AgroParisTech

Analyse en Composantes Principales

C. Duby, S. Robin

Table des matières

1 Introduction

L'Analyse en Composantes principales (ACP) fait partie du groupe des méthodes descriptives multidimensionnelles appelées méthodes factorielles. Ces méthodes qui sont apparues au début des années 30 ont été surtout développées en France dans les années 60, en particulier par Jean-Paul Benzécri qui a beaucoup exploité les aspects géométriques et les représentations graphiques. Dans la mesure où ce sont des méthodes descriptives, elles ne s'appuient pas sur un modèle probabiliste, mais elles dépendent d'un modèle géométrique. L'ACP propose, à partir d'un tableau rectangulaire de données comportant les valeurs de p variables quantitatives pour n unités (appelées aussi individus), des représentations géométriques de ces unités et de ces variables. Ces données peuvent être issues d'une procédure d'échantillonnage ou bien de l'observation d'une population toute entière. Les représentations des unités permettent de voir s'il existe une structure, non connue a priori, sur cet ensemble d'unités. De façon analogue, les représentations des variables permettent d'étudier les structures de liaisons linéaires sur l'ensemble des variables considérées. Ainsi, on cherchera si l'on peut distinguer des groupes dans l'ensemble des unités en regardant quelles sont les unités qui se ressemblent, celles qui se distinguent des autres, *etc.* Pour les variables, on cherchera quelles sont celles qui sont très corrélées entre elles, celles qui, au contraire ne sont pas corrélées aux autres, *etc.*

Nous verrons après l'exposé de la méthode, quelles précautions il faut prendre pour interpréter correctement les représentations obtenues. Dans tous les cas, il ne faut pas oublier d'où sont issues les données utilisées et ce qu'elles représentent et signifient pour le problème que l'on se pose.

Enfin, comme pour toute méthode descriptive, réaliser une ACP n'est pas une fin en soi. L'ACP servira à mieux connaître les données sur lesquelles on travaille, à détecter éventuellement des valeurs suspectes, et aidera à formuler des hypothèses qu'il faudra étudier à l'aide de modèles et d'études statistiques inférentielles. On pourra aussi, a posteriori, se servir des représentations fournies par l'ACP pour illustrer certains résultats dans un but pédagogique.

2 Tableau de données

Les données sont les mesures effectuées sur n unités $\{u_1, u_2, ..., u_i, ...u_n\}$. Les p variables quantitatives qui représentent ces mesures sont $\{v_1, v_2, ..., v_j, ...v_p\}$.

Le tableau des données brutes à partir duquel on va faire l'analyse est noté $\mathbf{X}$ et a la forme suivante :

$$\mathbf{X} = \begin{array}{c} \\ u_1 \\ u_2 \\ . \\ u_i \\ . \\ u_n \end{array} \begin{array}{c} \begin{array}{cccccc} v_1 & v_2 & ... & v_j & ... & v_p \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & ... & x_{1j} & ... & x_{1p} \\ x_{21} & x_{22} & ... & x_{2j} & ... & x_{2p} \\ . & . & ... & . & ... & . \\ x_{i1} & x_{i2} & ... & x_{ij} & ... & x_{ip} \\ . & . & ... & . & ... & . \\ x_{n1} & x_{n2} & ... & x_{nj} & ... & x_{np} \end{bmatrix} \end{array} .$$

On peut représenter chaque unité par le vecteur de ses mesures sur les p variables :

$${}^tU_i = \begin{bmatrix} x_{i1} & x_{i2} & ... & x_{ij} & ... & x_{ip} \end{bmatrix} \quad \text{ce qui donne} \quad U_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ . \\ x_{ij} \\ . \\ x_{ip} \end{bmatrix} .$$

Alors U_i est un vecteur de $\mathbb{R}^p$.

De façon analogue, on peut représenter chaque variable par un vecteur de $\mathbb{R}^n$ dont les composantes sont les valeurs de la variable pour les n unités :

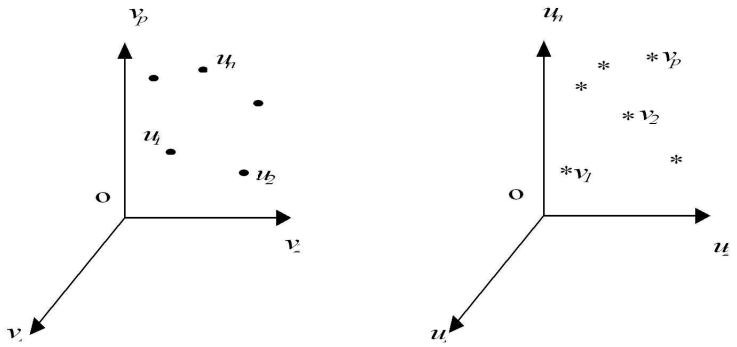
$$V_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ . \\ x_{ij} \\ . \\ x_{nj} \end{bmatrix} .$$

Pour avoir une image de l'ensemble des unités, on se place dans un espace affine en choisissant comme origine un vecteur particulier de $\mathbb{R}^p$, par exemple le vecteur dont toutes les coordonnées sont nulles. Alors, chaque unité sera représentée par un point dans cet espace. L'ensemble des points qui représentent les unités est appelé traditionnellement "nuage des individus".

En faisant de même dans $\mathbb{R}^n$, chaque variable pourra être représentée par un point de l'espace affine correspondant. L'ensemble des points qui représentent les variables est appelé "nuage des variables".

On constate, que ces espaces étant de dimension supérieure en général à 2 et même 3, on ne peut visualiser ces représentations. L'idée générale des méthodes factorielles est

de trouver un système d'axes et de plans tels que les projections de ces nuages de points sur ces axes et ces plans permettent de reconstituer les positions des points les uns par rapport aux autres, c'est-à-dire avoir des images les moins déformées possible.



Exemple : Budgets de l'état de 1872 à 1971 On présente ici un exemple tiré de [1]. Il s'agit de l'étude des différents postes du budget de l'état français de 1872 à 1971. Les valeurs sont données en pourcentage du budget global pour éliminer l'effet de l'évolution de la valeur du franc nominal au cours du temps. Les intitulés complets des variables sont :

`AN`	:	Année (irrégulièrement espacées de 1872 à 1971),
`PVP`	:	Pouvoirs publics,
`AGR`	:	Agriculture,
`CMI`	:	Commerce et industrie,
`TRA`	:	Travail,
`LOG`	:	Logement,
`EDU`	:	éducation,
`ACS`	:	Action sociale,
`ANC`	:	Anciens combattants,
`DEF`	:	Défense,
`DET`	:	Remboursement de la dette,
`DIV`	:	Divers.

Les colonnes (hormis `OBS` et `AN`) du tableau `Donnees brutes` du listing SAS présenté page 28 correspondent à la matrice $\mathbf{X}$. Les individus sont les années et les variables les pourcentages de budgets alloués aux différents postes. Chaque année est représentée par un point situé dans un espace de dimension 11 dont les coordonnées sont les pourcentages des différents postes dans le budget national de cette année-là. (En fait, le nuage des individus est compris dans un sous-espace de dimension 10 puisque la somme des différents pourcentages pour une année donnée vaut toujours 100%). Chaque poste budgétaire est représenté par un point situé dans un espace de dimension 24 dont les coordonnées sont les pourcentages accordés à ce poste aux différentes années.

3 Choix d'une distance

Pour faire une représentation géométrique, il faut choisir une distance entre deux points de l'espace. La distance utilisée par l'ACP dans l'espace où sont représentés les unités, est la distance euclidienne classique. La distance entre deux unités u_i et $u_{i'}$ est égale à :

$$d^2(u_i, u_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \ .$$

Avec cette distance, toutes les variables jouent le même rôle et les axes définis par les variables constituent une base orthogonale. à cette distance on associe un produit scalaire entre deux vecteurs :

$$\langle \overrightarrow{ou_i}, \overrightarrow{ou_{i'}} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j} = \ {}^t U_i \ U_{i'}$$

ainsi que la norme d'un vecteur :

$$\| \overrightarrow{ou_i} \|^2 = \sum_{j=1}^{p} x_{ij}^2 = \ {}^t U_i \ U_i$$

On peut alors définir l'angle α entre deux vecteurs par son cosinus :

$$\cos(\alpha) = \frac{\langle \overrightarrow{ou_i}, \overrightarrow{ou_{i'}} \rangle}{\| \overrightarrow{ou_i} \| \ \| \overrightarrow{ou_{i'}} \|} = \frac{\sum_{j=1}^{p} x_{ij} x_{i'j}}{\sqrt{\sum_{j=1}^{p} x_{ij}^2 \sum_{j=1}^{p} x_{i'j}^2}} = \frac{{}^t U_i \ U_{i'}}{\sqrt{({}^t U_i \ U_i) \ ({}^t U_{i'} \ U_{i'})}}.$$

4 Choix de l'origine

Le point o correspondant au vecteur de coordonnées toutes nulles n'est pas forcément une origine satisfaisante, car si les coordonnées des points du nuage des individus sont grandes, le nuage est éloigné de cette origine. Il apparaît plus judicieux de choisir une origine liée au nuage lui-même : le centre de gravité du nuage. Pour définir ce centre de gravité, il faut choisir un système de pondération des unités :

$\forall\ i = 1, ..., n$ $\quad p_i =$ poids de l'unité u_i tel que $\sum_{i=1}^{n} p_i = 1$. Par définition le centre de gravité est défini comme le point tel que :

$$\sum_{i=1}^{n} p_i \ \overrightarrow{Gu_i} = \overrightarrow{0}$$

Pour l'ACP on choisit de donner le même poids $\frac{1}{n}$ à tous les individus.

Le centre de gravité G du nuage des individus est alors le point dont les coordonnées sont les valeurs moyennes des variables :

$$G = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{ij} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{ip} \end{pmatrix} = \begin{pmatrix} x_{\bullet 1} \\ \vdots \\ x_{\bullet j} \\ \vdots \\ x_{\bullet p} \end{pmatrix} .$$

Prendre G comme origine, conformément à la figure suivante, revient alors à travailler sur le tableau des données centrées :

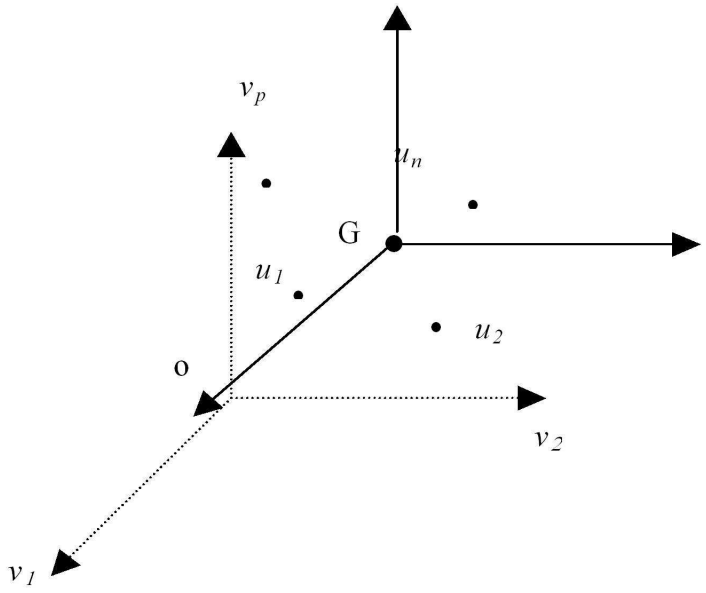
$$\mathbf{X}_c = \begin{bmatrix} x_{11} - x_{\bullet 1} & \cdots & x_{1j} - x_{\bullet j} & \cdots & x_{1p} - x_{\bullet p} \\ \vdots & & . & & \vdots \\ x_{i1} - x_{\bullet 1} & \cdots & x_{ij} - x_{\bullet j} & \cdots & x_{ip} - x_{\bullet p} \\ \vdots & & \vdots & & \vdots \\ x_{n1} - x_{\bullet 1} & \cdots & x_{nj} - x_{\bullet j} & \cdots & x_{np} - x_{\bullet p} \end{bmatrix}$$

et le vecteur des coordonnées centrées de l'unité u_i est :

$$U_{ci} = \begin{bmatrix} x_{i1} - x_{\bullet 1} \\ x_{i2} - x_{\bullet 2} \\ \vdots \\ x_{ij} - x_{\bullet j} \\ \vdots \\ x_{ip} - x_{\bullet p} \end{bmatrix}$$

celui des coordonnées centrées de la variable v_j est :

$$V_{cj} = \begin{bmatrix} x_{1j} - x_{\bullet j} \\ \vdots \\ x_{ij} - x_{\bullet j} \\ \vdots \\ x_{nj} - x_{\bullet j} \end{bmatrix} .$$



Exemple Le paragraphe `Simple Statistics` de la page 28 donne notamment les moyennes des différentes variables. La ligne `Mean` donne donc les coordonnées du barycentre :

$$G = \begin{pmatrix} 12.2125 \\ 1.9958 \\ \vdots \\ 1.1833 \end{pmatrix} .$$

Après changement d'origine la coordonnée de l'année `1872` sur l'axe défini par la variable `PVP` devient

$$18.0 - 12.2125 = 5.7875 .$$

Cette valeur est également la nouvelle coordonnée de la variable `PVP` sur l'axe défini par l'année`1872`.

5 Moments d'inertie

5.1 Inertie totale du nuage des individus

On note I_G le moment d'inertie du nuage des individus par rapport au centre de gravité G :

$$I_G = \frac{1}{n}\sum_{i=1}^{n} d^2(G, u_i) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - x_{\bullet j})^2 = \frac{1}{n}\sum_{i=1}^{n} {}^tU_{ci}\, U_{ci}\ .$$

Ce moment d'inertie totale est intéressant car c'est une mesure de la dispersion du nuage des individus par rapport à son centre de gravité. Si ce moment d'inertie est grand, cela signifie que le nuage est très dispersé, tandis que s'il est petit, alors le nuage est très concentré sur son centre de gravité.

Remarque On peut voir, en inversant l'ordre des signes sommes, que I_G peut aussi s'écrire sous la forme suivante :

$$I_G = \sum_{j=1}^{p}\left[\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - x_{\bullet j})^2\right] = \sum_{j=1}^{p}\mathrm{Var}(v_j)$$

où $\mathrm{Var}(v_j)$ est la variance empirique de la variable v_j. Sous cette forme, on constate que l'inertie totale est égale à la trace de la matrice de covariance $\mathbf{\Sigma}$ (cf. annexe A) des p variables v_j :

$$I_G = \mathrm{trace}(\mathbf{\Sigma}).$$

Exemple Le tableau `Covariance Matrix` de la page 29 donne la matrice $\mathbf{\Sigma}$ pour les 11 variables étudiées. La rubrique `Total Variance` donne la somme des termes de la diagonale de $\mathbf{\Sigma}$ (*i.e.* les $\mathrm{Var}(v_j)$) :

$$I_G = \mathrm{trace}(\mathbf{\Sigma}) = 310.57\ .$$

On peut remarquer que les variances sont très hétérogènes : $\mathrm{Var}(\texttt{DET}) = 148.69$, $\mathrm{Var}(\texttt{DIV}) = 1.052$. La contribution du remboursement de la dette à l'inertie totale est donc beaucoup plus forte, ce poste influera plus sur les résultats de l'analyse que le poste "Divers".

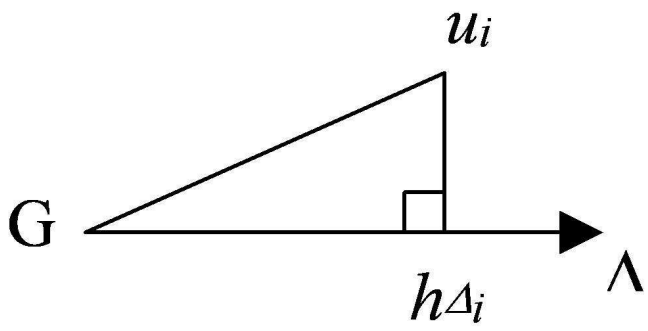
5.2 Inertie du nuage des individus par rapport à un axe passant par G

L'inertie du nuage des individus par rapport à un axe Δ passant par G est égale, par définition, à :

$$I_\Delta = \frac{1}{n}\sum_{i=1}^{n} d^2(h_{\Delta i}, u_i)$$

où $h_{\Delta i}$ est la projection orthogonale de u_i sur l'axe Δ.

Cette inertie mesure la proximité à l'axe Δ du nuage des individus.



5.3 Inertie du nuage des individus par rapport à un sous-espace vectoriel V passant par G

Cette inertie est, par définition, égale à :

$$I_V = \frac{1}{n}\sum_{i=1}^{n} d^2(h_{Vi}, u_i)$$

où h_{Vi} est la projection orthogonale de u_i sur le sous-espace V.

5.4 Décomposition de l'inertie totale

Si on note V^* le complémentaire orthogonal de V dans $\mathbb{R}^p$ et h_{V^*i} la projection orthogonale de u_i sur V^*, en appliquant le théorème de Pythagore, on peut écrire :

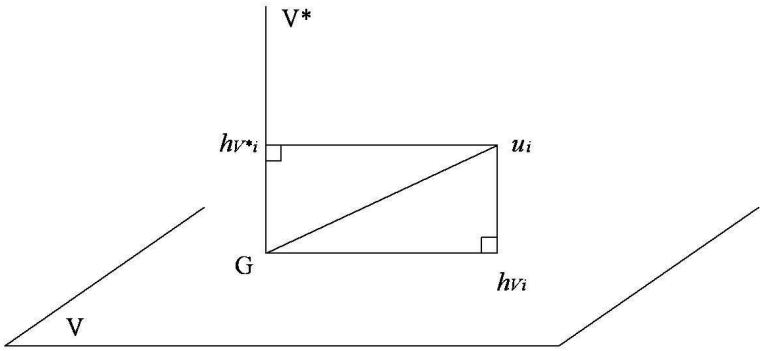
$$d^2(h_{Vi}, u_i) + d^2(h_{V^*i}, u_i) = d^2(G, u_i) = d^2(G, h_{Vi}) + d^2(G, h_{V^*i}).$$

On en déduit, c'est le théorème de Huygens, que :

$$I_V + I_{V^*} = I_G$$

Dans le cas particulier où le sous-espace est de dimension 1, c'est-à-dire est un axe, I_{V^*} est une mesure de l'allongement du nuage selon cet axe. On emploie pour I_{V^*} les expressions

"d'inertie portée par l'axe" ou bien "d'inertie expliquée par l'axe".



En projetant le nuage des individus sur un sous-espace V, on perd l'inertie mesurée par I_V, on ne conserve que celle mesurée par I_{V^*}.

De plus, si on décompose l'espace $\mathbb{R}^p$ comme la somme de sous-espaces de dimension 1 et orthogonaux entre eux :

$$\Delta_1 \oplus \Delta_2 \oplus ... \oplus \Delta_p$$

on peut écrire (cf. annexe page 47) :

$$I_G = I_{\Delta_1^*} + I_{\Delta_2^*} + ... + I_{\Delta_p^*}$$

6 Recherche de l'axe Δ_1 passant par G d'inertie minimum

On cherche un axe Δ_1 passant par G d'inertie I_{Δ_1} minimum car c'est l'axe le plus proche de l'ensemble des points du nuage des individus, et donc, si l'on doit projeter ce nuage sur cet axe, c'est lui qui donnera l'image la moins déformée du nuage. Si on utilise la relation entre les inerties donnée au paragraphe précédent, rechercher Δ_1 tel que I_{Δ_1} est minimum, est équivalent à chercher Δ_1 tel que $I_{\Delta_1^*}$ est maximum.

$$I_{\Delta_1} \text{ est minimum} \Longleftrightarrow I_{\Delta_1^*} \text{ est maximum}$$

On définit l'axe Δ_1 par son vecteur directeur unitaire $\overrightarrow{Ga_1}$.

Il faut donc trouver $\overrightarrow{Ga_1}$ tel que $I_{\Delta_1^*}$ est maximum sous la contrainte que $\left\|\overrightarrow{Ga_1}\right\|^2 = 1$.

6.1 Expressions algébriques de $I_{\Delta_1^*}$ et de $\left\|\overrightarrow{Ga_1}\right\|^2$

$$d^2(G, h_{\Delta_1 i}) = \left\langle \overrightarrow{Gu_i}, \overrightarrow{Ga_1} \right\rangle^2 = {}^t a_1 U_{ci}\ {}^t U_{ci}\ a_1$$

en utilisant la symétrie du produit scalaire. On en déduit :

$$I_{\Delta_1^*} = \frac{1}{n}\sum_{i=1}^{n} {}^t a_1 U_{ci}\ {}^t U_{ci}\ a_1 = {}^t a_1 \left[\frac{1}{n}\sum_{i=1}^{n} U_{ci}\ {}^t U_{ci}\right] a_1$$

Entre crochets on reconnaît la matrice de covariance empirique $\mathbf{\Sigma}$ des p variables (cf. annexe page 45).

$$I_{\Delta_1^*} = {}^t a_1 \mathbf{\Sigma}\ a_1$$

et

$$\left\|\overrightarrow{Ga_1}\right\|^2 = {}^t a_1\ a_1\ .$$

6.2 Recherche du maximum

Le problème à résoudre : trouver a_1 tel que ${}^t a_1 \mathbf{\Sigma}\ a_1$ soit maximum avec la contrainte ${}^t a_1\ a_1 = 1$ est le problème de la recherche d'un optimum d'une fonction de plusieurs variables liées par une contrainte (les inconnues sont les composantes de a_1). La méthode des multiplicateurs de Lagrange peut alors être utilisée (cf. annexe page 48).

Dans le cas de la recherche de a_1, il faut calculer les dérivées partielles de :

$$g(a_1) = g(a_{11}, a_{12}, ..., a_{1p}) = {}^t a_1 \mathbf{\Sigma}\ a_1 - \lambda_1({}^t a_1\ a_1 - 1).$$

En utilisant la dérivée matricielle (cf. annexe page 49), on obtient :

$$\frac{\partial g(a_1)}{\partial a_1} = 2\mathbf{\Sigma}\ a_1 - 2\lambda_1\ a_1 = 0.$$

Le système à résoudre est :

$$\left\{ \begin{array}{lcccll} \mathbf{\Sigma}\ a_1 & - & \lambda_1\ a_1 & = & 0 & (1) \\ {}^t a_1\ a_1 & - & 1 & = & 0 & (2) \end{array} \right.$$

De l'équation matricielle (1) de ce système on déduit que a_1 est vecteur propre de la matrice $\mathbf{\Sigma}$ associé à la valeur propre λ_1.

En multipliant à gauche par ${}^t a_1$ les deux membres de l'équation (1) on obtient :

$${}^t a_1 \mathbf{\Sigma}\ a_1 - \lambda_1\ {}^t a_1\ a_1 = 0$$

et en utilisant l'équation (2) on trouve que :

$${}^t a_1 \mathbf{\Sigma}\ a_1 = \lambda_1\ .$$

On reconnaît que le premier membre de l'équation précédente est égal à l'inertie $I_{\Delta_1^*}$ qui doit être maximum. Cela signifie que la valeur propre λ_1 est la plus grande valeur propre de la matrice de covariance $\mathbf{\Sigma}$ et que cette valeur propre est égale à l'inertie portée par l'axe Δ_1.

L'axe Δ_1 pour lequel le nuage des individus a l'inertie minimum a comme vecteur directeur unitaire le premier vecteur propre associé à la plus grande valeur propre de la matrice de covariance $\mathbf{\Sigma}$.

Exemple La ligne `AXE1` du tableau `Eigenvalues of the Covariance Matrix` de la page 29 nous donne

$$I_{\Delta_1^*} = \lambda_1 = 204.717.$$

et nous indique de plus (dans la colonne `Proportion`) que cette inertie représente

$$\frac{I_{\Delta_1^*}}{I_G} = \frac{204.717}{310.572} = 65.91\%$$

de l'inertie totale

Les coordonnées du vecteur propre a_1 correspondant se trouvent dans la colonne `AXE1` du tableau `Eigenvectors`.

7 Recherche des axes suivants

On recherche ensuite un deuxième axe Δ_2 orthogonal au premier et d'inertie minimum. On peut, comme dans le paragraphe précédent, définir l'axe Δ_2 passant par G par son vecteur directeur unitaire a_2. L'inertie du nuage des individus par rapport à son complémentaire orthogonal est égale à :

$$I_{\Delta_2^*} = {}^t a_2 \mathbf{\Sigma}\, a_2,$$

elle doit être maximum avec les deux contraintes suivantes :

$${}^t a_2\, a_2 = 1 \quad \text{et} \quad {}^t a_2\, a_1 = 0.$$

La deuxième contrainte exprime que le deuxième axe doit être orthogonal au premier et donc que le produit scalaire des deux vecteurs directeurs est nul. En appliquant la méthode des multiplicateurs de Lagrange, cette fois avec deux contraintes, on trouve que a_2 est le vecteur propre de $\mathbf{\Sigma}$ correspondant à la deuxième plus grande valeur propre. On peut montrer que le plan défini par les axes Δ_1 et Δ_2 est le sous-espace de dimension 2 qui porte l'inertie maximum.

On peut rechercher de nouveaux axes en suivant la même procédure. Les nouveaux axes sont tous vecteurs propres de $\mathbf{\Sigma}$ correspondant aux valeurs propres ordonnées. La matrice de covariance $\mathbf{\Sigma}$ étant une matrice symétrique réelle, elles possède p vecteurs propres réels, formant une base orthogonale de $\mathbb{R}^p$:

$$\left\{ \begin{array}{ccccccc} \Delta_1 & \perp & \Delta_2 & \perp & \dots & \perp & \Delta_p \\ a_1 & \perp & a_2 & \perp & \dots & \perp & a_p \\ \lambda_1 & \geq & \lambda_2 & \geq & \dots & \geq & \lambda_p \\ I_{\Delta_1^*} & \geq & I_{\Delta_2^*} & \geq & \dots & \geq & I_{\Delta_p^*} \end{array} \right.$$

On passera de la base orthogonale initiale des variables centrées à la nouvelle base orthogonale des vecteurs propres de $\mathbf{\Sigma}$. On appelle les nouveaux axes, axes principaux.

Exemple Les vecteurs propres a_2, a_3, *etc* sont donnés par les colonnes du tableau `Eigenvectors` page 29. Les valeurs propres correspondantes sont données dans le paragraphe `Eigenvalues of the Covariance Matrix` page 29. On note que la dernière valeur propre λ_{11} est nulle : cela provient du fait que le nuage est en fait compris dans un espace de dimension 10 (et non 11).

8 Contributions des axes à l'inertie totale

En utilisant le théorème de Huygens, on peut décomposer l'inertie totale du nuage des individus (voir annexe page 47) :

$$I_G = I_{\Delta_1^*} + I_{\Delta_2^*} + ... + I_{\Delta_p^*} = \lambda_1 + \lambda_2 + ... + \lambda_p$$

La contribution absolue de l'axe Δ_k à l'inertie totale du nuage des individus est égale à :

$$\text{ca}(\Delta_k / I_G) = \lambda_k$$

valeur propre qui lui est associée.

Sa contribution relative est égale à :

$$\text{cr}(\Delta_k / I_G) = \frac{\lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p}$$

On emploie souvent l'expression "pourcentage d'inertie expliquée par Δ_k".

On peut étendre ces définitions à tous les sous-espaces engendrés par les nouveaux axes. Ainsi, le pourcentage d'inertie expliqué par le plan engendré par les deux premiers axes Δ_1 et Δ_2 est égal à :

$$\text{cr}(\Delta_1 \oplus \Delta_2 / I_G) = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + ... + \lambda_p}$$

Ces pourcentages d'inertie sont des indicateurs qui rendent compte de la part de variabilité du nuage des individus expliquée par ces sous-espaces. Si les dernières valeurs propres ont des valeurs faibles, on pourra négliger la variabilité qu'expliquent les axes correspondants.

On se contente souvent de faire des représentations du nuage des individus dans un sous-espace engendré par les d premiers axes si ce sous-espace explique un pourcentage d'inertie proche de 1. On peut ainsi réduire l'analyse à un sous-espace de dimension $d < p$.

Exemple Le tableau `Eigenvalues of the Covariance Matrix` page 29 nous donne les inerties (valeurs propres) associées aux différents axes. On peut vérifier que leur somme est bien égale à l'inertie totale :

$$\sum_i \lambda_i = \sum_i I_{\Delta_i^*} = 310.57 \ .$$

La colonne `Cumulative` nous indique que les axes Δ_1, Δ_2 et Δ_3 portent à eux trois 90.96% de l'inertie totale. On utilise souvent la colonne `Difference` pour choisir le nombre d'axes à conserver dans l'étude. Ici on trouve une forte chute entre les axes Δ_4 et Δ_5 ($\lambda_4 - \lambda_5 = 10.88$) ce qui, joint à la forte part d'inertie portée par Δ_1, Δ_2 et Δ_3, suggère de conserver les 3 ou les 4 premiers axes dans l'analyse.

9 Représentation des individus dans les nouveaux axes

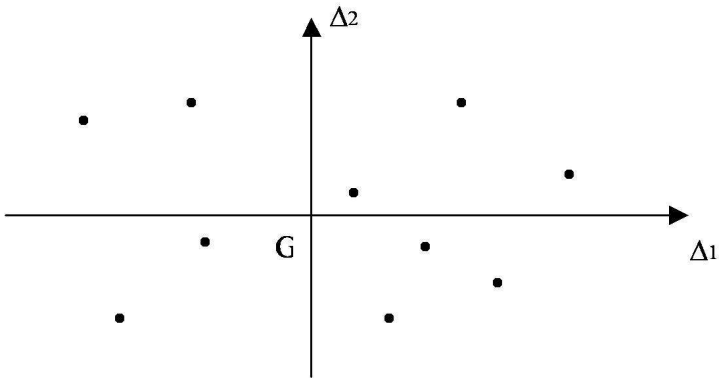
Pour faire la représentation des individus dans les plans définis par les nouveaux axes, il suffit de calculer les coordonnées des individus dans les nouveaux axes. Pour obtenir y_{ik}, coordonnée de l'unité u_i sur l'axe Δ_k, on projette orthogonalement le vecteur $\overrightarrow{Gu_i}$ sur cet axe et on obtient :

$$y_{ik} = \left\langle \overrightarrow{Gu_i}, \overrightarrow{a_k} \right\rangle = {}^t a_k \, U_{ci}$$

et

$$Y_i = {}^t\mathbf{A} \, U_{ci}$$

où Y_i est le vecteur des coordonnées de l'unité u_i et $\mathbf{A}$ est la matrice du changement de base ($\mathbf{A}$ matrice des vecteurs propres orthogonaux et de norme 1 est une matrice orthogonale, son inverse est égale à sa transposée cf. annexe page 51).



Remarque L'orientation des axes est complètement arbitraire et peut différer d'un logiciel à l'autre. Le signe des coordonnées des individus sur un axe n'a donc pas de signification. En revanche, la comparaison des signes peut s'interpréter. Si deux individus u_i et $u_{i'}$ ont sur un axe Δ, le premier une coordonnée positive et le second une coordonnée négative, cela signifie qu'ils s'opposent sur cet axe.

Exemple Le tableau `Coordonnees des individus sur les axes principaux` page 30 donne les coordonnées y_{ik} des différents individus dans le nouveau système d'axes. Le tableau `Eigenvectors` donne la matrice de changement de base $\mathbf{A}$. Le graphe page 33 montre la projection du nuage dans le premier plan principal dit `Plan 1-2`.

9.1 Qualité de la représentation des individus

Lorsque des points projections des individus sont éloignés sur un axe (ou sur un plan), on peut assurer que les points représentants ces individus sont éloignés dans l'espace. En

revanche, deux individus dont les projections sont proches sur un axe (ou sur un plan) peuvent ne pas être proches dans l'espace.

Pour interpréter correctement la proximité des projections de deux individus sur un plan, il faut donc s'assurer que ces individus sont bien représentés dans le plan. Pour que l'individu u_i soit bien représenté sur un axe (ou sur un plan, ou un sous-espace), il faut que l'angle entre le vecteur $\overrightarrow{Gu_i}$ et l'axe (ou le plan, ou le sous-espace) soit petit. On calcule donc le cosinus de cet angle, ou plutôt le carré de ce cosinus. En effet, en utilisant le théorème de Pythagore, on peut montrer que le carré du cosinus de l'angle d'un vecteur avec un plan engendré par deux vecteurs orthogonaux, est égal à la somme des carrés des cosinus des angles du vecteur avec chacun des deux vecteurs qui engendrent le plan. Cette propriété se généralise à l'angle d'un vecteur avec un sous-espace de dimension k quelconque. Si le carré du cosinus de l'angle entre $\overrightarrow{Gu_i}$ et l'axe (ou le plan, ou le sous-espace) est proche de 1, alors on pourra dire que l'individu u_i est bien représenté par sa projection sur l'axe (ou le plan, ou le sous-espace). Et si deux individus sont bien représentés en projection sur un axe (ou un plan, ou un sous-espace) et ont des projections proches, alors on pourra dire que ces deux individus sont proches dans l'espace. Le carré du cosinus de l'angle α_{ik} entre $\overrightarrow{Gu_i}$ et un axe Δ_k de vecteur directeur unitaire a_k est égal à :

$$\cos^2(\alpha_{ik}) = \frac{\left\langle \overrightarrow{Gu_i}, \overrightarrow{Ga_k} \right\rangle^2}{\left\| \overrightarrow{Gu_i} \right\|^2} = \frac{{}^t a_k U_{ci}\ {}^t U_{ci}\ a_k}{{}^t U_{ci}\ U_{ci}} = \frac{\left[\sum_{j=1}^{p}(x_{ij} - x_{\bullet j})\ a_{kj}\right]^2}{\sum_{j=1}^{p}(x_{ij} - x_{\bullet j})^2}$$

En utilisant le théorème de Pythagore on peut calculer le carré du cosinus de l'angle $\alpha_{ikk'}$ entre $\overrightarrow{Gu_i}$ et le plan engendré par deux axes $\Delta_k \oplus \Delta_{k'}$:

$$\cos^2(\alpha_{ikk'}) = \cos^2(\alpha_{ik}) + \cos^2(\alpha_{ik'})$$

Si, après l'étude des pourcentages d'inertie expliqués par les sous-espaces successifs engendrés par les nouveaux axes, on a décidé de ne retenir qu'un sous-espace de dimension $d < p$, on pourra calculer la qualité de la représentation d'un individu u_i en calculant le carré du cosinus de l'angle de $\overrightarrow{Gu_i}$ avec ce sous-espace.

Remarque Si un individu est très proche du centre de gravité dans l'espace, c'est-à-dire si $\left\| \overrightarrow{Gu_i} \right\|^2$ est très petit, le point représentant cet individu sur un axe (ou un plan, ou un sous-espace) sera bien représenté.

Exemple Le tableau `Qualite de la représentation des individus` page 31 donne les cosinus carrés $\cos^2(\alpha_{ik})$ pour les axes Δ_1, Δ_2, Δ_3 (`COS2-1`, `COS2_2`, `COS2_3`), pour le le plan 1-2 ($\cos^2(\alpha_{i12})$ = `COS2_12`) et pour l'espace 1-2-3 ($\cos^2(\alpha_{i123})$ = `COS2_123`).

La lecture de ce tableau nous montre que la plupart des années sont bien représentées dans le plan 1-2. On peut cependant noter que les années `1932` et `1935` ne sont pas très bien représentées. Leur apparente proximité sur le graphe page 33 ne peut donc pas être interprétée directement. On peut d'ailleurs noter en revenant au tableau `Coordonnées`

`des individus sur les axes principaux` que leurs coordonnées sur l'axe Δ_3 sont assez différentes (3.85 pour 1932 et 6.22 pour 1935). La `Projection dans l'espace 1-2-3` donnée page 34 montre effectivement que l'année `1935` est située au dessus de l'année `1932` (attention à l'effet de perspective !).

9.2 Interprétation des nouveaux axes en fonction des individus

Lorsqu'on calcule l'inertie $I_{\Delta_k^*}$ portée par l'axe Δ_k,on peut voir quelle est la part de cette inertie due à un individu u_i particulier.

9.2.1 Contribution absolue d'un individu à un axe

$I_{\Delta^* k}$ étant égale à $\frac{1}{n}\sum_{i=1}^{n} d^2(h_{\Delta_{ki}}, G)$, la contribution absolue de u_i à cette inertie est égale à :

$$\text{ca}(u_i/\Delta_k) = \frac{1}{n} d^2(h_{\Delta_{ki}}, G)$$

puisque tous les individus ont le même poids. Un individu contribuera d'autant plus à la confection d'un axe, que sa projection sur cet axe sera éloignée du centre de gravité du nuage. Inversement, un individu dont la projection sur un axe sera proche du centre de gravité contribuera faiblement à l'inertie portée par cette axe. On se sert de ces contributions pour interpréter les nouveaux axes de l'ACP en fonction des individus.

9.2.2 Contribution relative d'un individu à un axe

On peut aussi, pour un individu particulier u_i, donner sa contribution relative à l'inertie portée par cet axe :

$$\text{cr}(u_i/\Delta_k) = \frac{\frac{1}{n} d^2(h_{\Delta_{ki}}, G)}{I_{\Delta_k^*}} = \frac{\frac{1}{n}\left\langle \overrightarrow{Gu_i}, \overrightarrow{Ga_k} \right\rangle^2}{\lambda_k} = \frac{\frac{1}{n}\, {}^t a_k U_{ci}\, {}^t U_{ci}\, a_k}{\lambda_k}$$

L'examen de ces contributions permet d'interpréter les axes principaux avec les individus. On peut remarquer que $\sum_{i=1}^{n} \text{cr}(u_i/\Delta_k) = 1$.

Exemple Les contributions relatives $\text{cr}(u_i/\Delta_k)$ sont données dans le tableau `Contributions relatives de individus à l'inertie des axes` page 31. Pour chaque axe, on vérifie que la somme des contributions relatives vaut bien 1. La colonne `Norme2` représente $\left\|\overrightarrow{Gu_i}\right\|^2$. On peut noter les années les plus éloignées du barycentre : `1872`, `1926`, `1950`, `1968`. Le fait que les années du début (avant 1900) et de la fin (après 1960) de la période étudiée soient éloignées du barycentre nous rappelle les limites de notre analyse qui ne vaut que pour cette période et ne peut donner lieu à aucune extrapolation.

On peut remarquer que les coordonnées des individus sur le premier axe (page 30) respectent presque systématiquement l'ordre chronologique : on pourrait donc dire que Δ_1 représente le temps. Pour ce qui est de Δ_2, la plus forte contribution (et de loin) est

celle de l'année 1926. Cet axe semble opposer la période 1925-1935 (ainsi que la toute fin des années 60) à l'avant première guerre mondiale et à la reconstruction de l'après la seconde guerre mondiale. On peut sans doute parler d'une opposition entre les périodes de crise économique et les périodes de plus forte activité.

10 Représentation des variables

On peut envisager le problème de la représentation des variables de façon complètement symétrique de celui des individus. Les raisonnements se font dans $\mathbb{R}^n$ au lieu de $\mathbb{R}^p$. Mais dans l'ACP, au delà de la symétrie formelle entre les individus et les variables, on peut utiliser la dissymétrie liée à la sémantique : les variables n'ont pas la même signification que les individus. On peut alors faire le raisonnement suivant : on a représenté les individus dans l'espace des anciennes variables, et on a fait un changement de base dans cet espace. Les nouveaux axes sont des combinaisons linéaires des anciens axes et peuvent donc être considérés comme de nouvelles variables combinaisons linéaires des anciennes. On appelle communément ces nouvelles variables "composantes principales".

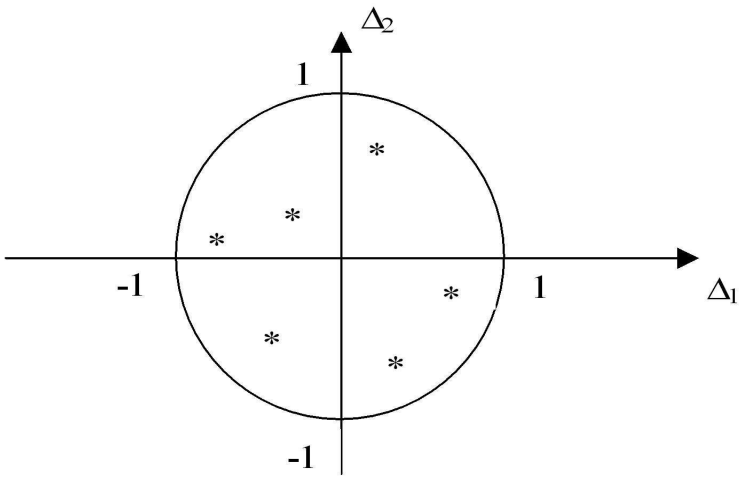
On note Z_1, Z_2, ..., Z_k, ..., Z_p les composantes principales, Z_k étant la nouvelle variable correspondant à l'axe Δ_k :

$$Z_k = \sum_{j=1}^{p} a_{kj} V_{cj} = \mathbf{X}_c \, a_k$$

et de façon générale :

$$\mathbf{Z} = [Z_1 \; Z_2 ... \; Z_k ... \; Z_p] = \mathbf{X}_c \mathbf{A} \; .$$

Il est alors intéressant de voir comment les anciennes variables sont liées aux nouvelles et pour cela on calcule les corrélations des anciennes variables avec les nouvelles. La représentation des anciennes variables se fera en prenant comme coordonnées des anciennes variables leurs coefficients de corrélation avec les nouvelles variables. On obtient alors ce que l'on appelle communément le "cercle des corrélations", dénomination qui vient du fait qu'un coefficient de corrélation variant entre -1 et +1, les repr ésentations des variables de départ sont des points qui se trouvent à l'intérieur d'un cercle de rayon 1 si on fait la représentation sur un plan.



On peut montrer que les variances, covariances et coefficients de corrélation empiriques des composantes principales entre elles ou avec les variables de départ sont :

$$\mathrm{Var}(Z_k) = \frac{1}{n} \, {}^t a_k \, {}^t\mathbf{X}_c \, \mathbf{X}_c \, a_k = {}^t a_k \, \mathbf{\Sigma} \, a_k = \lambda_k$$

$$\mathrm{Cov}(Z_k, V_{cj}) = \frac{1}{n}\ {}^t a_k\ {}^t\mathbf{X}_c\ V_{cj} = \frac{1}{n}\ {}^t a_k\ {}^t\mathbf{X}_c\ \mathbf{X}_c \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

$$= {}^t a_k\ \mathbf{\Sigma} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = \lambda_k\ {}^t a_k \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = \lambda_k\ a_{kj}\ .$$

Enfin :

$$\mathrm{Cor}(Z_k, V_{cj}) = \sqrt{\lambda_k}\frac{a_{kj}}{\sqrt{\mathrm{Var}(V_j)}}$$

où a_{kj} est la jème coordonnée du vecteur directeur unitaire a_k de Δ_k.

De façon générale, la matrice de covariance des composantes principales est égale à (voir annexe page 51) $\mathbf{\Sigma_Z}$:

$$\mathbf{\Sigma_Z} = \frac{1}{n}\ {}^t\mathbf{A}\ {}^t\mathbf{X}_c\ \mathbf{X}_c\ \mathbf{A} = {}^t\mathbf{A}\ \mathbf{\Sigma}\ \mathbf{A} = \mathbf{\Lambda}\ ,$$

où (voir annexe page 52) $\mathbf{\Lambda}$ est la matrice diagonale des valeurs propres de $\mathbf{\Sigma}$:

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & & (\mathbf{0}) \\ & \ddots & \\ (\mathbf{0}) & & \lambda_p \end{pmatrix}$$

et la matrice des covariances entre les composantes principales et les anciennes variables vaut

$$\mathrm{Cov}(\mathbf{Z}, \mathbf{V}) = \frac{1}{n}\ {}^t\mathbf{X}_c\ \mathbf{X}_c\ \mathbf{A} = \mathbf{\Sigma}\ \mathbf{A} = \mathbf{A}\ \mathbf{\Lambda}\ .$$

Si on remarque que la variance empirique d'une variable est égale au carré de la norme du vecteur qui la représente dans la géométrie euclidienne choisie et que le coefficient de corrélation empirique de deux variables est égal au produit scalaire des deux vecteurs qui les représentent (cf. annexe page 50), on pourra interpréter les angles des vecteurs comme des corrélations.

Exemple. Les coordonnées des nouvelles variables Z_k dans l'espace des individus se lisent en colonne dans le tableau `Coordonnes factorielles des individus` 30. Les corrélations des nouvelles variables avec les anciennes sont données page 32 dans le tableau `Corrélations variables axes`.

10.1 Interprétation des axes en fonction des anciennes variables

On peut interpréter les axes principaux en fonction des anciennes variables. Une ancienne variable V_j expliquera d'autant mieux un axe principal qu'elle sera fortement corrélée avec la composante principale correspondant à cet axe.

10.2 Qualité de la représentation des variables

Pour les mêmes raisons qui ont poussé à se préoccuper de la qualité de la représentation des individus, il faut se préoccuper de la qualité de la représentation des variables sur un axe, un plan ou un sous-espace. Une variable sera d'autant mieux représentée sur un axe que sa corrélation avec la composante principale correspondante est en valeur absolue proche de 1. En effet (cf. annexe page 50), le coefficient de corrélation empirique entre une ancienne variable V_{cj} et une nouvelle variable Z_k n'est autre que le cosinus de l'angle du vecteur joignant l'origine au point v_j représentant la variable sur l'axe avec cet axe.

Une variable sera bien représentée sur un plan si elle est proche du bord du cercle des corrélations, car cela signifie que le cosinus de l'angle du vecteur joignant l'origine au point représentant la variable avec le plan est, en valeur absolue, proche de 1, *etc.*

10.3 étude des liaisons entre les variables

Sur le graphique du cercle des corrélations, on peut aussi interpréter les positions des anciennes variables les unes par rapport aux autres en termes de corrélations. Deux points très proches du cercle des corrélations, donc bien représentées dans le plan, seront très corrélées positivement entre elles. Si elles sont proches du cercle, mais dans des positions symétriques par rapport à l'origine, elles seront très corrélées négativement.

Deux variables proches du cercle des corrélations et dont les vecteurs qui les joignent à l'origine forment un angle droit, ne seront pas corrélées entre elles.

Il faut, pour interpréter correctement ces graphiques des cercles de corrélation, se souvenir qu'un coefficient de corrélation est une mesure de liaison linéaire entre deux variables, et qu'il peut arriver que deux variables très fortement liées aient un coefficient de corrélation nul ou très faible, si leur liaison n'est pas linéaire.

Exemple La lecture directe du tableau des corrélations entre les anciennes et les nouvelles variables page 32 montre de fortes corrélations entre le premier axe et la Dette (négative), le Commerce, le Logement et l'Agriculture (positives). L'opposition entre ces 2 groupes de variables se retrouve dans le `Cercle des corrélations dans le plan1-2` page 35. Cette opposition entre `DET` et le groupe {`EDU`, `ACS`, `AGR`, `CMI`, `LOG`} (qui était déjà visible dans la matrice de covariance page 29) peut fournir une interprétation de l'axe Δ_1. L'axe Δ_2 est lui très lié à la variable `DEF` : ce poste budgétaire semble avoir une évolution assez différente des autres postes.

La proximité des variables `EDU`, `ACS`, `AGR`, `CMI` et `LOG` sur le graphe page 35 permet de conclure qu'il existe un lien fort entre ces variables mais cette interprétation ne vaut que

parce ces variables sont bien représentées dans le plan 1-2 (les points correspondants sont proches du cercle). La proximité des variables `PVP` et `TRA` ne peut pas être interprétée : ces variables sont mal représentées puisque les points correspondants sont éloignés du cercle. La lecture de la matrice de covariance page 29 montre d'ailleurs que le lien est assez faible : $\mathrm{Var}(\texttt{PVP}) = 4.801$, $\mathrm{Var}(\texttt{TRA}) = 6.0899$, $\mathrm{Cov}(\texttt{PVP}, \texttt{TRA}) = 1.2585$.

11 Analyse en composantes principales normée

Dans les paragraphes précédents, nous avons étudié l'ACP simple, pour laquelle, non seulement tous les individus ont le même poids dans l'analyse, mais aussi, toutes les variables sont traitées de façon symétrique (on leur fait jouer le même rôle) et les nouveaux axes sont issus de la matrice de covariance empirique des variables. Cela pose parfois des problèmes. Le premier reproche fait par des praticiens est que, si les anciennes variables sont hétérogènes, comme par exemple des poids, des tailles et des âges, quel sens peut-on donner aux composantes principales qui sont alors des combinaisons linéaires de variables hétéroclites ? Le deuxième reproche, est que, si on change d'unités sur ces variables, on peut changer complètement les résultats de l'ACP. Le dernier reproche vient du fait qu'une variable contribuera d'autant plus à la confection des premiers axes, que sa variance est forte.

Pour échapper à tous ces problèmes, on cherchera à normaliser les variables et à travailler sur des variables sans dimension. Il y a plusieurs façons de normaliser les variables, mais la plus couramment utilisée est celle qui consiste à diviser les valeurs des variables par leur écart-type, c'est-à-dire que l'on travaille sur des variables centrées et réduites.

Cela revient à faire la même analyse que pour l'ACP simple, mais à choisir une autre distance euclidienne entre les individus que la distance euclidienne classique. La distance choisie est alors :

$$d^2(u_i, u_{i'}) = \sum_{j=1}^{p} \frac{1}{\sigma_j^2}(x_{ij} - x_{i'j})^2 \ .$$

Cette nouvelle distance ne traite plus les variables de façon symétrique, mais elle permet de faire jouer un rôle plus équitable à chacune d'entre elles.

Si on reprend tous les calculs de l'ACP simple, mais en remplaçant les variables de départ par les variables centrées réduites, on voit que ce n'est plus la matrice de covariance, mais la matrice de corrélation $\mathbf{R}$ qui intervient pour la recherche des nouveaux axes (cf. annexe page 45). Les particularités de l'ACP normée par rapport à l'ACP simple proviennent du fait que la matrice de corrélation $\mathbf{R}$ n'a que des 1 sur sa diagonale principale. Cela entraîne que sa trace est toujours égale à p. On a vu que la trace de la matrice est égale à l'inertie totale du nuage calculée avec la distance euclidienne que l'on a choisie. L'inertie totale du nuage des individus dans $\mathbb{R}^p$ est donc toujours égale à p dans toute ACP normée. Cette particularité donne une règle supplémentaire pour choisir le nombre d'axes que l'on va garder pour les interprétations, fondée sur le raisonnement suivant : on a p valeurs propres dont la somme vaut p (puisque l'on a vu que l'inertie totale est aussi égale à la somme des valeurs propres) ; on peut ne considérer comme significatives que les valeurs propres dont la valeur est supérieure à 1, puisque la valeur moyenne des valeurs propres vaut 1 et leur somme vaut p. C'est bien sûr une règle empirique mais qui peut servir de guide pour le choix de la dimension du sous-espace que l'on veut garder.

Une autre particularité de l'ACP normée est que la représentation des variables avec les cercles de corrélation correspond exactement à la représentation des variables dans $\mathbb{R}^n$ que l'on aurait construite si l'on avait adopté la même démarche que celle qui a servi pour la représentation des individus dans $\mathbb{R}^p$.

Exemple. Les listings et les graphiques présentés de la page 36 à la page 43 sont les résultats d'une ACP normée sur les budgets de l'état français de 1872 à 1971. Le fait de normer l'ACP permet d'éliminer les effets de masse qui induisaient notamment la prédominance des variables `DET` (en moyenne 19.14% du budget total) et `DEF` (en moyenne 30.26%) sur les autres.

On note sur la sortie page 37 que SAS utilise cette fois la matrice des corrélations entre les variables. On lit bien, page 37, que la somme des valeurs propres vaut 11, c'est-à-dire le nombre de variables explicatives : (la dernière valeur propre est nulle pour les même raisons que dans l'ACP simple). Ici les trois premiers axes ne portent que 75.6% de l'inertie totale et il faut aller jusqu'au cinquième axe pour obtenir 91.0% (colonne `Cumulative` du tableau `Eigenvalues of the Correlation Matrix` page 37).

Le graphique page 41 n'appelle pas une interprétation très différente de celle de l'ACP simple. On peut cependant noter que l'axe Δ_2 est orienté dans l'autre sens : on rappelle que l'orientation des axes est arbitraire. Dans la projection dans l'espace 1-2-3 donnée page 42, on peut noter l'émergence de l'année 1968 le long de l'axe Δ_3.

Sur le graphique du cercle des corrélations donné page 43 on retrouve le groupe homogène de variables {`LOG`, `CMI`, `AGR`, `ACS`, `EDU`} opposé à la variable `DET`. Le poste `DEF` est par contre beaucoup moins bien représenté : une fois éliminé (par normalisation) l'effet de masse de ce poste, sa contribution au premier plan devient faible.

12 Individus et variables supplémentaires

Il arrive que l'on veuille faire apparaître dans les représentations graphiques certains individus sans qu'ils interviennent dans la détermination des axes. Cela peut être le cas de nouveaux individus que l'on veut simplement positionner par rapport aux autres sans que les positions de ceux-ci soient influencées par les nouveaux. On dit d'eux qu'ils sont des individus supplémentaires.

Il en est de même pour les variables. On peut, par exemple, vouloir représenter une variable qui dépend de façon synthétique des p variables choisies pour faire l'ACP, afin de mieux comprendre comment cette variable est liée aux anciennes, mais on ne souhaite pas qu'elle intervienne dans la confection des axes car ses liaisons avec les p variables de départ fausseraient la représentation si elle faisait partie intégrante de l'ACP. Elles sont appelées variables supplémentaires.

Pour représenter un individu supplémentaire, il suffit d'exprimer les coordonnées de cet individu dans la nouvelle base des axes principaux. Pour une variable supplémentaire, il suffit de calculer ses coefficients de corrélation empiriques avec les composantes principales. La plupart des logiciels proposent des options permettant de le faire.

13 Exemple : Budgets de l'état de 1872 à 1971

Le programme principal de l'A.C.P ainsi que les macros utilisées dans ce programme sont disponibles à l'adresse $T:\backslash math\backslash Macro\text{-}SAS\backslash ACP$. Les pages suivantes donnent les listings obtenus avec ce programme pour une ACP simple et une ACP normée.

Budgets de l'état de 1872 à 1971 : ACP simple

Donnees brutes

OBS	AN	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV
1	1872	18.0	0.5	0.1	6.7	0.5	2.1	2.0	0.0	26.4	41.5	2.1
2	1880	14.1	0.8	0.1	15.3	1.9	3.7	0.5	0.0	29.8	31.3	2.5
3	1890	13.6	0.7	0.7	6.8	0.6	7.1	0.7	0.0	33.8	34.4	1.7
4	1900	14.3	1.7	1.7	6.9	1.2	7.4	0.8	0.0	37.7	26.2	2.2
5	1903	10.3	1.5	0.4	9.3	0.6	8.5	0.9	0.0	38.4	27.2	3.0
6	1906	13.4	1.4	0.5	8.1	0.7	8.6	1.8	0.0	38.5	25.3	1.9
7	1909	13.5	1.1	0.5	9.0	0.6	9.0	3.4	0.0	36.8	23.5	2.6
8	1912	12.9	1.4	0.3	9.4	0.6	9.3	4.3	0.0	41.1	19.4	1.3
9	1920	12.3	0.3	0.1	11.9	2.4	3.7	1.7	1.9	42.4	23.1	0.2
10	1923	7.6	1.2	3.2	5.1	0.6	5.6	1.8	10.0	29.0	35.0	0.9
11	1926	10.5	0.3	0.4	4.5	1.8	6.6	2.1	10.1	19.9	41.6	2.3
12	1929	10.0	0.6	0.6	9.0	1.0	8.1	3.2	11.8	28.0	25.8	2.0
13	1932	10.6	0.8	0.3	8.9	3.0	10.0	6.4	13.4	27.4	19.2	0.0
14	1935	8.8	2.6	1.4	7.8	1.4	12.4	6.2	11.3	29.3	18.5	0.4
15	1938	10.1	1.1	1.2	5.9	1.4	9.5	6.0	5.9	40.7	18.2	0.0
16	1947	15.6	1.6	10.1	11.4	7.6	8.8	4.8	3.4	32.2	4.6	0.0
17	1950	11.2	1.3	16.5	12.4	15.8	8.1	4.9	3.4	20.7	4.2	1.5
18	1953	12.9	1.5	7.0	7.9	12.1	8.1	5.3	3.9	36.1	5.2	0.0
19	1956	10.9	5.3	9.7	7.6	9.6	9.4	8.5	4.6	28.2	6.2	0.0
20	1959	13.1	4.4	7.3	5.7	9.8	12.5	8.0	5.0	26.7	7.5	0.0
21	1962	12.8	4.7	7.5	6.6	6.8	15.7	9.7	5.3	24.5	6.4	0.1
22	1965	12.4	4.3	8.4	9.1	6.0	19.5	10.6	4.7	19.8	3.5	1.8
23	1968	11.4	6.0	9.5	5.9	5.0	21.1	10.7	4.2	20.0	4.4	1.9
24	1971	12.8	2.8	7.1	8.5	4.0	23.8	11.3	3.7	18.8	7.2	0.0

Principal Component Analysis

24 Observations
11 Variables

Simple Statistics

	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV
Mean	12.21250000	1.995833333	3.941666667	8.320833333	3.958333333	9.941666667	4.816666667	4.275000000	30.25833333	19.14166667	1.183333333
StD	2.19113983	1.645822785	4.489053043	2.467789018	4.181897164	5.223258615	3.408771365	4.154841553	7.30952100	12.19371138	1.025778837

Budgets de l'état de 1872 à 1971 : ACP simple

Covariance Matrix

	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV
PVP	4.8010937	-0.3049479	-0.0038542	1.2584896	0.3263542	-1.7171875	-0.9814583	-6.2534375	1.6184375	0.8965625	0.3356250
AGR	-0.3049479	2.7087326	4.4339236	-1.1203299	2.9989931	6.2868403	4.5200694	0.3028125	-5.3939236	-13.9460764	-0.4679861
CMI	-0.0038542	4.4339236	20.1515972	1.0299653	16.7271528	10.9511806	9.5055556	0.4197917	-17.5990972	-44.0209028	-1.6026389
TRA	1.2584896	-1.1203299	1.0299653	6.0899826	1.7142014	-2.7475347	-1.7082639	-3.2115625	2.8496181	-4.4637847	0.2895139
LOG	0.3263542	2.9989931	16.7271528	1.7142014	17.4882639	5.0754861	6.9531944	0.7768750	-11.5717361	-38.6545139	-1.8786111
EDU	-1.7171875	6.2868403	10.9511806	-2.7475347	5.0754861	27.2824306	15.5788889	3.4064583	-20.0082639	-42.6888194	-1.3322222
ACS	-0.9814583	4.5200694	9.5055556	-1.7082639	6.9531944	15.5788889	11.6197222	4.0816667	-14.1313889	-33.5931944	-1.8518056
ANC	-6.2534375	0.3028125	0.4197917	-3.2115625	0.7768750	3.4064583	4.0816667	17.2627083	-12.6597917	-2.5010417	-1.6087500
DEF	1.6184375	-5.3939236	-17.5990972	2.8496181	-11.5717361	-20.0082639	-14.1313889	-12.6597917	53.4290972	23.3196528	0.1530556
DET	0.8965625	-13.9460764	-44.0209028	-4.4637847	-38.6545139	-42.6888194	-33.5931944	-2.5010417	23.3196528	148.6865972	6.9286111
DIV	0.3356250	-0.4679861	-1.6026389	0.2895139	-1.8786111	-1.3322222	-1.8518056	-1.6087500	0.1530556	6.9286111	1.0522222

Total Variance = 310.57244792

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
AXE1	204.717	148.896	0.659161	0.65916
AXE2	55.822	33.853	0.179738	0.83890
AXE3	21.969	5.566	0.070736	0.90963
AXE4	16.402	10.880	0.052813	0.96245
AXE5	5.522	2.444	0.017780	0.98023
AXE6	3.078	1.966	0.009910	0.99014
AXE7	1.111	0.039	0.003578	0.99372
AXE8	1.072	0.440	0.003452	0.99717
AXE9	0.632	0.384	0.002034	0.99920
AXE10	0.248	0.248	0.000798	1.00000
AXE11	0.000	.	0.000000	1.00000

Eigenvectors

	AXE1	AXE2	AXE3	AXE4	AXE5	AXE6	AXE7	AXE8	AXE9	AXE10	AXE11
PVP	-.009986	-.069159	-.162955	0.356894	-.036514	-.722328	0.315600	0.296112	0.183891	0.087154	0.299001
AGR	0.086264	0.030954	0.074533	0.081767	0.197748	0.003959	-.692358	0.049476	0.384175	0.473551	0.294265
CMI	0.270902	0.052108	-.403435	0.020304	0.273428	0.428386	0.107430	0.585227	-.239286	-.008024	0.301817
TRA	0.011817	-.116355	-.205511	0.059113	-.872375	0.156839	-.135536	-.039858	-.157312	0.135954	0.302932
LOG	0.227541	-.034281	-.497724	-.179625	0.230674	-.060136	0.169328	-.691069	-.005996	0.110820	0.302758
EDU	0.272252	0.190359	0.578192	0.371798	0.001904	0.277975	0.408124	-.190950	-.010593	0.220458	0.296795
ACS	0.208579	0.108705	0.235933	0.042611	0.089847	-.323263	-.396196	-.081442	-.574535	-.425473	0.316078
ANC	0.038824	0.302146	0.234331	-.796670	-.125843	-.148635	0.173580	0.187864	0.148946	0.055611	0.298291
DEF	-.238664	-.843118	0.253182	-.175274	0.160258	0.084235	0.079818	0.027008	-.079659	0.012434	0.301647
DET	-.830484	0.355069	-.072886	0.102969	0.134578	0.103990	0.025134	-.076965	-.188117	0.073600	0.303350
DIV	-.036845	0.023729	-.029541	0.117105	-.053339	0.211487	-.042247	-.056324	0.585177	-.706422	0.299166

Budgets de l'état de 1872 à 1971 : ACP simple

Coordonnees factorielles des individus

OBS	AN	AXE1	AXE2	AXE3	AXE4	AXE5	AXE6	AXE7	AXE8	AXE9	AXE10	AXE11
1	1872	-22.6032	7.7833	-6.2842	5.84666	1.67038	-4.31829	-0.52764	0.89304	-0.61306	-0.59476	-0.007291
2	1880	-14.3507	0.6764	-5.9269	3.66778	-6.28825	0.00455	-1.64934	-0.88624	0.15063	0.53168	-0.004190
3	1890	-17.1203	0.1510	-0.8776	4.02146	2.11494	0.73798	0.99786	-0.20195	-0.29923	0.64376	0.009427
4	1900	-10.6692	-5.9665	0.1276	2.91792	2.11608	0.27690	0.88878	0.86426	1.41891	0.36573	-0.014059
5	1903	-11.8135	-6.0183	1.8347	2.18445	-0.14981	3.62660	-0.34517	-1.09320	0.69879	-0.09025	-0.003980
6	1906	-10.0079	-6.7626	1.9451	2.92932	0.71454	0.55068	0.58368	-0.02446	0.58230	0.25730	0.029304
7	1909	-7.7294	-5.8192	2.0601	3.42313	-0.56415	0.03605	0.00286	-0.12995	0.30462	-1.00546	-0.011683
8	1912	-5.0510	-10.7825	3.9904	2.07491	-0.59843	-0.09893	-0.40896	-0.05143	-0.55777	-0.50643	-0.004053
9	1920	-10.0906	-11.7210	-0.6369	-2.09099	-2.33287	-0.20867	-0.57461	-0.29382	-1.05409	0.50420	0.003139
10	1923	-15.4740	8.0060	-0.0216	-5.79686	2.79122	3.10823	-0.38938	1.52166	-0.86656	0.41344	-0.017713
11	1926	-19.0954	17.9626	-2.0596	-2.39987	1.91346	0.02618	0.81971	-1.13735	0.21267	-0.20414	0.018414
12	1929	-7.2359	5.9930	2.1734	-5.97156	-2.98898	0.01260	0.46385	0.61506	1.24181	-0.47715	-0.000819
13	1932	0.0929	5.1931	3.8534	-7.35286	-3.29270	-2.58843	0.49077	-0.46952	0.00319	-0.05374	-0.006690
14	1935	0.8304	3.5731	6.2281	-5.40484	-1.53990	0.32958	-0.51583	0.12599	0.27422	0.49304	0.002276
15	1938	-2.8864	-8.2856	6.3062	-4.04049	2.18281	-0.09341	0.04665	0.30450	-1.46065	-0.81299	0.007061
16	1947	13.7739	-7.7207	-4.7940	-0.87480	-1.84051	-1.32718	1.43230	3.24142	0.27079	0.37711	0.005757
17	1950	20.2552	1.9774	-14.2761	-1.85940	-0.93924	3.33528	0.87524	-0.27626	-0.57244	-0.58404	0.001365
18	1953	12.4391	-10.4486	-3.8572	-4.18382	2.16844	-1.56446	1.48211	-2.10152	0.37235	-0.10443	-0.006594
19	1956	15.0494	-2.1088	-3.4337	-2.55108	2.48663	-0.18341	-3.09383	0.14633	-0.42630	0.12063	0.007165
20	1959	14.3563	0.1833	-1.3065	-0.82710	3.12095	-2.14151	-0.29718	-1.33042	0.69343	0.66472	-0.016025
21	1962	16.4397	2.5731	1.8341	0.99374	1.38389	-1.49955	-0.73098	0.07645	0.05683	0.38882	0.016453
22	1965	21.1670	5.9850	2.6308	3.78147	-1.87055	0.13015	-0.38760	0.00701	0.49380	-0.64280	0.010346
23	1968	20.9944	6.8884	4.4457	4.64318	1.58662	1.50702	-0.93670	0.78027	0.93359	-0.19915	-0.007952
24	1971	18.7293	8.6882	6.0444	6.86967	-1.84455	0.34204	1.77342	-0.57986	-1.85782	0.51493	-0.009657

Budgets de l'état de 1872 à 1971 : ACP simple

Contributions relatives des individus a l'inertie des axes

OBS	AN	CONT_1	CONT_2	CONT_3	CONT_4	CONT_5	CONT_6	CONT_7	CONT_8	CONT_9	CONT_10	CONT_11	NORME2
1	1872	0.10399	0.04522	0.07490	0.08684	0.02105	0.25246	0.01044	0.03099	0.02479	0.05948	0.01732	668.40
2	1880	0.04192	0.00034	0.06662	0.03417	0.29837	0.00000	0.10200	0.03053	0.00150	0.04753	0.00572	298.33
3	1890	0.05966	0.00002	0.00146	0.04108	0.03375	0.00737	0.03733	0.00159	0.00590	0.06968	0.02896	316.63
4	1900	0.02317	0.02657	0.00003	0.02163	0.03379	0.00104	0.02962	0.02903	0.13278	0.02249	0.06441	166.20
5	1903	0.02840	0.02704	0.00638	0.01212	0.00017	0.17806	0.00447	0.04645	0.03220	0.00137	0.00516	198.90
6	1906	0.02039	0.03414	0.00718	0.02180	0.00385	0.00411	0.01277	0.00002	0.02236	0.01113	0.27984	159.82
7	1909	0.01216	0.02528	0.00805	0.02977	0.00240	0.00002	0.00000	0.00066	0.00612	0.16998	0.04448	111.01
8	1912	0.00519	0.08678	0.03020	0.01094	0.00270	0.00013	0.00627	0.00010	0.02052	0.04312	0.00535	163.11
9	1920	0.02072	0.10255	0.00077	0.01111	0.04107	0.00059	0.01238	0.00336	0.07328	0.04274	0.00321	251.25
10	1923	0.04874	0.04784	0.00000	0.08536	0.05879	0.13079	0.00568	0.08999	0.04952	0.02874	0.10224	357.99
11	1926	0.07422	0.24084	0.00805	0.01463	0.02763	0.00001	0.02519	0.05027	0.00298	0.00701	0.11049	703.01
12	1929	0.01066	0.02681	0.00896	0.09059	0.06741	0.00000	0.00807	0.01470	0.10170	0.03828	0.00022	139.95
13	1932	0.00000	0.02013	0.02816	0.13734	0.08181	0.09071	0.00903	0.00857	0.00000	0.00049	0.01458	113.90
14	1935	0.00014	0.00953	0.07357	0.07421	0.01789	0.00147	0.00998	0.00062	0.00496	0.04087	0.00169	84.54
15	1938	0.00170	0.05124	0.07543	0.04147	0.03595	0.00012	0.00008	0.00360	0.14070	0.11113	0.01625	140.74
16	1947	0.03861	0.04449	0.04359	0.00194	0.02556	0.02385	0.07692	0.40834	0.00484	0.02391	0.01080	291.00
17	1950	0.08350	0.00292	0.38655	0.00878	0.00666	0.15060	0.02872	0.00297	0.02161	0.05735	0.00061	634.97
18	1953	0.03149	0.08149	0.02822	0.04447	0.03548	0.03314	0.08236	0.17164	0.00914	0.00183	0.01417	310.20
19	1956	0.04610	0.00332	0.02236	0.01653	0.04666	0.00046	0.35889	0.00083	0.01198	0.00245	0.01673	265.24
20	1959	0.04195	0.00003	0.00324	0.00174	0.07350	0.06209	0.00331	0.06879	0.03171	0.07429	0.08369	225.63
21	1962	0.05501	0.00494	0.00638	0.00251	0.01445	0.03044	0.02003	0.00023	0.00021	0.02542	0.08821	286.09
22	1965	0.09119	0.02674	0.01313	0.03633	0.02640	0.00023	0.00563	0.00000	0.01608	0.06947	0.03488	509.40
23	1968	0.08971	0.03542	0.03749	0.05477	0.01899	0.03075	0.03290	0.02366	0.05748	0.00667	0.02061	536.73
24	1971	0.07140	0.05634	0.06929	0.11988	0.02567	0.00158	0.11792	0.01307	0.22763	0.04458	0.03039	520.72
		1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	7453.74

Qualite de la representation des individus (cosinus carres)

OBS	AN	COS2_1	COS2_2	COS2_12	COS2_3	COS2_1_3
1	1872	0.76437	0.09063	0.85500	0.05908	0.91408
2	1880	0.69031	0.00153	0.69184	0.11775	0.80959
3	1890	0.92571	0.00007	0.92578	0.00243	0.92821
4	1900	0.68491	0.21419	0.89910	0.00010	0.89920
5	1903	0.70164	0.18210	0.88374	0.01692	0.90067
6	1906	0.62671	0.28616	0.91287	0.02367	0.93654
7	1909	0.53819	0.30505	0.84324	0.03823	0.88147
8	1912	0.15641	0.71279	0.86920	0.09763	0.96683
9	1920	0.40526	0.54680	0.95206	0.00161	0.95367
10	1923	0.66887	0.17904	0.84791	0.00000	0.84791
11	1926	0.51868	0.45896	0.97764	0.00603	0.98368
12	1929	0.37411	0.25663	0.63073	0.03375	0.66448
13	1932	0.00008	0.23678	0.23685	0.13037	0.36722
14	1935	0.00816	0.15102	0.15917	0.45884	0.61801
15	1938	0.05920	0.48779	0.54698	0.28257	0.82955
16	1947	0.65196	0.20484	0.85680	0.07898	0.93578
17	1950	0.64614	0.00616	0.65229	0.32097	0.97327
18	1953	0.49881	0.35194	0.85076	0.04796	0.89872
19	1956	0.85390	0.01677	0.87066	0.04445	0.91512
20	1959	0.91343	0.00015	0.91358	0.00757	0.92115
21	1962	0.94467	0.02314	0.96781	0.01176	0.97957
22	1965	0.87954	0.07032	0.94986	0.01359	0.96344
23	1968	0.82121	0.08841	0.90962	0.03682	0.94644
24	1971	0.67366	0.14496	0.81863	0.07016	0.88879

Budgets de l'état de 1872 à 1971 : ACP simple

Correlations variables - axes

Correlation Analysis

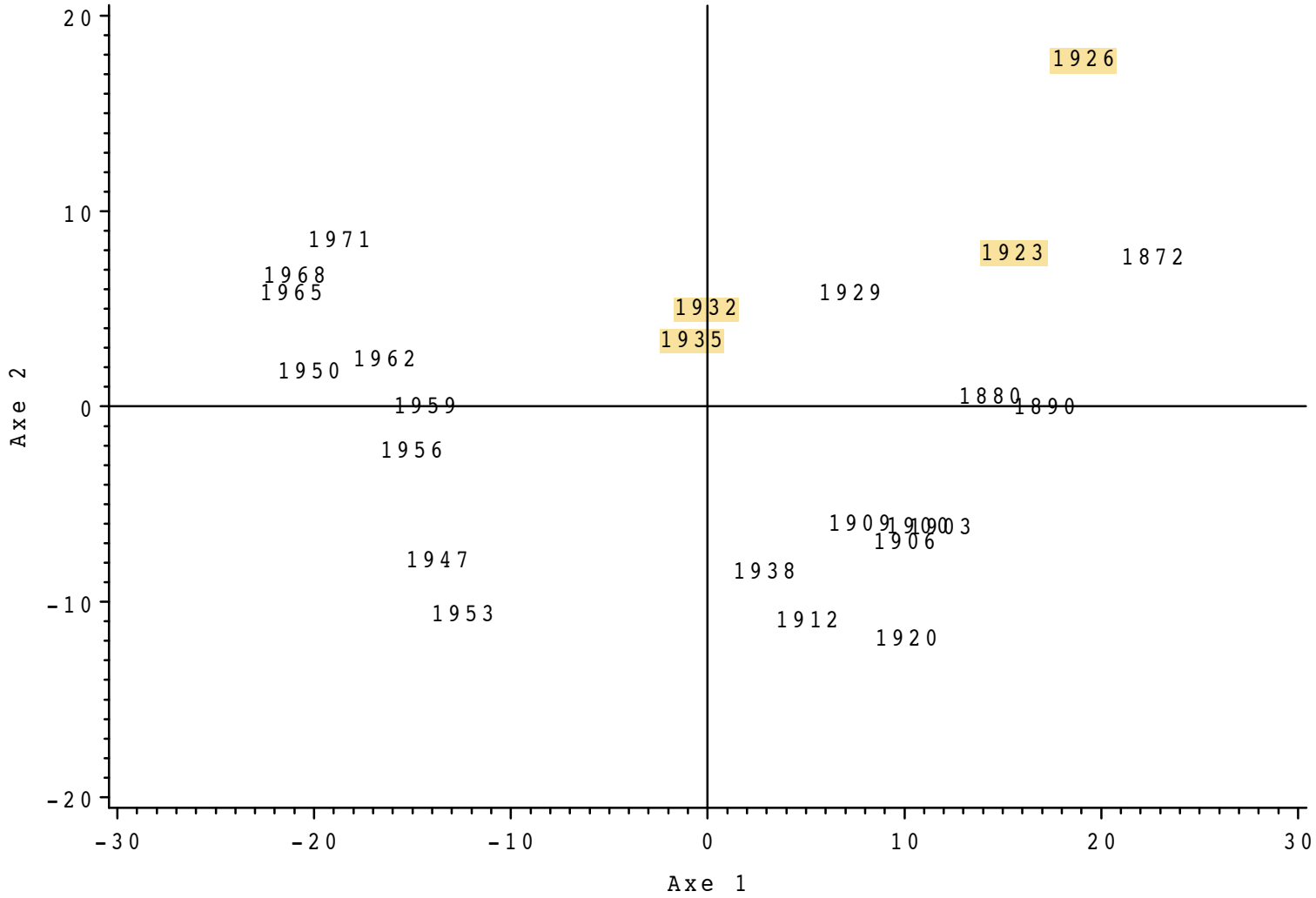
11 'WITH' Variables:	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV
10 'VAR' Variables:	AXE1	AXE2	AXE3	AXE4	AXE5	AXE6	AXE7	AXE8	AXE9	AXE10	

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 24

	AXE1	AXE2	AXE3	AXE4	AXE5	AXE6	AXE7	AXE8	AXE9	AXE10
PVP	-0.06521	-0.23582	-0.34858	0.65966	-0.03916	-0.57833	0.15184	0.13993	0.06671	0.01980
	0.7621	0.2673	0.0951	0.0005	0.8558	0.0031	0.4788	0.5143	0.7568	0.9268
AGR	0.74994	0.14052	0.21226	0.20121	0.28234	0.00422	-0.44346	0.03113	0.18554	0.14324
	0.0001	0.5125	0.3194	0.3458	0.1813	0.9844	0.0300	0.8852	0.3854	0.5043
CMI	0.86345	0.08673	-0.42123	0.01832	0.14313	0.16741	0.02523	0.13499	-0.04237	-0.00089
	0.0001	0.6870	0.0404	0.9323	0.5046	0.4343	0.9069	0.5294	0.8442	0.9967
TRA	0.06851	-0.35227	-0.39033	0.09701	-0.83070	0.11150	-0.05790	-0.01672	-0.05067	0.02743
	0.7504	0.0914	0.0593	0.6520	0.0001	0.6040	0.7881	0.9382	0.8141	0.8988
LOG	0.77851	-0.06125	-0.55785	-0.17396	0.12962	-0.02523	0.04268	-0.17111	-0.00114	0.01319
	0.0001	0.7762	0.0046	0.4162	0.5461	0.9069	0.8430	0.4240	0.9958	0.9512
EDU	0.74577	0.27229	0.51884	0.28828	0.00086	0.09336	0.08237	-0.03785	-0.00161	0.02101
	0.0001	0.1980	0.0094	0.1719	0.9968	0.6643	0.7020	0.8606	0.9940	0.9224
ACS	0.87549	0.23826	0.32441	0.05063	0.06194	-0.16637	-0.12252	-0.02474	-0.13397	-0.06214
	0.0001	0.2622	0.1220	0.8143	0.7737	0.4372	0.5684	0.9087	0.5326	0.7730
ANC	0.13370	0.54333	0.26435	-0.77656	-0.07117	-0.06276	0.04404	0.04682	0.02849	0.00666
	0.5334	0.0061	0.2119	0.0001	0.7410	0.7708	0.8381	0.8280	0.8949	0.9753
DEF	-0.46717	-0.86179	0.16235	-0.09711	0.05152	0.02022	0.01151	0.00383	-0.00866	0.00085
	0.0214	0.0001	0.4485	0.6517	0.8110	0.9253	0.9574	0.9858	0.9680	0.9969
DET	-0.97448	0.21756	-0.02802	0.03420	0.02593	0.01496	0.00217	-0.00654	-0.01226	0.00300
	0.0001	0.3072	0.8966	0.8739	0.9043	0.9447	0.9920	0.9758	0.9546	0.9889
DIV	-0.51392	0.17283	-0.13498	0.46235	-0.12219	0.36170	-0.04342	-0.05685	0.45344	-0.34283
	0.0102	0.4193	0.5294	0.0229	0.5695	0.0824	0.8404	0.7919	0.0261	0.1010

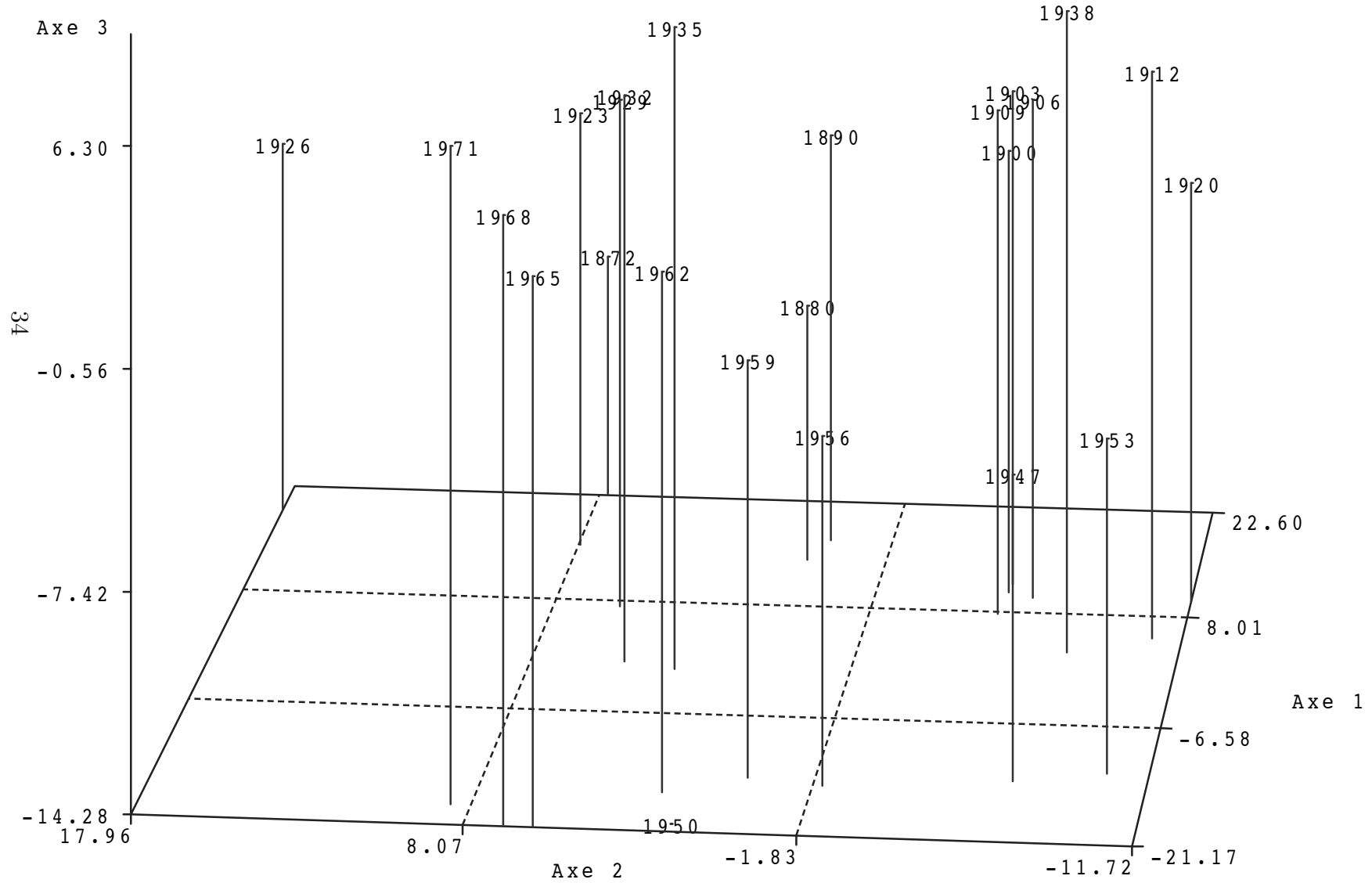
Budgets de l'etat de 1872 a 1971 : ACP simple

Projection dans le plan 1-2



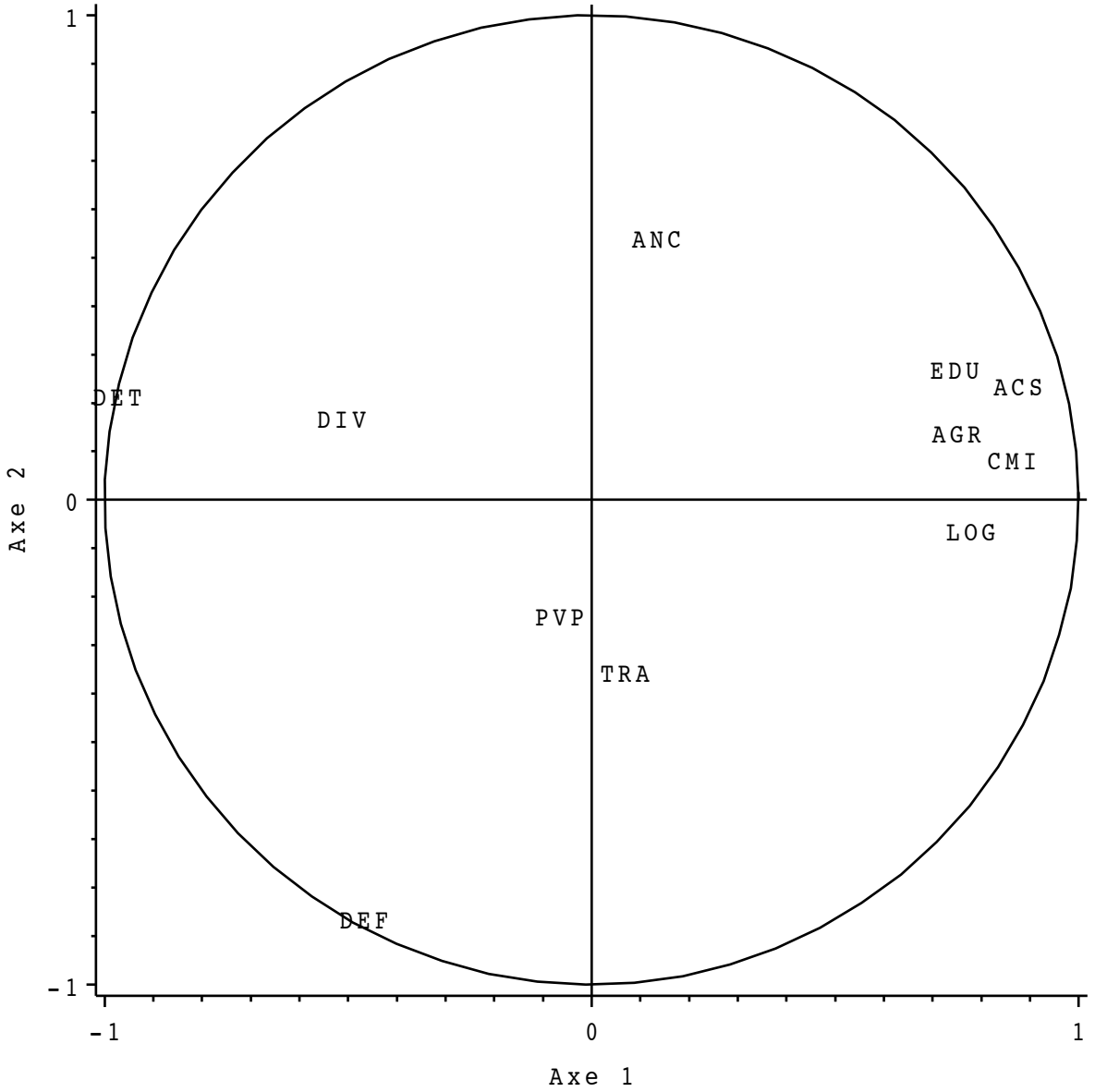
Budgets de l'etat de 1872 a 1971 : ACP simple

Projection dans l'espace 1-2-3



Budgets de l'etat de 1872 a 1971 : ACP simple

Cercle des correlations dans le plan 1-2



Budgets de l'état de 1872 à 1971 : ACP normée

Donnees brutes

OBS	AN	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV
1	1872	18.0	0.5	0.1	6.7	0.5	2.1	2.0	0.0	26.4	41.5	2.1
2	1880	14.1	0.8	0.1	15.3	1.9	3.7	0.5	0.0	29.8	31.3	2.5
3	1890	13.6	0.7	0.7	6.8	0.6	7.1	0.7	0.0	33.8	34.4	1.7
4	1900	14.3	1.7	1.7	6.9	1.2	7.4	0.8	0.0	37.7	26.2	2.2
5	1903	10.3	1.5	0.4	9.3	0.6	8.5	0.9	0.0	38.4	27.2	3.0
6	1906	13.4	1.4	0.5	8.1	0.7	8.6	1.8	0.0	38.5	25.3	1.9
7	1909	13.5	1.1	0.5	9.0	0.6	9.0	3.4	0.0	36.8	23.5	2.6
8	1912	12.9	1.4	0.3	9.4	0.6	9.3	4.3	0.0	41.1	19.4	1.3
9	1920	12.3	0.3	0.1	11.9	2.4	3.7	1.7	1.9	42.4	23.1	0.2
10	1923	7.6	1.2	3.2	5.1	0.6	5.6	1.8	10.0	29.0	35.0	0.9
11	1926	10.5	0.3	0.4	4.5	1.8	6.6	2.1	10.1	19.9	41.6	2.3
12	1929	10.0	0.6	0.6	9.0	1.0	8.1	3.2	11.8	28.0	25.8	2.0
13	1932	10.6	0.8	0.3	8.9	3.0	10.0	6.4	13.4	27.4	19.2	0.0
14	1935	8.8	2.6	1.4	7.8	1.4	12.4	6.2	11.3	29.3	18.5	0.4
15	1938	10.1	1.1	1.2	5.9	1.4	9.5	6.0	5.9	40.7	18.2	0.0
16	1947	15.6	1.6	10.1	11.4	7.6	8.8	4.8	3.4	32.2	4.6	0.0
17	1950	11.2	1.3	16.5	12.4	15.8	8.1	4.9	3.4	20.7	4.2	1.5
18	1953	12.9	1.5	7.0	7.9	12.1	8.1	5.3	3.9	36.1	5.2	0.0
19	1956	10.9	5.3	9.7	7.6	9.6	9.4	8.5	4.6	28.2	6.2	0.0
20	1959	13.1	4.4	7.3	5.7	9.8	12.5	8.0	5.0	26.7	7.5	0.0
21	1962	12.8	4.7	7.5	6.6	6.8	15.7	9.7	5.3	24.5	6.4	0.1
22	1965	12.4	4.3	8.4	9.1	6.0	19.5	10.6	4.7	19.8	3.5	1.8
23	1968	11.4	6.0	9.5	5.9	5.0	21.1	10.7	4.2	20.0	4.4	1.9
24	1971	12.8	2.8	7.1	8.5	4.0	23.8	11.3	3.7	18.8	7.2	0.0

Principal Component Analysis

24 Observations
11 Variables

Simple Statistics

	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV
Mean	12.21250000	1.995833333	3.941666667	8.320833333	3.958333333	9.941666667	4.816666667	4.275000000	30.25833333	19.14166667	1.183333333
StD	2.19113983	1.645822785	4.489053043	2.467789018	4.181897164	5.223258615	3.408771365	4.154841553	7.30952100	12.19371138	1.025778837

Budgets de l'état de 1872 à 1971 : ACP normée

Correlation Matrix

	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV
PVP	1.0000	-.0846	-.0004	0.2327	0.0356	-.1500	-.1314	-.6869	0.1011	0.0336	0.1493
AGR	-.0846	1.0000	0.6001	-.2758	0.4357	0.7313	0.8057	0.0443	-.4484	-.6949	-.2772
CMI	-.0004	0.6001	1.0000	0.0930	0.8910	0.4671	0.6212	0.0225	-.5363	-.8042	-.3480
TRA	0.2327	-.2758	0.0930	1.0000	0.1661	-.2132	-.2031	-.3132	0.1580	-.1483	0.1144
LOG	0.0356	0.4357	0.8910	0.1661	1.0000	0.2324	0.4878	0.0447	-.3786	-.7580	-.4379
EDU	-.1500	0.7313	0.4671	-.2132	0.2324	1.0000	0.8750	0.1570	-.5241	-.6702	-.2486
ACS	-.1314	0.8057	0.6212	-.2031	0.4878	0.8750	1.0000	0.2882	-.5672	-.8082	-.5296
ANC	-.6869	0.0443	0.0225	-.3132	0.0447	0.1570	0.2882	1.0000	-.4169	-.0494	-.3775
DEF	0.1011	-.4484	-.5363	0.1580	-.3786	-.5241	-.5672	-.4169	1.0000	0.2616	0.0204
DET	0.0336	-.6949	-.8042	-.1483	-.7580	-.6702	-.8082	-.0494	0.2616	1.0000	0.5539
DIV	0.1493	-.2772	-.3480	0.1144	-.4379	-.2486	-.5296	-.3775	0.0204	0.5539	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
AXE1	4.97236	2.92172	0.452033	0.45203
AXE2	2.05064	0.76047	0.186422	0.63845
AXE3	1.29017	0.29711	0.117288	0.75574
AXE4	0.99306	0.28470	0.090278	0.84602
AXE5	0.70835	0.15020	0.064396	0.91042
AXE6	0.55815	0.35390	0.050741	0.96116
AXE7	0.20425	0.07905	0.018568	0.97973
AXE8	0.12520	0.06239	0.011382	0.99111
AXE9	0.06281	0.02781	0.005710	0.99682
AXE10	0.03500	0.03500	0.003182	1.00000
AXE11	0.00000	.	0.000000	1.00000

Eigenvectors

	AXE1	AXE2	AXE3	AXE4	AXE5	AXE6	AXE7	AXE8	AXE9	AXE10	AXE11
PVP	-.077718	0.516585	0.300916	-.107402	-.404401	0.537823	0.061979	0.332364	0.194031	-.071792	0.122925
AGR	0.367027	0.003811	0.322831	-.154056	0.038983	-.293732	0.769430	-.012856	0.090144	-.202374	0.090871
CMI	0.373592	0.238303	-.124641	0.258277	-.179742	-.260294	-.202583	-.245291	0.620424	0.265645	0.254240
TRA	-.061420	0.440474	-.330723	0.282102	0.649712	0.291728	0.275157	-.099139	-.000165	-.028325	0.140279
LOG	0.323623	0.277749	-.338881	0.208289	-.311932	-.236101	-.080438	0.213674	-.561863	-.287386	0.237586
EDU	0.352832	-.095637	0.374006	-.115808	0.377314	0.151641	-.474486	-.069307	0.025456	-.482172	0.290917
ACS	0.418512	-.070420	0.146475	-.150915	0.123794	0.232519	0.002425	0.090739	-.357080	0.727772	0.202165
ANC	0.129569	-.563765	-.330197	0.203075	-.014167	0.261901	0.144533	0.527343	0.292046	-.089294	0.232540
DEF	-.274545	0.151074	-.228493	-.639441	0.183438	-.348330	-.082867	0.295056	0.122651	0.074713	0.413733
DET	-.398625	-.210436	0.141175	0.179893	-.209127	0.080306	0.141026	-.409903	-.156697	0.008613	0.694120
DIV	-.245903	0.078312	0.472199	0.506346	0.218932	-.377401	-.090149	0.476981	-.023836	0.155432	0.057559

Budgets de l'état de 1872 à 1971 : ACP normée

Coordonnees factorielles des individus

OBS	AN	AXE1	AXE2	AXE3	AXE4	AXE5	AXE6	AXE7	AXE8	AXE9	AXE10	AXE11
1	1872	-2.90044	1.02429	1.56420	0.48725	-2.05734	1.21852	0.30696	-0.00673	-0.05264	0.35289	-.0013768
2	1880	-2.76714	2.01204	-0.16868	1.48363	1.23605	0.71308	0.93078	-0.26963	-0.22648	-0.13092	-.0007923
3	1890	-2.41640	0.22399	0.76550	-0.26772	-0.70961	-0.12154	-0.32784	-0.45696	0.02342	-0.22915	0.0017976
4	1900	-2.05679	0.75486	1.00709	-0.52320	-0.50306	-0.61651	-0.09622	0.27766	0.34795	-0.25476	-.0026481
5	1903	-2.33805	0.16694	0.62263	0.18086	1.21338	-1.45201	-0.13159	-0.02886	-0.14124	-0.11346	-.0007492
6	1906	-1.98507	0.62579	0.69303	-0.71000	0.15170	-0.37484	-0.16577	0.06905	0.08612	-0.17400	0.0054998
7	1909	-1.90731	0.81166	0.98733	-0.20093	0.59516	-0.25234	-0.29918	0.39996	-0.09540	0.24825	-.0022141
8	1912	-1.43080	0.76797	0.19556	-1.29024	0.78145	-0.07780	-0.13081	0.02835	-0.09753	0.22312	-.0007630
9	1920	-2.13873	0.95660	-1.74582	-1.06392	0.62703	0.32372	0.28880	-0.38875	-0.10212	-0.01595	0.0006095
10	1923	-1.14368	-2.88292	-0.86815	0.43627	-0.54206	-0.84056	0.05041	-0.70681	0.37582	0.06195	-.0032948
11	1926	-1.67455	-2.61050	0.49638	1.76239	-1.19394	0.07340	-0.28263	0.04773	-0.25440	-0.07030	0.0034623
12	1929	-1.17359	-1.83068	-0.61190	1.15625	0.69458	0.30872	0.02705	0.68071	0.21008	0.01038	-.0001847
13	1932	0.27089	-1.95889	-1.46254	0.04132	0.34437	1.40821	0.09301	0.42346	-0.12827	-0.02835	-.0012627
14	1935	0.65896	-2.29628	-0.66335	-0.30717	0.82324	0.22300	0.38609	0.02122	0.09462	-0.14261	0.0004272
15	1938	-0.40235	-1.34280	-0.84992	-1.84912	0.06308	-0.23950	-0.50333	-0.05058	-0.06825	0.40597	0.0013370
16	1947	1.08894	2.25683	-1.27788	-0.22653	-0.37665	0.86107	-0.10681	0.16341	0.82156	-0.03259	0.0010713
17	1950	2.37084	2.17613	-1.91767	2.65983	-0.18484	-0.90121	-0.64898	-0.21389	-0.03832	0.14763	0.0002623
18	1953	1.20336	1.13454	-1.66117	-0.75364	-0.96024	-0.40847	-0.52278	0.51788	-0.41873	-0.18344	-.0012353
19	1956	2.92698	0.23013	-0.58915	-0.44521	-0.50939	-0.94524	1.09962	-0.27412	-0.10371	0.21796	0.0013386
20	1959	2.68561	0.13932	0.07215	-0.69076	-1.21132	-0.18021	0.39770	0.17638	-0.26291	-0.29334	-.0030065
21	1962	3.05409	-0.11197	0.58714	-0.64513	-0.41926	0.24536	0.40288	-0.03630	-0.00890	-0.05974	0.0030861
22	1965	3.14213	0.30960	1.41291	0.76374	0.93077	0.22039	-0.03700	0.31130	-0.03356	0.14386	0.0019154
23	1968	3.69437	-0.46870	2.29718	0.28211	0.47550	-0.73271	0.17895	0.10598	0.20925	0.04036	-.0015139
24	1971	3.23873	-0.08796	1.11514	-0.28011	0.73141	1.54747	-0.90929	-0.79044	-0.13637	-0.12376	-.0017657

Budgets de l'état de 1872 à 1971 : ACP normée

Contributions relatives des individus a l'inertie des axes

OBS	AN	CONT_1	CONT_2	CONT_3	CONT_4	CONT_5	CONT_6	CONT_7	CONT_8	CONT_9	CONT_10	CONT_11	NORME2
1	1872	0.07049	0.02132	0.07902	0.00996	0.24897	0.11084	0.01922	0.00002	0.00184	0.14825	0.01754	18.085
2	1880	0.06416	0.08226	0.00092	0.09236	0.08987	0.03796	0.17673	0.02420	0.03402	0.02041	0.00581	16.979
3	1890	0.04893	0.00102	0.01892	0.00301	0.02962	0.00110	0.02192	0.06949	0.00036	0.06251	0.02991	7.434
4	1900	0.03545	0.01158	0.03276	0.01149	0.01489	0.02837	0.00189	0.02566	0.08031	0.07726	0.06490	6.994
5	1903	0.04581	0.00057	0.01252	0.00137	0.08660	0.15739	0.00353	0.00028	0.01323	0.01533	0.00519	9.546
6	1906	0.03302	0.00796	0.01551	0.02115	0.00135	0.01049	0.00561	0.00159	0.00492	0.03604	0.27994	5.550
7	1909	0.03048	0.01339	0.03148	0.00169	0.02084	0.00475	0.01826	0.05324	0.00604	0.07336	0.04537	6.050
8	1912	0.01715	0.01198	0.00124	0.06985	0.03592	0.00045	0.00349	0.00027	0.00631	0.05926	0.00539	5.034
9	1920	0.03833	0.01859	0.09843	0.04749	0.02313	0.00782	0.01701	0.05030	0.00692	0.00030	0.00344	10.412
10	1923	0.01096	0.16887	0.02434	0.00799	0.01728	0.05274	0.00052	0.16626	0.09369	0.00457	0.10047	12.211
11	1926	0.02350	0.13847	0.00796	0.13032	0.08385	0.00040	0.01630	0.00076	0.04293	0.00588	0.11095	14.554
12	1929	0.01154	0.06810	0.01209	0.05609	0.02838	0.00712	0.00015	0.15420	0.02928	0.00013	0.00032	7.526
13	1932	0.00061	0.07797	0.06908	0.00007	0.00698	0.14804	0.00176	0.05968	0.01091	0.00096	0.01476	8.358
14	1935	0.00364	0.10714	0.01421	0.00396	0.03986	0.00371	0.03041	0.00015	0.00594	0.02421	0.00169	7.148
15	1938	0.00136	0.03664	0.02333	0.14347	0.00023	0.00428	0.05168	0.00085	0.00309	0.19620	0.01654	6.593
16	1947	0.00994	0.10349	0.05274	0.00215	0.00834	0.05535	0.00233	0.00889	0.44775	0.00126	0.01062	9.561
17	1950	0.04710	0.09622	0.11877	0.29684	0.00201	0.06063	0.08592	0.01522	0.00097	0.02594	0.00064	22.445
18	1953	0.01213	0.02615	0.08912	0.02383	0.05424	0.01246	0.05575	0.08926	0.11631	0.04006	0.01412	7.902
19	1956	0.07179	0.00108	0.01121	0.00832	0.01526	0.06670	0.24666	0.02501	0.00714	0.05656	0.01658	11.661
20	1959	0.06044	0.00039	0.00017	0.02002	0.08631	0.00242	0.03226	0.01035	0.04585	0.10244	0.08366	9.558
21	1962	0.07816	0.00025	0.01113	0.01746	0.01034	0.00449	0.03311	0.00044	0.00005	0.00425	0.08814	10.504
22	1965	0.08273	0.00195	0.06447	0.02447	0.05096	0.00363	0.00028	0.03225	0.00075	0.02464	0.03395	13.583
23	1968	0.11437	0.00446	0.17042	0.00334	0.01330	0.04008	0.00653	0.00374	0.02904	0.00194	0.02121	20.076
24	1971	0.08790	0.00016	0.04016	0.00329	0.03147	0.17876	0.16867	0.20793	0.01234	0.01823	0.02886	16.234
		=======	=======	=======	=======	=======	=======	=======	=======	=======	=======	=======	=======
		1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	264.000

Qualite de la representation des individus (cosinus carres)

OBS	AN	COS2_1	COS2_2	COS2_12	COS2_3	COS2_1_3
1	1872	0.46517	0.05801	0.52319	0.13529	0.65848
2	1880	0.45098	0.23843	0.68941	0.00168	0.69109
3	1890	0.78539	0.00675	0.79214	0.07882	0.87096
4	1900	0.60489	0.08148	0.68637	0.14502	0.83139
5	1903	0.57263	0.00292	0.57555	0.04061	0.61615
6	1906	0.71000	0.07056	0.78056	0.08654	0.86710
7	1909	0.60130	0.10889	0.71019	0.16113	0.87133
8	1912	0.40668	0.11716	0.52385	0.00760	0.53145
9	1920	0.43931	0.08789	0.52719	0.29272	0.81991
10	1923	0.10712	0.68064	0.78776	0.06172	0.84949
11	1926	0.19267	0.46824	0.66091	0.01693	0.67784
12	1929	0.18300	0.44530	0.62830	0.04975	0.67805
13	1932	0.00878	0.45910	0.46788	0.25592	0.72379
14	1935	0.06075	0.73770	0.79845	0.06156	0.86001
15	1938	0.02455	0.27348	0.29803	0.10956	0.40759
16	1947	0.12403	0.53272	0.65675	0.17080	0.82755
17	1950	0.25043	0.21098	0.46141	0.16384	0.62525
18	1953	0.18325	0.16289	0.34614	0.34921	0.69535
19	1956	0.73469	0.00454	0.73923	0.02977	0.76900
20	1959	0.75456	0.00203	0.75659	0.00054	0.75714
21	1962	0.88797	0.00119	0.88917	0.03282	0.92199
22	1965	0.72684	0.00706	0.73390	0.14697	0.88086
23	1968	0.67982	0.01094	0.69077	0.26285	0.95362
24	1971	0.64613	0.00048	0.64660	0.07660	0.72320

Budgets de l'état de 1872 à 1971 : ACP normée

Correlations variables - axes

Correlation Analysis

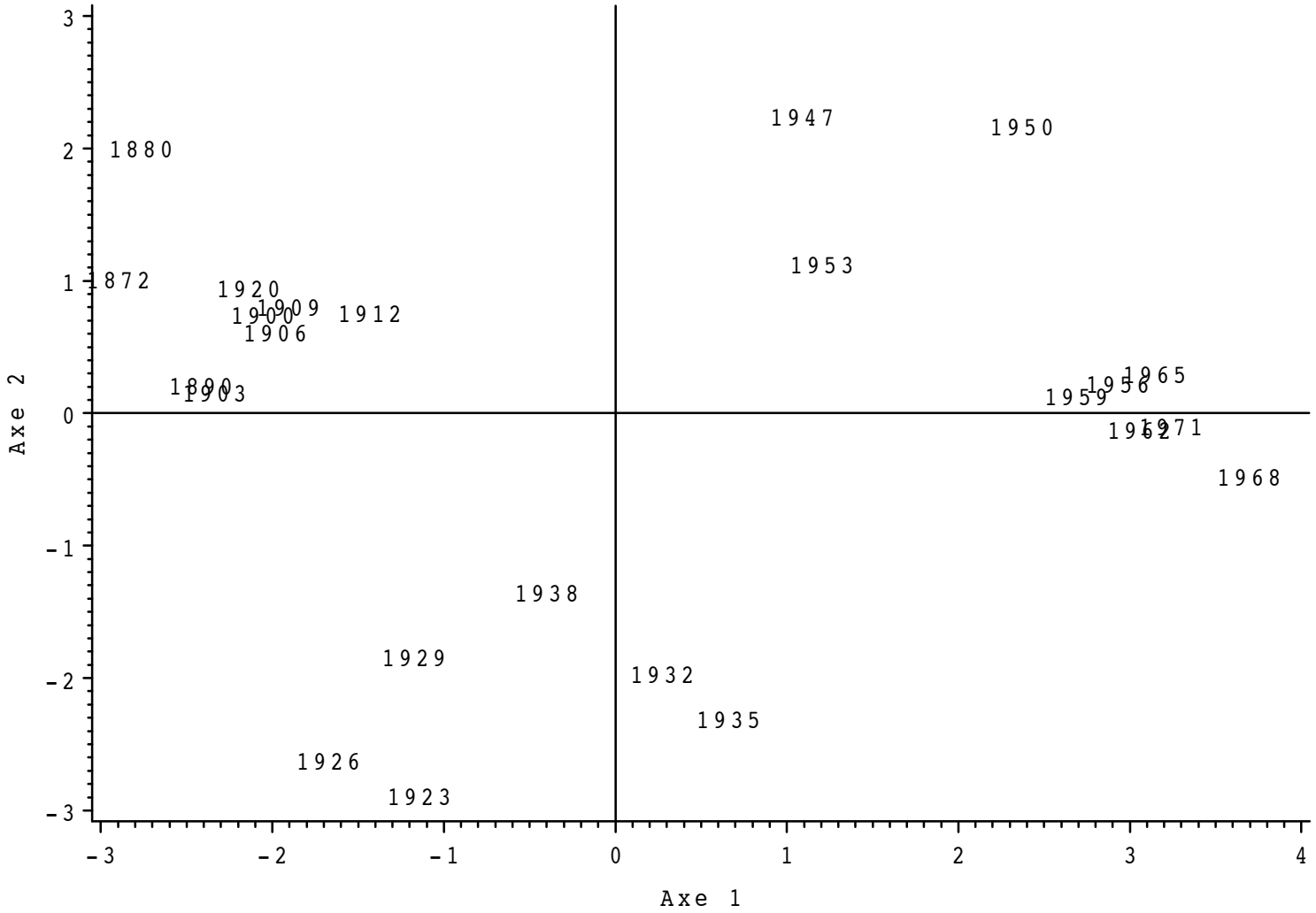
11 'WITH' Variables:	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV
10 'VAR' Variables:	AXE1	AXE2	AXE3	AXE4	AXE5	AXE6	AXE7	AXE8	AXE9	AXE10	

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 24

	AXE1	AXE2	AXE3	AXE4	AXE5	AXE6	AXE7	AXE8	AXE9	AXE10
PVP	-0.17330	0.73975	0.34180	-0.10703	-0.34036	0.40180	0.02801	0.11760	0.04863	-0.01343
	0.4180	0.0001	0.1021	0.6186	0.1036	0.0516	0.8966	0.5842	0.8215	0.9503
AGR	0.81843	0.00546	0.36669	-0.15352	0.03281	-0.21945	0.34774	-0.00455	0.02259	-0.03786
	0.0001	0.9798	0.0780	0.4739	0.8790	0.3029	0.0959	0.9832	0.9165	0.8606
CMI	0.83306	0.34125	-0.14157	0.25738	-0.15128	-0.19446	-0.09156	-0.08679	0.15549	0.04970
	0.0001	0.1027	0.5093	0.2247	0.4804	0.3625	0.6705	0.6868	0.4681	0.8176
TRA	-0.13696	0.63076	-0.37565	0.28112	0.54682	0.21795	0.12436	-0.03508	-0.00004	-0.00530
	0.5234	0.0010	0.0705	0.1833	0.0057	0.3063	0.5626	0.8707	0.9998	0.9804
LOG	0.72164	0.39774	-0.38492	0.20756	-0.26253	-0.17639	-0.03635	0.07561	-0.14082	-0.05377
	0.0001	0.0543	0.0633	0.3304	0.2152	0.4097	0.8661	0.7255	0.5116	0.8030
EDU	0.78677	-0.13695	0.42482	-0.11541	0.31756	0.11329	-0.21444	-0.02452	0.00638	-0.09021
	0.0001	0.5234	0.0385	0.5913	0.1305	0.5981	0.3143	0.9094	0.9764	0.6751
ACS	0.93323	-0.10084	0.16637	-0.15039	0.10419	0.17371	0.00110	0.03211	-0.08949	0.13616
	0.0001	0.6392	0.4372	0.4830	0.6280	0.4169	0.9959	0.8816	0.6775	0.5258
ANC	0.28892	-0.80731	-0.37506	0.20237	-0.01192	0.19566	0.06532	0.18659	0.07319	-0.01671
	0.1709	0.0001	0.0709	0.3430	0.9559	0.3595	0.7617	0.3826	0.7339	0.9382
DEF	-0.61220	0.21634	-0.25954	-0.63722	0.15439	-0.26024	-0.03745	0.10440	0.03074	0.01398
	0.0015	0.3099	0.2207	0.0008	0.4713	0.2194	0.8621	0.6273	0.8866	0.9483
DET	-0.88888	-0.30134	0.16035	0.17927	-0.17601	0.06000	0.06374	-0.14504	-0.03927	0.00161
	0.0001	0.1524	0.4542	0.4019	0.4107	0.7806	0.7673	0.4989	0.8554	0.9940
DIV	-0.54833	0.11214	0.53635	0.50459	0.18426	-0.28195	-0.04074	0.16877	-0.00597	0.02908
	0.0055	0.6019	0.0069	0.0119	0.3887	0.1819	0.8501	0.4305	0.9779	0.8927

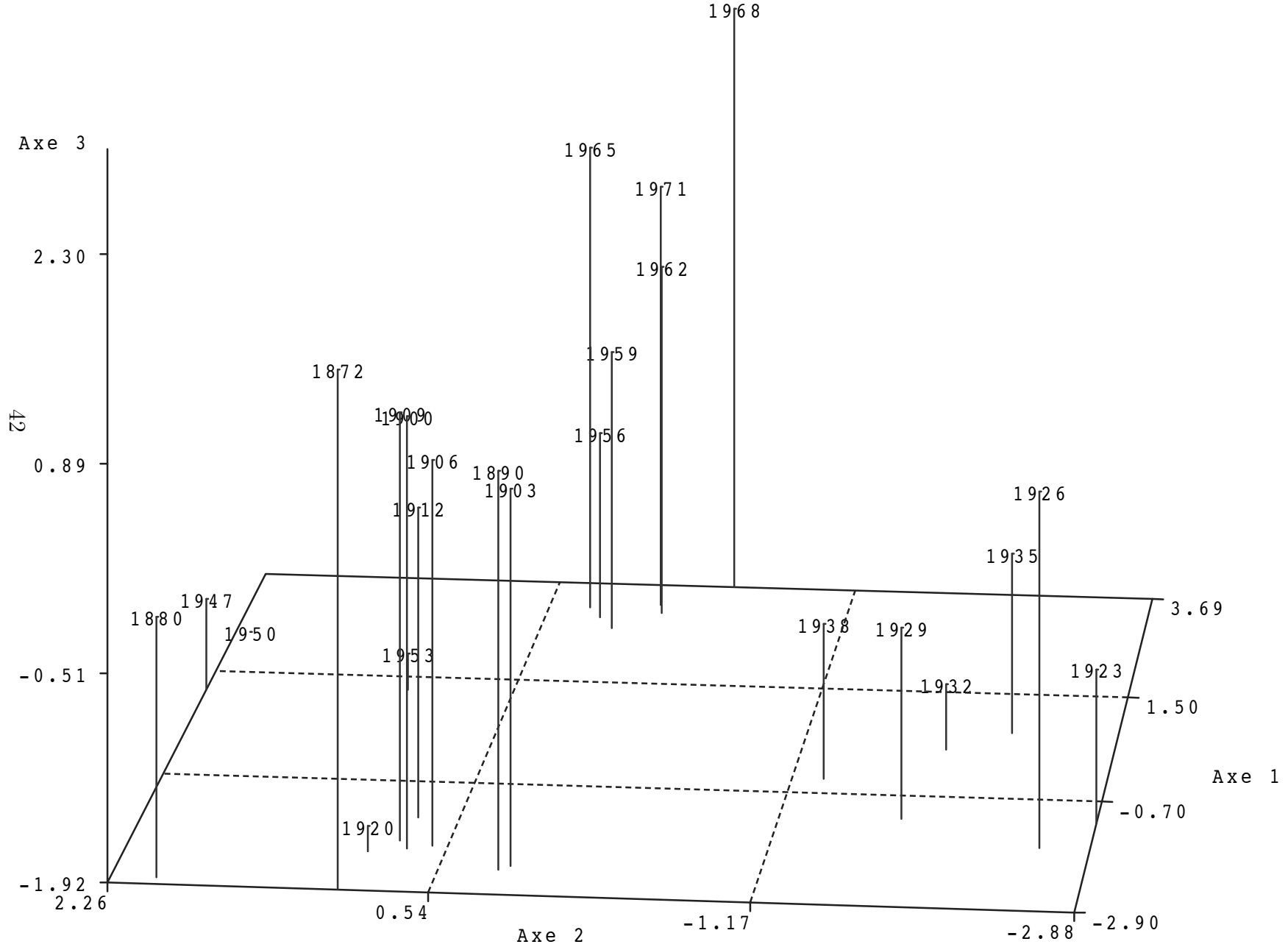
Budgets de l'etat de 1872 a 1971 : ACP normee

Projection dans le plan 1-2



Budgets de l'etat de 1872 a 1971 : ACP normee

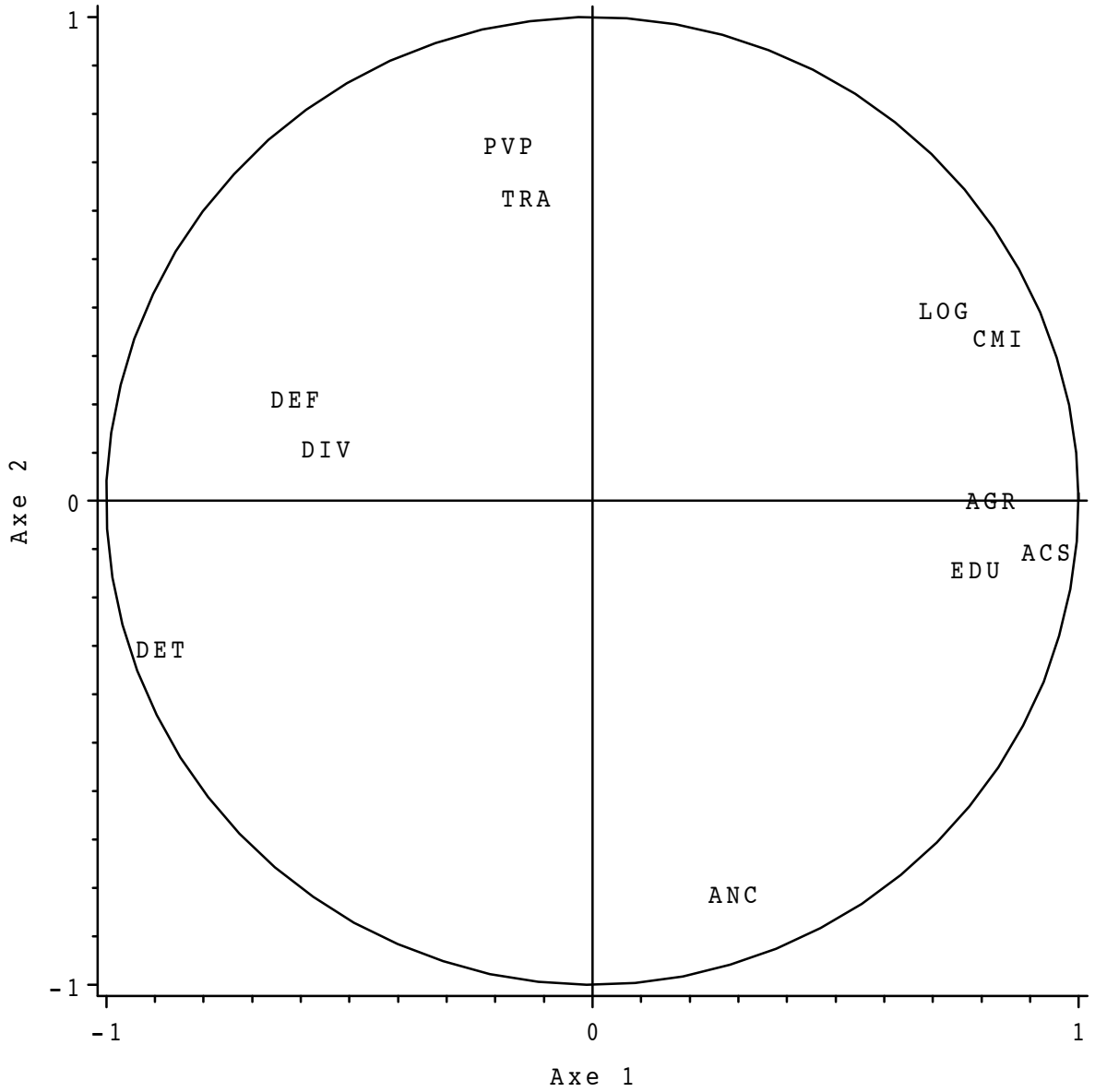
Projection dans l'espace 1-2-3



42

Budgets de l'etat de 1872 a 1971 : ACP normee

Cercle des correlations dans le plan 1-2



Références

[1] Bouroche, J.M., Saporta, G. (1980), *L'analyse des données*, PUF, Col. "Que sais-je ?"

[2] Saporta, G. (1990), *Probabilités, Analyse des données et Statistiques*, éditions Technip.

A Matrices de covariance et de corrélation empiriques

On appelle matrice de covariance empirique de p variables quantitatives

$$v_1, v_2, ..., v_j, ..., v_p$$

mesurées sur un ensemble de n unités, la matrice à p lignes et p colonnes contenant sur sa diagonale principale les variances empiriques des p variables, et ailleurs, les covariances empiriques de ces variables deux à deux :

$$\mathbf{\Sigma} = \begin{bmatrix} \mathrm{Var}(v_1) & \mathrm{Cov}(v_1, v_2) & \dots & \mathrm{Cov}(v_1, v_j) & \dots & \mathrm{Cov}(v_1, v_p) \\ \mathrm{Cov}(v_2, v_1) & \mathrm{Var}(v_2) & \dots & \mathrm{Cov}(v_2, v_j) & \dots & \mathrm{Cov}(v_2, v_p) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathrm{Cov}(v_j, v_1) & \mathrm{Cov}(v_j, v_2) & \dots & \mathrm{Var}(v_j) & \dots & \mathrm{Cov}(v_j, v_p) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(v_p, v_1) & \mathrm{Cov}(v_p, v_2) & \dots & \mathrm{Cov}(v_p, v_j) & \dots & \mathrm{Var}(v_p) \end{bmatrix}$$

avec

$$\begin{array}{lcl} \mathrm{Var}(v_j) & = & \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - x_{.j})^2 \\ \mathrm{Cov}(v_j, v_{j'}) & = & \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - x_{\bullet j})\ (x_{ij'} - x_{\bullet j'}) \end{array}$$

et :

$$x_{\bullet j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

Si

$$U_{ci} = \begin{bmatrix} x_{i1} - x_{\bullet 1} \\ x_{i2} - x_{\bullet 2} \\ \vdots \\ x_{ij} - x_{\bullet j} \\ \vdots \\ x_{ip} - x_{\bullet p} \end{bmatrix} = \begin{bmatrix} x'_{i1} \\ x'_{i2} \\ \vdots \\ x'_{ij} \\ \vdots \\ x'_{ip} \end{bmatrix}$$

est le vecteur des valeurs centrées des p variables mesurées sur la i-ème unité, on peut voir que :

$$\frac{1}{n}\sum_{i=1}^{n} U_{ci}\ {}^tU_{ci} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x'^2_{i1} & \dots & \frac{1}{n}\sum_{i=1}^{n} x'_{i1}x'_{ij} & \dots & \frac{1}{n}\sum_{i=1}^{n} x'_{i1}x'_{ip} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x'_{ij}x'_{i1} & \dots & \frac{1}{n}\sum_{i=1}^{n} x'^2_{ij} & \dots & \frac{1}{n}\sum_{i=1}^{n} x'_{ij}x'_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x'_{ip}x'_{i1} & \dots & \frac{1}{n}\sum_{i=1}^{n} x'_{ip}x'_{ij} & \dots & \frac{1}{n}\sum_{i=1}^{n} x'^2_{ip} \end{bmatrix}$$

où $x'_{ij} = x_{ij} - x_{\bullet j}$.

On retrouve bien la matrice de covariance empirique $\mathbf{\Sigma}$.

Si :

$$\mathbf{X}_c = \begin{bmatrix} {}^tU_{c1} \\ \vdots \\ {}^tU_{ci} \\ \vdots \\ {}^tU_{cn} \end{bmatrix},$$

on peut aussi écrire :

$$\mathbf{\Sigma} = \frac{1}{n}\ {}^t\mathbf{X}_c\ \mathbf{X}_c.$$

Cette matrice de covariance est une matrice symétrique. Elle est définie positive si les p variables ne sont pas liées linéairement. On peut remarquer que sa trace est égale à la somme des variances empiriques des p variables.

Si on veut travailler avec des variables centrées et réduites, on passe du tableau des valeurs centrées au tableau des valeurs centrées et réduites de la façon suivante :

$$\mathbf{X}_{cr} = (\mathbf{D}_{\mathbf{\Sigma}})^{-\frac{1}{2}}\ \mathbf{X}_c,$$

où

$$(\mathbf{D}_{\mathbf{\Sigma}})^{-\frac{1}{2}} = \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \frac{1}{\sigma_j} & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \frac{1}{\sigma_p} \end{bmatrix}$$

est la matrice diagonale qui a sur sa diagonale principale les inverses des écarts-type empirique des variables.

Si on calcule la matrice de covariance à partir d'un tableau de données centrées et réduites, on obtient la matrice des corrélations empiriques :

$$\mathbf{R} = \frac{1}{n}\ {}^t\mathbf{X}_{cr}\ \mathbf{X}_{cr} = (\mathbf{D}_{\mathbf{\Sigma}})^{-\frac{1}{2}}\ \mathbf{\Sigma}\,(\mathbf{D}_{\mathbf{\Sigma}})^{-\frac{1}{2}}\ .$$

B Décomposition de l'inertie totale

Si on décompose l'espace $\mathbb{R}^p$ comme la somme de sous-espaces de dimension 1 et orthogonaux entre eux :

$$\Delta_1 \oplus \Delta_2 \oplus ... \oplus \Delta_p$$

on peut écrire :

$$I_G = \frac{1}{n}\sum_{i=1}^{n} d^2(G, u_i) = I_{\Delta_1^*} + I_{\Delta_2^*} + ... + I_{\Delta_p^*}$$

En effet, en appliquant le théorème de Pythagore, on a :

$$d^2(G, u_i) = d^2(G, h_{\Delta_2 \oplus ... \oplus \Delta_p i}) + d^2(u_i, h_{\Delta_2 \oplus ... \oplus \Delta_p i})$$

où $h_{\Delta_2 \oplus ... \oplus \Delta_p i}$ est la projection orthogonale de u_i sur le sous-espace $\Delta_1^* = \Delta_2 \oplus ... \oplus \Delta_p$ complémentaire orthogonal de Δ_1 dans $\mathbb{R}^p$. Et comme tous les axes sont orthogonaux :

$$d^2(u_i, h_{\Delta_2 \oplus ... \oplus \Delta_p i}) = d^2(G, h_{\Delta_1 i})$$

On peut, de la même façon, décomposer $d^2(G, h_{\Delta_2 \oplus ... \oplus \Delta_p i})$:

$$d^2(G, h_{\Delta_2 \oplus ... \oplus \Delta_p i}) = d^2(G, h_{\Delta_3 \oplus ... \oplus \Delta_p i}) + d^2(G, h_{\Delta_2 i})$$

et aboutir de proche en proche à :

$$d^2(G, u_i) = d^2(G, h_{\Delta_1 i}) + d^2(G, h_{\Delta_2 i}) + ... + d^2(G, h_{\Delta_p i})$$

En utilisant cette expression du carré de la distance de u_i au centre de gravité, on retrouve la décomposition de l'inertie totale comme somme des inerties expliquées par les p axes orthogonaux.

C Méthode des multiplicateurs de Lagrange

Pour chercher les optimums d'une fonction

$$f(t_1, t_2, ..., t_p)$$

de p variables liées par une relation

$$l(t_1, t_2, ..., t_p) = cte$$

on calcule les dérivées partielles de la fonction

$$g(t_1, t_2, ..., t_p) = f(t_1, t_2, ..., t_p) - \lambda(l(t_1, t_2, ..., t_p) - cste)$$

par rapport à chacune des variables. En annulant ces p dérivées partielles et en ajoutant la contrainte, on obtient un système de $p + 1$ équations à $p + 1$ inconnues. Les $p + 1$ inconnues sont les valeurs des variables t_i $(i = 1, ..., n)$ et de λ appelé le "multiplicateur de Lagrange". L'existence de solutions à ce système est une condition nécessaire mais pas suffisante à l'existence d'un optimum pour la fonction f.

On peut généraliser ce problème au cas où les p variables sont soumises à c contraintes. On construit la fonction $g(t_1, t_2, ..., t_p)$ en rajoutant une combinaison linéaire des c contraintes, dont les coefficients $\lambda_1, \lambda_2, ..., \lambda_c$ sont les multiplicateurs de Lagrange. On doit alors résoudre un système de $p + c$ équations à $p + c$ inconnues.

D Dérivée d'une forme quadratique par rapport à un vecteur

Pour rechercher les axes principaux, on a dû calculer les dérivées partielles de ${}^t a\mathbf{\Sigma}\ a$ et de ${}^t a\ a$ par rapport aux composantes $a_1, a_2, \ldots, a_p$ du vecteur a. Il est commode de prendre la convention suivante. On note $\frac{\partial g(a)}{\partial a}$ le vecteur de $\mathbb{R}^p$ dont les composantes sont les dérivées partielles de $g(a)$ par rapport à chacune des composantes du vecteur a de $\mathbb{R}^p$:

$$\frac{\partial g(a)}{\partial a} = \begin{bmatrix} \frac{\partial g(a)}{\partial a_1} \\ \vdots \\ \frac{\partial g(a)}{\partial a_j} \\ \vdots \\ \frac{\partial g(a)}{\partial a_p} \end{bmatrix}.$$

On peut montrer que :

$$\frac{\partial({}^t a\mathbf{\Sigma}\ a)}{\partial a} = 2\mathbf{\Sigma}\ a.$$

En effet :

$$\frac{\partial({}^t a\mathbf{\Sigma}\ a)}{\partial a} = \begin{bmatrix} \frac{\partial\ {}^t a}{\partial a_1}\mathbf{\Sigma}\ a \\ \vdots \\ \frac{\partial\ {}^t a}{\partial a_j}\mathbf{\Sigma}\ a \\ \vdots \\ \frac{\partial\ {}^t a}{\partial a_p}\mathbf{\Sigma}\ a \end{bmatrix} + \begin{bmatrix} {}^t a\mathbf{\Sigma}\frac{\partial a}{\partial a_1} \\ \vdots \\ {}^t a\mathbf{\Sigma}\frac{\partial a}{\partial a_j} \\ \vdots \\ {}^t a\mathbf{\Sigma}\frac{\partial a}{\partial a_p} \end{bmatrix}.$$

On peut remarquer que dans cette dernière expression, les éléments des deux vecteurs sont égaux ligne à ligne, puisque chacun est la transposée de l'autre et qu'ils sont de dimension 1×1. Il en résulte que :

$$\frac{\partial({}^t a\mathbf{\Sigma}\ a)}{\partial a} = 2\frac{\partial\ {}^t a}{\partial a}\mathbf{\Sigma}\ a,$$

et la dérivée de ${}^t a$ par rapport à a est égale à :

$$\frac{\partial\ {}^t a}{\partial a} = \begin{bmatrix} \frac{\partial\ {}^t a}{\partial a_1} \\ \vdots \\ \frac{\partial\ {}^t a}{\partial a_j} \\ \vdots \\ \frac{\partial\ {}^t a}{\partial a_p} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \end{bmatrix} = \mathbf{I}_p$$

matrice identité de dimension p.

De la même manière, on peut montrer que $\frac{\partial\ {}^t a\ a}{\partial a} = 2a$.

E Correspondance entre statistique et géométrie

Les méthodes factorielles et leurs représentations géométriques utilisent la relation entre géométrie euclidienne et statistiques empiriques. Les statistiques élémentaires empiriques calculées sur n unités ont chacune leur correspondant géométrique dans un repère donné. Pour un ensemble quelconque de variables $x_{1,}x_{2,}\ldots,x_m$:

– Variance et carré de la norme :

$$n\mathrm{Var}(x_l) = \sum_{i=1}^{n}(x_{il} - x_{.l})^2 = \|\overrightarrow{ox_l}\|^2$$

– Covariance et produit scalaire :

$$n\mathrm{Cov}(x_l, x_{l'}) = \sum_{i=1}^{n}(x_{il} - x_{\bullet l})(x_{il'} - x_{\bullet l'}) = \langle\overrightarrow{ox_l},\ \overrightarrow{ox_{l'}}\rangle$$

– Coefficient de corrélation linéaire et cosinus d'angle :

$$\mathrm{Cor}(x_l, x_{l'}) = \frac{\mathrm{Cov}(x_l, x_{l'})}{\sqrt{\mathrm{Var}(x_l)\mathrm{Var}(x_{l'})}} = \frac{\langle\overrightarrow{ox_l},\ \overrightarrow{ox_{l'}}\rangle}{\|\overrightarrow{ox_l}\|\ \|\overrightarrow{ox_{l'}}\|} = \cos(\overrightarrow{ox_l},\ \overrightarrow{ox_l'})$$

F Matrices orthogonales

Une matrice $\mathbf{A}$ est orthogonale si tous ses vecteurs colonnes sont orthogonaux et de norme 1. Si on écrit cette matrice suivant ses vecteurs colonnes :

$$\mathbf{A} = [\mathbf{C}_1 \mathbf{C}_2 \dots \mathbf{C}_j \dots \mathbf{C}_p]$$

alors

$$\|\mathbf{C}_j\|^2 = 1, \ \forall \ j = 1, \dots, p$$

et

$$\langle \mathbf{C}_j, \mathbf{C}_{j'} \rangle = 0 \ \forall \ j = 1, \dots, p \text{ et } \forall \ j' = 1, \dots, p.$$

On peut alors voir que, pour une telle matrice, on a :

$$\begin{aligned} {}^t\mathbf{A}\ \mathbf{A} &= \begin{bmatrix} {}^t\mathbf{C}_1 \\ \vdots \\ {}^t\mathbf{C}_j \\ \vdots \\ {}^t\mathbf{C}_p \end{bmatrix} [\mathbf{C}_1 \dots \mathbf{C}_j \dots \mathbf{C}_p] \\ &= \begin{bmatrix} {}^t\mathbf{C}_1\mathbf{C}_1 & \dots & {}^t\mathbf{C}_1\mathbf{C}_j & \dots & {}^t\mathbf{C}_1\mathbf{C}_p \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ {}^t\mathbf{C}_j\mathbf{C}_1 & \dots & {}^t\mathbf{C}_j\mathbf{C}_j & \dots & {}^t\mathbf{C}_j\mathbf{C}_p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ {}^t\mathbf{C}_p\mathbf{C}_1 & \dots & {}^t\mathbf{C}_p\mathbf{C}_j & \dots & {}^t\mathbf{C}_p\mathbf{C}_p \end{bmatrix} \end{aligned}$$

${}^t\mathbf{A}\ \mathbf{A} = \mathbf{I}_p$ matrice identité de dimension p.

Il en résulte que l'inverse de $\mathbf{A}$ est égale à sa transposée :

$${}^t\mathbf{A} = \mathbf{A}^{-1}.$$

G Diagonalisation d'une matrice symétrique réelle

Les matrices de covariance ou de corrélation sont des matrices symétriques réelles car tous leurs termes sont réels On dispose pour ces matrices de bonnes propriétés qui permettent de les diagonaliser.

1. Les valeurs propres d'une matrice symétrique réelle $\mathbf{A}$ sont réelles :
   Pour une valeur propre λ correspondant au vecteur propre V, on a
   $\mathbf{A}\ V = \lambda V \Rightarrow\ {}^t\overline{V}\ \mathbf{A}\ V = \lambda {}^t\overline{V}\ V$, où $\overline{V}$ est le vecteur conjugué de V. Or ${}^t\overline{V}\ V$ et ${}^t\overline{V}\ \mathbf{A}\ V$ sont des réels puisque $\mathbf{A}$ l'est. Cela implique que λ est obligatoirement réelle.

2. Les vecteurs propres associés à deux valeurs propres différentes sont orthogonaux :
   Si V_1 et V_2 sont deux vecteurs propres associés respectivement aux valeurs propres λ_1 et λ_2, avec $\lambda_1 \neq \lambda_2$, alors
   $\mathbf{A}\ V_1 = \lambda_1 V_1$ et $\mathbf{A}\ V_2 = \lambda_2 V_2$
   $\Longrightarrow\ {}^t V_2 \mathbf{A}\ V_1 =\ \lambda_1\ {}^t V_2 V_1 =\ {}^t V_1 \mathbf{A}\ V_2 =\ \lambda_2\ {}^t V_1 V_2$
   $\Longrightarrow\ {}^t V_2 V_1 = 0$, et donc $V_1 \perp V_2$.
   On peut montrer que même dans les cas où il y a des valeurs propres multiples, on peut constituer une base orthogonale de vecteurs propres, car si une valeur propre est d'ordre d, il lui est associé un sous-espace propre de dimension d.

3. Si on se place dans une base orthonormée de vecteurs propres de $\mathbf{A}$, la matrice du changement de base est :

$$\mathbf{P} = \left[\ V_1\ \ \cdots\ \ V_j\ \ \cdots\ \ V_p\ \right]$$

Alors :

$${}^t\mathbf{P}\,\mathbf{A}\,\mathbf{P} = \begin{bmatrix} {}^t V_1 \\ \vdots \\ {}^t V_j \\ \vdots \\ {}^t V_p \end{bmatrix} \mathbf{A} \left[\ V_1\ \ \cdots\ \ V_j\ \ \cdots\ \ V_p\ \right]$$

$$= \begin{bmatrix} {}^t V_1 \\ \vdots \\ {}^t V_j \\ \vdots \\ {}^t V_p \end{bmatrix} \left[\ \mathbf{A}V_1\ \ \cdots\ \ \mathbf{A}V_j\ \ \cdots\ \ \mathbf{A}V_p\ \right]$$

$$= \begin{bmatrix} {}^t V_1 \\ \vdots \\ {}^t V_j \\ \vdots \\ {}^t V_p \end{bmatrix} \left[\ \lambda_1 V_1\ \ \cdots\ \ \lambda_j V_j\ \ \cdots\ \ \lambda_p V_p\ \right]$$

$$= \begin{bmatrix} \lambda_1 & & & & \\ & \ddots & & & 0 \\ & & \lambda_j & & \\ & 0 & & \ddots & \\ & & & & \lambda_p \end{bmatrix} = \mathbf{\Lambda}$$

$\mathbf{\Lambda}$ est la matrice diagonale des valeurs propres de $\mathbf{A}$.