Instituto de Matemática e Estatística
Universidade de São Paulo

Lista 3

MAE0330 - Análise Multivariada de Dados

Profª Lucia Pereira Barroso

Bruno Groper Morbin - *nºUSP 11809875*

Luigi Pavarini de Lima - *nºUSP 11844642*

São Paulo
30 setembro, 2022

1 Análise dos dados originais

1.1 Dados Nominais

As duas primeiras variáveis do conjunto de dados são do tipo nominal, e apresentaram a seguinte composição:

Tabela 1: Medidas resumo para cada combinação Localidade-Variedade com dados originais (não padronizados)

	Localidade 1 : Variedade Arkan							
	Área B	Perímetro B	Comprimento B	Largura B	Área A	Perímetro A	Comprimento A	Largura A
Mínimo	2.543,00	203,00	78,00	38,00	2.569,0	204,00	79,00	40,00
Mediana	2.773,00	216,00	88,00	42,00	2.853,0	215,50	88,00	43,50
Máximo	3.074,00	232,00	95,00	71,00	3.228,0	232,00	101,00	50,00
Média	2.765,44	215,92	87,39	42,67	2.863,0	217,14	87,69	43,39
Desvio Padrão	155,79	6,97	4,61	5,16	174,5	7,65	4,54	2,03
Observações	36,00	36,00	36,00	36,00	36,0	36,00	36,00	36,00

	Localidade 1 : Variedade Arthur							
	Área B	Perímetro B	Comprimento B	Largura B	Área A	Perímetro A	Comprimento A	Largura A
Mínimo	2.550,00	205,00	82,00	41,00	2.343,00	198,00	82,00	38,00
Mediana	2.922,50	216,00	88,00	45,00	2.714,50	215,00	86,50	42,00
Máximo	3.253,00	233,00	127,00	86,00	3.087,00	230,00	99,00	59,00
Média	2.920,67	216,22	88,81	47,67	2.746,03	214,75	87,56	42,81
Desvio Padrão	172,22	7,28	7,65	10,36	194,85	7,74	3,76	3,67
Observações	36,00	36,00	36,00	36,00	36,00	36,00	36,00	36,00

	Localidade 2 : Variedade Arkan							
	Área B	Perímetro B	Comprimento B	Largura B	Área A	Perímetro A	Comprimento A	Largura A
Mínimo	2.601,00	212,00	84,00	38,00	2.701,00	210,00	85,00	39,00
Mediana	2.975,50	225,50	93,00	43,00	3.072,00	227,00	93,00	44,00
Máximo	3.772,00	246,00	103,00	75,00	3.682,00	254,00	104,00	50,00
Média	3.012,00	225,42	92,92	43,44	3.093,06	227,20	93,54	44,40
Desvio Padrão	263,96	8,44	4,28	5,28	233,94	9,12	4,08	2,44
Observações	50,00	50,00	50,00	50,00	50,00	50,00	50,00	50,00

	Localidade 2 : Variedade Arthur							
	Área B	Perímetro B	Comprimento B	Largura B	Área A	Perímetro A	Comprimento A	Largura A
Mínimo	2.546,00	207,00	82,00	37,00	2.479,0	206,00	84,0	37,00
Mediana	3.069,50	220,00	89,00	46,00	2.860,5	218,00	90,0	43,00
Máximo	3.588,00	239,00	100,00	53,00	3.409,0	238,00	101,0	51,00
Média	3.077,06	220,48	90,10	45,76	2.866,3	219,24	90,8	43,30
Desvio Padrão	230,63	7,41	4,02	2,87	203,3	7,42	3,8	2,69
Observações	50,00	50,00	50,00	50,00	50,0	50,00	50,0	50,00

1.2 Medidas Resumo

Inicialmente, observamos a base de dados e avaliamos descritivamente as variáveis quantitativas afim de decidir se a análise será feita sobre a matriz de convariâncias ou de correlações.

Tabela 2: Medidas resumo das variáveis qualitativas - conjunto de dados 'trigos.xls'

	Área B	Perímetro B	Comprimento B	Largura B	Área A	Perímetro A	Comprimento A	Largura A
Mínimo	2.543,00	203,00	78,00	37,00	2.343,00	198,00	79,00	37,00
1º Quartil	2.779,00	213,00	86,75	42,00	2.738,75	213,00	87,00	42,00
Mediana	2.949,00	219,00	90,00	44,00	2.892,50	220,00	90,00	43,00
3º Quartil	3.114,00	225,25	93,00	46,00	3.050,50	226,00	93,00	45,00
Máximo	3.772,00	246,00	127,00	86,00	3.682,00	254,00	104,00	59,00
Média	2.960,19	220,07	90,08	44,84	2.906,35	220,17	90,27	43,53
Desvio Padrão	243,59	8,49	5,52	6,42	240,83	9,31	4,71	2,78
Observações	172,00	172,00	172,00	172,00	172,00	172,00	172,00	172,00

Com isso, seguimos para separar as variáveis quantitativas do banco de dados em uma nova variável. Em seguida,considerando os níveis descritivos printados acima e afim de evitar qualquer desequilíbrio de forma que uma variável se sobressaia a outra, já que observamos escalas bem diferentes entre as variáveis (como destacado nos desvios padrão), decidimos padronizar a base e consequentemente trabalharemos a seguir com a matriz de correlação.

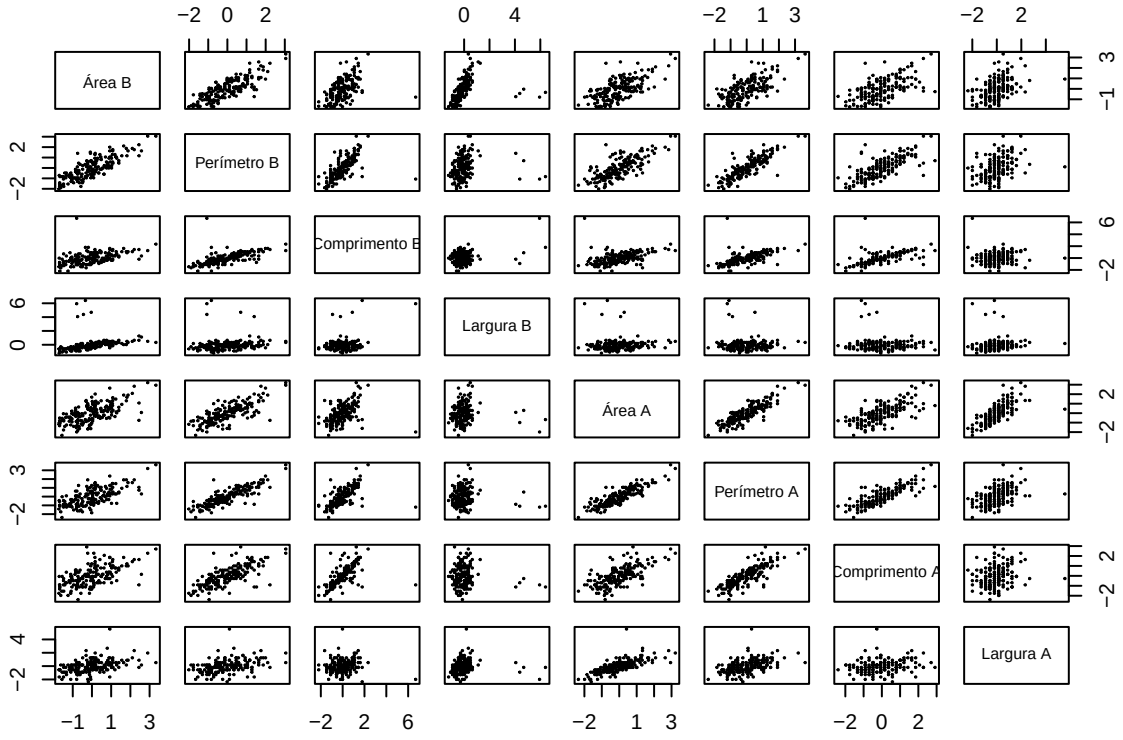

Figura 1: Gráfico de dispersão entre as variáveis quantitativas (dados padronizados)

Tabela 3: Medidas resumo das variáveis qualitativas padronizadas

	Área B	Perímetro B	Comprimento B	Largura B	Área A	Perímetro A	Comprimento A	Largura A
Mínimo	0,02	0,33	-0,22	-0,95	-1,24	-1,71	-1,68	-0,72
Mediana	-0,05	-0,13	-0,01	-0,13	-0,06	-0,02	-0,06	-0,19
Máximo	0,02	0,33	-0,22	-0,95	-1,24	-1,71	-1,68	-0,72
Média	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Desvio Padrão	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

```
Matriz de correlações:

                Área B Perímetro B Comprimento B   Largura B      Área A Perímetro A Comprimento A  Largura A
Área B       1.0000000   0.8083462     0.4533735  0.26989735  0.64367994  0.67311435    0.57401775 0.42956798
Perímetro B  0.8083462   1.0000000     0.6138876  0.11837179  0.75141227  0.85635765    0.72168864 0.39036315
Comprimento B 0.4533735  0.6138876     1.0000000  0.32676677  0.40647179  0.57902951    0.59331145 0.06039870
Largura B    0.2698973   0.1183718     0.3267668  1.00000000 -0.00044263 -0.04050293   -0.07667231 0.04386148
Área A       0.6436799   0.7514123     0.4064718 -0.00044263  1.00000000  0.86483730    0.68207728 0.72287369
Perímetro A  0.6731143   0.8563577     0.5790295 -0.04050293  0.86483730  1.00000000    0.82772860 0.45710884
Comprimento A 0.5740178  0.7216886     0.5933115 -0.07667231  0.68207728  0.82772860    1.00000000 0.19753686
Largura A    0.4295680   0.3903632     0.0603987  0.04386148  0.72287369  0.45710884    0.19753686 1.00000000
```

2 Aplicando Análise de Componentes Principais (PCA)

Seguindo então para análise com a matriz de correlações temos o seguinte:

```
CP <- princomp(trigop) # Aplicando nos dados quantitativos padronizados
summary(CP, loadings=T, cutoff = 0)
```

```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6     Comp.7      Comp.8
Standard deviation     2.1500775 1.1207990 1.0212391 0.67024857 0.51352357 0.40917052 0.29655364 0.251457851
Proportion of Variance 0.5812334 0.1579421 0.1311285 0.05648253 0.03315608 0.02104995 0.01105729 0.007950103
Cumulative Proportion  0.5812334 0.7391755 0.8703040 0.92678658 0.95994265 0.98099260 0.99204990 1.000000000

Loadings:
             Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
Área B        0.381  0.131  0.189  0.690  0.198  0.428  0.320  0.028
Perímetro B   0.429  0.068 -0.064  0.288  0.280 -0.517 -0.587 -0.187
Comprimento B 0.307  0.477 -0.255 -0.528  0.518  0.221  0.118 -0.058
Largura B     0.060  0.712  0.524 -0.073 -0.427 -0.144 -0.061  0.053
Área A        0.415 -0.266  0.166 -0.211 -0.221 -0.192  0.422 -0.649
Perímetro A   0.436 -0.116 -0.156 -0.079 -0.141 -0.389  0.337  0.694
Comprimento A 0.383 -0.014 -0.414 -0.043 -0.576  0.459 -0.369 -0.022
Largura A     0.254 -0.399  0.630 -0.322  0.178  0.285 -0.328  0.234
```

Logo de cara, avaliando somente a proporção acumulada do percentual de explicação, notamos que na segunda componente principal obtivemos um percentual de 73,91% e para a terceira obtivemos um percentual de 87,03% consideramos ambos os valores razoáveis para a explicação do modelo. Recorrendo ao critério de Kaiser, decidimos por manter 3 componentes principais, já que o desvio padrão de cada das 3 primeiras componentes é maior que 1 o que equivale dizer que a raiz dos autovalores são maiores que 1 e consequentemente são estatisticamente mais significativos ao modelo e juntos explicam 87,03% total do modelo. Contudo, resta analisar se todas variáveis estão bem representadas por essas 3 componentes atráves do cálculo da comunalidade parcial.

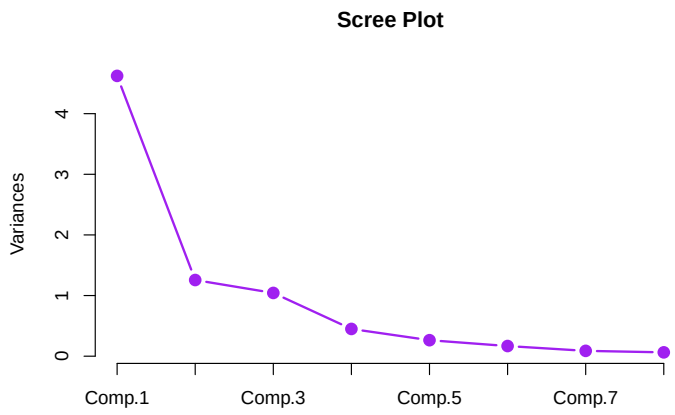

Figura 2: Scree Plot para observar variâncias de cada componente da análise

2.1 Comunalidade

Dessa forma, com 3 componentes principais, seguiremos para calcular o percentual da variância de cada variável original explicada por cada componente, ou seja, a soma das comunalidades para cada componente.

```
Comunalidade por cada Componente em cada variável original:
```

	Área B	Perímetro B	Comprimento B	Largura B	Área A	Perímetro A	Comprimento A	Largura A
Y1	0.6754	0.8540	0.4372	0.0165	0.8001	0.8833	0.6834	0.2999
Y2	0.0216	0.0058	0.2878	0.6404	0.0892	0.0170	0.0003	0.2016
Y3	0.0375	0.0043	0.0682	0.2884	0.0289	0.0254	0.1799	0.4165
Y4	0.2152	0.0376	0.1260	0.0024	0.0201	0.0028	0.0008	0.0469
Y5	0.0104	0.0208	0.0711	0.0483	0.0129	0.0053	0.0881	0.0084
Y6	0.0309	0.0449	0.0083	0.0035	0.0062	0.0255	0.0355	0.0137
Y7	0.0091	0.0305	0.0012	0.0003	0.0157	0.0100	0.0121	0.0095
Y8	0.0001	0.0022	0.0002	0.0002	0.0268	0.0306	0.0000	0.0035

```
Comunalidade Acumulada:
```

	Área B	Perímetro B	Comprimento B	Largura B	Área A	Perímetro A	Comprimento A	Largura A
Y1	0.6754	0.8540	0.4372	0.0165	0.8001	0.8833	0.6834	0.2999
Y2	0.6970	0.8597	0.7250	0.6569	0.8893	0.9004	0.6837	0.5015
Y3	0.7345	0.8640	0.7932	0.9453	0.9182	0.9258	0.8635	0.9180
Y4	0.9496	0.9016	0.9192	0.9477	0.9383	0.9286	0.8644	0.9649
Y5	0.9600	0.9224	0.9903	0.9960	0.9512	0.9339	0.9524	0.9733
Y6	0.9909	0.9673	0.9986	0.9995	0.9574	0.9593	0.9879	0.9870
Y7	0.9999	0.9978	0.9998	0.9998	0.9732	0.9694	1.0000	0.9965
Y8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

```
Comunalidade parcial (acumulada de Y1,Y2 e Y3):
```

	Percentual
Área B	0.7347
Perímetro B	0.8643
Comprimento B	0.7932
Largura B	0.9455

```
Área A           0.9184
Perímetro A      0.9259
Comprimento A    0.8637
Largura A        0.9179
```

Observação: Nota-se que a comunalidade da tabela do meio com a última não coincidem precisamente, devido às aproximações de casas decimais, porém o método de cálculo é o mesmo.

Notamos então que todos os percentuais estão acima de 75% indicando então que para todas as variáveis, considerando as 3 primeiras componentes, temos um percentual de explicação que consideramos bom para seguir com a análise.

2.2 Interpretação das componentes mantidas

Seguimos então para calcular a correlação entre as variáveis originais e as componentes.

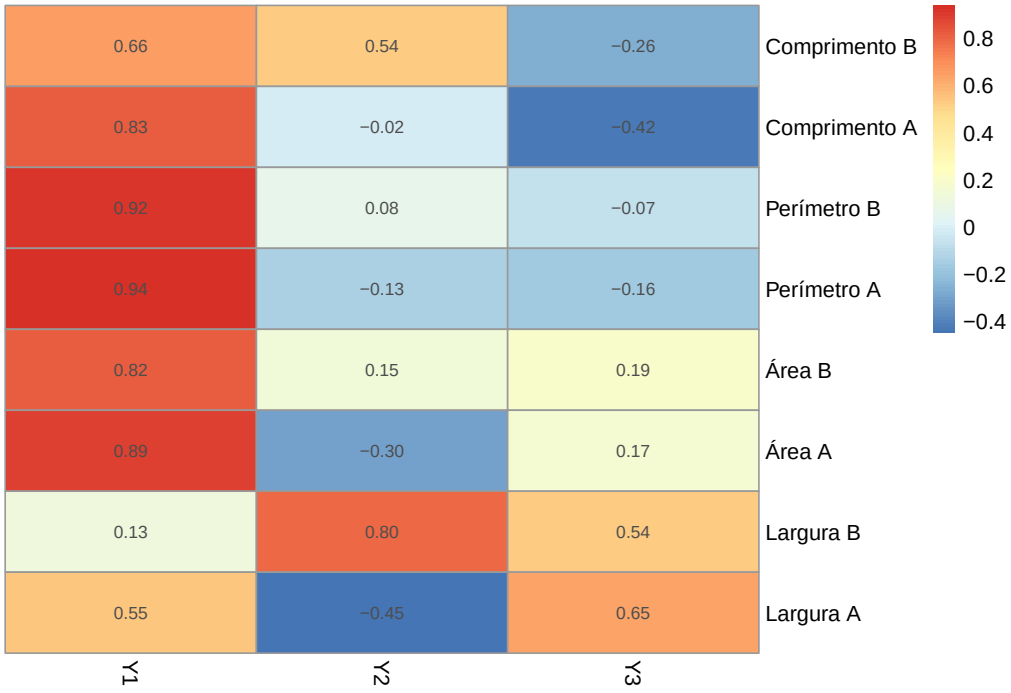

Figura 3: Gráfico de calor sobre as correlações entre as componentes e as variáveis

Vemos então que há uma alta correlação ($> 0,60$) das variáveis `Comprimento B`, `Comprimento A`, `Perímetro B`, `Perímetro A`, `Área B`, `Área A` com a Componente 1. Já `Largura B` tem correlação alta com a Componente 2 e `Largura A` com a Componente 3. Com os resultados descritos acima, podemos interpetar que a CP1 (Y1) ficou responsável em explicar, de forma geral, as variáveis Comprimento, Perímetro e Área. Já a CP2 (Y2) está mais voltada a explicar a Largura B e analogamente CP3 (Y3) está voltada a explicar Largura A.

2.3 Cálculo das componentes

A seguir, será calculado o valor das componentes mantidas através da matriz de cargas (*Loadings*) transposta:

```
Amostra tomada para cálculo da componentes assinada por ‘obs‘ (extraída do conjunto padronizado):

      Área B   Perímetro B Comprimento B     Largura B        Área A   Perímetro A Comprimento A     Largura A
  0.01973862   -0.12597530   -0.19597225   -0.28605935    1.23592988    0.62575162   -0.26915921    1.24586304
```

```
obs <- trigop[1,1:8]

Y1 = 0.381*obs[1] + 0.429*obs[2] + 0.307*obs[3] + 0.060*obs[4] + 0.415*obs[5] + 0.436*obs[6] + 0.383*obs[7] +
  0.254*obs[8]


Y2 = 0.131*obs[1] + 0.068*obs[2] + 0.477*obs[3] + 0.712*obs[4] + -0.266*obs[5] + -0.116*obs[6] + -0.014*obs[7] +
  -0.399*obs[8]


Y3 = 0.189*obs[1] + -0.064*obs[2] + -0.255*obs[3] + 0.524*obs[4] + 0.166*obs[5] + -0.156*obs[6] + -0.414*obs[7] +
  0.630*obs[8]

cat(Y1)
```

```
0.8752498
```

```
cat(Y2)
```

```
-1.197809
```

```
cat(Y3)
```

```
0.9157436
```

```
CP$scores[1,1:3]
```

```
    Comp.1     Comp.2     Comp.3
 0.8750144 -1.1978949  0.9160756
```

Nota: novamente, os valores divergem pela precisão dos cálculos, porém em suma representam o mesmo procedimento e são considerados coincidentes, verificando os cálculos.

3 Apêndice

3.1 Código

```
library(readxl)
library(ggplot2)
library(showtext)
library(tidyverse)
library(kableExtra)
library(knitr)

trigo <- read_excel("trigo.xls")
colnames(trigo) <- c("Localidade","Variedade","Área B", "Perímetro B","Comprimento B", "Largura B","Área A", "Perímetro A",
summary(trigo[,3:10])

trigo1 <- trigo[,3:10]
trigop <-scale(trigo1)

plot(as.data.frame(trigop), cex=0.15, mar=c(0,0,0,0))
```

```
cor(trigo1)

CP <- princomp(trigop) # Aplicando nos dados quantitativos padronizados
summary(CP, loadings=T, cutoff = 0)

screeplot(CP, col="purple", pch=16, type="lines", cex=1.5, lwd=2, main="Scree Plot")

# Calculando os autovetores e autovalores
de <- eigen(cor(trigop))
#Seprando e arredondando 4 casas decimais dos autovalores
B <- round(de$values[1:3],4)
#Seprando o quadrado e arredondando 4 casas decimais dos autovetores
A <- round((de$vectors[,1:3])^2,4)
#Calculando a soma de suas comunalidades para 3 CP

m <- matrix(rep(0,times=8*8),ncol=8,nrow=8)
for (i in 1:8){
  for (j in 1:8){
    m[i,j] <- ((de$values[i])*((de$vectors[j,i])^2))/(sum((de$vectors[i,]^2)*de$values))
  }
}
colnames(m) <- colnames(trigo1)
rownames(m) <- c('Y1','Y2','Y3','Y4','Y5','Y6','Y7','Y8')

cat("Comunalidade por cada Componente em cada variável original:\n")
round(m,4)

cat("\n")
cat("Comunalidade Acumulada:\n")
round(cumsum(as.data.frame(m)),4)

cat("\n")
cat("Comunalidade parcial (acumulada de Y1,Y2 e Y3):\n")
round(data.frame("Percentual"=A %*% B, row.names = colnames(trigop)),4)

library(FactoMineR)
library("pheatmap")
acpa <- PCA(trigop,graph = FALSE)
table <- round(acpa$var$coord[,1:3],4)
colnames(table) <- c("Y1","Y2","Y3")
pheatmap(table,display_numbers=TRUE,treeheight_row = 0,treeheight_col =0 )

obs <- trigop[1,1:8]

Y1 = 0.381*obs[1] + 0.429*obs[2] + 0.307*obs[3] + 0.060*obs[4] + 0.415*obs[5] + 0.436*obs[6] + 0.383*obs[7] +
  0.254*obs[8]


Y2 = 0.131*obs[1] + 0.068*obs[2] + 0.477*obs[3] + 0.712*obs[4] + -0.266*obs[5] + -0.116*obs[6] + -0.014*obs[7] +
  -0.399*obs[8]


Y3 = 0.189*obs[1] + -0.064*obs[2] + -0.255*obs[3] + 0.524*obs[4] + 0.166*obs[5] + -0.156*obs[6] + -0.414*obs[7] +
  0.630*obs[8]

cat(Y1)
cat(Y2)
cat(Y3)
CP$scores[1,1:3]
```