Descomposició en valors singulars: introducció i aplicacions

Problemes per a la ciència de dades

PID_00262388

Francesc Pozo Montero
Jordi Ripoll Missé

Universitat Oberta de Catalunya

Francesc Pozo Montero

Llicenciat en Matemàtiques per la Universitat de Barcelona (2000) i doctor en Matemàtica Aplicada per la Universitat Politècnica de Catalunya (2005). Ha estat professor associat a la Universitat Autònoma de Barcelona i professor associat, col·laborador i actualment professor agregat a la Universitat Politècnica de Catalunya. A més, és cofundador del Grup d'Innovació Matemàtica E-learning (GIMEL), responsable de diversos projectes d'innovació docent i autor de diverses publicacions. Com a membre del grup de recerca consolidat CoDAlab, centra la recerca en la teoria de control i les aplicacions en enginyeria mecànica i civil, com també en l'ús de la ciència de dades per al monitoratge de la integritat estructural i per al monitoratge de la condició, sobretot en turbines eòliques.

Jordi Ripoll Missé

Llicenciat en Matemàtiques i doctor en Ciències Matemàtiques per la Universitat de Barcelona (2005). Professor col·laborador de la Universitat Oberta de Catalunya des del 2011 i professor del Departament d'Informàtica, Matemàtica Aplicada i Estadística de la Universitat de Girona (UdG) des del 1996, on actualment és professor agregat i desenvolupa tasques de recerca en l'àmbit de la biologia matemàtica (models amb equacions en derivades parcials i dinàmica evolutiva). També ha estat professor i tutor de la UNED en dues etapes, primer al centre associat de Terrassa i actualment al de Girona. Ha participat en nombrosos projectes d'innovació docent, especialment pel que fa a l'aprenentatge de les matemàtiques en línia.

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats per la professora: Cristina Cano Bastidas (2019)

Primera edició: febrer 2019

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

Índex

1. Problemes: anàlisi de components principals (PCA)

En aquest apartat presentem un problema amb dades reals que es pot considerar i treballar mitjançant l'anàlisi de components principals. En la resolució proposem fer servir una llibreria especial de `R`, de manera que es pugui complementar el codi que també podeu trobar en el material "Descomposició en valors singulars: introducció i aplicacions. Estudi de cas i guia de resolució en R".

1. Considereu les dades proporcionades pel conjunt `mtcars`, lliurement disponible a `R`, extretes de la revista estatunidenca *Motor Trend* (1974). Aquest conjunt de dades presenta, per a 32 models de cotxes dels anys 1973 i 1974, el consum de combustible i deu aspectes més. Les variables mesurades són:

1) **mpg** (consum de combustible): en milles per galó dels Estats Units; els cotxes més potents i més pesats tendeixen a consumir més combustible.

2) **cyl** (nombre de cilindres): els cotxes més potents solen tenir més cilindres.

3) **disp** (desplaçament): en polzades cúbiques (en anglès, *cubic inch*); el volum combinat dels cilindres del motor.

4) **hp** (potència bruta): és una mesura de la potència generada pel cotxe.

5) **drat** (relació de l'eix posterior): es descriu com un gir de l'eix de transmissió corresponent a un gir de les rodes. Els valors més alts disminuiran l'eficiència del combustible.

6) **wt** (pes): corresponent a 1.000 lliures.

7) **qsec** (temps d'1/4 de milla): velocitat i acceleració dels cotxes.

8) **vs** (bloc del motor): indica si el motor del vehicle té forma de V o una forma recta més comuna.

9) **am** (transmissió): indica si la transmissió de l'automòbil és automàtica (0) o manual (1).

10) **gear** (nombre de marxes endavant): els cotxes esportius solen tenir més marxes.

11) **carb** (nombre de carburadors): associats a motors més potents.

Amb aquest conjunt de dades procedirem de la manera següent:

a) Carregueu el conjunt de dades `mtcars` amb aquesta instrucció:

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

Observeu els models de cotxe i els valors de les seves variables.

b) Calculeu amb R l'anàlisi de components principals i anomeneu l'estructura resultant `mtcars.pca`. Descarteu per a l'anàlisi les variables binàries **vs** i **am**. No us oblideu d'escalar les vostres dades.

c) Executeu la instrucció:

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

Quina és la quantitat d'informació (proporció de variància) retinguda per la primera component principal? Quina és la quantitat d'informació retinguda per les dues primeres components principals?

d) Executeu la instrucció:

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

Així tindreu accés a l'estructura `mtcars.pca`. Quina és la mitjana aritmètica de la variable **cyl**?

e) Abans de continuar amb la representació gràfica de les dues primeres components principals, carregueu `devtools` amb aquesta instrucció:

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

A continuació, executeu la instrucció:

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

El gràfic resultant us mostrarà 32 punts en el pla generat per les dues primeres components principals, així com les nou variables originals que contribueixen a les components principals. Quines són les variables originals que més contribueixen a la primera component principal? Quines són les variables originals que més contribueixen a la segona component principal? I quines són les variables originals que menys contribueixen a la segona component principal?

Amb la instrucció

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

obtindreu el mateix gràfic, però veureu el nom de cada un dels models de cotxe.

f) A continuació, agruparem els models de cotxe segons l'origen geogràfic del fabricant. Per exemple, el model Mazda RX4 és japonès, el Hornet 4 Drive és estatunidenc i el model Merc 240D (Mercedes 240D) és europeu. El vector complet amb els orígens dels models de cotxe és:

```
> mtcars.country <- c(rep("Japó", 3), rep("EUA",4),
rep("Europa", 7),rep("EUA",3),"Europa", rep("Japó", 3),
rep("EUA",4), rep("Europa", 3), "EUA", rep("Europa", 3))
```

De manera que, amb la instrucció:

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

obtenim el mateix gràfic, però amb els models de cotxe agrupats pel seu origen. A més, hem fet incloure una el·lipse que engloba els models de cotxe dins de cada grup geogràfic. Què destacaríeu —pel que fa a les variables originals— dels models de cotxe estatunidencs? I dels models de cotxe japonesos? I dels europeus?

g) Tenim un nou model de cotxe, del qual no coneixem l'origen, amb les característiques següents:

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

Volem projectar aquest nou model de cotxe per intentar esbrinar-ne l'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,"Desconegut")
> 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")
```

A la primera línia del codi definim les característiques del nou model; a la segona, escalem les nostres dades respecte de les dades originals, i a la tercera, projectem el model de cotxe sobre l'espai vectorial generat per les components principals. A les dues línies següents, s'amplia el conjunt de dades projectades `mtcars.pca$x` amb el nou punt. A la sisena línia de codi especifiquem la categoria del cotxe nou com a "Desconegut". Finalment, representem gràficament tots els models de cotxe, incloent-hi el nou. Segons la posició del nou model, quin podria ser el seu origen?

2. Problemes: descomposició en valors singulars (SVD)

La descomposició en valors singulars és una tècnica molt útil en el camp de la ciència de dades, ja que permet descompondre o factoritzar una matriu $\mathbf{A}$ com a suma d'altres matrius de rang 1. La matriu $\mathbf{A}$ presenta aquesta forma:

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

en què r és el rang de la matriu $\mathbf{A}$, σ_i són els valors singulars (valors positius ordenats en ordre decreixent) i $\mathbf{u}_i$ i $\mathbf{v}_i$ són els vectors singulars per l'esquerra i per la dreta, respectivament.

Una de les aplicacions més importants de la descomposició en valors singulars és la compressió d'imatges. En aquest sentit, per exemple, una imatge de $m \times n$ píxels —en escala de grisos— es representa per una matriu, en què cada element representa la intensitat del color gris de cada píxel. Per emmagatzemar aquesta matriu necessitem $m \cdot n$ nombres. Suposem que el rang d'aquesta matriu és r. Aleshores, si considerem la descomposició en valors singulars:

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

ara és necessari emmagatzemar

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

nombres. En funció del rang r de la matriu, això ja pot suposar un cert nivell de compressió o de reducció de la dimensionalitat. Però podem anar encara més enllà, ja que podem considerar l'aproximació donada per la ν-èsima suma parcial, en què $\nu \leq r$:

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

En aquest cas, ja només cal emmagatzemar

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

nombres. Quina és la magnitud de l'error que es comet quan fem aquesta aproximació? Segons el mòdul "Descomposició en valors singulars: introducció i aplicacions", la diferència en norma 2 entre aquestes dues matrius és igual a:

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

és a dir, serà petita en funció de la magnitud del valor singular $\sigma_{\nu+1}$. Fixeu-vos que, per exemple, si $m = n = 1000$, $r = 300$ i $\nu = 50$, podem reduir la quantitat d'informació que hem d'emmagatzemar en un 89.995% amb un error en norma 2 que seria igual a la magnitud del valor singular $\sigma_{\nu+1}$ —previsiblement, molt petit.

L'objectiu dels problemes que hi ha a continuació és —a més de calcular de manera precisa, automatitzada o mitjançant programari la descomposició en valors singulars— aprofundir en la comprensió del mètode per mitjà de la descomposició de les matrius que es presenten com a suma de matrius de rang 1. Podreu observar que:

1) La solució no és única. És a dir, podeu obtenir solucions diferents de les que es proposen en l'apartat 3.

2) En els problemes plantejats, no es calculen els valors singulars. En particular, això és conseqüència del fet que no busquem que els vectors $\mathbf{u}_i$ i $\mathbf{v}_i$ siguin unitaris.

3) Tant els vectors $\mathbf{u}_i$ com els vectors $\mathbf{v}_i$ són la base dels subespais vectorials generats per les columnes i per les files, respectivament. Ara bé, aquestes bases no són necessàriament ortonormals, és a dir, els vectors poden no ser ortogonals ni tenir norma 1.

1. Considereu aquesta matriu:

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

en què $a_{ij} = i \cdot j$. Aquesta matriu correspon a una imatge de 4×4 píxels. Calculeu el rang r de la matriu $\mathbf{A}$. Reescriviu $\mathbf{A}$ com la suma de r matrius de rang 1 de la forma $\mathbf{u}\mathbf{v}^\intercal$. No feu servir cap programari per resoldre aquest problema.

2. Considereu aquesta matriu:

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

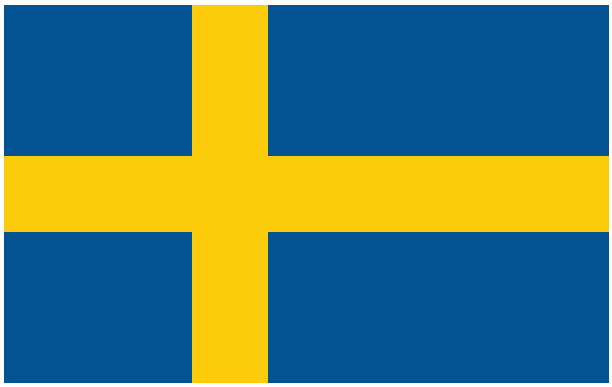
en què $b_{ij} = i + j$. Com en el cas anterior, aquesta matriu correspon a una imatge de 4×4 píxels. Calculeu el rang r de la matriu $\mathbf{A}$. Reescriviu $\mathbf{A}$ com la suma de r matrius de rang 1 de la forma $\mathbf{u}\mathbf{v}^\intercal$. No és necessari que $\mathbf{u}_1^\intercal \mathbf{u}_2 = \mathbf{v}_1^\intercal \mathbf{v}_2 = 0$. No feu servir cap programari per resoldre aquest problema.

3. Considereu aquesta matriu:

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

en què 1 és el color blau i 2 és el color groc (or) de la bandera de Suècia (vegeu la figura 1). Comproveu que la matriu tingui rang 2. Descomponeu $\mathbf{S}$ de manera que $\mathbf{S} = \mathbf{u}_1\mathbf{v}_1^\intercal + \mathbf{u}_2\mathbf{v}_2^\intercal$. La descomposició no és única. Comproveu si, en el vostre cas, $\mathbf{u}_1^\intercal\mathbf{u}_2 = \mathbf{v}_1^\intercal\mathbf{v}_2 = 0$. No feu servir cap programari per resoldre aquest problema.

Figura 1. Bandera de Suècia



Font: https://www.countryflags.com

4. Considereu la matriu

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

que representa la bandera de Benín (Àfrica), tal com mostra la figura 2, en què 1 és el color verd, 2 és el color groc i 3 és el color vermell. Comproveu que la matriu tingui rang 2. Descomponeu $\mathbf{S}$ de manera que $\mathbf{S} = \mathbf{u}_1\mathbf{v}_1^\intercal + \mathbf{u}_2\mathbf{v}_2^\intercal$. La descomposició no és única. Comproveu si, en el vostre cas, $\mathbf{u}_1^\intercal\mathbf{u}_2 = \mathbf{v}_1^\intercal\mathbf{v}_2 = 0$. No feu servir cap programari per resoldre aquest problema.

Figura 2. Bandera de Benín

Font: https://www.countryflags.com

3. Solucions dels problemes: anàlisi de components principals (PCA)

1. Ara respondrem les preguntes plantejades utilitzant el codi en R proposat.

a) Si fem servir les instruccions

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

obtindrem:

	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

Podem veure els 32 models de cotxe i els valors de cadascuna de les onze variables.

b) Amb la instrucció

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

generem l'estructura que conté l'anàlisi de components principals. Fixeu-vos que hem descartat les variables 8 i 9, que corresponen a **vs** i **am**. Les opcions `center = TRUE` i `scale = TRUE` tenen la funció d'escalar les nostres dades.

c) Amb la instrucció

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

obtenim:

```
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 component principal reté un 62.84% d'informació (*proportion of variance*). Les dues primeres components principals retenen un 85.98% d'informació (*cumulative proportion*).

d) Amb la instrucció

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

obtenim:

```
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" ...
```

Les mitjanes aritmètiques de les variables són a `mtcars.pca$center`.
Si **cyl** és la segona variable, podem observar que la seva mitjana aritmètica és 6.19. Aquest valor també el podem obtenir si escrivim:

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

ja que obtenim:

```
   cyl
6.1875
```

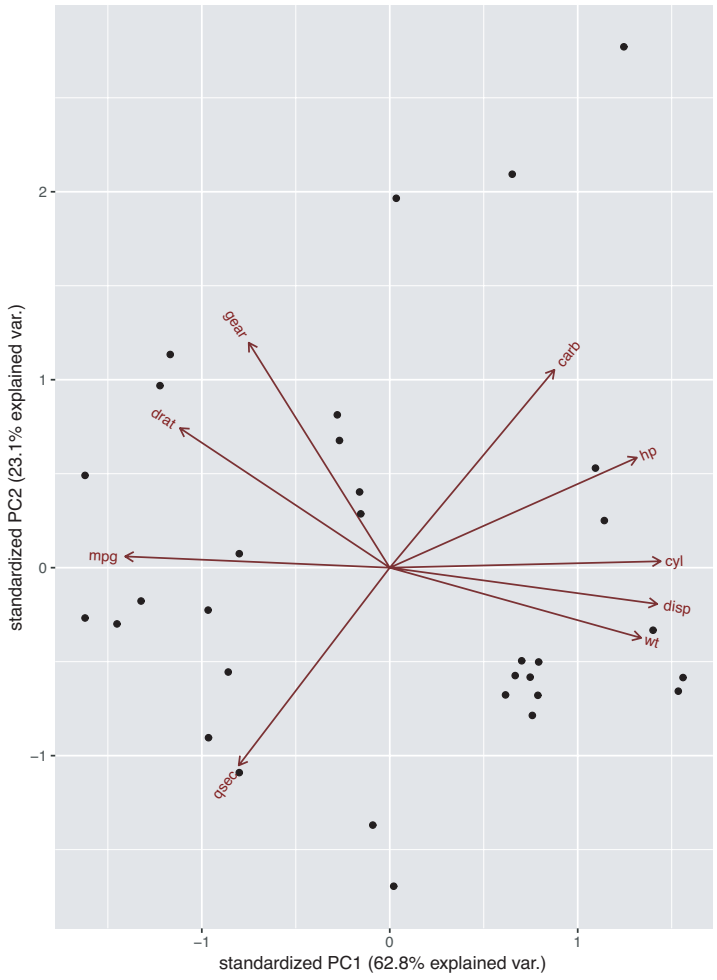
e) Amb les instruccions

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

obtenim la gràfica que es pot veure a la figura 3. Podem observar que les variables que més contribueixen a la primera component principal són **hp**, **cyl** i **disp** (en sentit positiu) i **mpg** (en sentit negatiu). Les variables que més contribueixen a la segona component principal són **gear** i **carb** (en sentit positiu) i **qsec** (en sentit negatiu). Les variables que tenen una influència més petita sobre la segona component principal són **cyl** i **mpg**, ja que es tracta de dos vectors que són pràcticament paral·lels a l'eix de la primera component principal (eix d'abscisses).

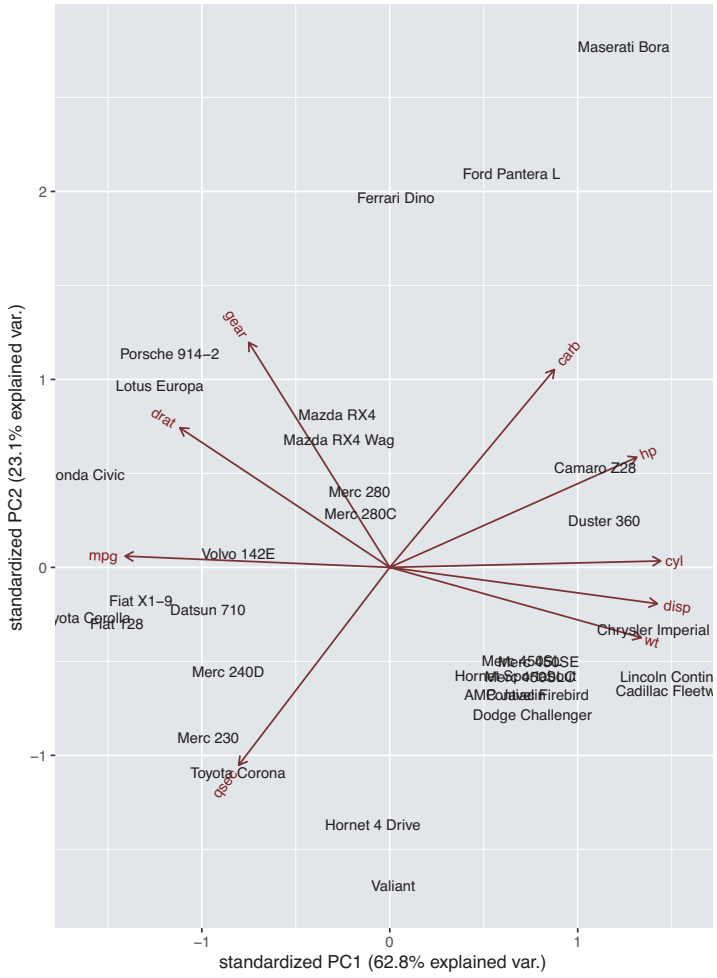
A la figura 4 podem veure el mateix gràfic, però amb els noms dels models de cotxe al pla generat per les dues primeres components principals.

Figura 3. Projecció sobre el pla generat per les dues components principals dels 32 models de cotxe. Hi podem veure la contribució de cada variable original a cadascuna de les dues components principals.



Font: elaboració pròpia

Figura 4. Projecció sobre el pla generat per les dues components principals dels 32 models de cotxe. Hi podem veure la contribució de cada variable original a cadascuna de les dues components principals.



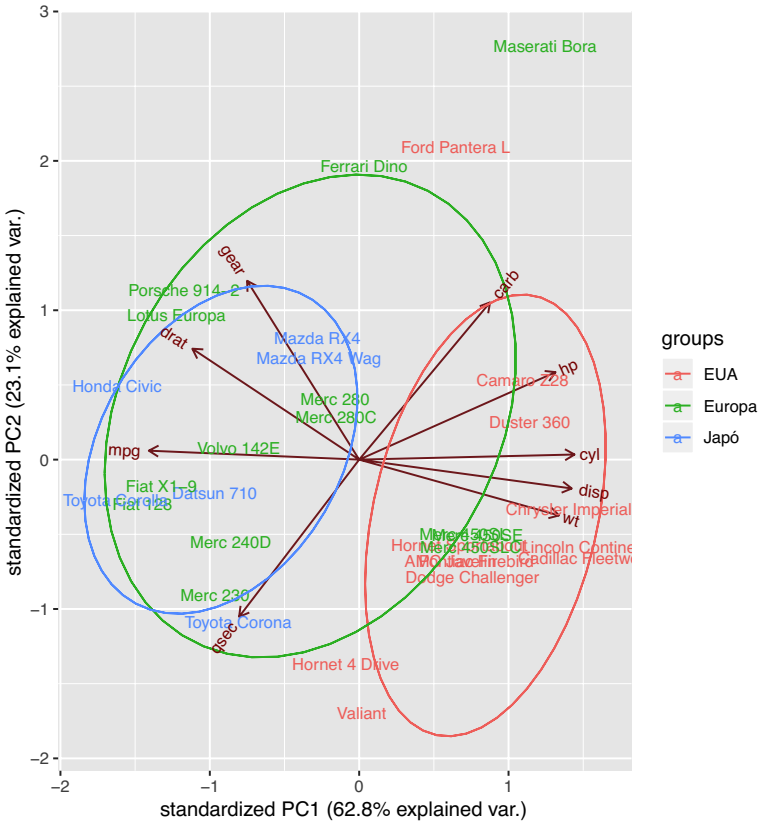
Font: elaboració pròpia

f) Amb les instruccions

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

obtenim la gràfica que es pot veure a la figura 5, en què els models de cotxe estan agrupats per origen geogràfic (Japó, Estats Units o Europa). Hi podem observar una cosa interessant: els cotxes americans formen un clúster diferent a la dreta. Mirant els eixos, es pot veure que els cotxes americans es caracteritzen per valors alts per a **cyl** (nombre de cilindres), **disp** (cubicatge) i **wt** (pes). Els cotxes japonesos, d'altra banda, destaquen per un **mpg** elevat (és a dir, un consum baix). Els automòbils europeus se situen una mica al mig de la zona i estan menys agrupats que qualsevol dels altres dos grups.

Figura 5. Projecció sobre el pla generat per les dues components principals dels 32 models de cotxe, agrupats per la regió d'origen (Japó, Estats Units o Europa).



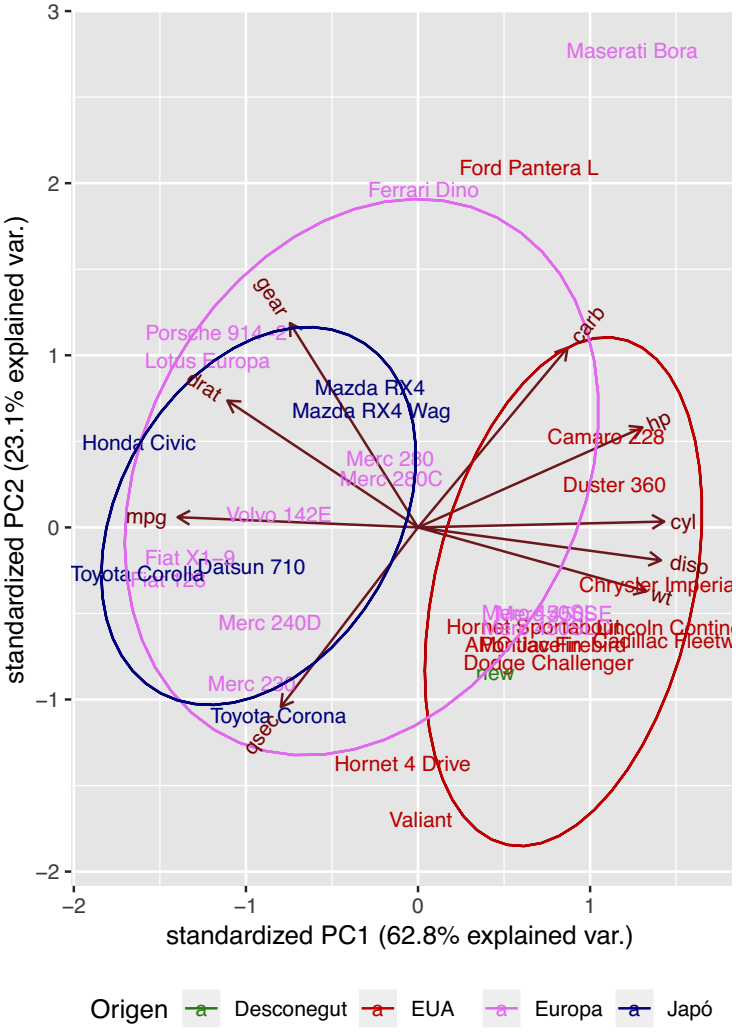
Font: elaboració pròpia

g) Amb les instruccions

```
> 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,"Desconegut")
> 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")
```

obtenim la gràfica que es pot veure a la figura 6. Fixeu-vos que el nou model de cotxe, anomenat *new*, cau a la zona d'influència dels models de cotxe dels Estats Units. Per tant, si haguéssim d'inferir l'origen del model d'aquest vehicle, podríem dir que és dels Estats Units.

Figura 6. Projecció sobre el pla generat per les dues components principals dels 32 models de cotxe, agrupats per la regió d'origen (Japó, Estats Units o Europa). Hi podem veure la ubicació del model de cotxe *new*.



Font: elaboració pròpia

4. Solucions dels problemes: descomposició en valors singulars (SVD)

1. El rang de la matriu és 1, ja que la segona fila és igual que la primera multiplicada per 2; la tercera fila és igual que la primera multiplicada per 3, i la quarta fila, igual que la primera multiplicada per 4. Per tant, si la segona, la tercera i la quarta files són una combinació lineal de la primera, el rang és igual que 1. Per tant, només cal transformar la matriu $\mathbf{A}$ com a producte de dos vectors $\mathbf{uv}^\intercal$. Com hem vist, aquesta matriu té les quatre files proporcionals. Per tant, podem pensar en:

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

De la mateixa manera, el vector $\mathbf{u}$ serà determinat per:

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

Es pot comprovar fàcilment 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 rang de la matriu

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

és 2, ja que el determinant de la matriu és zero. També és zero el determinant de tots els menors d'ordre 3. En canvi, el menor d'ordre 2

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

de la matriu $\mathbf{B}$ té determinant $-1 \neq 0$.

Per definició d'aquesta matriu, cada element $b_{ij} = i + j$ és la suma de la posició de la seva fila i de la seva columna. Això ens permet pensar que podem separar la matriu $\mathbf{B}$ en dues matrius:

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

en què $c_{ij} = i$ i $d_{ij} = j$. És a dir, els elements de la matriu $\mathbf{C}$ són iguals que la posició de la seva fila, mentre que els elements de la matriu $\mathbf{D}$ són iguals que la posició de la seva columna. En efecte:

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

Fixeu-vos que les files de $\mathbf{C}$ són proporcionals, mentre que les files de $\mathbf{D}$ són iguals. És evident que, ara, totes dues matrius tenen rang 1. Per tant, si volem descompondre $\mathbf{C}$, podem fer el següent:

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

De manera similar, si volem descompondre $\mathbf{D}$, podem fer:

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

En aquest cas, $\mathbf{u}_1^\intercal\mathbf{u}_2 = \mathbf{v}_1^\intercal\mathbf{v}_2 = 10 \neq 0$, la qual cosa implica que els vectors que hem considerat no són ortogonals.

3. Fixeu-vos que, en el cas de la matriu

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

la primera i la tercera files són iguals. Això implica que el rang no és 3, que seria el màxim rang possible d'aquesta matriu. Si considerem el menor d'ordre 2

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

de la matriu $\mathbf{S}$, té determinant $-2 \neq 0$. Per tant, el rang és $r = 2$. És fàcil descompondre la matriu $\mathbf{S}$ com a suma de dues matrius, cadascuna de les quals té rang 1, si pensem en les dues files que són iguals (la primera i la tercera) i la segona fila, que és diferent. En efecte:

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

Això ens permet expressar $\mathbf{S}_1$ com a $\mathbf{u}_1\mathbf{v}_1^\intercal$, en què

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

Observeu que, en aquest cas, $\mathbf{u}_1 \in \mathbb{R}^3$ i $\mathbf{v}_1 \in \mathbb{R}^4$.

De la mateixa manera, això ens permet expressar $\mathbf{S}_2$ com a $\mathbf{u}_2\mathbf{v}_2^\intercal$, en què

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

Fixeu-vos que, en aquest cas, també: $\mathbf{u}_2 \in \mathbb{R}^3$ i $\mathbf{v}_2 \in \mathbb{R}^4$.

L'elecció dels vectors $\mathbf{u}_1$ i $\mathbf{u}_2$ fa, en aquesta ocasió, que siguin ortogonals, ja que

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

En canvi, no són ortogonals els vectors $\mathbf{v}_1$ i $\mathbf{v}_2$, ja que

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

4. Observeu que, en el cas de la matriu

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

les dues files són diferents. Això implica que el rang és 2, que és el màxim rang possible d'aquesta matriu (recordeu que el rang màxim és el mínim entre el nombre de files i el nombre de columnes). Si considerem el menor d'ordre 2

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

de la matriu $\mathbf{B}$, té determinant $1 \neq 0$. Per tant, com hem dit, el rang és $r = 2$. És fàcil descompondre la matriu $\mathbf{B}$ com a suma de dues matrius, cadascuna de les quals té rang 1. En efecte:

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

Aleshores, això ens permet expressar $\mathbf{B}_1$ com a $\mathbf{u}_1\mathbf{v}_1^\intercal$, en què

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

Fixeu-vos que, en aquest cas, $\mathbf{u}_1 \in \mathbb{R}^2$ i $\mathbf{v}_1 \in \mathbb{R}^3$.

De la mateixa manera, això ens permet expressar $\mathbf{B}_2$ com a $\mathbf{u}_2\mathbf{v}_2^\intercal$, en què

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

En aquest cas, també, $\mathbf{u}_2 \in \mathbb{R}^2$ i $\mathbf{v}_2 \in \mathbb{R}^3$.

L'elecció dels vectors $\mathbf{u}_1$ i $\mathbf{u}_2$ fa, en aquesta ocasió, que no siguin ortogonals, ja que

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

Tampoc no són ortogonals els vectors $\mathbf{v}_1$ i $\mathbf{v}_2$, ja que

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