Tema 35

1. Análisis Factorial. — Introducción (Objetivo) < Exploratorio / Confirmatorio

2. Formulación del Problema. < Formulación / Hipótesis / Propiedades (Comunalidades)

3. Técnicas de resolución. → Métodos para la extracción de factores
   ↳ ~~Determinación~~ nº factores

4. Relación con el Análisis de Componentes Principales. - C.Pérez (pg 193) / D.Peña (pg 380)

5. Rotaciones. < Interpretación geométrica / Ortogonales / Oblicuas / Puntuaciones factoriales

6. Adecuación y Validación de hipótesis.
   ↳ Contrastes previos
   ~~Contrastes a posteriori~~

1. ANÁLISIS FACTORIAL

El análisis factorial tiene por objeto explicar un conjunto de variables observadas por un pequeño número de variables latentes, o no observadas, que llamaremos factores.
Para ello trata de encontrar dimensiones comunes (factores) que liguen las p variables observadas. Concretamente, se trata de encontrar un conjunto de K factores (K<p) no directamente observables $F_1, F_2, \ldots F_K$ que expliquen suficientemente a las variables observadas perdiendo el mínimo de información, de modo que sean fácilmente interpretables (principio de interpretabilidad) y que sean los menos posibles, es decir, K pequeño (principio de parsimonia). Además, los factores han de extraerse de forma ~~todo~~ que sean ortogonales.
En consecuencia, el análisis factorial es una técnica de reducción de datos que examina la interdependencia de variables y proporciona conocimiento de la estructura subyacente de los datos.
El análisis factorial puede aplicarse como una herramienta exploratoria o como un modelo para contrastar teorías, ~~en este~~ herramienta confirmatoria.
En este segundo caso, el número de factores se supone conocido a priori. ~~y se establecen restricciones sobre los~~
El A.F. surge por el interés de Karl Pearson y Charles Spearman en comprender las dimensiones de la inteligencia humana en los años 30 del siglo XX, y muchos de sus avances se han producido en el área de la psicometría.

2. FORMULACIÓN DEL PROBLEMA

Consideramos las variables observables $X_1, X_2, \ldots, X_q$ como variables tipificadas.
El modelo factorial formaliza la relación entre variables observables y factores de la siguiente forma:

$$X_1 = \lambda_{11} F_1 + \lambda_{12} F_2 + \ldots + \lambda_{1K} F_K + u_1$$
$$X_2 = \lambda_{21} F_1 + \lambda_{22} F_2 + \ldots + \lambda_{2K} F_K + u_2$$
$$\vdots$$
$$X_p = \lambda_{p1} F_1 + \lambda_{p2} F_2 + \ldots + \lambda_{pK} F_K + u_p$$

d· donde:

$F_1, F_2, \ldots F_K$ son los factores comunes
$u_1, u_2, \ldots u_p$ son los factores únicos o fact. específicos
λ_{jh} es el peso del factor h en la variable j, o carga factorial, o saturación.

La expresión matricial es:

$$X = \Lambda F + u$$

Si las variables no están tipificadas

$$X = \mu + \Lambda F + u$$

la media de las var.

Todas las variables origi
por todos los factores comu
variable ~~en~~ tiene un fac
esa variable que recoge
ha sido explicada por

las
la
sena
no

Hipótesis básicas

- los factores comunes siguen una distribución $N_m(0, I)$, es decir, son variables de media cero e independientes entre sí y con dist. normal.

- El vector de factores únicos (o específicos) es un vector de perturbaciones no observadas, con distribución $u \overset{d}{\rightsquigarrow} N_p(0, \psi)$ t.q. ψ es diagonal (e.d., las var. de los fact únicos pueden ser ≠s y dichos factores han de estar incorrelacionados entre sí.)

- las perturbaciones (factores únicos) están incorreladas con los factores comunes.

Como consecuencia de estas tres hipótesis X tiene distribución normal al ser suma de var. normales.

- $X = \Lambda F + u$, Λ = matriz de carga cuyos coeficientes constantes relacionan las variables y los factores. Λ contiene las covarianzas entre los fact. y las var. obs. y si X son var. estandarizadas, ~~se~~ son las correlaciones

- la matriz de covarianzas de las observaciones verifica:

$$V = \underbrace{\Lambda \Lambda'}_{\text{Parte común}} + \underbrace{\psi}_{\text{Parte específica de cada var.}} \quad \text{(descomposición de la varianza)}$$

$$\Rightarrow \underbrace{\sigma_i^2}_{\text{Var. de } x_i} = \sum_{j=1}^{m} \lambda_{ij}^2 + \psi_i^2 \quad , \quad i = 1, 2, \ldots, p$$

donde:

$$h_i^2 = \sum_{j=1}^{m} \lambda_{ij}^2$$ ≡ Comunalidad: suma de los efectos de los factores

φ_i^2 ≡ efecto de la perturbación.

Varianza observada = Variabilidad común (Comunalidad) + Variabilidad específica

$$\underbrace{(1 =)}_{\text{Var. tipificadas}} \quad \sigma_i^2 = h_i^2 + \varphi_i^2$$

- Unicidad del modelo: el modelo factorial está indeterminado ante rotaciones. Esta indeterminación se resuelve imponiendo restricciones sobre las componentes de la matriz de carga.

- Número máximo de factores:
A partir de la ecuación $V = \Lambda\Lambda' + \Psi$, sustituyendo la matriz teórica de covar. (V) por la muestral (S), se obtiene, al identificar si es posible resolverlo de manera única, una restricción en el número de factores posibles:

$$(p-k_1)^2 \geq p+k$$

Esta ec. implica que si p es pequeña ($p \leq 10$), el número máx. de factores debe ser menor que la mitad del nº de var. menos uno.

3. TÉCNICAS DE RESOLUCIÓN. (MÉTODOS PARA LA EXTRACCIÓN DE FACTORES).

3. 1 Método del factor principal

El método del factor principal es un método para estimar la matriz de carga basado en componentes principales. Evita tener que resolver las ecuaciones de máxima verosimilitud, que son más complejas. Tiene la ventaja de que la dimensión del sistema puede identificarse de forma aproximada.

Su base es la siguiente: supongamos que se puede obtener una estimación inicial de la matriz de varianzas de las perturbaciones ($\hat{\varphi}$). Entonces:

* estimación de las comunalidades.

$$S - \hat{\varphi} = \Lambda\Lambda'$$

$$S - \hat{\varphi} = HGH' = (HG^{1/2})(HG^{1/2})'$$

$S - \hat{\varphi}$ es simétrica

donde: $H_{p\times p}$ y ortogonal

$G_{p\times p}$, diagonal y contiene las raíces características de $S-\hat{\varphi}$

El modelo factorial establece que G debe ser diagonal del tipo:

$$G = \begin{bmatrix} G_{1(m\times m)} & 0_{m\times(p-1)} \\ 0_{(p-m)\times m} & 0_{(p-m)\times(p-m)} \end{bmatrix}$$

ya que $rg(S-\hat{\varphi}) = m$

Si llamamos $H_{1(p\times m)}$ a la matriz que contiene los vectores propios asociados a los valores propios no nulos de G_1 $\Rightarrow$

$$\hat{\Lambda} = H_1 G_1^{1/2}$$

tq. las columnas de $\hat{\Lambda}$ son ortogonales entre sí.

y el problema queda resuelto.

Los estimadores obtenidos serán consistentes pero no eficientes, como en máxima verosimilitud. Tampoco son invariantes ante transformaciones lineales como la MV.

* Estimación de las comunalidades

Estimar los términos diagonales de φ (φ_{ii}), equivale a definir valores para los términos diagonales, h_i^2, de $\Lambda\Lambda'$ ($h_i^2 = s_i^2 - \hat{\varphi}_{ii}$).

Hay dos alternativas:

a) Tomar $\hat{\varphi}_{ii} = 0$, que es el valor máximo de la comunalidad ($\hat{h}_i^2 = s_i^2$) (ó $h_i^2 = 1$ si correlaciones) por lo que podemos comenzar con un sesgo importante.

b) Tomar $\hat{h}_j^2 = s_j^2 - s_j^2(1 - R_j^2) = s_j^2 R_j^2$

donde R_j^2 = coeficiente de correlación múltiple entre x_j y el resto de las variables

Este método suele proporcionar una estimación sesgada a la baja.

3.2 Método Alpha

Este método determina la matriz factorial (Λ) especificando un número K de factores comunes y formando la matriz $T_{(p\times K)}$ con los autovectores unitarios correspondientes a los K primeros vectores propios de la matriz $H^{-1}\Lambda\Lambda' H^{-1}$, donde H^2 es la matriz de comunalidades (comunalidades en la diag. principal).

Si $D_\lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots \lambda_K)$ entonces, al diagonalizar $H^{-1}\Lambda\Lambda' H^{-1}$:

$$H^{-1}\Lambda\Lambda' H^{-1} = T D_\lambda T' \Rightarrow \Lambda\Lambda' = H T D_\lambda T' H' \quad (H' = H) \Rightarrow$$

$$\Rightarrow \Lambda = H T D_\lambda^{1/2}$$

3.3. Método del centroide

En el método del centroide se elige el primer factor de modo que pase por el centro de gravedad (centroide $C = \left(\frac{1}{p} \sum_{j=1}^{p} \lambda_{j1}, \ldots, \frac{1}{p} \sum_{j=1}^{p} \lambda_{jk} \right)$) de las variables sin unicidades $X_i' = X_i - U_i$. Entonces los pesos o saturaciones del primer factor serán:

$$\sum_{j=1}^{p} r_{hj} = \lambda_{h1} \sum_{j=1}^{p} \lambda_{j1} \Rightarrow \lambda_{h1} = \frac{S_h}{\sum_{j=1}^{p} \lambda_{j1}} = \frac{S_h}{\sqrt{T}} \qquad h = 1 \ldots p$$

Considerando ahora las correlaciones $r_{ij}' = r_{ij} - \lambda_{i1} \lambda_{j1}$; $i,j = 1 \ldots p$, las componentes de las var. en los restantes factores serán:

$$\begin{array}{c} (\lambda_{12} \ldots \lambda_{1k}) \\ \vdots \\ (\lambda_{p2} \ldots \lambda_{pk}) \end{array}$$

Ahora se elige el segundo factor de modo que pase por el origen y por C. Repitiendo el proceso anterior se obtiene:

$$\lambda_{h2} = \frac{S_h S_{1h}}{\sqrt{T_1}} \qquad h = 1 \ldots p$$

El proceso para obtener los demás factores es exactamente el mismo.

3.4. Método de las componentes principales

Se dispone de una muestra n acerca de p variables correlacionadas y se obtiene a partir de ellas un número $k \leq p$ de var. incorrelacionadas $Z_1 \ldots Z_p$ que sean combinación lineal de las variables iniciales, y que expliquen la mayor parte de ~~la~~ su variabilidad

$$Z = \alpha X \Rightarrow X = \alpha^{-1} Z$$

> En Pérez pag 168 pone matriz transpuesta no inversa

Puede que las componentes z_j no estén tipificadas, condición exigida a los factores $\Rightarrow$

$$\Rightarrow Y_j = \frac{z_j}{\sqrt{s_j}} \Rightarrow$$

$$\Rightarrow X_j = \alpha_{1j} Y_1 \sqrt{s_1} + \dots + \alpha_{pj} Y_p \sqrt{s_p} \Rightarrow$$

$$\Rightarrow \lambda_{ji} = \underbrace{\alpha_{ij}\sqrt{s_j}}_{= z_{ij}} \equiv \text{Cargas factoriales.}$$

Se observa que los K factores se estiman mediante las K primeras componentes principales tipificadas, pudiéndose estimar la comunalidad:

$$\hat{h}_j^2 = \lambda_{j1}^2 + \dots + \lambda_{jk}^2$$

y el factor único como:

$$u_j = r_{k+1,j} Y_{k+1} + r_{m+2,j} Y_{k+2} + \dots + r_{pj} Y_p$$

y la especificidad o parte de la varianza debida al factor único se estima como:

$$\hat{\psi}_j^2 = 1 - h_j^2$$

35. <u>Método de las componentes principales iteradas o ejes principales</u>

Método iterativo similar al de comp principales

- Cálculo de la matriz de correlación muestral: $R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & & \vdots \\ \vdots & & \ddots & \\ r_{p1} & \cdots & \cdots & 1 \end{bmatrix}$
- Estimación inicial de las comunalidades de cada var mediante el coef. de determinación obtenido en la regresión de cada var. sobre el resto de var.
- Se sustituye en R la diag principal por la estimación de la comunalidad de cada var: $R^* = \begin{bmatrix} h_1^2 & r_{12} & \cdots & r_{1p} \\ r_{21} & h_2^2 & \cdots & r_{2p} \\ \vdots & & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & h_p^2 \end{bmatrix}$

- Se calculan las raíces características y los vectores característicos asociados a R^* $\Rightarrow$ se obtienen las cargas factoriales λ_{jh}
- Se determinan los K factores a retener como en comp. principales y se calcula la comunalidad para cada var. con los K fact. retenidos:

$$h_j^2 = \lambda_{j1}^2 + \dots + \lambda_{jK}^2$$

y la especificidad:

$$\psi_j^2 = 1 - h_j^2 \quad j = 1, \dots p$$

(!) Si $\exists\ \psi_j^2 < 0 \Rightarrow$ el método no es aplicable

- Se itera el proceso partiendo de la nueva matriz R^* con las comunalidades recién estimadas
- El proceso se detiene cuando la diferencia entre las comunalidades estimadas para cada var. en dos iteraciones sucesivas sea menor que una cantidad prefijada.

3.6 Método de Máxima Verosimilitud

Es un método estrictamente estadístico basado en la teoría de la inferencia.
Se añade la hipótesis $\bar{x}_1, \bar{x}_2, \dots \bar{x}_n$ m.a.s. $\xrightarrow{d} N_p(\bar{\mu}, \Sigma)$ con $\Sigma = \Lambda\Lambda' + \Omega$.

$$L(\Sigma, \bar{x}_1, \dots \bar{x}_n) = \frac{1}{(2\pi)^{np/2} |\Sigma|^{n/2}} e^{-\frac{n}{2}\,\text{traza}(\Sigma^{-1}S)}$$

Para obtener la estimación de Λ se maximiza $\ln[L(\Sigma, \bar{x}_1, \dots \bar{x}_n)]$

3.7. Métodos MINRES, ULS y GLS

- El mét. MINRES o análisis factorial de correlaciones calcula la matriz factorial Λ que minimiza las diferencias entre la correlación observada y la deducida del modelo factorial, utilizando el criterio de los mínimos cuadrados

$$\min \bar{r}_{hs} = r_{hs} - \sum_{t=1}^{k} \lambda_{ht} \lambda_{jt} \quad h \neq s$$

- El mét. de los mínimos cuadrados no ponderados (ULS) minimiza la fc. : $U(L, \Omega) = \text{traza}\,(S - \Sigma)^2/2$

- El mét. de los mínimos cuadrados generalizados (GLS) minimiza la fc : $G(L, \Omega) = \text{traza}\,(I - S^{-1}\Sigma)/2$

Determinación nº factores:
Si se usa mét. comp. principales
⇒ Graf sedim / % var. / autov >0
Si se usa otro método ↑ no.
Hay contrastes de verosimilitud y criterios de selección para determinar nº de factores.
ver Peña pg. 368-371

4. RELACIÓN CON EL ANÁLISIS DE COMPONENTES PRINCIPALES.

En análisis factorial y componentes principales las variables tienen que ser cuantitativas y los factores o componentes deben de ser suficientes para resumir la mayor parte de la información contenida en las variables originales.

La diferencia entre análisis en componentes principales y análisis factorial radica en que en el análisis factorial se trata de encontrar variables sintéticas latentes, inobservables y aún no medidas cuya existencia se sospecha en las variables originales y que permanecen a la espera de ser halladas, mientras que en el análisis en componentes principales se obtienen variables sintéticas combinación de las originales y cuyo cálculo es posible basándose en aspectos matemáticos independientes de su interpretabilidad práctica.

En el análisis en componentes principales la varianza de cada variable original se explica completamente por las variables cuya combinación lineal la determinan sus componentes. Pero esto no ocurre en el análisis factorial.

En el análisis factorial sólo una parte de la varianza de cada variable original se explica completamente por las variables cuya combinación lineal la determinan los factores comunes. Esta parte de la variabilidad de cada variable original explicada por los factores comunes se denomina comunalidad, mientras que la parte de varianza no explicada por los factores comunes se denomina unicidad (comunalidad + unicidad = 1) y representa la parte de variabilidad propia f_i de cada variable x_i. Cuando la comunalidad es unitaria (unicidad nula) el análisis en componentes principales coincide con el factorial. Es decir, el análisis en componentes principales es un caso particular del análisis factorial en el que los factores comunes explican el 100% de la varianza total.

A parte de estas diferencias, ambas técnicas tienen una interpretación diferente: en componentes principales tratamos de representar gráficamente los datos, mientras que en análisis factorial suponemos que los factores generan la variables observadas.

5. ROTACIONES.

Rotaciones ortogonales

La matriz de carga no está identificada ante multiplicaciones por matrices ortogonales, que equivalen a rotaciones. En análisis factorial está definido el espacio de las columnas de la matriz de carga, pero cualquier base de este espacio puede ser una solución. Para elegir entre las posibles soluciones, se tiene en cuenta la interpretación de los factores. Intuitivamente, será más fácil interpretar un factor cuando se asocia a un bloque de variables observadas. Esto ocurrirá si las columnas de la matriz de carga, que representan el efecto de cada factor sobre las variables observadas, contienen valores altos para ciertas variables y pequeños para otras. Esta idea puede plantearse de distintas formas que dan lugar a distintos criterios para definir la rotación.

- Método Varimax: El método Varimax obtiene los efes de los factores maximizando la suma de varianzas de las cargas factoriales al cuadrado dentro de cada factor. Una propiedad de este método es que, después de aplicado, queda inalterada, tanto la varianza total explicada por los factores, como la comunalidad de cada una de las variables. La nueva matriz corresponde también a factores ortogonales y tiende a simplificar la matriz factorial por columnas, siendo muy adecuada cuando el número de factores es pequeño.

- Método Quartimax: Este método maximiza la suma de las cuartas potencias de todas las cargas factoriales. La nueva matriz corresponde también a factores ortogonales y tiende a simplificar la matriz factorial por filas, siendo muy adecuada cuando el número de factores es elevado.
- Métodos Ortomax: Estos métodos consideran una solución intermedia a los métodos Varimax y Quartimax. Realmente sólo existen dos métodos distintos para conseguir rotaciones ortogonales que se aproximen a la estructura simple.

Rotaciones oblicuas

El modelo factorial está indeterminado no sólo ante rotaciones ortogonales sino ante rotaciones oblicuas. Es decir, el modelo puede establecerse con factores incorrelados o correlados. La solución obtenida de la estimación de la matriz de cargas factoriales corresponde siempre a factores incorrelados, pero podemos preguntarnos si existe una solución con factores correlados que tenga una interpretación más interesante. El problema de las rotaciones oblicuas es que los factores, al estar correlados, no pueden interpretarse independientemente.

- Método Oblimax: Para obtener la matriz de cargas factoriales se empieza rotando un par de factores, esto se repite para todos los pares completando un ciclo, ciclos que se repiten hasta que me alcanza el óptimo.
- Método Quartimin: El proceso numérico para hallar la matriz de cargas factoriales exige una larga iteración, en la que cada paso es la obtención de los vectores propios de una matriz simétrica.
- Métodos Oblimin: Se trata de la adaptación de la rotación Varimax al caso oblicuo.

Puntuaciones factoriales

El análisis factorial es en muchas ocasiones un paso previo a otros análisis, en los que se sustituye el conjunto de variables originales por los factores obtenidos. Por ejemplo en el caso de estimación de modelos afectados de multicolinealidad. Por ello, es necesario conocer los valores que toman los factores en cada observación.

Las puntuaciones exactas para los factores se obtienen únicamente en el caso de que se haya aplicado el análisis de componentes principales para la extracción de factores. En los demás casos es necesario realizar estimaciones para obtenerlas. Estas estimaciones se pueden realizar por distintos métodos. Los procedimientos más conocidos son los de mínimos cuadrados, regresión , Anderson-Rubin y Barlett.

En el método de regresión las puntuaciones de los factores obtenidas pueden estar correlacionadas, aun cuando se asume que los factores son ortogonales, y la varianza de las puntuaciones de cada factor puede no ser igual a 1. Con el métodos de Anderson-Rubin se obtienen puntuaciones de factores que están incorrelacionadas y que tienen varianza 1. Finalmente, en el método de Barlett se aplica el método de máxima verosimilitud, haciendo el supuesto de que los factores tienen una distribución normal con media y matriz de covarianzas dadas.

6. ADECUACIÓN Y VALIDACIÓN DE HIPÓTESIS. (Contrastes en el modelo factorial).

Para estudiar la adecuación de un conjunto de variables observadas para realizar un análisis factorial se pueden observar diferentes medidas estadísticas, así como realizar contrastes previos a la extracción de los factores en los que trata de analizarse la validez de la aplicación del análisis factorial a un conjunto de variables observables.

Matriz de correlación muestral

La matriz de correlación muestral proporciona la correlación de las variables dos a dos observada en la muestra recogida. Si la correlación de cada par de variables es alta (mayor de 0,6) entonces las variables observadas serán adecuadas para realizar un análisis factorial.

Determinante de la matriz de correlación muestral

Si el determinante de la matriz de correlación muestral es cercano a cero entonces las variables observadas serán adecuadas para realizar un análisis factorial.

Contrastes de independencia de las variables observadas

Estos contrastes estudian si las variables observadas son independientes dos a dos. Se realizan tantos contrastes como pares de variables haya.

H_o: X_j y X_h son independientes

H_1: X_j y X_h no son independientes

Si se acepta la hipótesis nula en un número alto de contrastes entonces las variables observadas no son adecuadas para realizar un análisis factorial.

Contraste de esfericidad de Barlett

Estudia si las variables observadas están correlacionadas entre sí. Para ello compara el determinante de la matriz de correlación poblacional con el determinante de la matriz identidad (=1).

H_o: $|R_p| = 1$

H_1: $|R_p| \neq 1$

Si se acepta la hipótesis nula, es decir, R_p es significativamente igual a 1, se concluirá que las variables no están correlacionadas entre sí en cuyo caso no tendría sentido aplicar el análisis factorial. En caso contrario, si la matriz de correlación poblacional es significativamente distinta a 1, sí tendría sentido realizar el análisis.

(Ver desarrollo del contraste en César Pérez pg. 175)

Medida KMO de Kaiser, Meyer y Olkin de adecuación muestral global al modelo factorial

La medida KMO es un indicador global de la correlación existente entre todas las variables observadas. Está basada en los coeficientes de correlación observados de cada par de variables y en sus coeficientes de correlación parcial mediante la expresión siguiente:

$$KMO = \frac{\sum_j \sum_{h \neq j} r_{jh}^2}{\sum_j \sum_{h \neq j} r_{jh}^2 + \sum_j \sum_{h \neq j} a_{jh}^2}$$

r_{jh} son los coeficientes de correlación observados entre las variables X_j y X_h

a_{jh} son los coeficientes de correlación parcial entre las variables X_j y X_h

En el caso de que exista adecuación de los datos a un modelo de análisis factorial, el término del denominador será pequeño y, en consecuencia, la medida KMO será próxima a la unidad. Valores de KMO por debajo de 0,6 (0,5 según autores) no serán aceptables, considerándose inadecuados los datos para un modelo de análisis factorial. Para valores superiores a 0,6 (0,5 según autores) se considera aceptable la adecuación de los datos a un modelo de análisis factorial. Mientras más cerca estén de 1 los valores de KMO mejor es la adecuación de los datos a un modelo factorial, considerándose ya excelente la adecuación para valores de KMO próximos a 0,9.

$$MSA_j = \frac{\sum_{h \neq j} r_{jh}^2}{\sum_{h \neq j} r_{jh}^2 + \sum_{h \neq j} a_{jh}^2}$$

Medida MSA de adecuación muestral individual al modelo factorial

La medida MSA es un indicador de la correlación de ~~una~~ cada variable con todas las demás observadas. Está basada en la medida KMO. Se define de la siguiente forma:

MSA_1

MSA_k

Si el valor de MSA_j se aproxima a la unidad, la variable X_j será adecuada para su tratamiento en el análisis factorial con el resto de las variables.