

1-Práctica de Análisis de Componentes Principales en R

Vamos a realizar un Análisis de Componentes principales sobre los resultados obtenidos en la competición de heptathlon femenino en los juegos de Seúl 1988, estos datos se encuentran en el paquete HSAUR2 y corresponden a los datos de 25 atletas sobre 8 variables: 100m vallas, salto de altura, lanzamiento de peso, 200m lisos, salto de longitud, lanzamiento de jabalina, 800m y puntuación total obtenida.

1. Instalamos el paquete HSAUR2

```
> install.packages("HSAUR2")
```

2. Sacamos los datos de la tabla heptathlon del paquete HSAUR2

```
> data("heptathlon", package="HSAUR2")
```

	hurdles	highjump	shot	run200m	longjump	javelin	run800m	score
Joyner-Kersey (USA)	12.69	1.86	15.80	22.56	7.27	45.66	128.51	7291
John (GDR)	12.85	1.80	16.23	23.65	6.71	42.56	126.12	6897
Behmer (GDR)	13.20	1.83	14.20	23.10	6.68	44.54	124.20	6858
Sablovskaitė (URS)	13.61	1.80	15.23	23.92	6.25	42.78	132.24	6540
Choubenkova (URS)	13.51	1.74	14.76	23.93	6.32	47.46	127.90	6540
Schulz (GDR)	13.75	1.83	13.50	24.65	6.33	42.82	125.79	6411
Fleming (AUS)	13.38	1.80	12.88	23.59	6.37	40.28	132.54	6351
Greiner (USA)	13.55	1.80	14.13	24.48	6.47	38.00	133.65	6297
Lajbnerova (CZE)	13.63	1.83	14.28	24.86	6.11	42.20	136.05	6252
Bouraga (URS)	13.25	1.77	12.62	23.59	6.28	39.06	134.74	6252
Wijnsma (HOL)	13.75	1.86	13.01	25.03	6.34	37.86	131.49	6205
Dimitrova (BUL)	13.24	1.80	12.88	23.59	6.37	40.28	132.54	6171
Scheider (SWI)	13.85	1.86	11.58	24.87	6.05	47.50	134.93	6137
Braun (FRG)	13.71	1.83	13.16	24.78	6.12	44.58	142.82	6109
Ruotsalainen (FIN)	13.79	1.80	12.32	24.61	6.08	45.44	137.06	6101
Yuping (CHN)	13.93	1.86	14.21	25.00	6.40	38.60	146.67	6087
Hagger (GB)	13.47	1.80	12.75	25.47	6.34	35.76	134.93	5975
Brown (USA)	14.07	1.83	12.69	24.83	6.13	44.34	146.43	5972
Mulliner (GB)	14.39	1.71	12.68	24.92	6.10	37.76	138.02	5746
Hautenaue (BEL)	14.04	1.77	11.81	25.61	5.99	35.68	133.90	5734
Kytola (FIN)	14.31	1.77	11.66	25.69	5.75	39.48	133.35	5686
Geremias (BRA)	14.23	1.71	12.95	25.50	5.50	39.64	144.02	5508
Hui-Ing (TAI)	14.85	1.68	10.00	25.23	5.47	39.14	137.30	5290
Jeong-Mi (KOR)	14.53	1.71	10.83	26.61	5.50	39.26	139.17	5289
Launa (PNG)	16.42	1.50	11.78	26.16	4.88	46.38	163.43	4566

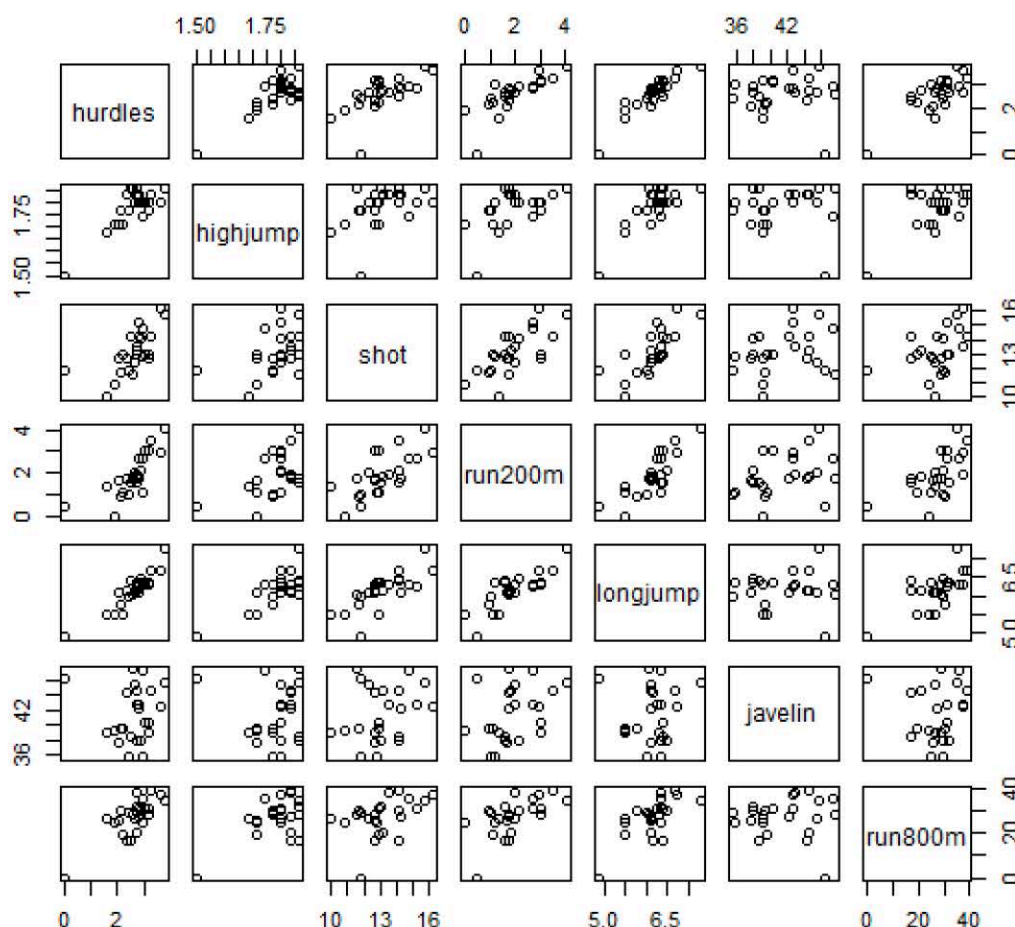
3. Recodificamos las pruebas relativas a las 3 carreras, vallas, 200m y 800m, restando al mayor valor en cada carrera, cada uno de los tiempos de las 35 atletas.

```
> heptathlon$hurdles <- max(heptathlon$hurdles) - heptathlon$hurdles
> heptathlon$run200m <- max(heptathlon$run200m) - heptathlon$run200m
> heptathlon$run800m <- max(heptathlon$run800m) - heptathlon$run800m
```

	hurdles	highjump	shot	run200m	longjump	javelin	run800m
Joyner-Kersey (USA)	3.73	1.86	15.80	4.05	7.27	45.66	34.92
John (GDR)	3.57	1.80	16.23	2.96	6.71	42.56	37.31
Behmer (GDR)	3.22	1.83	14.20	3.51	6.68	44.54	39.23
Sablovskaitė (URS)	2.81	1.80	15.23	2.69	6.25	42.78	31.19
Choubenkova (URS)	2.91	1.74	14.76	2.68	6.32	47.46	35.53
Schulz (GDR)	2.67	1.83	13.50	1.96	6.33	42.82	37.64
Fleming (AUS)	3.04	1.80	12.88	3.02	6.37	40.28	30.89
Greiner (USA)	2.87	1.80	14.13	2.13	6.47	38.00	29.78
Lajbnerova (CZE)	2.79	1.83	14.28	1.75	6.11	42.20	27.38
Bouraga (URS)	3.17	1.77	12.62	3.02	6.28	39.06	28.69
Wijnsma (HOL)	2.67	1.86	13.01	1.58	6.34	37.86	31.94
Dimitrova (BUL)	3.18	1.80	12.88	3.02	6.37	40.28	30.89
Scheider (SWI)	2.57	1.86	11.58	1.74	6.05	47.50	28.50
Braun (FRG)	2.71	1.83	13.16	1.83	6.12	44.58	20.61
Ruotsalainen (FIN)	2.63	1.80	12.32	2.00	6.08	45.44	26.37
Yuping (CHN)	2.49	1.86	14.21	1.61	6.40	38.60	16.76
Hagger (GB)	2.95	1.80	12.75	1.14	6.34	35.76	24.95
Brown (USA)	2.35	1.83	12.69	1.78	6.13	44.34	17.00
Mulliner (GB)	2.03	1.71	12.68	1.69	6.10	37.76	25.41
Hautenaue (BEL)	2.38	1.77	11.81	1.00	5.99	35.68	29.53
Kytola (FIN)	2.11	1.77	11.66	0.92	5.75	39.48	30.08
Geremias (BRA)	2.19	1.71	12.95	1.11	5.50	39.64	19.41
Hui-Ing (TAI)	1.57	1.68	10.00	1.38	5.47	39.14	26.13
Jeong-Mi (KOR)	1.89	1.71	10.83	0.00	5.50	39.26	24.26
Launa (PNG)	0.00	1.50	11.78	0.45	4.88	46.38	0.00

4. Vamos a realizar un diagrama de dispersión de la matriz de los resultados de las pruebas de las 25 atletas en las 7 pruebas, podemos observar que existe una relación positiva entre cada par de pruebas, con la excepción en el lanzamiento de jabalina, donde parece que no existe relación con el resto de las 6 pruebas restantes.

```
> score <- which(colnames(heptathlon) == "score")
> plot(heptathlon[, -score])
```



5. Comprobemos estos resultados con la matriz de correlaciones

```
> round(cor(heptathlon[, -score]), 2)
```

	hurdles	highjump	shot	run200m	longjump	javelin	run800m
hurdles	1.00	0.81	0.65	0.77	0.91	0.01	0.78
highjump	0.81	1.00	0.44	0.49	0.78	0.00	0.59
shot	0.65	0.44	1.00	0.68	0.74	0.27	0.42
run200m	0.77	0.49	0.68	1.00	0.82	0.33	0.62
longjump	0.91	0.78	0.74	0.82	1.00	0.07	0.70
javelin	0.01	0.00	0.27	0.33	0.07	1.00	-0.02
run800m	0.78	0.59	0.42	0.62	0.70	-0.02	1.00

Efectivamente, estos valores confirman que casi todos los emparejamientos entre cada par de pruebas se correlacionan positivamente, con una muy alta correlación entre el salto de longitud y los 100 metros vallas, algunos menos, como salto de altura y lanzamiento de peso, y la jabalina que presenta una correlación cercana a cero con el resto de pruebas, una posible explicación a este resultado puede ser que el entrenamiento para las otras seis pruebas, no aporta demasiado al de la jabalina, que es una prueba eminentemente técnica.

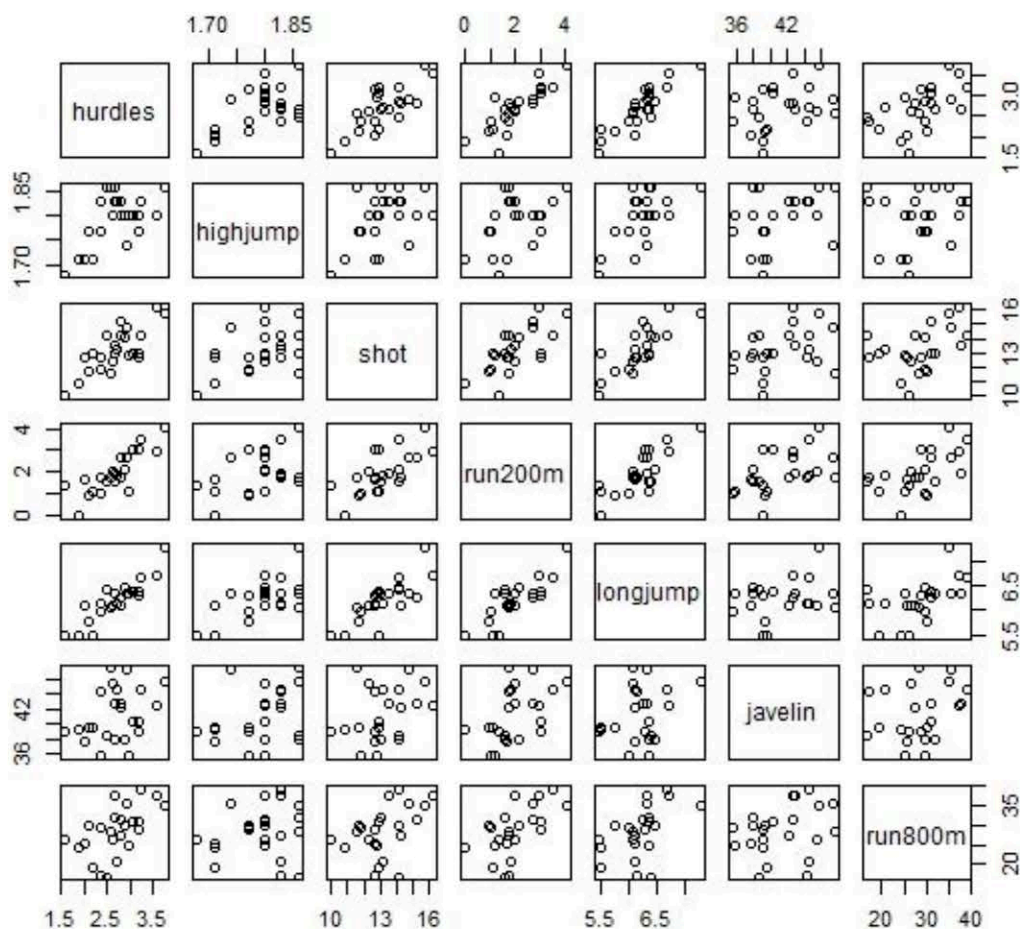
Si analizamos más detenidamente el diagrama de dispersión de la matriz, observamos que en todas las pruebas excepto en jabalina, hay un valor outlier, Launa, la participante por PNG, (Papua Nueva Guinea), que tiene peores marcas que el resto de participantes en las otras seis pruebas, de hecho es la última clasificada en el ranking de la puntuación total de las 25, pero sorprendentemente tiene la 3ª mejor marca en jabalina.

6. Puede ser interesante excluir a esta participante en un nuevo estudio del diagrama de dispersión de la matriz y de la matriz de correlaciones, para ver si los resultados son sensiblemente distintos.

```
> heptathlon <- heptathlon[-grep("PNG", rownames(heptathlon)),]
> heptathlon
```

	hurdles	highjump	shot	run200m	longjump	javelin	run800m
Joyner-Kersey (USA)	3.73	1.86	15.80	4.05	7.27	45.66	34.92
John (GDR)	3.57	1.80	16.23	2.96	6.71	42.56	37.31
Behmer (GDR)	3.22	1.83	14.20	3.51	6.68	44.54	39.23
Sablovskaitė (URS)	2.81	1.80	15.23	2.69	6.25	42.78	31.19
Choubenkova (URS)	2.91	1.74	14.76	2.68	6.32	47.46	35.53
Schulz (GDR)	2.67	1.83	13.50	1.96	6.33	42.82	37.64
Fleming (AUS)	3.04	1.80	12.88	3.02	6.37	40.28	30.89
Greiner (USA)	2.87	1.80	14.13	2.13	6.47	38.00	29.78
Lajbnerova (CZE)	2.79	1.83	14.28	1.75	6.11	42.20	27.38
Bouraga (URS)	3.17	1.77	12.62	3.02	6.28	39.06	28.69
Wijnsma (HOL)	2.67	1.86	13.01	1.58	6.34	37.86	31.94
Dimitrova (BUL)	3.18	1.80	12.88	3.02	6.37	40.28	30.89
Scheider (SWI)	2.57	1.86	11.58	1.74	6.05	47.50	28.50
Braun (FRG)	2.71	1.83	13.16	1.83	6.12	44.58	20.61
Ruotsalainen (FIN)	2.63	1.80	12.32	2.00	6.08	45.44	26.37
Yuping (CHN)	2.49	1.86	14.21	1.61	6.40	38.60	16.76
Hagger (GB)	2.95	1.80	12.75	1.14	6.34	35.76	24.95
Brown (USA)	2.35	1.83	12.69	1.78	6.13	44.34	17.00
Mulliner (GB)	2.03	1.71	12.68	1.69	6.10	37.76	25.41
Hautenaue (BEL)	2.38	1.77	11.81	1.00	5.99	35.68	29.53
Kytölä (FIN)	2.11	1.77	11.66	0.92	5.75	39.48	30.08
Geremias (BRA)	2.19	1.71	12.95	1.11	5.50	39.64	19.41
Hui-Ing (TAI)	1.57	1.68	10.00	1.38	5.47	39.14	26.13
Jeong-Mi (KOR)	1.89	1.71	10.83	0.00	5.50	39.26	24.26

```
> plot(heptathlon[, -score])
```



```
> round(cor(heptathlon[, -score]), 2)
```

	hurdles	highjump	shot	run200m	longjump	javelin	run800m
hurdles	1.00	0.58	0.77	0.83	0.89	0.33	0.56
highjump	0.58	1.00	0.46	0.39	0.66	0.35	0.15
shot	0.77	0.46	1.00	0.67	0.78	0.34	0.41
run200m	0.83	0.39	0.67	1.00	0.81	0.47	0.57
longjump	0.89	0.66	0.78	0.81	1.00	0.29	0.52
javelin	0.33	0.35	0.34	0.47	0.29	1.00	0.26
run800m	0.56	0.15	0.41	0.57	0.52	0.26	1.00

Efectivamente las correlaciones cambian sustancialmente y en el diagrama de dispersión de la matriz, no se observan ningunos valores extremos. A partir de aquí vamos a trabajar con la exclusión de la participante de Papua Nueva Guinea.

- Al estar los resultados de las 7 pruebas en diversas escalas, (metros, segundos), vamos a realizar el análisis a partir de la matriz de correlaciones, es decir vamos a realizar un Análisis de Componentes Principales Normalizado.

Una de las funciones que realiza el ACP es la función *prcomp* que lleva en este caso el valor *TRUE* en el argumento *scale* para asegurarnos que el análisis se realiza a través de la matriz de correlaciones, ver la información de esta función con esta orden

```
> help("prcomp")
```

```
> heptathlon_pca <- prcomp(heptathlon[, -score], scale = TRUE)
> heptathlon_pca
```

Standard deviations:

```
[1] 2.0793370 0.9481532 0.9109016 0.6831967 0.5461888 0.3374549 0.2620420
```

Rotation:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
hurdles	-0.4503876	0.05772161	-0.1739345	0.04840598	-0.19889364	0.84665086	-0.06961672
highjump	-0.3145115	-0.65133162	-0.2088272	-0.55694554	0.07076358	-0.09007544	0.33155910
shot	-0.4024884	-0.02202088	-0.1534709	0.54826705	0.67166466	-0.09886359	0.22904298
run200m	-0.4270860	0.18502783	0.1301287	0.23095946	-0.61781764	-0.33279359	0.46971934
longjump	-0.4509639	-0.02492486	-0.2697589	-0.01468275	-0.12151793	-0.38294411	-0.74940781
javelin	-0.2423079	-0.32572229	0.8806995	0.06024757	0.07874396	0.07193437	-0.21108138
run800m	-0.3029068	0.65650503	0.1930020	-0.57418128	0.31880178	-0.05217664	0.07718616

- Podemos sacar un resumen del análisis para observar algunos detalles más

```
> summary(heptathlon_pca)
```

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	2.0793	0.9482	0.9109	0.68320	0.54619	0.33745	0.26204
Proportion of Variance	0.6177	0.1284	0.1185	0.06668	0.04262	0.01627	0.00981
Cumulative Proportion	0.6177	0.7461	0.8646	0.93131	0.97392	0.99019	1.00000

- Los pesos de la 1ª componente principal son

```
> a1 <- heptathlon_pca$rotation[,1]
```

```
> a1
```

hurdles	highjump	shot	run200m	longjump	javelin	run800m
-0.4503876	-0.3145115	-0.4024884	-0.4270860	-0.4509639	-0.2423079	-0.3029068

Se puede observar como los 100 metros vallas y el salto de longitud, reciben los mayores pesos y la jabalina el menor. Necesitamos reescalar apropiadamente los datos para calcular la 1ª componente principal.

- Podemos obtener el centro y la escala usada por la función *prcomp* a través de *heptathlon_pca*

```
> center <- heptathlon_pca$center
```

```
> center
```

hurdles	highjump	shot	run200m	longjump	javelin	run800m
2.687500	1.793750	13.173333	2.023750	6.205417	41.278333	28.516667

```
> scale <- heptathlon_pca$scale
```

```
> scale
```

hurdles	highjump	shot	run200m	longjump	javelin	run800m
0.51456398	0.05232112	1.49714995	0.93676972	0.40165938	3.46870690	6.14724800

11. Para calcular los valores de la 1ª componente principal para cada atleta (coordenadas de los individuos), vamos a aplicar la función *scale* a los datos y multiplicar por los pesos de la 1ª componente

```
> hm <- as.matrix(heptathlon[, -score])
> drop(scale(hm, center = center, scale = scale) %*% heptathlon_pca$rotation[, 1])
```

Joyner-Kersey (USA)	John (GDR)	Behmer (GDR)	Sablovskaitė (URS)
-4.757530189	-3.147943402	-2.926184760	-1.288135516
Choubenkova (URS)	Schulz (GDR)	Fleming (AUS)	Greiner (USA)
-1.503450994	-0.958467101	-0.953445060	-0.633239267
Lajbnerova (CZE)	Bouraga (URS)	Wijnsma (HOL)	Dimitrova (BUL)
-0.381571974	-0.522322004	-0.217701500	-1.075984276
Scheider (SWI)	Braun (FRG)	Ruotsalainen (FIN)	Yuping (CHN)
0.003014986	0.109183759	0.208868056	0.232507119
Hagger (GB)	Brown (USA)	Mulliner (GB)	Hautenauve (BEL)
0.659520046	0.756854602	1.880932819	1.828170404
Kytoja (FIN)	Geremias (BRA)	Hui-Ing (TAI)	Jeong-Mi (KOR)
2.118203163	2.770706272	3.901166920	3.896847898

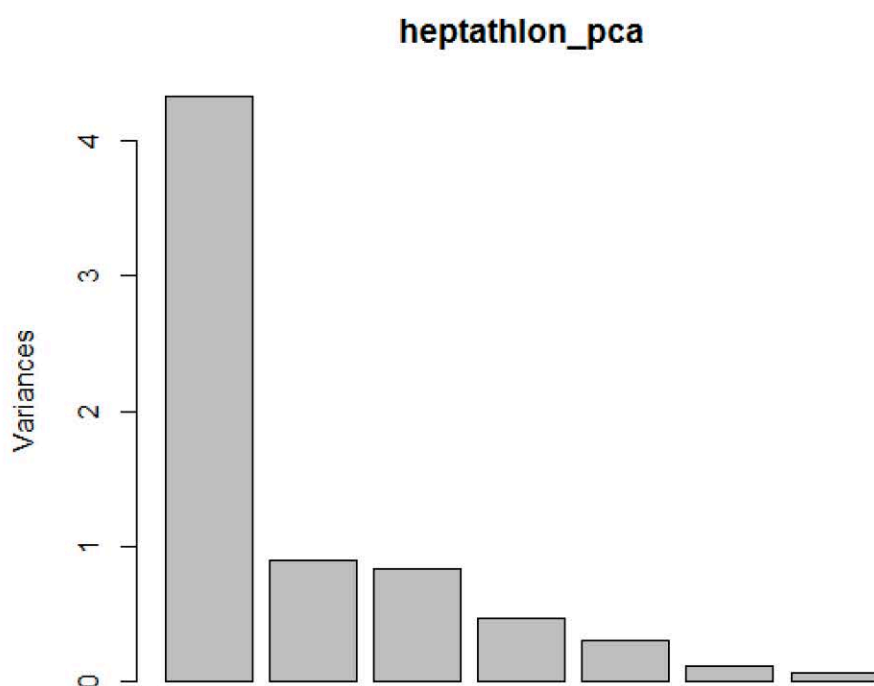
12. También podemos extraer las coordenadas de los individuos respecto al 1º eje, de la matriz que contiene todas las coordenadas de los individuos

```
> predict(heptathlon_pca)[, 1]
```

Joyner-Kersey (USA)	John (GDR)	Behmer (GDR)	Sablovskaitė (URS)
-4.757530189	-3.147943402	-2.926184760	-1.288135516
Choubenkova (URS)	Schulz (GDR)	Fleming (AUS)	Greiner (USA)
-1.503450994	-0.958467101	-0.953445060	-0.633239267
Lajbnerova (CZE)	Bouraga (URS)	Wijnsma (HOL)	Dimitrova (BUL)
-0.381571974	-0.522322004	-0.217701500	-1.075984276
Scheider (SWI)	Braun (FRG)	Ruotsalainen (FIN)	Yuping (CHN)
0.003014986	0.109183759	0.208868056	0.232507119
Hagger (GB)	Brown (USA)	Mulliner (GB)	Hautenauve (BEL)
0.659520046	0.756854602	1.880932819	1.828170404
Kytoja (FIN)	Geremias (BRA)	Hui-Ing (TAI)	Jeong-Mi (KOR)
2.118203163	2.770706272	3.901166920	3.896847898

13. Vamos a realizar un diagrama de barras de la varianza explicada por las componentes principales, es decir un diagrama de barras de los autovalores, que se obtienen elevando al cuadrado las desviaciones típicas obtenidas con la función *heptathlon_pca*

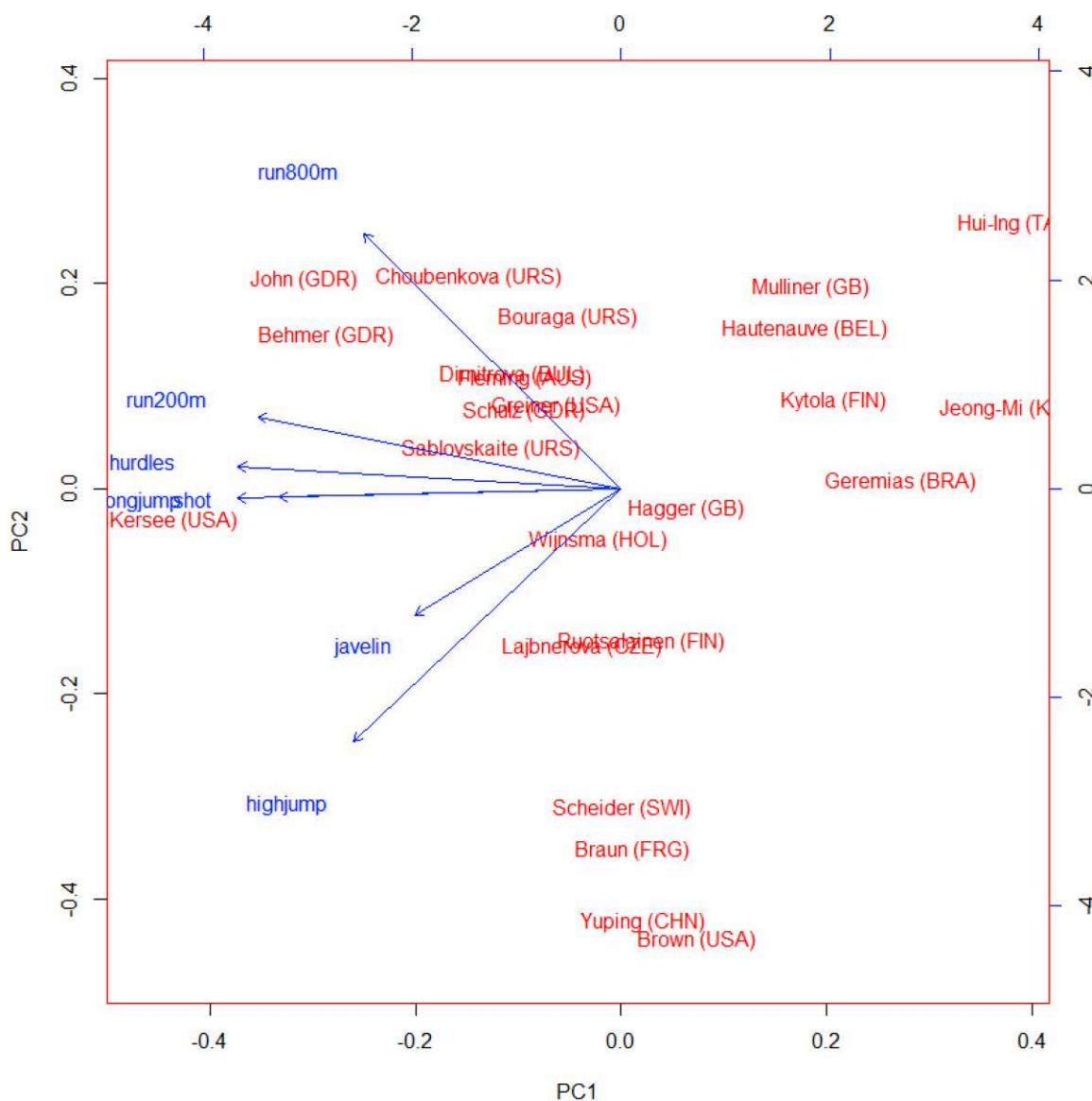
```
> plot(heptathlon_pca)
```



Las 2 primeras componentes representan el 75% de la varianza total por lo que podemos realizar una representación Biplot.

14. Vamos a realizar una gráfica de las 24 atletas sobre los 2 primeros ejes principales con la técnica Biplot

```
> biplot(heptathlon_pca, col = c("red", "blue"))
```



15. En la figura anterior tenemos no solo las coordenadas de las 24 atletas sobre los 2 primeros ejes principales sino también las coordenadas de las variables representadas por vectores, se conoce como representación punto-vector, nos da información de las varianzas y covarianzas de las variables, y de las distancias entre los individuos.

La gráfica muestra como la ganadora de la medalla de oro, Joyner-Kersey acumula la mayor puntuación en las pruebas de salto de longitud, 100 metros vallas y 200m.

16. La correlación entre la puntuación final dada a cada atleta en el heptathlon y la primera componente principal se calcula por

```
> cor(heptathlon$score, heptathlon_pca$x[,1])
[1] -0.9931168
```

Lo que significa que la 1ª componente principal está de acuerdo con la puntuación asignada a los atletas en la prueba del heptathlon.

17. Un diagrama de dispersión de la puntuación oficial dada a los atletas y la 1ª componente principal la tenemos en la gráfica:

```
> plot(heptathlon$score, heptathlon_pca$x[,1])
```

