



DEPARTAMENTO DE ESTATÍSTICA

03 maio 2023

## **Entrega 3 - Lista 4**

Prof. Dr. George von Borries

Análise Multivariada 1

Aluno: Bruno Gondim Toledo | Matrícula: 15/0167636

6. Exercício 32 da Lista 4 Suponha que um pesquisador padronizou os dados de um estudo através da transformação de Mahalanobis ( $\mathbf{Z} = \mathbf{XS}^{-1/2}$ ), em que  $\mathbf{S}$  é a matriz de variância-covariâncias amostrais. Seria razoável aplicar componentes principais nos dados transformados? Justifique sua resposta.

Sim, visto que esta transformação remove as correlações entre as variáveis e dá igual peso a cada uma delas na análise, considerando portanto a covariância entre as variáveis. Esta é uma técnica que costuma ser mais útil quando as variáveis estão apresentadas em escalas diferentes. As distâncias calculadas por esta transformada são associadas a correspondência entre elas, mas não representam um valor interpretável. No caso, se trata mais da intensidade entre elas do que uma distância real.

Como a transformação considera a matriz  $S^{-1/2}$  é a decomposição da matriz de variância-covariâncias amostrais  $S$  em sua forma de autovalores e autovetores; calculando a matriz de transformação que multiplica cada vetor de dados pelo inverso da raiz quadrada dos autovalores, a informação acaba se concentrando nos primeiros autovalores, o que permite uma redução de dimensionalidade buscado na análise de componentes principais.

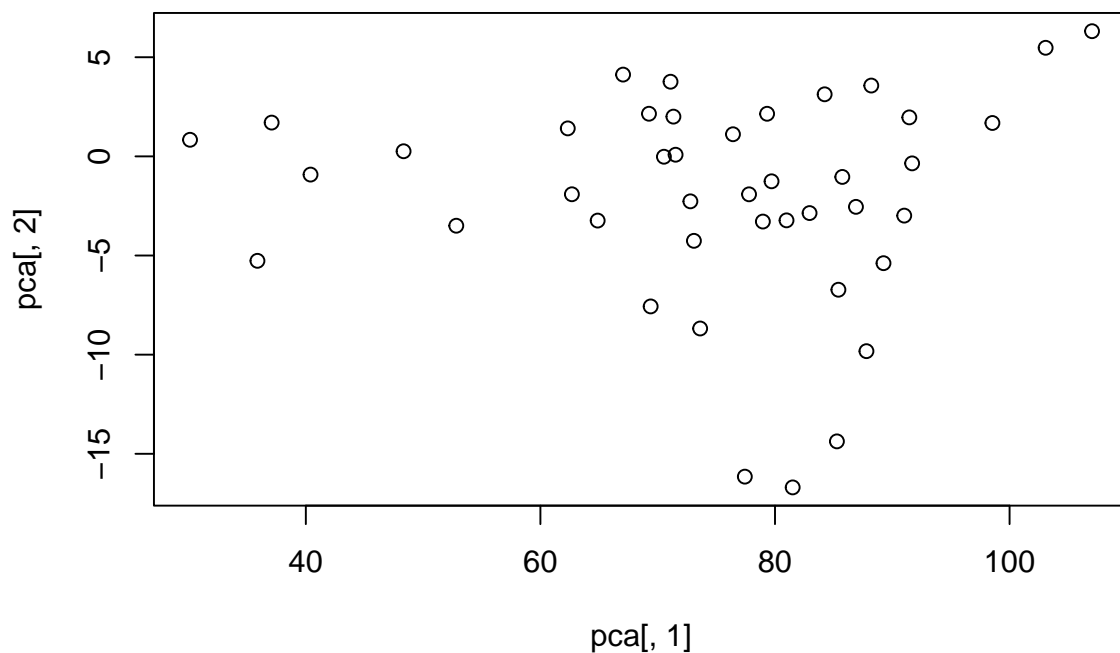
7. Exercício 37 da Lista 4 - Johnson e Wichern - Exercício 8.12. Dados no arquivo Air Pollution (T1-5.DAT). Os dados correspondem a 42 medidas de poluição do ar observadas na área de Los Angeles em um mesmo horário.

- (a) Resumir os dados em em menos de 7 dimensões (se possível) através de análise de componentes principais utilizando a matrix de covariâncias  $\mathbf{S}$  e apresentar suas conclusões.

```
## $values
## [1] 304.2578640 28.2761046 11.4644830 2.5243296 1.2795247 0.5287288
## [7] 0.2096157
```

Aqui, notamos que a maior parte da informação está concentrada na primeira dimensão. As dimensões 2 e 3 também são relevantes. Da quarta em diante, apenas uma fração da informação está contida.

Irei então trabalhar com 3 dimensões.

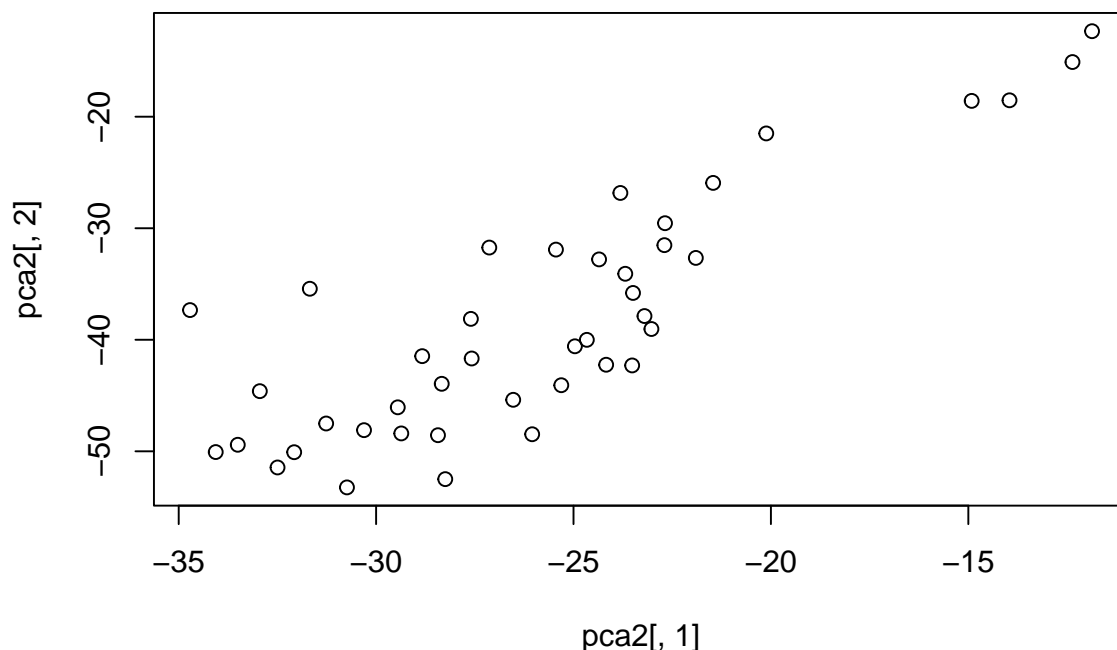


Notamos pelo gráfico acima que aparenta haver um *cluster* formado no primeiro quadrante quando analisamos as duas primeiras componentes. Apesar disso, nota-se pontos em praticamente todo o gráfico, exceto no terceiro quadrante, o que indica que estas duas componentes estão conseguindo explicar bem a variabilidade dos dados.

- (b) Resumir os dados em em menos de 7 dimensões (se possível) através de análise de componentes principais utilizando a matrix de correlações  $\mathbf{R}$  e apresentar suas conclusões.

```
## $values
## [1] 2.3367826 1.3860007 1.2040659 0.7270865 0.6534765 0.5366888 0.1558989
```

Utilizando a matriz de correlações, notamos que agora a informação está mais espalhada entre as dimensões. Portanto, provavelmente não será possível diminuir tanto o número de dimensões utilizando esta técnica em detrimento da técnica utilizada anteriormente.



Aqui, tentando analisar as duas primeiras componentes, notamos não só uma clusterização no terceiro quadrante, como uma menor dispersão dos dados no gráfico. O que leva a interpretação que estas duas dimensões provavelmente não estão sendo suficientes para explicar a variabilidade desse conjunto de dados.

(c) A escolha da matriz para análise faz alguma diferença? Explique.

Sim; conforme citado anteriormente, no caso da matriz de covariâncias, a informação acabou se concentrando melhor nas primeiras dimensões, permitindo assim uma redução maior sem muita perda de informação da variabilidade dos dados. No caso da matriz de correlações, a informação acabou ficando pulverizada entre as dimensões, não sendo possível por exemplo analisar visualmente com o gráfico a dispersão dos dados com fidedignidade.

(d) Os dados podem ser resumidos em 3 ou menos dimensões?

Utilizando o comando *prcomp*, obtemos o seguinte resultado:

```
## Importance of components:
##              PC1      PC2      PC3      PC4      PC5      PC6      PC7
## Standard deviation 17.443  5.31753  3.38592  1.58881  1.13116  0.72714  0.4578
## Proportion of Variance 0.873  0.08113  0.03289  0.00724  0.00367  0.00152  0.0006
## Cumulative Proportion 0.873  0.95408  0.98697  0.99421  0.99788  0.99940  1.0000
```

Onde vemos que 87.3% da variância está contida apenas na primeira componente. Conforme elucidado pelo Prof. von Borries em aula, não existe um valor exato para se alcançar na redução, mas levando ao limite, este valor (87.3%) já é um valor excelente, em outras palavras, a depender do tipo de análise, é possível trabalhar com apenas uma componente neste exemplo: A primeira!

Sendo um pouco mais conservador, ao manter a segunda componente, já temos 95.4% da variância explicada, então para quase qualquer análise esta seria uma redução perfeitamente aceitável.

Poderíamos parar por aqui e ficar com um trabalho mediano, porém se formos atentos e analisarmos o conjunto de dados

```
## # A tibble: 6 x 7
##       X1      X2      X3      X4      X5      X6      X7
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1      8     98      7      2     12      8      2
## 2      7    107      4      3      9      5      3
## 3      7    103      4      3      5      6      3
## 4     10     88      5      2      8     15      4
## 5      6     91      4      2      8     10      3
## 6      8     90      5      2     12     12      4
```

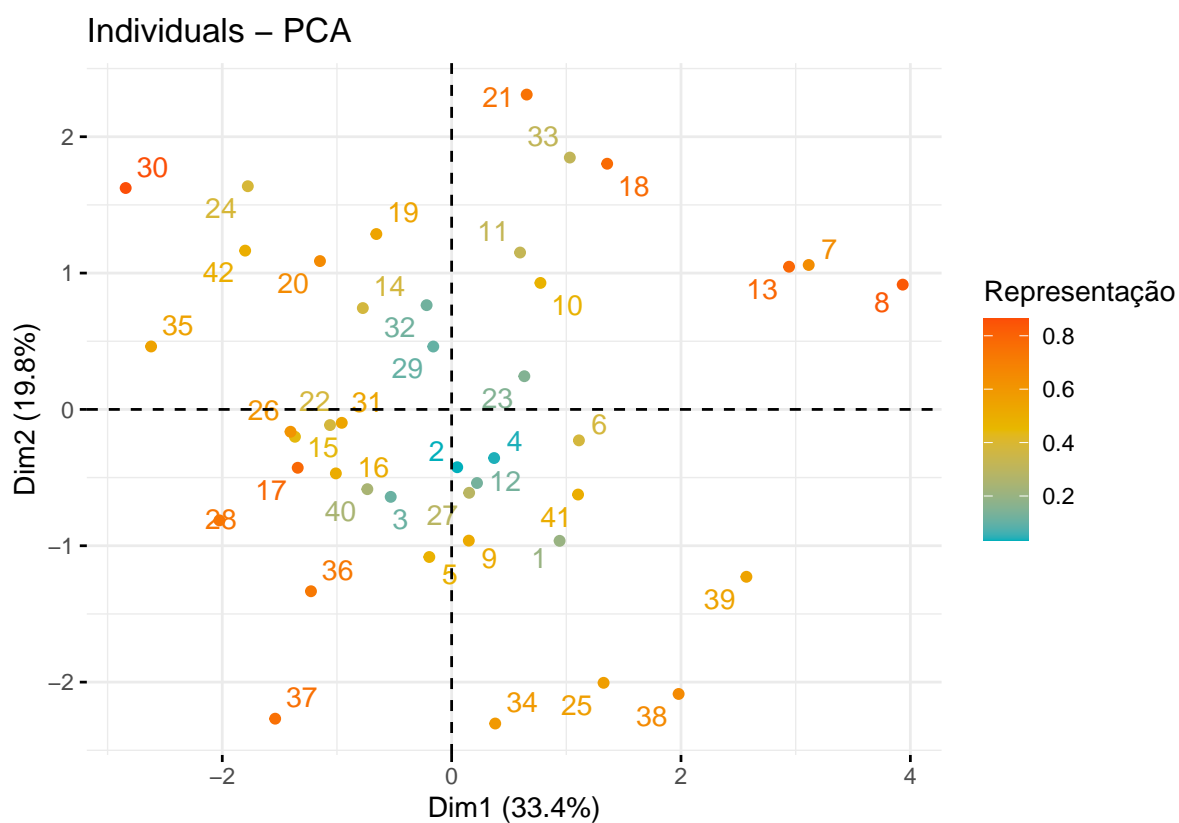
Notamos que os dados aparentam estar em escalas diferentes. Portanto, iremos utilizar a matriz de correlações, com a variável padronizada.

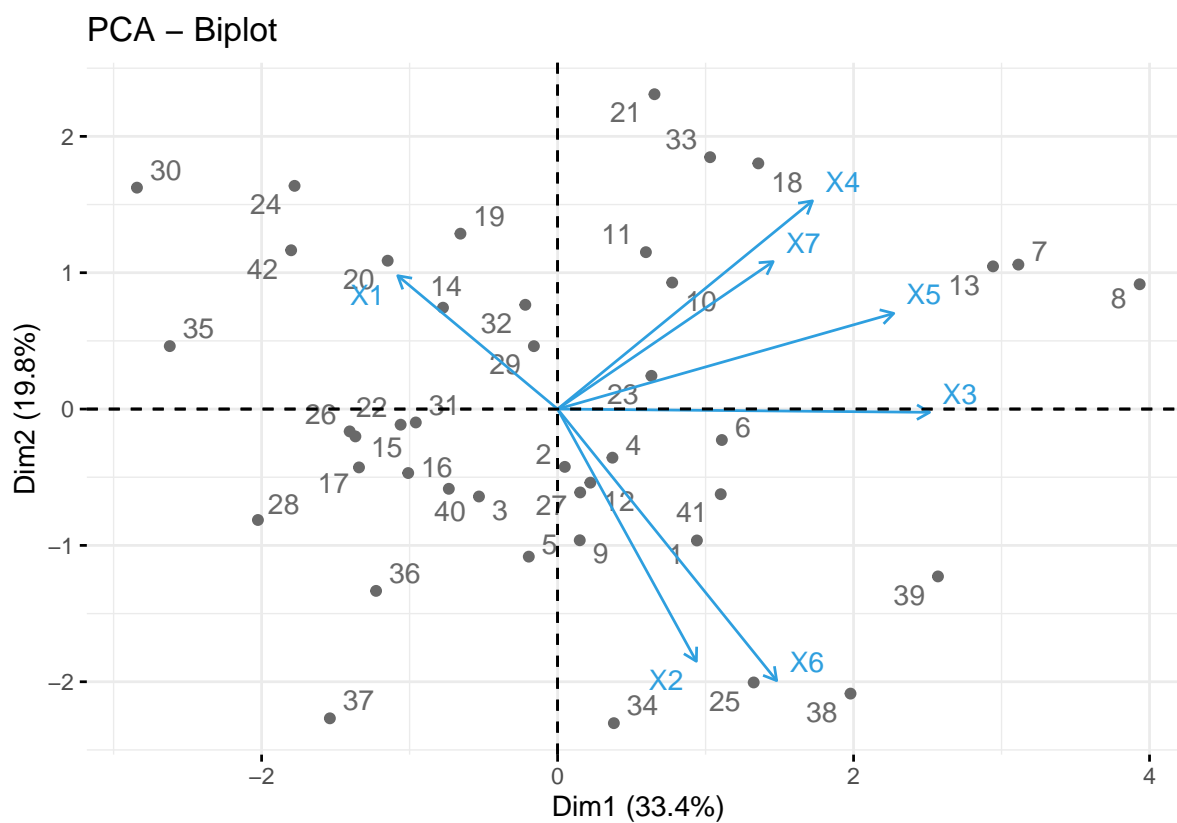
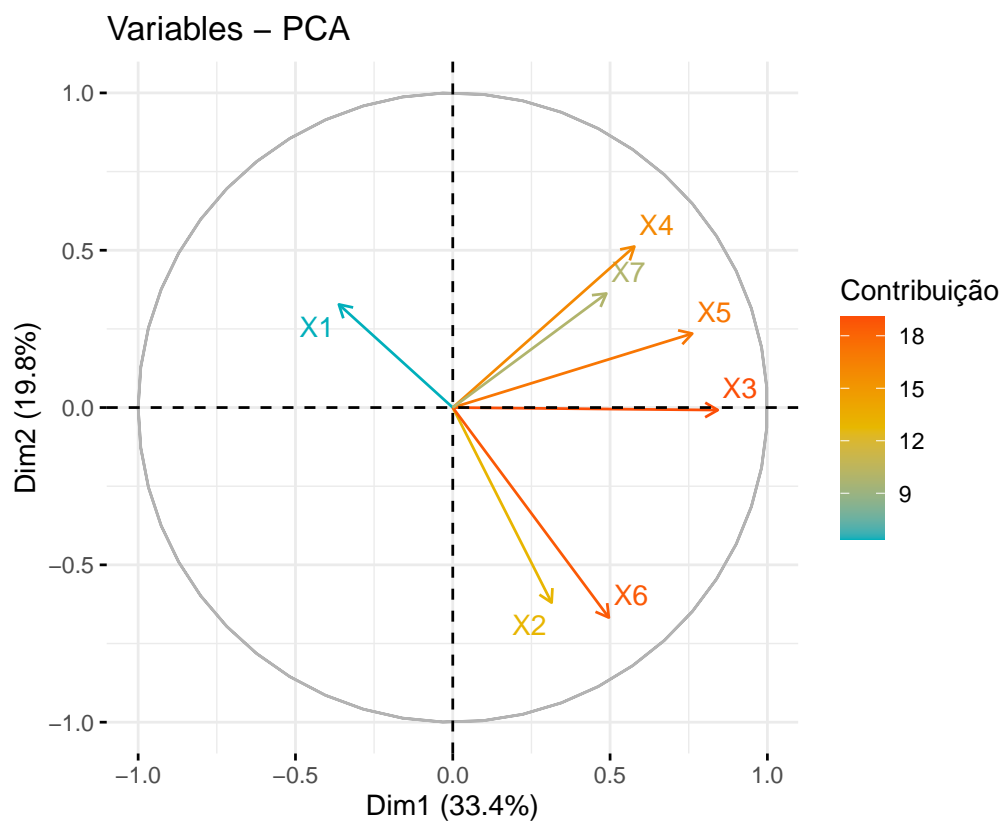
## Importance of components:

```
##              PC1      PC2      PC3      PC4      PC5      PC6      PC7
## Standard deviation 1.5287 1.1773 1.0973 0.8527 0.80838 0.73259 0.39484
## Proportion of Variance 0.3338 0.1980 0.1720 0.1039 0.09335 0.07667 0.02227
## Cumulative Proportion 0.3338 0.5318 0.7038 0.8077 0.90106 0.97773 1.00000
```

Por esta análise, notamos que até a terceira componente, acumulamos 70,38% da variação do conjunto. Portanto, notamos que na realidade, utilizar 3 ou menos dimensões iremos perder muito mais informação do que especulado anteriormente. Ainda assim, a depender da análise, este valor (70,38%) ainda pode ser excelente para algumas análises.

Abaixo algumas representações visuais desta análise [1]





Referências:

[1] SILVA, Adilane Ribeiro da. Análise de Componentes Principais (PCA): cálculo e aplicação no R. dezembro 17, 2020. Disponível em: <https://site.statplace.com.br/blog/analise-de-componentes-principais-pca-calculo-e-aplicacao-no-r/>