Descomposición en valores singulares: introducción y aplicaciones

Problemas para la ciencia de datos

PID_00262389

Francesc Pozo Montero
Jordi Ripoll Missé

Universitat Oberta de Catalunya

Francesc Pozo Montero

Licenciado en Matemáticas por la Universidad de Barcelona (2000) y doctor en Matemática Aplicada por la Universidad Politécnica de Cataluña (2005). Ha sido profesor asociado de la Universidad Autónoma de Barcelona y profesor asociado, colaborador y actualmente profesor agregado en la Universidad Politécnica de Cataluña. Además, es cofundador del Grupo de Innovación Matemática E-learning (GIMEL), responsable de varios proyectos de innovación docente y autor de varias publicaciones. Como miembro del grupo de investigación consolidado CoDAlab, centra su investigación en la teoría de control y las aplicaciones en ingeniería mecánica y civil, así como en el uso de la ciencia de datos para la monitorización de la integridad estructural y para la monitorización de la condición, sobre todo en turbinas eólicas.

Jordi Ripoll Missé

Licenciado en Matemáticas y doctor en Ciencias Matemáticas por la Universidad de Barcelona (2005). Profesor colaborador de la Universitat Oberta de Catalunya desde 2011 y profesor del Departamento de Informática, Matemática Aplicada y Estadística de la Universidad de Girona (UdG) desde 1996, donde actualmente es profesor agregado y desarrolla tareas de investigación en el ámbito de la biología matemática (modelos con ecuaciones en derivadas parciales y dinámica evolutiva). También ha sido profesor y tutor de la UNED en dos etapas, primero en el centro asociado de Terrassa y actualmente en el de Girona. Ha participado en numerosos proyectos de innovación docente, especialmente en cuanto al aprendizaje de las matemáticas en línea.

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por la profesora: Cristina Cano Bastidas (2019)

Primera edición: febrero 2019

Av. Tibidabo, 39-43, 08035 Barcelona
Diseño: Manel Andreu
Realización editorial: Oberta UOC Publishing, SL

Índice

1. Problemas: análisis de componentes principales (PCA) 5

2. Problemas: descomposición en valores singulares (SVD).... 9

3. Soluciones de los problemas: análisis de componentes principales (PCA) 12

4. Soluciones de los problemas: descomposición en valores singulares (SVD) 18

1. Problemas: análisis de componentes principales (PCA)

En este apartado presentamos un problema con datos reales que se puede considerar y trabajar mediante el análisis de componentes principales. En la resolución proponemos usar una librería especial de `R`, de forma que se pueda complementar el código que también podréis encontrar en el material "Descomposición en valores singulares: introducción y aplicaciones. Estudio de caso y guía de resolución en R".

1. Considerad los datos proporcionados por el conjunto `mtcars`, libremente disponible en `R`, extraídos de la revista estadounidense *Motor Trend* (1974). Este conjunto de datos presenta, para 32 modelos de coches de los años 1973 y 1974, el consumo de combustible y diez aspectos más. Las variables medidas son:

1) **mpg** (consumo de combustible): en millas por galón de los Estados Unidos; los coches más potentes y más pesados tienden a consumir más combustible.

2) **cyl** (número de cilindros): los coches más potentes suelen tener más cilindros.

3) **disp** (desplazamiento): en pulgadas cúbicas (en inglés, *cubic inch*); el volumen combinado de los cilindros del motor.

4) **hp** (potencia bruta): es una medida de la potencia generada por el coche.

5) **drat** (relación del eje posterior): se describe como un giro del eje de transmisión correspondiente a un giro de las ruedas. Los valores más altos disminuirán la eficiencia del combustible.

6) **wt** (peso): correspondiente a 1.000 libras.

7) **qsec** (tiempo de 1/4 de milla): velocidad y aceleración de los coches.

8) **vs** (bloque del motor): indica si el motor del vehículo tiene forma de V o una forma recta más común.

9) **am** (transmisión): indica si la transmisión del automóvil es automática (0) o manual (1).

10) **gear** (número de marchas hacia adelante): los coches deportivos suelen tener más marchas.

11) **carb** (número de carburadores): asociados a motores más potentes.

Con este conjunto de datos procederemos de la siguiente manera:

a) Cargad el conjunto de datos `mtcars` con esta instrucción:

```
> data("mtcars")
> mtcars
```

Observad los modelos de coche y los valores de sus variables.

b) Calculad con `R` el análisis de componentes principales y denominad `mtcars.pca` a la estructura resultante. Descartad para el análisis las variables binarias **vs** y **am**. No os olvidéis de escalar vuestros datos.

c) Ejecutad la instrucción:

```
> summary(mtcars.pca)
```

¿Cuál es la cantidad de información (proporción de varianza) retenida por la primera componente principal? ¿Cuál es la cantidad de información retenida por las dos primeras componentes principales?

d) Ejecutad la instrucción:

```
> str(mtcars.pca)
```

Así tendréis acceso a la estructura `mtcars.pca`. ¿Cuál es la media aritmética de la variable **cyl**?

e) Antes de seguir con la representación gráfica de las dos primeras componentes principales, cargad `devtools` con esta instrucción:

```
> install.packages("devtools")
> library(devtools)
> install_github("vqv/ggbiplot")
```

A continuación, ejecutad la instrucción:

```
> ggbiplot(mtcars.pca)
```

La gráfica resultante os mostrará 32 puntos sobre el plano generado por las dos primeras componentes principales, así como las nueve variables originales que contribuyen a las componentes principales. ¿Cuáles son las variables originales que más contribuyen a la primera componente principal? ¿Cuáles son las variables originales que más contribuyen a la segunda componente principal? ¿Y cuáles son las variables originales que menos contribuyen a la segunda componente principal?

Con la instrucción

```
> ggbiplot(mtcars.pca, labels = rownames(mtcars))
```

obtendréis la misma gráfica, pero veréis el nombre de cada uno de los modelos de coche.

f) A continuación, agruparemos los modelos de coche según el origen geográfico del fabricante. Por ejemplo, el modelo Mazda RX4 es japonés, el Hornet 4 Drive es estadounidense y el modelo Merc 240D (Mercedes 240D) es europeo. El vector completo con los orígenes de los modelos de coche es:

```
> mtcars.country <- c(rep("Japón", 3), rep("EE. UU.", 4),
rep("Europa", 7),rep("EE. UU.",3),"Europa", rep("Japón", 3),
rep("EE. UU.",4), rep("Europa", 3), "EE. UU.", rep("Europa", 3))
```

De forma que, con la instrucción:

```
> ggbiplot(mtcars.pca, ellipse = TRUE, labels = rownames(mtcars),
 groups = mtcars.country)
```

obtenemos la misma gráfica, pero con los modelos de coche agrupados por su origen. Además, hemos incluido una elipse que engloba los modelos de coche dentro de cada grupo geográfico. ¿Qué destacaríais —en cuanto a las variables originales— de los modelos de coche estadounidenses? ¿Y de los modelos de coche japoneses? ¿Y de los europeos?

g) Tenemos un nuevo modelo de coche, cuyo origen desconocemos, con las características siguientes:

– **mpg**: 18
– **cyl**: 7
– **disp**: 300
– **hp**: 158
– **drat**: 3
– **wt**: 3.4
– **qsec**: 18
– **vs**: 0
– **am**: 0
– **gear**: 3
– **carb**: 2

Queremos proyectar este nuevo modelo de coche para intentar conocer su origen.

```
> s<-c(18,7,300,158,3,3.4,18,0,0,3,2)
> s.sc<-(s[c(1:7,10,11)] -mtcars.pca$center)/mtcars.pca$scale
> u<-s.sc%*%mtcars.pca$rotation
> mtcars.plus.pca<-mtcars.pca
> mtcars.plus.pca$x<-rbind(mtcars.plus.pca$x,u)
> mtcars.countryplus<-c(mtcars.country,"Desconocido")
> ggbiplot(mtcars.plus.pca, ellipse = TRUE, circle = FALSE, var.axes=TRUE,
  labels=c(rownames(mtcars), "new"), groups=mtcars.countryplus)+
  scale_colour_manual(name="Origen",
  values= c("forest green", "red3", "violet", "dark blue"))+
  theme(legend.position = "bottom")
```

En la primera línea del código definimos las características del nuevo modelo; en la segunda, escalamos nuestros datos respecto a los datos originales, y en la tercera, proyectamos el modelo de coche sobre el espacio vectorial generado por las componentes principales. En las dos líneas siguientes, se amplía el conjunto de datos proyectados `mtcars.pca$x` con el nuevo punto. En la sexta línea de código especificamos la categoría del coche nuevo como "Desconocido". Finalmente, representamos gráficamente todos los modelos de coche, incluido el nuevo. Según la posición del nuevo modelo, ¿cuál podría ser su origen?

2. Problemas: descomposición en valores singulares (SVD)

La descomposición en valores singulares es una técnica muy útil en el campo de la ciencia de datos, ya que permite descomponer o factorizar una matriz $\mathbf{A}$ como suma de otras matrices de rango 1. La matriz $\mathbf{A}$ presenta esta forma:

$$\mathbf{A} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r^T = \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^T$$

donde r es el rango de la matriz $\mathbf{A}$, σ_i son los valores singulares (valores positivos ordenados en orden decreciente) y $\mathbf{u}_i$ y $\mathbf{v}_i$ son los vectores singulares por la izquierda y por la derecha, respectivamente.

Una de las aplicaciones más importantes de la descomposición en valores singulares es la compresión de imágenes. En este sentido, por ejemplo, una imagen de $m \times n$ píxeles —en escala de grises— se representa por una matriz, en la que cada elemento representa la intensidad del color gris de cada píxel. Para almacenar esta matriz necesitamos $m \cdot n$ números. Supongamos que el rango de esta matriz es r. Entonces, si consideramos la descomposición en valores singulares:

$$\mathbf{A} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r^T = \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^T$$

ahora es necesario almacenar

$$r \cdot (1 + m + n)$$

números. En función del rango r de la matriz, esto ya puede suponer un cierto nivel de compresión o de reducción de la dimensionalidad. Pero podemos ir aún más allá, ya que podemos considerar la aproximación dada por la ν-ésima suma parcial, donde $\nu \leq r$:

$$\mathbf{A}_\nu = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T + \cdots + \sigma_\nu \mathbf{u}_\nu \mathbf{v}_\nu^T = \sum_{j=1}^{\nu} \sigma_j \mathbf{u}_j \mathbf{v}_j^T$$

En este caso, ya solo es necesario almacenar

$$\nu \cdot (1 + m + n)$$

números. ¿Cuál es la magnitud del error que se comete cuando realizamos esta aproximación? Según el módulo "Descomposición en valores singulares: introducción y aplicaciones", la diferencia en norma 2 entre estas dos matrices es igual a:

$$\|\mathbf{A} - \mathbf{A}_\nu\|_2 = \sigma_{\nu+1},$$

es decir, será pequeña en función de la magnitud que tenga el valor singular σ_{v+1}. Notad que, por ejemplo, si $m = n = 1000$, $r = 300$ y $v = 50$, podemos reducir la cantidad de información que debemos almacenar en un 89.995 % con un error en norma 2 que sería igual a la magnitud del valor singular σ_{v+1} —previsiblemente, muy pequeño.

El objetivo de los problemas que hay a continuación es —además de calcular de forma precisa, automatizada o con *software* la descomposición en valores singulares— profundizar en la comprensión del método mediante la descomposición de las matrices que se presentan como suma de matrices de rango 1. Podréis observar que:

1) La solución no es única. Es decir, podéis obtener soluciones diferentes a las propuestas en el apartado 3.

2) En los problemas planteados, no se calculan los valores singulares. En particular, eso es consecuencia del hecho de que no buscamos que los vectores $\mathbf{u}_i$ y $\mathbf{v}_i$ sean unitarios.

3) Tanto los vectores $\mathbf{u}_i$ como los vectores $\mathbf{v}_i$ son la base de los subespacios vectoriales generados por las columnas y por las filas, respectivamente. Ahora bien, estas bases no son necesariamente ortonormales, es decir, los vectores pueden no ser ortogonales ni tener norma 1.

1. Considerad esta matriz:

$$\mathbf{A} = (a_{ij}) = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 6 & 8 \\ 3 & 6 & 9 & 12 \\ 4 & 8 & 12 & 16 \end{bmatrix}$$

donde $a_{ij} = i \cdot j$. Esta matriz corresponde a una imagen de 4×4 píxeles. Calculad el rango r de la matriz $\mathbf{A}$. Reescribid $\mathbf{A}$ como suma de r matrices de rango 1 de la forma $\mathbf{u}\mathbf{v}^\intercal$. No uséis ningún *software* para resolver este problema.

2. Considerad esta matriz:

$$\mathbf{B} = (b_{ij}) = \begin{bmatrix} 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \\ 5 & 6 & 7 & 8 \end{bmatrix}$$

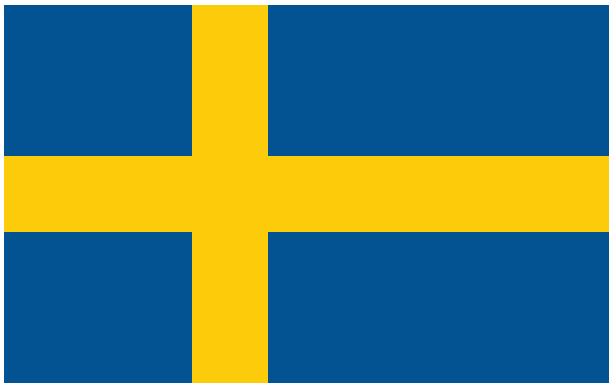
donde $b_{ij} = i + j$. Como en el caso anterior, esta matriz corresponde a una imagen de 4×4 píxeles. Calculad el rango r de la matriz $\mathbf{A}$. Reescribid $\mathbf{A}$ como suma de r matrices de rango 1 de la forma $\mathbf{u}\mathbf{v}^\intercal$. No es necesario que $\mathbf{u}_1^\intercal \mathbf{u}_2 = \mathbf{v}_1^\intercal \mathbf{v}_2 = 0$. No uséis ningún *software* para resolver este problema.

3. Considerad esta matriz:

$$\mathbf{S} = (s_{ij}) = \begin{bmatrix} 1 & 2 & 1 & 1 \\ 2 & 2 & 2 & 2 \\ 1 & 2 & 1 & 1 \end{bmatrix}$$

donde 1 es el color azul y 2 es el color amarillo (oro) de la bandera de Suecia (figura 1). Comprovad que la matriz tenga rango 2. Descomponed $\mathbf{S}$ de forma que $\mathbf{S} = \mathbf{u}_1\mathbf{v}_1^\intercal + \mathbf{u}_2\mathbf{v}_2^\intercal$. La descomposición no es única. Comprovad si, en vuestro caso, $\mathbf{u}_1^\intercal\mathbf{u}_2 = \mathbf{v}_1^\intercal\mathbf{v}_2 = 0$. No uséis ningún *software* para resolver este problema.

Figura 1. Bandera de Suecia



Fuente: https://www.countryflags.com

4. Considerad la matriz

$$\mathbf{B} = (s_{ij}) = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 3 & 3 \end{bmatrix}$$

que representa la bandera de Benín (Àfrica), como se muestra en la figura 2, donde 1 es el color verde, 2 es el color amarillo y 3 es el color rojo. Comprovad que la matriz tenga rango 2. Descomponed $\mathbf{S}$ de forma que $\mathbf{S} = \mathbf{u}_1\mathbf{v}_1^\intercal + \mathbf{u}_2\mathbf{v}_2^\intercal$. La descomposición no es única. Comprovad si, en vuestro caso, $\mathbf{u}_1^\intercal\mathbf{u}_2 = \mathbf{v}_1^\intercal\mathbf{v}_2 = 0$. No uséis ningún *software* para resolver este problema.

Figura 2. Bandera de Benín

Fuente: https://www.countryflags.com

3. Soluciones de los problemas: análisis de componentes principales (PCA)

1. Ahora vamos a responder a las cuestiones planteadas utilizando el código en R propuesto.

a) Si usamos las instrucciones

```
> data("mtcars")
> mtcars
```

obtendremos:

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

Podemos ver los 32 modelos de coche y los valores de cada una de las once variables.

b) Con la instrucción

```
> mtcars.pca <- prcomp(mtcars[,c(1:7,10,11)],center = TRUE, scale =TRUE)
```

generamos la estructura que contiene el análisis de componentes principales. Notad que hemos descartado las variables 8 y 9, que corresponden a **vs** y **am**. Las opciones `center = TRUE` y `scale = TRUE` tienen la función de escalar nuestros datos.

c) Con la instrucción

```
> summary(mtcars.pca)
```

obtenemos:

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7    PC8     PC9
Standard deviation     2.3782 1.4429 0.71008 0.51481 0.42797 0.35184 0.32413 0.2419 0.14896
Proportion of variance 0.6284 0.2313 0.05602 0.02945 0.02035 0.01375 0.01167 0.0065 0.00247
Cumulative proportion  0.6284 0.8598 0.91581 0.94525 0.96560 0.97936 0.99103 0.9975 1.00000
```

La primera componente principal retiene un 62.84 % de información (*proportion of variance*). Las dos primeras componentes principales retienen un 85.98 % de información (*cumulative proportion*).

d) Con la instrucción

```
> str(mtcars.pca)
```

obtenemos:

```
List of 5
 $ sdev    : num [1:9] 2.378 1.443 0.71 0.515 0.428 ...
 $ rotation: num [1:9, 1:9] -0.393 0.403 0.397 0.367 -0.312 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:9] "mpg" "çyl" "disp" "hp" ...
  .. ..$ : chr [1:9] "PC1" "PC2" "PC3" "PC4" ...
 $ center  : Named num [1:9] 20.09 6.19 230.72 146.69 3.6 ...
  ..- attr(*, "names")= chr [1:9] "mpg" "çyl" "disp" "hp" ...
 $ scale   : Named num [1:9] 6.027 1.786 123.939 68.563 0.535 ...
  ..- attr(*, "names")= chr [1:9] "mpg" "çyl" "disp" "hp" ...
 $ x       : num [1:32, 1:9] -0.664 -0.637 -2.3 -0.215 1.587 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:32] "Mazda RX4" "Mazda RX4 Wag" "Datsun 710" "Hornet 4 Drive" ...
  .. ..$ : chr [1:9] "PC1" "PC2" "PC3" "PC4" ...
```

Las medias aritméticas de las variables están en `mtcars.pca$center`. Si **cyl** es la segunda variable, podemos observar que su media aritmética es 6.19. Este valor también se puede obtener escribiendo:

```
> mtcars.pca$center[2]
```

puesto que obtenemos:

```
   cyl
6.1875
```

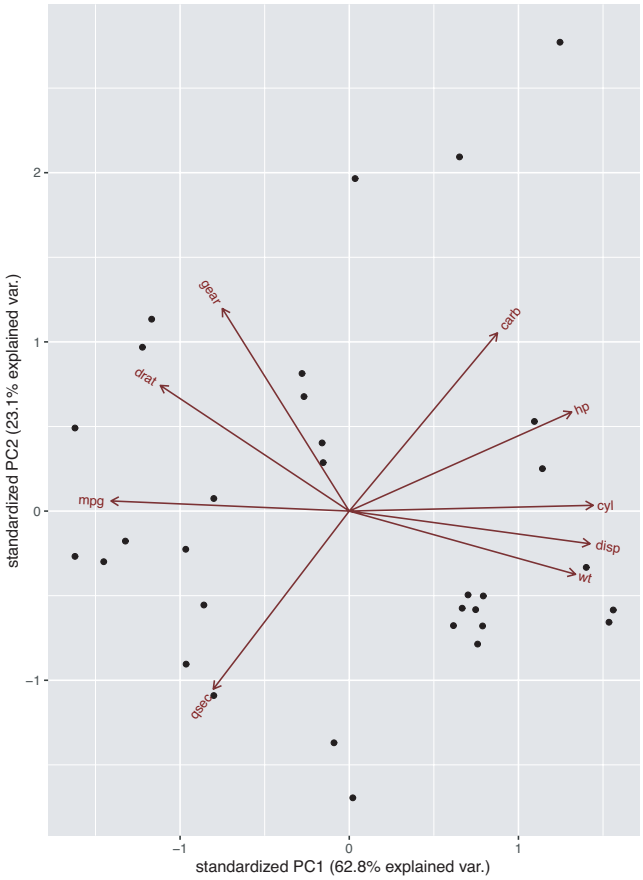
e) Con las instrucciones

```
> install.packages("devtools")
> library(devtools)
> install_github("vqv/ggbiplot")
> ggbiplot(mtcars.pca)
```

obtenemos la gráfica que se puede ver en la figura 3. Podemos observar que las variables que más contribuyen a la primera componente principal son **hp**, **cyl** y **disp** (en sentido positivo) y **mpg** (en sentido negativo). Las variables que más contribuyen a la segunda componente principal son **gear** y **carb** (en sentido positivo) y **qsec** (en sentido negativo). Las variables que tienen una influencia más pequeña sobre la segunda componente principal son **cyl** y **mpg**, ya que se trata de dos vectores prácticamente paralelos al eje de la primera componente principal (eje de abscisas).

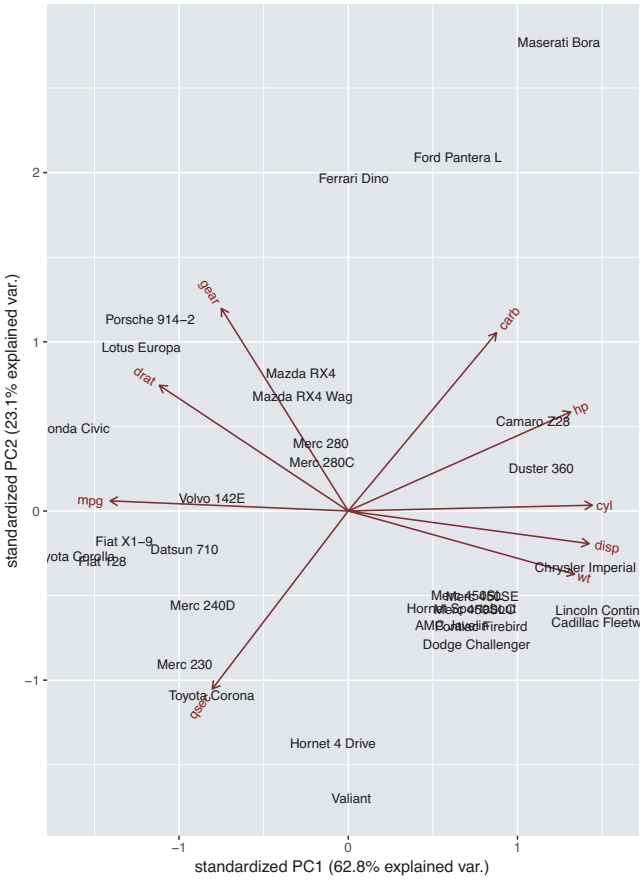
En la figura 4 podemos ver la misma gráfica, pero con los nombres de los modelos de coche sobre el plano generado por las dos primeras componentes principales.

Figura 3. Proyección sobre el plano generado por las dos componentes principales de los 32 modelos de coche. Podemos ver la contribución de cada variable original en cada una de las dos componentes principales.



Fuente: elaboración propia

Figura 4. Proyección sobre el plano generado por las dos componentes principales de los 32 modelos de coche. Podemos ver la contribución de cada variable original en cada una de las dos componentes principales.



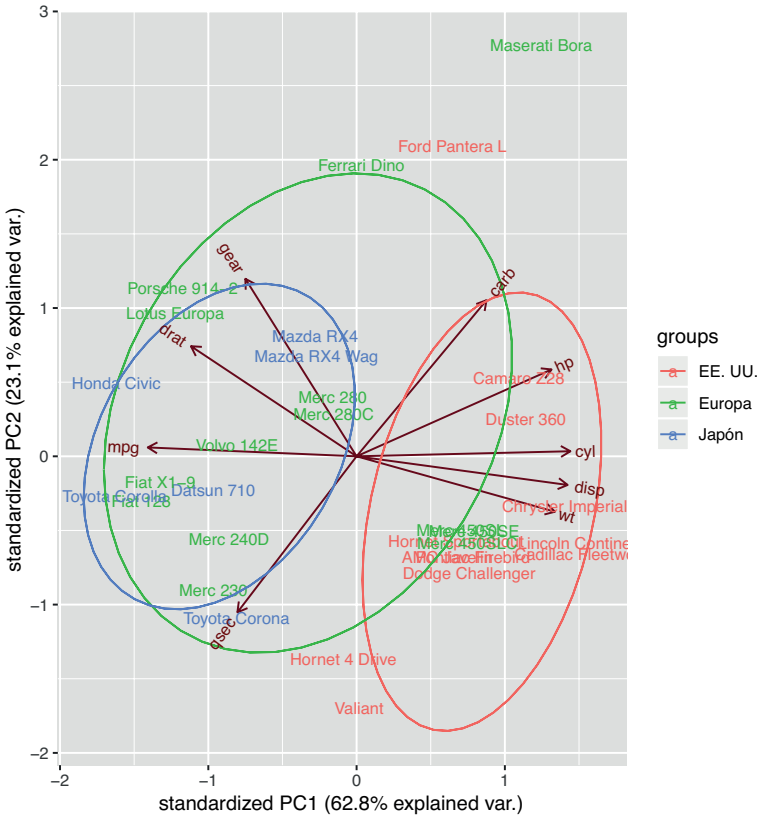
Fuente: elaboración propia

f) Con las instrucciones

```
> mtcars.country <- c(rep("Japón", 3), rep("EE. UU.",4),
rep("Europa", 7),rep("EE. UU.",3),"Europa", rep("Japón", 3),
rep("EE. UU.",4), rep("Europa", 3), "EE. UU.", rep("Europa", 3))
> ggbiplot(mtcars.pca, ellipse=TRUE,
  labels=rownames(mtcars), groups=mtcars.country)
```

obtenemos la gráfica que se puede ver en la figura 5, donde los modelos de coche están agrupados por origen geográfico (Japón, Estados Unidos o Europa). Podemos observar una cosa interesante: los coches americanos forman un clúster diferente a la derecha. Al observar los ejes, vemos que los coches americanos se caracterizan por valores altos para **cyl** (número de cilindros), **disp** (cubicaje) y **wt** (peso). Los coches japoneses, por otro lado, destacan por un **mpg** elevado (es decir, un consumo bajo). Los automóviles europeos se sitúan por la zona del medio y están menos agrupados que cualquiera de los otros dos grupos.

Figura 5. Proyección sobre el plano generado por las dos componentes principales de los 32 modelos de coche, agrupados por la región de origen (Japón, Estados Unidos o Europa).



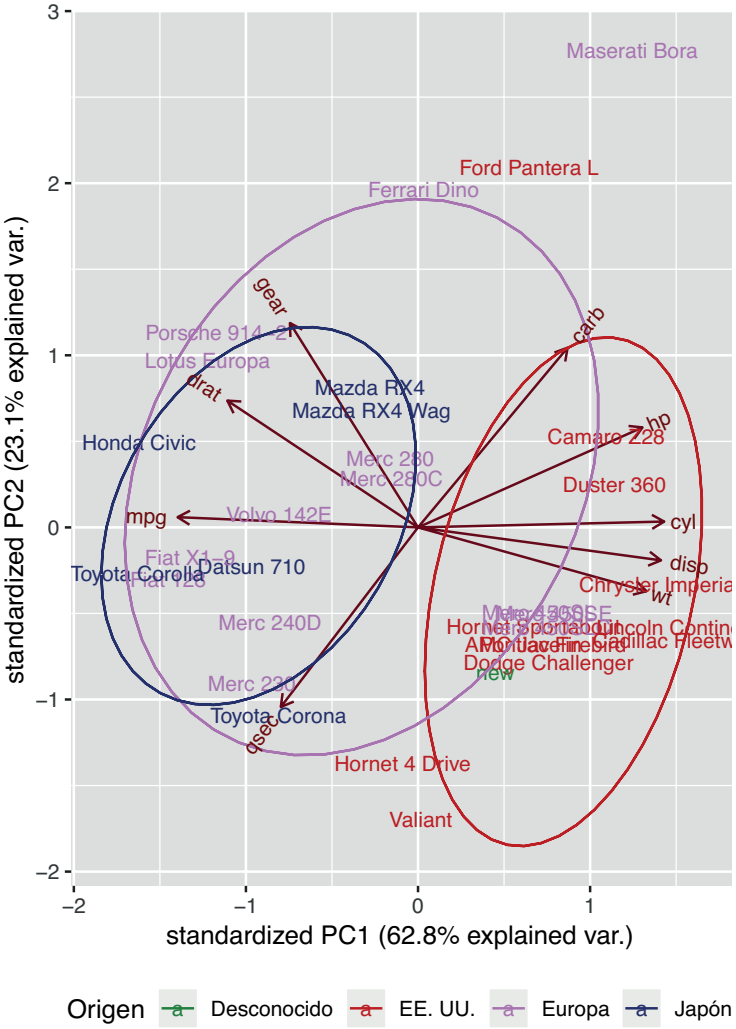
Fuente: elaboración propia

g) Con las instrucciones

```
> s<-c(18,7,300,158,3,3.4,18,0,0,3,2)
> s.sc<-(s[c(1:7,10,11)] -mtcars.pca$center)/mtcars.pca$scale
> u<-s.sc%*%mtcars.pca$rotation
> mtcars.plus.pca<-mtcars.pca
> mtcars.plus.pca$x<-rbind(mtcars.plus.pca$x,u)
> mtcars.countryplus<-c(mtcars.country,"Desconocido")
> ggbiplot(mtcars.plus.pca, ellipse = TRUE, circle = FALSE, var.axes=TRUE,
  labels=c(rownames(mtcars), "new"), groups=mtcars.countryplus)+
  scale_colour_manual(name="Origen",
  values= c("forest green", "red3", "violet", "dark blue"))+
  theme(legend.position = "bottom")
```

obtenemos la gráfica que se puede ver en la figura 6. Notad que el nuevo modelo de coche, llamado *new*, cae en la zona de influencia de los modelos de coche de los Estados Unidos. Por lo tanto, si tuviéramos que inferir el origen del modelo de este vehículo, afirmaríamos que es de los Estados Unidos.

Figura 6. Proyección sobre el plano generado por las dos componentes principales de los 32 modelos de coche, agrupados por la región de origen (Japón, Estados Unidos o Europa). Podemos ver la ubicación del modelo de coche *new*.



Fuente: elaboración propia

4. Soluciones de los problemas: descomposición en valores singulares (SVD)

1. El rango de la matriz es 1, ya que la segunda fila es igual que la primera multiplicada por 2; la tercera fila es igual que la primera multiplicada por 3, y la cuarta fila, igual que la primera multiplicada por 4. Por lo tanto, si la segunda, la tercera y la cuarta filas son una combinación lineal de la primera, el rango es igual a 1. Por lo tanto, solo hay que transformar la matriz $\mathbf{A}$ como producto de dos vectores $\mathbf{uv}^\intercal$. Como hemos visto, esta matriz tiene las cuatro filas proporcionales. Por lo tanto, podemos pensar en:

$$\mathbf{v}^\intercal = \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix}$$

Del mismo modo, el vector $\mathbf{u}$ será determinado por:

$$\mathbf{u} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$$

Se puede comprobar fácilmente que:

$$\mathbf{uv}^\intercal = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 \cdot \mathbf{v}^\intercal \\ 2 \cdot \mathbf{v}^\intercal \\ 3 \cdot \mathbf{v}^\intercal \\ 4 \cdot \mathbf{v}^\intercal \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 6 & 8 \\ 3 & 6 & 9 & 12 \\ 4 & 8 & 12 & 16 \end{bmatrix} = \mathbf{A}$$

2. El rango de la matriz

$$\mathbf{B} = (b_{ij}) = \begin{bmatrix} 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \\ 5 & 6 & 7 & 8 \end{bmatrix}$$

es 2, ya que el determinante de la matriz es cero. También es cero el determinante de todos los menores de orden 3. En cambio, el menor de orden 2

$$\begin{bmatrix} 2 & 3 \\ 3 & 4 \end{bmatrix}$$

de la matriz $\mathbf{B}$ tiene determinante $-1 \neq 0$.

Por definición de esta matriz, cada elemento $b_{ij} = i + j$ es la suma de la posición de su fila y de su columna. Esto nos permite pensar que podemos separar la matriz $\mathbf{B}$ en dos matrices:

$$\mathbf{B} = \mathbf{C} + \mathbf{D}$$

donde $c_{ij} = i$ y $d_{ij} = j$. Es decir, los elementos de la matriz $\mathbf{C}$ son iguales que la posición de su fila, mientras que los elementos de la matriz $\mathbf{D}$ son iguales que la posición de su columna. En efecto:

$$\mathbf{B} = \begin{bmatrix} 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \\ 5 & 6 & 7 & 8 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \\ 4 & 4 & 4 & 4 \end{bmatrix} + \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{bmatrix} = \mathbf{C} + \mathbf{D}$$

Notad que las filas de $\mathbf{C}$ son proporcionales, mientras que las filas de $\mathbf{D}$ son iguales. Es evidente que, ahora, las dos matrices tienen rango 1. Por lo tanto, si queremos descomponer $\mathbf{C}$, podemos hacer lo siguiente:

$$\mathbf{C} = \mathbf{u}_1\mathbf{v}_1^\intercal = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}$$

De forma similar, para descomponer $\mathbf{D}$ podemos hacer:

$$\mathbf{D} = \mathbf{u}_2\mathbf{v}_2^\intercal = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix}$$

En este caso, $\mathbf{u}_1^\intercal\mathbf{u}_2 = \mathbf{v}_1^\intercal\mathbf{v}_2 = 10 \neq 0$, lo que implica que los vectores que hemos considerado no son ortogonales.

3. Observad que, en el caso de la matriz

$$\mathbf{S} = \begin{bmatrix} 1 & 2 & 1 & 1 \\ 2 & 2 & 2 & 2 \\ 1 & 2 & 1 & 1 \end{bmatrix},$$

la primera y la tercera filas son iguales. Esto implica que el rango no es 3, que sería el máximo rango posible de esta matriz. Si consideramos el menor de orden 2

$$\begin{bmatrix} 1 & 2 \\ 2 & 2 \end{bmatrix}$$

de la matriz $\mathbf{S}$, tiene determinante $-2 \neq 0$. Por lo tanto, el rango es $r = 2$. Es fácil descomponer la matriz $\mathbf{S}$ como suma de dos matrices, cada una con rango 1, si pensamos en las dos filas que son iguales (la primera y la tercera) y la segunda fila, que es diferente. En efecto:

$$\mathbf{S} = \begin{bmatrix} 1 & 2 & 1 & 1 \\ 2 & 2 & 2 & 2 \\ 1 & 2 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 2 & 2 & 2 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \mathbf{S}_1 + \mathbf{S}_2$$

Esto nos permite expresar $\mathbf{S}_1$ como $\mathbf{u}_1\mathbf{v}_1^\intercal$, donde

$$\mathbf{S}_1 = \begin{bmatrix} 1 & 2 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 1 \end{bmatrix} = \mathbf{u}_1\mathbf{v}_1^\intercal = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 & 1 \end{bmatrix}$$

Notad que, en este caso, $\mathbf{u}_1 \in \mathbb{R}^3$ y $\mathbf{v}_1 \in \mathbb{R}^4$.

Del mismo modo, esto nos permite expresar $\mathbf{S}_2$ como $\mathbf{u}_2\mathbf{v}_2^\intercal$, donde

$$\mathbf{S}_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 2 & 2 & 2 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \mathbf{u}_2\mathbf{v}_2^\intercal = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 2 & 2 & 2 & 2 \end{bmatrix}$$

Observad que, en este caso, también: $\mathbf{u}_2 \in \mathbb{R}^3$ y $\mathbf{v}_2 \in \mathbb{R}^4$.

La elección de los vectores $\mathbf{u}_1$ y $\mathbf{u}_2$ hace, en esta ocasión, que sean ortogonales, ya que

$$\mathbf{u}_1^\intercal \mathbf{u}_2 = 0.$$

En cambio, no son ortogonales los vectores $\mathbf{v}_1$ y $\mathbf{v}_2$, puesto que

$$\mathbf{v}_1^\intercal \mathbf{v}_2 = 10 \neq 0.$$

4. Observad que, en el caso de la matriz

$$\mathbf{B} = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 3 & 3 \end{bmatrix},$$

las dos filas son diferentes. Esto implica que el rango es 2, el máximo rango posible de esta matriz (recordad que el rango máximo es el mínimo entre el número de filas y el número de columnas). Si consideramos el menor de orden 2

$$\begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix}$$

de la matriz $\mathbf{B}$, tiene determinante $1 \neq 0$. Por lo tanto, como hemos dicho, el rango es $r = 2$. Es fácil descomponer la matriz $\mathbf{B}$ como suma de dos matrices, cada una con rango 1. En efecto:

$$\mathbf{B} = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 3 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 2 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} = \mathbf{B}_1 + \mathbf{B}_2$$

Entonces, esto nos permite expresar $\mathbf{B}_1$ como $\mathbf{u}_1\mathbf{v}_1^\intercal$, donde

$$\mathbf{B}_1 = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 2 \end{bmatrix} = \mathbf{u}_1\mathbf{v}_1^\intercal = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 2 \end{bmatrix}$$

Notad que, en este caso, $\mathbf{u}_1 \in \mathbb{R}^2$ y $\mathbf{v}_1 \in \mathbb{R}^3$.

Del mismo modo, esto nos permite expresar $\mathbf{B}_2$ como $\mathbf{u}_2\mathbf{v}_2^\intercal$, donde

$$\mathbf{B}_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} = \mathbf{u}_2\mathbf{v}_2^\intercal = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}$$

En este caso, también, $\mathbf{u}_2 \in \mathbb{R}^2$ y $\mathbf{v}_2 \in \mathbb{R}^3$.

La elección de los vectores $\mathbf{u}_1$ y $\mathbf{u}_2$ hace, en esta ocasión, que no sean ortogonales, ya que

$$\mathbf{u}_1^\intercal \mathbf{u}_2 = 1 \neq 0.$$

Tampoco son ortogonales los vectores $\mathbf{v}_1$ y $\mathbf{v}_2$, puesto que

$$\mathbf{v}_1^\intercal \mathbf{v}_2 = 4 \neq 0.$$