UNIDAD DIDÁCTICA 2: Reducción de la dimensión
Tema 2: Análisis Factorial Exploratorio

TÉCNICAS ESTADÍSTICAS PARA EL APRENDIZAJE I

Máster Universitario en Estadística Computacional
y Ciencia de Datos para la Toma de Decisiones

UNIVERSITAS Miguel Hernández

Índice

Introducción

Introducción

- En muchas áreas de Ciencias Sociales, de Psicología, Sociología y ciencias afines, surgen muchas ocasiones en las que no es posible medir directamente ciertos conceptos (variables) de interés. Dos ejemplos obvios son la inteligencia y la clase social.

- En estos casos, se recoge información sobre variables para obtener indicadores de los conceptos a medir, y luego se trata de descubrir si las relaciones entre esas variables observadas son consistentes con esas otras variables no medibles de interés.

Diferencias y similitudes Análisis Factorial y Análisis de Componentes Principales

- El **punto de partida** del Análisis factorial (AF) **es el mismo** que el del Análisis de componentes principales (ACP).

- El AF busca esencialmente nuevas variables o factores que expliquen los datos, obtener unos **pocos factores no medibles que resulten interpretables**, pero **el formato matemático del AF y el de ACP es diferente**. En el ACP, en realidad, sólo se hacen transformaciones ortogonales de las variables originales, haciendo hincapié en la varianza de las nuevas variables. En el AF, por el contrario, interesa más explicar la estructura de las covarianzas entre las variables.

Introducción

- El **ACP tiene solución única**, la que proporciona, como hemos visto, los valores propios y los vectores propios, sin embargo, **en el AF, no hay una única solución, hay distintas formas de extracción de los factores** y cada una de ellas da lugar a un resultado distinto.

- El **ACP es un herramienta descriptiva**, mientras que el **AF presupone un modelo estadístico formal** de generación de la muestra dada.

- Al igual que en el ACP, para efectuar el AF, es necesario que **las variables originales estén correladas**.

Pruebas de adecuación del Análisis Factorial

Prueba de esfericidad de Barlett

La prueba de esfericidad de Bartlett contrasta si la matriz de correlaciones es una matriz identidad. El estadístico de Bartlett se obtiene a partir de una transformación del determinante de la matriz de correlaciones y cuanto mayor sea, más improbable es que la matriz sea una matriz identidad.

$$H_0 : P = I$$
$$H_1 : P \neq I$$

$$\chi^2_{\left(0.5\left(p^2-p\right)\right)} = -n\left[n-1-\frac{1}{6}\left(2p+5\right)\right]\ln\left|R\right|$$

KMO

La medida de la adecuación muestral de **Kaiser-Meyer-Olkin** (*Coeficiente KMO*) contrasta si las correlaciones parciales entre las variables son pequeñas. Toma valores entre 0 y 1, e indica que el Análisis Factorial es más adecuado cuanto mayor sea su valor.

Para:

KMO $\geq$ 0.9, el test es muy bueno

KMO $\geq$ 0.8: Notable

KMO $\geq$ 0.7: Mediano

KMO $\geq$ 0.6: Bajo

KMO $<$ 0.5: Muy bajo

$$KMO_j = \frac{\sum\sum_{i\neq j} r_{ij}^2}{\sum\sum_{i\neq j} r_{ij}^2 + \sum\sum_{i\neq j} a_{ij}^2}$$

$R = \left[r_{ij} \right]$ Matriz de correlaciones

$A = \left[a_{ij} \right]$ Matriz de correlaciones parciales

Modelo Factorial

Modelo Factorial

El AF está basado en si las covarianzas o correlaciones entre un conjunto de variables observadas $X' = (X_1, \ldots, X_p)$ pueden ser explicadas entre un conjunto de variables latentes no observables $f_1, \ldots, f_k$ con $k < p$.

$$X_1 = \lambda_{11} f_1 + \lambda_{12} f_2 + \ldots + \lambda_{1k} f_k + u_1$$

$$X_2 = \lambda_{21} f_1 + \lambda_{22} f_2 + \ldots + \lambda_{2k} f_k + u_2$$

$$\ldots = \ldots$$

$$X_p = \lambda_{p1} f_1 + \lambda_{p2} f_2 + \ldots + \lambda_{pk} f_k + u_p,$$

$$\mathbf{X} = \Delta \mathbf{f} + \mathbf{u}$$

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

Δ es una matriz de parámetros a estimar.
f son vectores aleatorios (se denominan **factores o componentes comunes**)
u se denominan componentes específicas o **factores únicos**.

Ejemplo:

Las asignaturas clásicas de la enseñanza media, se dividen, en líneas generales, en asignaturas de Ciencias o de Letras, las primeras con contenido más racional y empírico, las segundas con contenido más humanístico y artístico. Consideremos las siguientes 5 asignaturas:
Ciencias Naturales (CNa),Matemáticas (Mat), Francés (Fra), Latín (Lat), Literatura (Lit). Supongamos que están influidas por dos factores comunes o variables latentes: Letras (MR1) y Ciencias (MR2) y en otras palabras, suponemos que MR1 (Letras) y MR2 (Ciencias) son dos variables no observables, que de manera latente influyen sobre las cinco asignaturas. Las calificaciones de n = 20 alumnos en las asignaturas y en los factores se encuentran en la siguiente tabla:

	MR1	MR2
CienciasNaturales	0.44	0.64
Matematicas	0.04	1.00
Frances	0.97	0.07
Latin	0.81	0.17
Literatura	0.86	0.30

Las dos primeras asignaturas están más influidas por el factor MR2 (Ciencias), y las tres últimas por el factor MR1 (Letras).

Matemáticas tiene un coeficiente de 1.0 con Ciencias y sólo 0.04 con Letras.

Hipótesis básicas

Hipótesis básicas

1. Por simplicidad, asumiremos que el vector de medias para X es el vector nulo.

$$E(X) = 0.$$

2. Δ es una matriz (p x k) de constantes desconocidas (k < p). Contiene los coeficientes que describen como los factores f, afectan a las variables observadas, x, y se denomina matriz de carga (o matriz de saturaciones). Relacionan las variables y los factores.

Hipótesis básicas

3. f es un vector (k x 1). Supondremos que son variables de media cero. Variables estandarizadas que están incorreladas entre sí.

$$E(f) = 0 \qquad V(f) = I \qquad Cor(\mathbf{f}_i, \mathbf{f}_j) = 0$$

4. u es un vector (p x 1) que recoge el efecto de todas las variables distintas de los factores que influyen sobre x. Supondremos que E(u) = 0 y que

$$V(u) = \Psi = \mathrm{diag}\{\psi_1, \ldots, \psi_p\} \quad Cor(u_i, \mathbf{u}_j) = 0 \qquad Cor(\mathbf{f}_i, \mathbf{u}_j) = 0$$

Hipótesis básicas

La matriz de covarianzas entre las observaciones verifica:

$\Sigma =$ E[XX′] = E [(Δ**f** + **u**) (Δ**f** + **u**)′] **=** **E[** Δ**ff′**Δ' + uf′Δ' + Δ**f u′ + uu′]** =

= Δ **E[ff′]**Δ' + E[uf′] Δ'+ ΔE[fu′]+ E[uu′]

Como **E[ff′] = I,** E[uf′]=0, E[fu′]=0, se tiene que:

$$\Sigma = \Delta\Delta' + E[uu'] = \Delta\Delta' + \psi.$$

$\Delta\Delta'$ es una matriz simétrica de rango $k < p$. Esta matriz contiene la parte común al conjunto de las variables y depende de las covarianzas entre las variables y los factores.

ψ es diagonal y contiene la parte específica de cada variable, que es independiente del resto.

Hipótesis básicas

Esta descomposición implica que las varianzas de las variables observadas (V(X)) pueden expresarse cómo:

$$\sigma_i^2 = \sigma_{ii} = \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_i,$$ si las variables están estandarizadas, tendremos $$1 = \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_i$$

$$h_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2,$$ **comunalidad** (variabilidad de X explicada por los factores comunes)

ψ_i **unicidad o varianza específica**

(variabilidad de X que no es capaz de explicar los factores comunes)

Hipótesis básicas

Los valores $\boldsymbol{\lambda_{ij}}$ **se denominan cargas factoriales**. De hecho, cuando las variables X_i están estandarizadas, estas cargas representan las correlaciones entre los factores y las variables observadas:

$$Cor(X_i, f_j) = Cov(X_i, f_j) = \lambda_{ij},$$

y por tanto la importancia relativa del factor f_j en la explicación de la variable X_i.

Ejemplo:

Las cargas factoriales (saturaciones), comunalidades y unicidades del ejemplo de las asignaturas:

	MR1	MR2	h2	u2
CienciasNaturales	0.44	0.64	0.60	0.3990
Matematicas	0.04	1.00	1.00	0.0042
Frances	0.97	0.07	0.95	0.0461
Latin	0.81	0.17	0.69	0.3070
Literatura	0.86	0.30	0.83	0.1712

h_{CN}^2=0.60, el 60% de la variabilidad de Ciencias Naturales está explicada por los factores comunes

El 40% restante de su variabilidad no es capaz de explicarlo por los factores comunes (unicidad)

Estimación de los factores

Estimación de los factores

Estimar los parámetros de los valores de $\hat{\Delta}$ y $\hat{\Psi}$ tales que la matriz de covarianzas muestral S sea aproximadamente:

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

Cuando la matriz X está estandarizada , tendremos $P \approx \hat{\Delta}\hat{\Delta}' + \hat{\Psi}$

Llamaremos matriz de correlaciones reducida, a la matriz: $P^* = P - \hat{\Psi}$

$$P* = \begin{pmatrix} h_1^2 & \rho_{12} & \dots & \rho_{1p} \\ \rho_{21} & h_2^2 & \dots & \rho_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{p1} & \rho_{p2} & \dots & h_p^2 \end{pmatrix} \quad \text{donde} \quad P^* = \Delta\Delta'$$

Existen varios métodos de estimación de los términos anteriores: el método de los factores principales, método de máxima verosimilitud, método de mínimos cuadrados no ponderados, método imagen, etc.

Método de los factores principales

Técnica basada en autovalores y autovectores pero en lugar de operar sobre la matriz de correlaciones se opera sobre la matriz de correlaciones reducida.

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

La diagonal de P^* contiene las comunalidades estimadas.

El procedimiento es iterativo:

1. Se parte de unas comunalidades estimadas a partir del coeficiente de determinación de cada variable sobre el resto de variables, (R_i^2), se sustituye la diagonal de la matriz de correlaciones por esta primera estimación de las comunalidades.
2. Se efectúa un análisis de componentes principales sobre esta matriz P^* (se calculan los valores y vectores propios de P^* , y se determina el número de factores a retener.
3. Se obtienen las cargas factoriales a partir de $\hat{\lambda}_{ij} = \sqrt{v_i} W_{ij}$ (v_i: valor propio, W_i: vector propio). Se vuelve a estimar una nueva comunalidad, $\hat{h}_i^2 = \hat{\lambda}_{i1}^2 + \hat{\lambda}_{i2}^2 + \cdots + \hat{\lambda}_{ik}^2$ y unicidad $\hat{\psi}_i^2 = 1 - \hat{h}_i^2$
4. Se sustituye en la diagonal de la matriz de correlaciones reducida y se vuelve al paso 2.

El proceso se detiene cuando la diferencia entre la comunalidad entre dos iteraciones es menor que una cantidad dada.

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

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

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.

Método de Mínimos Cuadrados no ponderados

Método de extracción que minimiza la suma de los cuadrados de las diferencias entre las matrices de correlaciones observada y reproducida.

Estimación del número de factores

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.

Criterio 1. Elegir m como el número de factores necesarios para alcanzar un porcentaje de varianza explicada entre el 70% y el 80%. Dicha varianza explicada viene dada por tr(S) o tr(R). De esta forma, la varianza explicada por cada factor viene dada por $\lambda j/tr(S)$ o $\lambda j/p$, donde λj es el j-ésimo valor propio de S.

Criterio 2. Elegir como m el número de valores propios mayores que el promedio de los p valores propios. Este promedio es igual a 1 cuando trabajamos con la matriz de correlaciones.

Rotación de factores

Rotación de factores

Se consideran posibles rotaciones de los factores que hagan más interpretables los factores obtenidos. De hecho, en la solución no rotada pueden aparecer los siguientes problemas: Variables que pesan mucho en varios factores. Algunas variables tienen cargas positivas en un factor y negativos en otro.

En el Análisis Factorial no existe una solución única para determinar la matriz de pesos, de hecho, se puede multiplicar por una matriz ortogonal M de orden k×k de modo que

$$\begin{aligned} \mathbf{x} &= \Lambda\mathbf{f} + \mathbf{u} = \\ &= (\Lambda M)(M'\mathbf{f}) + \mathbf{u}, \end{aligned}$$

y este nuevo modelo verifica las mismas propiedades que el anterior: tiene como factores $f^* = M'f$ y como matriz de pesos ΛM. En este caso, la matriz de covarianzas de las variables originales es

$$\Sigma = (\Lambda M)(\Lambda M)' + \Psi,$$

que como $MM' = I$, se reduce a que $\Sigma = \Lambda\Lambda' + \Psi$ como antes; de este modo se explica de manera equivalente la matriz de covarianzas de las variables originales.

Rotación de factores

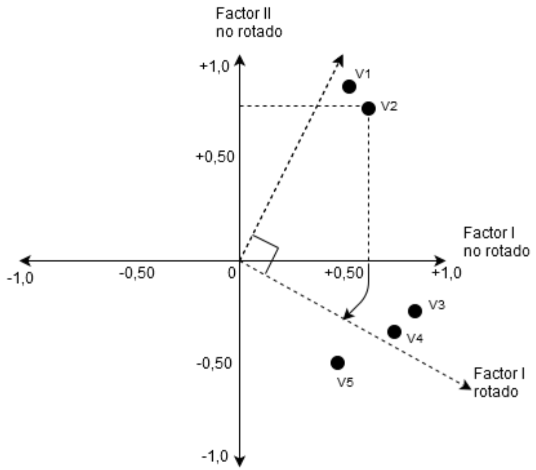
Puede ser que la solución sea más interpretable mediante el uso de alguna matriz ortogonal, lo que lleva al concepto de rotación de los factores.

Según Thurstone, la intención fundamental al realizar una rotación es encontrar una estructura simple:

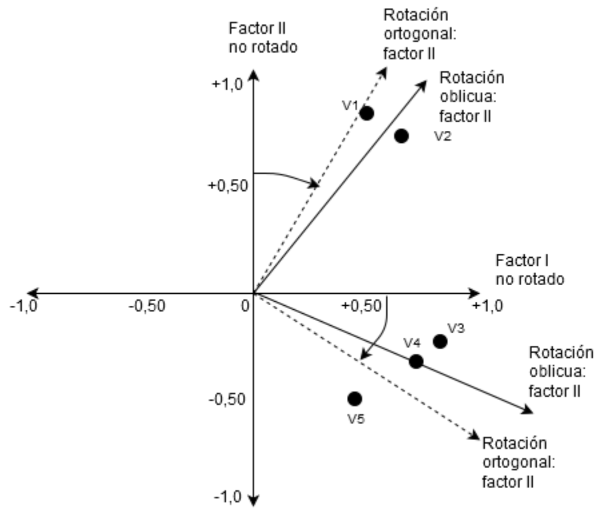
- Cuando se consigue una estructura simple, las variables observadas se encuentran en grupos mutuamente excluyentes de modo que los pesos son altos en unos pocos factores y bajos en el resto.

- Es decir, esencialmente, que las variables observadas caigan en grupos mutuamente exclusivos y que sus cargas sean altas en un único factor, quizás moderadas o bajas en pocos factores, y despreciables en los restantes.

Tipos de rotaciones: Ortogonales y Oblicuas



Rotación ortogonal



Rotación oblicua

Entre las rotaciones ortogonales se encuentran dos tipos principales: Rotación Varimax y Rotación Cuartimax
Entre las rotaciones oblicuas las más importantes se encuentran: Rotación Covarimín, Rotación Quartimín y Rotación Oblimín.

El hecho de rotar los factores siempre es controvertido ya que se pueden elegir los ejes que resulten de mayor conveniencia. Sin embargo, se puede considerar que una rotación es sólo un medio para conseguir unos ejes que permitan describir los puntos de la muestra de la manera más simple posible.

Rotación de factores

Rotación Varimax:

El objetivo de esta rotación es obtener unos nuevos ejes, de forma que, haya pocas cargas con valores altos y muchos próximos a cero. Esto hace que haya factores con correlaciones altas con un número pequeño de variables y correlaciones nulas en el resto, quedando así redistribuida la varianza de los factores.

Rotación Cuartimax:

Trata de que una variable dada esté muy correlacionada con un factor y muy poco correlacionada con el resto de factores. Se usa menos frecuentemente que la anterior.

Rotación Oblimín:

Trata de encontrar una estructura simple sin que importe el hecho de que las rotaciones sean ortogonales, esto es, **las saturaciones no representan ya la correlaciones entre los factores** y las variables. **En este caso se analiza la "matriz estructura" y la correlación entre factores**.

Medición de los factores

Regresión por mínimos cuadrados:

Si interpretamos el modelo del AF, **X** = Δ**f** + **u,** como un modelo lineal, donde X son las observaciones, Δ es la matriz de diseño, f es el vector de parámetros y e = u es el término de error, el criterio de los mínimos cuadrado nos da:

$$f = (\Delta'\Delta)^{-1}\,\Delta' x$$

Otros métodos:

- Método propuesto por M. S. Bartlett
- Método propuesto por T. W. Anderson y H. Rubin

Bibliografía

Bibliografía

- Aldás Manzano, J., & Uriel Jimenez, E. (2017). Análisis multivariante aplicado con R. Ediciones Paraninfo, SA.
- Cuadras, C. M. (2007). Nuevos métodos de análisis multivariante. Barcelona, Spain: CMC Edicions.
- Hair, Anderson, Tatham, Black. (2001). Análisis Multivariante.
- Husson, F., Lê, S., & Pagès, J. (2011). Exploratory multivariate analysis by example using R (Vol. 15). Boca Raton: CRC press.
- Peña, D. (2002). Análisis de datos multivariantes. McGraw-Hill.