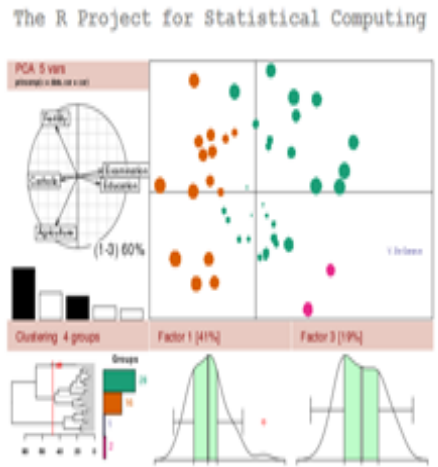
Analyse en Composante Principale (ACP)

HAMDANE Cinda



Sommaire

I. Feedback sur le QCM 1

II. Les étapes en analyse de données

III. Rappels sur l'Analyse en Composante Principale (ACP)

A) Définition

B) Principe

C) L'interprétation

D) Cas pratique

IV. TP 3 sur l'ACP

FEEDDBACK

Graphiques

- ✓ **Plot()** : graphique classique
- ✓ **Barplot()** : graphique en barre
- ✓ **Boxplot()** : Graphique sous forme de Boite à moustache
- ✓ **Hist()** : Histogramme

Fonctions

- ✓ Summary() : permet d'obtenir la description statistique par variable dans un jeu de données
- ✓ Table() : compte le nombre d'occurrence par catégorie
- ✓ Var() : permet de calculer la variance
- ✓ Sd() : permet de calculer l'écart-type

Vous pouvez retrouver les commandes R sur : http://revue.sesamath.net/IMG/pdf/RCarte_Commandes-R.pdf

II. Les étapes en analyse de données

(1) Importation des packages

(2) Chargement des données

(3) Visualisation des données manquantes

(4) Transformation des variables

(5) Statistiques Descriptives

(6) Représentation graphique

(7) Analyse en composante principale (ACP)

II. Les étapes en analyse de données

(1) Importation des packages **(Explication du choix)**

(2) Chargement des données **(+ Extraction d'un sous-ensemble)**

(3) Visualisation des données manquantes **(Elimination ou Imputation)**

(4) Transformation des variables **(factor/ numeric/ character)**

(5) Statistiques **(variables quantitatives/qualitatives/descriptives/croisement)**

(6) Représentation graphique **(Box plot + Normalisation)**

(7) Analyse en Composante Principale (ACP)

III. Analyse en Composante Principale

Qu'est-ce qu'une ACP ?

Quelles sont ces objectifs ?

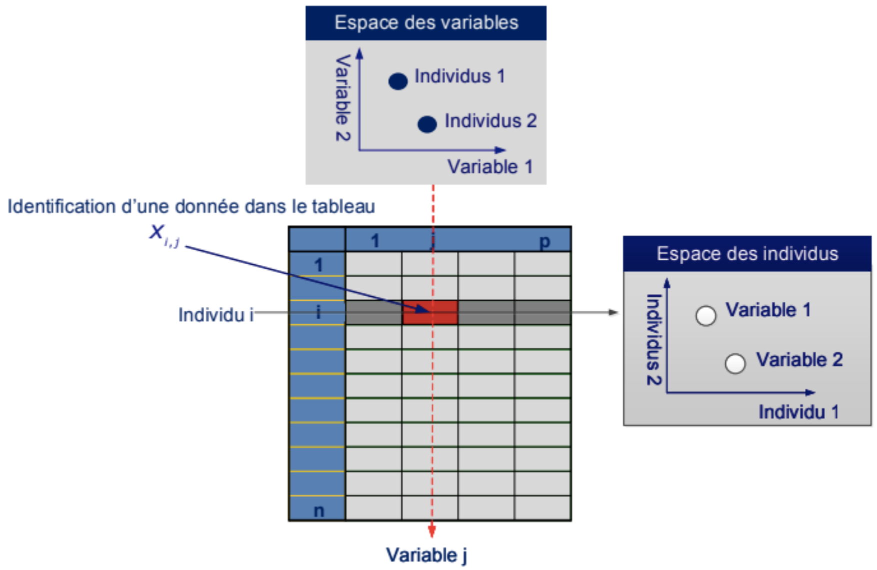
II. Analyse en Composante Principale

A) Définition

- Méthode d'analyse multidimentionnelle des données
- Décrire un jeu de données ac un nb +++ individus & variables quantitatives
- Elle se fait toujours sur <u>les variables quantitatives</u>

II. Analyse en Composante Principale

<u>Données</u> : Si On considère i individus observés sur j variables quantitatives



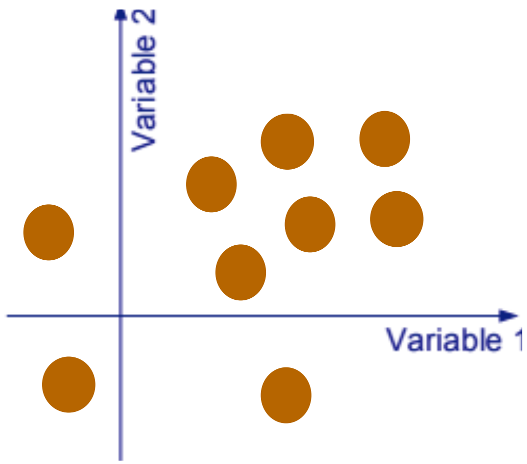
INDIVIDU = Élément de l'espace R^i
VARIABLE = Élément de l'espace R^j

<u>Rq</u> : La représentation des variables et individus est possible qd dim =2 ou dim =3

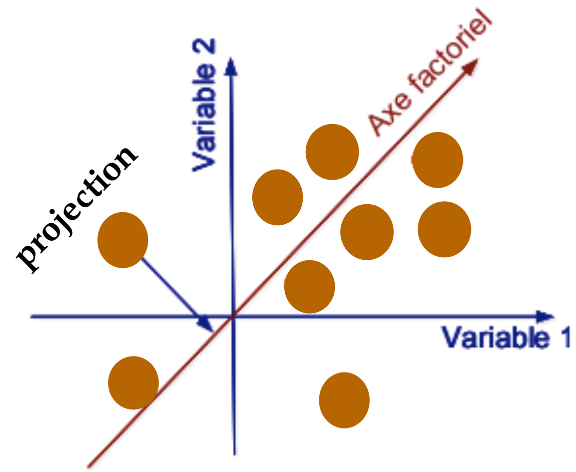
II. Analyse en Composante Principale

B) Principe : Si On considère i individus observés sur j variables quantitatives

→ Sur un plan de dimension 2

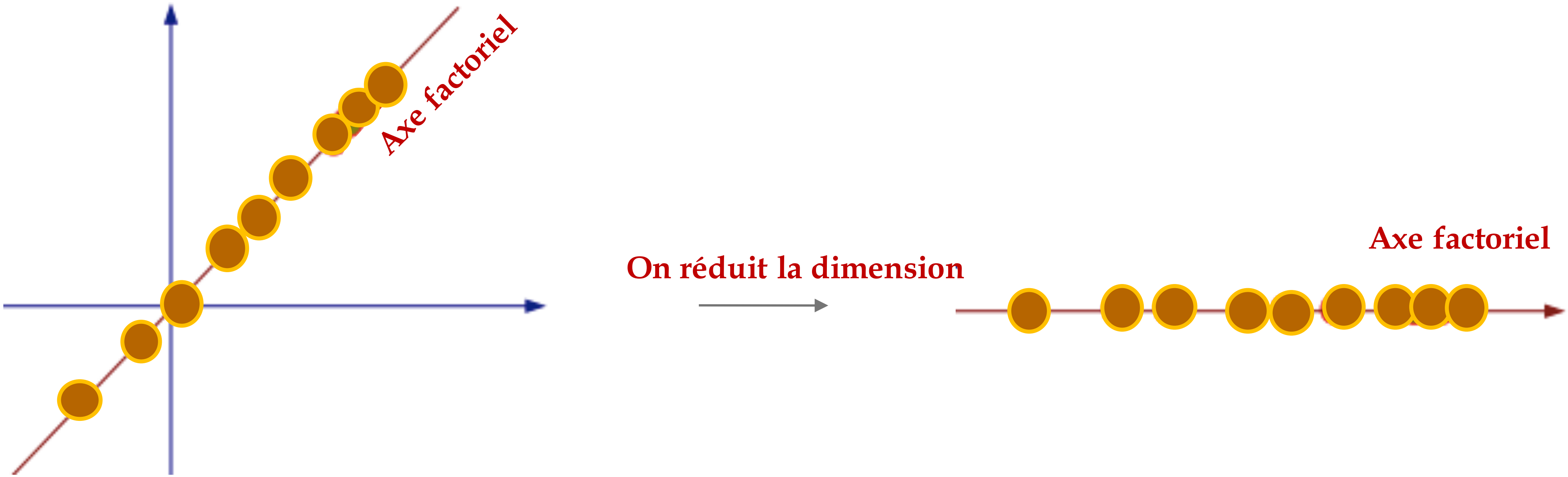


On souhaite représenter tout les individus dans un espace de dimension moindre tout en gardant un maximum d'information



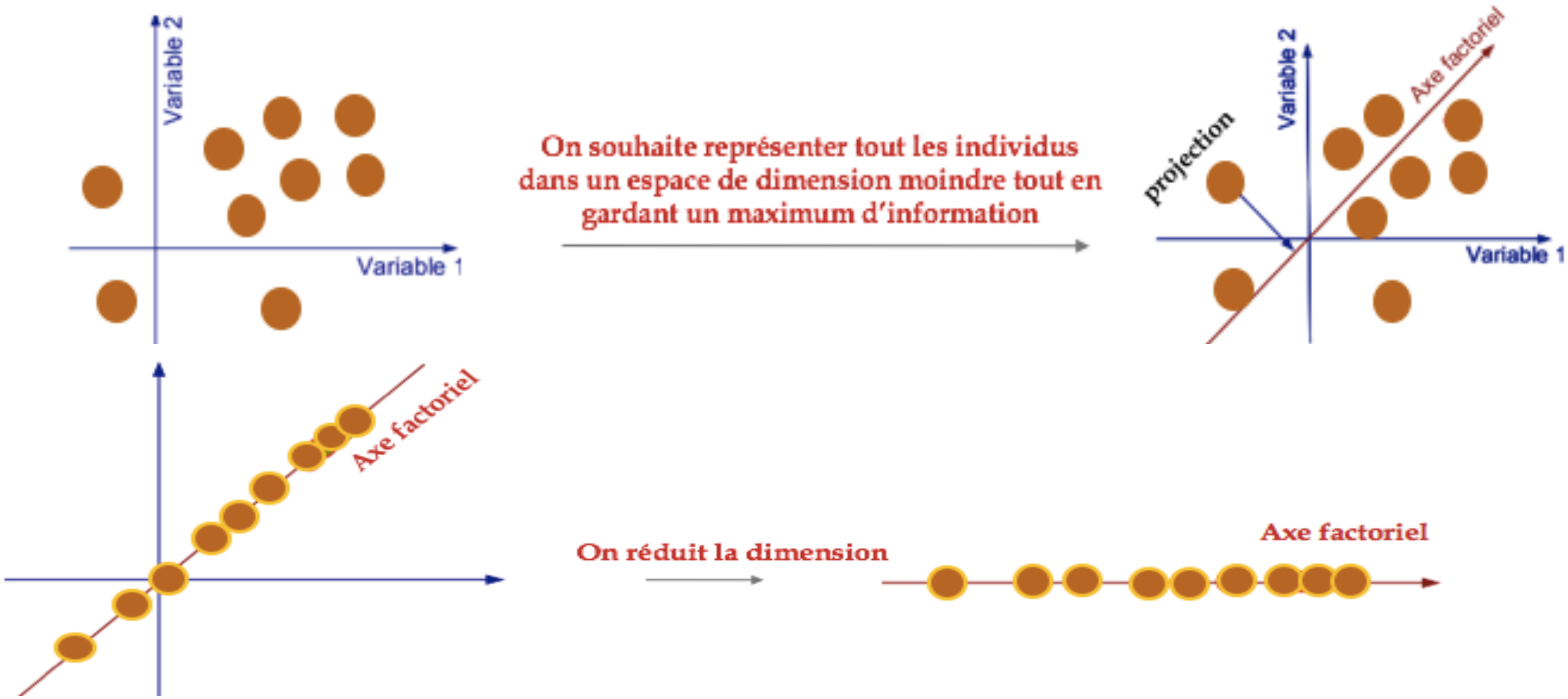
II. Analyse en Composante Principale

B) Principe



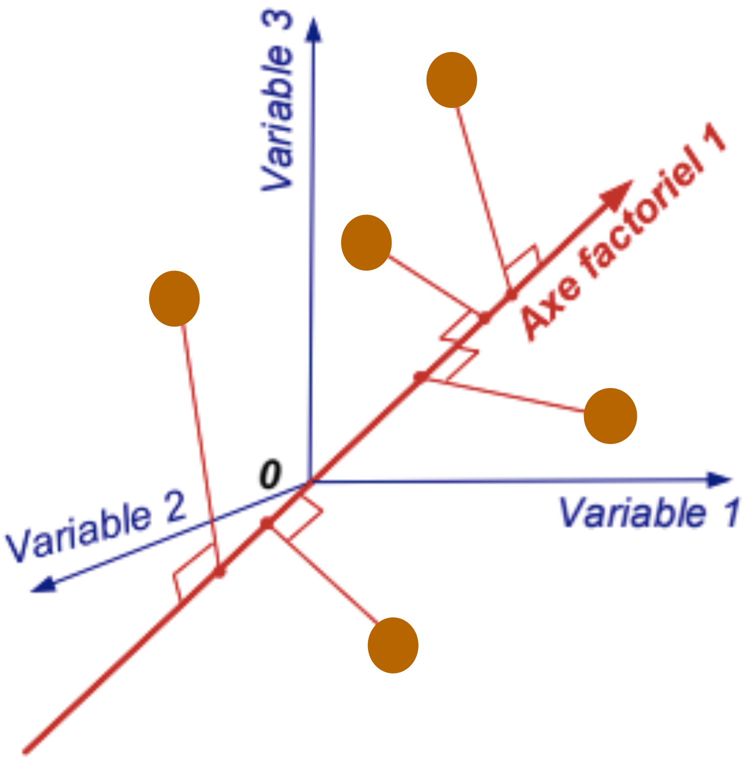
II. Analyse en Composante Principale

B) Principe



II. Analyse en Composante Principale

B) Principe

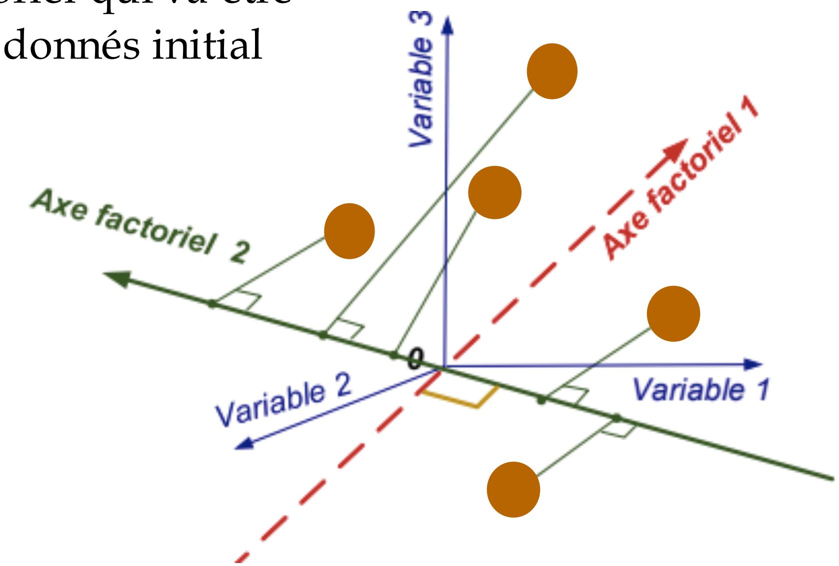


Etape 1 : on recherche le premier axe factoriel qui permettra d'expliquer aux mieux les données de départ

II. Analyse en Composante Principale

B) Principe

Etape 2 jusqu'à n : on détermine le second axe factoriel qui va être perpendiculaire au premier et qui décrit au mieux le jeu de donnés initial



II. Analyse en Composante Principale

B) Principe

- On ne peut pas représenter graphiquement si $p > 3$

- Nous allons rechercher des **axes factoriels/principaux** qui représentent des « **combinaisons linéaires** » des var. Initiales ➡ Permet ainsi de décrire l'ensemble des individus en tenant compte de l'ensemble des variables

- Ces axes factoriels sont des projections

- Nous allons rechercher des axes de projection des points qui permettront d'obtenir la meilleure « **VISUALISATION** » du nuage de points dans des espaces de + faibles dimensions

- Les axes qui seront étudiés deux à deux

II. Analyse en Composante Principale

Soit un tableau à p variables et à n individus

- Recherche des « espaces de dimensions plus petites » dans lesquels il est possible d'observer au mieux les variables et les individus

- **Impossible de visualiser dès que $p > 3$**

- **L'ACP permet de réduire la dimentionnalité des données tout en gardant le maximum d'information**

II. Analyse en Composante Principale

C) Interprétation des résultats

- ✓ Comment les variables sont-elles structurées entre elles ?

  (**Corrélation** entre les variables) = liaison entre les variables

  Quelles sont les variables qui sont associées ou non ?

  Quelles sont les variables qui vont dans le même sens ?

  Quelles sont les variables qui vont dans des sens opposés ?

- ✓ Est-ce que les individus se ressemblent ?

  (**Notion de distance** entre individus)

  Comment est la répartition des individus qui se ressemblent ou bien qui sont dissemblables ?

➡ **Exploration les liaisons entre variables et les ressemblances entre individus**

Les étapes en ACP

(1) Calculer la matrice de corrélation

(2) Appliquer la fonction ACP sur les données normées

- ✓ Calculer l'inertie
- ✓ Déterminer les valeurs propres et les vecteurs propres
- ✓ Représentation de la projection sur le plan principal

(3) Interprétations

Exemple introductif

- C’est une enquête sur la consommation de produit alimentaire par les ménages selon les catégories socio-professionnelles
- **Les individus ?**
- **Les variables ?**

	pain	fruit	viande	volaille	lait	légume	alcool
MA2	332	354	1437	526	247	428	427
EM2	293	388	1527	567	239	559	258
CA2	372	562	1948	927	235	767	433
MA3	406	341	1507	544	324	563	407
EM3	386	396	1501	558	319	608	363
CA3	438	689	2345	1148	243	843	341
MA4	534	367	1620	638	414	660	407
EM4	460	484	1856	762	400	699	416
CA4	385	621	2366	1149	304	789	282
MA5	655	423	1848	759	495	776	486
EM5	584	548	2056	893	518	995	319
CA5	515	887	2630	1167	561	1097	284

Exemple introductif

- C’est une enquête sur la consommation de produit alimentaire par les ménages selon les catégories sociaux-professionnelles

- Les individus sont représentés par **Catégories socio-professionnelles par Nb d’enfants**

  EM : Employés **MA :** Travailleurs manuels **CA :** Cadres

- Les variables sont représentées par **Indices des dépenses annuelles selon les différents types d’aliments**

	pain	fruit	viande	volaille	lait	légume	alcool
MA2	332	354	1437	526	247	428	427
EM2	293	388	1527	567	239	559	258
CA2	372	562	1948	927	235	767	433
MA3	406	341	1507	544	324	563	407
EM3	386	396	1501	558	319	608	363
CA3	438	689	2345	1148	243	843	341
MA4	534	367	1620	638	414	660	407
EM4	460	484	1856	762	400	699	416
CA4	385	621	2366	1149	304	789	282
MA5	655	423	1848	759	495	776	486
EM5	584	548	2056	893	518	995	319
CA5	515	887	2630	1167	561	1097	284

Exemple introductif

- C'est une enquête sur la consommation de produit alimentaire par les ménages selon les catégories socio-professionnelles

<u>Statistiques :</u> La fonction summary() permet d'obtenir la description statistique du jeu/tableau de données.

Pour une variable donnée, **la fonction** renvoie 5 valeurs :

- le minimum (Min.)	le premier quartile (1st Qu.)	la médiane (Median)
- la moyenne (Mean)	le troisième quartile (3rd Qu.)	le maximum (Max)

	pain	fruit	viande	volaille	lait	legume	alcool
Min.	293	341	1437	526	235	428	258
1st Qu.	381.8	382.8	1522	564.8	246	596.8	310.2
Median	422	453.5	1852	760.5	321.5	733	385
Mean	446.7	505	1887	803.2	358.2	732	368.6
3rd Qu.	519.8	576.8	2128	982.2	434.2	802.5	418.8
Max.	655	887	2630	1167	561	1097	486
sd	107.148	165.092	395.75	249.561	117.127	189.18	71.7818

Exemple introductif

➢ **<u>Matrice des corrélations :</u>** permet d'évaluer la dépendance entre plusieurs variables c'est-à-dire prises deux à deux

➢ Plus le coefficient est proche des valeurs -1 et 1 et plus la corrélation linéaire entre les variables est forte

	pain	fruit	viande	volaille	lait	legume	alcool
pain	1	0.19614	0.32127	0.24801	0.85557	0.59311	0.30376
fruit	0.19614	1	0.95948	0.92554	0.33219	0.85625	-0.4863
viande	0.32127	0.95948	1	0.98179	0.37459	0.88108	-0.4372
volaille	0.24801	0.92554	0.98179	1	0.23289	0.82678	-0.4002
lait	0.85557	0.33219	0.37459	0.23289	1	0.6628	0.00688
legume	0.59311	0.85625	0.88108	0.82678	0.6628	1	-0.3565
alcool	0.30376	-0.4863	-0.4372	-0.4002	0.00688	-0.3565	1

(1) Comment les variables sont liées entre elles ?

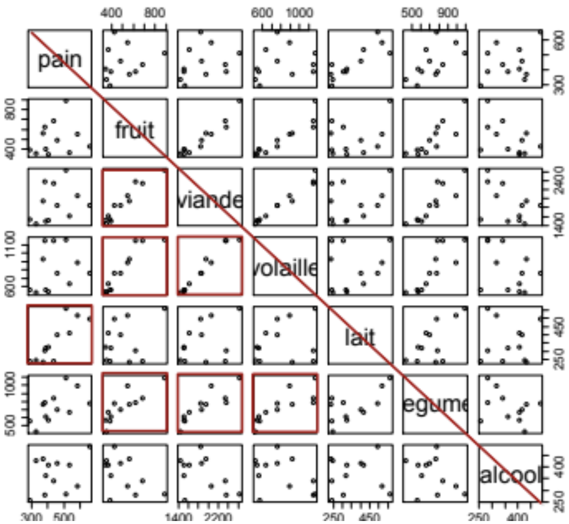
(2) Quel est le comportement des différentes catégories socio-professionnelles entre elles ?

(3) Comment se comportent les diverses catégories socio-professionnelles par rapport à la consommation alimentaire ?

Exemple introductif

- **<u>Matrice des corrélations :</u>** matrice des coefficients calculés sur plusieurs variables prises deux à deux
- Plus le coefficient est proche des valeurs -1 et 1 et plus la corrélation linéaire entre les variables est forte
- Le diagramme de dispersion est un outil graphique utile pour visualiser la présence d'une corrélation entre deux variables

	pain	fruit	viande	volaille	lait	legume	alcool
pain	1	0.19614	0.32127	0.24801	0.85557	0.59311	0.30376
fruit	0.19614	1	0.95948	0.92554	0.33219	0.85625	-0.4863
viande	0.32127	0.95948	1	0.98179	0.37459	0.88108	-0.4372
volaille	0.24801	0.92554	0.98179	1	0.23289	0.82678	-0.4002
lait	0.85557	0.33219	0.37459	0.23289	1	0.6628	0.00688
legume	0.59311	0.85625	0.88108	0.82678	0.6628	1	-0.3565
alcool	0.30376	-0.4863	-0.4372	-0.4002	0.00688	-0.3565	1



Que peut-on dire de ces résultats ?

Cas pratique 1

- Nous allons travailler sur le jeu de données **mtcars** : variables contenant l'aspect et les performances d'un ensemble de 32 voitures (modèles entre 1973-1974)

1) Afficher les statistiques descriptives puis le Boxplot des données : ***Que remarquez-vous ?***

2) Afficher le diagramme de dispersion : ***Que pouvez-vous dire ?***

3) Calculer la matrice de corrélation ***à l'aide de la fonction corr***

4) Afficher la matrice de corrélation ***à l'aide du package corrplot et de la fonction corrplot***

5) Interprétez les résultats

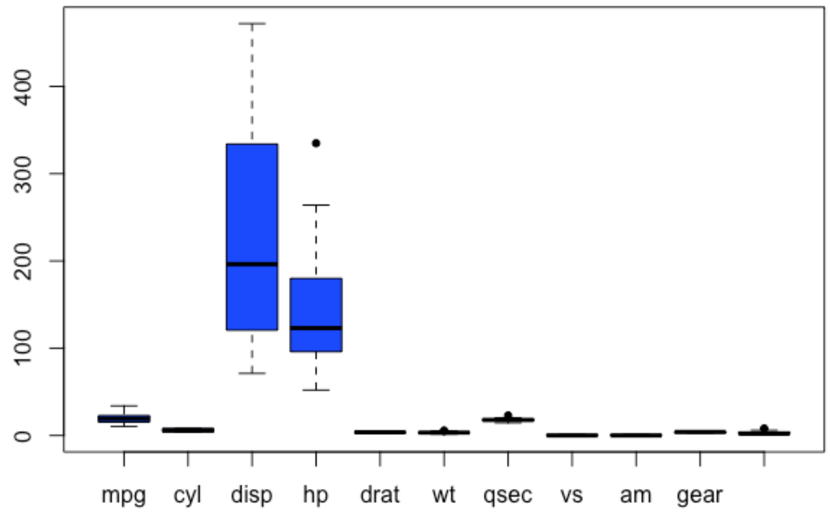
Cas pratique 1

- Nous allons travailler sur le jeu de données **mtcars** : variables contenant l'aspect et les performances d'un ensemble de 32 voitures (modèles entre 1973-1974)

1) Afficher les statistiques descriptives puis le Boxplot des données mtcars

Que remarquez-vous ?

```
> sapply(mtcars, summary)
              mpg    cyl     disp       hp     drat      wt     qsec     vs      am   gear   carb
Min.     10.40000 4.0000  71.1000  52.0000 2.760000 1.51300 14.50000 0.0000 0.00000 3.0000 1.0000
1st Qu.  15.42500 4.0000 120.8250  96.5000 3.080000 2.58125 16.89250 0.0000 0.00000 3.0000 2.0000
Median   19.20000 6.0000 196.3000 123.0000 3.695000 3.32500 17.71000 0.0000 0.00000 4.0000 2.0000
Mean     20.09062 6.1875 230.7219 146.6875 3.596563 3.21725 17.84875 0.4375 0.40625 3.6875 2.8125
3rd Qu.  22.80000 8.0000 326.0000 180.0000 3.920000 3.61000 18.90000 1.0000 1.00000 4.0000 4.0000
Max.     33.90000 8.0000 472.0000 335.0000 4.930000 5.42400 22.90000 1.0000 1.00000 5.0000 8.0000
```



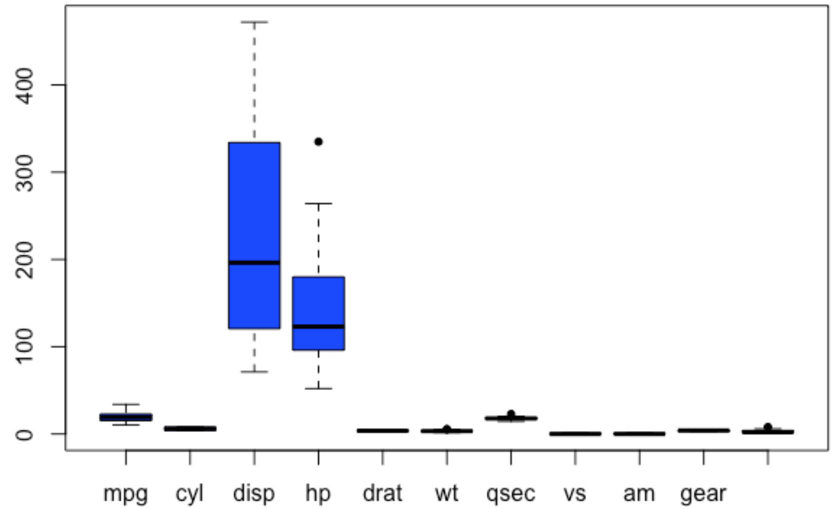
→ On observe seulement la boite à moustache pour les variables disp et hp

→ Les boites à moustaches pour les variables mpg, cyl, drat, wt, qsec, vs, am, gear ne sont pas visibles

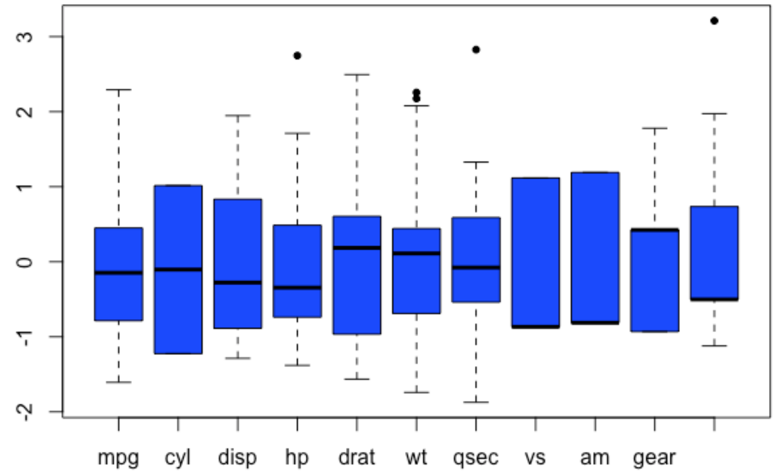
Normalisation des données

- Pour pouvoir comparer les variables entre elles, il est nécessaire de normaliser les données
  = c'est-à-dire centrer les données par la moyenne et réduire les données par l'écart-type

- **ACP est indépendante des métriques**



$$z_{i,j} = \frac{x_{i,j} - E(x_{.j})}{\sqrt{V(x_{.j})}}$$

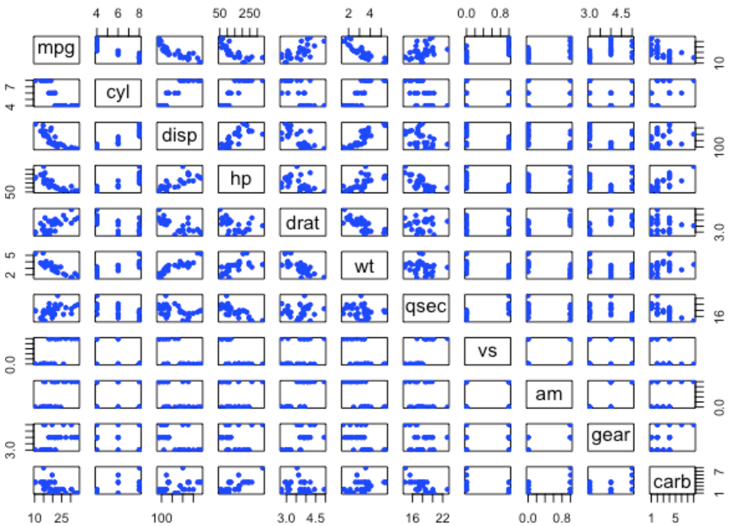


Cas pratique

- Nous allons travailler sur le jeu de données **mtcars** : variables contenant l'aspect et les performances d'un ensemble de 32 voitures (modèles entre 1973-1974)

1) Afficher les statistiques descriptives puis le Boxplot des données : ***Que remarquez-vous ?***

2) Afficher le diagramme de dispersion : ***Que pouvez-vous dire ?***



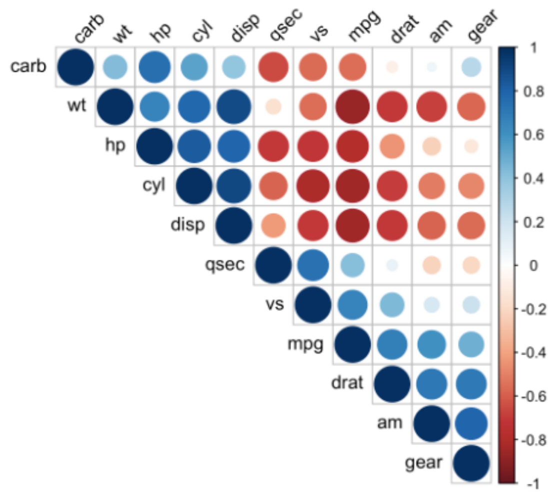
Cas pratique

- Nous allons travailler sur le jeu de données **mtcars** : variables contenant l'aspect et les performances d'un ensemble de 32 voitures (modèles entre 1973-1974)

3) Calculer la matrice de corrélation : **corr(mtcars)**

4) Afficher le graphique de matrice de corrélation à l'aide de la fonction corrplot :
library(corrplot) corrplot(matrice_cor, type="upper", order="hclust", tl.col="black", tl.srt=45)

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1.00	-0.85	-0.85	-0.78	0.68	-0.87	0.42	0.66	0.60	0.48	-0.55
cyl	-0.85	1.00	0.90	0.83	-0.70	0.78	-0.59	-0.81	-0.52	-0.49	0.53
disp	-0.85	0.90	1.00	0.79	-0.71	0.89	-0.43	-0.71	-0.59	-0.56	0.39
hp	-0.78	0.83	0.79	1.00	-0.45	0.66	-0.71	-0.72	-0.24	-0.13	0.75
drat	0.68	-0.70	-0.71	-0.45	1.00	-0.71	0.09	0.44	0.71	0.70	-0.09
wt	-0.87	0.78	0.89	0.66	-0.71	1.00	-0.17	-0.55	-0.69	-0.58	0.43
qsec	0.42	-0.59	-0.43	-0.71	0.09	-0.17	1.00	0.74	-0.23	-0.21	-0.66
vs	0.66	-0.81	-0.71	-0.72	0.44	-0.55	0.74	1.00	0.17	0.21	-0.57
am	0.60	-0.52	-0.59	-0.24	0.71	-0.69	-0.23	0.17	1.00	0.79	0.06
gear	0.48	-0.49	-0.56	-0.13	0.70	-0.58	-0.21	0.21	0.79	1.00	0.27
carb	-0.55	0.53	0.39	0.75	-0.09	0.43	-0.66	-0.57	0.06	0.27	1.00



Rappels ACP

- On s'intéresse à :
  - Comment les variables sont structurées entre elles ? Notion **de liaisons entre les variables**
  - La ressemblance entre les individus : **Distance entre les individus**

  → **POSITIONNEMENT des variables et des individus**

- L'ACP est une méthode géométrique dont le positionnement des variables et des individus sur les plans factoriels se fera à partir du calcul de la distance euclidienne

- Initialement, les données sont dépendantes de leur mesure et elles ont donc des poids différents → **NORMALISATION**

- Les données doivent apporter la « **même quantité d'information** » : normalisation/standardisation des données

  = modification de l'échelle **MAIS** pas de modification de la structure des données

Les étapes en ACP

(1) Calculer la matrice des corrélation

(2) Appliquer la fonction ACP sur les données normées

- ✓ Déterminer les valeurs propres
- ✓ Calculer l'inertie
- ✓ Représentation de la projection sur le plan principal

(3) Interprétations

Les étapes en ACP

(1) Calculer la matrice des corrélation

(2) Appliquer la fonction ACP sur les données normées

```
PCA(data_frame, scale.unit=T, graph=T)
```

Arguments :

- ′data' : données sur laquelle on va faire l'ACP
- 'scale.unit' : Permet de normaliser les données si cela n'est pas déjà fait
- 'quali.sup' : vecteur indiquant les index des variables supplémentaires qualitatives
- 'graph' : ici True pour l'afficher

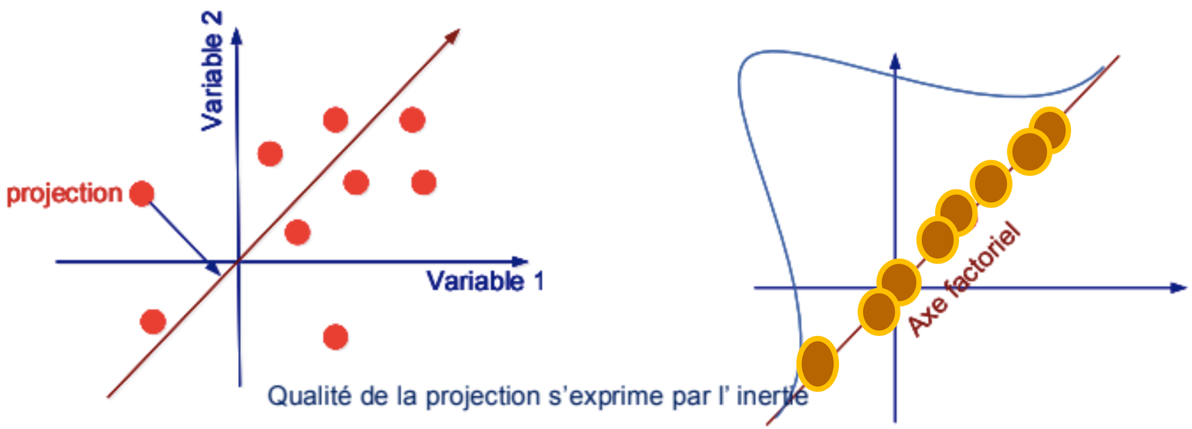
Les étapes en ACP

(1) Calculer la matrice des corrélation

(2) Appliquer la fonction ACP sur les données normées

Calculer l'inertie

✓ La moyenne de la dispersion des points représente la variance



✓ L'inertie est le pourcentage de variance expliquée par un axe factoriel

✓ Il s'agit donc de la quantité d'information exprimée par un axe

Les étapes en ACP

(1) Calculer la matrice des corrélation

(2) Appliquer la fonction ACP sur les données normées

Valeurs propres = **Correspond à la variance totale expliquée par chaque axe/composante**

Vecteurs propres = **Correspond aux coefficients factoriels**

Calculer l'inertie = **C'est donc la quantité d'information fournie par les axes principaux/factorielles**

Les étapes en ACP

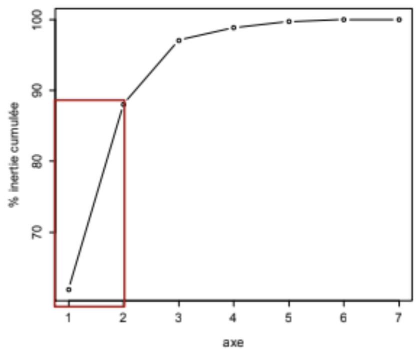
(1) Calculer la matrice des corrélation

(2) Appliquer la fonction ACP sur les données normées

Calculer l'inertie

= C'est donc la quantité d'information fournie par les axes principaux/factorielles

axe	% inertie cumulée
1	61.9
2	88.1
3	97.1
4	98.9
5	99.7
6	100.0
7	100.0



✓ Les deux premiers axes «représentent » 88.1 % de la quantité d'information initiale (ensemble des observations)

✓ Passage de 11 variables à deux composantes (axes principaux) qui seront à interpréter selon les variables et les individus

*Source : CM Bertrand Roudier

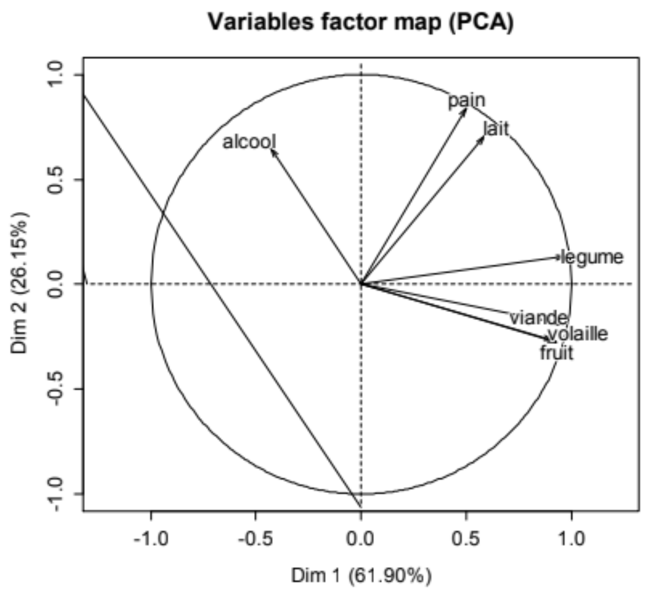
Les étapes en ACP

(2) Appliquer la fonction ACP sur les données normées

Coordonnées

Représentation des variables dans le plan factoriel (axe 1- axe 2)

	axes 1	axes 2	axes 3	axes 4	axes 5
pain	0.50	0.84	-0.01	-0.19	0.01
fruit	0.93	-0.28	0.12	0.20	-0.02
viande	0.96	-0.19	0.16	-0.02	0.10
volaille	0.91	-0.27	0.28	-0.12	0.05
lait	0.58	0.71	-0.35	0.16	0.08
legume	0.97	0.13	-0.05	-0.01	-0.19
alcool	-0.43	0.65	0.62	0.11	-0.02

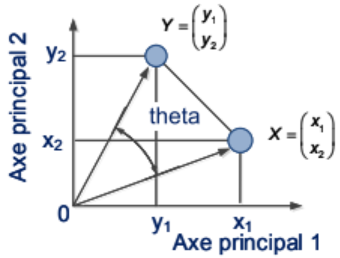


*Source : CM Bertrand Roudier

Les étapes en ACP

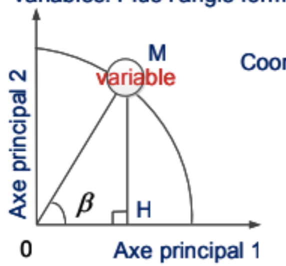
(2) Appliquer la fonction ACP sur les données normées

Qualité de représentation des variables



$$\cos(\theta) = \frac{\langle X,Y \rangle}{\|X\|\|Y\|} \longrightarrow \frac{x_1y_1 + x_2y_2}{\sqrt{x_1^2 + x_2^2}\sqrt{y_1^2 + y_2^2}} = \frac{\sum_{i=1}^{2} x_i y_i}{\sqrt{\sum_{i=1}^{2} x_i^2}\sqrt{\sum_{i=1}^{2} y_i^2}} = r$$

Le cosinus de l'angle formé par les vecteurs V1 et V2 correspond à la **corrélation** entre les deux variables. Plus l'angle formé entre deux variables est « petit », meilleure sera la corrélation



Coordonnées des points sur l'axe principal $\cos\beta = \frac{OH}{OM}$

$\cos^2\beta = OH^2$ est appelée qualité de représentation

Plus l'angle béta est faible, meilleure est la représentation de la variable sur l'axe factoriel

*Source : CM Bertrand Roudier

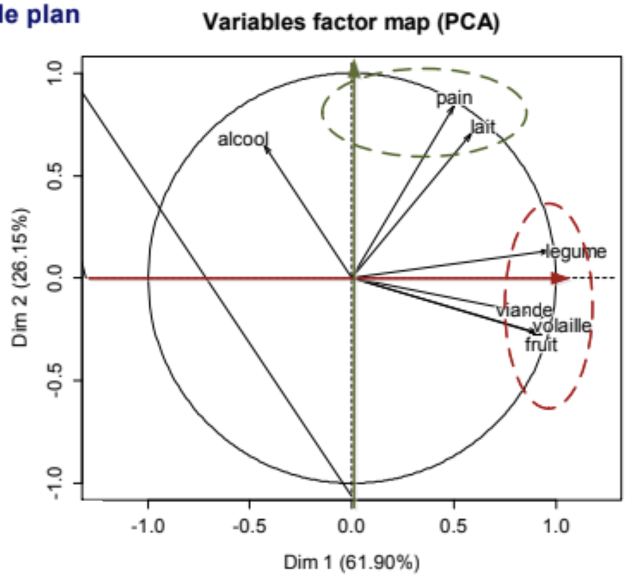
Les étapes en ACP

(2) Appliquer la fonction ACP sur les données normées

Qualité de représentation des variables

	axe 1	Axe 2	Sum qtl
pain	0.25	0.71	0.96
fruit	0.86	0.08	0.94
viande	0.93	0.04	0.96
volaille	0.83	0.07	0.90
lait	0.34	0.50	0.84
legume	0.94	0.02	0.96
alcool	0.18	0.42	0.60

le plan



<u>Exemple variable légume :</u>

- ✓ La qualité de représentation est de 94% sur l'axe 1 et de 2% sur l'axe 2
- ✓ Dans le plan, la qualité de représentation est de 96%
- ✓ Cette variable permet d'interpréter avec d'autres variable (viande 93% volaille 83% fruit 86%) l'axe 1

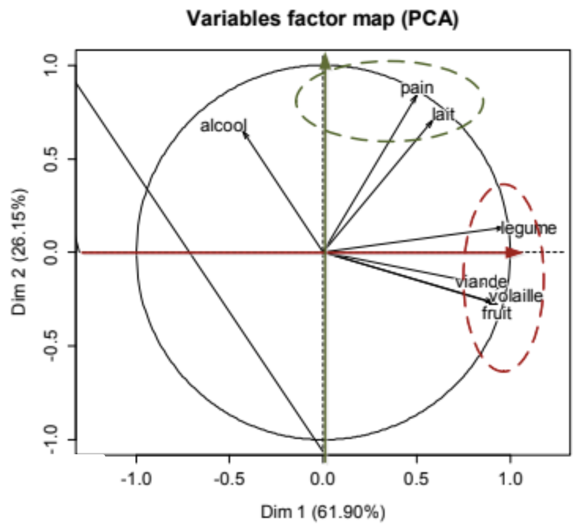
*Source : CM Bertrand Roudier

Les étapes en ACP

(2) Appliquer la fonction ACP sur les données normées

Qualité de représentation des variables

	axe 1	Axe 2	Sum qtl
pain	0.25	0.71	0.96
fruit	0.86	0.08	0.94
viande	0.93	0.04	0.96
volaille	0.83	0.07	0.90
lait	0.34	0.50	0.84
legume	0.94	0.02	0.96
alcool	0.18	0.42	0.60



Interprétations

- ✓ L'axe 1 va opposer les catégories socio-professionnelles qui consomment de manière préférentielle ces aliments (légume, viande, volaille, fruit) à ceux qui en consomment peu ou pas
- ✓ L'axe 2 opposent les catégories socio-professionnelle qui consomment préférentiellement du lait et du pain à ceux qui n'en consomment peu ou pas
- ✓ **L'interprétation des deux axe est indépendante l'une de l'autre**

*Source : CM Bertrand Roudier

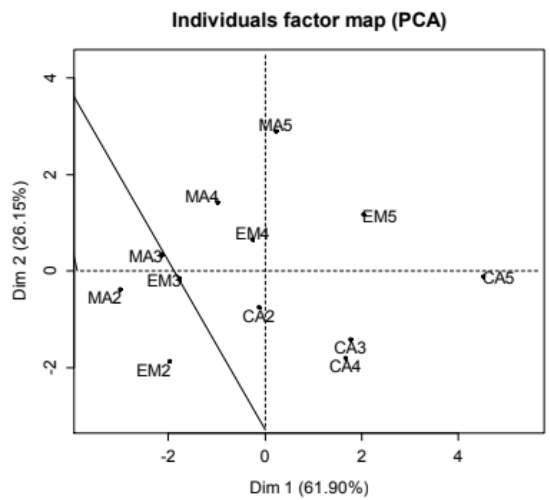
Les étapes en ACP

(2) Appliquer la fonction ACP sur les données normées

Coordonnées & Qualité de représentation des individus sur le plan factoriel

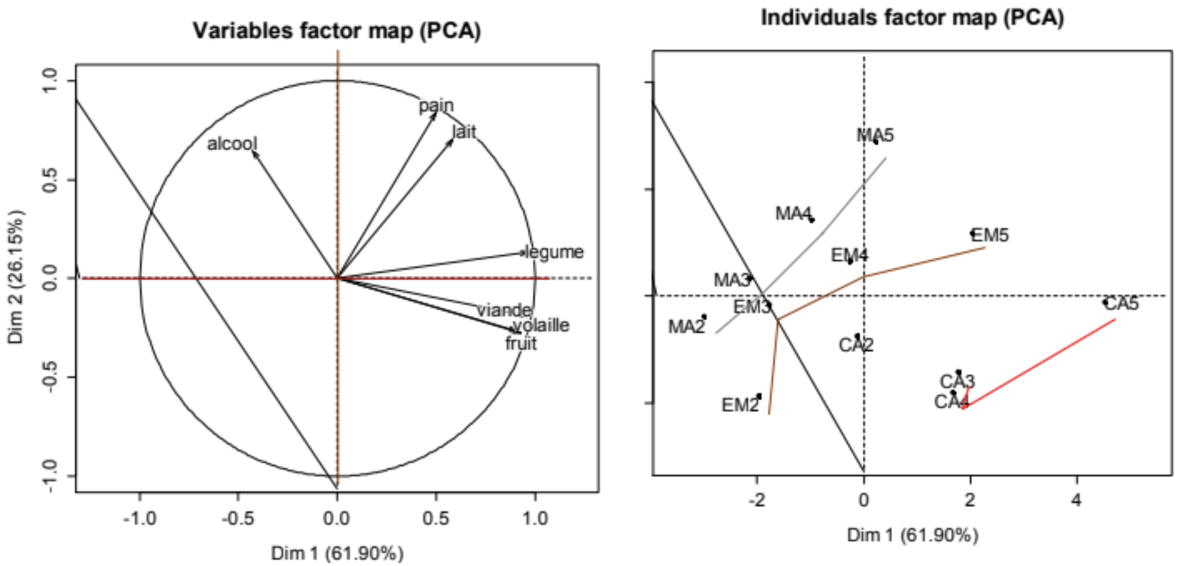
	axes 1	axes 2
MA2	-2.99	-0.38
EM2	-1.97	-1.87
CA2	-0.12	-0.76
MA3	-2.13	0.34
EM3	-1.77	-0.17
CA3	1.77	-1.42
MA4	-0.97	1.43
EM4	-0.26	0.66
CA4	1.67	-1.81
MA5	0.23	2.90
EM5	2.04	1.18
CA5	4.51	-0.11

	Dim.1	Dim.2	Somme
MA2	0.94	0.02	0.96
EM2	0.42	0.38	0.80
CA2	0.00	0.19	0.19
MA3	0.97	0.02	0.99
EM3	0.89	0.01	0.90
CA3	0.48	0.31	0.79
MA4	0.30	0.65	0.94
EM4	0.10	0.61	0.70
CA4	0.43	0.50	0.93
MA5	0.01	0.94	0.95
EM5	0.60	0.20	0.81
CA5	0.96	0.00	0.96



Les étapes en ACP

(3) Interprétations



✓ Les catégories socio-professionnelles ont un comportement différent selon leur consommation en produit alimentaire

✓ Pour chaque classe, l'évolution de la consommation en produit alimentaire est fonction du nombre d'enfants

Cas pratique 2

- Nous allons travailler sur le jeu de données **mtcars** : variables contenant l'aspect et les performances d'un ensemble de 32 voitures (modèles entre 1973-1974)

6) Appliquer l'ACP sur le data frame « mtcars »

7) Afficher les valeurs propres

8) Calculer le % d'inertie : **Que peut-on dire de ce résultat ? Quelle est la quantité d'information conservée ?**

9) Afficher la représentation des variables sur l'axe 1, l'axe 2 et le plan factoriel (coordonnées & qualité)

10) Représenter les individus sur le plan factoriel (coordonnées & qualité)

11) Analyser les résultats obtenues à l'aide de l'ACP

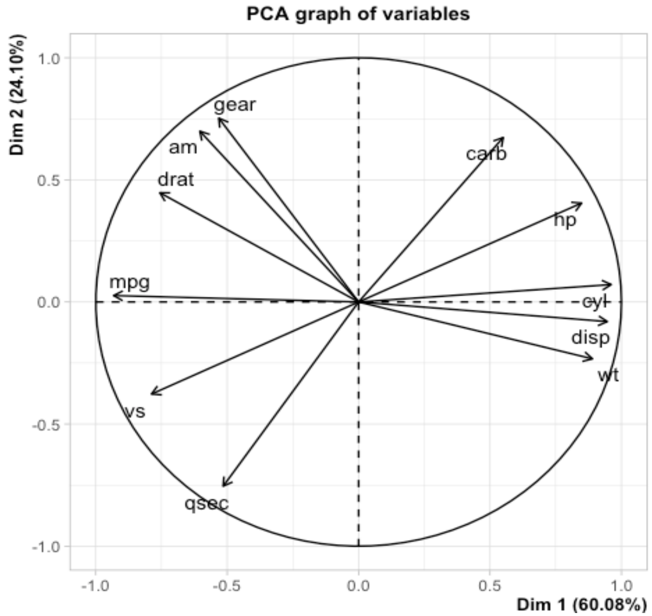
Cas pratique 2

- Nous allons travailler sur le jeu de données **mtcars** : variables contenant l'aspect et les performances d'un ensemble de 32 voitures (modèles entre 1973-1974)

6) Appliquer l'ACP sur le data frame « mtcars »

```
> PCA(mtcars, scale.unit=T, graph=T)
**Results for the Principal Component Analysis (PCA)**
The analysis was performed on 32 individuals, described by 11 variables
*The results are available in the following objects:
```

```
res.pca2 = PCA(mtcars, scale.unit=T, graph=T)
```



Cas pratique 2

- Nous allons travailler sur le jeu de données **mtcars** : variables contenant l'aspect et les performances d'un ensemble de 32 voitures (modèles entre 1973-1974)

7) Afficher les valeurs propres

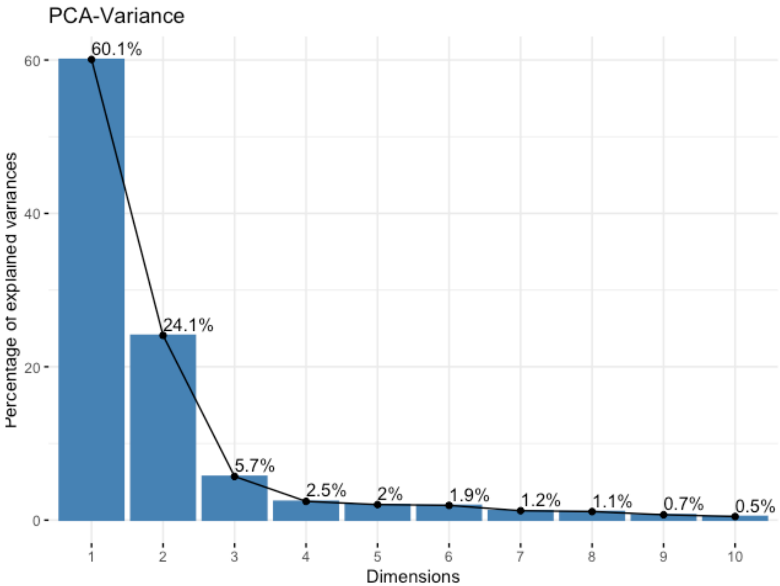
= Correspond à la variance totale expliquée par chaque axe/composante

```
> res.pca2$eig
        eigenvalue percentage of variance cumulative percentage of variance
comp 1  6.60840025             60.0763659                          60.07637
comp 2  2.65046789             24.0951627                          84.17153
comp 3  0.62719727              5.7017934                          89.87332
comp 4  0.26959744              2.4508858                          92.32421
comp 5  0.22345110              2.0313737                          94.35558
comp 6  0.21159612              1.9236011                          96.27918
comp 7  0.13526199              1.2296544                          97.50884
comp 8  0.12290143              1.1172858                          98.62612
comp 9  0.07704665              0.7004241                          99.32655
comp 10 0.05203544              0.4730495                          99.79960
comp 11 0.02204441              0.2004037                         100.00000
```

Cas pratique 2

- Nous allons travailler sur le jeu de données **mtcars** : variables contenant l'aspect et les performances d'un ensemble de 32 voitures (modèles entre 1973-1974)

8) Calculer le % d'inertie : **Que peut-on dire de ces résultats ?**



EBOULI DES VALEURS PROPRES

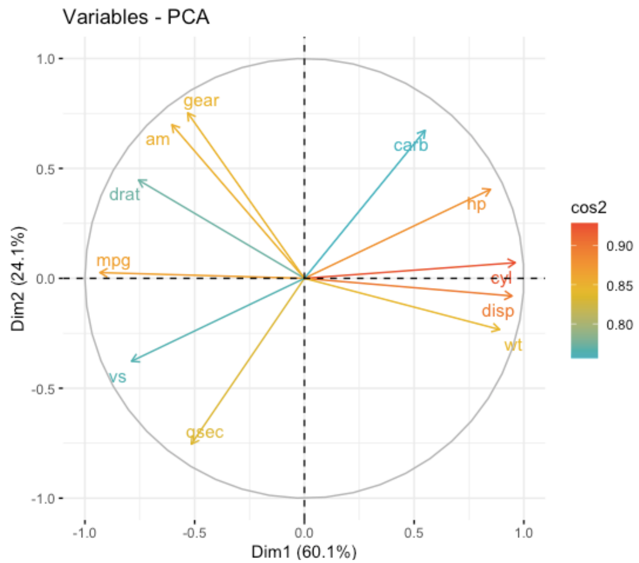
- ✓ Le premier axe conserve 60,1% de l'inertie du nuage. Le second axe conserve une part importante de l'inertie totale soit 24,1 %
- ✓ Les deux premiers axes permettent de représenter » 84.1 % de la quantité d'information initiale (ensemble des observations)
- ✓ Observation d'une chute « **effet coude** » qui est importante dès le 3ème axe
- ✓ **Nous sommes donc passé de 11 variables à deux composantes (axes principaux) qui seront à interpréter selon les variables et les individus**

Cas pratique 2

- Nous allons travailler sur le jeu de données **mtcars** : variables contenant l'aspect et les performances d'un ensemble de 32 voitures (modèles entre 1973-1974)

9) Afficher la représentation des variables sur l'axe 1, l'axe 2 et le plan factoriel (coordonnées & qualité)

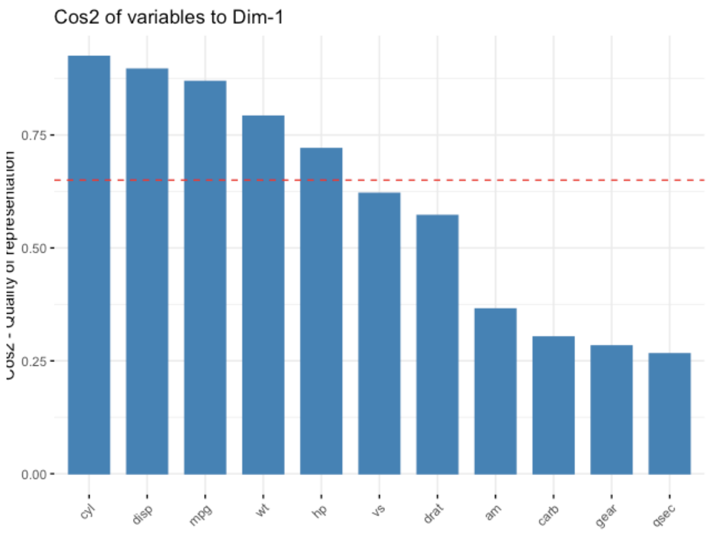
```
> res.pca2$var$coord
          Dim.1       Dim.2       Dim.3        Dim.4       Dim.5
mpg  -0.9319502  0.02625094 -0.17877989 -0.011703525  0.04861531
cyl   0.9612188  0.07121589 -0.13883907 -0.001345754  0.02764565
disp  0.9464866 -0.08030095 -0.04869285  0.133237930  0.18624400
hp    0.8484710  0.40502680  0.11088579 -0.035139337  0.25528375
drat -0.7561693  0.44720905  0.12765473  0.443850788  0.03655308
wt    0.8897212 -0.23286996  0.27070586  0.127677743 -0.03546673
qsec -0.5153093 -0.75438614  0.31929289  0.035347223 -0.07783859
vs   -0.7879428 -0.37712727  0.33960355 -0.111555360  0.28340603
am   -0.6039632  0.69910300 -0.16295845 -0.015817187  0.04244016
gear -0.5319156  0.75271549  0.22949350 -0.137434658  0.02284570
carb  0.5501711  0.67330434  0.41858505 -0.065832458 -0.17079759
```

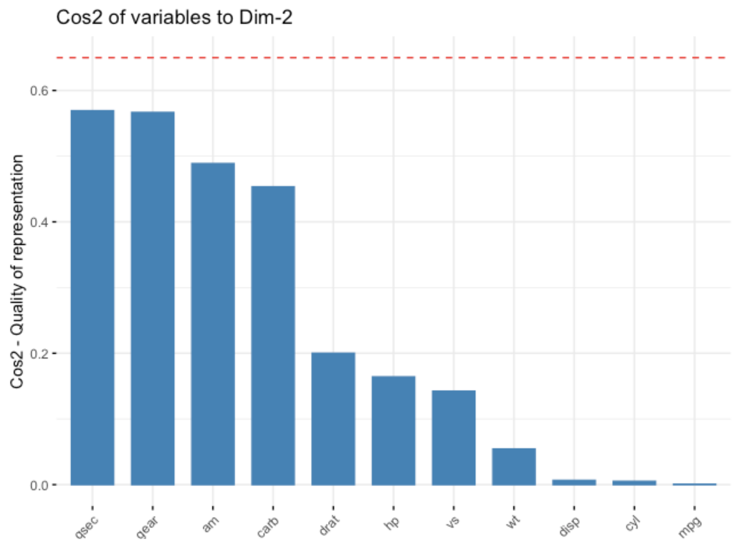


Cas pratique

9) Afficher la représentation des variables sur le plan factoriel (coordonnées & qualité)

```
> #3.7 Qualité de projection des variables sur les axes
> fviz_cos2(res.pca2, choice = "var", axes = 1) +
+     geom_hline (yintercept = 0.65, linetype = 2, color = "red") # axe
>
> #3.7 Qualité de projection des variables sur les axes
> fviz_cos2(res.pca2, choice = "var", axes = 2) +
+     geom_hline (yintercept = 0.65, linetype = 2, color = "red") # axe
```

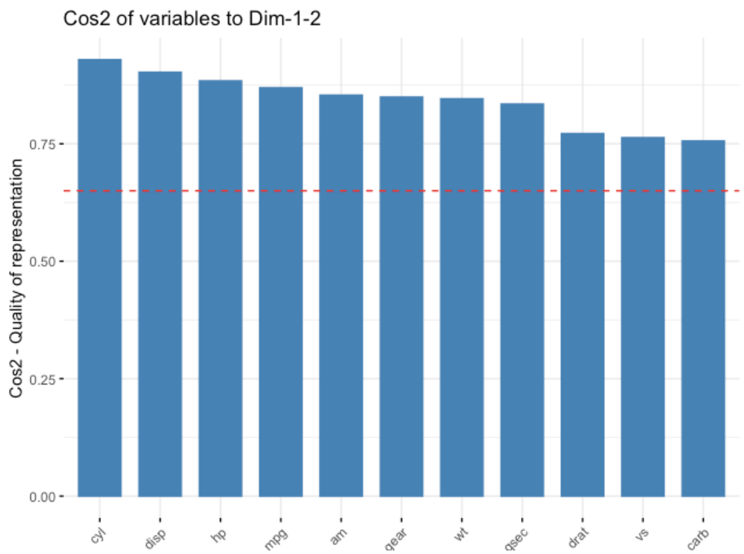




Cas pratique

➢ Nous allons travailler sur le jeu de données **mtcars** : variables contenant l'aspect et les performances d'un ensemble de 32 voitures (modèles entre 1973-1974)

9) Afficher la représentation des variables sur le plan factoriel (coordonnées & qualité)

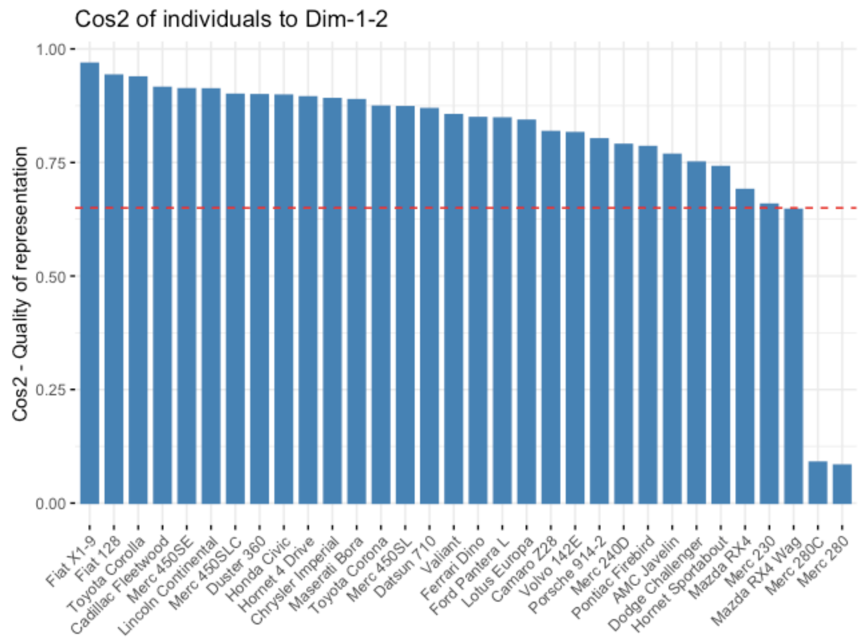


Cas pratique

- Nous allons travailler sur le jeu de données **mtcars** : variables contenant l'aspect et les performances d'un ensemble de 32 voitures (modèles entre 1973-1974)

10) Représenter les individus sur le plan factoriel (coordonnées & qualité)

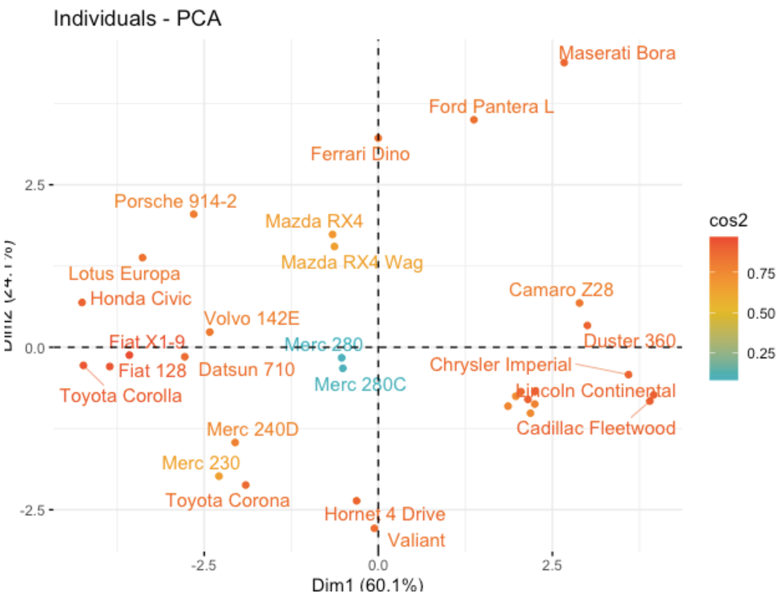
```
> res.pca2$ind$coord
                        Dim.1      Dim.2      Dim.3        Dim.4        Dim.5
Mazda RX4         -0.6572132031  1.7354457 -0.6011992  0.115521565 -0.960652698
Mazda RX4 Wag     -0.6293955058  1.5500334 -0.3823225  0.202307351 -1.032948665
Datsun 710        -2.7793970426 -0.1464566 -0.2412383 -0.249139146  0.405142889
Hornet 4 Drive    -0.3117707086 -2.3630190 -0.1357593 -0.511861672  0.557996837
Hornet Sportabout  1.9744889419 -0.7544022 -1.1344023  0.075653430  0.210836160
Valiant           -0.0561375337 -2.7859996  0.1638257 -0.990771095  0.215052237
Duster 360         3.0026742880  0.3348874 -0.3627592 -0.052353736  0.349349791
Merc 240D         -2.0553287289 -1.4651808  0.9438949 -0.144403291 -0.321718130
Merc 230          -2.2874083842 -1.9835265  1.7972411  0.291806624 -0.339021611
Merc 280          -0.5263812077 -0.1620126  1.4927700  0.067323643 -0.070738219
Merc 280C         -0.5092054932 -0.3238945  1.6835849  0.095867034 -0.151184659
Merc 450SE         2.2478104359 -0.6834740 -0.3753827 -0.131874803 -0.384669304
Merc 450SL         2.0478227622 -0.6832207 -0.4844640 -0.214367071 -0.361301912
```



Cas pratique

- Nous allons travailler sur le jeu de données **mtcars** : variables contenant l'aspect et les performances d'un ensemble de 32 voitures (modèles entre 1973-1974)

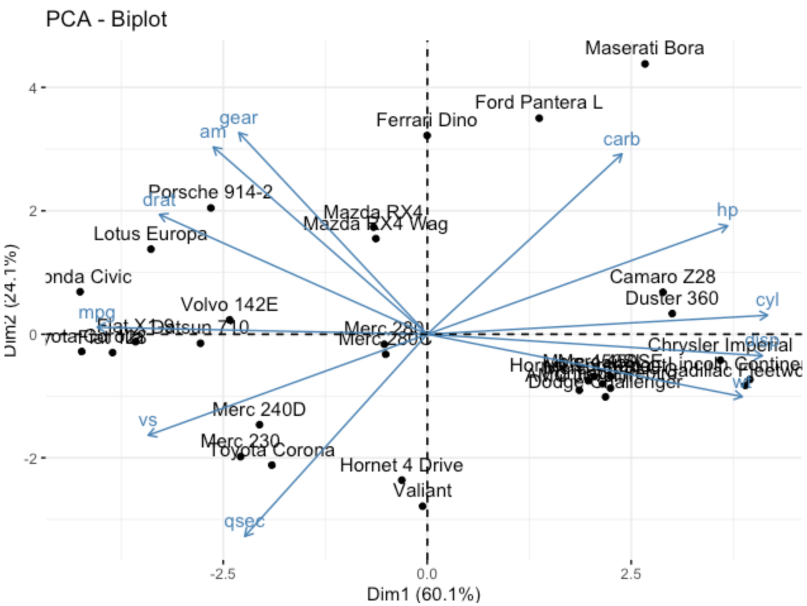
10) Représenter les individus sur le plan factoriel (coordonnées & qualité)



Cas pratique

➢ Nous allons travailler sur le jeu de données **mtcars** : variables contenant l'aspect et les performances d'un ensemble de 32 voitures (modèles entre 1973-1974)

11) Analyser les résultats obtenues à l'aide de l'ACP



Les étapes en ACP

(1) Calculer la matrice des corrélation

(2) Appliquer la fonction ACP sur les données normées

- ✓ Calculer l'inertie
- ✓ Déterminer les valeurs propres et les vecteurs propres
- ✓ Représentation de la projection sur le plan principal

(3) Observations + Interprétations

TP 3 ACP

<u>Consignes</u>

- ❑ Envoyer 1 FOIS votre rapport en pdf ou html sous le nom suivant TP3_ACP_Nom_Prenom
- ❑ Deadline fixée Lundi 17 Mai avant 23h59
- ❑ Expliquer le choix des packages et commenter le code utilisé
- ❑ Afficher le code et l'affichage des commandes
- ❑ La qualité des graphiques et l'interprétation des résultats comptera pour 50% de la notation

Pour aller plus loin…

Site Stdha : graphiques +++ élaborés

http://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique/73-acp-analyse-en-composantes-principales-avec-r-l-essentiel/

Méthodes d'analyses

AFC : Analyse factorielle des correspondances
variables qualitatives (tableaux de contingence)

AFCM : Analyse factorielle des correspondances en composante principale
variables qualitatives (tableaux disjonctifs)

AFD : Analyse factorielle discriminante

Méthodes d'analyses

AFD : Analyse factorielle discriminante

Méthode factorielle permet la réduction de la dimension des données : Explicative & PREDICTIVE

→ Objectif : Exploration des statistique de variables quantitatives et d'une variable qualitative

1- vérifier sur un graphique à plusieurs dimensions (dim= 2 ou 3) dimensions si les groupes auxquels appartiennent les observations sont biendifférents

2- permet de déterminer quelles sont les caractéristiques des groupes en se basant sur les variables explicatives

3- Contribue à prédire le groupe d'appartenance pour une nouvelle observation

→ **Principe** :

Dans l'espace des individus, il s'agit de projeter les individus dans une direction afin de mettre en avant les groupes

Etapes :

(1) Diagonaliser

Méthodes d’analyses

<u>AFD</u> : Analyse factorielle discriminante

Méthode factorielle permet la réduction de la dimension des données : Explicative & PREDICTIVE

→ **<u>Principe</u>** :

Dans l’espace des individus, il s’agit de projeter les individus dans une direction afin de mettre en avant les groupes

→ Packages MASS : fonctions lda et predict permet de réaliser une analyse discriminante

```
Taille de l'ensemble des groupes :
----------------------------------
[1] 74

Moyennes pour l'ensemble des groupes :
--------------------------------------
    [,1]
X1 177.3
X2 124.0
X3  50.4
X4 134.8
X5  13.0
X6  95.4

Matrice de variances pour l'ensemble des groupes :
-----------------------------------------
        X1    X2    X3      X4    X5      X6
X1  853.41  6.48 -7.64 -100.60 48.56 -237.29
X2    6.48 70.96 15.50   48.63 -2.19   58.43
X3   -7.64 15.50  7.47   16.66 -1.82   20.04
X4 -100.60 48.63 16.66  105.69 -5.49  114.60
X5   48.56 -2.19 -1.82   -5.49  4.53  -14.47
```


*<u>Sources figures</u> : http://www.math.u-bordeaux.fr/~mchave100p/wordpress/wp-content/uploads/2013/10/TP_AFD.pdf

Méthodes d'analyses

AFD : Analyse factorielle discriminante

Méthode factorielle permet la réduction de la dimension des données : Explicative & PREDICTIVE

→ **Statistiques descriptives pour chacun des groupes**

Groupe A
Groupe B
Groupe C

```
Matrice de correlations pour l'ensemble des groupes :
-----------------------------------------------------
      X1    X2    X3    X4    X5    X6
X1  1.00  0.03 -0.10 -0.33  0.78 -0.57
X2  0.03  1.00  0.67  0.56 -0.12  0.49
X3 -0.10  0.67  1.00  0.59 -0.31  0.52
X4 -0.33  0.56  0.59  1.00 -0.25  0.78
X5  0.78 -0.12 -0.31 -0.25  1.00 -0.48
X6 -0.57  0.49  0.52  0.78 -0.48  1.00


Taille de l'echantillon du groupe ' A ' :
----------------------------------------
[1] 21


Moyennes pour le groupe ' A ' :
-----------------------------

X1 183.1
X2 129.6
X3  51.2
X4 146.2
X5  14.1
X6 104.9

Matrice de variances pour le groupe ' A ' :
------------------------------------------
       X1    X2    X3    X4    X5    X6
X1 140.47 63.46 17.64 14.36 -4.96 13.54
X2  63.46 48.81 11.00  2.36 -1.73  2.95
X3  17.64 11.00  4.75  5.57 -0.50  5.22
X4  14.36  2.36  5.57 30.15 -0.92 14.88
X5  -4.96 -1.73 -0.50 -0.92  0.75 -1.89
X6  13.54  2.95  5.22 14.88 -1.89 36.41
```

Méthodes d’analyses

AFD : Analyse factorielle discriminante

Méthode factorielle permet la réduction de la dimension des données : Explicative & PREDICTIVE

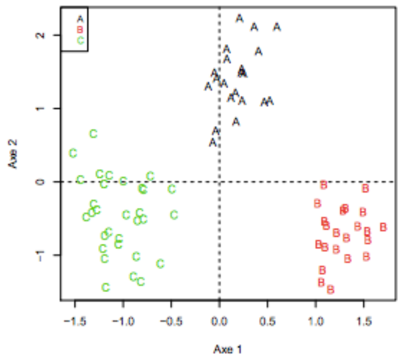
→ Fonction AFD sur Rstudio

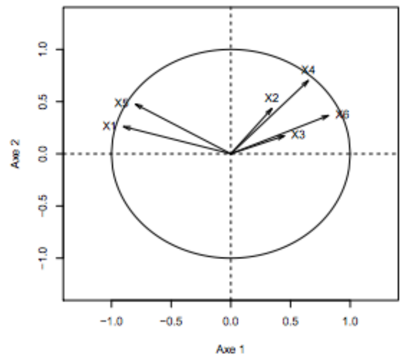
Resultats numériques de l’AFD :

```
Liste des pouvoirs discriminants :
----------------------------------
[1] 0.947 0.795

Matrice des facteurs discriminants :
------------------------------------
      u1    u2
X1 -0.02  0.01
X2  0.01  0.02
X3  0.03 -0.13
X4  0.02  0.09
X5 -0.07  0.24
X6  0.01  0.01

Matrice des variables discriminantes :
--------------------------------------
      s1    s2
1   0.08  1.82
2   0.25  1.48
3  -0.12  1.31
4   0.08  1.67
5   0.41  1.78
6   0.17  1.22
7   0.23  1.54
8   0.12  1.15
9   0.52  1.11
10  0.05  1.34
11  0.17  0.82
12  0.36  2.12
```

```
Matrice descorrelations avec les variables discriminantes :
-----------------------------------------------------------
      s1   s2
X1 -0.90 0.26
X2  0.34 0.43
X3  0.45 0.17
X4  0.65 0.70
X5 -0.80 0.48
X6  0.82 0.37
```

Graphiques de l’AFD :





Projet final

Consignes

- Envoyer en 1 FOIS votre rapport en pdf ou html par BINOME sous le nom suivant Projet_final_Nom1_Prenom1_Nom2_Prenom2

- Le rapport doit contenir vos codes, l'exécution de vos codes ainsi que vos analyses

- Deadline fixée Jeudi 27 Mai avant 23h59

- La qualité des graphiques et l'interprétation des résultats comptera pour 80% de la notation

Projet Final

Description

- C'est une enquête concernant le temps passé dans différentes activités au cours d'une journée (Budget/Temps)

- Le data frame contient 10 variables numériques et 4 variables catégorisées

  1) Les 10 variables numériques représentent le temps passé en :
  Profession, Transport, Ménage, Enfants, Courses, Toilette, Repas, Sommeil Télé, Loisirs

  2) Les 4 variables catégorisées sont:
  Le sexe : 1=Hommes 2=Femmes
  L'activité 1=Actifs 2=Non Act.9 =Non précisé
  L'état civil 1=Célibataires 2=Mariés 9=Non précisé
  Le Pays 1=USA 2=Pays de l'Ouest 3=Pays de l'Est 4=Yougoslavie

Projet Final

Description

- C'est une enquête concernant le temps passé dans différentes activités au cours d'une journée (Budget/Temps)
- Le data frame contient 10 variables numériques et 4 variables catégorisées

3) Le code suivant est utilisé pour identifier les lignes:
H: Hommes, F: Femmes, A: Actifs
N: NonActifs(ves), M: Mariés, C: Célibataires
U: USA, W: Pays de l'Ouest sauf USA, E : Est sauf (ex)Yougoslavie, Y: (ex) Yougoslavie

4) Les temps sont notés en centièmes d'heures :
La première case en haut à gauche du tableau (HAU) indique que les Hommes Actifs des USA passent en moyenne 6 heures et 6 minutes (6 heures +10/100 d'heure, soit 6 heures et 6mn) en activité professionnelle
Le total d'une ligne (sur ces 10 variables numériques) est 2400 (24 heures).

Projet Final

Missions

- Effectuez les statistiques descriptives de ces données
- Réalisez une ACP puis interprétez les résultats
- Quelles sont les critiques à apporter à cette analyse ?