Análisis Factorial y Componentes principales.

Análisis Factorial

El Análisis Factorial es un método estadístico para determinar si un número determinado de variables de interés $X_1, X_2, ..., X_p$ están linealmente relacionadas a un número más pequeño de factores *no observables* $F_1, F_2, ..., F_k$. El hecho de que dichos factores no puedan observarse, descarta inmediatamente la posibilidad de utilizar regresión sobre los mismos. La intención del método estriba en dos propósitos fundamentales:

- Reducción de dimensiones
- Crear índices con las variables originales que midan atributos similares.

En minería de datos, usaremos la técnica para el primer propósito, en virtud de que resumiremos la información de entrada para los modelos.

El modelo factorial presenta la siguiente estructura:

$$X_1 = l_{11}F_1 + l_{12}F_2 + ... + l_{1k}F_k + e_1$$
$$X_2 = l_{21}F_1 + l_{22}F_2 + ... + l_{2k}F_k + e_2$$

$$X_p = l_{p1}F_1 + l_{p2}F_2 + ... + l_{pk}F_k + e_p$$

Aquí, $F_1, F_2, ..., F_k$ son los ***factores comunes***, $e_1, e_2, ..., e_p$ son llamados ***factores específicos*** y l_{jh} son conocidos como **cargas** de la variable j en el factor h.

El modelo nos plantea que cada variable es representada como una combinación lineal de k factores comunes a todas las variables y de un factor único por cada variable. Es posible representar el modelo de forma matricial, esto es:

$$\begin{bmatrix} X_1 \\ X_2 \\ ... \\ X_P \end{bmatrix} = \begin{bmatrix} l_{11} & l_{12} & ... & l_{1k} \\ l_{21} & l_{22} & ... & l_{2k} \\ ... & ... & ... & ... \\ l_{p1} & l_{p2} & ... & l_{pk} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ ... \\ F_P \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ ... \\ e_P \end{bmatrix}$$

O en forma corta:

$$\vec{X} = L\vec{F} + e$$

Para poder aplicar el modelo, tenemos que tomar en cuenta ciertos supuestos sobre nuestros factores:

- $F_1, F_2, ..., F_k$ son variables estandarizadas de media cero y varianza unitaria
- Los factores no están correlacionados entre sí

El punto dos, implica que $E\left[\vec{F}\vec{F}^T\right] = I$ y $E\left[\vec{F}\right] = \vec{0}$

Por otra parte, supondremos que la matriz de covarianzas de los factores específicos es una matriz diagonal lo cual implica que son incorrelacionados, es decir $E\left[\vec{e}\vec{e}^T\right] = \Omega$ con Ω diagonal. Adicional a lo anterior, supondremos que $E\left[\vec{e}\right] = \vec{0}$. Por último supondremos también que $E\left[\vec{F}\ \vec{e}^T\right] = \vec{0}$.

Al ser $\vec{X}$ estandarizado, su matriz de covarianzas es equivalente a la matriz de correlaciones, esto es:

$$\begin{aligned} R_p &= E\left[\vec{X}\vec{X}^T\right] = E\left[L\vec{F} + \vec{e}\right]\left[L\vec{F} + \vec{e}\right]^T \\ &= LE\left[\vec{F}\vec{F}^T\right]L^T + E\left[\vec{e}\vec{e}^T\right] + LE\left[\vec{F}\vec{e}^T\right] + E\left[\vec{e}\vec{F}^T\right]L^T \\ &= LIL^T + \Omega \\ &= LL^T + \Omega \end{aligned}$$

De lo anterior, podemos ver que la varianza de cada variable estandarizada X_j está dada por:

$$V_j = 1 = l_{j1}{}^2 + l_{j2}{}^2 + ... + l_{jp}{}^2 + \omega_j{}^2$$

Sea

$$h_j{}^2 = l_{j1}{}^2 + l_{j2}{}^2 + ... + l_{jp}{}^2$$

La varianza de X_j podrá ser descompuesta como:

$$V_j = h_j{}^2 + \omega_j{}^2 \qquad j = 1,2,..., p$$

Donde $h_j{}^2$ es la parte de la varianza debida a los factores comunes conocida también como ***comunalidad***, *mientras que* $\omega_j{}^2$ es la parte de la varianza específica debida a los factores únicos y es denominada especificidad.

Una vez que hemos expuesto el modelo factorial, revisaremos como calcular los factores a través del método de componentes principales.

Antes de iniciar, presentemos brevemente en qué consiste la teoría de los componentes principales.

Análisis de Componentes Principales

Es un método estadístico multivariante que se clasifica como método de simplificación o de reducción de la dimensión. El ACP permite describir de forma sintética la estructura e interrelación de las variables originales.

El método tiene por objeto transformar un conjunto de variables en un nuevo conjunto denominado componentes principales. Los nuevos componentes tienen la característica de ser incorrelacionados (ortogonales) y se ordenan de acuerdo a la cantidad de información (varianza) que llevan incorporada. Las componentes principales se expresan como una combinación lineal de las variables originales. Revisemos un esbozo de su obtención.

Sean $X_1, X_2, ..., X_p$ un conjunto de variables de una muestra de tamaño n interrelacionadas entre sí, se busca obtener otro conjunto $Z_1, Z_2, ..., Z_k$ con $k \leq p$ tales que sean una combinación lineal del conjunto inicial y que expliquen la mayor parte de su variabilidad.

Obtengamos la primera componente:

$$Z_{1i} = u_{11}X_{1i} + u_{12}X_{2i} + ... + u_{1p}X_{pi}$$

Al tomar las n observaciones muestrales, tenemos:

$$\begin{bmatrix} Z_{11} \\ Z_{12} \\ ... \\ Z_{1n} \end{bmatrix} = \begin{bmatrix} X_{11} & X_{21} & ... & X_{p1} \\ X_{12} & X_{22} & ... & X_{p2} \\ ... & ... & ... & ... \\ X_{1n} & X_{2n} & ... & X_{pn} \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{12} \\ ... \\ u_{1p} \end{bmatrix}$$

O en notación matricial:

$$\vec{Z}_1 = X\ \vec{u}_1$$

Al suponer que las X_j están estandarizadas, podemos asumir que:

$$E\left[\vec{Z}_1\right] = E\left[X\ \vec{u}_1\right] = E\left[X\right]\vec{u}_1 = 0$$

Y la varianza sería:

$$\begin{aligned} V\left[\vec{Z}_1\right] &= \frac{1}{n}\sum_{i=1}^{n} {Z_{1i}}^2 \\ &= \frac{1}{n}{\vec{Z}_1}^T \vec{Z}_1 \\ &= \frac{1}{n}{\vec{u}_1}^T X^T X\ \vec{u}_1 \\ &= {\vec{u}_1}^T \left[\frac{1}{n} X^T X\right]\vec{u}_1 \\ &= {\vec{u}_1}^T\ V\ \vec{u}_1 \end{aligned}$$

Donde V es la matriz de covarianzas.

Para hallar $\vec{Z}_1$, necesitamos maximizar la varianza tal que la suma de los pesos u_{1j} al cuadrado sea igual a la unidad, en consecuencia tenemos un problema de optimización.

$$Max\quad V\left[\vec{Z}_1\right] = {\vec{u}_1}^T\ V\ \vec{u}_1$$

$$s.a.\quad \sum_{j=1}^{p} {u_{1i}}^2 = {\vec{u}_1}^T \vec{u}_1 = 1$$

Para resolverlo, recurrimos a los multiplicadores de Lagrange:

$$L = {\vec{u}_1}^T\ V\ \vec{u}_1 - \lambda\left({\vec{u}_1}^T \vec{u} - 1\right)$$

$$\frac{\partial L}{\partial \vec{u}_1} = 2V\vec{u}_1 - 2\lambda\vec{u}_1 = 0 \Rightarrow \left(V - \lambda I\right)\vec{u}_1 = 0$$

La ecuación anterior solo tiene solución si $\left|V - \lambda I\right| = 0$ y en consecuencia, λ es un valor propio de la matriz V. Al premultiplicar por ${\vec{u}_1}^T$, tenemos:

$${\vec{u}_1}^T\left(V - \lambda I\right)\vec{u}_1 = 0 \Rightarrow {\vec{u}_1}^T\ V\ \vec{u}_1 - \lambda{\vec{u}_1}^T\ I\ \vec{u}_1 = {\vec{u}_1}^T\ V\ \vec{u}_1 - \lambda = 0 \Rightarrow {\vec{u}_1}^T\ V\ \vec{u}_1 = \lambda = V\left[\vec{Z}_1\right]$$

Sabemos que $\lambda_1, \lambda_2, ..., \lambda_n$ pueden ordenarse de forma ascendente tal que $\lambda_1 > \lambda_2 > ... > \lambda_n$, de esta manera maximizaremos la varianza explicada tomando el mayor valor propio de V.

La metodología de componentes principales es uno de los métodos para estimar los factores, existen además otros métodos para su cálculo que mencionaremos pero no trataremos a detalle en el curso:

- Método de Turstone
- Método del Factor principal
- Método alpha
- Método del centroide
- Método de las componentes principales iteradas
- Método de máxima verosimilitud
- Método Minres
- Método ULS
- Método GLS

Por el método de componentes principales se realiza una comparación entre la matriz en reversa de los CP y el modelo factorial, siendo las k primeras componentes principales los factores y las cargas l_{jh} los coeficientes de correlación de la variable j-ésima y la componente h-ésima

Rotación de factores

Una de las razones por las que se realiza el análisis factorial recae en que los factores comunes tengan una interpretación clara. Lo anterior no es tarea sencilla, por ello se recurre al procedimiento de rotación de factores independientemente del método por el cual hayan sido obtenidos. En la solución inicial, cada uno de los factores comunes están correlacionados en mayor o menor medida con cada uno de los factores, cuando trabajamos con los factores rotados, tratamos de que cada una de las variables tenga una correlación lo más cercana posible a uno con uno de los factores y lo más cercana a cero con los demás.

Existen dos formas de rotar los factores: la rotación ortogonal y la rotación oblicua. La diferencia consiste en que la rotación ortogonal preserva la incorrelación de los factores al rotar los ejes mientras que en la oblicua no se busca la preservación de la incorrelación. A continuación listamos los principales métodos de rotación, para nuestro curso utilizaremos la rotación *varimax.*

- Rotación ortogonal
  - Varimax
  - Equamax
  - Quartimax
- Rotación oblicua
  - Oblimin
  - Oblimax
  - Promax
  - Quartimin
  - Biquartimin
  - Covarimin

Echémos un vistazo a la rotación varimax:

El método obtiene los ejes de los factores maximizando la suma de las varianzas de las cargas factoriales al cuadrado dentro de cada factor. Esto es:

$$W = p^2 SN^2 = p\sum_{i=1}^{k}\sum_{j=1}^{p}\left(\frac{l_{ji}^{\ 2}}{h_j^{\ 2}}\right)^2 - \sum_{i=1}^{k}\left[\sum_{j=1}^{p}\left(\frac{l_{ji}^{\ 2}}{h_j^{\ 2}}\right)\right]^2$$

Para realizar la maximización se debe hallar la matriz de rotación $T = \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix}$ donde φ es el ángulo al cual se giran los factores en un procedimiento iterativo.

Cerremos nuestro estudio con un ejemplo:

Suponga que un grupo de estudiantes ingresará a un programa de MBA el cual tiene como requisito tres cursos: Finanzas, Marketing y Política de Negocios. Sean Y_1, Y_2 y Y_3 las calificaciones de estos tres cursos respectivamente, observemos la siguiente tabla:

ID Estudiante	Finanzas	Mkt	Política
1	3	6	5
2	7	3	3
3	10	9	8
4	3	9	7
5	10	6	5

Supóngase que dichas calificaciones son funciones de dos factores subyacentes F_1 y F_2 tentativamente descritos como habilidad verbal y habilidad matemática respectivamente.