Análisis de Datos 1

TEMA 1: ANÁLISIS DE COMPONENTES PRINCIPALES

20 de febrero de 2023

Introducción

- La información de partida en el análisis multivariante es una tabla de datos correspondiente a p variables medidas en los n elementos (individuos) de un conjunto.
- Por tanto, la matriz de datos tiene dimensión $n \times p$

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} \vec{x}_1 & \vec{x}_2 & \dots & \vec{x}_p \end{pmatrix}$$

donde $\vec{x}_i = (x_{1i}, \dots, x_{ni})'$ es una muestra de tamaño n de la variable X_i.

Covarianzas

- La **media** de una muestra $\vec{x} = (x_1, \ldots, x_n)$ es

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- La **covarianza** entre las muestras $\vec{x}$ y $\vec{y}$ de las variables X y Y respectivamente viene dada por

$$S_{XY} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- Matriz de varianzas-covarianzas $p \times p$ de los datos $\mathbf{X}$

$$\mathbf{S} = \mathrm{Var}(\mathbf{X}) = [S_{X_j X_k}]_{jk}$$

La diagonal contiene las varianzas

$$S_{X_j}^2 = \mathrm{Var}(\vec{x}_j) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Correlaciones

- El **coeficiente de correlación** lineal de Pearson viene dado por

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} \in [-1, +1]$$

- Matriz de correlaciones $p \times p$

$$\mathbf{R} = [r_{X_j X_k}]_{jk}$$

La diagonal contiene unos: $r_{X_j X_j} = 1$

Combinación lineal

- Sea $\vec{a}$ un vector de p coeficientes.
- Considerar los datos trasformados linealmente $\vec{a}' \cdot \mathbf{x}$.
- Su varianza es

$$\text{Var}(\vec{a}' \cdot \mathbf{X}) = \vec{a}' \cdot \text{Var}(\mathbf{X}) \cdot \vec{a} = \vec{a}' \cdot \mathbf{S} \cdot \vec{a}$$

Obtención de componentes principales

- Buscar la combinación lineal que haga la varianza máxima, con la condición $\|\vec{a}\| = 1$:

$$\max\left\{\mathrm{Var}(\vec{a}' \cdot \mathbf{X}) \mid \|\vec{a}\| = 1\right\} = \max\left\{\vec{a}'\mathbf{S}\vec{a} \mid \|\vec{a}\| = 1\right\}$$

- Por el teorema de la descomposición espectral, existen
  - autovalores en orden decreciente $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geq \lambda_p$

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

  - autovectores ortonormales respectivos $\vec{u}_1, \ldots, \vec{u}_p$

$$\mathbf{U} = \begin{pmatrix} \vec{u}_1 & \ldots & \vec{u}_p \end{pmatrix}$$

tales que

$$\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$$

Obtención de componentes principales

- $\|\mathbf{U}'\vec{a}\| = \vec{a}'\mathbf{U}\mathbf{U}'\vec{a} = \vec{a}'\vec{a} = \|\vec{a}\|$
- Buscar la combinación lineal que haga la varianza máxima, con la condición $\|\vec{a}\| = 1$:

$$\begin{aligned}
\max\left\{\vec{a}'\mathbf{S}\vec{a} \mid \|\vec{a}\| = 1\right\} &= \max\left\{\vec{a}'\mathbf{U}\mathbf{\Lambda}\mathbf{U}'\vec{a} \mid \|\vec{a}\| = 1\right\} \\
&= \max\left\{\vec{b}'\mathbf{\Lambda}\vec{b} \;\middle|\; \|\vec{b}\| = \|\mathbf{U}'\vec{a}\| = 1\right\} \\
&= \max\left\{\sum_{i=1}^{p}\lambda_i b_i^2 \;\middle|\; \|\vec{b}\| = 1\right\} \\
&= \lambda_1
\end{aligned}$$

con $b_1 = 1$ y $b_2 = \cdots = b_p = 0$, luego $\vec{a} = \vec{u}_1$

Obtención de componentes principales

- Buscar la combinación lineal que haga la varianza máxima, con la condición $\|\vec{a}\| = 1$.
- Solución:
  El vector propio de **S** asociado a su mayor valor propio.
- Se buscan nuevas combinaciones, linealmente independientes de la primera y que tengan varianza máxima. Se obtienen los vectores propios asociados a los valores propios.
- Variabilidad explicada por k componentes:

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p}$$

- Elección del número de componentes:
  1. Se alcance un porcentaje fijado de la variabilidad global. Por ejemplo el 70 %.
  2. Si hay dos componentes con varianzas muy similares se deben elegir las dos o ninguna.

Interpretación del significado de componentes

- En cada componente se analiza el perfil de los individuos que tienen puntuaciones extremas.
- El signo en sí no se interpreta porque cada valor propio tiene como vectores propios dos con signo distinto.
- Cuanto más próximo a 0 sea el valor absoluto de los coeficientes de las variables, menos importantes son.