs à toutes les variables, et de p facteurs spécifiques à chaque variable. L'ensemble de ces facteurs communs et uniques reproduit les covariances des variables aléatoires initiales. Ainsi, le modèle d'A.F. permet de résumer au "mieux" l'information contenue dans p variables aléatoires, ou de détecter des facteurs sous-jacents communs dans une problématique particulière. Comme toute méthode factorielle, le point stratégique du modèle d'A.F. réside dans le choix du nombre q de facteurs communs, difficulté à laquelle nous avons souhaité apporter une aide. Cet aide n'est que partielle car nous avons vu que seule une interprétation attentive des résultats et des objectifs de l'étude permet de répondre au problème du choix de q.

Après une présentation synthétique du modèle d'A.F., nous avons décrit les techniques d'estimation et nous avons vu que lorsqu'on estime le modèle d'A.F. via les composantes principales, cela revient à faire une A.C.P.

L'accent a ensuite été mis sur les techniques de rotation des facteurs, qui peuvent s'avérer très utiles. Nous avons montré que, contrairement à ce qu'on peut lire dans certains travaux, la rotation en A.C.P. est également possible à condition d'effectuer convenablement la transformation.

Une application numérique a ensuite été mise en place sur des données concernant la criminalité de villes américaines. Ainsi, l'estimation du modèle d'A.F. couplée à une rotation de type Varimax nous a permis de résumer la criminalité des villes américaines à l'aide de trois facteurs communs : les crimes violents contre la personne, les crimes en rapport avec la maison et les crimes commis à l'extérieur. De plus, cette application a permis de clarifier le vocabulaire utilisé par les logiciels SAS, SPAD et SPSS, réelle source de confusion. L'exemple accompagné de nombreuses illustrations pourra servir de guide, tant pour l'implémentation que pour la lecture des résultats numériques.

7 Annexes

7.1 Annexe 1 : Etude des propriétés des facteurs communs estimés par la méthode des composantes principales

- Par construction, les facteurs communs estimés sont centrés.

- L'hypothèse (H_1) est vérifiée car $\hat{F}_q' M \hat{F}_q = U_q' M^{-1} U_q = I_q$.

- L'hypothèse (H_2) n'est pas nécessairement vérifiée. En effet, cette méthode d'estimation de la matrice Ξ ne garantit pas que $\hat{\Xi} = \hat{E}_q \hat{E}_q'$ soit diagonale. Cependant, en pratique, les termes en dehors de la diagonale de la matrice $\hat{\Xi}$ sont souvent négligeables. Ainsi, la solution trouvée avec cette méthode est souvent une approximation convenable, ce qui explique que cette méthode d'estimation du modèle d'A.F. est très utilisée. On pourrait préconiser à l'utilisateur d'examiner attentivement la matrice $\hat{E}_q \hat{E}_q'$ et de recommencer

l'estimation du modèle avec une autre méthode si les valeurs en dehors de la diagonale de cette matrice sont trop grandes.

• L'hypothèse (H_3) est vérifiée. En effet, en ne retenant que les vecteurs propres associés aux q plus grandes valeurs propres, on a :

$$\tilde{X} = \underbrace{M^{-1}U_q}_{\hat{F}_q}\underbrace{\Lambda_q V_q'}_{\hat{A}_q'} + \hat{E}_q$$

où $\hat{E}_q = M^{-1}U_e\Lambda_e V_e'$, avec U_e, Λ_e et V_e les matrices contenant respectivement les $r-q$ dernières colonnes de U, Λ et V.

On a alors :

$$\begin{aligned} \hat{E}_q' M \hat{F}_q &= \hat{E}_q' U_q \\ &= V_e \Lambda_e U_e' M^{-1} U_q \\ &= 0 \end{aligned} \tag{47}$$

car la matrice U est orthonormée.

• On peut vérifier que les valeurs de la matrice $\hat{A}_q$, notés $\hat{a}_j^\alpha$ sont les corrélations empiriques entre les variables x^j et les facteurs f^α :

$$\begin{aligned} \hat{a}_j^\alpha &= \sum_{i=1}^n z_i^j u_i^\alpha \\ &= \sum_{i=1}^n z_i^j \frac{1}{\sqrt{m}} f_i^\alpha \\ &= \sum_{i=1}^n \left(\frac{x_i^j - \overline{x}^j}{\sqrt{m} s^j} \right) \left(\frac{f_i^\alpha - \overline{f}^\alpha}{\sqrt{m}\sqrt{var(f^\alpha)}} \right) \\ &= corr(x^j, f^\alpha) \end{aligned} \tag{48}$$

7.2 Annexe 2 : Démonstration de la propriété des facteurs et des "loadings" après rotation

• Les facteurs après rotation sont toujours centrés.

• L'hypothèse (H_1) est vérifiée car $\hat{G}_q' M \hat{G}_q = T' \hat{F}_q' M \hat{F}_q T = I_q$.

• L'hypothèse (H_2) est vérifiée car la matrice des erreurs $\hat{E}_q$ n'est pas modifiée.

• L'hypothèse (H_3) est vérifiée car $\hat{E}_q' M \hat{G}_q = \hat{E}_q' M \hat{F}_q T = 0$.

• Les "loadings" après rotation, notés $\hat{b}_j^\alpha$, représentent les corrélations des variables x^j aux facteurs g^α après la rotation.

On a : $\hat{B}_q = \hat{A}_q T = Z'(U_q T) = Z'\breve{U}_q$ où $\breve{U}_q = U_q T$.
De plus, on a :

$$\hat{g}^\alpha = \sqrt{m}\breve{u}^\alpha$$

où $\hat{g}^\alpha$ est la $\alpha^{\text{ème}}$ colonne de la matrice $\hat{G}_q$, et $\breve{u}^\alpha$ est la $\alpha^{\text{ème}}$ colonne de la matrice $\breve{U}_q$.
On en déduit :

$$\begin{aligned}
\hat{b}_j^\alpha &= \sum_{i=1}^{n} z_i^j \breve{u}_i^\alpha \\
&= \sum_{i=1}^{n} z_i^j \frac{1}{\sqrt{m}} g_i^\alpha \\
&= \sum_{i=1}^{n} \left(\frac{x_i^j - \overline{x}^j}{\sqrt{m}s^j}\right)\left(\frac{g_i^\alpha - \overline{g^\alpha}}{\sqrt{m}\sqrt{var(g^\alpha)}}\right) \\
&= corr(x^j, g^\alpha). \qquad (49)
\end{aligned}$$

• La matrice de saturation après la rotation reproduit toujours le modèle de structure de covariance défini par (8) :

$$\hat{B}_q\hat{B}_q' + \hat{\Xi} = \hat{A}_q T T' \hat{A}_q' + \hat{\Xi} = \hat{A}_q\hat{A}_q' + \hat{\Xi} = R. \qquad (50)$$

• Les communalités sont inchangées :

$$\hat{h}_j^2 = \sum_{\alpha=1}^{q} (\hat{b}_j^\alpha)^2 = \sum_{\alpha=1}^{q} (\hat{a}_j^\alpha)^2 \qquad (51)$$

car $\hat{B}_q\hat{B}_q' = (\hat{A}_q T)(\hat{A}_q T)' = \hat{A}_q\hat{A}_q'$.

• La variance totale expliquée par les q facteurs communs n'est pas modifiée :

$$\sum_{j=1}^{p}\sum_{\alpha=1}^{q} (\hat{b}_j^\alpha)^2 = \sum_{i=1}^{p}\sum_{\alpha=1}^{q} (\hat{a}_j^\alpha)^2. \qquad (52)$$

Références

[1] Baccini A., Besse P. (2005), "Data mining I, Exploration Statistique" *http ://www.lsp.ups-tlse.fr/Besse/pub/Explo_stat.pdf.*

[2] Bouveyron C., (2006), *Modélisation et classification des données de grande dimension - Application à l'analyse d'images*, p 45-47, Thèse, Université Joseph Fourier - Grenoble 1.

[3] Fine J. (1993), "Problèmes d'indétermination en analyse en facteurs et analyse en composantes principales optimale", *Revue de Statistique Appliquée*, tome 41, n°4, p 45-72.

[4] Garnett J.-C.(1919), "General ability, cleverness and purpose", *British Journal of Psychiatry*, **9**, p 345-366.

[5] Harman H. H. (1960), *Modern Factor Analysis*, University of Chicago Press.

[6] Heywood H.B. (1931), "On finite sequences of real numbers", *Proceedings of the Royal Society, Series A*, **134**, p 486-501.

[7] Hotelling H. (1933), "Analysis of a complex of statistical variables into principal components", *Journal of Educational Psychology*, **24**, p 417-441.

[8] Jobson J.D. (1992), *Applied Multivariate Data Analysis, Volume II : Categorical and Multivariate Methods*, Springer-Verlag.

[9] Lawley D.N., Maxwell A.E. (1963), *Factor Analysis as a statistical method*, Butterworths London.

[10] Lebart L., Morineau A., Piron M. (1997), *Statistique exploratoire multidimensionnelle, 2e cycle, 2e édition*, Editions Dunod.

[11] Pearson K. (1901), "On lines and planes of closest fit to systems of points in space", *Philosophical Magazine*, **2**, p 559-572.

[12] Rencher, A.C. (2002), *Methods of Multivariate Analysis, Second Edition*, Wiley Series in Probability and Statistics.

[13] Seber G.A.F. (1984), *Multivariate observations*, Wiley Series in Probability and Mathematical Statistics.

[14] Spearman C. (1904), "General intelligence, objectively determined and measured", *American Journal of Psychology*, **15**, p 201-293.

[15] Tipping M.E, Bishop C.M. (1999), "Probabilistic Principal Component Analysis", *Journal of the Royal Statistical Society, Series B*, **61**, Part 3, p 611-622.