Análise de Correspondência

Laís Botelho e Mariana Neves

Definição: Técnica exploratória para mapeamento perceptual baseada em categorias de uma tabela de contingência (HAIR,2009).

Análise de Correspondência no R:

- No pacote FactoMineR:CA()
- No pacote ca:ca()
- No pacote ade4:dudi.coa()
- No pacote MASS:corresp()
- No pacote ExPosition:epCA()

Pacotes utilizados:

Utilizou-se o pacote **FactoMineR** para realizar as análises e o pacote **factoextra** para visualização em ggplot2.

```
#install.packages("factoextra")
#install.packages("FactoMineR")
#install.packages("gplots")
```

```
library(factoextra)
library(FactoMineR)
library(gplots)
```

Banco utilizado

```
data(housetasks)
str(housetasks)
```

```
## 'data.frame':    13 obs. of  4 variables:
##  $ Wife       : int  156 124 77 82 53 32 33 12 10 13 ...
##  $ Alternating: int  14 20 11 36 11 24 23 46 51 13 ...
##  $ Husband    : int  2 5 7 15 1 4 9 23 75 21 ...
##  $ Jointly    : int  4 4 13 7 57 53 55 15 3 66 ...
```

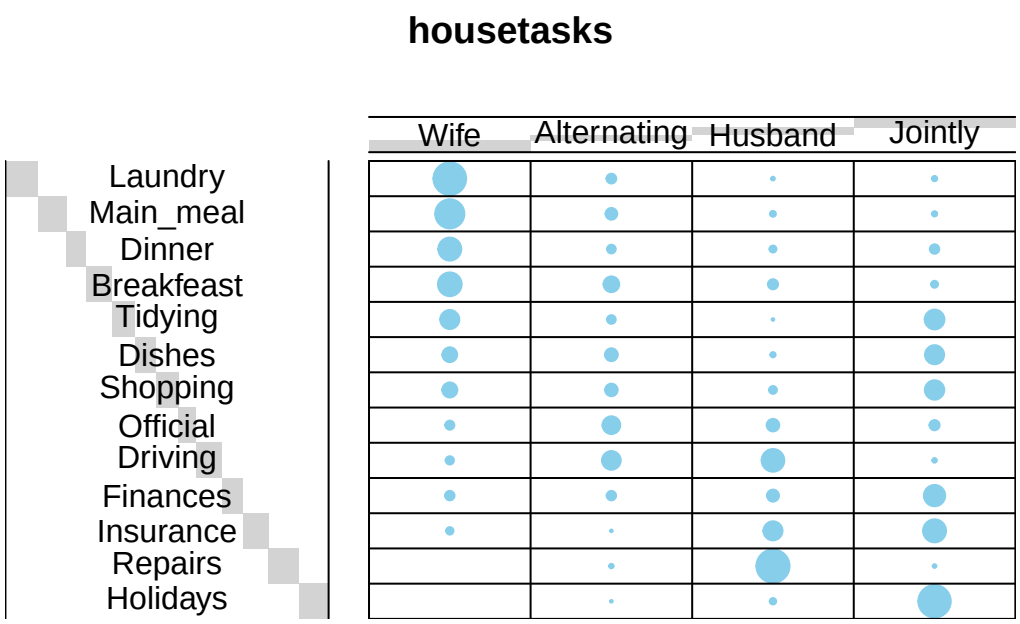
```
View(housetasks)
```

```
#Convertendo os dados do pacote em uma tabela
dt <- as.table(as.matrix(housetasks))
```

Tabela de contingência Gráfico balloonplot

Neste gráfico, a célula contém um ponto cujo tamanho reflete a magnitude relativa do componente correspondente.

```
balloonplot(t(dt), main ="housetasks", xlab ="", ylab="",
            label = FALSE, show.margins = FALSE)
```




```
res.ca <- CA(housetasks, graph = FALSE)
res.ca
```

1. Cálculo da análise de correspondência

```
## **Results of the Correspondence Analysis (CA)**
## The row variable has  13  categories; the column variable has 4 categories
## The chi square of independence between the two variables is equal to 1944.456 (p-value =  0 ).
## *The results are available in the following objects:
```

```
##
##    name              description
## 1  "$eig"            "eigenvalues"
## 2  "$col"            "results for the columns"
## 3  "$col$coord"      "coord. for the columns"
## 4  "$col$cos2"       "cos2 for the columns"
## 5  "$col$contrib"    "contributions of the columns"
## 6  "$row"            "results for the rows"
## 7  "$row$coord"      "coord. for the rows"
## 8  "$row$cos2"       "cos2 for the rows"
## 9  "$row$contrib"    "contributions of the rows"
## 10 "$call"           "summary called parameters"
## 11 "$call$marge.col" "weights of the columns"
## 12 "$call$marge.row" "weights of the rows"
```

2. Significância estatística: Teste qui-quadrado É utilizado para avaliar se há uma dependência significativa entre as linhas e colunas.

Neste exemplo, o qui-quadrado de independência entre as duas variáveis foi igual a 1944,456 (p-valor = 0), tendo, portanto, associação significativa.

```
summary(res.ca)
```

```
##
## Call:
## CA(X = housetasks, graph = FALSE)
##
## The chi square of independence between the two variables is equal to 1944.456 (p-value =  0 ).
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3
## Variance               0.543   0.445   0.127
## % of var.             48.692  39.913  11.395
## Cumulative % of var.  48.692  88.605 100.000
##
## Rows (the 10 first)
##               Iner*1000     Dim.1     ctr    cos2     Dim.2     ctr    cos2
## Laundry     |   134.160 |  -0.992  18.287   0.740 |   0.495   5.564   0.185 |
## Main_meal   |    90.692 |  -0.876  12.389   0.742 |   0.490   4.736   0.232 |
## Dinner      |    38.246 |  -0.693   5.471   0.777 |   0.308   1.321   0.154 |
## Breakfeast  |    41.124 |  -0.509   3.825   0.505 |   0.453   3.699   0.400 |
## Tidying     |    24.667 |  -0.394   1.998   0.440 |  -0.434   2.966   0.535 |
## Dishes      |    19.587 |  -0.189   0.426   0.118 |  -0.442   2.844   0.646 |
## Shopping    |    14.970 |  -0.118   0.176   0.064 |  -0.403   2.515   0.748 |
## Official    |    53.300 |   0.227   0.521   0.053 |   0.254   0.796   0.066 |
## Driving     |   101.509 |   0.742   8.078   0.432 |   0.653   7.647   0.335 |
## Finances    |    29.564 |   0.271   0.875   0.161 |  -0.618   5.559   0.837 |
##               Dim.3     ctr    cos2
## Laundry      -0.317   7.968   0.075 |
## Main_meal    -0.164   1.859   0.026 |
## Dinner       -0.207   2.097   0.070 |
## Breakfeast    0.220   3.069   0.095 |
## Tidying      -0.094   0.489   0.025 |
```

```
## Dishes        0.267   3.634   0.236 |
## Shopping      0.203   2.223   0.189 |
## Official      0.923  36.940   0.881 |
## Driving       0.544  18.596   0.233 |
## Finances      0.035   0.062   0.003 |
##
## Columns
##               Iner*1000     Dim.1     ctr    cos2     Dim.2     ctr    cos2
## Wife        |   301.019 |  -0.838  44.462   0.802 |   0.365  10.312   0.152 |
## Alternating |   117.824 |  -0.062   0.104   0.005 |   0.292   2.783   0.105 |
## Husband     |   381.373 |   1.161  54.234   0.772 |   0.602  17.787   0.208 |
## Jointly     |   314.725 |   0.149   1.200   0.021 |  -1.027  69.118   0.977 |
##               Dim.3     ctr    cos2
## Wife         -0.200  10.822   0.046 |
## Alternating   0.849  82.549   0.890 |
## Husband      -0.189   6.133   0.020 |
## Jointly      -0.046   0.495   0.002 |
```

3.Dimensões Segundo HAIR (2009), o número de dimensões a serem mantidas na solução se baseia em:

- Dimensões com inércia (autovalores) maiores que 0,2.
- Dimensões suficientes para atender os objetivos da pesquisa (geralmente duas ou três).

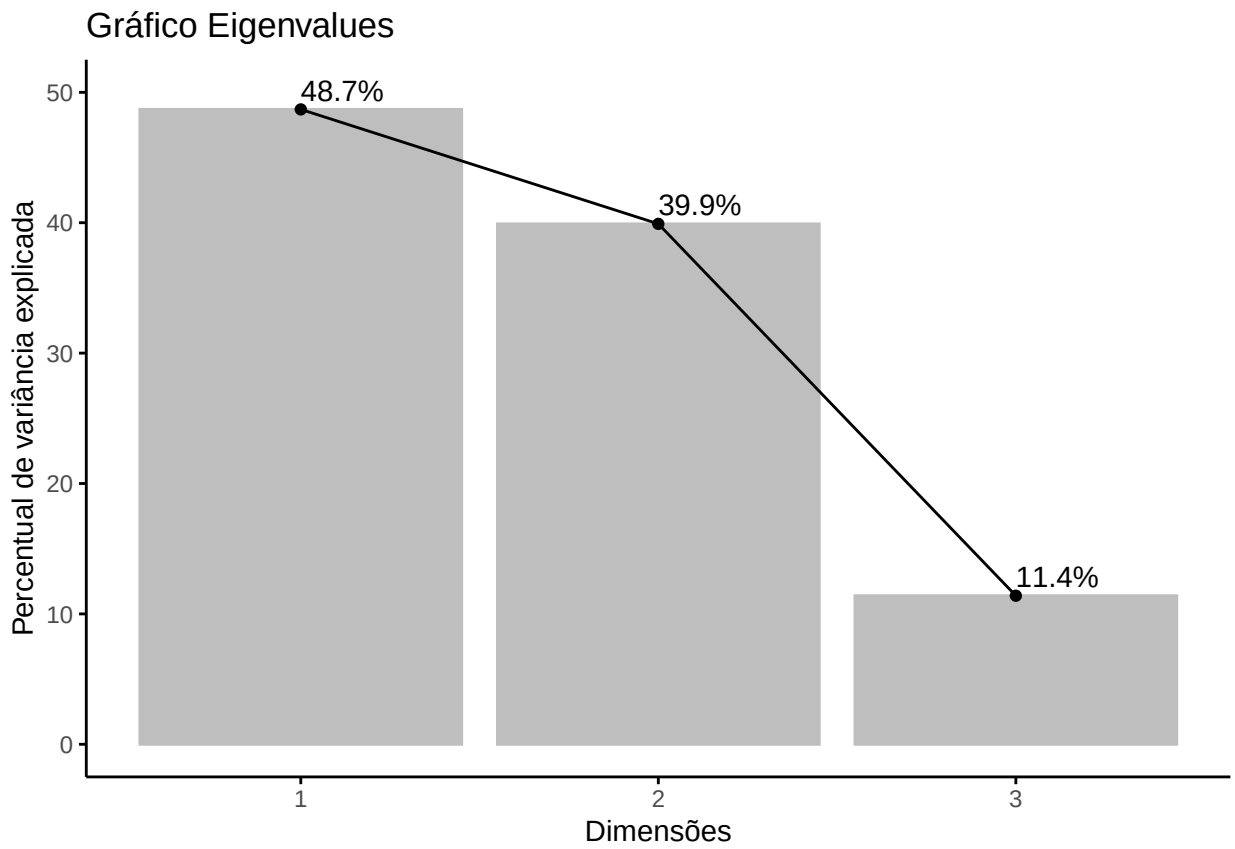
```
eig.val <- get_eigenvalue(res.ca)
eig.val
```

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1  0.5428893         48.69222                    48.69222
## Dim.2  0.4450028         39.91269                    88.60491
## Dim.3  0.1270484         11.39509                   100.00000
```

A dimensão 1 explica a maior variação, seguida pela dimensão 2 e assim por diante. A primeira dimensão representa 48.7% da variação. Cerca de 88,6% da variação é explicada pelas duas primeiras dimensões.

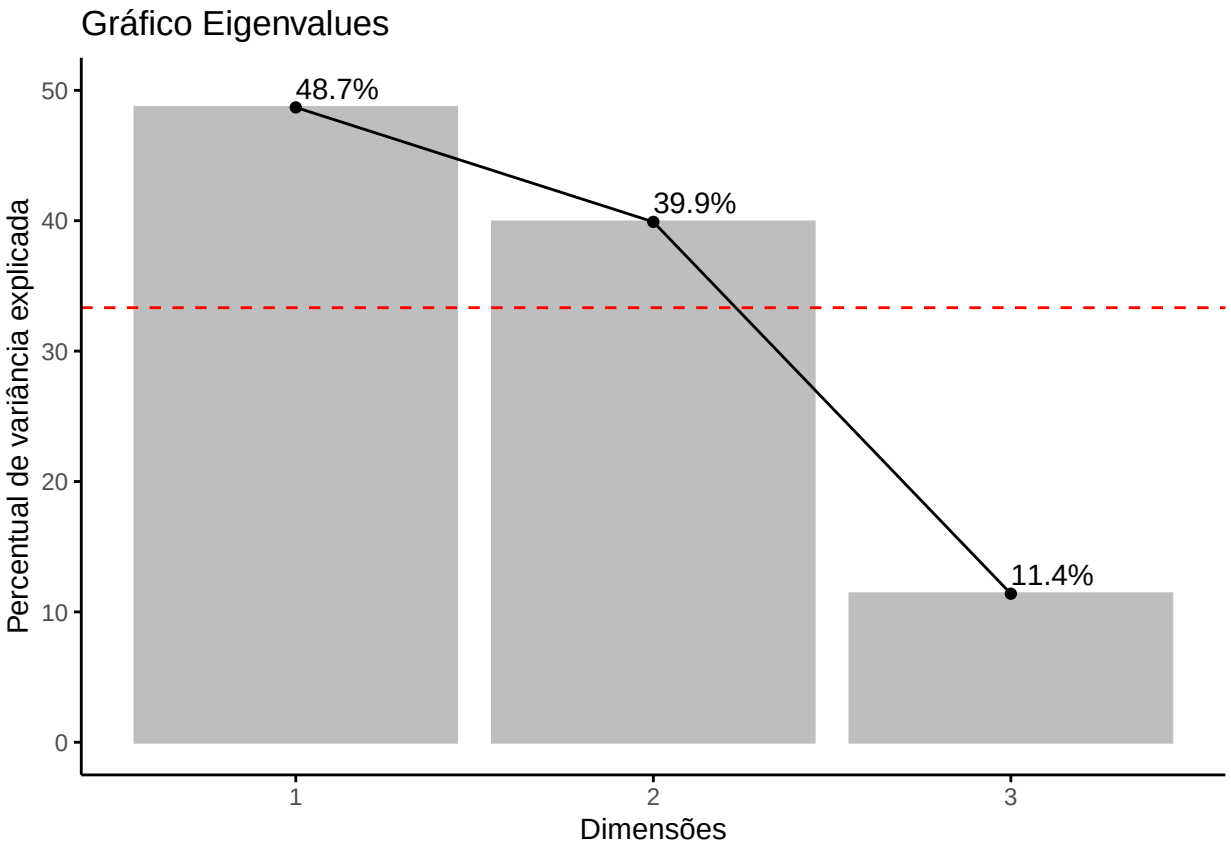
O número de dimensões pode ser visualizado através do Scree Plot. Este gráfico ordena as dimensões da maior para o menor. O ponto em que gráfico mostra uma curva (o chamado "cotovelo") pode ser considerado como indicando uma dimensionalidade ideal.

```
g1<- fviz_screeplot(res.ca, addlabels = TRUE, ylim = c(0, 50),
               main = "Gráfico Eigenvalues",
               barfill = "Gray", barcolor = "Gray",linecolor = "Black",
               xlab = "Dimensões",
               ylab= "Percentual de variância explicada",
               ggtheme = theme_classic())
g1
```



Screeplot com valor médio de eigen em linha tracejada:

```
g2<- g1+
 geom_hline(yintercept=33.33, linetype=2, color="red")
g2
```



Interpretação:

O gráfico demonstra que é possível utilizar apenas as dimensões 1 e 2 , pois a dimensão 3 explica apenas 11,4% da inércia total, abaixo do autovalor médio (33,33%). As dimensões 1 e 2 contribuem significativamente para a interpretação da natureza da associação entre as linhas e colunas, porque correspondem a cerca de 88,6% da variação.

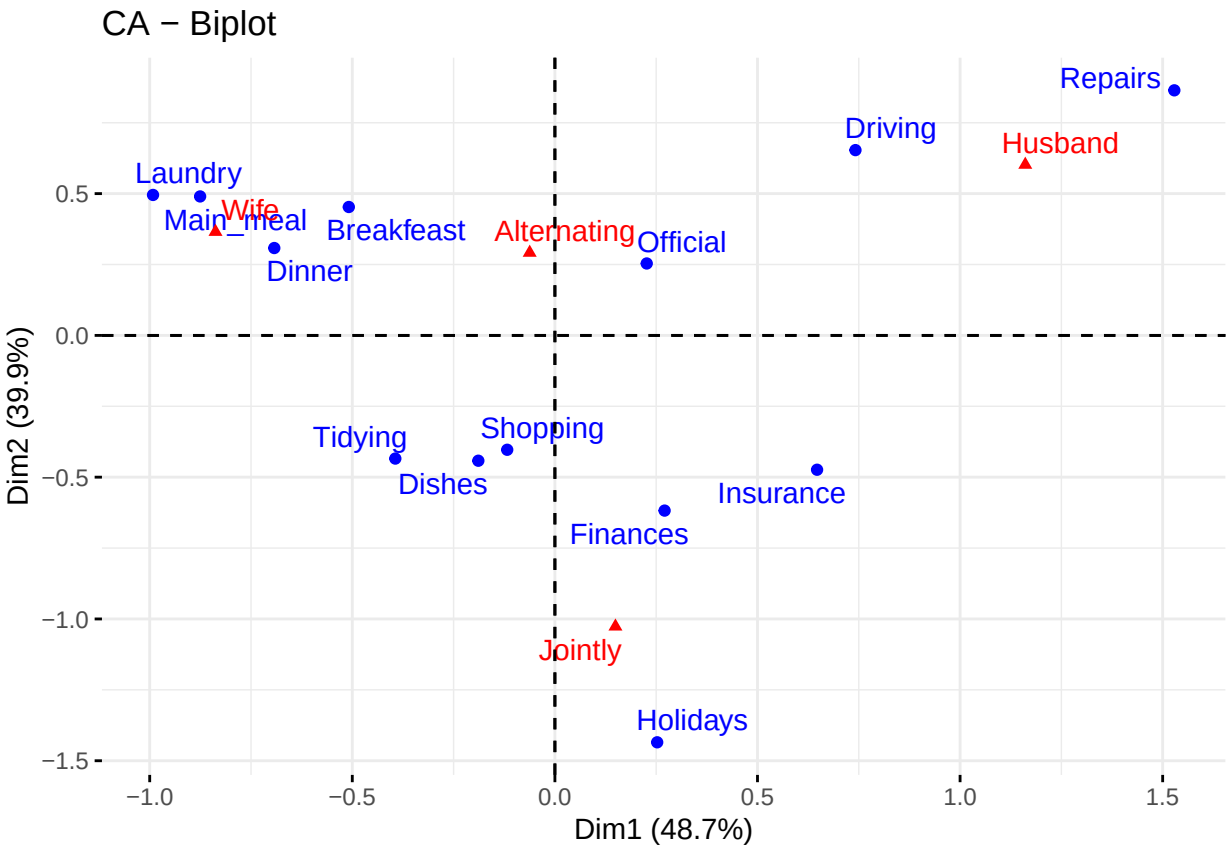
4.Avaliação da associação de colunas e linhas "CA cria um mapa perceptual usando a medida padronizada para estimar dimensões ortogonais sobre as quais as categorias podem ser colocadas para explicar melhor a intensidade de associação representada pelas distâncias qui-quadrado" (HAIR,2009).

Biplot

- Biplot é uma exibição gráfica de linhas e colunas segundo dimensões.
- A proximidade indica o nível de associação entre as categorias linha ou coluna (HAIR,2009).
- As linhas são representadas por pontos azuis e colunas por triângulos vermelhos.

Biplot de variáveis de linha e coluna:

```
fviz_ca_biplot(res.ca, repel = TRUE)
```



Interpretação:

Este gráfico mostra que as tarefas domésticas como preparo do jantar, café da manhã e lavanderia são feitas com mais frequência pela esposa. Condução e reparos são feitos pelo marido.

Obs: Este é um tipo de gráfico **Biplot simétrico**. Ele representa os perfis de linha e coluna simultaneamente em um espaço comum. Neste caso, apenas a distância entre os pontos de linha **OU** a distância entre os pontos da coluna podem ser realmente interpretadas.

4.1 Linhas Extraindo os resultados apenas para as linhas:

```
row <- get_ca_row(res.ca)
row
```

```
## Correspondence Analysis - Results for rows
##  ===================================================
##   Name       Description
## 1 "$coord"   "Coordinates for the rows"
## 2 "$cos2"    "Cos2 for the rows"
## 3 "$contrib" "contributions of the rows"
## 4 "$inertia" "Inertia of the rows"
```

Coordenadas de cada ponto de linha em cada dimensão (1, 2 e 3):

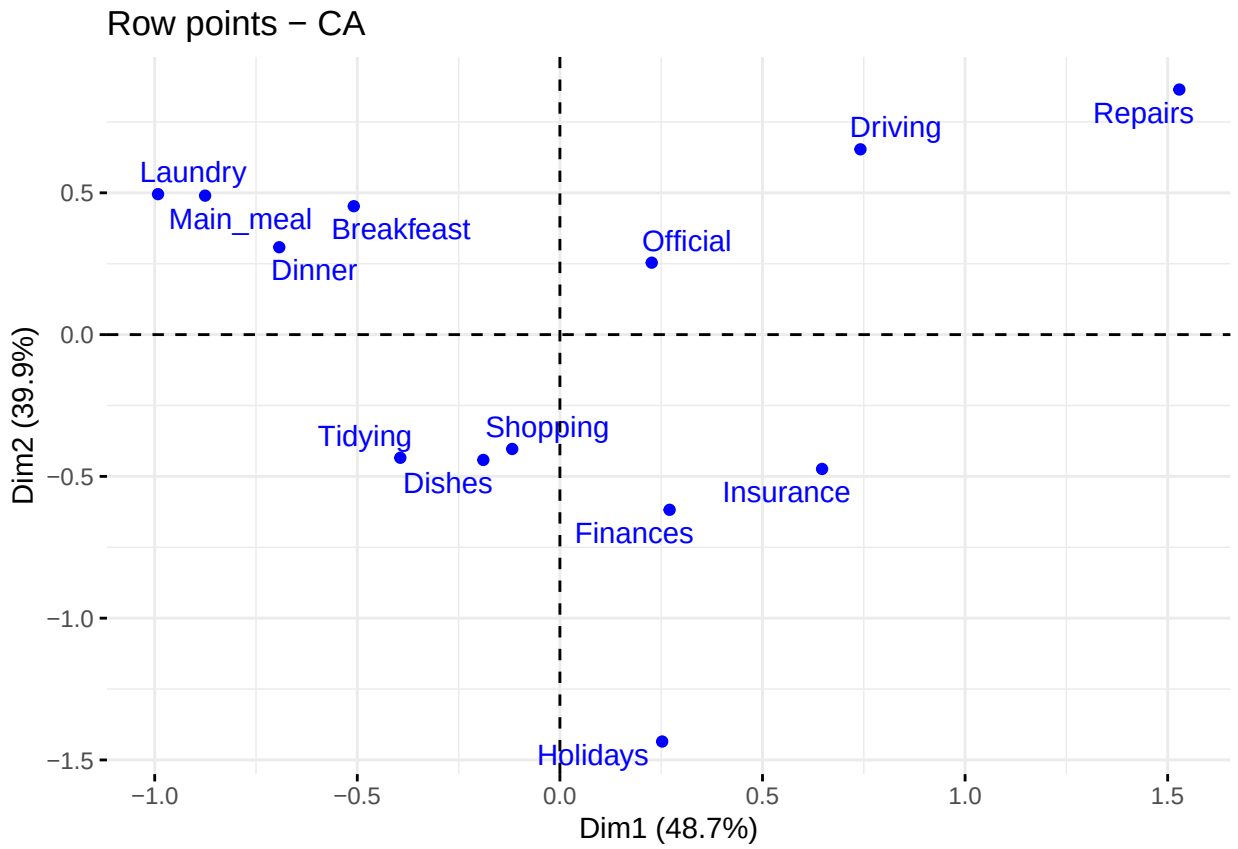
```
head(row$coord)
```

```
##                 Dim 1      Dim 2       Dim 3
## Laundry    -0.9918368  0.4953220 -0.31672897
## Main_meal  -0.8755855  0.4901092 -0.16406487
## Dinner     -0.6925740  0.3081043 -0.20741377
## Breakfeast -0.5086002  0.4528038  0.22040453
## Tidying    -0.3938084 -0.4343444 -0.09421375
## Dishes     -0.1889641 -0.4419662  0.26694926
```

Coordenadas de pontos de linha:

- Linhas com perfil semelhante são agrupadas.
- Linhas negativamente correlacionadas são posicionadas em lados opostos do quadrante.

```
fviz_ca_row(res.ca, repel = TRUE)
```



Qualidade de representação das linhas:

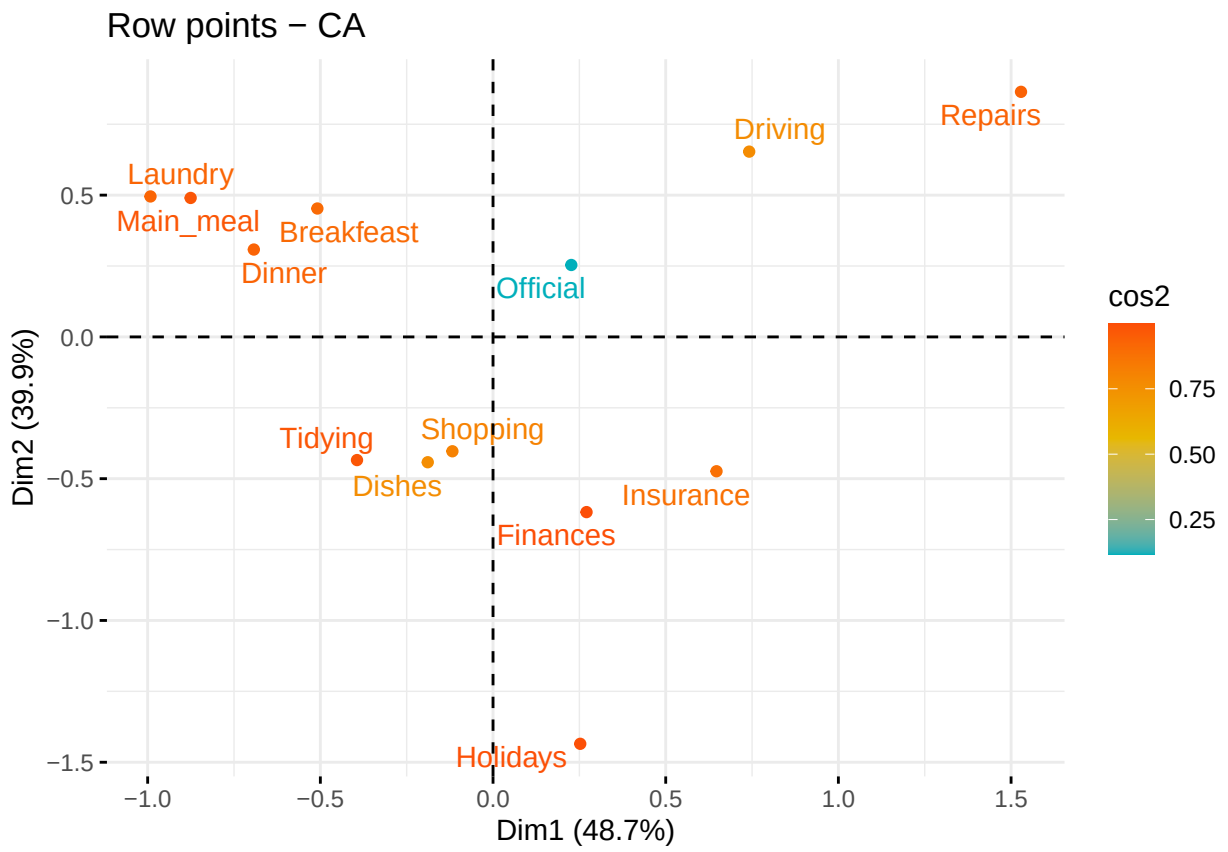
A soma do cos2 para linhas em todas as dimensões é igual a 1. Para alguns dos itens da linha, mais de 2 dimensões são necessárias para representar perfeitamente os dados.

```
head(row$cos2)
```

```
##                Dim 1     Dim 2      Dim 3
## Laundry    0.7399874 0.1845521 0.07546047
## Main_meal  0.7416028 0.2323593 0.02603787
## Dinner     0.7766401 0.1537032 0.06965666
## Breakfeast 0.5049433 0.4002300 0.09482670
## Tidying    0.4398124 0.5350151 0.02517249
## Dishes     0.1181178 0.6461525 0.23572969
```

Gráfico de variáveis de linha:

```
fviz_ca_row(res.ca, col.row = "cos2",
            gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
            repel = TRUE)
```



Contribuição de linhas (em %) para a definição das dimensões:

- As variáveis de linha com maior valor, contribuem mais para a definição das dimensões.
- As linhas que mais contribuem para Dim.1 e Dim.2 são as mais importantes para explicar a variabilidade no conjunto de dados.

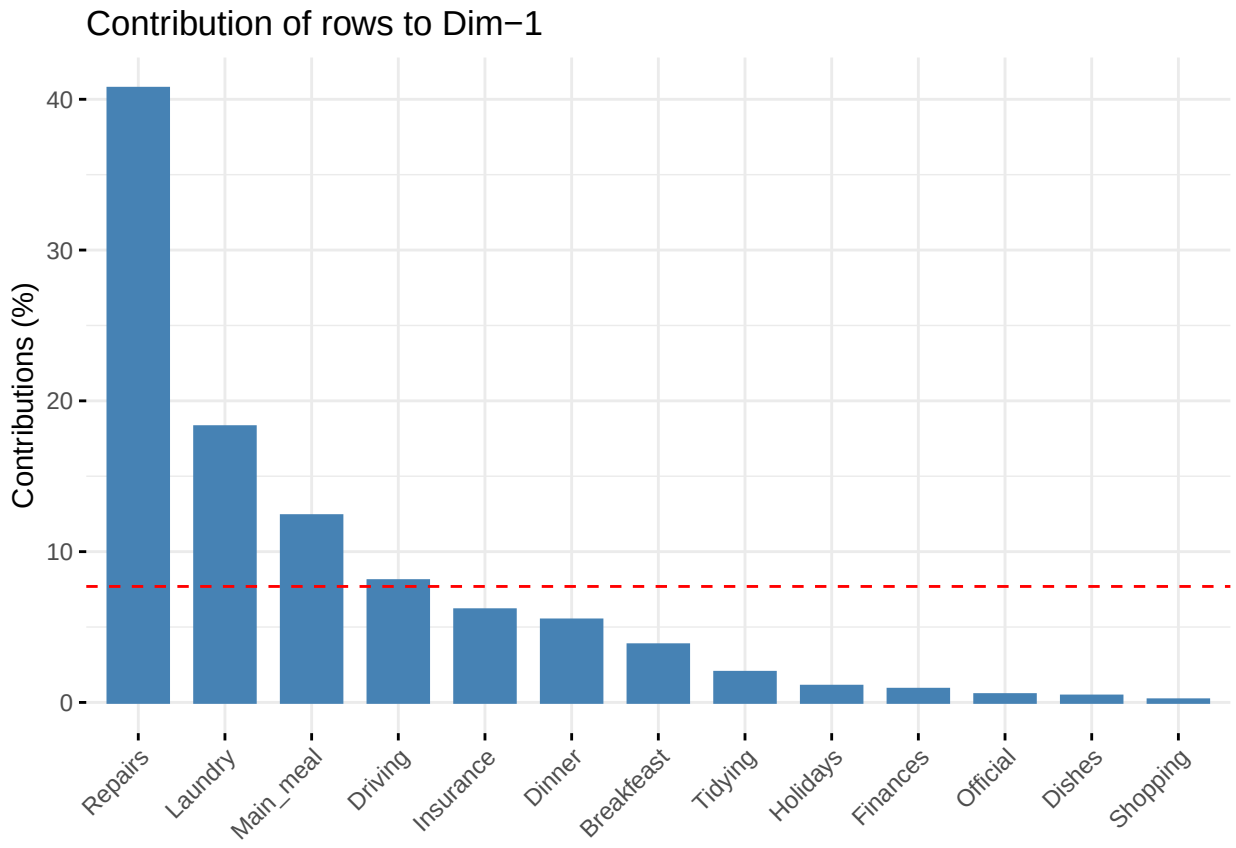
```
head(row$contrib)
```

```
##                Dim 1    Dim 2    Dim 3
## Laundry    18.2867003 5.563891 7.968424
```

```
## Main_meal  12.3888433 4.735523 1.858689
## Dinner      5.4713982 1.321022 2.096926
## Breakfeast  3.8249284 3.698613 3.069399
## Tidying     1.9983518 2.965644 0.488734
## Dishes      0.4261663 2.844117 3.634294
```

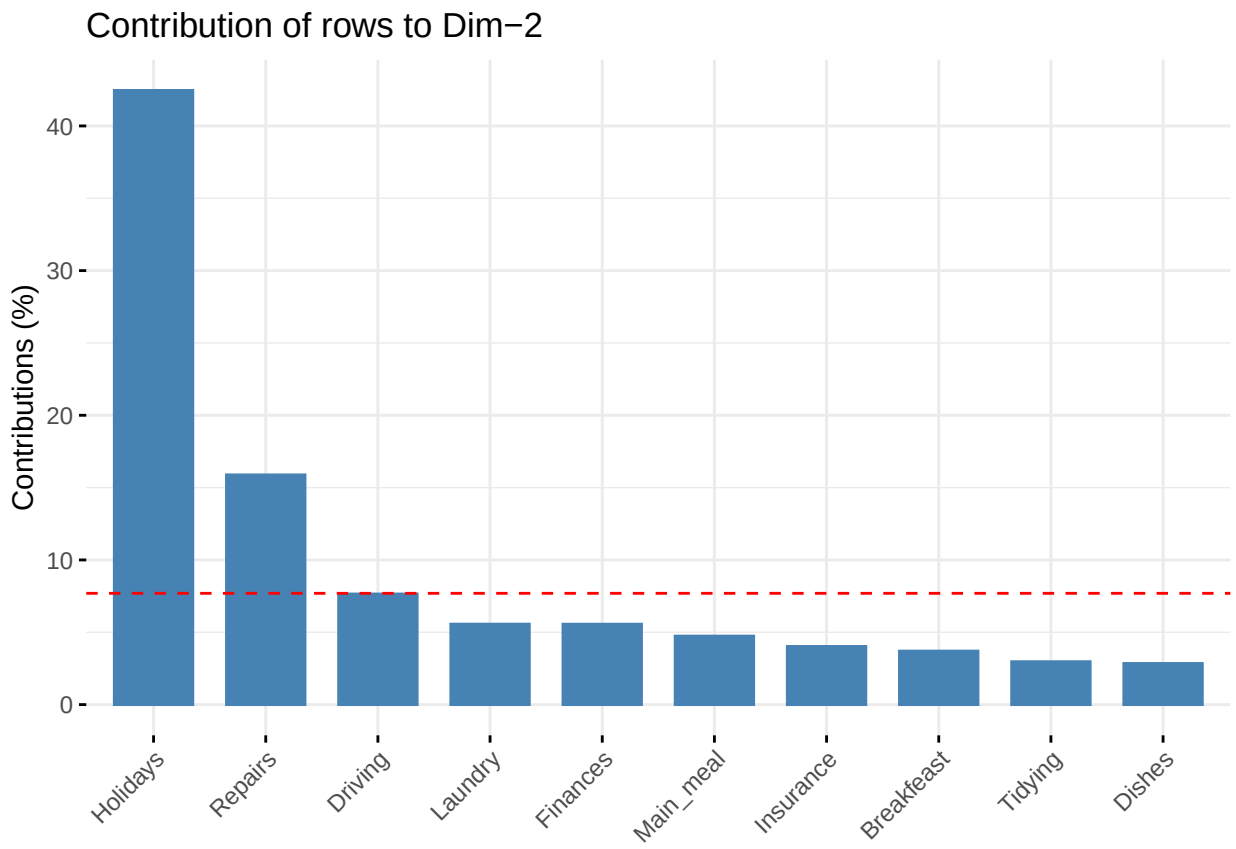
Contribuições de linhas para a dimensão 1:

```
fviz_contrib (res.ca, escolha = "linha", eixos = 1, topo = 10)
```



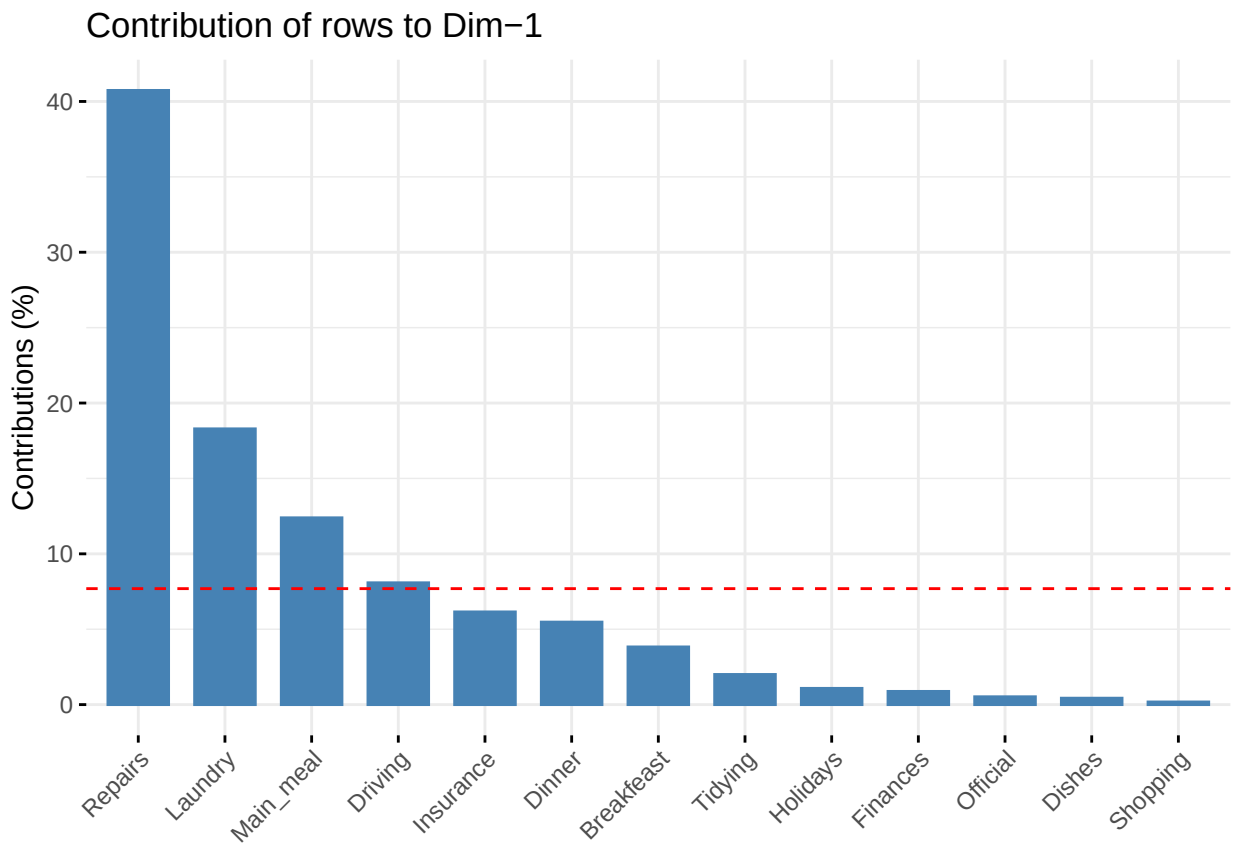
Contribuições de linhas para a dimensão 2:

```
fviz_contrib (res.ca, choice = "row", axes = 2, top = 10)
```



Contribuição total para a dimensão 1 e 2:

```
fviz_contrib (res.ca, escolha = "linha", eixos = 1: 2, topo = 10)
```



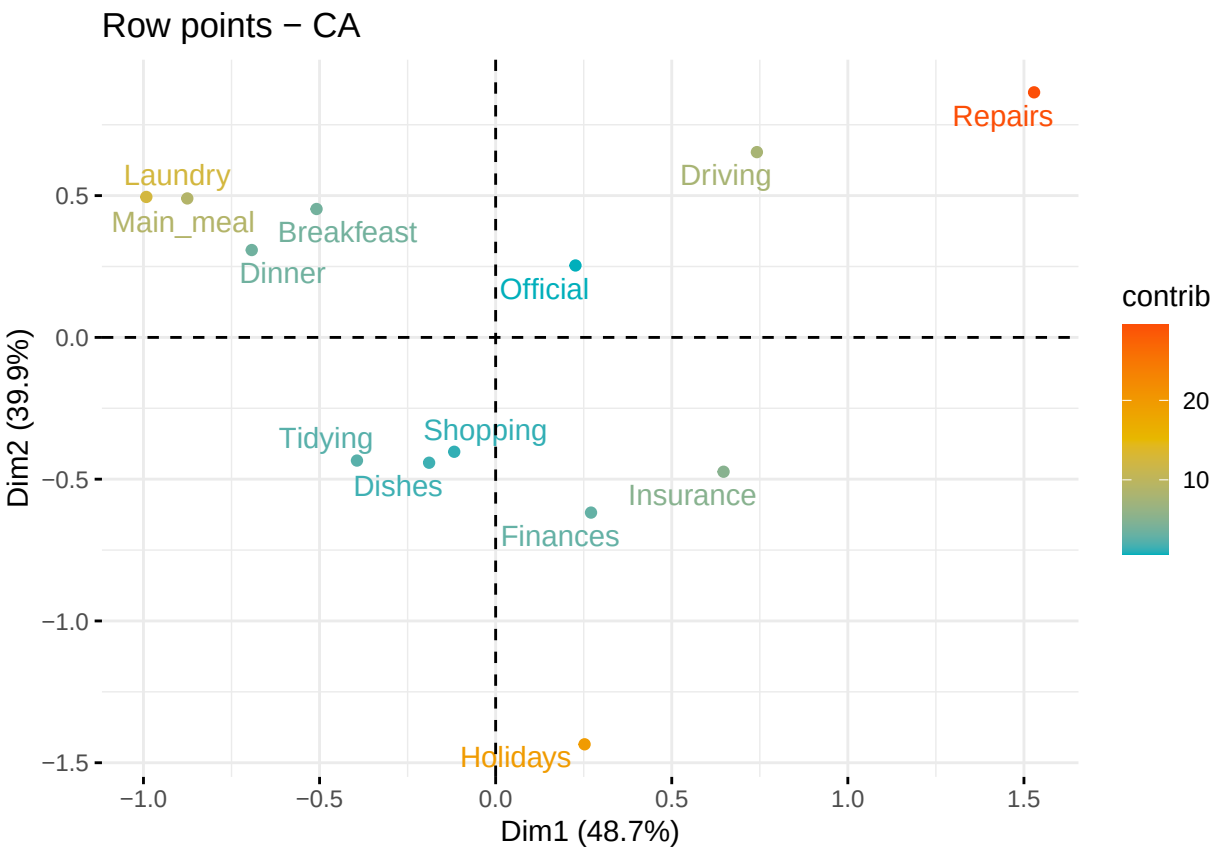
Obs: A linha vermelha tracejada no gráfico acima indica o valor médio esperado, se as contribuições forem uniformes.

Interpretação:

Os itens da linha Reparos, Lavanderia, Refeição Principal e Condução são os mais importantes na definição da primeira dimensão. Os itens da linha Feriados e reparos são os que mais contribuem para a dimensão 2.

Os pontos de linha mais importantes (em relação à contribuições) podem ser destacados no gráfico de dispersão da seguinte forma:

```
fviz_ca_row(res.ca, col.row = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE)
```



Interpretação:

As categorias de linha Reparo e Condução têm uma contribuição importante para o polo positivo da primeira dimensão, enquanto as categorias Lavanderia e Refeição principal têm grande contribuição para o polo negativo da primeira dimensão.

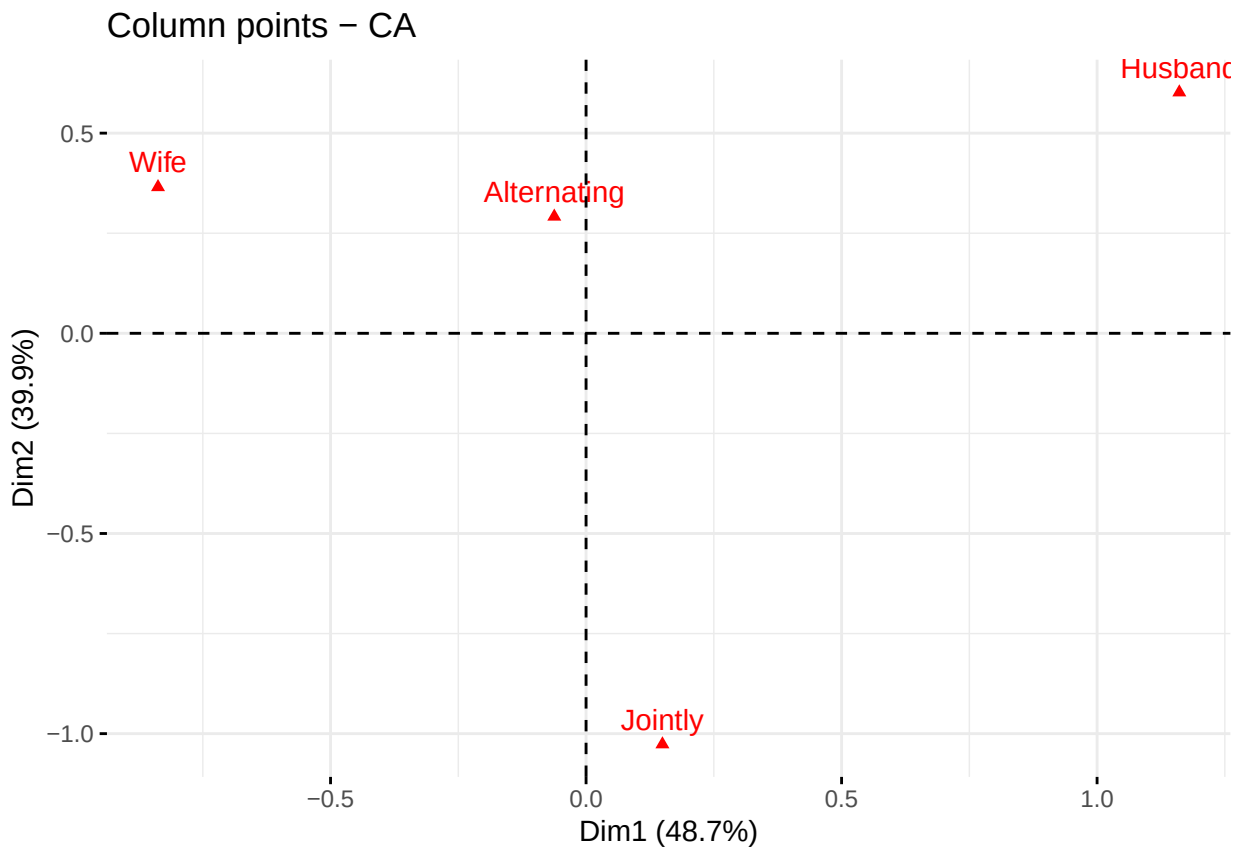
Isso significa que a dimensão 1 é definida principalmente pela oposição de Reparo e Condução (polo positivo), e Lavanderia e Refeição principal (polo negativo).

```
col <- get_ca_col(res.ca)
col
```

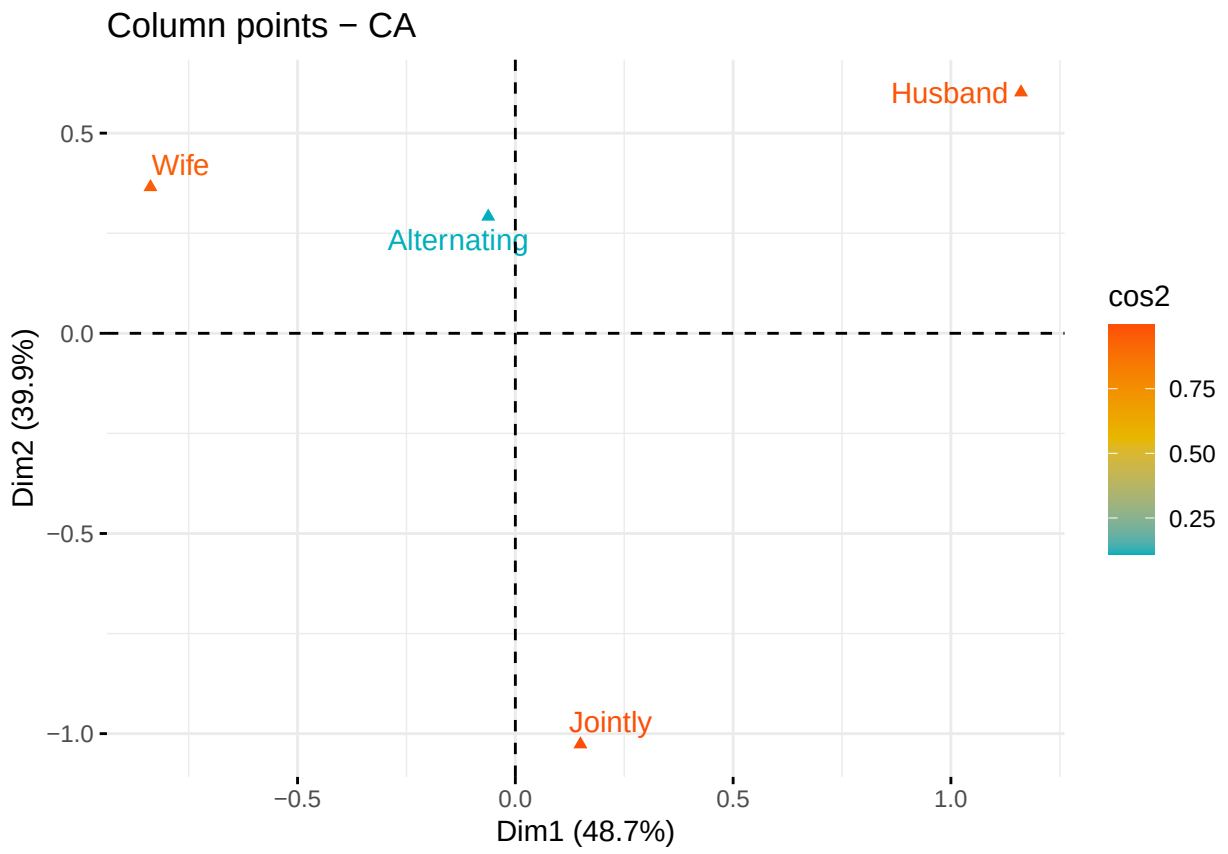
4.2 Colunas

```
## Correspondence Analysis - Results for columns
##  ==================================================
##   Name       Description
## 1 "$coord"   "Coordinates for the columns"
## 2 "$cos2"    "Cos2 for the columns"
## 3 "$contrib" "contributions of the columns"
## 4 "$inertia" "Inertia of the columns"
```

```
fviz_ca_col(res.ca)
```




```
fviz_ca_col(res.ca, col.col = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE)
```



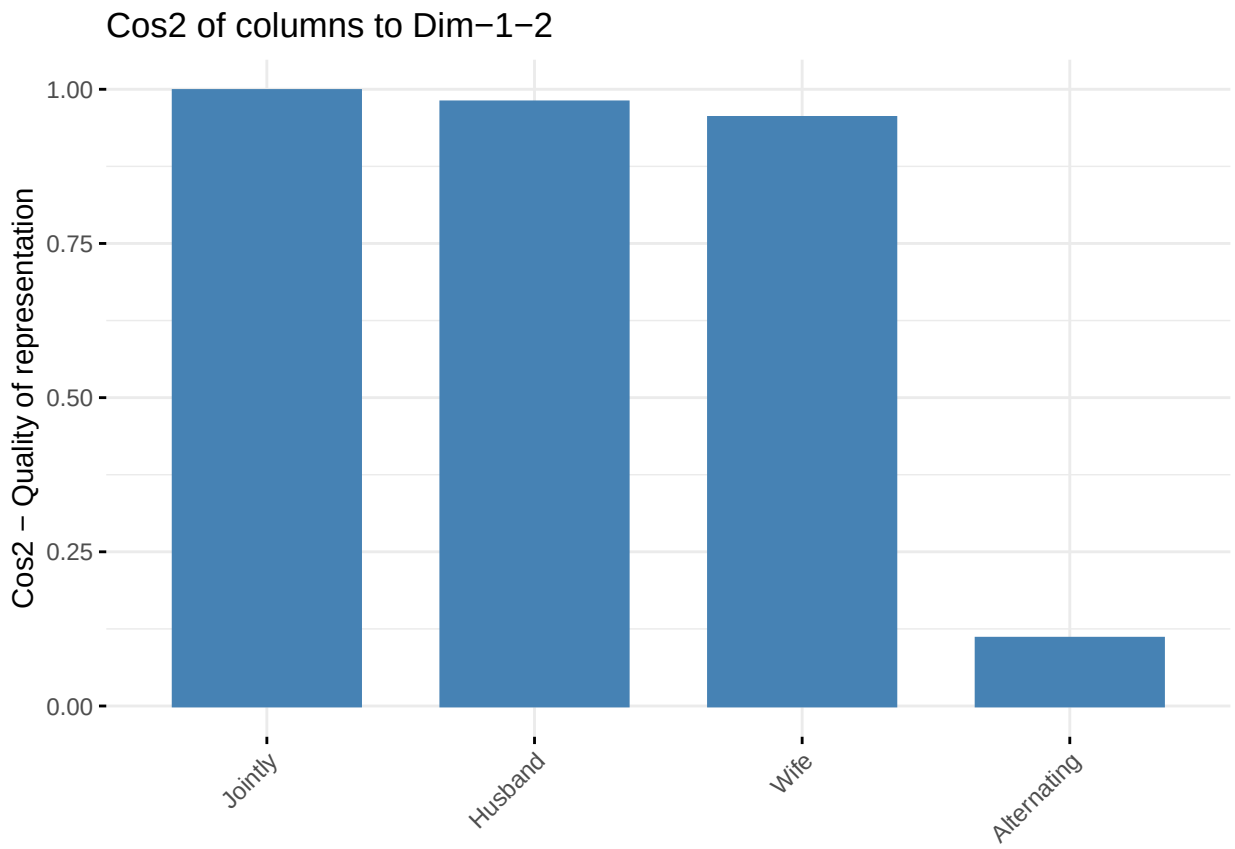
cos2 por coluna

O cos2 mede o grau de associação entre linhas/colunas e um eixo específico. Seu valor máximo é 1.

Interpretação

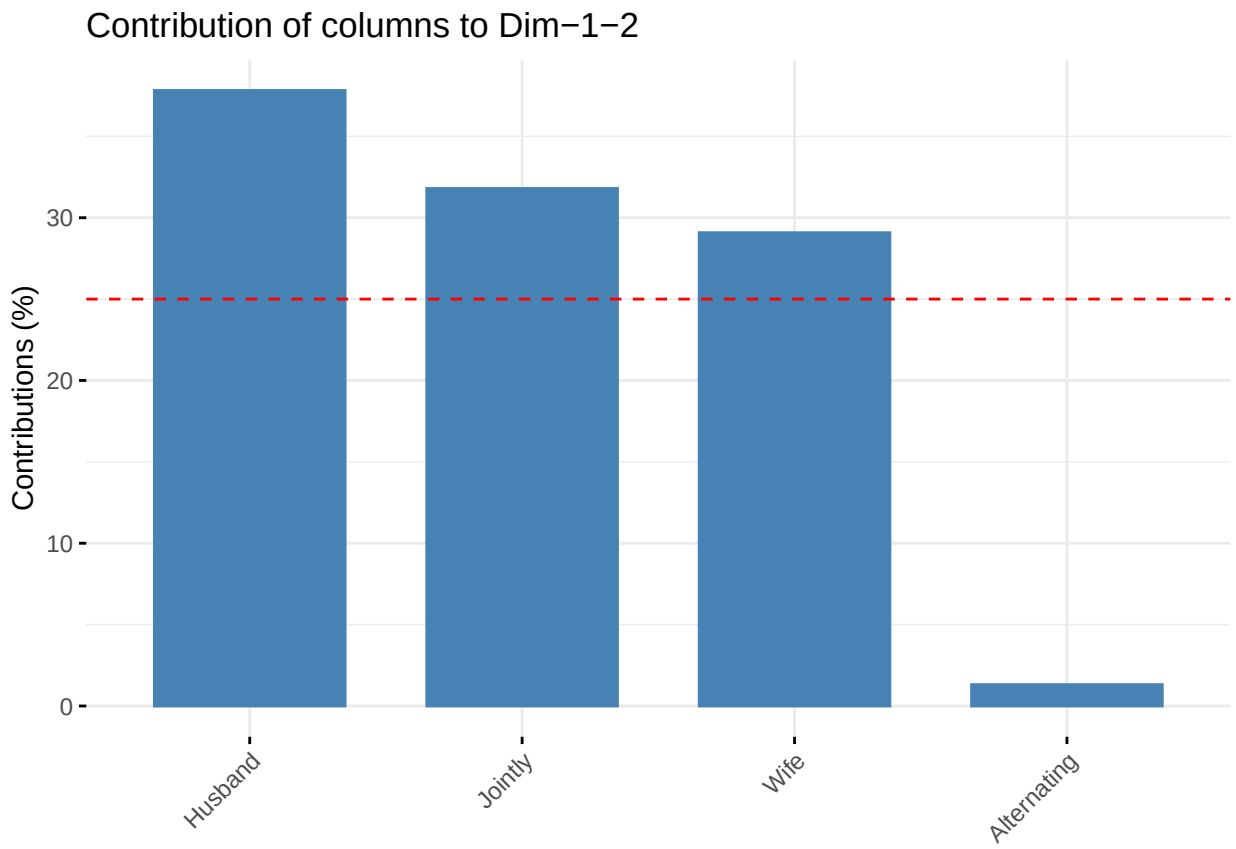
Apenas o item da coluna Alternando não é muito bem exibido nas duas primeiras dimensões. A posição deste item deve ser interpretada com cautela no espaço formado pelas dimensões 1 e 2.

```
fviz_cos2(res.ca, choice = "col", axes = 1:2)
```



Para visualizar a contribuição das linhas para as duas primeiras dimensões:

```
fviz_contrib(res.ca, choice = "col", axes = 1:2)
```

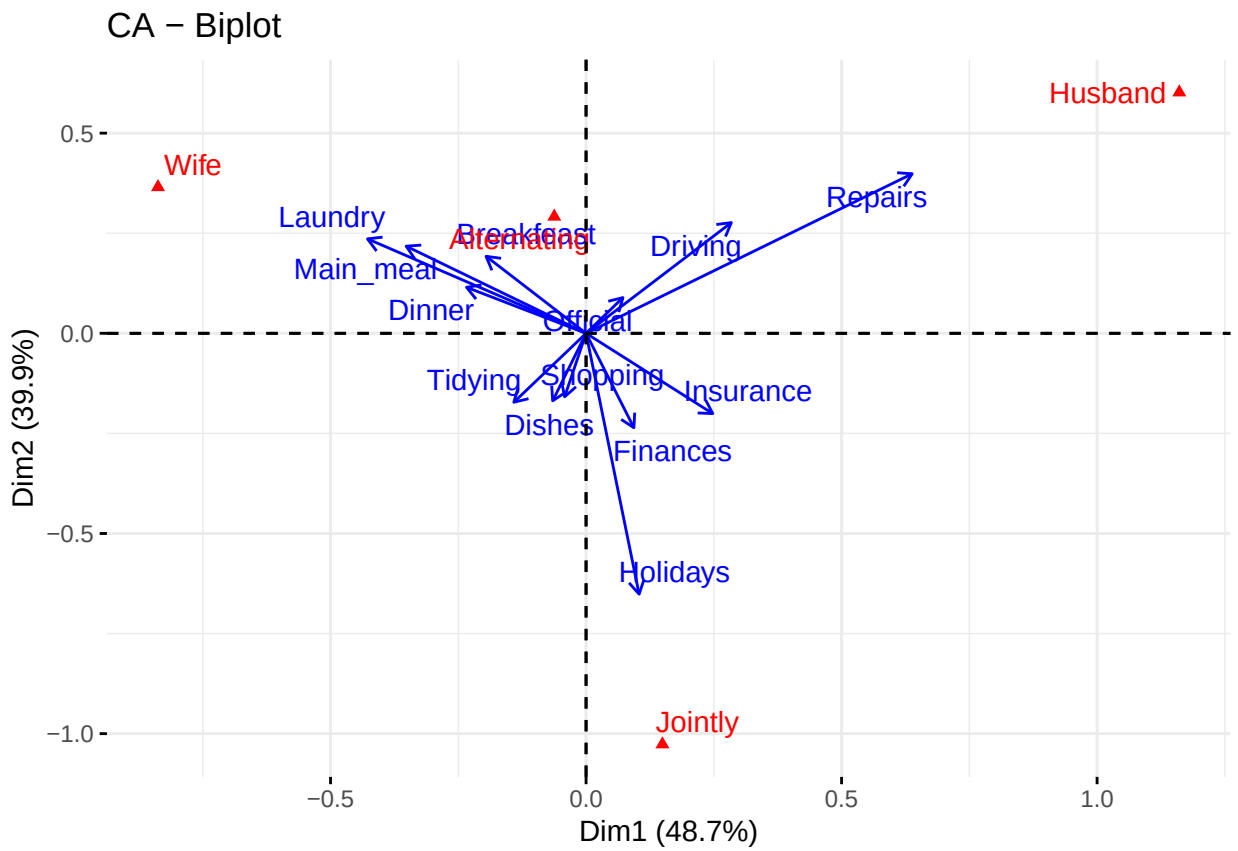


Biplot

Biplot contribuição

- Proposto por Michael Greenacre em 2013.
- Neste tipo de Biplot, os pontos que contribuem muito pouco para a solução, estão próximos ao centro do biplot e são relativamente sem importância para a interpretação.
- Quanto mais perto uma seta está (em termos de distância angular) para um eixo maior é a contribuição da categoria de linha nesse eixo em relação ao outro eixo. Se a seta estiver no meio do caminho entre os dois, sua categoria de linha contribui para os dois eixos na mesma medida.
- A posição dos pontos do perfil da coluna é inalterada em relação à do biplot convencional. No entanto, as distâncias dos pontos de linha da origem da parcela estão relacionadas às suas contribuições para o mapa fator bidimensional.

```
fviz_ca_biplot(res.ca, map ="colgreen", arrow = c(TRUE, FALSE),
               repel = TRUE)
```



Interpretação:

Interpretando a contribuição das linhas para os eixos.

Os reparos da categoria linha têm uma contribuição importante para o polo positivo da primeira dimensão, enquanto as categorias Lavanderia e Refeição principal têm grande contribuição para o polo negativo da primeira dimensão.

A dimensão 2 é definida principalmente pela categoria de linha Feriados.

A categoria de linha Driving contribui para os dois eixos na mesma medida.

Descrição da dimensão

```
res.desc <- dimdesc(res.ca, axes = c(1,2))
```

Descrição da dimensão 1 por pontos da linha:

```
head(res.desc[[1]]$row, 4)
```

```
##                coord
## Laundry    -0.9918368
## Main_meal  -0.8755855
## Dinner     -0.6925740
## Breakfeast -0.5086002
```

Descrição da dimensão 1 por pontos de coluna:

```
head(res.desc[[1]]$col, 4)
```

```
##                   coord
## Wife        -0.83762154
## Alternating -0.06218462
## Jointly      0.14942609
## Husband      1.16091847
```

Descrição da dimensão 2 por pontos de linha:

```
res.desc[[2]]$row
```

```
##                 coord
## Holidays   -1.4350066
## Finances   -0.6178684
## Insurance  -0.4737832
## Dishes     -0.4419662
## Tidying    -0.4343444
## Shopping   -0.4033171
## Official    0.2536132
## Dinner      0.3081043
## Breakfeast  0.4528038
## Main_meal   0.4901092
## Laundry     0.4953220
## Driving     0.6534143
## Repairs     0.8642647
```

Descrição da dimensão 1 por pontos de coluna:

```
res.desc[[2]]$col
```

```
##                  coord
## Jointly     -1.0265791
## Alternating  0.2915938
## Wife         0.3652207
## Husband      0.6019199
```

Para identificar facilmente pontos de linha e coluna que são os mais associados às principais dimensões. As variáveis linha/coluna são classificadas por suas coordenadas na saída.

6. Outliers Não há aparentes outliers em nossos dados. Se houver outliers nos dados, eles devem ser suprimidos ou tratados como pontos complementares ao re-executar a análise de correspondência.

7. Referência HAIR, J. F, *et al*, Análise multivariada de dados – 6. ed. Porto Alegre : Bookman, 2009.

Articles - Principal Component Methods in R: Practical Guide