université de BORDEAUX

geii Génie électrique et informatique industrielle / iut de BORDEAUX

**Année universitaire 2024-2025**

# COURS
# INFORMATIQUE SPÉCIALISÉE

## SEMESTRE 5
## BUT 3 Parcours AII

Auteur : Florent ARNAL
Adresse électronique : florent.arnal@u-bordeaux.fr
Site : http://flarnal.e-monsite.com

Le modèle de régression linéaire multiple est l'outil statistique le plus utilisé pour l'étude de données multidimensionnelles. Cas particulier de modèle linéaire, il constitue la généralisation naturelle de la régression simple vue en $S_4$.

Ce cours introduit le modèle linéaire multidimensionnel dans lequel une variable quantitative $Y$ est expliquée, modélisée, par plusieurs variables quantitatives $X_j$ $(j = 1, \cdots, p)$.
Après avoir explicité les hypothèses nécessaires et les termes du modèle, les notions d'estimation des paramètres du modèle (moindres carrés), de prévision par intervalle de confiance, la signification des tests d'hypothèse sont discutées de même que les outils de diagnostics (graphe des résidus).

Sera abordée, dans un second temps, la classification ascendante hiérarchique (CAH) permettant de faire des groupes d'individus "similaires".

Et de terminer par l'annalyse en composantes principales (ACP) permettant d'avoir une analyse asez fine d'individus et variables quantitatives avec des projections dans des espaces dépendant des variables les plus pertinents possibles.

# I Régression linéaire multiple

## I.1 Rappels de régression linéaire (simple)

On considère une série statistique double représentée par un nuage de points $M_i\,(x_i; y_i)_{1\leq i\leq n}$.
Soit $(D)$ une droite d'ajustement. Pour tout entier naturel $i$ tel que $1 \leq i \leq n$, on note $P_i$ le projeté de $M_i$ sur la droite $(D)$ parallèlement à l'axe des ordonnées.
Ajuster ce nuage de point par la méthode des moindres carrés, c'est déterminer la droite $(D)$ pour que la somme $\sum_{i=1}^{n} {M_iP_i}^2$ soit minimale.

L'objectif est de déterminer la droite d'équation $y = ax + b$ rendant minimale la somme $\sum_{i=1}^{n} {M_iP_i}^2$.

On rappelle que la variance observée associée à $X$ est définie par :

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} {x_i}^2 - \bar{x}^2$$

De même, la covariance observée de $X$ et $Y$ est définie par :

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

**Propriété 1** *La droite de régression de $Y$ en $X$ a pour équation $Y = aX + b$ avec*

$$a = \frac{s_{XY}}{s_X^2} \text{ et } b = \overline{y} - a\overline{x}$$

On définit également les résidus, notés usuellement $e_i$, par :

$$e_i = y_i - \hat{y}_i$$

Il est important de rappeler que la moyenne et la somme des résidus sont nulles.

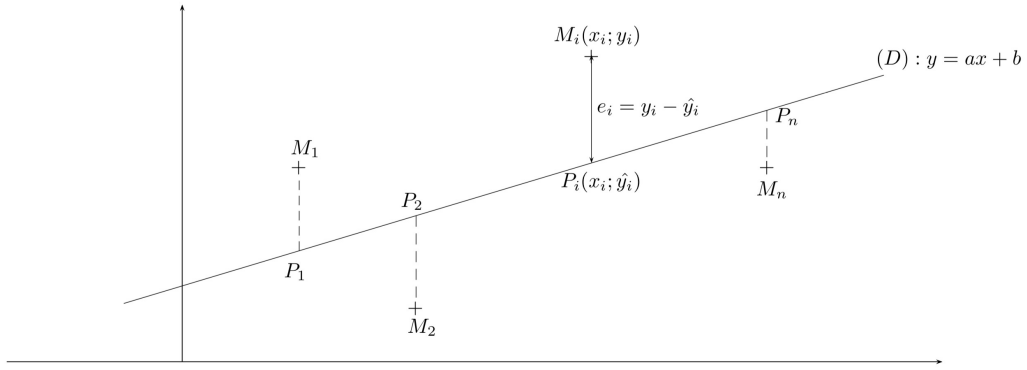



Figure 1: Illustration de la notion de résidus


florent.arnal@u-bordeaux.fr

## I.2 Une autre approche possible

On pose : $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ et $\varepsilon = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$.

Dans le cas bivarié, on cherche $a$ et $b$ tels que : $Y = aX + b \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \varepsilon$ où $a$ et $b$ permettent de minimiser $\sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2$.

Il s'avère que, pour tout $i \in [\![1; n]\!]$, on a : $y_i = ax_i + b + e_i$ avec $\sum_{i=1}^{n} e_i = 0$. Ainsi :

✍

On a donc, pour tout $i \in [\![1; n]\!]$ : $y_i - \bar{y} = a\,(x_i - \bar{x})$.

En notant $Y^c = \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix}$ et $X^c = \begin{pmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix}$, on a : $Y^c = aX^c + \varepsilon$. On admet que :

$$a = \left({}^t X^c X^c\right)^{-1} {}^t X^c Y$$

✍


florent.arnal@u-bordeaux.fr

# II Régression linéaire multiple

## II.1 Contexte

On considère désormais une variable quantitative $Y$ que l'on souhaite expliquer, modéliser, à l'aide d'une combinaison linéaire de plusieurs variables quantitatives $X_j$ $(j = 1, \cdots, p)$.

Si l'on dispose de $n$ observations, on notera : $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $X_1 = \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix}$, $\cdots$, $X_p = \begin{pmatrix} x_{1p} \\ x_{2p} \\ \vdots \\ x_{np} \end{pmatrix}$.

On cherche donc à déterminer des coefficients $\beta_0, \cdots, \beta_p$ tels que, pour tout $i \in [\![1; n]\!]$, on a :

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + e_i$$

Ces relations peuvent s'écrire ainsi :

$$Y = \beta_0 \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

La nullité de la moyenne des résidus conduit à $\bar{y} = \beta_0 + \beta_1 \bar{x_1} + \cdots + \beta_p \bar{x_p}$ ce qui conduit à

$$y_i - \bar{y} = \beta_1 (x_{1i} - \bar{x_1}) + \cdots + \beta_p (x_{pi} - \bar{x_p}) + e_i$$

On peut donc utiliser une nouvelle fois des variables centrées conduisant à

$$Y^c = \beta_1 {X_1}^c + \cdots + \beta_p {X_p}^c + \varepsilon$$

qui s'écrit également

$$Y^c = X^c \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \varepsilon$$

où

$$X^c = \begin{pmatrix} x_{11} - \bar{x_1} & x_{12} - \bar{x_2} & \cdots & \cdots & x_{1p} - \bar{x_p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} - \bar{x_1} & x_{n2} - \bar{x_2} & \cdots & \cdots & x_{np} - \bar{x_p} \end{pmatrix}$$

On va calculer ${}^t X^c X^c$ avec $X^c$ à $n$ lignes et $p$ colonnes. Ainsi, ${}^t X^c$ a $p$ lignes et $n$ colonnes ce qui permet de déterminer que ${}^t X^c X^c$ est une matrice carrée d'ordre $p$ ($p$ lignes et $p$ colonnes).
Déterminons les deux premiers coefficients de la première colonne :
✍ ${}^t X^c X^c =$


florent.arnal@u-bordeaux.fr

Ainsi, le premier coefficient est égal à :

Quant au deuxième coefficient, il est égal à :

Plus généralement, la matrice ${}^tX^cX^c$ peut s'écrire :

$$ {}^tX^cX^c = \begin{pmatrix} ns_1{}^2 & ns_{12} & \cdots & \cdots & ns_{1p} \\ ns_{12} & ns_2{}^2 & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ ns_{1p} & \cdots & \cdots & \cdots & ns_p{}^2 \end{pmatrix} $$

On a donc que :

$$ \frac{{}^tX^cX^c}{n} = \begin{pmatrix} s_1{}^2 & s_{12} & \cdots & \cdots & s_{1p} \\ s_{12} & s_2{}^2 & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{1p} & \cdots & \cdots & \cdots & s_p{}^2 \end{pmatrix} $$

Cette matrice (symétrique) est appelée matrice de variance-covariance.


florent.arnal@u-bordeaux.fr

Les coefficients de la régression $\beta_1, \cdots, \beta_p$ sont estimés par des réels, notés $\hat{\beta}_1, \cdots, \hat{\beta}_p$ tels que :

$$\begin{pmatrix} \hat{\beta}_1 \\ \beta_2 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = \left[{}^t X^c X^c\right]^{-1} {}^t X^c Y$$

Ces coefficients seront fournis par le logiciel R.

## II.2 Étude d'un exemple

On souhaite expliquer les valeurs de $Y$ en fonction de $X_1$ et $X_2$ à l'aide d'un modèle linéaire du type

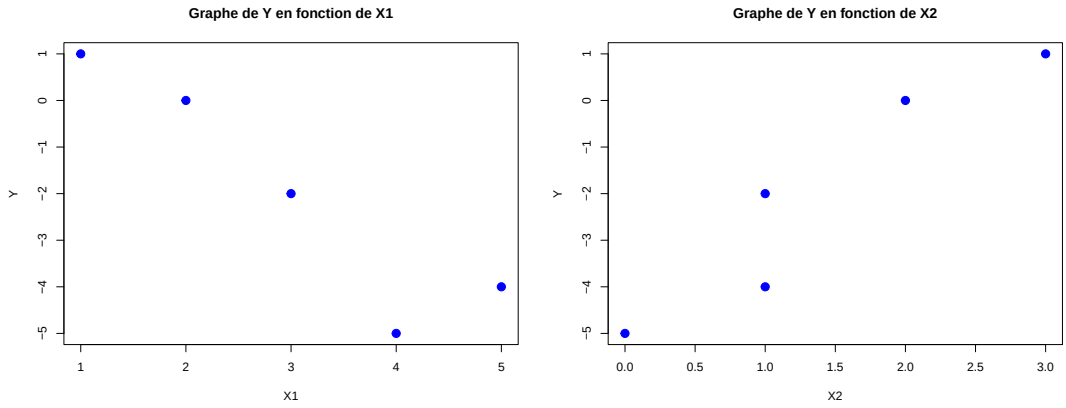
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X2 + \varepsilon$$

```
  X1 X2  Y
1  1  3  1
2  2  2  0
3  3  1 -2
4  4  0 -5
5  5  1 -4
```

On peut déterminer les coefficients de détermination des variables $Y$ et $X_1$ ainsi que celui des variables $Y$ et $X_2$.

```
> D1 = cor(X1,Y)^2
> D2 = cor(X2,Y)^2
> round(data.frame(D1, D2),3)
```

```
     D1    D2
1 0.865 0.895
```




Une régression de $Y$ en $X_1$ ou de $Y$ en $X_2$ pourrait être envisagée mais il peut s'avérer intéressant de combiner $X_1$ et $X_2$ pour expliquer $Y$.


florent.arnal@u-bordeaux.fr

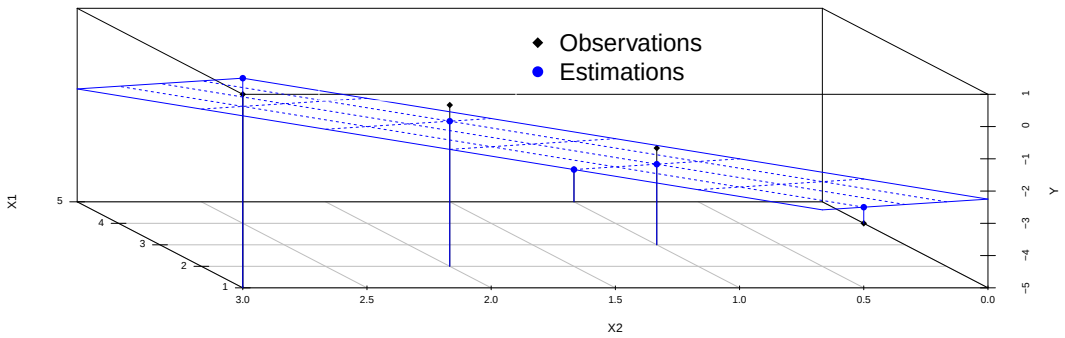
```
> modele = lm(Y~ X1 + X2)
> modele


Call:
lm(formula = Y ~ X1 + X2)

Coefficients:
(Intercept)           X1           X2
      -1.50        -0.75         1.25
```

Les estimations de $Y$ par le modèle sont données par :

$$\hat{Y} = -1.5 - 0.75X_1 + 1.25X_2$$

Il s'agit de l'équation du plan représenté ci-dessous.

```
> Estimations = predict(modele, newdata=data)
> data.frame(X1, X2, Y, Estimations)


  X1 X2  Y Estimations
1  1  3  1         1.5
2  2  2  0        -0.5
3  3  1 -2        -2.5
4  4  0 -5        -4.5
5  5  1 -4        -4.0


> # On peut également utiliser
> modele$fitted.values


   1    2    3    4    5
 1.5 -0.5 -2.5 -4.5 -4.0
```




Les résidus associés à ce modèle sont associés à la variable $\varepsilon$ telle que

$$\varepsilon = Y - \hat{Y}$$


florent.arnal@u-bordeaux.fr

```
> Y-Estimations
```

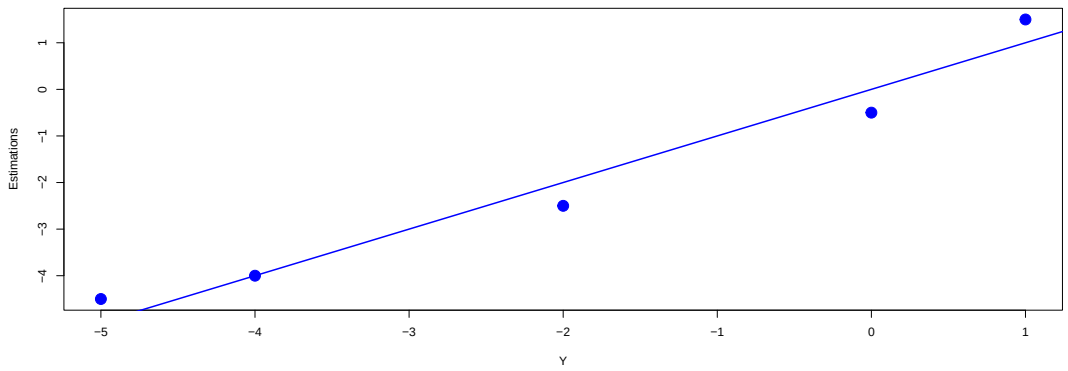
```
            1             2             3             4             5
-5.000000e-01  5.000000e-01  5.000000e-01 -5.000000e-01 -4.440892e-16
```

```
> # On peut utiliser directement avec R la fonction 'resid'
> residus = resid(modele)
> round(residus,1)
```

```
  1   2   3   4   5
-0.5  0.5  0.5 -0.5  0.0
```

On peut représenter les résidus voire représenter les valeurs prédites (estimations) en fonction des valeurs observées de $Y$ (ou inversement). On espère que ces valeurs sont proches (deux à deux) ce qui conduit à espérer des points proches de la bissectrice d'équation $Y = \hat{Y}$ de la forme $Y = b\hat{Y} + a$ avec $a = 0$ et $b = 1$.

```
> plot(Estimations~Y, pch=19, col="blue", cex=2)
> abline(a = 0, b = 1, col = "blue", lwd = 2)
```




## II.3 Conditions d'utilisation du modèle et pertinence du modèle

On s'intéresse à la recherche d'un modèle du type :

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

conduisant aux estimations suivantes :

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Il est impératif d'avoir plus d'observations, dont le nombre est noté $n$, que de coefficients à déterminer ce qui nous conduit à la relation

$$\boxed{n > p + 1}$$


florent.arnal@u-bordeaux.fr

Afin de déterminer la pertinence du modèle, on peut calculer le coeffcient de détermination $R^2$ défini par $R^2 = \dfrac{SCE_{\text{expliquée}}}{SCE_{\text{totale}}} = \dfrac{SCE_{\hat{Y}}}{SCE_Y} = \dfrac{s^2_{\hat{Y}}}{s^2_Y}$. Comme pour une régression simple, on a la relation :

$$\boxed{SCE_{\text{totale}} = SCE_{\text{expliquée}} + SCE_{\text{résiduelle}}}$$

avec $SCE_{\text{totale}} = SCE_Y$, $SCE_{\text{expliquée}} = SCE_{\hat{Y}}$ et $SCE_{\text{résiduelle}} = SCE_\varepsilon$.
✍

On a donc : $R^2 = 1 - \dfrac{SCE_{\text{résiduelle}}}{SCE_{\text{totale}}}$.
Cependant, plus $p$ est grand, plus ce coefficient est grand. En conséquence, il est recommandé d'utiliser le coefficient de détermination ajusté, noté $R^2_{\text{ajusté}}$, tel que

$$\boxed{R^2_{\text{ajusté}} = 1 - \dfrac{\dfrac{SCE_{\text{résiduelle}}}{n-p-1}}{\dfrac{SCE_{\text{totale}}}{n-1}} = 1 - \dfrac{n-1}{n-p-1}(1-R^2)}$$

Ce coefficient de détermination ajusté traduit donc à la fois la qualité de l'ajustement (liaison entre $Y$ et les $X_i$) et la complexité du modèle (nombre $p$ de variables explicatives).
Si $R^2_{\text{ajusté}}$ est proche de 1 alors le modèle est proche de la réalité.
Si $R^2_{\text{ajusté}}$ est proche de 0 alors le modèle explique très mal la réalité. Il faut alors trouver un meilleur modèle.

Les variables $\varepsilon_i$ doivent être indépendantes et distribuées suivant une même loi normale $\mathcal{N}(0;\sigma)$ pour $i \in [\![1;n]\!]$.

Il est donc nécessaire de vérifier :
- l'homogénéité (égalité) des variances des résidus (homoscédasticité) ;
- la normalité de la distribution des résidus.

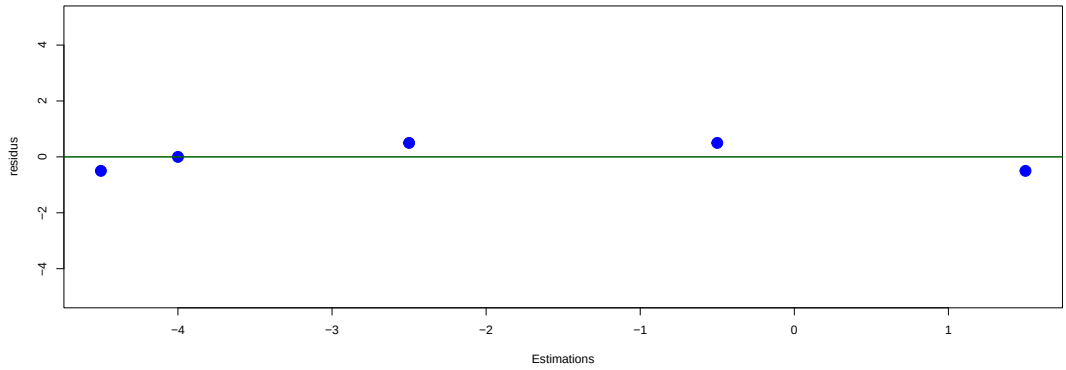
Afin de s'assurer de la normalité des résidus, on peut utiliser un test statitique tel que le test de Shapiro.

```
> shapiro.test(residus)


        Shapiro-Wilk normality test

data:  residus
W = 0.82083, p-value = 0.1185
```

La probabilité étant supérieure à 5%, on ne rejette pas l'hypothèse de normalité.

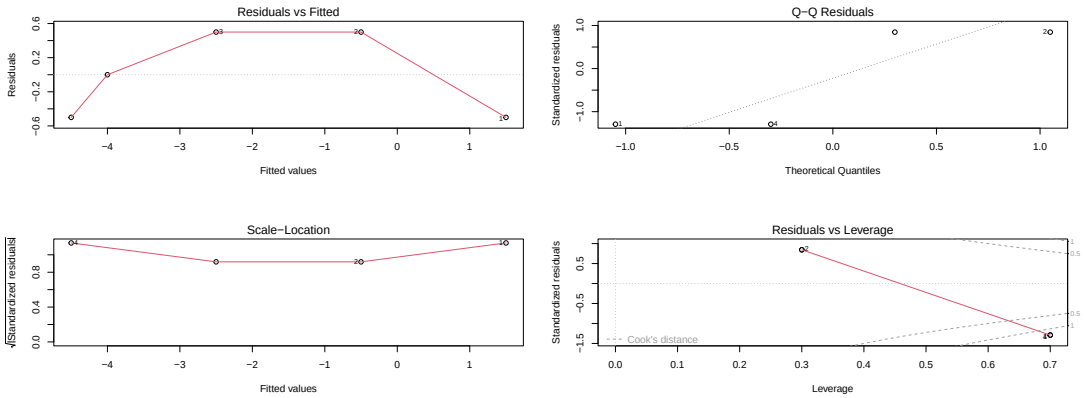
On peut également utiliser une représentation graphique des résidus en fonction des valeurs prédites par le modèle (de régression linéaire).

```
> plot(residus ~ Estimations, pch=19, col="blue", cex=2, ylim = c(-5,5))
> abline(h = 0, col = "darkgreen", lwd = 2, )
```




Chaque point représente l'écart entre la réponse (liée à $Y$) et la réponse prédite par le modèle. Il nous informe, notamment, sur l'indépendance des résidus.
On peut obtenir plusieurs graphiques liés aux résidus en utilisant le code suivant :

```
> par(mfrow=c(2,2))
> plot(modele)
> plot(residus ~ Estimations, pch=19, col="blue", cex=2, ylim = c(-5,5))
> abline(h = 0, col = "darkgreen", lwd = 2, )
```




Le graphique situé en haut, à droite, est un graphique Quantile-Quantile de normalité (QQ-norm). Si les points sont sensiblement alignés (nuage longiligne), on peut considérer la distribution gaussienne (normale).

On peut avoir des informations sur la pertinence du modèle en utilisant le $R^2$ ajusté mais également en regardant la probabilité associée à l'analyse de variance (ANOVA).


florent.arnal@u-bordeaux.fr

```
> summary(modele)


Call:
lm(formula = Y ~ X1 + X2)

Residuals:
         1          2          3          4          5
-5.000e-01  5.000e-01  5.000e-01 -5.000e-01 -2.918e-16

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.5000     1.9365  -0.775    0.520
X1           -0.7500     0.4031  -1.861    0.204
X2            1.2500     0.5590   2.236    0.155

Residual standard error: 0.7071 on 2 degrees of freedom
Multiple R-squared:  0.9615,        Adjusted R-squared:  0.9231
F-statistic:    25 on 2 and 2 DF,  p-value: 0.03846
```

Le test ANOVA permet de tester l'hypothèse $\mathcal{H}_0$ selon laquelle tous les coefficients du modèle sont égaux à 0 ce qui peut se traduire par l'absence de relation linéaire entre $Y$ et les variables $X_1, \cdots, X_p$.
Dans cet exemple, la probabilité étant supérieure à 5%, on peut considérer que le modèle linéaire est pertinent.

A noter que l'on peut également observer que des tests de Student sont effectués pour chacun des coefficients afin de tester l'hypothèse $\mathcal{H}_0$ de nullité du coefficient.
Pour chacun des tests, si la probabilité est supérieure à 5%, le coefficient n'est pas significativement différent de 0.
On peut alors considérer que la variable $Y$ ne dépend pas de la variable $X_i$. On peut dès lors la supprimer du modèle.

## II.4 Pour affiner le modèle : sélection de variables

Il peut être utile de réaliser une sélection de variables dans trois situations principales.
Tout d'abord, quand on construit des modèles, notamment prédictifs, de classification ou de régression, il peut être intéressant de choisir au mieux les variables que l'on peut utiliser afin de rendre le modèle plus interprétable.
Il est intéressant d'identifier les variables qui jouent un rôle prépondérant dans la description d'un phénomène plutôt que d'utiliser de multiples variables dont certaines peuvent dépendre d'autres. Ceci rend nortamment les méthodes d'apprentissage plus performantes.
En outre, l'utilisation d'un modèle simple, utilisant peu de variables, a plus de chances de pouvoir être généralisé à d'autres situations (échantillons) qu'un modèle basé sur des centaines de variables qui a toutes les chances d'ajuster parfaitement le jeu d'apprentissage mais d'avoir un faible pouvoir de généralisation.

Plusieurs méthodes existent pour sélectionner les variables les plus pertinentes, notamment les méthodes pas à pas.
Elles consistent à considérer, dans un premier temps, un modèle faisant intervenir toutes les variables explicatives puis de procéder par élimination ou ajout successif de variables en utilisant des tests de Student.
On peut citer :

- la méthode descendante lorsque l'on élimine des variables (à chaque étape, on élimine la variable la moins significative à l'aide de tests de Student),
- la méthode ascendante lorsque l'on ajoute des variables,
- une combinaison de ces deux méthodes (une variable considérée comme la plus significative à une étape de l'algorithme peut à une étape ultérieure devenir non significative en raison de ses corrélations avec d'autres variables introduites après coup dans le modèle).
  Après l'introduction d'une nouvelle variable dans le modèle, on réexaminer la significativité de chaque variable explicative anciennement admise dans le modèle et, après réexamen, si des variables ne sont plus significatives, alors on retire du modèle la moins significative d'entre elles.

Il est à noter que toutes ces procédures ne mènent pas forcément à la même solution quand elles sont appliquées au même problème.

Avec R, pour la méthode combinée, on utilise :

```
> step(modele, direction = "both")
```

# III Classification ascendante hiérarchique (CAH)

Dans cette partie, nous allons voir comment regrouper $N$ individus, caractérisés par des variables quantitatives ($p$), en groupes homogènes.
Pour ce faire, l'objectif est d'obtenir un ensemble de classes de moins en moins fines par regroupements successifs, en utilisant la distance euclidienne. La distance, $d(A, B)$, entre deux individus $A(x_1, \cdots, x_p)$ et $B(y_1, \cdots, y_p)$ est telle que

$$d^2(A, B) = \sum_{i=1}^{p} (x_i - y_i)^2$$

L'alogorithme de classement est le suivant :

- On regroupe les deux individus les plus proches créant aunsi une classe comprenant ces deux individus.
- On calcule ensuite la dissimilarité entre cette classe et les $N - 2$ autres objets en utilisant le critère d'agrégation.
  Puis on regroupe les deux objets ou classes d'objets dont le regroupement minimise le critère d'agrégation. La distance entre deux groupe $G_1$ et $G_2$ étant définie par

$$d(G_1, G_2) = \min_{i \in G_1, j \in G_2} d(i, j)$$

- On continue ainsi jusqu'à ce que tous les objets soient regroupés.

Ces regroupements successifs produisent un arbre binaire de classification (dendrogramme), dont la racine correspond à la classe regroupant l'ensemble des individus. Ce dendrogramme représente une hiérarchie de partitions. On peut alors choisir une partition en tronquant l'arbre à un niveau donné, le niveau dépendant soit des contraintes de l'utilisateur (l'utilisateur sait combien de classes il veut obtenir), soit de critères plus objectifs.
Exemple :
On considère quatre individus caractérisés par trois variables quantitatives : la masse (en kg), la taille (en cm) et l'âge (en années).
Les données sont les suivantes :

| | masse | taille | age |
|---|---|---|---|
| A | 55 | 160 | 20 |
| B | 60 | 160 | 21 |
| C | 75 | 180 | 20 |
| D | 80 | 190 | 22 |

La distance entre $A$ et $B$ est telle que :

$$d(A, B) = \sqrt{(55 - 60)^2 + (160 - 160)^2 + (20 - 21)^2} = \sqrt{26} \simeq 5,1$$


florent.arnal@u-bordeaux.fr

Il s'avère que les variables n'ont pas, en général, la même variabilité. Certaines variables à forte variabilité auraient alors un effet plus important que d'autres dans le calcul de la distance, ce qui augmenterait leur importance.
On convient donc de travailler de préférence sur des données **centrées réduites** en utilisant pour une variable $X_i$ la variable, notée $X_i{}^*$ définie par

$$X_i{}^* = \frac{X_i - \bar{x}_i}{s_i}$$

où $s_i$ est une estimation de l'écart-type de $X_i$.

De manière plus globale, on peut obtenir distances sur les variables centrées réduites entre tous les individus en utilisant :

```
> # Données centrées réuites
> data_CR = data.frame(scale(data))
> data_CR
```

```
      masse     taille        age
A -1.050210 -0.8333333 -0.7833495
B -0.630126 -0.8333333  0.2611165
C  0.630126  0.5000000 -0.7833495
D  1.050210  1.1666667  1.3055824
```

```
> distance = dist(data_CR)
> distance
```

```
         A        B        C
B 1.125780
C 2.145066 2.111142
D 3.574269 2.813261 2.232611
```

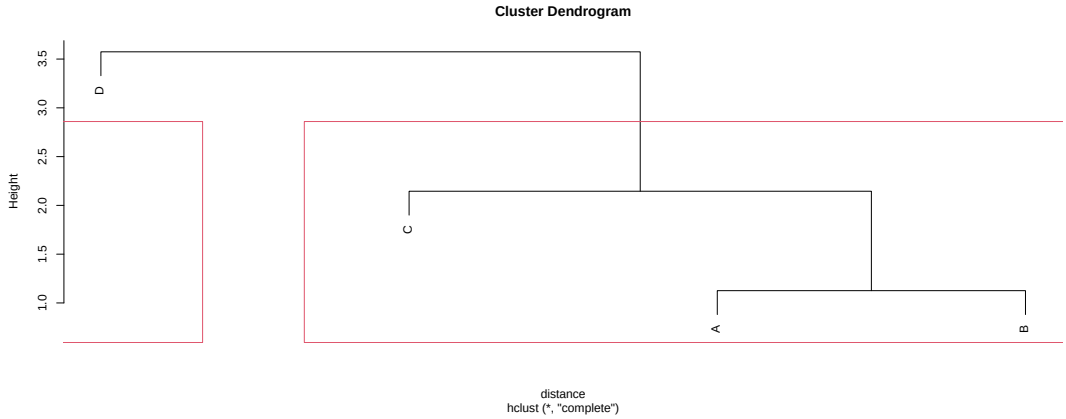
Le dendogramme peut alors être obtenu, en choisissant, si on le souhaite, des groupes ($k = 2$ ici) en utilisant ce qui suit :

```
> CAH = hclust(distance)
> CAH
```

```
Call:
hclust(d = distance)

Cluster method   : complete
Distance         : euclidean
Number of objects: 4
```


florent.arnal@u-bordeaux.fr

```
> plot(CAH)
> rect.hclust(CAH, k=2)
```




Les différents groupes créés peuvent également être identifiés.

```
> classe = cutree(CAH, k = 2)
> data$classe = as.factor(classe)
> data
```

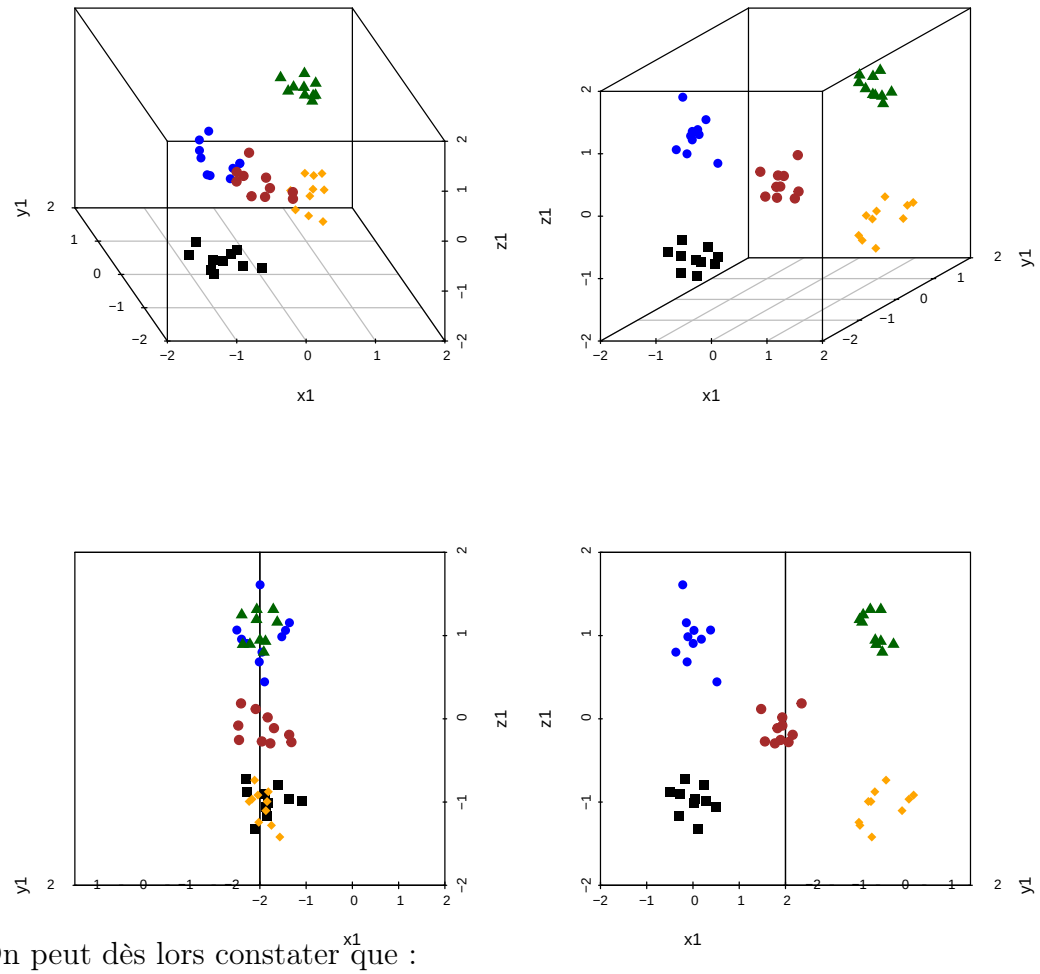
```
  masse taille age classe
A    55    160  20      1
B    60    160  21      1
C    75    180  20      1
D    80    190  22      2
```


florent.arnal@u-bordeaux.fr

# IV Analyse en Composantes principales (ACP)

## IV.1 Principe de l'ACP

On considère trois variables quantitatives $X$, $Y$ et $Z$ observées sur $n = 50$ individus.
Quatre représentations graphiques de ces observations sont données ci-dessous :




On peut dès lors constater que :

- Il est difficile d'obtenir des informations fiables sur des graphiques en 3D.
- On peut réussir à obtenir, via des projections sur un plan (en dimension 2), de "nouvelles" informations très intéressantes, qui ne dénaturent pas trop les informations initiales.

Avec les méthodes d'analyse de données multivariées telles que l'analyse en composante principale (ACP), nous allons tenter de trouver automatiquement les meilleurs façons de représenter les variables quantitatives issues d'un jeu de données.

Figure 2: De l'importance d'avoir le meilleur angle de vue

## IV.2 Principe de l'ACP

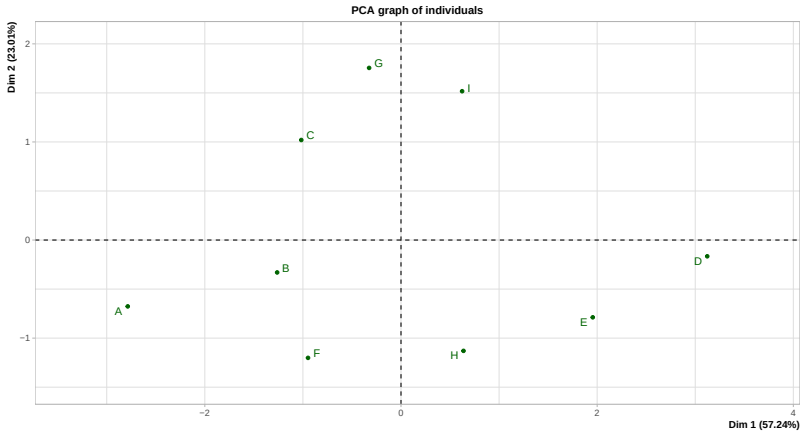
Une ACP (ou "PCA" pour Principal Components Analysis en anglais) est une analyse qui vise à décrire un jeu de données comprenant de multiples variables quantitatives.
Le principe d'une ACP est de réduire un jeu de données à $p$ variables $X_1, X_2, \cdots, X_p$ (i.e. $p$ dimensions) à quelques nouvelles variables (généralement 2 voire 3),que l'on appelle les composantes principales.
On obtient ainsi un nouveau jeu de données, qui structure et résume l'information contenue dans le jeu de données initial, bien plus facile à observer et décrire (cf. 2 dimensions).
Voici un exemple présentant les notes de 9 étudiants dans 5 modules : Mathématiques, Electronique, Physique, Energie, Culture et Communication.

| | Maths | Phys | ELN | Energie | CC |
|---|---|---|---|---|---|
| A | 6.0 | 6.0 | 5.0 | 5.5 | 8 |
| B | 8.0 | 8.0 | 8.0 | 8.0 | 9 |
| C | 6.0 | 7.0 | 11.0 | 9.5 | 11 |
| D | 14.5 | 14.5 | 15.5 | 15.0 | 8 |
| E | 14.0 | 14.0 | 12.0 | 12.0 | 10 |
| F | 11.0 | 10.0 | 5.5 | 7.0 | 13 |
| G | 5.5 | 7.0 | 14.0 | 11.5 | 10 |
| H | 13.0 | 12.5 | 8.5 | 9.5 | 12 |
| I | 9.0 | 9.5 | 12.5 | 12.0 | 18 |

Afin de visualiser les individus dans un plan adapté, on obtient les coordonnées suivantes liées aux composantes principales (nouvelles variables les plus pertinentes pour des projections).

| | Noms | Dim.1 | Dim.2 | Dim.3 |
|---|---|---|---|---|
| A | A | -2.785716 | -0.6764554 | -0.7368007 |
| B | B | -1.262490 | -0.3303384 | -0.5549504 |
| C | C | -1.016747 | 1.0198286 | -0.2880884 |

Graphiquement, on obtient, dans le plan constitué par les deux axes principaux (composantes principales), le nuage de points suivant :





florent.arnal@u-bordeaux.fr

## IV.3 Généralités sur l'ACP

L'objectif est de mettre en place des outils afin de déterminer un bilan :

- des ressemblances entre individus :
  - Quels sont les individus qui se ressemblent ?
  - Quels sont ceux qui sont différents ?
  - Existe-t-il des groupes homogènes d'individus ?

- des liens entre variables :
  - Quelles sont les variables qui sont liées positivement entre elles (corrélées) ?
  - Quelles sont celles qui sont liées "négativement" (anti-corrélées) ?
  - Existe-t-il des groupes de variables corrélées entre elles ?

### Matrices de corrélation et covariance

Nous avons vu dans la première partie que la matrice de variance-covariance, que l'on peut noter $V$, sur des données centrées est donnée par :

$$V = \frac{{}^tX^cX^c}{n} = \begin{pmatrix} {s_1}^2 & s_{12} & \cdots & \cdots & s_{1p} \\ s_{12} & {s_2}^2 & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{1p} & \cdots & \cdots & \cdots & {s_p}^2 \end{pmatrix}$$

Si l'on considère des données centrés réduites, la matrice $X^c$ devient

$$X_r^c = \begin{pmatrix} \frac{x_{11}-\bar{x_1}}{s_1} & \frac{x_{12}-\bar{x_2}}{s_2} & \cdots & \cdots & \frac{x_{1p}-\bar{x_p}}{s_p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{x_{n1}-\bar{x_1}}{s_1} & \frac{x_{n2}-\bar{x_2}}{s_2} & \cdots & \cdots & \frac{x_{np}-\bar{x_p}}{s_p} \end{pmatrix}$$

ce qui conduit à considérer la matrice des corrrélations :

$$R = \frac{{}^tX_r^cX_r^c}{n} = \begin{pmatrix} 1 & r_{12} & \cdots & \cdots & r_{1p} \\ r_{12} & 1 & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{1p} & \cdots & \cdots & \cdots & 1 \end{pmatrix}$$

### Espace des individus

L'inertie du nuage de points de $n$ individus $E_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$, associés à $p$ variables est définie par

$$I_G = \sum_{i=1}^{n} d^2(E_i, G)$$


florent.arnal@u-bordeaux.fr

Si les données sont centrées réduites, le centre de gravité du nuage des individus $E_i^* = \begin{pmatrix} \frac{x_{11}-\bar{x_1}}{s_1} \\ \frac{x_{12}-\bar{x_2}}{s_2} \\ \vdots \\ \frac{x_{1p}-\bar{x_p}}{s_p} \end{pmatrix}$ est l'origine $O$. Ainsi, l'inertie du nuage (indépendante des valeurs) est donnée par

$$I_O = \sum_{i=1}^{n} d^2(E_i^*, O) = \sum_{i=1}^{n} \|E_i^*\|^2 = p$$

**Espace des variables**

On définit la produit scalaire de deux variables centrées $X_i{}^* = \begin{pmatrix} x_{i1} - \bar{x_1} \\ x_{i2} - \bar{x_2} \\ \vdots x_{ip} - \bar{x_p} \end{pmatrix}$ et $X_j{}^* = \begin{pmatrix} x_{j1} - \bar{x_1} \\ x_{j2} - \bar{x_2} \\ \vdots x_{jp} - \bar{x_p} \end{pmatrix}$ par

$$< X_i{}^*, X_j{}^* >= \frac{1}{n} \sum_{k=1}^{n} (x_{ik} - \bar{x_k}) (x_{jk} - \bar{x_k}) = s_{ij}$$

De plus,

$$< X_i{}^*, X_j{}^* >= \|X_i{}^*\| \|X_j{}^*\| \cos(X_i{}^*, X_j{}^*) = \sqrt{s_i{}^2} \sqrt{s_j{}^2} \cos(X_i{}^*, X_j{}^*) = s_i s_j \cos(X_i{}^*, X_j{}^*)$$

Ainsi :

$$\cos(X_i{}^*, X_j{}^*) = \frac{s_{ij}}{s_i s_j} = r_{ij}$$

À retenir
Le cosinus de l'angle entre deux variables centrées correspond à leur coefficient de corrélation.
Le cosinus au carré correspond donc au coefficient de détermination.

**Détermination des axes principaux d'inertie**

Afin d'obtenir la meilleur qualité de projection, il faut diagonaliser la matrice $V$ des variances-covariances ou celle des corrélations.

À retenir :
La projection se fait sur l'espace (de dimension 2) engendré par les vecteurs propres obtenus lors de la diagonalisation.
Il s'avère que plus la valeur prore est grande, plus le vecteur propre associé est "pertinent".
On considère donc les vecteeurs propres, appelés axes principaux d'inertie, associées aux valeurs propres les plus grandes.
Les composantes principales $C_i$ sont les coordonnées des projections des individus sur les axes principaux.

Reprenons l'exemple précédent.

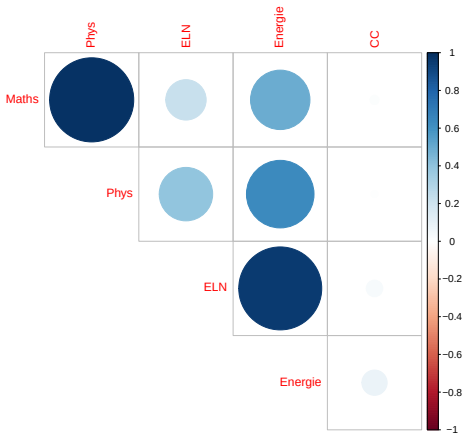
Toutes les variables étant associées au même type de mesures (notes), on peut travailler sur les données brutes.
Plus généralement, on privilégie des variables centrées réduites.

On peut obtenir la matrice des corrélations entre les différentes variables ainsi qu'un graphique associé à ces corrélations entre variables appelé corrélogramme.

```
> Mat_cor = cor(data)
> Mat_cor
```

```
             Maths        Phys        ELN    Energie          CC
Maths   1.00000000 0.982535729 0.22673193 0.49049826 0.011183835
Phys    0.98253573 1.000000000 0.39669324 0.63398550 0.006309933
ELN     0.22673193 0.396693238 1.00000000 0.95613107 0.038035989
Energie 0.49049826 0.633985503 0.95613107 1.00000000 0.088601589
CC      0.01118384 0.006309933 0.03803599 0.08860159 1.000000000
```

```
> library(corrplot)
> corrplot(Mat_cor, type = "upper", diag = F )
```



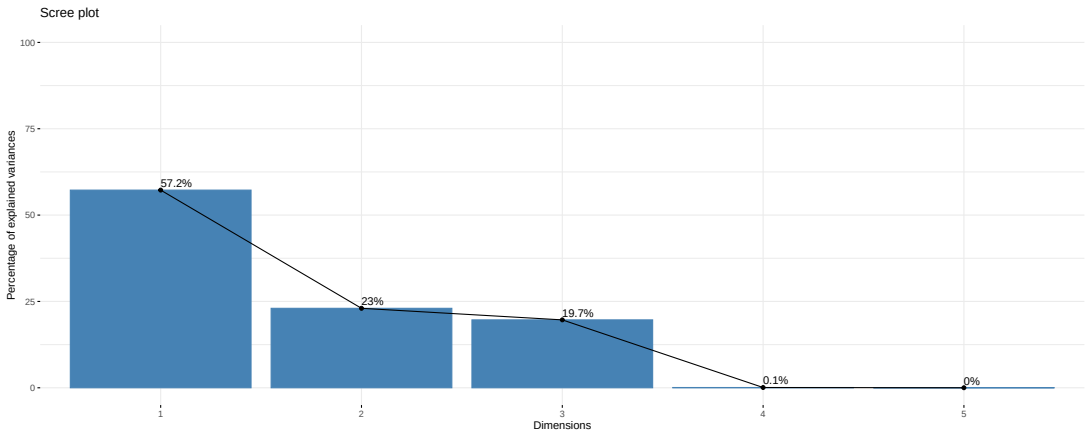

On peut constater que les notes de mathématiques et physique sont fortement corrélées.
Il en est de même entre les notes d'énergie et d'électronique. Dans une moindre mesure, il y a également une corrélation positive entre les notes de physique et d'énergie.
On ne note aucune corrélation significative entre les notes en Culture-Communication et celles des autres matières.

On peut dès lors déterminer les valeurs propres et ainsi, obtenir la qualité de la projection sur les axes associés.
On peut constater ici que seulement trois valeurs propres permettent d'avoir quasiment toute l'information.
En outre, les valeurs propres 2 et 3, associées aux axes 2 et 3 apportent sensiblement le même niveau d'information.

```
> eig.val = get_eigenvalue(ACP)
> eig.val
```

```
        eigenvalue variance.percent cumulative.variance.percent
Dim.1 2.8618175283     57.236350565                     57.23635
Dim.2 1.1506811283     23.013622566                     80.24997
Dim.3 0.9831407383     19.662814766                     99.91279
Dim.4 0.0039371205      0.078742411                     99.99153
Dim.5 0.0004234846      0.008469692                    100.00000
```

```
> fviz_eig(ACP, addlabels = TRUE, ylim = c(0, 100))
> fviz_pca_var(ACP, axes = c(1,2), col.var = "cos2",
+ gradient.cols = c("blue", "orange", "red"), repel = TRUE)
```




Le dernier graphique permet de visualiser l'importance du lien entre les variables initiales et les axes principaux choisis.

On peut obtenir davantage d'informations telles que les coordonnées des variables suivant les axes principaux, la contribution de chaque variable à la détermination d'un axe et le lien entre des variables et un axe principal.

```
> library("corrplot")
> var = get_pca_var(ACP)
> var$cor
```

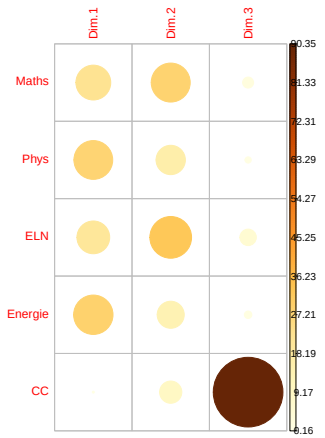
```
             Dim.1      Dim.2       Dim.3
Maths    0.80590635 -0.5713584  0.15344338
Phys     0.89700978 -0.4307915  0.09285091
ELN      0.75808764  0.6110469 -0.22570903
Energie  0.91025382  0.3974903 -0.10842325
CC       0.06668631  0.3275248  0.94248311
```

```
> var$contrib
```

```
            Dim.1     Dim.2      Dim.3
Maths    22.694844 28.370193  2.3948628
Phys     28.115928 16.127955  0.8769133
ELN      20.081534 32.448465  5.1818181
Energie  28.952301 13.730869  1.1957190
CC        0.155393  9.322519 90.3506868
```


florent.arnal@u-bordeaux.fr

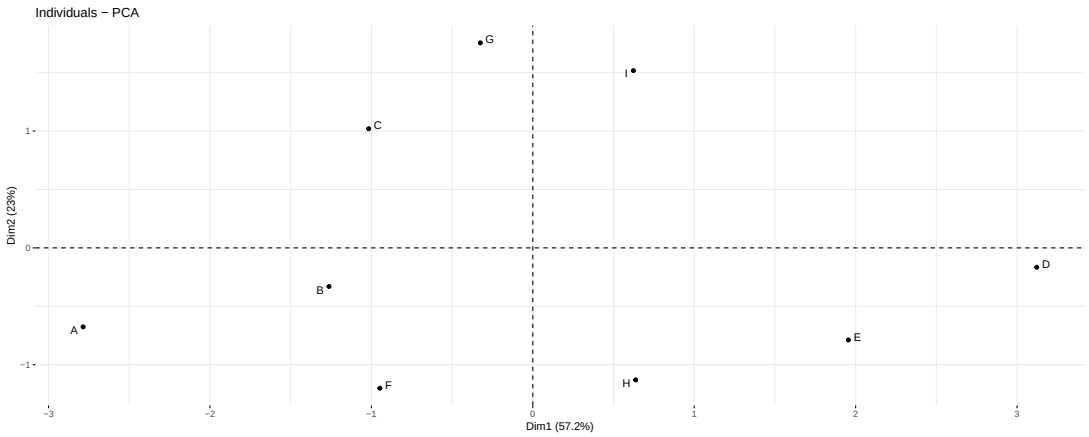
```
> corrplot(var$contrib, is.corr=FALSE)
> dimdesc(ACP, proba = 0.05)


$Dim.1

Link between the variable and the continuous variables (R-square)
================================================================================
        correlation      p.value
Energie   0.9102538 0.0006524958
Phys      0.8970098 0.0010423566
Maths     0.8059063 0.0087226651
ELN       0.7580876 0.0179277629

$Dim.2

$Dim.3

Link between the variable and the continuous variables (R-square)
================================================================================
   correlation      p.value
CC   0.9424831 0.0001419979
```




On peut enfin visualiser les individus dans un nouvel espace (plan) engendré par les axes principaux. D'après ce qui précède, l'axe 1 est fortement corrélé aux notes en maths, physique, électronique et énergie alors que l'axe 3 est fortement corrélé avec la note en Culture et Communication.

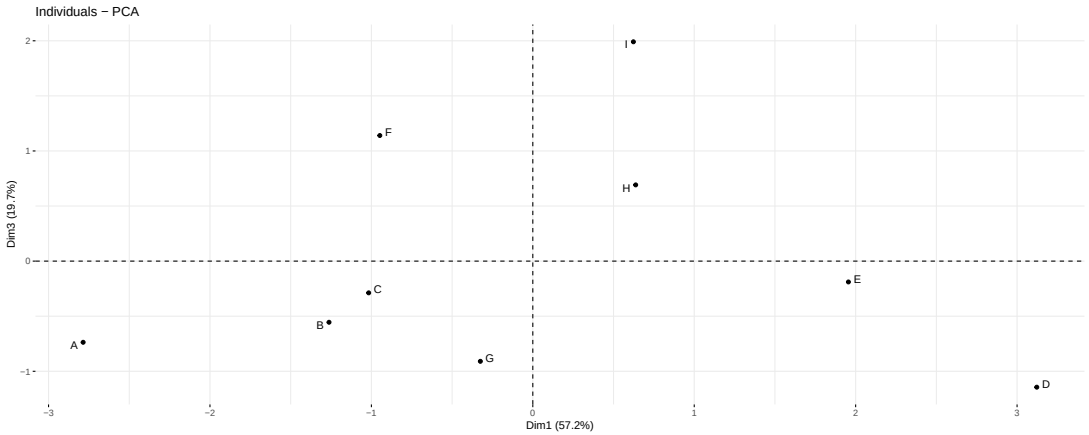

florent.arnal@u-bordeaux.fr

```
> fviz_pca_ind(ACP, axes = c(1,2), repel = TRUE )
```




Plusieurs informations peuvent être obtenues à partir du plan associé aux dimensions 1 et 2.
On peut noter, par exemple, que l'individu D se caractérise par de très bonnes notes en maths, physique, électronique et énergie.
Ce n'est pas le cas pour l'individu A.
On peut également projeter les individus dans le plan associé aux dimensions 1 et 3.

```
> fviz_pca_ind(ACP, axes = c(1,3), repel = TRUE)
```




On peut noter, par exemple, que l'individu I se distingue des autres par son niveau très satisfaisant en Culture et Communication.


florent.arnal@u-bordeaux.fr