BURGEVIN Valentin GR IMA

DURET Guillaume

Analyse en Composantes Principales (ACP):

2018-2019

# INTRODUCTION

Nous voulons, à l’aide d’un tableau de données affichant les habitudes alimentaires de 9 pays (qui représentent les données) pour 16 aliments (qui représentent les variables statistiques). Nous voulons savoir s’il est possible de retrouver un pays si on connaît ses habitudes alimentaires ce qui n’ai pas évident à la vue du tableau de valeur suivant :

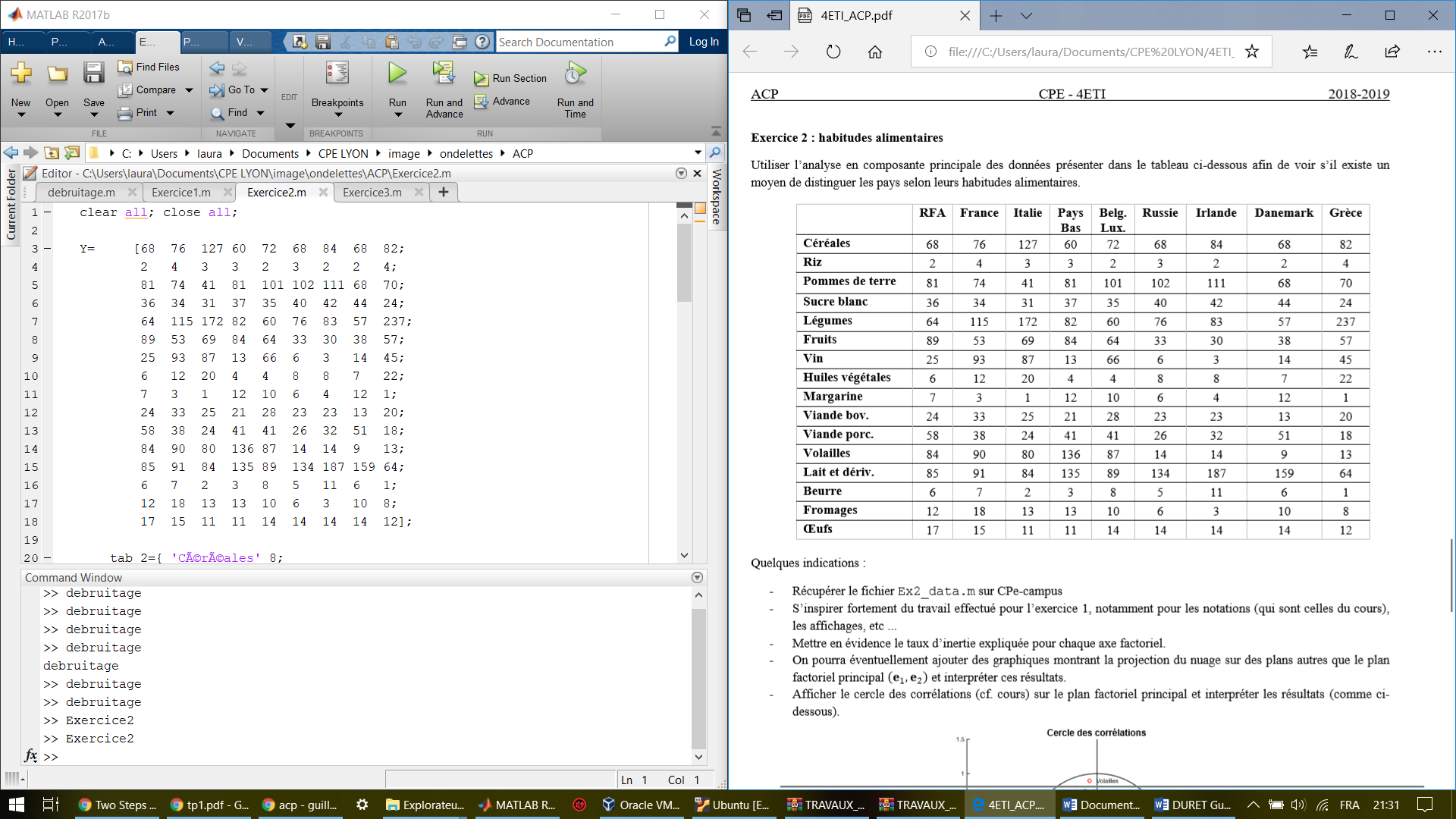


Figure 1: Tableau de données

Pour cela nous allons appliquer la méthode d'analyse en composantes principales afin d’approximer les données et de réduire le nombre de variables. En effet on peut avoir finalement seulement 2 variables qui correspondent aux 2ère composantes principales sur lesquelles les valeurs sont projeté en perdant le moins possible d’information. On parle de réduction de dimensionnalité.

# Methode

Nous écrivons le code suivant :

clear all; close all;

Y= [68 76 127 60 72 68 84 68 82;

2 4 3 3 2 3 2 2 4;

81 74 41 81 101 102 111 68 70;

36 34 31 37 35 40 42 44 24;

64 115 172 82 60 76 83 57 237;

89 53 69 84 64 33 30 38 57;

25 93 87 13 66 6 3 14 45;

6 12 20 4 4 8 8 7 22;

7 3 1 12 10 6 4 12 1;

24 33 25 21 28 23 23 13 20;

58 38 24 41 41 26 32 51 18;

84 90 80 136 87 14 14 9 13;

85 91 84 135 89 134 187 159 64;

6 7 2 3 8 5 11 6 1;

12 18 13 13 10 6 3 10 8;

17 15 11 11 14 14 14 14 12];

tab\_1={ 'RFA' 3;

'France' 6;

'Italie' 6;

'Pays Bas' 8;

'Belg. Lux.' 10;

'Russie' 6;

'Irlande' 7;

'Danemark' 8;

'Grèce' 5};

tab\_2={ 'Céréales' 8;

'Riz' 3;

'Pommes de terre' 14;

'Sucre blanc' 11;

'Légumes' 7;

'Fruits' 6;

'Vin' 3;

'Huiles végétales' 16;

'Margarine' 9;

'Viande bov.' 11;

'Viande porc.' 12;

'Volailles' 9;

'Lait et dériv.' 14;

'Beurre' 6;

'Fromages' 8;

'Oeufs' 5 };

Y = Y';

[n,m]=size(Y); %Nombres de donnees et de variables statistiques

X=zeros(n,m); %initialisation de la matrice des donnees centrees

moy=mean(Y); %Vecteur des moyennes de chaque colonne

for i = (1:m) %Calcul de la matrice des donnees centrees

X(:,i)=Y(:,i)-moy(i);

end

M=1/n\*(X'\*X); %Calcul de la matrice de covariance

[V,D]=eig(M); %Diagonalisation de M

lambda=flipud(diag(D)); %Valeurs propres dans l'ordre dÃ©croissant

P=fliplr(V); %Rangement des vecteurs propres dans l'ordre des valeurs propres

figure(2)

plot(lambda,'b-\*') %Affichage des valeurs propres

axis([-inf,inf,-inf,inf])

title('Affichage des valeurs propres');

En effet dans ce code nous réalisons une transposé du tableau de données pour bien avoir les individus (pays) sur les lignes et les variables statistiques autrement dit attributs (aliments) en colonne m. Cela nous permet d’obtenir la matrice des données.

Nous réalisons ensuite la matrice des données centrées en retirant à chaque élément de la matrice la moyenne de chaque colonne qui correspond à la moyenne de chaque aliment pour les pays.

On calcule ensuite la matrice de covariance à l’aide de la formule :

Avec X la matrice des données centrées.

Etant donné que l’on cherche à trouver les axes vectoriels qui minimise la perte d’information ; on veut maximiser l’inertie de chaque axe vectoriel.

Or d’après le théorème :

**Le premier axe factoriel e1 est le vecteur propre de la matrice M associé à la plus grande valeur propre.**

De ce fait, nous diagonalisation ensuite la matrice afin d’en extraire les valeurs propres maximal qui serviront ensuite à approximer au mieux la matrice.

Ainsi on les organise de façon décroissante à l’aide de la fonction flipud et on réorganise aussi les vecteurs propres à l’aide de la fonction fliplr

L’affiche des valeurs propres nous donne la figure suivante :

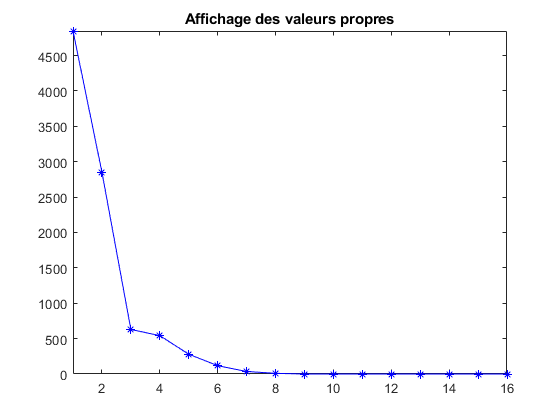


Figure 2: affichage des valeur propres dans l'ordre décroissant

On a 16 valeurs propres ce qui est cohérent au nombre de variables statistiques.

On peut aussi remarquer que les dernières valeurs propres sont très proches de zéro et qu’une majorité des informations sont contenues dans les premières valeurs propres

En effet on peut s’en rendre compte en calculant le taux d’inertie :

tau=zeros(1,m); %Initialisation du vecteur des taux d'inertie

for i=1:3 %Calcul des taux d'inertie

tau(i)=lambda(i)/sum(lambda);

end

Cela nous permet de montrer que 52.05% des informations est contenue dans la première valeur propre, 30.62% dans la seconde et 6.76% dans la troisième. On obtient donc environ 90% des informations contenues dans les 3ères valeurs propres.

Nous devons donc ensuite projeter les données sur les vecteurs propres obtenues :

Xstar=X\*P;

figure(3)

plot(Xstar(:,1),Xstar(:,2),'bo');

text(Xstar(:,1)+0.1,Xstar(:,2),tab\_1(:,1));

grid()

xlabel('e2 (30.62 %)')

ylabel('e3 (6,76 %)')

title('analyse en composante principale');

en composante principale');

En effet nous projetons les données centrées en multipliant la matrice de données centrée par la matrice de projection P composée des différant vecteurs propres.

Après affichage des valeurs de cette projection sur les 2 premier axes vectoriels (qui sont les deux premières colonnes de la matrice) on obtient :

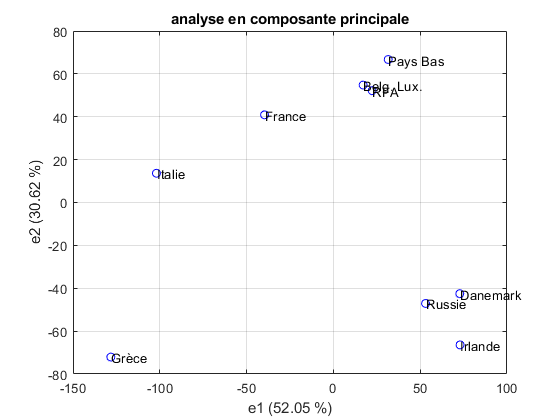


Figure 3:anakyse en composante principales pour les 2 premiers axes vectoriels

Sur cette figure on peut y voir la position des différents pays par rapport à la moyenne qui est (0,0) on peut donc remarquer le Grèce est généralement souvent en dessous de la moyenne pour chaque aliment à côté des autres pays. À l’opposé on peut remarquer les pays bas qui sont donc généralement au dessus de la moyenne.

Nous cherchons maintenant à savoir les éventuels liens qu’il existe entre les différentes variables à l’aide, pour cela nous calculons les coefficients de corrélation :

avec

sigma=std(Y);

Z=X\*diag(1./sigma);

r1=(1/(n\*sqrt(lambda(1))))\*Z'\*Xstar(:,1);

r2=(1/(n\*sqrt(lambda(2))))\*Z'\*Xstar(:,2);

figure(4);

hold on;

theta=0:0.1:360;

crlx=cos(theta);

crly=sin(theta);

plot(r1,r2,'ro')

axis equal

text(r1+0.05,r2,tab\_2(:,1));

plot(crlx,crly,'black');

grid()

xlabel('e1 (52.05 %)')

ylabel('e2 (30.62 %)')

title('Cercle des corrélations');

De plus dans le code nous affichons le cercle unité pour observer si les variables sont associées au plan factoriel ce qui nous permet d’obtenir la figure suivante :

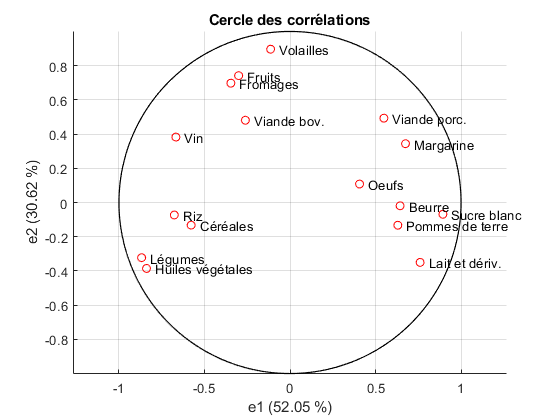


Figure 4: cercle des corrélations pour les deux premiers axes vectoriels

On note tout d’abord que plus une variable statistique et proche du cercle plus les axes principaux choisi ont de l’information sur cette variable. On peut donc voir par exemple que l’axe e1 possède une majorité de l’information sur le sucre blanc. A l’inverse l’œuf est l’aliment le moins représenté par les deux axes.

Du fait du choix des deux premiers axes vectoriels qui contiennent 80% de l’information, les points sont généralement proches du cercle unité.

Nous pouvons interpréter sur cette figure que plus deux points sont proches plus ils sont corrélés en effet par exemple pour les légumes et l’huile végétale si un pays achète l’un alors il y aura de grandes chances d’acheter l’autre. On peut ainsi dire la même chose pour le riz et les céréales ou encore les fruits et le fromage.

De plus si deux points sont proche du cercle mais séparé de 180° l’un de l’autre alors cela veut dire que si un pays achète l’un des produits alors il y a peu de chance qu’il achète l’autre et inversement. Dans notre exemple on a par exemple les légumes ou huiles végétales qui sont opposés à la viande de porc ou margarine.

Enfin si deux points sont proches du cercle mais écarté de 90° alors ces deux points sont décoléré. Par exemple le sucre blanc et la volaille n’ont pas de liens entre eux.

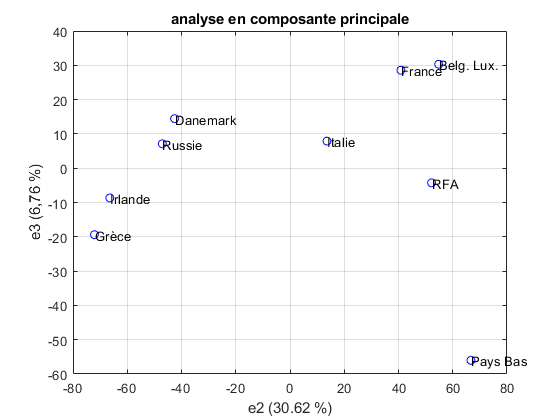


Figure 5 :analyse en composantes principales pour les 2ème et 3ème axes vectoriels

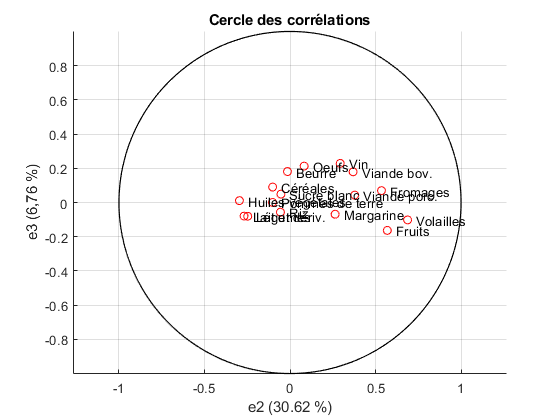


Figure 6: analyse en composantes principales pour les 2ème et 3ème axes vectoriels

On peut bien voir que pour les 2eme et 3eme axes nous avons environs 37% de l’information d’origine. En effet on peut bien voir que les points sont généralement loin du cercle unité ce qui prouve que les axes e2 et e3 possède pas beaucoup d’informations et les résultats peuvent être incohérent. Par exemple l’œuf et le vin sont ici proche alors qu’ils ne le sont pas du tout avec e1 et e2. Cette figure ne représente donc pas la réalité et ne peut être interprété.

# CONCLUSION

Nous avons appliqué la méthode d’analyse en composantes principales sur un tableau avec beaucoup de données qui n’est pas directement interprétable.

Cette méthode nous a permis de réduire à seulement deux variables et ensuite les afficher plus ensuite interpréter.

Bien que cette méthode approxime les données d’origines, dans l’exemple des données sur les habitudes alimentaires on a encore plus de 80% de l’information pour le nombre de variables réduit de 16 à 2.

# INTRODUCTION

# CONCLUSION