FACULTÉ DES SCIENCES
DÉPARTEMENT D'INFORMATIQUE

STT707 - Analyses des données

Projet - *Devoir 1*

préparé par

Gabriel Gibeau Sanchez (gibg2501)

présenté à

Bernard Colin

12 février 2021

Table des matières

1 Partie A

1.1 Problème 1

1.1.1 1)

Soit $\mathbb{I} = [1, 1, ..., 1],\ \mathbb{I} \in \mathbb{R}^n$. Nous avons également la métrique $D = diag(p_1, p_2, ..., p_n)$, où $\sum_{i=1}^{n} p_i = 1$. Pour calculer la norme $\|\mathbb{I}\|_D$ nous utilisons le produit scalaire :

$$\begin{aligned} < \mathbb{I}, \mathbb{I} >_D = {}^t\mathbb{I}D\mathbb{I} &= \begin{bmatrix} 1 & 1 & ... & 1 \end{bmatrix} \begin{bmatrix} p_1 & 0 & \dots & 0 \\ 0 & p_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & p_n \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ ... \\ 1 \end{bmatrix} \\ &= (1)p_1(1) + (1)p_2(1) + ... + (1)p_n(1) \\ &= \cancel{(1)} \sum_{i=1}^{n} p_i \end{aligned}$$

Or nous savons que $\sum_{i=1}^{n} p_i = 1$. En extrayant la racine carrée on retrouve évidement que :

$$\sqrt{\sum_{i=1}^{n} p_i} = 1$$

Alors, $\|\mathbb{I}\|_D = 1$ et donc, le vecteur $\mathbb{I}$ est bien de norme unitaire lorsqu'on lui applique la métrique D.

1.1.2 2)

Pour tout j quelconque tel que $j = 1, 2, ..., p$ on a ${}^t\xi^j = [x_1^j\, x_2^j \dots x_n^j]$ La projection orthogonal d'un vecteur sur autre correspond à la notion de produit scalaire. Ainsi, l'expression de la projection D-orthogonale de ξ^j sur l'axe de support $\Delta_{\mathbb{I}}$ est obtenue en multipliant les éléments de $\mathbb{I}$ par un nombre réel. Ce nombre correspond au rapport entre le produit scalaire des vecteurs et la norme au carré du vecteur sur lequel est faite la projection :

$$P_{\xi^j}(\Delta_{\mathbb{I}}) = \frac{\langle \xi^j, \mathbb{I} \rangle_D}{\|\mathbb{I}\|_D^2} * \mathbb{I}$$
$$= \left(\frac{\xi^j D \, \mathbb{I}}{\|\mathbb{I}\|_D^2} \right) * \mathbb{I}$$

Puisque $\|\mathbb{I}\|_D = 1$:

$$= \left(\xi^j D \, \mathbb{I} \right) * \mathbb{I}$$

De plus, la norme de la projection $P_{\xi^j}(\Delta_{\mathbb{I}})$ est donnée par :

$$\|P_{\xi^j}(\Delta_{\mathbb{I}})\| = \frac{\langle \xi^j, \mathbb{I} \rangle_D}{\|\mathbb{I}\|}$$
$$= cos(\xi^j, \mathbb{I}) * \|\xi^j\|$$

1.1.3 3)

Dans $\mathbb{R}^n$ nous avons :

$$\xi^j = \begin{bmatrix} x_1^j \\ x_2^j \\ \vdots \\ x_n^j \end{bmatrix}$$

Pour centrer les variables x_i^j il faut leur soustraire leurs moyennes respectives : $x_i^j - \bar{x}^j$, $\forall j$:

$$\widetilde{\xi}^j = \begin{bmatrix} x_1^j - \bar{x}^j \\ x_2^j - \bar{x}^j \\ \vdots \\ x_n^j - \bar{x}^j \end{bmatrix}, \quad j = 1, 2, ..., p$$

Ce qui correspond à une translation (addition d'un vecteur constant à tous les vecteurs $\in \mathbb{R}^p$). Ainsi, le centre de gravité correspond à l'origine :

$$g = \sum_{i=1}^{n} p_i x_i^j = 0 \, , g \in \mathbb{R}^p$$

1.1.4 4)

Nous avons $X = \{x_i^j\}, {}^t\mathbb{I} = [1, 1, ..., 1], D = diag(p_1, p_2, ..., p_n)$. Nous pouvons écrire le centre de gravité g comme étant le produit scalaire $\langle {}^tX, \mathbb{I}\rangle_D$ doté de la pondération D :

$$
\begin{aligned}
g = {}^tXD\mathbb{I} &= \begin{bmatrix} x_1^1 & x_1^2 & ... & x_1^n \\ x_2^1 & x_2^2 & ... & x_2^n \\ \vdots & \vdots & \ddots & \vdots \\ x_p^1 & x_p^2 & ... & x_p^n \end{bmatrix} \begin{bmatrix} p_1 & 0 & ... & 0 \\ 0 & p_2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & p_n \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} x_1^1 & x_1^2 & ... & x_1^n \\ x_2^1 & x_2^2 & ... & x_2^n \\ \vdots & \vdots & \ddots & \vdots \\ x_p^1 & x_p^2 & ... & x_p^n \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} \\
&= \begin{bmatrix} \sum_{j=1}^n p_i x_1^i \\ \sum_{j=1}^n p_i x_2^i \\ \vdots \\ \sum_{j=1}^n p_i x_p^i \end{bmatrix} \\
g &= \begin{bmatrix} \bar{x}^1 \\ \bar{x}^2 \\ \vdots \\ \bar{x}^p \end{bmatrix}
\end{aligned}
$$

Pour centrer les données de X il faut soustraire le vecteur du centre de gravité à chaque observation à l'instar de ce qui est fait au point 1.1.3 :

$$
\widetilde{X} = \begin{bmatrix} x_1 - g \\ x_2 - g \\ \vdots \\ x_n - g \end{bmatrix}
$$

Donc :

$$
\begin{aligned}
\widetilde{X} &= X - (\mathbb{I})({}^t\mathbb{I}DX) \\
&\text{Avec } {}^t\mathbb{I}DX = {}^tg
\end{aligned}
$$

2 Partie B

2.1 Examen préliminaire des données

2.1.1 Centralisation et réduction des données

Un examen préliminaire des données révèle qu'elle ont des ordres de grandeurs différents :

	POPU	DENS	TATO	AASP	AIND	PNB	PIBA	FBCF	RECC	RESO	TESC	IMPT	EXPT	CAL	LOG	ELEC	EDUC	TV
count	18.00	18.00	18.00	18.00	18.00	18.00	18.00	18.00	18.00	18.00	18.00	18.00	18.00	18.00	18.00	18.00	18.00	18.00
mean	36310.33	125.22	0.89	17.30	37.87	2154.56	9.42	23.21	34.40	3010.61	6.25	9942.39	10028.11	3014.44	7.92	3748.39	4.73	185.22
std	49975.88	114.97	0.36	11.87	6.37	1030.79	5.67	4.38	7.27	3440.05	1.22	8725.85	10442.20	228.86	2.49	3128.85	1.62	93.75
min	2921.00	2.00	0.25	2.90	22.50	600.00	2.90	16.70	21.20	290.00	3.50	1231.00	554.00	2460.00	4.00	607.00	1.40	9.00
25%	7522.00	27.00	0.70	8.35	34.70	1547.50	5.90	20.80	30.90	626.25	5.81	2854.50	2039.50	2910.00	6.45	1959.25	4.20	137.00
50%	11428.50	89.50	0.82	14.90	37.80	2080.00	7.00	23.15	35.55	1984.50	6.25	7941.50	7831.50	2990.00	8.05	2690.50	4.80	187.50
75%	53171.25	216.00	1.05	23.75	41.12	2710.00	13.85	25.07	38.05	3652.50	7.00	14552.25	14703.50	3175.00	8.95	3799.50	5.80	240.75
max	203213.00	352.00	1.85	48.20	49.10	4660.00	20.30	35.20	48.10	12305.00	9.00	30052.00	37988.00	3450.00	13.40	12976.00	7.40	392.00

FIGURE 1 – Statistiques descriptives du jeu de données initial

De plus, le lexique sur les acronymes utilisés nous apprends que les différentes variables ont des unités différentes (million de dollar américain, pourcentage du PIB, nombre de calories par jour par habitant etc.). Pour ces raisons, il est nécessaire de centrer et standardiser les données. Pour ce faire on utilise l'objet *StandardScaler* de *SciKit-learn* qui effectue la soustraction de la moyenne et la division par l'écart-type sur chaque composante.

2.2 Analyse en composantes principales

Par la suite, il est possible d'appliquer une réduction de dimensions du jeu de données à l'aide de l'analyse en composante principale (ACP). Dans un premier temps, l'ensembles de variables a été examiné en ACP. Il en ressort que les 4 premières composantes principales sont associées à 80.4% variance totale :

	0	1	2	3
	0.39748	0.16740	0.12858	0.11082

FIGURE 2 – 4 première valeurs propres associées aux composantes de l'ACP de l'ensembe de données

On peut également constater en examinant la figure 3 que les plus grandes différence entre la variation expliquée surviennent entre la 1$^{\text{ière}}$ et la 2$^{\text{ième}}$ composante, ainsi qu'entre la 4$^{\text{ième}}$ et la 5$^{\text{ième}}$ composantes, ce qui suggère la présence d'une structure dans les données.

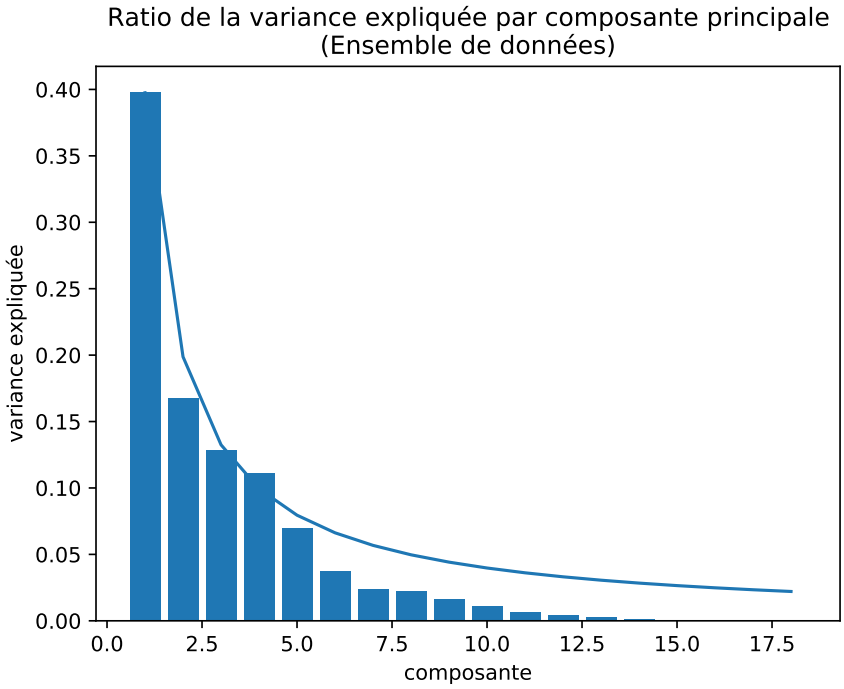

FIGURE 3 – Diagramme à bar montrant le rapport de la variation expliquée entre les composantes principales - Ensemble des variables

Évidemment, pour représenter graphiquement la projection des données sur les composantes principales (CP) on ne peut utiliser que 3 variables. Il convient donc de choisir les 3 premières composantes qui expliquent cumulativement 69.8% de la variance du nuage de points. L'ensemble des données projetées sur les CP sont présentées à la figure suivante.

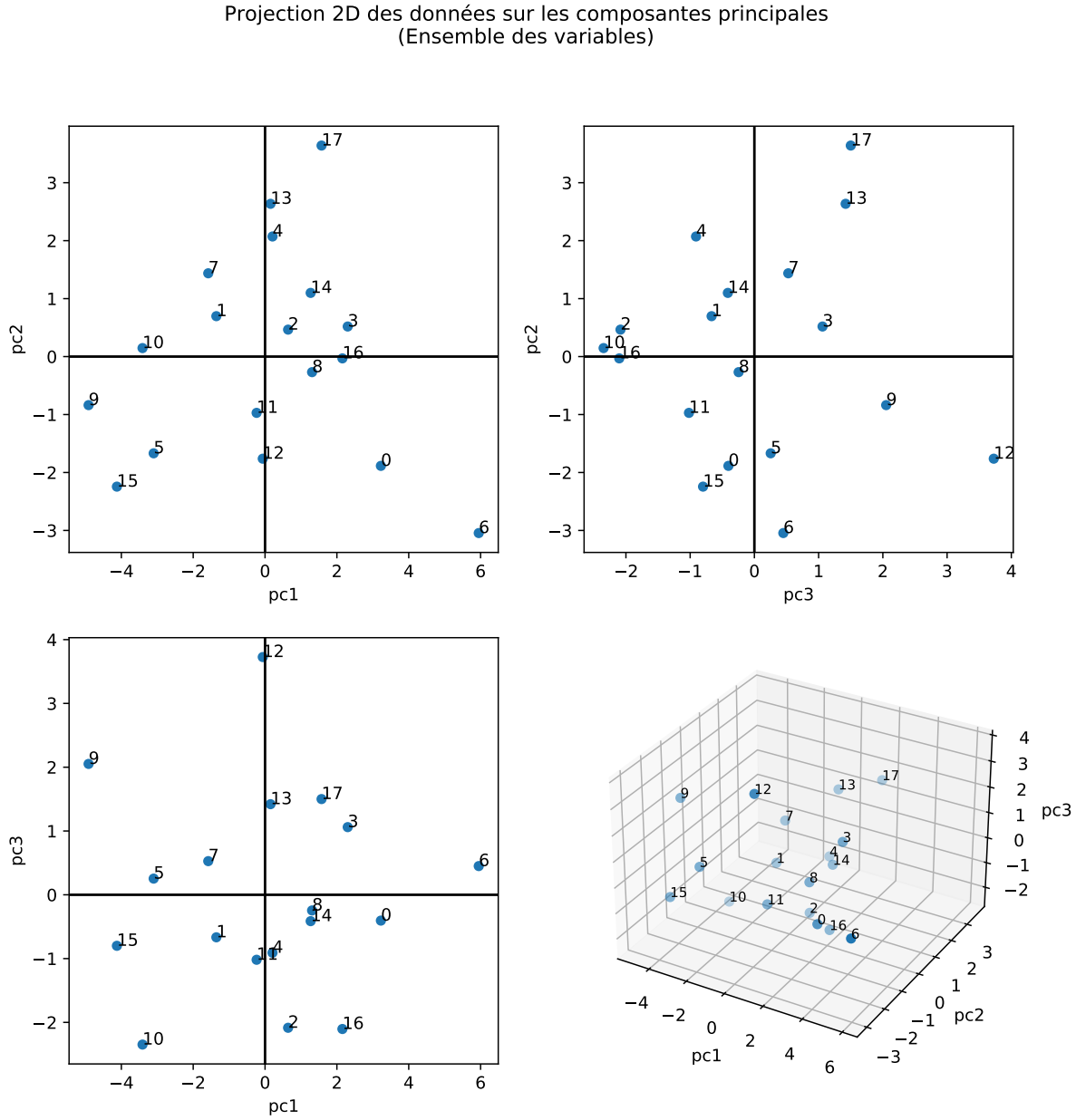

FIGURE 4 – 0-ALLEMAGNE, 1-AUTRICHE, 2-BELGIQUE-LUXEMBOURG, 3-CANADA, 4-DANEMARK, 5-ESPAGNE, 6-USA, 7-FINLANDE, 8-FRANCE, 9-GRECE, 10-IRLANDE, 11-ITALIE, 12-JAPON, 13-NORVEGE, 14-PAYS-BAS, 15-PORTUGAL, 16-ANGLETERRE, 17-SUEDE

On peut également effectuer l'ACP sur les variables de structure et de consommation séparément afin de voir s'il y a des structures sous-jacentes.

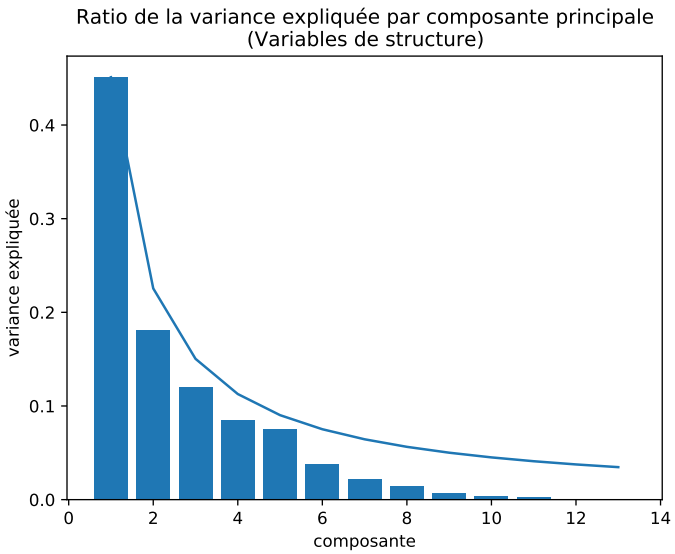

FIGURE 5 – Diagramme à barres montrant le rapport de la variation expliquée entre les composantes principales - Variables de structure

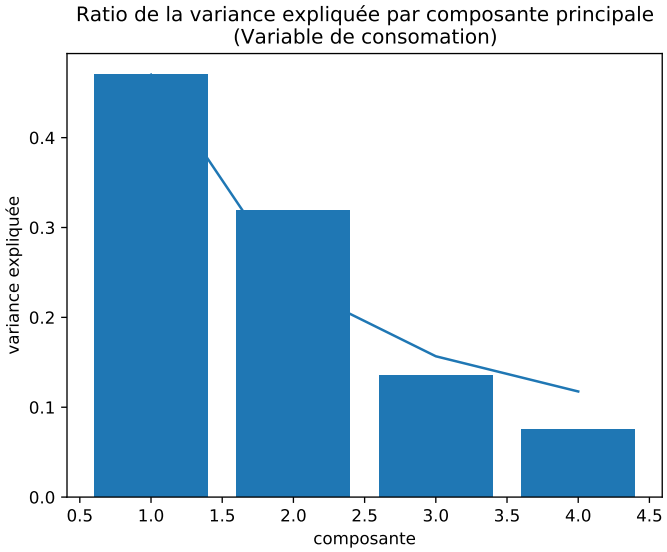

FIGURE 6 – Diagramme à bar montrant le rapport de la variation expliquée entre les composantes principales - Variables de consommation

On voit à la figure 5 que la cassure est plus nette entre les différent ratio de variance pour les données de structure. Conséquemment, on voit que la projection des données de structure sur les CP (figure 7) est plus étendue que les projections de l'ensemble des variables (figure 4) et des variables de consommation (figure 8) :

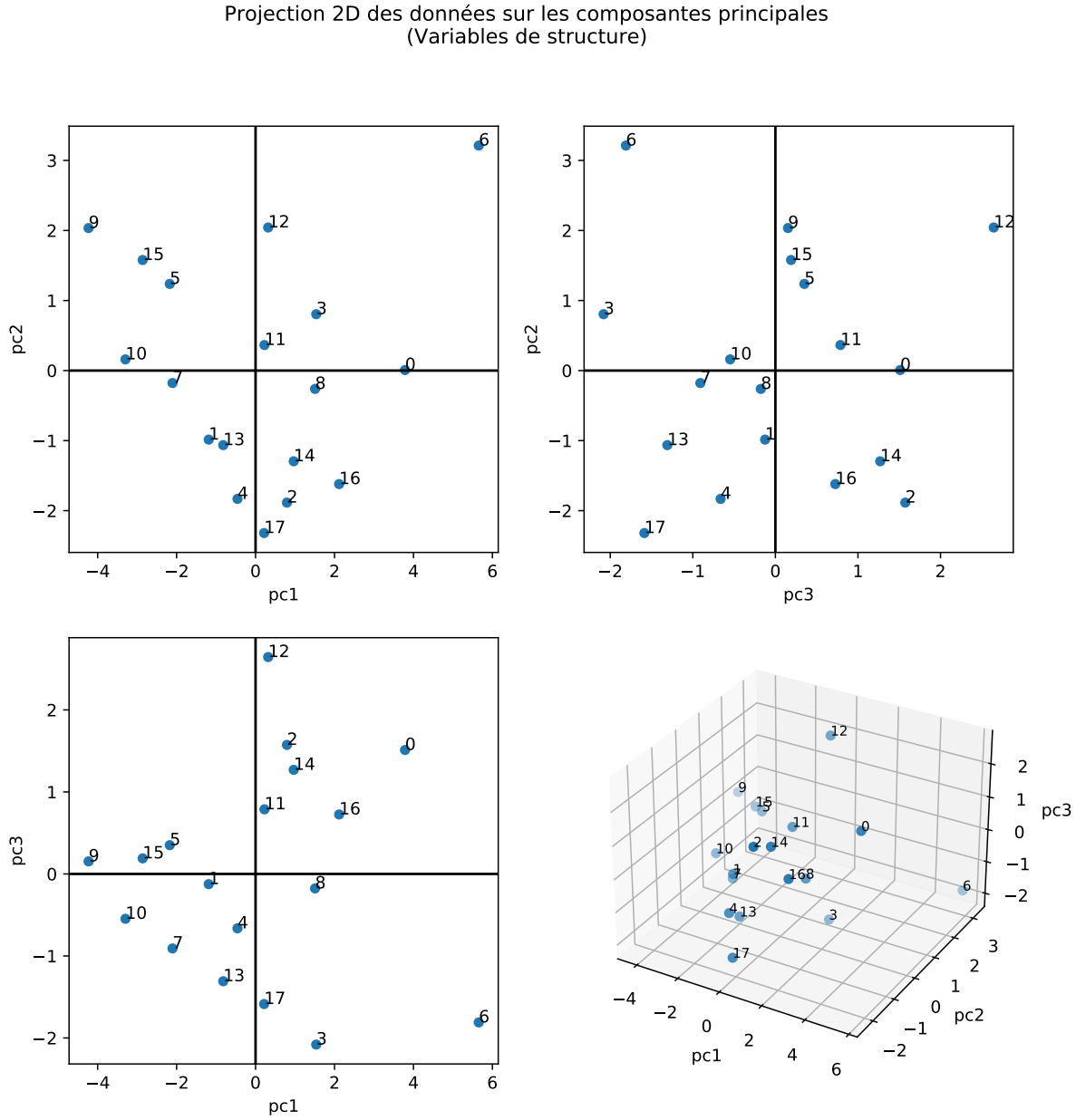

FIGURE 7 – 0-ALLEMAGNE, 1-AUTRICHE, 2-BELGIQUE-LUXEMBOURG, 3-CANADA, 4-DANEMARK, 5-ESPAGNE, 6-USA, 7-FINLANDE, 8-FRANCE, 9-GRECE, 10-IRLANDE, 11-ITALIE, 12-JAPON, 13-NORVEGE, 14-PAYS-BAS, 15-PORTUGAL, 16-ANGLETERRE, 17-SUEDE

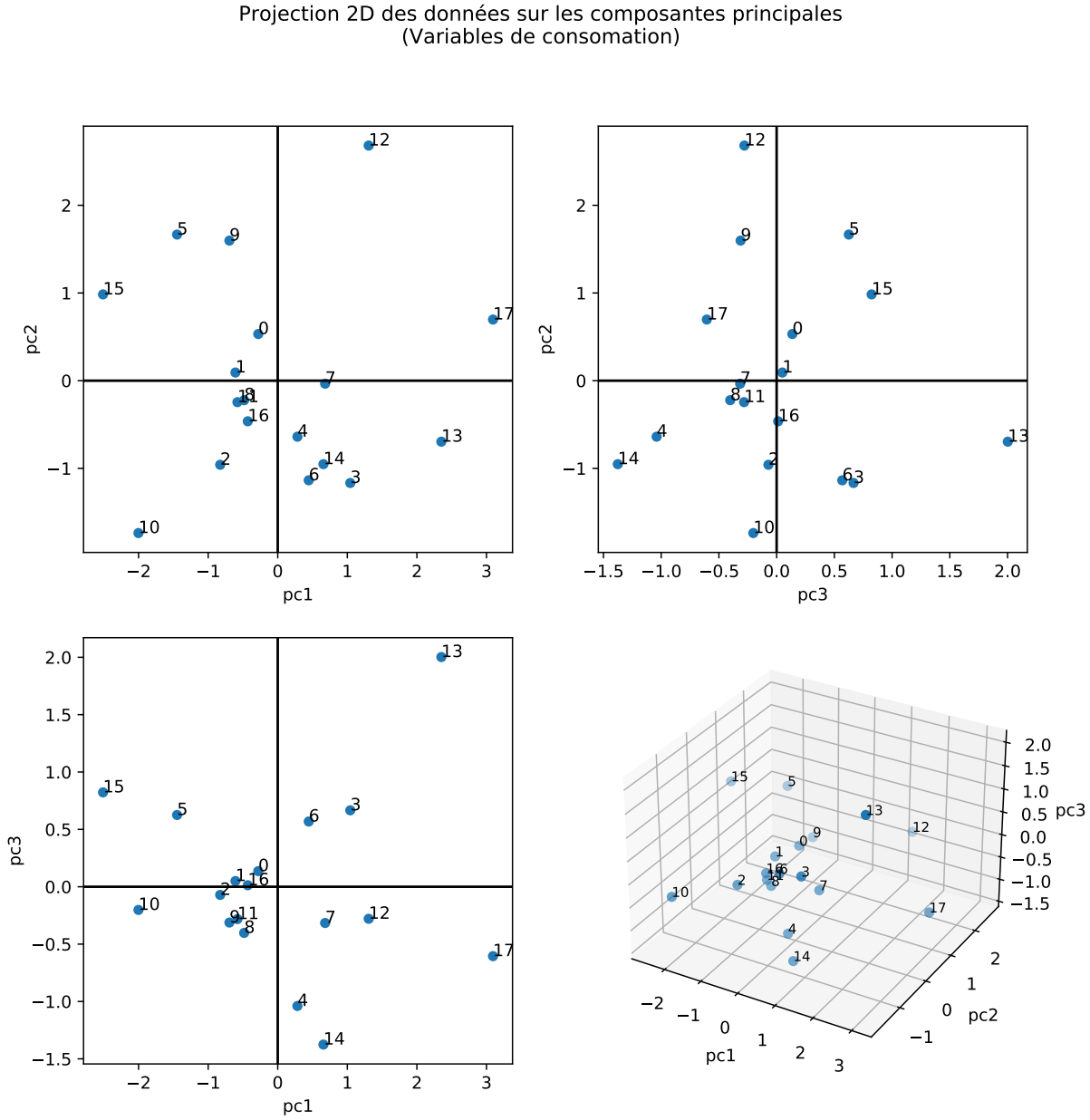

FIGURE 8 – 0-ALLEMAGNE, 1-AUTRICHE, 2-BELGIQUE-LUXEMBOURG, 3-CANADA, 4-DANEMARK, 5-ESPAGNE, 6-USA, 7-FINLANDE, 8-FRANCE, 9-GRECE, 10-IRLANDE, 11-ITALIE, 12-JAPON, 13-NORVEGE, 14-PAYS-BAS, 15-PORTUGAL, 16-ANGLETERRE, 17-SUEDE

2.3 Interprétation des axes des composantes principales - Variables de structure

La 1$^{\text{ière}}$ CP est majoritairement déterminée par les variables 12 et 13 des données initiales, nommément "Importations totales de marchandises en million de dollars (US)" (IMPT) et "Exportation totales de marchandises en million de dollars (US)" (EXPT). Ainsi, "Volume du commerce extérieur" semble un nom raisonnable pour l'axe PC1.De la même manière, la 2$^{\text{ième}}$ CP est majoritairement déterminée par les variables "9) RECC : Recettes courantes en pourcentage du PNB" et "1) POPU : Population totale en milliers d'habitants", et la 3$^{\text{ième}}$ CP est majoritairement déterminée par "2) DENS : Densité de la population au km 2" et "6) PNB : Produit national brut en dollars (US) par habitant". On peut donc attribuer les noms respectifs de "Revenus relatif par habitant" et de manière plutôt abstraite "Densité spatiale de la monnaie".

TABLE 1 – Interprétation des axes de projection

Composante	Interprétation
PC1	Volume du commerce extérieur
PC2	Revenus relatif par habitant
PC3	Densité spatiale de la monnaie

2.4 Création des variables en supplémentaires

Il est possible d'utiliser les données de consommation pour créer des classes virtuelles dans lesquelles les pays peuvent être répartit. En l'occurrence un "indice de consommation" à été créé en additionnant toutes les valeurs brutes des variables de consommation pour chaque pays, de manière à avoir un portrait global du volume de consommation absolu pour chaque pays. Ce faisant on émet l'hypothèse que les variables de consommation disponibles reflètent fidèlement le volume de consommation général de chaque pays.
On divise ensuite les pays en 4 classes correspondantes au 4 quartiles de la distribution des variables de consommation additionnées, ce qui nous permet d'assigner les même classes dans les données de structures, et de projeter en supplémentaire la moyenne de chacune de ces classes de consommateurs. Les classes sont numérotées de 0 à 3, 0 étant les pays les moins consommateurs et 3 les pays les plus consommateurs.

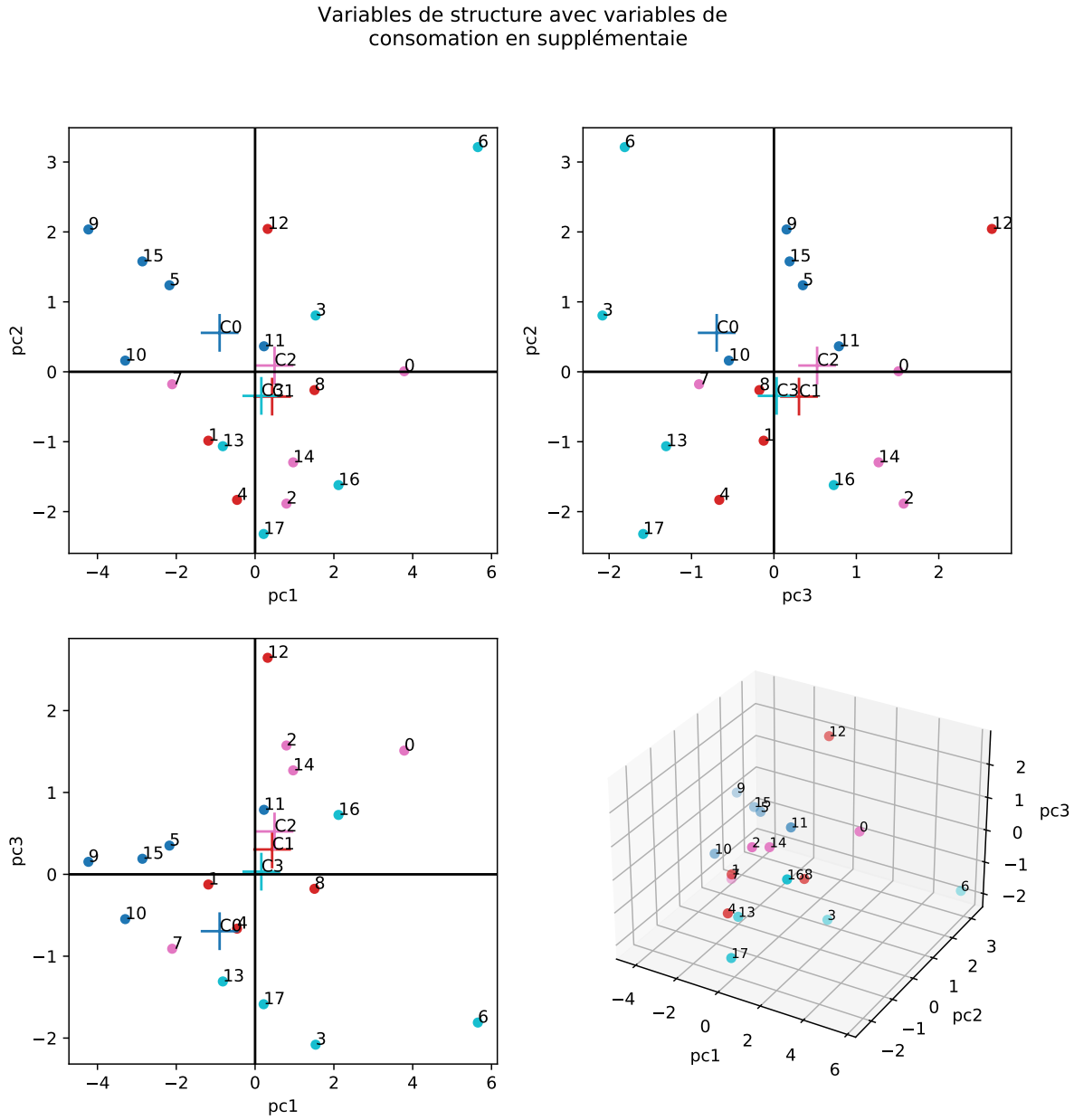

FIGURE 9 – 0-ALLEMAGNE, 1-AUTRICHE, 2-BELGIQUE-LUXEMBOURG, 3-CANADA, 4-DANEMARK, 5-ESPAGNE, 6-USA, 7-FINLANDE, 8-FRANCE, 9-GRECE, 10-IRLANDE, 11-ITALIE, 12-JAPON, 13-NORVEGE, 14-PAYS-BAS, 15-PORTUGAL, 16-ANGLETERRE, 17-SUEDE

2.5 Interprétation des données de structure en ACP

On voit donc que les pays de la classe 0, les pays les moins consommateurs, sont distribués d'avantage vers l'extrémité négative de l'axe PC1 et vers l'extrémité positive de l'axe PC2. Ce qui pourrait être reformulé en langage naturel comme : "Les payes qui ont un faible volume de commerce international et un faible revenu relatif par habitant consomment moins".
On constate également qu'à l'autre extrême, les pays les plus consommateurs ont en commun un d'être distribués vers l'extrémité négative de la CP 3, ce qui pourrait être interprété comme : "Les pays où les gens ont de l'argent et de l'espace consomment le plus". Également, ils ont généralement des volumes de commerce international légèrement supérieurs à la moyenne.
Les classes 1 et 2 sont moins faciles à interpréter puisqu'elle se rapprochent d'avantage des moyennes pour chaque CP. On peut cependant dire de la classe 2 qu'elle est plutôt distribuée vers les valeurs négatives de la $2^{\text{ième}}$ CP. On peut donc en dire que les pays à consommation moyenne-elevé ont en général des revenus relatifs par habitant égaux ou supérieurs à la moyenne. La classe 1 suit la même tendance, à l'exception notable du Japon.