

IN 540
Métodos Estadísticos
para Economía y Gestión

Cap. VII

Análisis Factorial

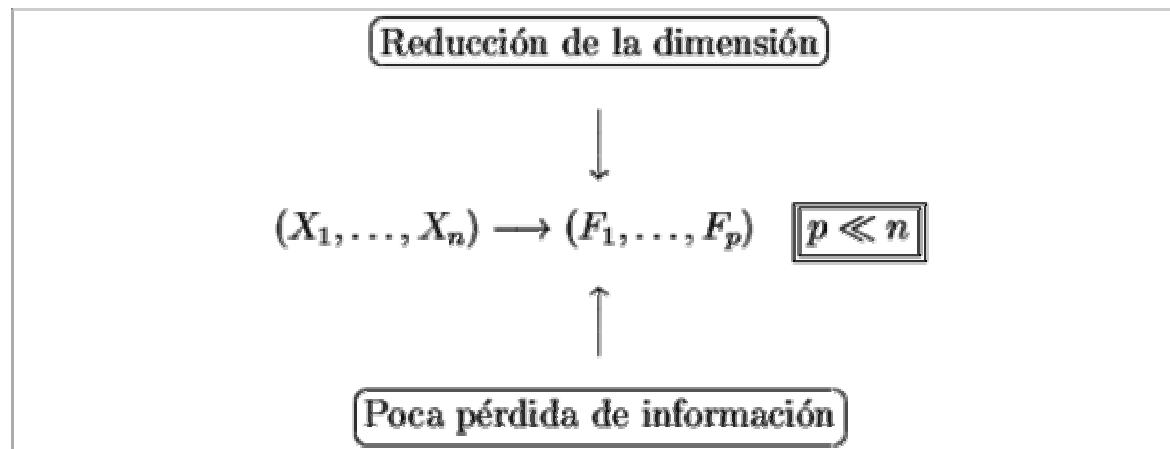
7.1 Introducción

El Análisis Factorial es una técnica multivariada que nos permite reducir el “tamaño” de un problema sin “demasiada pérdida de información”.

Supongamos que disponemos de un problema en el que queremos estudiar el comportamiento de X_1, X_2, \dots, X_n variables. Si n es un valor muy alto, dicho estudio será bastante dificultoso, por lo tanto parece razonable buscar una técnica que nos permitiera trabajar con un número de variables considerablemente menor.

En la mayoría de los casos, no todas las variables aportarán información relevante e incluso puede haber varias variables que midan conceptos “similares”.

Por ejemplo, si disponemos de un conjunto de variables económicas en las que haya dos grupos claramente diferenciados, aunque utilicemos un número considerable de variables, realmente estaremos midiendo solamente dos “factores” diferentes, por lo tanto si conseguimos utilizar esos dos factores simplificaríamos mucho nuestro problema.



El modelo matemático consiste en expresar las variables originales del estudio (X_1, \dots, X_n) como combinaciones lineales de una nuevas variables (F_1, \dots, F_p) de la forma:

$$\begin{aligned} X_1 &= \alpha_{11} F_1 + \alpha_{12} F_2 + \dots + \alpha_{1p} F_p + U_1 \\ X_2 &= \alpha_{21} F_1 + \alpha_{22} F_2 + \dots + \alpha_{2p} F_p + U_2 \\ &\vdots \\ X_n &= \alpha_{n1} F_1 + \alpha_{n2} F_2 + \dots + \alpha_{np} F_p + U_n \end{aligned} \tag{1}$$

Para que la reducción de la dimensión del problema sea efectiva, lo ideal es que p fuera una cantidad *muy inferior* a n .

Usualmente en el análisis factorial a las variables F_i se les denomina **factores comunes** y a las variables U_i **factores únicos**

- Estos factores, en general, no los conocemos "a priori", por lo que esta metodología es una técnica exploratoria de la estructura del problema, aunque también la podemos utilizar como una técnica confirmatoria si tenemos una hipótesis previa sobre la estructura del problema.
- Los factores son variables **no observables**.
- Los factores deben aglutinar **informaciones redundantes**.
- Para que la técnica tenga éxito y la reducción de la dimensión sea significativa las variables deben tener ciertos niveles de **correlación**.

Ejemplo 1 Supongamos que a unos individuos se les pasa una batería de 100 preguntas para medir unas ciertas habilidades.

En un principio, el análisis estadístico consistiría en el estudio pormenorizado de las 100 cuestiones considerándolas como variables.

Ante tal magnitud de datos, todos estaremos de acuerdo en que es bastante difícil extraer conclusiones valiosas.

Sin embargo, si hemos diseñado bien la batería de preguntas, habrá grupos de preguntas que midan la misma habilidad.

Mediante un análisis factorial podemos explorar qué preguntas miden una determinada habilidad o, incluso, validar nuestra prueba, ya que todas las preguntas referentes a una determinada habilidad deberían estar incluidas en el mismo factor.

Así pues, en este ejemplo, las habilidades (que no son variables observables) serían los factores.

Si nos planteamos el problema de esta manera nos podemos hacer las siguientes preguntas:

1. ¿Con cuántos factores no quedamos?
2. ¿Cómo se interpretan esos factores?
3. ¿Qué se entiende por poca pérdida de información?

7.2 Metodología general

Los pasos a seguir en un análisis factorial son los siguientes:

1. Estudiar la matriz de correlaciones.
2. Extracción de factores.
3. Rotación de los factores para facilitar la interpretación.
4. Representaciones gráficas.

Por lo tanto, un análisis factorial consiste en el estudio de la matriz de correlaciones (covarianzas) de forma que:

- La mayor parte de la correlación (covarianza) entre las variables es explicadas por los **factores comunes**.
- Cualquier porción de varianza no explicada se asigna a una variable de error (**factor único**).

En análisis factorial se asume que la matriz de correlación (covarianzas) tiene una cierta estructura y puede dividirse en dos partes:

- La parte generada por los factores comunes.
- La parte generada por los errores o factores únicos.

Idea intuitiva

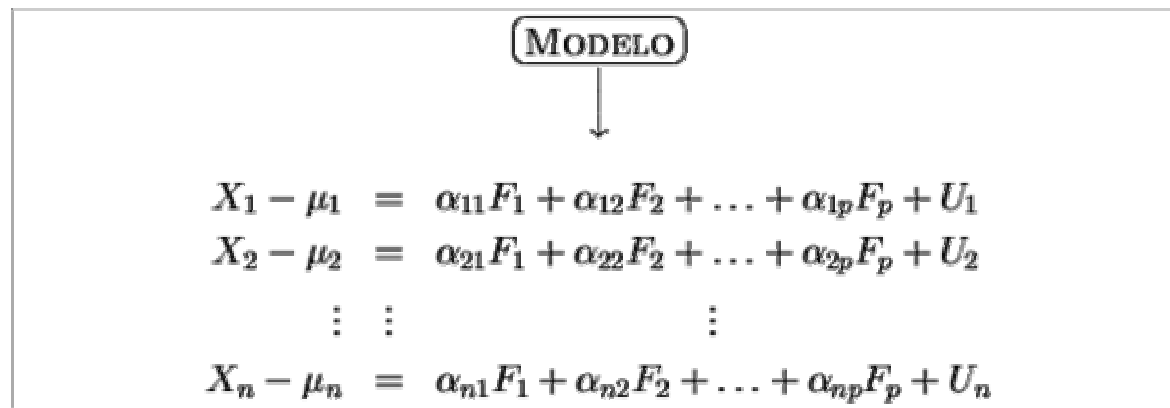
El análisis factorial es un procedimiento que agrupa variables de tal forma que:

- Las variables de cada grupo están **altamente correlacionadas**.
- Los grupos están **relativamente incorrelacionados**.

7.3 Modelo matemático

Veamos ahora el modelo matemático para solucionar este problema. En el análisis factorial aparecen varios tipos de variables:

- n variables **observables** X_1, \dots, X_n , con vector de medias μ ($n \times 1$) y matriz de covarianzas Σ ($n \times n$).
- p variables **no observables** F_1, \dots, F_p donde $p \ll n$
- n variables **no observables** U_1, \dots, U_n



En notación matricial:

$$\mathbf{X} - \mu = \mathbf{A}\mathbf{F} + \mathbf{U} \quad (2)$$

donde

- **X** - μ : vector ($n \times 1$).
- **F**: vector ($p \times 1$) de FACTORES COMUNES **linealmente independientes**.
- **A**: matriz ($n \times p$) cuyos elementos se conocen como CARGAS FACTORIALES.
- **U**: vector ($n \times 1$) de FACTORES 'UNICOS o ERRORES.

7.4 Condiciones del modelo

Las condiciones que tenemos que exigirle a este modelo son:

1. Los FACTORES COMUNES tienen media 0.
2. Los FACTORES COMUNES varianza 1 y no están correlacionados entre sí.
3. Los FACTORES ÚNICOS tienen media 0 y varianza $\sigma_{u_i}^2$.
4. Los FACTORES COMUNES no están correlacionados con los FACTORES ÚNICOS.

La traducción matemática de estas condiciones nos lleva a las siguientes expresiones:

1. $E(\mathbf{F}) = 0$.

2. $E(\mathbf{FF}^t) = \mathbf{I}_{p \times p}$.

3. $E(\mathbf{U}) = 0$ y $E(\mathbf{UU}^t) = \Phi_{n \times n}$

($\Phi_{n \times n}$ es una matriz diagonal, donde los elementos de la diagonal principal son $\sigma_{u_i}^2$).

4. $E(\mathbf{UF}^t) = 0$.

Si se dan todas estas suposiciones la matriz de covarianzas de \mathbf{X} se puede expresar de la forma

$$\Sigma = \mathbf{AA}' + \Phi \quad (3)$$

Con \mathbf{X} de la forma $\mathbf{X} = \mathbf{AF} + \mathbf{U}$.

Si utilizamos esta expresión, la varianza de cada variable X_i se puede escribir como:

$$\sigma_i^2 = \sum_{j=1}^n a_{ij}^2 + \sigma_{u_i}^2$$

$$\text{Var. } X_i = \text{Var. Factores o Comunalidad} + \text{Var. Especifica}$$

Si \mathbf{X} está estandarizada, es decir, $E(\mathbf{X}) = 0$ y $\text{Var}(\mathbf{X}) = \mathbf{1}$, entonces los elementos de \mathbf{A} representan la **correlación** existente entre las variables y los factores.

En este caso (variables estandarizadas), el modelo nos queda de la forma

$$\begin{aligned} \mathbf{X} &= \mathbf{AF} + \mathbf{U} \\ \mathbf{R} &= \mathbf{AA}' + \mathbf{\Phi} \end{aligned}$$

(4)

por lo tanto se cumplen estas propiedades importantes que nos dan una interpretación semántica de los coeficientes que aparecen en un análisis factorial:

- $\mu = 0$, $\sigma_i^2 = 1$ y Σ es la matriz de correlación \mathbf{R} , por ser variables normalizadas.
- $\sum_{j=1}^n a_{ij}^2$ representa la proporción de **varianza explicada** por los factores comunes.
- Las cargas factoriales son las correlaciones entre las variables y los factores.

7.5 Estudio de la matriz de correlación

Dado que el análisis factorial se basa en el estudio de la matriz de correlaciones hay que ver si se dan las condiciones adecuadas para poder aplicar un análisis de este tipo.

¿Cuáles son las condiciones adecuadas para poder aplicar un análisis factorial?

La respuesta a esta pregunta la podemos tener en la siguiente afirmación:

Las variables han de estar altamente correlacionadas y la matriz de correlación ha de tener una cierta estructura.

La comprobación de esta condición se puede hacer utilizando las siguientes técnicas:

- Mediante la visualización de la matriz de correlaciones. Este método es un poco primitivo pero nos puede dar una idea inicial de por donde podemos movernos. La dificultad principal estriba en que si el número de variables del modelo es alto es difícil extraer conclusiones ya que la matriz de correlaciones será muy grande y, por lo tanto, difícil de visualizar en su conjunto.
- Mediante el estudio de unos determinados coeficientes como, por ejemplo, el *Determinante de la matriz de correlaciones*: Valor bajo \Rightarrow alta adecuación del AF.

- *Test de esfericidad de Barlett* $H_0 : \mathbf{R} = \mathbf{I}$ frente a $H_1 : \mathbf{R} \neq \mathbf{I}$.
- La medida de *Kaiser-Meyer-Olkin (KMO)*, basada en estudios de coeficientes de correlación parcial:
KMO alto \Rightarrow alta adecuación del AF (por debajo de 0.5 no es aceptable el AF).
- *Coeficiente de correlación múltiple*: Valor alto \Rightarrow alta adecuación del AF. Este coeficiente coincide con la comunalidad inicial cuando el método de extracción no es el de Componentes principales.
- *La matriz de correlación anti-imagen*: Esta matriz está formada por los negativos de los coeficientes de correlación parcial: Elementos no diagonales pequeños \Rightarrow alta adecuación del AF. (La diagonal principal de esta matriz es la *medida de adecuación muestral (MSA)*: Si estos coeficientes son bajos es desaconsejable el AF. Se puede considerar eliminar variables con MSA bajo, siempre que estos valores sean bajos en unas pocas variables.)

7.6 Extracción de los factores

Después de comprobar que el análisis factorial podría ser una buena solución para analizar nuestro problema, hemos de pasar a la parte técnica de la extracción de dichos factores.

Observación:

Para resolver el análisis factorial debemos estimar los coeficientes de las matrices \mathbf{A} y Φ del problema planteado en la ecuación (3) (en el caso de variables arbitrarias) o en las ecuaciones (4) (en el caso de variables estandarizadas), es decir,

$$\begin{cases} \mathbf{X} = \mathbf{AF} + \mathbf{U} \\ \mathbf{R} = \mathbf{AA}' + \Phi \end{cases}$$

Si analizamos detenidamente estas ecuaciones podemos observar que Σ es una matriz que contiene $n(n + 1)/2$ coeficientes conocidos (por ser simétrica), mientras que los coeficientes desconocidos (correspondientes a \mathbf{A} y a Φ) son $np + n$. Es decir, si $p > (n - 1)/2$, habrá más incógnitas que ecuaciones, por lo que el sistema no tendrá solución única.

Aún en el caso de que hubiera más ecuaciones que incógnitas, el sistema no tiene solución única.

Este problema nos lleva a que haya que diseñar algoritmos para la estimación de estos coeficientes, que diferirán entre sí en la metodología en la búsqueda de las soluciones. Muchos de estos algoritmos son técnicas iterativas, es decir, parten de una solución inicial y en sucesivos pasos se va mejorando la solución hasta llegar a un resultado óptimo (si es posible).

Los métodos de extracción de factores usualmente incluidos en los paquetes estadísticos son:

- Componentes principales.
- Ejes principales.
- Mínimos cuadrados.
- Análisis Alfa.
- Análisis Imagen.
- Máxima verosimilitud.

En este momento se plantea la cuestión de cuál de estos métodos es el mejor.

Realmente no hay una respuesta convincente a esta pregunta por lo que la táctica usual es probar con todos y ver qué método explica mejor nuestro problema.

El método más popular y el que todos los paquetes traen por defecto es el de componentes principales. Si sabemos que nuestros datos se han extraído de una población normal multivariado deberemos elegir el método de máxima verosimilitud.

Una vez que hemos decidido el método de extracción de los factores debemos decidir también el número de factores con el que nos vamos a quedar. Esta elección es importante ya que nuestro objetivo es la reducción de la dimensión del problema sin demasiada pérdida de información.

En los paquetes estadísticos aparecen, esencialmente, dos criterios:

- Considerar los autovalores (eigenvalues) mayores que la unidad. Este es el método por defecto.
- Fijar número de factores según varianza explicada. En este caso nos quedaremos con una cantidad de factores de forma que la varianza explicada por el modelo sea satisfactoria para el investigador.

7.7 Rotación de los factores

El análisis factorial es **inútil** si no se pueden interpretar los factores. Esto es importante ya que los factores son variables **no observables** y si no los podemos interpretar, no podemos extraer conclusiones sobre entes que no sabemos lo que significan.

Para facilitar la interpretación vamos a utilizar la siguiente propiedad que tienen las cargas factoriales

*La estructura del problema **no cambia** si se le aplica una rotación ortogonal (transformación rígida que respeta los ángulos).*

En efecto, si tenemos una matriz ortogonal \mathbf{T} , es decir, $\mathbf{TT}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$, entonces

$$\mathbf{X} = \mathbf{AF} + \mathbf{U} \Rightarrow \mathbf{X} = \mathbf{ATT}'\mathbf{F} + \mathbf{U}$$

($\mathbf{X} = \mathbf{BG} + \mathbf{U}$ donde $\mathbf{B} = \mathbf{AT}$ y $\mathbf{G} = \mathbf{T}'\mathbf{F}$).

La descomposición de matriz de regresión asociada al modelo antes de aplicarle la rotación es

$$\mathbf{R} = \mathbf{AA}' + \Phi$$

y la descomposición asociada al modelo después de efectuar la rotación es

$$\mathbf{R} = \mathbf{BB}' + \Phi = \mathbf{AT(AT)'} + \mathbf{U} = \mathbf{ATT}'\mathbf{A}' + \Phi = \mathbf{AA}' + \Phi$$

Es decir, la descomposición de la matriz de regresión es invariante frente a las transformaciones ortogonales.

Es importante hacer notar que si se aplica una rotación ortogonal los factores siguen estando no correlacionados.

Esta propiedad nos indica que la solución al problema no es única (en este sentido) por lo que debemos escoger la **mejor**.

¿En que sentido podremos hablar de *mejor solución*?

Para el investigador la mejor solución será aquella que mejor explique su problema. Debemos escoger la solución en la que las cargas factoriales sean altas en algunas variables y bajas en otras.

De forma gráfica lo podemos ver en la Figura 1:

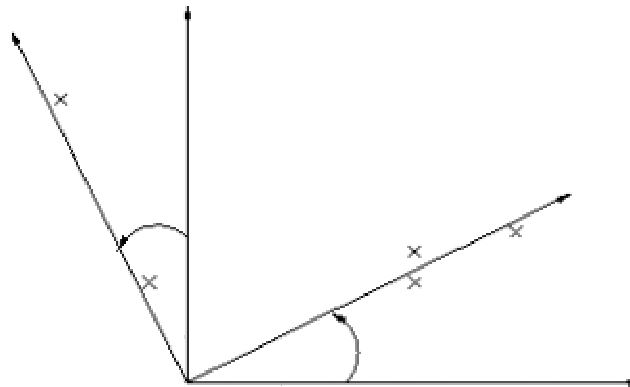


Figura 1: Rotación Ortogonal

Al rotar la solución puntos que tenían valores intermedios en ambos factores pasan a tener valores altos en uno y bajos en otro. Puesto que las cargas factoriales representan la correlación entre las variables y los factores, este hecho nos permitirá “explicar” el significado semántico de los factores.

7.8 Tipos de rotaciones ortogonales

Hay diversos métodos para la realización de rotaciones ortogonales. Cada uno de estos métodos prima la maximización de las cargas en un sentido.

- **Varimax** Hace que se optimice por factores, maximizando la suma de las varianzas de las cargas factoriales al cuadrado dentro de cada factor.
- **Quartimax** Hace que se optimice por variables
- **Equimax** Es un híbrido de las anteriores en intenta optimizar a la vez por variables y por factores.

7.9 Tipos de rotaciones no ortogonales

Hay otro tipo de rotaciones que no son ortogonales y que se conocen como rotaciones oblicuas. Estos métodos se utilizan si las rotaciones ortogonales no producen resultados satisfactorios.

Los más utilizados son los métodos **oblimin** y **promax**.

En este caso después de rotar los factores **sí** están correlacionados y las cargas **no** se pueden interpretar como la correlación entre variables y factores. Por lo tanto habrá que considerar la matriz de correlación entre los factores para comprobar lo que nos alejamos de la no correlación entre factores.

7.10 Puntuaciones factoriales

Después de estimar la matriz de cargas factorial y de haberle dado una explicación semántica a los factores habremos obtenido un conjunto de variables (factores) más reducido que el que teníamos originalmente.

Una vez hecho esto, podemos usar este nuevo conjunto de variables para posteriores análisis estadísticos para lo cual tendremos que ver que valor toma cada individuo o unidad experimental en cada uno de los factores. Por ejemplo, si estamos hablando de un test de capacidad, nos puede interesar ver que puntuación tiene un determinado individuo en la capacidad visual o en su capacidad verbal. Realmente, los valores observados serán los obtenidos en la batería de preguntas y los que nos interesan son los valores obtenidos en los factores contruidos a partir de las preguntas.

A estas puntuaciones las denominaremos **puntuaciones factoriales** y su cálculo no es sencillo ya que disponemos del modelo

$$\mathbf{X} = \mathbf{AF} + \mathbf{U}$$

donde **A** es estimada y **U** desconocida, por lo que, dado un conjunto de observaciones no podemos obtener los valores e **F** de forma explícita.

Lo que en realidad se hace es estimarlos para lo que hay varios métodos propuestos en la literatura. Por ejemplo, los que utiliza el SPSS son:

- El método de Thompson o de regresión.
- El método de Barlett o de mínimos cuadrados ponderados.
- El método de Anderson-Rubin.