

# Análise de correlações canônicas

(em linguagem R)

2025

O conjunto de dados dos recordes nacionais masculinos de oito provas de pista encontra-se na página <https://pages.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). Quanto menor for o tempo recorde, mais alto é o desempenho.

```
## Leitura dos dados
dados <- read.table("https://pages.stat.wisc.edu/~rich/JWMULT06dat/T8-6.DAT")
n <- nrow(dados)
p <- ncol(dados) - 1
cat("\n n =", n, ", p =", p, "\n")
```

```
##
## n = 54 , p = 8
```

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

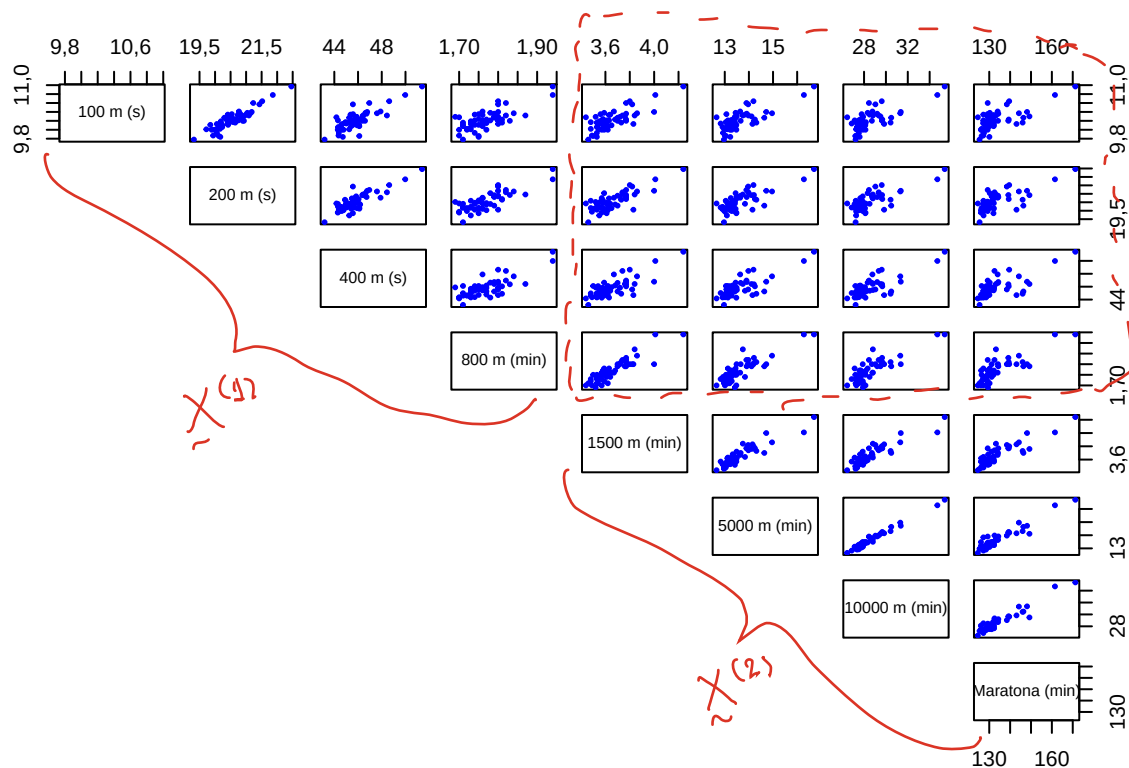
```
provas <- c("100 m (s)", "200 m (s)", "400 m (s)", "800 m (min)",
            "1500 m (min)", "5000 m (min)", "10000 m (min)",
            "Maratona (min)")
```

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
```

Uma matriz de gráficos de dispersão permite representar graficamente os dados. Notamos um padrão de associação linear positiva entre as variáveis.

```
pairs(dados, lower.panel = NULL, pch = 20, col = "blue", cex = 0.6)
```



Consideramos dois conjuntos de variáveis identificados como provas de distância mais curta (100 m, 200 m, 400 m e 800 m) e provas de distância mais longa (1500 m, 5000 m, 10000 m e Maratona). Estes conjuntos correspondem às colunas 1–4 e 5–8, respectivamente.

```
i1 <- 1:4 # X(1)
i2 <- 5:8 # X(2)
matcor <- cor(dados)
print(matcor[i1, i2], digits = 3)
```

```
##           1500 m (min) 5000 m (min) 10000 m (min) Maratona (min)
## 100 m (s)      0,766      0,740      0,715      0,676
## 200 m (s)      0,795      0,761      0,748      0,721
## 400 m (s)      0,772      0,780      0,766      0,713
## 800 m (min)    0,896      0,861      0,843      0,807
```

No pacote CCA temos a função `cc` que implementa análise de correlações canônicas. As variáveis são padronizadas (por quê?).

```
library(CCA)
dados <- scale(dados) # padronização
mac <- cc(dados[, i1], dados[, i2])
names(mac)
```

```
## [1] "cor"      "names"    "xcoef"    "ycoef"    "scores"
```

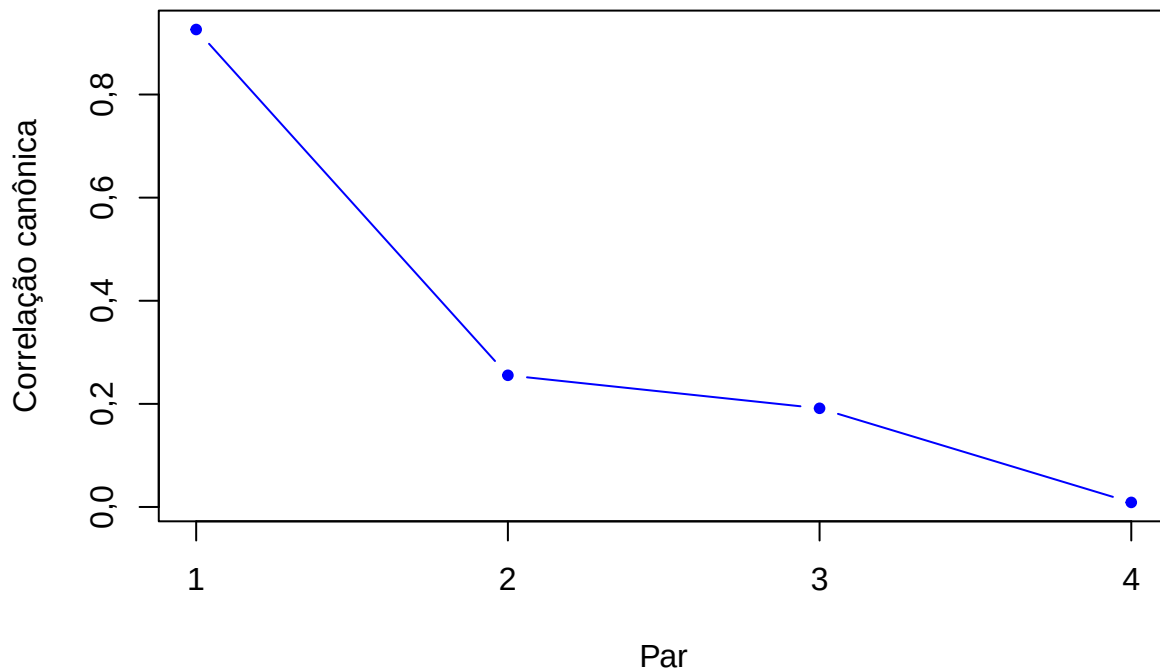
```
cat("\n Correlações canônicas =", mac$cor, "\n")
```

```
##
```

```
## Correlações canônicas = 0,9260961 0,2554187 0,1913998 0,008881891
```

```
plot(1:length(mac$cor), mac$cor, type = "b", xlab = "Par", pch = 20, col = "blue",
     ylab = "Correlação canônica", axes = FALSE)
axis(1, 1:length(mac$cor), 1:length(mac$cor))
axis(2)
```

```
box()
```



Os vetores de coeficientes do primeiro par de variáveis canônicas são dados abaixo.

```
## a1
```

```
mac$xcoef[, 1]
```

```
## 100 m (s) 200 m (s) 400 m (s) 800 m (min)
## -0,3698043 0,2169679 -0,1542022 -0,7646822
```

```
## b1
```

```
mac$ycoef[, 1]
```

```
## 1500 m (min) 5000 m (min) 10000 m (min) Maratona (min)
## -0,6970880 -0,7262171 0,2189074 0,2025012
```

**Exercício 1.** Procure interpretar o primeiro par de variáveis canônicas.

Tomando  $A = [a_1 : a_2 : \dots : a_p]$  e  $B = [b_1 : b_2 : \dots : b_q]$ , pode ser provado que as matrizes de correlações entre  $U$  e  $X^{(1)}$  e entre  $V$  e  $X^{(2)}$ , são

$$A\Sigma_{11}V_{11}^{-1/2} \quad \text{e} \quad B\Sigma_{22}V_{22}^{-1/2}, \quad (1)$$

respectivamente, em que  $V_{11}^{-1/2}$  e  $V_{22}^{-1/2}$  são matrizes diagonais com elementos  $1/\sqrt{\sigma_{11j}}$ ,  $j = 1, \dots, p$  e  $1/\sqrt{\sigma_{22j}}$ ,  $j = 1, \dots, q$ , respectivamente. Estimativas das correlações entre o primeiro par de variáveis canônicas e as variáveis originais são mostradas abaixo.

```
## cor(X(1), U1)
```

```
mac$scores$corr.X.xscores[, 1]
```

```
## 100 m (s) 200 m (s) 400 m (s) 800 m (min)
## -0,8397350 -0,8609882 -0,8553316 -0,9734437
```

```
## cor(X(2), V1)
```

```
mac$scores$corr.Y.yscores[, 1]
```

```
## 1500 m (min) 5000 m (min) 10000 m (min) Maratona (min)
## -0,9876215 -0,9575917 -0,9338567 -0,8861700
```

Considerando variáveis padronizadas, para o  $k$ -ésimo par de variáveis canônicas ( $U_k, V_k$ ), as proporções da variância total que são explicadas por cada variável canônica em separado são dadas por

$$\frac{1}{p} \sum_{j=1}^p \{\text{cor}(U_k, X_j^{(1)})\}^2 \quad \text{e} \quad \frac{1}{q} \sum_{j=1}^q \{\text{cor}(V_k, X_j^{(2)})\}^2, \quad (2)$$

respectivamente, para  $k = 1, \dots, \min(p, q)$ . Tomando o primeiro par de variáveis canônicas, as estimativas (em %) são dadas abaixo.

```
## U1
sum(mac$scores$corr.X.xscores[, 1]^2) / length(i1) * 100
```

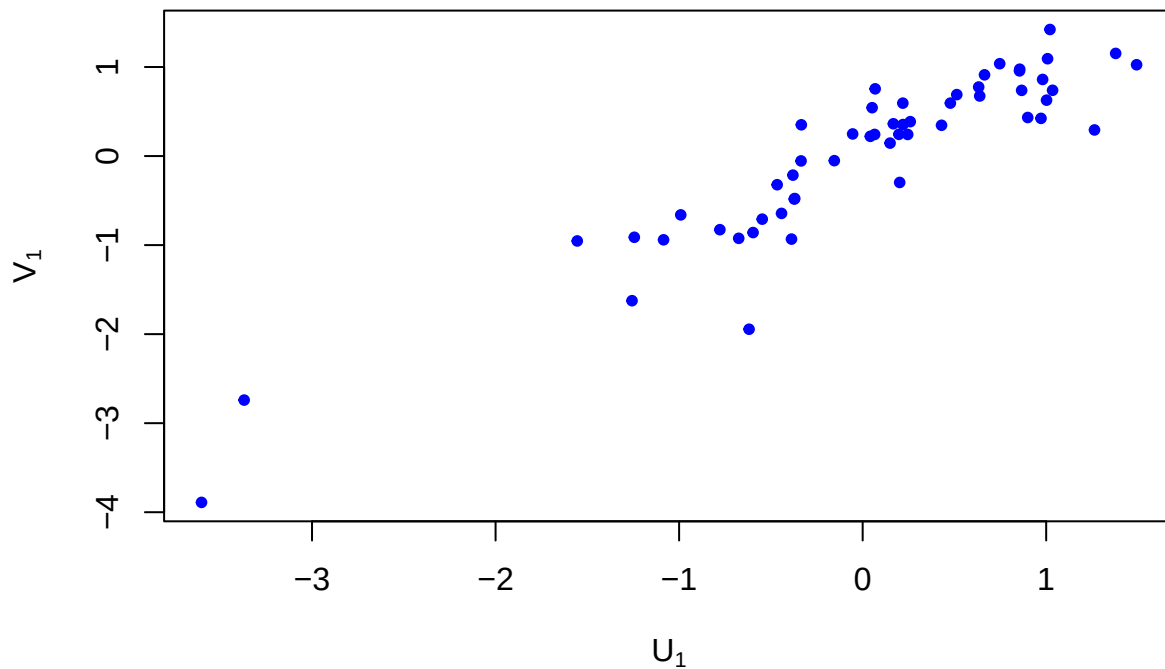
```
## [1] 78,14101
```

```
## V1
sum(mac$scores$corr.Y.yscores[, 1]^2) / length(i2) * 100
```

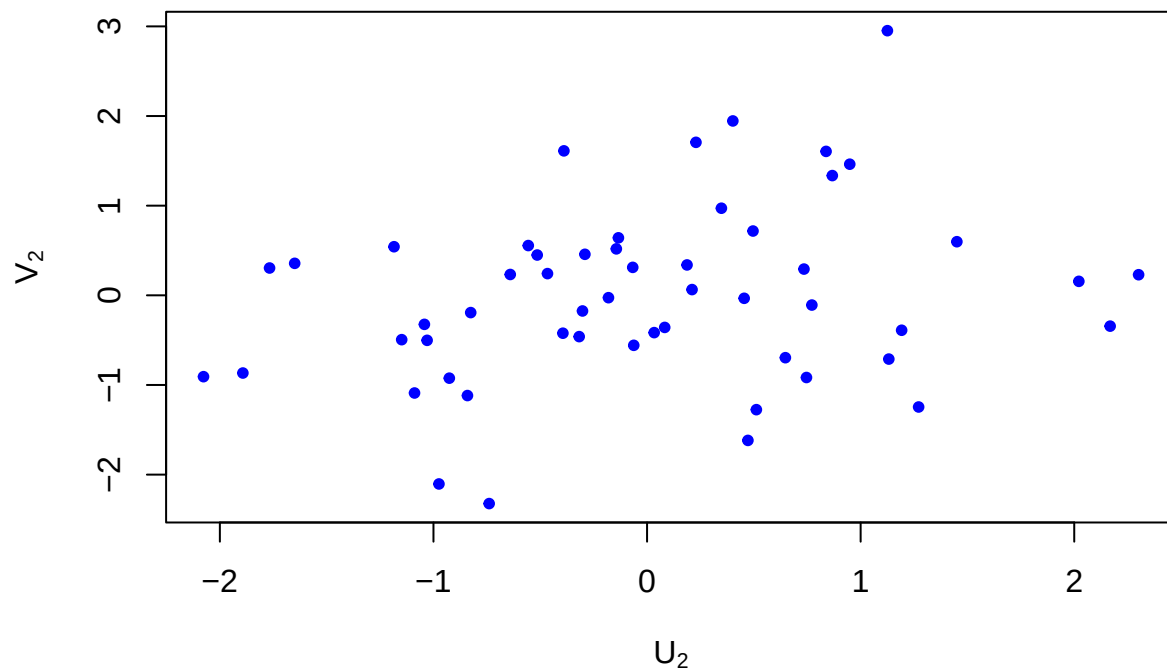
```
## [1] 88,74409
```

Em seguida mostramos os gráficos de dispersão dos pares de escores das variáveis canônicas 1, 2 e 4.

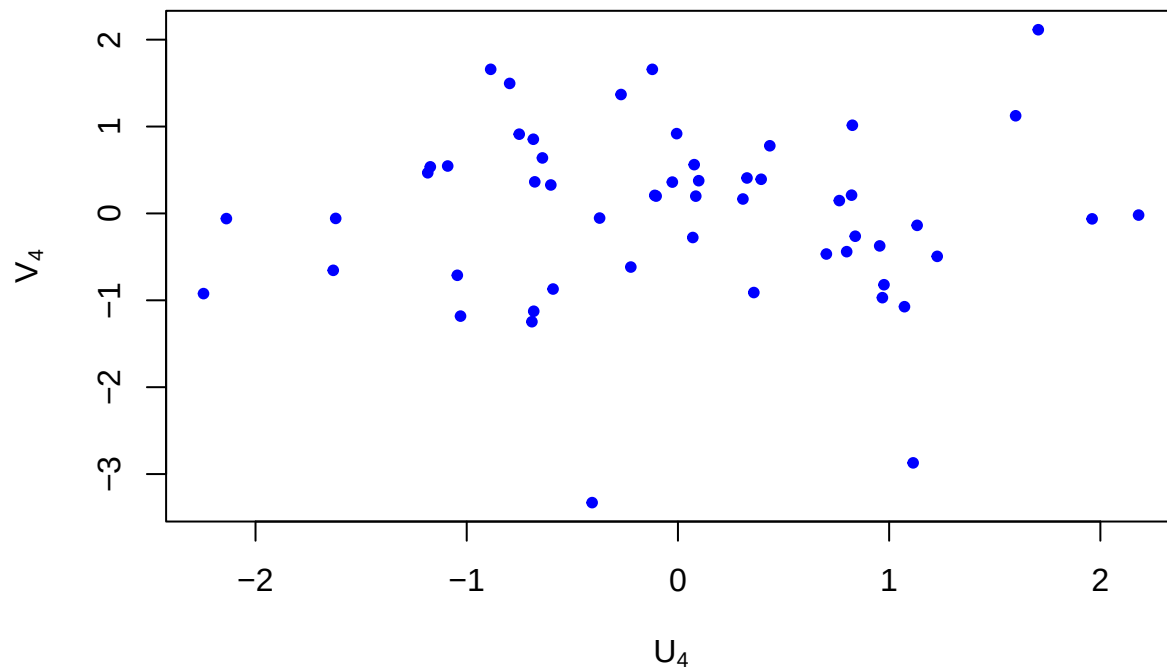
```
plot(mac$scores$xscores[, 1], mac$scores$yscores[, 1], pch = 20, col = "blue",
      xlab = expression(U[1]), ylab = expression(V[1]))
```



```
plot(mac$scores$xscores[, 2], mac$scores$yscores[, 2], pch = 20, col = "blue",
      xlab = expression(U[2]), ylab = expression(V[2]))
```



```
plot(mac$scores$xscores[, 4], mac$scores$yscores[, 4], pch = 20, col = "blue",
     xlab = expression(U[4]), ylab = expression(V[4]))
```



**Exercício 2.** Refaça o exemplo em linguagem Python.