Análise fatorial em R

1. 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 Statisitical Analysis*, sixth ed. Upper Saddle River, NJ: Pearson/Prentice Hall). Os dados encontram-se na página http://www.stat.wisc.edu/~rich/JWMULT06dat/T1-9.dat. Os dados são de 2005 e dizem respeito aos recordes femininos em 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), 3000 m (min) e maratona (min). O comando para leitura dos dados na linha abaixo deve ser completado. A coluna com os nomes dos países é excluída.

```
dados <- …
dados <- dados[, -1]
p <- ncol(dados)
```

Algumas medidas resumo são apresentadas em seguida. Por enquanto as variáveis são identificadas pelos códigos V2 (100 m) a V8 (maratona).

```
summary(dados)
```

```
       V2              V3              V4              V5       
 Min.   :10.49   Min.   :21.34   Min.   :47.60   Min.   :1.890  
 1st Qu.:11.12   1st Qu.:22.57   1st Qu.:49.97   1st Qu.:1.970  
 Median :11.32   Median :22.98   Median :51.65   Median :2.005  
 Mean   :11.36   Mean   :23.12   Mean   :51.99   Mean   :2.022  
 3rd Qu.:11.57   3rd Qu.:23.61   3rd Qu.:53.12   3rd Qu.:2.070  
 Max.   :12.52   Max.   :25.91   Max.   :61.65   Max.   :2.290  
       V6              V7              V8       
 Min.   :3.840   Min.   : 8.100   Min.   :135.2  
 1st Qu.:4.003   1st Qu.: 8.543   1st Qu.:143.5  
 Median :4.100   Median : 8.845   Median :148.4  
 Mean   :4.189   Mean   : 9.081   Mean   :153.6  
 3rd Qu.:4.338   3rd Qu.: 9.325   3rd Qu.:157.7  
 Max.   :5.420   Max.   :13.120   Max.   :221.1  
```

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

Inicialmente calculamos a matriz de correlações amostral.

```
matcor <- cor(dados)
print(matcor, digits = 2)
```

```
     V2   V3   V4   V5   V6   V7   V8
V2 1.00 0.94 0.87 0.81 0.78 0.73 0.67
V3 0.94 1.00 0.91 0.82 0.80 0.73 0.68
V4 0.87 0.91 1.00 0.81 0.72 0.67 0.68
V5 0.81 0.82 0.81 1.00 0.91 0.87 0.85
V6 0.78 0.80 0.72 0.91 1.00 0.97 0.79
V7 0.73 0.73 0.67 0.87 0.97 1.00 0.80
V8 0.67 0.68 0.68 0.85 0.79 0.80 1.00
```

Todas as correlações amostrais são positivas e variam de 0,67 a 0,97, correspondendo aos pares (V4, V7) e (V6, V7), respectivamente. Os gráficos de dispersão e as correlações amostrais estão representadas na Fig. 1.

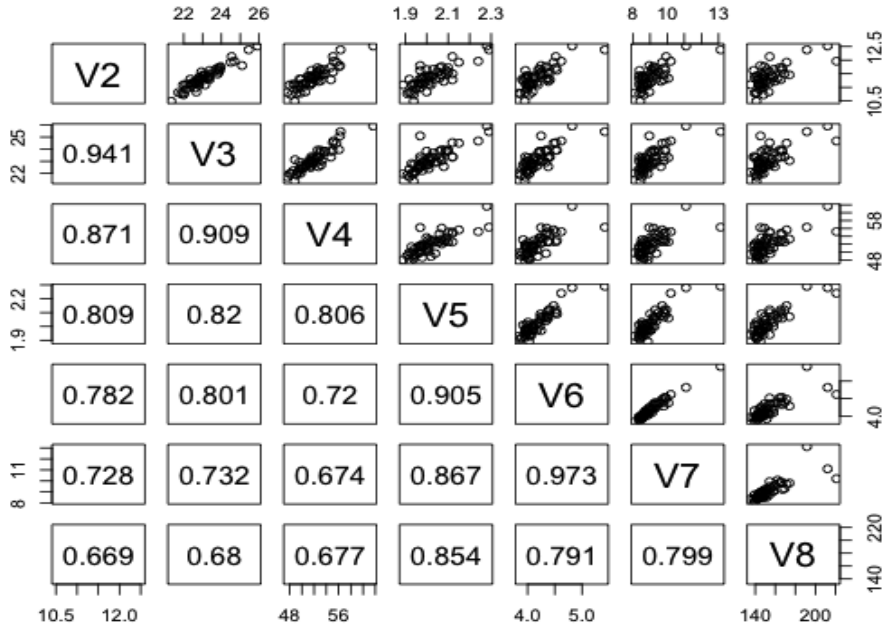


Figura 1. Matriz de gráficos de dispersão e correlações amostrais.

Uma matriz de correlações também pode ser representada em um corgrama, que está implementado em R na função `corrgram` do pacote corrgram. No comando abaixo o primeiro argumento é a matriz de correlações (poderia ser a matriz de dados $n \times p$). O segundo argumento informa que estamos usando uma matriz de correlações (`type = "cor"`).

```
library(corrgram)
corrgram(matcor, type = "cor", lower.panel = panel.shade, upper.panel = panel.pie)
```

Na Fig. 2 todas as correlações estão em cor azul porque são positivas, com tons mais fortes para as correlações mais altas. Para estas, o ângulo (sentido horário) no gráfico de setores do painel superior (`upper.panel = panel.pie`) é maior.

2. Medidas de adequação amostral

São utilizadas para avaliar se é adequado analisar os dados com a técnica de análise fatorial. Uma delas é o teste de esfericidade de Bartlett. Caso a hipótese nula não seja rejeitada, a técnica não é recomendada.

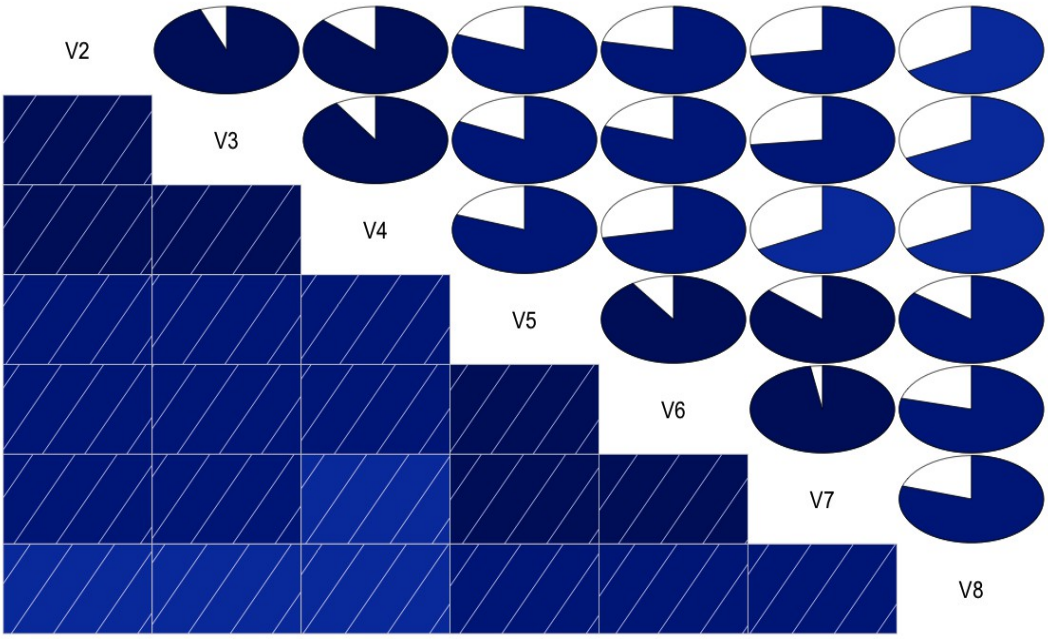


Figura 2. Corgrama das variáveis.

Outras medidas são baseadas em coeficientes de correlação entre os pares de variáveis X_j e X_m, $j \neq m$. Os resíduos da regressão de X_j como variável resposta e X_l, $l \neq j$, $l \neq m$, como variáveis explicativas são denotados por d_{ji}, $i = 1, ..., n$. Os resíduos da regressão de X_m como variável resposta e X_l, $l \neq j$, $l \neq m$, como variáveis explicativas são denotados por d_{mi}, $i = 1, ..., n$. O coeficiente de correlação parcial amostral entre X_j e X_m é definido como sendo o coeficiente de correlação linear entre os resíduos (d_{ji}, d_{mi}), $i = 1, ..., n$. É denotado por r_{pjm}, com $r_{pjj} = 0$, $j = 1, \ldots, p$. Neste cálculo as variáveis X_l, $l \neq j$, $l \neq m$, são mantidas fixas (dizemos que controlamos estas variáveis). Se houver uma associação linear expressiva entre as variáveis, que é desejável para uma aplicação bem sucedida da técnica de análise fatorial, $| r_{pjm} |$ deve ser pequeno comparado a $| r_{jm} |$, sendo que r_{jm} representa o coeficiente de correlação linear entre X_j e X_m. Fazendo $\boldsymbol{T} = \boldsymbol{R}^{-1}$, em que $\boldsymbol{R}$ é a matriz de correlações amostral, pode ser provado que

$$r_{pjm} = -\frac{t_{jm}}{\sqrt{t_{jj}\ t_{mm}}}.$$

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

$$\mathrm{KMO} = \frac{\sum_{j=1}^{p} \sum_{m=1, m\neq j}^{p} r_{jm}^2}{\sum_{j=1}^{p} \sum_{m=1, m\neq j}^{p} r_{jm}^2 + \sum_{j=1}^{p} \sum_{m=1, m\neq j}^{p} r_{pjm}^2}.$$

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

$$\text{MAA}_j = \frac{\sum_{l=1, l\neq j}^{p} r_{jl}^2}{\sum_{l=1, l\neq j}^{p} r_{jl}^2 + \sum_{l=1, l\neq j}^{p} r_{\text{p}jl}^2}, \quad j = 1, \ldots, p.$$

A Tabela 1 apresenta uma síntese (sugestão) da adequação amostral (em tradução livre).

Tabela 1. Adequação amostral segundo a medida KMO.

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 `partial.cor` do pacote Rcmdr.

```
partial.cor <- function (X, ...)
{
   R <- cor(X, ...)
   RI <- solve(R)
   D <- 1/sqrt(diag(RI))
   Rp <- -RI * (D %o% D)
   diag(Rp) <- 0
   rownames(Rp) <- colnames(Rp) <- colnames(X)
   Rp
}
matcorp <- partial.cor(dados)
```

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 aceitável (> 0,5) e meritória. As medidas MAA são calculadas para cada variável.

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

```
 MAA 1 = 0.8881977  MAA 5 = 0.7401403
 MAA 2 = 0.7824067  MAA 6 = 0.7599955
 MAA 3 = 0.8587344  MAA 7 = 0.8773649
 MAA 4 = 0.8462426
```

Todas as variáveis têm adequação superior a 0,7.

3. Análise fatorial

Os fatores serão obtidos (ou extraídos) aplicando o método dos componentes principais com a matriz de correlações amostral.

```
acpcor <- prcomp(dados, 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 Fig. 3 indica que dois componentes a reter é um número adequado.

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

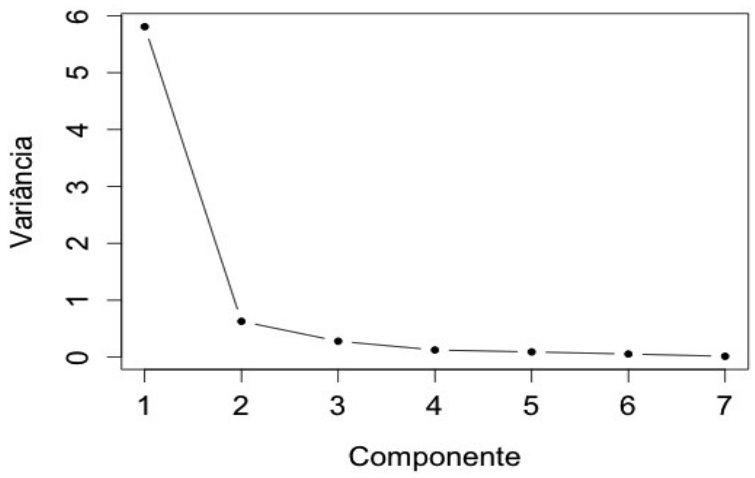


Figura 3. Gráfico de escarpa.

A análise fatorial será baseada em dois fatores, com cargas fatoriais dadas abaixo.

```
k <- 2
carfat <- acpcor$rotation[, 1:k] %*% diag(acpcor$sdev[1:k])
colnames(carfat) <- paste("Fator", 1:k, sep = " ")
```

```
      Fator 1    Fator 2
V2 -0.9103780  0.3228503
V3 -0.9234990  0.3279673
V4 -0.8869307  0.3642220
V5 -0.9513832 -0.1278522
V6 -0.9380805 -0.2450762
V7 -0.9063506 -0.3355481
V8 -0.8560043 -0.3086096
```

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

Com as cargas fatoriais passamos à estimação das comunalidades e das variâncias específicas.

```
comum <- rowSums(carfat^2)
vespec <- diag(matcor) - comum
estimat <- cbind(comum, vespec, diag(matcor))
rownames(estimat) <- colnames(dados)
colnames(estimat) <- c("Comunalidade", "Variância única", "Variância")
```

```
   Comunalidade Variância única Variância
V2    0.9330205      0.06697954         1
V3    0.9604130      0.03958704         1
V4    0.9193037      0.08069628         1
V5    0.9214762      0.07852376         1
V6    0.9400574      0.05994263         1
V7    0.9340639      0.06593610         1
V8    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

$$\boldsymbol{E} = \boldsymbol{R} - (\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\Lambda}}' + \widehat{\boldsymbol{\Psi}}),$$

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

com elementos na diagonal principal iguais a 0 (por quê?) e com a propriedade

$$\sum_{j=1}^{p}\sum_{m=1}^{p} E_{jm}^2 \leq \widehat{\lambda}_{k+1}^2 + \cdots + \widehat{\lambda}_p^2.$$

Neste exemplo, o lado esquerdo da expressão acima vale 0,049 (`sum(resid^2)`) e no lado direito obtemos 0,564 (`sum(acpcor$sdev[(k + 1):p]^2)`), correspondendo a cerca de 8% da variância total (por quê?). Temos indicações de um ajuste satisfatório do modelo.

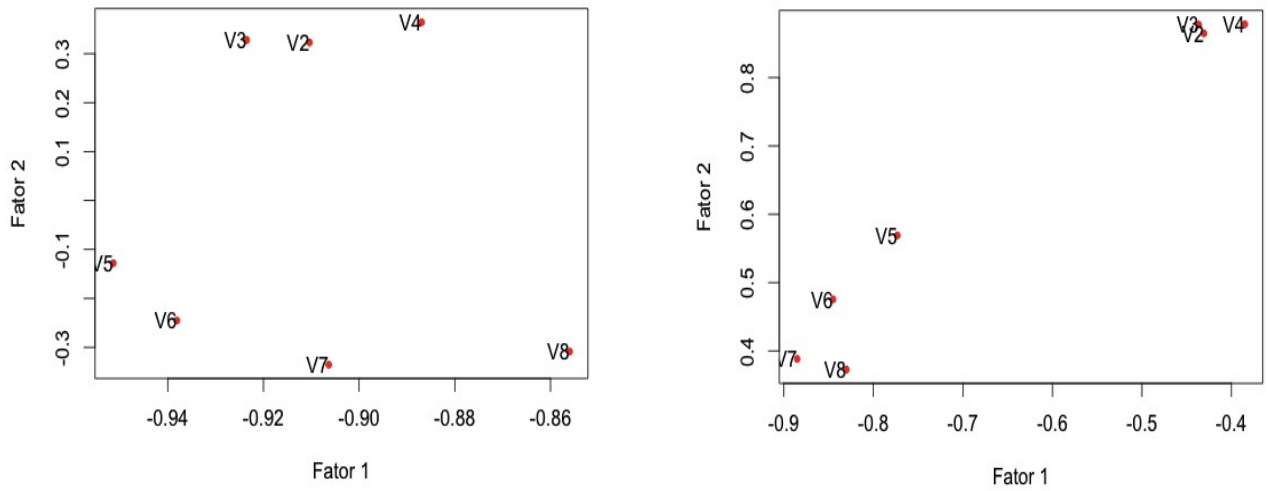
Visando auxiliar na interpretação dos fatores, realizamos uma rotação pelo método varimax. A função `varimax` encontra-se no pacote stats.

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

Os gráficos da Fig. 4 mostram as estimativas das cargas fatoriais das variáveis sem e com rotação pelo método varimax, respectivamente.

```
plot(carfat, pch = 20, col = "red", xlab = "Fator 1", ylab = "Fator 2")
text(carfat, rownames(carfat), adj = 1)

plot(carfatr$loadings, pch = 20, col = "red", xlab = "Fator 1", ylab = "Fator 2")
text(carfatr$loadings, rownames(carfat), adj = 1)
```



(a) (b)

Figura 4. Cargas dos fatores 1 e 2 antes (a) e após rotação pelo método varimax (b).

Comparando os gráficos da Fig. 4, percebemos que após a rotação pelo método varimax há uma separação mais nítida das variáveis em relação aos fatores. Na Fig. 4(b) as variáveis V5, ..., V8 têm cargas mais altas (em valor absoluto) no fator 1, ao passo que V2, V3 e V4 têm cargas mais altas no fator 2. Consultando a descrição das variáveis, os fatores podem ser denominados como desempenho em provas de distâncias mais curtas e desempenho em provas de distâncias mais longas.

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

Nota 3. Obtenha uma solução com o método de rotação promax (função `promax`) e compare com a solução aqui apresentada.

Nota 4. A função `factanal` do pacote stats em R permite realizar uma análise fatorial a partir de uma matriz de dados $n \times p$ ou a partir de uma matriz de covariâncias (ou de correlações) $p \times p$. O método de estimação é o de máxima verossimilhança aplicado a uma distribuição normal multivariada.

Nota 5. Procure reproduzir os resultados utilizando outros pacotes estatísticos (por exemplo, SAS, Minitab, SPSS e Statistica).