Ana-données – IS – Les réalités sociales françaises à l'aune européenne (Correction)

mardi 19 mars 2024 — durée : 1 heure 30 minutes — documents **non** *autorisés*

1 Reconstitution de la matrice de corrélation (5 points)

On se place dans le cadre de l'ACP sur p variables centrées réduites : on note $\mathbf{Z}$ la table des coordonnées centrées réduites des individus et on suppose un poids uniforme $\frac{1}{n}$. On rappelle que dans ce cas la matrice de variance covariance de $\mathbf{Z}$ est $\mathbf{R} = \frac{1}{n}\mathbf{Z}'\mathbf{Z}$, où $\mathbf{Z}'$ est la matrice transposée de $\mathbf{Z}$. On rappelle aussi la formule de reconstruction $\mathbf{Z} = \sum_{\ell=1}^{p} \mathbf{c}_\ell \mathbf{a}'_\ell$, ou les $\mathbf{a}_k$ sont les axes principaux orthonormés et les composantes principales $\mathbf{c}_k$ satisfont $\mathrm{Var}(\mathbf{c}_k) = \lambda_k$ et sont décorrélées entre elles. On cherche à exprimer $\mathbf{R}$ en fonction des éléments de l'ACP.

Question 1 *Montrez que* $\mathbf{R} = \frac{1}{n}\sum_{k=1}^{p}\sum_{\ell=1}^{p} \mathbf{a}_k\mathbf{c}'_k\mathbf{c}_\ell\mathbf{a}'_\ell$.

On écrit simplement que

$$\mathbf{R} = \frac{1}{n}\mathbf{Z}'\mathbf{Z} = \frac{1}{n}\left(\sum_{k=1}^{p}\mathbf{c}_k\mathbf{a}'_k\right)' \times \left(\sum_{\ell=1}^{p}\mathbf{c}_\ell\mathbf{a}'_\ell\right) = \frac{1}{n}\left(\sum_{k=1}^{p}\mathbf{a}_k\mathbf{c}'_k\right) \times \left(\sum_{\ell=1}^{p}\mathbf{c}_\ell\mathbf{a}'_\ell\right) = \frac{1}{n}\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathbf{a}_k\mathbf{c}'_k\mathbf{c}_\ell\mathbf{a}'_\ell.$$

Question 2 *En déduire que* $\mathbf{R} = \sum_{\ell=1}^{p}\lambda_\ell\mathbf{a}_\ell\mathbf{a}'_\ell$.

D'après les propriétés de variance/covariance de $\mathbf{c}_k$, et en utilisant le fait que $\mathbf{c}_k$ est de moyenne nulle, on peut écrire

$$\mathrm{cov}(\mathbf{c}_k, \mathbf{c}_\ell) = \frac{1}{n}\mathbf{c}'_k\mathbf{c}_\ell = \begin{cases}\lambda_k & \text{si } k = \ell,\\ 0 & \text{sinon.}\end{cases}$$

Il ne reste donc dans la double somme de la question précédente que les termes où $k = \ell$ et

$$\mathbf{R} = \sum_{\ell=1}^{p}\mathbf{a}_k\frac{1}{n}\mathbf{c}'_k\mathbf{c}_k\mathbf{a}'_k = \sum_{\ell=1}^{p}\lambda_k\mathbf{a}_k\mathbf{a}'_k.$$

2 Introduction au jeu de données (2 points)

Le Centre d'Analyse Stratégique à publié en octobre 2007 l'étude « Les réalités sociales françaises à l'aune européenne » décrit comme « *un panorama permettant de positionner la France au sein de l'Union* ». On s'intéresse ici à une partie de ces données, qui sont disponibles pour tous les pays de l'Europe à 25. Les variables considérées sont :

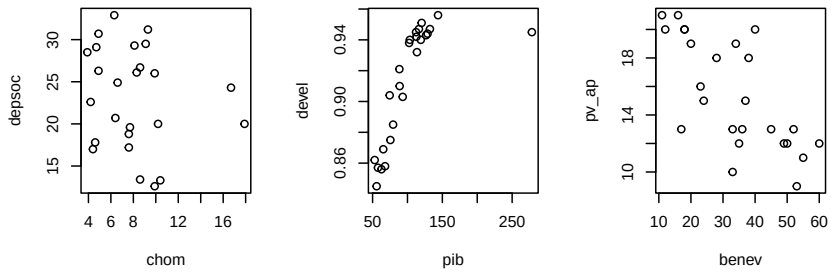
- `benev` : part des personnes exerçant une activité bénévole dans une association ;
- `chom` : taux de chômage des personnes entre 15 et 74 ans ;
- `depedu` : dépenses publiques d'éducation en % du PIB, tous niveaux confondus ;
- `depsoc` : dépenses de protection sociale en % du PIB ;
- `devel` : indice de développement humain par pays (*indice composite, calculé par la moyenne de trois indices : santé/longévité, niveau d'éducation et niveau de vie*) ;
- `pib` : Produit Intérieur Brut par habitant ;
- `pv_apr` : taux de pauvreté après transferts sociaux autre que pensions de vieillesse et de survie ;
- `pv_av` : taux de pauvreté avant transferts sociaux autre que pensions de vieillesse et de survie ;
- `trv_pv` : taux de pauvreté des travailleurs à temps complet.

Les pays concernés sont : `at` (Autriche), `be` (Belgique), `cy` (Chypre), `cz` (République Tchèque), `de` (Allemagne), `dk` (Danemark), `ee` (Estonie), `es` (Espagne), `fi` (Finlande), `fr` (France), `gr` (Grèce), `hu` (Hongrie), `ie` (Irlande), `it` (Italie), `lt` (Lituanie), `lu` (Luxembourg), `lv` (Lettonie), `mt` (Malte), `nl` (Pays-Bas), `pl` (Pologne), `pt` (Portugal), `se` (Suède), `si` (Slovénie), `sk` (Slovaquie), `uk` (Royaume Uni). Parmi ces pays, ceux qui ont rejoint l'Union Européenne en 2004 sont : `cy`, `cz`, `ee`, `hu`, `lt`, `lv`, `mt`, `pl`, `si`, `sk`.

On donne ci-dessous les données collectées, les corrélations des variables et la représentation des couples de variables (`chom`,`depsoc`), (`pib`, `devel`) et (`benev`, `pv_ap`).

	benev	chom	depedu	depsoc	devel	pib	pv_ap	pv_av	trv_pv
at	60	4.7	5.45	29.1	0.944	128.8	12	24	6
be	37	8.1	5.99	29.3	0.945	112.3	15	28	3
cy	23	4.6	6.71	17.8	0.903	93.2	16	22	6
cz	33	7.7	4.42	19.6	0.885	79.4	10	21	3
de	52	9.1	4.60	29.5	0.932	113.6	13	24	4
dk	49	4.9	8.47	30.7	0.943	126.7	12	31	4
ee	38	8.6	5.09	13.4	0.858	67.9	18	24	6
es	18	10.2	4.25	20.0	0.938	102.4	20	24	10
fi	50	8.6	6.43	26.7	0.947	116.4	12	28	3
fr	36	9.3	5.81	31.2	0.942	112.8	13	26	5
gr	18	9.9	4.22	26.0	0.921	88.4	20	23	12
hu	17	6.4	5.43	20.7	0.869	65.3	13	29	8
ie	40	4.4	4.75	17.0	0.956	143.8	20	32	5
it	34	8.3	4.59	26.1	0.940	103.4	19	24	8
lt	11	10.4	5.20	13.3	0.857	57.7	21	26	8
lu	45	4.2	3.93	22.6	0.945	278.3	13	23	9
lv	20	9.9	5.08	12.6	0.845	55.8	19	26	8
mt	24	7.6	4.99	18.8	0.875	75.5	15	21	5
nl	55	3.9	5.18	28.5	0.947	132.2	11	22	6
pl	16	17.9	5.41	20.0	0.862	53.0	21	30	13
pt	12	6.6	5.31	24.9	0.904	74.5	20	26	12
se	53	6.3	7.35	32.9	0.951	120.3	9	29	4
si	35	7.6	5.96	17.2	0.910	88.8	12	26	4
sk	33	16.7	4.21	24.3	0.856	62.7	13	22	9
uk	28	4.9	5.29	26.3	0.940	119.1	18	31	6

	benev	chom	depedu	depsoc	devel	pib	pv_ap	pv_av	trv_pv
benev	1.00	-0.41	0.28	0.58	0.62	0.56	-0.71	0.02	-0.63
chom	-0.41	1.00	-0.28	-0.19	-0.56	-0.55	0.32	-0.07	0.46
depedu	0.28	-0.28	1.00	0.34	0.23	-0.05	-0.36	0.50	-0.45
depsoc	0.58	-0.19	0.34	1.00	0.70	0.36	-0.50	0.14	-0.23
devel	0.62	-0.56	0.23	0.70	1.00	0.69	-0.30	0.23	-0.35
pib	0.56	-0.55	-0.05	0.36	0.69	1.00	-0.32	0.02	-0.18
pv_ap	-0.71	0.32	-0.36	-0.50	-0.30	-0.32	1.00	0.16	0.66
pv_av	0.02	-0.07	0.50	0.14	0.23	0.02	0.16	1.00	-0.09
trv_pv	-0.63	0.46	-0.45	-0.23	-0.35	-0.18	0.66	-0.09	1.00



Question 3 *Parmi toutes les variables, quel est le couple de variables qui sont les plus corrélées entre elles ? Les moins corrélées entre elles ? Les plus opposées ?*

Les variables les plus corrélées entre elles sont `devel` et `depsoc` $(0,70)$. Le niveau de dépenses sociales est très lié au niveau de développement humain.

Les variables les moins corrélées entre elles sont `pv_av` et `benev` $(0,02)$, mais aussi `pv_av` et `pib` avec la même valeur. D'une manière générale, `pv_av` est très peu corrélé avec les autres variables, sauf peut-être `depedu`.

Les variables les plus opposées sont `pv_ap` et `benev` $(-0,71)$. On pourrait imaginer que la présence de nombreux bénévoles permet d'aider à améliorer le niveau de vie des populations pauvres.

Question 4 *Pour chacun des couples (`chom`, `depsoc`), (`pib`, `devel`) et (`benev`, `pv_ap`), commentez la répartition des valeurs, identifiez les éventuels individus « anormaux » et expliquez le lien avec les corrélations mesurées.*

Le nuage correspondant à (`chom`, `depsoc`), est assez uniformément réparti, mis à part deux points qui correspondent à des valeurs très importantes du taux de chômage : en regardant les données, on voit qu'il s'agit de `sk` $(16,7\%)$ et `pl` $(17,9\%)$. Ceci est cohérent avec la corrélation faible.

Le couple (`pib`, `devel`) a une dépendance presque parfaitement linéaire, à ceci près qu'un point a une valeur très élevée de `pib`. Il s'agit de `lu`, qui présente un PIB par habitant record de 278. Il est intéressant de noter que ce point à lui seul semble expliquer une corrélation moyenne de $0,69$ (en fait, si on enlève l'individu `lu` et qu'on recalcule la corrélation de (`pib`, `devel`), on obtient 0.95).

Finalement, le couple (`benev`, `pv_ap`) exhibe une dépendance moyenne compatible avec sa corrélation de $-0,71$. Le taux de corrélation ne recouvre pas la même réalité que (`pib`, `devel`). Il n'y a pas vraiment ici de point singulier.

3 Une première analyse en composantes principales (5 points)

On fait une analyse en composantes principales des données centrées réduites. On donne ci-dessous, pour les 4 premiers axes principaux : les variances des composantes principales, le tableau des corrélations avec des variables, le tableau des coordonnées des individus sur les axes, et enfin le tableau des contributions (en %) des individus à chacun des axes.

Variances

```
[1] 4.05 1.48 1.29 0.93
```

Corrélations

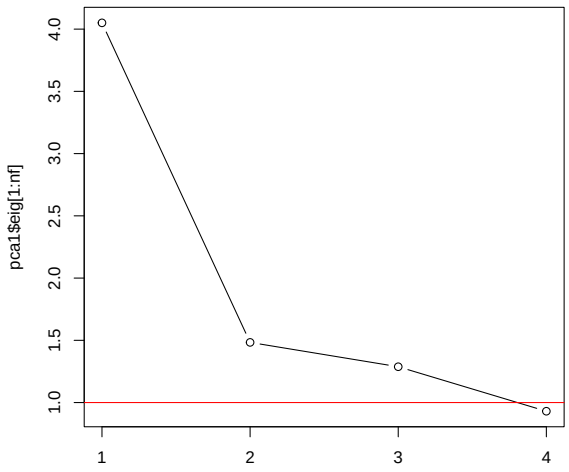
	Comp1	Comp2	Comp3	Comp4
benev	0.87	-0.14	0.18	0.12
chom	-0.66	0.12	0.16	0.62
depedu	0.49	0.77	0.09	-0.03
depsoc	0.70	0.02	-0.15	0.63
devel	0.81	-0.14	-0.46	0.09
pib	0.66	-0.48	-0.40	-0.14
pv_ap	-0.73	0.06	-0.57	-0.14
pv_av	0.18	0.75	-0.52	-0.04
trv_pv	-0.70	-0.21	-0.48	0.29

	Axis1	Axis2	Axis3	Axis4
at	2.39	-0.89	0.26	0.34
be	1.58	0.85	-0.22	0.36
cy	-0.24	0.27	0.69	-1.42
cz	-0.05	-1.03	2.40	-0.54
de	1.46	-0.92	0.69	0.94
dk	3.26	2.54	-0.32	0.07
ee	-1.63	0.02	1.15	-1.10
es	-1.74	-1.04	-1.24	0.22
fi	2.22	0.94	0.39	0.38
fr	1.38	0.27	0.05	1.10
gr	-1.93	-1.18	-1.24	1.05
hu	-1.06	1.11	0.25	-0.53
ie	0.75	0.31	-2.06	-1.75
it	-0.27	-0.91	-0.86	0.49
lt	-3.16	0.75	0.05	-0.91
lu	1.87	-3.44	-1.54	-0.75
lv	-2.86	0.61	0.56	-0.95
mt	-1.09	-0.61	1.54	-0.75
nl	2.32	-1.47	0.52	0.14
pl	-3.75	1.55	-1.03	1.80
pt	-1.80	0.19	-1.46	0.18
se	3.39	1.48	0.29	0.83
si	0.40	0.64	1.13	-0.86
sk	-2.11	-0.80	1.69	2.18
uk	0.70	0.77	-1.71	-0.51

	Axis1	Axis2	Axis3	Axis4
at	5.7	2.1	0.2	0.5
be	2.5	2.0	0.2	0.5
cy	0.1	0.2	1.5	8.7
cz	0.0	2.9	17.9	1.3
de	2.1	2.3	1.5	3.8
dk	10.5	17.4	0.3	0.0
ee	2.6	0.0	4.1	5.2
es	3.0	2.9	4.8	0.2
fi	4.9	2.4	0.5	0.6
fr	1.9	0.2	0.0	5.2
gr	3.7	3.8	4.8	4.7
hu	1.1	3.3	0.2	1.2
ie	0.5	0.3	13.2	13.1
it	0.1	2.2	2.3	1.0
lt	9.9	1.5	0.0	3.6
lu	3.4	31.9	7.4	2.4
lv	8.1	1.0	1.0	3.9
mt	1.2	1.0	7.4	2.4
nl	5.3	5.8	0.8	0.1
pl	13.9	6.5	3.3	13.9
pt	3.2	0.1	6.6	0.1
se	11.3	5.9	0.3	3.0
si	0.2	1.1	3.9	3.2
sk	4.4	1.7	8.9	20.4
uk	0.5	1.6	9.1	1.1

Question 5 *Faites une représentation graphique des valeurs propres. Combien de composantes principales faut-il retenir ? Quel est le pourcentage d'inertie totale expliquée par le sous-espace principal correspondant ?*

Les valeurs propres sont bien sûr données par les variances des composantes principales. On obtient l'histogramme suivant pour les 4 premières valeurs propres :



Le critère de Kaiser ($\lambda > 1$) conduit à considérer les 3 premières valeurs propres. L'inertie totale est 9 (car il y a 9 variables) et l'inertie portée par les 3 premiers axes est $4,05 + 1,48 + 1,29 = 6,82$. Cela représente 76% de l'inertie totale. C'est un qualité satisfaisante.

Question 6 *Y a-t-il un effet de taille ? Doit-on (peut-on) faire quelque chose pour changer la situation ?*

Il n'y a pas ici d'effet de taille, puisque seules 3 des variables (`chom`, `pv_ap`, `trv_pv`) ont une corrélation négative avec le premier axe. Comme elles représentent un « mauvais » indicateur du pays, on pourrait imaginer de les inverser pour mettre tout le monde du côté positif de l'axe. Ce n'est en fait pas possible, puisque la variable `pv_av`, qui a ces mêmes propriétés, est corrélée positivement (même si la valeur est très faible) avec l'axe 1. Par ailleurs il n'est pas facile de savoir si avoir une forte valeur d'une variable comme `benev` est un bon signe pour un pays (c'est bien, mais c'est une dimension différente).

Question 7 *Quelles sont les variables qui déterminent les axes que l'on retient (préciser les critères utilisés) ?*

On regarde les variables qui ont les corrélations les plus fortes avec les axes, avec autant que possible une limite commune sur la corrélation. Pour les deux premiers axes, on sélectionnera celle qui sont plus grandes que $0,70$ en valeur absolue ; par contre comme ce n'est pas possible pour le 3e axe, on se contentera d'une limite de $0,5$.

On représente les variables dans les tableaux suivants

Axe 1

$\ominus$	$\oplus$
`pv_ap` $(-0,73)$	`benev` $(0,87)$
`trv_pv` $(-0,70)$	`devel` $(0,81)$
	`depsoc` $(0,70)$

Axe 2

$\ominus$	$\oplus$
	`depedu` $(0,77)$
	`pv_av` $(0,75)$

Axe 3

$\ominus$	$\oplus$
`pv_ap` $(-0,57)$	
`pv_av` $(-0,52)$	

On ne prend pas de valeurs limites sur le 3e axe, qui est déjà difficile à interpréter.

Question 8 *Expliquez pourquoi* `lu` *a l'air de poser un problème dans l'ACP et pourquoi ce n'est pas surprenant au vu de la question 4. Que doit-on faire pour gérer cette situation ?*

Sur le second axe, on voit que `lu` a une contribution absolue qui est très importante (>30% du total). C'est donc un individu sur-représenté.

On connaît déjà la valeur exceptionnelle du `pib` dans ce pays, qui est riche et a très peu d'habitants. On peut aussi noter que les dépenses d'éducation (`depedu`) sont très faibles, probablement parce qu'on les a calculées en points de `pib`. Comme cette variable est l'une des plus importante de l'axe 2 $(0, 77)$, on peut craindre que `lu` n'ait déformé les axes.

Pour éviter ce problème, il est possible de retirer `lu` de la liste des individus pour l'analyse et de le traiter après coup comme un individu supplémentaire.

4 Une nouvelle analyse en composantes principales (8 points)

On fait une analyse en composantes principales des données centrées réduites en retirant `lu`. On donne ci-dessous, pour les 4 premiers axes principaux : les variances des composantes principales, le tableau des corrélations avec des variables, les coordonnés de `lu`, les coordonnées des autres individus sur les axes, leurs contributions aux axes (en %) et enfin leurs qualités de représentation par chacun des axes (en % encore).

Variances

```
[1] 4.56 1.40 1.13 0.91 0.49
```

Corrélations

	Comp1	Comp2	Comp3	Comp4
benev	0.85	0.28	-0.07	0.12
chom	-0.65	0.08	-0.08	0.65
depedu	0.57	-0.27	0.63	0.21
depsoc	0.72	-0.07	-0.36	0.51
devel	0.83	-0.29	-0.42	-0.07
pib	0.89	-0.20	-0.31	-0.17
pv_ap	-0.68	-0.57	-0.25	-0.26
pv_av	0.26	-0.82	0.35	0.14
trv_pv	-0.74	-0.35	-0.36	0.22

Individu supplémentaire

	Axis1	Axis2	Axis3	Axis4
lu	3.3	-0.26	-3.61	-2.11

	Axis1	Axis2	Axis3	Axis4
at	2.57	0.77	-0.97	-0.11
be	1.78	-0.49	0.11	0.36
cy	-0.15	0.40	0.96	-1.23
cz	-0.24	2.61	0.35	-0.60
de	1.54	1.16	-1.09	0.46
dk	3.60	-1.36	1.76	0.81
ee	-1.73	0.91	0.93	-0.84
es	-1.56	-0.68	-1.72	-0.48
fi	2.41	0.00	0.57	0.61
fr	1.54	0.03	-0.42	0.93
gr	-1.88	-0.52	-2.07	0.23
hu	-1.18	-0.30	1.35	0.00
ie	1.28	-2.08	-0.47	-1.97
it	-0.15	-0.28	-1.62	-0.21
lt	-3.22	-0.46	0.98	-0.58
lv	-2.98	0.06	1.19	-0.51
mt	-1.20	1.59	0.38	-0.79
nl	2.49	1.25	-1.27	-0.47
pl	-3.79	-1.71	0.35	2.23
pt	-1.80	-1.35	-0.62	-0.09
se	3.57	-0.25	0.88	1.23
si	0.41	0.68	1.29	-0.42
sk	-2.35	1.88	-0.54	2.08
uk	1.02	-1.85	-0.28	-0.66

	Axis1	Axis2	Axis3	Axis4
at	6.05	1.75	3.50	0.05
be	2.91	0.73	0.04	0.57
cy	0.02	0.47	3.41	6.93
cz	0.05	20.30	0.45	1.62
de	2.16	4.01	4.41	0.96
dk	11.86	5.53	11.40	2.99
ee	2.75	2.48	3.19	3.19
es	2.22	1.37	10.98	1.04
fi	5.32	0.00	1.18	1.71
fr	2.16	0.00	0.65	3.98
gr	3.22	0.82	15.90	0.25
hu	1.28	0.27	6.71	0.00
ie	1.49	12.89	0.81	17.77
it	0.02	0.23	9.75	0.20
lt	9.46	0.62	3.56	1.52
lv	8.11	0.01	5.26	1.20
mt	1.32	7.55	0.54	2.82
nl	5.69	4.62	5.96	1.01
pl	13.15	8.65	0.45	22.70
pt	2.97	5.45	1.44	0.03
se	11.66	0.18	2.88	6.89
si	0.16	1.36	6.13	0.81
sk	5.03	10.52	1.10	19.78
uk	0.95	10.18	0.30	1.97

	Axis1	Axis2	Axis3	Axis4
at	72.4	6.4	10.3	0.1
be	67.8	5.2	0.3	2.7
cy	0.4	2.9	16.6	27.5
cz	0.7	84.7	1.5	4.4
de	39.7	22.7	20.1	3.5
dk	67.1	9.6	15.9	3.4
ee	44.1	12.2	12.6	10.3
es	35.6	6.8	43.6	3.4
fi	81.1	0.0	4.5	5.2
fr	60.0	0.0	4.4	22.2
gr	39.9	3.1	48.7	0.6
hu	23.5	1.5	30.5	0.0
ie	12.5	33.1	1.7	29.8
it	0.7	2.5	87.2	1.5
lt	85.8	1.7	8.0	2.8
lv	81.9	0.0	13.1	2.4
mt	27.9	49.2	2.8	12.0
nl	59.5	14.9	15.4	2.1
pl	62.6	12.7	0.5	21.7
pt	40.2	22.7	4.8	0.1
se	83.3	0.4	5.1	9.9
si	5.6	15.0	54.4	5.9
sk	39.0	25.1	2.1	30.8
uk	18.0	59.2	1.4	7.5

Question 9 *Combien de composantes principales faut-il retenir ? La qualité globale de l'analyse a-t-elle été modifiée ?*

Le critère de Kaiser ($\lambda > 1$) conduit toujours à considérer les 3 premières valeurs propres. L'inertie portée par les 3 premiers axes est maintenant $4,56 + 1,40 + 1,13 = 6,99$. Cela représente 79% de l'inertie totale, soit un petit peu plus que précédemment. La qualité de l'analyse n'est que peu améliorée.

Question 10 *Quelles sont les variables qui déterminent les axes que l'on retient ? On gardera les critères de la question 7.*

On conserve les mêmes critères qu'à la question précédente pour faciliter la comparaison : pour les deux premiers axes, on sélectionnera donc les variables dont les corrélations sont plus grandes que $0,7$ en valeur absolue ; par contre pour le 3e axe, on se contentera d'une limite de $0,5$.

On représente les variables dans les tableaux suivants

Axe 1

-	+
`trv_pv` $(-0,74)$	`pib` $(0,89)$
[`pv_ap` $(-0,68)$]	`benev` $(0,85)$
	`devel` $(0,83)$
	`depsoc` $(0,72)$

Axe 2

-	+
`pv_av` $(-0,82)$	

Axe 3

-	+
	`depedu` $(0,63)$

Question 11 *Commentez les modifications les plus importantes des corrélations entre les deux analyses.*

Les changements principaux qu'on peut voir sont :

- `pib` est apparu dans l'axe 1 comme la variable la plus importante. Sa corrélation était artificiellement basse à cause de `lu`. `pv_ap` disparaît, mais c'est anecdotique, il a peu changé. C'est plutôt mieux en termes d'interprétation de voir le PIB dans cet axe.
- L'axe 2 semble changer de signe, ce qui ne veut rien dire comme on le sait. La disparition de `depedu` peut être liée au fait que `lu` a des dépenses d'éducation normales qui sont rapportées à son `pib` anormal (et donc paraissaient faibles).
- Du coup, `depedu` qui a disparu de l'axe 2 réapparaît sur l'axe 3.

Les axes 2 et 3 ont donc été redéfinis.

Question 12 *Quels sont les pays qui déterminent les axes que l'on retient (précisez les critères utilisés) ?*

Il faut comparer les contributions des pays à leur poids, c'est-à-dire $1/24 = 4.17\%$ (en effet, il n'y a plus que 24 pays). On décide de garder ceux dont les contributions sont supérieure à 2 fois leur poids, c'est-à-dire 8.33. Le signe n'est pas donné par le tableau des contributions, il est donc nécessaire de se reporter aux coordonnées des individus. Les pays qui déterminent les trois premières composantes principales sont alors :

Axe 1

-	+
`pl` (13.15)	`dk` (11.86)
`lt` (9.46)	`se` (11.66)
[`lv` (8.11)]	

Axe 2

-	+
`ie` (12.89)	`cz` (20.30)
`uk` (10.18)	`sk` (10.52)
`pl` (8.65)	

Axe 3

-	+
`gr` (15.90)	`dk` (11.40)
`es` (11.98)	
`it` (9.75)	

Pour le premier axe, `lv` est noté entre crochet pour indiquer qu'on a choisi de le garder parce que sa contribution n'est qu'à $2,6\%$ de la limite choisie. On aurait pu aussi l'ignorer.

Question 13 *Quelle interprétation peut-on donner des axes que l'on retient ?*

Le premier axe représente le niveau de développement du pays ; il oppose des pays d'Europe du nord (Danemark, Suède) à des pays de l'est (Pologne, Lettonie, Lituanie) – si on gardait plus de données ce serait encore plus clair.

Le second axe met en valeur des pays où le taux de pauvreté avant transferts sociaux est très important, c'est-à-dire qui produisent le plus d'inégalité (on a pris en compte la richesse en axe 1). Ce sont la Pologne, mais aussi, de manière plus surprenante, le Royaume-uni et l'Irlande. Par contre, la République Tchèque et la Slovaquie (soit l'ancienne Tchécoslovaquie) ont de bons résultats de ce point de vue. Il est intéressant de noter qu'on retrouve groupés des pays qui ont une histoire commune.

Enfin, le troisième axe mesure l'effort éducatif des pays. Le Danemark a une dépense forte (en points de `pib`), contrairement à l'Italie, le Grèce et l'Espagne (Europe du sud ?). On remarque dans les données que la Slovaquie, qui dépense encore moins que la Grèce, ne figure pas dans la liste ; la raison n'est pas évidente.

Question 14 *Quels sont les deux individus qui sont le plus mal représentés par l'espace principal que l'on retient (expliquez ce que vous faites) ?*

La « qualité de représentation » des individus se mesure par le cosinus carré de l'angle que forme le point par rapport à sa projection (en centrant autour du barycentre des individus). Si c_{ik} est la coordonnée de l'individu i sur l'axe k, alors la qualité de représentation d'un individu par les 3 premiers axes principaux est :

$$\frac{c_{i1}^2 + c_{i2}^2 + c_{i3}^2}{c_{i1}^2 + c_{i2}^2 + \cdots + c_{i9}^2}.$$

Il faut donc ajouter les trois premières colonnes du tableau des qualités de représentation (ce ne sont pas des valeurs cumulées ici). Les données obtenues sont :

```
          at   be   cy   cz   de   dk ee es   fi   fr   gr   hu   ie   it   lt   lv mt   nl   pl   pt   se si   sk   uk
Axis1:3 89.2 73.3 19.9 86.9 82.6 92.7 69 86 85.6 64.5 91.7 55.6 47.3 90.4 95.5 95.1 80 89.8 75.7 67.8 88.8 75 66.2 78.5
```

Les deux individus les plus mal représentés sur l'espace $(1, 2, 3)$ sont Chypre (`cy`, $19, 9$) et l'Irlande (`ie`, $47, 3$). On remarque que `ie` est mal représenté alors qu'il détermine l'axe 2. Par contre, `cy` est proche du centre gravité sur les 3 premiers axes, et il est possible que la représentation ne soit pas aussi mauvaise qu'on l'imagine.

Question 15 *Expliquez pourquoi les coordonnées de* `lu` *sont données à part. Comment interpréter sa position sur les premiers axes ?*

Les coordonnées de `lu` sont données à part parce qu'il est traité comme individu supplémentaire. Ses coordonnées ont donc dû être calculées après l'ACP elle-même.

Si on regarde les coordonnées de `lu` sur les 3 premiers axes, on voit que

— `lu` est lié positivement avec l'axe 1 : le Luxembourg est en effet un pays très développé et son PIB par habitant en particulier est très important ;

— il n'y a pas de lien visible sur le second axe ;

— sur le 3e axe, `lu` est très nettement du coté négatif ; cet effet est probablement dû à ce que `depedu` est relatif au PIB, et que `pib` est très élevé pour ce pays (`lu` a les dépenses d'éducation en points de PIB les plus basses de tous les pays ! une valeur en dépense par habitant serait peut-être plus normale).

Notons qu'il n'est pas souhaitable d'utiliser les arguments de type « contribution aux axes » pour les individus supplémentaires.