Clustering (5 points)

Données :

x1	x2
0	3
1	2
2	1
3	3
3	4
4	1

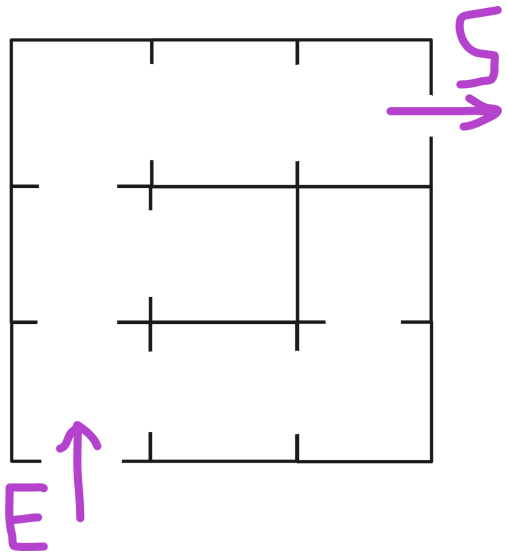
Appliquer en détaillant raisonnablement chaque étape...

1. L'algorithme des k-means avec $K = 2$. (2 points)
   Centres initiaux : (0,2) et (3,1), puis (1,2) et (2,0).
   Donnez les centres finaux ainsi que les compositions des clusters.
2. L'algorithme de clustering hiérarchique "complete linkage" (3 points)
   Dessinez le dendrogramme final. Est-il unique ?

Chaînes de Markov (5 points)

1. Considérant le mini labyrinthe ci-dessous, modélisez la situation par une chaîne de Markov en supposant un joueur se déplaçant au hasard (uniformément, sans biais). (2 points)
2. Peut-on passer de n'importe quel état à n'importe quel autre ? (Justifiez).
   La chaîne est-elle régulière ? (Rappel : régulière si $\exists n \,/\, P^n > 0$). (1 point)
3. Comment calculer le temps moyen passé dans le labyrinthe avant de sortir ? Décrivez au moins les étapes. (2 points)

Note: E pour "Entrée" et S pour "Sortie".



Arbres de décision (5 points)

Le jeu de données suivant issu de Wikipedia indique si une banque accordera ou non un prêt selon trois critères :

- Savings : symbolique "ordonnée" Low, Medium, High ;
- Assets : symbolique "ordonnée" Low, Medium, High ;
- Income : numérique en milliers de dollars (par an).

Construisez un arbre de décision **binaire** en utilisant l'indice de Gini :
$I_G = 1 - \sum_{k=1}^{K} p_k^2$, avec p_k proportion de la classe k dans les données du noeud courant.
Détaillez les étapes (4 points) et dessinez l'arbre (1 point).
Hint: variables à considérer = Income, Savings, Income, Assets (dans cet ordre).

Customer	Savings	Assets	Income ($1000s)	Credit risk
1	Medium	High	75	Good
2	Low	Low	50	Bad
3	High	Medium	25	Bad
4	Medium	Medium	50	Good
5	Low	Medium	100	Good
6	High	High	25	Good
7	Low	Low	25	Bad
8	Medium	Medium	75	Good

Analyse en Composantes Principales (5 points)

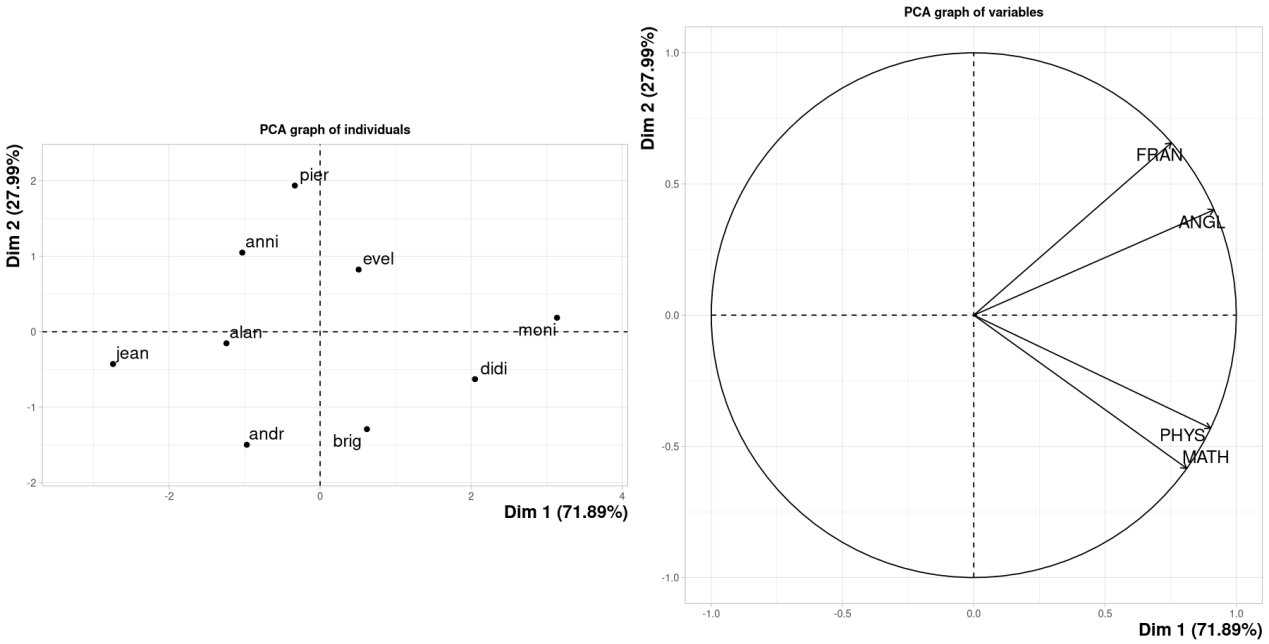
Le jeu de données suivant (provenant de cette présentation) contient les notes de 9 étudiants dans 4 matières :

Nom	MATH	PHYS	FRAN	ANGL
jean	6.00	6.00	5.00	5.50
alan	8.00	8.00	8.00	8.00
anni	6.00	7.00	11.00	9.50
moni	14.50	14.50	15.50	15.00
didi	14.00	14.00	12.00	12.50
andr	11.00	10.00	5.50	7.00
pier	5.50	7.00	14.00	11.50
brig	13.00	12.50	8.50	9.50
evel	9.00	9.50	12.50	12.00

On effectue une ACP (sur les colonnes numériques), puis on affiche le cercle des corrélations ainsi que le nuage des individus.

In [1]:
```
library(FactoMineR) ; options(repr.plot.width=15, repr.plot.height=9)
data <- matrix(
    c(6,8,6,14.5,14,11,5.5,13,9, 6,8,7,14.5,14,10,7,12.5,9.5,
    5,8,11,15.5,12,5.5,14,8.5,12.5, 5.5,8,9.5,15,12.5,7,11.5,9.5,12), ncol=4)
rownames(data) <- c("jean","alan","anni","moni","didi","andr","pier","brig","eve
colnames(data) <- c("MATH", "PHYS", "FRAN", "ANGL")
res <- PCA(data)
```

In [2]:
```
library(gridExtra)
grid.arrange(
    plot(res, choix="ind", cex.axis=1.5, cex=1.3),
    plot(res, choix="var", cex.axis=1.5, cex=1.3), ncol=2)
```



1. A-t-on intérêt à choisir plus de deux axes ? Justifiez. (1 point)
2. Que pouvez-vous déduire de la représentation des variables à droite ? Justifiez. (2 points)
3. En observant les individus extrêmes et en vous aidant du cercle des corrélations, interprétez les axes 1 et 2. (2 points)