1 Objectif

Stratégies de détermination du nombre d'axes en ACP (Analyse en Composantes Principales).

L'analyse en composantes principales (ACP) est une technique exploratoire très populaire. Il s'agit de résumer l'information contenue dans un fichier en un certain nombre de variables synthétiques, combinaisons linéaires des variables originelles. On les appelle « composantes principales », ou « axes factoriels », ou tout simplement « facteurs ». Nous devons les interpréter pour comprendre les principales idées forces que recèlent les données.

Le choix du nombre de facteurs est très important. L'enjeu est de distinguer d'une part l'information pertinente (le « signal »), véhiculée par les axes que l'on choisit de retenir ; et d'autre part, l'information résiduelle – le « bruit » issu des fluctuations d'échantillonnage – traduite par les derniers facteurs que l'on choisit de négliger. Et c'est là que le bât blesse. On nous dit généralement qu'il faut conserver les facteurs intéressants, pourvu qu'ils soient interprétables. Difficile d'être plus approximatif. Pourtant, pour un expert, ce n'est pas vraiment un problème. Attention, l'expertise ne se limite pas à la méthode. Elle englobe une bonne connaissance du domaine (savoir ce qui est possible d'obtenir ou pas) et des données manipulées (pouvoir détecter rapidement les anomalies ou les évidences). Bref, il dispose de tous les garde-fous nécessaires pour produire des résultats valables.

Pour le néophyte, l'à peu près n'est absolument pas gérable. C'est un peu le cas de mes étudiants. Ils connaissent bien les statistiques. Ils sont en train d'apprendre les techniques factorielles. Mais pour ce qui est des applications pratiques sur des fichiers de données, ils ne sont ni médecins, ni spécialistes du marketing, etc. L'expertise métier leur fait défaut pour encadrer les résultats. Dès lors, ils ont besoin de repères numériques forts lorsqu'ils mènent une analyse. Et je me suis rendu compte que la très grande majorité des étudiants se contentaient très souvent de la règle de Kaiser-Guttman (valeur propre > 1), et parfois de la règle du coude associée au « scree plot » mettant en lumière la décroissance des valeurs propres (« éboulis » des valeurs propres). Pourtant d'autres règles existent. Elles sont malheureusement peu connues, peu diffusées, et de ce fait peu utilisées.

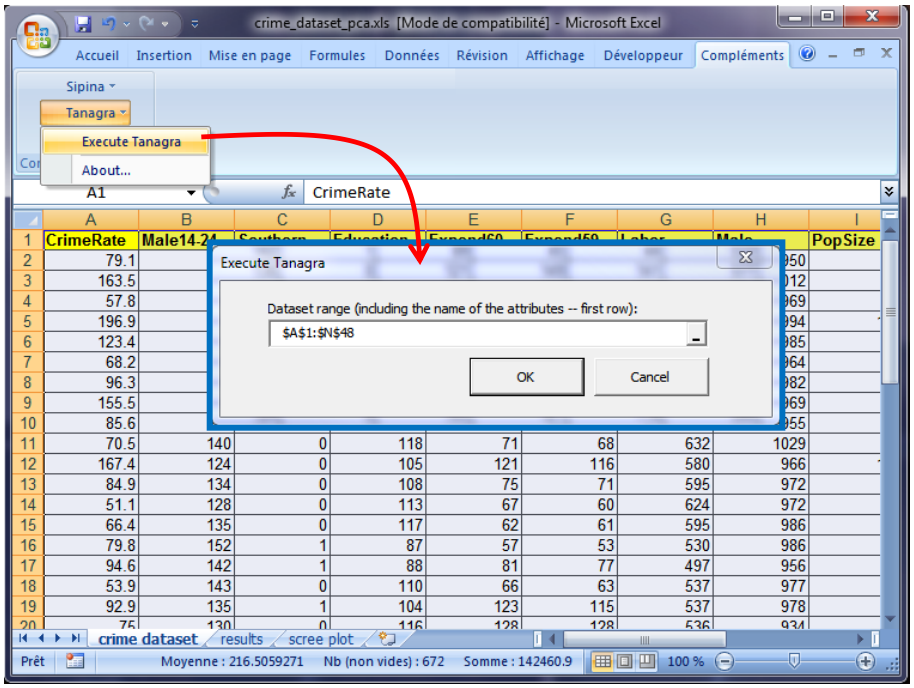
Dans ce tutoriel, nous présentons plusieurs méthodes de détermination du nombre adéquat de facteurs. Nous nous concentrerons tout d'abord sur les procédures simples, facilement opérationnelles. Les techniques de ré-échantillonnage, efficaces certes, mais gourmandes en ressources surtout lorsque la taille des fichiers augmente, feront l'objet d'une description à part. Nous détaillerons les calculs à partir des résultats d'une **ACP normée** menée sur une base relativement réduite. Nous travaillerons dans un premier temps avec le couple TANAGRA + tableur Excel puis, dans un second temps, nous décrirons la même analyse menée à l'aide de la fonction PRINCOMP du logiciel R. Ce document a été inspiré par plusieurs articles référencés en bibliographie.

2 Données – Analyse en composantes principales

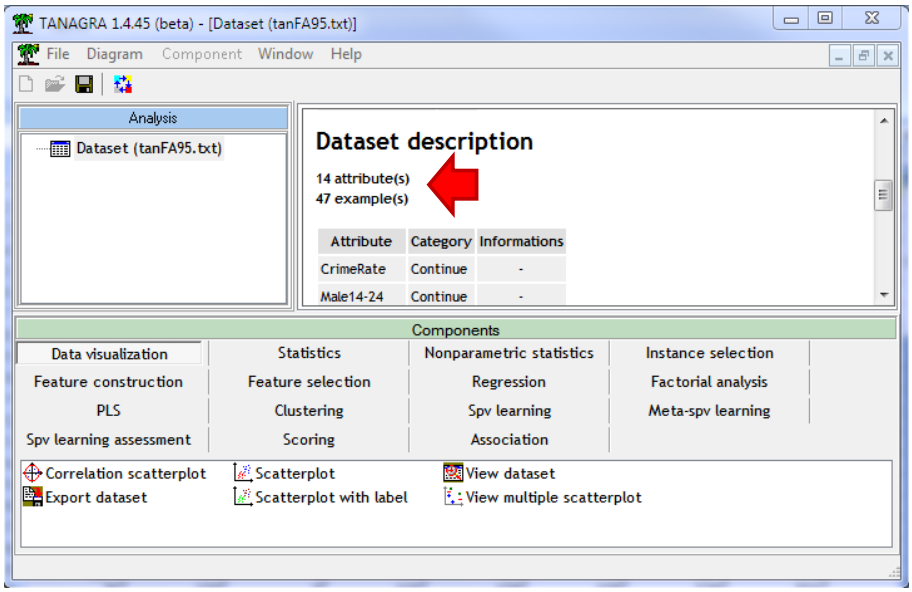
Nous travaillons sur le fichier « crime_dataset_pca.xls ». Il décrit « **p = 14** » indicateurs relatifs à la criminalité pour « **n = 47** » états des USA dans les années 60. Les données proviennent du site DASL (http://lib.stat.cmu.edu/DASL/Datafiles/USCrime.html). Nous l'avions déjà traité dans un didacticiel

présentant la rotation VARIMAX en analyse en composantes principales[1]. Nous nous contenterons donc d'un descriptif succinct de la mise en œuvre de l'ACP pour nous concentrer sur les stratégies de détermination du nombre « **k** » de facteurs.

Après avoir chargé les données dans le tableur Excel, nous les transmettons à Tanagra via la macro-complémentaire « tanagra.xla »[2].



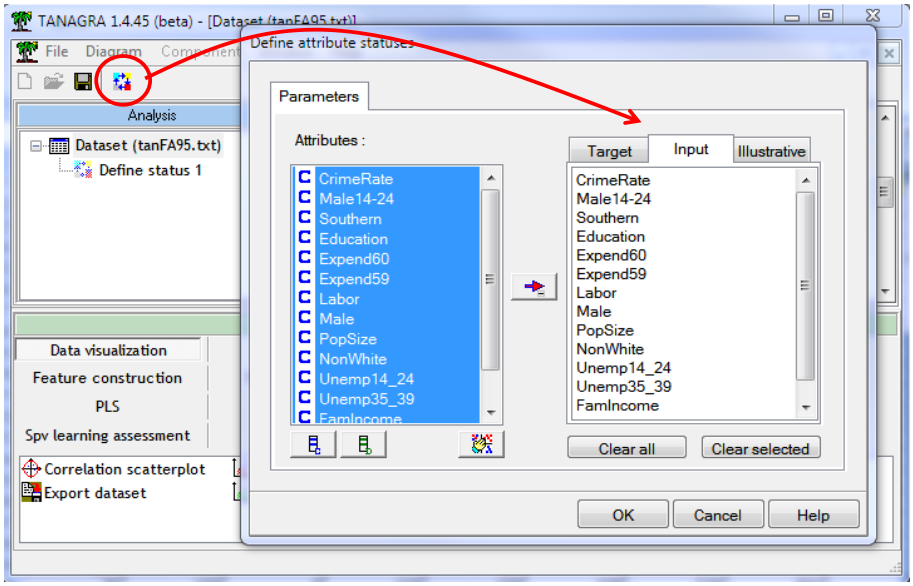
Tanagra est automatiquement démarré. Nous disposons de 47 observations et 14 variables.



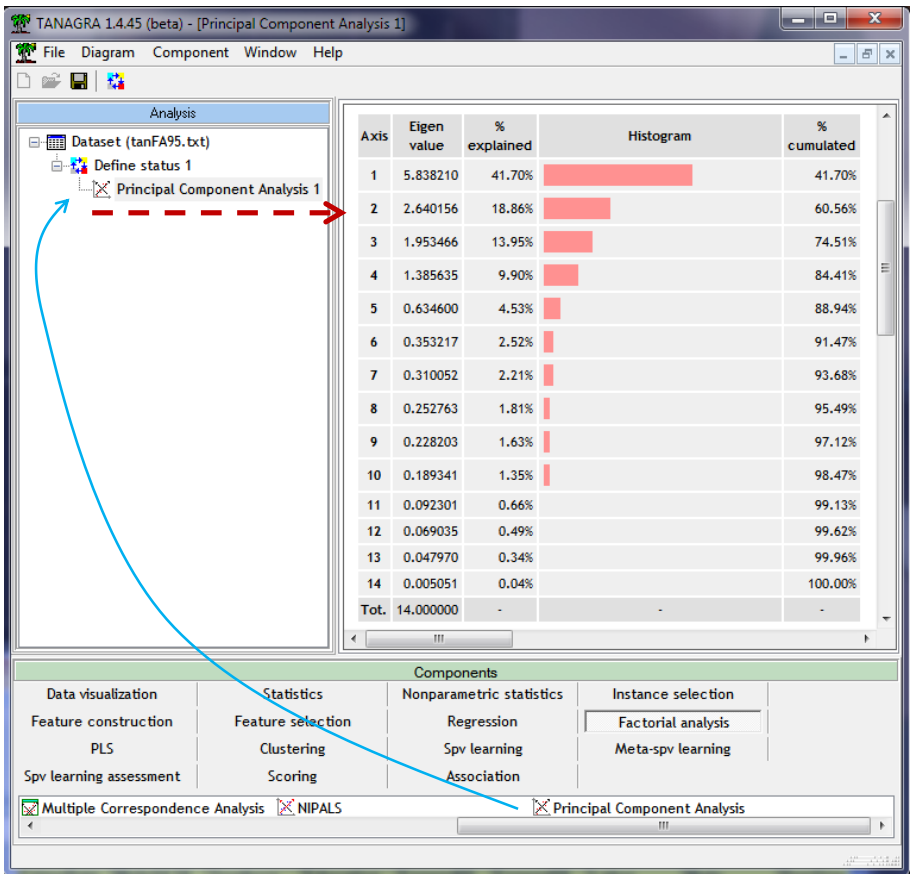
Nous insérons le composant DEFINE STATUS pour définir les variables de l'étude (CRIMERATE…INCUNDERMED).

[1] « Rotation VARIMAX en ACP » - http://tutoriels-data-mining.blogspot.fr/2008/04/rotation-varimax-en-acp.html

[2] Voir http://tutoriels-data-mining.blogspot.fr/2010/08/ladd-in-tanagra-pour-excel-2007-et-2010.html pour l'installation de la macro complémentaire sous Excel 2007 et 2010.



Puis nous insérons le composant PRINCIPAL COMPONENT ANALYSIS (onglet FACTORIAL ANALYSIS). Par défaut, et c'est ce que nous souhaitons mener, Tanagra effectue une ACP normée c.-à-d. les variables sont réduites, nous diagonalisons la matrice de corrélation.



Axis	Eigen value	% explained	Histogram	% cumulated
1	5.838210	41.70%		41.70%
2	2.640156	18.86%		60.56%
3	1.953466	13.95%		74.51%
4	1.385635	9.90%		84.41%
5	0.634600	4.53%		88.94%
6	0.353217	2.52%		91.47%
7	0.310052	2.21%		93.68%
8	0.252763	1.81%		95.49%
9	0.228203	1.63%		97.12%
10	0.189341	1.35%		98.47%
11	0.092301	0.66%		99.13%
12	0.069035	0.49%		99.62%
13	0.047970	0.34%		99.96%
14	0.005051	0.04%		100.00%
Tot.	14.000000	-	-	-

Tanagra fournit – entres autres[3] – le tableau des valeurs propres rangées de manière décroissante.

[3] Voir http://tutoriels-data-mining.blogspot.fr/2008/03/acp-description-de-vhicules.html pour une description détaillée des sorties de Tanagra (cercle des corrélations, projections des individus dans les plans factoriels, etc.).

Pour rappel, la valeur propre associée à un axe correspond à la fraction d'inertie qu'il retranscrit. Plus elle est élevée, plus le facteur est important dans la lecture des résultats. L'enjeu justement est de déceler à partir de quel stade l'information restituée peut être considérée négligeable. L'affaire n'est pas facile. En effet, plusieurs éléments entrent en ligne de compte (Jackson, 1993) : le nombre d'observations « **n** » ; le nombre de variables « **p** » de l'analyse ; le ratio « **n:p** »entre le nombre d'observations et le nombre de variables ; le degré de liaison (la corrélation) entre les variables ; l'existence éventuelle de blocs de variables corrélées dans le tableau de données.

Le ratio « n:p » est particulièrement important, il détermine la stabilité des résultats. Certaines références affirment qu'une ACP n'est vraiment viable que s'il est supérieur à 3 (Grossman et al., 1991). Nous avons 47/14 = 3.36 dans notre fichier. Nous pouvons travailler en confiance.

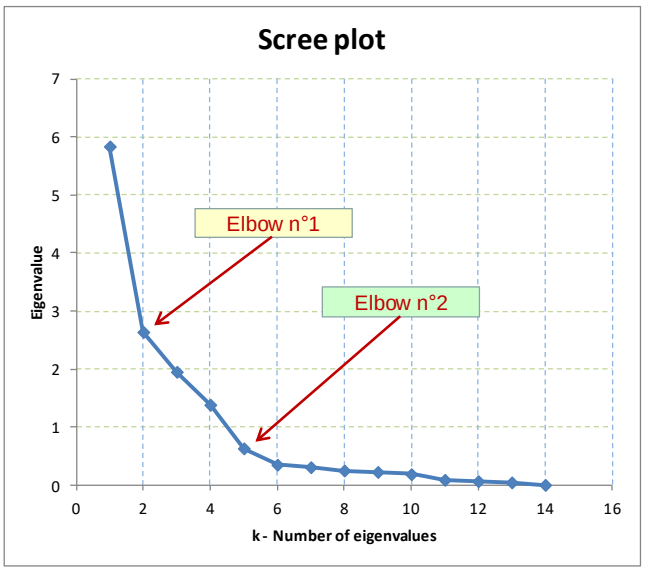
3 Scree plot

3.1 Scree plot

Cattell (1966, 1977) propose d'étudier la courbe de décroissance des valeurs propres (λ_k). L'idée est de détecter les « coudes » (les « cassures ») signalant un changement de structure. Cette approche est intéressante parce qu'elle est nuancée. Elle permet de dépasser l'arbitraire purement numérique. Mais elle est compliquée à mettre en œuvre parce qu'elle est justement soumise à notre appréciation. La détection n'est pas toujours évidente. Il faut répondre à plusieurs questions : Où est situé le coude ? Est-ce qu'il est unique ? Est-ce que nous l'incluons ou pas dans la sélection ?

En règle générale, le coude est très marqué lorsque nous traitons des variables fortement corrélées. Lorsqu'elles le sont faiblement ou lorsqu'il y a des blocs de variables corrélées, plutôt qu'une solution unique « évidente », nous devons faire face à plusieurs scénarios. Concernant l'intégration du coude dans la sélection, Cattel lui-même a varié. Dans un premier temps (Cattell, 1966), il conseillait de ne sélectionner que les facteurs qui sont avant le coude ; puis, dans un second temps (1977), il préconise finalement de l'intégrer. Tout dépend de la valeur associée au coude en réalité. Si elle est faible, on peut négliger le facteur. Nous devons le sélectionner en revanche si elle est élevée.

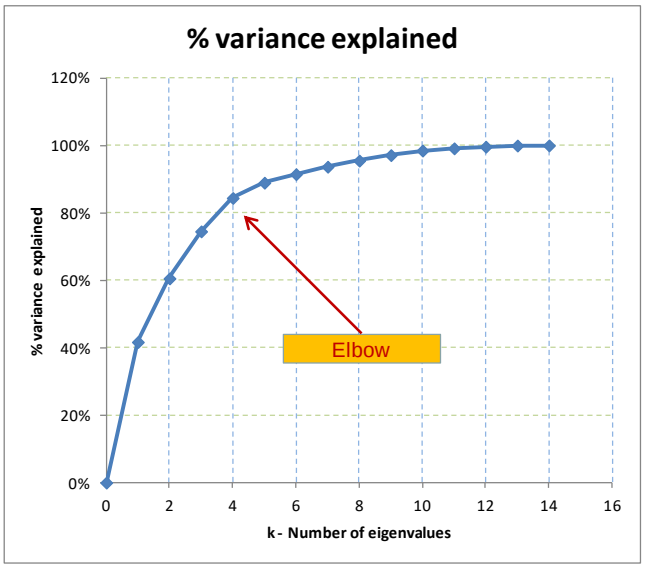
Pour ce qui est des données CRIME, nous avons :



Il semble y avoir deux coudes ? En k = 2 et k = 5. Laquelle choisir ? Si nous prenons la première solution, nous devons prendre k = 2 facteurs c.-à-d. inclure le coude dans la sélection. En effet, la composante explique 18.86% de la variabilité totale ($\lambda_2 = 2.64$). On peut difficilement le négliger.

Si nous optons pour la seconde solution, il est clair que la 5ème composante ne doit pas incluse dans la sélection parce que $\lambda_5 = 0.6346$ (4.53% de l'inertie totale) est trop faible pour porter de l'information pertinente. Nous choisirons donc k = 4 composantes dans ce second cas de figure.

Pour préciser la lecture, il semble intéressant de compléter le « scree plot » par un second graphique décrivant l'évolution de l'inertie expliquée par les axes. Il s'agit ni plus ni moins de la courbe des valeurs propres cumulées, en pourcentage. Un « coude » devrait y être visible également. Mais la partie subséquente doit être horizontale, indiquant ainsi un apport négligeable des facteurs restants. Pour nos données nous obtenons :



Ce graphique complète le scree plot ; à deux, Ils s'avèrent souvent décisifs. Clairement, une analyse en (k = 2) composantes n'est pas une bonne idée. Le gain informationnel en intégrant 3 axes reste substantiel, idem pour le passage de 3 à 4. En revanche, il est faible lors du passage de 4 à 5 facteurs. **On peut penser que (k = 4) axes semble être la solution la plus appropriée pour nos données**.

3.2 Proportion d'inertie expliquée

On est parfois tenté d'utiliser explicitement la part de variance expliquée pour déterminer le nombre de facteurs. La règle serait alors : « sélectionner suffisamment d'axes pour expliquer au moins x% de l'inertie totale ». Cette stratégie est descendue en flammes par toutes les références que j'ai pu consulter. Pour la simple raison qu'elle ne tient pas compte du tout des corrélations entre les variables. Dans notre exemple, mettons que l'on souhaite capter 95% de l'information disponible, nous serions amenés à retenir 8 facteurs. Tout en sachant pertinemment que la variabilité est quasi nulle à partir du 5ème.

Néanmoins, après coup, après avoir choisi le nombre d'axes à l'aide d'une des approches décrites dans ce document, il peut être très intéressant de pouvoir situer la quantité d'information – à l'aide de la fraction d'inertie expliquée - que restituent les facteurs sélectionnés.

4 Règles de Kaiser simples et améliorées

4.1 Règle de Kaiser - Guttman

La règle de Kaiser repose sur une idée simple. Dans une ACP normée, la somme des valeurs propres étant égale au nombre de variables, leur moyenne vaut 1. Nous considérons par conséquent qu'un axe est intéressant si sa valeur propre est supérieure 1.

Il existe d'autres manières de considérer ce seuil : un axe est intéressant s'il contribue plus qu'une des variables prise individuellement ; ou encore, si les variables étaient deux à deux orthogonales, les valeurs propres issues de l'analyse seraient toutes égales à 1.

Dans notre exemple, nous retiendrons ainsi les « **k = 4** » premiers axes factoriels, ils restituent 84.41% de l'information (l'inertie) disponible.

Kaiser critical value	1

Axis	Eigen value	% explained	% cumulated
1	5.838210	41.70%	41.70%
2	2.640156	18.86%	60.56%
3	1.953466	13.95%	74.51%
4	1.385635	9.90%	84.41%
5	0.634600	4.53%	88.94%
6	0.353217	2.52%	91.47%
7	0.310052	2.21%	93.68%
8	0.252763	1.81%	95.49%
9	0.228203	1.63%	97.12%
10	0.189341	1.35%	98.47%
11	0.092301	0.66%	99.13%
12	0.069035	0.49%	99.62%
13	0.047970	0.34%	99.96%
14	0.005051	0.04%	100.00%
Tot.	14	-	-

Un seuil abrupt, sans nuances, paraît toujours un peu arbitraire. C'est en cela que les règles purement numériques sont gênantes. Nous noterons cependant que ce résultat confirme l'inspection de l'éboulis des valeurs propres.

4.2 Règle de Karlis – Saporta - Spinaki (2003)

On pense généralement que le seuil 1 est trop permissif c.-à-d. nous retenons plus d'axes factoriels qu'il n'en faut. Il n'est réellement fondé que si les variables sont fortement corrélées, autrement il faudrait le relever. Une règle plus restrictive consiste à le définir comme suit : moyenne des valeurs propres + 2 fois leur écart-type (Saporta, 2006 ; page 172). Elle rappelle la définition de la valeur critique d'un test unilatéral de conformité à 5%, où la statistique suit asymptotiquement une loi normale.

La règle d'acceptation devient maintenant :

$$\lambda > 1 + 2\sqrt{\frac{p-1}{n-1}}$$

Par rapport à la règle de Kaiser, elle est plus restrictive. Cela va dans le sens souhaité. Notons également qu'elle dépend du ratio « n:p », déterminant dans la qualité des résultats de l'ACP. Nous serons d'autant plus exigeants – nous serons enclins à accepter moins de facteurs - que le nombre de variables « p » est élevé par rapport aux observations disponibles « n ».

Dans notre exemple,

$$1+2\sqrt{\frac{p-1}{n-1}}=1+2\sqrt{\frac{14-1}{47-1}}=2.063$$

Avec un seuil égal à **2.063**, nous ne retiendrions que les **k= 2** premiers axes factoriels.

n	47
p	14

Karlis et al. critical value	**2.063**

Axis	Eigen value	% explained	% cumulated
1	5.838210	41.70%	41.70%
2	2.640156	18.86%	60.56%
3	1.953466	13.95%	74.51%
4	1.385635	9.90%	84.41%
5	0.634600	4.53%	88.94%
6	0.353217	2.52%	91.47%
7	0.310052	2.21%	93.68%
8	0.252763	1.81%	95.49%
9	0.228203	1.63%	97.12%
10	0.189341	1.35%	98.47%
11	0.092301	0.66%	99.13%
12	0.069035	0.49%	99.62%
13	0.047970	0.34%	99.96%
14	0.005051	0.04%	100.00%
Tot.	14	-	-

Rappelons que k = 2 faisait partie des solutions plausibles lorsque nous avions étudié l'éboulis des valeurs propres. Notons aussi que la 3ème valeur propre λ_3 = 1.953466 est assez proche du seuil.

5 Tests de Bartlett

5.1 Existence de facteurs pertinents

Ce test sert à détecter l'existence de relations intéressantes entre les variables de notre fichier[4]. Formellement, l'hypothèse nulle (H0) du test de sphéricité de Bartlett correspond à : « toutes les valeurs propres sont identiques, elles sont égales à 1 ». Si on rejette H0, on sait qu'il existe au moins un facteur pertinent dans l'ACP. Nous ne pouvons pas spécifier leur nombre toutefois.

La statistique de test est basée sur le déterminant du coefficient de corrélation R, elle s'écrit :

$$C=-\left(n-1-\frac{2p+5}{6}\right)\times\ln|R|=-\left(n-\frac{2p+11}{6}\right)\times\ln|R|$$

[4] Voir http://tutoriels-data-mining.blogspot.fr/2012/05/acp-sous-r-indice-kmo-et-test-de.html

Sous H0, elle suit une loi du χ^2 à [p x (p-1) / 2] degrés de liberté.

Il n'est pas nécessaire de calculer explicitement la matrice de corrélation. En effet, son déterminant est égal au produit des valeurs propres. La formule peut être réécrite comme suit :

$$C = -\left(n - 1 - \frac{2p+5}{6}\right) \times \ln|R| = -\left(n - \frac{2p+11}{6}\right) \times \sum_{k=1}^{p} \ln(\lambda_k)$$

En ce qui nous concerne, le déterminant est égal à

$$|R| = \prod_k \lambda_k = (5.838210 \times \cdots \times 0.005051) = 4.889 \times 10^{-8}$$

Nous pouvons calculer facilement la statistique :

n	47
p	14

Axis	Eigen value	% explained	% cumulated
1	5.838210	41.70%	41.70%
2	2.640156	18.86%	60.56%
3	1.953466	13.95%	74.51%
4	1.385635	9.90%	84.41%
5	0.634600	4.53%	88.94%
6	0.353217	2.52%	91.47%
7	0.310052	2.21%	93.68%
8	0.252763	1.81%	95.49%
9	0.228203	1.63%	97.12%
10	0.189341	1.35%	98.47%
11	0.092301	0.66%	99.13%
12	0.069035	0.49%	99.62%
13	0.047970	0.34%	99.96%
14	0.005051	0.04%	100.00%
Tot.	14	-	-

|R|	4.889E-08

C	681.76
d.f.	91
p-value	2.941E-91

Manifestement, il y a des relations exploitables dans ce fichier.

<u>Note :</u> Il y a quand même une petite restriction à l'utilisation de ce test, il a tendance à être systématiquement significatif lorsque la taille de l'échantillon augmente.

5.2 Détection du nombre de facteurs pertinents

Une variante du test de sphéricité de Bartlett peut être utilisée pour déterminer le nombre d'axes. Formellement, elle n'est valable que pour l'ACP non-normée, c.-à-d. calculée à partir de la matrice de variance covariance (Saporta, 2006, page 171 ; Grossman et al., 1991). Il semble néanmoins qu'elle soit applicable pour l'ACP normée, tout en sachant qu'elle s'avère conservatrice dans ce cas c.-à-d. elle a tendance à sélectionner moins de facteurs qu'il ne faudrait (Jackson, 1993 ; Neto et al., 2004). Moi je n'ai pas d'a priori, nous allons voir ce qu'il en est sur notre fichier de données.

Le principe du test est le suivant : nous choisissons « k » facteurs dans l'ACP parce que les (p-k) suivants ont la même valeur propre. Ces valeurs correspondent à la partie horizontale située après le coude dans le graphique « scree plot ». Voici l'hypothèse nulle :

$$H_0 : \lambda_{k+1} = \cdots = \lambda_p$$

Malheureusement, nos références diffèrent sur la définition de la statistique de test. Après moult vérifications, j'ai décidé de prendre la formulation de Saporta (2006, page 171) parce que d'une part elle est très intuitive, d'autre part elle permet de retrouver la valeur de la statistique de test de la section précédente, lorsque nous décidons d'éprouver l'égalité des « p » valeurs propres. Les autres descriptions ne sont pas cohérentes de ce point de vue.

Sous H0, les (p-k) valeurs propres sont identiques. Dans ce cas leur moyenne arithmétique $\bar{\lambda}$ est égale à leur moyenne géométrique $\tilde{\lambda}$. La statistique permettant de tester l'égalité des (p-k) valeurs propres consécutifs au k^{ème} facteur s'écrit :

$$c_k = \left(n - \frac{2p+11}{6}\right) \times (p-k) \times \ln\left(\frac{\bar{\lambda}}{\tilde{\lambda}}\right)$$

Où $\bar{\lambda} = \frac{1}{p-k}\sum_{i=k+1}^{p}\lambda_i$

Et $\ln\left(\tilde{\lambda}\right) = \frac{1}{p-k}\sum_{i=k+1}^{p}\ln\left(\lambda_i\right)$

Sous H0, elle suit une loi du χ^2 à $\frac{(p-k+2)(p-k-1)}{2}$ degrés de liberté.

La statistique est en accord avec celle du test précédent. En effet, pour k=0 c.-à-d. nous testons l'égalité de toutes les valeurs propres (toutes égales à 1 par conséquent), nous avons $\bar{\lambda} = 1$ et

$$\begin{aligned} c_0 &= \left(n - \frac{2p+11}{6}\right) \times p \times \left[-\frac{1}{p}\ln\left(\prod_{i=1}^{p}\lambda_i\right)\right] \\ &= \left(n - \frac{2p+11}{6}\right) \times p \times \left[-\frac{1}{p}\sum_{i=1}^{p}\ln\left(\lambda_i\right)\right] \\ &= -\left(n - \frac{2p+11}{6}\right) \times \sum_{i=1}^{p}\ln\left(\lambda_i\right) \\ &= C \end{aligned}$$

Il y a des problèmes malheureusement du côté des degrés de liberté. Lorsque k= 0, nous avons d.f. = (p+2)(p-1)/2, différent du p x (p - 1) / 2 décrit plus haut. Pourtant, toutes les références s'accordent sur cette expression. Cela reste un mystère.

Nous avons monté une feuille Excel (page suivante) pour détecter le nombre de facteurs à retenir pour l'ACP. Nous devons nous arrêter à k = 12 dans notre tableau, nous testons dans ce cas l'égalité des 13ème et 14ème valeurs propres.

Pour k = 0, nous testons l'égalité de toutes les valeurs propres, nous retrouvons bien C = c_0 = 681.7625. Les degrés de liberté diffèrent cependant. Ici nous avons d.f. = 104 (contre 91).

Manifestement, ce test n'est pas adapté. Il nous annonce que toutes les valeurs propres sont significativement différentes, quelle que soit la valeur de « k ». Ce résultat est en totale contradiction avec les constatations effectuées jusqu'ici (et dans ce qui viendra par la suite).

k	i	lambda	lambda_barre	lambda_tilde	ln(l_barre/l_tilde)	c_k	d.f.	p-value
0	1	5.838210	1.0000	0.3005	1.2024	681.7625	104	9.99E-86
1	2	2.640156	0.6278	0.2392	0.9651	508.1433	90	1.36E-59
2	3	1.953466	0.4601	0.1958	0.8545	415.2914	77	7.57E-48
3	4	1.385635	0.3244	0.1588	0.7140	318.0966	65	5.05E-35
4	5	0.634600	0.2183	0.1279	0.5344	216.4189	54	2.56E-21
5	6	0.353217	0.1720	0.1071	0.4741	172.8204	44	3.56E-17
6	7	0.310052	0.1493	0.0922	0.4821	156.2073	35	3.00E-17
7	8	0.252763	0.1264	0.0775	0.4884	138.4702	27	6.13E-17
8	9	0.228203	0.1053	0.0637	0.5030	122.2381	20	1.10E-16
9	10	0.189341	0.0807	0.0493	0.4926	99.7420	14	5.31E-15
10	11	0.092301	0.0536	0.0352	0.4189	67.8603	9	3.99E-11
11	12	0.069035	0.0407	0.0256	0.4643	56.4095	5	6.69E-11
12	13	0.047970	0.0265	0.0156	0.5325	43.1292	2	4.31E-10
13	14	0.005051						

Je n'ai pas effectué de vérifications pour une ACP non normée. Mais j'ai quand même quelques doutes quant à cette procédure. Il semble que le test (global) de Bartlett soit surtout intéressant pour déceler la présence de facteurs pertinents. Mieux vaut se tourner vers d'autres approches pour déterminer leur nombre (Neto et al., 2004).

6 Test des « bâtons brisés »

Ce test est du à Frontier (1976) et Legendre-Legendre (1983). Il repose sur l'idée que si l'inertie totale était dispatchée aléatoirement sur les axes, la distribution des valeurs propres suivrait la loi des « bâtons brisés » (broken-stick). Elle a été tabulée semble-t-il. Mais il est très facile de calculer la valeur critique pour le choix de « k » facteurs. Elle s'écrit :

$$b_k = \sum_{i=k}^{p} \frac{1}{i}$$

Si nous souhaitons raisonner en termes de part d'inertie expliquée, nous utiliserons le seuil

$$b'_k = \frac{1}{p} \sum_{i=k}^{p} \frac{1}{i}$$

Les calculs sont très faciles à mettre en œuvre sur un tableur.

k	Eigen value	1/i	b_k
1	5.838210	1.000	3.252
2	2.640156	0.500	2.252
3	1.953466	0.333	1.752
4	1.385635	0.250	1.418
5	0.634600	0.200	1.168
6	0.353217	0.167	0.968
7	0.310052	0.143	0.802
8	0.252763	0.125	0.659
9	0.228203	0.111	0.534
10	0.189341	0.100	0.423
11	0.092301	0.091	0.323
12	0.069035	0.083	0.232
13	0.047970	0.077	0.148
14	0.005051	0.071	0.071

Pour tester le premier axe (k = 1) par exemple, nous calculons le seuil comme suit :

$$b_1 = \left(\frac{1}{1} + \frac{1}{2} + \cdots + \frac{1}{14}\right) = 3.252$$

Comme λ_1=5.83821, il est validé.

Ainsi, la procédure des bâtons brisés nous annonce que **k = 3** facteurs sont pertinents dans cette analyse. Signalons quand même que le 4ème axe a été éliminé de justesse. Est-ce qu'il faut réellement l'écarter ? Ce dilemme pèsera toujours sur les approches purement numériques.

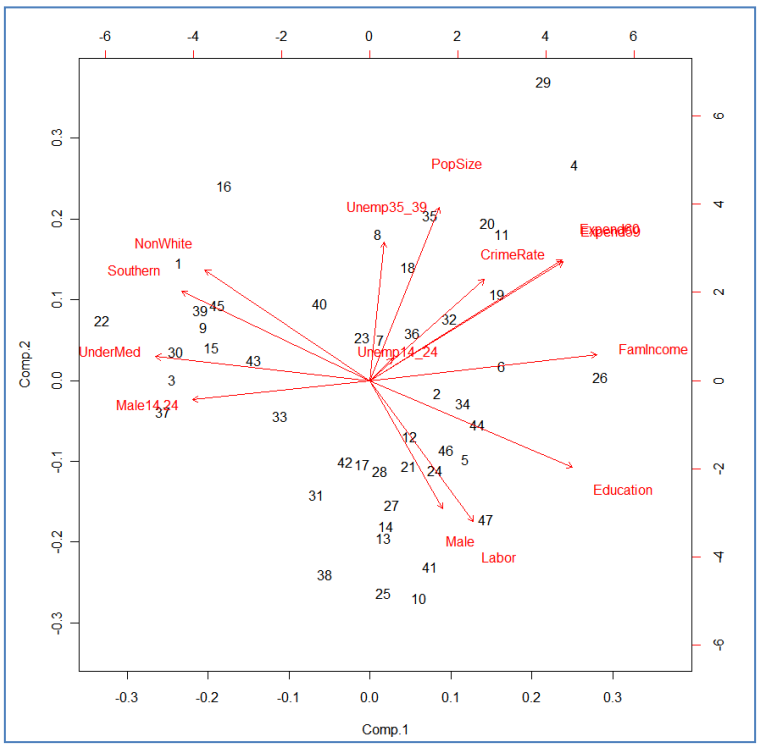
Quoiqu'il en soit, le test des bâtons brisés est performant (Jackson, 1993). On lui reproche simplement de ne tenir compte ni de « n » (le nombre d'observations) ni du ratio « n:p » dans la définition des valeurs critiques (Franklin et al., 1995).

7 Analyse en composantes principales avec R

Nous réalisons la même analyse sous R pour préparer la présentation des techniques de ré échantillonnage dans la section suivante. Nous avons utilisé le code :

```
rm(list=ls())
#importing the data file
library(xlsx)
crime.data <- read.xlsx(file="crime_dataset_pca.xls",sheetIndex=1,header=T)
#performing the pca with princomp
crime.pca <- princomp(crime.data,cor=T)
eig.val <- crime.pca$sdev^2
print(eig.val)
biplot(crime.pca)
```

La séquence des valeurs propres est bien la même que celle de Tanagra. A titre de pure curiosité, nous montrons la représentation « biplot » dans le premier plan factoriel.



8 Techniques de ré-échantillonnage

8.1 Calculer les seuils autrement – L'analyse parallèle

L'analyse parallèle cherche à obtenir les seuils de significativité en s'affranchissant des hypothèses statistiques. On cherche à calculer les valeurs propres pour un fichier de données avec les mêmes caractéristiques « n » et « p », mais où les variables seraient indépendantes. La procédure repose sur une approche Monte Carlo (Neto et al, 2004) :

1. Générer aléatoirement un jeu de données de mêmes dimensions « n » et « p » que notre fichier. Chaque variable suit une loi normale N(0,1).
2. Lancer l'ACP, recueillir la séquence des « p » valeurs propres (λ_k).
3. Répéter T fois les étapes (1) et (2).
4. Calculer la moyenne des valeurs propres (μ_k) pour chaque facteur.
5. On considère qu'un facteur est pertinent si ($\lambda_k > \mu_k$).

Si on veut être plus exigeant, nous calculons le quantile d'ordre 0.95 [$q_k^{0.95}$] (pour un test à 5%) à l'étape n°4, et nous utilisons ce seuil à l'étape n°5.

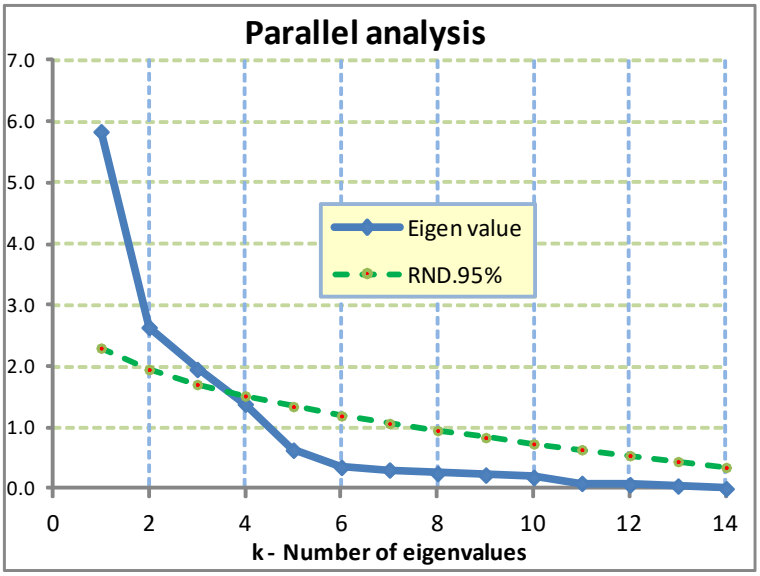
Voici le code R pour générer les seuils critiques, avec T = 1000.

```
#*****************
#PARALLEL ANALYSIS
#*****************

#n : number of instance, p : number of variables
n <- nrow(crime.data)
p <- ncol(crime.data)

#generation of a dataset
gendata <- function(n,p){
  df <- list()
  for (k in 1:p){
    x <- rnorm(n)
    df[[k]] <- x
  }
  df <- data.frame(df)
  colnames(df) <- 1:p
  return(df)
}

#pca on gendata
pca.gendata <- function(n,p){
  data.gen <- gendata(n,p)
  pca <- princomp(data.gen,cor=T)
  eig <- pca$sd^2
  return(eig)
}

set.seed(1)
```

```
#repeating T times the analysis
T <- 1000
res <- replicate(T, pca.gendata(n,p))

#computing the mean of the eigenvalues
rnd.mean <- apply(res,1,mean)
print(rnd.mean)
#computing the 0.95 percentile
rnd.95 <- apply(res,1,quantile,probs=(0.95))
print(rnd.95)
```

Nous avons repris les seuils (quantile d'ordre 0.95) et nous avons positionné chaque valeur propre dans un graphique.

Axis	Eigen value	RND.Mean	RND.95%
1	5.838210	2.051060	2.303778
2	2.640156	1.771630	1.952445
3	1.953466	1.563366	1.708953
4	**1.385635**	**1.388553**	**1.521725**
5	0.634600	1.239614	1.347275
6	0.353217	1.098187	1.195431
7	0.310052	0.978249	1.071068
8	0.252763	0.862912	0.955985
9	0.228203	0.752236	0.840423
10	0.189341	0.649536	0.737475
11	0.092301	0.551690	0.642607
12	0.069035	0.457871	0.545664
13	0.047970	0.366953	0.448322
14	0.005051	0.268144	0.351839



Les 3 premières composantes sont significatives à 5%. La 4ème est plus mitigée. La valeur propre λ_4=1.3856 est en-deçà du quantile. Mais elle est très proche de la moyenne μ_4=1.3885. On retrouve ici le doute que nous avions émis pour le test des bâtons brisés. La décision est autrement plus évidente pour la 5ème composante et les suivantes, nous pouvons les éliminer sans états d'âme.

Tables statistiques. Pour les réfractaires à la programmation, il existe des tables statistiques dans certains ouvrages. Les valeurs critiques sont générées selon la même démarche, mis à part qu'elles sont basées sur des réplications plus nombreuses. Dans (Husson et al, 2009 ; pages 204 à 207), des tables à 95% ont été élaborée avec T = 10000 essais. On lit par exemple que le premier axe est significatif si la proportion de variance restituée est supérieure[5] à 0.165. Ramené sur notre base, le seuil devient 0.165 x 14 = 2.31, très proche de ce que nous avions trouvé (2.30).

Calcul de la p-value du test. Plutôt que de comparer λ_k avec le quantile d'ordre 0.95 des valeurs produites par la simulation, nous pouvons obtenir directement une sorte de p-value du test en

[5] La valeur n'est pas disponible pour n =47. Nous lisons 0.168 pour n = 45, et 0.162 pour n = 50. Nous avons pris une valeur intermédiaire. Avec le programme R ci-dessus, il est possible de calculer les valeurs critiques pour n'importe quelle configuration « n » et « p ». Le nombre de réplications peut être modifié également.

comptabilisant la proportion de ces dernières situées au delà de λ_k. Si elle est plus petite que α=0.05, on peut considérer que l'axe est significatif.

Voici le code programme R correspondant

```
#computing the proportion of values upper than eig.val
prop <- rep(0,length(eig.val))
names(prop) <- names(eig.val)
for (k in 1:length(eig.val)){
 prop[k] <- length(which(res[k,] > eig.val[k]))/T
}
print(prop)
```

Nous obtenons :

```
R Console
> print(prop)
 Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8  Comp.9 Comp.10 Comp.11
  0.000   0.000   0.000   0.501   1.000   1.000   1.000   1.000   1.000   1.000   1.000
Comp.12 Comp.13 Comp.14
  1.000   1.000   1.000
>
```

Le résultat est forcément cohérent avec la décision basée sur le quantile. Les trois premiers facteurs sont clairement significatifs. Pour le 4ème, nous ne disposions pas de cette information précédemment, près de 50% des valeurs issues de la simulation sont supérieures à λ_4.

Analyse parallèle pour l'ACP non normée. La méthode peut être étendue à l'ACP non normée. Il suffit pour cela d'utiliser la moyenne et l'écart-type empirique de chaque variable lors de l'appel de **rnorm()** durant la génération des données.

8.2 Technique de randomisation

L'analyse parallèle est robuste par rapport à la normalité (utilisée lors de la génération des données) (Neto et al., 2004). Néanmoins, elle peut paraître inutilement restrictive. Nous pouvons nous en affranchir en utilisant une stratégie de randomisation pour la génération des données non corrélées, en exploitant les observations disponibles.

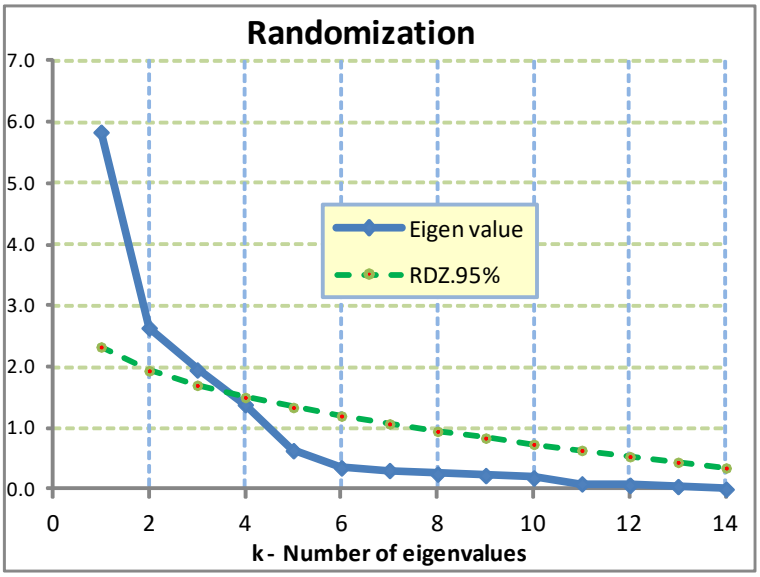
Nous formons toujours un échantillon de taille (n x p). Pour chaque variable, traitée de manière indépendante, nous mélangeons aléatoirement les valeurs. De fait, la corrélation entre les variables, si elle existait, est totalement altérée. La procédure est répétée T fois. Les valeurs critiques sont définies à partir du quantile d'ordre 0.95 (ou de la moyenne si on souhaite être moins restrictif).

```
#**************
# RANDOMIZATION
#**************
set.seed(1)
one.randomization <- function(dataset){
 dataset.rdz <- data.frame(lapply(dataset,function(x){sample(x,length(x),replace=F)}))
 pca.rdz <- princomp(dataset.rdz,cor=T)
 eig.rdz <- pca.rdz$sd^2
```

```
 return(eig.rdz)
}

#repeat the procedure
res.rdz <- replicate(T,one.randomization(crime.data))

#mean
rdz.mean <- apply(res.rdz,1,mean)
print(rdz.mean)
#quantile
rdz.95 <- apply(res.rdz,1,quantile,probs=(0.95))
print(rdz.95)
```

Nous obtenons une autre version du tableau de l'analyse parallèle, très peu différente en vérité, sauf pour le 4ème axe qui dépasse maintenant de très peu la moyenne des valeurs simulées.

Axis	Eigen value	RDZ.Mean	RDZ.95%
1	5.838210	2.062149	2.330069
2	2.640156	1.767845	1.940688
3	1.953466	1.561557	1.703605
4	**1.385635**	1.385119	1.505952
5	0.634600	1.236817	1.341472
6	0.353217	1.100492	1.202589
7	0.310052	0.974988	1.073624
8	0.252763	0.860179	0.951342
9	0.228203	0.752086	0.842079
10	0.189341	0.648836	0.739747
11	0.092301	0.552711	0.640388
12	0.069035	0.458954	0.537746
13	0.047970	0.367645	0.445864
14	0.005051	0.270622	0.352655



8.3 Calculer la significativité des valeurs propres par bootstrap

Appréhender la variabilité associée à chaque valeur propre est l'objectif de la méthode bootstrap. Pour ce faire, nous collectons les différentes versions de λ_k en répétant T fois la procédure suivante : on effectue un tirage avec remise des observations pour obtenir un échantillon de taille « n » ; nous calculons l'ACP sur ces données ; nous stockons la valeur propre pour chaque axe.

La décision est basée sur une variante de la règle de Kaiser. Un facteur est jugé pertinent si le quantile d'ordre 0.05 ($\lambda_k^{0.05}$) (la borne basse de l'intervalle de confiance) des valeurs propres bootstrap est supérieure à 1.

Le code R est relativement simple :

```
#**********
# BOOTSTRAP
#**********

#creating one replication of the dataset
```

```
one.replication <- function(dataset){
 n <- nrow(dataset)
 index <- sort(sample.int(n,replace=T))
 out.dataset <- dataset[index,]
 return(out.dataset)
}

#performing a pca on a replication of the dataset
pca.replication <- function(dataset){
 one.dataset <- one.replication(dataset)
 pca <- princomp(one.dataset,cor=T)
 eig <- pca$sd^2
 return(eig)
}

#bootstraping pca
res.boot <- replicate(T,pca.replication(crime.data))

#quantile 0.05
boot.05 <- apply(res.boot,1,quantile,probs=(0.05))
print(boot.05)
```

Les précédents résultats sont confirmés. Le doute subsiste toujours pour le 4ème axe ($\lambda_4^{0.05} = 0.966$).

Axis	Boot 0.05
1	5.176478
2	2.297267
3	1.649235
4	0.966243
5	0.459941
6	0.309183
7	0.228693
8	0.169048
9	0.120807
10	0.078729
11	0.047379
12	0.030524
13	0.016199
14	0.001778

8.4 Vérifier l'empiètement des valeurs propres par bootstrap

Cette procédure repose sur une autre vision du test d'égalité des valeurs propres. On considère qu'un facteur est pertinent si sa valeur propre est significativement supérieure à celle du facteur suivant ($\lambda_k > \lambda_{k+1}$). Dans les faits, il s'agit de vérifier qu'il n'y a pas d'empiètement entre leurs distributions, ou qu'il est suffisamment faible pour qu'on le considère négligeable. Nous utilisons le bootstrap pour estimer les distributions empiriques, et la règle de décision devient : le facteur « k » est sélectionné si $\lambda_k^{0.05} > \lambda_{k+1}^{0.95}$ c.-à-d. si la borne basse de l'intervalle de confiance de la kème valeur propre est supérieure à la borne haute de l'intervalle de confiance de la suivante.

Nous pouvons exploiter les résultats de la sous-section précédente pour calculer les bornes hautes.

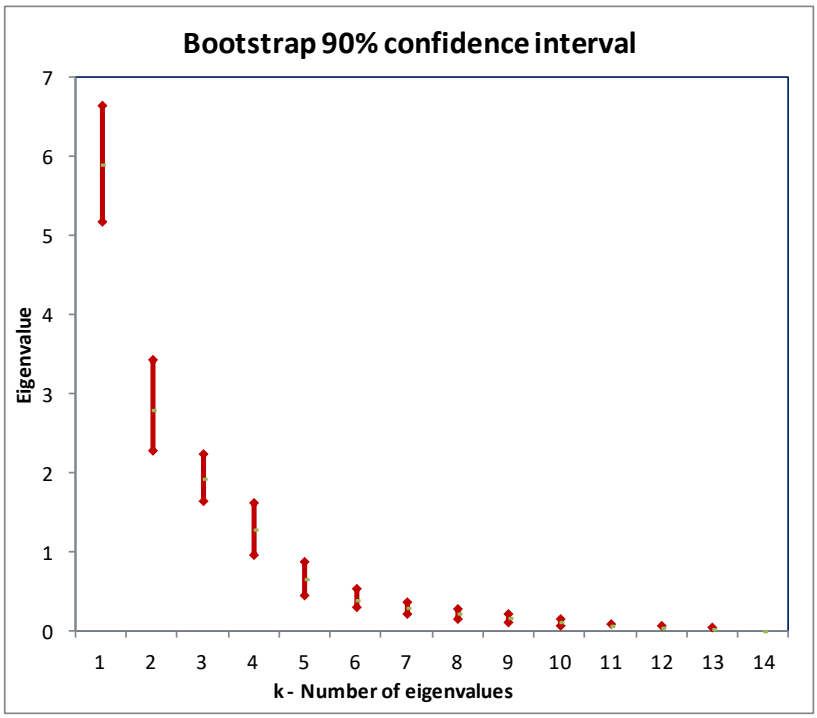
```
#quantile 0.95
boot.95 <- apply(res.boot,1,quantile,probs=(0.95))
print(boot.95)
```

Nous avons placé les valeurs dans un tableau en décalant les bornes hautes et basses pour rendre plus aisées les comparaisons.

Axis	Boot 0.05	Boot 0.95	Axis
		6.6547	1
1	5.1765	3.4381	2
2	2.2973	2.2418	3
3	1.6492	1.6303	4
4	0.9662	0.8790	5
5	0.4599	0.5374	6
6	0.3092	0.3787	7
7	0.2287	0.2945	8
8	0.1690	0.2263	9
9	0.1208	0.1652	10
10	0.0787	0.0975	11
11	0.0474	0.0653	12
12	0.0305	0.0426	13
13	0.0162	0.0053	14
14	0.0018		

Outre les 3 premiers, le 4ème axe est maintenant jugé pertinent avec cette procédure. En effet : $\lambda_4^{0.05} = 0.9662 > \lambda_5^{0.95} = 0.8790$.

La présentation est plus intuitive avec un graphique.



Visiblement, le décalage entre les distributions s'amenuise lorsqu'on s'éloigne des premiers axes.

9 Interprétation du 4ème axe

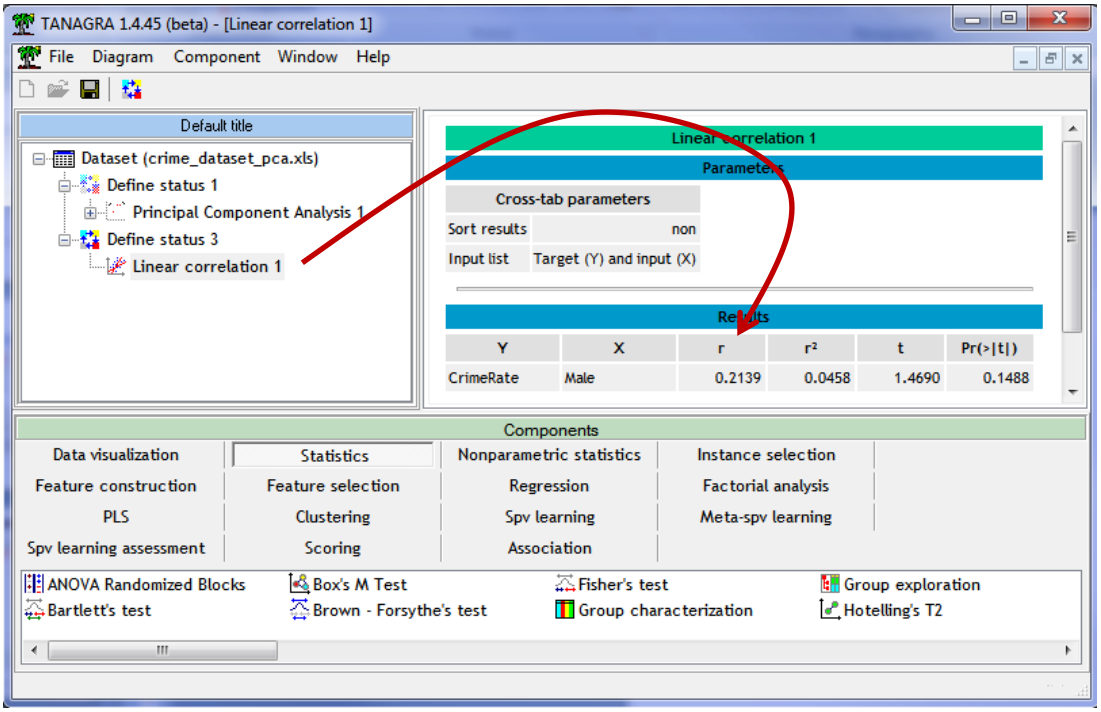
Les 3 premiers axes sont incontestables. Le 5ème et les suivants peuvent être éliminés. Reste le 4ème. Mais qu'y a-t-il derrière ce facteur ?

Reprenons les résultats détaillés de Tanagra.

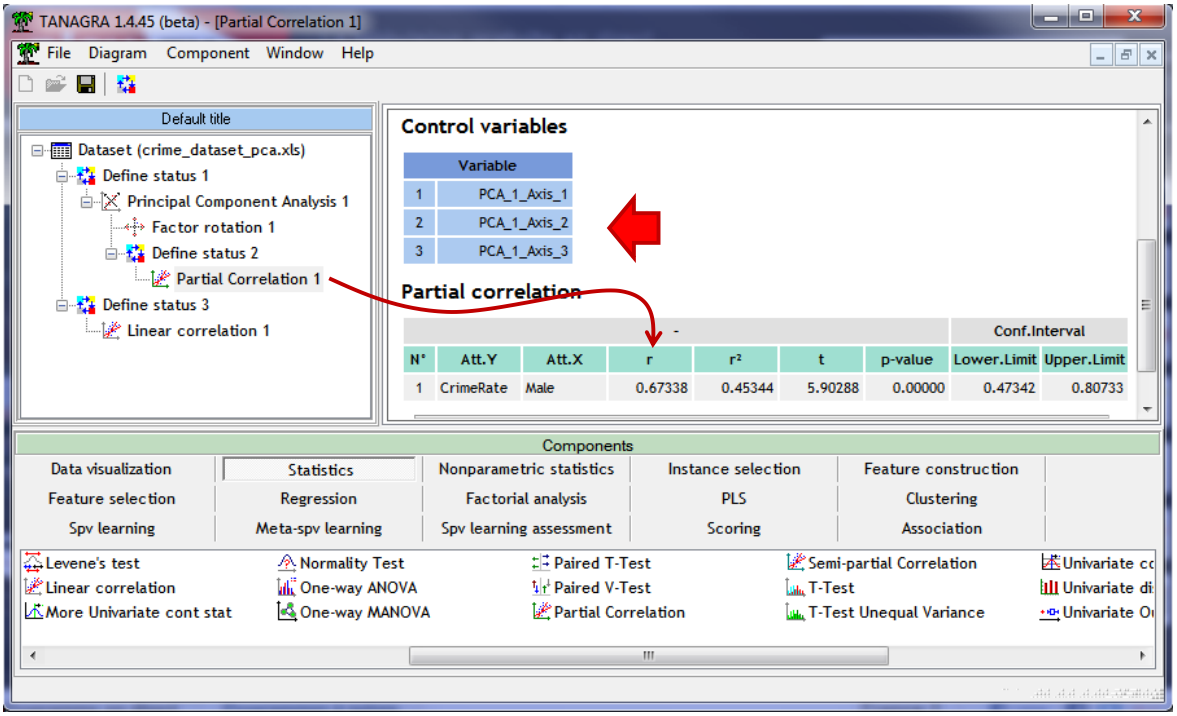
Attribute	Axis_1		Axis_2		Axis_3		Axis_4	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)
CrimeRate	0.4721	22 % (22 %)	-0.4198	18 % (40 %)	0.2710	7 % (47 %)	-0.6288	40 % (87 %)
Male14-24	-0.7332	54 % (54 %)	0.0781	1 % (54 %)	0.2781	8 % (62 %)	-0.3600	13 % (75 %)
Southern	-0.7788	61 % (61 %)	-0.3680	14 % (74 %)	0.1530	2 % (77 %)	-0.1726	3 % (80 %)
Education	0.8375	70 % (70 %)	0.3591	13 % (83 %)	0.0767	1 % (84 %)	-0.0701	0 % (84 %)
Expend60	0.7952	63 % (63 %)	-0.5002	25 % (88 %)	0.2084	4 % (93 %)	-0.1400	2 % (95 %)
Expend59	0.7991	64 % (64 %)	-0.4915	24 % (88 %)	0.2117	4 % (92 %)	-0.1144	1 % (94 %)
Labor	0.4283	18 % (18 %)	0.5836	34 % (52 %)	0.3219	10 % (63 %)	-0.2945	9 % (71 %)
Male	0.3001	9 % (9 %)	0.5307	28 % (37 %)	-0.2615	7 % (44 %)	-0.6774	46 % (90 %)
PopSize	0.2875	8 % (8 %)	-0.7152	51 % (59 %)	0.1597	3 % (62 %)	0.1789	3 % (65 %)
NonWhite	-0.6819	47 % (47 %)	-0.4572	21 % (67 %)	0.2470	6 % (74 %)	-0.2809	8 % (81 %)
Unemp14-24	0.0952	1 % (1 %)	-0.0937	1 % (2 %)	-0.9321	87 % (89 %)	-0.2159	5 % (93 %)
Unemp35-39	0.0598	0 % (0 %)	-0.5733	33 % (33 %)	-0.7451	56 % (89 %)	-0.1624	3 % (91 %)
FamIncome	0.9378	88 % (88 %)	-0.1075	1 % (89 %)	0.0306	0 % (89 %)	0.0642	0 % (90 %)
IncUnderMed	-0.8864	79 % (79 %)	-0.0986	1 % (80 %)	0.0410	0 % (80 %)	-0.2442	6 % (86 %)
Var. Expl.	5.8382	42 % (42 %)	2.6402	19 % (61 %)	1.9535	14 % (75 %)	1.3856	10 % (84 %)

Il indiquerait une liaison entre la présence des hommes (MALE : « The number of males per 1000 females ») et la criminalité (CRIMERATE : « # of offenses reported to police per million population »). Diantre, les hommes, parce qu'ils manquent de femmes, seraient d'humeur belliqueuse ? Ah oui, il ne fallait surtout pas passer à côté de cette information effectivement.

Attention, la liaison est mesurée en contrôlant l'effet des 3 premiers facteurs. Cette nuance n'est pas toujours comprise par les étudiants. En effet, si nous mesurons la corrélation brute entre MALE et CRIME, nous avons r(MALE,CRIME) = 0.2139, non significative à 5%.



En revanche, lorsque nous calculons la corrélation partielle[6], conditionnellement aux trois premiers facteurs, r(MALE, CRIME / FACT.1, FACT.2, FACT.3) = 0.67338. Elle est très significative[7].



Indubitablement, cet axe est porteur d'information. Mais, il faut bien connaître le contexte de l'étude – et les Etats-Unis des années 60 - pour produire une interprétation viable de ce phénomène.

10 Conclusion

Ce document repose beaucoup sur les articles accessibles en ligne référencés en bibliographie, celui de Jackson (1993) notamment que je trouve particulièrement bien écrit. Que leurs auteurs en soient remerciés. Je me démarque essentiellement sur deux points : (1) les techniques sont détaillées sur un exemple[8] ; (2) tous les outils (feuille Excel, code source du programme R) sont accessibles, le lecteur peut reproduire les calculs, voire les appliquer sur ses propres fichiers (le code R a été rendu aussi générique que possible).

J'ai beaucoup apprécié écrire ce tutoriel. Jusqu'à présent, pour déterminer le nombre de facteurs, je me contentais de mixer la lecture de la « scree plot », de la courbe de l'inertie cumulée et du niveau des valeurs propres. Ces éléments sont intégrés dans les sorties actuelles de Tanagra.

Je connaissais la règle de Karlis et al. (2003) parce que j'avais lu en détail le livre de Saporta (2006) à sa sortie. Je connaissais vaguement le test des bâtons brisés parce qu'un collègue m'en avait parlé

[6] Voir http://tutoriels-data-mining.blogspot.fr/2008/06/corrlation-partielle.html

[7] Le calcul des degrés de liberté est très approximative ici puisque les facteurs sont des variables synthétiques.

[8] Il est dommage que dans les articles dits « scientifiques », on ne prenne pas la peine d'expliciter les techniques à partir d'un exemple didactique. Cela aiderait énormément le lecteur. Certes, il faut reconnaître que la place est limitée dans les revues et les « proceedings » (et en plus, on nous fait payer très cher les extra-pages, c'est déprimant), on est malheureusement obligé d'aller à l'essentiel sans prendre le temps de se faire comprendre.

lors d'une discussion à bâtons rompus (*ok, ok, elle est très facile celle là*). Mais tant que je ne l'avais pas explicitement évalué sur un fichier, je ne pouvais pas réellement me faire un avis. Ces deux tests seront certainement incorporés dans Tanagra dans un futur proche.

Je suis moins enthousiaste en revanche concernant les approches basées sur les techniques de ré-échantillonnage. Leur qualité scientifique n'est pas en cause. Mais leur mise en œuvre sur des grands fichiers, dans une phase exploratoire où l'on cherche des solutions en tentant différentes analyses, peut se révéler rapidement rédhibitoire car très gourmande en temps de calcul. A l'inverse, il n'y a pas de problèmes si l'on travaille sur des petites bases. Pour notre fichier, les 1000 réplications bootstrap n'ont pris que quelques secondes.

Finalement, mis à part le test de Bartlett, toutes les stratégies se valent à peu près sur notre fichier. Comme d'habitude en statistique exploratoire, il n'y a pas de vérités absolues. Il y a seulement des pistes que nous devons étudier attentivement en le reliant aux caractéristiques (objectifs, contexte) de notre étude et de nos données. Comme d'habitude également, nous ne pouvons nous passer de l'expertise métier pour valider les résultats.

11 Bibliographie

A. Crawford, S. Green, R. Levy, W. Lo, L. Scott, D. Svetina, M. Thompson, "Evaluation of Parallel Analysis for Determining the Number of Factors", in Educational and Psychological Measurement, 70(6), pp. 885-901, 2010.

S. Franklin, D. Gibson, P. Robertson, J. Pohlmann, F. Fralish, "Parallel Analysis: A Method for Determining Significant Principal Components", in Journal of Vegetation Science, Vol. 6, Issue 1, pp. 99-106, 1995.

G. Grossman, D. Nickerson, M. Freeman, "Principal Component Anayses of Assemblage Structure: Utility of Tests based on Eigenvalues", in Ecology, 72(1), pp. 341-347, 1991.

F. Husson, S. Lê, J. Pagès, "Analyse de données avec R", PUF, 2009.

D. Jackson, "Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches", in Ecology, 74(8), pp. 2204-2214, 1993.

P. Neto, D. Jackson, K. Somers, "How Many Principal Components? Stopping Rules for Determining the Number of Aon-trivial Axes Revisited", in Computational Statistics & Data Analysis, 49(2005), pp. 974-997, 2004.

G. Saporta, « Probabilités, analyses des données et Statistiques », Dunod, 2006.