Un ejemplo de ACP paso a paso

Francesc Carmona
Departament d'Estadística

13 de enero de 2014

1. Introducción

Para ilustrar el procedimiento de cálculo, partamos de un ejemplo en el que disponemos de la valoración media que han hecho los encuestados sobre siete marcas de coche, con arreglo a tres características. En un estudio real hubiéramos considerado no sólo tres sino diez o veinte características, ya que el ACP tiene ventajas cuando la dimensión de la tabla que se pretende analizar es muy grande.
La siguiente tabla recoge las valoraciones medias que han concedido los encuestados a cada una de las marcas en las tres características consideradas. Así la marca A tiene una calificación media de 2 en la característica elegancia, de 3 en comodidad y 6 en deportividad.

	Características		
Marca	Elegancia	Comodidad	Deportividad
A	2	3	6
B	3	2	4
C	4	5	4
D	5	5	4
E	8	9	6
F	9	7	7

El objetivo del estudio es poner de relieve los factores que diferencian al máximo las marcas entre sí, determinar las marcas que el conjunto de encuestados considera semejantes y conocer las características causantes de este parecido o las que diferencian. Se trata de obtener un mapa sobre el cual se posicionan las marcas y características, permitiendo ver las relaciones entre ellas.

2. Preliminares

Antes de aplicar el ACP debe comprobarse si es necesario, es decir, si la correlación entre las variables analizadas es lo suficientemente grande como para justificar la factorización de la matriz de coeficientes de correlación. Esta comprobación puede hacerse mediante el test de Bartlett (1950), que parte de la hipótesis nula de que la matriz de coeficientes de correlación no es significativamente distinta de la matriz identidad. Bartlett calcula un estadístico basado en el valor del determinante de la matriz de coeficientes de correlación del siguiente modo:

$$-[n-1-(2k+5)/6]\ln|\mathbf{R}| \sim \chi^2_{(k^2-k)/2}$$

donde k es el número de variables de la matriz, n el tamaño de la muestra y $\mathbf{R}$ la matriz de correlaciones. En nuestro ejemplo la matriz de correlaciones entre las características es:

```
             elegancia comodidad deportividad
elegancia        1.000     0.892        0.585
comodidad        0.892     1.000        0.519
deportividad     0.585     0.519        1.000
```

y la prueba de esfericidad de Bartlett para esta matriz de correlaciones es:

```
 Bartlett's sphericity test

chi.square =  6.341 , df =  3 , p-value =  0.0961431
```

Con este resultado no deberíamos continuar nuestro análisis ya que con un nivel de significación del 0,05 no rechazamos la hipótesis nula de esfericidad. Sin embargo, la distribución ji-cuadrado asociada es asintótica y supone la normalidad multivariante de los datos. En nuestro caso podemos dudar de la normalidad conjunta y, sobre todo, el tamaño muestral es muy pequeño $n = 6$.

Nota

El test de Bartlett tiene otro un gran inconveniente. Tiende a ser estadísticamente significativo cuando el tamaño muestral n crece. Algunos autores advierten que únicamente se utilice cuando la razón n/k sea menor que 5.

El índice de Kaiser-Meyer-Olkin o medida de adecuación muestral KMO tiene el mismo objetivo que el test de Bartlett, se trata de saber si podemos factorizar las variables originales de forma eficiente.
El punto de partida también es la matriz de correlaciones entre las variables observadas. Sabemos que las variables pueden estar más o menos correlacionadas, pero la correlación entre dos de ellas puede estar influenciada por las otras. Así pues, utilizaremos la correlación parcial[1] para medir la relación entre dos variables eliminando el efecto del resto. El índice KMO compara los valores de las correlaciones entre las variables y sus correlaciones parciales. Si el índice KMO está próximo a 1, el ACP se puede hacer. Si el índice es bajo (próximo a 0), el ACP no será relevante. Algunos autores han definido una escala para interpretar el índice KMO de un conjunto de datos.
El siguiente resultado nos muestra la medida de adecuación muestral KMO para nuestros datos y el valor en la escala.

```
$overall
[1] 0.6317966

$report
[1] "The KMO test yields a degree of common variance mediocre."

$individual
   elegancia    comodidad deportividad
   0.5811766    0.5965991    0.8592540
```

Además de la medida KMO global que en nuestro caso es "mediocre", también se han calculado las medidas por variable de manera que podamos detectar aquellas que no están relacionadas con las demás. Para mejorar nuestro análisis deberíamos añadir más variables como hemos dicho al principio (y más observaciones). Se recomienda un mínimo de tres variables por factor.

3. Las componentes principales

El siguiente paso consiste en la obtención de los valores y vectores propios de la matriz de covarianzas muestral o de la matriz de coeficientes de correlación que se obtienen a partir de la matriz de datos. La elección de una u otra matriz para realizar el ACP es una cuestión controvertida. En este caso vamos a utilizar la matriz de correlaciones.

```
Importance of components:
                          Comp.1    Comp.2     Comp.3
Standard deviation     1.5312421 0.7421283 0.32333168
Proportion of Variance 0.7815674 0.1835848 0.03484779
Cumulative Proportion  0.7815674 0.9651522 1.00000000
```

La varianza asociada a cada factor (el cuadrado de las desviaciones estándar) viene expresada por su valor propio o raíz característica de la matriz de coeficientes de correlación (en este caso) o de la matriz de covarianzas.

```
Variances:


   Comp.1    Comp.2    Comp.3
2.3447023 0.5507544 0.1045434
```

[1] http://en.wikipedia.org/wiki/Partial_correlation

Los otros elementos importantes en un ACP son los vectores propios asociados a cada valor propio.

```
Loadings:
             Comp.1 Comp.2 Comp.3
elegancia    -0.619 -0.290  0.730
comodidad    -0.604 -0.419 -0.678
deportividad -0.502  0.861

               Comp.1 Comp.2 Comp.3
SS loadings     1.000  1.000  1.000
Proportion Var  0.333  0.333  0.333
Cumulative Var  0.333  0.667  1.000
```

Cada columna representa una combinación lineal (loadings) de las variables originales que proporcionan las componentes principales o factores. Así la primera componente se obtiene con la siguiente combinación:

$$F_1 = -0.619 * \text{elegancia} - 0.604 * \text{comodidad} - 0.502 * \text{deportividad}$$

Observamos que la primera componente tiene todos los coeficientes negativos. De manera que, aunque no es obligatorio, por necesidades de interpretación y estéticas cambiaremos todos esos coeficientes (de la primera componente) de signo. En consecuencia también debemos cambiar las puntuaciones o scores de la primera componente.

```
Loadings:
             Comp.1 Comp.2 Comp.3
elegancia     0.619 -0.290  0.730
comodidad     0.604 -0.419 -0.678
deportividad  0.502  0.861

               Comp.1 Comp.2 Comp.3
SS loadings     1.000  1.000  1.000
Proportion Var  0.333  0.333  0.333
Cumulative Var  0.333  0.667  1.000
```

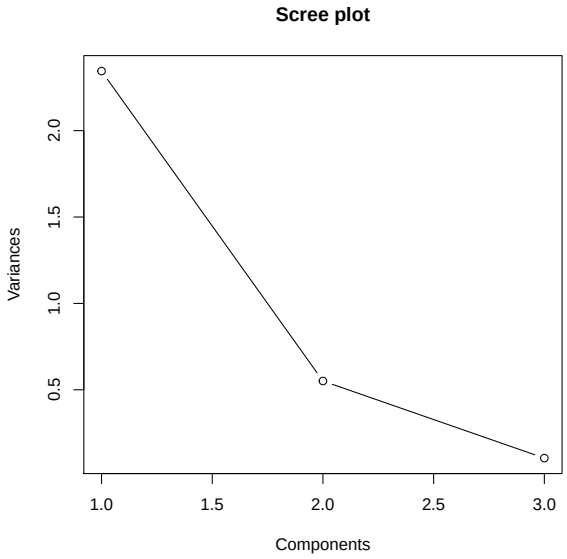

Figura 1: Gráfico de sedimentación.

La determinación del número de factores a retener es, en parte, arbitraria y queda a juicio del investigador. Un criterio es retener los factores con valor propio superior a 1.
También podemos representar un gráfico de sedimentación (scree plot) de los valores propios como el de la figura 1 y considerar el número de componentes en el que el descenso se estabiliza.

En este caso nos inclinamos por retener los dos primeros ya que explican un $96,52\,\%$ de la varianza y permiten una representación gráfica en dos dimensiones.

Como los factores no son directamente observables, su denominación es, en cierto modo, subjetiva, aunque se basa en las cargas de los factores con las variables originales. La **carga del factor** es la correlación existente entre una variable original y un factor, obtenido por combinación lineal de las variables originales. Estas correlaciones se pueden calcular como producto de los coeficientes o loadings y las desviaciones de cada componente:

```
Correlations:
```

```
             Comp.1 Comp.2 Comp.3
elegancia     0.948 -0.215  0.236
comodidad     0.925 -0.311 -0.219
deportividad  0.769  0.639 -0.027
```

Con las dos primeras columnas de correlaciones como coordenadas se dibuja el círculo de correlaciones 2 que permite interpretar los ejes o componentes principales.

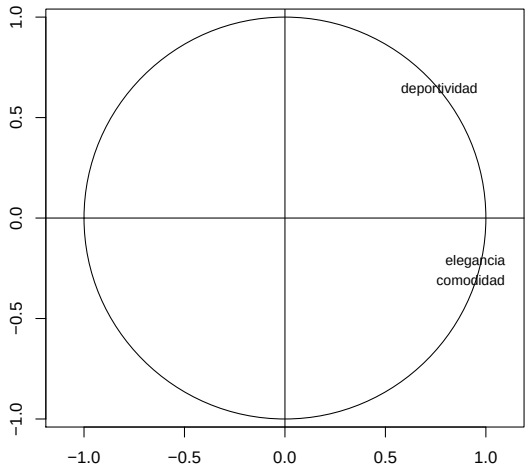

Figura 2: Círculo de correlaciones.

La **comunalidad** es un valor que se obtiene en el análisis factorial, para cada una de las variables originales, sumando los cuadrados de las correlaciones o cargas de los factores retenidos con la variable para la que se calcula y que expresa la proporción de varianza de la variable extraída o explicada con m factores, donde m es el número de factores retenidos. Si m es igual al número total de variables la comunalidad será igual a 1.
Los **cosenos** son las correlaciones al cuadrado y su acumulación proporciona las comunalidades.

```
Cosinus:
```

```
             Comp.1 Comp.2
elegancia     0.898  0.046
comodidad     0.855  0.097
deportividad  0.591  0.408
```

```
Communalities:
```

```
             Comp.1 Comp.2 Comp.3
elegancia     0.898  0.944      1
comodidad     0.855  0.952      1
deportividad  0.591  0.999      1
```

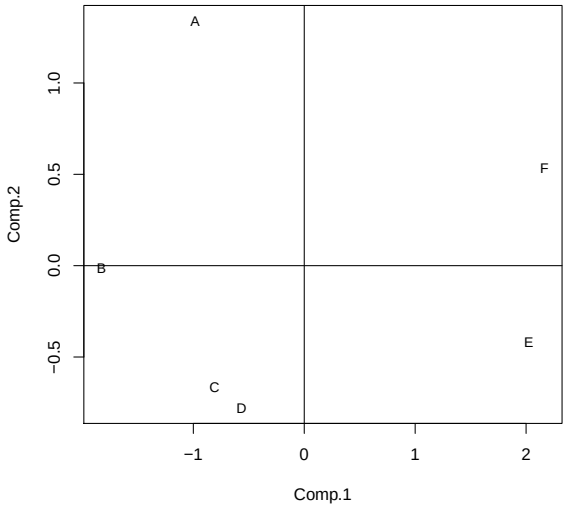

Figura 3: Representación de las marcas en dos dimensiones.

4. Resultados e interpretación del ACP

El principal resultado es el gráfico de puntuaciones de la figura 3 donde se representan las observaciones o marcas en los ejes formados por las dos primeras componentes o factores principales.
La nube de puntos-individuos está centrada en el origen, puesto que hemos centrado los datos iniciales. No ocurre lo mismo con la nube de variables en la figura 2. Los puntos-variables pueden, como en este caso, estar situados todos en el mismo lado, por ejemplo, $F_1 > 0$. Esto se debe a que las características están correlacionadas positivamente, y cuando un individuo (marca) toma valores altos en una característica, también los obtiene altos en las otras.
Se observa que las coordenadas de los puntos-variables son inferiores en valor absoluto a 1. Ello obedece a que las variables han sido tipificadas, con lo cual su distancia al origen es la unidad, y al proyectarlas sobre los ejes se puede producir una contracción y acercarse al origen, pero nunca un alejamiento.
El factor o componente principal es una variable artificial que se obtiene como combinación lineal de las tres características consideradas. Cada una de las marcas toma un valor en esta variable, su proyección.
La coordenada de un punto-variable sobre el factor es el coeficiente de correlación de éste (variable artificial) con la variable. Así,

$$\text{cor}(\text{elegancia}, F_1) = 0{,}95, \quad \text{cor}(\text{comodidad}, F_1) = 0{,}92, \quad \text{cor}(\text{deportividad}, F_1) = 0{,}77$$

El factor se interpreta en función de las variables más correlacionadas con él. En consecuencia, el primer factor combina la elegancia y la comodidad y en menor medida la deportividad, opone las marcas que toman valores altos en estas características a aquellas que toman valores bajos. Es un factor que podríamos llamar de *prestigio*. De izquierda a derecha ordena las marcas de menor a mayor prestigio.
Se observa en este ejemplo un fenómeno frecuente en el ACP. El primer factor es un factor de *tamaño* o talla. Opone los individuos que toman valores altos en todas las características correlacionadas positivamente con él, a los que toman valores bajos.
El segundo eje está muy correlacionado con la característica deportividad. Opone las marcas que la poseen a las que no. En el ACP clásico es un factor de *forma*.

La representación simultánea en la figura 4 de las dos nubes de puntos sobre el plano formado por los dos primeros ejes facilita la interpretación.

¡Atención!

En el gráfico **biplot** no se busca la proximidad entre observaciones y variables. ¡Son las direcciones lo que es importante!

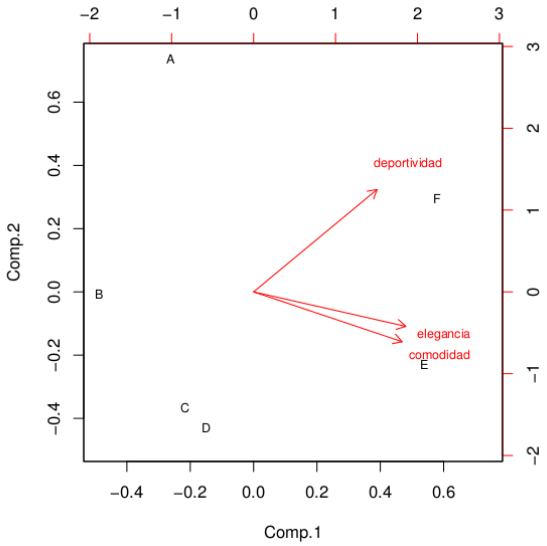

Figura 4: Gráfico biplot que combina la representación de las observaciones y de las variables.

En el primer cuadrante tenemos que $F_1 > 0$, luego se caracteriza por elegancia y comodidad; además $F_2 > 0$, por tanto, se caracteriza por ser deportivo. En consecuencia, la marca F situada en este cuadrante posee las tres características estudiadas, y en este sentido será la mejor.
En el cuarto cuadrante $F_1 > 0$, la marca situada en él, E, se caracteriza por la elegancia y la comodidad, pero no es deportiva.
En el tercer cuadrante se sitúan C y D, que son semejantes, pero no se caracterizan por ninguna de estas variables. Toman valores muy bajos para todas las características consideradas, y son las peores.
En el segundo cuadrante se sitúa la marca A, que si bien no es elegante ni cómoda, sí es deportiva $F_2 > 0$.

5. Rotación Varimax

Con el fin de facilitar la interpretación del significado de los factores seleccionados se suele llevar a cabo una rotación de los ejes factoriales. Uno de los métodos más corrientes es el Varimax, desarrollado por Kaiser (1958), que efectúa una rotación ortogonal de los ejes factoriales. El objetivo de la rotación Varimax es conseguir que la correlación de cada una de las variables sea lo más próxima a 1 con sólo uno de los factores y próxima a cero con todos los demás.
Recordemos que las correlaciones de las variables con las componentes obtenidas han sido las siguientes:

```
Correlations:
```

```
             Comp.1 Comp.2
elegancia     0.948 -0.215
comodidad     0.925 -0.311
deportividad  0.769  0.639
```

Con la rotación varimax de dos factores se obtienen las siguientes:

```
               RC1   RC2
elegancia    0.915 0.327
comodidad    0.947 0.234
deportividad 0.306 0.952
```

En el gráfico 5 se observa la rotación de las variables de forma que ahora el primer eje se identifica más con la elegacia y la comodidad, mientras que el segundo eje coincide con la deportividad.

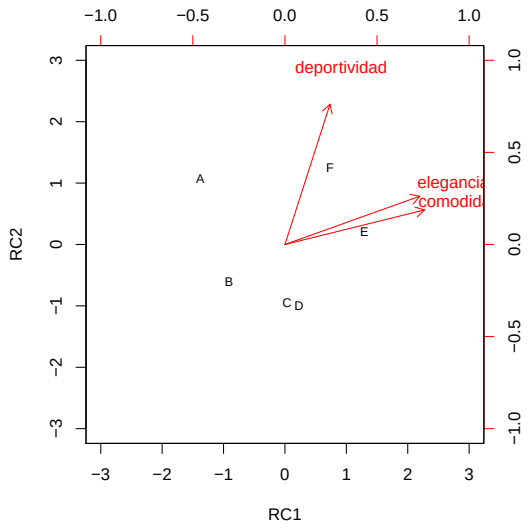

Figura 5: Gráfico biplot con la rotación Varimax de los ejes.

Con la rotación varimax de todos los factores se obtiene el mejor resultado, ya que practicamente asimila cada variable con un eje.

```
Loadings:
             RC1   RC2   RC3
elegancia    0.705 0.312 0.637
comodidad    0.938 0.253 0.237
deportividad 0.252 0.952 0.173

                RC1   RC2   RC3
SS loadings    1.44 1.068 0.492
Proportion Var 0.48 0.356 0.164
Cumulative Var 0.48 0.836 1.000
```

Comentario

Éste es un estudio comparativo de las marcas, no evaluativo. Pueden ser todas muy buenas o muy malas, pero el estudio determina únicamente las diferencias entre ellas, no el valor; éste se aprecia estudiando los valores iniciales.

Referencias

[1] Abascal, Elena y Grande, Ildefonso. *Métodos multivariantes para la investigación comercial*, Ariel Economía, Barcelona, 1989.

[2] Rakotomalala, Ricco. *Tutoriels Tanagra: ACP – Description de véhicules*
`http://tutoriels-data-mining.blogspot.fr/2008/03/acp-description-de-vhicules.html`

[3] Santesmases, Miguel. *Diseño y análisis de encuestas en investigación social y de mercados*, Ed. Pirámide, Madrid, 2009.

[4] Saporta, Gilbert. *Probabilités, Analyse de données et Statistique*, Dunod, 2011.