EST-46114 Métodos Multivariados y Datos Categóricos

Sesion 08 - Analisis de Factores - Parte 1/3

Juan Carlos Martinez-Ovando

Maestria en Ciencia de Datos

Objetivos

- Estudiaremos los fundamentos del analisis de factores latentes.
- Revisaremos procedimientos para su implementacion.

Preambulo

Consideremos que los `datos` estan definidos de la siguiente forma:

- $\boldsymbol{X}$ - matrix de dimension $(n \times p)$, n *observaciones/casos* y p *variables/dimensiones*

Se *supone* que el orden de las observaciones/casos **no es informativo**, i.e. los renglones con permutables (a.k.a. *supuesto de intercambiabilidad en* n).

- Consideramos cada *caso* como $\boldsymbol{x}_i$ como un vector de dimension $(p \times 1)$.

El modelo de factores se contruye localmente para cada observacion/caso.

Forma basica

La forma basica del modelo de factores considera que, paara todo entero $k < p$ **conocido/predefinido**, se tiene

$$\boldsymbol{x}_i|\boldsymbol{f}_i \sim \mathrm{N}(\boldsymbol{x}_i|\lambda \boldsymbol{f}_i, \Sigma),$$

donde

- $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ - matriz positivo definida de dimension $(k \times k)$
- λ - matriz de dimension $(p \times k)$ de *cargas de factores*
- $\boldsymbol{f}_i$ - *vector de factores latentes* de dimension $(k \times 1)$, con

$$\boldsymbol{f}_i \sim \mathrm{N}(\boldsymbol{f}_i|\mathbf{0}, \mathbf{1}),$$

con

- $\mathbf{0}$ - vector nulo k-dimensional
- $\mathbf{1}$ - matriz identidad k-dimensional

Caracteristicas

- Varianzas locales, $j = 1, \ldots, p$,

$$\text{var}(x_{ij} | \boldsymbol{f}_i) = \sigma_i^2,$$

- Covarianzas locales, $j \neq l$,

$$\text{cov}(x_{ij}, x_{il} | \boldsymbol{f}_i) = 0,$$

- Covarianzas marginales,

$$\text{var}(\boldsymbol{x}_i | \lambda, \Sigma) = \Omega = \lambda\lambda' + \Sigma$$

- Varianzas locales marginales, $j = 1, ..., p$,

$$\text{var}(x_{ij} | \lambda, \Sigma) = \sum_{l=1}^{k} \lambda_{lj}^2 + \sigma_j^2,$$

- Covarianzas locales marginales, $j \neq l$,

$$\text{cov}(x_{ij}, x_{il}) = \sum_{m=1}^{k} \lambda_{jm}\lambda_{lm},$$

para todo i.

Representacion

El modelo en cuestion considera a la coleccion $(\boldsymbol{f}_i)_{i=1}^{n}$ como **variables latentes**, i.e. el *modelo* es

$$\boldsymbol{x}_i|\lambda,\Sigma \sim \mathrm{N}(\boldsymbol{x}_i|\mathbf{0},\Omega),$$

para todo i.
Es decir, el modelo para las variables observables es,

$$\boldsymbol{x}_i|\lambda,\Sigma = \int \mathrm{N}(\boldsymbol{x}_i|\lambda\boldsymbol{f}_i,\Sigma)\mathrm{N}(\boldsymbol{f}_i|\mathbf{0},\mathbf{1})d\boldsymbol{f}_i,$$

suponiendo que en is las *observaciones/casos* son independientes estocasticamente.

Invarianza

El modelo de factores, para un k predefinido, es **invariante** ante transformaciones ortogonales de la forma,

$$\lambda^* = \lambda P', \text{ y } \boldsymbol{f}_i^* = P\boldsymbol{f}_i,$$

con P una matriz ortogonal de dimension $(k \times k)$.

Identificabilidad

Dado el resultado de invarianza, el modelo puede resultar ser **estructuralmente no identificable**.
Una solucion, dentro de varias alternativas, consiste en restringir los pesos en la matriz λ, de la forma

$$\lambda = \begin{pmatrix} \lambda_{11} & 0 & 0 & \cdots & 0 \\ \lambda_{21} & \lambda_{22} & 0 & \cdots & 0 \\ \lambda_{31} & \lambda_{32} & \lambda_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda_{k1} & \lambda_{k2} & \lambda_{k3} & \cdots & \lambda_{kk} \\ \lambda_{(k+1)1} & \lambda_{(k+1)2} & \lambda_{(k+1)3} & \cdots & \lambda_{(k+1)k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \lambda_{p1} & \lambda_{p2} & \lambda_{p3} & \cdots & \lambda_{pk} \end{pmatrix}.$$

Reduccion de parametros

El *numero total de parametros* posiblemente involucrados en Ω es

$$\frac{p(p+1)}{2},$$

para medir la variabilidad involucrada en las p-dimensiones originales.
Mientras que con la representacion de factores, el **numero de parametros efectivo** se reduce a

$$p(k+1) - \frac{k(k-1)}{2}.$$

Modelo saturado

En el caso del **modelo saturado,** en el que $k = p$, el analisis de factores debe cumplir con la restriccion

$$\frac{p(p+1)}{2} - p(k+1) + \frac{k(k-1)}{2} \geq 0,$$

lo cual define una **cota estructural** para la determinacion de k.

Casos

(C1) $p = 6$ implica que $k \leq 3$
(C2) $p = 12$ implica que $k \leq 7$
(C3) $p = 20$ implica que $k \leq 14$
(C4) $p = 50$ implica que $k \leq 40$

Ordenamiento de factores

A diferencia de PCA, el analisis de factores no genera *variables latentes ordenadas*. El ordenamiento puede realizarse mediante un ejercicio de rotacion con una matriz A, transformando el modelo con

$$\lambda^* = A\lambda.$$

> ***Observemos,*** *que la matriz λ^* no necesariamente tiene la estructura triangular inferior citada antes, para identificabilidad.*

Sin embargo, se puede encontrar una matriz ortogonal P tal que

$$A\lambda P',$$

sea triangular inferior y asi recuperar la misma estructura de codependencia del modelo original.

> *Esta consideracion es importante al hacer inferencia sobre k –numero de factores–.*

Siguiente sesion

- Estudio del analisis de factores latentes dinamicos/factores latentes en volatilidad.
- Inferencia sobre el numero de factores.
- Comparacion inferencial entre PCA y analisis de factores.
- Prediccion con factores latentes.

Lectura complementaria

- Stock & Watson (2002) *Forecasting using principal components from a large number of predictors.* `est46114_s08_suplemento1.pdf`
- West (2003) *Bayesian factor regression models in the "large p, small n" paradigm.* `est46114_s08_suplemento2.pdf`
- Pitt & Shephard (1999) *Time varying covariances: A factor stochastic volatility approach.* `est46114_s08_suplemento3.pdf`

Table of Contents

Objetivos

Formulacion

Comentarios

Complementos