ME731 - Métodos em Análise Multivariada
– Análise de Componentes Principais II –

Prof. Carlos Trucíos
ctrucios@unicamp.br
ctruciosm.github.io

Instituto de Matemática, Estatística e Computação Científica,
Universidade Estadual de Campinas

Aula 11

Agenda I

Introdução

Introdução

Na aula anterior...

- O método de análise de componentes principais, bem como algumas propriedades deste foram apresentados.
- Foi desenvolvida a intuição e o procedimento teórico para obter as componentes principais (o problema resume-se a obter os autovalores e autovetores).
- Discutimos as limitações de utilizar ACP quando as variáveis estão em diferentes unidades.

Introdução

Na aula de hoje. . .

- Discutiremos como resolver o problem de ACP quando as variáveis estão em diferentes unidades.
- Apresentaremos outras propriedades.
- Discutremos como selecionar o número de componentes principais a serem retidos.
- Aprenderemos a interpretar as componentes

ACP vía matriz de correlação

ACP vía matriz de correlação

Lembre-se

Para obter as componentes, maximizamos

$$aSa = \sum_{i=1}^{p} a_i^2 s_i^2 + \sum_{i \neq j} a_i a_j s_{ij}.$$

Se, por exemplo s_1^2 for muito maior do que as outras variâncias,para maximizar a expressão acima, fazemos a_1 tão grande quanto pudermos (em casos extremos, a primeira componente será, basicamente, x_1)

ACP vía matriz de correlação

Lembre-se

Para obter as componentes, maximizamos

$$aSa = \sum_{i=1}^{p} a_i^2 s_i^2 + \sum_{i \neq j} a_i a_j s_{ij}.$$

Se, por exemplo s_1^2 for muito maior do que as outras variâncias,para maximizar a expressão acima, fazemos a_1 tão grande quanto pudermos (em casos extremos, a primeira componente será, basicamente, x_1)

Solução: Transformar todas as variáveis para terem $s_i^2 = 1 \quad \forall i$.

ACP vía matriz de correlação

Sejam $x_i, \cdots, x_n$ n observações p-dimensionais centradas. Seja $\mathbf{D} = diag(\mathbf{S})$, então

$$
\begin{aligned}
y_1 &= \mathbf{D}^{-1/2} x_1, \\
&\vdots \\
y_n &= \mathbf{D}^{-1/2} x_n,
\end{aligned}
$$

terão variância unitária.

ACP vía matriz de correlação

Note que $y_1, \cdots, y_n$ escrito em forma matricial é igual a

$$Y = \mathbf{D}^{-1/2} X.$$

Então, a matriz de covariância de $y_1, \cdots, y_n$ é igual à matriz de correlação de $x_1, \cdots, x_n$.

ACP vía matriz de correlação

Note que $y_1, \cdots, y_n$ escrito em forma matricial é igual a

$$Y = \mathbf{D}^{-1/2}X.$$

Então, a matriz de covariância de $y_1, \cdots, y_n$ é igual à matriz de correlação de $x_1, \cdots, x_n$.

$$\mathbf{S}_y = \frac{1}{n}\sum_{i=1}^{n} y_i y_i' = \frac{1}{n}\sum_{i=1}^{n} \mathbf{D}^{-1/2} x_i x_i' \mathbf{D}^{-1/2} = \mathbf{D}^{-1/2}\mathbf{S}_x\mathbf{D}^{-1/2} = \mathbf{R}_x$$

ACP vía matriz de correlação

Utilizando o algoritmo descrito na última aula (mas sobre $y_1, \cdots, y_n$):

Algoritmo

1. Calcular a matriz de covariância $\mathbf{S}_y$ (que é igual a $\mathbf{R}_x$)
2. Decompor $\mathbf{S}_y$ ($\equiv \mathbf{R}_x$) em autovalores e autovetores

$$\mathbf{S}M = M\Lambda$$

   em que Λ é uma matriz diagonal de autovalores ordenados de maior a menor e M a matriz de autovetores assoaciados aos autovalores
3. As componentes serão

$$\mathbf{z} = \underbrace{\mathbf{y}}_{\mathbf{D}^{-1/2}X} M$$

Aplicamos APC vía matriz de correlação!

ACP vía matriz de correlação

```
library(dplyr)
library(palmerpenguins)
dados <- penguins %>%
  select(ends_with("_mm"), "body_mass_g") %>% na.omit()
# Implementação "raíz"
M <- eigen(cor(dados))$vectors
z <- scale(dados, center = TRUE, scale = TRUE) %*% M
# Implementação "nutella"
componentes_prcomp <- prcomp(dados, scale. = TRUE)
componentes_princomp <- princomp(dados, cor = TRUE)
```

ACP vía matriz de correlação

```
head(z, 1)
```

```
##              [,1]        [,2]       [,3]      [,4]
## [1,] -1.840748 -0.04763243 -0.2324536 0.5231365
```

```
head(componentes_prcomp$x, 1)
```

```
##            PC1         PC2       PC3       PC4
## [1,] -1.840748 -0.04763243 0.2324536 0.5231365
```

```
head(componentes_princomp$scores, 1)
```

```
##         Comp.1     Comp.2     Comp.3   Comp.4
## [1,] -1.843445 0.04770222 -0.2327942 0.523903
```

ACP vía matriz de correlação

```
head(z, 1)
```

```
##              [,1]        [,2]       [,3]      [,4]
## [1,] -1.840748 -0.04763243 -0.2324536 0.5231365
```

```
head(componentes_prcomp$x, 1)
```

```
##              PC1         PC2       PC3       PC4
## [1,] -1.840748 -0.04763243 0.2324536 0.5231365
```

```
head(componentes_princomp$scores, 1)
```

```
##          Comp.1     Comp.2     Comp.3   Comp.4
## [1,] -1.843445 0.04770222 -0.2327942 0.523903
```

O que aconteceu com `componentes_princomp`?

ACP vía matriz de correlação

For practitioners

- Quando as variáveis originais estão em distintas unidades, é melhor utilizar a matriz de correlação.
- Quando as variáveis estão na mesma unidade, ambas opções são possíveis (dependerá se a diferença das variâncias das variáveis é algo informativo que precisa ser levado em conta ou não)
- Quando existe alta correlação positiva entre todas as variáveis, os coeficientes da combinação linear para obter a primeira componente serão todos positivos.
- Para facilitar a interpretação, alguns preferem zerar os coeficientes pequenos e aredondar os coefientes maiores. Isto só faz sentido se a diferença da proporção de variância pela componente formada com os autovetores originais e os "arredondados"for pequena.

Mais propriedades

Mais propriedades

- $Cor(x_j, z_k) = a_{jk}\sqrt{\lambda_k}/s_j$

Mais propriedades

- $Cor(x_j, z_k) = a_{jk}\sqrt{\lambda_k}/s_j$

Demostração:

$$Cov(x_j, z_k) = \frac{1}{n}\sum_{i=1}^{n} x_{ij} z_{kj} = \frac{1}{n} x_j' z_k = \frac{1}{n} \underbrace{[0, \cdots, 1, 0, \cdots, 0]}_{\delta'} x' x a_k = \delta' S a_k$$

Mais propriedades

- $Cor(x_j, z_k) = a_{jk}\sqrt{\lambda_k}/s_j$

Demostração:

$$Cov(x_j, z_k) = \frac{1}{n}\sum_{i=1}^{n} x_{ij}z_{kj} = \frac{1}{n}x_j' z_k = \frac{1}{n}\underbrace{[0, \cdots, 1, 0, \cdots, 0]}_{\delta'} x'xa_k = \delta' S a_k$$

Mas, sabemos que $Sa_k = \lambda_k a_k$.

Mais propriedades

- $Cor(x_j, z_k) = a_{jk}\sqrt{\lambda_k}/s_j$

Demostração:

$$Cov(x_j, z_k) = \frac{1}{n}\sum_{i=1}^{n} x_{ij}z_{kj} = \frac{1}{n}x_j' z_k = \frac{1}{n}\underbrace{[0, \cdots, 1, 0, \cdots, 0]}_{\delta'} x'xa_k = \delta' S a_k$$

Mas, sabemos que $Sa_k = \lambda_k a_k$.

$$Cov(x_j, z_k) = \delta'\lambda_k a_k = \lambda_k \underbrace{\delta' a_k}_{ajk} = \lambda_k a_{jk}$$

Mais propriedades

- $Cor(x_j, z_k) = a_{jk}\sqrt{\lambda_k}/s_j$

Demostração:

$$Cov(x_j, z_k) = \frac{1}{n}\sum_{i=1}^{n} x_{ij}z_{kj} = \frac{1}{n}x_j'z_k = \frac{1}{n}\underbrace{[0, \cdots, 1, 0, \cdots, 0]}_{\delta'}x'xa_k = \delta' Sa_k$$

Mas, sabemos que $Sa_k = \lambda_k a_k$.

$$Cov(x_j, z_k) = \delta'\lambda_k a_k = \lambda_k \underbrace{\delta' a_k}_{ajk} = \lambda_k a_{jk}$$

$$Cor(x_j, z_k) = \frac{Cov(x_j z_k)}{\sqrt{Var(x_j)}\sqrt{Var(z_k)}} = \frac{\lambda_k a_{jk}}{s_j\sqrt{\lambda_k}} = \sqrt{\lambda_k}a_{jk}/s_j$$

Mais propriedades

- $Var(z) = \Lambda$

Mais propriedades

- $Var(z) = \Lambda$

Demostração:

$$Var(z) = \frac{1}{n} z'z = \frac{1}{n} M'x'xM = M'SM = M'\Lambda M = \Lambda \underbrace{M'M}_{I} = \Lambda$$

Mais propriedades

- $Var(z) = \Lambda$

Demostração:

$$Var(z) = \frac{1}{n} z'z = \frac{1}{n} M'x'xM = M'SM = M'\Lambda M = \Lambda \underbrace{M'M}_{\mathbf{I}} = \Lambda$$

- $Cor(z) = \mathbf{I}$

Número de Componentes

Número de Componentes

1. **Gráfico de sedimentação:** Um dos critérios mais conhecidos. Consiste em fazer um gráfico de λ_i vs. i. A ideia é descartas as componentes com λs pequenos e aproximadamente do mesmo tamanho.

Número de Componentes

1. **Gráfico de sedimentação:** Um dos critérios mais conhecidos. Consiste em fazer um gráfico de λ_i vs. i. A ideia é descartas as componentes com λs pequenos e aproximadamente do mesmo tamanho.

Número de Componentes

2. Escolher as componentes que sejam necessárias até atinguir uma proporção pre-determinada de variância explicada pelas componentes.

Número de Componentes

2. Escolher as componentes que sejam necessárias até atinguir uma proporção pre-determinada de variância explicada pelas componentes.

3. **Critério de Kaisser:** Utilizamos tantas componentes quanto autovalores satisfazendo $\lambda_i > \bar{\lambda}$ (no caso das variáveis serem padronizadas, satisfazendo $\lambda_i > 1$).

Número de Componentes

2. Escolher as componentes que sejam necessárias até atinguir uma proporção pre-determinada de variância explicada pelas componentes.

3. **Critério de Kaisser:** Utilizamos tantas componentes quanto autovalores satisfazendo $\lambda_i > \bar{\lambda}$ (no caso das variáveis serem padronizadas, satisfazendo $\lambda_i > 1$).

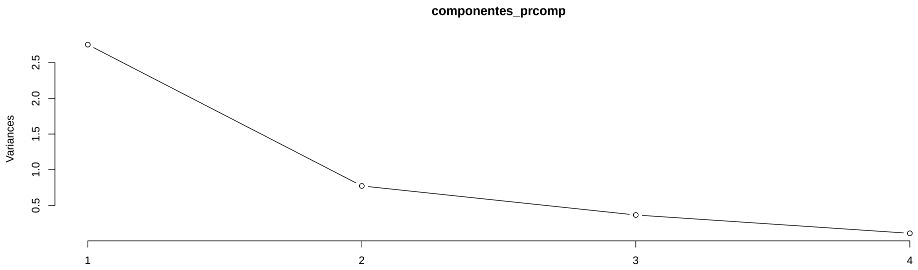
4. **Critério da razão de autovalores:** $r = \max_i \dfrac{\lambda_i}{\lambda_{i+1}}$.

Número de Componentes

2. Escolher as componentes que sejam necessárias até atinguir uma proporção pre-determinada de variância explicada pelas componentes.
3. **Critério de Kaisser:** Utilizamos tantas componentes quanto autovalores satisfazendo $\lambda_i > \bar{\lambda}$ (no caso das variáveis serem padronizadas, satisfazendo $\lambda_i > 1$).
4. **Critério da razão de autovalores:** $r = \max_i \dfrac{\lambda_i}{\lambda_{i+1}}$.
5. Critérios de Bai e Ng (2002) e Hallin e Liska (2007).

Número de Componentes

```r
screeplot(componentes_prcomp, type = "lines")
```



Número de Componentes

```
# Critério de Kaisser
componentes_prcomp$sdev^2 > mean(componentes_prcomp$sdev^2)
```

```
## [1]  TRUE FALSE FALSE FALSE
```

```
# Critério da razão de autovalores
componentes_prcomp$sdev[1:3]^2/componentes_prcomp$sdev[2:4]^2
```

```
## [1] 3.564654 2.115117 3.366471
```

Número de Componentes

```r
# Critérios de Bai e Ng (2002) e Hallin e Liska (2007)
POET::POETKhat(t(dados))
```

```
## $K1HL
## [1] 1
##
## $K2HL
## [1] 1
##
## $K1BN
## [1] 1
##
## $K2BN
## [1] 1
##
```

Número de Componentes

Todos os critérios sempre vão apontar para o mesmo número de componentes a serem retidos?

Número de Componentes

Todos os critérios sempre vão apontar para o mesmo número de componentes a serem retidos? Não!.

Interpretação

Interpretação

- **Proporção de variância explicada pelas componentes:** quanto da variabilidade total está sendo explicada pelas k primeiras componentes?

$$\frac{\lambda_1 + \cdots + \lambda_k}{\sum_{i=1}^{p} \lambda_i}$$

Interpretação

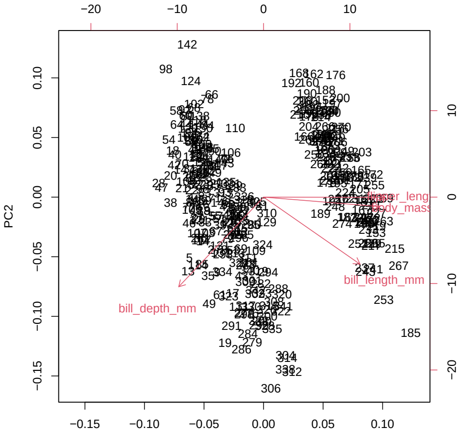
- **Proporção de variância explicada pelas componentes:** quanto da variabilidade total está sendo explicada pelas k primeiras componentes?

$$\frac{\lambda_1 + \cdots + \lambda_k}{\sum_{i=1}^{p} \lambda_i}$$

- **Autovetores:** são os "pesos" que cada variável tem na construção da componente. Variáveis com "pesos" maiores são mais importantes nessa componente. Os autovetores também nos dão o sinal da correlação entre a componente e as variáveis.

Interpretação

Biplot: Bi(dois)plot(gráfico), representa em um único gráfico informação a respeito das observações e das variáveis.



Interpretação

Biplot:

Interpretação

Biplot:

1. Escolher 2 componentes (geralmente as 2 primeiras) e projetar os x_i $(i = 1, \cdots, n)$ no plano gerado pelas componentes $(a_{11}x_{i1} + \cdots + a_{1p}x_{ip}, a_{21}x_{i1} + \cdots + a_{2p}x_{ip})$.

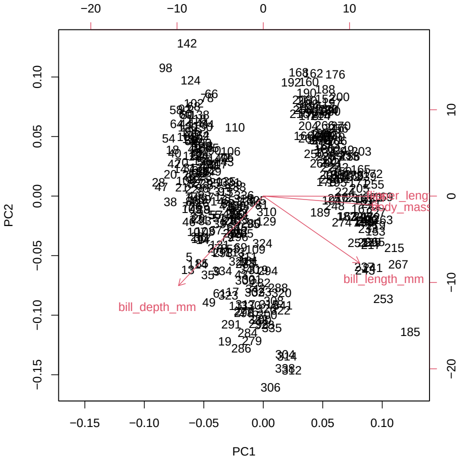
Interpretação

Biplot:

1. Escolher 2 componentes (geralmente as 2 primeiras) e projetar os x_i $(i = 1, \cdots, n)$ no plano gerado pelas componentes $(a_{11}x_{i1} + \cdots + a_{1p}x_{ip}, a_{21}x_{i1} + \cdots + a_{2p}x_{ip})$.
2. Representar (no mesmo plano) as variáveis originais. Isto é feito utilizando como coordenadas seu coeficiente de correlação com cada uma das componentes.

Interpretação

```
biplot(componentes_prcomp)
```



Observações:

- ACP trabalaha com correlações **lineares**, ante qualquer suspeita de relações não lineares uma opção é aplicar uma transformação ($\log(\cdot)$) para atinguir lineariedade.
- ACP gera novas variáveis (componentes) não correlacionadas (mas não podemos dizer que são independentes).

Observações:

- ACP trabalaha com correlações **lineares**, ante qualquer suspeita de relações não lineares uma opção é aplicar uma transformação ($\log(\cdot)$) para atinguir leariedade.
- ACP gera novas variáveis (componentes) não correlacionadas (mas não podemos dizer que são independentes).

Algumas alternativas ao ACP são:

1. Análise de Componentes Independentes
2. Análise de Componentes Condicionalmente não Correlacionados.
3. Análise de Componentes de Volatilidades
4. Autoencoder
5. t-SNE (t-distributed stochastic neighbor embedding)
6. UMAP (Uniform manifold approximation and projection)
7. etc

Referências

Referências

- Härdle, W. K., & Simar, L. (2019). Applied Multivariate Statistical Analysis. Fifth Editon. Springer Nature. Capítulo 11.
- Johnson, R. A., & Wichern, D. W. (2007). Applied multivariate statistical analysis. Sixth Edition. Pearson Prentice Hall. Capítulo 4.
- Mardia, K. V., Kent, J. T., & Bibby, J, M. (1979). Multivariate Analysis. Academic Press. Capítulo 8.
- Peña, D. (2002). Análisis de Datos Multivariantes. Mc Graw Hill. Capítulo 5.