Laboratorio: Análisis de Componentes Principales con R

Beatriz Pateiro López

En esta sesión práctica veremos como llevar a cabo un Análisis de Componentes Principales con R. El análisis de componentes principales se concibe como una técnica de reducción de la dimensión, pues permite pasar de una gran cantidad de variables interrelacionadas a unas pocas componentes principales. El método consiste en buscar combinaciones lineales de las variables originales que representen lo mejor posible a la variabilidad presente en los datos. De este modo, con unas pocas combinaciones lineales, que serán las componentes principales, sería suficiente para entender la información contenida en los datos. Al mismo tiempo, la forma en que se construyen las componentes, y su relación con unas u otras variables originales, sirven para entender la estructura de correlación inherente a los datos. Por último, las componentes principales, que forman un vector aleatorio de dimensión menor, pueden ser empleadas en análisis estadísticos posteriores, como por ejemplo en regresión.

1 Análisis de componentes principales con `princomp`

Trabajaremos a lo largo de este laboratorio con el siguiente ejemplo. Se ha examinado a 25 alumnos, aspirantes a ingresar en el Máster Interuniversitario en Big Data, de 5 materias diferentes: Programación (cuyo resultado se almacena en la variable `prog`), Ingeniería de Computadores (`ingcom`), Ingeniería del Software (`ingsof`), Sistemas de la Información (`sist`) y Estadística (`estad`). Las puntuaciones obtenidas se encuentran en el fichero `aspirantes.txt` y se muestran a continuación.

```
> dat <- read.table("aspirantes.txt", header = TRUE)  # Lectura de los datos
> dat
```

```
##    prog ingcom ingsof sist estad
## 1    36     58     43   36    37
## 2    62     54     50   46    52
## 3    31     42     41   40    29
## 4    76     78     69   66    81
## 5    46     56     52   56    40
## 6    12     42     38   38    28
## 7    39     46     51   54    41
## 8    30     51     54   52    32
## 9    22     32     43   28    22
## 10    9     40     47   30    24
## 11   32     49     54   37    52
## 12   40     62     51   40    49
## 13   64     75     70   66    63
```

```
## 14   36     38     58   62    62
## 15   24     46     44   55    49
## 16   50     50     54   52    51
## 17   42     42     52   38    50
## 18    2     35     32   22    16
## 19   56     53     42   40    32
## 20   59     72     70   66    62
## 21   28     50     50   42    63
## 22   19     46     49   40    30
## 23   36     56     56   54    52
## 24   54     57     59   62    58
## 25   14     35     38   29    20
```

El objetivo de este estudio es obtener un ranking global de alumnos para la entrada en el máster, a través de una puntuación global, extraída como cierta combinación lineal de las calificaciones en las cinco materias examinadas.

Vamos a realizar entonces un Análisis de Componentes Principales con R. El comando básico de R que ejecuta el análisis de componentes principales es `princomp`. También se pueden obtener resultados similares con el comando `prcomp`. La diferencia principal entre uno y otro es que princomp diagonaliza la matriz S, mientras que `prcomp` diagonaliza la matriz S_c. Esto modifica los autovalores en la proporción $n/(n-1)$, pero no supone ningún cambio en los autovectores. Hemos optado por el comando `princomp`. A continuación se muestran las instrucciones que permiten realizar un análisis de componentes principales para el ejemplo, utilizando R. En primer lugar, como salida del objeto se muestran las desviaciones típicas de las componentes, que son las raíces cuadradas de los autovalores de S.

```
> test.pca <- princomp(dat)
> test.pca
```

```
## Call:
## princomp(x = dat)
##
## Standard deviations:
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
## 28.489680  9.035471  6.600955  6.133582  3.723358
##
##  5  variables and  25 observations.
```

Si hacemos un `summary` del objeto, obtenemos además la proporción de varianza explicada y sus valores acumulados.

```
> summary(test.pca)
```

```
## Importance of components:
##                            Comp.1     Comp.2     Comp.3     Comp.4
## Standard deviation     28.4896795 9.03547104 6.60095491 6.13358179
## Proportion of Variance  0.8212222 0.08260135 0.04408584 0.03806395
## Cumulative Proportion   0.8212222 0.90382353 0.94790936 0.98597332
##                            Comp.5
## Standard deviation     3.72335754
## Proportion of Variance 0.01402668
## Cumulative Proportion  1.00000000
```

Además de esta información, el objeto `test.pca` almacena otra información relevante.

```
> names(test.pca)
```

```
## [1] "sdev"     "loadings" "center"   "scale"    "n.obs"    "scores"
```

```
## [7] "call"
```

Como hemos visto, podemos obtener las desviaciones típicas de las componentes. De este modo podemos obtener las varianzas de las componentes, que son los autovalores.

```
> test.pca$sdev
```

```
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
## 28.489680  9.035471  6.600955  6.133582  3.723358
```

```
> test.pca$sdev^2
```

```
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
## 811.66184  81.63974  43.57261  37.62083  13.86339
```

Recordamos que si extraemos una componente principal, obtenemos la variable aleatoria unidimensional

$$z_1 = {v_1}' x.$$

Si en lugar de recoger todo el vector aleatorio, sólo aportamos z_1 reduciendo la información a una variable unidimensional, junto con la simplificación se produce una reducción de variabilidad, que pasa a ser

$$\mathrm{Var}(z_1) = \lambda_1.$$

Decimos que el cociente

$$\frac{\lambda_1}{\lambda_1 + \cdots + \lambda_d}$$

es la proporción de variabilidad explicada por la primera componente principal.

Si en lugar de una única componente principal extraemos r componentes resulta

$$\frac{\lambda_1 + \cdots + \lambda_r}{\lambda_1 + \cdots + \lambda_r + \lambda_{r+1} + \cdots + \lambda_d}$$

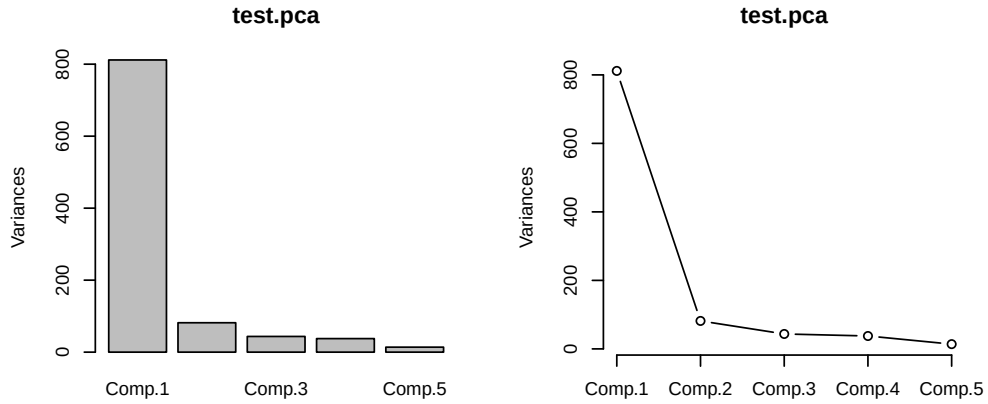
es la proporción de variabilidad explicada por las r primeras componentes principales.

Debemos decidir entre la simplificación que supone la reducción de la dimensión y la pérdida de información resultante de la variabilidad no explicada. Como criterios para tomar esta decisión, se suelen emplear los siguientes:

- Criterio de la varianza explicada. Consiste en retener el número de componentes que conjuntamente expliquen una proporción de varianza establecida, habitualmente un 90% o 95% del total.
- Gráfico de sedimentación (*screeplot*). Representar en un gráfico los valores propios $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ en orden decreciente, y buscar un "codo" en el gráfico, entendiendo por codo un punto a partir del cual los valores propios son claramente más pequeños que los anteriores, y muy similares entre sí.
- Retener un número preestablecido de componentes principales. Por ejemplo, es costumbre retener dos componentes, pues se pueden representar fácilmente en el plano.

Podemos hacer un gráfico de las varianzas de las componentes con `screeplot`.

```
> screeplot(test.pca)  # Representación de los autovalores (gráfico de sedimentación)
> screeplot(test.pca, type = "lines")  # El mismo gráfico con líneas
```



Además, podemos obtener los autovectores asociados (columnas de la matriz `test.pca$loadings`).

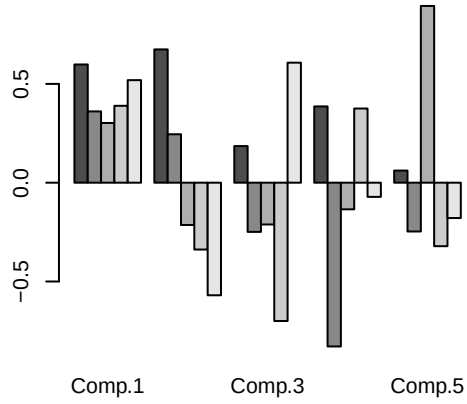
```
> test.pca$loadings
```

```
##
## Loadings:
##        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## prog    0.598  0.675  0.185  0.386
## ingcom  0.361  0.245 -0.249 -0.829 -0.247
## ingsof  0.302 -0.214 -0.211 -0.135  0.894
## sist    0.389 -0.338 -0.700  0.375 -0.321
## estad   0.519 -0.570  0.607        -0.179
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## SS loadings       1.0    1.0    1.0    1.0    1.0
## Proportion Var    0.2    0.2    0.2    0.2    0.2
## Cumulative Var    0.2    0.4    0.6    0.8    1.0
```

Observamos que la primera componente principal es la variable con máxima varianza y tiene todos sus coeficientes positivos. Se puede interpretar como una componente de *tamaño* que determina la "calificación general de los estudiantes". La primera componente ordena las estudiantes según su tamaño (calificación general), del más pequeño al más grande. La segunda componente tiene coeficientes positivos y negativos y se interpreta como uan componente de *forma* que ordena a los estudiantes contrastando los buenos en Programación e Ingeniería de Computadores frente a los buenos en el resto de materias.

Representamos los coeficientes de las componentes principales mediante el siguiente gráfico.

```
> barplot(loadings(test.pca), beside = TRUE)
```



Podemos obtener también el número de observaciones, el vector de medias y los escalados aplicados a cada variable.

```
> test.pca$n.obs
```

```
## [1] 25
```

```
> test.pca$center
```

```
##   prog ingcom ingsof   sist  estad
##  36.76  50.60  50.68  46.04  43.80
```

```
> test.pca$scale
```

```
##   prog ingcom ingsof   sist  estad
##      1      1      1      1      1
```

Por último, obtenemos los `scores`, que son el resultado de XP, siendo P la matriz que tiene como columnas los autovectores de S.

```
> test.pca$scores
```

```
##            Comp.1      Comp.2      Comp.3      Comp.4     Comp.5
##  [1,]  -7.5403215  10.2167650   2.5374713  -8.6708997 -4.3011164
##  [2,]  20.3610372  13.3460340   8.9820585   6.4124949 -1.3548711
##  [3,] -19.5031539   6.5552439  -1.6414327   5.0042015 -2.2980498
##  [4,]  65.9652730   1.3136646   5.1988159  -5.2093505 -1.0457668
##  [5,]   9.7780565   6.0680143  -9.1921582   2.9249303 -2.1069944
##  [6,] -33.0739529  -4.3722312  -3.7345190  -2.6038323 -5.3246839
##  [7,]   1.4212177  -0.7833317  -5.7800406   7.8225646 -0.4967347
##  [8,]  -6.7011638  -0.4667798 -13.3936497  -0.3040531  2.6525120
##  [9,] -36.1916160   5.6543609   2.9062814   5.5458471  6.5171961
## [10,] -38.0586178  -3.8266815  -2.5248244  -6.0338259  6.3212284
## [11,]  -1.6837296  -5.9262762  10.1237093  -4.9410716  4.5102357
## [12,]   6.4961788   3.9922267   5.0805842 -10.8805481 -1.3208797
## [13,]  48.6656614   2.5248899  -7.4227003  -6.1976282  3.0745497
## [14,]  12.8648106 -20.9375616   1.3328311  13.8459667  1.2286291
## [15,]  -5.1279619 -14.2990082  -2.9194088   2.7778854 -9.4299722
## [16,]  14.7627376   1.9542134   2.1019732   6.8802981  0.7262472
```

```
## [17,]   0.5206666   0.3328352  12.2272438   5.5083908  5.1001830
## [18,] -55.8465116   0.7024956   1.3349377  -4.9981377 -2.2873123
## [19,]   1.2809636  24.1913016   1.8618184   5.1875656 -3.1248129
## [20,]  44.0731110  -1.0133086  -8.2094067  -5.5695807  3.6879677
## [21,]   2.7282968 -15.4819784  13.1612877  -5.6870830 -3.1343155
## [22,] -22.3031608  -2.8413762  -5.9443642  -4.0929718  2.9545412
## [23,]  10.4526967  -7.6936163  -3.2010012  -3.0864389 -0.6469043
## [24,]  28.7147101  -2.0746379  -2.7048913   5.2002364 -0.7510021
## [25,] -42.0552279   2.8647426  -0.1806153   1.1650401  0.8501260
```

2 Análisis de componentes principales a través de la descomposición espectral de la matriz de covarianzas

Comprobamos que, efectivamente, los resultados prporcionados por la función `princomp` coindicen con los obtenidos a partir de la descomposición espectral de la matriz de covarianzas. En primer lugar, calculamos el número de observaciones y el vector de medias.

```
> n <- nrow(dat)
> apply(dat, 2, mean)
```

```
##   prog ingcom ingsof   sist  estad
##  36.76  50.60  50.68  46.04  43.80
```

A continuación, calculamos la matriz de covarianzas muestral, sus autovalores y autovectores y comprobamos que coinciden con los resultados de `test.pca$sdev^2` y `test.pca$loadings`

```
> S <- cov(dat) * (n - 1)/n
> auto <- eigen(S)
> lambda <- auto$values
> lambda
```

```
## [1] 811.66184  81.63974  43.57261  37.62083  13.86339
```

```
> v <- auto$vectors
> v
```

```
##            [,1]       [,2]       [,3]        [,4]        [,5]
## [1,] -0.5982782  0.6745404  0.1852556 -0.38597894  0.06131111
## [2,] -0.3607532  0.2450733 -0.2490064  0.82871854 -0.24701742
## [3,] -0.3021774 -0.2140882 -0.2114109  0.13484564  0.89441442
## [4,] -0.3890403 -0.3384022 -0.6999921 -0.37537871 -0.32129949
## [5,] -0.5188995 -0.5697232  0.6074477  0.07178665 -0.17892129
```

A continuación, calculamos la proporción de varianza explicada y sus valores acumulados.

```
> lambda/sum(lambda)
```

```
## [1] 0.82122218 0.08260135 0.04408584 0.03806395 0.01402668
```

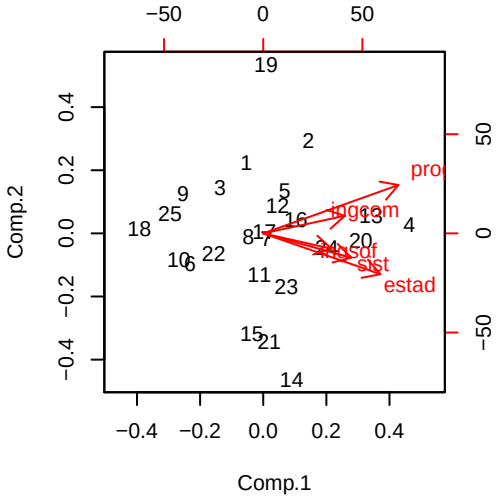
```
> cumsum(lambda/sum(lambda))
```

```
## [1] 0.8212222 0.9038235 0.9479094 0.9859733 1.0000000
```

3 El biplot

El biplot es una representación gráfica simultánea de los individuos (mediante puntos) y las variables (mediante flechas), en un mismo sistema de coordenadas bidimensional construido en base a las dos primeras componentes principales. Permite interpretar el significado de las componentes (la primera en el eje horizontal y la segunda en el eje vertical) en base a las direcciones de las flechas. A su vez, se valoran como parecidos los individuos cuyos puntos están próximos en el biplot. De igual modo tendrán correlación positiva las variables con flechas semejantes. Asimismo, los individuos que se encuentran en la dirección de cierta flecha tendrán observaciones altas en la variable representada por la flecha.

```
> biplot(test.pca)
```



4 Las componentes principales y los cambios de escala

Un problema importante del análisis de componentes principales es que sus resultados dependen de la escala de medida de las variables originales. Así, si se cambia de escala una variable (dejando las demás fijas) las componentes principales se modifican, cambiando incluso la posible interpretación de las mismas. En concreto, si se aumenta la escala de una variable, ésta verá incrementada su varianza y su aportación proporcional a la varianza total, atrayendo de este modo a la primera componente principal, que se asemejará a esta variable.

Claro está que si se realiza el mismo cambio de escala en todas las variables, entonces los resultados del análisis de componentes principales se mantendrán idénticos.

Este problema se puede solventar de dos maneras:

- Midiendo todas las variables en la misma escala (siempre que sean de la misma naturaleza).
- Aplicando el análisis de componentes principales a las variables estandarizadas. Esto último equivale a trabajar con la matriz de correlaciones, en lugar de la matriz de covarianzas.

El archivo `decatlon.txt` contiene los resultados de 33 participantes en la prueba combinada de atletismo durante una competición. Las filas corresponden a los participantes y las columnas a los resultados en las diez pruebas (cuatro carreras, tres lanzamientos y tres saltos). Las variables correspondientes son: 100 metros (X100), salto de longitud (long), lanzamiento de peso (poid), salto de altura (haut), 400 metros (X400), 110-metros vallas (X110), lanzamiento de disco (disq), salto con pértiga (perc), lanzamiento de jabalina (jave) y 1500 metros (X1500).

1. Realiza un Análisis de Componentes Principales con los datos de decatlón. Justifica el uso de la matriz de covarianzas o de la matriz de correlaciones muestrales para llevar a cabo el análisis.
2. Haz una interpretación de las dos primeras componentes principales. ¿Cuál es la proporción de variabilidad explicada por las dos primeras componentes principales?
3. Realiza el biplot correspondiente y comenta la gráfica obtenida.