COMPONENTES PRINCIPALES

Análisis en componentes principales es una técnica estadística que tiene por principal objetivo el de lograr reducir la dimensionalidad de los datos cuando nos encontramos frente a datos multivariados.

Más explícitamente, si tenemos un vector aleatorio que observa p variables que llamamos $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$ intentamos construir r variables que llamaremos $Z_1, Z_2, ..., Z_r$ con $r < p$ de modo que perdamos la menor cantidad de información posible al pasar de p variables a r variables.

A las $Z_1, Z_2, ..., Z_r$ les llamaremos las r componentes principales.

Le llamamos $\sum$ a la matriz de varianzas y covarianzas de X y supondremos que X es un vector centrado, es decir que $\mathbb{E}(X_1) = \mathbb{E}(X_2) = ... = \mathbb{E}(X_p) = 0$ (esto se debe a que algunos de los resultados teóricos que se utilizan en la teoría están basadas en subespacios vectoriales por lo que es necesario que los vectores estén centrados).

Es conveniente tener en cuenta las siguientes propiedades de toda matriz de varianzas y covarianzas.

1. $\sum$ es una matriz simétrica por lo que (teorema mediante) todos sus valores propios son reales.

2. Cuando una matriz es simétrica, los vectores propios asociados a valores propios distintos son ortogonales (es decir que si A es simétrica $p \times p$ y $Au = \alpha u$, $Av = \beta v$ con $\alpha \neq \beta$ y u, v no nulos, entonces se cumple que $u \perp v$ o lo que es lo mismo $u^T v = 0$, o bien $\langle u, v \rangle = u_1 v_1 + u_2 v_2 + ... + u_p v_p = 0$).

3. $\sum$ es definida positiva o semidefinida positiva (teorema mediante), lo que significa que todos sus valores propios son ≥ 0.Será importante ordenar los valores propios de la matriz $\sum$ por lo que les llamaremos $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$.

4. $tr(\sum) = \mathbb{V}(X_1) + \mathbb{V}(X_2) + ... + \mathbb{V}(X_p) = \lambda_1 + \lambda_2 + ... + \lambda_p$ (la traza de una matriz es por definición la suma de los valores de la diagonal, la propiedad dice que la misma coincide con la suma de los valores propios de $\sum$).

El criterio para construir las variables Z_i se realiza en pasos (primero Z_1, luego Z_2 y así sucesivamente).

Paso 1. Construcción de la primer componente principal.

$Z_1 = a_{11} X_1 + a_{12} X_2 + ... + a_{1p} X_p$ donde el vector $a_1 = (a_{11}, a_{12}, ..., a_{1p})$ se elige de forma que

$$a_1 = \arg \max_{\|a_1\|=1} \mathbb{V}(Z_1)$$

Por aplicación directa del teorema de la esfera unidad llegamos a que a_1 debe ser vector propio de la matriz $\sum$ (matriz de varianzas y covarianzas del vector X) de norma 1 correspondiente al mayor valor propio de $\sum$. Además, el teorema de la esfera unidad nos dice que esa máxima varianza es igual a λ_1.

Paso 2. Construcción de la segunda componente principal.

$Z_2 = a_{21}X_1 + a_{21}X_2 + ... + a_{2p}X_p$ donde el vector $a_2 = (a_{21}, a_{22}, ..., a_{2p})$ se elige de forma que

$$a_2 = \arg \max_{\|a_2\|=1,\ a_2 \perp a_1} \mathbb{V}(Z_2).$$

Nuevamente, aplicando el teorema de la esfera unidad restringida al complemento ortogonal del subespacio propio asociado al valor propio λ_1, se llega a que a_2 es vector propio de $\sum$ asociado al siguiente valor propio más grande (le llamamos λ_2). Además la varianza máxima buscada es λ_2.

Paso 3. Construcción de la tercer componente principal.

$Z_3 = a_{31}X_1 + a_{32}X_2 + ... + a_{3p}X_p$ donde el vector $a_3 = (a_{31}, a_{32}, ..., a_{3p})$ se elige de forma que

$$a_3 = \arg \max_{\|a_3\|=1,\ a_3 \perp a_1,\ a_3 \perp a_2} \mathbb{V}(Z_3)$$

a_3 es vector propio de $\sum$ asociado al siguiente valor propio más grande (le llamamos λ_3) (por aplicación del teorema de la esfera unidad restringida al complemento ortogonal de la suma entre subespacio propio asociado al valor propio λ_1 y el subespacio propio asociado a λ_2. Además la varianza máxima es igual a λ_3.

Pasos 4 a r se siguen con la misma idea.

Sintetizando.

Primer componente principal es $Z_1 = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p$ siendo $a_1 = (a_{11}, a_{12}, ..., a_{1p})$ vector propio de $\sum$ asociado al mayor valor propio λ_1. Además $\mathbb{V}(Z_1) = \lambda_1$.

Segunda componente principal es $Z_2 = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p$ siendo $a_2 = (a_{21}, a_{22}, ..., a_{2p})$ vector propio de $\sum$ asociado al segundo mayor valor propio λ_2. Además $\mathbb{V}(Z_2) = \lambda_2$.

Así sucesivamente con $Z_3, Z_4, ..., Z_r$.

Observaciones y comentarios.

1. Las componentes principales son centradas e incorrelacionadas (ver punto siguiente) entre sí.

2. La condición $a_1 \perp a_2$ equivale a la condición $\mathbb{COV}(Z_1, Z_2) = 0$. Esto se debe a que (abreviando la notación) si $Z_1 = a_1X_1 + a_2X_2 + ... + a_pX_p$ y $Z_2 = b_1X_1 + b_2X_2 + ... + b_pX_p$ siendo a vector propio de $\sum$ con valor

propio $\lambda > 0$ y le llamamos σ_{ij} al elemento i, j de la matriz $\sum$ (covarianza entre X_i y X_j) entonces

$$\mathbb{COV}(Z_1, Z_2) = \mathbb{COV}(a_1X_1 + a_2X_2 + ... + a_pX_p, b_1X_1 + b_2X_2 + ... + b_pX_p) =$$

$$\sum_{i,j=1}^{p} a_i b_j \sigma_{ij} = b^T \sum a = b^T \lambda a = \lambda b^T a.$$

De la igualdad anterior se deduce que $\mathbb{COV}(Z_1, Z_2) = 0$ si y sólo si $a \perp b$.

3. La restricción para maximizar las varianzas de $\|a\| = 1$ es necesaria para homogeneizar y además si no acotamos los valores de los coeficientes podemos lograr varianzas tan grandes como se desee.

4. Los coeficientes que definen las componentes principales no son únicos. Por ejemplo, dados los coeficientes que definen la primer componente principal $a_1 = (a_{11}, a_{12}, ..., a_{1p})$, si les cambiamos de signo a los mismos, obtenemos la misma varianza para Z_1 y siguen teniendo norma 1. Incluso podría haber infinitos coeficientes que alcancen el máximo.

5. Si definimos la matriz $A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rp} \end{pmatrix}$ incluyendo los coeficientes obtenidos que definen las componentes principales, se tiene que las filas de esta matriz tienen norma 1 y además son ortogonales entre sí. Estas condiciones equivalen a la igualdad $A^T A = I$.

6. Las igualdades que definen las componentes principales pueden ser escritas en forma matricial como $Z = AX$ siendo $X = (X_1, X_2, ..., X_p)^T$ y $Z = (Z_1, Z_2, ..., Z_r)^T$.

7. En el caso en que consideremos $r = p$ componentes principales, entonces la matriz A queda cuadrada y en ese caso, las filas tienen norma 1 y son ortogonales entre sí, pero también las columnas tienen norma 1 y son ortogonales entre sí. Eso equivale a que A es invertible y además $A^{-1} = A^T$ y por lo tanto $A^T A = AA^T = I$.

8. En el caso $r = p$, dado que A es invertible, se tiene que la igualdad $Z = AX$ equivale a $X = A^T Z$, lo que implica que, por un lado no reducimos la dimensionalidad ni perdemos información (porque dada Z podemos recuperar X) pero pasamos a variables incorrelacionadas y por otro lado tenemos una fórmula sencilla para recuperar las X_i ya que $X_i = \sum_{j=1}^{p} a_{ji} Z_j$.

Elección de un valor adecuado de r.

Cuando realmente necesitamos reducir dimensionalidad y trabajar con un $r < p$, el método más utilizado es el llamado método del "codo" que consiste en

considerar la proporción de variabilidad de las r componentes principales con respecto a la variabilidad total que es $tr\left(\sum\right) = \mathbb{V}\left(X_1\right) + \mathbb{V}\left(X_2\right) + ... + \mathbb{V}\left(X_p\right)$.

Es decir que consideramos los cocientes

$$\frac{\mathbb{V}\left(Z_1\right) + \mathbb{V}\left(Z_2\right) + ... + \mathbb{V}\left(Z_r\right)}{\mathbb{V}\left(X_1\right) + \mathbb{V}\left(X_2\right) + ... + \mathbb{V}\left(X_p\right)} = \frac{\lambda_1 + \lambda_2 + ... + \lambda_r}{\lambda_1 + \lambda_2 + ... + \lambda_p}.$$

Si graficamos esta proporción en función de r tendremos una función que irá creciendo tomando valores entre 0 y 1. En algún momento suele visualizarse que a partir de determinado valor de r esta proporción queda cerca de 1 y crece demasiado lentamente (es donde se produce el "codo", elbow en inglés, en el gráfico). Si tomamos un valor de r que cumpla con esas condiciones, tendremos un valor de r (que muchas veces es notoriamente menor que p) y con el cual la proporción de la variabilidad explicada por las r componentes principales es aproximadamente el 100% de la variabilidad total. De esta forma muchas veces logramos reducir claramente la dimensionalidad de los datos perdiendo muy poca información.

Recordar que para aplicar este criterio es necesario que o bien todas las variables X_i sea comparables entre sí en magnitudes, de lo contrario es conveniente estandarizar los datos ya que si hay mucha variación en el valor de las varianzas, las componentes principales estarán gobernadas por aquellas variables que tienen las mayores varianzas.

FUNDAMENTO TEÓRICO

Teorema de la esfera unidad.

Si A es una matriz real simétrica, y le llamamos $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ a sus valores propios y le llamamos u_1 a un vector propio de A asociado a λ_1 tal que $\|u_1\| = 1$, entonces

$$\max_{\|x\|=1} x^T A x = \lambda_1 \text{ y además } \arg\max_{\|x\|=1} x^T A x = u_1.$$

Dem.

Por teorema espectral sabemos que A es diagonalizable en una base ortonormal, lo que equivale a decir que existe una matriz ortogonal B ($B^T = B^{-1}$) tal que $B^T A B = D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}$. Entonces $A = BDB^T$ por lo que

$$x^T A x = x^T B D B^T x =$$

$$\left(B^T x\right) D \left(B^T x\right) \underset{y = B^T x}{=} y^T D y = \lambda_1 y_1^2 + \lambda_2 y_2^2 + ... + \lambda_p y_p^2 \leq \lambda_1 \left(y_1^2 + y_2^2 + ... + y_p^2\right).$$

Teniendo en cuenta que la transformación $y = B^T x$ es biyectiva y que la condición $\|x\| = 1$ se transforma en $\|y\| = 1$ (esto se debe a que $\|y\|^2 = \left\|B^T x\right\|^2 = \left(B^T x\right)^T B^T x = x^T B B^T x = x^T x = \|x\|$), se tiene que $\max_{\|x\|=1} x^T A x = \max_{\|y\|=1} y^T D y$. Pero si tomamos y tal que $\|y\| = 1$, tenemos que $y^T D y = \lambda_1 y_1^2 + \lambda_2 y_2^2 + ... + \lambda_p y_p^2 \leq \lambda_1 \left(y_1^2 + y_2^2 + ... + y_p^2\right) = \lambda_1$. Entonces $\max_{\|x\|=1} x^T A x \leq \lambda_1$. Para terminar la demostración veremos que la cota superior λ_1 se realiza en u_1. Efectivamente, como u_1 es vector propio (de norma 1) de A asociado al valor propio λ_1, se tiene que $A u_1 = \lambda_1 u_1$ entonces $x^T A x = u_1^T \lambda_1 u_1 = \lambda_1 u_1^T u_1 = \lambda_1 \|u_1\| = \lambda_1$. Esto termina la demostración porque llegamos a que $\max_{\|x\|=1} x^T A x = \lambda_1$ y además el máximo se alcanza en u_1.

Notas.

1. Si bien no nos interesa (dentro del tema de componentes principales), en el teorema de la esfera unidad se prueba (con demostración completamente análoga, basta cambiar las desigualdades) que $\min_{\|x\|=1} x^T A x = \lambda_p$ y además $\arg\min_{\|x\|=1} x^T A x = u_p$ siendo u_p vector propio de A asociado al valor propio λ_p.

2. El teorema de la esfera unidad justifica el paso 1 en la metodología para la obtención de las componentes principales, es decir que la combinación lineal de las X_i con varianza máxima sobre los coeficientes que tienen norma 1, es el mayor valor propio de la matriz $\sum$ y además, este máximo se encuentra en un vector propio asociado de norma 1.

3. De la demostración del teorema surge que en caso que la dimensión del subespacio propio asociado al vector propio λ_1 tuviera dimensión mayor a 1 (es decir que existen dos o más vectores propios linealmente independientes asociados a λ_1) el método no sugiere tomar ninguno en particular, de todas formas esto no ocurre en la práctica (salvo casos muy particulares).

JUSTIFICACIÓN DEL PASO 2 (los demás pasos se llevan a cabo análogamente).

Para ello adaptamos el teorema de la esfera unidad restringido al complemento ortogonal del subespacio de vectores propios asociados al valor propio λ_1.

Probaremos el siguiente resultado.

Teorema (de la esfera unidad restringido al subespacio ortogonal al subespacio propio generado por u_1.)

Si A es una matriz real simétrica, y le llamamos $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ a sus valores propios y le llamamos u_1 a un vector propio de A asociado a λ_1 tal que $\|u_1\| = 1$ y u_2 a un vector propio de A asociado a λ_2 tal que $\|u_2\| = 1$, entonces

$$\max_{\|x\|=1,\ x \perp u_1} x^T A x = \lambda_2 \text{ y además } \arg \max_{\|x\|=1,\ x \perp u_1} x^T A x = u_2.$$

Dem.

Como en la demostración del teorema de la esfera unidad, hacemos el cambio de variable $y = B^T x$ obteniendo que $x^T Ax = y^T Dy$. Ya vimos que la condición $\|x\| = 1$ se transforma en $\|y\| = 1$.

Por otro lado, veremos que la condición $x \perp u_1$ se transforma en $y_1 = 0$. Efectivamente, sabemos que en las columnas de B van vectores propios asociados a los valores propios $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ y además tienen norma 1 (porque la matriz es ortogonal). Escribimos entonces $B = \begin{pmatrix} \vdots & \vdots & \cdots & \vdots \\ u_1 & u_2 & \cdots & u_p \\ \vdots & \vdots & \cdots & \vdots \end{pmatrix}$ por lo que la condición $x \perp u_1$ equivale a $By \perp u_1$ o sea $(By)^T u_1 = 0$ lo que equivale a $y^T B^T u_1 = 0$, pero $B^T u_1 = \begin{pmatrix} \cdots & u_1^T & \cdots \\ \cdots & u_2^T & \vdots \\ \vdots & & \vdots \\ \cdots & u_p^T & \cdots \end{pmatrix} \begin{pmatrix} \vdots \\ u_1 \\ \vdots \end{pmatrix} = \begin{pmatrix} u_1^T u_1 \\ u_2^T u_1 \\ \vdots \\ u_p^T u_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ entonces la condición $y \perp B^T u_1$ son aquellos y tales que $y_1 = 0$.

Entonces

$$\max_{\|x\|=1,\ x \perp u_1} x^T Ax = \max_{\|y\|=1,\ y_1=0} y^T Dy =$$

$$\max_{\|y\|=1,\ y_1=0} \left(\lambda_2 y_2^2 + ... + \lambda_p y_p^2\right) \leq \lambda_2 \left(\left(\lambda_2 y_2^2 + ... + \lambda_p y_p^2\right)\right) = \lambda_2.$$

Además el máximo se realiza en u_2 ya que el vector u_2 (segunda columna de B) verifica las condiciones $\|x\| = 1$, $x \perp u_1$ y además $u_2^T Au_2 = u_2^T \lambda_2 u_2 = \lambda_2 u_2^T u_2 = \lambda_2$, lo que concluye la prueba del teorema.

Observación.

La demostración nos permite dar cuenta que podría darse el caso en que $\lambda_2 = \lambda_1$, en este caso necesariamente existirá otro vector propio u_2 asociado al mismo valor propio λ_1 que es ortogonal con u_1.