Estimarea parametrilor modelului. Extragerea factorilor

Estimarea parametrilor modelului constă în determinarea coeficienților factoriali și implicit a factorilor comuni și a factorilor specifici. Algoritmul de extragere a factorilor pornește de la ipoteza unui model cu un singur factor comun, după care se testează discrepanța dintre matricea de covarianță a variabilelor observate și cea produsă prin model:

$V \approx \hat{V} = L \cdot L^T + \psi$.

Dacă testul este respins, atunci se estimează un model cu doi factori comuni șamd până când testul discrepanței este trecut.

Există mai multe metode de extragere a factorilor, în funcție de criteriile de testare a discrepanței dintre cele două matrice de covarianță. În continuare vor fi prezentate metodele:

- metoda probabilității maxime (the maximum likelihood method)
- metoda celor mai mici pătrate (MINRES - MINimum RESiduals)
- metoda componentelor principale (principal component analysis)

Metoda probabilității maxime

Determinarea factorilor prin metoda probabilității maxime presupune ca datele să urmeze o distribuție normală multivariată de medie 0 (tabelul de observații este centrat) și covarianță $\hat{V}$.

Densitatea de repartiție pentru cazul multidimensional este:

$f(x) = P(X=x) = (2\pi)^{-\frac{m}{2}} |V|^{-\frac{1}{2}} \cdot e^{-\frac{1}{2}(x-\mu)^T V^{-1}(x-\mu)}$,

unde x este un vector cu m elemente reprezentând cele m valori luate de o instanță pentru cele m variabile ale modelului, V este matricea de covarianță a colectivității studiate, iar μ este centrul de greutate.

Dacă scriem funcția pentru o instanță din matricea de observații X, aceasta va fi:

$f(x_i) = (2\pi)^{-\frac{m}{2}} \left|L \cdot L^T - \psi\right|^{-\frac{1}{2}} \cdot e^{-\frac{1}{2} x_i^T \left(\Lambda\Lambda^T - \psi\right)^{-1} x_i}$,

unde $L \cdot L^T - \psi$ este matricea de covarianță estimată.

Forma logaritmică a funcției este:

$\ln(f(x_i)) = -\frac{m}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\left|L \cdot L^T - \psi\right|\right) - \frac{1}{2} x_i^T \left(L \cdot L^T - \psi\right)^{-1} x_i$.

Metoda probabilității maxime maximizează suma log-probabilităților pentru toate cele n instanțe ale modelului:

$$\underset{L}{Maxim}\left(-\frac{m \cdot n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\left|L \cdot L^T - \psi\right|\right) - \frac{1}{2}\sum_{i=1}^{n} x_i^T \left(L \cdot L^T - \psi\right)^{-1} x_i \right).$$

Metoda celor mai mici pătrate

Este abreviată MINRES (MINimum RESiduals). Sunt determinați coeficienții factoriali astfel încât suma pătratelor diferențelor dintre valorile matricei de covarianță și cele ale matricei de covarianță produse prin model să fie minime. Sunt urmărite valorile aflate sub diagonala principală. Dacă metoda este aplicată pentru o matrice de observații, X, standardizată, varianța unei variabile X_j este așa cum am văzut:

$Var(X_j) = \sum_{k=1}^{q} l_{jk}^2 + \psi_j = 1.$

Deci comunalitatea variabilei este mai mică sau egală cu 1: $\sum_{k=1}^{q} l_{jk}^2 \le 1$.

Coeficienții factoriali sunt determinați rezolvând următoarea problemă de optim:

$$\begin{cases} \underset{L}{Minim} \sum_{i=2}^{m}\sum_{j=1}^{i-1}\left(v_{ij} - \sum_{k=1}^{q} l_{ik} l_{jk} \right)^2 \\ \sum_{k=1}^{q} l_{jk}^2 \le 1, \quad j = \overline{1,m} \end{cases}$$

Metoda componentelor principale. Legătura cu analiza în componente principale

Analiza factorială și analiza în componente principale sunt deseori confundate. Unele aplicații specializate folosesc analiza în componente principale ca motor algoritmic pentru analiza factorială.

Am notat cu C matricea $n \times m$ a componentelor principale așezate pe coloane: C_1, C_2, ..., C_m. Notăm cu A matricea $m \times m$ a coeficienților a_k așezați de asemenea pe coloane: a_1, a_2, ..., a_m.

Deoarece $C_k = X \cdot a_k$, pentru k =1,2,...,m, atunci matriceal avem C=X·A.

Matricea A are coloanele ortogonale două câte două, deci $A^{-1} = A^{T}$.

Rezultă că $X = C \cdot A^{-1} = C \cdot A^{T}$, sau altfel scris $X = \sum_{j=1}^{m} C_j \cdot a_j^{T}$.

Abaterea standard a unei componente principale este $\sigma(C_j)$ și este egală cu rădăcină pătrată din varianţa componentei , adică $\sqrt{\alpha_j}$. Putem așadar scrie: $X = \sum_{j=1}^{m} \frac{C_j}{\sqrt{\alpha_j}} \cdot a_j^{T} \sqrt{\alpha_j} = \sum_{j=1}^{m} C_j^{S} \cdot R_j^{T}$, unde C_j^{S} reprezintă scorurile sau componentele standardizate, iar R_j este vectorul coloană al coeficienților de corelație dintre variabilele observate și componenta C_j.

Această relație arată cum tabelul X poate fi reconstituit pornind de la scoruri și corelațiile factoriale. După cum știm corelațiile factoriale sunt din ce in ce mai mici de la o componentă la cealaltă. Dacă luăm în considerare numai primele q componente, tabelul X va fi aproximat prin cantitatea $\sum_{j=1}^{q} C_j^{S} \cdot R_j^{T}$, în timp ce cantitatea $\sum_{j=q+1}^{m} C_j^{S} \cdot R_j^{T}$ poate fi considerată ca o cantitate reziduală, neglijabilă, cu un aport nesemnificativ în reconstituirea tabelului X.
Prin urmare:

$$X = \sum_{j=1}^{q} C_j^{S} \cdot R_j^{T} + \sum_{j=q+1}^{m} C_j^{S} \cdot R_j^{T} = C^{S} \cdot R^{T} + e,$$

unde C^{S} este matricea primelor q scoruri, R este matricea corelațiilor dintre primele q componente și variabilele observate iar e este aportul rezidual la reconstituirea lui X.
În concluzie, prin metoda componentelor principale factorii sunt scorurile din analiza în componente principale iar coeficientii factoriali sunt corelațiile dintre variabilele observate și componentele principale.

Estimarea numărului de factori. Testul Bartlett

Analiza factorială pornește de la asumarea unui anumit număr de factori, q. Acest număr poate fi insuficient. Un model bun este un model în care matricea de covarianță estimată, produsă de model, se apropie cât mai mult de matricea de covarianță a variabilelor observate. Pentru testarea modelului se folosesc teste statistice. Un astfel de test este testul Bartlett (*goodness-of-fit test*)

O descriere perfectă a datelor printr-un număr de factori q, presupune ca $\hat{V} = V$.

Dacă înmulțim relația la stânga cu $\hat{V}^{-1}$ rezultă: $\hat{V}^{-1}\hat{V} = \hat{V}^{-1}V$.

Deci $\hat{V}^{-1}V = \left(L \cdot L^{T} + \psi\right)^{-1} V = I$, unde I este matricea unitate de ordin m.

Statistica testului se calculează astfel:

$$\chi_C^2 = \left(n - 1 - \frac{2m + 4q - 5}{6}\right) \cdot \left(trace\left(\left(L \cdot L^{T} + \psi\right)^{-1} V\right) - \log\left(\left|\left(L \cdot L^{T} + \psi\right)^{-1} V\right|\right) - m\right).$$

Se observa că statistica testului este 0 pentru o descriere perfectă a datelor prin q factori. Urma matricei $\left(L \cdot L^{T} + \psi\right)^{-1} V$ este m (urma matricei identitate), iar logaritm din determinant este 0 (determinantul matricei unitate este 1). Prin urmare cu cât statistica are valori mai mici cu atât cresc șansele ca factorii să descrie adecvat datele și invers, cu cât statistica are valori mai mari cu atât se diminuează șansele ca factorii să descrie adecvat datele.

Ipoteze:
H0. Factorii modelului descriu adecvat datele
H1. Factorii nu descriu adecvat datele

Ipoteza H0 este respinsă dacă $\chi_C^2 > \chi^2(\alpha, r)$,

unde $r = \frac{(m-q)^2 - m - q}{2}$ reprezintă numărul gradelor de libertate.

Rotația factorilor

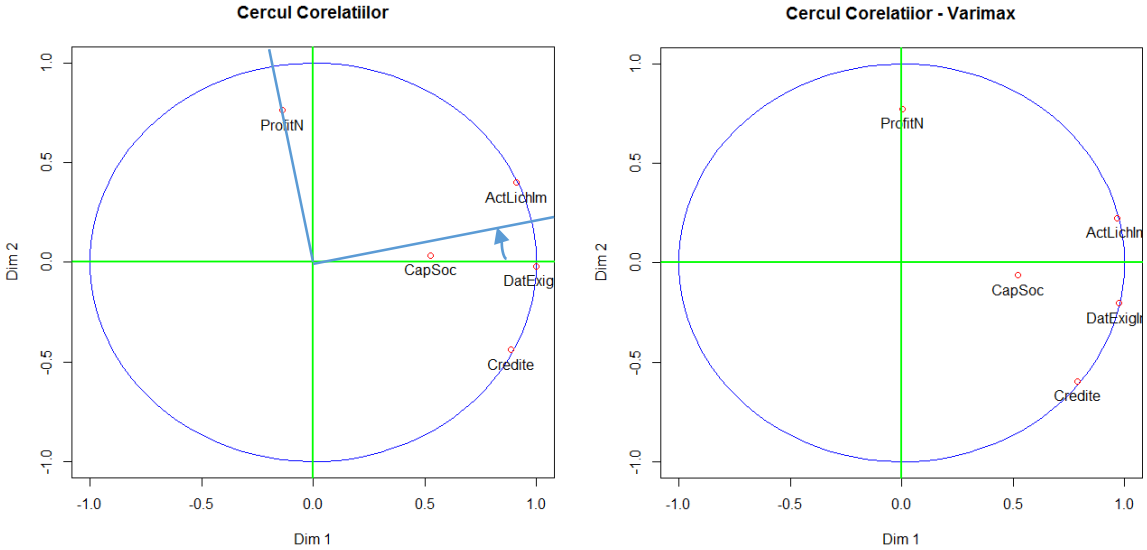
Prin rotația factorilor se încearcă obținerea unei matrice de coeficienți factoriali (L) cât mai ușor de interpretat, care să ușureze cât mai mult interpretarea factorilor. Sistemul de axe ortogonale reprezentat de factori este rotit în jurul originii într-o altă poziție. Prin rotația factorilor se minimizează numărul de variabile cu corelații factoriale mari pentru fiecare factor, simplificând astfel interpretarea factorilor. Una din metodele cele mai utilizate de rotație a factorilor este metoda Varimax. Varimax a fost introdusă de Kaiser în 1958. Prin rotație fiecare variabilă inițială (cauzală, observată) tinde să fie asociată cu un singur factor și fiecare factor cu un număr redus de variabile inițiale. Formal, Varimax caută prin rotații succesive ale sistemului de axe reprezentat de factori, să maximizeze varianța totală explicată de factori.
Acest lucru poate fi obținut prin maximizarea relației:

$$\upsilon = \sum_{k=1}^{q}\sum_{j=1}^{m}\left(l_{jk}^2 - \overline{\overline{L}}_k^2\right)^2 ,$$

unde l_{kj} sunt corelațiile factoriale (factor loadings) iar $\overline{\overline{L}}_k = \frac{1}{m}\sum_{i=1}^{m} l_{ik}^2$, $k = \overline{1,q}$ sunt mediile pătratelor corelațiilor factoriale pe fiecare factor.

Exemplu

În figura următoare sunt prezentate corelațiile factoriale înainte și după aplicarea Varimax. Se poate observa că prin Varimax s-a executat o rotație a axelor în jurul originii în sens trigonometric.



Unghiul de rotație poate fi calculat aplicând funcțiile arccosinus sau arcsinus pe valorile din matricea de rotație. O matrice de rotație pentru două axe, în sens trigonometric, în jurul originii, cu un unghi oarecare, θ, este:

$$R = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} .$$

Pe exemplul de mai sus matricea de rotație furnizată prin funcția *varimax* din R, este: $\begin{bmatrix} 0.982845 & -0.184432 \\ 0.184432 & 0.982845 \end{bmatrix}$, de unde rezultă un unghi de 10.62803 grade.