

Análisis Factorial Exploratorio

Sesión N° 6
Análisis Avanzado de Datos II

Profesor Gabriel Sotomayor López

Universidad Diego Portales

Contenidos Sesión N° 6

Análisis Factorial Exploratorio (y Análisis de Componentes Principales)

Etapas del análisis factorial

Preparación de los datos y evaluación de supuestos

Contenidos Sesión N° 6

La extracción de factores iniciales

La obtención de la matriz factorial y su interpretación

La evaluación del modelo factorial.

Cálculo de las puntuaciones factoriales

Análisis Factorial Exploratorio (y ACP)

Variables latentes en Ciencias Sociales

En ciencias sociales muchos de los conceptos que usamos no pueden medirse directamente: el autoritarismo, la conciencia de clase, el capital cultural, etc.

Estos corresponden a variable latentes: variables que no puede ser directamente observada o medida, pero se puede inferir a través de la medición de otras variables relacionadas.

Es decir, una variable latente es una construcción teórica o abstracta que no se puede medir directamente, pero que se asume que está influyendo en los datos observados.

Variables latentes en Ciencias Sociales

La medición de variables latentes requiere de que contemos con un **modelo de medición**.

Permiten evaluar la manera en que se relacionan un conjunto de variables observadas y si estas efectivamente miden un concepto teórico de interés:

- Análisis factorial exploratorio
- Análisis factorial confirmatorio

Análisis Factorial Exploratorio

Conjunto de técnicas de análisis **Se busca la síntesis de la información proporcionada por “p” variables observadas (o indicadores), con la menor pérdida posible de información, en un número inferior de “k” variables no observadas** (factores comunes). Esta serie menor de variables latentes ha de caracterizarse por aglutinar variables empíricas que estén bastante correlacionadas entre sí y escasamente correlacionadas con aquellas variables empíricas que conforman otra estructura latente (o dimensión del concepto que se analice).

Análisis Factorial Exploratorio

Tiene dos objetivos principales:

1. Analizar la correlación existente en una serie de variables, con el propósito de descubrir si comparten alguna estructura latente (no directamente observable).
2. La obtención de puntuaciones factoriales, variables típicas o, en su caso, variables sucedáneas, para cada factor. Éstas actuarán en representación de los factores o componentes en análisis posteriores.

Análisis Factorial Exploratorio

$$X_1 = \lambda_{11}F_1 + \lambda_{12}F_2 + \dots + \lambda_{1K}F_K + e_1$$

$$X_2 = \lambda_{21}F_1 + \lambda_{22}F_2 + \dots + \lambda_{2K}F_K + e_2$$

$$X_p = \lambda_{p1}F_1 + \lambda_{p2}F_2 + \dots + \lambda_{pK}F_K + e_p$$

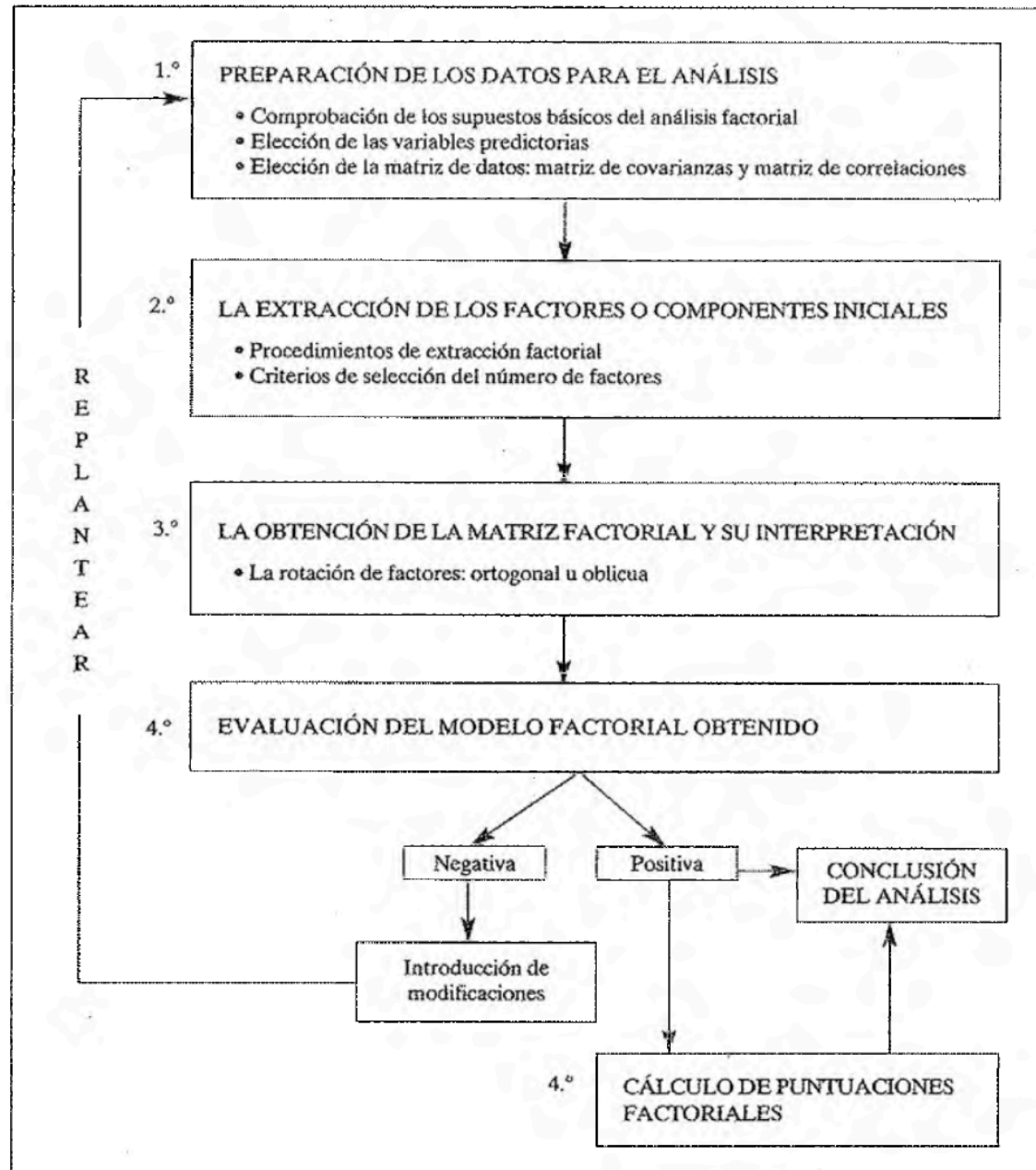
Cada variable observada X_i se expresa mediante una combinación lineal de un número pequeño de factores comunes latentes y un factor único, también latente. Estos últimos representan la parte de la varianza de la variable observada que “no” es explicada por los factores comunes. La elección de la letra “e” para denotar al factor único procede de su consideración como “término de error”.

AFE y ACP

Además del análisis factorial, otra técnica que se usa frecuentemente para la reducción de dimensiones de conjuntos de variables es el Análisis de Componentes Principales.

Se diferencia principalmente porque este último trabaja con la varianza total de las variables, mientras que el AFE utiliza solo la varianza común de las variables, la comunalidad.

Etapas del Análisis Factorial Exploratorio



Preparación de los datos y evaluación de supuestos

Supuestos del AFE

- Nivel de medición
- Tamaño muestral
- Normalidad
- Colinealidad
- Multicolinealidad
- Tratamiento de casos perdidos y casos atípicos

Nivel de medición

Tabla 1. Clasificación de modelos con variables latentes de acuerdo con los niveles de medición

| | | Variables manifiestas | |
|--------------------|-------------|--|---|
| | | Continuas | Categóricas |
| Variables latentes | Continuas | Análisis factorial (<i>Factor analysis</i>) | Análisis de rasgo latente (<i>Latent trait analysis</i>) |
| | Categóricas | Análisis de perfil latente (<i>Latent profile analysis</i>) | Análisis de clases latentes (<i>Latent class analysis</i>) |

El AFE requiere que las variables sean continuas, o variables ordinales que trataremos como tales. Esto puede resolverse también con el tipo de matriz de correlaciones a utilizar

Tamaño Muestral

Como piso mínimo se requieren 5 casos por cada variable que se incluya en el análisis, sin embargo, es preferible contar con al menos 20. Tamaños muestrales mayores ayudarán a obtener estimaciones muestrales estables.

Cómo mínimo se esperan unos 200 o 300 casos en total.

Debe considerarse los casos perdidos y aquellos que puedan ser eliminados por ser casos atípicos.

Normalidad (Multivariante)

Todas las variables observadas y sus combinaciones lineales han de estar distribuidas normalmente. Es decir, se espera que exista normalidad univariada en cada variable y normalidad multivariante.

El uso de procedimientos de extracción habituales en AFC, como los llamados “máxima verosimilitud” (ML) o “mínimos cuadrados” exige el cumplimiento del supuesto de normalidad multivariable.

Cuando no se cumple habrá que buscar que cada variable no sea extremadamente asimétrica (se sugiere coeficientes de asimetría que se encuentren dentro de un intervalo de ± 2).

Colinealidad

Para la extracción de factores comunes debe existir varianza común entre las variables, de lo contrario es poco probable encontrar estructuras latentes relevantes.

Como mínimo, se espera la existencia de correlaciones de al menos 0,3 entre las variables.

Según el nivel de medida pueden utilizarse correlaciones de Pearson (variables continuas o ordinales de suficientes categorías) o policóricas (variables ordinales)

Correlaciones policóricas

Las correlaciones policóricas son una medida de asociación entre variables ordinales que se basa en la teoría de correlación de Pearson y tiene en cuenta la naturaleza discreta de los datos.

Supone una variable subyacente continua y teóricamente normal.

Se recomienda un mínimo de 50 observaciones.

Sólo se interpreta sentido y fuerza, no un p de significación.

En el caso de variables dicotómicas se utilizan correlaciones tetracóricas.

Multicolinealidad

Además de las correlaciones bivariadas existen pruebas estadísticas que nos permiten establecer la existencia de multicolinealidad en el conjunto de variables.

Test de esfericidad de Barlett: Esta prueba se utiliza para determinar la existencia de multicolinealidad en un conjunto de datos. Si el valor-p es significativo (generalmente $<0,05$), entonces se rechaza la hipótesis nula de que la matriz de correlación es igual a una matriz de identidad, lo que indica la presencia de multicolinealidad. **Su correcta interpretación requiere de la existencia de normalidad multivariante**, por lo que en caso de que esta no exista, debe priorizarse la interpretación de la prueba KMO.

Multicolinealidad: KMO

La **prueba KMO (Kaiser-Meyer-Olkin)** se utiliza para evaluar la presencia de multicolinealidad en un conjunto de variables. La prueba produce un valor de MSA (Medida de adecuación de la muestra) para el conjunto de datos y para cada variable individual en el conjunto de datos.

Existen diferentes criterios para interpretar los valores del índice KMO en el AFE , pero algunos de los umbrales comúnmente utilizados son:

- Excelente adecuación: KMO mayor que 0,9.
- Buena adecuación: KMO entre 0,8 y 0,9.
- Adecuación aceptable: KMO entre 0,7 y 0,8.
- Inadecuada adecuación: KMO menor que 0,7.

Es importante tener en cuenta que estos umbrales son solo una guía general y que la interpretación de los resultados del AFE debe basarse en varios criterios, incluyendo el juicio del investigador, la coherencia con la teoría y la interpretación de los patrones de carga factorial.

Tratamiento de variables

Antes de realizar un AFE debemos revisar dos decisiones respecto a los datos: La existencia de **casos atípicos** y de **casos perdidos**.

Casos perdidos:

- Proporción de casos perdidos (10% aprox)
- Distribución aleatoria o no de los casos perdidos
- Imputación de datos: media, regresión lineal, criterio de investigación, imputación múltiple

Tratamiento de variables

Antes de realizar un AFE debemos revisar dos decisiones respecto a los datos: La existencia de **casos atípicos** y de **casos perdidos**.

Casos atípicos:

Debe realizarse un diagnóstico de casos atípicos multivariantes, es decir, aquellos que se alejan del centro medio de las observaciones en un espacio multidimensional. Esto se mide con la distancia de Mahalanobis (distancias mayores a 0,001 se consideran atípicas)

Tratamiento de variables: estandarización

- La estandarización favorece la comparabilidad de variables y es comúnmente utilizada en análisis factorial.
- Se realiza mediante la transformación de las variables a puntajes Z.
- La varianza de las variables depende de su unidad de medida, por lo que la estandarización permite comparar variables con diferentes grados de heterogeneidad.
- La elección de la matriz de entrada para obtener un modelo factorial depende de si las variables están en su unidad original o estandarizadas.

Análisis Factorial Exploratorio

Sesión N° 6
Análisis Avanzado de Datos II

Profesor Gabriel Sotomayor López

Universidad Diego Portales