Instituto de Matemática e Estatística
Universidade de São Paulo

Lista 4

MAE0330 - Análise Multivariada de Dados

Profª Lucia Pereira Barroso

Bruno Groper Morbin - nºUSP 11809875

Luigi Pavarini de Lima - nºUSP 11844642

São Paulo
12 outubro, 2022

Análise Descritiva

Verificamos que todas variáveis do conjunto de dados são qualitativas, e pela tabela a seguir, observa-se que a magnitude das variáveis é a mesma. Além disso, sabe-se que o valor de todas variáveis consiste à atribuição de notas para diferentes partes do corpo variando de 0 a 5.

Tabela 1: Medidas resumo das variáveis qualitativas - conjunto de dados 'BCAQ.xls'

	bel_coxas	bel_barriga	bel_estomago	bel_bumbum	bel_rosto	apa_coxas	apa_estomago	apa_bumbum	apa_rosto	apa_costelas
Mínimo	0,00	0,00	0,00	0,00	0,0	0,00	0,00	0,00	0,00	0,00
1º Quartil	0,00	1,00	0,00	0,00	0,0	1,00	0,75	0,00	1,00	1,00
Mediana	1,00	2,00	2,00	1,00	1,0	2,00	2,00	2,00	2,00	2,00
3º Quartil	3,00	4,00	3,00	3,00	4,0	3,00	4,00	3,00	4,00	4,00
Máximo	5,00	5,00	5,00	5,00	5,0	5,00	5,00	5,00	5,00	5,00
Média	1,81	2,31	1,93	1,60	1,8	1,99	2,20	1,88	2,26	2,09
Desvio Padrão	1,77	1,59	1,77	1,77	1,8	1,62	1,78	1,64	1,62	1,66
Observações	80,00	80,00	80,00	80,00	80,0	80,00	80,00	80,00	80,00	80,00

Portanto, concluímos que podemos considerar a matriz de covariâncias para a análise fatorial a seguir.

Método de estimação

Iremos decidir entre a escolha do método por máxima verossimilhança ou pelo método relacionado com componentes principais (PCA).

Primeiramente, para verificar se conjunto de dados é condizente com as suposições do método de máxima verossimilhança, observa-se o gráfico QQ-Plot com a distribuição Normal.

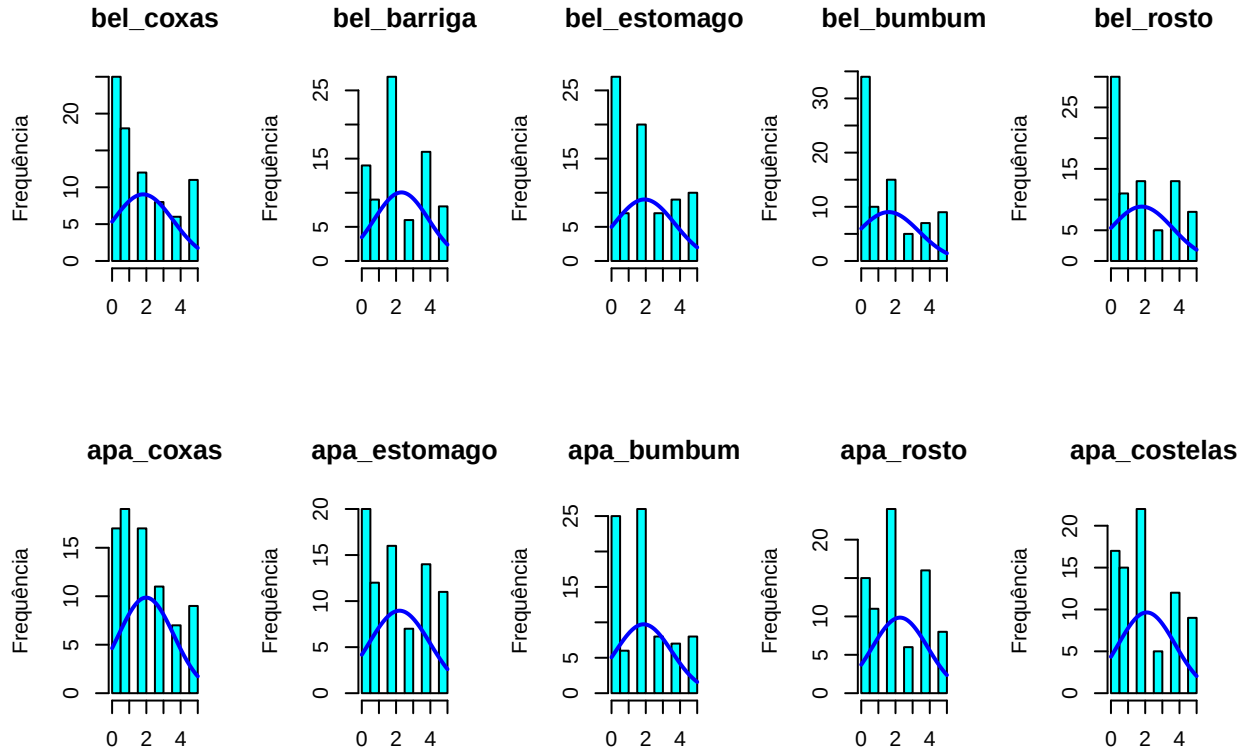


Figura 1: Histrograma das variáveis, com sobreposição da curva Normal.

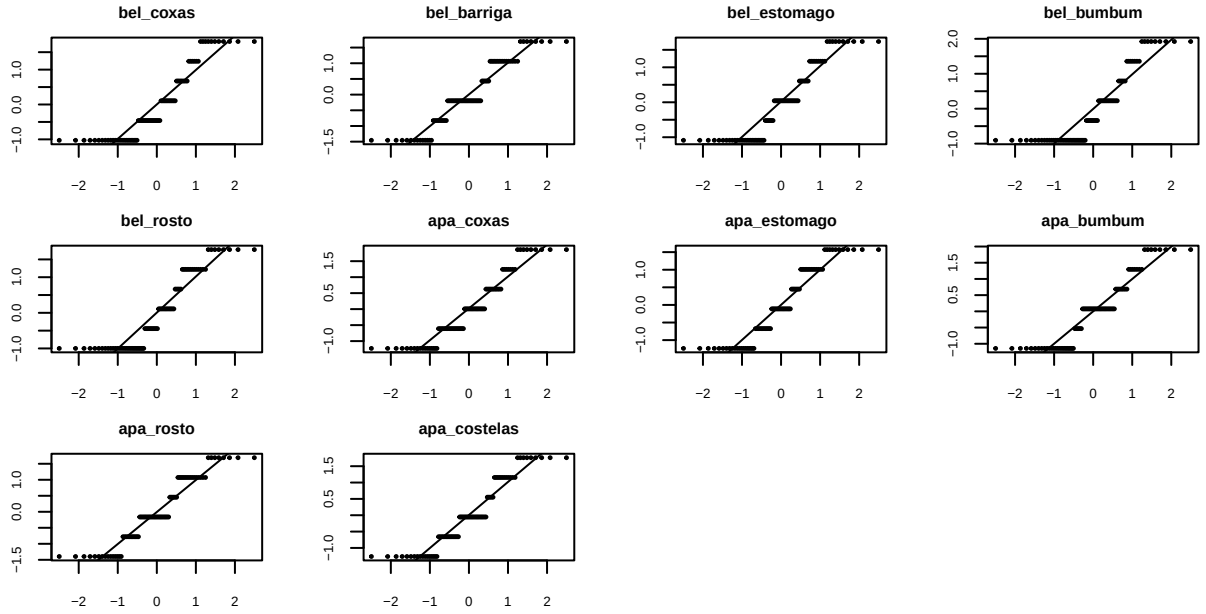


Figura 2: Gráfico Quantil-Quantil para identificação da hipótese de normalidade.

Pelos gráficos, já percebemos que as variáveis não são seguem distribuição Normal individualmente, o que implica que não se tem Normal multivariada. A seguir, realizamos o teste não paramétrico para inferir sobre a hipótese de normalidade multivariada para as variáveis em questão.

Tabela 2: Teste de Normal Multivariada por Mardia

Test	Statistic	p value	Result
Mardia Skewness	545.825997624532	1.15396488147671e-29	NO
Mardia Kurtosis	13.5679940347611	0	NO
MVN			NO

Resultando em p-valor $< 5\%$, o que nos leva a rejeitar a hipótese de normalidade multivariada.

Decidimos prosseguir então pelo método por PCA (também conhecido como *Principal Factor*) para Análise Fatorial.

Análise Fatorial

Viabilidade da análise

Avaliamos o critério de Kaiser para medir a adequação da análise fatorial para os dados em questão.

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = DADOS)
Overall MSA =  0.87
MSA for each item =
   bel_coxas  bel_barriga bel_estomago   bel_bumbum    bel_rosto    apa_coxas apa_estomago
        0.93         0.85         0.81         0.89         0.90         0.84         0.89
  apa_bumbum    apa_rosto apa_costelas
        0.81         0.89         0.94
```

Nesse sentido, como obtivemos um $MSA = 0.87$ considerado ótimo pela tabela de interpretação, e além disso, cada variável apresentou MSA superior a 0.8, portanto está confirmada a viabilidade da análise fatorial para esse conjunto de dados. Caso uma variável apresentasse MSA baixo, tiraríamos a mesmo, continuando com a análise das demais.

Tabela 3: Interpretação da KMO.

KMO	Interpretação
0,90 - 1,00	Excelente
0,80 - 0,90	Ótimo
0,70 - 0,80	Bom
0,60 - 0,70	Regular
0,50 - 0,60	Ruim
0,00 - 0,50	Inadequado

Rotação

Consideramos usar a rotação VARIMAX já que a mesma busca a variabilidade máxima das cargas fatoriais, o que corrobora com a caracterização e interpretação de quais variáveis são explicadas por cada fator. Ou seja, esse tipo de rotação ressalta as diferenças entre as variáveis de acordo com os fatores. Além disso, a rotação VARIMAX mantém a ortogonalidade dos fatores.

Número de fatores

Scree-Plot

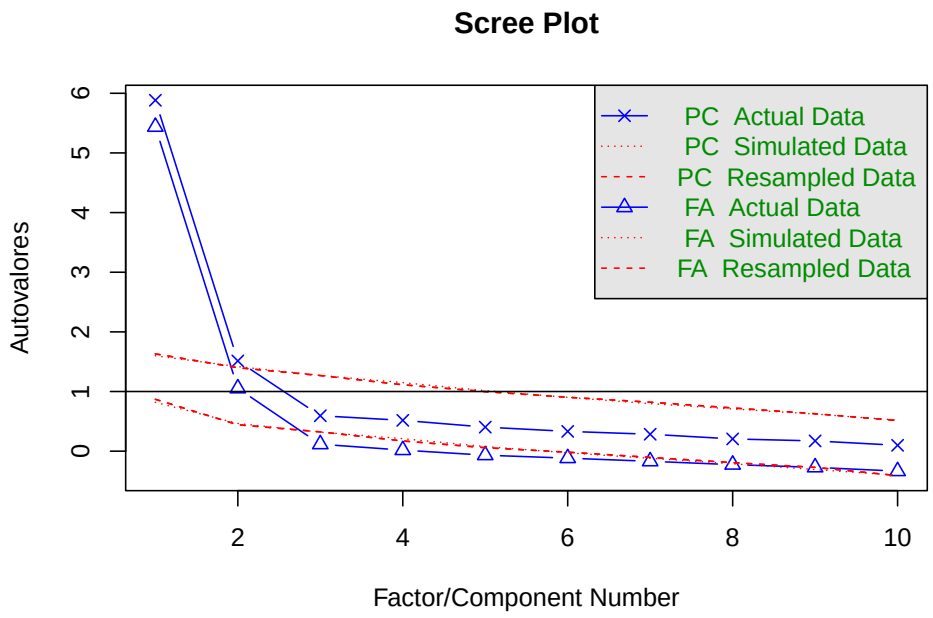


Figura 3: Scree-Plot

```
Parallel analysis suggests that the number of factors =  2  and the number of components =  2
```

Pelo gráfico acima, consideramos razoável a priori a escolha de 2 fatores, pois a partir dessa quantidade o incremento entre autovalores não parece ser relevante.

4 Fatores

```
> fac0 <- fa(DADOS, nfactors = 4, rotate = "varimax", covar = T, fm = "pa")
```

```
> fac0$Vaccounted
```

```
                           PA1       PA2        PA3        PA4
SS loadings           10.3452544 8.8472919 2.00329475 0.67937382
Proportion Var         0.3566804 0.3050341 0.06906895 0.02342323
Cumulative Var         0.3566804 0.6617145 0.73078342 0.75420665
Proportion Explained   0.4729213 0.4044437 0.09157829 0.03105678
Cumulative Proportion  0.4729213 0.8773649 0.96894322 1.00000000
```

3 Fatores

```
> fac1 <- fa(DADOS, nfactors = 3, rotate = "varimax", covar = T, fm = "pa")
```

```
> fac1$Vaccounted
```

```
                          PA1       PA2        PA3
SS loadings           9.6214438 8.8629411 2.75372510
Proportion Var        0.3317251 0.3055736 0.09494205
Cumulative Var        0.3317251 0.6372987 0.73224075
Proportion Explained  0.4530273 0.4173131 0.12965961
Cumulative Proportion 0.4530273 0.8703404 1.00000000
```

2 Fatores

```
> fac2 <- fa(DADOS, nfactors = 2, rotate = "varimax", covar = T, fm = "pa")
```

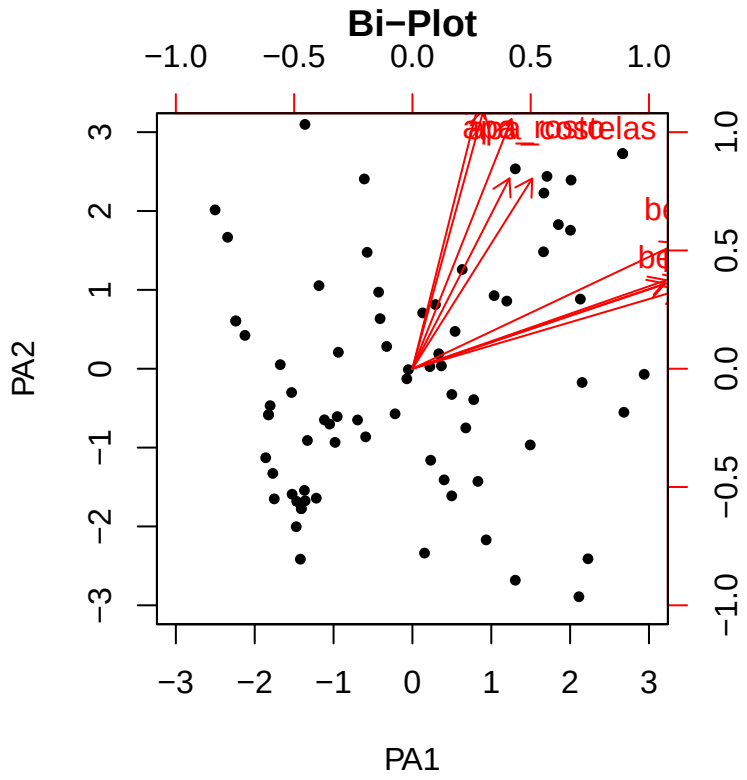
```
> fac2$Vaccounted
```

```
                           PA1       PA2
SS loadings           10.7743425 8.9075089
Proportion Var         0.3714743 0.3071102
Cumulative Var         0.3714743 0.6785846
Proportion Explained   0.5474253 0.4525747
Cumulative Proportion  0.5474253 1.0000000
```

Decisão

Por fim, consideramos razoável a escolha de 2 fatores uma vez que tanto o Scree-Plot quanto a proporção acumulada da variância explicada que já se faz acima de 80% considerando os diferentes cenários de número de fatores analisados.

Visualização



Interpretação dos fatores

```
> fac2$loadings
```

```
Loadings:
             PA1   PA2
bel_coxas    1.423 0.477
bel_barriga  1.410 0.415
bel_estomago 1.405 0.664
bel_bumbum   1.341 0.468
bel_rosto    1.335 0.450
apa_coxas    0.354 1.419
apa_estomago 0.524 1.319
apa_bumbum   0.376 1.368
apa_rosto    0.513 1.006
apa_costelas 0.633 1.005

                 PA1   PA2
SS loadings    10.774 8.908
Proportion Var  1.077 0.891
Cumulative Var  1.077 1.968
```

Observando as cargas fatoriais da saída do código acima, temos então a seguinte descrição dos fatores:

- **Fator 1 (PA1):** Representa as variáveis relacionadas à nota de frequência de *Beliscar* sendo elas atribuídas à *Coxa, Barriga, Estômago, Bumbum e Rosto*.

- **Fator 1 (PA2):** Representa as variáveis relacionadas à nota de frequência de *Apalpar* sendo elas atribuídas à *Coxa, Estômago, Bumbum, Rosto e Costelas*.

Escores

Tabela 4: Escores da análise

	PA1	PA2		PA1	PA2
1	2.6661488	2.7280812	41	-1.3314662	-0.9095208
2	2.6661488	2.7280812	42	-1.4092435	-1.7742530
3	-2.2403310	0.6056542	43	-1.5322428	-0.3015481
4	1.6623100	1.4824444	44	-1.8247872	-0.5842994
5	-0.4286213	0.9708122	45	-0.2198876	-0.5719266
6	1.8510703	1.8276144	46	-1.8607704	-1.1288269
7	-1.5256908	-1.5913589	47	-0.9539468	-0.6073064
8	0.9355536	-2.1703325	48	0.4986478	-1.6122578
9	-1.3628318	3.0984121	49	-3.4869623	4.1755150
10	0.1299505	0.7079511	50	-1.8048138	-0.4685402
11	-0.9809165	-0.9338443	51	2.1296824	0.8828520
12	-0.9393633	0.2085921	52	-1.4208180	-2.4153631
13	2.1114038	-2.8914503	53	0.4996915	-0.3261923
14	0.2922567	0.8137473	54	0.1534828	-2.3378456
15	3.9917566	0.1654085	55	-1.3692965	-1.5427346
16	0.6765476	-0.7505087	56	-2.1238837	0.4227601
17	2.0042532	1.7562557	57	0.3678356	0.0366902
18	-0.0513277	-0.0119352	58	-1.7703093	-1.3275225
19	2.6821578	-0.5522929	59	-1.0257179	3.5965415
20	3.4972363	0.3481740	60	2.1520566	-0.1743920
21	0.2315809	-1.1604719	61	0.4022867	-1.4093524
22	3.7381091	-0.5674385	62	-1.1849582	1.0536258
23	0.5407116	0.4735708	63	-2.3436587	1.6670620
24	1.1964767	0.8576021	64	0.3340191	0.1896813
25	0.7774054	-0.3922605	65	-0.6101741	2.4065879
26	2.0120952	2.3927132	66	2.9393336	-0.0707287
27	1.3049431	2.5350021	67	-1.1160812	-0.6482801
28	0.6316618	1.2583349	68	-0.5743870	1.4764284
29	1.4942117	-0.9668411	69	-0.5929867	-0.8639344
30	-1.3601327	-1.6747833	70	-2.2403310	0.6056542
31	1.3049793	-2.6820105	71	-0.6956614	-0.6498673
32	0.2209134	0.0266807	72	-1.2192900	-1.6417797
33	1.6676481	2.2273409	73	-0.4098176	0.6353674
34	0.8303299	-1.4287341	74	2.2266378	-2.4102261
35	1.7065819	2.4394028	75	-1.4732544	-2.0033610
36	-1.7520325	-1.6504880	76	-1.4092435	-1.7742530
37	1.0359919	0.9271475	77	-2.5028365	2.0139047
38	-1.6760866	0.0517332	78	-1.0499485	-0.7020797
39	-0.0713012	-0.1276944	79	-1.4674671	-1.6828060
40	-1.8247872	-0.5842994	80	-0.3264432	0.2825148

Código

```
library(tidyverse)
library(kableExtra)
library(knitr)
library(readxl)

DADOS <- read_excel("BCAQ.xls")

names<-colnames(DADOS)

par(mfrow= c(2,5))
for (i in names){
  h<-hist(DADOS[[i]],main = i,xlab = '',col="cyan", ylab = "Frequência")
  xfit<-seq(min(DADOS[[i]]),max(DADOS[[i]]),length=length(DADOS[[1]]))
  yfit<-dnorm(xfit,mean=mean(DADOS[[i]]),sd=sd(DADOS[[i]]))
  yfit <- yfit*diff(h$mids[1:2])*length(DADOS[[i]])
  lines(xfit, yfit, col="blue", lwd=2)
}

par(mar = c(2, 2, 2, 2))
par(mfrow= c(3,4))
for (i in 1:length(DADOS)){
  qqnorm(scale(DADOS[[i]]), pch=19, main=colnames(DADOS)[i], ylab="Variável Padronizada",\
         xlab="Normal Padrão", cex.lab=0.6, cex=0.3, cex.main=0.8, cex.axis=0.6)
  qqline(rnorm(10000),rnorm(10000))
}

library(MVN)
options(knitr.kable.NA = '')
kbl(mvn(DADOS,mvnTes="mardia")$multivariateNormality,
    booktabs = T,
    caption = "Teste de Normal Multivariada por Mardia",
    format.args = list(big.mark=".", decimal.mark=",")
    )%>%
    kable_styling(latex_options = c("HOLD_position"), position="center")

library(psych)
KMO(DADOS)

fa.parallel(DADOS,
            fm="pa", # método de Componetes Principais
            fa="both", # Análise Fatorial
            main="Scree Plot", ylab="Autovalores")

fac0 <- fa(DADOS, nfactors=4, rotate="varimax", covar=T, fm="pa")
fac0$Vaccounted

fac1 <- fa(DADOS, nfactors=3, rotate="varimax", covar=T, fm="pa")
fac1$Vaccounted

fac2 <- fa(DADOS, nfactors=2, rotate="varimax", covar=T, fm="pa")
fac2$Vaccounted

biplot(fac2,cex = 0.2, main="Bi-Plot")

fac2$loadings
```

```
fac2$scores
```