Université de Rennes 2
Statistiques des données
M1-GEO

Analyse en composantes principales

Ouvrages recommandés

Ces livres sont à la BU. Pour les acheter, venir au bureau A-240 ou envoyer un mail : nicolas.jegou@uhb.fr

1. P.A. Cornillon et *al.*, "Statistiques avec R.", PUR, 2008
   Présentation du logiciel : objets, graphiques, programmation. Quinze méthodes statistiques classiques présentées avec R. Indispensable pour l'aspect logiciel.
2. J. Pagès, "Statistiques générales pour utilisateurs. 1-Méthodologie", PUR, 2005
   Transcription du cours donné à Agrocampus Rennes. Estimation, analyse de variance et régression puis introduction aux plans d'expérience et à l'ACP. Introduction à la statistique pratique, trés pédagogique et trés bien écrit.
3. F. Husson et J. Pagès, "Statistiques générales pour utilisateurs. 2-Exercices et corrigés", PUR, 2005
   Exercices et corrigés en lien avec l'ouvrage précédent. Quelques TP sur R proposés.
4. J. Pagès, F. Husson, S. Lê, "Analyse de données avec R", PUR, 2009. ACP, AFC, AFM illustrées avec le logiciel R.

Les éléments de cours donnés ci-dessous sont très largement inspirés de ces ouvrages et les données utilisées pour l'illustrer sont issues des 1ère et 4e références. Chacun trouvera dans ces livres tous les compléments et démonstrations nécéssaires à une meilleure compréhension de ce cours incomplet. Par ailleurs, l'analyse en composantes principales fait appel à quelques notions géométriques essentielles comme la notion de norme de vecteurs et de produit scalaire. On pourra trouver ces notions de base dans n'importe quel manuel scolaire niveau première et terminale scientifique.

1 Motivations, notations

L'analyse en composantes principales (ACP) est une méthode classique de l'un des grands champs de la statistique appelé analyse de données (data analysis en anglais). Plutôt que cette dénomination peut-être trop générale, certains préfèrerons parler de statistique exploratoire multidimensionnelle. L'analyse des données regroupe un ensemble de méthodes dont les deux principales caractéristiques sont d'être descriptives et multidimensionnelles.

– Multidimensionnelle s'oppose à unidimensionnelle : on suppose donc que l'on disposera de plusieurs variables sur les individus concernés.
– Exploratoire s'oppose à inférentielle. Le but est de faire emerger des laisons entre les variables et de former des groupes d'individus se ressemblant. Par contre, la population observée n'est pas supposée être issue d'une population plus large dont elle constituerait un échantillon. En ce sens, l'analyse de données peut être vue comme une généralisation de la statistique descriptive.

1.1 Les données

Elles se présentent dans un tableau ou matrice à n lignes et p colonnes que l'on notera X. Chacune des n lignes représente un individu et chacune des p colonnes une variable. A l'intersection de la $i^{\text{ème}}$ ligne et de la $j^{\text{ème}}$ colonne, on trouve x_{ij} valeur de l'individu i pour la variable j (cf. figure 1).
Pour une ACP, la variables sont quantitatives : la matrice X est donc constituée de valeurs numériques. Nous avons fait figurer au bas du tableau figure 1, la moyenne et l'écart-type des variables. Avec ces notations, la

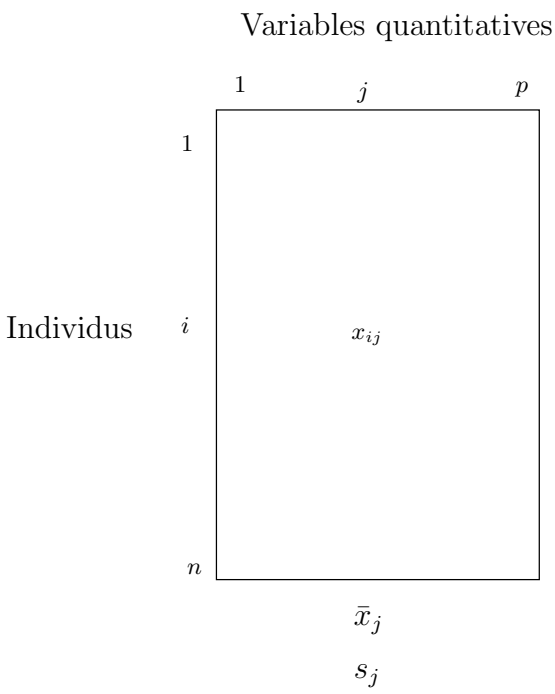

FIGURE 1 – Notations de l'ACP.

moyenne de la $j^{\text{ème}}$ variable est

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

et son écart-type :

$$s_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}.$$

Nous présentons maintenant l'exemple qui servira d'illsutration.

1.2 Exemple

Nous utilisons les données issues d'un jeu qui sera étudié plus en détail en TD et qui donnent les températures moyennes relevées dans 35 grandes villes Européennes. Les cinq premières lignes de X sont les températures mensuelles suivantes :

```
> Dataset[1:5,1:12]
          Janvier Fevrier Mars Avril  Mai Juin Juillet Aout Septembre Octobre
Amsterdam     2.9     2.5  5.7   8.2 12.5 14.8    17.1 17.1      14.5    11.4
Athenes       9.1     9.7 11.7  15.4 20.1 24.5    27.4 27.2      23.8    19.2
Berlin       -0.2     0.1  4.4   8.2 13.8 16.0    18.3 18.0      14.4    10.0
Bruxelles     3.3     3.3  6.7   8.9 12.8 15.6    17.8 17.8      15.0    11.1
Budapest     -1.1     0.8  5.5  11.6 17.0 20.2    22.0 21.3      16.9    11.3
          Novembre Decembre
Amsterdam      7.0      4.4
Athenes       14.6     11.0
Berlin         4.2      1.2
Bruxelles      6.7      4.4
Budapest       5.1      0.7
```

Chaque individu est ici une ville. Il est caractérisé par 12 valeurs qui correspondent aux observations des variables quantitatives i.e aux moyennes de température observées chaque mois. Il y a 12 variables : température moyenne en janvier, température moyenne en février... Chacune d'elle est caractérisée par les mesures qui ont été faites sur les 35 individus i.e dans les 35 villes d'Europe.

1.3 Objectifs

La matrice X peut être analysée à travers ses lignes (les individus) ou à travers ses colonnes (les variables) ce qui induit plusieurs types de questions. L'idée sera de résumer l'information portée par X en gardant à l'esprit cette dualité.

Il existe une variabilité du point de vue des températures entre les individus. Après avoir indiqué le profil moyen, la question est de savoir quels sont les individus proches de cet individu moyen et quels sont ceux qui en sont éloignés. Le concept clé est donc celui de ressemblance. Peut-on former des groupes d'individus proches les uns des autres et qui seraient éloignés des autres individus ? Quelles sont les variables (i.e les mois) qui expliquent le plus la variabilité inter-individus ?

L'autre aspect majeur de l'ACP consiste à étudier les liaisons entre variables. Certaines variables sont elles très liées entre elles ? Quelles sont les variables qui expliquent le plus ou le moins la variabilité inter-individus ?

Pour progresser dans la compréhension, il convient d'essayer à la fois de trouver des représentations appropriées aux données et de se doter de mesures permettant de quantifier la proximité entre les individus et la liaison entre les variables. C'est l'objet de la section suivante.

2 Les deux nuages

2.1 Le nuage des individus N_p

En régression multiple, nous avions vu qu'un triplet d'observations pouvait être représenté dans l'espace usuel de dimension 3 (on dit et note $\mathbb{R}^3$). Par analogie, un individu sera ici caractérisé par une ligne du tableau, i.e. ses p coordonnées et pourra être considéré comme un élément (ou un point) de $\mathbb{R}^p$. Le nuage des individus correspond donc à la représentation des n individus suivant leurs coordonnées $(x_{i1}, \cdots, x_{ip})_{i=1,\cdots,n}$ dans $\mathbb{R}^p$, espace de dimension p. Le problème est que dès que $p > 3$, c'est-à-dire dès que le nombre de variables est supérieur à 3, les individus ne sont plus représentables dans l'espace usuel.

Mesure de proximité entre individus On vient de voir qu'un des objectifs de l'ACP est de déterminer quels individus sont proches les uns des autres et en particulier de savoir si l'on peut former des groupes d'individus suivant leur proximité. Intuitivement, deux individus sont proches si leurs coordonnées dans $\mathbb{R}^p$ sont proches c'est-à-dire si les observations faites sur les p variables sont proches. Dans notre exemple, deux villes seront proches si leurs températures mensuelles sont proches. Pour quantifier cette proximité, il faut associer à l'espace $\mathbb{R}^p$ une mesure de cette proximité i.e une mesure de distance entre les individus.

Dans $\mathbb{R}^3$, une mesure du carré de la distance entre deux points $M(x_1, x_2, x_3)$ et $M'(x'_1, x'_2, x'_3)$ est la somme des carrés des différences de leurs coordonnées :

$$d^2(M, M') = (x'_1 - x_1)^2 + (x'_2 - x_2)^2 + (x'_3 - x_3)^2.$$

La mesure que nous utilisons en ACP est la généralisation de celle-ci en dimension p. Ainsi, on peut mesurer la distance entre deux individus $(x_{i1}, \cdots, x_{ip})$ et $(x_{l1}, \cdots, x_{lp})$ en calculant

$$d^2(i, k) = \sum_{j=1}^{p} (x_{ij} - x_{kj})^2. \tag{1}$$

Par exemple, la distance entre Amsterdam (individu 1) et Athenes (individu 2) est :

```
> sum((Dataset[1,1:12]-Dataset[2,1:12])^2)
[1] 786.72
```

La distance entre Amsterdam (individu 1) et Berlin (individu 3) est :

```
> sum((Dataset[1,1:12]-Dataset[3,1:12])^2)
[1] 42.49
```

Ainsi, le profil de température de Berlin est plus proche de celui d'Amsterdam que celui d'Athènes.

Une mesure de l'information portée par le nuage : la somme des distances inter-individus La distance entre deux individus mesure donc la différence existant entre eux. Analyser la variabilité entre les individus revient donc à étudier l'ensemble des distances inter-individus. Ainsi, on peut voir la somme des distances inter-individus comme une mesure de l'information portée par le nuage. En effet, la somme des distances inter-individus quantifie en quelque sorte la forme du nuage. Si les points sont tous proches les uns des autres, cette quantité sera faible alors que des points très éloignés des autres auront tendance à l'augmenter.

Avec les notations précédentes, la somme des distances inter-individus s'écrit

$$\sum_i \sum_k d^2(i,k) = \sum_i \sum_k \sum_j (x_{ij} - x_{kj})^2.$$

Un objectif de l'ACP sera de décomposer une quantité dérivant de cette somme (l'inertie) en faisant apparaître des individus ou des groupes d'individus y contribuant de manière particulière. On cherchera en particulier à déterminer quelles directions de l'espace y contribuent le plus, autrement-dit, on cherchera à savoir dans quelles directions de l'espace les déformations ou les allongements du nuage sont les plus importants.

2.2 Centrage et réduction des données

Avec le nuage N_p dont on vient de parler, on peut aussi représenter le point dont les coordonnées sont les moyennes pour chacune des variables. Ce point, appelé point moyen du nuage et noté G (par analogie au centre de gravité en mécanique), a pour coordonnées :

$$G = (\bar{x}_1, \cdots, \bar{x}_j, \cdots, \bar{x}_p).$$

Dans notre exemple, le point moyen est défini par les 12 coordonnées suivantes :

```
> apply(Dataset[,1:12],MARGIN=2,FUN=mean)

  Janvier   Fevrier      Mars     Avril       Mai      Juin   Juillet      Aout
 1.345714  2.217143  5.228571  9.282857 13.911429 17.414286 19.622857 18.980000
Septembre   Octobre  Novembre  Decembre
15.631429 11.002857  6.065714  2.880000
```

Ce point regroupe les moyennes mensuelles calculées sur les 35 villes.
On choisit en général de placer le centre du repère associé à la représentation des individus au point G. C'est l'opération de centrage des données. Cela revient à considérer les valeurs $x_{ij} - \bar{x}_j$ au lieu de x_{ij}. Notons que cette opération ne change en rien la représentation du nuage puisque le nuage des individus est inchangé.

Imaginons maintenant des données où les deux premières variables seraient des mesures de longueurs comparables mais que la première soit exprimée en centimètres et que la seconde soit exprimée en mètres. Tout naturellement, les premières coordonnées des individus seront plus grandes que les secondes donnant un trop grande importance à la première variable. Réduire les données, c'est-à-dire diviser les observations par les

écart-types de chaque variable permet de se prémunir de ce genre d'inconvénients. L'opération de réduction, qui revient à considérer

$$\frac{x_{ij} - \bar{x}_j}{s_j}$$

au lieu de $x_{ij} - \bar{x}_j$, modifie la forme du nuage en harmonisant sa variabilité dans toutes les directions des vecteurs de base. Une ACP faite sur les données centrées réduites est dite normée. Sauf mention du contraire, les ACP que nous ferons seront toutes normées. Notons d'ailleurs que, par défaut, le logiciel R effectue des ACP normées et travaille sur des données comme en figure 2.

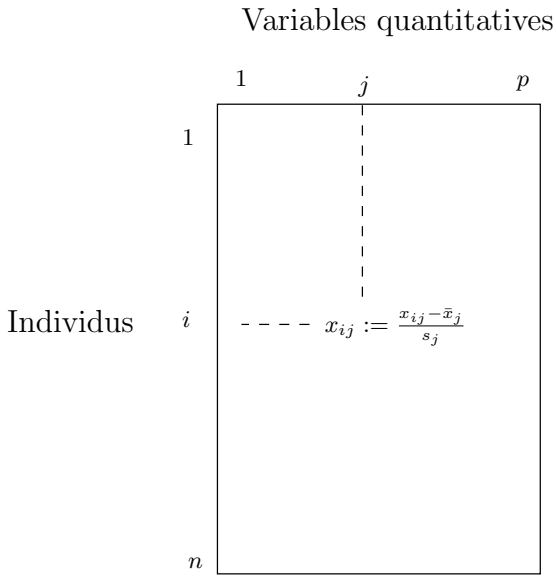

FIGURE 2 – Centrage et réduction.

On peut obtenir les données centrées-réduites avec R en appliquant la fonction scale :

```
> round(scale(Dataset[,1:12]),2)
               Janvier Fevrier  Mars Avril   Mai  Juin Juillet  Aout Septembre
Amsterdam         0.28    0.05  0.10 -0.28 -0.43 -0.79   -0.71 -0.50     -0.28
Athenes           1.41    1.36  1.33  1.61  1.89  2.13    2.18  2.20      1.99
Berlin           -0.28   -0.39 -0.17 -0.28 -0.03 -0.43   -0.37 -0.26     -0.30
Bruxelles         0.36    0.20  0.30 -0.10 -0.34 -0.55   -0.51 -0.32     -0.15
Budapest         -0.44   -0.26  0.06  0.61  0.94  0.84    0.66  0.62      0.31


               Octobre Novembre Decembre
Amsterdam         0.09     0.20     0.31
Athenes           1.90     1.87     1.63
Berlin           -0.23    -0.41    -0.34
Bruxelles         0.02     0.14     0.31
Budapest          0.07    -0.21    -0.44
```

L'analyse des données centrées-réduites fournit une information sur les individus en général beaucoup plus facile à lire que dans la matrice X initiale. On voit par exemple ici que les températures à Athènes sont nettement au dessus de la moyenne des 35 villes et ce pendant toute l'année. A l'inverse, les valeurs observées par Amsterdam sont beaucoup plus proches du profil moyen. Notons qu'ici, l'écart à la moyenne se mesure

en “nombres d'écart-types”. Par exemple, au mois d'août, la température à Athènes est supérieure à la moyenne d'environ 2 écart-types (des températures observées au mois d'août).

2.3 Le nuage des variables N_n

Nous avons dit qu'une variable était définie par une colonne dans la matrice X (cf. figure 1 ou figure 2). On assimilera une variable non pas à un point mais à un vecteur défini par l'ensemble des observations faites pour cette variable sur les individus. Ainsi, la $j^{\text{ème}}$ variable est définie par le vecteur formé de la $j^{\text{ème}}$ colonne de X c'est-à-dire par le n-uplet : $(x_{1j}, \cdots, x_{nj})$. Un jeu de données tel que présenté en figure 1 ou figure 2 comporte ainsi p variables assimilables à des vecteurs ayant n coordonnées. Le nuage des variables peut donc être considéré comme un ensemble de p vecteurs représentés dans un espace de dimension n.

Remarque 1 *Le fait d'identifier les colonnes à des vecteurs plutôt qu'à des points comme dans le cas des individus vient du fait que l'on va chercher à mesurer des corrélations entre variables et que ces corrélations peuvent être interprétées comme des mesures du degré de colinéarité entre les vecteurs.*

Une conséquence importante de l'opération de centrage-réduction des variables est que les vecteurs colonnes de la matrice X transformée comme en figure 2 ont tous des normes identiques (on dit alors des vecteurs qu'ils sont normés). On peut se ramener à une situation où les normes valent toutes 1 auquel cas, il s'en suit que les extrémités de ces vecteurs sont tous à une distance 1 de l'origine. Dans $\mathbb{R}^3$, la conséquence graphique serait que les extrémités de ces vecteurs seraient toutes situées sur la sphère de $\mathbb{R}^3$ centrée sur l'origine et de rayon 1 (on parle de sphère unité). Dans $\mathbb{R}^n$, on peut envisager une situation comparable : les vecteurs ont leur extrémité sur la sphère unité mais dans un espace de dimension n. On retrouve là l'idée que l'on accorde à toutes les variables la même importance.

Nous avons dit que l'un des objectifs de l'ACP était de recencer les liaisons entre les variables. L'ACP se borne à mesurer l'éventuelle relation linéaire entre les variables via leur coefficient de corrélation. Cette quantité, que nous allons définir tout de suite, peut être interprétée géométriquement car elle correspond au produit scalaire de deux vecteurs. Nous faisons au préalable quelques rappels sur cette notion de produit scalaire.

Rappels : produit scalaire Soient deux vecteurs $\vec{u}$ et $\vec{v}$, le produit scalaire de $\vec{u}$ et $\vec{v}$, noté $\langle \vec{u}, \vec{v} \rangle$ est défini par

$$\langle \vec{u}, \vec{v} \rangle = ||\vec{u}|| \times ||\vec{v}|| \cos(\vec{u}, \vec{v}).$$

C'est donc une quantité qui tient compte à la fois de la norme des vecteurs ainsi que de l'angle qu'ils forment. Deux vecteurs formant un angle aigu donneront un produit scalaire positif alors que pour deux vecteurs formant un angle obtu, le produit scalaire sera négatif. Entre ces deux cas, notons que deux vecteurs orthogonaux auront un produit scalaire nul.
On a une autre définition du produit scalaire de deux vecteurs en lien avec leurs coordonnées. Si l'on considère deux vecteurs $\vec{u}$ et $\vec{v}$ de $\mathbb{R}^3$ repérés par leurs coordonnées $\vec{u} = (u_1, u_2, u_3)$ et $\vec{v} = (v_1, v_2, v_3)$, le produit scalaire $\langle \vec{u}, \vec{v} \rangle$ peut s'écrire

$$\langle \vec{u}, \vec{v} \rangle = u_1 v_1 + u_2 v_2 + u_3 v_3.$$

Si l'on considère deux vecteurs $\vec{u}$ et $\vec{v}$ de norme 1 et de coordonnées respectives $\vec{u} = (u_1, u_2, u_3)$ et $\vec{v} = (v_1, v_2, v_3)$, on a donc

$$\langle \vec{u}, \vec{v} \rangle = \cos(\vec{u}, \vec{v}) = \sum_{i=1}^{3} u_i v_i.$$

Autrement dit, pour des vecteurs normés, le produit scalaire donne une mesure de l'angle qu'ils forment via le cosinus de cet angle et ce produit scalaire correspond à la somme du produit terme à terme de leurs coordonnées.

Une mesure de liaison entre deux variables : le coefficient de corrélation linéaire Si l'on considère les observations de deux variables X et Y : $(x_1, \cdots, x_n)$ et $(y_1, \cdots, y_n)$, le coefficient de corrélation linéaire est défini par le rapport entre leur covariance empirique et le produit de leurs écart-types :

$$r_{X,Y} = \frac{\text{cov}(X,Y)}{s_X s_Y} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_X}\right)\left(\frac{y_i - \bar{y}}{s_Y}\right). \tag{2}$$

Si nous considérons deux variables j et j' associées aux données que nous étudions, leur coefficient de corrélation linéaire s'écrit donc

$$r_{j,j'} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{ij} - \bar{x}_j}{s_j}\right)\left(\frac{x_{ij'} - \bar{x}_{j'}}{s_{j'}}\right). \tag{3}$$

Notons qu'un coefficient de corrélation est toujours compris entre -1 et 1. On voit par ailleurs, qu'au coefficient $1/n$ près, $r_{j,j'}$ correspond au produit scalaire entre deux vecteurs colonnes de la matrice X des données centrées réduites. Comme nous avons dit que les vecteurs colonnes j et j' avaient tous les deux la même norme, ce coefficient donne en fait une mesure du cosinus de l'angle formé par ces vecteurs. Plus précisement, $r_{j,j'}$ correspond exactement au cosinus de l'angle formé par ces deux vecteurs comme le montre le raisonnement ci-dessous.

Quelques preuves

Tout d'abord, montrons que les vecteurs colonnes de la matrice des données centrées réduites ont pour norme $\sqrt{n}$. Considérons pour cela une colonne quelconque notée X_j :

$$X_j = \left(\frac{x_{1j} - \bar{x}_j}{s_j}, \cdots, \frac{x_{nj} - \bar{x}_j}{s_j}\right).$$

Le carré de la norme de X_j est la somme des carrés de ses coordonnées :

$$\begin{aligned} ||X_j||^2 &= \sum_{i=1}^{n}\left(\frac{x_{ij} - \bar{x}_j}{s_j}\right)^2 \\ &= \frac{1}{s_j^2}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \\ &= \frac{n s_j^2}{s_j^2} \\ &= n. \end{aligned}$$

Ainsi $||X_j||^2 = n$ donc $||X_j|| = \sqrt{n}$. Considérons maintenant le produit scalaire de deux vecteurs colonne X_j et $X_{j'}$ de cette même matrice. Par définition,

$$\langle X_j, X_{j'}\rangle = ||X_j|| \times ||X_{j'}|| \cos(j, j')$$

et comme $||X_j|| = ||X_{j'}|| = \sqrt{n}$, il vient

$$\cos(j, j') = \frac{1}{n}\langle X_j, X_{j'}\rangle.$$

Comme le produit scalaire $\langle X_j, X_{j'}\rangle$ s'écrit aussi

$$\begin{aligned} \langle X_j, X_{j'}\rangle &= \sum_{i=1}^{n}\left(\frac{x_{ij} - \bar{x}_j}{s_j}\right)\left(\frac{x_{ij'} - \bar{x}_{j'}}{s_{j'}}\right) \\ &= n \times r(j, j') \end{aligned}$$

on a par identification

$$r(j, j') = \cos(j, j').$$

■

On peut interpréter assez facilement un coefficient de corrélation. Pour illustrer, on considère quelques vecteurs de $\mathbb{R}^2$ en figure 3. On dira que deux variables sont corrélées positivement si, lorsque l'une a tendance à prendre des valeurs supérieures à sa moyenne sur certains individus, l'autre a tendance à prendre également des valeurs supérieures à sa moyenne sur ces mêmes individus. Ainsi, géométriquement, lorsque les coordonnées de l'une seront grandes, les coordonnées de l'autre le seront aussi. On comprend donc que deux variables fortement corrélées pourront être représentées par des vecteurs presque colinéaires et de même sens comme les vecteurs u_1 et u_2. L'angle entre les deux étant de mesure presque nulle, le cosinus vaut presque 1. Si deux variables sont corrélées négativement c'est que quand l'une prend des valeurs supérieures à la moyenne sur certains individus, l'autre a tendance à prendre au contraire des valeurs inférieures à sa moyenne sur les mêmes individus. Cela donne lieu à des coordonnées plutôt opposées et un angle presque plat : $\cos(j, j') \approx -1$. C'est le cas pour u_1 et u_3 ou u_2 et u_3. Lorsque les vecteurs sont presque orthogonaux, la connaissance des coordonnées d'un vecteur ne donne pas d'information particulière sur les coordonnées de l'autre : c'est le cas entre u_1 et u_4 par exemple où $\cos(j, j') = r_{j,j'} \approx 0$.

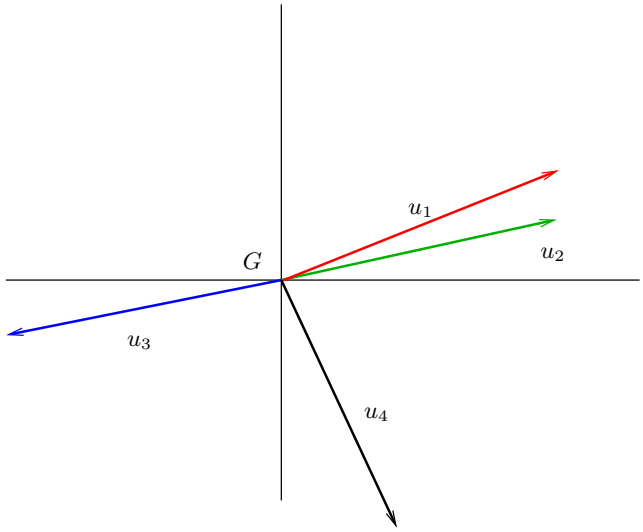

FIGURE 3 – Interprétation du coefficient de corrélation.

La fonction cor() de R donne directement la matrice des corrélations c'est-à-dire la matrice carrée de taille $p \times p$ regroupant les coefficients de corrélation de toutes les variables prises deux à deux :

```
> round(cor(Dataset[,1:12]),3)
```

	Janvier	Fevrier	Mars	Avril	Mai	Juin	Juillet	Aout	Septembre
Janvier	1.000	0.990	0.956	0.831	0.636	0.565	0.574	0.645	0.814
Fevrier	0.990	1.000	0.979	0.880	0.692	0.624	0.623	0.691	0.850
Mars	0.956	0.979	1.000	0.945	0.796	0.720	0.716	0.780	0.910
Avril	0.831	0.880	0.945	1.000	0.943	0.888	0.862	0.895	0.968
Mai	0.636	0.692	0.796	0.943	1.000	0.973	0.942	0.939	0.940
Juin	0.565	0.624	0.720	0.888	0.973	1.000	0.984	0.965	0.928
Juillet	0.574	0.623	0.716	0.862	0.942	0.984	1.000	0.987	0.932
Aout	0.645	0.691	0.780	0.895	0.939	0.965	0.987	1.000	0.961
Septembre	0.814	0.850	0.910	0.968	0.940	0.928	0.932	0.961	1.000
Octobre	0.912	0.930	0.964	0.962	0.877	0.833	0.838	0.885	0.975
Novembre	0.967	0.973	0.973	0.922	0.790	0.737	0.739	0.793	0.922
Decembre	0.994	0.983	0.957	0.851	0.677	0.609	0.617	0.681	0.841

	Octobre	Novembre	Decembre
Janvier	0.912	0.967	0.994
Fevrier	0.930	0.973	0.983
Mars	0.964	0.973	0.957
Avril	0.962	0.922	0.851
Mai	0.877	0.790	0.677
Juin	0.833	0.737	0.609
Juillet	0.838	0.739	0.617
Aout	0.885	0.793	0.681
Septembre	0.975	0.922	0.841
Octobre	1.000	0.981	0.934
Novembre	0.981	1.000	0.982
Decembre	0.934	0.982	1.000

Une remarque s'impose au regard de cette matrice : pour notre exemple, toutes les valeurs sont positives. Ainsi, si l'on considère deux variables quelconques (ici deux mois quelconques), les individus (ici les villes) prenant des valeurs supérieures à la moyenne pour une variable prendront en général des valeurs au dessus de la moyenne pour l'autre également. En bref, disons que les villes présentant des températures élévées (resp. basse) un mois en particulier ont tendance à présenter des températures élevées (resp. basses) toute l'année. Ici une première analyse est assez simple à faire du fait de cette particularité mais la plupart du temps, ce n'est pas le cas. Par ailleurs, il n'en demeure pas moins qu'il faut trouver quelles variables ou groupes de variables expliquent le plus la variabilité entre les individus.

2.4 L'inertie : l'information à expliquer

Par définition, l'inertie I des données est

$$I = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(\frac{x_{ij} - \bar{x}_j}{s_j} \right)^2 \tag{4}$$

C'est donc, au coefficient $1/n$ près, la somme des carrés de toutes les cellules de la matrice X des données centrées réduites. En cela, il est bien clair que c'est une mesure de l'information portée par les données. Cependant, on peut également en faire deux interprétations : une en lien avec le nuage N_p des individus et l'autre en lien avec le nuage N_n des variables.

Interprétation en lien avec le nuage N_p des individus Conisdérons un individu i quelconque. La quantité $\sum_{j=1}^{p} \left(\frac{x_{ij} - \bar{x}_j}{s_j} \right)^2$ représente la distance entre cet individu et le centre de gravité du nuage. Par conséquent, l'inertie peut être vue comme la somme (au coefficient $1/n$ près) des carrés des distances au centre de gravité pour tous les individus. En cela, l'inertie renseigne sur la "forme" du nuage des individus.

Interprétation en lien avec le nuage N_n des variables Il est possible d'intervertir les signes $\sum$ dans I. Une autre écriture de I est donc :

$$I = \frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n} \left(\frac{x_{ij} - \bar{x}_j}{s_j} \right)^2 .$$

La quantité $\sum_{i=1}^{n} \left(\frac{x_{ij} - \bar{x}_j}{s_j} \right)^2$ correspond au carré de la norme de la variable j : c'est donc le carré de la longueur du vecteur la représentant dans l'espace N_n. Comme on a vu plus haut que ces carrés de longueur

valaient toutes n. Il vient ainsi la simplification suivante pour I :

$$\begin{aligned} I &= \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n}\left(\frac{x_{ij}-\bar{x}_j}{s_j}\right)^2 \\ &= \frac{1}{n}\sum_{j=1}^{p} n \\ &= np/n \\ &= p. \end{aligned}$$

L'inertie (pour une ACP normée) est donc toujours égale au nombre de variables. Cette propriété est certes simple à retenir mais c'est l'écriture (4) qui permet avant tout de comprendre le lien entre I et l'information portée par les données. L'ACP consiste en fait en une décomposition de cette inertie dans des directions privilégiées des espaces propres aux représentations des individus et des variables. C'est que nous expliquons dans les sections qui suivent.

3 Représentations simplifiées des nuages N_p et N_n

Les propriétés géométriques des nuages induisent que leur visualisation permettrait de répondre aux questions posées : variabilité des individus (via les distances inter-individus dans N_p) ; liaisons entre variables (via les angles inter-variables dans N_n). Le problème est que ces nuages évoluent dans des espaces de dimension supérieure à 3 rendant leur visualisation directe impossible. L'idée de l'ACP est de fournir, pour chacun des nuages, une représentation simplifiée.

3.1 Meilleures représentations de N_p

Imaginons une forme géométrique complexe, dans un espace de dimension élevée disons de dimension 3 pour pouvoir visualiser. Pensons pour cela à l'image d'un chameau. En figure 4, nous proposons deux représentations simplifiées de cette image : des représentations en dimension 2. Deux vues viennent naturellement en tête : la vue de face et la vue de profil.

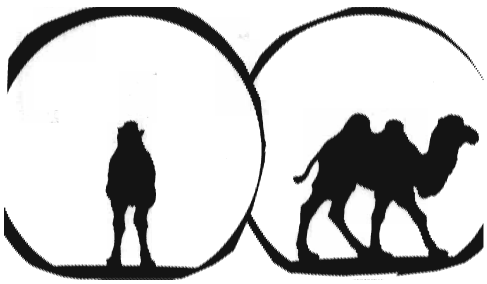

FIGURE 4 – Quelle représentation choisir pour le chameau ?

Quelle est la meilleure représentation simplifiée ? A l'évidence, c'est la vue de profil. La raison est que l'image projetée du chameau dans ce plan est plus proche de l'image initiale dans le sens où la variabilité des points servant à sa représentation est plus grande et donc restitue mieux la variabilité des points d'origine en dimension 3. Réduire la dimension pour obtenir une représentation plus simple du nuage N_p tout en conservant le plus possible de variabilité est le principe appliqué en ACP.

Meilleure représentation axiale de N_p On cherche tout d'abord la meilleure représentation axiale de N_p. plus précisement, on cherche la direction de $\mathbb{R}^p$ ($\mathbb{R}^p$ est l'espace de représentation des individus) de sorte à ce que les distances entre les points initiaux M_i soient les plus proches possibles de leurs projetés orthogonaux et ce d'un point de vue global i.e. en tenant compte de tous les points M_i. On illustre cela en figure 5.

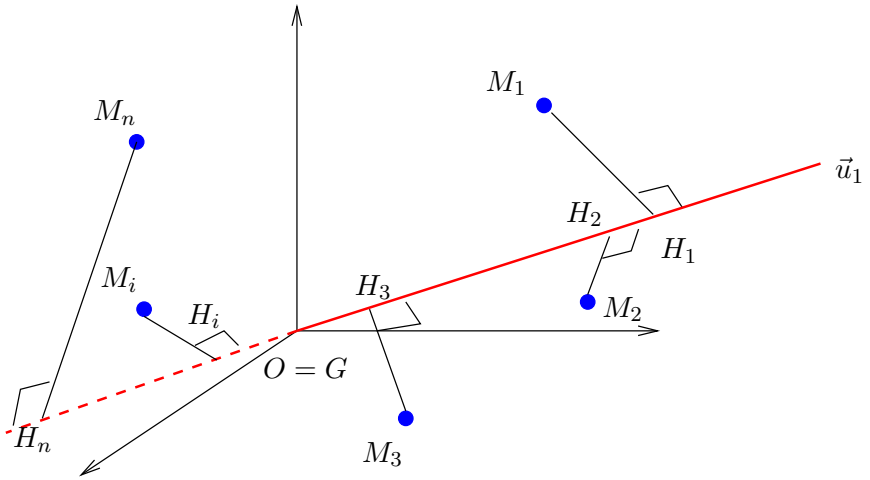

FIGURE 5 – Recherche du meilleur axe de projection.

Plus formellement, on cherche la direction $\vec{u}_1$ de $\mathbb{R}^p$ telle que $\sum_{i=1}^n OH_i^2$ soit maximum, ce qui revient au même. Comme on a vu que l'inertie correspondait (au coefficient $1/n$ près) à la somme des longueurs OM_i^2, on dira qu'on cherche $\vec{u}_1$ telle que l'inertie projetée est maximum.

Meilleure représentation plane de N_p On projette cette fois sur un plan $\mathcal{P}$ avec la même idée qu'au dessus. On cherche $\mathcal{P}$ tel que $\sum_{i=1}^n OH_i^2$ soit maximum, les H_i désignant encore les projetés orthogonaux des M_i sur $\mathcal{P}$.

On peut montrer que$\mathcal{P}$ contient $\vec{u}_1$: le "meilleur" plan contient donc le "meilleur" axe. Du coup, on caractérise $\mathcal{P}$ par $\vec{u}_1$ et par un second $\vec{u}_2$ qui est à la fois orthogonal à $\vec{u}_1$ et dans $\mathcal{P}$. Le vecteur $\vec{u}_2$ ainsi construit est le vecteur de $\mathbb{R}^p$ orthogonal à $\vec{u}_1$ et qui maximise l'inertie projetée. Autrement dit, la direction donnée par $\vec{u}_2$ est celle qui maximise l'inertie projetée dans le sous-espace de $\mathbb{R}^p$ orthogonal à $\vec{u}_1$.

Suite d'axes de représentations de N_p On construit ainsi de manière itérative une suite d'axes de directions $\vec{u}_1, \vec{u}_2, \vec{u}_3, \cdots, \vec{u}_p$ telle que
– $\vec{u}_1$ donne la direction de $\mathbb{R}^p$ qui maximise l'inertie projetée.
– $\vec{u}_2$ donne la direction du reste de l'espace qui maximise l'inertie projetée.
– ...
– ...
Il faut remarquer que ce qui vient d'être décrit comporte p itérations. A l'issue de cette opération, on dispose donc de p vecteurs orthogonaux deux à deux qui permettent donc de reconstituer l'espace $\mathbb{R}^p$ où évoluent les individus. On reconstitue ainsi la totalité de l'inertie en récupérant la part qui a été projetée dans chacune des directions.

Les projetés des 35 villes de notre exemple sont représentés en figure 6. Nous voyons que le premier axe restitue 86.87% de l'inertie et le second 11.42%. A eux deux, nous avons donc 98.29% de l'information soit la quasi-totalité. Il est donc inutile ici d'aller chercher encore de l'information dans les axes suivants.
Les points les plus éloignés du centre sont ceux qui se projetent le mieux dans ce plan. On remarque des villes comme Rejkjavik dans la partie gauche et un groupe formé de Palerme, Seville et Athènes à droite. Ces villes sont donc bien représentées par projection sur ce plan. Celles de droite sont en particulier bien représentées sur le premier axe car leurs premières coordonnées sont grandes. Nous approfondirons cette interprétation plus tard en lien avec l'analyse sur les variables et verrons en TD toute une gamme de résultats que l'on peut obtenir avec le logiciel.

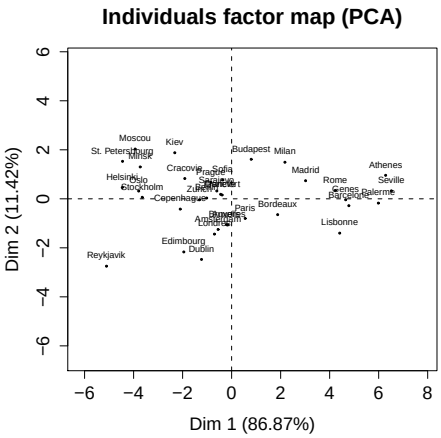

FIGURE 6 – Projection des individus sur le premier plan factoriel.

3.2 Meilleures représentations de N_n

La démarche précédente s'applique au nuage N_n représentant les variables. Nous avions dit qu'en centrant et réduisant les données, on pouvait représenter les variables par des vecteurs d'extrémité située sur la sphère unité. Projeter les variables sur un plan s'interprète donc comme en figure 7.
Les variables sont d'autant mieux projetées sur ce premier plan factoriel que l'extrémité du vecteur projeté s'approche du cercle unité. Pour notre exemple, les projections des 12 variables sur le premier plan factoriel sont données en figure 8.
Nous remarquons que toutes les variables sont bien projetées sur ce premier plan. On retrouve le même pourcentage d'inertie expliquée par ce premier plan que lorsqu'on a projeté les individus. Le fait remarquable de ces données est que toutes les variables sont corrélées positivement (ce qu'on a déjà mentionné). On peut distinguer 3 groupes de variables : septembre, avril, octobre sont très corrélées au premier axe. Le groupe formé des variables mai, juin, juillet et août s'oppose sur le second axe au groupe janvier, février, mars, novembre décembre.

4 Interprétation duale des deux représentations

Pour faire une interprétation correcte de l'ensemble des résultats, il faut comprendre de façon plus fine ce que représentent les axes orthogonaux dans les représentations faites en figures 6 et 8.

Indications Dans le nuage N_p, un individu est repéré par ses p coordonnées qui représentent les valeurs prises par cet individu sur chacune des variables. Ainsi, représentons l'espace $\mathbb{R}^p$, un individu M et la direction offrant la meilleure représentation axiale du nuage symbolisée par $\vec{u}_1$ comme en figure 9.
Les p vecteurs colonnes des données sont représentées (pour $p = 3$ bien sûr) par les vecteurs V_1, V_2, V_3. Les coordonnées (x_1, x_2, x_3) de M représentent donc les observations pour l'individu M associées à chacune de ces variables. La direction $\vec{u}_1$ peut être vue comme une combinaison linéaires des variables V_i : $\vec{u}_1 = \sum_j \alpha_j V_j$.

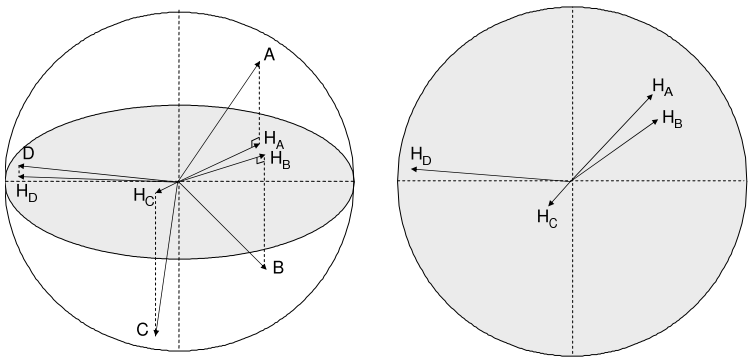

FIGURE 7 – Projection des variables sur un plan.

En cela c'est une variable synthétique, variable synthétique qui rend compte au mieux de la variabilité des individus. Il se trouve que, de l'autre point de vue, c'est aussi la variable qui est la plus corrélée à l'ensemble des p variables. L'interprétation est la même pour les autres directions $\vec{u}_2, \cdots, \vec{u}_p$ de l'espace.

Nous faisons une interprétation dans un cas simplifié illustré en figure 10.
Nous voyons que l'axe 1 est très corrélé aux variables 1 et 2. En effet, ces variables sont bien projetées dans le plan et l'angle formé avec $\vec{u}_1$ est faible : nous dirons que ces deux variables contribuent à la création de l'axe 1. L'axe 2 lui, semble devoir sa création avant tout à la variable V_3 pour les mêmes raisons. Nous voyons que l'individu M_1 est bien projeté sur le premier plan factoriel et que sa coordonnée sur l'axe 1 est grande : cet individu prend des grandes valeurs sur pour les variables V_1 et V_2. L'individu M_2 est également bien projeté sur ce plan mais c'est sur l'axe 2 que sa coordonnée est grande. Du fait de la corrélation négative entre V_3 et le second axe, on s'attend à ce que M_2 prenne des valeurs sensiblement inférieures à la moyenne pour cette variable.
La variable V_4 est mal représentée sur le plan : cela signifie qu'elle se situe quelque part dans l'orthogonal de ce plan. Le point M_3 est projeté près du centre de gravité sur le plan : cela ne signifie pas qu'il est nécessairement proche du centre de gravité mais simplement que son projeté sur le plan se retrouve près de G. Si il en est éloigné, sans doute faut-il sans doute en chercher les raisons dans une autre direction de l'espace. Peut-être est-ce lié à la variable V_4 ? Pour le savoir, il faut aller analyser les représentations suivantes i.e. celles faites avec $\vec{u}_3$ et $\vec{u}_4$ comme axes de base.

Interprétation dans l'exemple des températures Dans notre exemple, le premier plan factoriel contient presque toute l'information. L'analyse des figures 6 et 8 fournira donc assez d'éléments d'explication. Le graphe des variables montre que toutes les variables sont positivement corrélées (c'est ce qu'on appelle un "effet taille") et dont on a déjà parlé. C'est le trait essentiel de ce jeu. Des villes comme Athènes, Séville et Palerme qui sont très bien représentées à droite sur cet axe présentent des températures chaudes toute l'année en particulier pour ces mois de mi-saison. De l'autre côté, des villes comme Helsinki ou Reykjavik ont des températures froides toute l'année en particulier pour certains de ces mois.

Une information supplémentaire peut être observée en lien avec le second axe. Nous voyons que deux groupes de variables qui s'opposent dans sa construction : les mois de mai à août d'un côté et novembre à mars de l'autre. Certaines villes comme Dublin ou Edimbourg auront en effet des températures particulièrement clémentes en hiver mais moins chaudes que la moyenne au printemps-été. D'autres au contraire comme

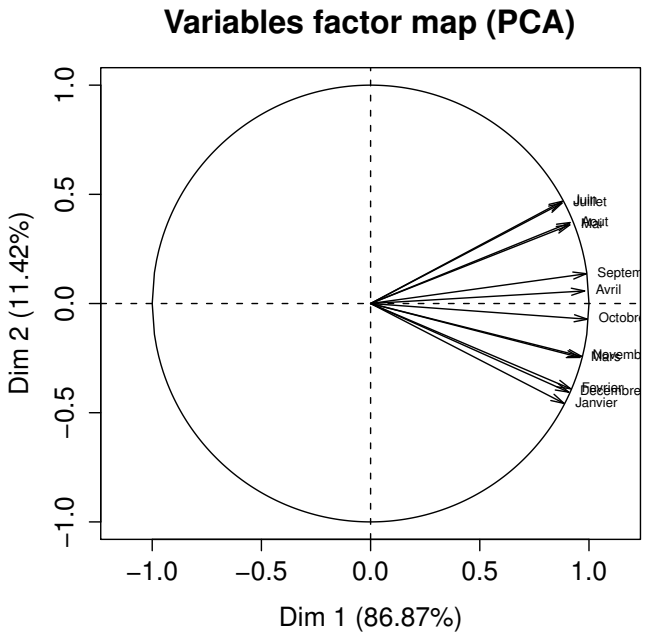

FIGURE 8 – Projection des variables.

Moscou ou Kiev auront des hivers rigoureux et des températures plus normales au printemps et en été.

Ces analyses constituent une première approche du travail qui serait à faire sur ces données. Nous verrons en TD d'autres indicateurs et d'autres méthodes qui viendront compléter cette étude sommaire.

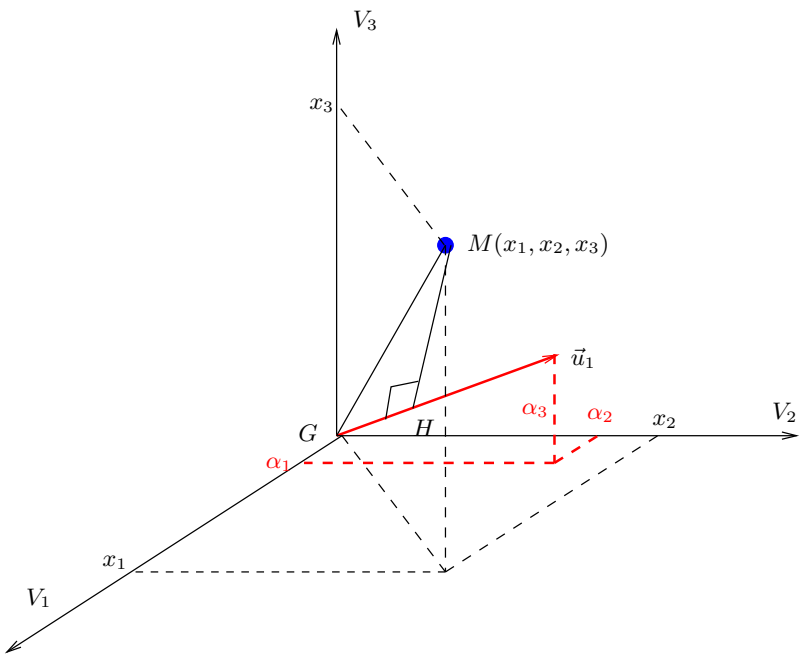

FIGURE 9 – Interprétation des axes de l'ACP.

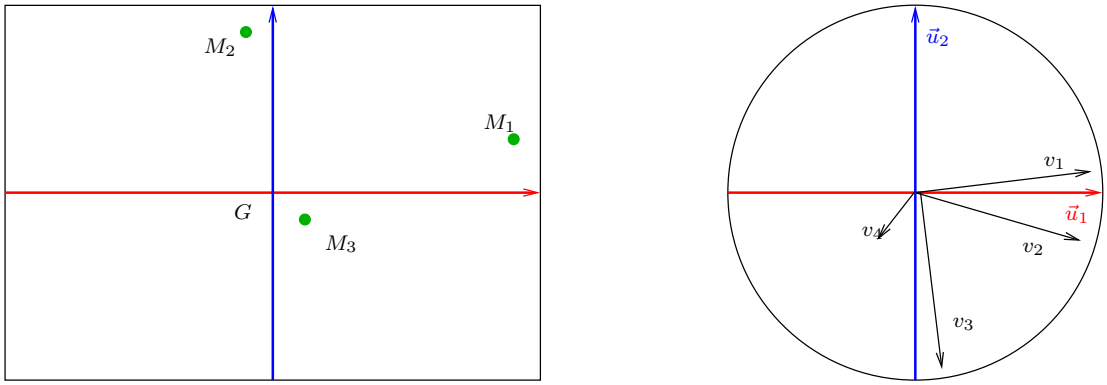

FIGURE 10 – Interprétation duale l'ACP.

Ci-dessous les données centrées réduites pour une lecture plus aisée des résultats.

```
> tp
             Janvier Fevrier  Mars Avril   Mai  Juin Juillet  Aout Septembre
Amsterdam       0.28    0.05  0.10 -0.28 -0.43 -0.79   -0.71 -0.50     -0.28
Athenes         1.41    1.36  1.33  1.61  1.89  2.13    2.18  2.20      1.99
Berlin         -0.28   -0.39 -0.17 -0.28 -0.03 -0.43   -0.37 -0.26     -0.30
Bruxelles       0.36    0.20  0.30 -0.10 -0.34 -0.55   -0.51 -0.32     -0.15
Budapest       -0.44   -0.26  0.06  0.61  0.94  0.84    0.66  0.62      0.31
Copenhague     -0.32   -0.48 -0.81 -0.91 -0.86 -0.61   -0.71 -0.64     -0.57
Dublin          0.63    0.51  0.14 -0.39 -1.07 -1.24   -1.29 -1.17     -0.71
Helsinki       -1.30   -1.53 -1.63 -1.62 -1.13 -1.03   -0.68 -1.09     -1.44
Kiev           -1.32   -1.31 -1.14 -0.49  0.12  0.12   -0.06 -0.13     -0.47
Cracovie       -0.92   -0.77 -0.68 -0.36 -0.22 -0.15   -0.34 -0.37     -0.47
Lisbonne        1.66    1.65  1.56  1.37  0.85  0.60    0.53  0.78      1.16
Londres         0.37    0.36  0.06 -0.26 -0.61 -0.70   -0.76 -0.67     -0.40
```

Madrid	0.66	0.80	0.86	0.77	0.64	1.02	1.42	1.43	1.01
Minsk	-1.50	-1.53	-1.47	-1.02	-0.46	-0.46	-0.62	-0.72	-0.98
Moscou	-1.93	-1.79	-1.49	-0.86	-0.28	-0.25	-0.37	-0.61	-1.08
Oslo	-1.03	-1.09	-1.20	-1.28	-1.10	-0.76	-0.76	-0.96	-1.10
Paris	0.43	0.27	0.43	0.11	-0.06	-0.28	-0.17	-0.08	0.11
Prague	-0.48	-0.37	-0.33	-0.13	0.12	0.06	-0.09	-0.08	-0.18
Reykjavik	-0.30	-0.39	-0.91	-1.68	-2.26	-2.44	-2.38	-2.25	-1.88
Rome	1.05	1.09	1.08	1.16	1.19	1.29	1.34	1.37	1.28
Sarajevo	-0.50	-0.26	-0.07	0.00	-0.03	-0.12	-0.20	-0.08	-0.10
Sofia	-0.55	-0.37	-0.19	0.11	0.12	0.09	0.11	0.14	0.04
Stockholm	-0.88	-1.04	-1.34	-1.52	-1.44	-0.85	-0.68	-0.80	-0.96
Anvers	0.32	0.12	0.20	-0.10	-0.31	-0.58	-0.48	-0.37	-0.23
Barcelone	1.41	1.47	1.35	1.27	1.07	1.14	1.28	1.37	1.48
Bordeaux	0.77	0.82	0.78	0.69	0.33	0.27	0.22	0.27	0.48
Edimbourg	0.28	0.25	-0.11	-0.57	-1.23	-1.33	-1.38	-1.26	-0.86
Francfort	-0.21	-0.08	0.04	0.11	0.12	0.03	-0.17	-0.18	-0.20
Geneve	-0.23	-0.06	-0.03	0.03	-0.03	-0.03	-0.06	-0.13	-0.15
Genes	1.34	1.18	1.27	1.19	1.10	1.08	1.36	1.51	1.50
Milan	-0.04	0.25	0.57	0.87	1.04	1.17	1.17	1.02	0.80
Palerme	1.66	1.69	1.66	2.00	2.13	1.92	1.36	0.89	1.62
Seville	1.70	1.74	1.82	1.79	1.77	1.80	1.98	2.07	2.11
St. Petersbourg	-1.73	-1.84	-1.84	-1.60	-1.19	-0.61	-0.34	-0.56	-1.01
Zurich	-0.37	-0.28	-0.19	-0.21	-0.31	-0.37	-0.45	-0.48	-0.37

	Octobre	Novembre	Decembre
Amsterdam	0.09	0.20	0.31
Athenes	1.90	1.87	1.63
Berlin	-0.23	-0.41	-0.34
Bruxelles	0.02	0.14	0.31
Budapest	0.07	-0.21	-0.44
Copenhague	-0.51	-0.43	-0.32
Dublin	-0.30	0.14	0.51
Helsinki	-1.34	-1.31	-1.04
Kiev	-0.81	-1.07	-1.30
Cracovie	-0.56	-0.76	-0.92
Lisbonne	1.48	1.67	1.65
Londres	-0.19	0.05	0.31
Madrid	0.67	0.58	0.51
Minsk	-1.20	-1.31	-1.43
Moscou	-1.37	-1.57	-1.79
Oslo	-1.23	-1.22	-1.16
Paris	0.35	0.27	0.47
Prague	-0.37	-0.50	-0.52
Reykjavik	-1.50	-0.96	-0.54
Rome	1.27	1.23	1.09
Sarajevo	-0.12	-0.21	-0.42
Sofia	-0.07	-0.23	-0.46
Stockholm	-1.04	-0.96	-0.90
Anvers	0.11	0.16	0.37
Barcelone	1.50	1.54	1.43
Bordeaux	0.58	0.53	0.65
Edimbourg	-0.53	-0.17	0.17
Francfort	-0.28	-0.26	-0.24

```
Geneve            -0.28    -0.26    -0.30
Genes              1.57     1.34     1.43
Milan              0.49     0.18    -0.06
Palerme            1.71     1.93     1.84
Seville            1.94     1.85     1.67
St. Petersbourg   -1.34    -1.42    -1.65
Zurich            -0.49    -0.47    -0.52
attr(,"scaled:center")
  Janvier   Fevrier      Mars     Avril       Mai      Juin   Juillet      Aout
 1.345714  2.217143  5.228571  9.282857 13.911429 17.414286 19.622857 18.980000
Septembre   Octobre  Novembre  Decembre
15.631429 11.002857  6.065714  2.880000
attr(,"scaled:scale")
  Janvier   Fevrier      Mars     Avril       Mai      Juin   Juillet      Aout
 5.502157  5.498956  4.863040  3.806456  3.273582  3.320271  3.574673  3.727939
Septembre   Octobre  Novembre  Decembre
 4.109728  4.323226  4.566820  4.967411
```

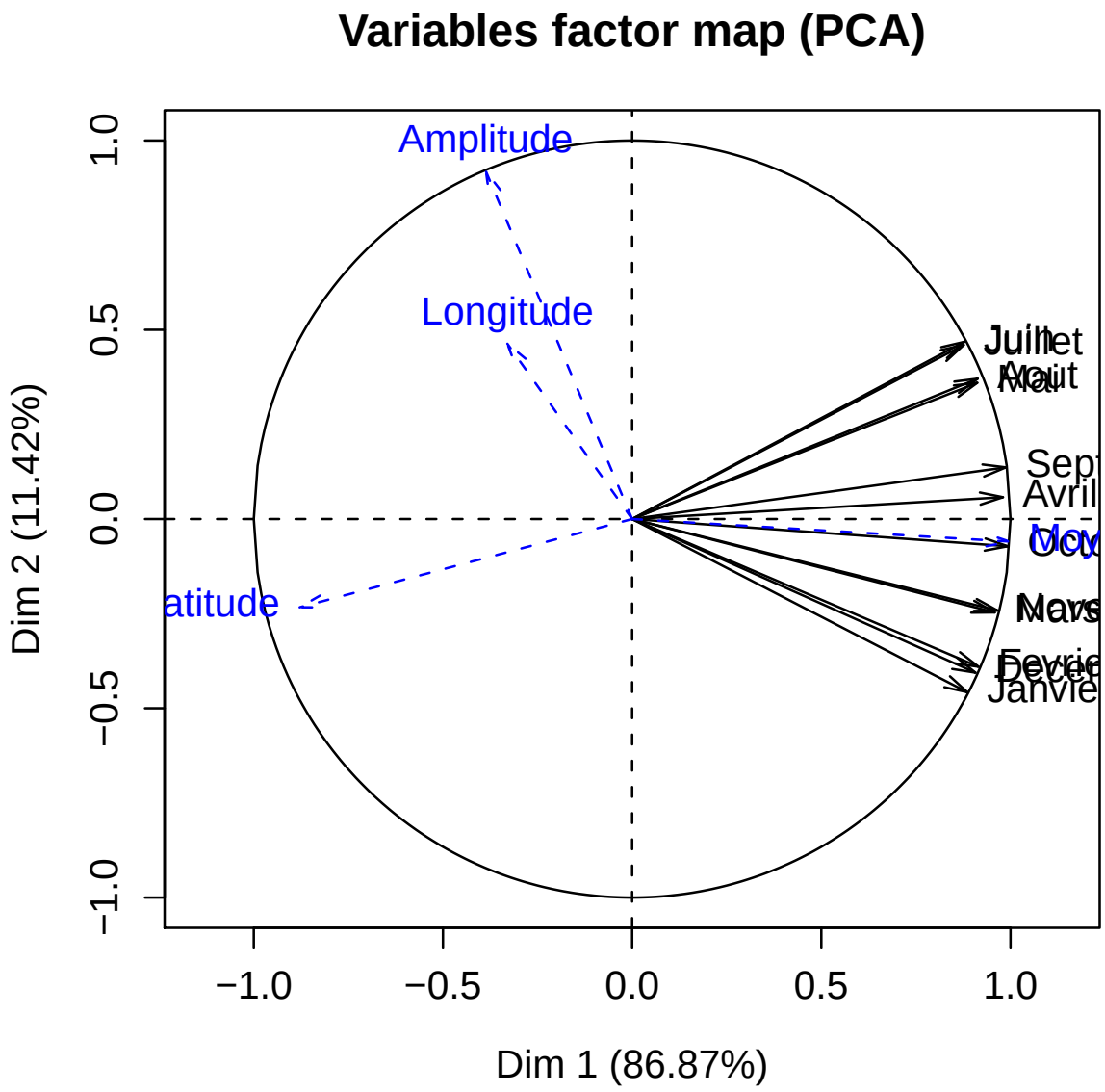

FIGURE 11 – Variables supplementaires.