ANALISIS DE CORRESPONDENCIAS

Paquetes

Hay muchos paquetes que permiten realizar un Análisis de Correspondencias, por ejemplo:

- CA() [FactoMineR package]
- ca() [ca package],
- dudi.coa() [ade4 package],
- corresp() [MASS package],
- epCA() [ExPosition package]
- CA [MultBiploR]

Utilizaremos *FactoMiner* para el análisis y *factoextra* para la visualización mediante *ggplot2*. Asegúres se de que tiene instalados los tre paquetes antes de comenzar.

Los datos

Partimos de una tabla de freccuencias $\{\bf{F}} = ({f_{ij}})$ con I filas y J columnas correspondeintes alas categorías de dos variables nominales (u ordinales). BENZECRI (1973) propone el AFC como una técnica multivariante para representar las filas y las columnas de una tabla de contingencia como puntos de un espacio vectorial de baja dimensión (un plano, por ejemplo) en el cual se pueden interpretar las posiciones de los puntos. GIFI (1990) establece que el Análisis de Correspondencias es una técnica multivariante para estudiar relaciones de dependencia entre variables categóricas, a partir de una tabla de contingencia.

```
library("FactoMineR")
```

```
## Warning: package 'FactoMineR' was built under R version 3.5.2
```

```
library("factoextra")
```

```
## Loading required package: ggplot2
```

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://go
o.gl/13EFCZ
```

Los datos que utilizaremos son los siguientes

```
data(housetasks)
housetasks
```

```
##            Wife Alternating Husband Jointly
## Laundry     156           14       2       4
## Main_meal   124           20       5       4
## Dinner       77           11       7      13
## Breakfeast   82           36      15       7
## Tidying      53           11       1      57
## Dishes       32           24       4      53
## Shopping     33           23       9      55
## Official     12           46      23      15
## Driving      10           51      75       3
## Finances     13           13      21      66
## Insurance     8            1      53      77
## Repairs       0            3     160       2
## Holidays      0            1       6     153
```

Las 13 filas de la tabla son distintas tareas domésticas.

Las columnas especifican quién realiza las tareas: solo por la mujer, alternativamente, solo por el marido o conjuntamente.

Dibujo directo de la tabla y test chi-cuadrado

Como la tabla es pequeña, es fácil interpretar alguna las frecuencias directamente. Haremos una representación gráfica en la que las frecuencias están relacionadas con el tamaño de los puntos.

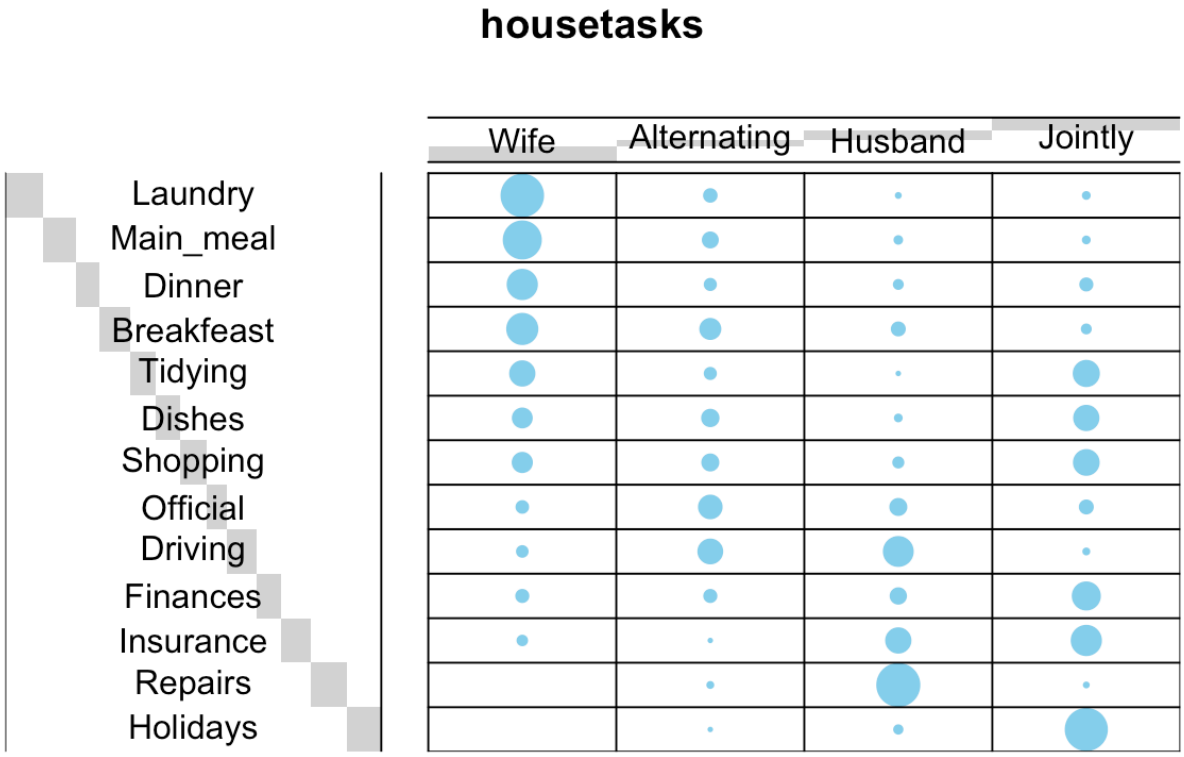
```
library("gplots")
```

```
## Warning: package 'gplots' was built under R version 3.5.2
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
dt <- as.table(as.matrix(housetasks))
balloonplot(t(dt), main ="housetasks", xlab ="", ylab="",
            label = FALSE, show.margins = FALSE)
```



Para contrastar la hipótesis de que no hay asociación entre las tareas y quién las lleva a cabo, utilizamos el test chi- cuagrado.

```
chisq <- chisq.test(housetasks)
chisq
```

```
##
##  Pearson's Chi-squared test
##
## data:  housetasks
## X-squared = 1944.5, df = 36, p-value < 2.2e-16
```

Encontramos una asociación significativa entre ambas variables. Podemos entender el Análisis de Correspondencias como la búsqueda de las causas de la significación. Si no tenemos una relación significativa, las diferencias que encontremos pueden deberse al azar.

Código para calcular en Análisis de Correspondencias

El comando del paquete *FactoMiner* para llevar a cabo el Análisis de Correspondencias es *CA* que tiene la siguiente sintaxis simplificada

```
CA(X, ncp = 5, graph = TRUE)
```

donde

- X : un marco de datos conteniendo la tabla de contingencia
- ncp : número de dimensiones a retener en los resultados.
- graph : un valor lógico. Si es TRUE a se muestra el gráfico.

Para nuestra tabla de datos

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

El objeto con los resultados contiene múltiples listas y matrices como las que hemos visto en teoría

```
print(res.ca)
```

```
## **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"
```

Visualización e Interpretación de los resultados.

Usaremos las siguientes funciones [en *factoextra*] para yudar a la visualización e interpretación del Análisis de Correspondencias:

- get_eigenvalue(res.ca)

  : Extrae los autovalores/varianzas de cada dimensión o eje

- fviz_eig(res.ca)

  : Visualiza los valores propios
- get_ca_row(res.ca), get_ca_col(res.ca)

  : Extrae los resultados de filas y columnas.
- fviz_ca_row(res.ca), fviz_ca_col(res.ca)(res.ca)

  : Visualiza los resultados de filas y columnas.
- fviz_ca_biplot(res.ca)

  : realiza la representación conjunta de las fias y las columnas fviz_ca_biplot(res.ca): Make a biplot of rows and columns. In the next sections, we'll illustrate each of these functions.

Valores propios/Varianzas

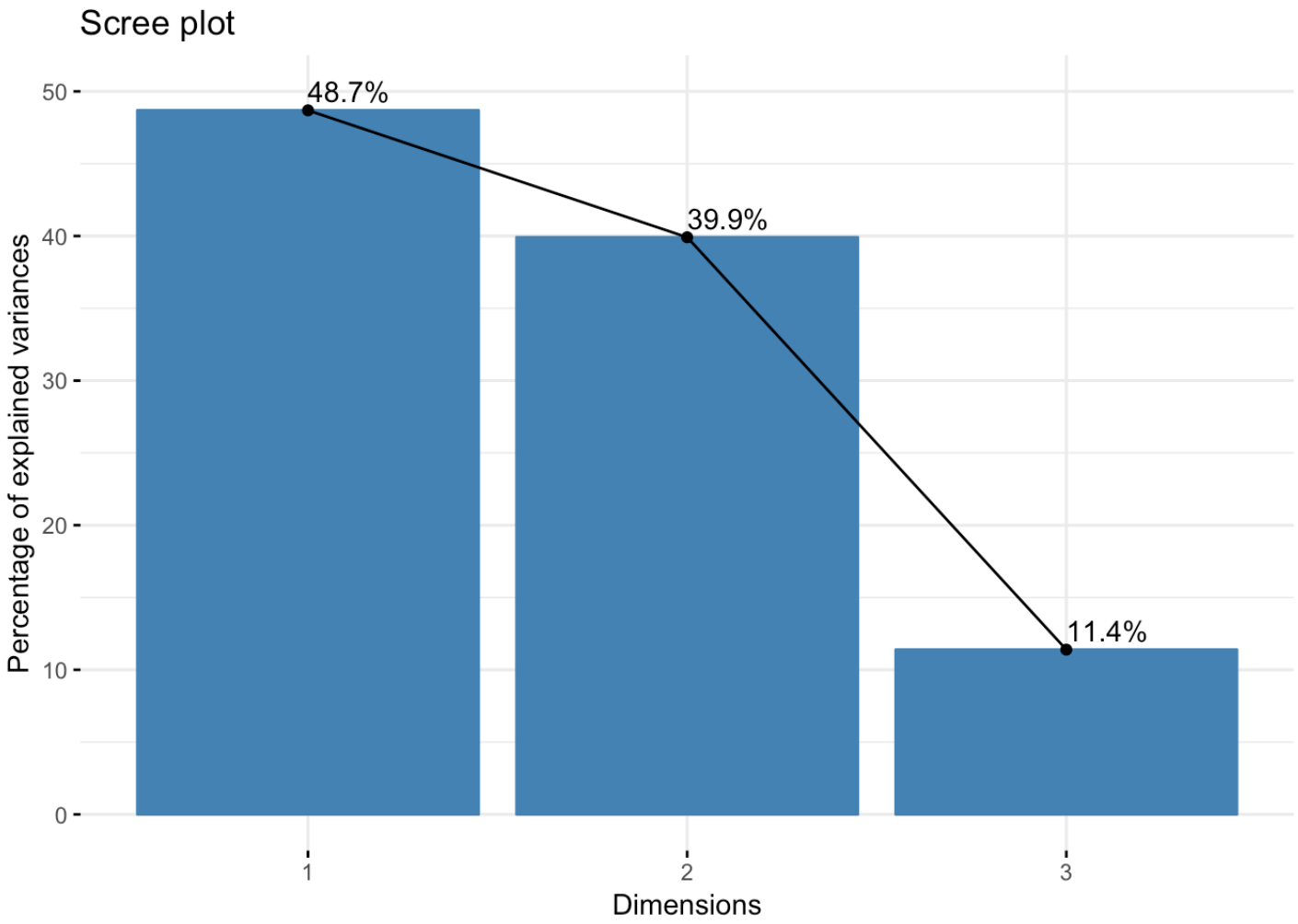
Los valores propios aparecen en la abla siguiente

```
library("factoextra")
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
```

Los dos primeros ejes recogen el 88.6% de la inercia, lo que puede considerarse suficiente para explicar la estructura de los datos. Podemos representarlos en un gráfico con la siguiente instrucción.

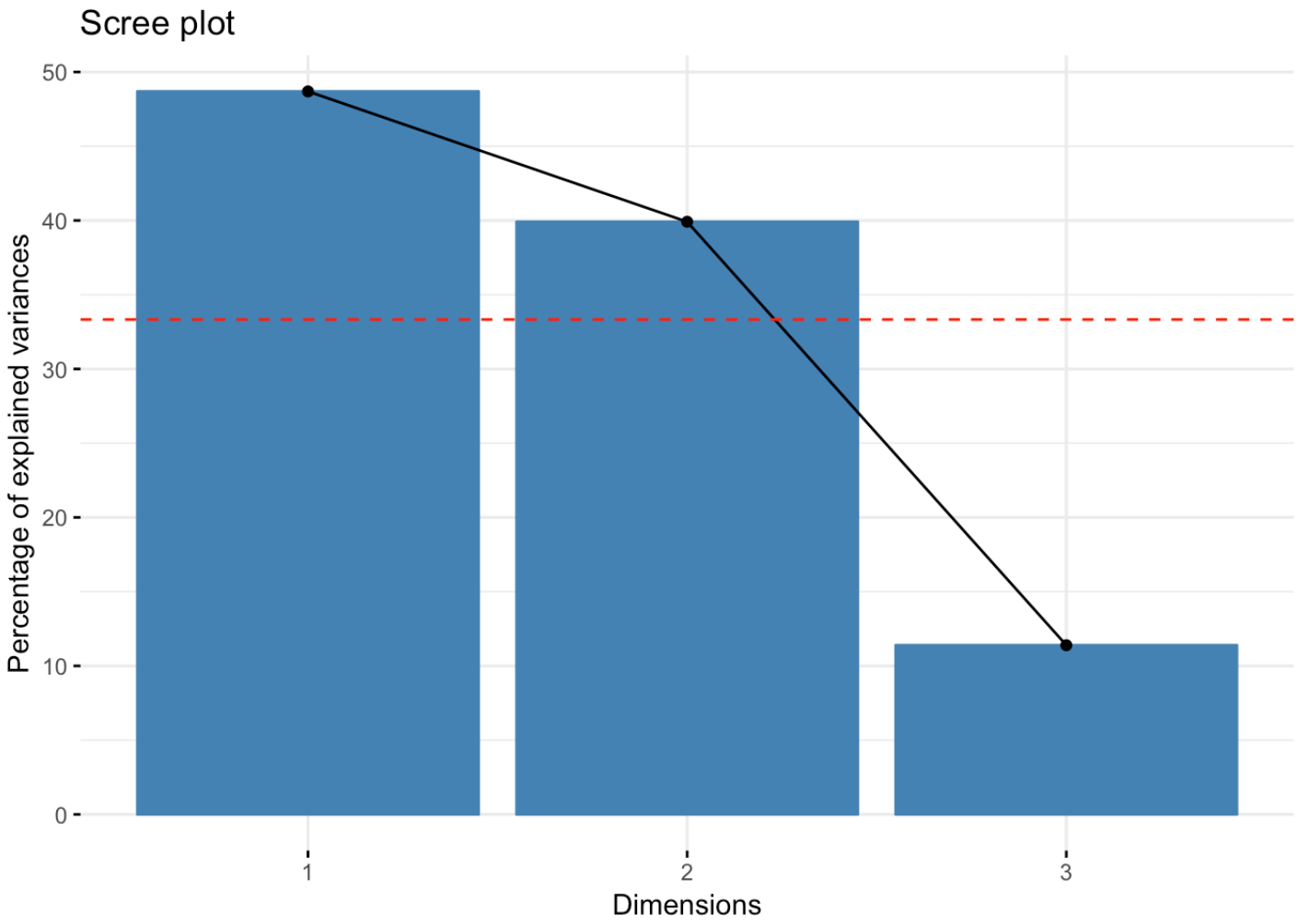
```
fviz_screeplot(res.ca, addlabels = TRUE, ylim = c(0, 50))
```



Una posible regla para saber cuantos ejes tomar es quedarse con aquellos que tengan valores por encima de la media. Nuestros datos contienen 13 filas y 4 columnas. Si los datos fueran aleatorios, el valor esperado del autovalor para cada eje sería 1/(nrow(housetasks)-1) = 1/12 = 8.33% en términos de las filas.

De la misma manera, la media debería recoger el 1/(ncol(housetasks)-1) = 1/3 = 33.33% en términos de las 4 columnas. Cualquier eje que contribuya más del máximo de ambas cantidades, debería ser considerado importante. Representamos el valor en la figura siguiente.

```
fviz_screeplot(res.ca) +
 geom_hline(yintercept=33.33, linetype=2, color="red")
```

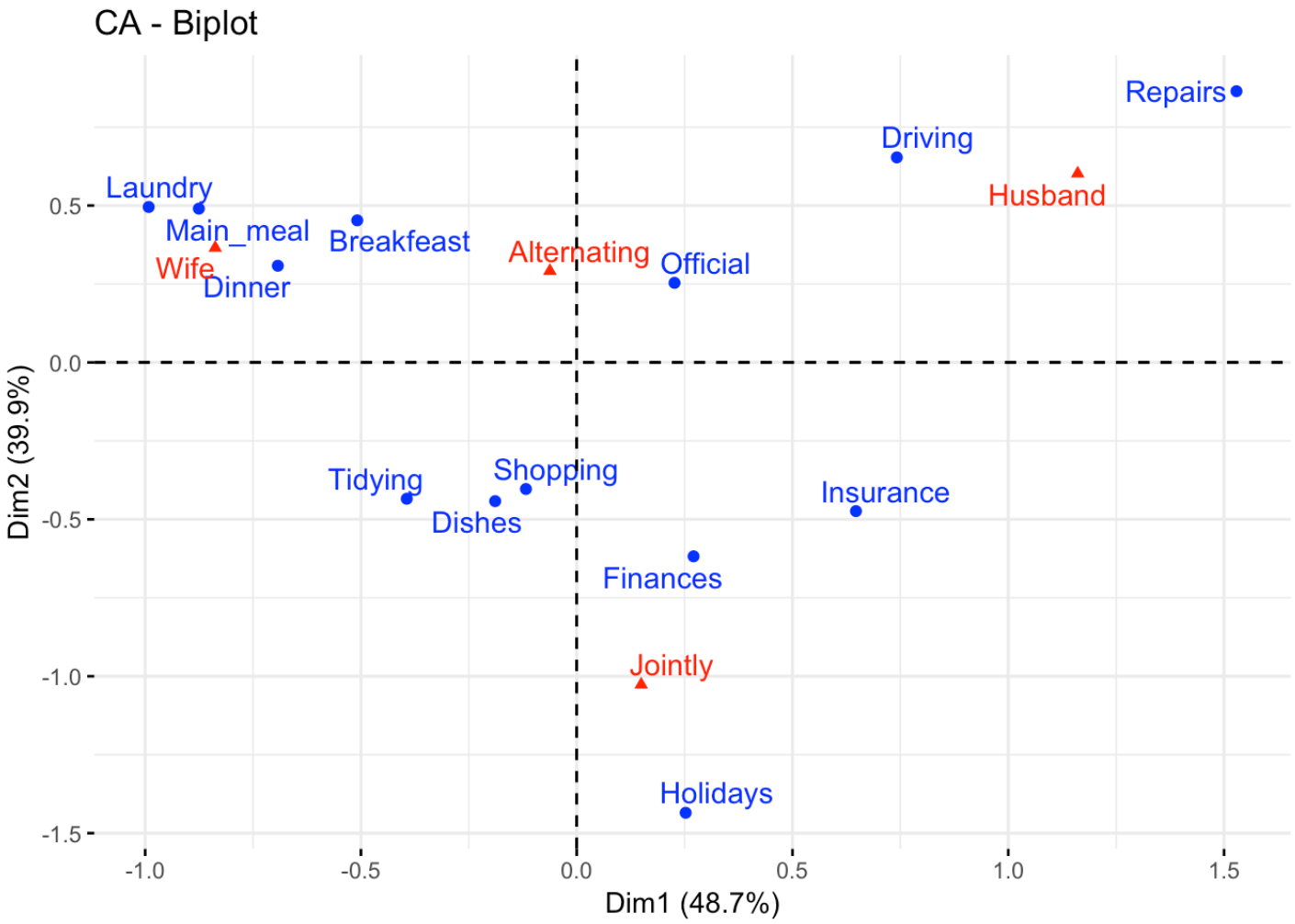


Los dos primeros complen el criterio establecido.

Biplot

La función *fviz_ca_biplot()* [factoextra package] puede usarse para la representación simultánea de las filas y las columnas de la tabla.

```
# repel= TRUE to avoid text overlapping (slow if many point)
fviz_ca_biplot(res.ca, repel = TRUE)
```



Se trata de una representación simétrica que muestra el patrón global de comportamiento de los datos.

La distancia entre dos puntos fila o dos puntos columna nos da una medida de su similitud (o disimilitud). Los puntos fila con un perfil similar aparecerán cercanos en el gráfico. Lo mismo ocurre con dos puntos columna.

En el gráfico podemos ver que:

- Las tareas como la cena, el desayuno, la lavandería son realizadas más a menudo por la mujer.
- Conducir o las tareas de reparación son realizadas con mayor frecuencia por el marido.
- El resto de las tareas, especialmente la planificación de las vacaciones, se realizan conjuntamente.
- No queda muy claro cuales son las tareas que se realizan alternativamente

Para la interpretación tenemos que tener en cuenta algunas consideraciones:

- El gráfico simétrico representa a las filas y a las columnas en un espacio común. En este caso, solamente la distancia entre filas o entre columnas puede ser realmente interpretada.
- La distancia entre una fila y una columna no es interpretable directamente. Solamente podemos realizar consideraciones generales. La interpretación es baricéntrica, que es un poco más complicada que las distancias.

El paso siguiente es determinar filas y que columnas contribuyen más a las dimensiones.

Coordenadas de los puntos fila

Mostramos las coordenadadas de las filas:

```
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"
```

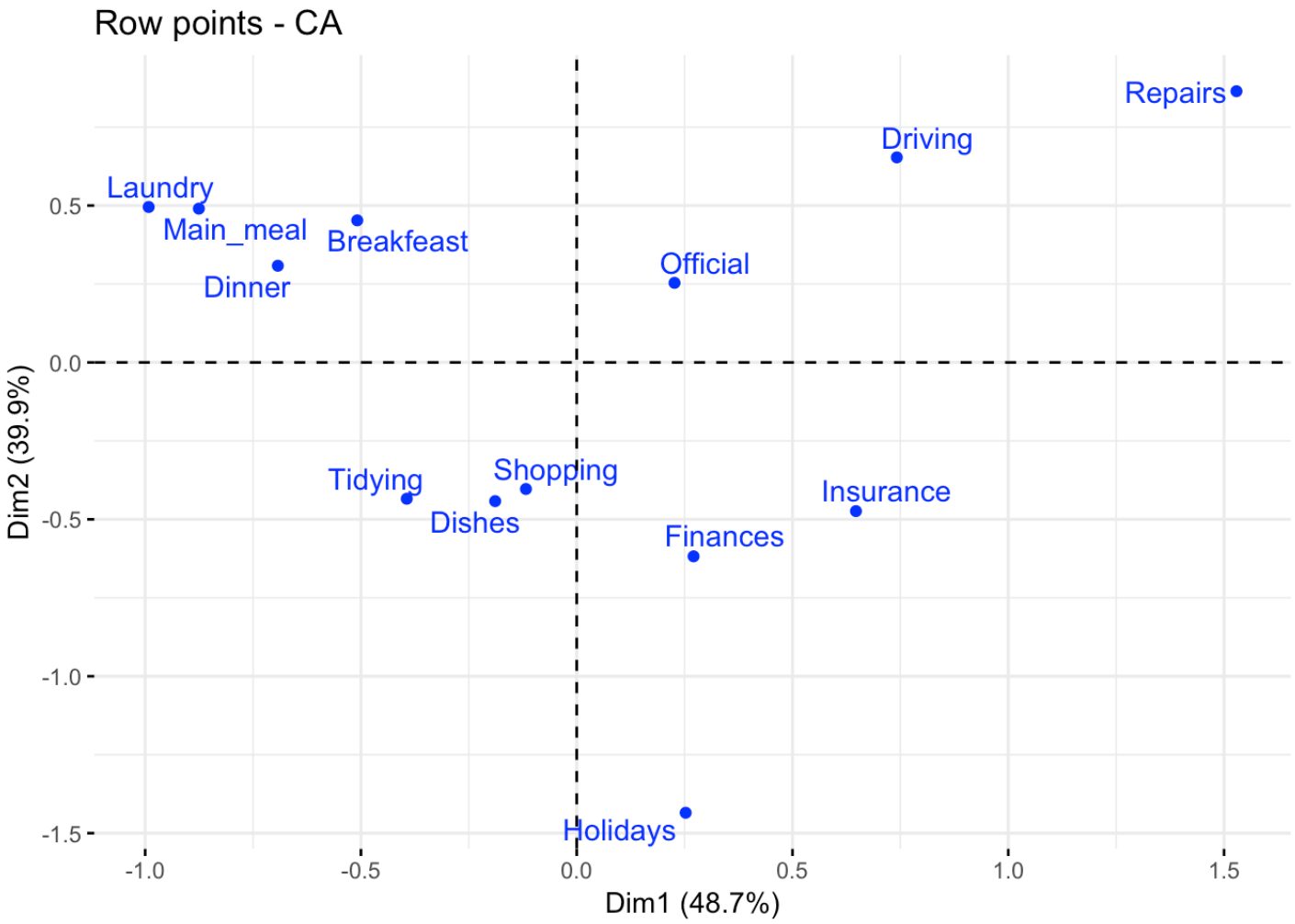
```
head(row$coord)
```

```
##                 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
```

De la misma manera podemos obtener las contribuciones (del elemento al factor), los cosenos al cuadrado (o contribuciones del factor al elemnto) o la inercia recogida por cada uno de los puntos

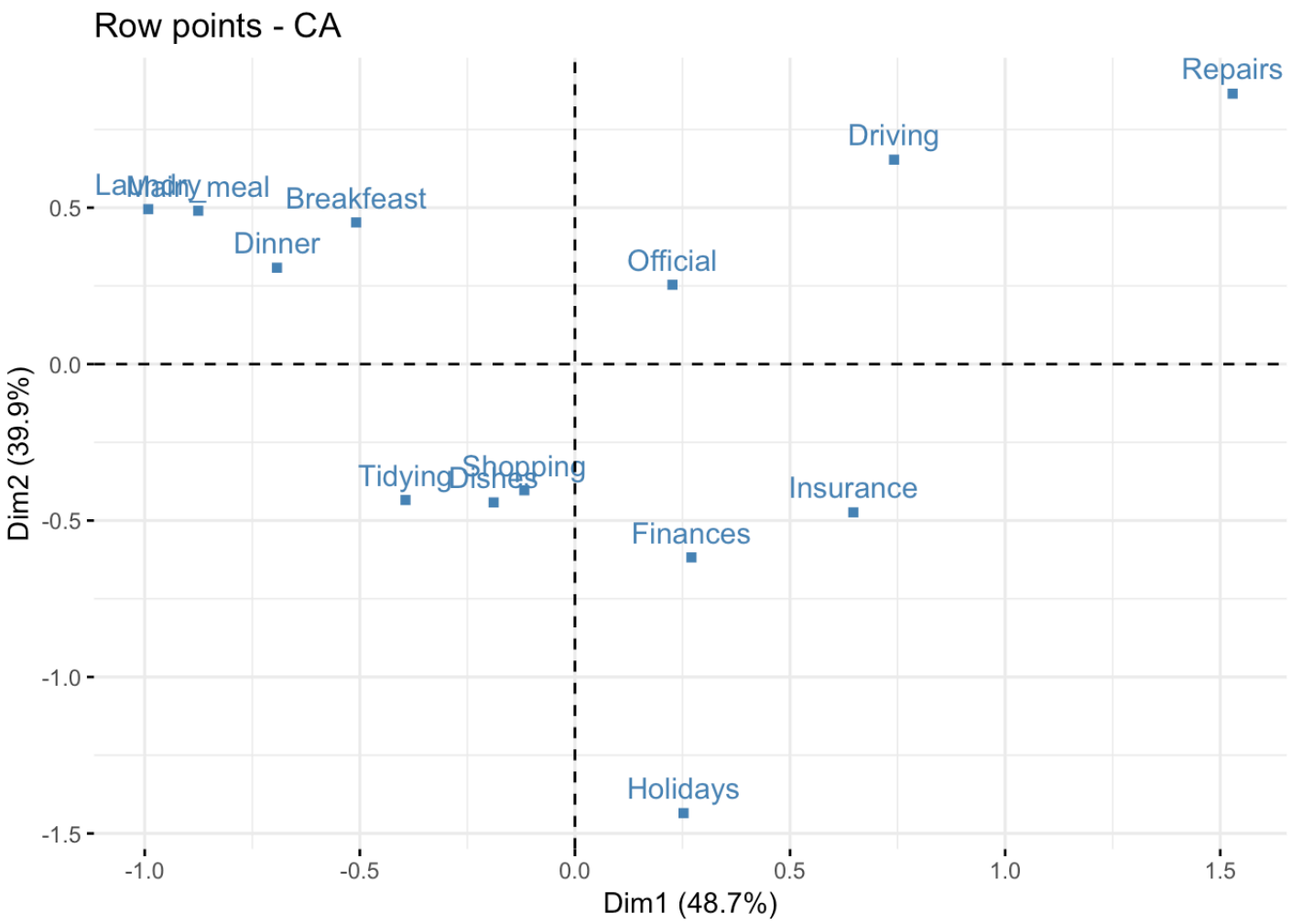
La función *fviz_ca_row()* [en *factoextra*] nos permite visualizar solamente las coordenadas de las filas:

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



Es posible cambiar el color y la forma de los puntos usando los argumentos *col.row* y *shape.row* como sigue:

```
fviz_ca_row(res.ca, col.row="steelblue", shape.row = 15)
```



Los gráficos anteriores muestran las relaciones entre los puntos fila:

- Filas con perfiles similares aparecen cercanas.
- Filas negativamente correlacionadas apuntarán en direcciones opuestas.
- La distancia al origen de los puntos fila está relacionada con su calidad de represntación en el mapa factorial, es decir, puntos cercanos al origen están mal representados mientras que puntos alejados están bien representados.

Calidad de representación de las filas

El resultado del análisis muestra que, la tabla de contingencia se ha representado satisfactoriamente en el espacio de baja dimensión mediante el uso del análisis de correspondencias. Las dos primeras dimensiones consiguen explicar el 88.6% de la inercia (o variación) total contenida en los datos.

Sin embargo, no todos los puntos están igualmente representados en las dos dimensiones.

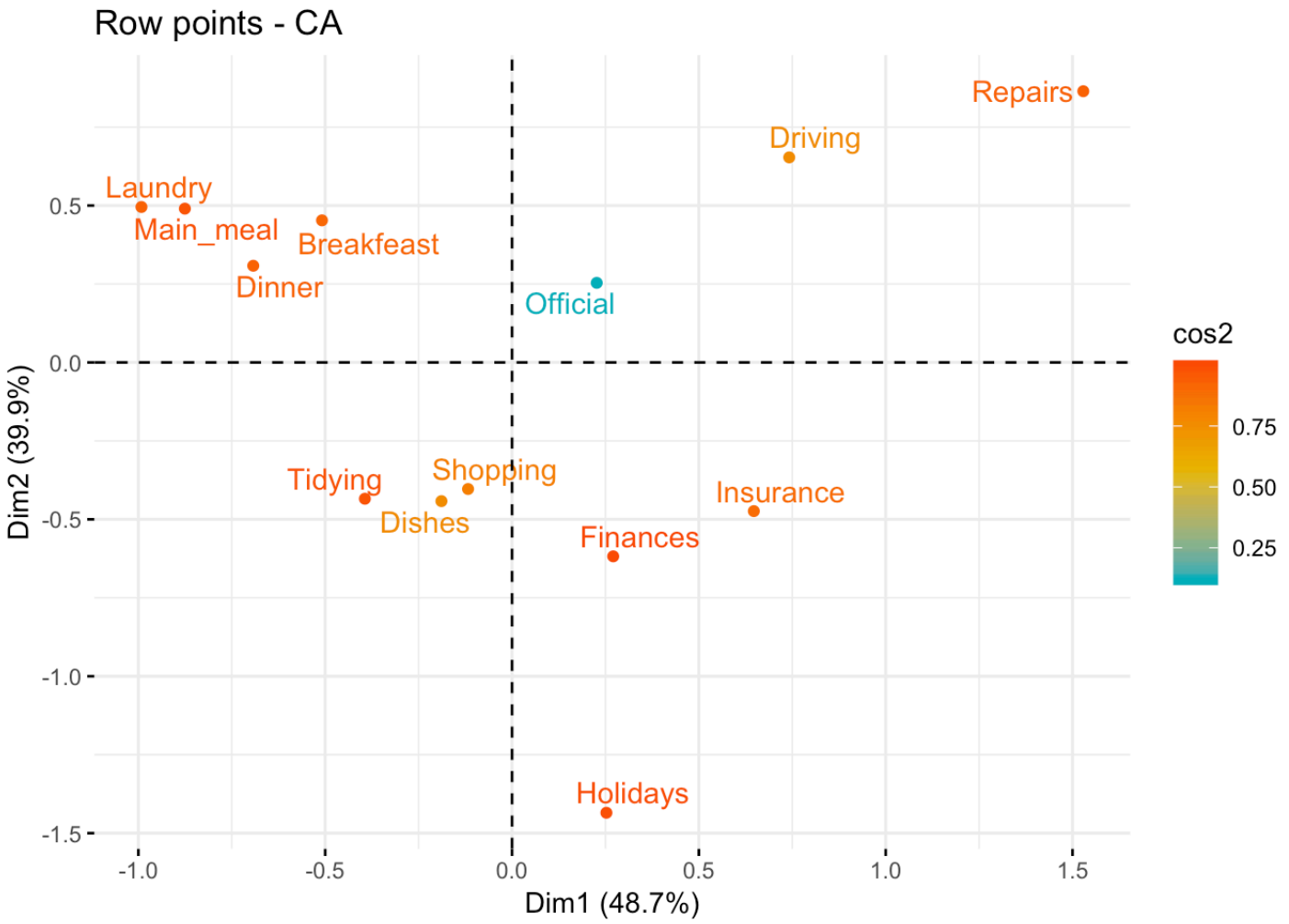
La calidad de representación se denomina también $\cos^2$, ya que se puede calcular como el coseno al cuadrado del ángulo que forman el vector que representa a la fila en el espacio en dimensión completa y el plano de la representación. También puede entenderse como el grado de asociación entre la fila y el plano.

```
row$cos2
```

```
##                 Dim 1      Dim 2       Dim 3
## Laundry    0.73998741 0.18455213 0.075460467
## Main_meal  0.74160285 0.23235928 0.026037873
## Dinner     0.77664011 0.15370323 0.069656660
## Breakfeast 0.50494329 0.40023001 0.094826699
## Tidying    0.43981243 0.53501508 0.025172490
## Dishes     0.11811778 0.64615253 0.235729693
## Shopping   0.06365362 0.74765514 0.188691242
## Official   0.05304464 0.06642648 0.880528877
## Driving    0.43201860 0.33522911 0.232752289
## Finances   0.16067678 0.83666958 0.002653634
## Insurance  0.57601197 0.30880208 0.115185951
## Repairs    0.70673575 0.22587147 0.067392778
## Holidays   0.02979239 0.96235977 0.007847841
```

Se pueden interpretar como correlaciones al cuadrado, es decir, cuanto más próximas estén a 1 major es la calidad de representación. En el gráfico siguiente vamos acolocar una escala de colores que depende de la calidad de representación.

```
# Color by cos2 values: quality on the factor map
fviz_ca_row(res.ca, col.row = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE)
```

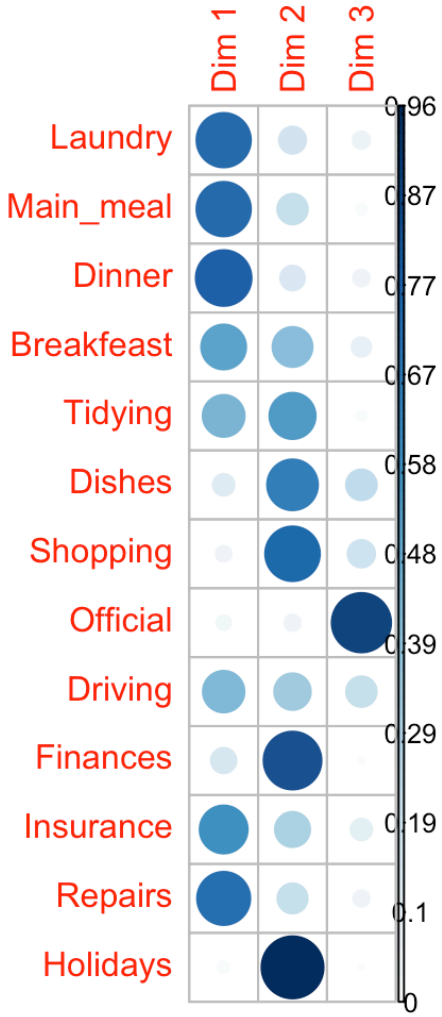


Podemos visualizar los cosenos al cuadrado de los tres ejes

```
library("corrplot")
```

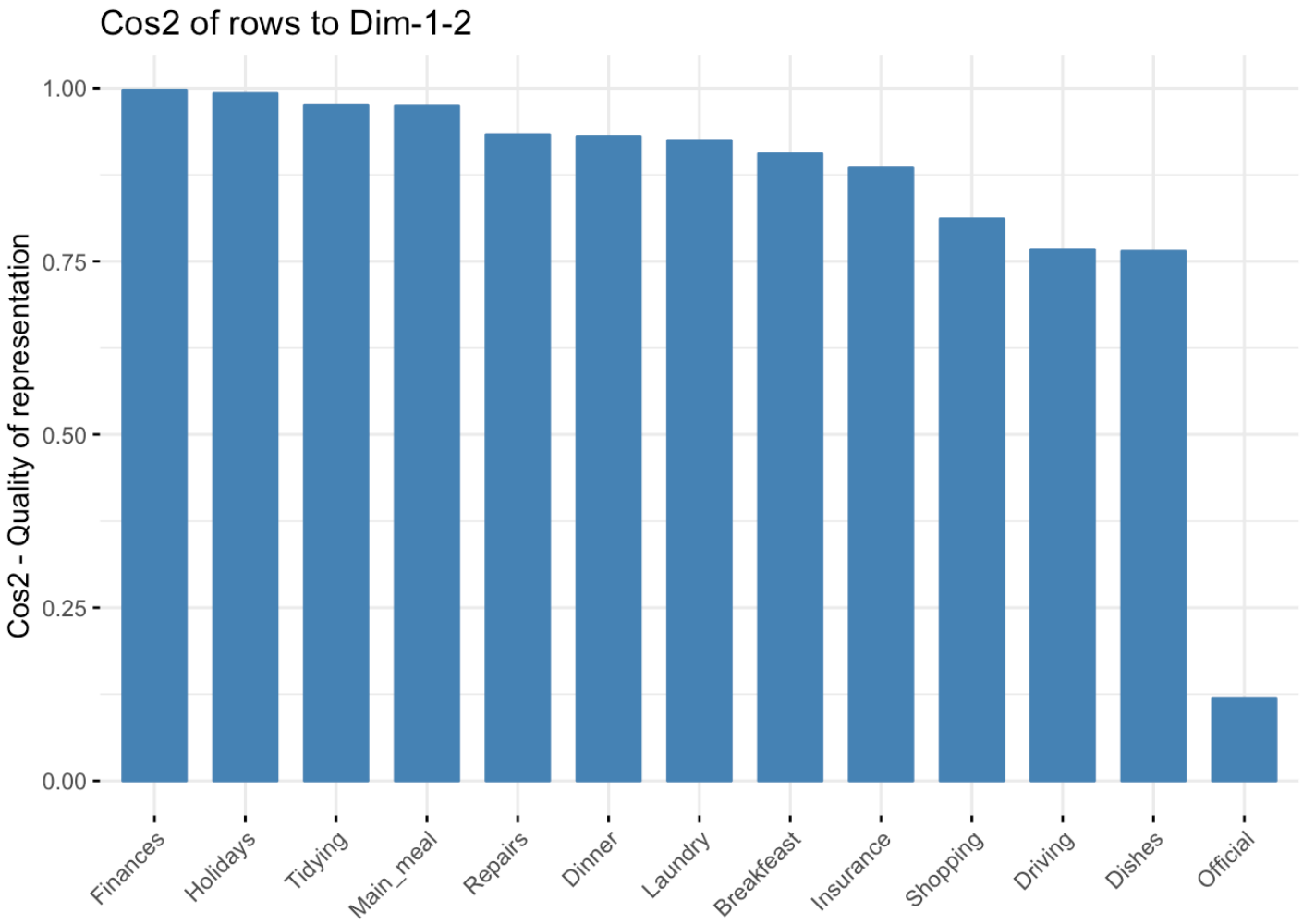
```
## corrplot 0.84 loaded
```

```
corrplot(row$cos2, is.corr=FALSE)
```



También los podemos visualizar con la siguiente instrucción.

```
# Cos2 of rows on Dim.1 and Dim.2
fviz_cos2(res.ca, choice = "row", axes = 1:2)
```



Contribución de las filas

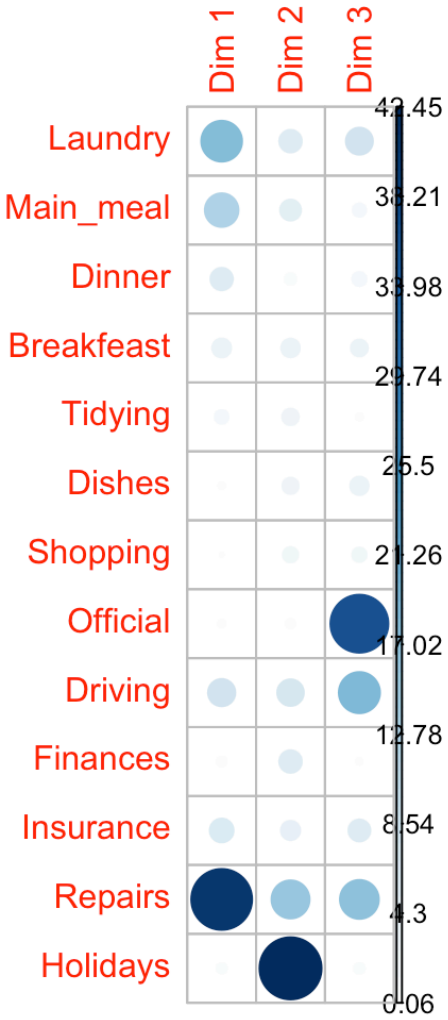
La contribución de las filas (en %) a la definición de las diemensiones puede

```
row$contrib
```

```
##                Dim 1      Dim 2       Dim 3
## Laundry    18.2867003  5.5638913  7.96842443
## Main_meal  12.3888433  4.7355230  1.85868941
## Dinner      5.4713982  1.3210221  2.09692603
## Breakfeast  3.8249284  3.6986131  3.06939857
## Tidying     1.9983518  2.9656441  0.48873403
## Dishes      0.4261663  2.8441170  3.63429434
## Shopping    0.1755248  2.5151584  2.22335679
## Official    0.5207837  0.7956201 36.94038942
## Driving     8.0778371  7.6468564 18.59638635
## Finances    0.8750075  5.5585460  0.06175066
## Insurance   6.1470616  4.0203590  5.25263863
## Repairs    40.7300940 15.8806509 16.59639139
## Holidays    1.0773030 42.4539986  1.21261994
```

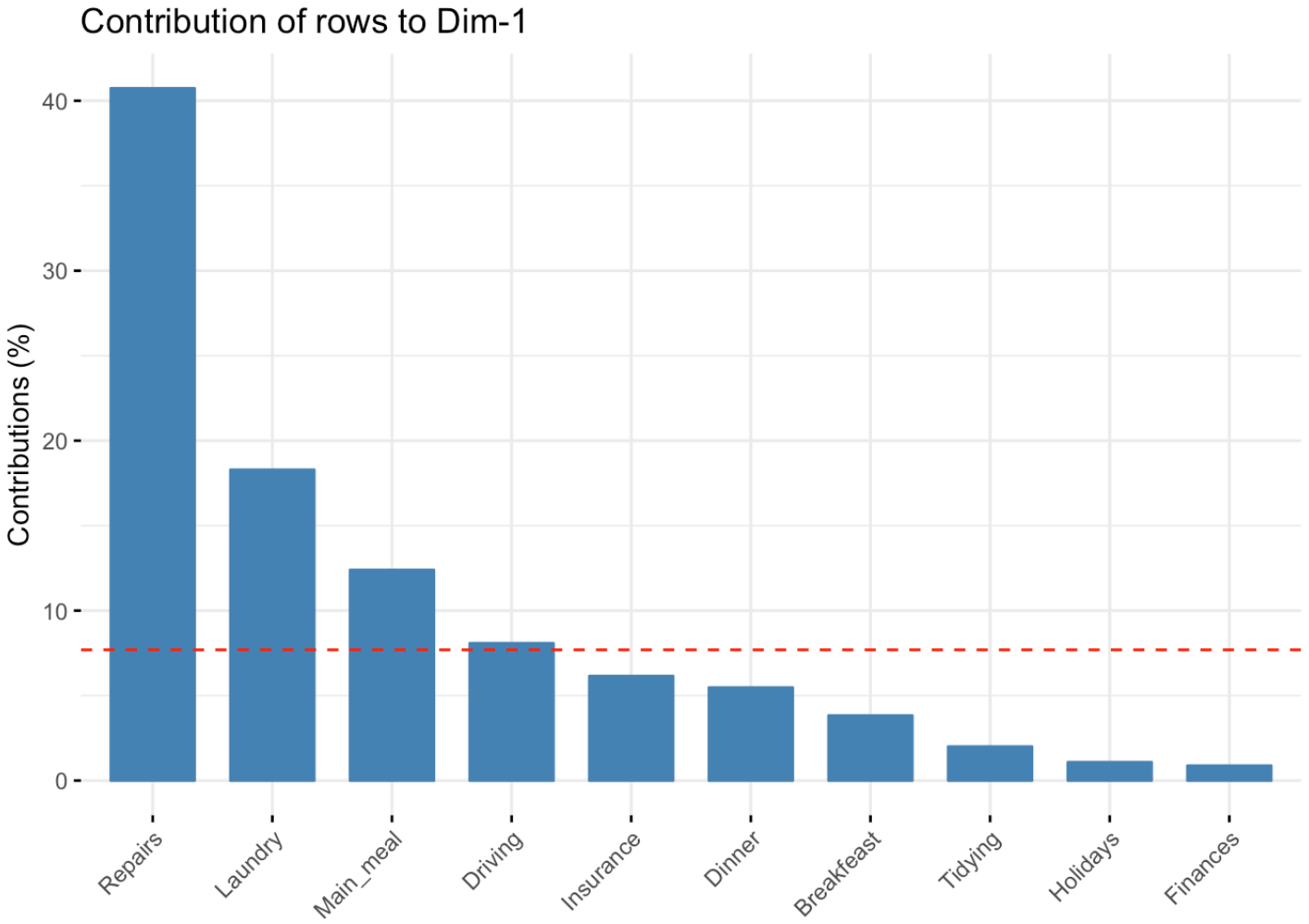
Las filas con los valores más altos son las que más contribuyen a la definición de las dimensiones. Podemos visualizarlas

```
library("corrplot")
corrplot(row$contrib, is.corr=FALSE)
```

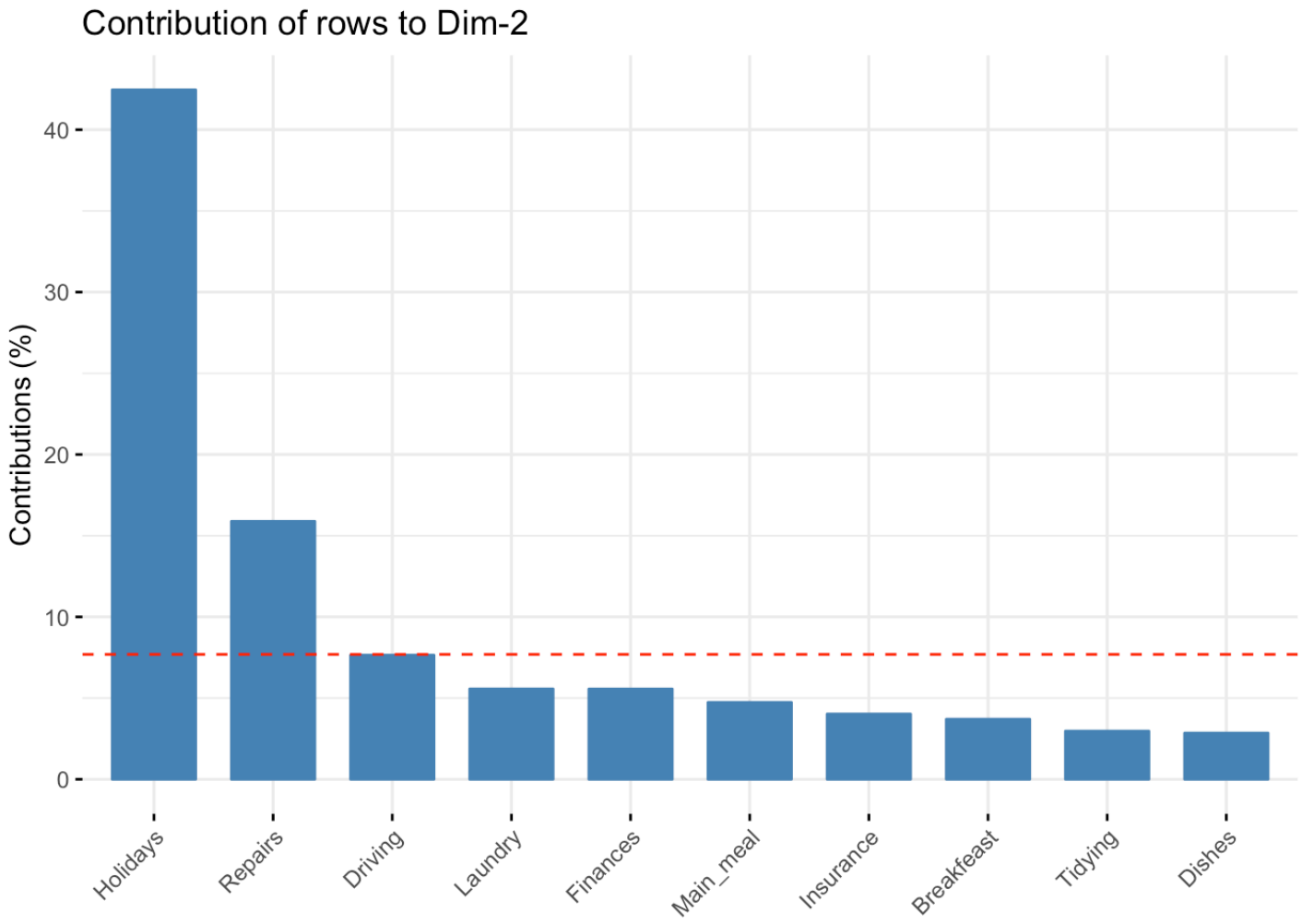


O también como

```
# Contributions of rows to dimension 1
fviz_contrib(res.ca, choice = "row", axes = 1, top = 10)
```




```
# Contributions of rows to dimension 2
fviz_contrib(res.ca, choice = "row", axes = 2, top = 10)
```

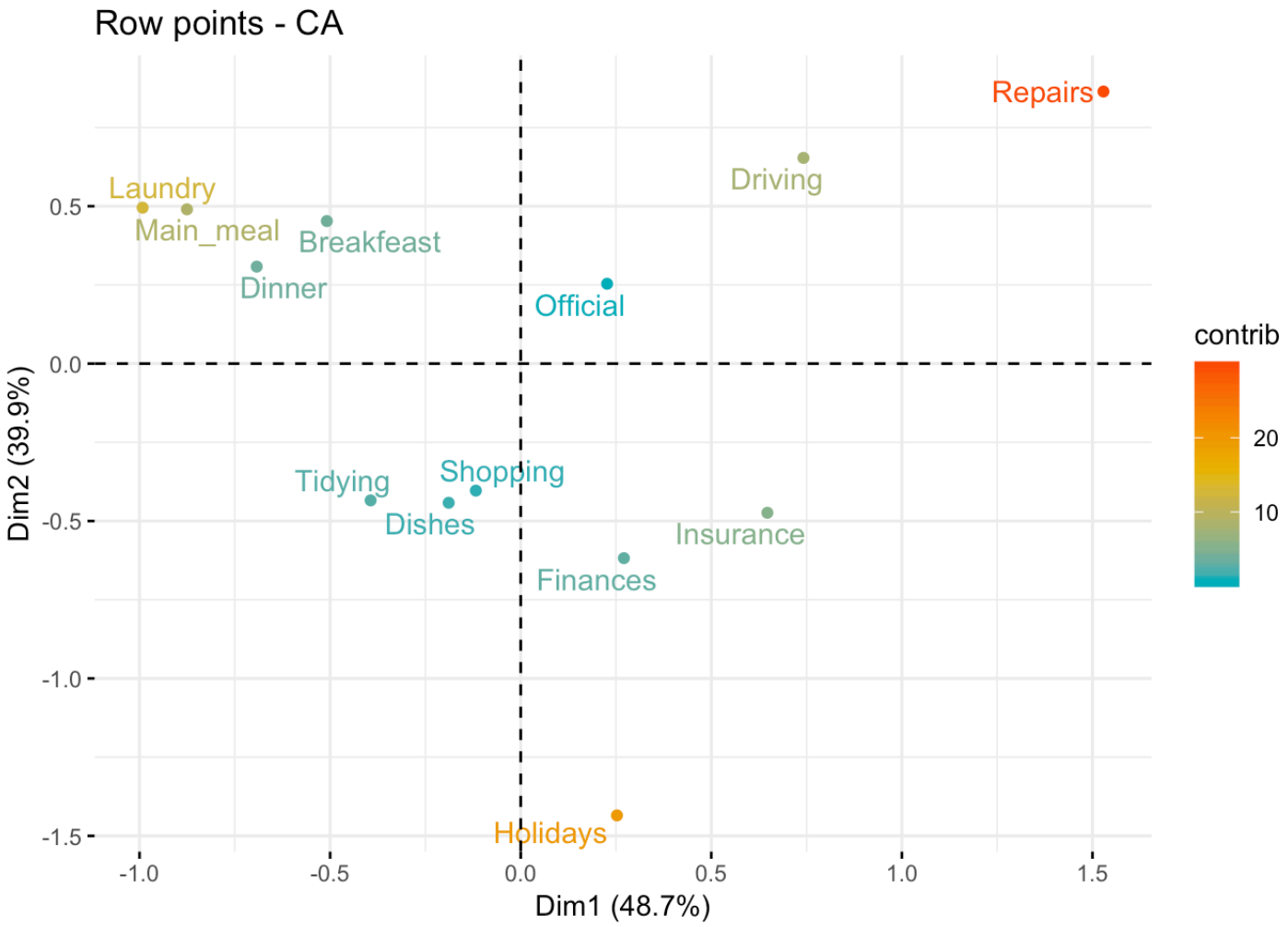


Puede verse que:

- Las filas *Repairs*, *Laundry*, *Main_meal* y *Driving* son las más importantes para la definición de la primera dimensión.
- Las filas *Holidays* y *Repairs* son las que más contribuyen a la dimensión 2.

Podemos resaltar estos puntos en el diagrama de dispersión como sigue:

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



Información para las columnas

Se deja como ejercicio al lector

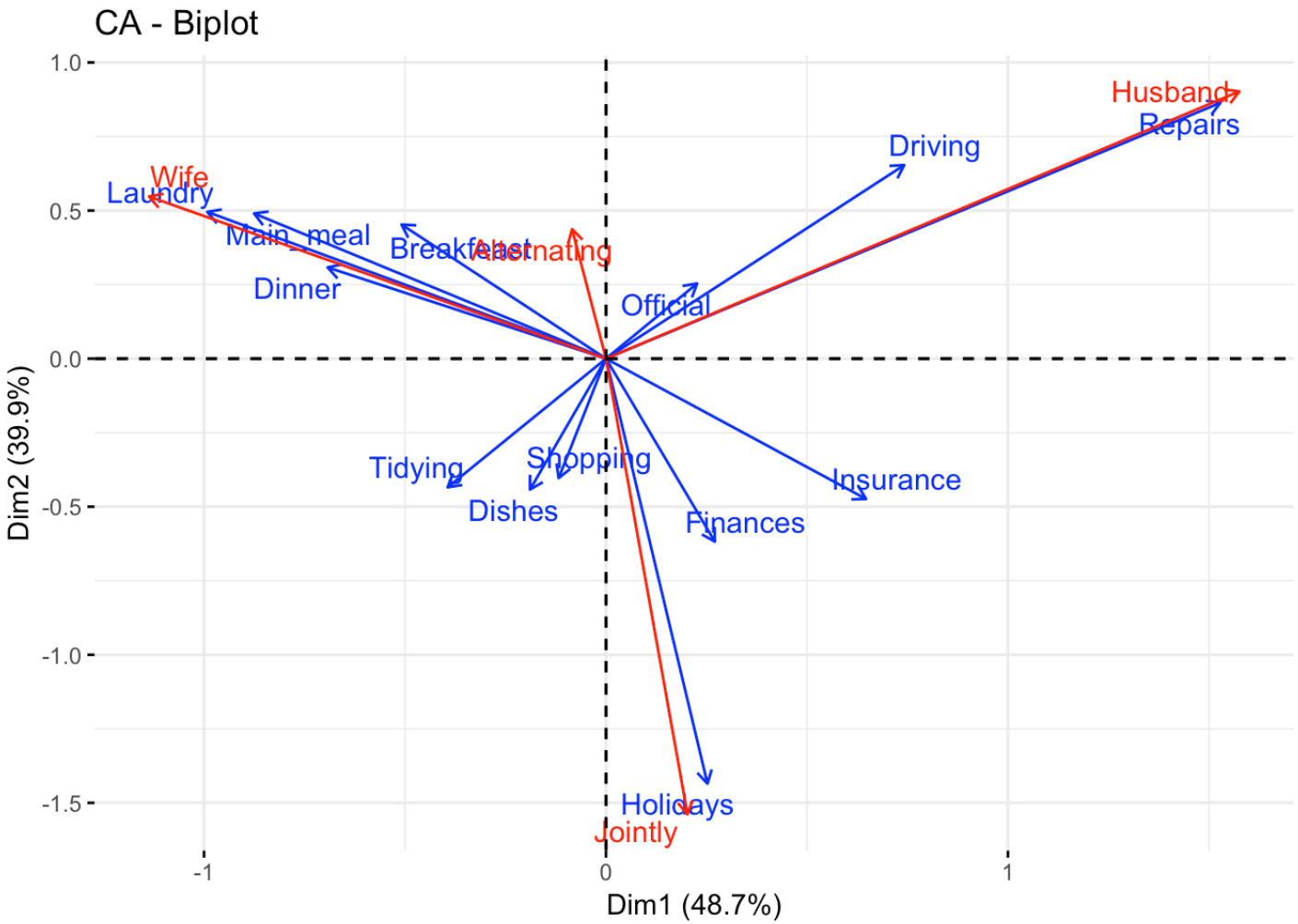
Biplots

Además de los biplots simétricos existen otras alternativas con propiedades diferentes.

Para poder interpretar directamente la elación entre los puntos fila y columna es necesario representar unos en el espacio de los otros y no mediante la representación simétrica que fue la propuesta por los autores originales. Explicaremos más estos conceptos cuando hablemos con más detalle de los métodos biplot.

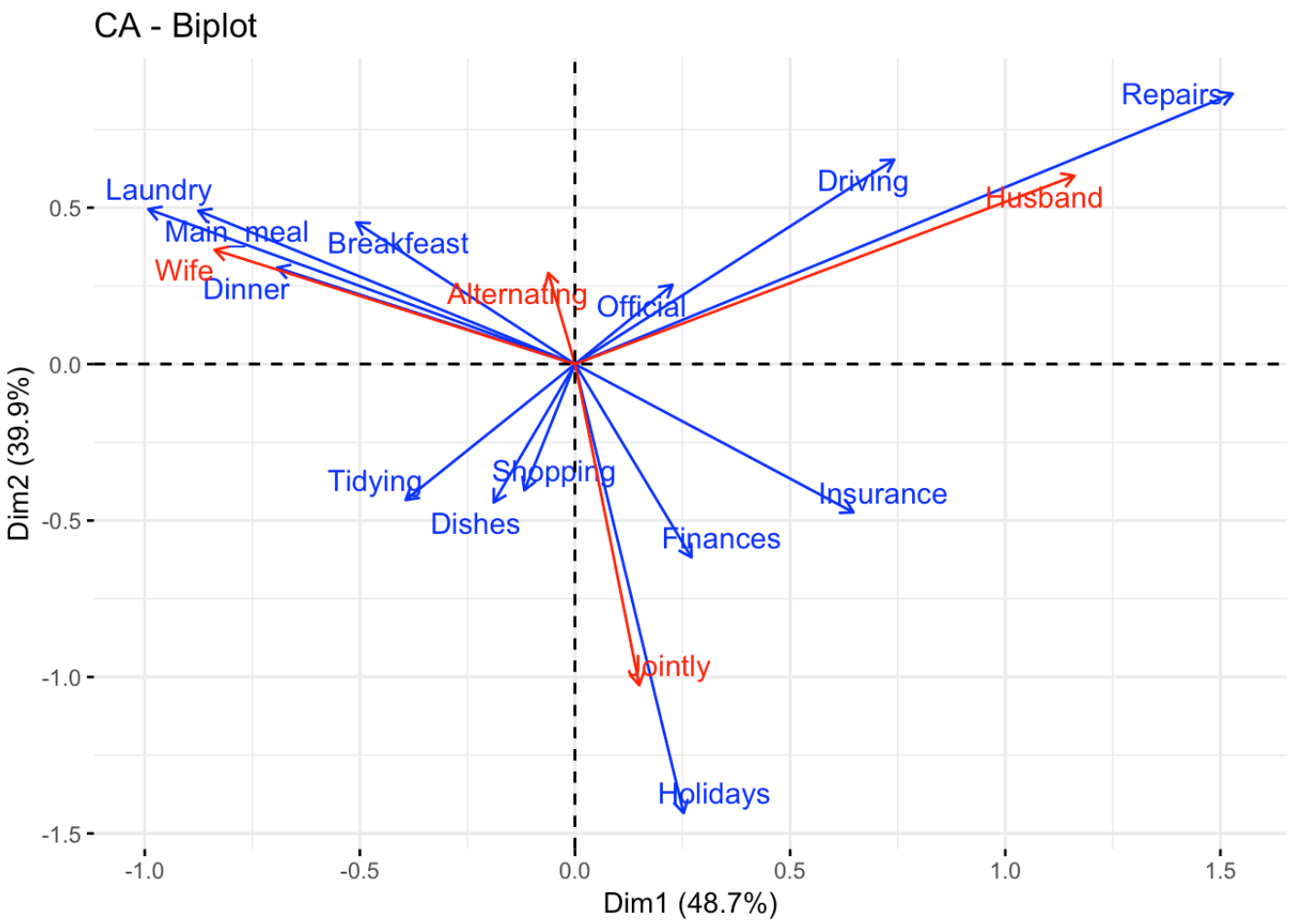
Podemos representar las columnas en el espacio de las filas, es la opción *rowprincipal* que representaremos a continuación.

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



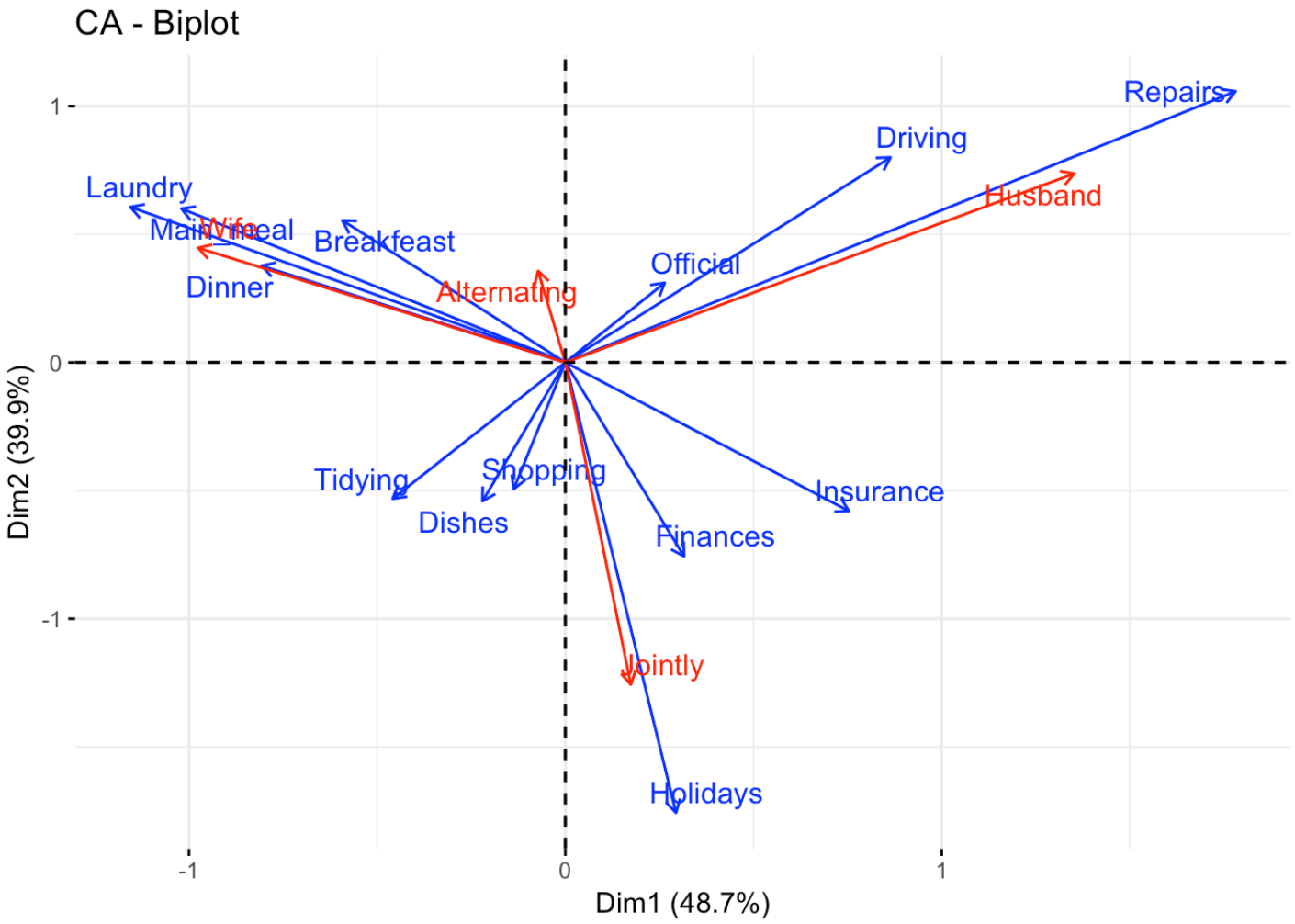
Podemos representar las filas en el espacio de las columnas, es la opción *columnprincipal* que representaremos a continuación.

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



Incluso un biplot simétrico pero que no conserva ni la métrica de las filas ni la de las columnas

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


CA - Biplot
Repairs
Driving
Husband
Laundry
Official
Breakfeast
Dinner
Alternating
Tidying
Shopping
Dishes
Insurance
Finances
Jointly
Holidays
1
0
-1
Dim2 (39.9%)
-1
0
1
Dim1 (48.7%)