ANÁLISIS FACTORIAL

Autor:
Ramón Mahía Casado

1.- IDEA CONCEPTUAL BÁSICA

- (1) Parte de un conjunto amplio de variables que presentan interrelaciones importantes..

- (2) se asume que las relaciones existen porque las variables son manifestaciones comunes de factores no "observables" de forma directa...y

- (3) se pretende llegar a un cálculo de esos factores:
    - (a)- resumiendo información
    - (b)- clarificando las relaciones entre ellas y
    - (c) sin pérdida excesiva de información.

2.- DEFINICIÓN

- **(1)** Proporciona la estructura interna, las dimensiones subyacentes, el transformado de un conjunto amplio de variables, elaborando una estructura más simple, con menos dimensiones, que proporcione la misma información y permita globalizar así el entendimiento del fenómeno.

- **(2)** Simplifica la modelización convirtiendo, por eliminación de redundancias expresadas en altas correlaciones entre variables, un amplio conjunto de variables en factores "estructurales".

3.- DIFERENCIAS CON OTRAS TÉCNICAS

- **No es una técnica de dependencia** (no hay selección a priori de dependiente y exógenas), es una Técnica de Interdependencia

- **No es una técnica de agrupación:** Aunque puede aplicarse con fines de agrupación sobre matrices de correlaciones entre objetos/sujetos (Factorial Q), lo habitual es su aplicación sobre matrices de correlaciones entre variables (Factorial R).

4.- MODELO FACTORIAL EXPLORATORIO (Un ejemplo):

Se desea extrapolar de la provincia al municipio, un modelo de regresión explicativo del nivel de renta disponible función de una serie de manifestaciones de esa renta. Para ello, se parte de un amplio conjunto de variables provinciales y para los 8.000 municipios españoles.

- Recaudación de los distintos impuestos directos e indirectos
- Tasa de paro y actividad
- Generación neta de empleo
- Kilómetros de carreteras de cada tipo en servicios
- Kilómetros de línea férrea en servicio
- Número de vehículos de distintos tipos por habitante
- Líneas telefónicas por cada 100 habitantes
- Camas hospitalarias por cada 1000 habitantes
- Empresas creadas y cerradas en el año
- Índice de precios al consumo
- Índice de precios industriales
- Índice de comercio al por menor
- Licencias fiscales concedidas
- ….etc…

Con el fin de poder abordar con grados de libertad suficiente la estimación del modelo de renta, la información relativa a estas variables se intenta resumir en tres factores, sin perder excesiva información y logrando una incorrelación muy conveniente. El factorial arrojó tres factores cuyos significados se asociaron a:

Factor 1: Factor de renta y riqueza personal - familiar
Factor 2: Factor de salud y desarrollo del mercado laboral
Factor 3: Factor de desarrollo infraestructural

5.- MODELO FACTORIAL CONFIRMATORIO (ejemplo):

Se desea medir la capacidad de abstracción, analítica y memoria de los alumnos.

Se observaron 10 notas de cada alumno de un determinado grupo de estudiantes universitarios. Entre estas notas, o al menos entre algunas de ellas, se observan correlaciones

elevadas que, en cierta medida, provienen de aptitudes globales del alumno que no se observan directamente:

- Nota en álgebra
- Nota en cálculo
- Nota en estadística
- Nota en derecho mercantil
- Nota en derecho laboral
- Nota en contabilidad financiera y de sociedades
- Nota en análisis de costes
- Nota en comunicación comercial
- Nota en actuariales
- Nota en econometría

Un análisis factorial permitió que la información relativa a estas variables se resumiese en tres únicos factores de fondo, sin pérdida excesiva de información y logrando, de nuevo, una incorrelación muy conveniente. Cada uno de estos tres factores se interpretó como:

F2 - Factor de CAPACIDAD DE ABSTRACCIÓN
F3 - Factor de MEMORIA
F4 - Factor de CAPACIDAD ANÁLÍTICA

Independientemente de estos tres factores relacionados con grupos de variables (notas) se identificó, claro está, una factor común que podríamos llamar inteligencia en general y un factor específico para cada asignatura (su propia dificultad y componentes de tipos aleatorio relativos a las distintas formas de evaluación).

6.- MODELO FACTORIAL TEÓRICO

$$X_{ij} = a_{i1} \cdot F_{1j} + a_{i2} \cdot F_{2j} + a_{i3} \cdot F_{3j} + \ldots\ldots + d_i \cdot U_{ij}$$

X_{ij} = Valor normalizado de la variable "i" para el sujeto "j"
Nota en Matemáticas (i) del alumno (j)

F_{1j} = Valor del Factor 1 para el sujeto "j"
Valor del factor CAPACIDAD DE ABSTRACCIÓN del alumno "j"

a_{i1} = Relación entre variable "i" y factor 1

Relación entre las Matemáticas y la CAPACIDAD DE ABSTRACCIÓN

F_{2j} = Valor del Factor 2 para el sujeto "j"

Valor del factor MEMORIA del alumno "j"

a_{i2} = Relación entre variable "i" y factor 2

Relación entre las Matemáticas y la MEMORIA

…

…

…

$d_i \cdot U_{ij}$ = Parte aleatoria independiente de los factores:

Donde:

- "Di" es "la/s particularidad/es" de la nota en Matemáticas
- "Uij" es la forma en que esa peculiaridad afecta al alumno "j". (P.ej. "di" puede hacer referencia a la concentración que se requiere en un examen de matemáticas y Uij a la capacidad de concentración del alumno):

Si los factores están normalizados (esperanza núla y varianza unitaria) y son independientes los unos de los otros pueden obtenerse los siguientes resultados:

A) a_{ik} SERÁ EL COEFICIENTE DE CORRELACIÓN SIMPLE ENTRE LA VARIABLE "i" Y EL FACTOR "k":

$$a_{ik} = \frac{1}{N} \sum_{j} X_{ij} \cdot F_{kj}$$

- **Cargas factoriales:** Coeficientes básicos para determinación contenido conceptual de los factores en análisis exploratorio.

- **Matriz de cargas:** Se denomina así a la matriz que recoge las cargas entre todas las variables originales y la selección final de factores.

B) LA VARIANZA DE LA VARIABLE OBSERVADA "i" PUEDE DESCOMPONERSE EN UNA PARTE EXPLICADA POR LOS FACTORES COMUNES AL RESTO DE VARIABLES Y OTRA EXPLICADA POR EL FACTOR ESPECÍFICO:

$$Var\ (X_i) = \sum_{k=1}^{m} a_{ik}^2 + d_i^2$$

- **Comunalidad:** Uno de los términos más clásicos del análisis factorial expresa la parte de cada variable (su variabilidad) que puede ser explicada por los factores comunes a todas ellas.

- **Especificidad:** Es el término opuesto a comunalidad ya que expresa la parte específica de cada variable que escapa a los factores comunes.

C) EL COEFICIENTE DE CORRELACIÓN ENTRE DOS VARIABLES DEPENDERÁ EXCLUSIVAMENTE DE LA FORMA EN QUE AMBAS VARIABLES COMPARTAN FACTORES COMUNES:

$$Cov\ (X_i X_s) = \sum_{k=1}^{m} a_{ik} \cdot a_{sk}$$

7.- PASOS A COMPLETAR

7.1.- SELECCIÓN DE VARIABLES

Dimensión conceptual: Variables en relación con el fenómeno de análisis. Aún en el caso de un análisis exploratorio, tener claro el modelo factorial teórico ayuda a la selección conceptual de las mismas. "Basura dentro - Basura Fuera"

Dimensión técnica:

(1) Deben ser métricas, aunque se admite la presencia (no generalizada) de ficticias (0,1).

(2) Un número elevado no garantiza un mejor análisis, es más, debe optarse por la minimización del número inicial.

(3) Las correlaciones son la base del planteamiento.

3.A) Deben existir altas correlaciones en general para encontrar factores comunes.

3.B) Todas deben presentar, al menos, alguna relación fuerte: variables aisladas del resto constituirán factores aislados.

7.2.- SELECCIÓN DE LA MUESTRA

Amplitud: Cuanto mayor ratio observaciones/variables, mejor. (*receta: Nº observaciones 5 veces mayor que el de variables*). Una ratio reducida aumenta las posibilidades de encontrar correlaciones espurias, propias de la muestra, no de la población general.

Heterogeneidad: Evidente pero a veces se olvida: una muestra de objetos/sujetos homogénea no contiene información.

7.3.- EXTRACCIÓN DE FACTORES:

Nos referimos al cálculo analítico de los factores a partir de las variables originales.

La extracción implicará:

(1) Decidir el método analítico - matemático de cálculo de los mismos.

1.A) Factorial por **componentes principales**: El análisis explora toda la varianza de cada variable: la común al resto, la específica y la debida a errores de observación.

- Recomendable para reducción de datos
- Recomendable en conjuntos con varianza común elevada

1.B) Factorial Común: El análisis explora sólo la parte común al resto, de la varianza de cada variable.

- Recomendable en análisis confirmatorio de dimensiones latentes (objetivo de reducción en 2º plano)

- Recomendable cuando las puntuaciones factoriales no son importantes (no van a usarse); el método adolece de **indeterminación de factores.**

(2) Seleccionar el número de factores que son necesarios para captar una cantidad razonable de información de los datos originales.

2.A) Valor de los Autovalores o Raíces Latentes
2.B) Selección a proiri (modelo teórico conocido)
2.C) Utilidad práctica (conceptual) de los factores
2.D) % global varianza original explicada
2.E) Contraste de caída en la comunalidad acumulada

7.4.- INTERPRETACIÓN Y ROTACIÓN:

La matriz de cargas, factorial o de componentes relacionarán factores y variables para aproximarnos a su significado. (**Matriz de estructura:** Matriz que contiene los coeficientes de correlación entre factores y variables originales. Para factores ortogonales coincide con la de cargas).

- A la hora de valorar si una carga expresa relación o no (es suficientemente elevada), deberemos ser tanto más exigentes cuanto:
  - Menor sea el tamaño muestral
  - Menos variables se incluyan en el factorial
  - El factor analizado sea de los últimos extraídos

Si esta misión es difícil, **la rotación elimina ambigüedades,** ayudando a hacerlo:

- **Rotaciones ortogonales (Varimax, Equamax, Cuartimax):** Es conceptualmente menos realista, pero maximiza la varianza "explicada" y, en ocasiones, la ortogonalidad resulta útil.

- **Rotaciones oblicuas (Oblimin):** Es más realista (es difícil suponer ortogonalidad conceptualmente), suele ofrecer resultados más claros y además aporta información sobre la relación entre factores.

8.- DETALLE TÉCNICO

8.1.- ALGO MÁS SOBRE EL ANÁLISIS DE CORRELACIONES

- **Triple condición de análisis:**

  A.- En general, la mayor parte de las variables deben estar relacionadas de forma importante

  B.- Todas las variables deben estar relacionadas con, al menos, otra del conjunto

  C.- Las correlaciones parciales no indican presencia de factores subyacenetes comunes.

- **Matriz Anti - Imagen:** Matriz de correlaciones parciales.

- **Test de "esfericidad" de Bartlett:** Test paramétrico basado en el determinante transformado de la matriz de correlaciones: permite contrastar la doble hipótesis de que los elementos de la diagonal principal de la matriz son la unidad y el resto cero.

- **Test Kaiser – Mayer – Olkin:** Ratio sencilla entre correlaciones simples entre parciales + simples debe ser cercano a 1. La misma medida puede elaborarse para una sola variable atendiendo sólo a sus relaciones con el resto de variables (Test MAS_i).

8.2.- ALGO MÁS SOBRE LA EXTRACCIÓN POR COMPONENTES PRINCIPALES (MARCO GENERAL)

"P" variables iniciales:

$$X' = [X_1, X_2 \ldots\ldots X_p]$$

Construiremos **p** componentes principales guiados por: (1) función lineal de las variables originales, (2) que absorban el máximo de variación de las variables X y (3) que estén incorrelacionados.

$$Y_{ij} = \hat{b}_{i1} X_{1j} + \hat{b}_{i2} X_{2j} + \ldots\ldots + \hat{b}_{ip} X_{pj} \; ; \qquad j = 1, 2, \ldots\ldots, n$$

$$Y_i = X \hat{b}_i$$

$$Y_i = \begin{bmatrix} Yi1 \\ Yi2 \\ . \\ . \\ Yin \end{bmatrix}; \quad X = \begin{bmatrix} X_{11} X_{21} \ldots\ldots X_{p1} \\ X_{12} X_{22} \ldots\ldots X_{p2} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ X_{1n} X_{2n} \ldots\ldots X_{pn} \end{bmatrix}; \quad \hat{b}_i = \begin{bmatrix} \hat{b}_{i1} \\ \hat{b}_{i2} \\ . \\ . \\ \hat{b}_{ip} \end{bmatrix}$$

La variación de la variable Y_i, será:

$$Y_i' \; Y_i = \hat{b}_i' S \hat{b}_i \quad \text{donde } \boldsymbol{S = X'X}$$

8.3.- DETALLE OBTENCIÓN 1ER Y 2º COMPONENTES PRINCIPALES

- **El primer componente es:** $Y_1 = X\hat{b}_1$

..... y debemos procurar que se maximice: $Y_1'Y_1 = \hat{b}_1' S \hat{b}_1$

Para abordar el proceso debemos exigir: $\hat{b}'_1 \hat{b}_1 = 1$

..... por tanto al final: *Max* $Z = \hat{b}'_1 S \hat{b}_1 - \hat{l}_1(\hat{b}'_1 \hat{b}_1 - 1)$ ..... o sea:

$$\frac{\partial Z}{\partial \hat{b}_1} = 2S\hat{b}_1 - 2\hat{l}_1\hat{b}_1 = 0$$

$$S\hat{b}_1 - \hat{l}_1\hat{b}_1 = 0$$

$$(S - \hat{l}_1 I)\hat{b}_1 = 0$$

Huyendo de la solución trivial tenemos: $\left| S - \hat{l}_1 I \right| = 0$ .....

..... a partir de aquí, hallamos $\hat{l}_1$ que sustituida en $(S - \hat{l}_1 I)\hat{b}_1 = 0$ nos da $\hat{b}_1$

- **El segundo componente es**: $Y_2 = X\hat{b}_2$

..... y de nuevo debemos procurar maximizar: $Y'_2 Y_2 = \hat{b}'_2 S \hat{b}_2$

.... sujeto de nuevo a la $\hat{b}'_2 \hat{b}_2 = 1$ a la que ahora añadimos la ausencia de correlación con el primer componente: $Y'_2 Y_1 = 0$....o lo que es igual $\hat{b}'_2 S \hat{b}_1 = 0$ que puede escribirse también como $\hat{b}'_2 \hat{b}_1 = 0$

Por tanto, la función a maximizar queda:

$$Max \quad Z = \hat{b}'_2 S \hat{b}_2 - \hat{l}_2(\hat{b}'_2 \hat{b}_2 - 1) - \mu_1(\hat{b}'_2 \hat{b}_1)$$

tras hallar la primera derivada y realizar una serie de reducciones, tenemos que:

$$S\hat{b}_2 - \hat{l}_2\hat{b}_2 = 0 \text{ o sea ..... } (S - \hat{l}_2 I)\hat{b}_2 = 0$$

que se resuelve como para el 1er componente.