Universidade de Brasília

DEPARTAMENTO DE ESTATÍSTICA

25 junho 2023

Lista 6 - Análise fatorial exploratória

Prof. Dr. George von Borries

Análise Multivariada 1

Aluno: Bruno Gondim Toledo | Matrícula: 15/0167636

Questão 43

Ex. 9.1 | Johnson & Wichern

Show that the covariance matrix

$$\rho = \begin{bmatrix} 1 & .63 & .45 \\ .63 & 1 & .35 \\ .45 & .35 & 1 \end{bmatrix}$$

for the $p = 3$ standardized random variables Z_1, Z_2, and Z_3 can be generated by $m = 1$ factor model

$$Z_1 = .9F_1 + \epsilon_1 Z_2 = .7F_1 + \epsilon_2 Z_3 = .5F_1 + \epsilon_3$$

where $Var(F_1) = 1, Cov(\epsilon, F_1) = 0$, and

$$\Psi = Cov(\epsilon) = \begin{bmatrix} .19 & 0 & 0 \\ 0 & .51 & 0 \\ 0 & 0 & .75 \end{bmatrix}$$

That is, write ρ in the form of $\rho = \mathbf{LL^T} + \mathbf{\Psi}$.

Solução:

Do modelo geral, $\mathbf{X} = \mu + \mathbf{LF} + \epsilon$, temos que $\mathbf{L} = [\ .9 \ \ .7 \ \ .5]$, logo,

$$\mathbf{L^T} = \begin{bmatrix} .9 \\ .7 \\ .5 \end{bmatrix}$$

Multiplicando as matrizes $\mathbf{L}$ e $\mathbf{L^T}$, obtemos

$$\mathbf{LL^T} = \begin{bmatrix} .81 & .63 & .45 \\ .63 & .49 & .35 \\ .45 & .35 & .25 \end{bmatrix}$$

Somando essa matriz a matriz $\mathbf{\Psi}$, obtemos

$$\mathbf{LL^T} + \mathbf{\Psi} = \begin{bmatrix} 1 & .63 & .45 \\ .63 & 1 & .35 \\ .45 & .35 & 1 \end{bmatrix}$$

Que é precisamente a matriz ρ, o que demonstra o resultado $\square$

Questão 44

Ex. 9.2 | Johnson & Wichern

Use the information in Exercise 9.1.

(a) Calculate communalities $h_i^2, i = 1, 2, 3$ and interpret these quantities.

(b) Calculate $Corr(Z_i, F_1)$ for $i = 1, 2, 3$. Which variable might carry the greatest weight in "naming" the common factor? Why?

Soluções:

a) As comunalidades são dadas por: $\sum_{j=1}^{n} l_{ij}^2 = h_i^2$. Para a matriz $\mathbf{L} = [.9 \quad .7 \quad .5]$, temos que as comunalidades são:

$h_1^2 = .81$
$h_2^2 = .49$
$h_3^2 = .25$.

Como as comunalidades são quantidades de variâncias de cada variável explicada pelos fatores, quanto maior for a comunalidade, maior será o poder de explicação daquela variável pelo fator. A comunalidade h_i^2 assume valores no intervalo [0,1]. Desejamos, em geral, valores acima de 0.5. Neste caso, temos que $h_1^2 > 0.5$, enquanto $h_2^2, h_3^2 < 0.5$. Entretanto $h_2^2 \approx 0.5$, temos que h_2^2 também pode ser utilizada.

b) Como $Cov(\mathbf{X}, \mathbf{F} = \mathbf{L})$, e $Cov(\mathbf{X_i}, \mathbf{F_j}) = \ell ij$ (Resultado 2., pag. 484 J&W) [1], e $Cov(\mathbf{X_1}, \mathbf{F_1}) = Corr(\mathbf{X_1}, \mathbf{F_1})$ (pag. 486 J&W) [1] sabemos que $Cor(\mathbf{Z_i}, \mathbf{F_1})$, para $i = 1, 2, 3$ será $\ell_{i1} = [.9 \quad .7 \quad .5] = \mathbf{L}$. Isso indica que a variável Z_1 carrega a maior carga fatorial, dado seu maior valor absoluto (comparando também com o último resultado encontrado sobre comunalidade).

Questão 45

Ex. 9.3 | Johnson & Wichern

Ex. 9.3 | Johnson & Wichern

The eigenvalues and eigenvectors of the correlation matrix ρ in Exercise 9.1 are

$$\begin{aligned}
\lambda_1 &= 1.96, & e_1' &= [.625, \ .594, \ .507] \\
\lambda_2 &= .68, & e_2' &= [-.219, \ -.491, \ .843] \\
\lambda_3 &= .36, & e_3' &= [.749, \ -.638, \ -.177]
\end{aligned}$$

(a) Assuming an $m = 1$ factor model, calculate the loading matrix $\mathbf{L}$ and matrix of specific variances Ψ using the principal component solution method. Compare the results with those in Exercise 9.1.

(b) What proportion of the total population variance is explained by the first common factor?

Soluções

a) Sabemos que $\Sigma = \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + ... + \lambda_p e_p e_p' = [\sqrt{\lambda_1} e_1 | \sqrt{\lambda_2} e_2 | ... | \sqrt{\lambda_p} e_p] \cdot [\sqrt{\lambda_1} e_1 | \sqrt{\lambda_2} e_2 | ... | \sqrt{\lambda_p} e_p]^T$ (pag. 488 J&W) [1]. Visto que iremos trabalhar com $m = 1$, iremos considerar apenas o autovalor $\lambda_1 = 1.96$ e o autovetor $e_1' = [.625, \ .593, \ .507]$ tal que:

$\sqrt{1.96}[.625 \quad .593 \quad .507] = \mathbf{L}$;

e

$\Sigma = \mathbf{L}\mathbf{L}^\mathbf{T} + \mathbf{\Psi} =$

```
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.7280211 0.6213917
## [2,] 0.7280211 1.0000000 0.5905707
## [3,] 0.6213917 0.5905707 1.0000000
```

Onde a matriz $\mathbf{\Psi}$ foi obtida simplesmente somando o valor necessário para a matriz $\mathbf{L}\mathbf{L}^\mathbf{T}$ ter o valor 1 na diagonal principal, ou seja, $\mathbf{\Psi} =$

```
##           [,1]      [,2]     [,3]
## [1,] 0.2339845 0.0000000 0.000000
## [2,] 0.0000000 0.3080887 0.000000
## [3,] 0.0000000 0.0000000 0.495927
```

b) A proporção da variância explicada pelo primeiro fator comum é $\frac{\lambda_1}{\sum_{i=1}^{3} \lambda_i} = \frac{1.96}{1.96+0.68+0.36} = \frac{1.96}{3} \approx 0.65$.

Questão 46

Ex. 9.6 | Johnson & Wichern

Verificar as identidades

a) $(\mathbf{I} + \mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L})^{-1}\mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L} = \mathbf{I} - (\mathbf{I} + \mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L})^{-1}$

Sol:

$(\mathbf{I} + \mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L})(\mathbf{I} + \mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L})^{-1}\mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L} = (\mathbf{I} + \mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L})[\mathbf{I} - (\mathbf{I} + \mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L})^{-1}]$
$\rightarrow \mathbf{I}\mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L} = \mathbf{I} + \mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L} - \mathbf{I}$
$\rightarrow \mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L} = \mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L}$
$\square$

b) $(\mathbf{L}\mathbf{L}' + \mathbf{\Psi})^{-1} = \mathbf{\Psi}^{-1} - \mathbf{\Psi}^{-1}\mathbf{L}(\mathbf{I} + \mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L})^{-1}\mathbf{L}'\mathbf{\Psi}^{-1}$

Sol:

$(\mathbf{L}\mathbf{L}' + \mathbf{\Psi})^{-1}(\mathbf{L}\mathbf{L}' + \mathbf{\Psi}) = [\mathbf{\Psi}^{-1} - \mathbf{\Psi}^{-1}\mathbf{L}(\mathbf{I} + \mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L})^{-1}\mathbf{L}'\mathbf{\Psi}^{-1}](\mathbf{L}\mathbf{L}' + \mathbf{\Psi})$
$\rightarrow \mathbf{I} = [\mathbf{\Psi}^{-1} - \mathbf{\Psi}^{-1}\mathbf{L}(\mathbf{I} + \mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L})^{-1}\mathbf{L}'\mathbf{\Psi}^{-1}](\mathbf{L}\mathbf{L}' + \mathbf{\Psi})$
Aproveitando o resultado encontrado em a) para $\mathbf{L}(\mathbf{I} + \mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L})^{-1}\mathbf{L}'\mathbf{\Psi}^{-1}$, obtemos:
$\rightarrow \mathbf{I} = \mathbf{\Psi}^{-1}(\mathbf{L}\mathbf{L}' + \mathbf{\Psi}) - \mathbf{\Psi}^{-1}\mathbf{L}(\mathbf{I} - (\mathbf{I} + \mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L})^{-1})\mathbf{L}' - \mathbf{\Psi}^{-1}\mathbf{L}(\mathbf{I} + \mathbf{L}'\mathbf{\Psi}^{-1})^{-1}$
Expandindo e reorganizando, obtemos:
$\rightarrow \mathbf{I} = [\mathbf{I}]_{(1)}[\mathbf{\Psi}^{-1}\mathbf{L}\mathbf{L}']_{(2)} - [\mathbf{\Psi}^{-1}\mathbf{L}\mathbf{L}']_{(3)} + [\mathbf{\Psi}^{-1}\mathbf{L}(\mathbf{I} + \mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L})^{-1}\mathbf{L}']_{(4)} - [\mathbf{\Psi}^{-1}\mathbf{L}(\mathbf{I} + \mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L})^{-1}\mathbf{L}']_{(5)}$
Assim, fica claro que podemos cancelar $_{(2)}$ com $_{(3)}$; e $_{(4)}$ com $_{(5)}$, sobrando apenas a identidade $\mathbf{I} = \mathbf{I}$
$\square$

c) $\mathbf{L}'(\mathbf{L}\mathbf{L}' + \mathbf{\Psi})^{-1} = (\mathbf{I} + \mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L})^{-1}\mathbf{L}'\mathbf{\Psi}^{-1}$

Questão 47

Ex. 9.14 | Johnson & Wichern (Reproduza dos resultados do)

Dado que a estimativa de máxima verossimilhança das cargas fatoriais é dada por $\hat{\mathbf{L}} = \hat{\mathbf{\Psi}}^{1/2}\hat{\mathbf{E}}\hat{\mathbf{\Delta}}^{1/2}$; verificar que $\hat{\mathbf{L}}'\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{L}} = \hat{\mathbf{\Delta}}$

Sol:

Como $\hat{\mathbf{L}} = \hat{\mathbf{\Psi}}^{1/2}\hat{\mathbf{E}}\hat{\mathbf{\Delta}}^{1/2}$; então $\hat{\mathbf{L}}' = \hat{\mathbf{\Psi}}^{1/2}\hat{\mathbf{E}}'\hat{\mathbf{\Delta}}^{1/2}$. Como $\hat{\mathbf{\Psi}}^{1/2}\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{\Psi}}^{1/2} = \mathbf{I}$; $\hat{\mathbf{E}}'\hat{\mathbf{E}} = \mathbf{I}$; e $\hat{\mathbf{\Delta}}^{1/2}\hat{\mathbf{\Delta}}^{1/2} = \hat{\mathbf{\Delta}}$; organizando a equação, teremos:
$\rightarrow \hat{\mathbf{L}}'\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{L}} = \hat{\mathbf{\Delta}}^{1/2}\hat{\mathbf{E}}'\hat{\mathbf{\Psi}}^{1/2}\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{\Psi}}^{1/2}\hat{\mathbf{E}}\hat{\mathbf{\Delta}}^{1/2}$
$\rightarrow \hat{\mathbf{L}}'\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{L}} = \hat{\mathbf{\Delta}}^{1/2}\hat{\mathbf{E}}'\hat{\mathbf{E}}\hat{\mathbf{\Delta}}^{1/2} = \hat{\mathbf{\Delta}}$.
$\square$

Questão 48

Ex. 9.10 | Johnson & Wichern

a) Variância específica:

```r
t(t(sol$uniquenesses))
```

```
##            [,1]
## [1,] 0.32912803
## [2,] 0.11572227
## [3,] 0.08309722
## [4,] 0.07959191
## [5,] 0.08165456
## [6,] 0.07194759
```

b) Comunalidades:

```
t(t(sol$communality))
```

```
##           [,1]
## [1,] 0.6708720
## [2,] 0.8842777
## [3,] 0.9169028
## [4,] 0.9204081
## [5,] 0.9183454
## [6,] 0.9280524
```

c) Proporção da variância explicada por cada fator:

```
h2 <- c(.67,.88,.92,.92,.92,.93)
t(t(h2))
```

```
##      [,1]
## [1,] 0.67
## [2,] 0.88
## [3,] 0.92
## [4,] 0.92
## [5,] 0.92
## [6,] 0.93
```

d) Matriz de resíduos:

```
psi47d <- matrix(rep(0,36),6,6)
diag(psi47d) <- sol[["uniquenesses"]]
corm - (sol$loadings[,1:2] %*% t(sol$loadings[,1:2]) + psi47d)
```

```
##             [,1]        [,2]         [,3]         [,4]        [,5]        [,6]
## [1,]  0.00000000 -0.19460300 -0.037323920 -0.035565293 -0.03074747 -0.03048646
## [2,] -0.19460300  0.00000000  0.028615783  0.022712676  0.01238982  0.01529278
## [3,] -0.03732392  0.02861578  0.000000000  0.009157101 -0.03639513 -0.04338177
## [4,] -0.03556529  0.02271268  0.009157101  0.000000000 -0.04486829 -0.03012616
## [5,] -0.03074747  0.01238982 -0.036395127 -0.044868292  0.00000000  0.01486252
## [6,] -0.03048646  0.01529278 -0.043381774 -0.030126155  0.01486252  0.00000000
```

Questão 49

Ex. 9.12 | Johnson & Wichern

Questão 50

Ex. 9.25 | Johnson & Wichern

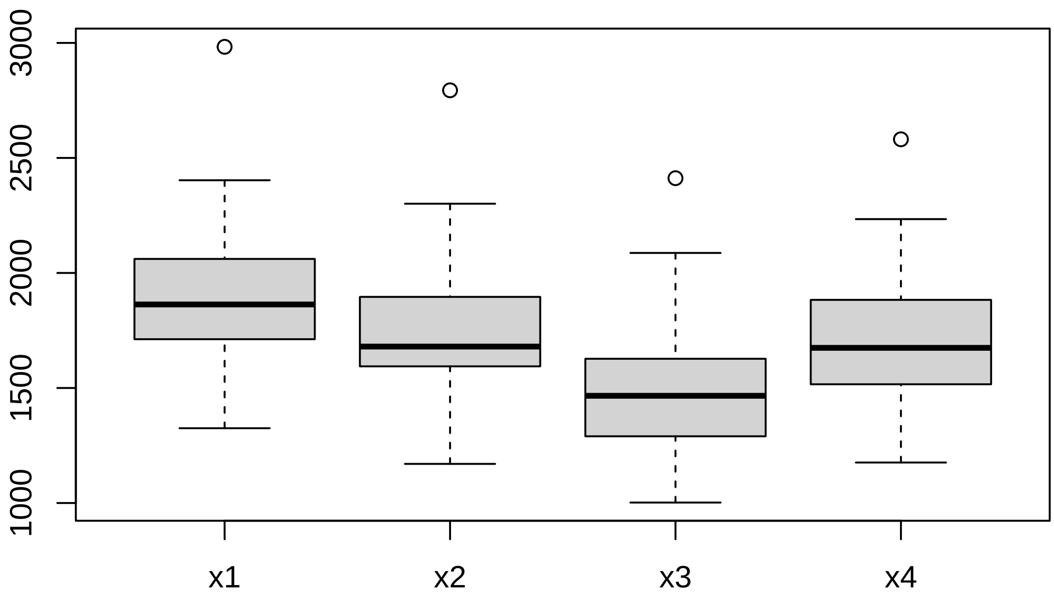
Escores:

```
##           Factor1
##  [1,] -0.03355809
##  [2,]  1.55569643
##  [3,]  0.73421647
##  [4,] -0.81826312
##  [5,]  0.30565376
##  [6,] -0.52229581
##  [7,] -0.04835030
##  [8,]  0.57992839
##  [9,]  3.32638788
```

```
## [10,] -0.52487996
## [11,] -0.49637359
## [12,]  0.46314882
## [13,] -0.07309942
## [14,] -0.13024549
## [15,] -0.18374837
## [16,] -0.03768372
## [17,] -1.84016710
## [18,] -1.42858939
## [19,] -0.16474754
## [20,] -0.54372769
## [21,]  1.10886136
## [22,] -0.18692428
## [23,] -0.82697020
## [24,]  0.53980327
## [25,] -0.32490918
## [26,] -0.69680811
## [27,]  0.72390424
## [28,] -0.64782459
## [29,]  1.48607536
## [30,] -1.29451004
```

(Possíveis) Outliers:

```
boxplot(stiff)
```



A análise dos boxplots das variáveis nos leva a crer que existem outliers nos dados, no caso, um em cada variável (pareado, nas 4 variáveis o outlier se refere a 9ª observação). Olhando para os escores computados acima, notamos que o escore da 9ª observação se destaca dos demais, sendo o maior.

Questão 51

Ex. 9.19 | Johnson & Wichern

A firm is attempting to evaluate the quality of its sales staff and is trying to find an examination or series of test that may reveal the potential for good performance in sales. The firm has selected a random sample of 50 sales people and has evaluated each on 3 measures of performance: growth of sales, profitability of sales, and new-account sales. These measures have been converted to a scale, on which 100 indicates "average" performance. Each of the 50 individuals took each 4 test, which purported to measure creativity, mechanical reasoning, abstract reasoning, and mathematical ability, respectively. The $n = 50$ observations on $p = 7$ variables are listed in Table 9.12.

(a) Assume an orthogonal factor model for the standardized variables $Z_i = \frac{(X_i - \mu_i)}{\sqrt{\sigma_{ii}}}, i = 1, 2, ..., 7$. Obtain either the principal component solution or the maximum likehood solution for $m = 2$ and $m = 3$ common factors.

(b) Given your solution in (a), obtain the rotated loadings for $m = 2$ and $m = 3$. Compare the two sets of rotated loadings. Interpret the $m = 2$ and $m = 3$ factor solutions.

(c) List the estimated communalities, specific variances, and $\hat{\mathbf{L}}\hat{\mathbf{L}}^T + \hat{\mathbf{\Psi}}$ for the $m = 2$ and $m = 3$ solutions. Compare the results. Which choice of m do you prefer at this point? Why?

(d) Conduct a test of $H_0 : \mathbf{\Sigma} = \mathbf{LL}' + \mathbf{\Psi}$ versus $H_1 : \mathbf{\Sigma} \neq \mathbf{LL}' + \mathbf{\Psi}$ for both $m = 2$ and $m = 3$ at the $\alpha = .01$ level. With these results and those in Parts b and c, which choice of m appears to be the best?

(e) Suppose a new salesperson, selected at random, obtains the test scores $\mathbf{x}' = [x_1, x_2, ..., x_7] = [110, 98, 105, 15, 18, 12, 35]$. Calculate the salesperson's factor score using the weighted least squares method and the regression method.

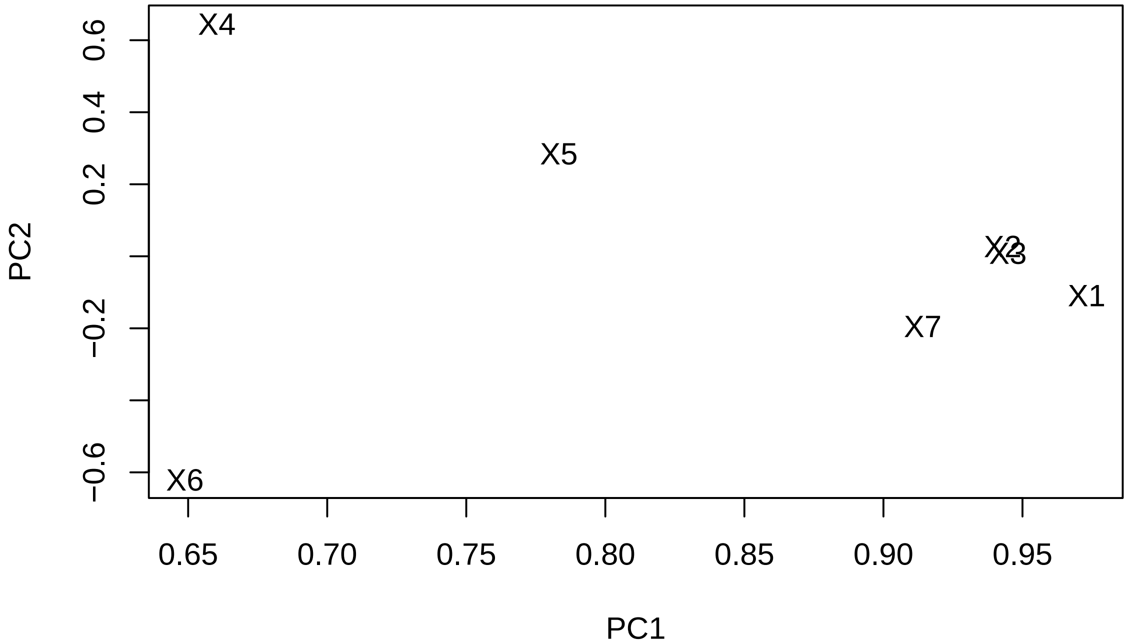
Soluções

a)

Solução por componentes principais para m = 2:

```
## Principal Components Analysis
## Call: principal(r = dados, nfactors = 2, rotate = "none", covar = FALSE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##     PC1   PC2   h2    u2 com
## X1 0.97 -0.11 0.96 0.041 1.0
## X2 0.94  0.03 0.89 0.110 1.0
## X3 0.94  0.01 0.89 0.107 1.0
## X4 0.66  0.65 0.85 0.147 2.0
## X5 0.78  0.28 0.69 0.305 1.3
## X6 0.65 -0.62 0.81 0.194 2.0
## X7 0.91 -0.19 0.87 0.127 1.1
##
##                        PC1  PC2
## SS loadings           5.03 0.93
## Proportion Var        0.72 0.13
## Cumulative Var        0.72 0.85
## Proportion Explained  0.84 0.16
## Cumulative Proportion 0.84 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.08
##  with the empirical chi square  11.93  with prob <  0.15
##
```

```
## Fit based upon off diagonal values = 0.99
```



Solução por máxima verossimilhança para m = 3:

```
##
## Call:
## factanal(factors = 3, covmat = cor(dados), rotation = "none")
##
## Uniquenesses:
##    X1    X2    X3    X4    X5    X6    X7
## 0.039 0.034 0.088 0.005 0.447 0.005 0.038
##
## Loadings:
##    Factor1 Factor2 Factor3
## X1  0.901   0.381
## X2  0.775   0.600
## X3  0.931   0.202
## X4  0.733  -0.118   0.666
## X5  0.689   0.225   0.169
## X6  0.757  -0.132  -0.636
## X7  0.762   0.608  -0.110
##
##                Factor1 Factor2 Factor3
## SS loadings      4.445   0.998   0.901
## Proportion Var   0.635   0.143   0.129
## Cumulative Var   0.635   0.778   0.906
##
## The degrees of freedom for the model is 3 and the fit was 1.4186
```

b)

Rotação varimax na solução por componentes principais para m = 2:

```
## Principal Components Analysis
## Call: principal(r = dados, nfactors = 2, rotate = "varimax", covar = FALSE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##     RC1   RC2   h2    u2 com
## X1 0.79  0.58 0.96 0.041 1.8
## X2 0.67  0.66 0.89 0.110 2.0
## X3 0.68  0.65 0.89 0.107 2.0
## X4 0.04  0.92 0.85 0.147 1.0
## X5 0.38  0.74 0.69 0.305 1.5
## X6 0.90 -0.01 0.81 0.194 1.0
## X7 0.80  0.48 0.87 0.127 1.6
##
##                        RC1  RC2
## SS loadings           3.13 2.84
## Proportion Var        0.45 0.41
## Cumulative Var        0.45 0.85
## Proportion Explained  0.52 0.48
## Cumulative Proportion 0.52 1.00
##
## Mean item complexity =  1.6
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.08
##  with the empirical chi square  11.93  with prob <  0.15
##
## Fit based upon off diagonal values = 0.99
```

Rotação quartimax na solução por máxima verossimilhança para m = 3:

```
##
## Call:
## factanal(factors = 3, covmat = cor(dados), rotation = "quartimax")
##
## Uniquenesses:
##    X1    X2    X3    X4    X5    X6    X7
## 0.039 0.034 0.088 0.005 0.447 0.005 0.038
##
## Loadings:
##    Factor1 Factor2 Factor3
## X1  0.964           0.160
## X2  0.976          -0.114
## X3  0.891   0.281   0.196
## X4  0.544   0.830
## X5  0.698   0.256
## X6  0.572  -0.102   0.811
## X7  0.971  -0.139
##
##                Factor1 Factor2 Factor3
## SS loadings      4.731   0.869   0.745
## Proportion Var   0.676   0.124   0.106
## Cumulative Var   0.676   0.800   0.906
##
## The degrees of freedom for the model is 3 and the fit was 1.4186
```

Interpretação: No caso da solução para $m = 2$, foi utilizada a rotação varimax que busca explicitar a relação entre os fatores, enquanto na solução para $m = 3$ optei por utilizar a rotação quartimax, que busca minimizar o número de fatores necessários para explicar a variável. Em ambos os casos, notamos que a variância acumulada na primeira carga aumenta, em especial para a rotação quartimax para $m = 3$. Como assumimos a ortogonalidade do modelo ao início do problema, optamos por essas duas rotações ortogonais. Como sabemos, a rotação da base serve para encontrar a posição de maior variância explicada pela elipsoide projetada no plano; não alterando portanto a informação contida nos dados.

c)

Solução: Para permitir a comparabilidade, o ideal agora é solucionar pelo mesmo método. Irei optar por componentes principais em ambos os casos. Para $m = 2$, temos as comunalidades $h_i^2 \approx [.95\ .88\ .89\ .85\ .69\ .80\ .87]$, variância específica $[.04\ .11\ .10\ .14\ .30\ .19\ .12]$ e $\hat{\mathbf{L}}\hat{\mathbf{L}} + \hat{\boldsymbol{\Psi}} =$

```
##            X1        X2        X3         X4        X5         X6        X7
## X1 1.0000000 0.9144237 0.9183475 0.57280177 0.7314249 0.69837236 0.9104147
## X2 0.9144237 1.0000000 0.8910297 0.64086858 0.7466052 0.59420052 0.8564286
## X3 0.9183475 0.8910297 1.0000000 0.62957641 0.7425467 0.60746435 0.8619032
## X4 0.5728018 0.6408686 0.6295764 1.00000000 0.7012578 0.02760441 0.4785904
## X5 0.7314249 0.7466052 0.7425467 0.70125775 1.0000000 0.33135247 0.6608599
## X6 0.6983724 0.5942005 0.6074644 0.02760441 0.3313525 1.00000000 0.7132694
## X7 0.9104147 0.8564286 0.8619032 0.47859044 0.6608599 0.71326939 1.0000000
```

Enquanto para $m = 3$ temos as comunalidades $h_i^2 = [.96\ .98\ .91\ .95\ .69\ .98\ .96]$, variância específica $[.03\ .01\ .08\ .04\ .30\ .01\ .03]$ e $\hat{\mathbf{L}}\hat{\mathbf{L}} + \hat{\boldsymbol{\Psi}} =$

```
##           X1        X2        X3        X4        X5        X6        X7
## X1 1.0000000 0.9310552 0.9106726 0.5558314 0.7312023 0.6756780 0.9267290
## X2 0.9310552 1.0000000 0.8460429 0.5413962 0.7453005 0.4611764 0.9520558
## X3 0.9106726 0.8460429 1.0000000 0.6754796 0.7431488 0.6688506 0.8177744
## X4 0.5558314 0.5413962 0.6754796 1.0000000 0.7025890 0.1633383 0.3810153
## X5 0.7312023 0.7453005 0.7431488 0.7025890 1.0000000 0.3331328 0.6595801
## X6 0.6756780 0.4611764 0.6688506 0.1633383 0.3331328 1.0000000 0.5827824
## X7 0.9267290 0.9520558 0.8177744 0.3810153 0.6595801 0.5827824 1.0000000
```

Se compararmos com a matriz ρ de variância-covariância:

```
##           X1        X2        X3        X4        X5        X6        X7
## X1 1.0000000 0.9260758 0.8840023 0.5720363 0.7080738 0.6744073 0.9273116
## X2 0.9260758 1.0000000 0.8425232 0.5415080 0.7459097 0.4653880 0.9442960
## X3 0.8840023 0.8425232 1.0000000 0.7003630 0.6374712 0.6410886 0.8525682
## X4 0.5720363 0.5415080 0.7003630 1.0000000 0.5907360 0.1469074 0.4126395
## X5 0.7080738 0.7459097 0.6374712 0.5907360 1.0000000 0.3859502 0.5745533
## X6 0.6744073 0.4653880 0.6410886 0.1469074 0.3859502 1.0000000 0.5663721
## X7 0.9273116 0.9442960 0.8525682 0.4126395 0.5745533 0.5663721 1.0000000
```

Neste caso, notamos que com $m = 2$ já é possível obter $\approx 85\%$ da variância, o que me parece ser um valor bem razoável pelo tamanho da redução. Portanto, eu optaria pela solução com $m = 2$ para problemas não tão conservadores, em que é satisfatório este valor. Talvez se fosse um estudo muito crítico e com necessidade de ser conservador, pudesse ser considerada a solução para $m = 3$.

d) Para testar as hipóteses:

$$\begin{cases} H_0 : \boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi} \\ H_1 : \boldsymbol{\Sigma} \neq \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi} \end{cases}$$

Sob um nível de significância $\alpha = .01$, iremos portanto testar a hipótese de número de Fatores (m) adequado. Note que este teste assume a normalidade multivariada dos dados amostrais, pressuposto este que não será testado. Conforme as notas de aula [2] references a análise fatorial, a estatística do teste (razão de verossimilhança) é:

$$(n - \frac{2p + 4m + 11}{6})\ ln\ (\frac{|\hat{L}\hat{L}^T + \hat{\Psi}|}{|S|}),$$

que tem distr. aproximada $\chi^2_\ell, \ell = [(p-m)^2 - p - m]/2$, e $\hat{\mathbf{L}}, \hat{\mathbf{\Psi}}$ são as estimativas de máxima verossimilhança.

Calculando as estatísticas de teste para m=2 e m=3 obtemos respectivamente os p-valores $1.1921862 \times 10^{-21}$ e $1.4550886 \times 10^{-13}$, em outras palavras, rejeitamos h_0 em ambos os casos para $\alpha = 0.01$.

Como este é um teste sensível a normalidade, talvez os resultados encontrados não sejam muito precisos sem testar o pressuposto de normalidade multivariada dos dados, o que pode influenciar no poder do teste, visto que as matrizes são similares a matriz de correlação e o teste rejeita a igualdade com tamanha certeza. Dito isso, este teste não auxiliou muito na tomada de decisão do número de fatores m adequado.

e)

Solução Pelo método de mínimos quadrados ponderados, para o caso da solução $m = 2$, temos que $\hat{\mathbf{f}}_\mathbf{j} = (\hat{\mathbf{L}}'\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{L}})^{-1}\hat{\mathbf{L}}'\hat{\mathbf{\Psi}}^{-1}(\mathbf{x_j} - \hat{\mu})$ (pag. 515 J&W) [1]. Portanto, teremos:

```r
# m=2
# Seja L =

L <- q51a1$loadings[,1:2]

# Então L' será

LT <- t(L)

# Já a matriz Psi, é dada por:

PSI <- matrix(rep(0,49),7,7)

diag(PSI) <- q51a1[["uniquenesses"]]

# E seja o vetor de médias de x1,...,x7=

XB <- c(mean(dados$X1),mean(dados$X2),mean(dados$X3),mean(dados$X4),
        mean(dados$X5),mean(dados$X6),mean(dados$X7))

# e seja o vetor xj:

XJ <- c(110,98,105,15,18,12,35)

# então f_j^chapeu será

FJH <-(solve(LT %*% solve(PSI) %*% L) %*% LT %*% solve(PSI))%*%(XJ-XB)
FJH
```

```
##          [,1]
## PC1  5.138916
## PC2 -2.285755
```

Já pelo método de regressão, teremos que $\hat{\mathbf{f}}_\mathbf{j} = \hat{\mathbf{L}}'\hat{\mathbf{\Sigma}}^{-1}(\mathbf{x_j} - \bar{\mathbf{x}})$, naturalmente com $\mathbf{\Sigma} = \mathbf{LL}' + \mathbf{\Psi}$. (pag. 516 J&W) [1]. Portanto, teremos:

```r
FJH2 <- LT%*%solve(L%*%LT+PSI)%*%(XJ-XB)
FJH2
```

```
##          [,1]
## PC1  5.055698
## PC2 -1.968998
```

Questão 52

Ex. 9.21 | Johnson & Wichern

Variância específica e loadings para PCA; utilizando a rotação Varimax:

```
fa2$uniquenesses
```

```
##            wind solar radiation              CO              NO             NO2
##       0.7614844       0.5168845       0.2902699       0.4049175       0.3651199
##              O3              HC
##       0.3083168       0.6302237
```

```
fa2$loadings
```

```
##
## Loadings:
##                 RC1    RC2
## wind            -0.124 -0.472
## solar radiation         0.691
## CO               0.701  0.468
## NO               0.763 -0.112
## NO2              0.766  0.220
## O3                      0.830
## HC               0.607
##
##                  RC1   RC2
## SS loadings    2.052 1.671
## Proportion Var 0.293 0.239
## Cumulative Var 0.293 0.532
```

Variância específica e loadings para método de máxima verossimilhança; utilizando a rotação Varimax:

```
emv$uniquenesses
```

```
##            wind solar radiation              CO              NO             NO2
##       0.9070224       0.8953343       0.2126417       0.4983564       0.6144170
##              O3              HC
##       0.0050000       0.9152467
```

```
emv$loadings
```

```
##
## Loadings:
##                 Factor1 Factor2
## wind            -0.176  -0.249
## solar radiation          0.319
## CO               0.797   0.391
## NO               0.692  -0.152
## NO2              0.602   0.152
## O3                       0.997
## HC               0.251   0.147
##
##                Factor1 Factor2
## SS loadings      1.573   1.379
## Proportion Var   0.225   0.197
## Cumulative Var   0.225   0.422
```

Questão 53

Ex. 9.22 | Johnson & Wichern

a) Escore dos fatores para m = 2 para:

Método regressivo:

```
##          Factor1     Factor2
##  [1,]  1.37594214 -0.22919418
##  [2,]  0.09278943 -0.78705235
##  [3,] -0.16999889 -0.61089179
##  [4,] -0.18592875  1.00054350
##  [5,] -0.44302587  0.10409148
##  [6,]  0.22402630  0.46948785
##  [7,]  1.65041657  1.01739742
##  [8,]  1.64112921  0.83604418
##  [9,] -0.05971006  0.29086104
## [10,]  0.42184189 -0.06745231
```

Método mínimos quadrados ponderados:

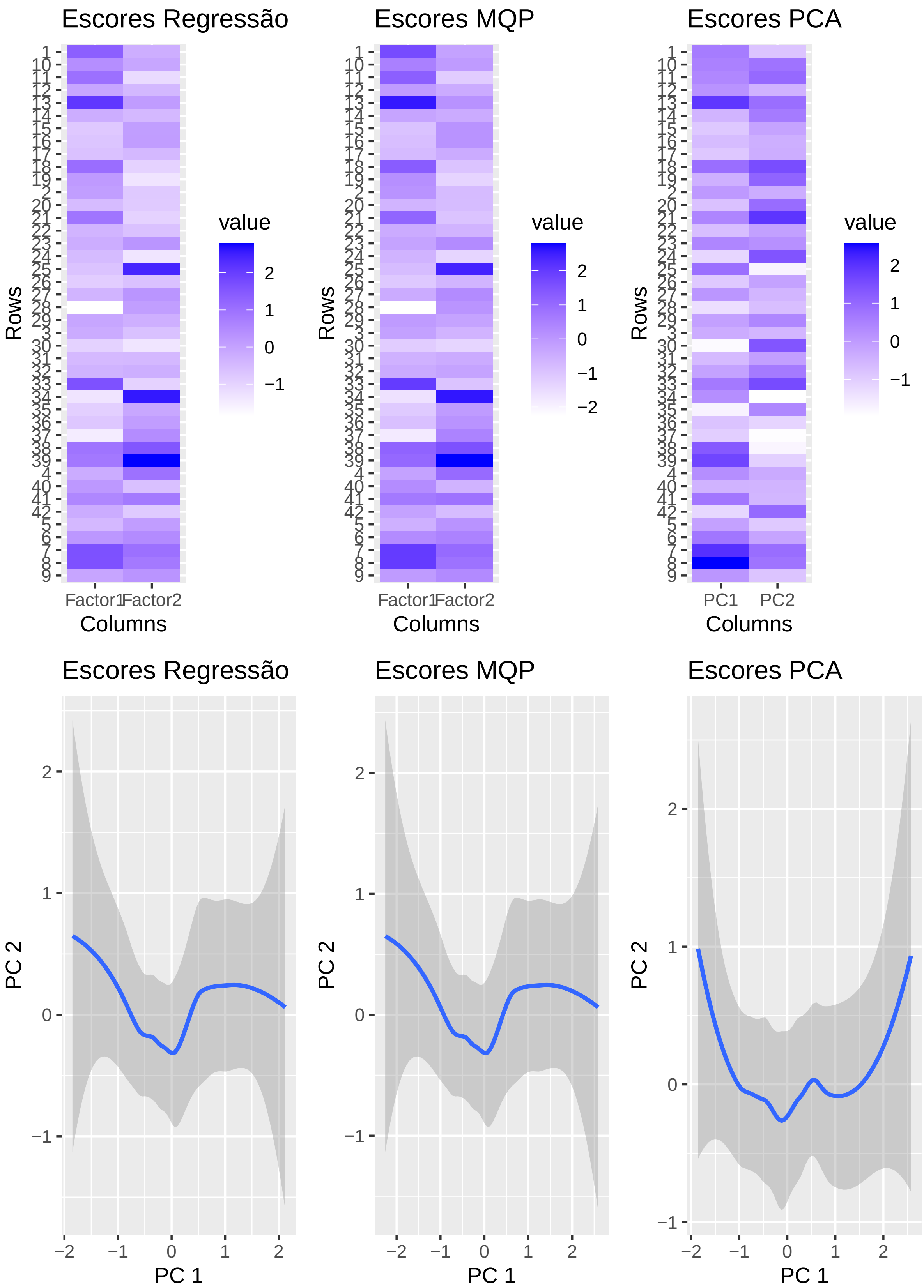
```
##          Factor1    Factor2
##  [1,]  1.67867247 -0.2303396
##  [2,]  0.11320466 -0.7909856
##  [3,] -0.20740150 -0.6139447
##  [4,] -0.22683619  1.0055436
##  [5,] -0.54049898  0.1046117
##  [6,]  0.27331584  0.4718341
##  [7,]  2.01353588  1.0224818
##  [8,]  2.00220515  0.8402222
##  [9,] -0.07284727  0.2923146
## [10,]  0.51465418 -0.0677894
```

b) Encontrar os escores dos fatores pelo método de análise de componente principal:

```
##              PC1        PC2
##  [1,]  0.61590690 -0.8186024
##  [2,]  0.03194177 -0.3601512
##  [3,] -0.34752031 -0.5448119
##  [4,]  0.24250269 -0.3029306
##  [5,] -0.12728999 -0.9194076
##  [6,]  0.72612096 -0.1927793
##  [7,]  2.03685873  0.8998212
##  [8,]  2.57309013  0.7773197
##  [9,]  0.09802174 -0.8173583
## [10,]  0.50663975  0.7880266
```

c) Comparar os três métodos:

Visualizando graficamente os escores:



Pela visualização dos gráficos, podemos perceber que os escores do método regressivo e de mínimos quadrados ponderados se assemelham bastante, enquanto os escores da PCA diferem um pouco desses últimos dois, mas ainda se mantendo numa escala parecida.

Questão 54

Ex. 9.23 | Johnson & Wichern

Variância específica para m=1, matriz = R

```
m1d$uniquenesses
```

```
##          wind solar radiation              CO              NO             NO2 
##     0.8689433       0.9012532       0.2903344       0.6667908       0.4204310 
##            O3              HC 
##     0.7538594       0.7616052
```

Variância específica para m=2, matriz = R

```
m2d$uniquenesses
```

```
##          wind solar radiation              CO              NO             NO2 
##     0.7614844       0.5168845       0.2902699       0.4049175       0.3651199 
##            O3              HC 
##     0.3083168       0.6302237
```

Variância específica para m=1, matriz = S

```
m1s$uniquenesses
```

```
##          wind solar radiation              CO              NO             NO2 
##     2.4693349       0.3820131       1.4619008       1.1755959      11.1845248 
##            O3              HC 
##    27.1325706       0.4768462
```

Variância específica para m=2, matriz = S

```
m2s$uniquenesses
```

```
##          wind solar radiation              CO              NO             NO2 
##   2.305046298     0.000511124     1.181578894     1.170665935    10.545075347 
##            O3              HC 
##   0.343002153     0.460801978
```

Sob qualquer métrica, os valores encontrados são diferentes. Mas com atenção especial para a variância específica onde é bem possível encontrar as diferenças. A diferença entre usar a matriz **R** e a matriz **S** fica bem explícita, e isso se dá por conta da escala das variáveis (uma das variáveis tem a escala bastante diferente das demais e "puxa" a variabilidade para ela).

Referências:

[1] JOHNSON, Richard A; WICHERN, Dean W. APPLIED MULTIVARIATE STATISTICAL ANALYSIS. 6ª Edição. Pearson, 2007.

[2] von Borries, George. Material de aula disponível no Aprender3; Notas de aula e códigos. Análise Multivariada 1. Universidade de Brasília, 2023.