Lista 3 - MAE0330

Guilherme NºUSP: 8943160 e Leonardo NºUSP: 9793436

Exercício 1

Considere a seguinte matriz de correlação

$$\rho = \begin{pmatrix} 1,000 & -0,488 & 0,150 \\ -0,488 & 1,000 & -0,130 \\ 0,150 & -0,130 & 1,000 \end{pmatrix}$$

de três variáveis padronizadas Z_1, Z_2 e Z_3.

(a) Mostre que ρ pode ser decomposta segundo um modelo fatorial com $m = 1$ fator dado por

$$Z_1 = 0,75F_1 + \epsilon_1,$$

$$Z_2 = -0,65F_1 + \epsilon_2,$$

e

$$Z_3 = 0,20F_1 + \epsilon_3$$

com $Var(F_1) = 1$ e $Cov(\epsilon_j, F_1) = 0$, $j = 1, 2, 3$. Obtenha a matriz Ψ com as variâncias específicas.

Resolução

Como sabemos que a matrix ρ pode ser ser decomposta em $\rho = LL^T + \Psi$ onde $L = \begin{pmatrix} 0.75 \\ -0.65 \\ 0.20 \end{pmatrix}$, assim $\Psi = \rho - LL^T$ então:

$$\Psi = \begin{pmatrix} 1,000 & -0,488 & 0,150 \\ -0,488 & 1,000 & -0,130 \\ 0,150 & -0,130 & 1,000 \end{pmatrix} - \begin{pmatrix} 0.75 \\ -0.65 \\ 0.20 \end{pmatrix} \begin{pmatrix} 0.75 & -0.65 & 0.20 \end{pmatrix} =$$

$$\begin{pmatrix} 1,000 & -0,488 & 0,150 \\ -0,488 & 1,000 & -0,130 \\ 0,150 & -0,130 & 1,000 \end{pmatrix} - \begin{pmatrix} 0,562 & 0,488 & 0,150 \\ 0,488 & 0,423 & 0,130 \\ 0,150 & 0,130 & 0,040 \end{pmatrix} \Rightarrow \Psi = \begin{pmatrix} 0,438 & 0 & 0 \\ 0 & 0,577 & 0 \\ 0 & 0 & 0,960 \end{pmatrix}$$

(b) Obtenha as comunalidades e interprete-as.

Resolução

Comulalidade de $Z_1 : C_1^2 = 0.75^2 = 0.562$

Comulalidade de $Z_2 : C_2^2 = (-0.65)^2 = 0.422$

Comulalidade de $Z_3 : C_3^2 = 0.2^2 = 0.04$

Onde cada comulalidade representa a proporção de variabilidade da variável Z_j ($j = 1, 2, 3$) explicada pelo fator. Logo o fator explica, principalmente, a variabilidade das duas primeiras variáveis.

(c) Calcule a correlação entre Z_j e F_1 , $j = 1, 2, 3$. Discuta.

Resolução

Sabe-se que $corr(Z_j, F_1) = l_{1j}$, j=1,2,3, logo:

$Corr(Z_1, F_1) = 0.75$

$Corr(Z_2, F_1) = -0.65$

$Corr(Z_3, F_1) = 0.20$

Utilizando o resultado do item (b), o fator explica principalmente Z_1 e Z_2, nota-se então que as correlações dessas variáveis com o fator possuem sinais opostos, logo conclui-se que o fator explica a proporcionalidade inversa dessas variáveis.

Exercício 2

Considere ainda a matriz de correlação do exercício anterior.

(a) Obtenha os autovalores e autovetores correspondentes.

Resolução

Table 1: Autovalores

λ_1	λ_2	λ_3
1.558	0.93	0.512

Table 2: Autovetores

e_1	e_2	e_3
0.670	-0.214	-0.711
-0.663	0.259	-0.702
0.334	0.942	0.032

(b) Considerando um modelo fatorial com $m = 1$ fator, obtenha a matriz de cargas fatoriais $\mathbf{L}$ e a matriz de variâncias específicas Ψ usando o método das componentes principais. Compare com os resultados do exercício anterior.

Resolução

Considerando um modelo fatorial com $m = 1$ fator, a matriz de cargas fatoriais $\mathbf{L}$ pode ser obtida pelo método de componentes principais $\hat{L} = \sqrt{\lambda_1} * e_1$, em que λ_1 é o primeiro autovalor e e_1 é o primeiro autovetor, assim

$$\hat{L} = \sqrt{(1.558)} \begin{pmatrix} 0.670 \\ -0.663 \\ 0.334 \end{pmatrix} \Rightarrow \hat{L} = \begin{pmatrix} 0.836 \\ -0.827 \\ 0.417 \end{pmatrix}$$

Além disso a matriz de variâncias específicas Ψ pode ser estimada por:

Como sabemos que a matrix ρ pode ser ser decomposta em $\rho = \hat{L}\hat{L^T} + \hat{\Psi}$ onde $\hat{L} = \begin{pmatrix} 0.836 \\ -0.827 \\ 0.417 \end{pmatrix}$, assim $\hat{\Psi} = \rho - \hat{L}\hat{L^T}$ então:

$$\hat{\Psi} = \begin{pmatrix} 0.20 & 0 & 0 \\ 0 & 0.09 & 0 \\ 0 & 0 & 0.07 \end{pmatrix}$$

Onde podemos notar que os valores de $\hat{L}$ são maiores que o L dado no exercício anterior, o que resulta em uma maior variabilidade explicada pelo fator e uma correlação mais forte entre as variáveis e o fator. E pela matriz de variâncias específicas $\hat{\Psi}$ possui valores menores que a matriz Ψ dada no exercício anterior, representando que a variância não explicada pelo fator diminuiu.

(c) Obtenha a proporção da variabilidade total dos dados explicada pelo fator.

Resolução

A proporção da variabilidade total dos dados explicada pelo fator 1 é: $\frac{\lambda_1}{p} = \frac{1.558}{3} = 0.52 = 52\%$

Em que λ_1 é o primeiro autovalor e p é o número de variáveis originais no caso são 3.

Exercício 3

As cargas fatoriais associadas a 6 variáveis padronizadas e as cargas fatoriais rotacionadas (varimax) estão apresentadas a seguir:

	Fatores		Fatores Rotac.	
Variáveis	F_1	F_2	F_1	F_2
Y_1	0.602	0.2	0.484	0.411
Y_2	0.467	0.154	0.375	0.319
Y_3	0.926	0.143	0.603	0.717
Y_4	1	0	0.519	0.855
Y_5	0.874	0.476	0.861	0.499
Y_6	0.894	0.327	0.744	0.594

(a) Obtenha as comunalidades e as variâncias específicas para as cargas fatoriais sem e com rotação.

Resolução

Considerando primeiramente as cargas fatoriais sem a rotação temos:

Comulalidade de $Y_1 : C_1^2 = 0.602^2 + 0.2^2 = 0.403$

Comulalidade de $Y_2 : C_2^2 = 0.467^2 + 0.154^2 = 0.242$

Comulalidade de $Y_3 : C_3^2 = 0.926^2 + 0.143^2 = 0.878$

Comulalidade de $Y_4 : C_4^2 = 1^2 + 0^2 = 1$

Comulalidade de $Y_5 : C_5^2 = 0.874^2 + 0.476^2 = 0.99$

Comulalidade de $Y_6 : C_6^2 = 0.894^2 + 0.327^2 = 0.906$

Agora as variâncias especifícas, por se tratar de variáveis padronizadas, temos:

$\Psi_1 = 1 - C_1^2 = 1 - 0.403 = 0.597$

$\Psi_2 = 1 - C_2^2 = 1 - 0.242 = 0.758$

$\Psi_3 = 1 - C_3^2 = 1 - 0.878 = 0.122$

$\Psi_4 = 1 - C_4^2 = 1 - 1 = 0$

$\Psi_5 = 1 - C_5^2 = 1 - 0.99 = 0.01$

$\Psi_6 = 1 - C_6^2 = 1 - 0.906 = 0.094$

Agora com as cargas fatoriais com a rotação temos:

Comulalidade de $Y_1 : C_1^2 = 0.484^2 + 0.411^2 = 0.403$

Comulalidade de $Y_2 : C_2^2 = 0.375^2 + 0.319^2 = 0.242$

Comulalidade de $Y_3 : C_3^2 = 0.603^2 + 0.717^2 = 0.878$

Comulalidade de $Y_4 : C_4^2 = 0.519^2 + 0.855^2 = 1$

Comulalidade de $Y_5 : C_5^2 = 0.861^2 + 0.499^2 = 0.99$

Comulalidade de $Y_6 : C_6^2 = 0.744^2 + 0.594^2 = 0.906$

Agora as variâncias especifícas, analogamente ao caso sem rotação, temos:

$\Psi_1 = 1 - C_1^2 = 1 - 0.403 = 0.597$

$\Psi_2 = 1 - C_2^2 = 1 - 0.242 = 0.758$

$\Psi_3 = 1 - C_3^2 = 1 - 0.878 = 0.122$

$\Psi_4 = 1 - C_4^2 = 1 - 1 = 0$

$\Psi_5 = 1 - C_5^2 = 1 - 0.99 = 0.01$

$\Psi_6 = 1 - C_6^2 = 1 - 0.906 = 0.094$

Nota-se que as comunalidades e as variâncias específicas são iguais para as cargas fatoriais sem e com rotação.

(b) Qual é proporção da variância total dos dados explicada por cada fator?

Resolução

A proporção de variância total por cada fator (caso não rotacionado) é dado pela tabela abaixo:

Variável	l_{i1}^2	l_{i2}^2
Y_1	0.36	0.04
Y_2	0.22	0.02
Y_3	0.86	0.02
Y_4	1	0
Y_5	0.76	0.23
Y_6	0.80	0.11
Y_j	4	0.42
%	40	4.2

A proporção de variância explicada por cada fator (caso rotacionado) é dado pela tabela abaixo:

Variável	l_{i1}^2	l_{i2}^2
Y_1	0.23	0.17
Y_2	0.14	0.1
Y_3	0.36	0.51
Y_4	0.27	0.73
Y_5	0.74	0.25
Y_6	0.55	0.35
Y_j	2.29	2.11
%	22.9	21.1

Nota-se que com a rotação a proporção da variância total dos dados explicada por cada fator é maior que sem a rotação.

(c) Para uma observação com valores observados das variáveis originais (já padronizados) iguais a (0.8, -0.2, 1.3, -0.6, 1.5, -0.7), obtenha os escores fatoriais utilizando os fatores rotacionados.

Resolução

Os escores fatoriais utilizando os fatores rotacionados são:

$\hat{F}_1 = 0.8 * 0.484 + (-0.2) * 0.375 + 1.3 * 0.603 + (-0.6) * 0.519 + 1.5 * 0.861 + (-0.7) * 0.744 = 1.5554$
$\hat{F}_2 = 0.8 * 0.411 + (-0.2) * 0.319 + 1.3 * 0.717 + (-0.6) * 0.855 + 1.5 * 0.499 + (-0.7) * 0.594 = 1.0168$

Exercício 4

Os dados no arquivo **T1-9.dat** são referentes a recordes nacionais femininos de corrida para diversos países. As colunas são referentes aos tempos recordes nas seguintes modalidades, res- pectivamente:

- 100 m (segundos);
- 200 m (segundos);
- 400 m (segundos);
- 800 m (minutos);
- 1500 m (minutos);
- 5000 m (minutos);
- 10.000 m (minutos);
- Maratona (minutos).

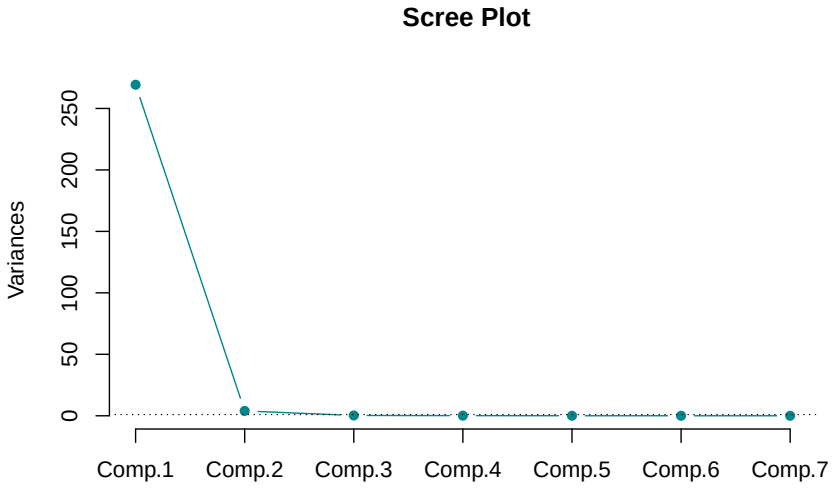
(a) Faça uma análise fatorial com a matriz de covariância dos dados.

Resolução

Fazendo a análise fatorial com a matriz de covariância dos dados, pelo método das componentes principais e sem rotação, temos:

Table 3: Matriz de Covariâncias

	100m	200m	400m	800m	1500m	3000m	Maratona
100m	0.155	0.345	0.891	0.028	0.084	0.234	4.334
200m	0.345	0.863	2.193	0.066	0.203	0.554	10.385
400m	0.891	2.193	6.745	0.182	0.509	1.427	28.904
800m	0.028	0.066	0.182	0.008	0.021	0.061	1.220
1500m	0.084	0.203	0.509	0.021	0.074	0.216	3.540
3000m	0.234	0.554	1.427	0.061	0.216	0.665	10.706
Maratona	4.334	10.385	28.904	1.220	3.540	10.706	270.270



Segundo o critério de Kaiser que considera que o número de fatores tem autovalores maiores que 1, assim observando o Scree plot acima, dois Fatores é suficiente para explicar a maior parte da variavilidade dos dados.

```
## Principal Components Analysis
## Call: principal(r = Cov, nfactors = 2, rotate = "none", covar = T)
## Unstandardized loadings (pattern matrix) based upon covariance matrix
##             PC1   PC2      h2      u2   H2      U2
## 100m       0.27  0.23 1.2e-01 0.03072 0.80 2.0e-01
## 200m       0.64  0.58 7.5e-01 0.11414 0.87 1.3e-01
## 400m       1.79  1.88 6.7e+00 0.02015 1.00 3.0e-03
## 800m       0.07  0.03 6.3e-03 0.00171 0.79 2.1e-01
## 1500m      0.22  0.07 5.2e-02 0.02182 0.71 2.9e-01
## 3000m      0.65  0.16 4.5e-01 0.21236 0.68 3.2e-01
## Maratona 16.44 -0.24 2.7e+02 0.00026 1.00 9.5e-07
##
##                        PC1  PC2
## SS loadings           274.36 4.02
## Proportion Var          0.98 0.01
## Cumulative Var          0.98 1.00
## Proportion Explained    0.99 0.01
## Cumulative Proportion   0.99 1.00
```

```
##
##  Standardized loadings (pattern matrix)
##          item  PC1   PC2   h2      u2
## 100m        1 0.68  0.58 0.80 2.0e-01
## 200m        2 0.69  0.63 0.87 1.3e-01
## 400m        3 0.69  0.72 1.00 3.0e-03
## 800m        4 0.83  0.30 0.79 2.1e-01
## 1500m       5 0.80  0.27 0.71 2.9e-01
## 3000m       6 0.80  0.19 0.68 3.2e-01
## Maratona    7 1.00 -0.01 1.00 9.5e-07
##
##                  PC1  PC2
## SS loadings     4.38 1.46
## Proportion Var  0.63 0.21
## Cumulative Var  0.63 0.83
## Cum. factor Var 0.75 1.00
##
## Mean item complexity =  1.5
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.02
##
## Fit based upon off diagonal values = 1
```

Onde podemos notar que pela variabilidade total dos dados, o primeiro fator representa uma variabilidade de 274.36 enquanto o segundo fator, apenas 4.02.

Sendo as comunalidades dadas por:

Table 4: Comunalidades

100m	200m	400m	800m	1500m	3000m	Maratona
0.124	0.749	6.725	0.006	0.052	0.453	270.27

Onde cada comulalidade representa a proporção de variabilidade de X_j $(j = 1, ..., 7)$ explicada pelos fatores, em que a variável Maratona se destaca, obtendo uma grande parcela da variância explicada pelos fatores.

E as variâncias específicas:

Table 5: Variâncias especifícas

100m	200m	400m	800m	1500m	3000m	Maratona
0.031	0.114	0.02	0.002	0.022	0.212	0

Onde cada variância específica representa a parcela da variância não explicada pelos fatores, tendo destaque para a Maratona que teve toda a sua variância explicada.

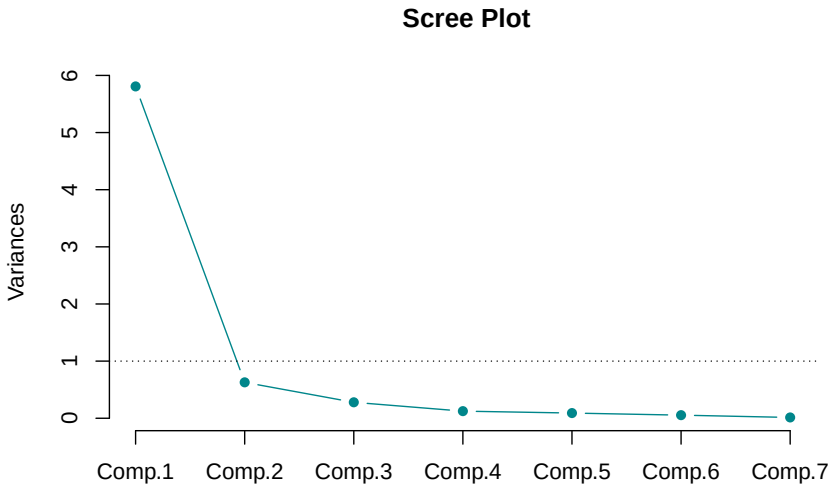
(b) Faça uma análise fatorial com a matriz de correlação dos dados.

Resolução

Fazendo a análise fatorial com a matriz de correlação dos dados, pelo método das componentes principais e sem rotação, temos:

Table 6: Matriz de Correlações

	100m	200m	400m	800m	1500m	3000m	Maratona
100m	1.000	0.941	0.871	0.809	0.782	0.728	0.669
200m	0.941	1.000	0.909	0.820	0.801	0.732	0.680
400m	0.871	0.909	1.000	0.806	0.720	0.674	0.677
800m	0.809	0.820	0.806	1.000	0.905	0.867	0.854
1500m	0.782	0.801	0.720	0.905	1.000	0.973	0.791
3000m	0.728	0.732	0.674	0.867	0.973	1.000	0.799
Maratona	0.669	0.680	0.677	0.854	0.791	0.799	1.000



Segundo o critério de Kaiser que considera que o número de fatores tem autovalores maiores que 1, assim observando o Scree plot acima, um fator é suficiente para explicar a maior parte da variavilidade dos dados.

```
## Principal Components Analysis
## Call: principal(r = Cor, nfactors = 1, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##           PC1   h2    u2 com
## 100m     0.91 0.83 0.171   1
## 200m     0.92 0.85 0.147   1
## 400m     0.89 0.79 0.213   1
## 800m     0.95 0.91 0.095   1
## 1500m    0.94 0.88 0.120   1
## 3000m    0.91 0.82 0.178   1
## Maratona 0.86 0.73 0.267   1
##
##                 PC1
```

```
## SS loadings    5.81
## Proportion Var 0.83
##
## Mean item complexity =  1
## Test of the hypothesis that 1 component is sufficient.
##
## The root mean square of the residuals (RMSR) is  0.08
##
## Fit based upon off diagonal values = 0.99
```

Onde podemos notar que pela variabilidade total dos dados, o fator representa uma variabilidade de 5.81.

Sendo as comunalidades dadas por:

Table 7: Comunalidades

100m	200m	400m	800m	1500m	3000m	Maratona
0.829	0.853	0.787	0.905	0.88	0.822	0.733

Onde cada comulalidade representa a proporção de variabilidade de X_j $(j = 1, ..., 7)$ explicada pelos fatores, em que todas as variáveis possuem proporção entre 0.7 e 1.

E as variâncias específicas:

Table 8: Variâncias especifícas

100m	200m	400m	800m	1500m	3000m	Maratona
0.171	0.147	0.213	0.095	0.12	0.178	0.267

Onde cada variância específica representa a parcela da variância não explicada pelos fatores, em que todas as variáveis possuem valores entre 0 e 0.3.

Exercício 5

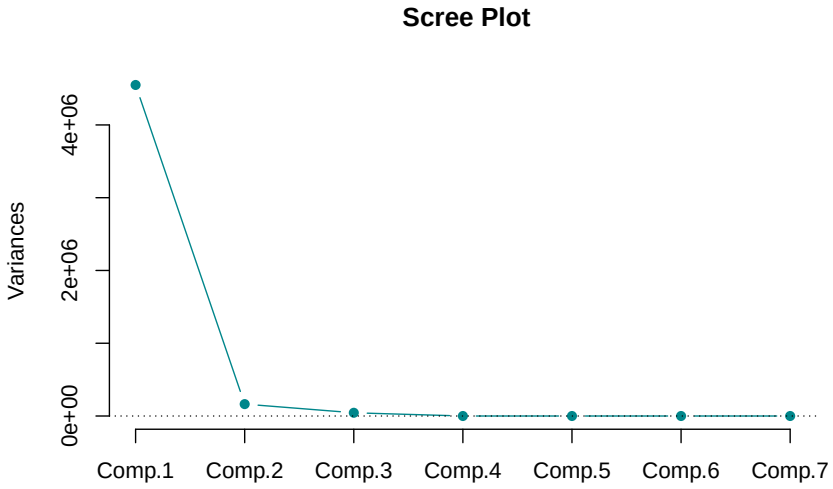
Ainda com os dados do arquivo **T1-9.dat**, transforme os tempos recordes em velocidades (na unidade metros por segundo). A maratona corresponde a um percurso de 42.195 metros (ou 26,2 milhas). Faça análise fatorial com a matriz de covariância dos dados e com a matriz de correlação. Discuta os resultados.

Resolução

Fazendo a transformação dos tempos recordes em velocidades (m/s) então a análise fatorial com a matriz de covariâncias dos dados transformados pelo método das componentes principais e sem rotação, temos:

Table 9: Matriz de Covariâncias

	100m	200m	400m	800m	1500m	3000m	Maratona
100m	0.091	0.096	0.097	234.348	296.135	331.507	0.291
200m	0.096	0.115	0.114	269.823	345.748	379.740	0.336
400m	0.097	0.114	0.138	291.445	344.010	390.358	0.367
800m	234.348	269.823	291.445	952856.488	1120446.569	1292820.821	1222.148
1500m	296.135	345.748	344.010	1120446.569	1604971.690	1862543.296	1535.443
3000m	331.507	379.740	390.358	1292820.821	1862543.296	2288532.851	1899.923
Maratona	0.291	0.336	0.367	1222.148	1535.443	1899.923	2.162



Segundo o critério de Kaiser que considera que o número de fatores tem autovalores maiores que 1, assim observando o Scree plot acima, três fatores é suficiente para explicar a maior parte da variavilidade dos dados.

```
## Principal Components Analysis
## Call: principal(r = Cov, nfactors = 3, rotate = "none", covar = T)
## Unstandardized loadings (pattern matrix) based upon covariance matrix
##              PC1     PC2     PC3      h2      u2   H2      U2
## 100m        0.23    0.07    0.03 6.0e-02 3.1e-02 0.66 3.4e-01
## 200m        0.27    0.08    0.06 8.2e-02 3.3e-02 0.72 2.8e-01
## 400m        0.28    0.12    0.00 8.9e-02 4.9e-02 0.65 3.5e-01
## 800m      907.05  357.27  -49.78 9.5e+05 3.8e-07 1.00 4.0e-13
## 1500m    1254.89  -25.86  171.92 1.6e+06 3.1e-07 1.00 1.9e-13
## 3000m    1495.84 -194.95 -114.04 2.3e+06 4.3e-07 1.00 1.9e-13
## Maratona    1.27    0.16   -0.30 1.7e+00 4.4e-01 0.80 2.0e-01
##
##                            PC1       PC2      PC3
```

```
## SS loadings           4635006.64 166317.02 45039.32
## Proportion Var              0.96      0.03     0.01
## Cumulative Var              0.96      0.99     1.00
## Proportion Explained        0.96      0.03     0.01
## Cumulative Proportion       0.96      0.99     1.00
##
##  Standardized loadings (pattern matrix)
##          item  PC1   PC2   PC3   h2      u2
## 100m        1 0.77  0.23  0.11 0.66 3.4e-01
## 200m        2 0.79  0.24  0.18 0.72 2.8e-01
## 400m        3 0.74  0.31  0.01 0.65 3.5e-01
## 800m        4 0.93  0.37 -0.05 1.00 4.0e-13
## 1500m       5 0.99 -0.02  0.14 1.00 1.9e-13
## 3000m       6 0.99 -0.13 -0.08 1.00 1.9e-13
## Maratona    7 0.86  0.11 -0.20 0.80 2.0e-01
##
##                  PC1  PC2  PC3
## SS loadings     5.34 0.37 0.11
## Proportion Var  0.76 0.05 0.02
## Cumulative Var  0.76 0.82 0.83
## Cum. factor Var 0.92 0.98 1.00
##
## Mean item complexity =  1.2
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.01
##
## Fit based upon off diagonal values = 1
```

Onde podemos notar que com os 3 fatores a variabilidade total proporcional é de 100%.

Sendo as comunalidades dadas por:

Table 10: Comunalidades

100m	200m	400m	800m	1500m	3000m	Maratona
0.06	0.082	0.089	952856.5	1604972	2288533	1.724

Onde cada comulalidade representa a proporção de variabilidade de X_j $(j = 1, ..., 7)$ explicada pelos fatores, em que as variáveis de 800m, 1500m e 3000m possuem valores muito maiores que os demais.

E as variâncias específicas:

Table 11: Variâncias especifícas

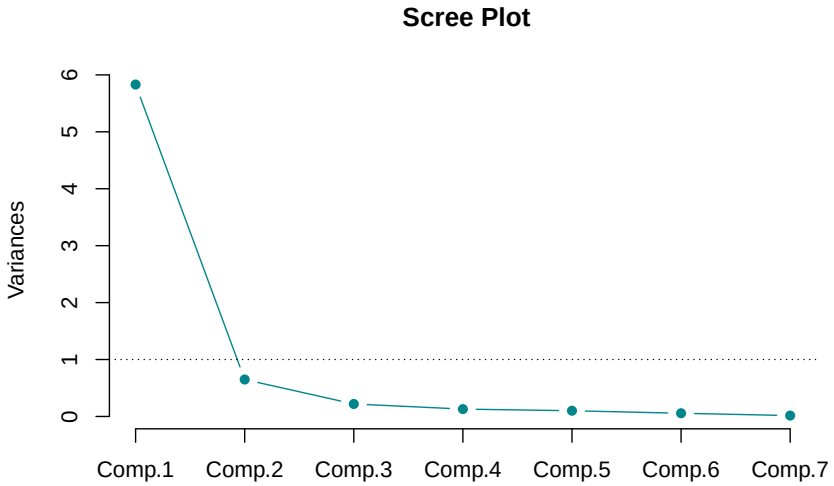
100m	200m	400m	800m	1500m	3000m	Maratona
0.0309	0.0325	0.0485	0	0	0	0.4383

Onde cada variância específica representa a parcela da variância não explicada pelos fatores, em que todas as variáveis possuem valores entre menores que 0.5, então a variabilidade foi quase toda explicada pelos fatores.

Fazendo o mesmo com a matriz de correlações, temos:

Table 12: Matriz de Correlações

	100m	200m	400m	800m	1500m	3000m	Maratona
100m	1.000	0.938	0.865	0.797	0.776	0.728	0.658
200m	0.938	1.000	0.905	0.816	0.805	0.741	0.675
400m	0.865	0.905	1.000	0.804	0.731	0.695	0.671
800m	0.797	0.816	0.804	1.000	0.906	0.875	0.852
1500m	0.776	0.805	0.731	0.906	1.000	0.972	0.824
3000m	0.728	0.741	0.695	0.875	0.972	1.000	0.854
Maratona	0.658	0.675	0.671	0.852	0.824	0.854	1.000



Segundo o critério de Kaiser que considera que o número de fatores tem autovalores maiores que 1, assim observando o Scree plot acima, um fator é suficiente para explicar a maior parte da variavilidade dos dados.

```
## Principal Components Analysis
## Call: principal(r = Cor, nfactors = 1, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##           PC1   h2   u2 com
## 100m     0.90 0.81 0.19   1
## 200m     0.92 0.85 0.15   1
## 400m     0.89 0.79 0.21   1
## 800m     0.95 0.90 0.10   1
## 1500m    0.94 0.89 0.11   1
## 3000m    0.92 0.85 0.15   1
## Maratona 0.87 0.75 0.25   1
##
##                 PC1
## SS loadings    5.83
## Proportion Var 0.83
##
```

```
## Mean item complexity =  1
## Test of the hypothesis that 1 component is sufficient.
##
## The root mean square of the residuals (RMSR) is  0.08
##
## Fit based upon off diagonal values = 0.99
```

Onde podemos notar que o fator possui uma proporção de variabilidade total de 0.83.

Sendo as comunalidades dadas por:

Table 13: Comunalidades

100m	200m	400m	800m	1500m	3000m	Maratona
0.814	0.848	0.787	0.899	0.889	0.845	0.749

Onde cada comulalidade representa a proporção de variabilidade de X_j $(j = 1, ..., 7)$ explicada pelos fatores, em que os valores estão contidos no intervalo de 0.7 e 0.9.

E as variâncias específicas:

Table 14: Variâncias especifícas

100m	200m	400m	800m	1500m	3000m	Maratona
0.186	0.152	0.213	0.101	0.111	0.155	0.251

Onde cada variância específica representa a parcela da variância não explicada pelos fatores, onde os valores estão entre 0.1 e 0.3.

Exercício 6

Os vetores $X^{(1)}$ e $X^{(2)}$ apresentam os seguintes vetores de média e matriz de covariância:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix} = \left[\begin{array}{c} -3 \\ 2 \\ \hdashline 0 \\ 1 \end{array}\right]$$

e

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} = \left[\begin{array}{cc:cc} 8 & 2 & 3 & 1 \\ 2 & 5 & -1 & 3 \\ \hdashline 3 & -1 & 6 & -2 \\ 1 & 3 & -2 & 7 \end{array}\right]$$

(a) Calcule as correlações canônicas.

Resolução

As correlações canônicas em valor absoluto são:

Table 15: Correlações canônicas

0.552	0.49

Em que 0.552 é a correlação canônica entre U_1 e V_1 e 0.49 é a correlação canônica entre U_2 e V_2.

(b) Obtenha os pares de variáveis canônicas (U_1, V_1) e (U_2, V_2).

Resolução

Os pares de variáveis canônicas (U_1, V_1) e (U_2, V_2) são:

$U_1 = 0.32X_{11} - 0.36X_{12}$

$U_2 = 0.19X_{11} + 0.3X_{12}$

E

$V_1 = 0.36X_{21} - 0.09X_{22}$

$V_2 = 0.23X_{21} + 0.38X_{22}$

(c) Obtenha os autovalores de $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ e compare com os autovalores de $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}^{-1/2}$.

Resolução

Para a matriz $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ obtemos os seguintes autovalores:

Table 16: Autovalores

λ_1	λ_2
0.305	0.24

Para a matriz $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}^{-1/2}$ obtemos os seguintes autovalores:

Table 17: Autovalores

λ_1	λ_2
0.305	0.24

Em que podemos notar que os autovalores são identicos para as duas expressões.

Exercício 7

Quatro diferentes testes foram aplicados em $n = 140$ crianças da sétima série nos Estados Unidos. Os testes aplicados foram:

- Leitura:

– $X_1^{(1)}$: Velocidade

– $X_2^{(1)}$: Capacidade de interpretação

- Matemática:

– $X_1^{(2)}$: Velocidade

– $X_2^{(2)}$: Capacidade ou habilidade

A seguinte matriz de correlação foi obtida com os dados:

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \boldsymbol{R}_{21} & \boldsymbol{R}_{22} \end{bmatrix} = \left[\begin{array}{cc:cc} 1,0000 & & & \\ 0,6328 & 1,0000 & & \\ \hdashline 0,2412 & -0,0553 & 1,0000 & \\ 0,0586 & 0,0655 & 0,4248 & 1,0000 \end{array}\right].$$

(a) Determine as correlações canônicas amostrais.

Resolução

As correlações canônicas em valor absoluto são:

Table 18: Correlações canônicas

0.395	0.069

(b) Teste a hipótese $H_0 : \Sigma_{12} = 0$ com nível de significância de 5%. Se a hipótese nula for rejeitada, teste a hipótese da primeira correlação canônica apenas ser igual a zero.

Resolução

Testando a hipótese $H_0 : \Sigma_{12} = 0$ utlizando a estatística de razão de verossimilhança, temos que:

$$-2ln\Lambda = nln(\frac{|S_{11}||S_{22}|}{|S|})$$

Em que S_{11} e S_{22} são as submatrizes que estimam as submatrizes populacionais Σ_{11} e Σ_{22} respectivamente e S é a matriz que estima a matriz Σ, assim:

$|S_{11}| = 0.599$, $|S_{22}| = 0.819$ e $|S| = 0.413$

Obtendo a seguinte estatística de teste:

```
## [1] 24.34903
```

Como visto em aula, a estatística $-2ln\Lambda \sim \chi^2_{p*q}$ quando $n \to \infty$ e como temos p e q varíaveis iguais a 2 temos o seguinte *p-value* do teste:

```
## [1] 6.798461e-05
```

Assim, com un nível de significância de 5% temos evidências estatísticas para rejeitar H_0.

Como o teste acima rejeitou H_0, iremos testar se apenas a primeira correlação é igual a zero:

$$\begin{cases} H_0 : \rho_1 = 0 \\ H_1 : \rho_1 \neq 0 \end{cases}$$

Para testar a hipótese utilizaremos a estatística de teste:

$$t_{teste} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

que possui distribuição assintótica t student com n-2 graus de liberdade.

logo,

$$t_{teste} = \frac{0.395}{\sqrt{\frac{1-0.395^2}{140-2}}} = 5.0509$$

Como $n = 140$, é possível aproximar para a normal, deste modo o p-valor é de:

```
## [1] 4.572062e-07
```

Desse modo, rejeita-se H_0, logo a primeira correlação canônica não é igual a 0 a um nível de significância de 5%.

(c) Obtenha as variávies canônicas (utilizando-se os dados padronizados).

Resolução

Os pares de variáveis canônicas (U_1, V_1) e (U_2, V_2) são:

$U_1 = 1.26Z_{11} - 1.03Z_{12}$

$U_2 = 0.29Z_{11} + 0.78Z_{12}$

E

$V_1 = -1.10Z_{21} + 0.45Z_{22}$

$V_2 = 0.02Z_{21} - 1.01Z_{22}$

(d) Obtenha um tabela com as correlações entre as variáveis canônicas e as variáveis originais.

Resolução

$$cor(U, X_1^{(1)}) = AR_{11}$$
$$cor(U, X_2^{(1)}) = AR_{12}$$
$$cor(V, X_1^{(2)}) = BR_{21}$$
$$cor(V, X_2^{(2)}) = BR_{22}$$

resolvendo, respectivamente:

```
##            [,1]        [,2]
## [1,] -0.4410068 -0.39673292
## [2,]  0.1713205 -0.07246798
```

```
##             [,1]         [,2]
## [1,] -0.06556456 -0.031419986
## [2,]  0.10405231  0.001466118
```

```
##               [,1]         [,2]
## [1,] -0.1012267305  0.005347241
## [2,]  0.0003200176 -0.030528241
```

```
##             [,1]       [,2]
## [1,] -0.46082726 -0.3812017
## [2,] -0.06883403 -0.3454383
```

(e) Interprete as variáveis canônicas.

Resolução

U_1 representa a oposição entre a velocidade de leitura e a capacitação de interpretação, U_2 se refere a associação entre a velocidade de leitura e a capacitação de interpretação. V_1 explica a velocidade em matemática em oposição à sua capacidade ou habilidade. E por último, V_2 diz sobre a capacidade ou habilidade matemática.

Exercício 8

Os dados disponíveis no arquivo **T7-7.dat** são referentes a propriedades de polpa (ou pasta) de celulose utilizada para fabricação do papel e também algumas propriedades do papel produzido com a polpa. Os dados são de 62 observações e as variáveis observadas são:

- Propriedades do papel:

– $X_1^{(1)}$: BL (*breaking length*);

– $X_2^{(1)}$: EM (*elastic modulus*);

– $X_3^{(1)}$: SF (*stress ar failure*);

– $X_4^{(1)}$: BS (*burst strength*).

- Propriedades da polpa de celulose:

– $X_1^{(2)}$: AFL (*arithmetic fiber length*);

– $X_2^{(2)}$: LFF (*long fiber fraction*);

– $X_3^{(2)}$: FFF (*fine fiber fraction*);

– $X_4^{(2)}$: ZST (*zero span tensile*).

Obtenha os pares de variáveis canônicas e as correlações canônicas. O primeiro par de variáveis canônicas é uma boa medida sumária das variáveis que representam? Justifique a resposta. Teste a significância das correlações canônicas e interprete os pares de variáveis canônicas com correlações significativas com nível de significância igual a 5%.

Resolução

Os pares de de variáveis canônicas são:

U_1	V_1	U_2	V_2
1.505*BL	0.159*AFL	-3.496*BL	0.689*AFL
0.212*EM	-0.632*LFF	-1.543*EM	1.003*LFF
-1.998*SF	-0.325*FFF	1.076*SF	0.005*FFF
-0.676*BS	-0.818*ZST	3.768*BS	-1.562*ZST

E as correlações canônicas são:

Table 19: Correlações Canônicas

0.917	0.817	0.265	0.092

Podemos notar que U_1 e U_2 explicam a diferença entre as propriedades BL e EM do papel e as propriedades SF e BS do papel. V_1 por sua vez, explica a diferença entre a propriedade AFL e as demais propriedades da polpa de celulose. Por fim, V_2 explica a diferença entre a variável ZST e as demais propriedades da polpa de celulose. Nota-se também que a correlação entre U_1 e V_1 é alta (0.917), assim como a correlação entre U_2 e V_2 (0.817).

Fazendo o teste de significância das correlações canônicas, temos:

```
##
## Canonical correlation analysis of:
##   4   X  variables:  BL, EM, SF, BS
##   with    4   Y  variables:  AFL, LFF, FFF, ZST
##
##      CanR   CanRSQ    Eigen percent    cum                              scree
## 1 0.91733 0.841493 5.308872 71.7468  71.75 ******************************
## 2 0.81693 0.667370 2.006340 27.1147  98.86 ***********
## 3 0.26539 0.070429 0.075766  1.0239  99.89
## 4 0.09168 0.008406 0.008477  0.1146 100.00
##
## Test of H0: The canonical correlations in the
## current row and all that follow are zero
```

```
##
##      CanR LR test stat approx F numDF  denDF   Pr(> F)
## 1 0.91733      0.04860  17.5022    16 165.61 < 2.2e-16 ***
## 2 0.81693      0.30660   9.3119     9 134.01 6.688e-11 ***
## 3 0.26539      0.92176   1.1642     4 112.00    0.3305
## 4 0.09168      0.99159   0.4832     1  57.00    0.4898
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

E com o teste acima pode se notar que os dois primeiros pares de variáveis canônicas são significantes para a análise.