Função prcomp

1. Introdução

Apresentamos alguns exemplos de utilização da função prcomp do pacote stats em R. Esta função permite realizar uma análise de componentes principais a partir de uma matriz de dados *n* x *p*. Na seção 2 os resultados dos comandos em R são destacados em cor azul.

Iniciamos com o gráfico de dispersão da Fig. 1, que indica uma associação linear entre as variáveis. As variáveis X_1 e X_2 apresentam amplitudes não muito diferentes. Os desvios padrão, iguais a 1,08 e 1,40, respectivamente, também são próximos. O coeficiente de correlação linear é 0,91.

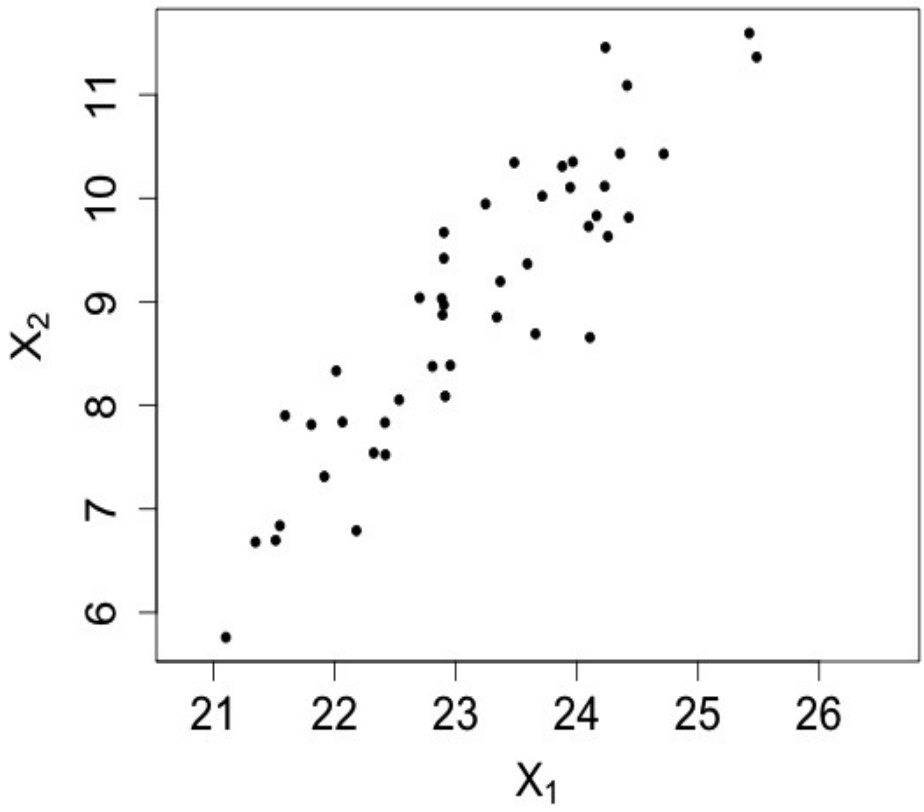

Figura 1. Gráfico de dispersão de X_1 e X_2.

Na Fig. 2 as observações (X_1 e X_2) passaram por uma transformação e são representadas como Y_1 e Y_2 em um outro sistema de eixos ortogonais. Os desvios padrão de Y_1 e Y_2 são 1,73 e 0,37, respectivamente. Sendo assim, a variação em Y_1, quantificada pela variância, é cerca de 22 vezes a variação em Y_2. A transformação $(X_1, X_2) \rightarrow (Y_1, Y_2)$ é construída utilizando os assim chamados componentes principais. As diferenças $(X_1 - \overline{X}_1, X_2 - \overline{X}_2)$ são projetadas nos dois eixos, indicados por Y_1 e Y_2, sendo que Y_1 corresponde ao eixo de maior variação. De fato, estão representados os escores dos componentes principais (Y_1^* e Y_2^*).

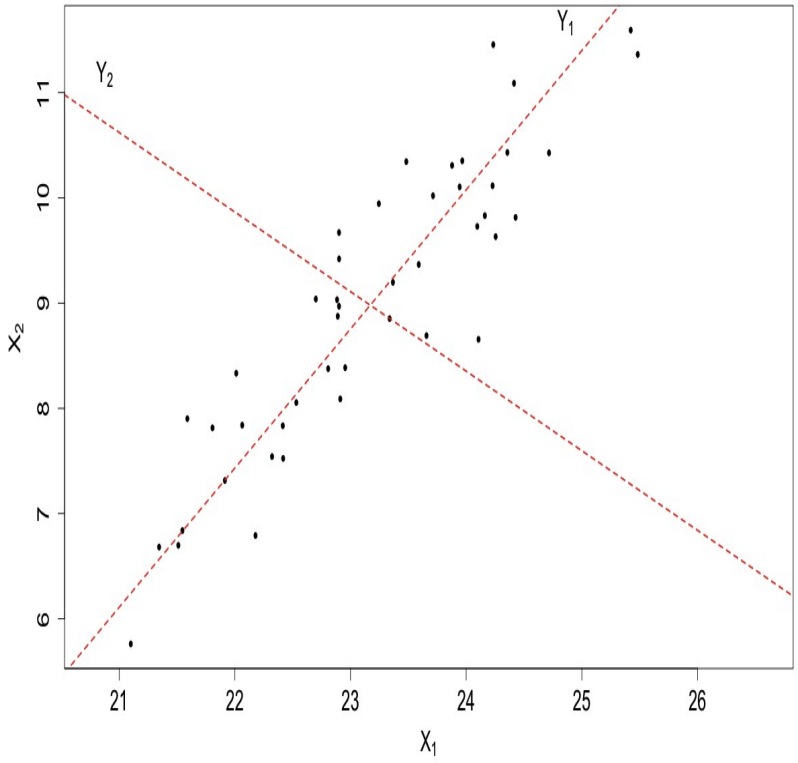

Figura 2(a). Gráfico de dispersão de X_1 e X_2 com mudança de eixos.

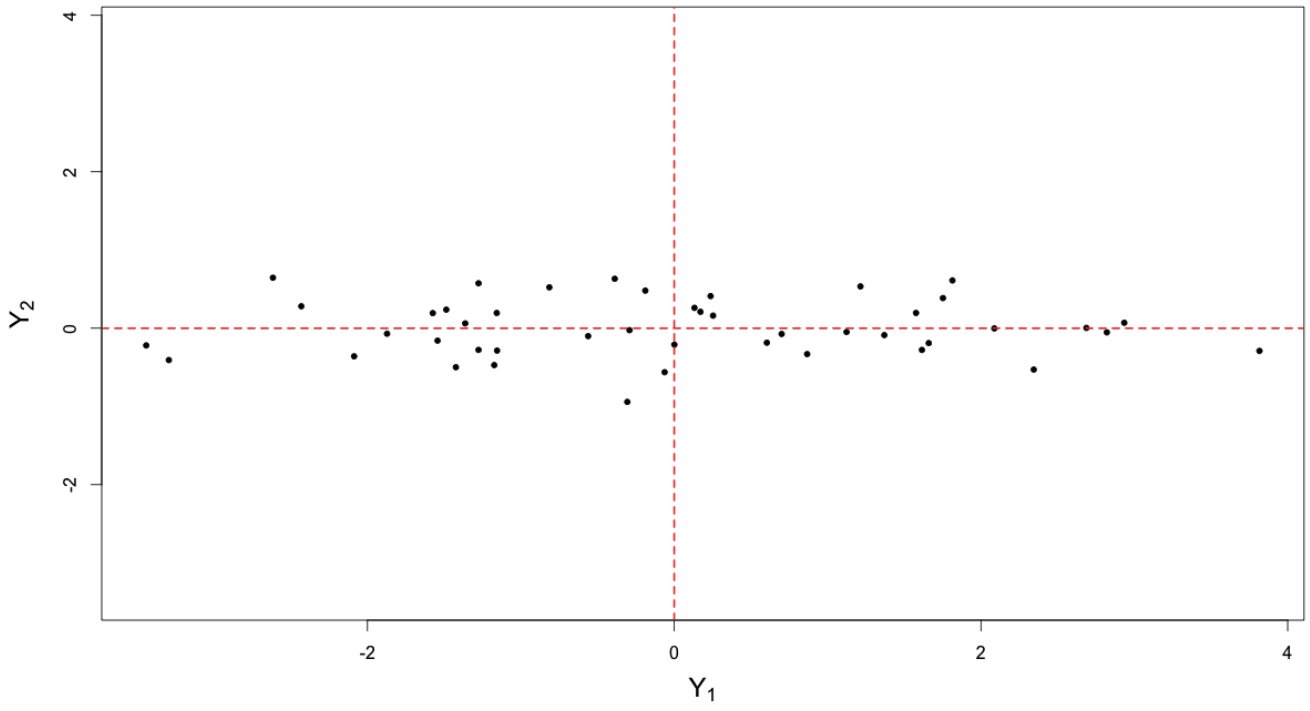

Figura 2(b). Gráfico de dispersão de Y_1 e Y_2.

2. Exemplo

O conjunto de dados dos recordes nacionais masculinos de oito provas de pista encontram-se na página **http://www.stat.wisc.edu/~rich/JWMULT06dat/T8-6.DAT** (vide também Johnson and Wichern, 2007, *Applied Multivariate Statistical Analysis*, 6th ed., p. 477). Os dados são de 2005 e dizem respeito a 54 países, listados na primeira coluna do arquivo. As demais colunas contêm os resultados das seguintes provas (unidades): 100 m (s), 200 m (s), 400 m (s), 800 m (min), 1500 m (min), 5000 m (min), 10000 m (min) e maratona (min).

Como primeira etapa, complete a linha abaixo.

```
dados <- ... ?

dim(dados)

[1] 54  8
```

O vetor `provas` contém os nomes e unidades dos resultados das provas.

```
provas <- c("100 m (s)", "200 m (s)", "400 m (s)", "800 m (min)",
            "1500 m (min)", "5000 m (min)", "10000 m (min)",
            "Maratona (min)")
```

Os nomes dos países são armazenados em um vetor, a primeira coluna da folha de dados é eliminada e os nomes das provas são anexados à folha de dados.

```
paises <- dados[, 1]
dados <- dados[, -1]
names(dados) <- provas
```

Apresentamos algumas medidas resumo destas variáveis.

```
summary(dados)

    100 m (s)       200 m (s)       400 m (s)
 Min.   : 9.78   Min.   :19.32   Min.   :43.18
 1st Qu.:10.10   1st Qu.:20.17   1st Qu.:44.91
 Median :10.20   Median :20.43   Median :45.58
 Mean   :10.22   Mean   :20.54   Mean   :45.83
 3rd Qu.:10.32   3rd Qu.:20.84   3rd Qu.:46.32
 Max.   :10.97   Max.   :22.46   Max.   :51.40

    800 m (min)     1500 m (min)    5000 m (min)
 Min.   :1.690   Min.   :3.440   Min.   :12.66
 1st Qu.:1.730   1st Qu.:3.550   1st Qu.:13.15
 Median :1.760   Median :3.610   Median :13.42
 Mean   :1.768   Mean   :3.653   Mean   :13.62
 3rd Qu.:1.800   3rd Qu.:3.737   3rd Qu.:13.91
 Max.   :1.940   Max.   :4.240   Max.   :16.70
```

```
   10000 m (min)   Maratona (min)
  Min.   :26.46   Min.   :124.5
  1st Qu.:27.55   1st Qu.:128.3
  Median :27.92   Median :130.3
  Mean   :28.54   Mean   :133.5
  3rd Qu.:28.98   3rd Qu.:134.2
    Max.   :35.38   Max.   :171.3
```

```
apply(dados, 2, sd)
```

```
       100 m (s)      200 m (s)      400 m (s)     800 m (min)
     0.2212983      0.5485466      1.4387341      0.0524521

  1500 m (min)   5000 m (min)  10000 m (min) Maratona (min)
     0.1517694      0.7608385      1.6791572      8.9518354
```

As variáveis são medidas em unidades diferentes. A análise de componentes principais será baseada na matriz de correlações amostral, que é dada abaixo.

```
matcor <- cor(dados)
print(matcor, digits = 3)
```

```
               100 m (s) 200 m (s) 400 m (s) 800 m (min)
100 m (s)          1.000     0.915     0.804       0.712
200 m (s)          0.915     1.000     0.845       0.797
400 m (s)          0.804     0.845     1.000       0.768
800 m (min)        0.712     0.797     0.768       1.000
1500 m (min)       0.766     0.795     0.772       0.896
5000 m (min)       0.740     0.761     0.780       0.86
10000 m (min)      0.715     0.748     0.766       0.843
Maratona (min)     0.676     0.721     0.713       0.807

              1500 m (min) 5000 m (min) 10000 m (min)
100 m (s)            0.766        0.740         0.715
200 m (s)            0.795        0.761         0.748
400 m (s)            0.772        0.780         0.766
800 m (min)          0.896        0.861         0.843
1500 m (min)         1.000        0.917         0.901
5000 m (min)         0.917        1.000         0.988
10000 m (min)        0.901        0.988         1.000
Maratona (min)       0.878        0.944         0.954

               Maratona (min)
100 m (s)              0.676
200 m (s)              0.721
400 m (s)              0.713
800 m (min)            0.807
1500 m (min)           0.878
5000 m (min)           0.944
10000 m (min)          0.954
Maratona (min)         1.000
```

Uma matriz de gráficos de dispersão permite representar graficamente os dados.

```
library(lattice)
splom(dados, pch = 20, col = "black", xlab = "")
```

Nos gráficos da Fig. 3 notamos que as maiores correlações amostrais são observadas nos painéis próximos à diagonal secundária (vide os valores das correlações logo acima). Alguma explicação?

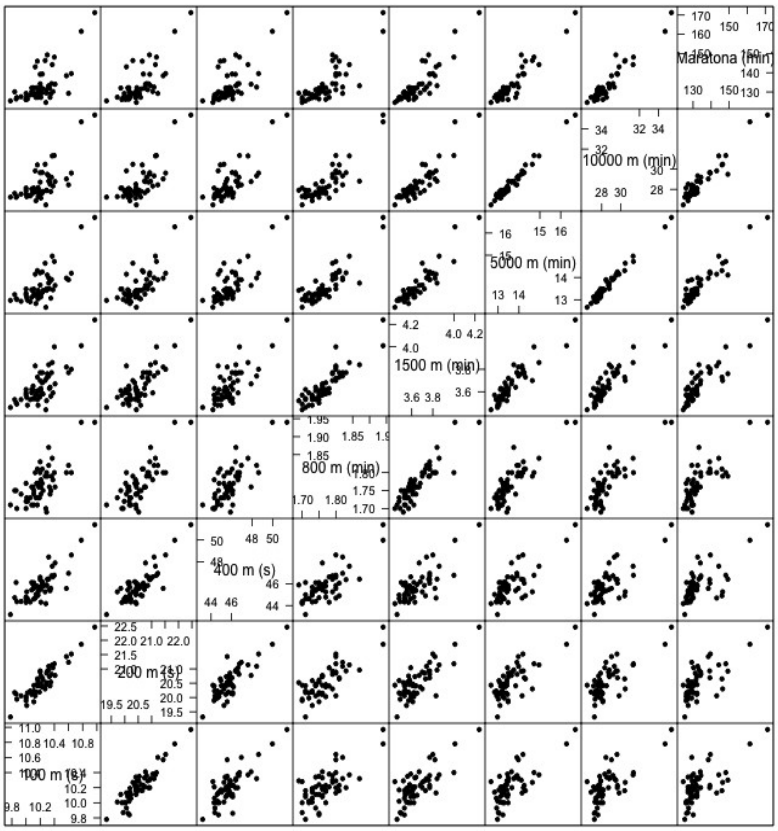

Figura 3. Gráficos de dispersão dos recordes nacionais masculinos.

A fim de utilizar a matriz de correlações amostral na análise, devemos especificar o argumento `scale` como `TRUE` na chamada da função prcomp.

```
acpcor <- prcomp(dados, scale = TRUE)
summary(acpcor)
```

```
Importance of components:
                          PC1    PC2     PC3     PC4    PC5
Standard deviation     2.5891 0.7990 0.47700 0.45371 0.3124
Proportion of Variance 0.8379 0.0798 0.02844 0.02573 0.0122
Cumulative Proportion  0.8379 0.9177 0.94615 0.97188 0.9841
                           PC6     PC7     PC8
Standard deviation     0.26587 0.21666 0.09858
Proportion of Variance 0.00884 0.00587 0.00121
Cumulative Proportion  0.99292 0.99879 1.00000
```

Os desvios padrão dos componentes principais estão no componente `sdev`. A soma dos quadrados deles é igual a 8, conforme esperado ($p = 8$).

```
sum(acpcor$sdev^2)
```

```
[1] 8
```

O primeiro componente principal responde por cerca de 84% da variância total dos dados padronizados, ao passo que se tomarmos os dois primeiros componentes atingimos cerca de 92% da variância total. O primeiro componente principal tem variância 6,70 (`acpcor$sdev[1]^2`), bem maior do que a média das variâncias (igual a 1). Além disso, o gráfico da Fig. 4 indica que o número de componentes a reter é dois.

```
plot(1:ncol(dados), acpcor$sdev^2, type = "b", xlab = "Componente",
     ylab = "Variância", pch = 20, cex.axis = 1.3, cex.lab = 1.3)
```

Nota 1. Verifique os resultados dos comandos `screeplot(acpcor)` e `plot(acpcor)`.

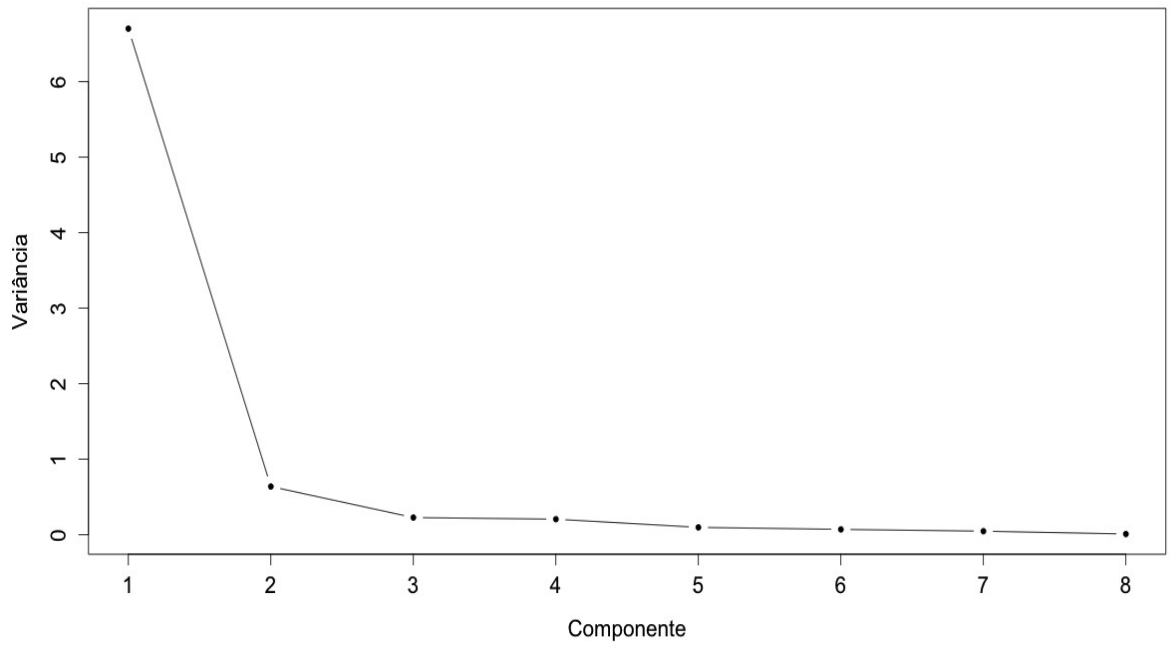

Figura 4. Gráfico de escarpa para os dados dos recordes nacionais masculinos.

Os coeficientes dos componentes principais (isto é, os autovetores $\boldsymbol{e}_1$, $\boldsymbol{e}_2$, ..., $\boldsymbol{e}_p$ da matriz de correlações amostral) estão armazenados no componente `rotation` (ou seja, `acpcor$rotation`). Os dois primeiros são mostrados abaixo.

```
acpcor$rotation[, 1:2]
                     PC1         PC2
100 m (s)      -0.3323877 -0.52939911
200 m (s)      -0.3460511 -0.47039050
400 m (s)      -0.3391240 -0.34532929
800 m (min)    -0.3530134  0.08945523
1500 m (min)   -0.3659849  0.15365241
5000 m (min)   -0.3698204  0.29475985
10000 m (min)  -0.3659489  0.33360619
Maratona (min) -0.3542779  0.38656085
```

As correlações entre os dois primeiros componentes principais e as variáveis são estimadas abaixo.

```
print(acpcor$sdev[1:2] * t(acpcor$rotation[, 1:2]), digits = 3)
```

Nota 2. Qual seria a expressão se fosse utilizada a matriz de covariâncias amostral?

```
    100 m (s) 200 m (s) 400 m (s) 800 m (min) 1500 m (min)
PC1    -0.861    -0.896    -0.878     -0.9140       -0.948
PC2    -0.423    -0.376    -0.276      0.0715        0.123

   5000 m (min) 10000 m (min) Maratona (min)
PC1      -0.957        -0.947         -0.917
PC2       0.236         0.267          0.309
```

O primeiro componente tem correlação forte com todas as variáveis, ao passo que o segundo componente tem correlação amostral próxima de 0 com o recorde na prova de 800 m.

Como os elementos de $\boldsymbol{e}_1$ têm o mesmo sinal, podemos interpretar o primeiro componente principal como sendo um índice de desempenho atlético em provas de corrida. Quanto maior o valor deste índice, mais alto o desempenho do país (Por quê?). O segundo componente principal pode ser interpretado como um contraste entre resultados de provas de distâncias mais curtas (100, 200 e 400 m) e as demais.

Muitas vezes, convencionamos que o elemento e_{11} deve ser positivo. Neste exemplo, o primeiro autovetor deveria ser multiplicado por -1. Após esta mudança, algumas interpretações devem ser adaptadas.

Se especificarmos `retx` como `TRUE`, obtemos diretamente os escores dos componentes principais (componente `x` de `acpcor`).

```
acpcor <- prcomp(dados, scale = TRUE, retx = TRUE)
```

Nota 3. Qual o resultado do comando `cor(acpcor$x)`? Surpresa?

Tendo em vista que o primeiro componente principal pode ser visto como um indicador de desempenho atlético, utilizamos os escores deste componente para classificar os países.

```
escores1 <- acpcor$x[, 1]
names(escores1) <- paises
ordem <- order(escores1, decreasing = TRUE)
barplot(escores1[ordem], ylab = "Escore do CP1", las = 2)
box()
```

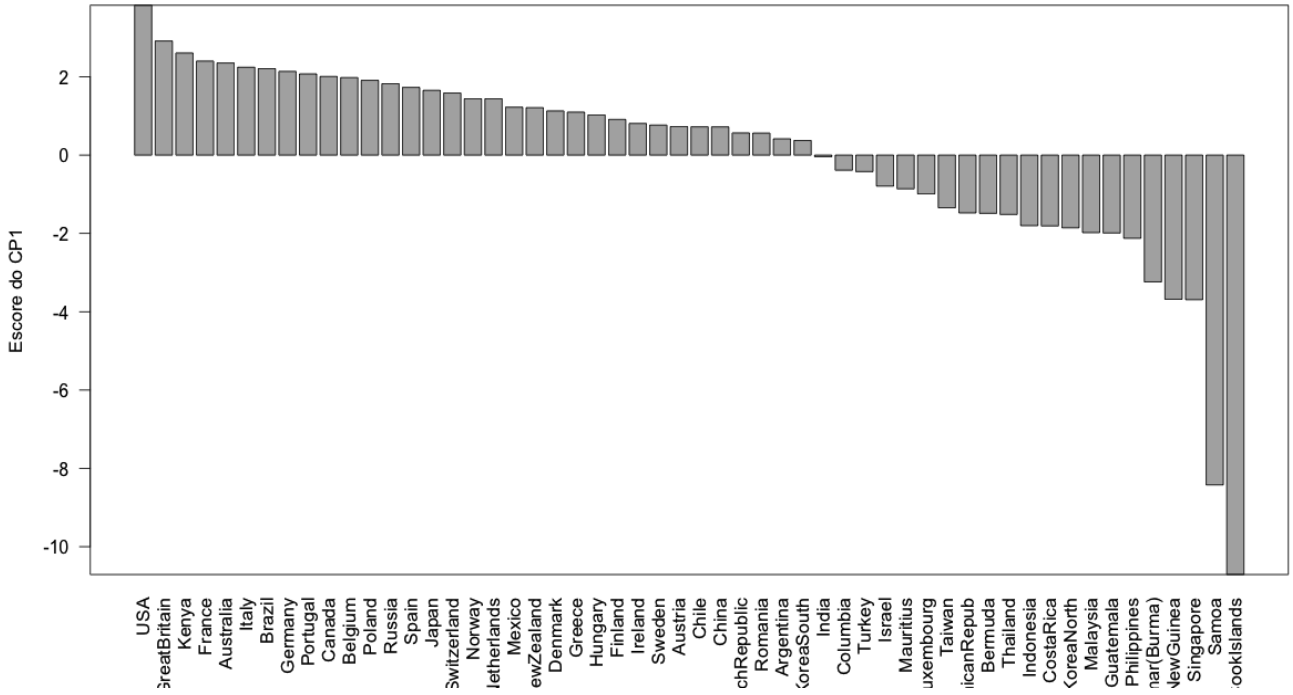

Figura 5. Escores do primeiro componente principal para os dados dos recordes nacionais masculinos.

O gráfico de dispersão dos escores dos dois primeiros componentes baseados na matriz de correlações juntamente com os respectivos autovetores é mostrado na Fig. 6. Este gráfico é chamado de *biplot*. É uma representação bidimensional de dados multivariados.

```
biplot(acpcor, xlab = "CP1", ylab = "CP2",cex.lab = 1.5, cex.axis = 1.5)
```

As observações (os países) na Fig. 6 são identificadas pelos seus números, ao passo que os autovetores permitem representar as variáveis. Cada observação é representada pelo par de escores dos dois primeiros componentes principais (Y_{1i}^*, Y_{2i}^*), $i = 1, ..., n$. Cada variável é representada por um vetor na direção conectando os pontos (0,0) e (e_{1j}, e_{2j}), $j = 1, \ldots, p$. Os comprimentos destes vetores são proporcionais às variâncias das variáveis (lembrando que neste exemplo as variáveis foram padronizadas). As duas escalas (Y^* e e) podem não ser compatíveis, de modo que pode haver necessidade de mudança de escala para adequar a representação gráfica. Os ângulos entre os vetores estão relacionados às correlações entre as variáveis, sendo que quanto menor o ângulo, mais correlacionadas estão. As posições dos pontos (países) no gráfico indicam semelhanças e diferenças entre eles.

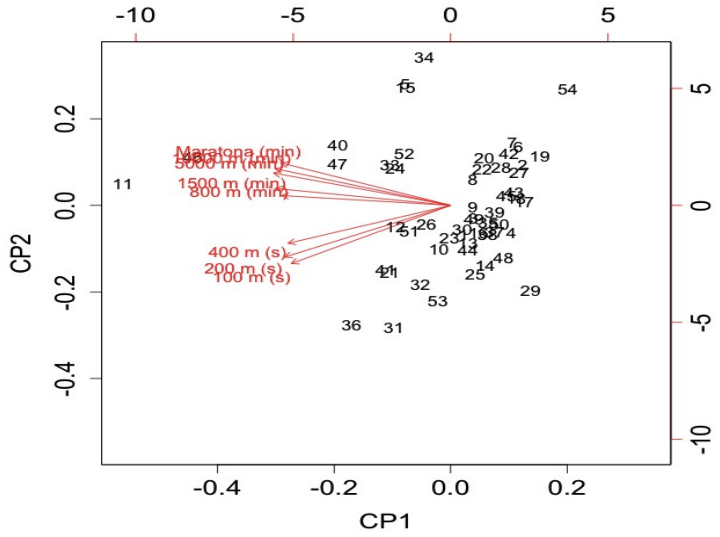

Figura 6. *Biplot* para os dados dos recordes nacionais masculinos.

Os países à esquerda são aqueles com menor desempenho atlético. Alguns grupos de países podem ser identificados.

Os últimos componentes principais podem ser úteis na identificação de observações aberrantes. A análise deve ser efetuada sem estas observações e comparada com os resultados da análise incluindo todas as observações. O gráfico da Fig. 7 não sugere a existência de observações aberrantes.

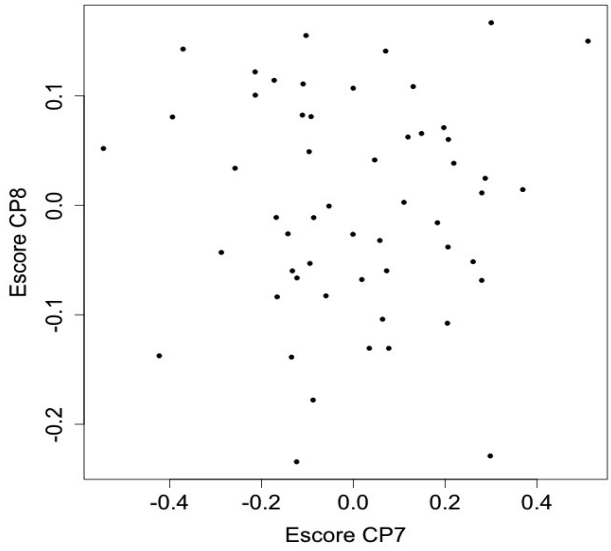

Figura 7. Gráfico de dispersão dos escores dos dois últimos componentes para os dados dos recordes nacionais masculinos.

Nota 4. Procure reproduzir os resultados utilizando outros pacotes estatísticos (por exemplo, SAS, Minitab, SPSS e Statistica).