Componentes principales (II)

Eliseo Martínez Herrera

1. Propiedades de los componentes

Los componentes principales tienen las siguientes propiedades:

1 La suma de las varianzas de los componentes es igual a la varianza de las variables originales. En fecto, puesto que $Var(z_j) = \lambda_j$, y la suma de los valores propios de $\mathbf{S}$ es la traza de $\mathbf{S}$ y además por construcción de esta matriz se tiene que

$$tr(\mathbf{S}) = \sum_{j=1}^{p} Var(x_j)$$

entonces

$$tr(\mathbf{S}) = \sum_{j=1}^{p} Var(x_j) = \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} Var(z_j)$$

De modo que las nuevas variables, los componentes principales, tienen conjuntamente la misma variabilidad que las variables originales.

2 La proporción de la variabilidad explicada por un componente es el cociente entre su varianza, que es el vector propio que lo define, y la suma de los valores propios de la matriz $\mathbf{S}$. Eso es

$$\frac{\lambda_h}{\sum_{j=1}^{p} \lambda_j}$$

es la proporción de la varianza explicada por el componente h.

3 Las covarianzas entre cada componente principal y las variables columnas de $\mathbf{X}$ vienen dadas por el producto de las coordenadas del vector propio y el autovalor propio asociado, esto es

$$Cov(z_i, x_1, \ldots, x_p) = \lambda_i \mathbf{a}_i = \lambda_i \begin{pmatrix} a_{i1} & \cdots & a_{ip} \end{pmatrix}$$

donde $\mathbf{a}_i$ es el i-ésimo autovector que define al i-ésimo componente.

4 El coeficiente de correlación lineal entre la i-ésima componente y la j-ésima variable columna de $\mathbf{X}$, $\rho(z_i, x_j)$ está dado por

$$\rho(z_i, x_j) = \frac{Cov(z_i, x_j)}{\sqrt{Var(z_i)\, Var(x_j)}} = \frac{\lambda_i\, a_{i\,j}}{\sqrt{\lambda_i\, s_j^2}} = a_{i\,j}\, \frac{\sqrt{\lambda_i}}{s_j}$$

5 Los r componentes principales $(r < p)$ proporcionan la predicción lineal óptima con r variables del conjunto de valores de variables $\mathbf{X}$. De otra forma, si queremos aproximar la matriz $\mathbf{X}$, de rango p, por otra matriz $\widehat{\mathbf{X}}_r$ de rango $r < p$, la aproximación óptima es

$\widehat{\mathbf{X}}_r = \mathbf{X}\mathbf{A}_r\mathbf{A}_r^t$ donde la matriz $\mathbf{A}_r$ es de $p \times r$ y sus columnas son los vectores propios asociados a los r mayores valores propios de la matriz $\mathbf{S}$.

6 Si estandarizamos los componentes principales, dividiendo cada uno por su desviación típica, se obtiene la estandarización multivariante de los datos originales. En efecto, la matriz de componentes $\mathbf{Z}$ se obtienen mediante la ecuación

$$\mathbf{Z} = \mathbf{X} \cdot \mathbf{A}$$

Si estandarizamos los componentes $\mathbf{Z}$ por sus desviaciones típicas, debemos considerar la matriz diagonal $\mathbf{D}$ formada en su diagonal principal por las varianzas de los componentes, luego invirtiendo y sacando la raíz cuadrada para obtener la desviación estándar como denominador, obtenemos la estandarización

$$\mathbf{Y} = \mathbf{Z} \cdot \mathbf{D}^{-\frac{1}{2}}$$

y entonces se concluye que

$$\mathbf{Y} = \mathbf{Z} \cdot \mathbf{D}^{-\frac{1}{2}} = \mathbf{X} \cdot \mathbf{A} \cdot \mathbf{D}^{-\frac{1}{2}}$$

que es la estandarización de las variables originales.

2. Análisis normado o con correlaciones

Para determinar el componente principal asociado al vector de dirección principal $\mathbf{a}$, se debe maximizar la función

$$M = \mathbf{a}^t\mathbf{S}\mathbf{a}$$

sujeta a la condición $\mathbf{a}^t\mathbf{a} = 1$. La función M puede escribirse como

$$M = \sum_{i=1}^{p} a_i^2 s_i^2 + 2\sum_{i=1}^{p}\sum_{j=i+1}^{p} a_i a_j s_{i\,j} \tag{1}$$

Supongamos que, a modo de ejemplo, la varianza s_1^2 es mucho mayor que las demas varianzas, una manera de maximizar M es sencillamente es hacer tan grande como se pueda la coordenada a_1 asociada a esta variable x_1. Si una variable original tiene una varianza mucho mayor que las demás, el primer componente coincidirá muy aproximadamente con esta variable, en efecto recuerde que el primer componente satisface para cada observación i:

$$z_{1i} = x_{i1}\, a_1 + \cdots + x_{ip}\, a_p$$

De modo que si una variable tiene una varianza mucho mayor que las demás, el primer componente principal coincidirá con esta variable.

De tal modo que esta pripiedad dependerá del tamaño de escala que esté utilizando una determinada variable, de tal manera que la maximización de (1) dependerá decisivamente de la escala a usar en cada variables. Es decir las escalas con valores más grandes tendrán mayor peso en el análisis. Una manera de evitar este riesgo consiste en estandarizar las variables antes de calculas los componentes principales. Una vez estandarizadas las

variables, se tiene que la función que se debe maximizar es

$$M' = 1 + 2\sum_{i=1}^{p}\sum_{j=i+1}^{p} a_i a_j r_{i\,j}$$

siendo r_{ij} en coeficiente de correlación entre las variables i y j. En consecuencia, esta maximización dependerá de las correlaciones y no de las varianzas.

Los componnete principales normados se obtienen calculando los vectores y valores propios de la matriz $\mathbf{R}$, matriz de los coeficientes de correlación lineal. Si llamamos λ_i^R a las raíces características de esa matriz, que vamos a suponer es no singular, se verifica que

$$\sum_{i=1}^{p} \lambda_i^R = traza(R) = p$$

Las propiedades de los componnetes principales extraídos de $\mathbf{R}$ son:

1 La proporción de variación explicada por λ_i^R será

$$\frac{\lambda_i^R}{p}$$

2 Las correlaciones entre cada componente z_j y las variables $\mathbf{X}$ originales vienen dados directamente por $\mathbf{a}^t{}_j\sqrt{\lambda_j}$ siendo $z_j = \mathbf{X}\mathbf{a}_j$.

Cuando las variables originales de $\mathbf{X}$ están en distintas unidades conviene aplicar el análisis de los componentes principales emanados de la matriz $\mathbf{R}$ de correlación. Cuando las variables originales tienen las mismas unidades ambas alternativas son posibles. Si las diferencias entre las variables son informativas y queremos considerar este hecho en el análisis no conviene estandarizar las variables. Por el contrario, si las diferencias entre las varianzas no son relevantes, simplemente se elimina del análisis considerando la matriz de correlaciones.

Ejemplo. Este conjunto de datos llamados INVEST y que puede ser obtenido el el Internet *http://www.mhe.es/universidad/ciencias_matematicas/pena/ficheros.html* presenta 21 observaciones de 8 variables. Las observaciones corrsponden a los países de la OCDE y las variables son el número de publicaciones científicas recogidas en el trienio 1982-84 en ocho bases de datos de producción científica. Las variables se han llamado según la orientación de la base de datos: InterA (por interdisciplinaria), Inter F (por interdisciplinaria), Agric., Biolo., Medic., Quimic., Ingen. y Física. Fuente: Caballero y Peña (1987).

Pasando estos datos a logaritmo natural para "suavizarlos" y utilizando el software MINITAB, eligiendo esta vez la opción de cálculo de los componentes principales generados por la matriz de correlación, obtenemos el siguiente resultado

Eigenanalysis of the Correlation Matrix

Autoval	7.3683	0.2407	0.1793	0.0984	0.0497	0.0422	0.0120	0.0094
propor.	0.921	0.030	0.022	0.012	0.006	0.005	0.002	0.001
acum.	0.921	0.951	0.974	0.986	0.992	0.997	0.999	1.000

Y los vectores propios de las ocho componentes son

Vectores propios de las ocho componentes

Variables	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
X_1	-0,362	-0.089	-0.334	-0.088	-0.126	-0.382	-0.356	0.673
X_2	-0,336	0.618	0.608	-0.302	-0.158	-0.066	0.036	0.116
X_3	-0,360	-0.041	-0.015	-0.263	0.837	0.080	-0.250	-0.171
X_4	-0,355	-0.366	-0.126	-0.491	-0.179	0.376	0.546	0.104
X_5	-0,364	-0.149	-0.106	-0.141	-0.432	.0.219	-0.350	-0.676
X_6	-0,339	-0.527	0.585	0.500	0.001	0.048	-0.042	0.108
X_7	-0,352	0,363	-0,307	0,422	-0,116	0,655	-0,171	0,021
X_8	-0,359	0,207	-0,236	0,378	0,162	-0,472	0,597	-0,155

La interpretación de estos resultados son como sigue: la primera tabla muestra en la primera fila los autovalores ordenados en forma decreciente, y puesto que son los autovalores de la matriz de correlación $\mathbf{R}$, la suma de estos autovalores debe ser igual $p = 8$ (el número de variables); la segunda fila describe para cada autovalor la proporción de varianza explicada; y la tercera fila describe la proporción acumulada. La segunda tabla tiene la siguiente lectura. Cada columna, indicadas por C_1 hasta la C_8, son los autovectores asociados a los respectivos autovalores ya ordenados de mayor a menor en la primera table. Los componentes en rigor están denotados por $C_1, .., C_8$ y las variables originales están denotadas por X_1 hasta X_8, y su interpretación, a modo de ejemplo para la primera componente, es la siguiente:

$$\begin{aligned} C_1 &= -0.362X_1 - 0,336X_2 - 0,360X_3 - 0,355X_4 \\ &\quad -0,364X_5 - 0,339X_6 - 0,352X_7 - 0,359X_8 \end{aligned}$$

3. Interpretación de los componentes

Toda vez que hemos pasado de las variables originales determinadas por las columnas de X, ¿cuál es la interpretación de los componentes principales?

En primer lugar. la interpretación tiene un sentido cuando solo elegimos las primeras componentes, puesto que en definitiva el objetivo es "reducir variables". En segundo lugar, los componentes elegidos, en rigor los primeros componentes elegidos estára asociado a la proporción de varianza acumulada, puesto que si, a modo de ejemplo, los tres primeros componentes ya explican el 90% de la variabilidad total, tendrá sentido entonces buscar la interpretación de estos tres primeros componentes. En tercer lugar, es claro que entre las variables originales tiene que existir un cierto grado de colinealidad o correlación, puesto que si las variables originales, en un caso extremo, no están correlacionadas su matriz de varianza o de correlación serán diagonales y poca información podrán entregar y evidentemente no podremos reducir variables. Un vez pasados estos unmbrales, estamos en condiciones de realizar una interpretación sobre los componentes principales.

Si existe una alta correlación entre las variables originales, por lo regular el primer componente tendrá sus coordenadas del mismo signo, de modo que su interpretación es de ser un promedio ponderado de todas las variables originales. De otra forma, el primer

componente entrega el factor global de tamaño. En términos más simples, las observaciones pueden ser unicadas en un orden descendente de los individuos evaluados conforme a las puntuaciones obtenidas según este componente. Los restantes componentes, que tendrán coordenadas positivas y negativas, se interpretarán como factores de "forma", puesto que se van a contraponer grupos de variables originales frente a otros grupos de variables. De otra forma, estos factores de forma serán medias ponderadas de dos grupos de variables con distinto signo, y además en cada una de las medias habrán variables que dominan unas màs que otras y serán estas que darán sentido a la "forma" que describe el componente.

Daremos un breve y sencillo ejemplo dídactico que intente reflejar esta interpretación.
Ejemplo. Supongamos que nuestras unidades muestrales o individuos a ser observados serán 6 rectángulos, sobre los cuales mediremos dos atributos: longitud de la base, y altura del rectángulo (medidas en las mismas unidades de longitud, por ejemplo centímetros). De modo que, para este ejemplo, $p = 2$. Supongamos que las observaciones obtenidas son las indicadas por la siguiente tabla,

rectángulo	X_1(base)	X_2 (altura)
1	2	2
2	1.5	0.5
3	0.7	0.5
4	0.5	1.5
5	0.5	0.7
6	0.7	0.7

Aplicaremos a estos datos logaritmo en base 10 para facilitar la interpretación de sus componentes. la matriz de las observaciones queda entonces como

$$\mathbf{X} = \begin{pmatrix} .30103 & .30103 \\ .17609 & -.30103 \\ -.1549 & -.30103 \\ -.30103 & .17609 \\ -.30103 & -.1549 \\ -.1549 & -.1549 \end{pmatrix}$$

Con cualquier software matemático, por ejemplo con el DERIVE, obtenemos la matriz de varianzas y covarianzas

$$\mathbf{S} = \begin{pmatrix} 0.06387086891 & 0.01407125000 \\ 0.01407125000 & 0.06387086891 \end{pmatrix}$$

Los autovalores de esta matriz son

$$\lambda_1 = 0.07794211891 \ ; \ \lambda_2 = 0.04979961890$$

cuyos autovectores asociados correspondientes son

$$\begin{aligned} \mathbf{a}_1 &= \begin{pmatrix} 0.7071067804 & 0.7071067804 \end{pmatrix} \\ \mathbf{a}_2 &= \begin{pmatrix} 0.7071067804 & -0.7071067804 \end{pmatrix} \end{aligned}$$

Entonces las primeras componentes son

$$z_1 = 0.7071067804\, X_1 + 0.7071067804\, X_2$$

$$z_2 \quad = \quad 0.7071067804\, X_1 - 0.7071067804\, X_2$$

recordando que X_i es el logaritmo en base 10 de la i-ésima variable original, $i = 1, 2$. De manera que las evaluaciones de las primeras componentes para las 6 observaciones se obtienen del siguiente producto matricial

$$\begin{pmatrix} .30103 & .30103 \\ .17609 & -.30103 \\ -.1549 & -.30103 \\ -.30103 & .17609 \\ -.30103 & -.1549 \\ -.1549 & -.1549 \end{pmatrix} \begin{pmatrix} 0.7071067804 & 0.7071067804 \\ 0.7071067804 & -0.7071067804 \end{pmatrix}$$

$$= \begin{pmatrix} .42572 & 0 \\ -8.8346 \times 10^{-2} & .33737 \\ -.32239 & .10333 \\ -8.8346 \times 10^{-2} & -.33737 \\ -.32239 & -.10333 \\ -.21906 & 0 \end{pmatrix}$$

En resumen

rectángulo	z_1	z_2
1	0.426	0
2	-0.088	0.337
3	-0.322	0.103
4	-0.088	-0.337
5	-0.322	-0.103
6	-0.219	0

Si ordenamos los rectángulos según la primera componente:

rectángulo	z_1
1	0.426
2	-0.088
4	-0.088
6	-0.219
3	-0.322
5	-0.322

y ordenando los rectángulos según la segunda componente

rectángulo	z_2
2	0.337
3	0.103
1	0
6	0
5	-0.103
4	-0.337

Notemos que el primer ordenamiento es por el factor "tamaño". En efecto, los rectángulos 1, 2, 4, 6, 3, 5 tienen las siguientes áreas respectivas: 4, 0.75, 0.75, 0.49, 0.35, 0.35. ¡El primer componente los ha ordenado por el "tamaño" o área!

El segundo ordenamiento es un poco más complicado de ver, pero relaciona la base con la altura. Esto es, considera primero los rectángulos cuya base es superior a la altura. Observemos que los rectángulos 1 y 6 (que son los que dan la pista) tienen igual base y altura, mientras que los rectángulos 2 y 3 su base es mayor que la altura, siendo el rectángulo 2 el que tiene mayor diferencia positiva entre la base y la altura, los rectángulos 5 y 4 son los que su altura es mucho mayor que su base, siendo el rectángulo 4 el de mayor diferencia negativa entre la base y la altura. ¡El segundo componente los ha ordenado según su "forma"!.

4. Selección del número de componentes

¿Cuántos componentes principales seleccionar?

Se tienen los siguientes tres criterios para determinar el número de componentes:

- Realizar un gráfico de los puntos $(\lambda_i, i)\,, i = 1, ..., p$, que a menudo se llama *gráfico de sedimentación*, y comenzar eligiendo componentes hasta que los restantes puntos estén a la misma altura de un autovalos λ_k. La idea es buscar un "codo" o cambio brusco de pendiente a la cual a partir de este codo la pendiente es aproximadamente un plano horizontal. De otra forma, buscar el valor de k de tal forma que los demás autovalores, λ_j con $j > k$, tengan casi el mismo valor, y ese valor de k indica el número de componentes a considerar.
- Seleccionar componentes de tal forma que entre ellas la proporción de varianza acumulada satisfaga un requerimiento a priori, como por ejemplo el 80 o 90%. Sin embargo, este criterio no debe usarse a rajatabla, puesto que es posible que el primer componente alcance por sí solo el 90%, y puede existir otros componentes que nos expliquen la "forma" de las variables, que con este criterio lo perderíamos.
- Desechar aquellos componentes asociados a valores propios que son inferiores a una cota establecida como puede ser la varianza media de los componentes, esto es $\sum \lambda_i/p$. Y en caso que estemos trabajando con la matriz de correlación $\mathbf{R}$, que será lo más frecuente, este valor es 1, de tal manera que solamente consideraremos aquellas componentes asociados a los autovalores mayores que 1. Cuando las variables originales son pocas, es posible que un solo autovalor cumpla este requisito, y podríamos caer en la arbritrariedad del punto anterior. Por lo general este criterio se utiliza cuando el número de variables originales es suficientemente grande y nos permite encontar por lo menos 3 componentes principales cuyos autovalores satisfacen la cota de 1. Se debe usar con cuidado.