Análisis de Componentes Principales

Mathias Bourel

DMMC - Facultad de Ciencias Económicas y Administración, Universidad de la República, Uruguay
IMERL - Facultad de Ingeniería, Universidad de la República, Uruguay

8 de junio de 2016

Introducción

Suponemos que tenemos nuestra matriz de datos $X \in \mathcal{M}_{n\times p}$ **centrada**, es decir que la media de cada columna es 0. Queremos encontrar un subespacio de dimensión menor que p que represente de manera adecuada los datos. Más precisamente querremos encontrar un subespacio de dimensión menor que p tal que cuando proyectamos los individuos sobre él, la estructura se distorciona lo menos posible.

Consideremos una recta por el origen (subespacio de dimensión 1) generada por un vector $a_1 \in \mathbb{R}^p$ unitario. Si consideramos un individuo $\mathbf{x}_i$ su proyección sobre el subespacio generado por a_1 es

$$z_i a_1 = \frac{a_1'\mathbf{x}_i}{||a_1||^2} a_1 = a_1'\mathbf{x}_i a_1$$

DIBUJO

Si queremos minimizar $\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} ||\mathbf{x}_i - z_i a_1||^2 = \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{z}_i a_1)'(\mathbf{x}_i - \mathbf{z}_i a_1)$

observamos que por el teorema de Pitágoras

$$\mathbf{x}'_i\mathbf{x}_i = z_i^2 + r_i^2$$

y entonces

$$\sum_{i=1}^{n} \mathbf{x}'_i\mathbf{x}_i = \sum_{i=1}^{n} z_i^2 + \sum_{i=1}^{n} r_i^2$$

Como el término $\sum_{i=1}^{n} \mathbf{x'}_i \mathbf{x}_i$ es constante, minimizar $\sum_{i=1}^{n} r_i^2$ equivale a maximizar $\sum_{i=1}^{n} z_i^2$ que no es otra cosa que la varianza muestral **de los datos proyectados** dado que los datos son centrados. En efecto

$$\sum_{i=1}^{n} z_i = \sum_{i=1}^{n} a_1' \mathbf{x}_i = a_1' \left(\sum_{i=1}^{n} \mathbf{x}_i \right) = a_1' \overline{\mathbf{x}} = 0$$

Idea principal

Reducir el número de variables sin perder (demasiada) información.
Mayor información relacionado con mayor variabilidad.
Objetivo:

$$x_1, \ldots, x_p \text{ correladas} \to z_1, \ldots, z_l \textbf{ incorreladas}$$

donde $z_1, \ldots, z_l$ son combinaciones lineales de $x_1, \ldots, x_p$:

$$z_j = a_{j1}x_1 + a_{j2}x_2 + \cdots + a_{jp}x_p = \mathbf{X}a_j \ \ \forall j = 1, \ldots, l, \ \ l < p$$

En un primer momento:

$$x_1, \dots, x_p \text{ correladas} \to z_1, \dots, z_p \textbf{ incorreladas}$$

donde $z_1, \dots, z_p$ son combinaciones lineales de $x_1, \dots, x_p$:

$$z_j = a_{j1}x_1 + a_{j2}x_2 + \cdots + a_{jp}x_p = \mathbf{X}a_j \ \ \forall j = 1, \dots, p$$

1. Vamos a imponer que $||a_j'|| = 1 \ \ \forall j = 1, \dots, p$
2. Vamos a buscar a_1 tal que z_1 tenga la mayor varianza y $||a_1|| = 1$.
3. Vamos a buscar a_2 tal que z_2 sea incorrelada con z_1, convarianza menor que z_1 y $||a_2|| = 1$.
4. ...

Sea Σ la matriz de covarianzas de $\mathbf{X}$ (habitualmente se usa la matriz de correlaciones).

1. Como $z_1 = \mathbf{X}a_1$ entonces la ser las variables originales con media cero entonces también el vector z_1 tiene media cero y su varianza es $var(z_1) = \frac{1}{n} z_1^{'} z_1 = \frac{1}{n} a_1^{'} \mathbf{X}'\mathbf{X} a_1 = a_1^{'} \Sigma a_1$. Para maximizar $var(z_1)$ de manera que $||a_1|| = 1$:

$$L(a_1) = \overbrace{a_1' \mathbf{\Sigma} a_1}^{var(z_1)} - \lambda(a_1' a_1 - 1)$$

$$\frac{\partial L(a_1)}{\partial a_1} = 0 \Rightarrow 2\mathbf{\Sigma} a_1 - 2\lambda I a_1 = 0$$

$$\Rightarrow (\mathbf{\Sigma} - \lambda I) a_1 = 0 \Rightarrow \det(\mathbf{\Sigma} - \lambda I) = 0 \text{ para } a_1 \neq 0$$

$$\Rightarrow \lambda \text{ es valor propio de } \mathbf{\Sigma} \text{ asociado al vector propio } a_1$$

Recordar que $\mathbf{\Sigma}$ es diagonalizable en una base ortonormal pues es simétrica.

Cálculo de las componentes

Al ser la matriz de covarianzas Σ semidefinida positiva y de tamaño $p \times p$, consideramos $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ los valores propios de Σ.

$$Var(z_1) = Var(\mathbf{X}a_1) = a_1'\Sigma a_1 = a_1'\lambda_1 a_1 = \lambda_1 a_1' a_1 = \lambda_1$$

Para maximizar la varianza, tomo entonces el mayor valor propio λ_1 de Σ y el correspondiente vector propio $a_1' = (a_{11}, a_{12}, \ldots, a_{1p})'$ (normalizado) y entonces

$$z_1 = \mathbf{X}a_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p$$

es la combinación lineal de los $x_1, \ldots, x_p$ con la mayor varianza.

2 Queremos ahora encontrar $z_2 = \mathbf{X}a_2$ tal que $\begin{cases} Cov(z_2, z_1) = 0 \\ ||a_2|| = 1 \end{cases}$

$$0 = Cov(z_2, z_1) = a_2'\Sigma a_1 = a_2'\lambda_1 a_1 \Leftrightarrow a_2' a_1 = 0$$

Maximizamos entonces la varianza de z_2 de manera que $||a_2|| = 1$ y que $a_2' a_1 = 0$.

$$L(a_2) = \overbrace{a_2'\mathbf{\Sigma} a_2}^{var(z_2)} - \lambda(a_2' a_2 - 1) - \delta a_2' a_1$$

$$\frac{\partial L(a_2)}{\partial a_2} = 0 \Rightarrow 2\mathbf{\Sigma} a_2 - 2\lambda a_2 - \delta a_1 = 0$$

Multiplicando por a_1' se tiene

$$2a_1'\mathbf{\Sigma} a_2 - \delta = 0 \Rightarrow \delta = 2a_1'\mathbf{\Sigma} a_2 = 2a_2'\mathbf{\Sigma} a_1 = 0$$

Cálculo de las componentes

$$\frac{\partial L(a_2)}{\partial a_2} = 0 \Leftrightarrow 2\mathbf{\Sigma} a_2 - 2\lambda a_2 = 0 \Leftrightarrow \left(\mathbf{\Sigma} - \lambda I\right) a_2 = 0$$

Elijo entonces λ el 2do mayor valor propio de $\mathbf{\Sigma}$ con vector propio asociado a_2.

Repetimos este procedimiento p veces, obteniendo los vectores $a_1, a_2, \ldots, a_p$ y se obtiene una matriz ortogonal $A = \begin{pmatrix} a_1 & a_2 & \ldots & a_p \end{pmatrix}$

- Observar que se puede escribir (poniendo las caracteristicas en filas):

$$\begin{pmatrix} z_1' \\ z_2' \\ \vdots \\ z_p' \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \dots & a_{pp} \end{pmatrix} \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_p' \end{pmatrix}$$

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

- O si no:

$$\begin{pmatrix} z_1 & z_2 & \dots & z_p \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & \dots & x_p \end{pmatrix} \begin{pmatrix} a_{11} & a_{21} & \dots & a_{p1} \\ a_{12} & a_{22} & \dots & a_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1p} & a_{2p} & \dots & a_{pp} \end{pmatrix}$$

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

A las columnas de Z se le llaman *componentes principales* de $\mathbf{X}$.

- Como $Var(z_1) = \lambda_1$, $Var(z_2) = \lambda_2$, $\ldots$, $Var(z_p) = \lambda_p$ y son incorreladas:

$$\Sigma_Z = Var(Z) = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \lambda_p \end{pmatrix} \underbrace{=}_{Z=\mathbf{X}A} A' Var(\mathbf{X}) A.$$

Entonces:

$$\mathbf{\Sigma} = A \Sigma_Z A'$$

Porcentajes de variabilidad

$$\sum_{i=1}^{p} Var(z_i) = \sum_{i=1}^{p} \lambda_i = tr\left(\Sigma_Z\right) = tr\left(A'\Sigma_X A\right) = tr\left(\Sigma_X AA'\right) = tr\left(\Sigma_X\right)$$

Porcentaje de variabilidad de la variable i:

$$\frac{Var(z_i)}{\sum_{i=1}^{p} Var(z_i)} = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \qquad \left(\text{con matriz correlaciones } \frac{\lambda_i}{p}\right)$$

Porcentaje de variabilidad de las m primeras variables i:

$$\sum_{j=1}^{m} \frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i} \qquad \text{donde } m < p$$

Nos quedamos con un número mucho menor de componentes que recogen un porcentaje amplio de la variabilidad total (fijada por el usuario). En general no se elige más de 3.

- Cada eje de $\mathbb{R}^p$ representa una de las p variables.
- Supongamos que tenemos N individuos, y nos focalizamos en el individuo n, entonces las coordenadas de $\mathbf{x_n}'$ son los datos de las p variables para este individuo.
- $\mathbf{z_n}' = \mathbf{x_n}'\mathbf{A}$ son las coordenadas del individuo $\mathbf{x_n}'$ en el nuevo sistema de referencia determinado por las componentes principales.
- Podemos entonces pensar que “proyectamos” la nube de la población dada por $\mathbf{X}$ sobre un subespacio de dimensión la cantidad de componentes principales que retendremos.

Como

$$Cov(z_j, x_i) = Cov(z_j, \sum_{k=1}^{p} a_{ik} z_k) = a_{ij} Var(z_j) = \lambda_j a_{ij}$$

entonces la correlación es:

$$Cor(z_j, x_i) = \frac{\lambda_j a_{ij}}{\sqrt{\lambda_j}}$$

1. Se calculan las componentes principales sobre variables originales estandarizadas (media 0 y varianza 1). Tomo entonces las componentes principales sobre la matriz de correlaciones y se le da la misma importancia a todas las variables.
2. Si las variables $x_1, \ldots, x_p$ ya son incorreladas, entonces no tiene sentido hacer componentes principales. Si se hace se obtiene las mismas variables ordenadas de mayor a menor varianza. Para ver eso se hace el test de esfericidad de Bartlett (package psych) o el indice de Kayser-Meyer-Olkin (KDO).
3. Si Σ tiene un valor propio con multiplicidad mayor que 1 se toma vectores propios ortgonales en el subespacio propio correspondiente.
4. Se conservan en general dos o tres componentes.

Referencias

1. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2013.
2. Daniel Peña, Análisis Multivariante, Mac Graw Hill, 2002.