

# Tema 4: Análisis Factorial

En numerosas áreas de Psicología y de Ciencias del Comportamiento no es posible medir directamente las variables que interesan; por ejemplo, los conceptos de *inteligencia* y de *clase social*. En estos casos es necesario recoger medidas indirectas que estén relacionadas con los conceptos que interesan. Las variables que interesan reciben el nombre de *variables latentes* y la metodología que las relaciona con variables observadas recibe el nombre de Análisis Factorial.

El modelo de Análisis Factorial es un modelo de regresión múltiple que relaciona variables latentes con variables observadas.

El Análisis Factorial tiene muchos puntos en común con el análisis de componentes principales, y busca esencialmente nuevas variables o *factores* que expliquen los datos. En el análisis de componentes principales, en realidad, sólo se hacen transformaciones ortogonales de las variables originales, haciendo hincapié en la varianza de las nuevas variables. En el análisis factorial, por el contrario, interesa más explicar la estructura de las covarianzas entre las variables.

Al igual que en el método de los componentes principales, para efectuar el análisis factorial, es necesario que las variables originales no estén incorreladas porque si lo estuvieran no habría nada que explicar de las variables.

Consideramos un conjunto de  $p$  variables observadas  $\mathbf{x}' = (x_1, x_2, \dots, x_p)$  que se asume relacionadas con un número dado de variables latentes  $f_1, f_2, \dots, f_k$ , donde  $k < p$ , medi-

ante una relación del tipo

$$\begin{aligned}x_1 &= \lambda_{11}f_1 + \cdots + \lambda_{1k}f_k + u_1 \\ &\vdots \\ x_p &= \lambda_{p1}f_1 + \cdots + \lambda_{pk}f_k + u_p\end{aligned}$$

o de modo más conciso

$$\mathbf{x} = \Lambda \mathbf{f} + \mathbf{u}.$$

donde

$$\Lambda = \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1k} \\ \vdots & \ddots & \vdots \\ \lambda_{p1} & \cdots & \lambda_{pk} \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} f_1 \\ \vdots \\ f_k \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_p \end{pmatrix}.$$

Los  $\lambda_{ij}$  son los *pesos factoriales* que muestran como cada  $x_i$  depende de factores comunes y se usan para interpretar los factores. Por ejemplo, valores altos relacionan un factor con la correspondiente variable observada y así se puede caracterizar cada factor.

Se asume que los términos residuales  $u_1, \dots, u_p$  están incorrelados entre sí y con los factores  $f_1, \dots, f_k$ . Cada variable  $u_i$  es particular para cada  $x_i$  y se denomina *variable específica*.

Dado que los factores no son observables, se puede fijar arbitrariamente su media en 0 y su varianza en 1, esto es, se consideran variables estandarizadas que están incorreladas entre sí, de modo que los pesos factoriales resultan ser las correlaciones entre las variables y los factores.

Así, con las suposiciones previas, la varianza de la variable  $x_i$  es

$$\sigma_i^2 = \sum_{j=1}^k \lambda_{ij}^2 + \psi_i$$

donde  $\psi_i$  es la varianza de  $u_i$ .

De este modo, la varianza de cada variable observada se puede descomponer en dos partes. La primera  $h_i^2$ , denominada *comunalidad*, es

$$h_i^2 = \sum_{j=1}^k \lambda_{ij}^2$$

y representa la varianza compartida con las otras variables por medio de los factores comunes. La segunda parte,  $\psi_i$ , se denomina varianza específica y recoge la variabilidad no compartida con las otras variables.

La definición del modelo implica que la covarianza entre las variables  $x_i$  y  $x_j$  es

$$\sigma_{ij} = \sum_{l=1}^k \lambda_{il} \lambda_{lj}.$$

Las covarianzas no dependen en absoluto de las variables específicas, de hecho, basta con los factores comunes. De este modo, la matriz de covarianzas  $\Sigma$  de las variables observadas es

$$\Sigma = \Lambda \Lambda' + \Psi$$

donde  $\Psi$  es una matriz diagonal cuyos componentes son las varianzas específicas:  $\Psi = \text{diag}(\psi_i)$ .

Lo contrario también se verifica: dada la descomposición de la varianza anterior, se puede encontrar un modelo factorial para las variables originales,  $\mathbf{x}$ , con  $k$  factores.

En la práctica se tienen que estimar los parámetros del modelo a partir de una muestra, de modo que el problema se centra en encontrar los valores  $\hat{\Lambda}$  y  $\hat{\Psi}$  tales que la matriz de covarianzas muestral  $S$  es aproximadamente

$$S \approx \hat{\Lambda} \hat{\Lambda}' + \hat{\Psi}$$

Se tienen dos métodos de estimación de los términos anteriores: el método de los factores principales y el método de máxima verosimilitud.

### **Método de los factores principales**

Es una técnica basada en autovalores y autovectores pero en lugar de operar sobre la matriz de covarianzas se opera sobre la llamada matriz de covarianzas *reducida*,

$$S^* = S - \hat{\Psi}$$

donde  $\hat{\Psi}$  es una matriz diagonal que contiene las estimas de  $\psi_i$ .

Los elementos diagonales de  $S^*$  contiene las comunalidades estimadas (las partes de las varianzas de cada variable explicada por los factores comunes). Al contrario que el análisis de componentes principales, el análisis factorial no pretende recoger toda la varianza observada de los datos, sino la que comparten los factores comunes. De hecho, el análisis factorial se centra más en recoger las covarianzas o correlaciones que aparecen entre las variables originales.

El procedimiento es iterativo: se parte de unas comunalidades estimadas a partir de las correlaciones entre las variables observadas y luego se efectua un análisis de componentes principales sobre la matriz  $S^*$ .

### **Método de la máxima verosimilitud**

Este método es el habitualmente preferido por los estadísticos. Asumiendo normalidad en los datos se define una distancia  $F$ , entre la matriz de covarianzas observada y los valores predichos de esta matriz por el modelo del análisis factorial. La expresión de dicha distancia es

$$F = \ln |\Lambda\Lambda' + \Psi| + \text{traza} \left( S |\Lambda\Lambda' + \Psi|^{-1} \right) - \ln |S| - p$$

Las estimaciones de los pesos factoriales se obtienen minimizando esta función, y esto es equivalente a maximizar la función de verosimilitud del modelo  $k$  factorial asumiendo normalidad.

### **Estimación del número de factores**

El hecho de tomar un número adecuado de factores  $k$  para representar las covarianzas observadas es muy importante: entre una solución con  $k$  ó con  $k + 1$  factores se pueden encontrar pesos factoriales muy diferentes, al contrario que en el método de componentes principales, donde los primeros  $k$  componentes son siempre iguales.

Una ventaja del método de máxima verosimilitud es que lleva asociado un test estadístico para estimar el número de factores.









































