EISTI Ecole Internationale des Sciences du Traitement de l'Information	**ING2-SIE : ANALYSE DES DONNEES : CORRIGE ACP-OCDE**
Rédigé par : Hervé de Milleville	Ref : *ING2-SIE-ADD-ACP-OCDE*
A l'intention de : Etudiants des ING2-SIE	Créé le : 01/04/2012

Présentation de l'étude

On fait une étude socio-économique sur 18 pays de l'OCDE. On cherche les différences les plus significatives entre ces pays. Chaque variable est une mesure quantitative sociologique ou économique. On utilisera donc une analyse en composantes principales.

Vous trouverez en annexe les données non normalisées.

Analyse : Les résultats avec SAS et l'interprétation des résultats

1. *Préalables*

Il est important dans l'interprétation que ce corpus de données date de 1974?

Dans SAS, nous avons importé le fichier de données dans le répertoire sasuser. Ce fichier s'appelle "ocde"

Nous avons exécuté les deux commandes princomp et factor :

```
ods graphics on;
proc princomp data=sasuser.ocde plots= score(ellipse ncomp=5);
id pays;
var popu--tv;
run;
ods graphics off;

PROC FACTOR data=sasuser.ocde method=prin;
var popu -- tv;
run;
```

2. *Choix des axes*

Résultats obtenus par SAS : Le tableau ci-dessous est présent dans les 2 procédures

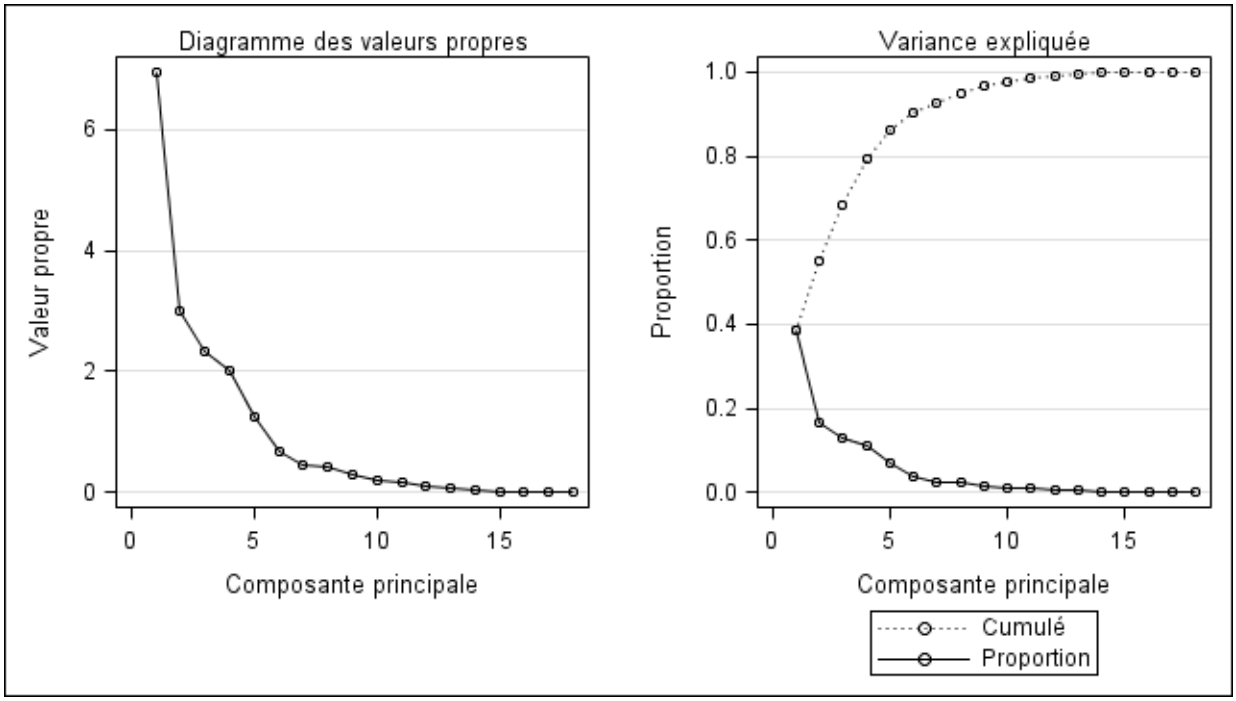
Valeurs propres de la matrice de corrélation

	Valeur propre	Différence	Proportion	Cumulé
1	6.95156132	3.94514420	0.3862	0.3862
2	3.00641712	0.66090913	0.1670	0.5532
3	2.34550799	0.34075570	0.1303	0.6835
4	2.00475229	0.75722689	0.1114	0.7949
5	1.24752540	0.57596317	0.0693	0.8642
6	0.67156223	0.21190948	0.0373	0.9015
7	0.45965275	0.04491107	0.0255	0.9271
8	0.41474168	0.12549771	0.0230	0.9501
9	0.28924397	0.08709489	0.0161	0.9662
10	0.20214908	0.02404779	0.0112	0.9774
11	0.17810129	0.08343955	0.0099	0.9873
12	0.09466174	0.00563801	0.0053	0.9925
13	0.08902373	0.06053082	0.0049	0.9975
14	0.02849292	0.01484650	0.0016	0.9991
15	0.01364642	0.01130984	0.0008	0.9998
16	0.00233658	0.00171309	0.0001	1.0000
17	0.00062349	0.00062349	0.0000	1.0000
18	0.00000000		0.0000	1.0000

Il y a plusieurs méthodes pour retenir les axes

1. On fixe nous même un seuil du % de dispersion à récupérer
2. On retient tous les axes dont la dispersion récupérée est strictement supérieure à 1/p (p : le nombre de variables). Dans notre cas, 1/p vaut 5,55%. Dans cette étude, on retiendrait 5 axes.
3. On étudie le diagramme des valeurs propres pour localiser un coude franc dans le graphique

Résultats de SAS : le diagramme est obtenu avec la procédure princomp



On retiendrait encore 5 axes.

4. On ne retient pas un axe quand on n'arrive pas à trouver une signification. On est souvent confronté à ce problème dès qu'on arrive au 4ème ou au delà.

Dans notre cas, on les 5 premiers axes. On n'interprétera que les 3 premiers axes.

3. *Interprétation des axes*

On cherche le tableau des corrélations variables-facteurs. Dans la procédure factor, SAS appelle ces corrélations pour un facteur, la représentation du facteur.

Pour chaque axe, on ne retient que les plus fortes corrélations en valeurs absolues.

Axe 1 :

En corrélation positive : TV (91,80%), IMPT (86,20%), EXPT (86,30%) et PNB/HAB (83,20%)
En corrélation négative : AASP (-86,14%) et PIBA (-90,20%)

Cet axe oppose les pays riches avec une économie ouverte vers l'extérieur (partie positive de l'axe) aux pays plutôt caractérisés par une forte agriculture (PIBA) traditionnelle (AASP). Ce propos est bien sûr relatif à l'ensemble des pays de l'étude.

Axe 2 :

En corrélation positive : RECC (78,80%) et EDUC (71,72%)
En corrélation négative : POPU (-62,99%) et RESO (-58,24%)

Une première remarque : les corrélations sont moins forte que pour le premier axe. C'est normal car cet axe explique nettement moins de dispersion que le premier axe.

RECC signifie recettes courantes de l'état et EDUC dépenses publiques d'éducation. RESO signifie réserves en dollars. Cet axe oppose les pays avec une politique économique plutôt étatique (RECC, EDUC) à une politique économique plutôt libérale (RESO). Il y a un problème dans cette interprétation car on n'a pas tenu compte de la variable POPU. Dans ce cas, il faut récupérer la corrélation entre RESO et POPU. On la lit dans la matrice de corrélations variables x variables et dans ce cas ρ(RESO, POPU) = 0.7834. Cette corrélation est très forte et positive. La conséquence est que chaque axe fortement corrélé avec la variable RESO sera aussi fortement corrélé (et dans le même sens) avec la variable POPU.
Il faut simplement constater qu'en 1974 les pays d'économie libérale avaient plutôt une population importante et les pays d'économie étatique avait plutôt une faible population. En se référant aux mapping (voir annexes) contenant l'axe 2, on voit les USA, la France et l'Allemagne sur la partie négative de l'axe et la Suède, la Norvège et le Danemark sur la partie positive de l'axe.

Axe 3 :
En corrélation positive : FBCF (66,82%) et LOG (71,65%)
En corrélation négative : CAL (-70,41%)

FBCF signifie Formation Brute de Capital Fixe(outils de production) et LOG nombre de logements achevés dans l'année. CAL signifie nombre de calories consommées par habitant.

On peut se lancer dans l'explication suivante : Cet axe rassemble des variables avec lesquelles il est difficile de donner une explication générale. Il doit probablement caractériser certains pays mal représentés sur les deux principaux axes. On peut constater que le Japon se situe sur la partie positive de l'axe. Or ce pays était réputé à l'époque pour être très productif (FBCF fort) et croissance démographique forte ($\Rightarrow$ LOG fort).

Corrélations variables/Facteurs
Résultats obtenus par SAS : Le tableau ci-dessous provient de la procédure factor

Représentation du facteur

	Factor1	Factor2	Factor3	Factor4	Factor5
POPU	0.64728	**-0.62992**	0.17412	-0.01409	0.16923
DENS	0.17470	-0.10803	-0.31363	0.81539	-0.18009
TATO	0.53398	-0.28031	0.38121	-0.12220	-0.03495
AASP	**-0.86145**	-0.32006	0.24376	-0.11269	0.12253
AIND	0.49560	0.18289	-0.48836	0.47481	-0.40270
PNB	0.83274	0.01741	0.24043	-0.30003	-0.05343
PIBA	**-0.90201**	-0.21784	0.04445	-0.19465	0.16814
FBCF	-0.21785	0.11181	**0.66822**	0.59049	0.07958
RECC	0.46760	**0.78820**	-0.15055	-0.08028	-0.09902
RESO	0.71612	**-0.58240**	-0.04448	-0.03058	-0.03546
TESC	0.10630	0.30572	-0.25078	0.31330	0.80898
IMPT	**0.86202**	-0.45314	-0.03808	0.09263	0.12412
EXPT	**0.86306**	-0.46354	-0.00780	0.09153	0.10170
CAL	0.20341	0.09451	**-0.70411**	-0.37910	0.33685
LOG	0.28274	0.36277	**0.71650**	0.34227	0.22397
ELEC	0.51853	0.39877	0.39760	-0.46404	-0.24302
EDUC	0.45931	**0.71727**	0.09829	-0.04500	0.10633
TV	**0.91804**	0.18731	0.03681	-0.09655	0.21293

4. Caractéristiques des individus

Comme pour l'AFC, une fois l'interprétation des axes terminés, on se doit de caractériser chaque individu par rapport à l'ensemble des autres individus. On se sert pour cela des cos^2 entre les individus et chaque axe. Plus ce cos^2 est grand mieux l'individu est représenté sur cet axe. La méthode consiste pour chaque individu de chercher les axes sur lesquels il est bien représenté puis sur chacun de ces axes de regarder sa position et de se référer à l'interprétation des axes pour conclure. Cette étude n'a de sens que si le nombre d'individus est petit et que chaque individu est clairement identifié.

En l'absence de ces valeurs, on peut malgré tout se risquer à caractériser les USA en regardant ses positions sur les mappings (voir annexe).
Ce pays se trouve très éloigné sur la partie positive de l'axe 1 et assez éloigné sur la partie négative de l'axe 2. En revanche, il est quasiment proche de l'origine sur l'axe 3. Il est très sûrement très bien représenté sur l'axe 1, bien représenté sur l'axe 2 et peu représenté sur l'axe 3.
En conclusion, les USA sont caractérisés en tant que pays riche, ouvert vers l'extérieur et avec une économie plutôt libérale.

1. Annexes

Le corpus de données Pays x Variables socio-économiques

PAYS	POPU	DENS	TATO	AASP	AIND	PNB	PIBA	FBCF	RECC
D	60848	245	1,05	9,6	49,1	2520	3,6	24,4	37,9
A	7373	88,1	0,5	13,1	39,9	1690	7	23,2	37,5
B	9984	332	0,6	5,4	44,8	2352	5,4	23,1	35,1
CDN	21689	2	1,85	8,2	32,3	3460	5,9	21,7	35,2
DK	4893	114	0,75	11,9	38,5	28,6	8,9	22,6	37,1
E	32949	65	0,95	30,7	37,1	870	15	22	22,4
USA	203213	22	1,35	4,6	33,7	4660	2,9	16,7	31,5
SF	4786	14	0,7	24,5	34,6	1940	14,7	23,8	35,9
F	50235	91	1,05	15,1	40,6	2770	6	25,4	38,1
GR	8866	67	0,7	48,2	22,5	950	20,3	29,7	26,9
SE	2921	42	0,25	28,4	29,7	1040	19,7	19,9	30,7
I	54120	180	0,85	21,5	43,1	1520	11,3	20,5	33,3
JAP	102380	277	1,05	18,8	35	1630	8,7	35,2	21,2
N	3851	12	0,8	14,7	36,8	2530	6,5	25,3	43,4
NL	12873	352	1,25	7,5	41,3	2190	7	25,5	41,9
P	9583	105	0,9	31,5	35,5	600	17,7	18,4	24
RUN	55643	228	0,65	2,9	46,8	1970	3	17,3	39
S	7969	18	0,7	8,8	40,4	3230	5,9	23,6	48,1

PAYS	RESO	TESC	IMPT	EXPT	CAL	LOG	ELEC	EDUC	TV
D	10940	6,5	24926	29052	2990	8,6	3322	3	231
A	1563	5	2825	2412	2990	6,6	2647	4,4	134
B	2406	7	9984	10069	3150	5	2814	5,3	184
CDN	3846	6	13137	13754	3160	8,2	8199	5,7	279
DK	384	9	3800	2958	3180	9	2413	6	244
E	1518	6,5	4233	1900	2750	6,4	1245	2,1	84
USA	12306	5,75	36052	37988	3210	7,7	7013	5,1	392

SF	379	6	2023	1985	2900	7,9	3836	6,3	193
F	617	7,5	17392	15028	3160	8,2	2407	4,3	185
GR	290	6,5	1594	554	2910	10,1	823	2,4	9
SE	694	7,31	1413	891	3450	4	1577	4,2	111
I	4642	5,5	12450	11729	2940	5,1	1810	5,8	146
JAP	3072	6	15024	15990	2460	11,9	2734	4,5	190
N	607	4,5	2943	2203	2910	8,8	12976	5,8	175
NL	2621	6	10991	9965	3240	9,7	2565	6,7	197
P	1442	3,5	1231	823	2930	4,3	607	1,4	29
RUN	2469	7	19956	17515	3180	7,7	3680	4,2	263
S	506	7	5899	5698	2750	13,4	6803	7,4	288

La matrice de corrélations entre les variables

Matrice de corrélation : Ce tableau ci-dessous provient de la procédure factor

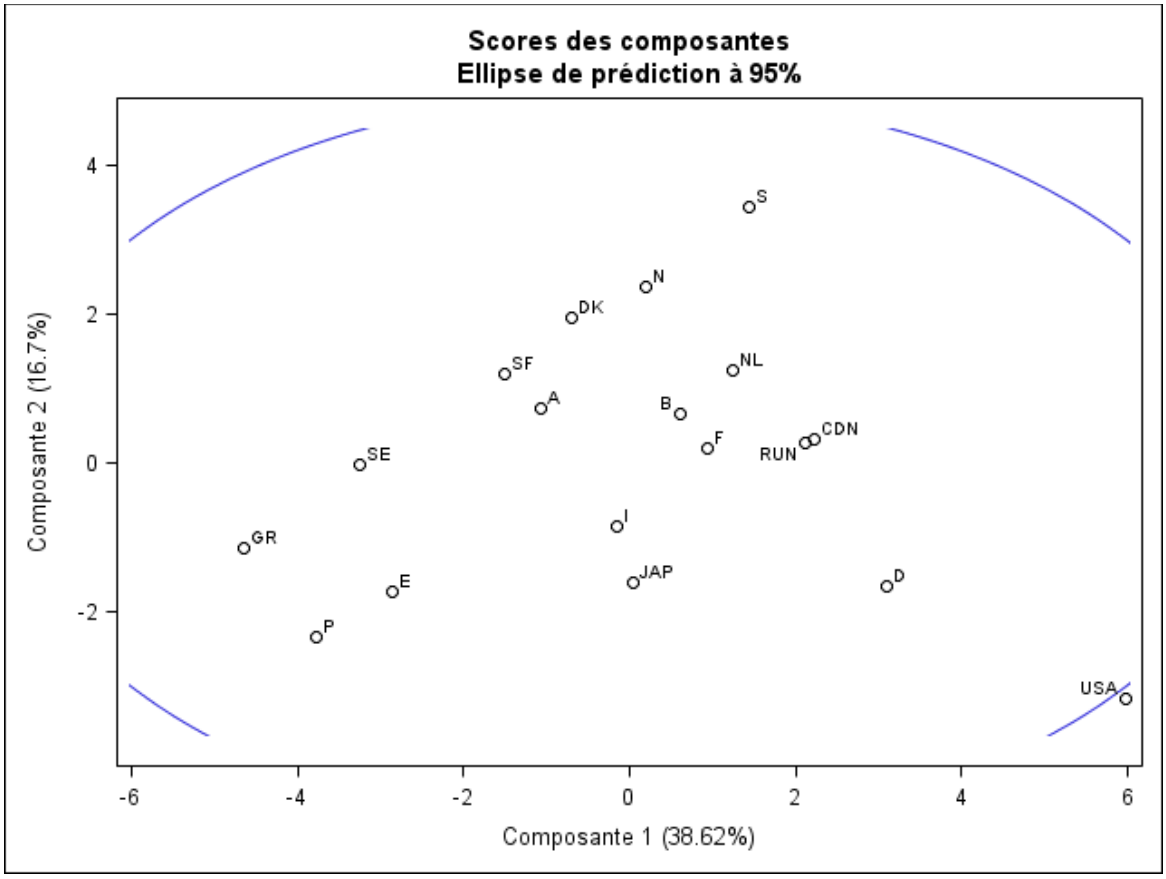
		POPU	DENS	TATO	AASP	AIND	PNB	PIBA	FBCF	RECC
POPU	POPU	1.0000	0.0618	0.4251	-.2989	0.0482	0.5444	-.4092	-.1182	-.2608
DENS	DENS	0.0618	1.0000	-.0068	-.3145	0.5745	-.1371	-.2974	0.2594	-.0264
TATO	TATO	0.4251	-.0068	1.0000	-.3018	-.0305	0.5271	-.3907	0.0330	-.0543
AASP	AASP	-.2989	-.3145	-.3018	1.0000	-.6829	-.6107	0.9320	0.2614	-.6228
AIND	AIND	0.0482	0.5745	-.0305	-.6829	1.0000	0.1567	-.6729	-.2183	0.4682
PNB	PNB	0.5444	-.1371	0.5271	-.6107	0.1567	1.0000	-.6812	-.1497	0.4294
PIBA	PIBA	-.4092	-.2974	-.3907	0.9320	-.6729	-.6812	1.0000	0.0912	-.5826
FBCF	FBCF	-.1182	0.2594	0.0330	0.2614	-.2183	-.1497	0.0912	1.0000	-.1554
RECC	RECC	-.2608	-.0264	-.0543	-.6228	0.4682	0.4294	-.5826	-.1554	1.0000
RESO	RESO	0.7834	0.1756	0.4777	-.4000	0.2748	0.5652	-.4782	-.2539	-.0563
TESC	TESC	-.0597	0.1109	-.1747	-.1872	0.1145	-.1149	-.1279	0.0713	0.2019

Matrice de corrélation

	RESO	TESC	IMPT	EXPT	CAL	LOG	ELEC	EDUC	TV
POPU	**0.7834**	-.0597	0.8682	0.8602	-.0525	0.1027	0.1243	-.0566	0.5498
DENS	0.1756	0.1109	0.2551	0.2508	0.0030	0.0006	-.3893	0.0016	-.0121
TATO	0.4777	-.1747	0.5212	0.5252	-.0304	0.2521	0.3263	0.0982	0.4301
AASP	-.4000	-.1872	-.5766	-.5691	-.2920	-.1751	-.4610	-.5817	-.8407
AIND	0.2748	0.1145	0.3669	0.3546	0.0489	-.0566	0.0040	0.1983	0.3526
PNB	0.5652	-.1149	0.6847	0.6911	0.1309	0.2792	0.6446	0.4203	0.7436
PIBA	-.4782	-.1279	-.6794	-.6689	-.1027	-.3256	-.5099	-.4594	-.8057
FBCF	-.2539	0.0713	-.1927	-.1637	-.5883	0.6113	-.0552	0.0354	-.2495
RECC	-.0563	0.2019	0.0634	0.0574	0.3115	0.3281	0.5277	0.7039	0.5027
RESO	1.0000	-.1318	0.8579	0.9029	0.1187	-.0400	0.1544	-.0741	0.5379
TESC	-.1318	1.0000	0.0897	0.0588	0.3169	0.2328	-.2227	0.2455	0.2914

1.1 Les mappings

Mapping des axes 1 et 2



Mapping des axes 1 et 3