

Reducción de la Dimensión

Dado un grupo de individuos, caracterizados por n variables aleatorias, resulta interesante, con vistas a analizar como se comporta cada individuo, respecto a esas variables, resumir la información que estas variables aportan, en otro grupo menor de variables. Este método también se conoce como Análisis de Componentes Principales, porque permite explicar la variabilidad del de las variables observadas, a través de otro grupo de variables que son combinaciones lineales de las originales que recogen la mayor variabilidad posible.

I.- Planteo del problema.

Supongamos k individuos I_i ($1 \leq i \leq k$), representados por los puntos P_i de R^n cuyas coordenadas son las realizaciones $x_i = (x_{i1}, x_{i2}, \dots, x_{in})$ de n variables aleatorias (X_1, X_2, \dots, X_n). El objetivo es construir una variedad lineal de dimensión $q < n$ que mejor se ajuste a la nube de puntos P_i . Podemos considerar que esta variedad lineal, será tal que que la suma de los cuadrados de las distancias de los puntos P_i a la variedad sea mínima.

Si su dimensión es $q < n$, la ecuación de la variedad afín es:

$$y = t + \beta_1 v_1 + \beta_2 v_2 + \dots + \beta_q v_q \quad (1)$$

donde v_1, v_2, \dots, v_q es una base del subespacio director asociado a la variedad y que lo elegimos ortonormal, es decir:

$$\langle v_i, v_j \rangle = \delta_{ij} \quad 1 \leq i, j \leq q \quad (2)$$

siendo el producto escalar, el producto escalar ordinario, es decir el definido por la matriz identidad.

Los vectores:

$$z_i = x_i - t \quad 1 \leq i, j \leq q \quad (3)$$

pueden descomponerse de forma única como:

$$z_i = p_i + p^i \quad 1 \leq i, j \leq q \quad (4)$$

donde p_i es un elemento del subespacio director de la variedad y p^i es un vector ortogonal a dicho subespacio.

La proyección de z_i en la variedad es:

$$p_i = \sum_{j=1}^q \Pi_j v_j = \sum_{j=1}^q \langle z_i, v_j \rangle v_j = \sum_{j=1}^q \langle x_i - t, v_j \rangle v_j \quad (5)$$

La distancia al cuadrado del punto P_i a la variedad es $\|p^i\|^2$ y su valor es:

$$\|p^i\|^2 = \|z_i\|^2 - \|p_i\|^2 \quad (6)$$

siendo

$$\begin{aligned} \|p_i\|^2 &= \langle p_i, p_i \rangle = \langle \sum_{j=1}^q \langle x_i - t, v_j \rangle v_j, \sum_{j=1}^q \langle x_i - t, v_j \rangle v_j \rangle = \\ &= \sum_{j=1}^q \langle x_i - t, v_j \rangle^2 \end{aligned} \quad (7)$$

Para lograr el objetivo propuesto debemos minimizar la función

$$\begin{aligned}\phi(t, v) &= \sum_{j=1}^k \|p^j\|^2 = \sum_{j=1}^k (\|z_j\|^2 - \|p_j\|^2) = \\ &= \sum_{i=1}^k \langle x_i - t, x_i - t \rangle - \sum_{j=1}^q \langle x_i - t, v_j \rangle^2\end{aligned}\quad (8)$$

con la condición

$$\langle v_i, v_j \rangle = \delta_{ij} \quad 1 \leq i, j \leq q \quad (9)$$

Si desarrollamos (8) obtenemos:

$$\phi(t, v) = \sum_{j=1}^k \left(\sum_{h,j=1}^n (x_{ij} - t_j)(x_{ij} - t_h) \right) - \sum_{j=1}^q \left(\sum_{h,j=1}^n (x_{ih} - t_h)v_{jk} \right)^2 \quad (10)$$

Matricialmente esta expresión la podemos poner como:

$$\begin{aligned}\phi(t, v) &= \text{traza}(X - \mathbf{1}')(X - \mathbf{1}')' - \sum_{j=1}^q \left(\sum_{h,j=1}^n (x_{ih} - t_h)v_{jk} \right)^2 = \\ &= \text{tr}(X - \mathbf{1}t')(X - \mathbf{1}t')' - \sum_{j=1}^q v_j'(X - \mathbf{1}t')'(X - \mathbf{1}t')v_j\end{aligned}\quad (11)$$

siendo

$$X = (x_{ij}) \quad 1 \leq i \leq k, \quad 1 \leq j \leq n, \quad \mathbf{1} = (1, \dots, 1)'$$

II.- Resolución del problema.

En primer lugar veremos que vector t y después que vectores v_1, v_2, \dots, v_q hacen mínima la función $\phi(t, v)$.

a) El vector

$$t = \bar{x} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n)' \quad \bar{x}_1 = \frac{1}{k} \sum_{h=1}^k x_{hi} \quad (12)$$

minimiza la función $\phi(t, v)$.

En efecto, cualquier vector t lo podemos expresar como $t = \bar{x} + a$, entonces la expresión (8) queda como:

$$\begin{aligned}\phi(t, v) &= \sum_{i=1}^k (\langle x_i - \bar{x} - a, x_i - \bar{x} - a \rangle - \sum_{j=1}^q \langle x_i - \bar{x} - a, v_j \rangle^2) = \\ &= \sum_{i=1}^k \|x_i - \bar{x}\|^2 - 2 \sum_{i=1}^k \langle x_i - \bar{x}, a \rangle + k\|a\|^2 - \sum_{i=1}^k \sum_{j=1}^q \langle x_i - \bar{x} - a, v_j \rangle^2\end{aligned}\quad (13)$$

Al ser

$$\bar{x} = \frac{1}{k} \sum_{ih=1}^k x_i$$

se tiene que

$$\phi(t, v) = \sum_{i=1}^k \|x_i - \bar{x}\|^2 + k\|a\|^2 - \sum_{j=1}^q \left(\sum_{i=1}^k \langle x_i, v_j \rangle^2 + k \langle a, v_j \rangle^2 - k \langle \bar{x}, v_j \rangle^2 \right) \quad (14)$$

La proyección de a sobre la variedad es:

$$\sum_{j=1}^q \langle a, v_j \rangle v_j \quad (15)$$

y su norma al cuadrado es:

$$\sum_{j=1}^q \langle a, v_j \rangle^2 \leq \|a\|^2 \quad (16)$$

siendo igual a la norma de a al cuadrado cuando $a = t - \bar{x}$, pertenezca al subespacio director asociado a la variedad lineal.

Por todo ello la expresión (14) será mínima cuando $a = 0$, ya que para este valor el segundo sumando de (14) es mínimo y el vector nulo pertenece al subespacio director.

El vector t será pues $t = \bar{x}$, por lo tanto la expresión de $\phi(t, v)$ mínima será $\phi(\bar{x}, v)$, o sea:

$$\begin{aligned} \phi(t, v) &= \text{tr} [(X - \mathbf{1}\bar{x}')(X - \mathbf{1}\bar{x}')'] - \sum_{j=1}^q v_j'(X - \mathbf{1}\bar{x}')'(X - \mathbf{1}\bar{x}')v_j = \\ &= \sum_{i=1}^k \|x_i - \bar{x}\|^2 - \sum_{j=1}^q v_j'(X - \mathbf{1}\bar{x}')'(X - \mathbf{1}\bar{x}')v_j \end{aligned} \quad (17)$$

b) A continuación pasaremos a calcular los vectores v_j , ($1 \leq j \leq q$) que minimizan (17) con las restricciones (9).

Estos vectores v_j son aquellos que maximizan

$$B = \sum_{j=1}^q v_j' R v_j \quad (18)$$

con las restricciones

$$\langle v_i, v_j \rangle = \delta_{ij} \quad 1 \leq i, j \leq q \quad (19)$$

siendo R

$$R = (X - \mathbf{1}\bar{x}')'(X - \mathbf{1}\bar{x}') \quad (20)$$

una matriz definida positiva.

Vamos a maximizar cada uno de los sumandos de (18), es decir vamos a encontrar los vectores que maximizan

$$f(v) = v' R v. \quad (21)$$

con la condición

$$\langle v, v \rangle = 1.$$

Sean w_1, w_2, \dots, w_n , vectores tales que

$$Rw_i = \lambda_i w_i. \quad 1 \leq i \leq n \quad (22)$$

con $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ y sujetos a la condición $\langle w_i, w_j \rangle = \delta_{ij}$ es decir, w_1, w_2, \dots, w_n es una base de vectores propios ortonormales. Entonces si

$$v = \sum_{i=1}^n \alpha_i w_i \quad (23)$$

se tiene

$$f(v) = \sum_{i,j=1}^n \alpha_i \alpha_j w_i' R w_j \quad 1 \leq i, j \leq n \quad (24)$$

Por (22) y como $\langle w_i, w_j \rangle = \delta_{ij}$, se tiene que:

$$f(v) = \sum_{i=1}^n \alpha_i^2 \lambda_i \leq \lambda_1 \sum_{i=1}^n \alpha_i^2 \quad (25)$$

y por la condición () se tiene que

$$\langle v, v \rangle = \sum_{i,j=1}^n \alpha_i \alpha_j w_i' R w_j = \sum_{i=1}^n \alpha_i^2 = 1 \quad (26)$$

por consiguiente $f(v) \leq \lambda_1$. Es decir, la función (24) está acotada por λ_1 . Como

$$f(w_1) = w_1' R w_1 = \lambda_1 w_1' w_1 = \lambda_1 \quad (27)$$

se tiene que el vector w_1 hace máximo el primer sumando de (18). Los vectores que hacen máximo el resto de los sumandos, son los vectores propios correspondientes a los $q - 1$ siguientes valores propios $\lambda_2, \dots, \lambda_q$ dados en (22).

En el caso de que algún valor propio de (22) sea múltiple se eligen tantos vectores del subespacio propio correspondiente, como orden de multiplicidad tenga el valor propio y que sean ortogonales con el producto escalar euclídeo usual. Por otra parte $Rw_i = \lambda_i w_i$.

En consecuencia la variedad lineal tal que la suma de los cuadrados de las distancias de los puntos P_i a dicha variedad es mínima viene definida por:

$$y = \bar{x} + \beta_1 w_1 + \dots + \beta_q w_q \quad (28)$$

donde \bar{x} es el vector de medias general y donde w_i , $1 \leq i \leq q$ son los vectores propios de R , correspondientes a los q primeros valores propios de R , correspondientes a los q primeros valores propios en orden decreciente en el caso de ser distintos y ortogonales. Si la multiplicidad de un valor propio es r entonces se eligen r vectores propios ortonormales del subespacio propio correspondiente.

III.- Propiedad Básica.

La suma de los cuadrados de las interdistancias de las proyecciones de los puntos P_i sobre la variedad (28) es máxima.

Sea F la variedad lineal q -dimensional (28), las proyecciones en F de los puntos P_i y P_j de coordenadas $x_i = (x_{i1}, \dots, x_{in})$ y $x_j = (x_{j1}, \dots, x_{jn})$, vienen dadas por:

$$\begin{aligned} p_i &= \alpha_{i1}, \dots, \alpha_{in} \\ p_j &= \alpha_{j1}, \dots, \alpha_{jn} \end{aligned} \quad (29)$$

siendo

$$\begin{aligned} \alpha_{ik} &= \langle x_i, w_k \rangle = x_i' w_k = w_k' x_i \\ \alpha_{jk} &= \langle x_j, w_k \rangle = x_j' w_k = w_k' x_j \end{aligned} \quad (30)$$

Por (29) y (30) tenemos que:

$$d^2(p_i, p_j) = (W'(x_i - x_j))'(W'(x_i - x_j)) = (x_i - x_j)' W W' (x_i - x_j) \quad (31)$$

donde W es la matriz que tiene por vectores columna w_1, w_2, \dots, w_q .

La suma de los cuadrados de las interdistancias en F vendrá dada por la expresión.

$$D = \sum_{i,j=1}^k d^2(p_i, p_j) = \sum_{i,j=1}^k (x_i - x_j)' W W' (x_i - x_j) \quad (32)$$

Por otra parte tenemos que:

$$\begin{aligned}
D &= \sum_{i,j=1}^k (x_i - x_j)' WW' (x_i - x_j) = \\
&= \sum_{i,j=1}^k (x_i' WW' x_i - x_i' WW' x_j - x_j' WW' x_i + x_j' WW' x_j) = \\
&= \sum_{i,j=1}^k (x_i' WW' x_i - \sum_{i,j=1}^k x_i' WW' x_j - \\
&\quad - \sum_{i,j=1}^k x_j' WW' x_i + \sum_{i,j=1}^k x_j' WW' x_j) = \\
&= k \sum_{i=1}^k (x_i' WW' x_i - k \sum_{j=1}^k \bar{x}' WW' x_j - \\
&\quad - k \sum_{i=1}^k \bar{x} WW' x_i + k \sum_{j=1}^k x_j' WW' x_j) = \\
&= 2k \sum_{i=1}^k (x_i' WW' x_i - 2k \sum_{i=1}^k \bar{x}' WW' x_i) = \\
&= 2k \sum_{i=1}^k (x_i' WW' x_i - 2k \sum_{i=1}^k \bar{x}' WW' x_i) = \\
&= 2k (\sum_{i=1}^k (x_i' WW' x_i - \sum_{i=1}^k \bar{x}' WW' x_i) + \\
&\quad + \sum_{i=1}^k (x_i' WW' \bar{x} - \sum_{i=1}^k \bar{x}' WW' \bar{x})) = \\
&= 2k (\sum_{i=1}^k [(x_i' - \bar{x}') WW' (x_i - \bar{x})]) = \\
&= 2k \operatorname{tr} [(X - \mathbf{1}\bar{x}') WW' (X - \mathbf{1}\bar{x}')] = \\
&= 2k \operatorname{tr} [(X - \mathbf{1}\bar{x}') WW' (X - \mathbf{1}\bar{x}')] = \\
&= 2k \operatorname{tr} [W' (X - \mathbf{1}\bar{x}') (X - \mathbf{1}\bar{x}')' W] = \\
&= 2k \operatorname{tr} [W' RW] = 2k (\lambda_1 w_1' w_1 + \dots + \lambda_q w_q' w_q) = \\
&= 2k (\lambda_1 + \dots + \lambda_q) =
\end{aligned} \tag{33}$$

Es decir

$$D = 2k (\lambda_1 + \dots + \lambda_q) \tag{34}$$

Como hemos visto que los sumandos de la forma $u' Ru$, con $\langle u, u \rangle = 1$, están acotados por $\lambda_1, \dots, \lambda_q$, queda demostrada la propiedad, pues cualquier otra base que no sea la de los vectores propios w_1, \dots, w_q genera una variedad en la cual la suma de los cuadrados de las interdistancias de las proyecciones de los puntos P_i son mayores.

IV.- Coordenadas Canónicas.

Las coordenadas de las medias poblacionales centradas en el nuevo subespacio de dimensión reducida q , referidas a los vectores propios de coordenadas w_i , $i = 1, \dots, q$, son:

$$Y' = W' (X - \mathbf{1}\bar{x}') \tag{35}$$

siendo Y una matriz cuyas filas son las coordenadas de los puntos P_i en la variedad final y la matriz W tiene por vectores columna las componentes de los vectores w_i , $i = 1, \dots, q$.

Martín Ríos.
Departamento de Estadística.
Facultad de Biología.
Universidad de Barcelona.