UFSCar - UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
DEPARTAMENTO DE ESTATÍSTICA

ESTATÍSTICA MULTIVARIADA II

Atividade 2 - Componentes Principais

Grupo 12
Gabriel Carvalho Pereira
Luiz Paulo Dal Sasso

São Carlos - SP
Março de 2021

Conteúdo

1 Introdução

.

2 Descrição do conjunto de dados

Nosso conjunto de dados contém 25 observações que representam distintos lobos de duas determinadas regiões, o Ártico e as Montanhas Rochosas. Para cada observação, foram coletadas informações relacionadas à largura e ao comprimento de partes do crânio dos distintos animais em estudo. As variáveis disponíveis são:

- **Localização do lobo (Localização)**: Variável qualitativa dicotômica que descreve o local onde o determinado lobo vive. Seus possíveis valores são *"ar"- Ártico* ou *"rm" - Montanhas Rochosas*
- **Gênero do lobo (Gênero)**: Variável qualitativa dicotômica que descreve o gênero do determinado lobo. Seus possíveis valores são *"m"- Macho* ou *"f" - Fêmea*
- **Largura Zigomática (X0)**: Variável quantitativa contúnua que descreve a largura zigomática do determinado lobo, isto é, a largura facial máxima. Seu intervalo de variação está entre *125mm* e *152mm*.
- **Comprimento Palatal (X1)**: Variável quantitativa contúnua que descreve o comprimento palatal do determinado lobo, isto é, o comprimento da espinha nasal caudal do palatino ao próstio. Seu intervalo de variação está entre *107mm* e *128mm*.
- **Comprimento pós Palatal (X2)**: Variável quantitativa contúnua que descreve o comprimento pós palatal do determinado lobo. Seu intervalo de variação está entre *91mm* e *111mm*.
- **Largura palatinas fora dos primeiros molares superiores (X3)**: Variável quantitativa contúnua que descreve a largura palatina fora dos primeiros molares superiores do determinado lobo. Seu intervalo de variação está entre *16,5mm* e *19mm*.
- **Largura palatinas fora dos segundos molares superiores (X4)**: Variável quantitativa contúnua que descreve a largura palatina fora dos segundos molares superiores do determinado lobo. Seu intervalo de variação está entre *30,10mm* e *37,20mm*.

- **Largura entre os forames pós-glenóides (X5)**: Variável quantitativa contúnua que descreve a largura entre os forames pós-glenóides do determinado lobo. Seu intervalo de variação está entre *61,60mm* e *70,30mm*.

- **Largura interorbital (X6)**: Variável quantitativa contúnua que descreve a largura interorbital do determinado lobo. Seu intervalo de variação está entre *40,70* e *52,70mm*.

- **Menor largura da caixa intercraniana (X7)**: Variável quantitativa contúnua que descreve a menor largura da caixa intercraniana do determinado lobo. Seu intervalo de variação está entre *34,10mm* e *45,60mm*.

- **Comprimento da coroa do primeiro molar superior (X8)**: Variável quantitativa contúnua que descreve o comprimento da coroa do primeiro molar superior do determinado lobo. Seu intervalo de variação está entre *16,50mm* e *19mm*.

Tabela 1: Descrição das variáveis

Variável	**Descrição**
Localização	Montanha Rochosa ou Ártico
Gênero	Masculino ou Feminino
X_0	Largura zigomática
X_1	Comprimento palatal
X_2	Comprimento pós palatal
X_3	Largura palatinas fora dos primeiros molares superiores
X_4	Largura palatinas fora dos segundos molares superiores
X_5	Largura entre os forames pós-glenóides
X_6	Largura interorbital
X_7	Menor largura da caixa intercraniana
X_8	Comprimento da coroa do primeiro molar superior

3 Seleção das variáveis

Como visto na seção anterior, exceto as variáveis *localização* e *gênero* dos lobos, todas as outras são quantitativas contínuas, e, a partir delas será construída a matriz de correlações a seguir:

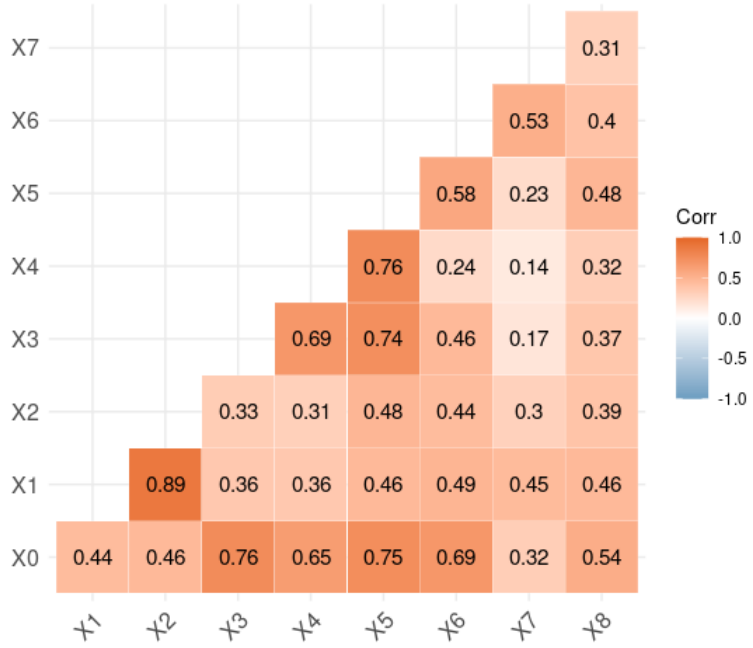

Figura 1: Representação da matriz de correlações

Dado que a matriz de correlações tem 1's em sua diagonal, a Figura 1 foi rearranjada de maneira a facilitar o entendimento da matriz de correlações. Pode-se perceber que a maior correlação é entre as variáveis *X1: Comprimento palatal* e *X2: Comprimento pós palatal*. Analisando-as conjuntamente,

Tabela 2: Medidas de posição das variáveis mais correlacionada

	Mínimo	**Mediana**	**Média**	**Máximo**	**Desvio Padrão**
X_1	107,0	117,0	117,4	128,0	6,061353
X_2	91,0	102,0	101,7	111,0	5,047442

Pela Tabela 2 temos algumas medidas de posição e o desvio padrão de ambas variáveis e pode-se perceber que tanto as medidas de posição como o desvio padrão são maiores para o *comprimento palatal* em comparação com o *comprimento pos palatal* dos lobos, sem a distinção por gênero ou localização.

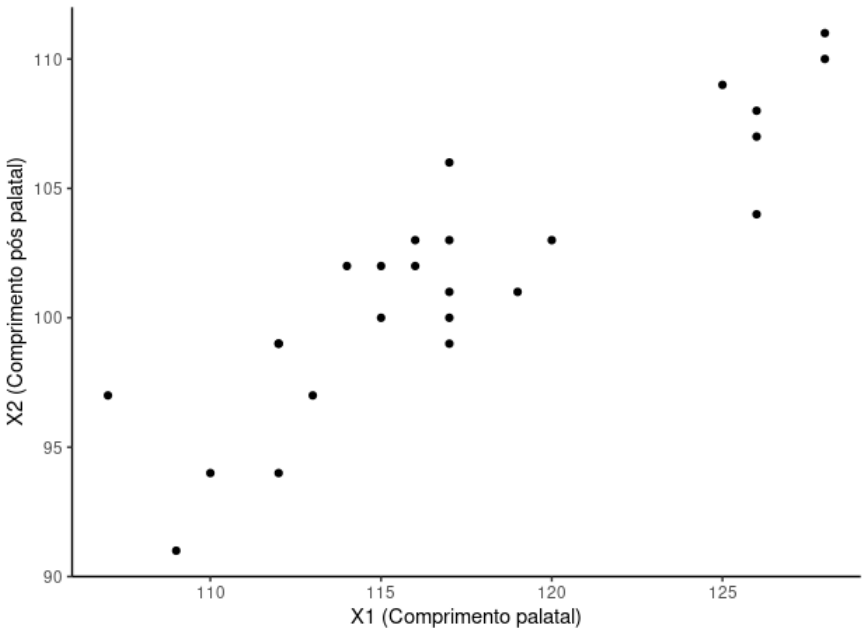

Figura 2: Gráfico de dispersão das variáveis mais correlacionadas

As variáveis possuem uma correlação linear crescente e bem definida, pode-se perceber pela Figura 2 que, no geral, quanto maior o *comprimento palatal*, maior será o *comprimento pos palatal* deste lobos.

4 Componentes Principais

Consideremos os seguintes dados das variáveis escolhidas na seção anterior,

Tabela 3: Valores assumidos pelas variáveis.

	X_1	X_2
1	126	104
2	128	111
3	126	108
4	125	109
5	126	107
6	128	110
7	116	102
8	120	103
9	116	103
10	117	99
11	115	100
12	117	106
13	117	101
14	117	103
15	119	101
16	115	102
17	117	100
18	114	102
19	110	94
20	112	94
21	109	91
22	112	99
23	112	99
24	113	97
25	107	97
Média	117,36	101,68
Variância	36,74	25,47666667

- **Variância total: 62,2166667**
- $\%$ **da variabilidade explicada por X1: 0,590517**
- $\%$ **da variabilidade explicada por X2: 0,409483**
- **Correlação: 0,8864389**

A coordenada das observações com respeito ao novo eixo X_1^* é uma combinação linear das coordenadas (antigas) do ponto com respeito aos eixos originais. Isto é,

$$X_1^* = X_1 cos(\theta) + X_2 sen(\theta)$$

e

$$X_2^* = -X_1 sen(\theta) + X_2 cos(\theta)$$

tal que X_1 e X_2 são as variáveis especificadas na Tabela 3. A variável X_1^* não explica toda a variabilidade dos dados, e a partir de uma segunda nova variável, X_2^*, explicará o máximo da variância que não foi explicada pela primeira, de tal maneira que o ângulo entre X_1 e X_1^* é o mesmo que entre X_2 e X_2^*.

Tabela 4: Ângulo θ e a $\%$ da variância total explicada utilizando tal ângulo.

Ângulo θ	$\%$
0	0.590517
10	0.7341434
20	0.8495286
30	0.9227556
40	0.9449919
50	0.9135557
60	0.8322386
70	0.7108485
80	0.5640271
90	0.409483

Podemos perceber pela Tabela 4 que o ângulo θ que maximiza a variância total explicada por X_1^* está entre 30^o e 50^o.

Utilizando o software R, geramos um vetor com todos os possíveis valores de θ entre 0 e 90 graus, com intervalos de 0,001 entre os ângulos. Após isso, calculamos para cada θ a porcentagem da variância que é explicada por X_1^* utilizando tal ângulo. Foi encontrado o valor de $\theta = 39,134^o$ como o valor que maximiza a variância explicada por X_1^*.

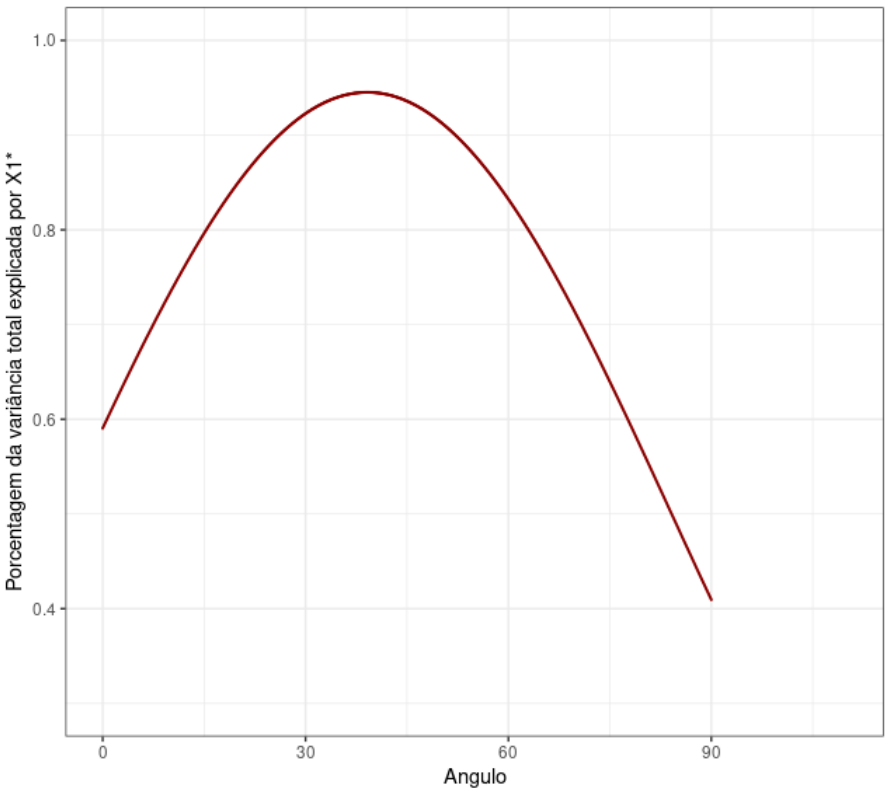

Figura 3: Gráfico da porcentagem da variância em função do ângulo entre os eixos.

Utilizando então o ângulo $\theta = 39,134^o$, temos as equações utilizadas para gerar as novas coordenadas a seguir:

$$X_1^* = 0,7736854X_1 + 0,6335699X_2$$

e

$$X_2^* = -0,6335699X_1 + 0,7736854X_2$$

Logo, os valores das variáveis selecionadas e transformadas estão na Tabela 5,

Tabela 5: Valores das variáveis

	X_1	X_2	X_1^*	X_2^*
1	126	104	163,3721	1,243675
2	128	111	169,339	5,414648
3	126	108	165,8948	4,34786
4	125	109	165,7495	5,754582
5	126	107	165,2642	3,571814
6	128	110	168,7083	4,638601
7	116	102	154,3503	5,99834
8	120	103	158,0852	4,251683
9	116	103	154,981	6,774386
10	117	99	153,2343	3,039525
11	115	100	152,3129	5,076923
12	117	106	157,6491	8,47185
13	117	101	154,4957	4,591618
14	117	103	155,757	6,14371
15	119	101	156,0478	3,330266
16	115	102	153,5743	6,629016
17	117	100	153,865	3,815571
18	114	102	152,7982	7,259691
19	110	94	144,6486	3,574023
20	112	94	146,2007	2,312672
21	109	91	141,9806	1,87656
22	112	99	149,3541	6,192904
23	112	99	149,3541	6,192904
24	113	97	148,8688	4,010135
25	107	97	144,2125	7,79419
Média	117,36	101,68	155,203924	4,89228588
Variância	36,74	25,47666667	58,80700122	3,409796019

As informações, das variáveis transformadas,

- **Variância total: 62,2166667**
- $\%$ **da variabilidade explicada por** X_1**: 0,9451951**
- $\%$ **da variabilidade explicada por** X_2**: 0,05480487**
- **Correlação: 0**

Para mostrar que a correlação entre as duas novas variáveis, X_1^* e X_2^*, é 0, utilizaremos do fato de que, sendo $< \cdot, \cdot >$ o produto interno e $|| \cdot ||$ a norma de um vetor, temos da álgebra linear que $\frac{<v_1, v_2>}{||v_1|| \cdot ||v_2||} = cos(\theta)$ em que θ é o ângulo entre v_1 e v_2. Consequentemente, podemos interpretar a $Cor(X_1^*, X_2^*)$ como o cosseno do ângulo entre essas duas variáveis. Dado que a correlação é igual a zero, tem-se que o vetor aleatório dessas duas variáveis é ortogonal ($\theta = \pi/2$).

Referências

[1] **The skull of Canis lupus**. Disponível em: http://www.naturalworlds.org/wolf/moretopics/wolf_skull.htm. World of the Wolf. Natural Worlds. Acesso em 14 de março de 2021

[2] MINGOTI, S. A. **Análise de Dados Através de Métodos de Estatística Multivariada uma abordagem aplicada**. Belo Horizonte: Editora UFMG, 2005.

[3] **Análise Multivariada.** Disponível em: https://www.ufjf.br/lates/files/2016/12/Conte%c3%bado-7-%e2%80%93-A_An%c3%a1lise-discriminante-AD.pdf. Acesso em 14/03/2021.

```r
library(ggcorrplot)
library(tidyverse)
library(dplyr)
library(doBy)
library(readxl)

d1 <- read_excel("d1 (2).xlsx")

d1$X3 = as.numeric(d1$X3)
d1$X4 = as.numeric(d1$X4)
d1$X5 = as.numeric(d1$X5)
d1$X6 = as.numeric(d1$X6)
d1$X7 = as.numeric(d1$X7)
d1$X8 = as.numeric(d1$X8)

corr <- cor(d1[-c(1,2)])

ggcorrplot(
           corr,
           type = "lower",
           outline.col = "white",
           colors = c("#6D9EC1", "white", "#E46726"),
           lab = T
           )

selected_data <- d1 %>% select(X1,X2)

summary(selected_data)

p <- ggplot(selected_data, aes(x=X1, y=X2)) +
  geom_point() +
  theme_classic()+
  labs(x = "X1 (Comprimento palatal)" , y = "X2 (Comprimento p s palatal)")
p

#medias e variancias das variaveis originais
media_x1 = mean(selected_data$X1)
media_x2 = mean(selected_data$X2)
var_x1 = var(selected_data$X1)
var_x2 = var(selected_data$X2)
```

```r
var_total = var_x1 + var_x2

#porcentagem da variancia total explicada por cada variavel
var_x1_pct = var_x1/var_total
var_x2_pct = var_x2/var_total

#vetor para encontrar theta que maximiza a vari?ncia da transforma??o
vec_theta = seq(0,90,0.1)
vec_pct = c()
for (num in 1:length(vec_theta)) {
  k = selected_data$X1 * cos(vec_theta[num]*pi/180) + selected_data$X2 *
    sin(vec_theta[num]*pi/180)
  vec_pct[num] = var(k)/var_total
}

#gr fico dos thetas
plot(vec_pct)

ggplot(data.frame(y = vec_pct, x = vec_theta), aes(x = vec_theta, y = vec_pct))
    +
  theme_bw() + geom_point(color = "darkred", size = 0.1) +
  ylim(.3,1) +
  xlim(0,110) +
  labs(x = "Angulo", y = "Porcentagem da vari ncia total explicada por X1*")

#theta que maximiza
max(vec_pct)ected_data$X1 * cos(vec_theta1[num]*pi/180) + selected_data$X2 *
    sin(vec_theta1[num]*pi/180)
  vec_pct1[num] = var(k)/var_total
}

#gr?fico dos thetas
plot(vec_pct1)

#theta que maximiza
max(vec_pct1)
theta_pct1 = cbind(vec_theta1,vec_pct1)
theta_pct1
theta_max = 39.134
var_max = 0.9451951
```

```
80
81 #x1*
82 x1_t = selected_data$X1 * cos(39.134*pi/180) + selected_data$X2 *
83   sin(39.134*pi/180)
84 x1_t
85
86 var(x1_t)
87 var(x1_t)/var_total
88
89 #x2*
90 x2_t =  -1*selected_data$X1 * sin(39.134*pi/180) + selected_data$X2 *
91   cos(39.134*pi/180)
92 var(x2_t)
93 var(x2_t)/var_total
94
95 cor(x1_t, x2_t)
96
97 transformed_data = cbind(x1_t, x2_t)
98 transformed_data
```

Listing 1: Códigos utilizados