Instituto de Matemática e Estatística
Universidade de São Paulo

Lista 9

MAE0330 - Análise Multivariada de Dados

Profª Lucia Pereira Barroso

Bruno Groper Morbin - nºUSP 11809875

Luigi Pavarini de Lima - nºUSP 11844642

São Paulo
19 de dezembro, 2022

Análise de Correlação Canônica

Análise descritiva

Tabela 1: Medidas resumo - conjunto de dados 'ginidh.txt'

	Peso	Ram	Memória	Pixel	Hardware	Tela	Câmera	Desempenho
Mínimo	135	3	32	269	5.9	5.3	6.5	4.3
1º Quartil	181.5	4	128	379	7.2	8.35	7.9	6.55
Mediana	191	6	128	395	7.7	8.4	8.4	8.5
3º Quartil	216.5	7	256	418.5	8.6	8.6	8.65	9.3
Máximo	518	12	512	476	9.5	9.4	9.5	10
Média	207.7	5.67	234.67	388.56	7.77	8.1	8.28	7.78
Desvio Padrão	67.8	2.15	163.41	62	1.02	1.13	0.62	1.89
Observações	27	27	27	27	27	27	27	27

Com este resumo descritivo, notamos logo de cara que as variáveis tem magnitudes diferentes, em epecial **Peso** com range de 135 a 518, **Memória** com um range de 32 a 512 e **Pixel** com um range 269 a 476 têm as maiores magnitudes do conjunto de variáveis.Além disso, podemos dizer que tanto **Peso** quando **Pixel** parecem simétricas e **Memória** tem uma leve assimetria a esquerda.

Sobre as demais variáveis, tem-se um range que começa sempre em torno de 5 e vão até valores próximos de 10 com destaque para **Ram** e **Desempenho** que são as variáveis com as maiores variabilidades ($DP \approx 2$). Vale comentar também, que todas estas demais variáveis aparentam ser simétricas já que $Mdia \approx Mediana$.

Dito, isso consideramos razoável seguirmos com a **Análise de Correlação Canônica** utilizando-se das variáveis normalizadas afim de facilitar nas interpretações e eventuais inferências.

Seguindo então vamos para a separação dos grupos e determinação do que se pede.

Aplicação do Método e Determinação da Quantitadade de Pares

Seja então os seguintes conjuntos selecionados e padronizados:

Tabela 2: Primeiras observações do Cojunto 1

Peso	Ram	Memória	Pixel
-0.2463613	-1.2412658	-1.2402355	-1.9122948
-0.0398767	-1.2412658	-1.2402355	-1.9122948
-0.1136212	-0.7757911	-0.6527555	-1.9284248
-0.4085993	0.1551582	-0.6527555	0.3620277
-0.3496037	-0.7757911	-0.6527555	0.3620277
-0.5708373	1.0861076	-0.6527555	0.0555587

Tabela 3: Primeiras observações do Cojunto 2

Hardware	Tela	Câmera	Desempenho
-1.8351742	-2.4757087	-2.8599152	-0.7289356
-1.4419226	-2.2986382	-0.2859915	-0.6760290
-1.3436097	-2.2986382	0.0357489	-0.6231224
-1.1469839	0.4459555	0.1966192	-1.8399745
-1.1469839	0.4459555	0.1966192	-1.8399745
-0.7537323	0.2688849	0.8401001	-1.1521885

Onde temos as seguintes matrizes de correlação:

Tabela 4: Matriz de Correlações - Σ11

	Peso	Ram	Memória	Pixel
Peso	1.000	0.386	0.129	-0.382
Ram	0.386	1.000	0.154	0.109
Memória	0.129	0.154	1.000	0.396
Pixel	-0.382	0.109	0.396	1.000

Tabela 5: Matriz de Correlações - Σ22

	Hardware	Tela	Câmera	Desempenho
Hardware	1.000	0.649	0.170	0.761
Tela	0.649	1.000	0.378	0.221
Câmera	0.170	0.378	1.000	-0.165
Desempenho	0.761	0.221	-0.165	1.000

Tabela 6: Matriz de Correlações - Σ12 e t(Σ21)

	Hardware	Tela	Câmera	Desempenho
Peso	-0.146	-0.335	0.083	0.158
Ram	-0.055	0.242	0.150	0.040
Memória	0.729	0.398	-0.063	0.673
Pixel	0.660	0.934	0.296	0.268

Aplicando então **Análise de correlação Canônica** pelo função cc(.) do pacote CCA temos os seguinte **Correlações Canônicas**

Tabela 7: Correlações Canônicas

$\sqrt{\lambda_1}$	$\sqrt{\lambda_2}$	$\sqrt{\lambda_3}$	$\sqrt{\lambda_4}$
0.952	0.771	0.511	0.27

Calculando os percentuais de correlação explicada.

$$\sqrt{\lambda_1} = 38.01\%$$
$$\sqrt{\lambda_1} + \sqrt{\lambda_2} = 68.82\%$$
$$\sqrt{\lambda_1} + \sqrt{\lambda_2} + \sqrt{\lambda_3} = 89.22\%$$
$$\sqrt{\lambda_1} + \sqrt{\lambda_2} + \sqrt{\lambda_3} + \sqrt{\lambda_4} = 100\%$$

Podemos notar então que utlizando os três primeiros pares de variáveis canônicas (U_i, V_i para i=1,2,3) obtém-se $\approx 90\%$ da correlação total explicada.

```
## ------------------------------------------------------------------------------------
```

```
## Wilks' Lambda, using F-approximation (Rao's F):
##               stat   approx df1      df2      p.value
## 1 to 4:  0.0262196 8.411048  16 58.68358 5.127232e-10
## 2 to 4:  0.2775389 3.761161   9 48.82535 1.179963e-03
## 3 to 4:  0.6853901 2.182953   4 42.00000 8.739121e-02
## 4 to 4:  0.9271908 1.727585   1 22.00000 2.022589e-01
```

```
## --------------------------------------------------------------------------------
```

Interpretando os testes de Hipóteses, concluímos que apenas as correalções entre (U_3, V_3) e (U_4, V_4) não são significativos ao nível de $\alpha = 0.05$, porém avaliando a primeira linha rejeitamos a hipótese de não existência de correlação entre os conjuntos 1 e 2. Além disso, apesar da não rejeição das hipóteses do pare (U_3, V_3) verificamos que a correlação total explicada sem esse par gira em torno de 70% e com esse par gira em torno de 90% portanto manteremos as três variáveis canônicas para a explicação da correlação total entre os conjuntos.

Cálculo das variáveis canônicas para uma observação

Tem-se que as variáveis canônicas são obtidas por $U = XA$, $V = YB$, onde $X \equiv CJ1$ (Conjunto 1), $Y \equiv CJ2$ (Conjunto 2), e A, B são as matrizes dos coeficientes obtidas pelo ajuste do modelo.

Logo,

$$A = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \end{array} \begin{array}{c} \begin{array}{cccc} U_1 & U_2 & U_3 & U_4 \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} \end{array}$$

$$B = \begin{array}{c} \\ Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{array} \begin{array}{c} \begin{array}{cccc} V_1 & V_2 & V_3 & V_4 \end{array} \\ \begin{pmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \\ b_{31} & b_{32} & b_{33} & b_{34} \\ b_{41} & b_{42} & b_{43} & b_{44} \end{pmatrix} \end{array}$$

Sendo:

- X_1 = Peso
- X_2 = RAM
- X_3 = Mémoria
- X_4 = Pixel
- Y_1 = Hardware
- Y_2 = Tela
- Y_3 = Câmera
- Y_4 = Desempenho

Logo, $U_k = a_{.k}^T X_{obs}$ e $V_k = b_{.k}^T Y_{obs}$, para $k = 1, 2, 3, 4$ e $a^T \equiv$ vetor da i-ésima coluna de A (análogo para b^T).

Portanto, tem-se:

$$\begin{aligned}
U_1 &= -0,068 * Peso + 0,148 * RAM + 0,145 * Memoria + 0,877 * Pixel \\
U_2 &= -0,052 * Peso - 0,598 * RAM + 0,957 * Memoria - 0,367 * Pixel \\
U_3 &= -0,879 * Peso - 0,214 * RAM - 0,185 * Memoria - 0,163 * Pixel \\
U_4 &= 0,935 * Peso - 0,932 * RAM - 0,603 * Memoria + 0,858 * Pixel \\
V_1 &= 0,057 * Hardware + 0,969 * Tela - 0,086 * Camera + 0,069 Desempenho \\
V_2 &= 1,668 * Hardware - 0,976 * Tela - 0,329 * Camera - 0,443 * Desempenho \\
V_3 &= 1,397 * Hardware - 0,213 * Tela - 0,724 * Camera - 1,755 * Desempenho \\
V_4 &= 1,102 * Hardware - 0,776 * Tela + 0,832 * Camera - 0,771 * Desempenho
\end{aligned}$$

Interpretação

Dito isso, temos os seguintes **Coeficientes das Variáveis Canônicas** ajustados pela função cc(.) do pacote CCA:

Tabela 8: Coeficientes das Variáveis Canônicas U

	U1	U2	U3	U4
Peso	-0.086	-0.052	-0.879	0.935
Ram	0.148	-0.598	-0.214	-0.932
Memória	0.145	0.957	-0.185	-0.603
Pixel	0.877	-0.367	-0.163	0.858

Tabela 9: Coeficientes das Variáveis Canônicas V

	V1	V2	V3	V4
Hardware	0.057	1.668	1.397	1.102
Tela	0.969	-0.976	-0.213	-0.776
Câmera	-0.086	-0.329	-0.724	0.832
Desempenho	0.069	-0.443	-1.755	-0.771

Avaliando então os coeficeintes da **Tabela 8** temos que o maior coeficiente, em módulo, da combinação linear de U_1 com as variáveis do $Cj1$ se deve a variável **Pixel** ($a_{4,1} = 0.877$), ou seja tem-se U_1 como representante da variável **Pixel**.

Analogamente, temos que o maior coeficiente da comb. linear de U_2 com as variáveis de $Cj1$ se deve a var. **Memória** ($a_{3,2} = 0.957$), ou seja U_2 é representante da variável **Memória**.

já para U_3 , ocorreu que o maior coeficiente da comb. linear de U_3 com as variáveis de $Cj1$ se deve a var. **Peso** ($a_{1,3} = -0.879$), ou seja U_3 é representante de **Peso**.

Agora em relação a **Tabela 9**, temos que o maior coeficiente, em módulo, da combinação linear de V_1 com as variáveis de $Cj2$ se deve a variável **Tela** ($b_{2,1} = 0.969$), ou seja V_1 é representante da variável **Tela**.

Analogamente, temos que o maior coeficiente da comb. linear de V_2 com as variáveis de $Cj2$ se deve a var. **Hardware** ($b_{1,2} = 1.668$), ou seja V_2 é representante da variável **Hardware**.

já para V_3 , ocorreu que o maior coeficiente da comb. linear de V_3 com as variáveis de $Cj2$ se deve a var. **Desempenho** ($b_{3,3} = -1.755$), ou seja V_3 é representante da variável **Desempenho**.

Além disso, temos os seguintes **Cargas Canônicas: Corr. entre Variáveis Canônicas e Variáveis Originais** ajustadas pela função cc(.) do pacote CCA:

Tabela 10: Cargas Canônicas U:Cj1

	U1	U2	U3	U4
Peso	-0.346	-0.019	-0.923	0.169
Ram	0.233	-0.511	-0.600	-0.571
Memória	0.504	0.712	-0.396	-0.286
Pixel	0.984	-0.034	0.075	0.160

Tabela 11: Cargas Canônicas V:Cj2

	V1	V2	V3	V4
Hardware	0.725	0.641	-0.201	0.153
Tela	0.989	-0.116	0.032	0.083
Câmera	0.279	-0.341	-0.278	0.854
Desempenho	0.341	0.665	-0.619	-0.240

Avaliando então as correlações da **Tabela 10** temos que a maior correlação de U_1 com as variáveis do $Cj1$ se deve a variável **Pixel** ($\rho(a)_{4,1} = 0.984$), ou seja a medida que U_1 cresce, **Pixel** também cresce.

Analogamente, temos que a maior correlação de U_2 com as variáveis de $Cj1$ se deve a var. **Memória** ($\rho(a)_{3,2} = 0.712$), ou seja quando U_2 cresce a variável **Memória** também cresce.

já para U_3 , ocorreu que as maiores correlações de U_3 com as variáveis de $Cj1$ se devem as var's **Peso** ($\rho(a)_{1,3} = -0.923$) e **Ram** ($\rho(a)_{1,3} = -0.6$) ou seja quando U_3 aumenta tanto **Peso** quanto **Ram** diminuem.

Agora em relação a **Tabela 11**, temos que as maiores correlações de V_1 com as variáveis de $Cj2$ se devem as variáveis **Hardware** ($\rho(b)_{1,1} = 0.725$) e **Tela** ($\rho(b)_{2,1} = 0.989$), ou seja quando V_1 aumenta tanto **Tela** quanto **Hardware** aumentam também.

Analogamente, temos que as maiores correlações de V_2 com as variáveis de $Cj2$ se devem a var's. **Hardware** ($\rho(b)_{1,2} = 0.641$) e **Desempenho** ($\rho(b)_{4,2} = 0.665$), ou seja quando V_2 aumenta as variáveis **Hardware** e **Desempenho** também aumentam.

já para V_3 , ocorreu que a maior correlação de V_3 com as variáveis de $Cj2$ se deve a var. **Desempenho** ($\rho(b)_{3,3} = -0.619$), ou seja quando V_3 aumenta a variável **Desempenho** diminui.

Tabela 12: Cargas Cruzadas V:Cj1

	V1	V2	V3	V4
Peso	-0.329	-0.015	-0.471	0.046
Ram	0.221	-0.394	-0.306	-0.154
Memória	0.480	0.549	-0.202	-0.077
Pixel	0.936	-0.026	0.039	0.043

Tabela 13: Cargas Cruzadas U:Cj2

	U1	U2	U3	U4
Hardware	0.689	0.495	-0.103	0.041
Tela	0.941	-0.089	0.016	0.023
Câmera	0.265	-0.263	-0.142	0.230
Desempenho	0.325	0.513	-0.316	-0.065

Avaliando então as correlações cruzadas da **Tabela 12** temos que a maior correlação de V_1 com as variáveis do $Cj1$ se deve a variável **Pixel** ($\rho(a)_{4,1} = 0.936$), ou seja a medida que V_1 cresce, **Pixel** também cresce.

Analogamente, temos que a maior correlação cruzada de V_2 com as variáveis de $Cj1$ se deve a var. **Memória** ($\rho(a)_{3,2} = 0.549$), ou seja quando V_2 cresce a variável **Memória** também cresce.

já para V_3 , ocorreu que a maior correlação cruzada de V_3 com as variáveis de $Cj1$ se devem as var's **Peso** ($\rho(a)_{1,3} = -0.471$) ou seja, apesar de não tão forte, ocorre que quando V_3 aumenta **Peso** diminue também.

Agora em relação a **Tabela 13**, temos que as maiores correlações cruzadas de U_1 com as variáveis de $Cj2$ se devem as variáveis **Hardware** ($\rho(b)_{1,1} = 0.689$) e **Tela** ($\rho(b)_{2,1} = 0.941$), ou seja quando U_1 aumenta tanto **Tela** quanto **Hardware** aumentam também.

Analogamente, temos que a maior correlação cruzada de U_2 com as variáveis de $Cj2$ se devem a var. **Desempenho** ($\rho(b)_{4,2} = 0.513$), ou seja quando U_2 aumenta a variável **Desempenho** também aumentam.

já para U_3, ocorreu que a maior correlação cruzada de U_3 com as variáveis de $Cj2$ se deve a var. **Desempenho** ($\rho(b)_{3,3} = -0.316$), ou seja quando U_3 aumenta a variável **Desempenho** diminui.

Código

```
library(readxl)
library(CCA)
library(CCP)
library(tidyverse)
library(kableExtra)
library(ascii)
library(knitr)

celular <- read_excel("Celular.xlsx")
celular1<- celular[c("Peso","Ram","Memória","Pixel","Hardware",
                     "Tela","Câmera","Desempenho")]
celular_p <- scale(celular[c("Peso","Ram","Memória","Pixel",
                           "Hardware","Tela","Câmera","Desempenho")])

cria_tabela_resumo <- function(database){
  medidas_resumo <- data.frame(row.names = c("Mínimo","1º Quartil","Mediana","3º Quartil","Máximo","Média","Desvio Padrão", "Observações"))
  for(i in 1:ncol(database)){
    media <- round(mean(unlist(database[,i]), na.rm = T),2)
    dp <- round(sqrt(var(database[,i],na.rm=T))[[1]],2)
    max <- round(max(database[i], na.rm=T),2)
    min <- round(min(database[i], na.rm=T),2)
    mediana <- round(median(unlist(database[,i]))[[1]],2)
    quartil1 <- round(quantile(unlist(database[,i]),0.25)[[1]],2)
    quartil3 <- round(quantile(unlist(database[,i]),0.75)[[1]],2)
    obs <- trunc(sum(!is.na(unlist(database[,i]))))

    nome_variavel <- (colnames(database))[i]
    medidas_resumo <- cbind(medidas_resumo,
                            data.frame(c(min,quartil1,mediana,quartil3,max,media,dp,obs))
                            )
    colnames(medidas_resumo)[i] <- nome_variavel

  }
  return(as.data.frame(sapply(medidas_resumo,as.character), row.names = rownames(medidas_resumo)))
}

tabela_resumo <- cria_tabela_resumo(celular1)
#https://haozhu233.github.io/kableExtra/awesome_table_in_pdf.pdf
#https://cran.r-project.org/web/packages/kableExtra/kableExtra.pdf
tabela_resumo%>%kbl(
  booktabs = T,
  caption = "Medidas resumo - conjunto de dados `ginidh.txt`",
  # digits = 2,
  align = c("r"),
  format.args = list(big.mark=".", decimal.mark=","),
  linesep = c(rep("",4),"\\addlinespace")
  )%>%
  kable_styling(font_size = 9,latex_options = c("HOLD_position"), position="center")

cj1<-as.matrix(celular_p[,1:4])
cj2<-as.matrix(celular_p[,5:8])

kable(head(as.data.frame(cj1),6), format = "markdown",caption = "Primeiras observações do Cojunto 1")
kable(head(as.data.frame(cj2),6), format = "markdown",caption = "Primeiras observações do Cojunto 2")

#print(ascii(head(as.data.frame(cj1),10)), type = 'pandoc')
#print(ascii(head(as.data.frame(cj2),10)), type = 'pandoc')

cov_cj1 <- cov(cj1)

cov_cj2 <- cov(cj2)

cov_cj1_cj2 <- cov(cj1,cj2)

cov_cj2_cj1 <- cov(cj2,cj1)

kable(head(as.data.frame(round(cov_cj1, 3))), format = "markdown",caption = "Matriz de Correlações - Σ11")

kable(head(as.data.frame(round(cov_cj2, 3))), format = "markdown",caption = "Matriz de Correlações - Σ22")

kable(head(as.data.frame(round(cov_cj1_cj2, 3))), format = "markdown",caption = "Matriz de Correlações - Σ12 e t(Σ21)")


cct1 <- cc(cj1,cj2)
#cct1
rho<-cct1$cor

tab2 <- as.data.frame(round(t(rho), 3))
colnames(tab2)<-str_c(str_c("$\\sqrt{\\lambda_", 1:4),"}$")

kable(head(tab2), escape = F,format = "markdown",caption = "Correlações Canônicas")

cat("\n","$\\sqrt{\\lambda_1}$:",  rho[1]/sum(rho)*100,"%" )
```

```
cat("\n","$\\sqrt{\\lambda_1}+\\sqrt{\\lambda_2}$:", (rho[1]+rho[2])/sum(rho)*100,"%" )
cat("\n","$\\sqrt{\\lambda_1}+\\sqrt{\\lambda_2}+\\sqrt{\\lambda_3}$:",(rho[1]+rho[2]+rho[3])/sum(rho)*100,"%" )
cat("\n","$\\sqrt{\\lambda_1}+\\sqrt{\\lambda_2}+\\sqrt{\\lambda_3}+\\sqrt{\\lambda_4}$:", (rho[1]+rho[2]+rho[3]+rho[4])/sum(rho)*100,"%" )

#testes
n <- dim(cj1)[1]
p <- dim(cj1)[2]
q <- dim(cj2)[2]
cat(str_c(rep("-",88),collapse = ""))
teste <- p.asym(rho, n, p, q, tstat ="Wilks")
cat(str_c(rep("-",88),collapse = ""))

tab1 <- as.data.frame(round(cct1$xcoef, 3))
colnames(tab1)<-str_c("U",1:4)

kable(head(tab1), format = "markdown",caption = "Coeficientes das Variáveis Canônicas U")

kable(head(as.data.frame(round(cct1$ycoef, 3))), format = "markdown",caption = "Coeficientes das Variáveis Canônicas V")

tab3<-as.data.frame(round(cct1$scores$corr.X.xscores, 3))
colnames(tab3)<-str_c("U",1:4)# U contra cj1
kable(tab3, format = "markdown",caption = "Cargas Canônicas U:Cj1")

kable(as.data.frame(round(cct1$scores$corr.Y.yscores, 3)), format = "markdown",caption = "Cargas Canônicas V:Cj2")#V contra cj2

kable(as.data.frame(round(cct1$scores$corr.X.yscores, 3)), format = "markdown",caption = "Cargas Cruzadas V:Cj1") # V contra cj1

tab4<-as.data.frame(round(cct1$scores$corr.Y.xscores, 3))
colnames(tab4)<-str_c("U",1:4)
kable(tab4, format = "markdown",caption = "Cargas Cruzadas U:Cj2") # U contraa cj2
```