Análisis de Componentes Principales

Cuando se recoge la información de una muestra de datos, lo más frecuente es tomar el mayor número posible de variables. Sin embargo, si tomamos demasiadas variables sobre un conjunto de objetos, por ejemplo 20 variables, tendremos que considerar $\binom{20}{2} = 180$ posibles coeficientes de correlación; si son 40 variables dicho número aumenta hasta 780. Evidentemente, en este caso es difícil visualizar relaciones entre las variables.

Otro problema que se presenta es la fuerte correlación que muchas veces se presenta entre las variables: si tomamos demasiadas variables (cosa que en general sucede cuando no se sabe demasiado sobre los datos o sólo se tiene ánimo exploratorio), lo normal es que estén relacionadas o que midan lo mismo bajo distintos puntos de vista. Por ejemplo, en estudios médicos, la presión sanguínea a la salida del corazón y a la salida de los pulmones están fuertemente relacionadas.

Análisis de Componentes Principales

Se hace necesario, pues, reducir el número de variables. Es importante resaltar el hecho de que el concepto de mayor información se relaciona con el de mayor variabilidad o varianza. Cuanto mayor sea la variabilidad de los datos (varianza) se considera que existe mayor información, lo cual está relacionado con el concepto de entropía.

Análisis de Componentes Principales

Componentes Principales

Estas técnicas fueron inicialmente desarrolladas por Pearson a finales del siglo XIX y posteriormente fueron estudiadas por Hotelling en los años 30 del siglo XX. Sin embargo, hasta la aparición de los ordenadores no se empezaron a popularizar.

Para estudiar las relaciones que se presentan entre p variables correlacionadas (que miden información común) se puede transformar el conjunto original de variables en otro conjunto de nuevas variables incorreladas entre sí (que no tenga repetición o redundancia en la información) llamado conjunto de componentes principales.

Las nuevas variables son combinaciones lineales de las anteriores y se van construyendo según el orden de importancia en cuanto a la variabilidad total que recogen de la muestra.

De modo ideal, se buscan $m < p$ variables que sean combinaciones lineales de las p originales y que estén incorreladas, recogiendo la mayor parte de la información o variabilidad de los datos.

Análisis de Componentes Principales

Si las variables originales están incorreladas de partida, entonces no tiene sentido realizar un análisis de componentes principales.

El análisis de componentes principales es una técnica matemática que no requiere la suposición de normalidad multivariante de los datos, aunque si esto último se cumple se puede dar una interpretación más profunda de dichos componentes.

Análisis de Componentes Principales

Cálculo de los Componentes Principales

Se considera una serie de variables $(x_1, x_2, ..., x_p)$ sobre un grupo de objetos o individuos y se trata de calcular, a partir de ellas, un nuevo conjunto de variables $y_1, y_2, ..., y_p$, incorreladas entre sí, cuyas varianzas vayan decreciendo progresivamente.

Cada y_j (donde $j = 1, \ldots, p$) es una combinación lineal de las $x_1, x_2, ..., x_p$ originales, es decir:

$$\begin{aligned} y_j &= a_{j1}x_1 + a_{j2}x_2 + ... + a_{jp}x_p = \\ &= \mathbf{a}'_j\mathbf{x} \end{aligned}$$

siendo $\mathbf{a}'_j = (a_{1j}, a_{2j}, ..., a_{pj})$ un vector de constantes, y

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

Análisis de Componentes Principales

Obviamente, si lo que queremos es maximizar la varianza, como veremos luego, una forma simple podría ser aumentar los coeficientes a_{ij}. Por ello, para mantener la ortogonalidad de la transformación se impone que el módulo del vector $\mathbf{a}'_i = (a_{1i}, a_{2i}, ..., a_{pi})$ sea 1. Es decir,

$$\mathbf{a}'_j \mathbf{a}_j = \sum_{k=1}^{p} a_{kj}^2 = 1$$

El primer componente se calcula eligiendo $\mathbf{a}_1$ de modo que y_1 tenga la mayor varianza posible, sujeta a la restricción de que $\mathbf{a}'_1 \mathbf{a}_1 = 1$. El segundo componente principal se calcula obteniendo $\mathbf{a}_2$ de modo que la variable obtenida, y_2 esté incorrelada con y_1.

Del mismo modo se eligen $y_1, y_2, \cdots, y_p$, incorrelados entre sí, de manera que las variables aleatorias obtenidas vayan teniendo cada vez menor varianza.

Análisis de Componentes Principales

Proceso de extracción de factores:

Queremos elegir $\mathbf{a}_1$ de modo que se maximice la varianza de y_1 sujeta a la restricción de que $\mathbf{a}_1'\mathbf{a}_1 = 1$

$$Var(y_1) = Var(\mathbf{a}_1'\mathbf{x}) = \mathbf{a}_1'\Sigma\mathbf{a}_1$$

El método habitual para maximizar una función de varias variables sujeta a restricciones el método de los *multiplicadores de Lagrange*.

El problema consiste en maximizar la función $\mathbf{a}_1'\Sigma\mathbf{a}_1$ sujeta a la restricción $\mathbf{a}_1'\mathbf{a}_1 = 1$.

Se puede observar que la incógnita es precisamente $\mathbf{a}_1$ (el vector desconocido que nos da la combinación lineal óptima).

Análisis de Componentes Principales

Así, construyo la función L:

$$L(\mathbf{a}_1) = \mathbf{a}_1' \Sigma \mathbf{a}_1 - \lambda(\mathbf{a}_1' \mathbf{a}_1 - 1)$$

y busco el máximo, derivando e igualando a 0:

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{a}_1} &= 2\Sigma \mathbf{a}_1 - 2\lambda I \mathbf{a}_1 = 0 \implies \\ (\Sigma - \lambda I)\, \mathbf{a}_1 &= 0. \end{aligned}$$

Esto es, en realidad, un sistema lineal de ecuaciones. Por el teorema de Roché-Frobenius, para que el sistema tenga una solución distinta de 0 la matriz $(\Sigma - \lambda I)$ tiene que ser singular. Esto implica que el determinante debe ser igual a cero:

$$|\Sigma - \lambda I| = 0$$

Análisis de Componentes Principales

y de este modo, λ es un autovalor de Σ. La matriz de covarianzas Σ es de orden p y si además es definida positiva, tendrá p autovalores distintos, $\lambda_1, \lambda_2, \ldots, \lambda_p$ tales que, por ejemplo, $\lambda_1 > \lambda_2 > \cdots > \lambda_p$.

Se tiene que, desarrollando la expresión anterior,

$$\begin{aligned} (\Sigma - \lambda I)\,\mathbf{a}_1 &= 0 \\ \Sigma\mathbf{a}_1 - \lambda I\mathbf{a}_1 &= 0 \\ \Sigma\mathbf{a}_1 &= \lambda I\mathbf{a}_1 \end{aligned}$$

entonces,

$$\begin{aligned} Var(y_1) &= Var(\mathbf{a}_1'\mathbf{x}) = \mathbf{a}_1'\Sigma\mathbf{a}_1 = \mathbf{a}_1'\lambda I\mathbf{a}_1 = \\ &= \lambda\mathbf{a}_1'\mathbf{a}_1 = \lambda \cdot 1 = \lambda. \end{aligned}$$

Análisis de Componentes Principales

Luego, para maximizar la varianza de y_1 se tiene que tomar el mayor autovalor, digamos λ_1, y el correspondiente autovector $\mathbf{a}_1$.

En realidad, $\mathbf{a}_1$ es un vector que nos da la combinación de las variables originales que tiene mayor varianza, esto es, si $\mathbf{a}_1' = (a_{11}, a_{12}, \ldots, a_{1p})$, entonces

$$y_1 = \mathbf{a}_1'\mathbf{x} = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p.$$

El segundo componente principal, digamos $y_2 = \mathbf{a}_2'\mathbf{x}$, se obtiene mediante un argumento parecido. Además, se quiere que y_2 esté incorrelado con el anterior componente y_1, es decir, $Cov(y_2, y_1) = 0$. Por lo tanto:

$$\begin{aligned} Cov(y_2, y_1) &= Cov(\mathbf{a}_2'\mathbf{x}, \mathbf{a}_1'\mathbf{x}) = \\ &= \mathbf{a}_2' \cdot E\left[(\mathbf{x} - \mu)(\mathbf{x} - \mu)'\right] \cdot \mathbf{a}_1 = \\ &= \mathbf{a}_2'\Sigma\mathbf{a}_1, \end{aligned}$$

es decir, se requiere que $\mathbf{a}_2'\Sigma\mathbf{a}_1 = 0$.

Análisis de Componentes Principales

Como se tenía que $\Sigma \mathbf{a}_1 = \lambda \mathbf{a}_1$, lo anterior es equivalente a

$$\mathbf{a}_2' \Sigma \mathbf{a}_1 = \mathbf{a}_2' \lambda \mathbf{a}_1 = \lambda \mathbf{a}_2' \mathbf{a}_1 = 0,$$

esto equivale a que $\mathbf{a}_2' \mathbf{a}_1 = 0$, es decir, que los vectores sean ortogonales.

De este modo, tendremos que maximizar la varianza de y_2, es decir, $\mathbf{a}_2 \Sigma \mathbf{a}_2$, sujeta a las siguientes restricciones

$$\begin{aligned} \mathbf{a}_2' \mathbf{a}_2 &= 1, \\ \mathbf{a}_2' \mathbf{a}_1 &= 0. \end{aligned}$$

Se toma la función:

$$L(\mathbf{a}_2) = \mathbf{a}_2' \Sigma \mathbf{a}_2 - \lambda(\mathbf{a}_2' \mathbf{a}_2 - 1) - \delta \mathbf{a}_2' \mathbf{a}_1$$

y se deriva:

$$\frac{\partial L(\mathbf{a}_2)}{\partial \mathbf{a}_2} = 2\Sigma \mathbf{a}_2 - 2\lambda \mathbf{a}_2 - \delta \mathbf{a}_1 = 0$$

Análisis de Componentes Principales

si se multiplica por $\mathbf{a}_1'$, entonces

$$2\mathbf{a}_1'\Sigma\mathbf{a}_2 - \delta = 0$$

porque

$$\begin{aligned}\mathbf{a}_1'\mathbf{a}_2 &= \mathbf{a}_2'\mathbf{a}_1 = 0\\ \mathbf{a}_1'\mathbf{a}_1 &= 1.\end{aligned}$$

Luego

$$\delta = 2\mathbf{a}_1'\Sigma\mathbf{a}_2 = 2\mathbf{a}_2'\Sigma\mathbf{a}_1 = 0,$$

ya que $Cov(y_2, y_1) = 0$.

De este modo, $\frac{\partial L(\mathbf{a}_2)}{\partial \mathbf{a}_2}$ queda finalmente como:

$$\begin{aligned}\frac{\partial L(\mathbf{a}_2)}{\partial \mathbf{a}_2} &= 2\Sigma\mathbf{a}_2 - 2\lambda\mathbf{a}_2 - \delta\mathbf{a}_1 = 2\Sigma\mathbf{a}_2 - 2\lambda\mathbf{a}_2 =\\ (\Sigma - \lambda I)\,\mathbf{a}_2 &= 0\end{aligned}$$

Análisis de Componentes Principales

Usando los mismos razonamientos que antes, elegimos λ como el segundo mayor autovalor de la matriz Σ con su autovector asociado $\mathbf{a}_2$.

Los razonamientos anteriores se pueden extender, de modo que al j-ésimo componente le correspondería el j-ésimo autovalor.

Entonces todos los componentes $\mathbf{y}$ (en total p) se pueden expresar como el producto de una matriz formada por los autovectores, multiplicada por el vector $\mathbf{x}$ que contiene las variables originales $x_1, \ldots, x_p$

$$\mathbf{y} = A\mathbf{x}$$

donde

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}, \quad A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$$

Análisis de Componentes Principales

Como

$$Var(y_1) = \lambda_1$$

$$Var(y_2) = \lambda_2$$

$$\dots$$

$$Var(y_p) = \lambda_p$$

la matriz de covarianzas de $\mathbf{y}$ será

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \lambda_p \end{pmatrix}$$

porque $y_1, \dots, y_p$ se han construido como variables incorreladas.

Análisis de Componentes Principales

Se tiene que

$$\Lambda = Var(Y) = A'Var(X)A = A'\Sigma A$$

o bien

$$\Sigma = A\Lambda A'$$

ya que A es una matriz ortogonal (porque $\mathbf{a}_i'\mathbf{a}_i = 1$ para todas sus columnas) por lo que $AA' = I$.

Análisis de Componentes Principales

Porcentajes de variablilidad

Vimos antes que, en realidad, cada autovalor correspondía a la varianza del componente y_i que se definía por medio del autovector $\mathbf{a}_i$, es decir, $Var(y_i) = \lambda_i$.

Si sumamos todos los autovalores, tendremos la varianza total de los componentes, es decir:

$$\sum_{i=1}^{p} Var(y_i) = \sum_{i=1}^{p} \lambda_i = traza\,(\Lambda)$$

ya que la matriz Λ es diagonal.

Pero, por las propiedades del operador traza,

$$traza(\Lambda) = traza(A'\Sigma A) = traza(\Sigma A' A) = traza(\Sigma),$$

porque $AA' = I$ al ser A ortogonal, con lo cual

$$traza(\Lambda) = traza(\Sigma) = \sum_{i=1}^{p} Var(x_i)$$

Análisis de Componentes Principales

Es decir, la suma de las varianzas de las variables originales y la suma de las varianzas de las componentes son iguales. Esto permite hablar del porcentaje de varianza total que recoge un componente principal:

$$\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} = \frac{\lambda_i}{\sum_{i=1}^{p} Var(x_i)}$$

(si multiplicamos por 100 tendremos el %).

Así, también se podrá expresar el porcentaje de variabilidad recogido por los primeros m componentes:

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} Var(x_i)}$$

donde $m < p$.

Análisis de Componentes Principales

En la práctica, al tener en principio p variables, nos quedaremos con un número mucho menor de componentes que recoja un porcentaje amplio de la variabilidad total $\sum_{i=1}^{p} Var(x_i)$. En general, no se suele coger más de tres componentes principales, a ser posible, para poder representarlos posteriormente en las gráficas.

Análisis de Componentes Principales

Cálculo de los componentes principales a partir de la matriz de correlaciones

Habitualmente, se calculan los componentes sobre variables originales estandarizadas, es decir, variables con media 0 y varianza 1. Esto equivale a tomar los componentes principales, no de la matriz de covarianzas sino de la matriz de correlaciones (en las variables estandarizadas coinciden las covarianzas y las correlaciones).

Así, los componentes son autovectores de la matriz de correlaciones y son distintos de los de la matriz de covarianzas. Si se actúa así, se da igual *importancia* a todas las variables originales.

Análisis de Componentes Principales

En la matriz de correlaciones todos los elementos de la diagonal son iguales a 1. Si las variables originales están tipificadas, esto implica que su matriz de covarianzas es igual a la de correlaciones, con lo que la variabilidad total (la traza) es igual al número total de variables que hay en la muestra. La suma total de todos los autovalores es p y la proporción de varianza recogida por el autovector j-ésimo (componente) es

$$\frac{\lambda_j}{p}.$$