Tema 4.- Análisis factorial (AF)

Asignatura: ESTADÍSTICA MULTIVARIANTE (Prácticas)

Grados en: Física y Matemáticas; Ingeniería Informática y Matemáticas; Matemáticas

(4º Curso - 1er semestre 2023-2024)

©Prof. Dr. José Luis Romero Béjar

Departamento de Estadística e Investigación Operativa
Facultad de Ciencias (Despacho 10)

Periodo de docencia: 11/09/2023 a 22/12/2023

Objetivo

El principal **objetivo** del Análisis Factorial (AF) es **captar la realidad de la forma más simple posible**, mediante la identificación de unas 'pocas' **variables latentes** que definan esta realidad.

Variables latentes

Una **variable latente** es una variable **no observable** que es **inferida** en función de otras variables observables a partir de un modelo matemático.

Se pueden encontrar **ejemplos** en distintos ámbitos de la ciencia:

- **Economía**: la calidad de vida es una variable latente que es inferida en función de otras por medio de un modelo matemático (AF, probit, logit, etc.).
- **Psicología:** las cinco variables que definen la personalidad: neruroticismo, extraversión, apertura a experiencias, amabilidad y responsabilidad, son variables latentes.

ACP vs. AF

El **AF** es un **conjunto de técnicas** que pretenden identificar factores ocultos (variables latentes) que, preferiblemente **correlacionen altamente con un grupo de las variables observables pero no con otras**, con el objetivo, mencionado antes, de explicar la realidad con el menor número de variables posibles, es decir **reducir la dimensión**.

En este sentido:

- El ACP y el AF **tienen en común** que ambos métodos buscan **reducir la dimensión del problema**.
- Parten de la **hipótesis común** de que las variables estén relativamente **correlaciondas**.
- El ACP y el AF **difieren**, en que mientras el primero busca combinaciones lineales de la variables aleatorias originales, por tanto **observables**, que **maximicen la varianza en cada dirección**, el segundo busca factores latentes, por tanto **no observables**, que **correlacionen en sentido máximo con ciertos grupos de las variables observadas**.

Planteamiento del problema

Sean $X_1, X_2, \ldots, X_p$ un conjunto de p variables aleatorias correladas. Se denota por $X = (X_1, X_2, \ldots, X_p)^t$ al vector aleatorio que forman. Se asume que X es centrado, $E[X] = 0$ y se denota por $\Sigma = E[XX^t]$ a su matriz de covarianzas. Finalmente, se

supone que el vector aleatorio se puede muestrear de la forma:

$$X = AF + L \tag{1}$$

donde,

- F_{kx1} es un vector aleatorio de $k \leq p$ **factores comunes** (no observables que correlacionan con un conjunto de variables observadas).
- L_{px1} es un vector aleatorio de p **factores específicos** (sólo correlacionan con la variable observada correspondiente).
- A_{pxk} es una matriz de constantes, la **matriz de pesos factorial**, que permitirá determinar la influencia de los factores en las variables observadas y viceversa.

Supuestos necesarios

Para resolver el problema (1) se realizan, sin pérdida de generalidad, los siguientes supuestos:

i. $E[F] = 0_{kx1}$

ii. $E[L] = 0_{px1}$

iii. $E[FL^t] = 0_{kxp}$

iv. $E[FF^t] = I_k$

v. $E[LL^t] = D_p = \begin{pmatrix} d_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & d_p \end{pmatrix}$, matriz diagonal.

Observación

Dada la aleatoriedad de los factores, si además de suponer que son centrados se asumen varianzas unitarias, entonces **la matriz factorial realmente representa las correlaciones de los factores con las variables observadas**.

Importante

- Los factores comunes F influyen en X a través de los coeficientes de la matriz factorial A.
- Los factores específicos en L sólo influyen en la variable homóloga (L_1 en X_1, L_2 en X_2, ...).
- Un modelo como (1) es indicado cuando se trabaja con un número grande de variables que pueden estar realmente causadas por unos pocos factores comunes.
- El modelo de AF se **asemeja a un modelo de regresión lineal** con la salvedad de que aquí la variable respuesta X es multivariante y que los regresores F son variables no observables.
- En cualquier caso los supuestos considerados permiten, en general, **obtener una solución**, si bien, **no única**.

A continuación se justificará que **el objetivo del AF es estimar las matrices** A y D.

Igualdad fundamental

$$\Sigma = E[XX^t] = AA^t + D \tag{2}$$

La demostración es muy sencilla **(ejercicio propuesto voluntario)** y se puede encontrar en la referencia bibliográfica *Tussell, 2016, p.66*.

Comunalidades

Si se escribe la ecuación (2) elemento a elemento, es fácil darse cuenta de que:

- $\sigma_i^2 = \sigma_{ii} = \sum_{j=1}^{k} a_{ij}^2 + d_i,\ i = 1, \ldots, p.$
- $\sigma_{ij} = \sum_{l=1}^{k} a_{il} a_{lj},\ i, j = 1, \ldots, p,\ i \neq j.$

La parte de la varianza de las variables aleatorias X_i identificada por los factores comunes se denomina **comunalidad** y se denota como:

$$h_i^2 = \sum_{j=1}^{k} a_{ij}^2 \ ,\forall i = 1, \ldots, p \tag{3}$$

Objetivo en la práctica

Teniendo en cuenta la igualdad fundamental (2) anterior, el **objetivo del AF**, desde un punto de vista computacional, será **hallar matrices** A **y** D **para una matriz de covarianzas,** Σ, dada cumpliendo la igualdad, de modo que A **tenga el menor número de columnas (factores) posibles**.

Observación

En la práctica no se conoce la matriz Σ de modo que **se trabaja con una estimación** S, a partir de la cual se trata de reconstruir Σ como un producto AA^t más una matriz diagonal.

Ejemplo: enunciado

El siguiente ejemplo sencillo ilustra el ajuste de un modelo factorial con un sólo factor.

Se considera un vector aleatorio que almacena las calificaciones de tres asignaturas distintas $X = (X_1, X_2, X_3)$. Partiendo de la estimación de la matriz de correlaciones, entre las calificaciones, siguiente:

$$S = \begin{pmatrix} 1 & 0.83 & 0.78 \\ 0.83 & 1 & 0.67 \\ 0.78 & 0.67 & 1 \end{pmatrix}$$

se pretende ajustar **un modelo factorial con un solo factor**.

Ejemplo: solución

Según la ecuación (1), el modelo para un factor tendrá la siguiente expresión:

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} F_1 + \begin{pmatrix} L_1 \\ L_2 \\ L_3 \end{pmatrix}$$

Esto implica, a través de la igualdad fundamental (2), que:

$$S = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} \begin{pmatrix} a_{11} & a_{21} & a_{31} \end{pmatrix} + \begin{pmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{pmatrix}$$

Sustituyendo y operando en la expresión anterior:

$$S = \begin{pmatrix} 1 & 0.83 & 0.78 \\ 0.83 & 1 & 0.67 \\ 0.78 & 0.67 & 1 \end{pmatrix} = \begin{pmatrix} a_{11}^2 & a_{11}a_{21} & a_{11}a_{31} \\ a_{21}a_{11} & a_{21}^2 & a_{21}a_{31} \\ a_{31}a_{11} & a_{31}a_{21} & a_{31}^2 \end{pmatrix} + \begin{pmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{pmatrix}$$

Ejemplo: solución

De la expresión anterior se obtiene el siguiente sistema de 6 ecuaciones con 6 incógnitas.

$$\begin{aligned} a_{11}^2 + d_1 &= 1 \\ a_{21}^2 + d_2 &= 1 \\ a_{31}^2 + d_3 &= 1 \\ a_{11}a_{21} &= 0.83 \\ a_{11}a_{31} &= 0.78 \\ a_{21}a_{31} &= 0.67 \end{aligned}$$

Por tanto el modelo ajustado con un sólo factor queda de la forma:

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} 0.983 \\ 0.844 \\ 0.793 \end{pmatrix} F_1 + \begin{pmatrix} L_1 \\ L_2 \\ L_3 \end{pmatrix}$$

Ejemplo: conclusión

- En este caso se ve como **la primera calificación es la que más influye** (es más influida) por el factor latente, aunque las otras dos también tienen una alta correlación con el factor.
- En un modelo con tan pocas variables no parece que ajustar dos factores mejore mucho la interpretación de los resultados, aunque **como ejercicio voluntario práctico** se propone **repetir este proceso para ajustar un modelo con dos factores** a este problema.

Situación idónea

Una **situación ideal** sería aquella en la que la **matriz factorial mostrara una alta correlación de cada uno de los factores con un grupo de variables observables** concretas y prácticamente nula con el resto.

Por ejemplo:

Suponiendo que $X = (X_1, \ldots, X_7)$ es un vector aleatorio al que se ha ajustado un modelo factorial con tres factores, que ha proporcionado la siguiente matriz de pesos factorial.

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

Una situación así indicaría que el primer factor influye en las variables X_1, X_2, X_3 pero no en las otras. El segundo factor en las variables X_4, X_5 y no en las otras. Finalmente el tercer factor solo influye en las variables X_6, X_7.

Algunos comentarios

- Como se dijo anteriormente, los problemas (1) y (2) no tienen solución única. Esto implica que, difícilmente, se encontrará como primera solución una representación tan sencilla de la realidad como la mostrada en el ejemplo anterior. En la medida de lo posible la **matriz factorial debería de aproximarse a una como la anterior**.
- Una matriz G_k **ortogonal** ($G^{-1} = G^t$) representa una isometría en el espacio vectorial euclídeo $\mathbb{R}^k$ (**rotaciones**, reflexiones o composición de ambas).
- Si se considera una matriz ortogonal, G de orden k, la ecuación (2) no cambia según la siguiente expresión:

$$\Sigma = E[XX^t] = AA^t + D = AGG^tA + D$$

Si se denota por $B = AG$, se obtiene una expresión equivalente a (2), de la forma:

$$\Sigma = E[XX^t] = BB^t + D$$

Más comentarios

- Encontrar las matrices A y D que resuelvan la ecuación (2) es equivalente a encontrar las matrices B y D que también la resuelvan.
- Esto implica que se puede **cambiar a una visión más simple de la realidad** sin más que **introducir una rotación adecuada**. Y por tanto la ecuación (1) se puede escribir como:

$$X = AGG^tF + L = BG^tF + L = BF_G + L$$

- La expresión anterior introduce el concepto de **rotación de los factores**.
- Existen dos tipos de rotaciones posibles: **ortogonales y oblicuas**.
- En este momento se consideran rotaciones **ortogonales, por lo simple de su interpretación, ya que los pesos de la matriz factorial representan las correlaciones entre las variables y los factores**. Esto no se cumple en el caso de las oblicuas.

Rotaciones

El **principal objetivo al realizar una rotación es encontrar una estructura simple**.

En este sentido la matriz factorial debería cumplir las siguientes **propiedades**:

- **Cada fila** de la matriz factorial debe **contener**, al menos, **un cero**.
- **Cada columna** de la matriz factorial debe **contener**, al menos, **k ceros**.
- **Cada par de columnas** de la matriz factorial debe **contener varias variables** cuyos **pesos sean nulos en una columna**, pero no en la otra.
- Si hay **más de cuatro factores** cada **par de columnas de la matriz factorial debe contener un número elevado de variables con pesos nulos** en ambas columnas.
- De manera recíproca, si hay **más de cuatro factores**, en cada par de columnas de la matriz factorial **sólo un número pequeño de variables debe contener pesos no nulos**.

Rotaciones: enfoque quartimax

- El enfoque **quartimax** escoge $A_G = AG$ para la que es máxima la varianza por **filas** de los cuadrados de los a_{ij}.
- Este enfoque trata que una **variable dada esté muy correlacionada con un factor** y muy poco correlacionada con el resto de factores.

Rotaciones: enfoque varimax

- El enfoque **varimax** escoge $A_G = AG$ para la que es máxima la varianza por **columnas** de los cuadrados de los a_{ij}.
- Este enfoque trata de que haya **factores con correlaciones altas con un número pequeño de variables y correlaciones nulas con el resto**. De esta forma queda redistribuida la varianza de los factores.
- Este es el **enfoque más usual** cuando se trabaja con rotaciones ortogonales de factores.

Objetivos

- Determinar el **número de factores** deseado.

  La elección del número adecuado de factores para representar covarianzas observadas es muy importante, ya que, **entre una solución con k factores y otra con $k+1$ factores se pueden obtener matrices factoriales muy diferentes**. Esto **no sucedía en el ACP**, ya que las direcciones principales son siempre las mismas, tomemos k ó $k+1$ de ellas.

- **Estimar inicialmente una matriz factorial** A, por algún método de los que se introducen a continuación, que después será rotada según si interesa simplificar la interpretación de la realidad o no.

Método del factor principal

- Técnica que como el ACP está basada en el **cálculo de valores y vectores propios**, pero en este caso, no sobre la matriz de covarianzas, sino sobre una **matriz de covarianzas reducida**

$$S^* = S - \hat{D},$$

donde $\hat{D}$ es una estimación de la matriz diagonal de varianzas específicas (varianzas de los factores específicos) y S, como antes, una estimación de la matriz de covarianzas.

- Al restar $\hat{D}$, se tiene que **la diagonal de la matriz S^* contiene las distintas comunalidades** (partes de las varianzas de cada variable explicada por los factores latentes).
- 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**.
- Finalmente, **el método del factor principal consiste en aplicar un análisis de componentes principales para la matriz S^***.
- **Los vectores propios** ahora representan las **columnas de la matriz factorial**.

Método de máxima verosimilitud

- Los datos deben de estar distribuidos según una **normal multivariante**.
- **Se define una matriz de distancias entre la matriz de covarianzas observada y sus valores predichos por el modelo de análisis factorial**. Esta distancia se define de la siguiente forma:

$$F = \ln |AA^t + D| + \text{traza}(S|AA^t + D|^{-1}) - \ln |S| - p$$

- Las estimación de la matriz de pesos factoriales, A, se obtiene **minimizando esta distancia**.
- Este problema de minimización es **equivalente a maximizar la función de verosimilitud** del modelo k factorial bajo el supuesto de normalidad.
- El método de máxima verosimilitud lleva asociado un **test estadístico para estimar el numero adecuado de factores**.

Práctica 2 de AF

En esta práctica, de entre 25 ítems de una prueba de personalidad, se van a **identificar las variables que corresponden a cada uno de los cinco aspectos de la personalidad** de un individuo. Las cinco caracteristicas que definen la personalidad de un individuo son: A - *Amabilidad*; C - *Conciencia o responsabilidad*; E - *Extraversion*; N - *Neuroticismo* y Ap - *Apertura a experiencias*.

Para la realización de la misma hay que **descargar y ejecutar** el archivo **Practica_2_AF.Rmd** disponible en la plataforma PRADO.

Aspectos tratados:

- Realizar un análisis exploratorio previo de los datos para identificar posibles **datos perdidos** y **valores extremos**.
- **Tomar decisiones y tratar** datos perdidos y valores extremos.
- Verificar los supuestos y realizar un **AF**.
- **Elección del número óptimo** de factores.
- Interpretación de distintas salidas **gráficas de interés** para este método.
- **Lenguaje R**: depuración de funciones.

[1] Anderson, T.W. (2003, 3ª ed.). An Introduction to Multivariate Statistical Analysis. John Wiley & Sons.

[2] Gutiérrez, R. y González, A. (1991). Estadística Multivariable. Introducción al Análisis Multivariante. Servicio de Reprografía de la Facultad de Ciencias. Universidad de Granada.

[3] Härdle, W.K. y Simar, L. (2015, 4ª ed.). Applied Multivariate Statistical Analysis. Springer.

[4] Johnson, R.A. y Wichern, D.W. (1988). Applied Multivariate Analysis. Prentice Hall International, Inc.

[5] Rencher, A.C. y Christensen, W.F. (2012, 3ª ed.). Methods of Multivariate Analysis. John Wiley & Sons.

[6] Salvador Figueras, M. y Gargallo, P. (2003). Análisis Exploratorio de Datos. Online en http://www.5campus.com/leccion/aed.

[7] Timm, N.H. (2002). Applied Multivariate Analysis. Springer.

[8] Vera, J.F. (2004). Análisis Exploratorio de Datos. ISBN: 84-688-8173-2.