Objectif

Montrer l'utilisation de la rotation VARIMAX en analyse en composantes principales.

Fichier

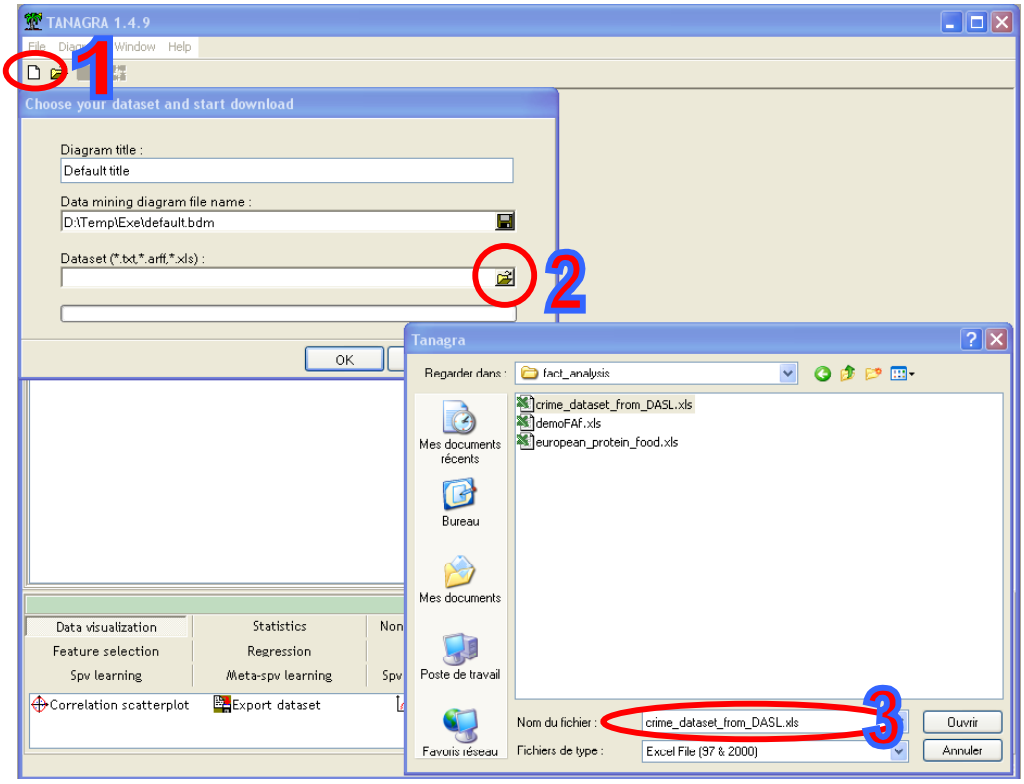
Nous traitons le fichier US CRIME DATAFILE en provenance du site DASL (http://lib.stat.cmu.edu/DASL/Datafiles/USCrime.html).

L'objectif est de comprendre les tenants et aboutissants du taux de criminalité aux USA en 1960. Les observations correspondent aux 47 Etats de 1960.

Rotation VARIMAX

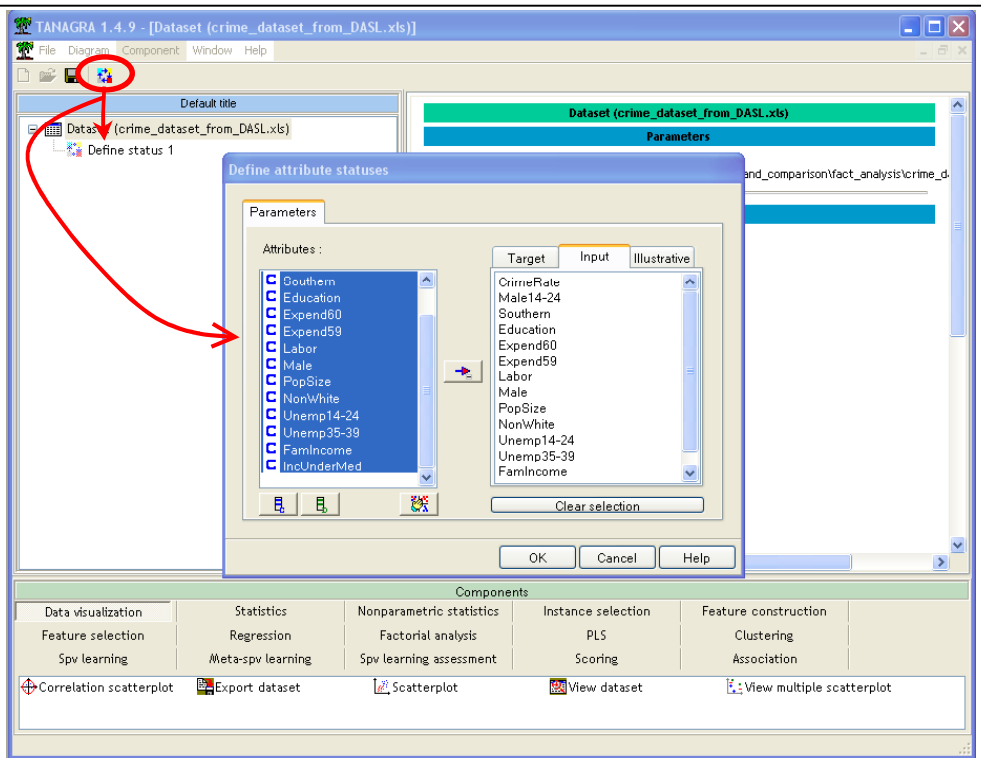
Charger les données

Nous devons dans un premier temps créer un diagramme et charger les données. Pour ce faire, nous cliquons sur le menu FILE/NEW. Nous sélectionnons le fichier CRIME_DATASET_FROM_DASL.XLS, au format EXCEL.



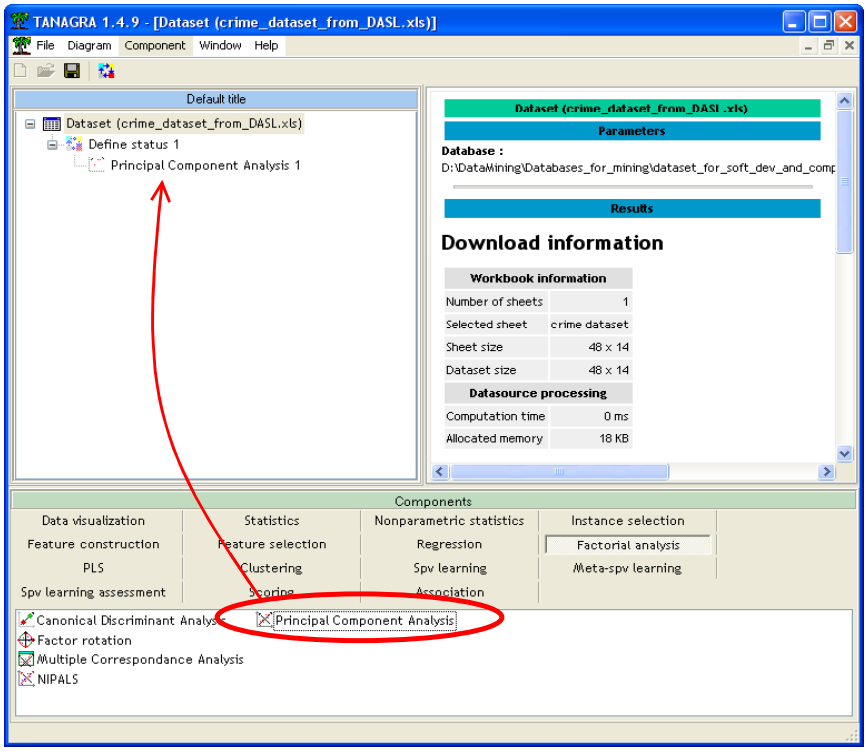
Définir le problème

L'étape suivante consiste à désigner, à l'aide du composant DEFINE STATUS, les variables actives (INPUT) : nous les sélectionnons toutes.



ACP

Nous plaçons l'analyse en composantes principales. Attention, l'idée n'est pas de procéder à une analyse fine dans ce didacticiel mais plutôt d'expliquer très sommairement pourquoi dans certains cas la rotation des axes de l'ACP peut s'avérer bénéfique, et comment procéder avec TANAGRA.



L'exécution (Menu VIEW) montre que les 4 premiers axes factoriels résument 84% de l'information disponible.

Eigen values

Matrix trace = 14.00

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	-	-	-

Pour interpréter les axes, nous nous tournons vers les corrélations entre les variables initiales et les axes factoriels (Voir **Factor Loadings and Communality Estimates –** Nous avons réduit l'affichage aux 4 premiers axes pour plus de lisibilité).

Factor Loadings [Communality Estimates]

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 %)*

Quelques éléments pour la lecture des résultats :

- « CORR » indique la corrélation entre la variable et l'axe factoriel. Les corrélations supérieures à 0.70 en valeur absolue sont signalées en rouge.
- « % » indique le COS^2, ici il s'agit du carré de CORR. Entre parenthèses, le cumul du COS^2. A priori, si nous sélectionnons tous les axes, ce cumul doit être égal à 100%.
- En bleu, le pourcentage d'inertie expliquée sur chaque axe. Les valeurs doivent coïncider avec celles lues dans le tableau EIGENVALUES.

Une des difficultés de l'ACP est l'interprétation des axes. Dans cet exemple, si nous nous tenons aux deux premiers axes factoriels, nous constatons que :

- Le premier axe oppose d'une part les états du sud (Southern) avec une forte proportion de personnes à bas salaires (IncUnderMed) et une jeunesse essentiellement masculine (Male14-24) aux, d'autre part, états composés de familles à hauts revenus (FamilyIncome) avec un niveau d'éducation élevé (Education), pour lesquelles les dépenses liées à la sécurité sont élevés (Expend).
- Le second axe n'est pas facile à lire. On croit y noter que les états avec beaucoup d'hommes actifs sont peu peuplés.

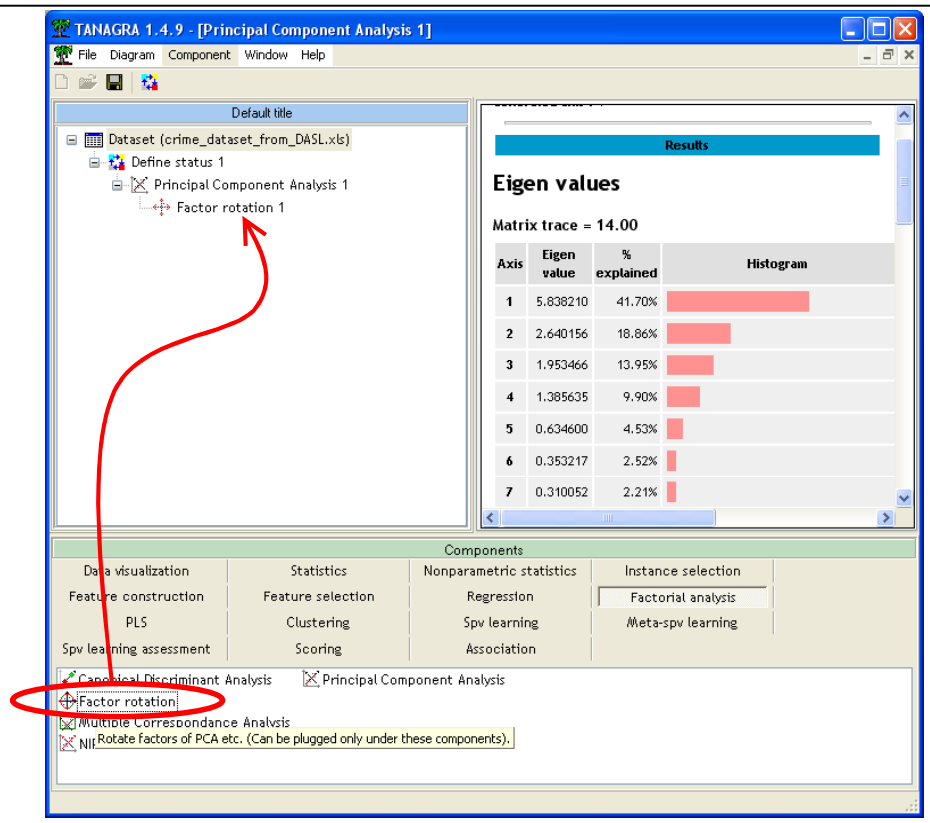
Voici un résultat assez étrange. Il semble indiquer que plus un état est peuplé de personnes à hauts revenus ayant suivi des études longues, plus les autorités consacrent de l'argent à la police, sans que le taux de criminalité ne soit dramatiquement élevé par ailleurs.

Rotation VARIMAX

Nous notons surtout qu'il y a de nombreuses variables avec des corrélations moyennes sur ces deux premiers axes factoriels (autour de 0.5 en valeur absolue), rendant l'interprétation des axes laborieuse. Le rôle des méthodes de rotation est justement de rendre les valeurs de ces corrélations plus tranchées en faisant pivoter les axes. De fait, leur lecture en sera facilitée.

La rotation VARIMAX fait tourner les axes en préservant leur orthogonalité. Elle cherche à maximiser la variance des corrélations dans chaque colonne de notre tableau ci-dessus. Nous pouvons le paramétrer en fixant le nombre d'axes à traiter.

Nous plaçons le composant à la suite de l'ACP.



Après exécution (menu VIEW), le composant renvoie le tableau des corrélations après et avant rotation.

Rotated Factor Loadings

Attribute	Axis_1		Axis_2	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)
CrimeRate	0.0784	1 % (1 %)	0.6269	39 % (40 %)
Male14-24	-0.5000	25 % (25 %)	-0.5420	29 % (54 %)
Southern	-0.8283	69 % (69 %)	-0.2365	6 % (74 %)
Education	0.8665	75 % (75 %)	0.2819	8 % (83 %)
Expend60	0.2684	7 % (7 %)	0.9003	81 % (88 %)
Expend59	0.2771	8 % (8 %)	0.8963	80 % (88 %)
Labor	0.7068	50 % (50 %)	-0.1567	2 % (52 %)
Male	0.5755	33 % (33 %)	-0.2014	4 % (37 %)
PopSize	-0.2551	7 % (7 %)	0.7274	53 % (59 %)
NonWhite	-0.8142	66 % (66 %)	-0.1056	1 % (67 %)
Unemp14-24	0.0098	0 % (0 %)	0.1332	2 % (2 %)
Unemp35-39	-0.3329	11 % (11 %)	0.4706	22 % (33 %)
FamIncome	0.6344	40 % (40 %)	0.6989	49 % (89 %)
IncUnderMed	-0.7316	54 % (54 %)	-0.5100	26 % (80 %)
Var. Expl.	4.4492	32 % (32 %)	4.0292	29 % (61 %)

vs. Unrotated Factor Loadings

Attribute	Axis_1		Axis_2	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)
CrimeRate	0.4721	22 % (22 %)	-0.4198	18 % (40 %)
Male14-24	-0.7332	54 % (54 %)	0.0781	1 % (54 %)
Southern	-0.7788	61 % (61 %)	-0.3680	14 % (74 %)
Education	0.8375	70 % (70 %)	0.3591	13 % (83 %)
Expend60	0.7952	63 % (63 %)	-0.5002	25 % (88 %)
Expend59	0.7991	64 % (64 %)	-0.4915	24 % (88 %)
Labor	0.4283	18 % (18 %)	0.5836	34 % (52 %)
Male	0.3001	9 % (9 %)	0.5307	28 % (37 %)
PopSize	0.2875	8 % (8 %)	-0.7152	51 % (59 %)
NonWhite	-0.6819	47 % (47 %)	-0.4572	21 % (67 %)
Unemp14-24	0.0952	1 % (1 %)	-0.0937	1 % (2 %)
Unemp35-39	0.0598	0 % (0 %)	-0.5733	33 % (33 %)
FamIncome	0.9378	88 % (88 %)	-0.1075	1 % (89 %)
IncUnderMed	-0.8864	79 % (79 %)	-0.0986	1 % (80 %)
Var. Expl.	5.8382	42 % (42 %)	2.6402	19 % (61 %)

Nous constatons :

- La quantité d'information traduite sur les deux premiers axes reste la même (61%) mais la répartition entre les deux axes a été fortement modifiée. Le second axe traduit maintenant 29% de l'inertie contre 19% auparavant.
- Les états du sud à forte population non-blanche et des salaires faibles sont clairement opposés aux régions à fort niveau d'éducation et actifs (Labor) sur le premier axe.
- Les dépenses de police (Expend) sont en réalité liées à la taille de la population (PopSize) et à la criminalité (Crime). Et ceci indépendamment de la dichotomie nord/sud et blancs/non-blancs du premier axe puisque les facteurs sont orthogonaux. Il reste que ces dépenses sont liées positivement avec le revenu moyen des familles (FamIncome).

Il faut néanmoins se méfier de ces techniques. Une telle différence de lecture des résultats avant et après rotation peut aussi être le reflet d'anomalies plus sournoises. On retrouve souvent deux causes possibles : soit une variable très importante manque dans l'ensemble de données, dans notre cas, la proportion des personnes vivant dans les zones urbaines et rurales peut jouer un rôle important, nous ne disposons pas de cette information ; soit quelques points atypiques faussent complètement les calculs en étirant exagérément le nuage de points.

Conclusion

L'ACP est un outil de visualisation très populaire. La difficulté à interpréter les axes factoriels est un des freins à son utilisation. Avec les méthodes de rotation d'axes, nous disposons d'un outil supplémentaire pour améliorer la lisibilité des résultats.