

Análise fatorial

2026

1. Dados

Os dados referem-se a $n = 54$ observações de $p = 7$ variáveis apresentadas na Tabela 1.9, p. 44 em Johnson and Wichern (2007, *Applied Multivariate Statistical Analysis*, sixth ed., Upper Saddle River, NJ: Pearson/Prentice Hall). Os dados são de 2005 e dizem respeito aos recordes femininos de provas de corrida em 54 países, listados na primeira coluna do arquivo. As demais colunas contêm os resultados das seguintes provas (e unidades): 100 m (s), 200 m (s), 400 m (s), 800 m (min), 1500 m (min), 3000 m (min) e maratona (min). Quanto menor for o tempo recorde, melhor é o resultado de cada prova.

Nota 1. A função `fa` do pacote `psych` efetua o ajuste de um modelo de análise fatorial. Esta função não será usada aqui.

```
## Leitura dos dados
dados <- read.delim("https://pages.stat.wisc.edu/~rich/JWMULT06dat/T1-9.dat",
                    header = FALSE )
n <- nrow(dados)
p <- ncol(dados) - 1
cat("\n n =", n, ", p =", p, "\n")
```

```
##
##  n = 54 , p = 7
```

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)",
            "3000 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.

```
países <- dados[, 1]
dados <- dados[, -1]
names(dados) <- provas
países[1:10]
```

```
## [1] "ARG" "AUS" "AUT" "BEL" "BER" "BRA" "CAN" "CHI" "CHN" "COL"
```

As variáveis são medidas em unidades diferentes. A análise será aplicada à matriz de correlações amostral.

```
summary(dados)
```

```
##      100 m (s)      200 m (s)      400 m (s)      800 m (min)      1500 m (min)
## Min.   :10.49   Min.   :21.34   Min.   :47.60   Min.   :1.890   Min.   :3.840
## 1st Qu.:11.12   1st Qu.:22.57   1st Qu.:49.97   1st Qu.:1.970   1st Qu.:4.003
## Median :11.32   Median :22.98   Median :51.65   Median :2.005   Median :4.100
## Mean   :11.36   Mean   :23.12   Mean   :51.99   Mean   :2.022   Mean   :4.189
## 3rd Qu.:11.57   3rd Qu.:23.61   3rd Qu.:53.12   3rd Qu.:2.070   3rd Qu.:4.338
## Max.   :12.52   Max.   :25.91   Max.   :61.65   Max.   :2.290   Max.   :5.420
##      3000 m (min)      Maratona (min)
```

```
## Min. : 8.100 Min. :135.2
## 1st Qu.: 8.543 1st Qu.:143.5
## Median : 8.845 Median :148.4
## Mean : 9.081 Mean :153.6
## 3rd Qu.: 9.325 3rd Qu.:157.7
## Max. :13.120 Max. :221.1
```

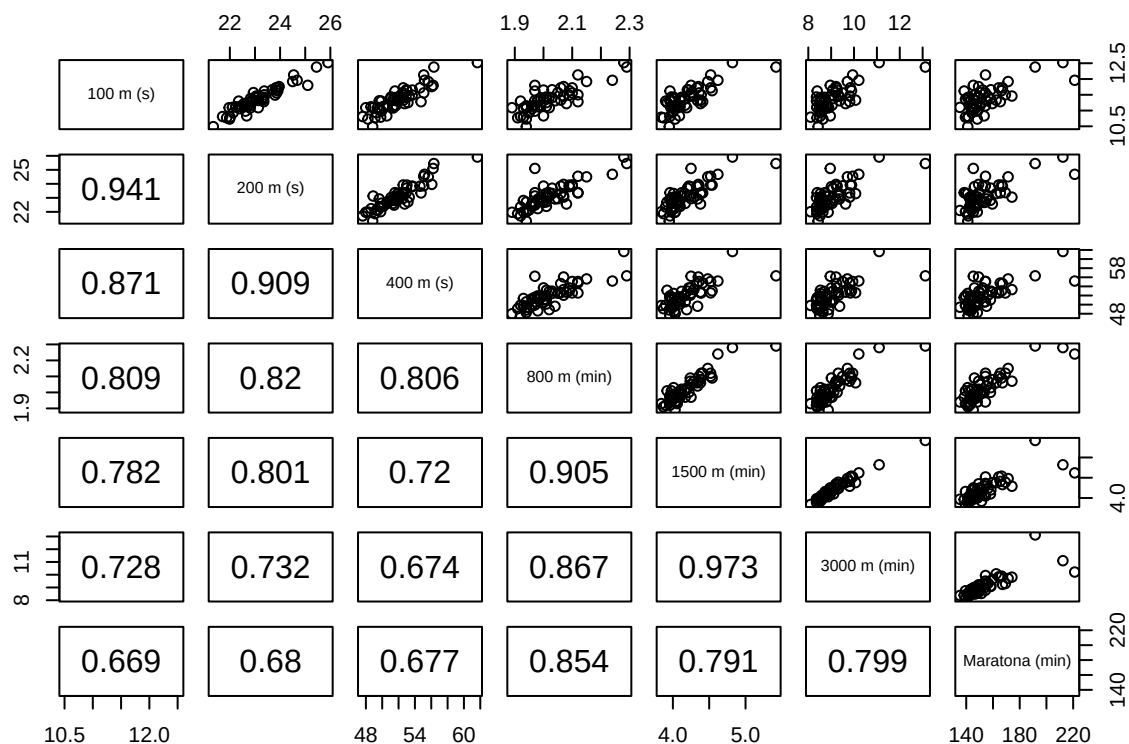
Médias e desvios padrão

```
xb <- apply(dados, MARGIN = 2, FUN = mean)
(desvp <- apply(dados, MARGIN = 2, FUN = sd))
```

```
##      100 m (s)      200 m (s)      400 m (s)      800 m (min)      1500 m (min)
##      0.39410116      0.92902547      2.59720188      0.08687304      0.27236502
##      3000 m (min) Maratona (min)
##      0.81532689      16.43989508
```

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

```
panel.cor = function(x, y, digits = 3) {
  par(usr = c(0, 1, 0, 1))
  r = cor(x, y)
  text(0.5, 0.5, round(r, digits), cex = 1.5)
}
pairs(dados, lower.panel = panel.cor)
```



Como as variáveis se referem a provas de corrida diferentes, a análise fatorial será baseada na matriz de correlações amostral.

```
matcor <- cor(dados)
```

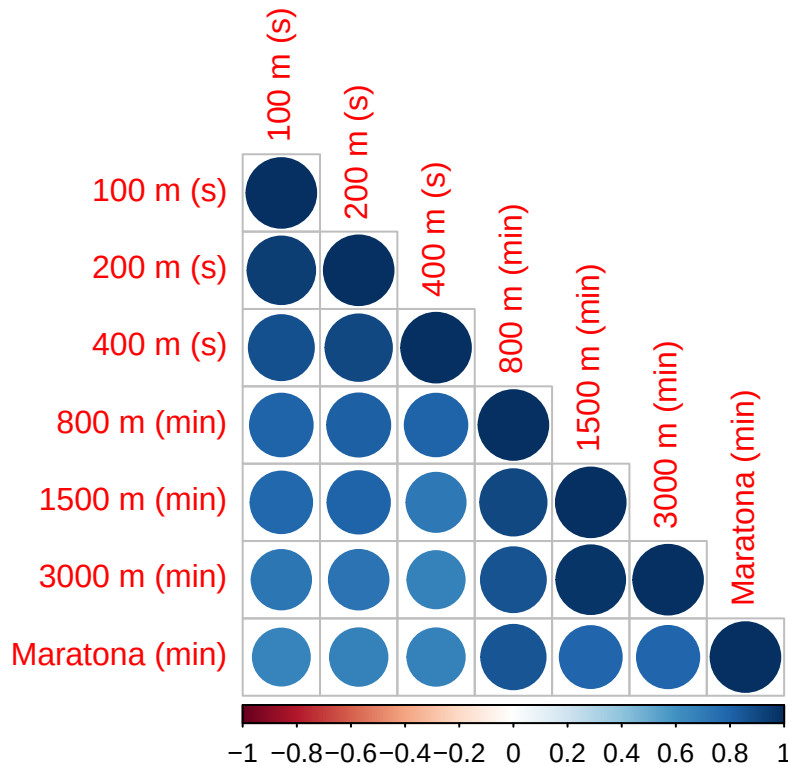
As correlações amostrais variam de 0,67 a 0,97.

```
range(matcor[matcor < 1.00])
```

```
## [1] 0.6689597 0.9733801
```

Uma matriz de correlações também pode ser representada em um mapa de calor, que está implementado em R na função `corrplot` do pacote `corrplot`.

```
library(corrplot)
corrplot(matcor, type = "lower")
```



Na figura acima todas as correlações estão representadas em cor azul porque são positivas, com tons mais fortes e círculos com diâmetros maiores para correlações mais altas.

2. Medidas de adequação amostral

São utilizadas como indicativo para avaliar se é adequado analisar os dados com a técnica de análise fatorial. Algumas medidas são baseadas em coeficientes de correlação entre os pares de variáveis X_j e X_k , $j \neq k$.

O coeficiente de correlação entre X_j e X_k após eliminar os efeitos de X_l , $l \neq j$, $l \neq k$ é chamado de coeficiente de correlação parcial entre X_j e X_k .

Os resíduos da regressão de X_j como variável resposta e X_l , $l \neq j$, $l \neq k$, como variáveis explicativas são denotados por d_{ji} , $i = 1, \dots, n$. Os resíduos da regressão de X_k como variável resposta e X_l , $l \neq j$, $l \neq k$, como variáveis explicativas são denotados por d_{ki} , $i = 1, \dots, n$. O coeficiente de correlação parcial amostral entre X_j e X_k é dado pelo coeficiente de correlação linear de Pearson entre os resíduos (d_{ji}, d_{ki}) , $i = 1, \dots, n$. É denotado por $r_{pa,jk}$, com $r_{pa,jj} = 1$, $j = 1, \dots, p$.

Se houver uma associação linear expressiva entre as variáveis (que é desejável para uma aplicação bem sucedida da técnica de análise fatorial), o valor de $|r_{pa,jk}|$ deve ser “pequeno” comparado a $|r_{jk}|$, sendo que r_{jk} representa o coeficiente de correlação linear de Pearson entre X_j e X_k . Denotando por $\mathbf{C}$ a inversa da

matriz de correlações amostral $\mathbf{P}$ ($\mathbf{C} = \mathbf{P}^{-1}$), pode ser provado que

$$r_{pa,jk} = -\frac{c_{jk}}{\sqrt{c_{jj} c_{kk}}}. \quad (1)$$

Uma medida global de adequação amostral é dada pela estatística KMO (Kaiser-Meyer-Olkin; vide Dziuban and Shirkey, 1974, *Psychological Bulletin* 81, 358–361), cuja expressão é

$$\text{KMO} = \frac{\sum_{j=1}^p \sum_{l=1, l \neq j}^p r_{jl}^2}{\sum_{j=1}^p \sum_{l=1, l \neq j}^p r_{jl}^2 + \sum_{j=1}^p \sum_{l=1, l \neq j}^p r_{pa,jl}^2}. \quad (2)$$

Uma medida de adequação amostral para a variável X_j é dada por

$$\text{MAA}_j = \frac{\sum_{l=1, l \neq j}^p r_{jl}^2}{\sum_{l=1, l \neq j}^p r_{jl}^2 + \sum_{l=1, l \neq j}^p r_{pa,jl}^2}, \quad j = 1, \dots, p. \quad (3)$$

A tabela abaixo apresenta uma síntese (sugestão) da adequação amostral (em tradução livre).

KMO	Adequação
> 0,9	Excelente
(0,8; 0,9]	Meritória
(0,7; 0,8]	Intermediária
(0,6; 0,7]	Medíocre
(0,5; 0,6]	Mísera
< 0,5	Inaceitável

As correlações parciais podem ser calculadas com a função `pcor` do pacote `ppcor`.

```
library(ppcor)
matcorp <- pcor(dados, method = "pearson")$estimate
```

Agora podemos obter as estatísticas de adequação KMO e MAA. Na primeira linha abaixo, `idiag` representa as posições da diagonal principal quando os elementos de uma matriz $p \times p$ são armazenados em um vetor.

```
idiag <- seq(1, by = p + 1, length = p)
somar2 <- sum((as.numeric(matcor)[-idiag])^2)
cat("\n KMO = ", somar2 / (somar2 + sum((as.numeric(matcorp)[-idiag])^2)))
```

```
##
## KMO = 0.8160765
```

A adequação amostral é meritória. As medidas MAA são calculadas para cada variável.

```
for (j in 1:p) {
  somar2j <- sum(matcor[j, -j]^2)
  cat("\n MAA", provas[j], "=", somar2j / (somar2j + sum(matcorp[j, -j]^2)))
}
```

```
##
## MAA 100 m (s) = 0.8881977
```

```
## MAA 200 m (s) = 0.7824067
## MAA 400 m (s) = 0.8587344
## MAA 800 m (min) = 0.8462426
## MAA 1500 m (min) = 0.7401403
## MAA 3000 m (min) = 0.7599955
## MAA Maratona (min) = 0.8773649
```

Todas as variáveis têm adequação amostral superior a 0,74.

3. Análise fatorial

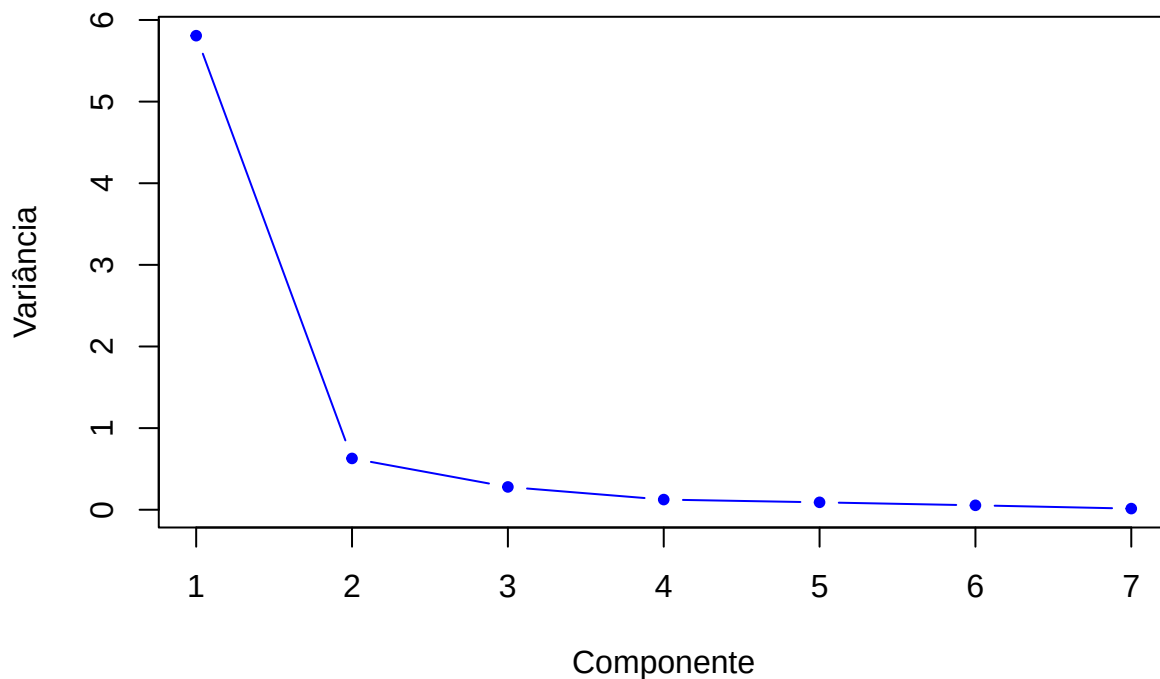
Os fatores serão obtidos (ou extraídos) aplicando o método dos componentes principais com a matriz de correlações amostral. Utilizamos a função `prcomp` com a matriz de correlações amostral (variáveis padronizadas usando `center` e `scale`).

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

```
## Importance of components:
##              PC1      PC2      PC3      PC4      PC5      PC6      PC7
## Standard deviation  2.4099 0.79290 0.5285 0.35292 0.3016 0.23349 0.11959
## Proportion of Variance 0.8297 0.08981 0.0399 0.01779 0.0130 0.00779 0.00204
## Cumulative Proportion 0.8297 0.91947 0.9594 0.97717 0.9902 0.99796 1.00000
```

O primeiro componente principal responde por cerca de 83% da variância total dos dados padronizados, ao passo que se tomarmos os dois primeiros componentes a proporção é cerca de 92% da variância total. O gráfico da figura abaixo indica que um ou dois componentes a reter é um número adequado.

```
plot(1:ncol(dados), acpcor$sdev^2, type = "b", xlab = "Componente", col = "blue",
     ylab = "Variância", pch = 20)
```



A análise fatorial será baseada em dois fatores ($m = 2$), com cargas fatoriais dadas abaixo.

```
m <- 2
# L^
```

```
carfat <- acpcor$rotation[, 1:m] %*% diag(acpcor$sdev[1:m])
colnames(carfat) <- paste("Fator", 1:m, sep = " ")
rownames(carfat) <- provas
carfat
```

```
##           Fator 1    Fator 2
## 100 m (s)    -0.9103780  0.3228503
## 200 m (s)    -0.9234990  0.3279673
## 400 m (s)    -0.8869307  0.3642220
## 800 m (min)  -0.9513832 -0.1278522
## 1500 m (min) -0.9380805 -0.2450762
## 3000 m (min) -0.9063506 -0.3355481
## Maratona (min) -0.8560043 -0.3086096
```

Exercício 1. Calcule as cargas fatoriais de outra forma, sem a função `diag`.

Com as estimativas das cargas fatoriais passamos à estimação das communalidades e das variâncias específicas, notando que o resultado de `diag(matcor)` é o vetor $(1, \dots, 1)^\top$ ($p \times 1$).

```
comum <- rowSums(carfat^2)
vespec <- diag(matcor) - comum
estimac <- cbind(comum, vespec, diag(matcor))
rownames(estimac) <- provas
colnames(estimac) <- c("Comunalidade", "Variância específica", "Variância")
estimac
```

```
##           Comunalidade Variância específica Variância
## 100 m (s)    0.9330205           0.06697954           1
## 200 m (s)    0.9604130           0.03958704           1
## 400 m (s)    0.9193037           0.08069628           1
## 800 m (min)  0.9214762           0.07852376           1
## 1500 m (min) 0.9400574           0.05994263           1
## 3000 m (min) 0.9340639           0.06593610           1
## Maratona (min) 0.8279832           0.17201676           1
```

As variâncias amostrais são iguais a 1 porque a análise foi baseada na matriz de correlações amostral. O menor valor de comunalidade é 0,83, de modo que temos indicações de um bom ajuste do modelo aos dados.

A matriz de resíduos do ajuste do modelo é dada por

$$\mathbf{E} = \mathbf{S} - (\hat{\mathbf{L}}\hat{\mathbf{L}}^\top + \hat{\mathbf{\Psi}}), \quad (4)$$

sendo que $\mathbf{S} = \mathbf{P}$ neste exemplo e notando que os elementos na diagonal principal de $\mathbf{E}$ são iguais a 0 (por quê?) e com a propriedade

$$\sum_{j=1}^p \sum_{k=1}^p E_{jk}^2 \leq \hat{\ell}_{m+1}^2 + \dots + \hat{\ell}_p^2. \quad (5)$$

```
resid <- matcor - (carfat %*% t(carfat) + diag(vespec))
```

Neste exemplo, o lado esquerdo da expressão em (5) vale 0,049 e no lado direito obtemos 0,564, correspondendo a cerca de 8% da variância total (no cálculo abaixo o denominador é igual a p porque usamos a matriz de correlações amostral). Temos indicações de um ajuste satisfatório do modelo.

```
sum(resid^2)
```

```
## [1] 0.04873268
```

```
sum(acpcor$sdev[(m + 1):p]^2)
```

```
## [1] 0.5636821
```

```
sum(acpcor$sdev[(m + 1):p]^2) / p * 100
```

```
## [1] 8.052602
```

3.1 Rotação dos fatores

Visando auxiliar na interpretação dos fatores, realizamos uma rotação $\hat{\mathbf{L}}^* = \hat{\mathbf{L}}\mathbf{B}$ pelo critério varimax (Kaiser, 1958, The varimax criterion for analytic rotation in factor analysis, *Psychometrika* 23, 187–200). Fazemos a mudança de escala

$$\tilde{\ell}_{jk}^* = \frac{\hat{\ell}_{jk}^*}{\hat{h}_j}, \quad j = 1, \dots, p, \quad k = 1, \dots, m, \quad (6)$$

notando que variáveis com estimativas de comunalidades mais baixas têm pesos maiores. O critério varimax calcula a transformação ortonormal $\mathbf{B}$ ($m \times m$) que maximiza

$$\frac{1}{p} \sum_{k=1}^m \left\{ \sum_{j=1}^p \tilde{\ell}_{jk}^{*4} - \frac{1}{p} \left(\sum_{j=1}^p \tilde{\ell}_{jk}^{*2} \right)^2 \right\}. \quad (7)$$

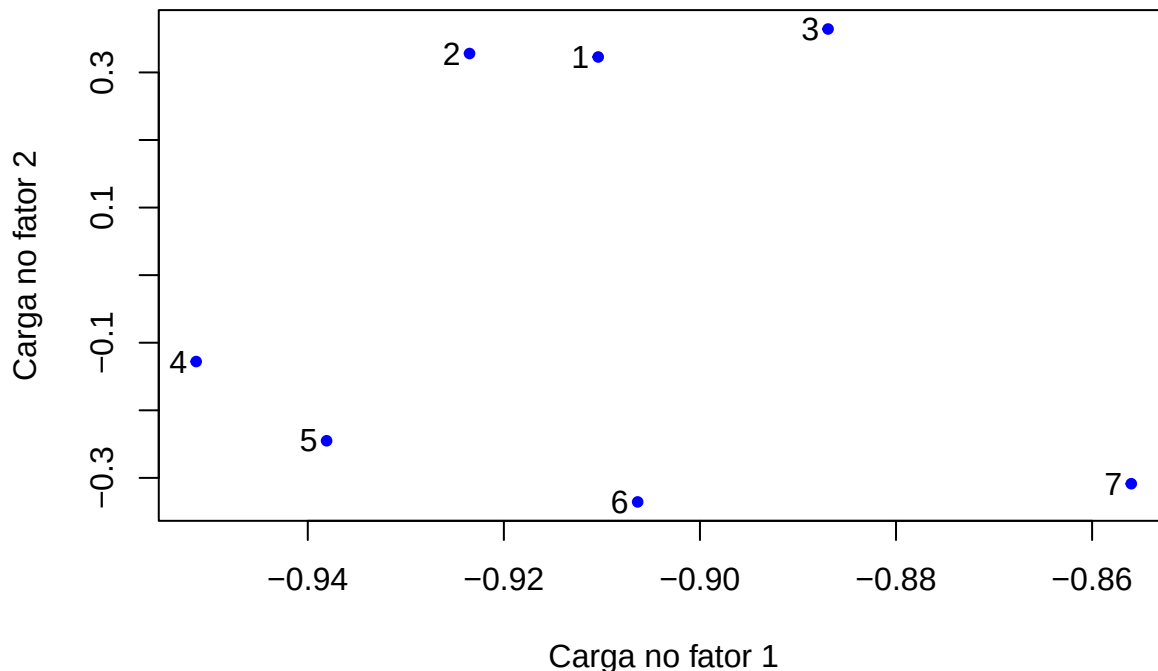
A expressão (7) é proporcional à soma das variâncias dos quadrados das cargas fatoriais (colunas de $\hat{\mathbf{L}}^*$) após mudança de escala (divisão por $\hat{h}_j$). Sendo assim, o critério varimax apresenta uma solução em que os quadrados das cargas dos fatores atingem dispersão máxima. Ao final, $\tilde{\ell}_{jk}^*$ é multiplicada por $\hat{h}_j$.

```
# Rotação varimax
```

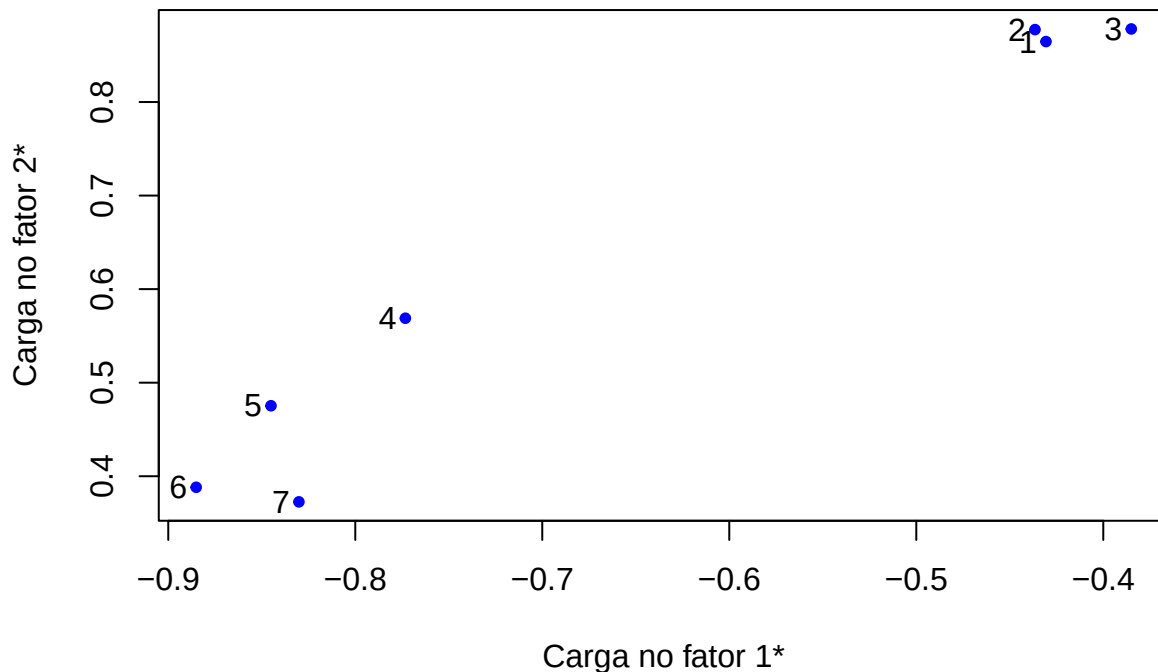
```
carfatr <- varimax(carfat)
```

Os gráficos da figura abaixo mostram as estimativas das cargas fatoriais das variáveis sem e com rotação pelo critério varimax, respectivamente.

```
plot(carfat, pch = 20, col = "blue", xlab = "Carga no fator 1",
      ylab = "Carga no fator 2")
text(carfat, as.character(1:p), adj = 1.5)
```



```
plot(carfatr$loadings, pch = 20, col = "blue", xlab = "Carga no fator 1*",
     ylab = "Carga no fator 2*")
text(carfatr$loadings, as.character(1:p), adj = 1.5)
```



Comparando os dois gráficos da figura acima, percebemos que após a rotação pelo critério varimax, (i) a variabilidade das cargas do fator F_1^* é maior do que no fator F_1 , (ii) a variabilidade das cargas do fator F_2^* é maior do que no fator F_2 e (iii) há uma separação mais nítida das variáveis em relação aos fatores. No gráfico logo acima (com rotação) as variáveis das provas de 800 m, 1500 m, 3000 m e Maratona (4 a 7) têm cargas mais altas (em valor absoluto) no fator F_1^* , ao passo que as variáveis das provas de 100 m, 200 m e 400 m têm cargas mais altas (em valor absoluto) no fator F_2^* .

Os fatores F_1^* e F_2^* podem ser denominados como desempenho em provas de distâncias mais longas e desempenho em provas de distâncias mais curtas, respectivamente.

As estimativas das cargas fatoriais com rotação varimax são mostradas abaixo.

```
# Com rotação varimax
print(carfatr$loadings)
```

```
##
## Loadings:
##          Fator 1 Fator 2
## 100 m (s)   -0.431  0.865
## 200 m (s)   -0.437  0.877
## 400 m (s)   -0.385  0.878
## 800 m (min) -0.773  0.569
## 1500 m (min) -0.845  0.475
## 3000 m (min) -0.885  0.388
## Maratona (min) -0.830  0.373
##
##          Fator 1 Fator 2
## SS loadings    3.309  3.128
## Proportion Var  0.473  0.447
```

```
## Cumulative Var    0.473    0.919
```

Nota 2. A rotação varimax preserva a variância acumulada dos dois primeiros fatores, embora as proporções individuais não se preservem. Compare o resultado acima com a página 5.

Para simplificar, após a rotação varimax, em cada variável poderíamos reter apenas a carga fatorial do fator com maior valor absoluto. Desta forma, teríamos a tabela abaixo. As cargas dos fatores devem ser padronizadas de modo que a norma da coluna j seja igual à norma da coluna j de $\hat{L}^*$ antes da simplificação, para $j = 1, \dots, m$.

Prova	Fator 1*	Fator 2*
100 m	0	0.865
200 m	0	0.877
400 m	0	0.878
800 m	-0.773	0
1500 m	-0.845	0
3000 m	-0.885	0
Maratona	-0.830	0

Exercício 2. Após a rotação varimax, as estimativas das communalidades (`rowSums(carfatr$loadings^2)`) coincidem com as estimativas já obtidas (vide o objeto `comum`). Por quê?

3.2 Escores dos fatores

Em seguida calculamos os escores dos dois fatores para cada país pelo método de regressão. Para que os escores dos fatores sejam compatíveis, multiplicamos por -1 as cargas do fator 2* de modo que nos dois fatores, quanto maior o valor do escore, mais alto o desempenho do país. No cálculo de `fchapeu` logo abaixo, a média amostral não é subtraída de `dadosz[i, ]` porque os dados foram padronizados ($\bar{z} = 0$).

```
dadosz <- t((t(dados) - xb) / desvp) # Padronização
P1 <- solve(matcor) # inversa da matriz de correlações
carfatr$loadings[, 2] <- -carfatr$loadings[, 2]
fchapeu <- matrix(NA, nrow = n, ncol = m)
for (i in 1:n) {
  fchapeu[i, ] <- t(carfatr$loadings) %*% P1 %*% dadosz[i, ]
}
```

Exercício 3. Os resultados abaixo surpreendem?

```
colMeans(fchapeu)
```

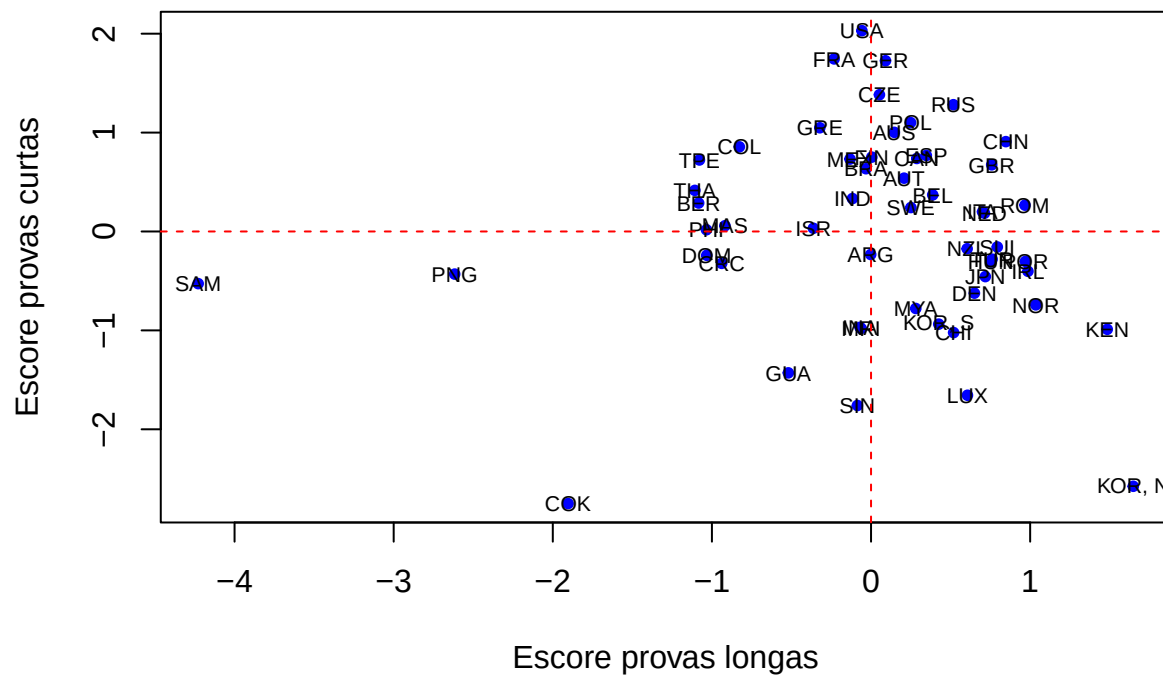
```
## [1] -4.641349e-16  2.220446e-16
```

```
cov(fchapeu)
```

```
##           [,1]           [,2]
## [1,] 1.000000e+00 3.258804e-15
## [2,] 3.258804e-15 1.000000e+00
```

O gráfico abaixo mostra os dois escores dos 54 países.

```
plot(fchapeu, pch = 20, col = "blue", xlab = "Escore provas longas",
     ylab = "Escore provas curtas")
text(fchapeu, países, cex = 0.7)
abline(h = 0, lty = 2, col = "red")
abline(v = 0, lty = 2, col = "red")
```



Exercício 4. Comente o gráfico acima.

Exercício 5. Calcule os escores dos fatores sem a rotação varimax e procure interpretar os resultados.

Exercício 6. Apresente uma solução com um modelo com um só fator.