Clase_1_PCA_autos

Diego Figueroa

2025-05-05

Clase 1 y 2 del Módulo II: Correlaciones y Reducción de Dimensiones

Carga de los datos

Los datos son cargados desde un archivo txt, y ocupamos la librería **FactoMineR** para realizar cálculos de PCA

```
Dataset <- read.table("C:/Users/diego/OneDrive/Escritorio/Diplomado Data
 ↪  Science/Diplomado PUCV/Módulo_2_Componentes_Principales/autos.txt",
                      header=TRUE,stringsAsFactors=TRUE, sep="",
                      na.strings="NA", dec=".", strip.white=TRUE,row.names="NOMBRE")
head(Dataset)
```

```
##                   CYL POT LAR ANCHO PESO VEL LUJO PRECIO
## ALFASUD-TI-1350  1350  79 393   161  870 165    B  30570
## AUDI-100-L       1588  85 468   177 1110 160   MB  39990
## SIMCA-1307-GLS   1294  68 424   168 1050 152    P  29600
## CITROEN-GS-CLUB  1222  59 412   161  930 151    P  28250
## FIAT-132-1600GLS 1585  98 439   164 1105 165    B  34900
## LANCIA-BETA-1300 1297  82 429   169 1080 160   MB  35480
```

Reducción de Dimensiones

La Reducción de Dimensiones busca eficientar el almacenamiento y procesamiento de los datos cuando nos enfrentamos a grandes cantidades de estos y una gran diversidad de features que explican lo que buscamos responder acerca de un fenómeno.

Sin embargo, en gran medida, todas las dimensiones son relevantes, por lo que buscamos encontrar un subset de estas que representen un nivel de información tal que explique lo más posible a una variable Y. Buscamos perder la menor cantidad de información con este método.

Lo que podríamos hacer es examinar representaciones bidimensionales (o un máximo de 3 dimensiones), que denotaremos como z_{1} y z_{2}, donde estarán representadas un alto porcentaje de las variables correlacionadas- Ambas explican un % de la varianza (la información de la base que dispongo).

Métodos para Reducir Dimensiones:

- PCA Robusto ()
- PCA (Aplica a transformaciones lineales)
- Kernel PCA (para casos de correlaciones entre variables no lineales)
- Singular value decomposition (SVD)[Para compresión de imágenes]
- Análisis Factorial (Es útil en marketing para ver grandes cantidades de variables que representarían o están relacionadas a pequeños conceptos englobadores, como 'satisfacción')
- Linear Discriminant Analysis (LDA)[supervising learning]

Componentes Principales (PCA)

Consiste en expresar un conjunto de combinaciones lineales de factores no correlacionados entre sí. Es importante resaltar el hecho de que el concepto de mayor información se relaciona con el de mayor variabilidad de varianza. Cuanto mayor sea la variabilidad de los datos (varianza) se considera que existe mayor ainformación, lo cual está relacionado al concepto de entropía.

Los PCA Máx_{varianza} ya que buscan ganar la mayor cantidad de datos que pueda; y Min_{residuales} La distancia del dato hacia el vector de la variable estimada

Procedimiento Se requiere en una primera etapa, y siempre con un uso más exploratorio de los datos, revisar la matriz de correlación. Aquí se busca cierto razonamiento de selección de variables que tenga cierta correlación (por ahora lineal) de los datos, ocupando la correlación de **Pearson**. Igualmente dejaremos el print de cómo se puede hacer con **Spearman** y **Kendal**

```
# Mide la relación lineal entre dos variables continuas. Y supone normalidad en los
 ↪  datos, funcionando poco en presencia de valores atípicos o relaciones no lineales
cor(Dataset[,c("CYL", "POT", "LAR", "ANCHO", "PESO",
 ↪  "VEL","PRECIO")],use="complete",method = "pearson")
```

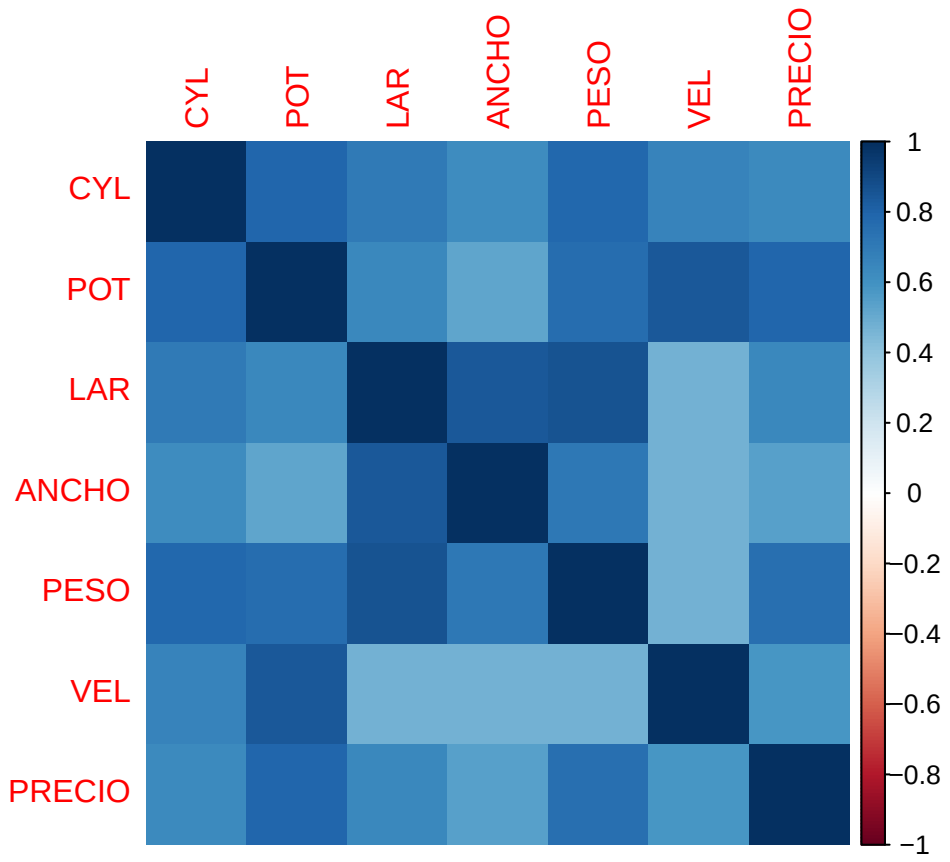
```
##              CYL       POT       LAR     ANCHO      PESO       VEL    PRECIO
## CYL    1.0000000 0.7966277 0.7014619 0.6297572 0.7889520 0.6649340 0.6385812
## POT    0.7966277 1.0000000 0.6413624 0.5208320 0.7652930 0.8443795 0.7987004
## LAR    0.7014619 0.6413624 1.0000000 0.8492664 0.8680903 0.4759285 0.6437569
## ANCHO  0.6297572 0.5208320 0.8492664 1.0000000 0.7168739 0.4729453 0.5466494
## PESO   0.7889520 0.7652930 0.8680903 0.7168739 1.0000000 0.4775956 0.7532948
## VEL    0.6649340 0.8443795 0.4759285 0.4729453 0.4775956 1.0000000 0.5817597
## PRECIO 0.6385812 0.7987004 0.6437569 0.5466494 0.7532948 0.5817597 1.0000000
```

```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.3
```

```
## corrplot 0.95 loaded
```

```r
corrplot(cor(Dataset[,c("CYL", "POT", "LAR", "ANCHO", "PESO",
 ↪  "VEL","PRECIO")],use="complete",method="pearson"),method="shade")
```




```r
# Mide la asociación monótona( cuanto una variable sube, la otra también sube o baja, no
 ↪  necesariamente de forma lineal)
# Se basa en rangos, por lo que es más robusta a outliers y no requiere de normalidad. I
 ↪  deal cuando la relación es no lineal, pero sí ordenada.
cor(Dataset[,c("CYL", "POT", "LAR", "ANCHO", "PESO",
 ↪  "VEL","PRECIO")],use="complete",method = "spearman")
```

```
##              CYL       POT       LAR     ANCHO      PESO       VEL    PRECIO
## CYL    1.0000000 0.7939976 0.8517562 0.7562306 0.8196384 0.6514392 0.6205473
## POT    0.7939976 1.0000000 0.7318856 0.6143464 0.6804773 0.8589333 0.7801371
## LAR    0.8517562 0.7318856 1.0000000 0.8527044 0.9147288 0.4665011 0.6329376
## ANCHO  0.7562306 0.6143464 0.8527044 1.0000000 0.7846445 0.4600945 0.6200171
## PESO   0.8196384 0.6804773 0.9147288 0.7846445 1.0000000 0.3731859 0.6818185
## VEL    0.6514392 0.8589333 0.4665011 0.4600945 0.3731859 1.0000000 0.5555663
## PRECIO 0.6205473 0.7801371 0.6329376 0.6200171 0.6818185 0.5555663 1.0000000
```

```r
# Similar a Spearman, pero usa pares concordantes/discordantes. Más conservadora y mejor
 ↪  para muestras pequeñas
cor(Dataset[,c("CYL", "POT", "LAR", "ANCHO", "PESO",
 ↪  "VEL","PRECIO")],use="complete",method = "kendall")
```

```
##              CYL       POT       LAR     ANCHO      PESO       VEL    PRECIO
## CYL    1.0000000 0.6267224 0.6776316 0.5840110 0.6600696 0.4799050 0.4983633
## POT    0.6267224 1.0000000 0.5133790 0.4557929 0.4816296 0.7534954 0.6113800
## LAR    0.6776316 0.5133790 1.0000000 0.7048409 0.7920835 0.3312021 0.4196744
## ANCHO  0.5840110 0.4557929 0.7048409 1.0000000 0.6196164 0.3724226 0.4549742
## PESO   0.6600696 0.4816296 0.7920835 0.6196164 1.0000000 0.2576997 0.4671154
## VEL    0.4799050 0.7534954 0.3312021 0.3724226 0.2576997 1.0000000 0.4446486
## PRECIO 0.4983633 0.6113800 0.4196744 0.4549742 0.4671154 0.4446486 1.0000000
```

Se necesitan aplicar test de Barlett y el cálculo del índice KMO. El primero se utiliza para PCA como Análisis Factorial para evaluar si una matriz de correlación de las variables es significativamente diferente a una matriz de identidad.

H_{0}: La matriz de correlación es una matriz de identidad H_{1}: La matriz de correlación no es una matriz de identidad, por lo que existe correlación significativa entre las variables.

```r
library("EFAtools")
```

```
## Warning: package 'EFAtools' was built under R version 4.3.3
```

```r
BARTLETT(cor(Dataset_sin_lujo),N=nrow(Dataset))
```

```
##
## v The Bartlett's test of sphericity was significant at an alpha level of .05.
##   These data are probably suitable for factor analysis.
##
##    ²(21) = 109.96, p < .001
```

En el segundo caso, el índice KMO busca evaluar ka adecuación de los datos para la reducción de dimensionalidad. Responde a la pregunta ¿Es apropiado aplicar PCA a este conjunto de datos? Su valor varía entre 0 y 1. Un KMO alto sugiere que las variables comparten una cantidad significativa de varianza común y que la correlaciones entre ellas no son debido a la influencia de otras variables. En este escenario, el PCA tiene más probabilidad de identificar componentes principales que representen de manera efectiva la estructura subyacente de los datos y lograr una reducción de dimensionalidad útil.

```r
KMO(cor(Dataset_sin_lujo),cor_method = 'spearman')
```

```
##
## -- Kaiser-Meyer-Olkin criterion (KMO) ---------------------------------------
##
## v The overall KMO value for your data is middling.
##   These data are probably suitable for factor analysis.
##
##   Overall: 0.79
##
##   For each variable:
##    CYL    POT    LAR  ANCHO   PESO    VEL PRECIO
##  0.929  0.717  0.827  0.792  0.762  0.651  0.900
```

Ahora, al ya establecer que es necesario el cálculo de los PCA, entonces ejecuto considerando todas las variables, pero de manera discrecional dejaremos el PRECIO fuera, ya que puede bien ser explicado por las otras 6 variables.

```
library(FactoMineR)
```

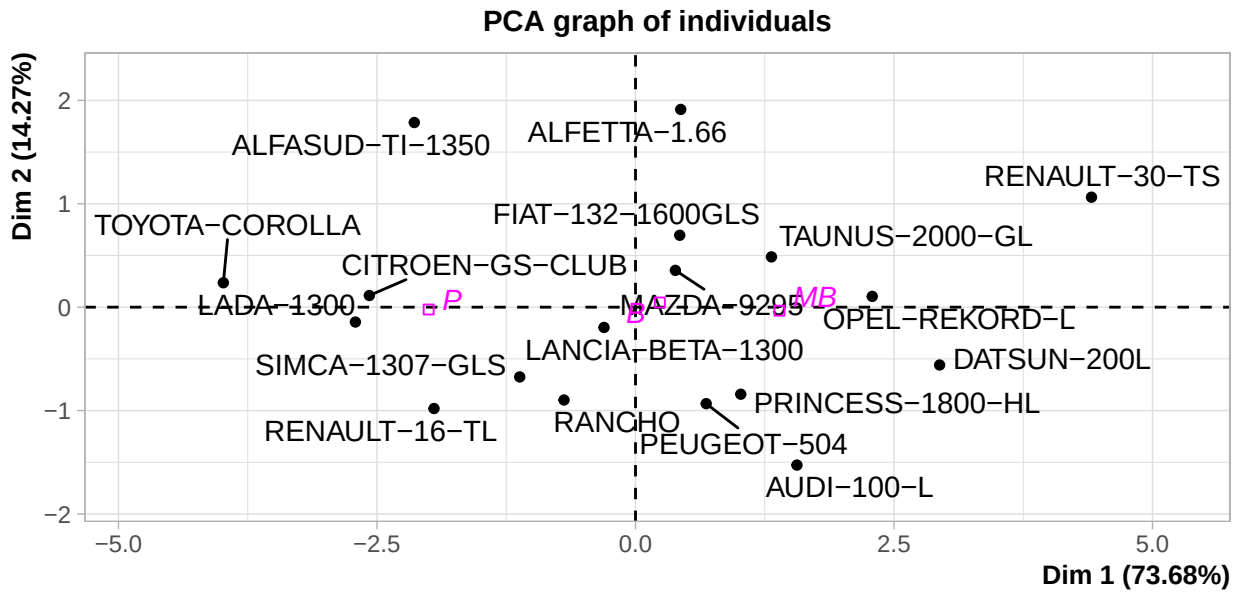
```
## Warning: package 'FactoMineR' was built under R version 4.3.3
```

```
Dataset.PCA<-Dataset[, c("CYL", "POT", "LAR", "ANCHO", "PESO", "VEL","PRECIO","LUJO")]

res<-PCA(Dataset.PCA , scale.unit=TRUE, ncp=5, quanti.sup=c(7: 7),quali.sup=c(8: 8),
 ↪  graph =FALSE)
```

Aquí estamos representando los datos en dos Dimensiones

```
print(plot.PCA(res, axes=c(1, 2), choix="ind", habillage="none",col.ind="black",
 ↪  col.ind.sup="blue", col.quali="magenta", label=c("ind","ind.sup",
 ↪  "quali"),new.plot=TRUE, title="PCA graph of individuals"))
```

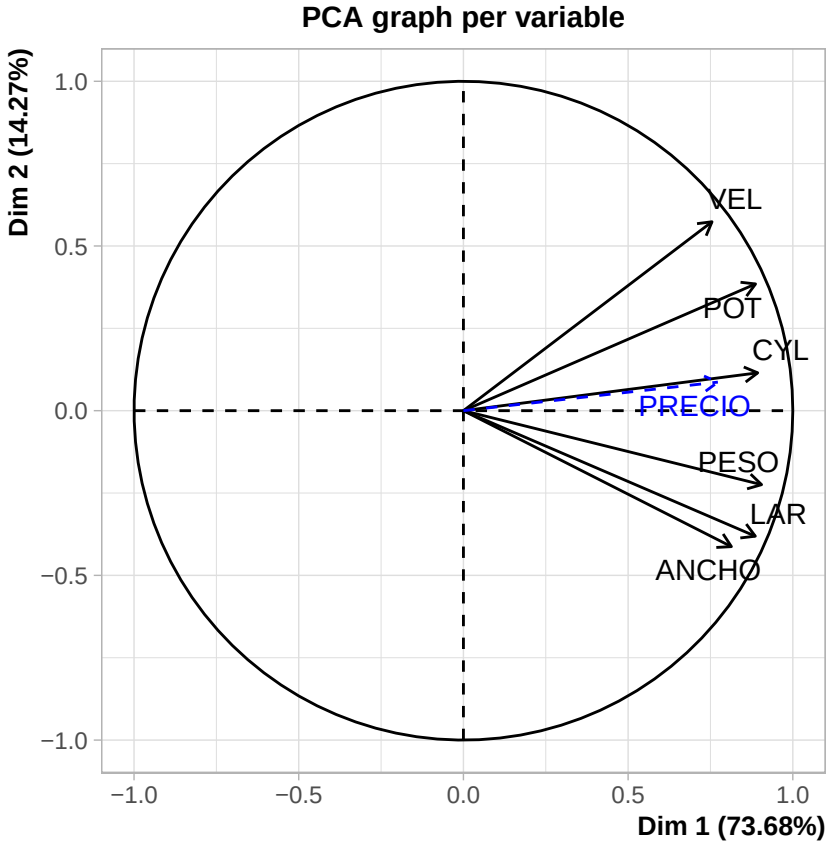


Mientras que aquí puedo ver qué variables representan o son aglomeradas por las dos dimensiones construidas. Aquí apróximadamente perdimos un 10% del total de información de la varianza con el método de PCA **100% - (73.68% + 14.27%)=12.05%**. Si revisamos el gráfico cada vector es una representación visual de la información obtenida del *feature* de cada auto. Vemos que mientras más se acerca a la esfera de radio 1, más cerca está de tener la información completa, pese al PCA. Aquí la que más se acerca a su máximo de información son las variables de LARGO, POTENCIA y VELOCIDAD, mientras que las que más pierden son CYL y PESO.

Adicionalmente, en la dirección que muestre el vector, será la relación directa o inversa que tendrán los datos. Por ejemplo,en el eje horizontal aquellos que se encuentren a la derecha del círculo, serán más veloces, más potentes, con más peso, largos y anchos. En cambio si van en la dirección contraria, significa que que son lentos, menos potentes, de menor peso, más cortos y más angostos. Aquí podríamos ver autos según el cuidado del medioambiente.

Ahora bien, en el eje vertical, debido a que la información considera solamente un 14% de la información, acaba siendo más por intuición. Arriba se encuentran autos más veloces y potentes, pero de menor peso, más chicos y angostos (podrían ser autos deportivos quizás), mientras que mirando hacia abajo es lo opuesto, pudiendo ser autos más friendly o familiares.

```
print(plot.PCA(res, axes=c(1, 2), choix="var", new.plot=TRUE,col.var="black",
 ↪  col.quanti.sup="blue", label=c("var", "quanti.sup"),lim.cos2.var=0, title="PCA graph
 ↪  per variable"))
```



Los eigenvalues (o valores propios) representan la varianza explicada por cada componente principal. Cuanto mayor sea un eigenvalue, más información (variabilidad) de los datos originales contiene ese componente. * Si sumamos todos los eigenvalues, obtenemos la varianza total del conjunto de datos estandarizados (equivale al número de variables cuando están estandarizadas). * El % de varianza explicada por cada dimensión nos dice qué proporción del comportamiento original de los datos es retenido en ese eje.

Interpretación Eigenvalues: ¿Cuánta información explica cada componente?

- El primer componente explica casi el 79% de la información del dataset, lo que indica que resume gran parte de la variabilidad.
- Con los primeros dos componentes, alcanzamos más del 90%, lo que sugiere que podemos visualizar y analizar los datos en un espacio bidimensional sin perder demasiada información.

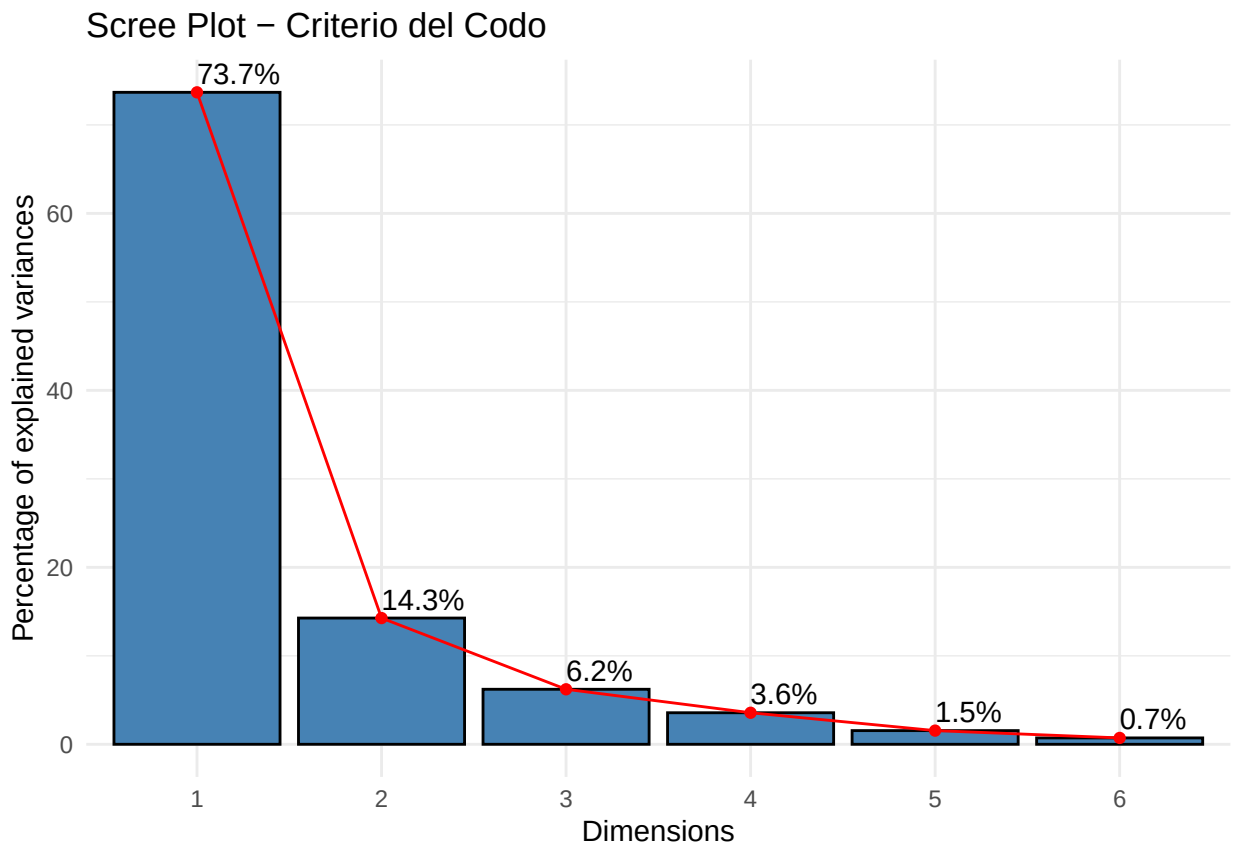
Importancia del % de varianza explicada:

- Nos ayuda a decidir cuántos componentes mantener. En ciencia de datos se suele usar el criterio del codo para elegir el punto donde el aumento de varianza explicada comienza a disminuir drásticamente.

```
## Warning: package 'factoextra' was built under R version 4.3.3
```

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

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```



- cos2: Indica qué tan bien se representa un individuo en un eje (como un R^2); valores cercanos a 1 indican una buena representación.
- ctr (contribución): Mide cuánto influye un individuo en la formación del eje/Dim_{i}. Valores altos indican que ese punto fue determinante para definir esa dimensión.

Ejemplo Individuos (observaciones): ¿Cómo se proyectan los autos?:

- El Toyota Corolla tiene un cos2 = 0.976 en Dim 1, lo que significa que su ubicación en ese eje es muy representativa de su perfil.
- También tiene una contribución importante a ese eje (16.292), lo que indica que ayuda a definir la dirección principal de la variabilidad en los datos.
- Otra manera de reforzar la presencia de valores atípicos que puedan afectar nuestro cálculo es por medio de las Distancias entre los puntos, donde observo en la siguiente tabla que el RENAULT-30-TS y el TOYOTA-COROLLA tienen una distancia promedio muy por sobre el resto de los autos. Lo que también informa en que contribuye o influencia en la construcción de la primera dimensión, donde respectivamente para la dimensión 1 influencia en un 22.43% y un 19.69%. Es bastante para ser outliers. Existen algunas soluciones como ocupar PCA más robustos que puedan trabajar con valores de outliers.

```r
summary(res, nb.dec = 3, nbelements=10, nbind = 10, ncp = 3, file="")
```

```
##
## Call:
## PCA(X = Dataset.PCA, scale.unit = TRUE, ncp = 5, quanti.sup = c(7:7),
##      quali.sup = c(8:8), graph = FALSE)
##
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6
## Variance               4.421   0.856   0.373   0.214   0.093   0.043
## % of var.             73.681  14.268   6.218   3.565   1.547   0.722
## Cumulative % of var.  73.681  87.949  94.166  97.732  99.278 100.000
##
## Individuals (the 10 first)
##                       Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2
## ALFASUD-TI-1350  |  2.868 | -2.139  5.749  0.556 |  1.786 20.693  0.388 |
## AUDI-100-L       |  2.583 |  1.561  3.064  0.365 | -1.527 15.133  0.349 |
## SIMCA-1307-GLS   |  1.469 | -1.119  1.575  0.580 | -0.675  2.953  0.211 |
## CITROEN-GS-CLUB  |  2.604 | -2.574  8.324  0.977 |  0.113  0.083  0.002 |
## FIAT-132-1600GLS |  1.081 |  0.428  0.230  0.157 |  0.696  3.140  0.414 |
## LANCIA-BETA-1300 |  1.065 | -0.304  0.116  0.082 | -0.196  0.250  0.034 |
## PEUGEOT-504      |  1.230 |  0.684  0.588  0.309 | -0.933  5.650  0.575 |
## RENAULT-16-TL    |  2.374 | -1.948  4.771  0.674 | -0.980  6.238  0.171 |
## RENAULT-30-TS    |  4.668 |  4.410 24.437  0.892 |  1.064  7.342  0.052 |
## TOYOTA-COROLLA   |  4.036 | -3.986 19.964  0.975 |  0.236  0.362  0.003 |
##                      Dim.3    ctr   cos2
## ALFASUD-TI-1350   0.572  4.870  0.040 |
## AUDI-100-L        1.315 25.762  0.259 |
## SIMCA-1307-GLS    0.457  3.104  0.097 |
## CITROEN-GS-CLUB   0.149  0.329  0.003 |
## FIAT-132-1600GLS -0.193  0.556  0.032 |
## LANCIA-BETA-1300  0.676  6.801  0.402 |
## PEUGEOT-504      -0.257  0.982  0.044 |
## RENAULT-16-TL    -0.620  5.716  0.068 |
## RENAULT-30-TS    -0.594  5.246  0.016 |
## TOYOTA-COROLLA   -0.303  1.368  0.006 |
##
## Variables
##                     Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr
## CYL              |  0.893 18.057  0.798 |  0.115  1.542  0.013 | -0.216 12.504
## POT              |  0.887 17.791  0.787 |  0.385 17.287  0.148 | -0.113  3.420
## LAR              |  0.886 17.763  0.785 | -0.381 16.959  0.145 |  0.041  0.457
## ANCHO            |  0.814 14.971  0.662 | -0.413 19.899  0.170 |  0.369 36.587
## PESO             |  0.905 18.534  0.819 | -0.225  5.889  0.050 | -0.296 23.464
## VEL              |  0.755 12.884  0.570 |  0.574 38.423  0.329 |  0.297 23.568
##                    cos2
## CYL               0.047 |
## POT               0.013 |
## LAR               0.002 |
## ANCHO             0.136 |
## PESO              0.088 |
## VEL               0.088 |
```

```
##
## Supplementary continuous variable
##                     Dim.1   cos2    Dim.2   cos2    Dim.3   cos2
## PRECIO           |  0.772  0.597 |  0.087  0.008 | -0.134  0.018 |
##
## Supplementary categories
##                      Dist    Dim.1   cos2 v.test    Dim.2   cos2 v.test
## B                 |  0.353 |  0.235  0.445  0.368 |  0.045  0.016  0.161 |
## MB                |  1.435 |  1.392  0.942  1.931 | -0.034  0.001 -0.107 |
## P                 |  2.004 | -2.000  0.996 -2.433 | -0.023  0.000 -0.062 |
##                    Dim.3   cos2 v.test
## B                  0.114  0.104  0.614 |
## MB                -0.075  0.003 -0.358 |
## P                 -0.070  0.001 -0.291 |
```

Interpretación: de las Variables activas: ¿Qué variables explican los ejes?

- El primer eje está principalmente determinado por variables como PESO, LARGO, ANCHO y POTENCIA, todas relacionadas con el tamaño y fuerza del auto.
- Estas variables tienen un cos2 cercano a 1 en Dim 1, lo que indica que están altamente correlacionadas con ese componente.

¿Qué son variables suplementarias Son variables no utilizadas para construir los ejes, pero que se proyectan sobre el espacio PCA para ver cómo se relacionan con la estructura hallada.

En este caso Variables suplementarias cuantitativas:

- VEL (Velocidad) no definió los ejes, pero está moderadamente asociada al primer componente (cos2 = 0.433).
- Esto sugiere que los autos más veloces tienden a estar asociados con dimensiones como el peso o la potencia.

¿Qué son categorías suplementarias? Son variables categóricas (factores) que no definen los ejes, pero permiten ver cómo se distribuyen sus niveles en el espacio de componentes. En este caso: la categoría LUJO.

Se evalúan mediante el v.test las Categorías suplementarias cualitativas (LUJO):

- Valores altos (en valor absoluto) indican que una categoría está fuertemente asociada con un eje y su posición no es aleatoria.

Esto permite estudiar si, por ejemplo, los autos de lujo están agrupados en una región específica del plano PCA (lo que implicaría un perfil técnico distinto a los no lujosos).

palabras clave: Contribución a la Dimensión i ; Alineación cercana a 1 por Eigenvalues; y Representación de la variable en la dimensión