Componentes principales

Eliseo Martínez Herrera

1. Introducción

Uno de los objetivos principales del análisis de datos multivariante es la reducción del número de la dimensionalidad, no queriendo con esto decir que basta encontrar una variable en función lineal, o de otro tipo de relación, para reducir la dimensión en 1. Más que eso, de una dimensión de p variables intentar reducir la dimensión a $r < p$, pero de tal forma que no ocurra una gran pérdida de información. El llamado análisis en componentes principales tiene ese objetivo. Dada n observaciones de p variables, se construye un método para reemplazar esas p variables por otro número reducido de nuevas variables y que en definitiva son combinaciones lineales de las originales, y que además estas nuevas variables expliquen en un gran porcentaje la variabilidad original. De otra forma, para fijar ideas supongamos, a modo de ejemplo, que tenemos un gran número de variables que las podemos reducir a un 20% del número original, de tal forma que estas nuevas variables no estén correlacionadas pero sin embargo pueden explicar más del 85% de la variabilidad original, teniendo en cuenta que este 20% de "nuevas" variables se formarán por combinaciones de clases de variables originales fuertemente correlacionadas. Y además estas "nuevas" variables, que reducen la dimensionalidad, en nuestro ejemplo en un 80%, son las llamadas variables "latentes" no correlacionadas, y que no son observadas directamente por los datos, o no se reflejan directamente en la matriz de datos.

2. Planteamiento del problema

Supongamos sin pérdida de generalidad que nuestra matriz de datos X ya está centrada, esto es $\bar{X} = X$, de modo que la matriz de varianza y covarianza está dada por $S = \frac{1}{n-1} X^t X$. Nuestro problema es encontrar un espacio de dimensión más reducida que represente adecuadamente los datos.

Se desea, entonces, encontrar un subespacio menor que p de tal forma que al proyectar los puntos sobre este nuevo subespacio, los puntos conserven su estructura con la menor distorsión posible.Supongamos que nuestros datos son puntos (o vectores) en el plano esto es $p = 2$, y queremos hacer una buena reducción a una dimensión, esto es a una recta. La Fig. 1 indica la dispersión de puntos en el plano, y una recta que, al parecer, proporciona un buen resumen de los datos, puesto que la recta pasa cerca de todos los puntos y la distancia entre ellos se mantiene aproximadamente entre las proyecciones a la recta.

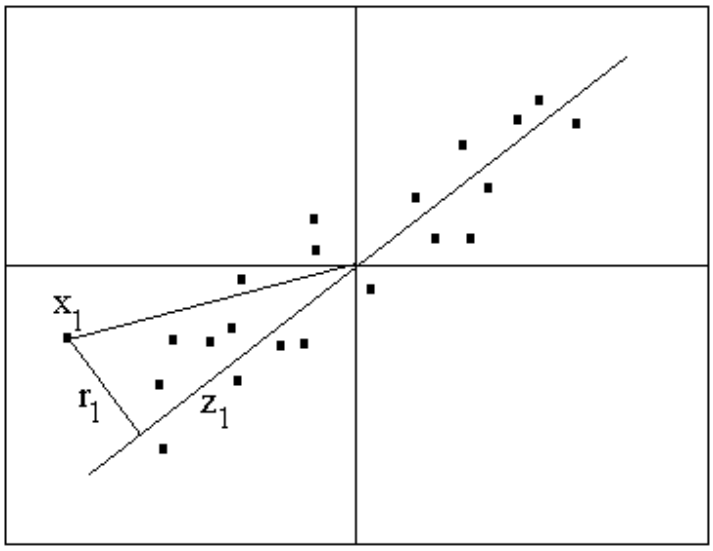

Fig. 1

Impondremos la condición de que la distancia de los puntos originales hacia las proyecciones de la recta sea la menor posible. Si consideramos el punto x_i cualquiera y una dirección fija, digamos a_1, de norma unitaria, entonces la proyección de x_i sobre la recta generada por el vector a_1 está dada obviamente por una expresión del tipo $z_i \cdot a_1$, donde

$$z_i = x_{i1}a_1 + \cdots + x_{ip}a_p \qquad (1)$$

y al factor escalar z_i a veces se le llama coeficiente de Fourier. Para convencerse de esto observe la Fig. 2.

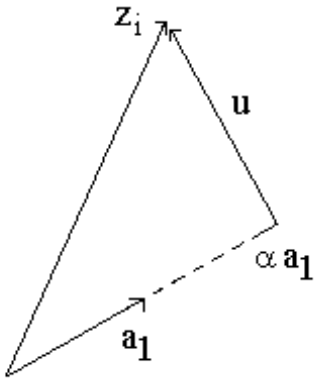

Fig. 2

En esta Fig. 2 observamos que el vector u es perpendicular al vector αa_1, donde de

momento el coeficiente escalar ® lo desconocemos, y por lo tanto se satisface que [1]:

$$u \pm a_1 = 0$$

Sin embargo el vector u es la diferencia entre a_1 y z_i, esto es

$$u = \circledR a_1 \; ¡ \; z_1$$

de modo que

$$\begin{aligned} (\circledR a_1 \; ¡ \; z_1) \pm a_1 &= 0 \\ \circledR a_1^2 \; ¡ \; z_1 \pm a_1 &= 0 \end{aligned}$$

y pusto que $a_1^2 = 1$ se concluye que efectivamente el coeficiente de proyección o de Fourier es el de la ecuación (1).

En consecuencia todos los puntos x_i tendrán un coeficiente de proyección de la forma:

$$z_i = x_i \pm a_1 = a_1^t x_i$$

donde hemos continuado la notación matricial en vez del producto escalar. Y omo lo establecimos en la Fig. 2 el vector que representa a esta proyección es $z_i \, a_1$. Definamos por r_i la distancia entre el punto x_i y el punto de la proyección correspondiente, esto es $z_i \, a_1$ y puesto que estamos exigiendo que las distancias r_i sean mínimas, esto es

$$\textit{minimizar} \sum_{i=1} r_i^2 = \sum_{i=1} \| x_i \; ¡ \; z_i \, a_i \|^2$$

donde la doble barra denota la norma de un vector.

La Fig. 1 muestra que al proyectar el punto sobre la recta se forma un triángulo rectángulo, donde la distancia del punto a la proyección es precisamente r_i, y la distancia de la proyección al origen es z_i, y puesto que la distancia de la hipotenusa esto es del punto al origen se obtiene mediante $(x_i^t x_i)^{1=2}$. De manera que podemos aplicar el teorema de Pitágoras, es decir

$$x_i^t x_i = z_i^2 + r_i^2$$

de manera que si sumamos sobre i, tenemos

$$\sum_{i=1} x_i^t x_i = \sum_{i=1} z_i^2 + \sum_{i=1} r_i^2$$

Puesto que el primer miembro es constante, en rigor son las filas de la matriz de datos centrada, de tal forma que las dos sumas de la derecha que son positivas están acotadas, y esto significa que minimizar la suma $\sum_{i=1}^n r_i^2$ es equivalente que maximizar la suma $\sum_{i=1}^n z_i^2$. Ahora bien, resulta sencillo verificar que las variable z_i tienen media cero. En efecto, en virtu de (1) tenemos que

$$\frac{1}{n} \sum_{i=1} z_i = \frac{1}{n} \left(a_1 \sum_{i=1} x_{i1} + ¢¢¢ + a_p \sum_{i=1} x_{ip} \right)$$

y puesto que las columnas de la matriz X las hemos supuesto centradas en sus medias, se tiene que los z_i tienen media cero, en consecuencia maximizar la suma de sus cuadrados equivale a maximizar su varianza. De otra forma, elegiremos un subespacio de proyección de tal forma que los puntos de proyección tengan varianza máxima. Y esta es la palanca

[1] El símbolo ± denota el producto escalar.

articuladora de los componentes principales. Para visualizar esta idea, suponga que hemos elegido otra recta de proyección a la indicada en la Fig. 1, que es como se indica en la Fig. 3. En ella, intuitivamente, podemos ver que los puntos proyectados tendrán muy poca variabilidad y por ende perderíamos mucha información si elegimos ese subespacio.

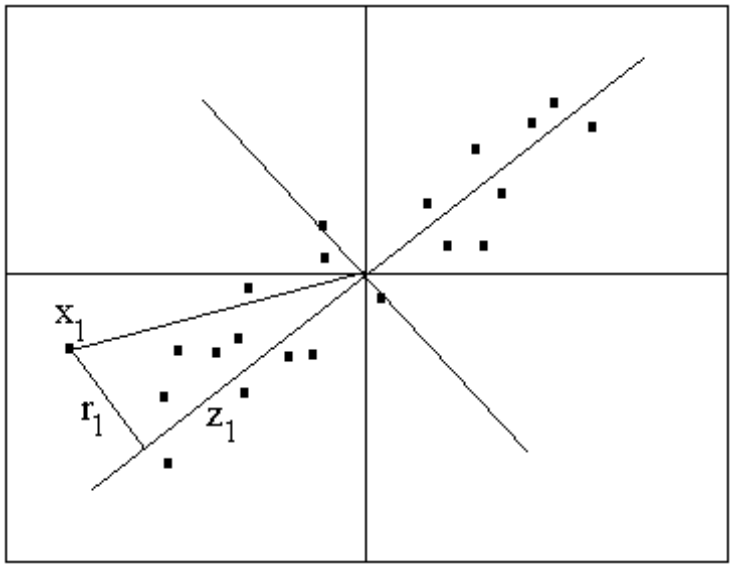

Fig. 3

3. Cálculo de los componentes

3.1 Cálculo del primer componente

En virtud de las ideas intuitivas, descriptivas y geométricas de la sección anterior, debemos encontrar una determinada dirección de una recta, definida por un vector a_1, de tal modo que las proyecciones de los puntos definidos por los vectores columnas de la matriz X, y que llamaremos z_1 al vector constituido por estas proyecciones, tenga varianza mínima entre sus componentes. De otra forma, llamaremos el primer componente principal a la combinación lineal de las variables originales, las columnas de X, que tengan varianza mínima. Esto significa que z_1 que satisface la ecuación (2) debe tener varianza mínima,

$$z_1 = X a_1$$

Como las columnas de X tienen media cero, también lo tendrá las componentes del vector z_1, de modo que su varianza es simplemente la suma cuadrática de sus componentes

dividido por $n - 1$, esto es

$$\frac{1}{n-1} z_1^t z_1 = \frac{1}{n-1} a_1^t X^t X a_1 = a_1^t S a_1 \quad (2)$$

Es claro que podemos maximizar la varianza hasta el infinito simplemente aumentando el módulo de a_1, y puesto que solamente a_1 nos indica la dirección de proyección, podemos, sin perdidad de generalidad, acotar a_1 de tal manera que tenga norma 1, esto es que $a_1^t a_1 = 1$. De tal forma que con esta restricción vamos a maximizar la expresión de la derecha de (2), y esto lo hacemos mediante la técnica de Lagrange. Para esto definimos la función vectorial objetivo a maximizar, que es

$$M = a_1^t S a_1 - \lambda (a_1^t a_1 - 1)$$

Derivamos esta expresión respecto de las componentes de a_1 e igualamos a cero. Se obtiene

$$\frac{\partial M}{\partial a_1} = 2Sa_1 - 2\lambda a_1 = 0$$

y la solución es

$$Sa_1 = \lambda a_1 \quad (3)$$

De esta forma la dirección elegida para proyectar las variables originales, es la indicada por el autovector a_1 de la matriz de varianza y covarianzas S, y además para que efectivamente la varianza de estas proyecciones sea máxima se tiene que λ debe ser el mayor autovalor de S, y, lo cual es algo extraordinariamente simple, el propio autovalor λ es la varianza de las componentes del vector z_1. De otra forma el radio espectral de la matriz simétrica semidefinida positiva S es la varianza del componnete principal z_1. En efecto, $a_1^t S a_1 = \lambda a_1^t a_1 = \lambda$

Por otro lado si definimos a los vectores columnas de la matriz X como $X_1; X_2; :::; X_p$, y además considerando que $a_1^t = (a_1; a_2; :::; a_p)$ entonces

$$z_1 = a_1 X_1 + a_2 X_2 + \cdots + a_p X_p \quad (4)$$

Y esta ecuación describle como el primer componente está en función lineal de las p variables originales, de tal modo que la variabilidad de z_1 es máxima. Ahora si aplicamos el operador varianza a la ecuación (4) y recordando que $Var(Xi) = s_i^2$, nos queda

$$\lambda = a_1^2 s_1^2 + a_2^2 s_2^2 + \cdots + a_p^2 s_p^2$$

Nota: Este es un resultado que tiene cierta elegancia. En efecto, si suuponemos por un momento que las columnas de la matriz X están estandarizadas, esto es dividida cada columna por sus desviaciones estándar, entonces $Var(X_i) = 1$, y puesto que a_1 es unitario, se concluye que $\lambda = 1$

Veamos un ejemplo sencillo.

Ejemplo de las acciones. Busque en la dirección de Internet

http://www.mhe.es/universidad/ciencias_matematicas/pena/ficheros.html

y en ese lugar baje el fichero llamado **Acciones**. Este conjunto de datos presenta 34 observaciones y 3 variables. Las observaciones corresponden a distintas acciones que cotizan en el mercado continuo español y las variables a tres medidas de rentabilidad de estas acciones durante un período de tiempo. Las variables son : X_1, la rentabilidad efectiva por dividen-

dos, X_2 es la proporción de beneficios que va a dividendos y X_3 la razón entre precio por acción y beneficios. Analizando estos datos mediante el programa **firstcomp** realizado en el DERIVE, y que está en la ubicación del Internet

http : //www.uantof.cl/facultades/csbasicas/Matematicas/academicos/emartinez/magister/firstcom.mth

Cargando en ese programa los datos del fichero **Acciones** obtendrá la siguiente matriz de varianza y covarianzas

$$S = \begin{pmatrix} 29{:}09562388 & 100{:}4406506 & -15{:}70281016 \\ 100{:}4406506 & 576{:}2292780 & -18{:}54627540 \\ -15{:}70281016 & -18{:}54627540 & 22{:}56599224 \end{pmatrix}$$

El mayor autovalor de esta matriz resulta ser

$$\lambda = 594{:}8566052$$

y el autovector unitario asociado a este autovalor es

$$a_1 = \begin{pmatrix} 0{:}1756683717 & 0{:}9837650412 & -0{:}03670104663 \end{pmatrix}$$

Dentro del mismo programa está definida los valores de las proyecciones z_i, esto es

$$z_i = \sum_{i=1}^{3} x_{ij}\, a_{1j}$$

$$z_i = x_{i1}\, 0{:}175683717 + x_{i2}\, 0{:}9837650412 + x_{i3}\,(-0{:}03670104663)$$

donde se verifica que

$$\frac{1}{33}\sum_{i=1}^{33} z_i = 594{:}8566052$$

■

3.2 Cálculo del segundo componente

Supongamos ahora que queremos proyectar los puntos de X en un plano, donde el plano estará definido por dos vectores a_1 y a_2, de tal forma que la suma de las varianzas de las proyeciones $z_1 = Xa_1$ y $z_2 = Xa_2$ sea máxima. De manera que debemos maximizar la función

$$\acute{A} = a_1^t S a_1 + a_2^t S a_2 - \lambda_1(a_1^t a_1 - 1) - \lambda_2(a_2^t a_2 - 1) \qquad (5)$$

donse se incorporan las restricciones que deben cumplir los vectores direccinales, esto es $a_i^t a_i = 1;\ i = 1; 2$. Derivando respecto de estas direcciones e igualando a cero obtenemos:

$$\frac{\partial \acute{A}}{\partial a_1} = 2Sa_1 - 2\lambda_1 a_1 = 0$$

$$\frac{\partial \acute{A}}{\partial a_2} = 2Sa_2 - 2\lambda_2 a_2 = 0$$

La solución de este sistema es

$$Sa_1 = \lambda_1 a_1$$

$$Sa_2 = \lambda_2 a_2$$

que indica que las direcciones a_1 y a_2 deben ser vectores propios de S. Sustituyendo estos valores en (5) se concluye que la función tiene su máximo en

$$\acute{A} = \lambda_1 + \lambda_2$$

de modo que λ_1 y λ_2 deben ser los mayores autovalores de la matriz S asociado a los autovectores a_1 y a_2 respectivamente.

El resultado importante que se desprende es que las proyecciones z_1 y z_2 tienen covarianza cero, puesto que los autovectores de una matriz simétrica semidefinida positiva son ortogonales, esto es $a_1^t a_2 = 0$, y por lo tanto $a_1^t S a_2 = 0$, que es la covarianza entre ambos vectores.

Veamos un ejemplo.

Encuesta de presupuestos familiares en España. En la misma dirección en el Internet del ejemplo de las acciones ubique el fichero de datos **EPF** (encuesta de presupuestos familiares). Estos datos corresponden a 51 observaciones y 9 variables. Las observaciones son las provincias españolas más Ceuta y Melilla, que aparecen unidas como una única provincia, y las variables los nueve epígrafes en los que se desglosa la Encuesta de Presupuestos Familiares en España. Las variables son: X_1 = alimentación, X_2 = vestido y calzado, X_3 = vivienda, X_4 = moviliario doméstico, X_5 = gastos sanitarios, X_6 = transporte, X_7 = enseñanza y cultura, X_8 = turismo y ocio, X_9 = otros gastos. Las unidades son gastos promedios anuales por familia en pesetas. Fuente: Encuesta de Presupuestos Familiares del año 1990/91.

Haciendo leves modificaciones del programa **firstcom** del DERIVE, podemos calcular los dos mayores autovalores de la matriz de varianzas y covarianzas con sus respectivos autovectores unitarios asociados. No obstante a la matriz de datos la vamos a "suavizar" aplicando logaritmo natural a cada una de sus entradas. Sobre este nuava matriz procedemos a calcular los mayores autovalores de la matriz S asociada. De hecho podemos calcular todos los autovalores, estos son [0.348, 0.010, 0.005, 0.0176, 0.027, 0.013, 0.032, 0.011, 0.006].

Podemos observar que el radio espectral es $\lambda_1 = 0{:}348$ y el segundo autovalor mayor es $\lambda_2 = 0{:}032$. Los autovectores unitarios asociados a estos autovalores respectivamente son

$$a_1 = \begin{pmatrix} 0{:}12 & 0{:}18 & 0{:}30 & 0{:}31 & 0{:}46 & 0{:}34 & 0{:}50 & 0{:}31 & 0{:}31 \end{pmatrix}$$

$$a_2 = \begin{pmatrix} 0{:}05 & 0{:}16 & -0{:}17 & 0{:}07 & -0{:}21 & 0{:}29 & -0{:}40 & -0{:}17 & 0{:}78 \end{pmatrix}$$

De tal manera que las proyecciones de las 51 observaciones en cada una de estas componentes principales vienen dada por las siguientes fórmulas

$$\begin{aligned} z_1 &= 0{:}12X_1 + 0{:}18X_2 + 0{:}30X_3 + 0{:}31X_4 + 0{:}46X_5 \\ &\quad +0{:}34X_6 + 0{:}50X_7 + 0{:}31X_8 + 0{:}31X_9 \\ z_2 &= 0{:}05X_1 + 0{:}16X_2 - 0{:}17X_3 + 0{:}07X_4 - 0{:}21X_5 \\ &\quad +0{:}29X_6 + -0{:}40X_7 - 0{:}17X_8 + 0{:}78X_9 \end{aligned}$$

¿Qué hemos conseguido con estas proyecciones o nuevas "puntuaciones"? Es en esta sim-

ple sumas y restas que dan puntuaciones a z_1 y a z_2 donde aparecen factores "latentes" que caracterizan a las provincias de España. Veamos la interpretación.

Podemos observar que la puntuación z_1 es una suma ponderada de todos los gastos, y donde hay mayor ponderación en X_7 (gastos enseñanza y cultura) y X_5 (gastos en salud). En cualquier caso, si ordenamos las 51 puntuaciones para z_1 de mayor a menor, quedarán ordenadas las provincias de España según el factor "renta". De manera que el primer componente descubre el factor renta, de otra manera nos indica un ordenamiento de las provincias según su renta. Se puede descubrir que las tres mayores puntuaciones corresponden a Navarra. Madrid y Barcelona, respectivamente.

La segunda componente la podemos arreglar de la siguiente manera equivalente

$$\begin{aligned} z_2 &= (0{:}05X_1 + 0{:}16X_2 + 0{:}07X_4 + 0{:}29X_6 + 0{:}78X_9) \\ &\quad - (0{:}17X_3 + 0{:}21X_5 + 0{:}40X_7 + 0{:}17X_8) \end{aligned}$$

Y podemos observar que esta variable proyectada es la diferencia entre dos promedios ponderados aproximadamente, en efecto

$$0{:}05 + 0{:}16 + 0{:}07 + 0{:}29 + 0{:}78 = 1{:}35$$

$$0{:}17 + 0{:}21 + 0{:}40 + 0{:}17 = 0{:}95$$

El primer promedio da mayor peso a X_9 (otros gastos) y X_6 (transporte). Suponiendo, por parte de la fuente de encuesta, que otros gastos incluye el envío de dinero a otras provincias, y pensando que este envío está asociado a gastos en transporte, de manera que podemos pensar que esta ponderación positiva descubre a las provincias que transfieren dinero fuera de ella. Y por otro lado, el segundo promedio, la que se resta, da mayor ponderación a X_7 (gastos en enseñanza y cultura) y X_5 (gastos en salud). De manera que esta puntuación z_2 va aseparar a las provincias que realizan una mayor transferencia de dinero a otras provincias, de aquellas que transfieren poco pero tienen altos gastos en educación y salud. Realizando este ordenamiento, es sugestivo que la última provincia en puntuar según este componente es Barcelona Es decir, realiza poca transferencia a otras provincias y gasta mucho en educación y salud (lo que confirma numéricamente una evidencia para quienes conocen la dinámica catalana). Quien recibe mayor puntuación en esta componente es la provincia de Zamora.

¡Este es el potencial que tiene el análisis en componentes principales!

3.3 Generalización

La manera de extender estas proyecciones a $r > 2$ es análoga. Las direcciones para las r rectas se llaman *direcciones principales* de los datos y las nuevas variables definidas por ellas (las fórmulas de las proyecciones) se llaman *componentes principales*.

Vamos a suponer, en general, que la matriz de los datos X, y por lo tanto la matriz de varianzas y covarianzas S, tiene rango p, de modo que existirán tantas componentes principales como valores propios o raíces características $¸_1; :::; ¸_p$ se tenga de la matriz S, que vienen definidas por las soluciones del polinomio característico

$$\det(S - ¸I) = 0$$

y sus vectores asociados son

$$(S - \lambda I)\, a_i = 0$$

Los valores λ_i son reales al ser la matriz S simétrica, y por ser una matriz definida positiva estos autovalores serán positivos. Además si λ_i y λ_j son dos raíces distintas sus autovectores asociados serán ortogonales. Además si S es semidefinida positiva de rango $r < p$ habría solamente r raíces características positivas, siendo los restantes $p - r$ igual a cero.

Supongamos que formamos la matriz cuadrada A, de dimensión igual al número de componentes principales que deseamos obtener o si se quiere igual al número de direcciones principales que vamos a suponer es r, definida como aquella que tiene por columnas los vectores direccionales, de otra forma sus columnas son los autovectores asociados, entonces las puntuaciones (componentes principales) que son los valores de las r componentes en los n individuos viene dada por la matriz Z de dimensión $n \times r$ que satisface la relación

$$Z = X \cdot A$$

Observe que $A^t A = I$, de tal forma que calcular los componentes principales equivale a aplicar una transformación ortogonal A a las variables X (ejes originales) para obtener unas nuevas variables incorrelacionadas entre sí.