TD 1 – Analyse en Composantes Principales

1 Exercice 1. Un peu de calcul matriciel

On considère la matrice

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 2 & 2 & 2 \\ 0 & 0 & 2 \end{pmatrix}$$

1. Calculer Z, la version centrée-réduite de la matrice X.
2. En déduire la matrice d'inertie de Z S associée à X.
3. Diagonaliser la matrice S en l'écrivant sous la forme

$$S = P\Lambda P^{-1},$$

où Λ est une matrice diagonale de termes diagonaux $\{\lambda_i\}_{i=1,2,3}$.

4. On rappelle que la trace d'une matrice carrée A de taille $n \times n$ correspond à la somme de ses termes diagonaux :

$$\mathrm{tr}(A) = \sum_{i=1}^{n} A_{i,i}.$$

Montrer que si U est une matrice de taille $n \times m$ et V une matrice de taille $m \times n$, alors

$$\mathrm{tr}(UV) = \mathrm{tr}(VU).$$

En déduire que

$$\mathrm{tr}(S) = \sum_{i=1}^{3} \lambda_i = p.$$

5. Calculer les parts d'inertie. Quelle(s) direction(s) aurait-on envie de retenir si la matrice X correspondait à un tableau de données ?
6. Calculer les composantes principales, et représenter les individus dans le plan des deux premières composantes.
7. Déterminer les corrélations entre les variables et les composantes et les représenter sur le cercle des corrélations dans le plan des deux premières composantes.

2 Exercice 2. Analyse sensorielle

On dispose de 10 mesures effectuées sur 10 cidres différents, afin d'évaluer leur qualité olfacto-gustative.

	odeur	sucre	acide	amer	astringence	suffocante	piquante	alcool	parfum	fruitee
cidre 1	2.14	1.86	3.29	2.29	2.00	0.14	2.29	1.86	1.29	1.29
cidre 2	2.43	0.79	2.71	2.57	2.00	0.43	2.57	2.86	0.43	0.14
cidre 3	2.71	3.14	2.57	2.57	1.43	0.14	2.14	0.86	2.29	1.71
cidre 4	3.00	3.71	2.14	2.07	1.57	0.00	1.29	1.00	3.14	3.14
cidre 5	3.43	1.29	2.86	3.14	2.17	1.00	1.86	2.86	1.14	0.29
cidre 6	3.14	0.86	2.86	3.79	2.57	0.14	1.71	3.29	0.14	0.00
cidre 7	3.14	1.14	2.86	2.86	2.00	0.43	1.71	1.86	0.14	0.00
cidre 8	2.43	3.71	3.21	1.57	1.71	0.00	1.00	0.57	2.57	2.86
cidre 9	5.10	2.86	2.86	3.07	1.79	1.71	0.43	1.43	0.57	2.71
cidre 10	3.07	3.14	2.57	3.00	2.00	0.00	0.43	1.29	2.57	3.07

1. Quel est l'objectif d'une ACP dans ce contexte ?
2. Les données sont dans un `data.frame` nommé `cidre` et on effectue l'ACP dans R.

```
> library(FactoMineR)
> cidre.acp = PCA(cidre, ncp = 10, scale.unit = TRUE, graph = FALSE)
```

▶ **Tableau**

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	5.154	51.544	51.544
comp 2	2.502	25.020	76.565
comp 3	1.097	10.965	87.530
comp 4	0.834	8.344	95.874
comp 5	0.194	1.937	97.812
comp 6	0.140	1.398	99.210
comp 7	0.049	0.491	99.701
comp 8	0.024	0.243	99.944
comp 9	0.006	0.056	100.000

Quelle est la part d'inertie expliquée par les deux premières composantes principales ? Combien de dimensions aurait-on envie de retenir ?

3. On affiche les deux sorties R ci-dessous

```
> cidre.acp$var$contrib[, 1:3] |>
+   print_table()
```

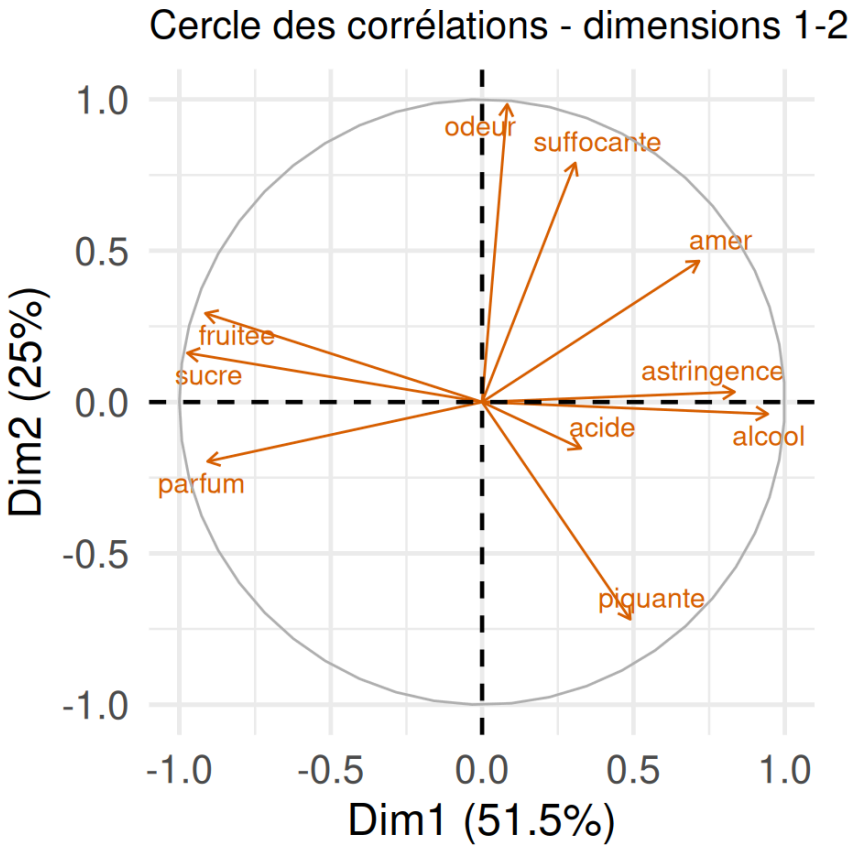
	Dim.1	Dim.2	Dim.3
odeur	0.134	38.701	0.002
sucre	18.400	1.051	0.045
acide	2.067	0.938	69.665
amer	9.969	8.679	13.298
astringence	13.492	0.045	0.732
suffocante	1.839	24.956	8.756
piquante	4.650	20.592	0.009
alcool	17.284	0.062	3.390
parfum	15.953	1.538	4.075
fruitee	16.212	3.438	0.028

```
> cidre.acp$var$cos2[, 1:3] |>
+   print_table()
```

	Dim.1	Dim.2	Dim.3
odeur	0.007	0.968	0.000
sucre	0.948	0.026	0.000
acide	0.107	0.023	0.764
amer	0.514	0.217	0.146
astringence	0.695	0.001	0.008
suffocante	0.095	0.624	0.096
piquante	0.240	0.515	0.000
alcool	0.891	0.002	0.037
parfum	0.822	0.038	0.045
fruitee	0.836	0.086	0.000

ainsi que les deux représentations graphiques suivantes.

▶ Dim 1-2



a. Quelles seraient les variables mal représentées dans le premier plan factoriel ? Pourquoi ?

b. Quelles seraient les variables les plus corrélées (positivement ou négativement) à la teneur en alcool ? Celles qui le seraient le moins ?

c. Quels seraient les caractères qui contribuent le plus au premier axe principal ?

d. Mêmes questions avec les deux autres axes.

4. On considère maintenant les sorties R

```
> cidre.acp$ind$contrib[, 1:3] |>
+   print_table()
```

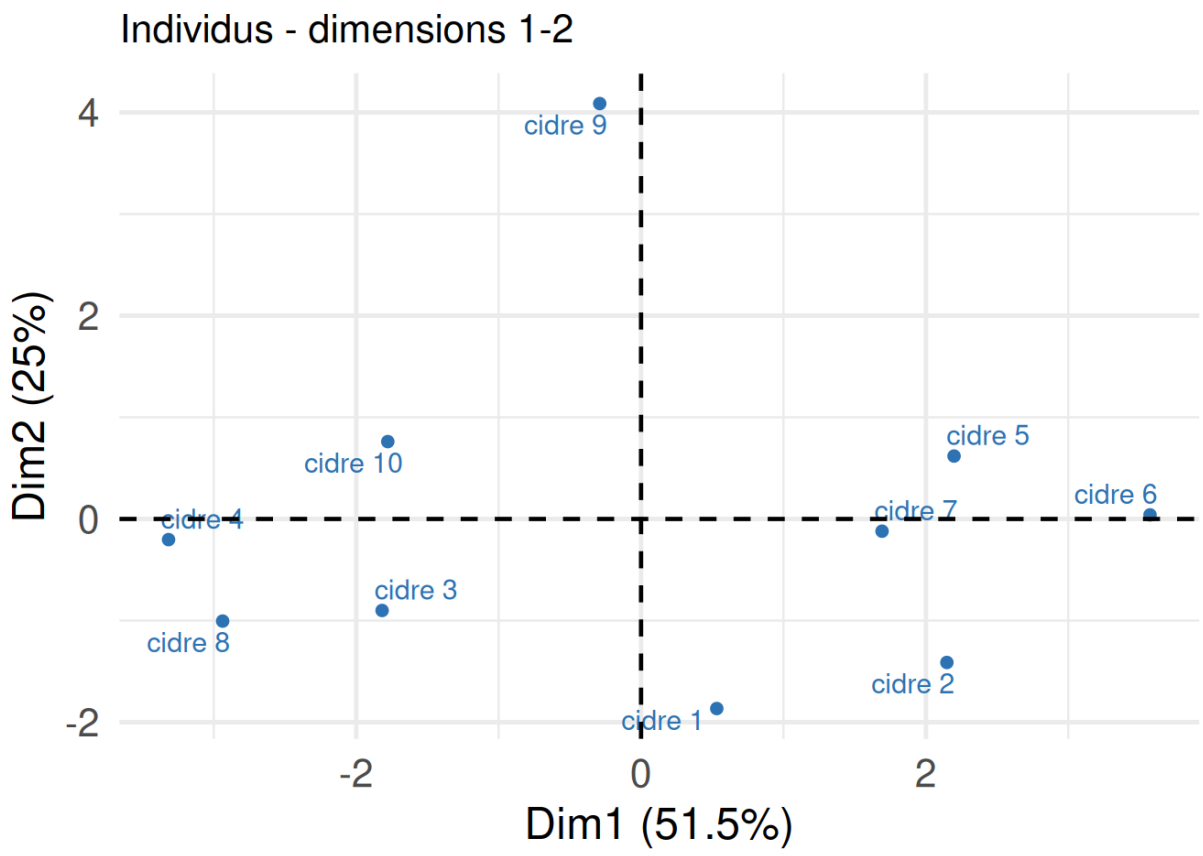
	Dim.1	Dim.2	Dim.3
cidre 1	0.548	13.910	17.790
cidre 2	8.945	7.970	0.557
cidre 3	6.413	3.242	2.194
cidre 4	21.352	0.165	23.318
cidre 5	9.369	1.531	0.008
cidre 6	24.775	0.006	8.027
cidre 7	5.554	0.058	0.610
cidre 8	16.744	4.037	27.229
cidre 9	0.163	66.766	9.380
cidre 10	6.135	2.315	10.887

```
> cidre.acp$ind$cos2[, 1:3] |>
+   print_table()
```

	Dim.1	Dim.2	Dim.3
cidre 1	0.047	0.580	0.325
cidre 2	0.566	0.245	0.008
cidre 3	0.492	0.121	0.036
cidre 4	0.776	0.003	0.180
cidre 5	0.781	0.062	0.000
cidre 6	0.824	0.000	0.057
cidre 7	0.697	0.004	0.016
cidre 8	0.639	0.075	0.221
cidre 9	0.005	0.919	0.057
cidre 10	0.424	0.078	0.160

ainsi que les représentations graphiques suivantes.

▶ Dim 1-2



a. Les individus sont-ils bien représentés ?

b. Quels individus seraient représentatifs de profils opposés par rapport au premier axe principal ? Même question pour les deux autres axes.

c. Proposez des profils de cidres en précisant leurs caractéristiques principales.

3 Exercice 3.

On va considérer le jeu de données `decathlon` disponible dans le paquet `FactoMineR`, qui s'intéresse en particulier aux performances d'athlètes pour les 10 épreuves du décathlon, dans deux événements différents : les Jeux Olympiques de 2004 et le Décastar de 2004. On ne va considérer ici que cette dernière épreuve, dont on stockera les mesures dans un `data.frame` nommé decastar

```
> require(FactoMineR)
>
> # Charger le jeu de données dans R
> data(decathlon)
>
> # Ne conserver que la compétition Décastar
> decastar = decathlon[decathlon$Competition == 'Decastar', ]
> # Supprimer la colonne « Competition », devenue inutile
> decastar = decastar[ , colnames(decastar) != 'Competition']
> print_table(decastar, vertical.headers = TRUE)
```

	100m	Long.jump	Shot.put	High.jump	400m	110m.hurdle	Discus	Pole.vault	Javeline	1500m	Rank	Points
SEBRLE	11.04	7.58	14.83	2.07	49.81	14.69	43.75	5.02	63.19	291.7	1	8217
CLAY	10.76	7.40	14.26	1.86	49.37	14.05	50.72	4.92	60.15	301.5	2	8122
KARPOV	11.02	7.30	14.77	2.04	48.37	14.09	48.95	4.92	50.31	300.2	3	8099
BERNARD	11.02	7.23	14.25	1.92	48.93	14.99	40.87	5.32	62.77	280.1	4	8067
YURKOV	11.34	7.09	15.19	2.10	50.42	15.31	46.26	4.72	63.44	276.4	5	8036
WARNERS	11.11	7.60	14.31	1.98	48.68	14.23	41.10	4.92	51.77	278.1	6	8030
ZSIVOCZKY	11.13	7.30	13.48	2.01	48.62	14.17	45.67	4.42	55.37	268.0	7	8004
McMULLEN	10.83	7.31	13.76	2.13	49.91	14.38	44.41	4.42	56.37	285.1	8	7995
MARTINEAU	11.64	6.81	14.57	1.95	50.14	14.93	47.60	4.92	52.33	262.1	9	7802
HERNU	11.37	7.56	14.41	1.86	51.10	15.06	44.99	4.82	57.19	285.1	10	7733
BARRAS	11.33	6.97	14.09	1.95	49.48	14.48	42.10	4.72	55.40	282.0	11	7708
NOOL	11.33	7.27	12.68	1.98	49.20	15.29	37.92	4.62	57.44	266.6	12	7651
BOURGUIGNON	11.36	6.80	13.46	1.86	51.16	15.67	40.49	5.02	54.68	291.7	13	7313

1. Quel est l'intérêt d'effectuer une ACP ici ? Sur quelles variables ?
2. Quelle opération est indispensable ici avant d'effectuer l'ACP ? Pourquoi ?
3. On affiche ci-dessous les résultats de l'ACP en termes de valeurs propres. Combien de directions aurait-on envie de conserver ?

```
> deca.acp = PCA(decastar[, 1:11], quanti.sup= 11, graph = FALSE )
```

▸ Tableau

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	3.114	31.144	31.144
comp 2	2.027	20.267	51.411
comp 3	1.390	13.903	65.314
comp 4	1.321	13.206	78.520
comp 5	0.861	8.610	87.129

4. Commentez les résultats numériques et graphiques ci-dessous en termes d'individus et de variables (suivez le modèle de l'exercice précédent)

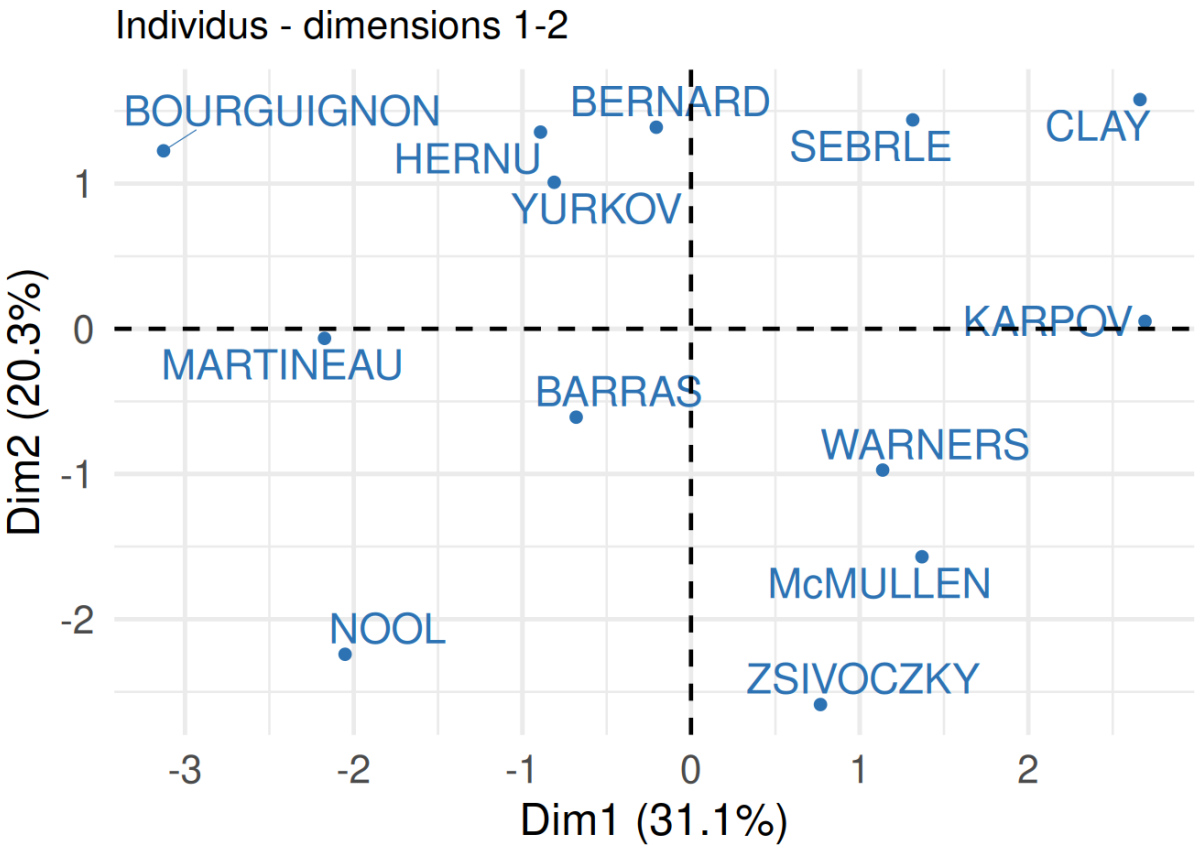
```
> deca.acp$ind$contrib[, 1:4] |>
+   print_table()
```

	Dim.1	Dim.2	Dim.3	Dim.4
SEBRLE	4.270	7.855	3.461	15.455
CLAY	17.502	9.457	0.545	6.486
KARPOV	17.897	0.010	9.553	8.709
BERNARD	0.105	7.311	18.165	0.054
YURKOV	1.626	3.863	9.496	41.154
WARNERS	3.189	3.592	1.903	4.393
ZSIVOCZKY	1.455	25.408	0.445	0.005
McMULLEN	4.632	9.353	0.004	9.583
MARTINEAU	11.663	0.016	36.313	2.157
HERNU	1.967	6.963	0.005	0.106
BARRAS	1.145	1.406	0.929	2.082
NOOL	10.391	19.064	17.597	0.385
BOURGUIGNON	24.158	5.701	1.583	9.431

```
> deca.acp$ind$cos2[, 1:4] |>
+   print_table()
```

	Dim.1	Dim.2	Dim.3	Dim.4
SEBRLE	0.222	0.266	0.080	0.341
CLAY	0.509	0.179	0.007	0.080
KARPOV	0.588	0.000	0.140	0.121
BERNARD	0.005	0.217	0.369	0.001
YURKOV	0.061	0.095	0.160	0.658
WARNERS	0.192	0.141	0.051	0.112
ZSIVOCZKY	0.069	0.789	0.009	0.000
McMULLEN	0.205	0.269	0.000	0.180
MARTINEAU	0.364	0.000	0.505	0.029
HERNU	0.105	0.242	0.000	0.002
BARRAS	0.175	0.140	0.063	0.135
NOOL	0.328	0.391	0.248	0.005
BOURGUIGNON	0.613	0.094	0.018	0.102

▸ Dim 1-2



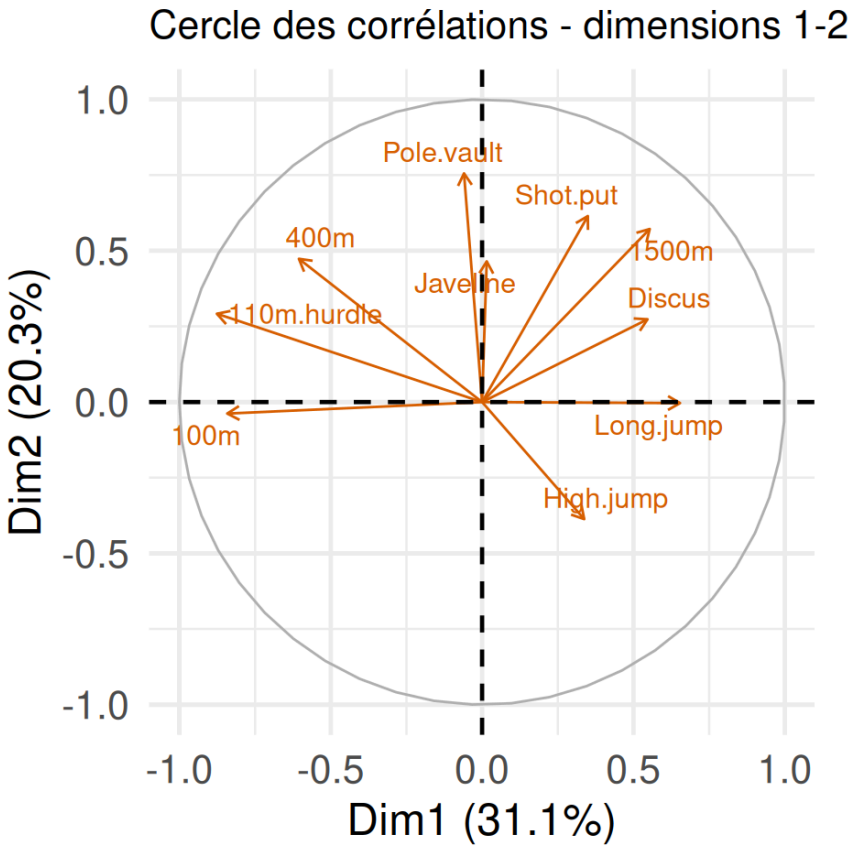

```
> deca.acp$var$contrib[, 1:4] |>
+   print_table()
```

	Dim.1	Dim.2	Dim.3	Dim.4
100m	22.754	0.071	11.314	0.093
Long.jump	13.724	0.001	12.685	2.104
Shot.put	3.918	18.637	19.584	5.351
High.jump	3.651	7.379	3.523	38.336
400m	11.757	11.053	1.625	3.935
110m.hurdle	24.644	4.201	1.891	4.483
Discus	9.585	3.691	32.724	0.161
Pole.vault	0.115	28.169	3.330	7.586
Javeline	0.007	10.663	12.013	34.142
1500m	9.845	16.134	1.312	3.810

```
> deca.acp$var$cos2[, 1:4] |>
+   print_table()
```

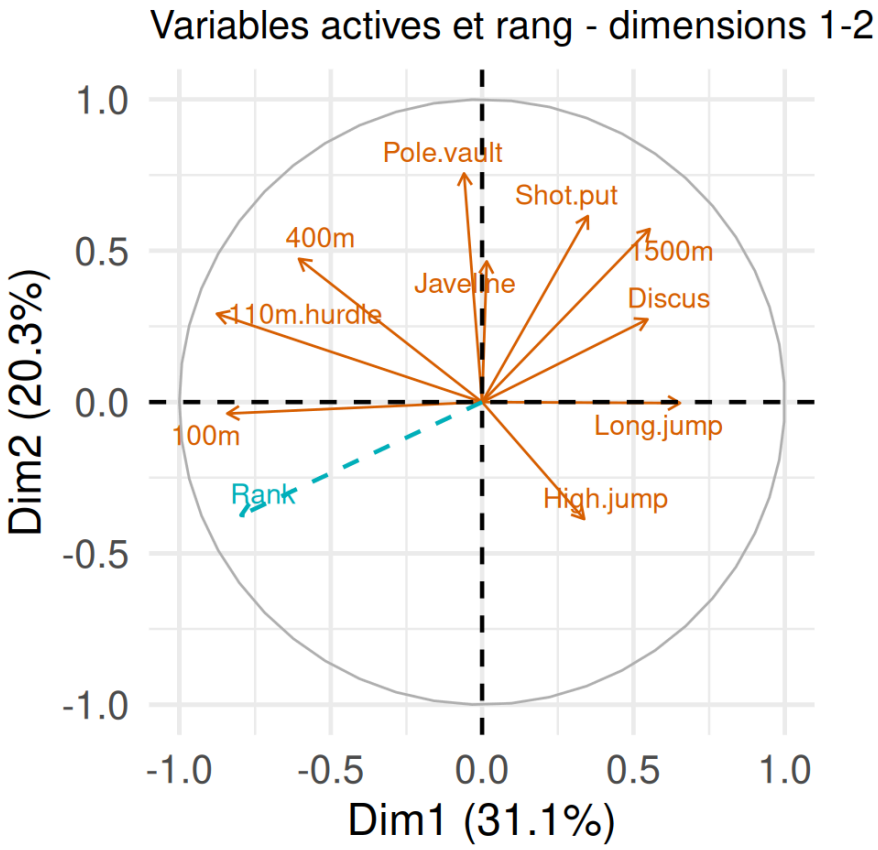
	Dim.1	Dim.2	Dim.3	Dim.4
100m	0.709	0.001	0.157	0.001
Long.jump	0.427	0.000	0.176	0.028
Shot.put	0.122	0.378	0.272	0.071
High.jump	0.114	0.150	0.049	0.506
400m	0.366	0.224	0.023	0.052
110m.hurdle	0.768	0.085	0.026	0.059
Discus	0.299	0.075	0.455	0.002
Pole.vault	0.004	0.571	0.046	0.100
Javeline	0.000	0.216	0.167	0.451
1500m	0.307	0.327	0.018	0.050

▶ **Dim 1-2**



5. On se propose à présent de mettre en parallèle le classement des athlètes avec les résultats obtenus. Commentez les graphiques ci-dessous.

▶ **Dim 1-2**


Variables actives et rang - dimensions 1-2
Pole.vault
Shot.put
400m
1500m
Javeline
Discus
110m.hurdle
100m
Long.jump
Rank
High.jump
1.0
0.5
0.0
-0.5
-1.0
Dim2 (20.3%)
Dim1 (31.1%)