Universidade Federal de São Carlos

Aluno: Rafael Setti RA 744870

José Matheus Badaró RA 636940

Professor: Pedro Ferreira Filho

Lista 2 - Estatística Multivariada 2

São Carlos - SP

2021

1 Exercício 1

1.1

Tendo como dado os autovalores da matriz de covariâncias, podemos sumarizar as variâncias explicadas por cada componente tal como:

Componente	**Autovalor**	**% da variância**	**% var acumulada**
CP_1	2,49	83%	83%
CP_2	0,42	14%	97%
CP_3	0,09	3%	100%

Tabela 1: Contribuição de cada componente para explicar a variância original do primeiro conjunto de variáveis

Dessa forma, é possível visualizar que o Componente Principal 1 é capaz de explicar 83% da variação dos dados e, junto ao Componente Principal 2, são capazes de explicar 97% da variação total dos dados.

Observando os autovetores, temos:

	CP_1	CP_2	CP_3
V_1	0.617	-0.001	0.787
V_2	0.557	-0.706	-0.437
V_3	0.556	0.708	-0.434

Tabela 2: Autovetores relativos aos Componentes e as variáveis originais

Como visto pela tabela elencando os autovetores para cada uma das variáveis, é possível visualizar que as 3 variáveis contribuem para a formação do componente principal 1, contudo, cada componente possui uma maior contribuição para cada uma das variáveis, embora o CP_1 seja capaz de explicar sozinho 83% da variação dos dados.

Dados os autovalores, tem-se os seguintes componentes principais:

$$Y_1 = 0.617V_1 + 0.557V_2 + 0.556V_3$$

$$Y_2 = -0.001V_1 - 0.706V_2 + 0.708V_3$$

$$Y_3 = 0.787V_1 - 0.437V_2 - 0.434V_3$$

1.2

Tendo os componentes principais definidos, é de nosso interesse posicionar as seguintes empresas de acordo com os componentes.

Assim, de acordo com as empresas fornecidas, formam-se os seguintes valores transformados de X, sendo eles:

$$Y_1 = (12801.906; 9081.845) \tag{1}$$

$$Y_2 = (6090.934; 3905.033) \tag{2}$$

Agora, basta entender onde esses novos pontos transformados de acordo com os componentes principais estarão posicionados em um gráfico do primeiro plano:

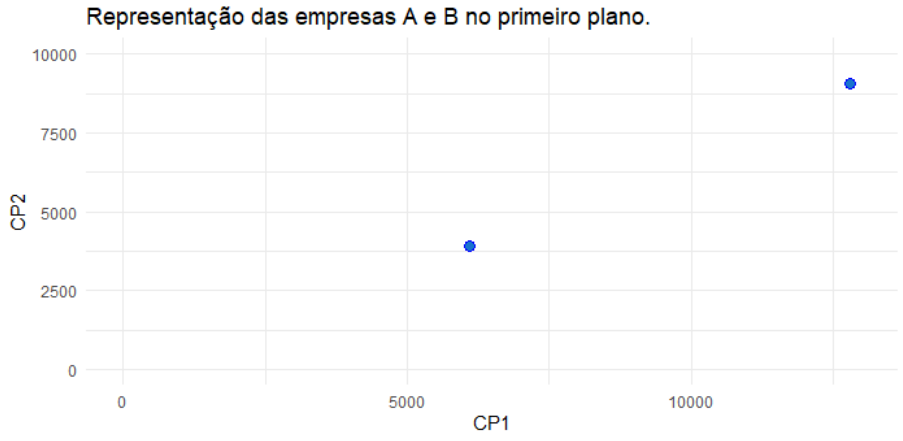

Figura 1: Empresas A e B representadas no primeiro plano

2 Exercício 2

2.1

Para ilustrar melhor os dados do problema, faz-se a seguinte tabela com os autovalores:

Componente	**Autovalor**	**% da Variância**	**% Var Acumulada**
CP1	2.629	65,725%	65,7250%
CP2	1.19	29,75%	95,4750%
CP3	0.13	3,250%	98,725%
CP4	0.051	1,275%	100%

Tabela 3: Contribuição de cada componente para explicar a variância original do primeiro conjunto de variáveis

E os autovetores são dados por:

	CP1	CP2	CP3	CP4
Sabor	0.49	0.54	-0.61	0.24
Aroma	0.48	0.76	0.42	0.01
Massa	0.50	-0.51	0.38	0.56
Recheio	0.51	-0.45	0.01	0.78

A partir da matriz de autovetores, obtém-se os componentes fatoriais a partir das seguintes combinações:

$Y_1 = 0.49X_1 + 0.48X_2 + 0.5X_3 + 0.51X_4$, explica 65.725% da variação total;

$Y_2 = 0.54X_1 + 0.76X_2 - 0.51X_3 - 0.45X_4$, explica 29.75% da variação total;

$Y_3 = -0.61X_1 + 0.42X_2 + 0.38X_3 + 0.01X_4$, explica 3.45% da variação total;

$Y_4 = 0.024X_1 - 0.001X_2 + 0.54X_3 - 0.78X_4$, explica 1.275% da variação total.

2.2

O que se espera de uma marca de lasanha com avaliação acima da média em todos os quesitos é um posicionamento mais próximo do componente principal 1 e um pouco mais elevado em relação ao eixo de CP1.

Isso porque a contribuição do primeiro componente principal é superior em relação aos outros 3, com 65.725% da variabilidade total sendo explicada por ela, seguido do 2° componente, com 29.75%.

Caso apareça uma marca com valores acima da média para Massa e Recheio e abaixo para sabor e aroma, a mesma será melhor explicada pelo segundo componente principal, no qual possui coeficientes positivos para as variáveis "Sabor" e "Aroma" e coeficientes negativos para "Massa" e "Recheio".

3 Exercício 3

A partir dos dados do problema e a partir da matriz de covariâncias, tem-se os seguintes autovalores:

Componente	**Autovalor**	**% da Variância**	**% Var acumulada**
CP1	3.63027606	90.756901%	90.756901%
CP2	0.19524697	4.881174%	95.63808%
CP3	0.13041172	3.260293%	98.89837%
CP4	0.04406525	1.101631%	100%

E os seguintes autovetores:

	CP1	CP2	CP3	CP4
X1	0.4991785	0.1531956	0.82505749	-0.215944711
X2	0.4877205	0.7561635	-0.43627548	-0.003018725
X3	0.5015674	-0.5317702	-0.35694167	-0.581586810
X4	0.5112531	-0.3492395	-0.03919805	0.784293026

Dessa forma, tem-se as componentes principais dadas por:

$$
\begin{aligned}
\boldsymbol{Y}_1 &= 0.4991785\boldsymbol{X}_1 + 0.4877205\boldsymbol{X_2} + 0.5015674\boldsymbol{X_3} + 0.5112531\boldsymbol{X_4} \\
\boldsymbol{Y}_2 &= 0.1531956\boldsymbol{X}_1 + 0.7561635\boldsymbol{X_2} - 0.5317702\boldsymbol{X_3} - 0.3492395\boldsymbol{X_4} \\
\boldsymbol{Y}_3 &= 0.82505749\boldsymbol{X}_1 - 0.43627548\boldsymbol{X_2} - 0.35694167\boldsymbol{X_3} - 0.03919805\boldsymbol{X_4} \\
\boldsymbol{Y}_4 &= -0.215944711\boldsymbol{X}_1 - 0.003018725\boldsymbol{X_2} - 0.581586810\boldsymbol{X_3} + 0.784293026\boldsymbol{X_4}
\end{aligned}
$$

Em seguida, é de nosso interesse visualizar o Scree Plot para entender a contribuição de cada componente para explicar a variação dos dados:

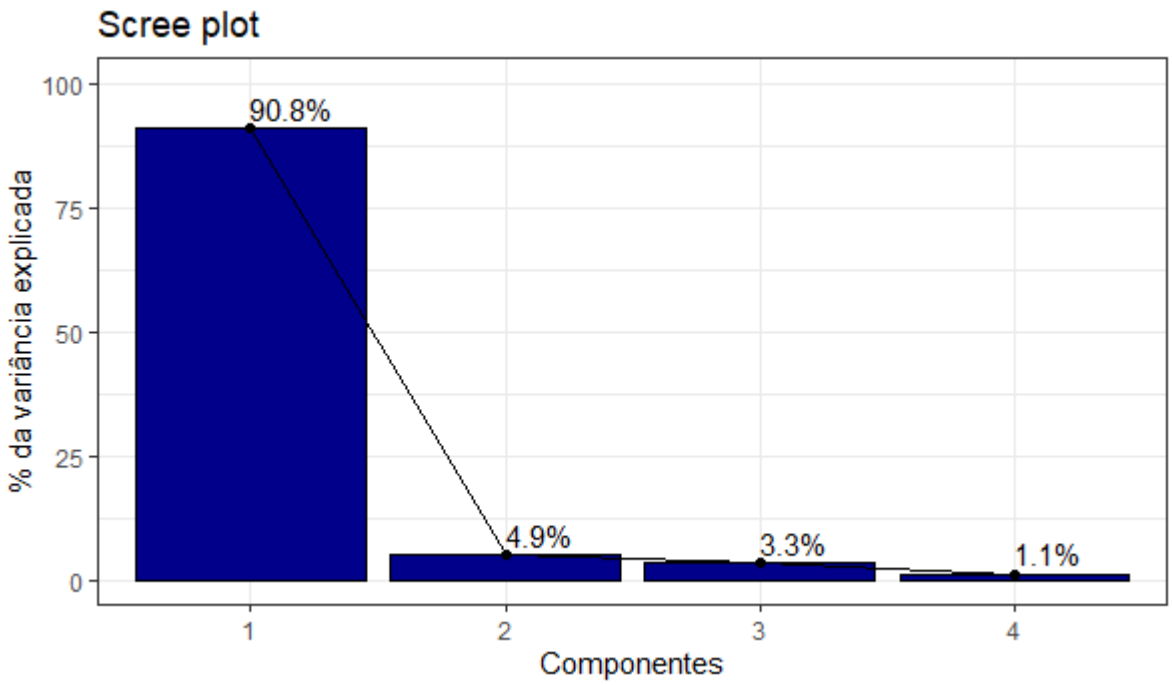

Figura 2: Scree Plot para os componentes principais identificados

A partir do "Scree Plot"visualizado, é possível perceber que os 2 primeiros principais componentes são capazes de explicar aproximadamente 96% da variância dos dados originais. Com somente a primeira componente sendo responsável por quase 91% da variação total.

Dessa forma, é possível explicar 96% da variação do sistema apenas com as 2 primeiras componentes principais, formando assim o primeiro plano.

Assim, é de nosso interesse identificar o quanto as variáveis são capazes de contribuir

para o primeiro plano a ser formado pelas 2 primeiras componentes principais.

Para isso, faz-se um gráfico de contribuição das variáveis para o primeiro plano:

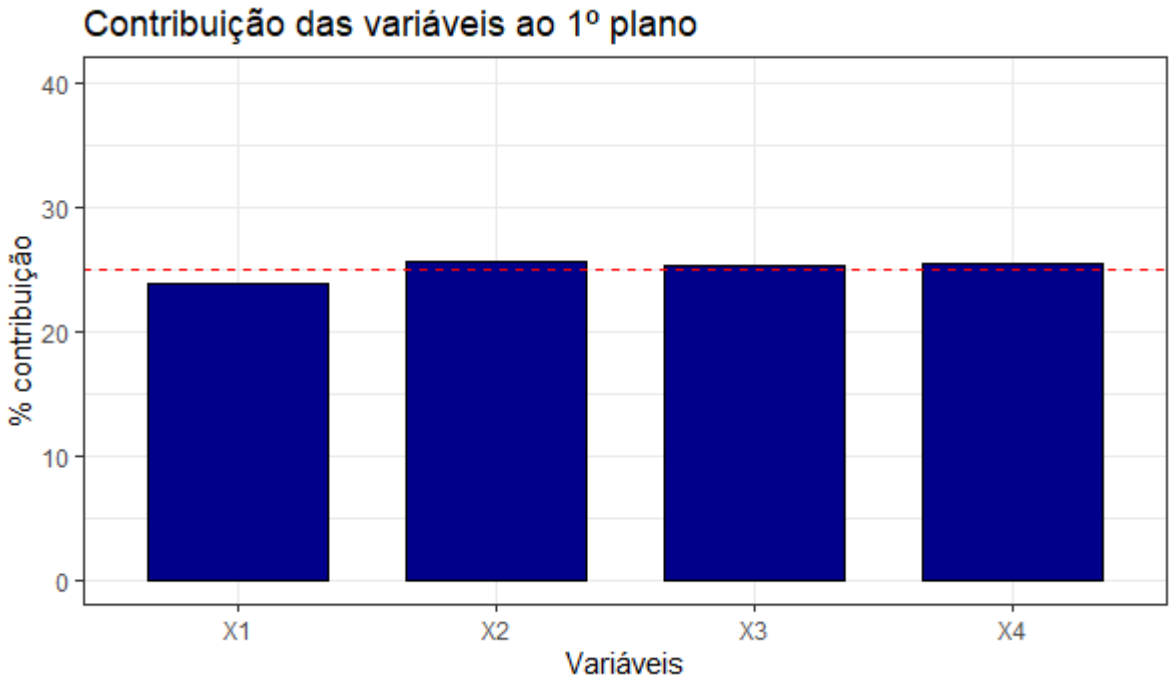

Figura 3: Contribuição das variáveis para o primeiro plano

A partir do gráfico de contribuição para o primeiro plano, é possível identificar uma contribuição semelhante para cada uma das variáveis, podendo ser causa da alta correlação encontrada para cada uma das variáveis presentes.

Por fim, é interessante visualizar os scores fatoriais gerados pela transformação em componentes principais dentro do primeiro plano formado pelas duas primeiras componentes principais (As componentes com maior poder de explicar a variação dos dados).

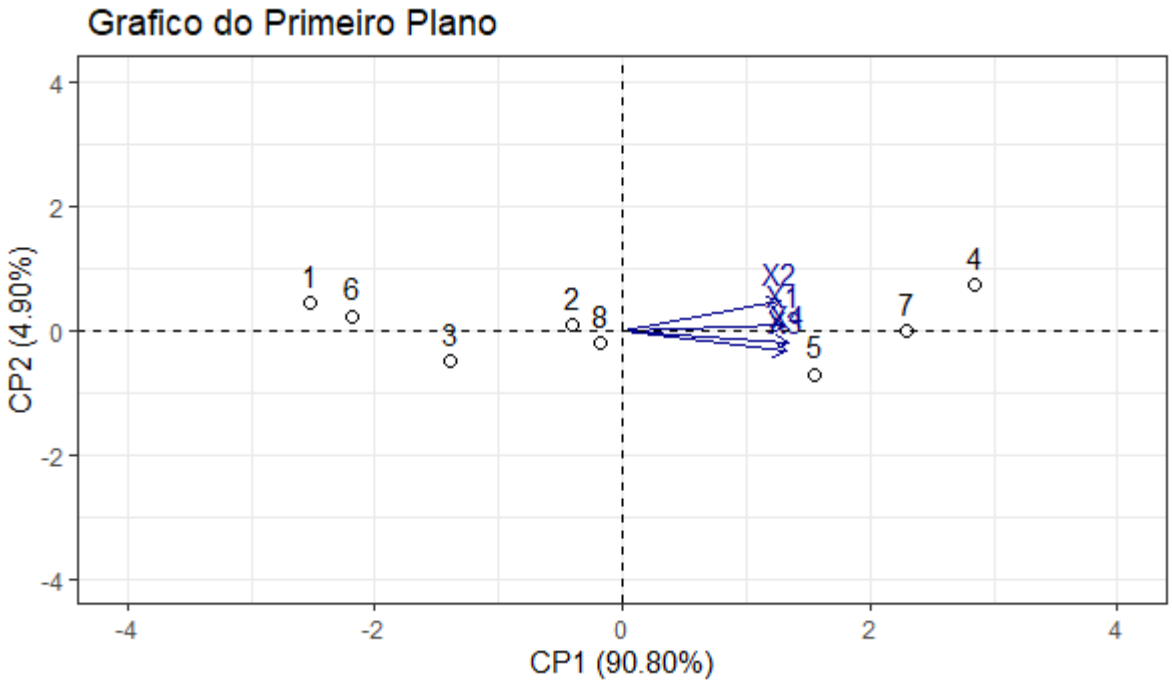

Figura 4: Observações distribuídas na escala dos Componentes Principais

A partir do gráfico, percebe-se que as variáveis presentes contribuem quase que somente para o componente principal 1, de forma que todas tem direção e sentido de acordo com o CP1.

4 Exercício 4

Primeiramente, faz-se um gráfico de correlação para todas as variáveis presentes no estudo a fim de entender a relação presente nas variáveis de estudo, de forma que:

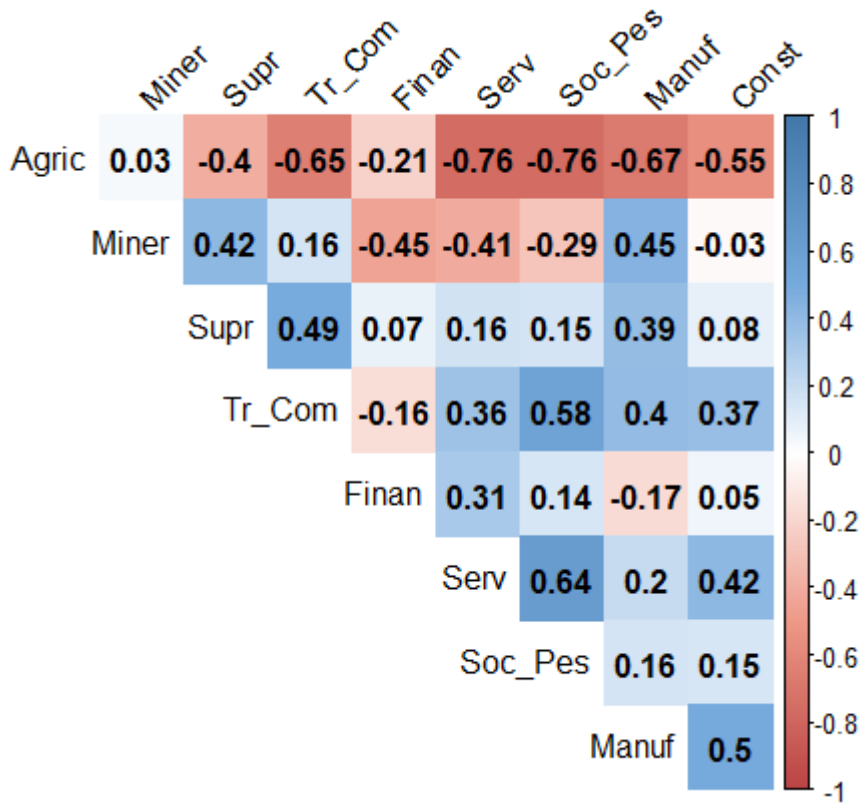

Figura 5: Matriz de Correlação para as variáveis quantitativas de estudo

Para ter uma noção de como nossas variáveis se correlacionam, faz-se uma matriz de correlação e, a partir do gráfico, vemos que a variável "Percentual de empregados na área agrícola"tem grande correlação negativa com as variáveis TrCom, Serv, SocPes, Manuf, Const. Já quanto as demais variáveis, nota-se que não há uma correlação tão significativa para as restantes.

Partindo para análise de componentes principais, é de nosso interesse entender o quanto nossos componentes são capazes de explicar a variação dos valores presente nos dados de estudo:

Componente	Autovalor	% da Variância	% Var acumulada
CP1	3.655385	40.61539%	40.61539%
CP2	2.144673	23.8297%	64.44509%
CP3	1.025562	11.39513%	75.84022%
CP4	0.941107	10.45675%	86.29697%
CP5	0.505636	5.61818%	91.91515%
CP6	0.352258	3.91397%	95.82912%
CP7	0.228081	2.53424%	98.36336%
CP8	0.147254	1.63615%	99.99951%
CP9	0.000044	0,00049%	100%

Tabela 4: Contribuição de cada componente para explicar a variância original do primeiro conjunto de variáveis

Graficamente, visualiza-se as contribuições de cada componente para a variação total através do "Scree Plot":

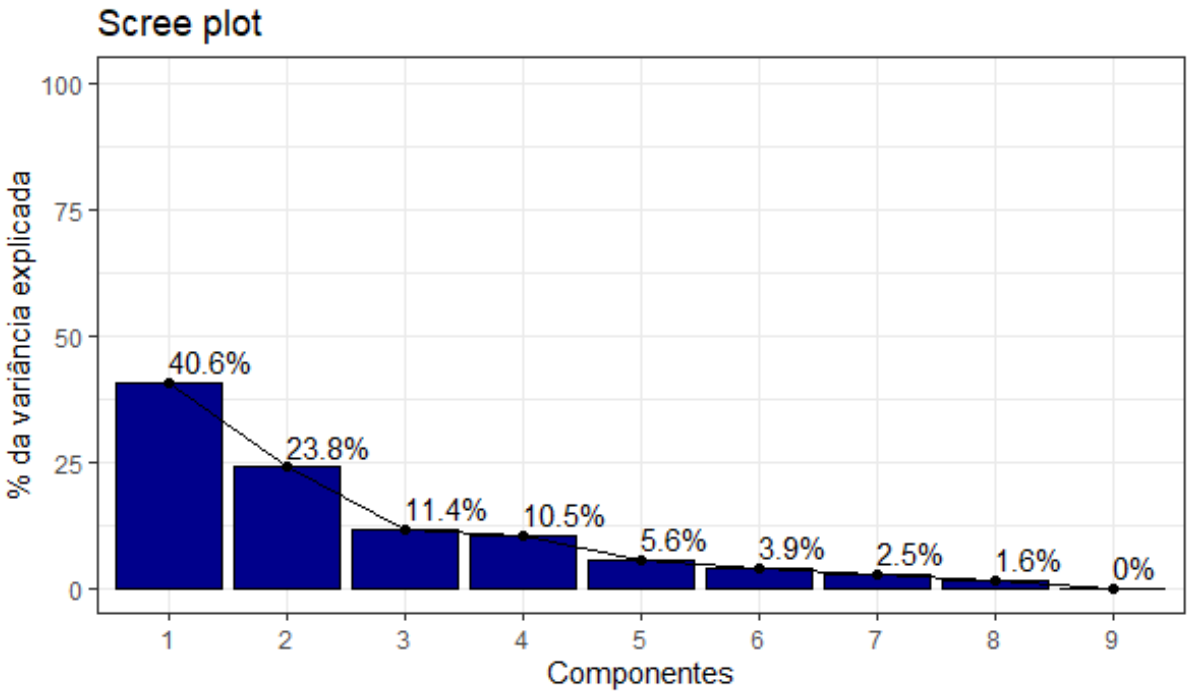

Figura 6: Scree Plot para os componentes principais identificados

A partir da tabela e do gráfico ”Scree Plot”, é possível identificar que, embora o ”cotovelo”

seja formado na terceira componente principal, é necessária também a inclusão da quarta componente principal, fazendo com que a composição das 4 primeiras componentes principais sejam suficientes para explicar aproximadamente 86% da variância original dos dados.

Em seguida, como permitido pela ACP, é de nosso interesse identificar a influência das variáveis em cada um dos componentes. Nesse caso, será considerado apenas o primeiro plano (Composto apenas pela CP1 e CP2, que são as mais significantes) para entender a contribuição de cada uma das variáveis. Assim, o gráfico de contribuição das variáveis no primeiro plano é dado por:

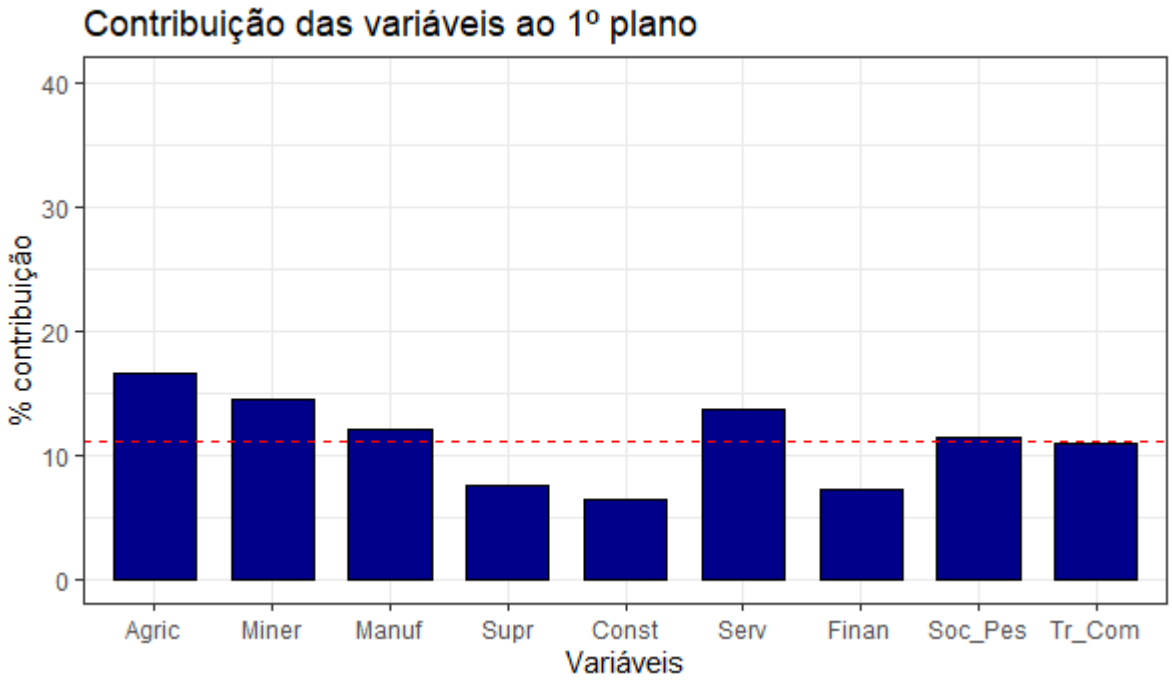

Figura 7: Gráfico de contribuição das variáveis para o primeiro plano

A partir do gráfico de contribuição para o primeiro plano, é possível notar que a maioria das variáveis possuem uma contribuição semelhante, seguidas pela primeira variável (Setor da agricultura). Contudo, visualiza-se 3 variáveis que não são tão significantes para a formação do primeiro plano, sendo elas "Supr", "Const"e "Finan".

É possível notar que as variáveis que mostram-se menos contributivas para o primeiro plano, o que pode ser explicado pela baixa correlação visualizada no ínicio.

Como o primeiro plano não é capaz de explicar uma porcentagem que consideremos suficiente para ACP, é de nosso interesse entender como as variáveis contribuem para cada

Componente principal obtido, assim, será feito gráficos de contribuição para cada uma das dimensões, de forma que:

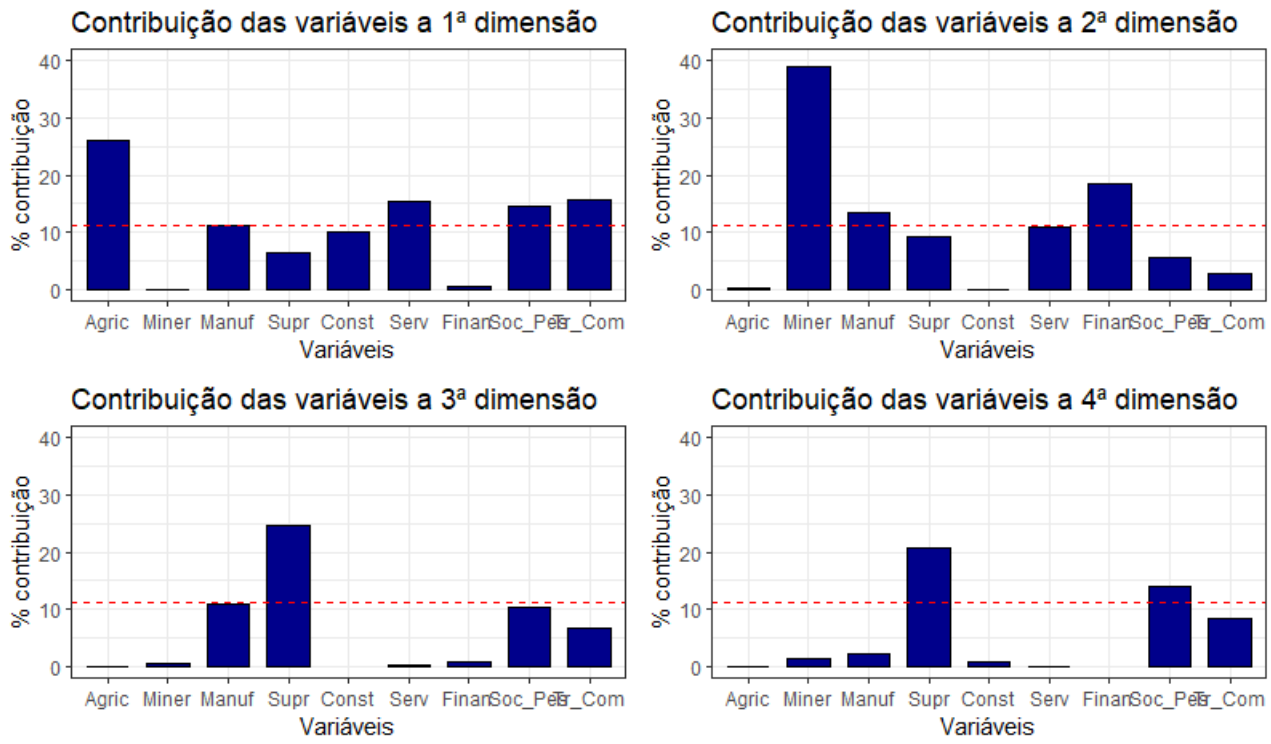

Figura 8: Gráfico de Contribuição para as 4 Componentes principais mais significantes

Nota-se que "Percentual de empregados na área de mineração" possui maior contribuição para a dimensão 2, enquanto que "Percentual de empregados na área agrícola" é a maior contribuição da primeira. Para as 3° e 4° dimensões, "Percentual de empregados na área de suprimentos" é a variável que se destaca em termos de contribuição.

Em seguida, para visualizar a influência das nossas variáveis nos principais componentes, faz-se o gráfico do primeiro plano:

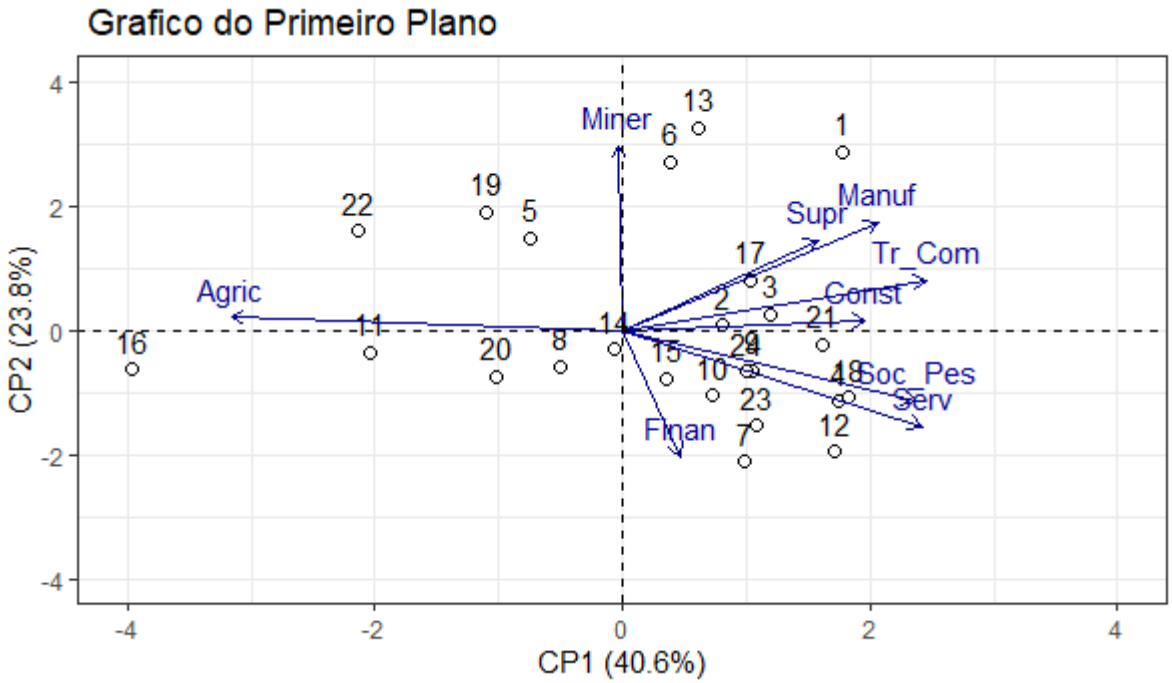

Figura 9: Observações distribuídas na escala dos Componentes principais

A partir do gráfico do primeiro plano, é possível visualizar que as variáveis contribuem de forma semelhante na maioria dos casos, contudo, é possível notar a variável "Miner" contribuindo apenas para a componente principal 2, e a variável "Agric" contribuindo somente para a CP1, contudo, de maneira negativa, o que pode ser indicado pela forte correlação negativa visualizada na matriz de correlações apresentada no início.
A variável "Finan" contribui maioritariamente para PC2 negativamente, e uma parcela pequena de contribuição positiva para PC1. Todas as restantes contribuem positivamente para PC1, e variam no quesito contribuição positiva ou negativa para PC2.

Deve-se destacar de fato em como o percentual de empregados em agrigultura e em mineração se contrapõem em relação ao restante das outras variáveis.

Códigos

```r
library(factoextra)
library(FactoMineR)
library(ggplot2)
library(tidyverse)
library(corrplot)
library(readxl)



# ex1


x_a = c(8209.5,538.16,13375.7)
x_b = c(4326.3,316.3,5837.1)
cp1 = c(0.617, 0.557, 0.556)
cp2 = c(-0.001,-0.706,0.708)

y_a = c(sum(x_a*cp1),sum(x_a*cp2))
y_b = c(sum(x_b*cp1),sum(x_b*cp2))

## ex 3



M = c("M1", "M2", "M3", "M4", "M5", "M6", "M7", "M8")
X1 = c(2.75, 3.9, 3.12, 4.58, 3.97, 3.01, 4.19, 3.82)
X2 = c(4.03, 4.12, 3.97, 4.86, 4.34, 3.98, 4.65, 4.12)
X3 = c(2.8, 3.4, 3.62, 4.34, 4.28, 2.9, 4.52, 3.62)
X4 = c(2.62, 3.52, 3.05, 4.82, 4.98, 2.82, 4.77, 3.71)
d3 = data.frame(M, X1, X2, X3, X4)
```

```r
cor(d3[,-1])

acp3 <- PCA(d3[,-1])

corrplot(cor(d3[,-1]), method = "color", col = col(200),
         type = "upper", order = "hclust",
         addCoef.col = "black", tl.col = "black",
         tl.srt = 45, sig.level = 0.01, diag = FALSE, tl.cex = 0.75)

fviz_eig(acp3, addlabels = T, barfill = 'darkblue',
         barcolor = 'black') + theme_bw() + labs(x = 'Componentes',
         y = '% da variância explicada') +
  ylim(c(0, 100))

fviz_contrib(acp3, choice = 'var', axes = 1:2, fill = 'darkblue',
             color = 'black', sort.val = 'none') +
  ylim(c(0, 40)) +
  labs(title = 'Contribuição das variáveis ao 1º plano', x = 'Variáveis',
       y = '% contribuição') +
  theme_bw()

fviz_pca_biplot(acp3, pointshape = 21, pointsize = 2,
                col.var = 'darkblue') +
  theme_bw() +
  labs(title = " Grafico do Primeiro Plano", x = 'CP1 (90.80%)',
       y = 'CP2 (4.90%)') +
  xlim(c(-4, 4)) +
  ylim(c(-4, 4))
```

```r
## ex 4

d4 <- read_excel("C:/Users/rafa_/Downloads/L24.xlsx")

d4$Agric <- as.numeric(d4$Agric)
d4$Miner <- as.numeric(d4$Miner)
d4$Manuf <- as.numeric(d4$Manuf)
d4$Supr <- as.numeric(d4$Supr)
d4$Const <- as.numeric(d4$Const)
d4$Serv <- as.numeric(d4$Serv)
d4$Finan <- as.numeric(d4$Finan)
d4$Soc_Pes <- as.numeric(d4$Soc_Pes)
d4$Tr_Com <- as.numeric(d4$Tr_Com)


#matriz de corr

CM <- cor(d4[,-1])

col <- colorRampPalette(c("#BB4444", "#EE9988",
                          "#FFFFFF", "#77AADD", "#4477AA"))

g_c <- corrplot(CM, method = "color", col = col(200),
         type = "upper", order = "hclust",
         addCoef.col = "black", tl.col = "black",
         tl.srt = 45, sig.level = 0.01, diag = FALSE, tl.cex = 0.75)

# ACP
```

```r
acp <- PCA(d4[,-1])

## scree plot

fviz_eig(acp, addlabels = T, barfill = 'darkblue',
         barcolor = 'black') + theme_bw() + labs(x = 'Componentes',
         y = '% da variância explicada') +
  ylim(c(0, 100))

## contribuição no primeiro plano
fviz_contrib(acp, choice = 'var', axes = 1:2, fill = 'darkblue',
             color = 'black', sort.val = 'none') +
  ylim(c(0, 40)) +
  labs(title = 'Contribuição das variáveis ao 1º plano', x = 'Variáveis',
       y = '% contribuição') +
  theme_bw()


## gráfico do primeiro plano

fviz_pca_biplot(acp, pointshape = 21, pointsize = 2,
                col.var = 'darkblue') +
  theme_bw() +
  labs(title = " Grafico do Primeiro Plano", x = 'CP1 (40.6%)',
       y = 'CP2 (23.8%)') +
  xlim(c(-4, 4)) +
  ylim(c(-4, 4))
```

```
 #contribuição por dimensão

p1 = fviz_contrib(acp, choice="var", axes=1, fill = 'darkblue',
             color = 'black', sort.val = 'none') + ylim(c(0, 40)) +
  labs(title = 'Contribuição das variáveis a 1ª dimensão',
       x = 'Variáveis',
       y = '% contribuição') +
  theme_bw()

p2 = fviz_contrib(acp, choice="var", axes=2,  fill = 'darkblue',
             color = 'black', sort.val = 'none') + ylim(c(0, 40)) +
  labs(title = 'Contribuição das variáveis a 2ª dimensão',
       x = 'Variáveis',
       y = '% contribuição') +
  theme_bw()

p3 = fviz_contrib(acp, choice="var", axes=3, fill = 'darkblue',
             color = 'black', sort.val = 'none') + ylim(c(0, 40)) +
  labs(title = 'Contribuição das variáveis a 3ª dimensão',
       x = 'Variáveis',
       y = '% contribuição') +
  theme_bw()

p4 = fviz_contrib(acp, choice="var", axes=4,  fill = 'darkblue',
             color = 'black', sort.val = 'none') + ylim(c(0, 40)) +
  labs(title = 'Contribuição das variáveis a 4ª dimensão',
       x = 'Variáveis',
       y = '% contribuição') +
  theme_bw()
```

```r
options(repr.plot.width=14, repr.plot.height=10)
plot_grid(p1,p2,p3,p4, ncol = 2, nrow = 2)
```