Análisis de Componentes Principales (Tercera Parte)

Mathias Bourel

IMERL - Facultad de Ingeniería, Universidad de la República, Uruguay

May 7, 2019

Interpretación de los resultados

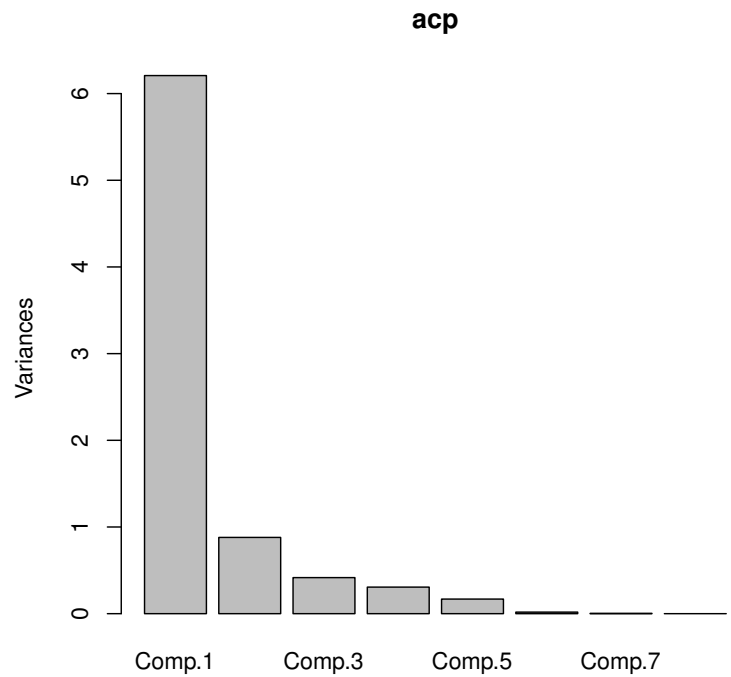
Para interpretar los resultados, debemos:

- Elegir la cantidad de ejes factoriales con los cuales nos quedamos.
- Hacer los gráficos.
- Dar un nuevo significado a las nuevas variables.
- Evaluar los resultados obtenidos.

Elección de la cantidad de ejes

Esencialmente hay dos criterios:

- Criterio del codo. Se seleccionan los ejes antes del decaimiento menor de la varianza.



- Criterio de Kaiser. Se seleccionan los ejes cuya inercia es mayor a la inercia media I/p. Si la matriz es centrada y reducida, se retienen los ejes cuyos valores propios son mayores que 1.

En la practica se retienen los ejes que el usuario sabe interpretar.

Interpretación de los ejes

Para cada eje seleccionado y cada nube, miramos:

- ¿Cuáles son las variables que más contribuyen a la formación del eje?
- ¿Qué individuos participan más en la formación del eje?

Herramienta de medición: contribuciones de puntos (individuos, si no son anónimos y variables) a la inercia de este eje.

Son los puntos cuya contribución es mayor que el promedio lo que hace posible dar significado al eje.

Interpretación de los ejes

Para cada eje que se retiene y cada nube, miramos cuales son los variables que participan más a la formación del eje y cuales son los individuos que contribuyen más a la formación del eje.

Se mide esta contribución respecto de la inercia del eje. Si estos individuos tienen una contribución superior a la media, los mismos dan un sentido al eje.

La contribución del individuo i a la construcción del eje k es

$$ctr_k(i) = \frac{I(i)}{I_k} = \frac{p_i c_{ik}^2}{\lambda_k}$$

La suma de las contribuciones da 1. Se suelen retener los individuos cuya contribución es mayor que $1/n$ en valor absoluto. Si todos los individuos tienen igual peso, entonces retenemos los individuos tales que $|c_{ik}| > \sqrt{\lambda_k}$

Interpretación de los ejes

Para cada eje que se retiene y cada nube, miramos cuales son los variables que participan más a la formación del eje y cuales son los individuos que contribuyen más a la formación del eje.

La contribución de la variable x_j a la construcción del eje k es

$$ctr_k(x_j) = \frac{I(x_j)}{I_k} = \frac{d_{jk}^2}{\lambda_k} = \frac{(\sqrt{\lambda_k} a_{jk})^2}{\lambda_k} = a_{jk}^2$$

La suma de las contribuciones da 1. Se suelen retener las variables cuya contribución es mayor que $1/p$ en valor absoluto, es decir tales que $|a_{jk}| > 1/\sqrt{p}$

Si la matriz de datos es estandarizada, son las variables proximas al borde de la circunferencia que contribuyen más a la construcción del eje, puesto que

$$d_{jk}^2 = r_{x_j, z_k}^2$$

Interpretación de los ejes

Una contribución excesiva de uno de los puntos a un eje debe considerarse con precaución ($\approx$ 25% de inercia).

Se debe garantizar que los puntos que más contribuyen al eje estén bien representados en el eje (de lo contrario, se deben colocar en elementos adicionales).
La contribución es solo una ayuda para la interpretación:

- La contribución de algunos puntos puede estar muy por debajo del umbral y respaldar la interpretación del eje que uno hubiera hecho sin ellos. Luego los incluimos en la interpretación.
- Por el contrario, cuando una contribución es muy fuerte en comparación con otras que están por encima del umbral, el punto determina el eje casi exclusivamente

La interpretación de las nuevas variables (ejes factoriales) se realizará utilizando los individuos y las variables que más contribuyen al eje con la siguiente regla: si una variable tiene una fuerte contribución positiva al eje, los individuos con una fuerte contribución positiva al eje se caracteriza por un alto valor de la variable.

Interpretación de los ejes, ejemplo

Interpretación del primer eje. Contribución de los individuos. $c_{i1} > \sqrt{\lambda_1} = 2.491575$

```
> acp1$eig
        eigenvalue percentage of variance cumulative percentage of variance
comp 1 6.207946839             77.59933549                           77.59934
comp 2 0.879681393             10.99601741                           88.59535
comp 3 0.415961123              5.19951404                           93.79487
comp 4 0.306454670              3.83068337                           97.62555
comp 5 0.168441497              2.10551872                           99.73107
comp 6 0.018067709              0.22584636                           99.95692
comp 7 0.003446769              0.04308461                          100.00000

> sqrt(acp1$eig[1,1])
[1] 2.491575
> acp1$ind$contrib
             Dim.1       Dim.2      Dim.3        Dim.4      Dim.5
PRodRu   22.889096  0.85862826 21.1831585 15.766778699 24.6686523
Asalrur  24.972938  2.84428985  3.7134275 34.237720045 18.0848730
Prof      4.363107  0.04865263  9.1884034 29.781885764 41.6653666
Ejsup    38.255441  0.44069383 31.8307079  0.009398506  4.9692885
Ejmoy     5.943573 10.42770773  0.9164726  6.919852720  0.1180624
Emp       1.309809  9.28909571  3.5736754  0.194884165  3.1475171
Obr       1.627714  0.47607181 28.7236966 12.375700062  6.3785789
Des       0.638321 75.61486019  0.8704580  0.713780038  0.9676612
```

PRodRu, Asalrur, Ejsup son los individuos que contribuyen más a la construcción del primer eje.

Interpretación de los ejes, ejemplo

```
> a$ind$coord
             Dim.1       Dim.2      Dim.3       Dim.4       Dim.5
PRodRu  -3.3715788 -0.24581608  0.8395890 -0.62172682  0.57655700
Asalrur -3.5217117 -0.44739860  0.3515271  0.91617942 -0.49365924
Prof     1.4720309  0.05851415 -0.5529570  0.85448454  0.74930243
Ejsup    4.3587865  0.17610682  1.0291875  0.01517950 -0.25877162
Ejmoy    1.7180777 -0.85664744 -0.1746349 -0.41188554  0.03988644
Emp      0.8065346 -0.80852679 -0.3448490 -0.06912202 -0.20594611
Obr     -0.8991001 -0.18303912 -0.9776683 -0.55082419 -0.29317809
Des     -0.5630391  2.30680707 -0.1701944 -0.13228491 -0.11419083
```

Interpretación de los ejes, ejemplo

Interpretación del primer eje. Contribución de las variables ($1/\sqrt{8} \approx 0.35$).

```
> a$var$contrib #devuelve porcentajes, la suma de cada columna es 100
           Dim.1      Dim.2      Dim.3       Dim.4     Dim.5
PC     15.312396  1.8995166  2.6151444  1.42443318  8.646269
OP     12.157378 19.4115329 10.2367831  4.74843319  7.045927
VC     12.193565  4.0675682 46.3259598  0.08342491  6.037657
OV     13.959573  6.7760892  0.5399656 15.72478652 11.944270
P       6.069888 55.3277802 31.0984451  0.54747464  3.087936
Veg    13.309507  1.6389460  0.1049827 26.92461130 44.781774
Uva    13.916742 10.6262892  6.4642966  0.40584971  7.372980
Platos 13.080951  0.2522777  2.6144227 50.14098653 11.083187

> sqrt(a$var$contrib[,1])# Son los loadings de la primera componente (en val. abs)
      PC       OP       VC       OV        P      Veg      Uva   Platos
3.913106 3.486743 3.491929 3.736251 2.463714 3.648220 3.730515 3.616760

> a$var$coord #los d_{jk}
            Dim.1       Dim.2       Dim.3       Dim.4      Dim.5
PC     -0.9749797  0.12926598  0.10429757  0.06606998  0.1206810
OP      0.8687483  0.41323074  0.20635173  0.12063082 -0.1089416
VC     -0.8700402  0.18916036  0.43897378 -0.01598936  0.1008460
OV      0.9309151  0.24414749  0.04739248 -0.21952071 -0.1418418
P      -0.6138529  0.69764474 -0.35966296 -0.04096049  0.0721205
Veg    -0.9089814  0.12007291  0.02089707  0.28724855 -0.2746472
Uva     0.9294859  0.30574089  0.16397854  0.03526677  0.1114413
Platos  0.9011429 -0.04710881 -0.10428318  0.39199413  0.1366334
```

Contribuyen más PC, VC por un lado OP, OV, Uva, Platos por otro a la construcción del primer eje.

Interpretación de los ejes, ejemplo

La primer componente mide la repartición de la consumición entre alimentos básicos (PC,VC,Veg) y alimentos más refinados (OP, OV, Uva, Platos) y contraponen los ejecutivos superiores a los trabajadores rurales.

El segundo eje es más característico de la consumición de papas, comida generalmente consumida por inactivos.

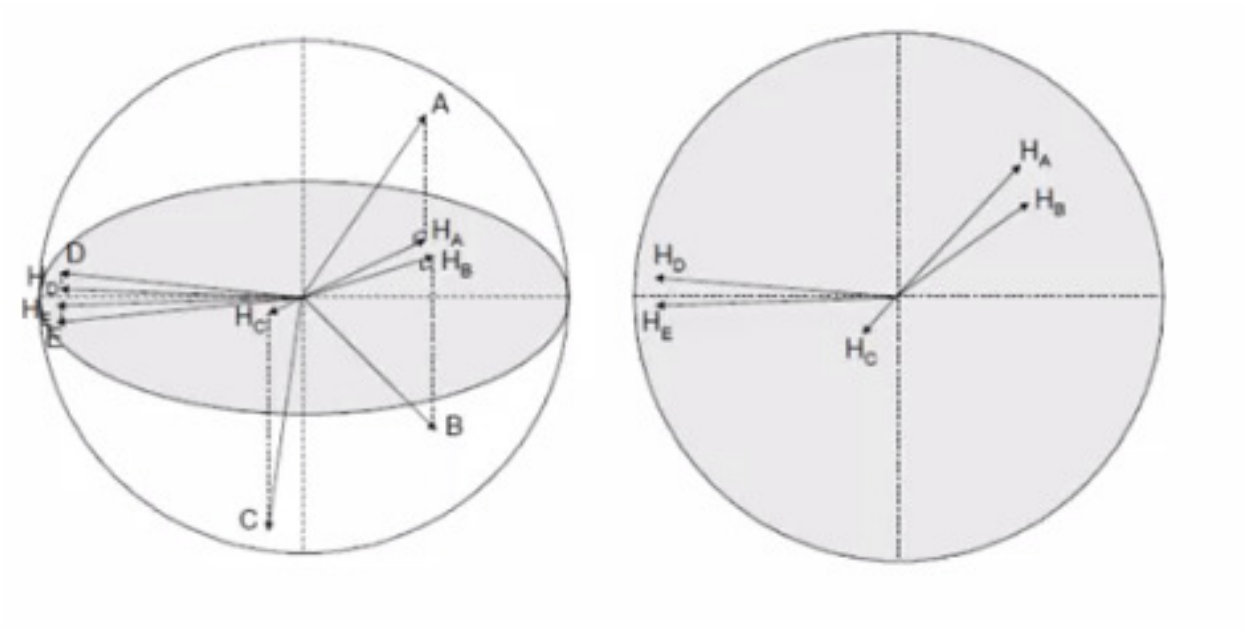
Variables e individuos bien representadas - Calidad de la proyección

Una vez que se interpretan los ejes, uno puede mirar los gráficos y analizar con mayor precisión las proximidades entre los puntos.

Las proximidades entre los puntos observados en un eje o un plano factorial deben corresponder a la realidad (y no deben ser creadas artificialmente por la proyección). Para poder interpretar las proximidades entre puntos, deben estar bien representados en el eje o el plano factorial.

Se dice que un punto está bien representado en un eje o un plano factorial si está cerca de su proyección en el eje o el plano. Si está lejos, se dice que está pobremente representado. El indicador de la calidad de la representación será el ángulo formado entre el punto y su proyección en el eje.

Variables e individuos bien representadas - Calidad de la proyección



$r(A, B) = \cos(A, B)$ y si $\cos(A, B) \approx \cos(H_A, H_B)$ si las variables están bien proyectadas. Sólo las variables bien proyectadas (cerca del eje y del borde del círculo) pueden ser correctamente interpretadas.
Lo mismo en cuanto a los individuos. Si dos individuos están mal proyectados, quizás estén lejos en el espacio de partido.

Calidad de representación de los puntos

Calidad de representación de los individuos.

Se mide si $\mathbf{x}_i$ es próximo a su proyección sobre el eje o el plano factorial con el angulo que forman. Sobre el eje k-esimo la calidad de representación de $\mathbf{x_i}$ es

$$cal_k(i) = \cos^2(\theta_{ik}) = \frac{p_i c_{ik}^2}{||\mathbf{x}_i||^2}$$

Cuando este coseno está cerca de 1, es decir el angulo es 0 o π el individuo está bien representado. En caso contrario el coseno está cerca de cero y el individuo mal representado.
Sobre el plano factorial determinado por a_{k_1} y a_{k_2}, la calidad de la representación es

$$cal_{k_1,k_2}(i) = cal_{k_1}(i) + cal_{k_2}(i)$$

Calidad de representación de los puntos

Calidad de representación de las variables.

De la misma manera la calidad de la representación de la variable x_j sobre el eje k es:

$$cal_k(x_j) = \cos^2(\theta_{jk}) = \frac{d_{jk}^2}{||x_j||^2}$$

y si la matriz de datos es centrada reducida $d_{jk}^2 = r_{jk}^2$.

```
> (a$var$coord[,1])^2==a$var$cos2[,1]
    PC     OP     VC     OV      P    Veg    Uva Platos
  TRUE   TRUE   TRUE   TRUE   TRUE   TRUE   TRUE   TRUE
```

- Una variable estará mejor representada en un eje si estáa cerca del borde del círculo de correlaciones y el eje, mientras que estará pobremente representada si se encuentracerca del origen.
- Solamente las variables bien representadas pueden ser interpretadas!
- Las variables que contribuyen más a la construcción del eje son aquellas que están mejor representadas y al revez: aquellas que contribuyen menos son las que no tienen una buena representación. Esto es porque $cal_k(x_j) = \frac{d_{jk}^2}{||x_j||^2}$ y $ctr_k(x_j) = \frac{d_{jk}^2}{\lambda_k}$.

Calidad de representación de los puntos

Mirar la calidad de las representaciones:

- hace posible resaltar posibles proximidades que no se han notado al interpretar los ejes. Interpretamos las proximidades de elementos bien representados en el plano factorial.
- se utiliza para identificar los puntos que no contribuyen significativamente a la inercia del eje, pero que están bien representados por este eje, es decir, que tienen características especÃficas del eje.

Más aún:

- La proximidad en el espacio factorial entre dos individuos bien representados refleja el parecido real de estos dos individuos en términos de los valores tomados por las variables. (Cuando la calidad de representación de dos individuos es buena, su proximidad observada es parecida a su proximidad real en el espacio).
  Recordar: la lectura directa de las proximidades en el gráfico puede ser errónea: no hay interpretación de las proximidades entre individuos mal representados.
- La proximidad entre dos variables en un eje da, si las dos variables están bien representadas en el eje (cerca del eje y el borde del círculo), una aproximación de su correlaci'on.
  - Dos variables cercanas están correlacionadas positivamente.
  - Dos variables opuestas están correlacionadas negativamente
  - Dos variables ortogonales no están correlacionadas.

Descripción de las dimensiones por las variables

A mayor correlación las variables están muy ligadas a los nuevos ejes. A la derecha tenemos los p-valores que nos indican si el coeficiente de correlación es significativamente distinto de cero.

```
> dimdesc(a,axes=c(1,2))
$Dim.1
$Dim.1$quanti
       correlation      p.value
OV       0.9309151 7.821882e-04
Uva      0.9294859 8.308315e-04
Platos   0.9011429 2.239726e-03
OP       0.8687483 5.110853e-03
VC      -0.8700402 4.966446e-03
Veg     -0.9089814 1.758745e-03
PC      -0.9749797 3.842664e-05


$Dim.2
$Dim.2$quanti
     correlation p.value

> dimdesc(a,axes=c(1,2),p=0.2)
$Dim.1
$Dim.1$quanti
       correlation      p.value
OV       0.9309151 7.821882e-04
Uva      0.9294859 8.308315e-04
Platos   0.9011429 2.239726e-03
OP       0.8687483 5.110853e-03
P       -0.6138529 1.054770e-01
VC      -0.8700402 4.966446e-03
Veg     -0.9089814 1.758745e-03
PC      -0.9749797 3.842664e-05


$Dim.2
$Dim.2$quanti
  correlation    p.value
P   0.6976447 0.05437981
```

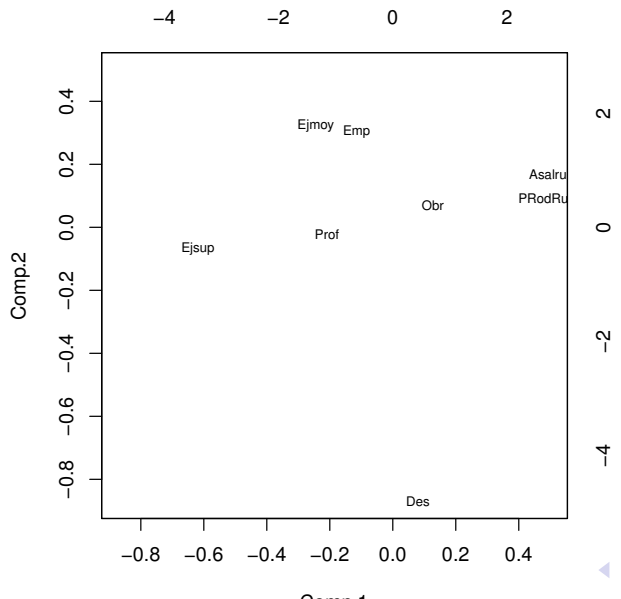
Calidad de la representación

Mirar la calidad de la representación permite mirar las proximidades de los puntos bien representados en el eje factorial.

1. La proximidad entre dos individuos bien representados indica un parecer entre estos individuos en cuanto a los valores que toman las variables. Si la calidad de representación entre estos dos individuos es buena, la proximidad que se observa es real. Si los individuos no están bien representados, no se puede interpretar nada en cuanto a su proximidad.
2. La proximidad entre dos variables bien representadas (cerca del borde del círculo y del eje) es una aproximación de su correlación: si están cercanas están correladas positivamente, si están opuestas negativamente, y si son ortogonales no están correladas.

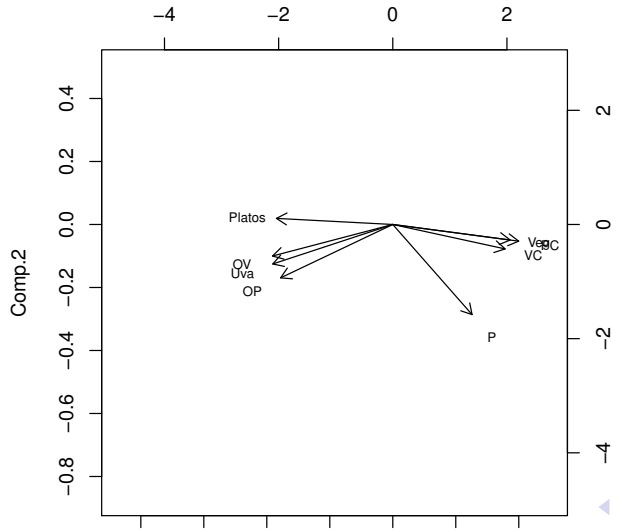
ejemplo

```
> a$ind$cos2
              Dim.1         Dim.2       Dim.3         Dim.4       Dim.5
PRodRu   0.88444010 0.0047013477 0.054844774 3.007468e-02 0.0258634411
Asalrur  0.89805821 0.0144939287 0.008947765 6.077961e-02 0.0176462090
Prof     0.57459845 0.0009079299 0.081079929 1.936150e-01 0.1488829157
Ejsup    0.94181776 0.0015374041 0.052507908 1.142223e-05 0.0033194718
Ejmoy    0.75288231 0.1871740871 0.007778640 4.327076e-02 0.0004057814
Emp      0.42778496 0.4299008580 0.078205503 3.142044e-03 0.0278924499
Obr      0.36060411 0.0149452245 0.426380807 1.353445e-01 0.0383422502
Des      0.05551846 0.9319290611 0.005072836 3.064650e-03 0.0022836142
```



ejemplo

```
> a$var$cos2
            Dim.1      Dim.2       Dim.3        Dim.4       Dim.5
PC      0.9505854 0.01670969 0.0108779840 0.0043652420 0.014563904
OP      0.7547236 0.17075964 0.0425810378 0.0145517953 0.011868265
VC      0.7569700 0.03578164 0.1926979829 0.0002556595 0.010169921
OV      0.8666029 0.05960800 0.0022460470 0.0481893426 0.020119107
P       0.3768154 0.48670819 0.1293574416 0.0016777616 0.005201366
Veg     0.8262471 0.01441750 0.0004366874 0.0825117286 0.075431090
Uva     0.8639440 0.09347749 0.0268889606 0.0012437454 0.012419159
Platos 0.8120585 0.00221924 0.0108749822 0.1536593947 0.018668686
```



Conclusión

El Análisis de Componentes Principales permite analizar las correlaciones entre variables. Se construyen nuevas variables no correladas con varianza importante.

Sin embargo si bien este análisis permite visualizar las correlaciones, las mismas son visibles unicamente sobre planos lo que complica la interpretabilidad si la cantidad de variables es grande y las relaciones entre las mismas más complejas.

El ACP no es adecuado para fenómenos no lineales que son de mayor tamaño. Para este tipo de problema, se han desarrollado otros métodos, como Kernel-PCA (Análisis de componentes principales por núcleo).

Referencias

- FactoMineR `https://www.youtube.com/watch?v=CTSbxU6KLbM&list=PLnZgp6epRBbTsZEFXi_p6W48HhNyqwxIu&index=3`
- D. Peña, *Analisis de Datos Multivariantes*, Mac Graw Hill, 2002.
- J. Blanco, *Introducción al Análisis Multivariado*, Instituto de Estadística, Facultad de Ciencias Económicas y Administración, Universidad de la República.