Análisis factorial

Daniel Czarnievicz

Análisis Factorial

Se parte de una matriz de datos de la forma:

$$\mathbf{X}_{I \times J} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1J} \\ x_{21} & x_{22} & \dots & x_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ x_{I1} & x_{I2} & \dots & x_{IJ} \end{bmatrix}$$

A partir de ella se definen dos espacios:

1. El espacio definido por la nube de las N_I filas, el cual está incluido en $\mathbb{R}^J$ (dado que cada fila constituye un vector con J componentes).

2. El espacio definido por la nube de las N_J columnas, el cual está incluido en $\mathbb{R}^I$ (dado que cada columna constituye un vector con I componentes).

El objetivo principal de un análisis factorial es eliminar la información redundante (reducción de dimensionalidad). Los resultados de un análisis factorial son: los ejes de incercia, y las coordenadas de los puntos sobre dichos ejes (llamados factores).

Desarrollo por N_I

Trabajamos primero con la nube de puntos N_I, definida por las filas de la matriz $\mathbf{X}_{I \times J}$. Cada individuo está representado por un vector $\mathbf{x}_i = (x_{i1},\, x_{i2},\, \dots,\, x_{iJ})' \in \mathbb{R}^J$. El objetivo es encontrar el conjunto de ejes ortonormados que maximicen la inercia proyectada sobre ellos. Al conjunto de dichas coordenadas sobre un eje de inercia se le llama *factor*.

Se definen las matrices diagonales $\mathbf{M}_{J \times J}$ y $\mathbf{D}_{I \times I}$ tales que:

$$\mathbf{M}_{J \times J} = \begin{bmatrix} m_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & m_J \end{bmatrix} \qquad \mathbf{D}_{I \times I} = \begin{bmatrix} p_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & p_I \end{bmatrix}$$

Estas matrices actuarán de pesos o métricas en cada análisis según la siguiente tabla:

	Espacio	Métrica	Pesos
Nube de filas	$\mathbb{R}^J$	$\mathbf{M}$	$\mathbf{D}$
Nube de columnas	$\mathbb{R}^I$	$\mathbf{D}$	$\mathbf{M}$

Dado que $\mathbf{M}$ es diagonal, la distancia entre dos puntos i y k de N_I se calcula como:

$$d^2(i,\, k) = \sum_{j=1}^{J} (x_{ij} - x_{kj})^2 m_j$$

Sea $\mathbf{u}_s$ un vector director de un eje cualquiera de $\mathbb{R}^J$. Definimos el vetor de coordenadas proyectadas sobre $\mathbf{u}_s$, al que llamaremos $\mathbf{F}_s(i)$, como:

$$\mathbf{F}_s(i) = \mathbf{x}_i' \mathbf{M} \mathbf{u}_s$$

Nótese que $\mathbf{x}_i'$ es de forma $1 \times J$, $\mathbf{M}$ es una matriz de forma $J \times J$, $\mathbf{u}_s$ es de forma $J \times 1$. $\mathbf{F}_s(i)$ es un escalar. Este número representa la proyección del i-ésimo individuo sobre el eje de inercia $\mathbf{u}_s$. Visto en forma matricial:

$$\mathbf{F}_s = \mathbf{X} \mathbf{M} \mathbf{u}_s \;\; \text{con} s = 1,\, \dots,\, J$$

Nótenes que omitimos la mención al i-ésimo individuo en $\mathbf{F}_s$. Esto se debe a que $\mathbf{X}$ es una matriz de tamaño $I \times J$, por lo que $\mathbf{F}_s$ es un vector de dimensión $I \times 1$, donde su i-ésima entrada es la proyección del i-ésimo individuo sobre el eje de inercia $\mathbf{u}_s$.

Podemos entonces definir la *inercia de la nube proyectada* como:

$$\text{inercia} = \mathbf{F}_s' \mathbf{D} \mathbf{F}_s$$

la cual podemos escribir, utilizando la definición de $\mathbf{F}_s$, como:

$$\text{inercia} = \mathbf{u}_s' \mathbf{M}' \mathbf{X}' \mathbf{D} \mathbf{X} \mathbf{M} \mathbf{u}_s$$

El objetivo, tal como fuera planteado, es hallar el eje $\mathbf{u} \in \mathbb{R}^J$ unitario en la métrica $\mathbf{M}$, que maximice la inercia. Es decir,

$$\max_u \{\text{inercia}\} = \max_u \left\{ \mathbf{u}_s' \mathbf{M}' \mathbf{X}' \mathbf{D} \mathbf{X} \mathbf{M} \mathbf{u}_s \right\}$$

Si la métrica es euclidea, $\mathbf{M}$ es la matriz identidad, y el problema se reduce a hallar un vector $\mathbf{u}$ tal que $\mathbf{u}'\mathbf{u} = 1$ que maximice la inercia:

$$\max_u \{\text{inercia}\} = \max_u \left\{ \mathbf{u}' \mathbf{X}' \mathbf{D} \mathbf{X} \mathbf{u} \right\}$$

Dado que $\mathbf{X}'\mathbf{D}\mathbf{X}$ es simétrica:

- $\mathbf{X}'\mathbf{D}\mathbf{X}$ es diagonalizable.
- $\exists\, \mathbf{U}$ matriz ortogonal (es decir, $\mathbf{U}'\mathbf{U} = \mathbf{U}\mathbf{U}' = \mathbf{I}$).
- $\mathbf{U} = ((u_j))$ son los vetores propios asociado al valor propio λ_j.
- se define $\mathbf{\Lambda}$ como la matriz diagonal de valores propios: $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_J)$ tales que

$$\mathbf{U}'\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{U} = \mathbf{\Lambda} \Rightarrow \mathbf{X}'\mathbf{D}\mathbf{X} = \mathbf{U}'\mathbf{\Lambda}\mathbf{U}$$

- sus vectores propios forman una base ortonormal en $\mathbf{R}^J$.

Desarrollo por N_J

Interpretación de resultados

Análisis de Componentes Principales

Partimos de una matriz de datos donde a n individuos se le miden p variables. Cada individuo puede entonces ser representado por un vector en $\mathbb{R}^p$.

Análisis de Correspondencias Simples

Análisis de Correspondencias Múltiple

Referencias

Beygelzimer, Alina, Sham Kakadet, John Langford, Sunil Arya, David Mount, and Shengqiao Li. 2018. *FNN: Fast Nearest Neighbor Search Algorithms and Applications*. https://CRAN.R-project.org/package=FNN.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 112. Springer.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rencher, Alvin C. 1998. *Multivariate Statistical Inference and Applications*. Wiley New York.

Wasserman, Larry. 2007. *All of Nonparametric Statistics*. Springer, New York.

Wickham, Hadley. 2017. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.