Dauphine | PSL
UNIVERSITÉ PARIS

Département MIDO, M1

Analyse des données : TD 2020-2021

M1

Patrice Bertrand et Denis Pasquignon

(version du 1er septembre 2020)

Patrice Bertrand et Denis Pasquignon
CEREMADE, Université Paris-Dauphine, 75016 Paris, France.
E-mail : `denis.pasquignon@ceremade.dauphine.fr`

Ce polycopié a été créé avec LaTeX; pour la mise en forme, nous avons adapté des fichiers de style fournis par la Société Mathématique de France, notamment la classe `smfbook`.

ANALYSE DES DONNÉES : TD 2020-2021

Patrice Bertrand et Denis Pasquignon

TD 1 : Révisions algèbre

Soit p un entier naturel non nul, on se place dans $\mathbb{R}^p$ considéré comme un espace euclidien muni d'un produit scalaire $< .,. >$.

Partie 1 : Projecteur de $\mathbb{R}^p$

Soit E un espace vectoriel et soient E_1 et E_2 deux sous-espaces vectoriels de E. On rappelle que E_1 et E_2 sont dits *supplémentaires*, et on note $E = E_1 \oplus E_2$, si pour tout x de E, il existe de façon unique deux vecteurs $x_1 \in E_1$ et $x_2 \in E_2$ tels que :

$$x = x_1 + x_2 \,.$$

On rappelle que le *projecteur* P sur E_1 parallèlement à E_2, est l'application qui à tout vecteur x associe le vecteur x_1. Par la suite, l'application identité est notée I.

1. Montrer que tout projecteur est linéaire et idempotent (i.e. $P^2 = P$).
2. Montrer que si P est un endomorphisme idempotent, alors P est un projecteur sur $\operatorname{Im} P$ de direction $\operatorname{Ker} P$.
3. De la relation $P^2 = P$, déduire que les valeurs propres sont 1 ou 0.

Partie 2 : Métrique de $\mathbb{R}^p$

Soit $\mathcal{E} = (e_1, \cdots, e_p)$ une base de $\mathbb{R}^p$, la métrique de $\mathbb{R}^p$ est la matrice notée M carrée d'ordre p de terme courant

$$\forall (i, i') \in [\![1, p]\!]^2, \;\; m_{ii'} =< e_i, e_{i'} > .$$

(a) Montrer que M est une matrice symétrique définie positive.

(b) Que vaut M si $\mathcal{E}$ est une base orthogonale ? orthonormale ?

(c) Soit x et y deux vecteurs de $\mathbb{R}^p$, montrer que

$$< x, y >= x' \, M \, y \qquad (1),$$

où l'on associe aux vecteurs x et y leurs représentation matricielle. Dans toute la suite, on confond l'écriture du vecteur et de la représentation matricielle dans une base donnée. Pour signaler la métrique on mettra M en indice $< x, y >_M$.

(d) Réciproquement montrer que si M est une matrice symétrique définie positive, alors la formule (1) précédente définie un produit scalaire sur $\mathbb{R}^p$.

Partie 3 : Projection orthogonale

Soit F un sous-espace vectoriel de $\mathbb{R}^p$, on note $F^\perp$ *l'orthogonal* de F selon la métrique M, c'est-à-dire le sous-espace vectoriel défini par :

$$F^\perp = \{y \in E \mid \forall x \in F, \; < x, y >_M = 0\}.$$

Rappelons que le projecteur M-orthogonal sur F est le projecteur sur F parallèlement à $F^\perp$. Dans la suite de ce texte, P désigne le projecteur M-orthogonal sur F.
On suppose que la dimension de F est q et que $\mathcal{H} = (h_1, \cdots, h_q)$ est une base orthonormale de F.

(a) Montrer que pour tout vecteur x de $\mathbb{R}^p$, on a

$$P(x) = \sum_{k=1}^{q} < x, h_k >_M \; h_k.$$

(b) Dans le cas particulier où $q = 1$, F est une droite vectorielle, montrer que si u est un vecteur unitaire de F alors

$$\forall x \in \mathbb{R}^p, \;\; P(x) =< x, u >_M \; u = u \, u' \, M x.$$

(c) En utilisant le théorème de Pythagore, montrer que le projecteur P, M-orthogonal sur F, vérifie l'équivalence suivante

$$\forall x \in E, \;\; \widehat{x} = Px \Longleftrightarrow \begin{cases} \| \, x - \widehat{x} \, \|_M^2 = \inf_{z \in F} (\| \, x - z \, \|_M^2), \\ \widehat{x} \in F \, . \end{cases}$$

Partie 4 : Matrice de Gram

Soit $\mathcal{U} = (u_1, \cdots, u_q)$ une famille de q vecteurs de $\mathbb{R}^p$. On appelle matrice de Gram de la famille $\mathcal{U}$ notée $G_{\mathcal{U}}$ la matrice carrée d'ordre q de terme courant

$$\forall (i, i') \in [\![1, q]\!], \;\; g_{ii'} = < u_i, u_{i'} >_M \; .$$

(a) Montrer que $G_{\mathcal{U}}$ est une matrice symétrique et que

$$G_{\mathcal{U}} = U' \, M \, U,$$

où U est la matrice associée à la famille $\mathcal{U}$.

(b) Soit r le rang de $\mathcal{U}$ c'est-à-dire la dimension de l'espace vectoriel généré par $\mathcal{U}$ noté Vect($\mathcal{U}$). Montrer que $G_{\mathcal{U}}$ est une matrice de rang r. En déduire que la famille $\mathcal{U}$ est libre si et seulement si $G_{\mathcal{U}}$ est inversible.

(c) Rappeler pourquoi $G_{\mathcal{U}}$ est diagonalisable et montrer que les valeurs propres de $G_{\mathcal{U}}$ sont positives ou nulles.

(d) Calculer la trace de $G_{\mathcal{U}}$ et en déduire que toute valeur propre λ de $G_{\mathcal{U}}$ vérifie

$$\lambda \leq \sum_{i=1}^{q} \|u_i\|_M^2.$$

(e) On suppose que $\mathcal{U}$ est une famille libre, montrer alors que la matrice de la projection orthogonale sur Vect($\mathcal{U}$) est

$$P = U \, G_{\mathcal{U}}^{-1} \, U' \, M.$$

(f) On considère une autre famille de vecteurs $\mathcal{V} = (v_1, \cdots, v_n)$ une famille de n vecteurs de $\mathbb{R}^p$. On pose

$$H = U' M V,$$

où V est la matrice associée à la famille $\mathcal{V}$. Déterminer le format de H et déterminer le terme courant de H en fonction des vecteurs des familles $\mathcal{U}$ et $\mathcal{V} = (v_1, \cdots, u_n)$

TD 2
Interprétation géométrique de la moyenne et de la covariance empiriques

Dans ce texte, on considère p variables dont on connait les valeurs sur un échantillon de n individus.

Définitions et notations

On notera x_i^j la valeur de la variable j $(1 \leq j \leq p)$ pour l'individu i $(1 \leq i \leq n)$. Il en résulte qu'une variable j est caractérisée par le vecteur x^j de $F = \mathbb{R}^n$, vecteur dont les composantes sont les x_i^j pour $1 \leq i \leq n$. De même, un individu i est caractérisé par le vecteur x_i de $E = \mathbb{R}^p$, vecteur dont les composantes sont les x_i^j pour $1 \leq j \leq p$.

Chaque individu i est muni d'un poids p_i tel que :

$$\sum_{i \in I} p_i = 1, \text{ avec } I = \{1, \ldots, n\}$$

Les poids p_i sont généralement égaux à $\dfrac{1}{n}$.

Rappelons les définitions suivantes :

— Moyenne (empirique) de la variable j : $\quad \overline{x}^j = \displaystyle\sum_{i \in I} p_i \, x_i^j$.

— Variable j centrée : $y_i^j = x_i^j - \overline{x}^j$

— Covariance (empirique) entre les variables j et j' :

$$v_{jj'} = \sum_{i \in I} p_i (x_i^j - \overline{x}^j)(x_i^{j'} - \overline{x}^{j'}) = \sum_{i \in I} p_i \, y_i^j \, y_i^{j'}.$$

— Variance (empirique) de la variable j : $\quad s_j^2 = v_{jj}$

— Corrélation (empirique) entre les variables j et j' : $\quad r_{jj'} = \dfrac{v_{jj'}}{s_j \, s_{j'}}$.

On note :

$$\overline{x} = \begin{pmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_p \end{pmatrix} \qquad y_i = \begin{pmatrix} y_i^1 \\ \vdots \\ y_i^p \end{pmatrix} \qquad y^j = \begin{pmatrix} y_1^j \\ \vdots \\ y_n^j \end{pmatrix}$$

vecteur des moyennes — individu i après centrage — variable j centrée

On note enfin j_n le vecteur de $\mathbb{R}^n$ dont toutes les composantes sont égales à 1, D_p la matrice $(n \times n)$ diagonale des poids p_i, X la matrice $(n \times p)$ des données x_i^j et Y la matrice $(n \times p)$ des données centrées y_i^j.

(a) Pour quelle métrique N de $\mathbb{R}^n$ la moyenne $\overline{x}^j$ peut-elle être considérée comme l'abscisse de la projection de x^j sur j_n ?

(b) A partir du résultat obtenu en **1.** et de la relation : $y_i^j = x_i^j - \overline{x}^j$ avec i variant de 1 à n, montrer que y^j est l'image de x^j, selon une transformation géométrique de $\mathbb{R}^n$ que l'on précisera.

(c) De même, à partir de la relation : $y_i^j = x_i^j - \overline{x}^j$ avec j variant de 1 à p, montrer que y_i est l'image de x_i, selon une transformation géométrique de $\mathbb{R}^p$ que l'on précisera.

(d) Interpréter à l'aide du produit scalaire de $\mathbb{R}^n$ défini par D_p, les quantités $v_{jj'}, s_j^2$ et $r_{jj'}$.

(e) Soit V la matrice variance empirique des p variables, c'est-à-dire la matrice $p \times p$ dont le terme général est $v_{jj'}$. Montrer que V définit une forme bilinéaire symétrique positive.

(f) Tout vecteur $u = \begin{pmatrix} u_1 \\ \vdots \\ u_p \end{pmatrix}$ de $\mathbb{R}^p$ définit une nouvelle variable $u^\star$ combinaison linéaire des variables x^j

$$u^\star = \sum_{j=1}^{p} u_j \, x^j.$$

(i) Montrer que $u'Vu$ est égal à la variance de la variable $u^\star$.

(ii) De même, si w désigne un deuxième vecteur de $\mathbb{R}^p$, montrer que $u'Vw$ est égal à la covariance empirique des variables associées aux vecteurs u et w.

(iii) En déduire que V peut être considérée comme une forme quadratique semi-définie positive sur le dual $(\mathbb{R}^p)^\star$.

(g) Montrer que V peut s'écrire sous la forme matricielle suivante :

$$V = Z' \, N \, Z,$$

où N est la métrique définie en **1.** et Z est une matrice à préciser. Si V est une matrice définie, on dit alors que V est la métrique induite par la métrique N et par l'application linéaire Z.

TD 3
Application du théorème des trois perpendiculaires

Soit E un espace vectoriel de dimension finie, muni d'une métrique M. Par la suite W désigne un sous-espace vectoriel de E et l'on note P_W le projecteur M-orthogonal sur W.

(a) **Inertie d'un nuage de points.** On rappelle qu'un nuage $\mathcal{M}$ de n points munis de masses p_i, peut être identifié à l'ensemble formé par les n vecteurs $x_i \in E$ représentant ces points :

$$\mathcal{M} = \{x_i \mid i = 1, ..., n\},$$

où E désigne ici l'espace vectoriel associé à l'espace affine $\mathcal{E}$ contenant les n points. On rappelle que le centre de gravité g de ce nuage est défini par :

$$g = \frac{1}{p} \sum_{i=1}^{n} p_i \, x_i, \text{ avec } p = \sum_{i=1}^{n} p_i.$$

Dans tout le texte, on suppose que $g = 0$ et que l'inertie totale du nuage $\mathcal{M}$ peut s'écrire sous la forme :

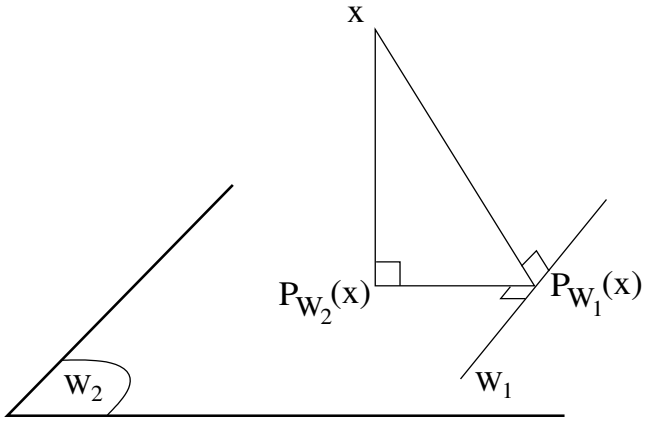
$$I_T(\mathcal{M}) = \sum_{i=1}^{n} p_i \, \|x_i - g\|^2 = \sum_{i=1}^{n} p_i \, \|x_i\|^2 .$$

De plus, on définit l'inertie du nuage $\mathcal{M}$ par rapport au sous espace vectoriel W comme étant :

$$I_W(\mathcal{M}) = \sum_{i=1}^{n} p_i \, \|x_i - P_W(x_i)\|^2.$$

En utilisant le fait que l'application linéaire $I - P_W$ est le projecteur M-orthogonal sur $W^\perp$, montrer les relations suivantes :
a) $I_{W^\perp}(\mathcal{M}) = I_T(P_W(\mathcal{M}))$,
b) $I_T(\mathcal{M}) = I_W(\mathcal{M}) + I_{W^\perp}(\mathcal{M})$.

(b) **Théorème des trois perpendiculaires.**
Si W_1 et W_2 désignent deux sous-espaces vectoriels de E, montrer que les trois conditions suivantes sont équivalentes :
(i) $W_1 \subseteq W_2$,
(ii) $W_2^\perp \subseteq W_1^\perp$,
(iii) $P_{W_1} = P_{W_1} \circ P_{W_2}$.
Remarque : cette propriété est appelée "théorème des trois perpendiculaires" ; en effet, la condition (iii) s'interprète géométriquement par l'existence de trois angles droits, comme le montre l'exemple suivant :



(c) **Inertie du nuage projeté**

(i) Soit $\mathcal{N}$ le nuage $\mathcal{M}$ projeté sur W, c'est-à-dire $\mathcal{N} = P_W(\mathcal{M})$, et soit Δu un axe du sous-espace vectoriel W. Montrer [(2)] que l'inertie de $\mathcal{N}$ par rapport à $(\Delta u)^\perp$ est égale à l'inertie

2. on utilisera de préférence le théorème des trois perpendiculaires.

du nuage $\mathcal{M}$ par rapport à $(\Delta u)^{\perp}$. En déduire que le premier axe principal du nuage $\mathcal{N}$ est l'axe de W à inertie minimum pour le nuage $\mathcal{M}$.

(ii) Montrer que :

$$I_{\Delta u}(\mathcal{N}) = I_{\Delta u}(\mathcal{M}) + I_T(\mathcal{N}) - I_T(\mathcal{M}),$$

avec :

— $I_{\Delta u}(\mathcal{N})$: inertie de $\mathcal{N}$ par rapport à Δu,
— $I_{\Delta u}(\mathcal{M})$: inertie de $\mathcal{M}$ par rapport à Δu,
— $I_T(\mathcal{N})$ inertie totale de $\mathcal{N}$,
— $I_T(\mathcal{M})$ inertie totale de $\mathcal{M}$.

TD 4
Exemple d'analyse en composantes principales

On considère le tableau de données suivant associé aux résultats de trois variables x, y et z mesurées sur un échantillon I de six individus. :

$I \setminus J$	1	2	3	4	5	6
x	1	0	0	2	1	2
y	0	0	1	2	0	3
z	0	1	2	1	0	2

On suppose que chaque individu i de I ($1 \leq i \leq 6$) est muni de la masse $1/6$. On note X le tableau associé.

1re Partie

On désire effectuer l'Analyse en composantes principales (A.C.P.) de X sur matrice variance (i.e. en supposant $\mathbb{R}^3$ muni de la métrique identité). On dit encore que l'on effectue une A.C.P. non normée.

(a) Quel est le tableau Y centré associé à X ?

(b) Donner la matrice variance V associée au tableau X.

(c) Montrer que le vecteur $\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$ est vecteur propre de V relatif à la valeur propre nulle. Qu'en déduit-on pour la représentation de J ?

(d) Calculer les axes factoriels non triviaux (i.e. relatifs à une valeur propre non nulle) associés au tableau X, les valeurs propres et les pourcentages d'inertie correspondants.

(e) A.C.P. (non normée) du nuage des individus.

(i) Dans un tableau dont les colonnes représentent les 6 individus, indiquer les résultats numériques suivants :

(A) les valeurs des variables centrées ; on indiquera sur 2 colonnes supplémentaires les coordonnées des deux axes factoriels non triviaux (ces résultats sont donnés sur les 3 premières lignes),

(B) les valeurs des deux composantes principales (2 lignes suivantes),

(C) la valeur du coefficient INR (1 ligne),

(D) les valeurs des contributions CTR et COR (4 lignes).

On rappellera la définition et l'intérêt des coefficients $\mathrm{INR}(i)$, $\mathrm{CTR}_\alpha(i)$ et $COR_\alpha(i)$.

(ii) Donner la représentation graphique du nuage des individus dans le plan euclidien des deux premiers axes principaux.

(iii) A.C.P. (non normée) du nuage des variables. On se limitera à calculer les covariances et les corrélations des trois variables x, y et z avec les deux facteurs non triviaux, et on donnera la représentation graphique de ces trois variables sur le cercle de corrélations.

2e Partie

On désire maintenant faire l'A.C.P. normée des données, c'est-à-dire on désire effectuer l'A.C.P. sur matrice de corrélation.

(a) Calculer la matrice de corrélation R.

(b) Donner le tableau centré réduit Y associé à R.

(c) . A.C.P. normée. On donnera les résultats suivants :

a) pour le nuage des individus, calculer les axes principaux d'inertie et les valeurs propres associées,

b) pour le nuage des individus et celui des variables, donner les composantes principales,

c) représenter les individus sur le plan des deux premiers axes factoriels et représenter également les variables sur le cercle des corrélations.

Pour les calculs, on pourra adopter la présentation de la question **5.a**. On comparera les résultats obtenus à ceux des questions **5.** et **6.**

3e Partie

On désire à présent faire l'A.C.P. des données en utilisant la métrique de $\mathbb{R}^3$ dont la matrice est

$$M = \mathrm{Diag}(a, b, c) = \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix}.$$

On dit encore que l'on effectue une A.C.P. avec la métrique M.

(a) Quelles conditions doivent vérifier les trois réels a, b, c pour que M soit une métrique ?

(b) Préciser la matrice dont les vecteurs propres sont les axes factoriels de l'A.C.P.

(c) Calculer cette matrice en fonction de a, b, c et montrer qu'elle est singulière. Que peut-on en déduire pour la représentation du nuage des points ? Montrer que les valeurs propres sont toutes distinctes.

Par la suite, sauf indication contraire, on prendra $a = c$.

(d) Calculer les composantes principales notées ψ_1 et ψ_2.

(e) Remplir le tableau en posant $d = \sqrt{2a + 4b}$, :

$I \setminus J$	x	y	z	ψ_1	ψ_2	$\mathrm{INR}(i)$	$\mathrm{CTR}_1(i)$	$\mathrm{CTR}_2(i)$	$COR_1(i)$	$COR_2(i)$
1	0	-1	-1							
2	-1	-1	0							
3	-1	0	1							
4	1	1	0							
5	0	-1	-1							
6	1	2	1							

Calculer en fonction des réels a et b les axes factoriels de l'A.C.P. ainsi que la part d'inertie du nuage qu'ils expliquent.

(f) Application Numérique. Effectuer la représentation graphique du nuage des points dans le premier plan principal en supposant que $M = \mathrm{Diag}(a = 1, b = 2, c = 1)$.

TD 5
A.F.T.D. : Analyse Factorielle d'un Tableau de Distances

On considère un nuage de n points, que l'on note $\mathcal{N} = \{P_1, \ldots, P_n\}$ dans l'espace $\mathbb{R}^p$ muni de la métrique M. Chacun de ces n points P_i est muni du poids p_i et on a

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

On note G le centre de gravité du nuage $\mathcal{N}$.

On suppose que pour tout i et i', on connaît le carré de la distance, noté $d_{ii'}$, entre les points P_i et $P_{i'}$.

$$d_{ii'} = \|P_iP_{i'}\|_M^2 = (P_iP_{i'})'\mathrm{M}(P_iP_{i'})$$

Par la suite, on désire effectuer une *Analyse Factorielle sur Tableau de Distances*, c'est-à-dire représenter (le mieux possible) le nuage $\mathcal{N}$ en ayant pour seules données les valeurs $d_{ii'}$, c'est-à-dire les carrés des distances entre les points. Rappelons que pour une A.C.P., les données sont constituées par les valeurs prises par les variables sur les n individus.

On pose pour tout i

$$d_{i\,\cdot} = \sum_{i'=1}^{n} p_{i'}\, d_{ii'}, \quad d_{\cdot\cdot} = \sum_{i=1}^{n} p_i\, d_{i\,\cdot} \quad \text{et} \quad \mathbf{D_p} = \mathrm{Diag}(p_i).$$

1ère partie : Etude dans le cas général.

Dans cette question, on suppose que l'on connaît la matrice Y (centrée) du type individus $\times$ variables et que l'on effectué l'ACP du nuage $\mathcal{N}$ avec la métrique M. On rappelle que compte tenu des notations précédentes, la matrice Y peut s'écrire sous la forme

$$Y = \big(GP_1, \ldots, GP_n\big)'.$$

(a) **Equation vérifiée par les composantes principales F.**

(i) Soit u le vecteur normé dirigeant l'axe factoriel associé à la valeur propre non nulle λ et soit F la composante principale qui lui est associée. Montrer que

$$YMY'D_pF = \lambda F \text{ avec } \|F\|_{D_p}^2 = \lambda.$$

(ii) Réciproquement soit F un vecteur propre de $YMY'D_p$ associé à la valeur propre non nulle λ et de norme au carré égale à λ, montrer que le vecteur $u = \frac{1}{\lambda}Y'D_pF$ est un vecteur unitaire pour la métrique M dirigeant l'axe factoriel associé à la valeur propre λ.

(iii) Montrer que YMY' est la matrice de Gram associée à la famille de vecteurs $\big(GP_1, \ldots, GP_n\big)$ c'est-à-dire la matrice carrée d'ordre n de terme courant

$$\forall (i,i') \in [\![1,n]\!],\ \ (YMY')_{ii'} = < GP_i, GP_{i'} >_{D_p}.$$

(iv) De la relation $\sum_{i=1}^{n} p_i\,(GP_i) = 0$, déduire que 0 est une valeur propre de YMY'.

(b) Calcul des termes de l'équation en fonction des valeurs des $d_{ii'}$.

(i) En appliquant le théorème de Huyghens (†), et en notant I_G l'inertie du nuage $\mathcal{N}$ par rapport à G, montrer que :

$$d_{\cdot\, i} = I_G + \|GP_i\|^2$$

(ii) Déduire de que :

$$I_G = \frac{1}{2}\, d_{\cdot\cdot}$$

†. i.e. la relation $I_A(\mathcal{N}) = I_G(\mathcal{N}) + \|GA\|^2$, où A désigne un point quelconque.

(iii) En utilisant le théorème de Pythagore généralisé [††], montrer que pour tout $i, i' \in \{1, \ldots, n\}$, on a :

$$(GP_i)' M\,(GP_{i'}) = \frac{1}{2}\big(d_{i\cdot} + d_{i'\cdot} - d_{\cdot\cdot} - d_{ii'}\big).$$

2 ème partie : Application.

On suppose $n = 4$, et que $\begin{cases} d_{12} = d_{23} = d_{34} = d_{41} = a^2 \\ d_{13} = d_{24} = b^2 \\ p_1 = p_2 = p_3 = p_4 = 1/4 \end{cases}$ où a et b sont deux réels positifs.

(a) Déterminer la matrice de Gram.

(b) Montrer qu'outre la valeur propre nulle, $Y M Y' D_p$ admet une valeur propre simple et une valeur propre double.

(c) Donner la représentation du nuage en projection sur l'axe factoriel correspondant à la valeur propre simple, et sur le plan factoriel correspondant à la valeur propre double.

(d) Indiquer les relations que doivent vérifier les nombres a et b pour que le nuage $\mathcal{N}$ soit représentable :

α) dans un espace euclidien,

β) dans un plan,

γ) sur une droite.

††. i.e. la relation $\|BC\|^2 = \|AB\|^2 + \|AC\|^2 - 2\,(AB)'\mathrm{M}\,(AC)$ vérifiée pour tout triangle ABC.

TD 6
Analyse en Composantes Principales sous R

Quelques commandes utiles

• Vecteur

— $V = c(1.3, 5, -2.1)$ crée le vecteur V comportant les trois éléments 1.3,5,-2.1
— $mean(V)$ calcule la moyenne des éléments de V et $sd(V)$ calcule l'écart-type des éléments de V.

• Matrices

— $M = matrix(c(1, 2, 3, 4, 5, 6), 2, 3)$ crée une matrice de format 2-3 égale à $\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$.
— $dim(M)$ donne le format de la matrice M.
— $M[3,]$ retourne la 3ième ligne de la matrice ou du tableau de données M et $M[,4]$ retourne la 4ième colonne de M, $M[1:2,]$ retourne la matrice composée des deux premières lignes de M et $M[,3:4]$ retourne la matrice composée des 3ième et 4ième colonnes de M
— $M = as.matrix(T)$ convertit le tableau de données T en matrice M.
— $A\%*\%B$ calcule le produit des matrices A et B au sens du produit matriciel.
— $A * B$ calcule le produit terme à terme entre les deux matrices.
— $t(A)$ transpose la matrice A.

• Matrices particulières

— $rep(1, 6)$ crée un vecteur avec 6 fois la valeur 1.
— La commande $diag$ permet de créer une matrice diagonale : $diag(1:6)$. Elle peut aussi permettre d'extraire les termes diagonaux d'une matrice carrée.

• Statistiques

— $colMeans(M)$, $rowMeans(M)$ calculent respectivement les moyennes par colonnes, par lignes.
— $cor(M)$ calcule la matrice de corrélation à partir de la matrice de données M.

• Diagonalisation

— $eigen(M)\$values$ retourne les valeurs propres de la matrice M
— $eigen(M)\$vectors$ retourne les vecteurs propres de M

• Graphique

— $plot(X1, X2)$ affiche sur un graphique les points aux coordonnées $(X1[i], X2[i])$
— $text(X1, X2, V)$ Si V est un vecteur comportant du texte, appeler cette commande après plot rajoutera le texte sur les points du graphique.

Si vous avez besoin d'aide sur une fonction, faites help(nom de la fonction).

Pour vous entrainer

(a) Créez le vecteur [1 2 3 4]
(b) Calculez sa moyenne
(c) Retranchez sa moyenne à chacune de ces composantes (ceci revient à effectuer un centrage)
(d) Vérifiez que sa moyenne est devenue nulle.
(e) Affichez sa transposée
(f) Calculez le carré de sa norme (en faisant le produit scalaire de ce vecteur avec lui-même)

Analyse par Composantes Principales sur un fichier de données

Le fichier sur lequel nous allons travailler est disponible en version complète sur mycourse au nom de temperature. Il s'agit d'un fichier décrivant les températures en degré celsius de différentes villes pour chaque mois de l'année.

Ce tableau de données comporte 15 villes soit 15 individus, et 12 mois soit 12 variables.

(a) Chargez ce fichier dans un tableau de données T. Pour ce faire, il faut écrire

```
T=read.csv("temperature.csv", sep = ";", dec= ",",row.names=1)
```

(b) Affichez ce tableau de données à l'écran

(c) Convertissez la partie numérique du tableau en matrice M Affichez M pour voir si le résultat correspond à ce que vous attendez.

(d) Pour centrer un tableau, le cours donne la formule suivante :

$$Y = X - 1_n\, g'.$$

A partir de cette formule, créer un code en R qui calcule le tableau centré Y.

(e) D'après le cours, la matrice de variance V est

$$V = Y' D_p Y.$$

Dans notre cas, tous les individus ont même poids. Créer un code pour calculer V.

(f) On choisit l'identité pour la métrique M. Déterminez l'inertie totale du nuage. Calculez les valeurs propres avec deux décimales. Retrouvez la valeur de l'inertie totale ainsi que le taux d'inertie expliqué par le premier axe puis par le plan factoriel associé aux deux premiers axes factoriels. Combien d'axes factoriels allez vous retenir ?

(g) Calculer les vecteurs propres associés aux axes factoriels de l'ACP.

(h) La composante principale associée à l'axe u_α est donnée par

$$\psi_\alpha = YMu_\alpha.$$

Créez une matrice P dont les deux premières colonnes sont les deux premières composantes principales.

(i) Affichez avec la commande *plot* les 15 individus projetés sur les deux premiers axes factoriels. Avec la commande *text*, rajoutez le nom de ces individus. Avec la commande axis, créer les axes

```
axis(1,-10:10,pos=0,labels=FALSE)
axis(2,-5:5,pos=0,labels=FALSE)
```

(j) Modifiez la matrice P pour retrouver une carte de France. Interprétez les deux axes factoriels.

(k) On pose

$$E_k = vect(u_1, \cdots, u_k).$$

On rappelle que la qualité de représentation de l'individu y_i sur E_k est

$$QLT(y_i, E_k) = \sum_{\alpha=1}^{k} \left(\frac{\psi_{i,\alpha}}{||y_i||_M}\right)^2.$$

Calculer pour tous les individus la qualité de représentation sur le premier axe.

(l) On rappelle que la contribution relative de l'individu y_i à l'inertie de l'axe α est

$$CTR_\alpha(y_i) = p_i \frac{(\psi_{i,\alpha})^2}{\lambda_\alpha}.$$

Calculer pour tous les individus la contribution relative au premier axe.

ACP avec le logiciel FactoMineR

Dans la suite de ce TP, nous utilisons le logiciel FactoMineR pour réaliser des analyses Factorielles. On commence par charger la librairie FactoMineR par la commande

```
library(FactoMineR)
```

puis on importe les données

```
T=read.csv("temperature.csv", sep = ";", dec= ",",row.names = 1)
```

(a) Afficher les statistiques de base pour chaque variable à l'aide de la commande summary

```
summary(T)
```

(b) Effectuer une analyse en composantes principales normées avec la commande

```
T.pca <- PCA(T)
```

L'option scale.unit = FALSE permet une ACP sur la matrice variance. A la suite de cette commande, les nuages des individus et des variables apparaissent.

(c) Utiliser les commandes suivantes pour obtenir les coordonnées des variables pour les deux premiers axes factoriels, puis Les valeurs propres, la distance des individus au centre de gravité et enfin les contributions

```
round(T.pca$var$coord[,1:2],2)
round(T.pca$eig,2)
round(T.pca$ind$dist,2)
round(T.pca$ind$contrib[,1:2],2)
round(T.pca$var$contrib[,1:2],2)
```

(d) On rajoute des individus supplémentaires dans le graphe. Par exemple, on considère

```
 "Amiens" <-c( 3.1,3.8 ,6.7,9.5,12.8,15.8,17.6,17.6,15.5,11.1,6.8,4.2)
T<-rbind(T, Amiens)
row.names(T)[16]<-"Amiens"

Moscow<- c(-9.2,-8,-2.5,5.9 ,12.8,16.8,18.4,16.6,11.2,4.9,-1.5,-6.2)
T<-rbind(T, Moscow)
row.names(T)[17]<-"Moscow"

Marrakech<-c(11.3 ,12.8,15.8,18.1,21.2,24.7,28.6,28.6,25,20.9,15.9,12.1)
T<-rbind(T, Marrakech)
row.names(T)[18]<-"Marrakech"
```

On lance la commande

```
T.pca <- PCA(T,ind.sup=16:18)
```

Comparez vos résultats obtenus à la main et ceux obtenus avec FactoMineR.
Commentez les résultats des deux analyses non normalisées et normalisées.

TD 7
Analyse Factorielle des Correspondances

Rappels

L'Analyse Factorielle des Correspondances (AFC) a pour but d'analyser (et même de visualiser) un tableau K de nombres positifs.

Nous nous plaçons dans le cas usuel où K est un *tableau de contingence* : étant donné deux ensembles, notés $I = \{1, 2, \ldots, p\}$ et $J = \{1, 2, \ldots, q\}$, chacun d'eux décrivant les modalités prises par une variable qualitative, le terme général k_{ij} de K est égal à l'effectif des individus ayant pris simultanément les modalités i et j relativement aux deux variables.

Le principe de l'AFC consiste à effectuer deux ACPs, l'une sur le nuage $\mathcal{N}(I)$ constitué des profils lignes de K, l'autre sur le nuage $\mathcal{N}(J)$ constitué des profils colonnes de K.

Le $i^{\text{ème}}$ individu du nuage $\mathcal{N}(I)$, appelé *profil de la $i^{ème}$ ligne de K*, est égal à la distribution (empirique) de J lorsque l'on suppose que la modalité i de l'autre variable est réalisée. Ce profil, qui est donc un vecteur de $\mathbb{R}^q$, est noté f_J^i et s'obtient en divisant les q effectifs de la $i^{\text{ème}}$ ligne de K par le total, noté $k_{i\cdot}$, de cette $i^{\text{ème}}$ ligne. Le profil f_J^i est muni du poids $f_{i\cdot} = \dfrac{k_{i\cdot}}{k}$, le nombre k étant le total des termes du tableau K, i.e. l'effectif total des individus. Par conséquent, le poids $f_{i\cdot}$ est aussi la probabilité (empirique) que la $i^{\text{ème}}$ modalité soit réalisée.

Rappelons que le nuage $\mathcal{N}(I)$ est muni de la métrique $D_{1/f_j} = \text{Diag}\big(1/f_{\cdot j}\big)_{j\in J}$, c'est-à-dire la métrique, appelée *métrique du Khi-deux*, qui est définie par la matrice diagonale d'ordre q dont le terme général est $\dfrac{1}{f_{\cdot j}}$.

Les caractéristiques du nuage $\mathcal{N}(J)$ sont définies de façon similaire : le $j^{\text{ème}}$ individu est le *profil de la $j^{ème}$ colonne de K*, noté f_I^j, et les coordonnées de ce vecteur de $\mathbb{R}^p$ sont obtenues en divisant les p effectifs de la $j^{\text{ème}}$ colonne de K par le total, noté $k_{\cdot j}$, de cette $j^{\text{ème}}$ colonne. Le poids associé à ce profil est égal à $f_{\cdot j} = \dfrac{k_{\cdot j}}{k}$, qui s'interprète comme la probabilité que j soit réalisée. La métrique du nuage, appelée aussi *métrique du Khi-deux*, est la métrique associée à la matrice $D_{1/f_I} = \text{Diag}\big(1/f_{i\cdot}\big)_{i\in I}$.

$1^{\text{ère}}$ Partie

(a) Calculer les coordonnées du centre de gravité, noté g_J, du nuage $\mathcal{N}(I)$; sans faire de calcul, donner par symétrie les coordonnées du centre de gravité, noté g_I, du nuage $\mathcal{N}(J)$.

(b) On note $d^2(i, i')$ la distance (du Khi-deux) entre i et i', c'est-à-dire la distance entre les profils lignes i et i' selon la métrique D_{1/f_J}.

  (i) Exprimer $d^2(i, i')$ en fonction des quantités $f_j^i, f_j^{i'}$ et $f_{\cdot j}$, où j varie de 1 à k.

  (ii) Par symétrie, donner sans calculs l'expression de la distance (du Khi-deux) entre j et j', c'est-à-dire la distance $d^2(j, j')$ entre les profils colonnes j et j' selon la métrique D_{1/f_I}.

(c) (i) En considérant le nuage $\mathcal{N}(I)$, calculer l'inertie totale I_T en fonction de $f_j^i, f_{i\cdot}$ et $f_{\cdot j}$, où i varie de 1 à p et j varie de 1 à q.

  (ii) Par symétrie et en considérant le nuage $\mathcal{N}(J)$, donner sans calculs une seconde expression de I_T en fonction des quantités $f_i^j, f_{i\cdot}$ et $f_{\cdot j}$.

(d) On note F_α (resp. G_α) la $\alpha^{\text{ème}}$ composante principale du nuage des profils lignes (resp. des profils colonnes).

(i) En utilisant les formules de transition, exprimer pour tout $1 \leq i \leq p$, $F_\alpha(i)$ en fonction de $\sqrt{\lambda_\alpha}, f_j^i$ et $G_\alpha(j)$, j variant de 1 à q.

(ii) En supposant que la valeur de λ_α est constante et égale à λ, en déduire que le profil centré de la $i^{\text{ème}}$ ligne, i.e. $f_J^i - g_J$, est une combinaison linéaire (que l'on précisera) des profils centrés des colonnes, i.e. des vecteurs $f_I^j - g_I$.

2ème Partie : Application 1

On désire effectuer l'AFC du tableau K_{IJ} suivant :

$I \setminus J$	A	B	C	D	E	F
i_1	1	0	0	1	1	2
i_2	0	1	0	1	2	1
i_3	0	0	2	1	1	1

(a) Calculer les marges de K_{IJ}.

(b) On considère le nuage des profils-colonnes de K_{IJ}.

(i) Déterminer le poids de chaque élément j de J.

(ii) Quelle est la métrique dont est muni l'espace $\mathbb{R}^3$?

(iii) Calculer le tableau des profils-colonnes de K_{IJ}, noté F_1 ainsi que le centre de gravité g du nuage associé.

(iv) Dans l'espace $\mathbb{R}^3$, on considère les points U_1, U_2, U_3 qui sont les extrémités des vecteurs de la base canonique de $\mathbb{R}^3$, i.e. les points de coordonnées respectives $(1,0,0)$, $(0,1,0)$ et $(0,0,1)$. Placer les profils des points A, B, C, D, E et F dans le triangle $U_1U_2U_3$ ainsi que le centre de gravité G ($\overrightarrow{OG} = g$).

(v) Calculer (avec la métrique du Khi-deux) la longueur des côtés du triangle ABC. Que peut-on dire de ce triangle ?

(vi) Combien y a-t-il d'axes factoriels non triviaux ?

(c) On note $H_\alpha(i)$ (resp. $G_\alpha(j)$) ($\alpha = 1, 2$) l'abscisse de la projection du profil de la $i^{\text{ème}}$ ligne (resp. $j^{\text{ème}}$ colonne) sur le $\alpha^{\text{ème}}$ axe factoriel issu de l'analyse des correspondances de K_{IJ} qui est associé à la valeur propre λ_α. De plus, la relation suivante est ici vérifiée :

$$H_1 = \sqrt{\frac{\lambda_1}{2}} \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix} = \sqrt{\lambda_1}\, \varphi_1^I \,.$$

(i) À l'aide de la formule de transition, déterminer les valeurs de $G_1(j)$ pour $j \in J$, et en déduire la valeur propre λ_1.

(ii) Donner le facteur φ_2^I de variance 1 (on supposera $\varphi_2^{i_1} > 0$).

(iii) Calculer de même les valeurs de $G_2(j)$ pour $j \in J$, et λ_2.

(iv) Déduire de **a)**, **b)** et **c)** les valeurs de $H_1(i)$ et $H_2(i)$ pour $i \in I$.

(v) Rappeler la définition et l'intérêt des contributions CTR_α et calculer ces contributions pour tout $\alpha \in \{1, 2\}$ et pour tous les éléments de I et J.

(vi) Même question qu'en **e)** en remplaçant CTR par COR.

(vii) Calculer les contributions INR pour tous les éléments de I et J.

(viii) Effectuer la représentation simultanée de I et J dans le plan des axes factoriels 1 et 2, et interpréter cette représentation.

(d) Parmi les trois axes (DA), (DB) et (DC) lequel est un axe de totale M-symétrie ? Que peut-on en déduire ?

3ème Partie : Application 2

On désire effectuer l'AFC du tableau K suivant :

$I \setminus J$	A	B	C	D	E	F	G
α	1	0	0	0	1	1	1
β	0	1	0	1	0	1	1
γ	0	0	1	1	1	0	1

(a) **Etude du nuage N(I)**

(i) Calculer les poids associés aux profils des lignes α, β et γ, ainsi que le carré de la distance (du Khi-deux) entre α et β, β et γ, α et γ.

(ii) En déduire que :

(A) les deux valeurs propres non triviales λ_1 et λ_2 issues de l'AFC de K, ont la même valeur que l'on notera par la suite λ.

(B) le centre de gravité g_J, que l'on précisera, est à égale distance des profils de α, β et γ.

(iii) Calculer la valeur de l'inertie totale I_T et en déduire la valeur de λ.

(b) **Etude du nuage N(J)**

(i) Calculer les poids des sept éléments de J, ainsi que le carré de la distance (du Khi-deux) entre A et B, B et C, C et A.

(ii) Montrer que le centre de gravité du nuage $\mathbf{N}(\mathbf{J})$ est égal au profil de la colonne G.

(c) **Représentation du nuage N(J)**

(i) En considérant le plan engendré par les trois points A, B, C, placer les trois points A, B, C, puis situer les quatre autres points D, E, F et G par rapport à A, B, C.

(ii) Placer sur le graphique le point a centre de gravité des quatre points A, E, F, G affectés tous les quatre de la masse 1.

(iii) Donner la valeur numérique du rapport $\dfrac{d(G,a)}{d(G,A)}$, où $d(G,a)$ (resp. $d(G,A)$) désigne la distance du Khi-deux entre G et a (resp. G et A).

(d) **Représentation du nuage N(I)**

(i) En utilisant le résultat de la question **4.** de la partie 1, calculer le profil centré $f_J^\alpha - g_J$ en fonction du profil centré $f_J^a - g_I$, i.e. le vecteur $G\alpha$ en fonction du vecteur Ga. De même, exprimer le vecteur GA en fonction du vecteur $G\alpha$. En déduire la valeur de λ.

(ii) Placer sur le graphique les points α, β et γ, et donner la valeur de la longueur $G\alpha$.

TD 8
ACM : Analyse factorielle des correspondances multiples

On réalise une ACM du questionnaire suivant : sept personnes $i_1, i_2, \cdots, i_7$ ont été interrogées. Les deux questions posées étaient :

- $Q1$: Quel temps avez vous eu lors de vos dernières vacances ?
  Les réponses possibles sont : a : excellent, b : bon, c : moyen.
- $Q2$: Où avez-vous passé vos dernières vacances ?
  Les réponses possibles sont : A : à la montagne, B : à la mer.

On pose

$$I = \{i_1, i_2, \cdots, i_7\}, \quad J_1 = \{a, b, c\}, J_2 = \{A, B\} \text{ et } J = J_1 \cup J_2.$$

On obtient les réponses suivantes

La personne i_1 était à la montagne et le temps excellent.
La personne i_2 était à la mer et le temps bon.
La personne i_3 était à la montagne et le temps moyen.
La personne i_4 était à la montagne et le temps bon.
La personne i_5 était à la mer et le temps excellent.
La personne i_6 était à la mer et le temps moyen.
La personne i_7 était à la montagne et le temps excellent.

(a) (i) Construire le tableau K et calculer f_I ainsi que f_J.

(ii) En déduire les matrices F_1 et F_2, matrice profil colonne et matrice profil ligne de l'AFC de K.

(iii) On note B_{JJ} le tableau de Burt et on pose

$$B_{JJ} = \begin{pmatrix} B_{J_1J_1} & B_{J_1J_2} \\ B_{J_2J_1} & B_{J_2J_2} \end{pmatrix},$$

où $B_{J_1J_1}$ est le tableau de Burt en croisant J_1 avec lui-même, et de même pour $B_{J_1J_2}$, $B_{J_2J_1}$, $B_{J_2J_2}$. Construire le tableau de Burt B_{JJ} et déterminer $B_{J_1J_2}$.

(iv) Expliquer pourquoi la matrice des profils colonnes de B_{JJ} est égale à la matrice des profils lignes de B_{JJ}. On notera cette matrice B_J^J.

(b) Dans cette question, on réalise **l'AFC de $B_{J_1J_2}$.**
On note $B_{J_1}^{J_2}$ le profil colonne de $B_{J_1J_2}$ et $B_{J_2}^{J_1}$ le profil ligne. On note μ_β la valeur propre non nulle associée à l'axe β et F_β (resp. G_β) la composante principale associée au nuage des profils lignes (resp. profils colonnes).

(i) Déterminer $B_{J_1}^{J_2}$ et $B_{J_2}^{J_1}$.

(ii) Vérifier que

$$B_{J_1}^{J_2\prime} B_{J_2}^{J_1\prime} = \begin{pmatrix} 7/12 & 5/12 \\ 5/9 & 4/9 \end{pmatrix}.$$

En déduire la valeur propre non triviale μ_β et montrer que la composante principale G_β associée est

$$G_\beta = \frac{1}{12\sqrt{3}} \begin{pmatrix} 3 \\ -4 \end{pmatrix}.$$

(iii) En déduire la composante principale F_β.

(c) Le but ce cette question est de montrer que l'on peut retrouver les résultats de l'AFC de B_{JJ} à partir de l'AFC de $B_{J_1J_2}$.

(i) Exprimer le profil colonne B_J^J comme une matrice par blocs en utilisant $B_{J_1}^{J_2}$, $B_{J_2}^{J_1}$ et les matrices identités I_2 et I_3.

(ii) A l'aide de produit par blocs, montrer que

$$((B_J^J)')^2 = \frac{1}{4} \begin{pmatrix} I_3 + B_{J_2}^{J_1\prime} B_{J_1}^{J_2\prime} & 2B_{J_2}^{J_1\prime} \\ 2B_{J_1}^{J_2\prime} & B_{J_1}^{J_2\prime} B_{J_2}^{J_1\prime} + I_2 \end{pmatrix}.$$

(iii) Sachant que

$$B_{J_2}^{J_1\prime} B_{J_1}^{J_2\prime} F_\beta = \mu_\beta F_\beta \text{ et } B_{J_1}^{J_2\prime} B_{J_2}^{J_1\prime} G_\beta = \mu_\beta G_\beta,$$

montrer que $\begin{pmatrix} F_\beta \\ G_\beta \end{pmatrix}$ et $\begin{pmatrix} F_\beta \\ -G_\beta \end{pmatrix}$ sont des vecteurs propres de $((B_J^J)')^2$ dont on précisera les valeurs propres notées en fonction de μ_β.

(iv) On pose $v = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$, on admet que $v \in \mathrm{Ker}(B_{J_1}^{J_2})'$, montrer que $\begin{pmatrix} v \\ 0 \end{pmatrix}$est un vecteur propre de $((B_J^J)')^2$ dont on précisera la valeur propre associée.

(v) En déduire que les valeurs propres de $((B_J^J)')^2$ sont

$$1,\ \ \lambda_1 = \frac{49}{4 \times 36},\ \ \lambda_2 = \frac{1}{4},\ \ \lambda_3 = \frac{25}{4 \times 36},\ \ \lambda_4 = 0.$$

(vi) Montrer que les composantes principales notées H_1, H_2 et H_3 associées aux valeurs propres non triviales $\lambda_1 > \lambda_2 > \lambda_3$ sont

$$H_1 = \frac{7}{2}\begin{pmatrix} F_\beta \\ G_\beta \end{pmatrix},\ \ H_2 = \frac{\sqrt{14}}{4}\begin{pmatrix} 0 \\ 1 \\ -1 \\ 0 \\ 0 \end{pmatrix} \quad \text{et } H_3 = \frac{5}{2}\begin{pmatrix} F_\beta \\ -G_\beta \end{pmatrix},\ \ .$$

(d) En utilisant le cours, nous pouvons retrouver les résultats de l'AFC de K à partir de l'AFC de B_{JJ}.

(i) Quelles sont les valeurs propres γ_β et montrer que les composantes principales $\psi_{\beta,J}$ du nuage des profils colonnes de l'AFC de K sont

$$\psi_{1,J} = \sqrt{7}\begin{pmatrix} 1/3 \\ -1/4 \\ -1/4 \\ 1/4 \\ -1/3 \end{pmatrix},\ \ \psi_{2,J} = \frac{\sqrt{7}}{2}\begin{pmatrix} 0 \\ 1 \\ -1 \\ 0 \\ 0 \end{pmatrix},\ \ \psi_{3,J} = \sqrt{5}\begin{pmatrix} 1/3 \\ -1/4 \\ -1/4 \\ -1/4 \\ 1/3 \end{pmatrix}.$$

(ii) En déduire que les composantes principales $\psi_{\beta,I}$ du nuage des profils lignes de l'AFC de K sont

$$\psi_{1,I} = \frac{7\sqrt{3}}{12}\begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \\ 0 \\ -1 \\ 1 \end{pmatrix},\ \ \psi_{2,I} = \frac{\sqrt{14}}{4}\begin{pmatrix} 0 \\ 1 \\ -1 \\ 1 \\ 0 \\ -1 \\ 0 \end{pmatrix},\ \ \psi_{3,I} = \frac{\sqrt{3}}{12}\begin{pmatrix} 1 \\ 1 \\ -6 \\ -6 \\ 8 \\ 1 \\ 1 \end{pmatrix}.$$

(iii) Représenter les deux nuages simultanément dans un plan factoriel au choix.