Compte-rendu Mathématiques appliquées sur R

Marc BIANCHINI

Résumé

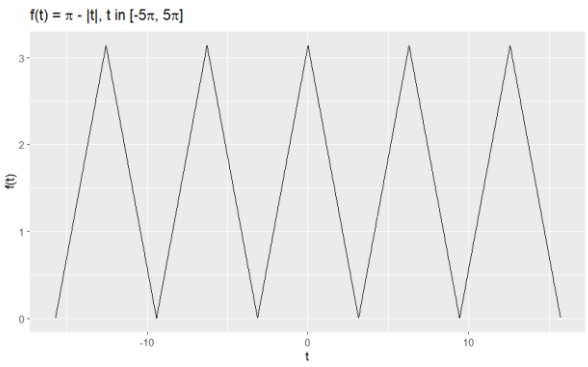
Jean-Marc ZHOU, Charles LOGEAIS, Nadejda DOROSENCO, Duncan LOPES,
Julian ALIZAY

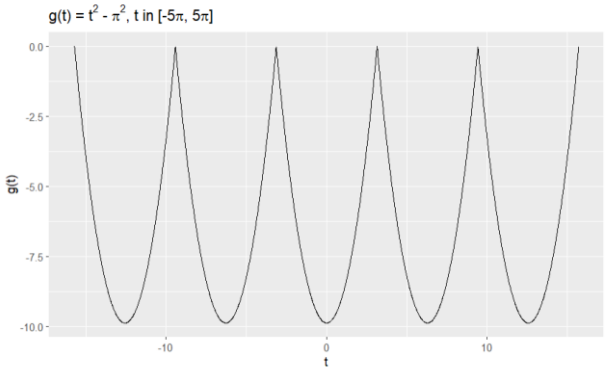
Table des matières

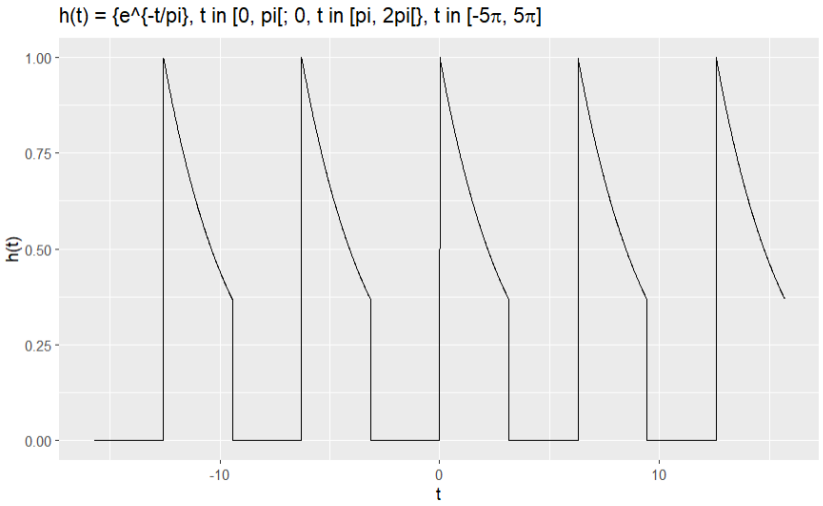
I. Série de Fourier

1. Question 1

Dans cette première question, notre tâche était de représenter graphiquement trois fonctions distinctes sur l'intervalle spécifié [−5π,5π]. L'objectif principal de cette question est d'examiner et de comprendre le comportement et les caractéristiques des fonctions dans cet intervalle spécifique.







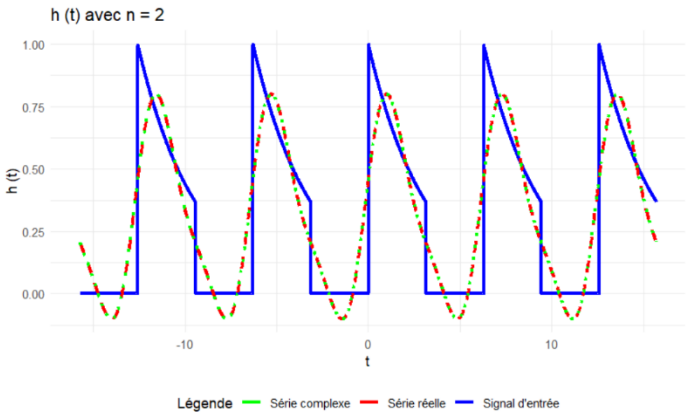
2. Question 2

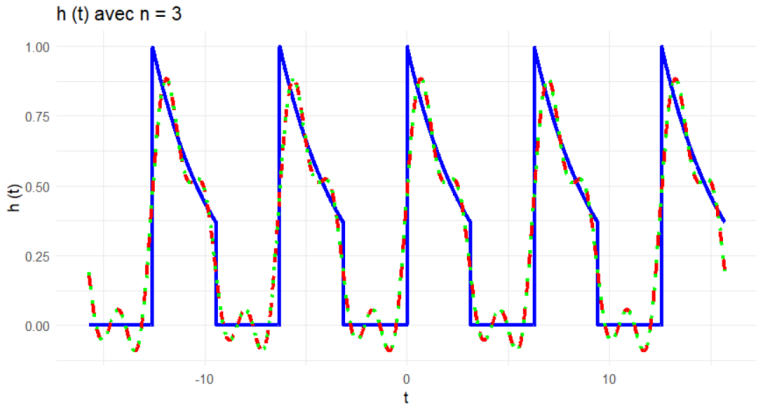
La question 2 exigeait le calcul des coefficients « a_0, a_n , b_n et c_n » en utilisant la fonction integrate de R. Les coefficients « a_0 » ont été calculés en appliquant la formule intégrale correspondante sur l'intervalle défini. Pour les coefficients « a_n, b_n et c_n », l'utilisateur a saisi une valeur de n, et les calculs ont été effectués en conséquence. Les captures d'écran ci-dessous illustrent les codes R utilisés et les résultats obtenus pour différentes valeurs de n. Pour n = 10 :

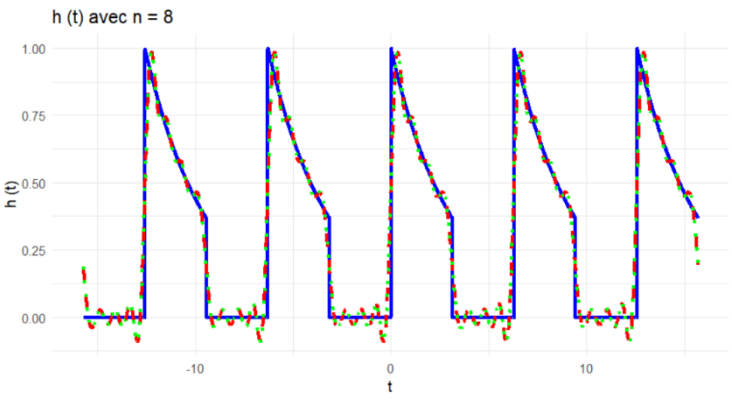
```
[1] "Pour f(t):"
[1] "a0: 1.5707963267949"
[1] "an: -4.50578579873998e-16"
[1] "bn: 0"
[1] "cn: -4.50578579873998e-16+0i"
[1] "Pour g(t):"
[1] "a0: -6.5797362673929"
[1] "an: 0.0400000000000025"
[1] "bn: 0"
[1] "cn: 0.0400000000000025+0i"
[1] "Pour h(t):"
[1] "a0: 0.316060279414279"
[1] "an: 0.000639823755249521"
[1] "bn: 0.0201006560908442"
[1] "cn: 0.0006398237552495-0.02010065609084421"
```

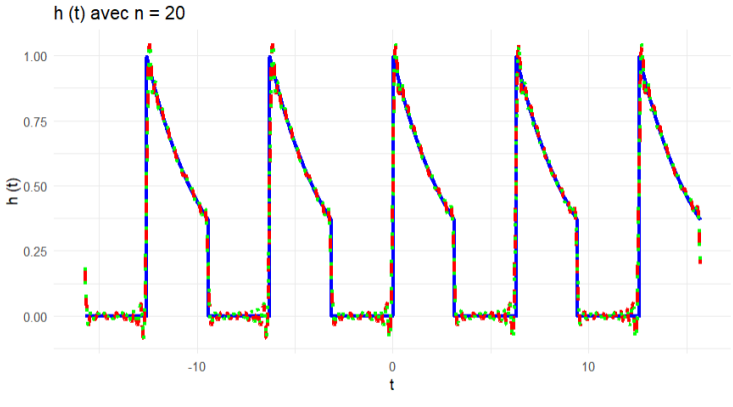
3. Question 3

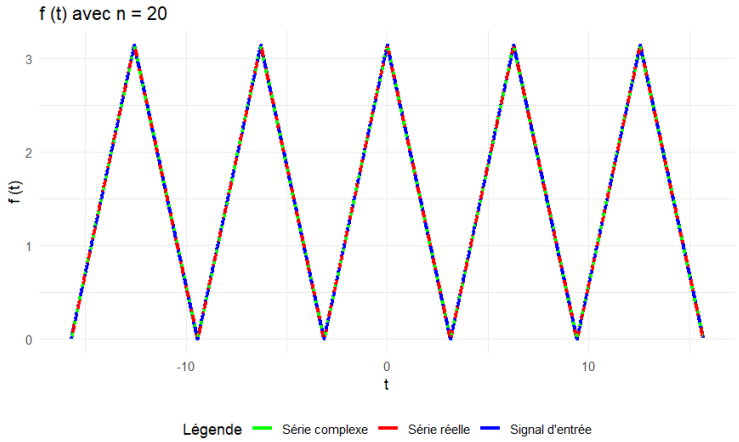
La question 3 nous a amenés à examiner les sommes des séries réelles et complexes, tronquées pour n=2,3,8 et 20, et à les superposer sur le signal d'entrée original. Les graphiques générés révèlent les différences et les similitudes entre les signaux d'entrée et les sommes tronquées des séries à différents niveaux de n. En augmentant la valeur de n, nous observons une convergence accrue des sommes tronquées vers le signal d'entrée original, reflétant ainsi la précision croissante de l'approximation. Les graphiques détaillés, présentant les signaux d'entrée et les sommes des séries superposées, sont fournis ci-dessous. Voici les 4 résultats pour la fonction H, nous mettons juste les derniers résultats pour les 2 autres fonctions.

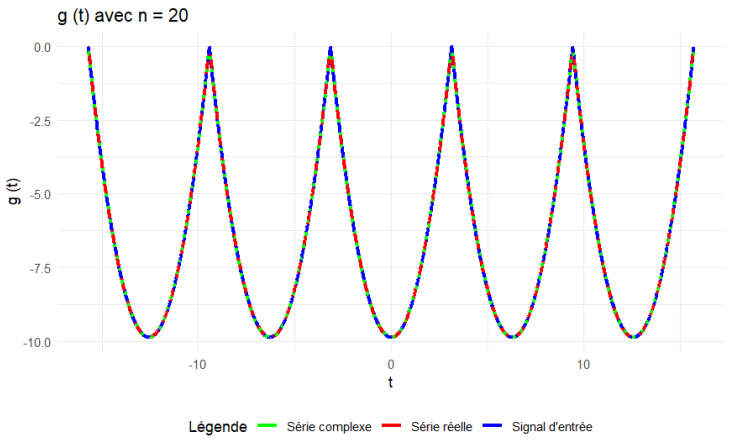






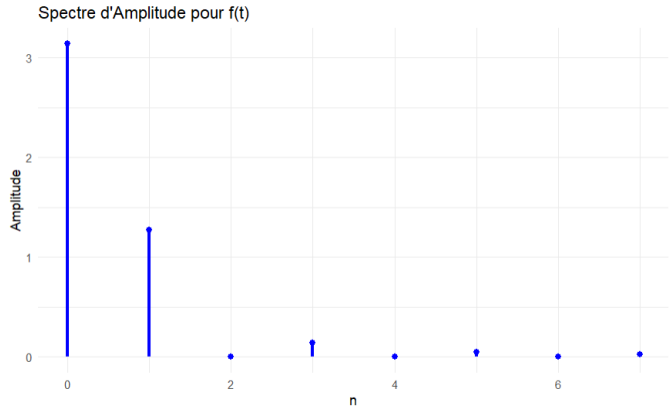


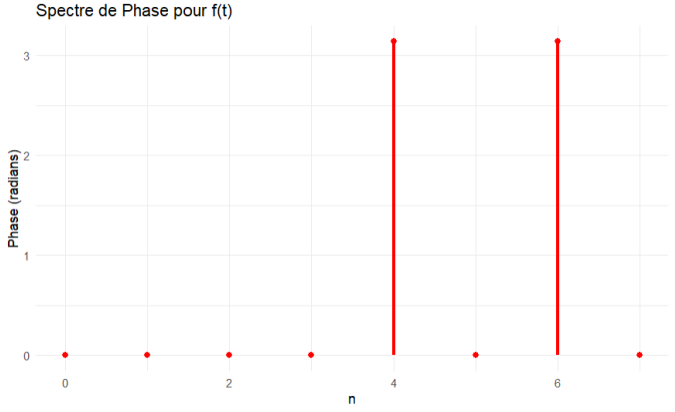


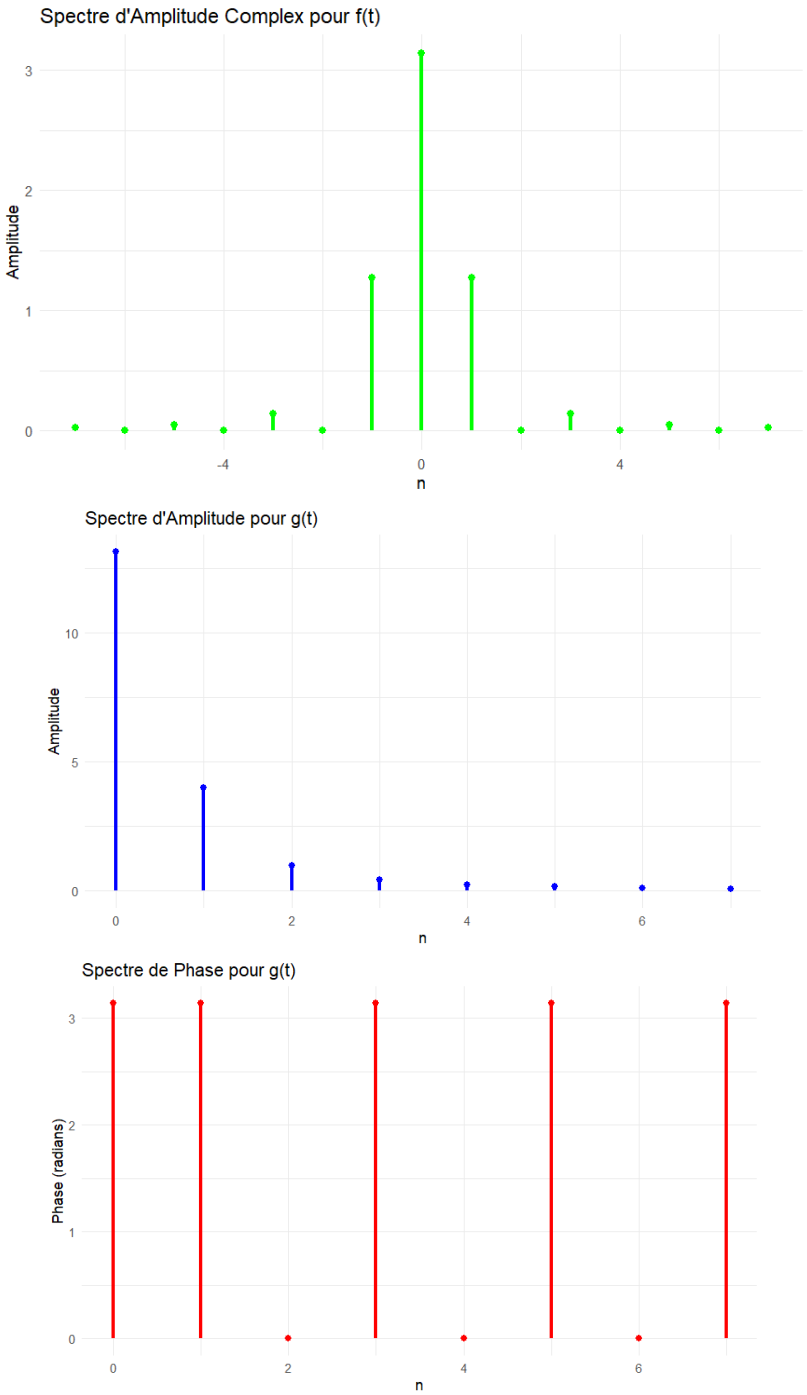


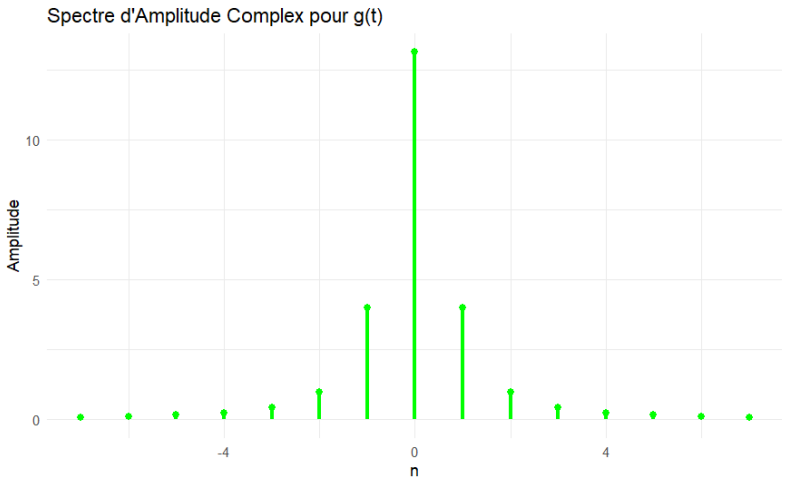
4. Question 4

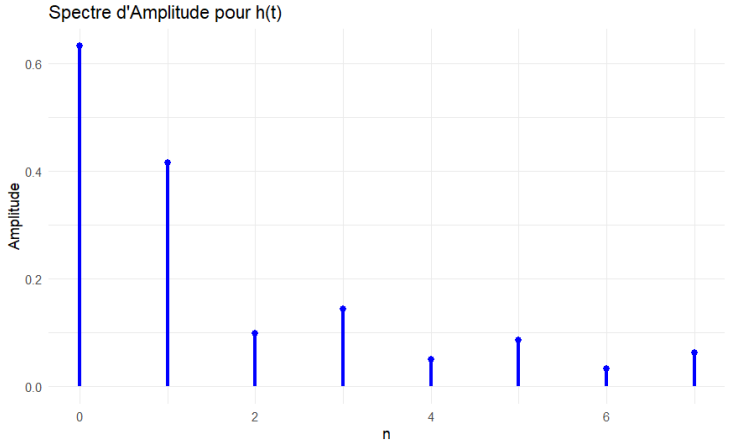
Les spectres d'amplitudes et de phase pour les séries réelles ont été représentés jusqu'à 7 pulsations, et les spectres d'amplitudes pour les séries complexes ont été illustrés entre -7 et 7 pulsations. L'analyse des spectres d'amplitude révèle que les amplitudes varient significativement avec le nombre de pulsations pour les trois fonctions, indiquant la prédominance de certaines fréquences dans les signaux. Les spectres de phase, d'autre part, offrent un aperçu des décalages de phase associés à chaque composant fréquentiel. En comparant les spectres d'amplitudes des séries réelles et complexes, des similitudes et des différences notables sont observées, suggérant que les deux types de séries capturent différentes caractéristiques des signaux.

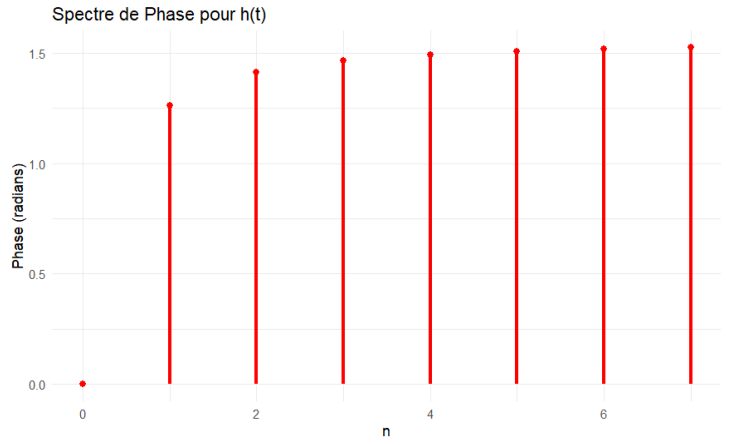


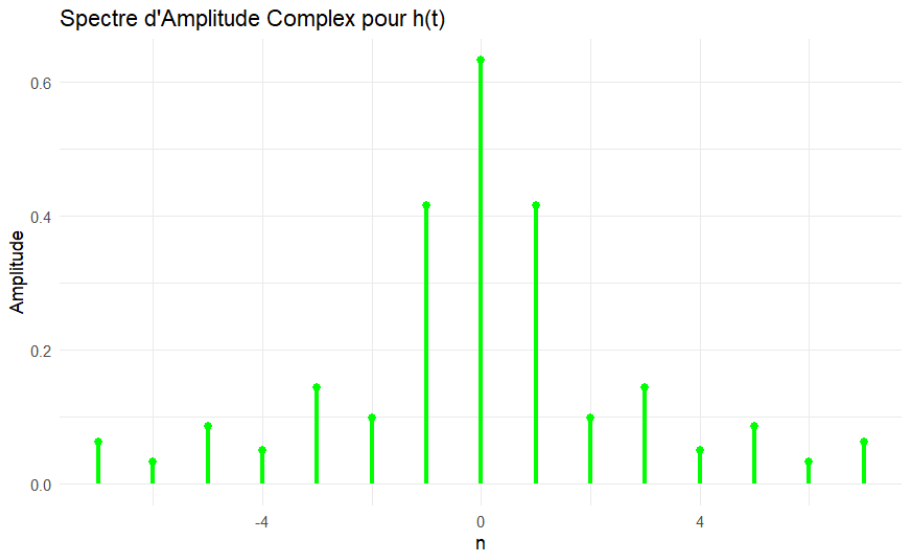



Spectre d'Amplitude Complex pour f(t)
Amplitude
n
Spectre d'Amplitude pour g(t)
Amplitude
n
Spectre de Phase pour g(t)
Phase (radians)
n


Spectre d'Amplitude Complex pour g(t)
Amplitude
10
5
0
-4
0
4
n


Spectre d'Amplitude pour h(t)
Amplitude
0.6
0.4
0.2
0.0
0
2
4
6
n


Spectre de Phase pour h(t)
Phase (radians)
1.5
1.0
0.5
0.0
0
2
4
6
n



II. Transformée de Fourier

A. Transformée de Fourier (TF) et spectre

Soient les signaux suivants :

$$x(t) = rect(\frac{t-\pi}{2\pi}), \text{ avec } rect(u) = \begin{cases} 1, & |u| \leqslant \frac{1}{2} \\ 0, & |u| > \frac{1}{2} \end{cases}$$

$$y(t) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}, \quad \text{et} \quad z(t) = \cos^2(2\pi f_0 t), \text{ tester avec } f_0 = 1, 2 \text{ et } 3$$

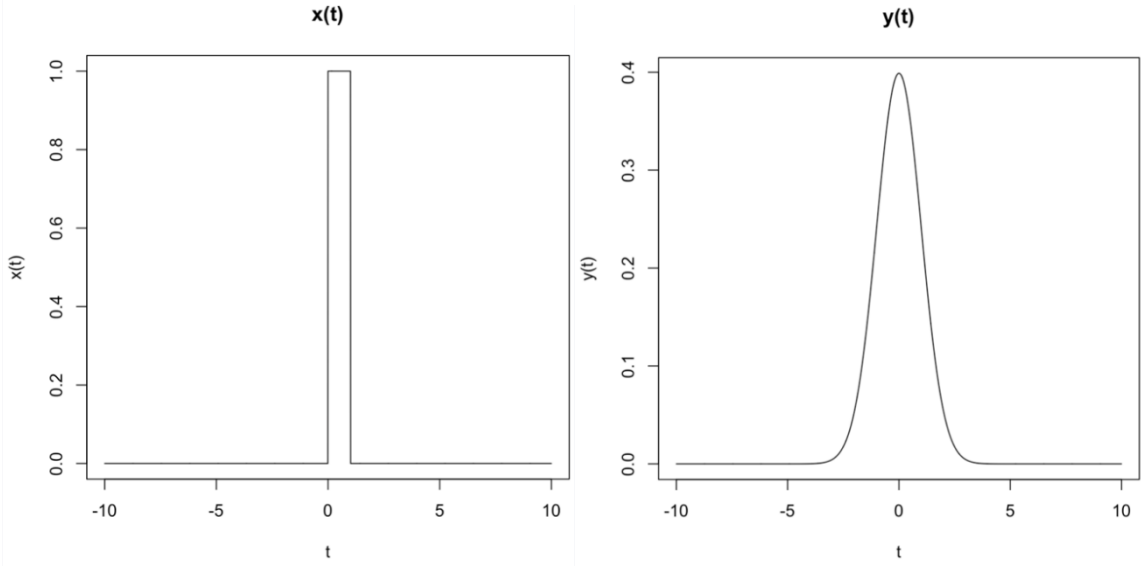
1) Pour tracer ces différents signaux nous avons défini t allant de -10 à 10 avec un pas de 0.01 puis nous avons défini des fonctions pour chaque signal :

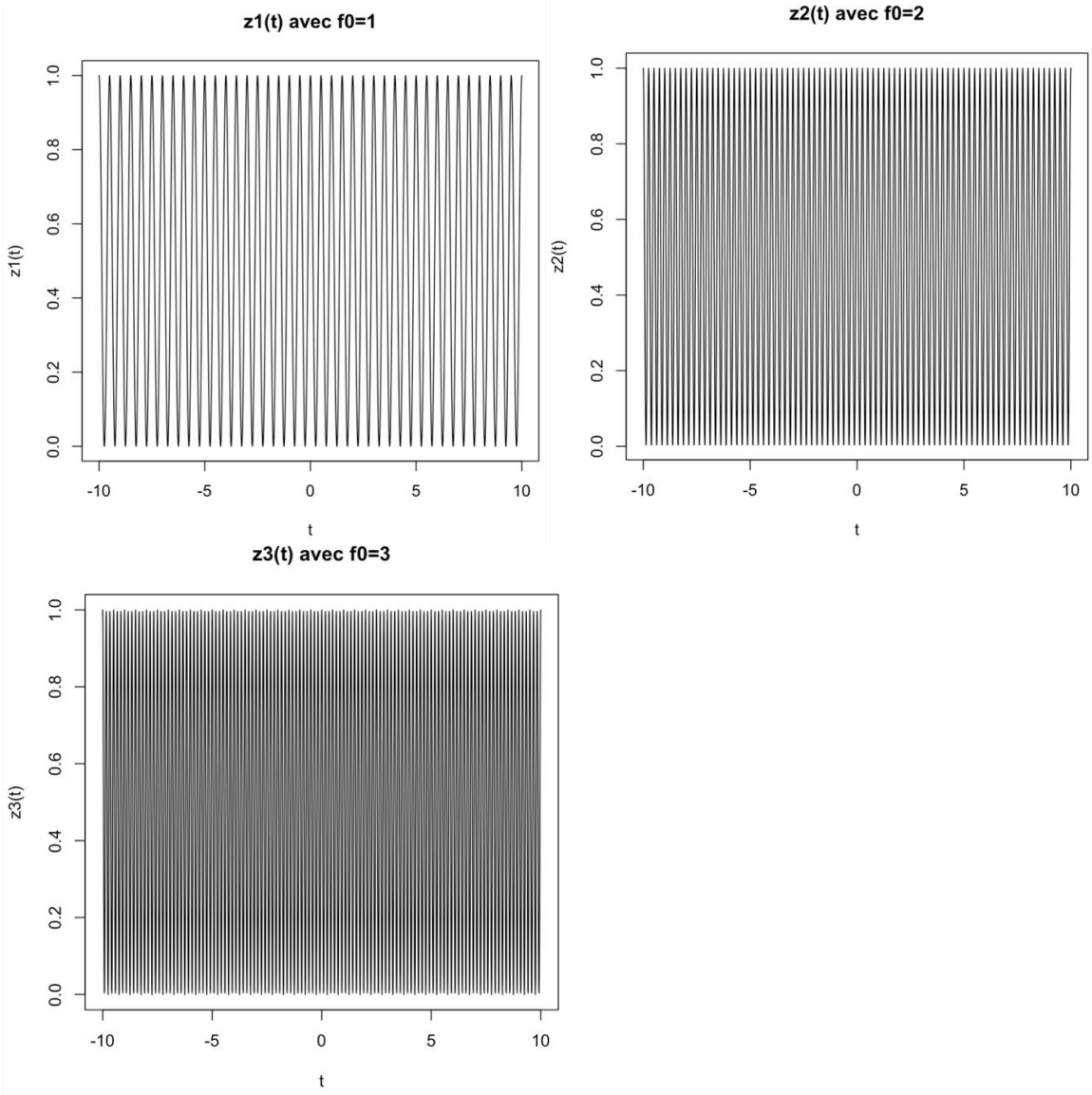
```r
t <- seq(-10, 10, by=0.01)

# Fonction rect(u)
rect <- function(u) {
  ifelse(abs(u) <= 0.5, 1, 0)
}

# Signal x(t)
x_t <- function(t) {rect((t - pi / (2 * pi)))}
# Signal y(t)
y_t <- function(t) exp(-t^2/2)/sqrt(2*pi)
# Signal z(t) pour différentes valeurs de f0
z1_t <- function(t) cos(2 * pi * 1 * t)^2
z2_t <- function(t) cos(2 * pi * 2 * t)^2
z3_t <- function(t) cos(2 * pi * 3 * t)^2
```

Nous avons ensuite utilisé la fonction plot pour afficher tous ces signaux :





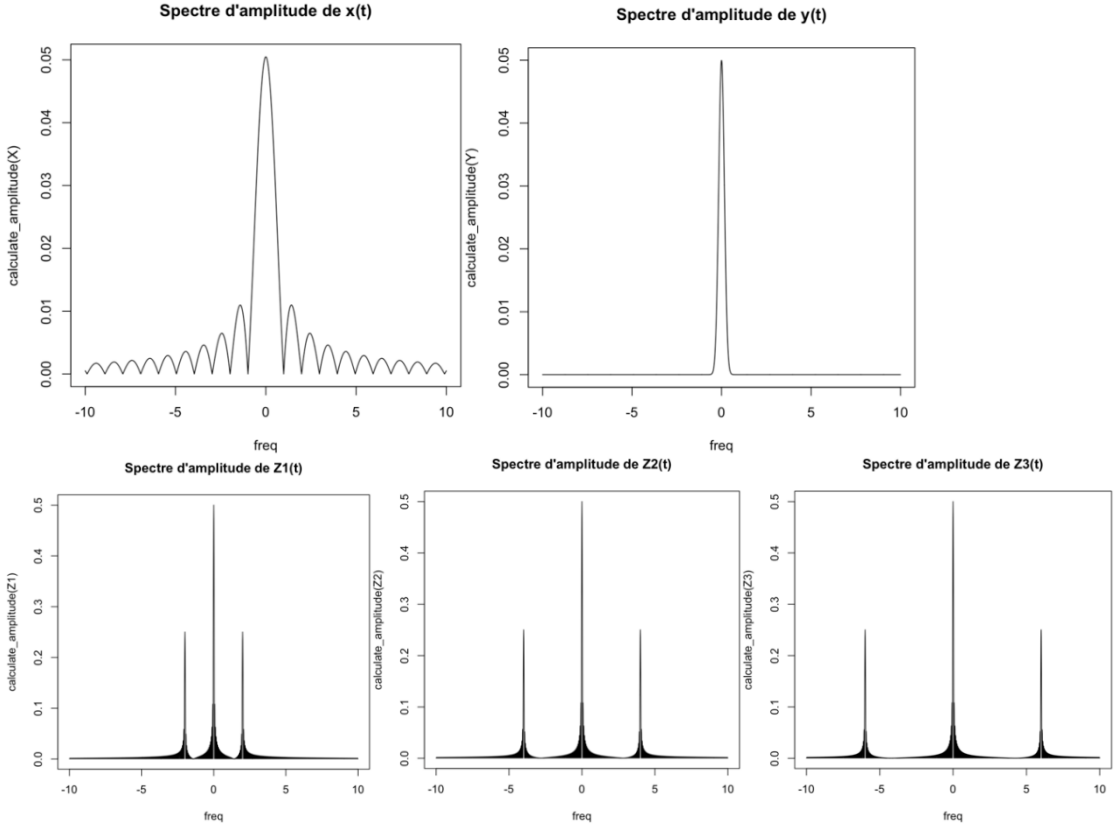
2) Nous avons ensuite implémenté une méthode de calcul de TF avec la méthode des sommes :

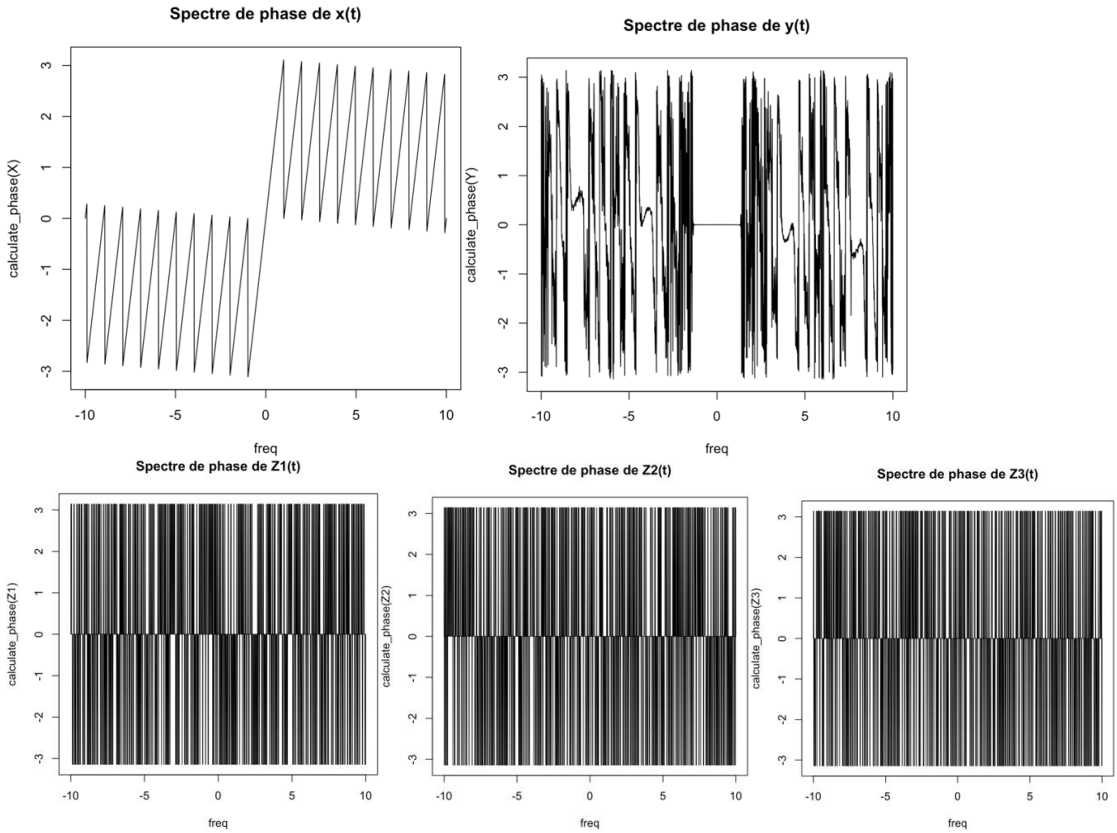
```r
TF <- function(signal, t, f) {
  # Define the real parts of the Fourier transform
  XRe <- function(f) {
    n <- length(t)
    real_part <- signal(t) * cos(2 * pi * f * t)
    result <- sum(real_part) / n
    return(result)
  }
  # Define the imaginary parts of the Fourier transform
  XIm <- function(f) {
    n <- length(t)
    imag_part <- signal(t) * sin(2 * pi * f * t)
    result <- sum(imag_part) / n
    return(result)
  }

  # Calculate the Fourier transform
  X <- complex(real = XRe(f), imaginary = XIm(f))

  return(X)
}
```

Enfin nous avons appliqué la TF à chacun des signaux pour pouvoir tracer leurs spectres d'amplitude et de phase et voici ce que nous avons obtenu :





Les spectres d'amplitude de x(t) et y(t) montrent des pics à 0, tandis que pour z(t), l'augmentation de la valeur de f0 entraine un éloignement des pics de gauche et droite de manière symétrique du pic qui est en 0, ce qui est normal étant donné que l'on a une fonction cosinus.

B. TF inverse

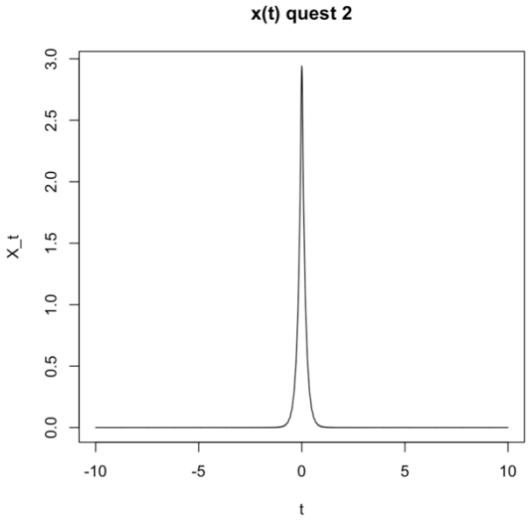
1) On a implémenté cette fois-ci une méthode de calcul d'une TF inverse toujours avec la méthode des sommes :

```
TF_inverse <- function(X, f, t) {
  delta_f <- f[2] - f[1]
  signal <- Re(sum(X * exp(1i*2*pi*f*t) * delta_f))
  return(signal)
}
```

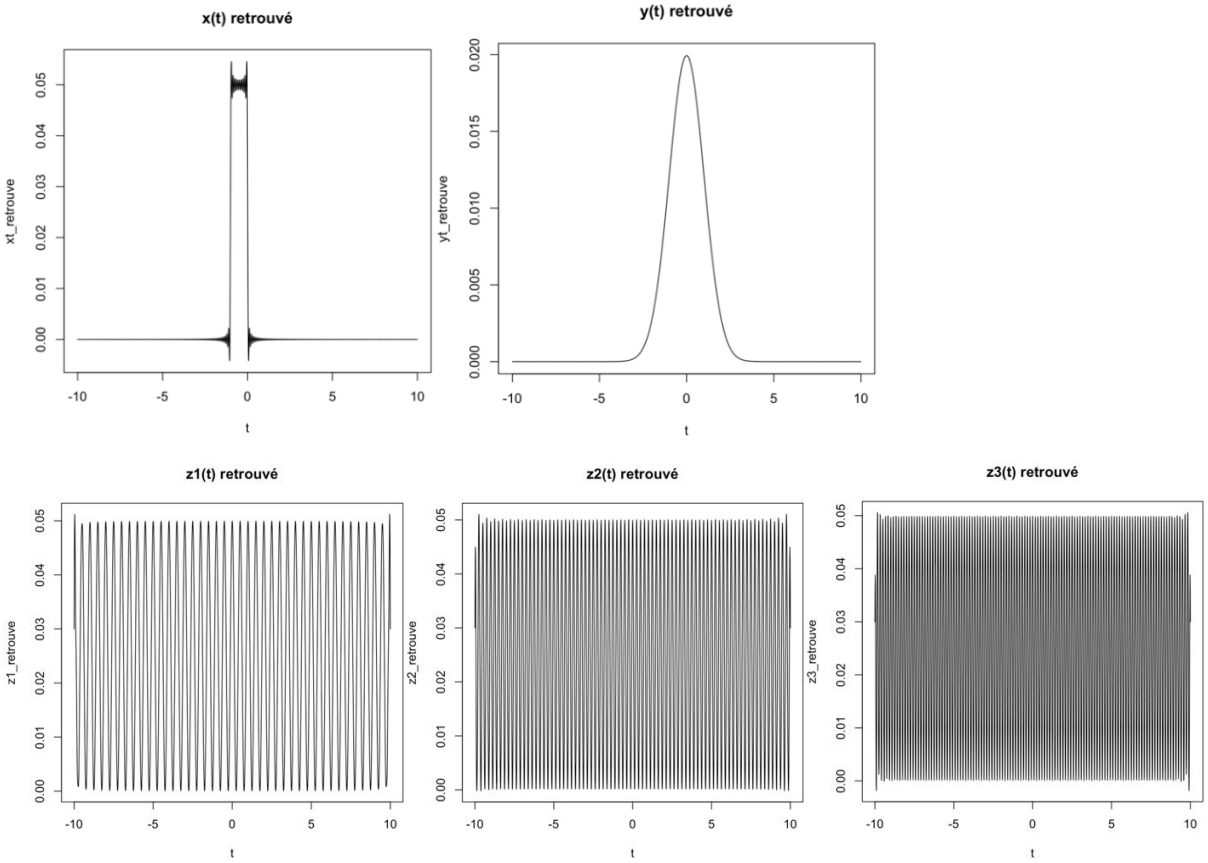
2)

$$\text{Soit } X(f) = \frac{1}{1+f^2}$$

On a trouvé grâce notre fonction de TF inverse, x(t) :



3) En appliquant la fonction de TF inverse sur les TF obtenus à l'exercice précédent, nous avons pu retrouver les signaux de départ :



C. <u>Transformée de Fourier d'un signal discret (TFTD) et transformée de Fourier discrète (TFD)</u>

La TFTD est la TF d'un signal échantillonné (une suite de nombres). Le calcul direct de la TFTD reste le même que pour le TF à la différence que l'on a une somme. La formule est la suivante :

TFTD :

$$X(f) = \sum_{n=-\infty}^{+\infty} x(n)exp(-2i\pi nf),$$

$$x(n) = \int_{-\frac{1}{2}}^{\frac{1}{2}} X(f)exp(2i\pi nf)df$$

1. TFD

Introduction
La transformation de Fourier discrète (TFD) est un outil essentiel pour l'analyse des signaux dans le domaine fréquentiel. Dans ce projet, nous nous intéressons à la TFD, qui permet de représenter un signal temporel dans le domaine fréquentiel. Pour valider notre approche, nous avons réalisé des tests sur deux signaux distincts.

---

1. Méthodologie
La TFD est définie par une relation qui convertit un signal temporel en une séquence de composantes fréquentielles. Pour valider notre implémentation, nous utilisons la méthode des moindres carrés pour comparer le signal original avec le spectre fréquentiel obtenu. Les signaux choisis pour les tests sont des combinaisons linéaires de fonctions sinus et cosinus.

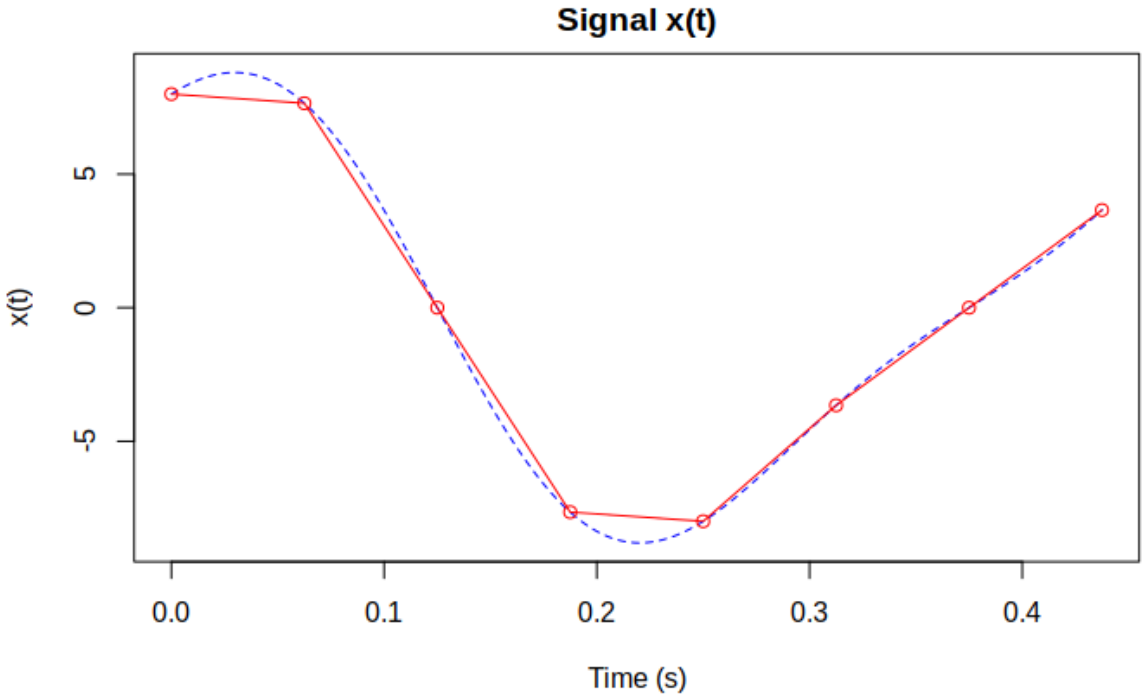
La formule pour calculer la TFD est :

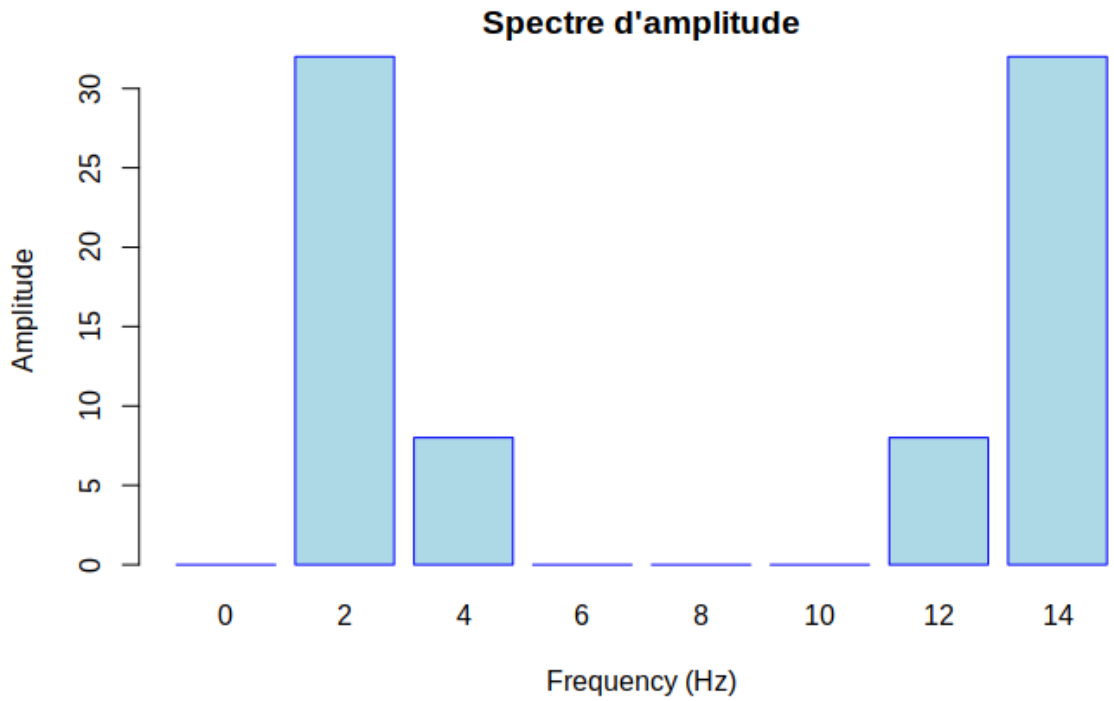
$$X(k) = \sum_{n=0}^{N-1} x(n)exp(-2i\pi kn/N), \quad 0 \leqslant k < N$$

Signal 1 : Description et Résultats
Pour le premier signal, la fréquence d'échantillonnage fe est de 16 Hz avec N=8 échantillons. Le signal est défini par l'équation :

x(t)=2sin(8πt)+8cos(4πt)

Résultats :



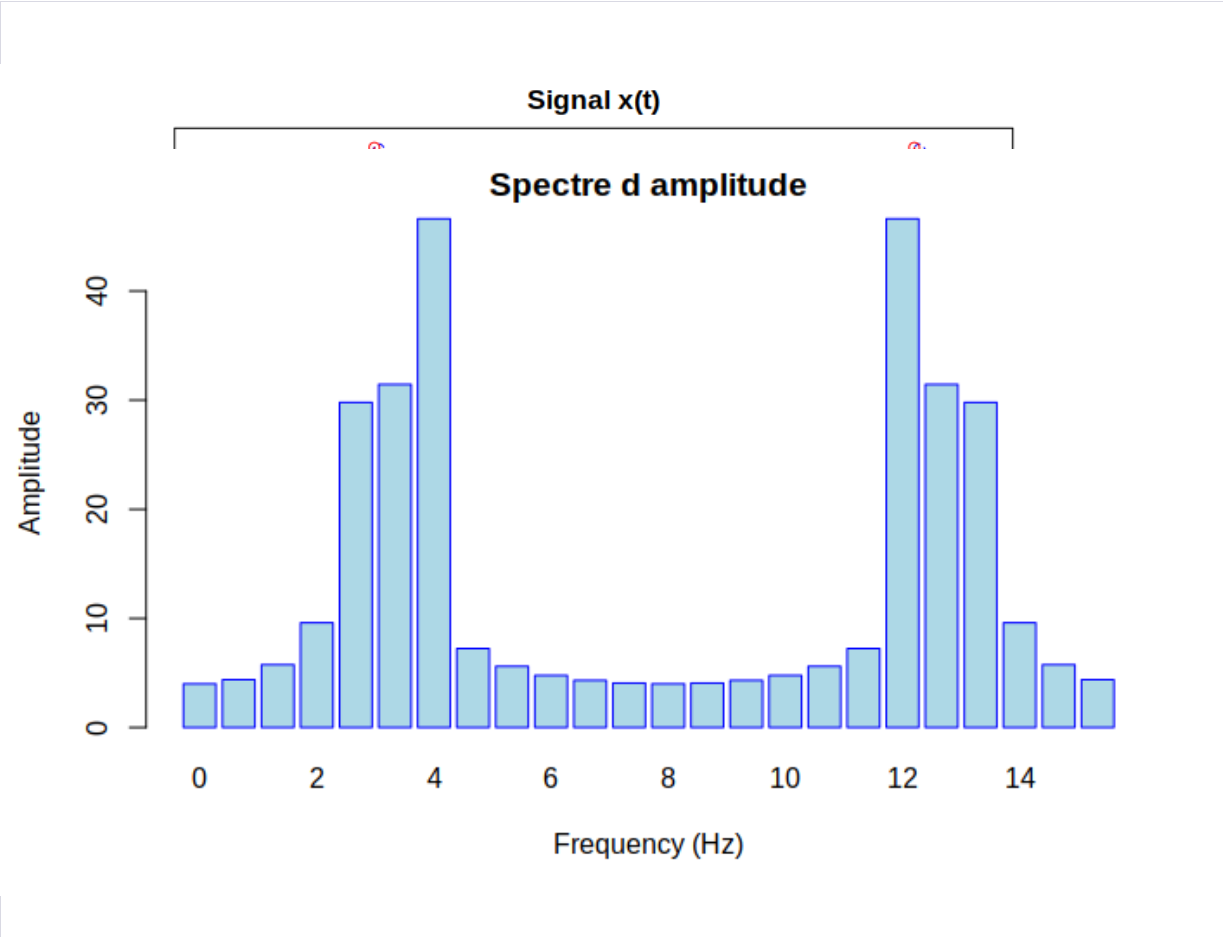


Signal 2 : Description et Résultats

Pour le second signal, la fréquence d'échantillonnage fe est de 16 Hz avec N=24 échantillons. Le signal est défini par l'équation :

$$x(t)=3\sin(8\pi t)+4\cos(6\pi t)$$

Résultats :



---

2. TFD inverse

Le TFD inverse est la fonction qui permet de retrouver le signal original après une TFD. À partir d'une séquence X[k], soit les composantes fréquentielles, nous retrouvons la séquence x[n].

Nous appliquons la formule suivante :

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] \cdot e^{j(2\pi/N) \cdot k \cdot n}$$

Nous effectuons quelques tests pour vérifier le bon fonctionnement de notre fonction.

```
[1] 10+0i -2+2i -2+0i -2-2i
[1] 1 2 3 4
```

Sortie console de notre fonction TFD inverse sur RStudio

Pour vérifier la différence entre le signal initial et notre nouveau signal initial, nous utilisons la méthode des moindres carrés. Testons notre fonction sur les signaux vus précédemment.

Signal 1 : nous obtenons une erreur à 10^{-10}.

$$x(t) = 2\sin(8\pi t) + 8\cos(4\pi t)$$

```
[1] "RMSE:  3.58941506137787e-10"
[1] "séquence originale x_n:"
[1]  8.000000e+00  7.656854e+00  7.347638e-16 -7.656854e+00 -8.000000e+00 -3.656854e+00 -7.347638e-16  3.656854e+00
[1] "séquence récupérée x_recovered:"
[1]  8.000000  7.656854  0.000000 -7.656854 -8.000000 -3.656854  0.000000  3.656854
```

Signal 2 : nous obtenons une erreur à 10^{-9}.

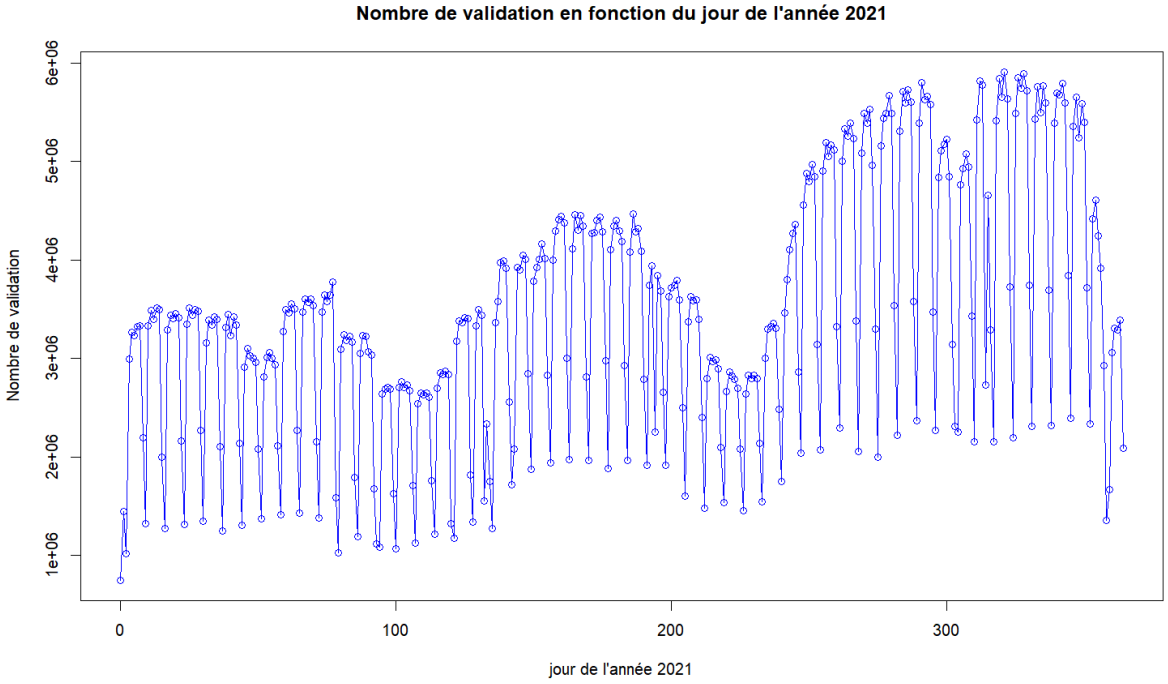
$$x(t) = 3\sin(8\pi t) + 4\cos(6\pi t)$$

```
[1] "RMSE:  2.38849178733197e-09"
[1] "Séquence originale x_n:"
 [1]  4.000000e+00  4.530734e+00 -2.828427e+00 -6.695518e+00 -1.469528e-15  6.695518e+00  2.828427e+00 -4.530734e+00 -4.000000e+00  1.469266e+00
[11]  2.828427e+00  6.955181e-01  0.000000e+00 -6.955181e-01 -2.828427e+00 -1.469266e+00  4.000000e+00  4.530734e+00 -2.828427e+00 -6.695518e+00
[21] -1.445307e-14  6.695518e+00  2.828427e+00 -4.530734e+00
[1] "Séquence récupérée x_recovered:"
 [1]  4.0000000  4.5307337 -2.8284271 -6.6955181  0.0000000  6.6955181  2.8284271 -4.5307337 -4.0000000  1.4692663  2.8284271  0.6955181
[13]  0.0000000 -0.6955181 -2.8284271 -1.4692663  4.0000000  4.5307337 -2.8284271 -6.6955181  0.0000000  6.6955181  2.8284271 -4.5307337
```

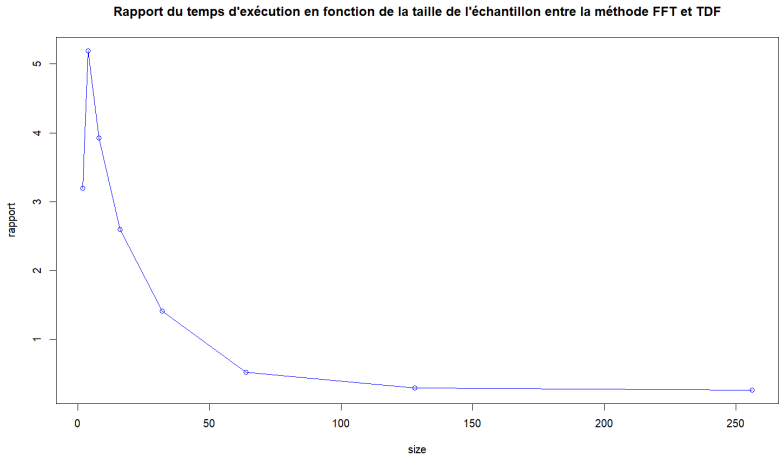
3. FFT

L'objectif de l'exercice est d'implémenter une version de la FFT (Fast Fourier Transform) qui est plus rapide que la TFD classique avec une complexité en O(nlog2n) au lieu de O(n2). On testera notre implémentation sur un cas réel afin de vérifier que l'on obtient bien les mêmes résultats entre la TDF et la FFT mais aussi pour la complexité annoncée.

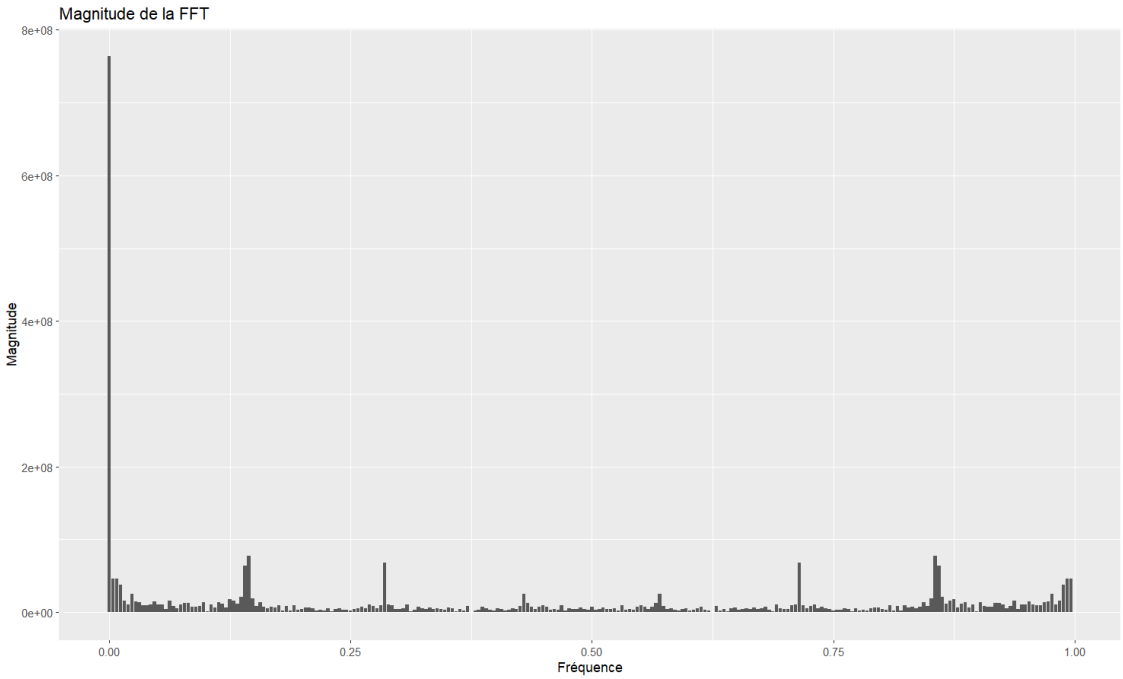
On a choisi comme jeu de test réel le nombre de validation de titre de transport par jour sur le réseau de transport d'île de France. Une représentation graphique des données ci-dessous.

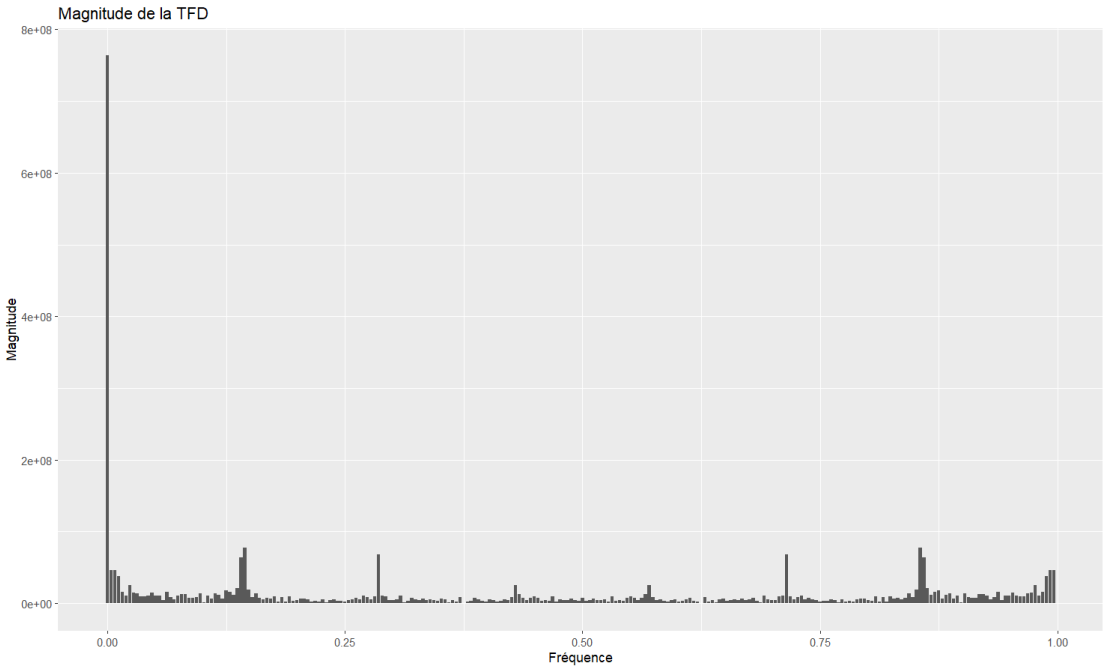


On a implémenté comme algorithme de FFT, cooley tukey qui a été publié en 1965. Elle possède la contrainte d'avoir une taille d'échantillon en puissance de 2. Afin de vérifier la complexité attendue on a tracé le rapport entre le temps d'exécution de la FFT sur la TDF pour 8 différentes valeurs d'échantillon N (2, 4, 8, 16, 32, 64, 128, 256). Théoriquement on devrait obtenir une courbe de la forme $\frac{log2n}{n}$. C'est bien ce que l'on obtient !



Ci-dessous les résultats de la TDF et FFT sur un échantillon de 256 jours





Nos résultats sont identiques et le temps d'exécution plus rapide pour la FFT ce qui confirme la bonne implémentation de nos algorithmes. Notre période d'échantillonnage étant de 1 jour, la fréquence la plus élevée que l'on puisse obtenir est de 0.5 (car il faut 2 points minimum pour observer une oscillation, théorème de shannon) ce qui correspond bien au milieu du graphique (il y a une symétrie en 0.5).

En observant nos données brutes, on peut remarquer une périodicité hebdomadaire flagrante. C'est-à-dire une fréquence de 1/7 = 0.14 qui est bien présente sur le graphique. 2

autres fréquences notables sont représentées sur le graphique pour 0.29 (période de 3.5 jours) et entre 0.004 et 0.011 (période entre 90 et 250 jours) plus difficilement observable depuis les données initiales.

III. Khi-deux

On souhaite réaliser une étude de marché pour ouvrir notre restaurant. Il est nécessaire d'étudier le nombre de clients minimum pour assurer la pérennité du business. Nous avons pour cela des fréquences observées pour les clients, soit des données réelles. Nous possédons aussi des fréquences théoriques que nous avons fixé.

Jour	Lundi	mardi	mercredi	Jeudi	Vendredi	Samedi
Théoriques (%)	10	10	15	20	30	15
Observés	30	14	34	45	57	20

Tableau de données des fréquences

Pour vérifier si ces fréquences sont pertinentes, nous allons réaliser un test paramétrique du khi-deux avec un seuil de signification α = 5%. Puisqu'il s'agit d'un test paramétrique, nous allons émettre des hypothèses, et vérifier avec la p-value retournée par le test du khi-deux quelle hypothèse nous allons privilégier. Afin de nous assurer de sa validité, il est important de posséder au moins 5 modalités dans nos variables (ici les jours de la semaine) afin que le résultat retourné par le test soit cohérent et scientifiquement valide.

H_0 : Les fréquences théoriques et observées ne sont pas corrélées.

H_1 : Les fréquences théoriques et observées sont corrélées.

Nous partons sur le principe que l'hypothèse H_0 est notre hypothèse de base.

Résultat : P-value : 0.04329 ou 4.329% < 5%

La p-value représente les chances de se tromper en rejetant l'hypothèse H_0, dans notre cas, nous avons 4.329% de nous tromper en rejetant H_0 et en acceptant l'hypothèse H_1. Nous avions établi un seuil au préalable pour lequel nous rejetons l'hypothèse H_0 qui est de 5%. Ainsi, nous affirmons que les fréquences théoriques et observées sont corrélées.

Nous avons de bonnes raisons de croire qu'il est possible d'ouvrir un restaurant, la distribution de clients dans le secteur concerné est favorable à un business de type alimentaire.

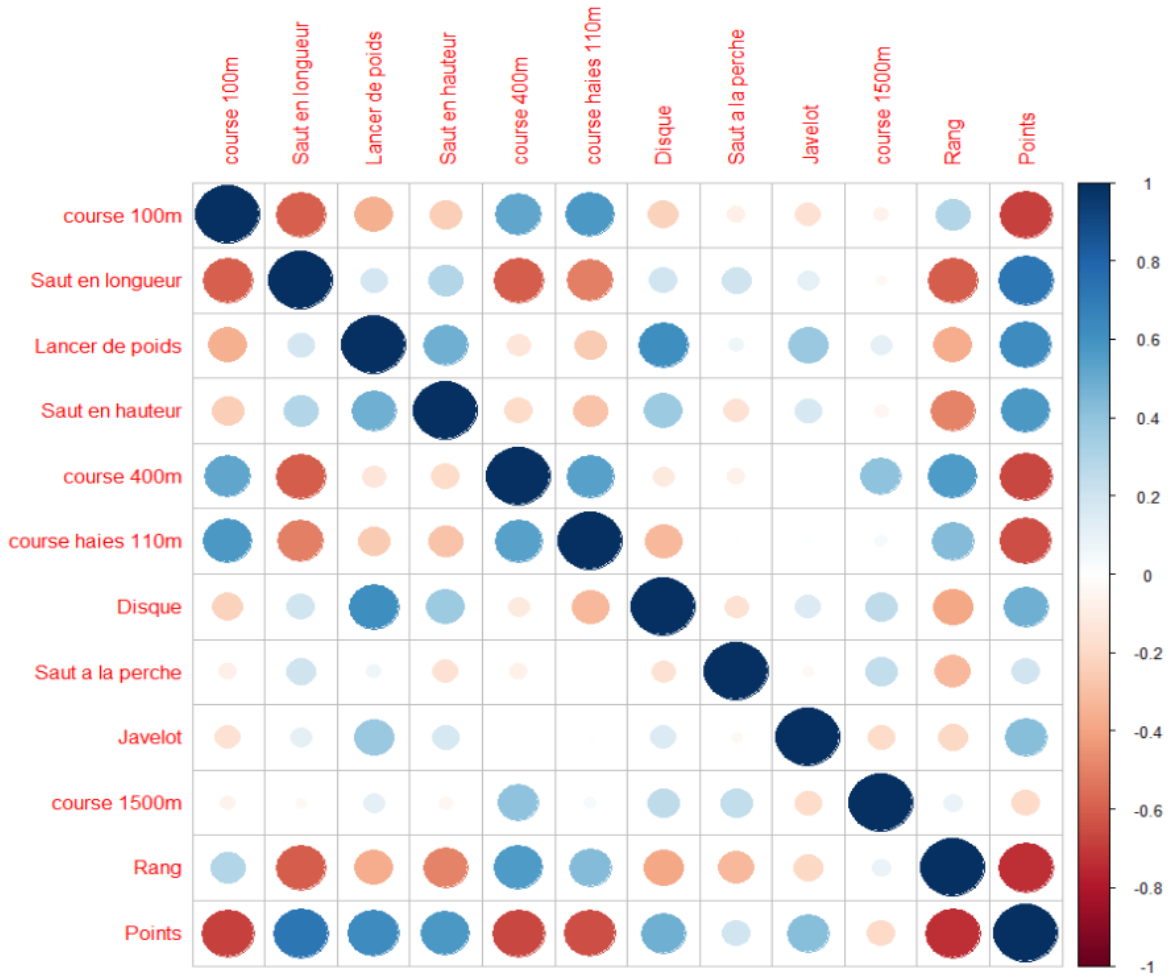
IV. ACP

Sur un dataset « decathlon » qui recense les performances de plusieurs athlètes sur des épreuves athlétiques. Notre dataset comporte 41 observations et 13 colonnes (12 quantitatives et 1 qualitative).

Nous souhaitons voir s'il existe des corrélations dans notre tableau de données. On rappelle que la matrice de corrélation ne prend en compte que des variables quantitatives, donc nous enlevons la variable qualitative COMPET qui indique si la performance a été réalisé durant des Jeux Olympiques ou un Decastar(compétition annuelle).

A. Analyse rapide

1. Matrice de corrélation



Matrice de corrélation du dataset decathlon

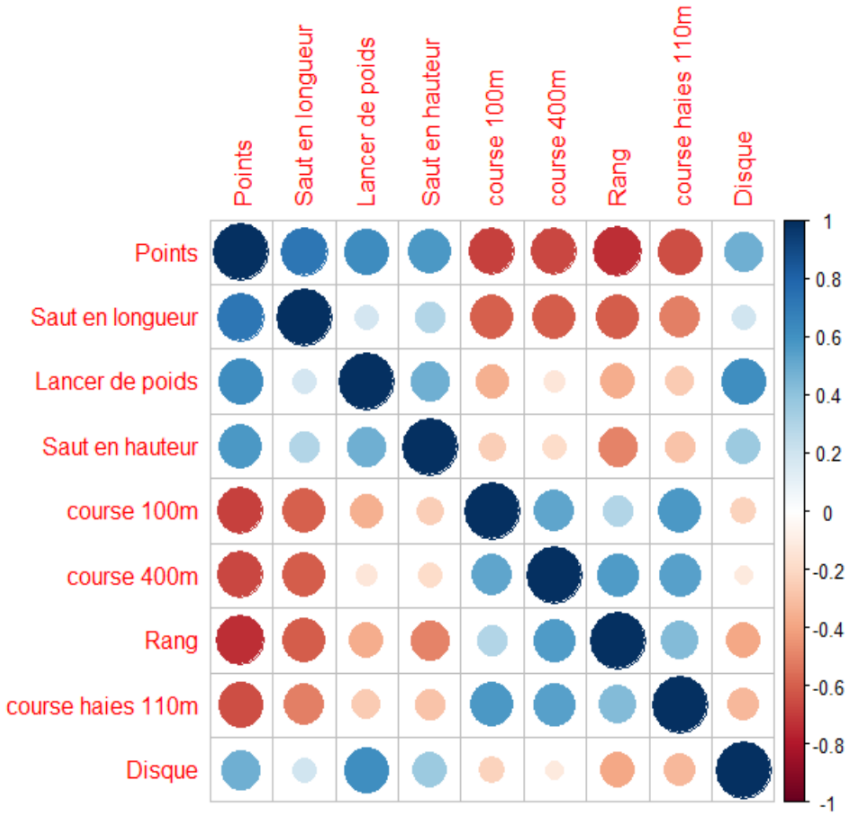
2. Relations entre les variables

La matrice de corrélation retourne des relations entre les variables allant de -1 à 1. Nous allons vous expliquer comment interpréter ce graphique :

- Lien proche de 1 : corrélation forte entre les variables
- Lien proche de 0 : corrélation faible entre les variables
- Lien proche de -1 : corrélation opposée forte entre les variables (c'est-à-dire lorsque la variable a est fortement positive, la variable opposée b sera fortement négative)

Voici notre interprétation du graphique :

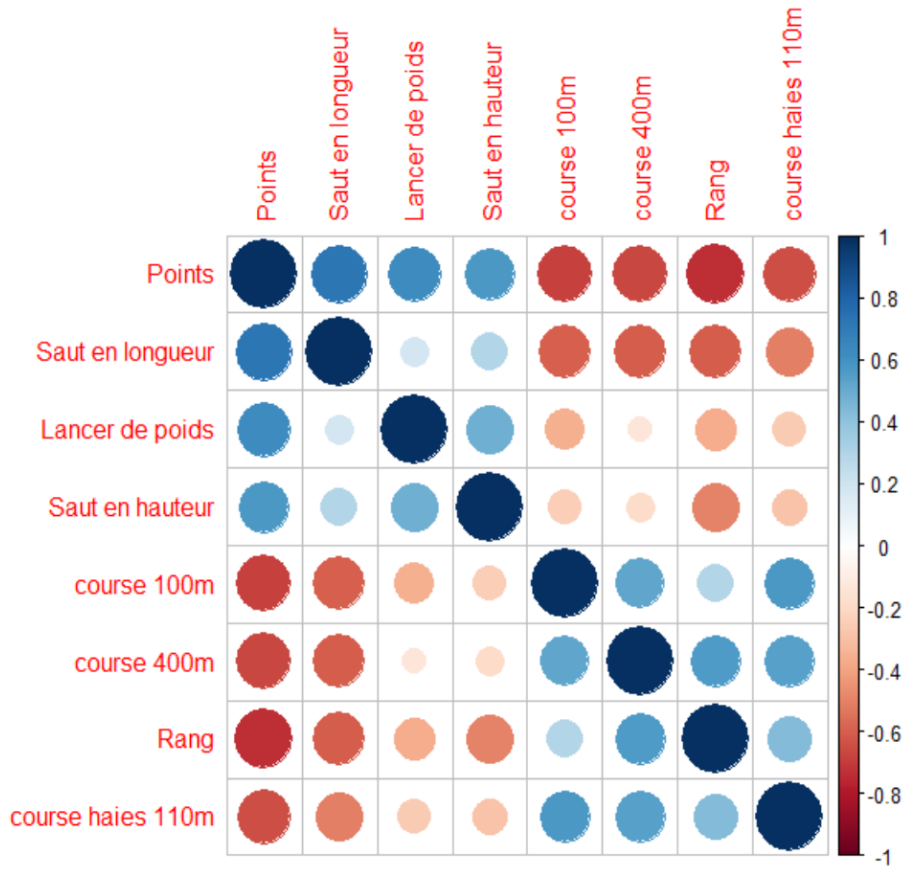
- Variables fortement corrélées positivement (autour de 0.5) :
  - (Points, Saut en longueur)
  - (Points, Lancer de poids)
  - (Points, Saut en hauteur)
  - (Course 100 mètres, Course 400 mètres)
  - (Rang, Course 400 mètres)
  - (Course à haies 110 mètres, Course 100 mètres)
  - (Lancer de disque, Lancer de poids)



Matrice des variables corrélées positivement

- Variables corrélées négativement (ou opposées) :
  - (Saut en longueur, Course 100 mètres)
  - (Course 400 mètres, Saut en longueur)
  - (Course à haies 110 mètres, Saut en longueur)

- (Rang, Saut en longueur)
- (Rang, Saut en hauteur)
- (Points, Course 100 mètres)
- (Points, Course 400 mètres)
- (Points, Course à haies 110 mètres)
- (Points, Rang)



Matrice des variables corrélées négativement

3. Groupement de variables

Comme vu précédemment, nous regroupons les variables de la matrice de corrélation selon leur signe. Des variables positivement corrélées représentent des variables qui suivent le même sens de variation, par exemple si la taille et le poids sont corrélés on peut en déduire que plus un humain est grand et plus son poids sera élevé.

Dans notre cas, sur la matrice de corrélation plus le point est bleu foncé et plus il y a une corrélation positive. Ainsi, lorsqu'un athlète est bon en saut en longueur il a aussi tendance à gagner beaucoup de points.

Le deuxième groupement important est celui des valeurs négativement corrélées. Dans le même principe, il s'agit d'un lien négatif entre deux variables : plus un individu est grand et

moins son poids sera élevé. Trouver ces variables est essentiel, ce sont des tendances dans un jeu de données et discriminants.

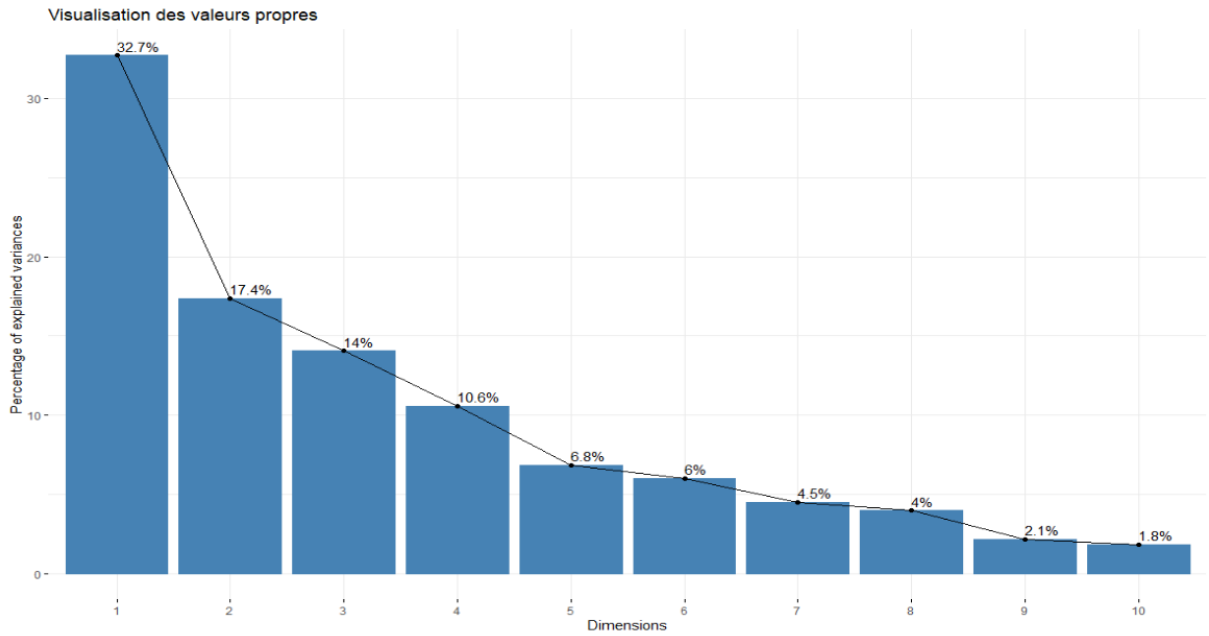
Dans nos données nous avons Rang et Points en tant que variables très négativement corrélées, ce qu'il faut comprendre c'est que plus un rang est élevé, moins de points l'athlète aura. Ce qui est logique car plus son rang dans le classement est bas et plus il a de points dans ce classement.

Nous avons un dernier groupement possible : les variables faiblement corrélées. Sur le graphique c'est lorsqu'on voit des points transparents. Lorsque les valeurs approximent le zéro, on comprend qu'il n'y a pas de rapport ou peu de rapport lorsque les valeurs quantitatives varient.

B. ACP

1. Valeurs propres

Nous allons réaliser notre ACP (Analyse des Composantes Principales) et déterminer les valeurs propres les plus significatives à l'aide de la méthode du coude.

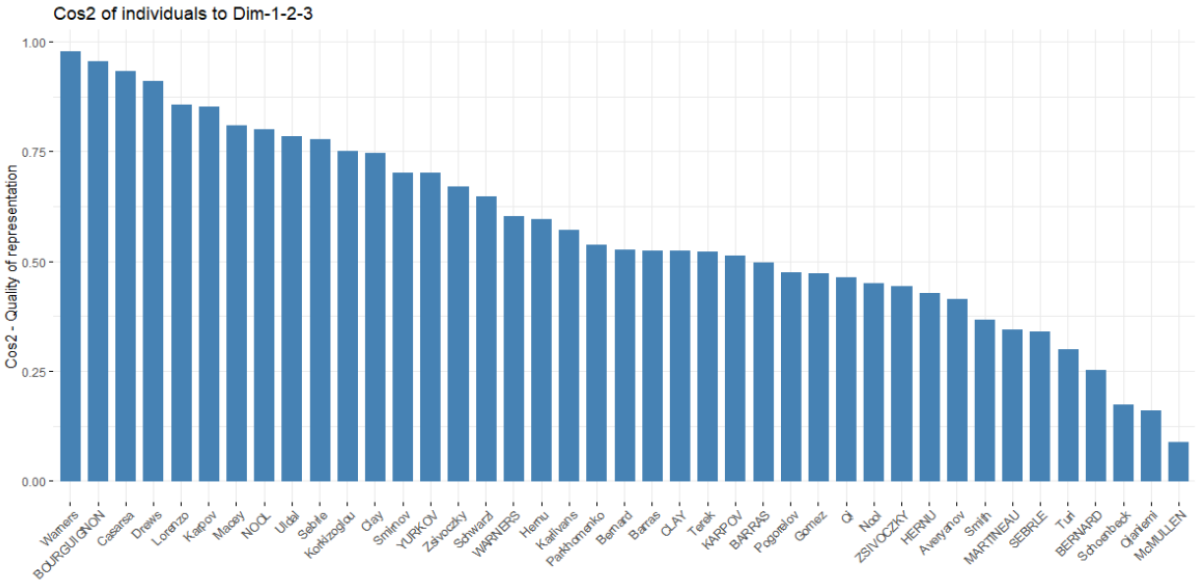


La méthode du coude consiste à sélectionner le nombre de valeurs propres idéal, lorsqu'il y a une cassure brute dans le pourcentage de variance expliquée. Ici on en voit une entre la valeur propre 1 et 2, mais effectuer une analyse sur seulement une valeur propre et 32.7% de variance expliquée ne serait pas assez pertinente. Alors nous allons en prendre 3 pour un total de 64.13% de variance expliquée. Nous devons faire un compromis entre complexité du modèle et la variance cumulée.

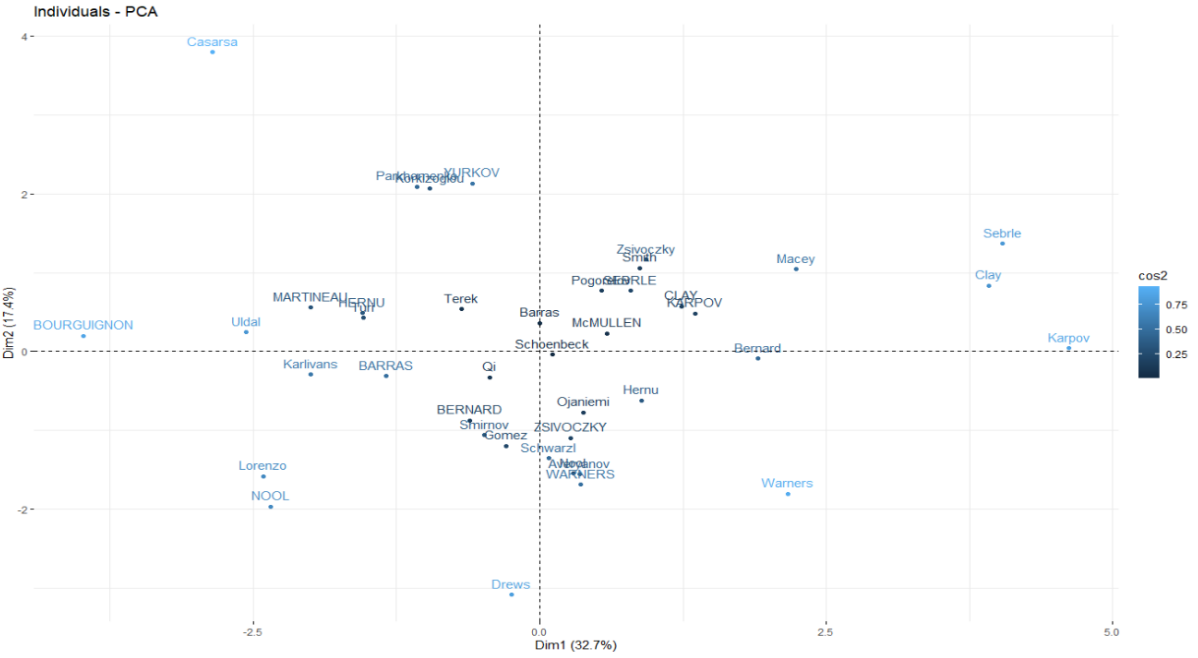
2. Composantes principales

Pour déterminer les projections des individus sur les 3 composantes principales, nous récupérons les coordonnées des athlètes dans le résultat de notre ACP. Voici une représentation par rapport à leur contribution sur les 3 premières composantes.

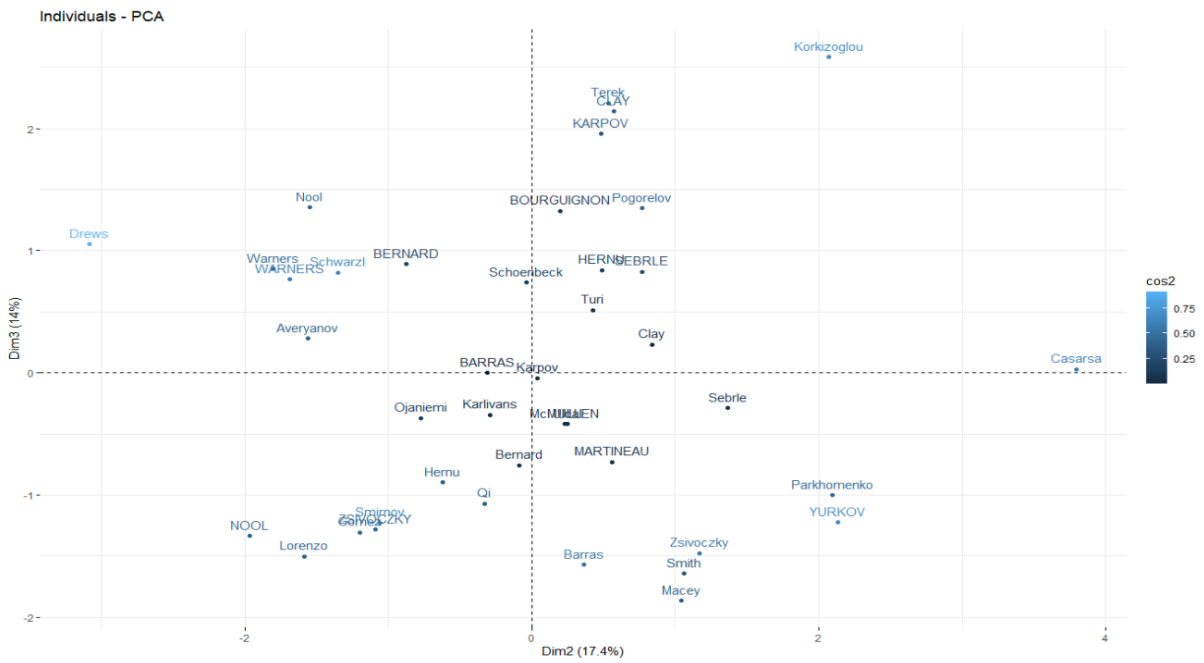
Ensuite une représentation des individus sur les 2 premières composantes principales.



On peut observer que Warners, Bourguignon et Casarsa sont les 3 athlètes qui contribuent le plus tandis que Schoenbeck, Ojaniemi et McMULLEN sont les athlètes qui contribuent le moins.



Nuage de points des individus sur les dimensions C1 et C2



Nuage de points des individus sur les dimensions C2 et C3

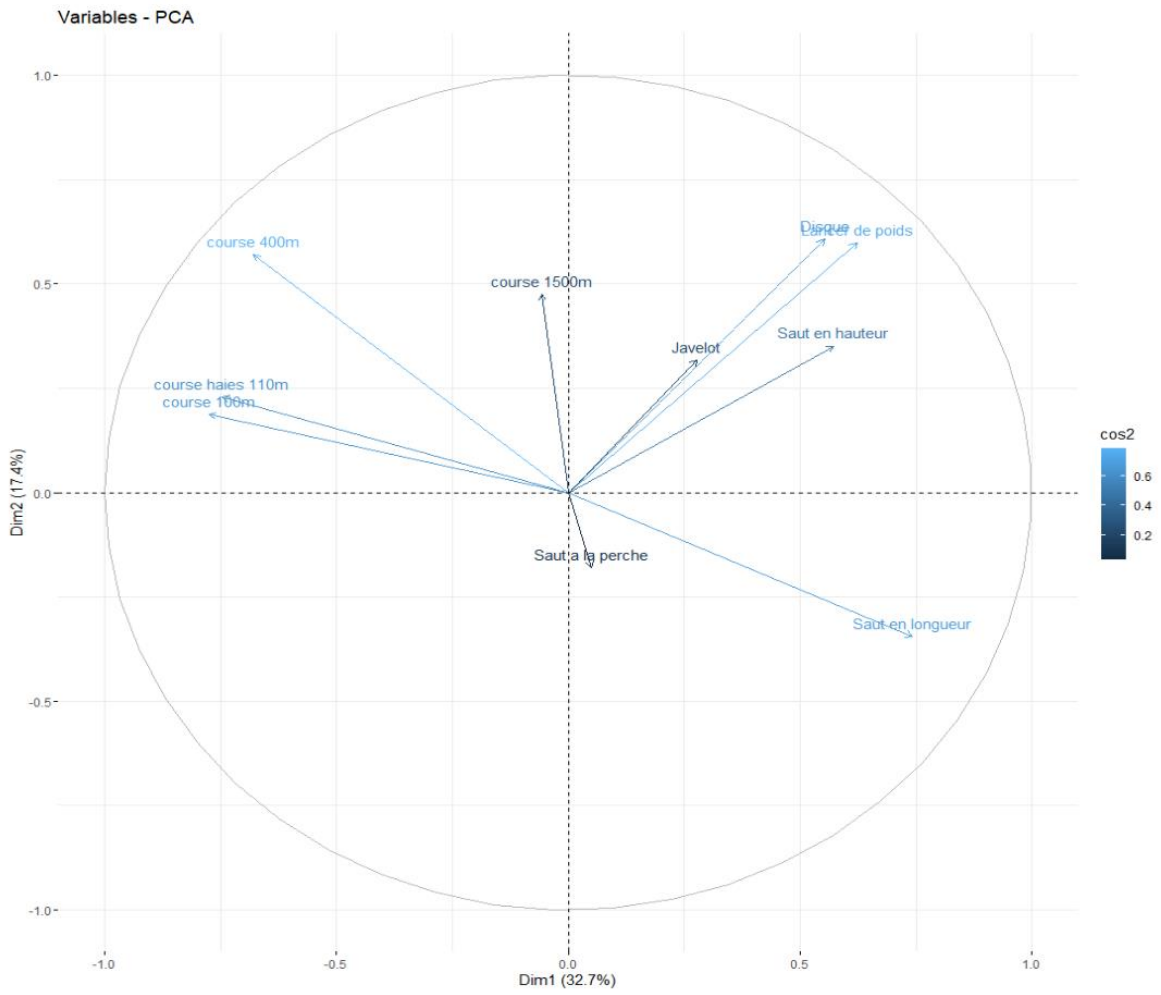
On remarque qu'il existe plusieurs athlètes atypiques dans notre jeu de données : prenons l'exemple de Casarsa. C'est l'un des pires athlètes en termes de points : il est à 7404 alors que le dernier athlète est à 7313 et la moyenne à 8005. Il est également le dernier dans le classement des jeux olympiques. Pourquoi est-il si particulier ? Il est relativement mauvais par rapport aux autres dans les courses mais se débrouille bien en lancer de disque. Il apparaît dans le graphe C1 et C2 tout en haut à gauche car il est mauvais en course mais correct en lancer de poids et lancer de disque.

Prenons un autre joueur : Bourguignon. C'est le pire athlète des Decastar, détenteur des 7313 points et classé 13ème. Il n'est pas spécialement bon ni en course ni aux épreuves de lancer de poids et de disque. C'est pourquoi il apparaît dans le premier graphique comme étant à l'extrémité gauche à cause de ses scores en courses, il se retrouve au centre car il appartient à la moyenne pour la dimension 2.

3. Corrélation entre les variables et les 3 composantes principales

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
course 100m	-0.77471983	0.1871420	-0.18440714	-0.03781826	0.30219639
Saut en longueur	0.74189974	-0.3454213	0.18221105	0.10178564	0.03667805
Lancer de poids	0.62250255	0.5983033	-0.02337844	0.19059161	0.11115082
Saut en hauteur	0.57194530	0.3502936	-0.25951193	-0.13559420	0.55543957
course 400m	-0.67960994	0.5694378	0.13146970	0.02930198	-0.08769157
course haies 110m	-0.74624532	0.2287933	-0.09263738	0.29083103	0.16432095
Disque	0.55246652	0.6063134	0.04295225	-0.25967143	-0.10482712
Saut a la perche	0.05034151	-0.1803569	0.69175665	0.55153397	0.32995932
Javelot	0.27711085	0.3169891	-0.38965541	0.71227728	-0.30512892
course 1500m	-0.05807706	0.4742238	0.78214280	-0.16108904	-0.15356189

Ce tableau nous indique le tableau des corrélations des variables par rapport aux composantes principales.



Cercle de corrélation des dimensions 1 et 2

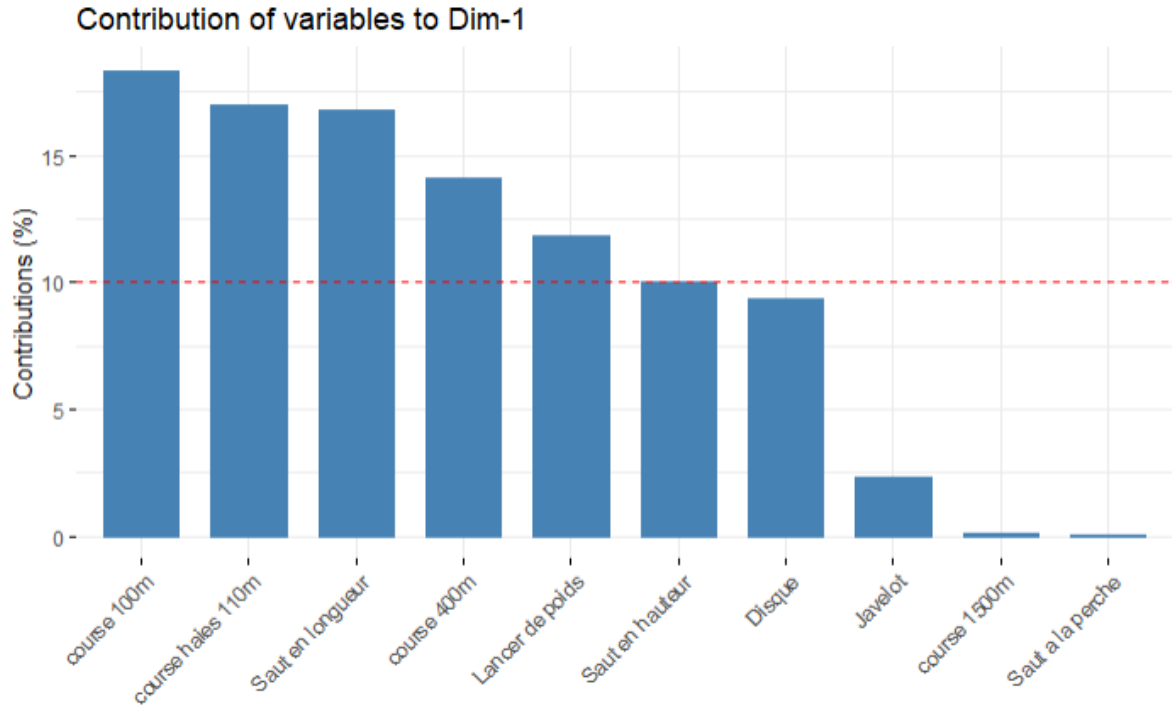
Cercle de corrélation des dimensions 2 et 3

Plus une variable est en bleu clair et mieux elle est représentée dans le cercle de corrélation.

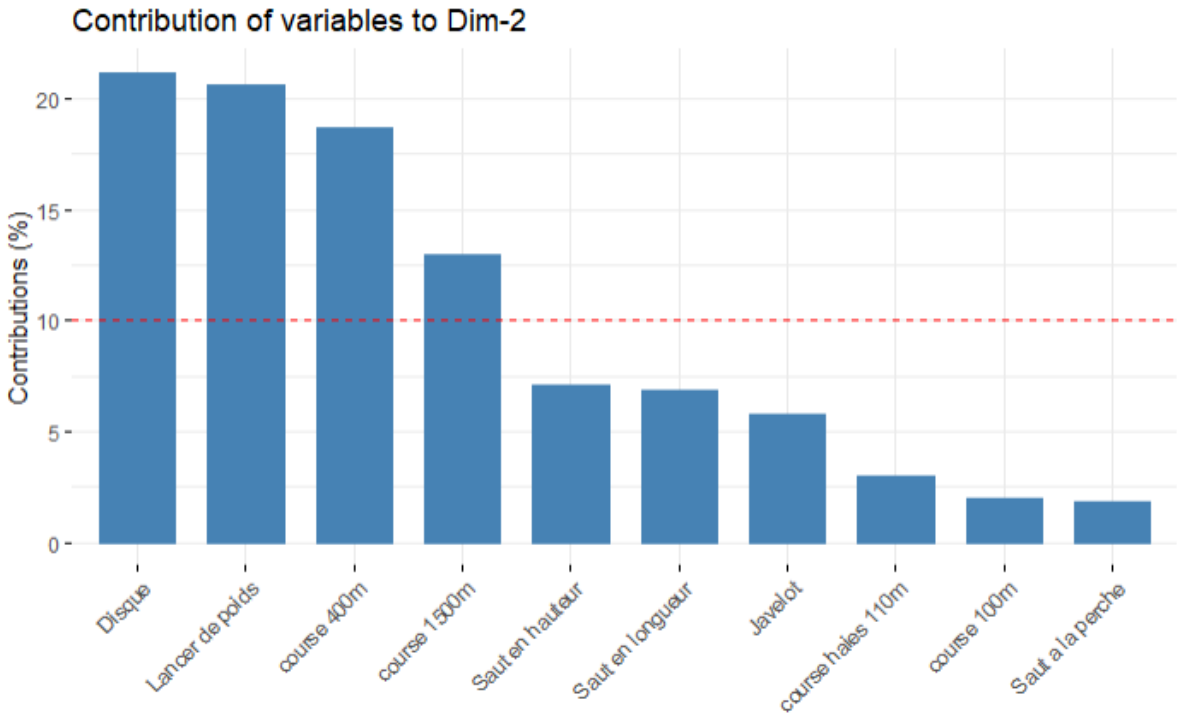
4. Contribution des composantes principales

On va s'intéresser à comment nos composantes principales sont construites. Pour cela, nous regardons la contribution des variables aux axes concernés (1, 2 et 3).

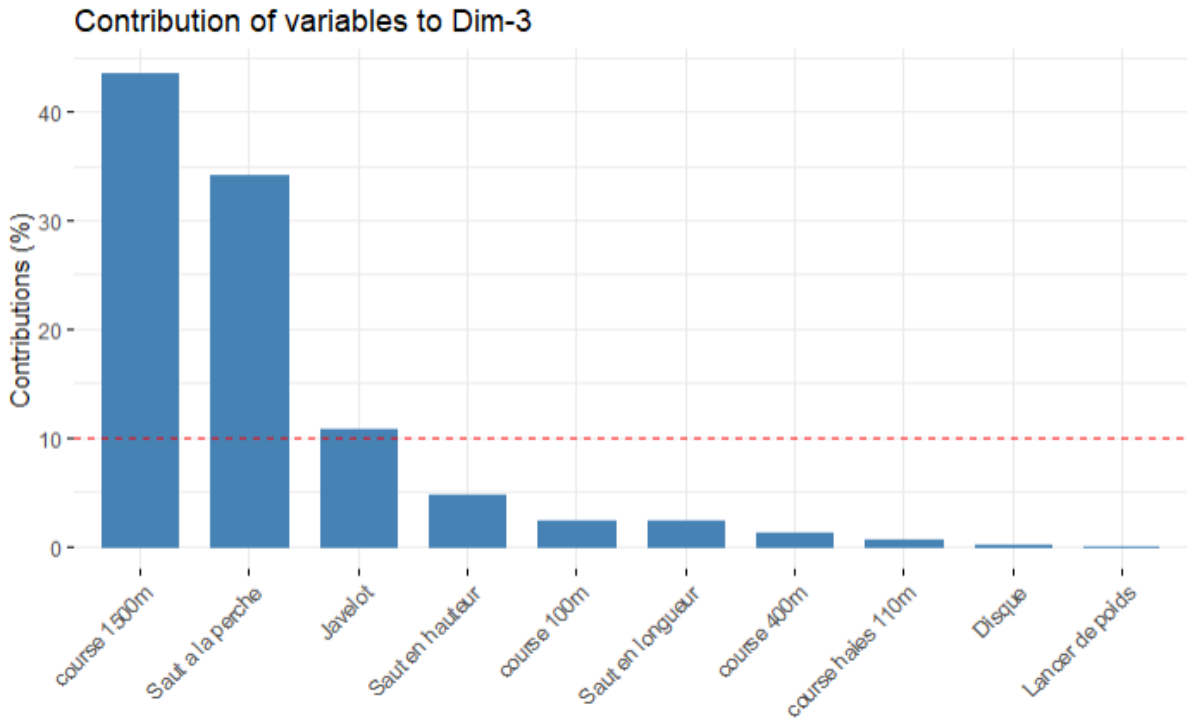
Regardons la dimension 1 : course 100m, course haie 110m, saut en longueur, course 400m, lancer de poids, saut en hauteur et disque. On observe que nous avons les courses à gauche de l'axe et les épreuves dont l'objectif est d'atteindre la plus grande distance à droite de l'axe.



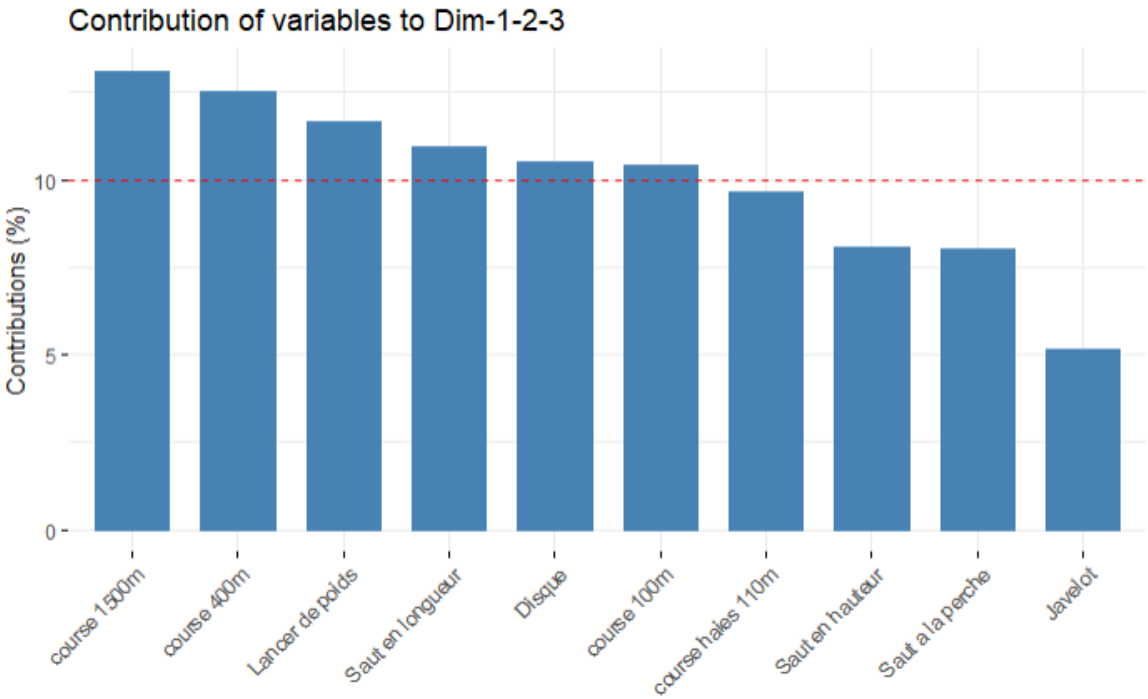
Pour la dimension 2 : Disque, Lancer de poids, course 400m et course 1500m contribuent majoritairement à cet axe.



Pour la dimension 3 : seulement course 1500m, Saut à la perche et Javelot pour le dernier axe.

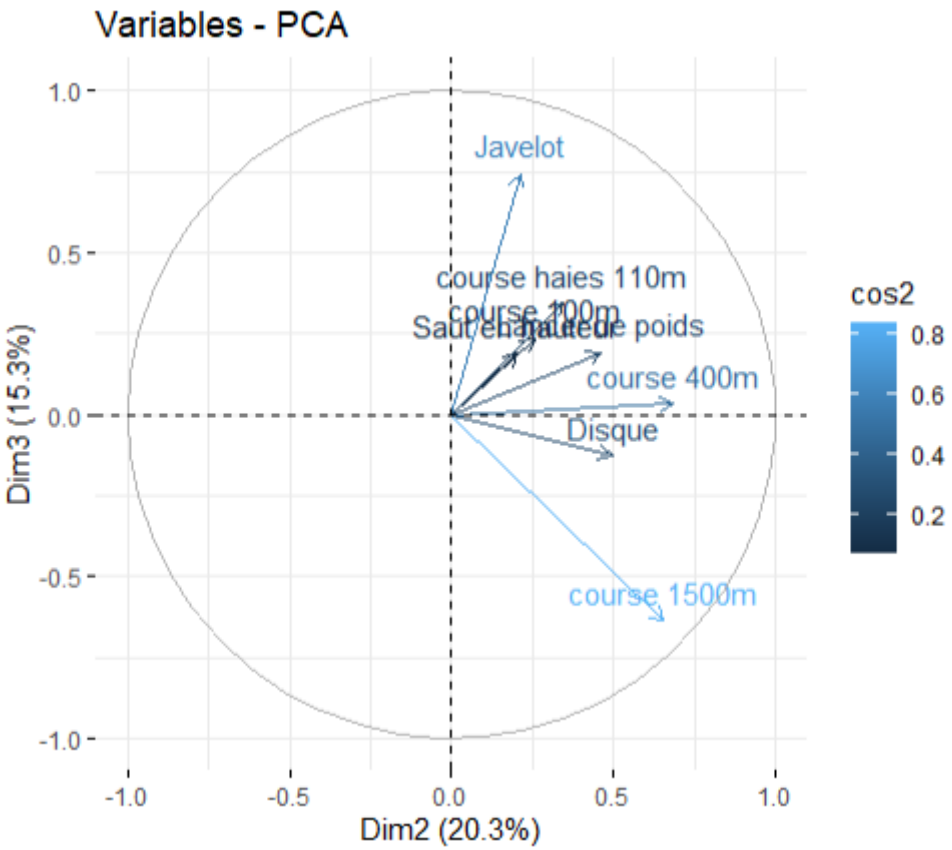


De manière générale voici la contribution des variables pour les 3 axes en même temps. Course 1500m, course 400m, lancer de poids, saut en longueur et disque sont les variables les plus importantes pour la construction de ces 3 composantes principales.



5. Effet de taille

Un effet de taille c'est le phénomène que nous retrouvons dans notre second cercle de corrélation par exemple, sur les dimensions 2 et 3. Lorsque beaucoup de variables sont regroupées dans une zone très proche et d'un seul côté du cercle. On distingue dans ce cercle de corrélation que (Saut en longueur, Saut à la perche) sont les seuls vraiment à gauche de l'axe vertical. Alors si nous enlevons ces 2 variables de l'ACP et qu'on réalise un nouveau cercle de corrélation, nous avons ceci :



Cercle de corrélation sur dimension 2 et 3

V. AFC

Tableau des données « sympathiques »

	SERI	GENE	GAI	HONN	INTL	SERV	COUR	COMP	DISC	TOTAL
PAYS	20	9	9	27	10	16	20	4	8	123
OUVR	42	10	22	51	18	28	38	12	22	243
VEND	11	2	5	14	8	7	5	8	6	66
COMM	8	9	12	23	14	16	14	12	12	120
EMPL	19	10	16	52	32	25	22	25	30	231
TECH	10	5	12	23	20	13	11	13	10	117
UNIV	2	8	7	6	15	6	6	9	4	63
LIBE	8	42	23	24	46	22	22	34	16	237
TOTAL	120	95	106	220	163	133	138	117	108	1200

Partie 1

1. Pour calculer le pourcentage de chaque caractéristique d'une personne sympathique on divise le nombre de vote pour cette caractéristique sur le total de vote.

```
      SERI GENE  GAI HONN INTL SERV COUR COMP DISC TOTAL
TOTAL  0.1 0.08 0.09 0.18 0.14 0.11 0.12  0.1 0.09     1
```

   Même principe pour les catégories professionnelles.

```
       TOTAL
PAYS    0.10
OUVR    0.20
VEND    0.06
COMM    0.10
EMPL    0.19
TECH    0.10
UNIV    0.05
LIBE    0.20
TOTAL   1.00
```

2. Sachant que chaque personne sondée a effectué 3 votes, on a 1200 votes/3 = **400 personnes**. La proportion d'employé pour qui être honnête rend sympathique correspond au nombre de vote honnête des employés diviser par le nombre d'employé => 52 / (231/3) = **0.68**. La proportion d'employé parmi les gens qui pensent qu'être honnête rend sympathique correspond au nombre de vote honnête des employés diviser par le nombre de vote honnête => 52 / 220 = **0.24**

Partie 2

3. Le nombre de valeur propre correspond à la variable possédant le moins de modalité moins un (min(8,9) - 1 = 7)

4.

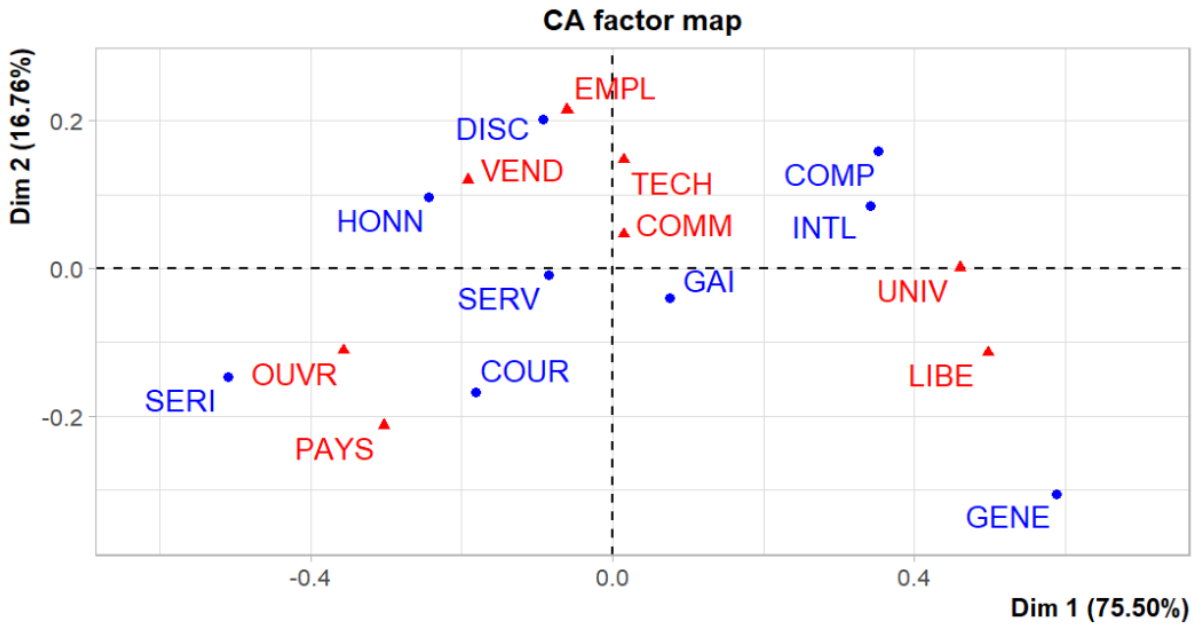
	Axis1(%)	Axis2(%)
PAYS	9.6	21.4
OUVR	26.3	11.6
VEND	2.1	3.6
COMM	0.0	1.0
EMPL	0.7	40.8
TECH	0.0	9.7
UNIV	11.4	0.0
LIBE	49.9	11.8

	Axis1(%)	Axis2(%)
seri	26.4	10.0
gene	28.0	34.0
gai	0.5	0.7
honn	11.1	7.8
intl	16.1	4.5
serv	0.8	0.0
cour	3.8	14.9
comp	12.4	11.3
disc	0.8	16.8

Les modalités qui définissent le plus le premier axe factoriel sont d'après le tableau de contribution aux axes : OUVRIER (26.3%) et LIBERAL (49.9%) pour les catégories professionnelles et SERIEUX (26.4%) et GENEREUX (28%) pour les caractéristiques d'une personne sympathique. De même pour le second axe : EMPLOYE (40.8%) et PAYSAN (21.4%), GENEREUX (34%) DISCIPLINE (16.8%) et COURTOIS (14.9%).

5. Pour effectuer la sélection des modalités les moins représentées, on va privilégier les modalités ayant la plus petite contribution suivant le premier axe car il possède une inertie de 75.5% par rapport au deuxième ayant une inertie de seulement 16.8%. De plus on ne considérera que les modalités qui n'ont pas été énuméré dans la question 4.
   Les modalités les moins bien représentés par le premier plan factoriel sont : VENDEUR (2.1% et 3.6%), COMMERCANT (0% et 1%), TECHNICIEN (0% et 9.7%) et GAI (0.5% et 0.7%), SERVIABLE (0.8% et 0%).

6.



OUVR et PAYS sont proches sur le graphique ce qui indique qu'ils ont une répartition similaire des votes des caractéristiques d'une personne sympathique ce qui est bien vérifié sur les données brutes. A l'opposé on trouve UNIV et LIBE qui ont une répartition inverse des votes, c'est-à-dire que les caractéristiques les plus voter pour OUVR et PAYS sont les moins voter pour UNIV et LIBE.

VEND et honn sont proches, ce qui indique que les vendeurs considèrent l'honnêteté comme la caractéristique la plus importante. A l'inverse la générosité est éloignée ce qui implique que ce critère est le moins importants pour les vendeurs.