Ivan Žežula

FAKTOROVÁ ANALÝZA

ÚLOHA

- FA predpokladá existenciu **latentných premenných**, ktoré sa nedajú priamo pozorovať, ale ovplyvňujú pozorovateľné (kvantitatívne) premenné.
- Cieľom je redukcia počtu premenných. Pozorované premenné chceme rozdeliť do skupín vzájomne korelovaných premenných, z ktorých každá je generovaná jedným faktorom (latentnou premennou). Pritom nerozlišujeme závislé a nezávislé premenné.
  - Príklad: známky žiakov z rôznych predmetov môžu byť z podstatnej časti určené faktormi ako inteligencia, usilovnosť, rodinná podpora.
- Z nekonečného množstva potenciálnych faktorov sa snažíme vybrať tie, ktoré sú ľahko interpretovateľné.

DRUHY ANALÝZY

- Prieskumová (exploratórna) FA hľadá v dátach faktory, pričom výskumník nemá predstavu o ich počte a štruktúre.
- Potvrdzovacia (konfirmatórna) FA sa snaží potvrdiť alebo zamietnuť apriórnu predstavu o počte a štruktúre faktorov. Obvykle sa robí pomocou zložitejšieho *modelu štruktúrnych rovníc* (SEM).

PRINCÍP

- Medzi pozorovanými (X_1,..., X_p) a latentnými (F_1,..., F_k) premennými sa predpokladá lineárny (regresný) vzťah:

$$X_i - \mu_i = l_{i1}F_1 + l_{i2}F_2 + \cdots + l_{ik}F_k + \varepsilon_i,$$

kde $k < p$ a μ_i je stredná hodnota X_i.
V maticovom tvare:

$$X - \mu = LF + \varepsilon.$$

Predpoklady:

1. F a ε sú nezávislé,
2. stredná hodnota F i ε je nulová,
3. chyby ε_i sú navzájom nekorelované,
4. faktory sú navzájom nekorelované.

PRINCÍP

- Ak $var\ X = \Sigma$ a $var\ \varepsilon = \Psi$, potom
$$\Sigma = LL' + \Psi.$$
Variančnú maticu pozorovaných premenných X sa teda snažíme rozložiť do tohto tvaru.
- K dispozícii máme iba odhad Σ, výberovú variančnú maticu S. Ak je počet pozorovaní malý, S sa nemusí dať rozložiť, aj keď Σ sa dá.
- Niekedy môže vyjsť rozklad s niektorými $\psi_i < 0$ (Heywoodov prípad). To sa nedá interpretovať, takýto rozklad nepovažujeme za riešenie.

NÁZVOSLOVIE

- Hodnota l_{ij} je **saturácia** (záťaž) j-tého faktora i-tou premennou (factor loading), matica saturácií je **faktorová štruktúra** (factor matrix, structure matrix);
- ak sú faktory korelované, **faktorová schéma** (pattern matrix) udáva koeficienty predstavujúce jedinečný príspevok premennej k príslušnému faktoru;
- vypočítané hodnoty f_j pre pozorované hodnoty x_i nazývame faktorové **skóry** (factor score);
- ε_i je chyba resp. **špecifický faktor**, F_j je **spoločný faktor**;

NÁZVOSLOVIE

- hodnota $\sigma_{ii} - \psi_i = {l_{i1}}^2 + \cdots + {l_{ik}}^2 = {h_i}^2$ je **komunalita** – časť rozptylu vysvetlená spoločnými faktormi, ψ_i je špecifický (reziduálny) rozptyl;
- **redukovaná variančná matica** je $\Sigma - \Psi$;
- **jedinečnosť premennej** je ψ_i, t.j. rozptyl premennej mínus komunalita premennej;
- **náhradnou premennou** faktora F_j je tá premenná X_i, ktorá má na ňom najvyššiu saturáciu;
- ak $V_j = \sum_{i=1}^{p} {l_{ij}}^2$ a $V = \sum_{i=1}^{p} {h_i}^2$, potom V_j/V je podiel celkového rozptylu vysvetlený faktorom F_j.

VLASTNOSTI

- Ak zmeníme mierku (meracie jednotky) pozorovaných premenných, matice Σ, LL' a Ψ sa zmenia rovnakým spôsobom, t.j. riešenie ostane v princípe rovnaké ⇒ stačí vychádzať z korelačnej matice $R = corr\ X$ (najčastejší spôsob).
- Dobre interpretovať vieme taký faktor, ktorý má vysoké saturácie nejakou podmnožinou premenných a nízke saturácie všetkými ostatnými. Ideálne je, ak premenné vysoko saturujúce jeden faktor nesaturujú iné faktory.
- V prieskumovej analýze sa niekedy používajú aj iné typy korelácie než Pearsonov koeficient, hoci teória ho predpokladá.

VLASTNOSTI

- Matice Σ, LL' a Ψ sa nezmenia, ak vezmeme LT miesto L a $T'F$ miesto F, kde T je ortogonálna matica. Hovoríme o **rotácii** faktorov. Snažíme sa nájsť rotáciu s najlepšou interpretáciou. Je ich ale nekonečne veľa!
- Ak matica T nie je ortogonálna (ale všeobecná regulárna), hovoríme o **šikmej rotácii.** Šikmá rotácia má za následok, že faktory sú korelované.
- Potrebný počet pozorovaní, aby výsledky boli spoľahlivé, je v desiatkach na každý odhadovaný parameter.

POSTUP RIEŠENIA

METÓDY EXTRAKCIE FAKTOROV

- **Metóda hlavných komponent** (PCA): urobí sa spektrálny rozklad matice S resp. R a z vlastných vektorov zodpovedajúcich najväčším vlastným číslam λ_j sa vytvorí odhad matice L. Zvyšok je odhadom Ψ. Keďže $\frac{V_j}{V} = \frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}$, kumulatívne percento vysvetleného rozptylu pre m faktorov je $\frac{\sum_{j=1}^{m} \lambda_j}{\sum_{i=1}^{p} \lambda_i}$.

METÓDY EXTRAKCIE FAKTOROV

- **Metóda hlavných faktorov** (PFA, Common Factor Analysis, Principal Axis Factoring): je iteratívna procedúra, pri ktorej sa opakovane odhadujú komunality a potom rozkladá redukovaná matica S resp. R, až kým sa odhady nestabilizujú.
  Varianty sa líšia počiatočnými odhadmi komunalít.
  Najčastejšie používané sú štvorec korelačného koeficientu X_i s ostatnými premennými (SMC) a maximálna absolútna korelácia X_i s ostatnými premennými (Maxrow).
  Riešenie *nie je invariantné* voči zmene mierky!

METÓDY EXTRAKCIE FAKTOROV

- **Metóda maximálnej vierohodnosti** (MLFA) predpokladá normalitu rozdelenia F i ε (a teda aj X. Odhady L a Ψ sú numerickým riešením vierohodnostných rovníc. Aby bolo riešenie jednoznačné, pridáva sa podmienka, že matica $L'\Psi^{-1}L$ musí byť diagonálna.
- **Metóda kanonických faktorov** (CFA, Rao's canonical factoring) extrahuje faktory, ktoré majú najväčšie kanonické korelácie s pozorovanými premennými X.

METÓDY EXTRAKCIE FAKTOROV

- **Projekčná metóda** (Image Factoring) robí rozklad korelačnej matice regresných predikcií, t.j. každá premenná je nahradená svojou predikciou pomocou regresie na ostatných premenných.
- **Metóda alfa-faktorov** (Alpha Factoring) predpokladá, že pozorované premenné sú náhodným výberom z priestoru všetkých možných premenných (všetky ostatné metódy berú premenné ako pevne dané, náhodným výberom sú jednotlivé pozorovania). Faktory sú extrahované tak, že je maximalizovaná ich (Cronbachova) α-spoľahlivosť.

METÓDY EXTRAKCIE FAKTOROV

- **Metóda najmenších štvorcov** (ULS) extrahuje faktory tak, že minimalizuje súčet štvorcov rozdielov medzi pozorovanou korelačnou maticou a reprodukovanou korelačnou maticou (diagonála je ignorovaná).
- **Metóda vážených najmenších štvorcov** (WLS, GLS) používa rovnaký princíp ako ULS, ale korelácie sú vážené prevrátenou hodnotou svojej jedinečnosti, t.j. $1/\psi_i$.

URČENIE POČTU FAKTOROV

- **Apriórne určenie** počtu je možné vtedy, ak ich počet vyplýva z povahy problému.
- **Kumulatívne percento rozptylu** vysvetleného faktormi je najčastejším kritériom pre určenie počtu extrahovaných faktorov. Odporúča sa, aby to bolo aspoň 60%.
- Určenie pomocou **sutinového grafu**, ak tento má jasný zlom.
- V prípade použitia PCA môžeme použiť **prahovú hodnotu pre vlastné čísla.** Vyberieme iba faktory s $\lambda_j \geq 1$ (Kaiserovo pravidlo), resp. $\geq \bar{\lambda}$ (ak pracujeme s variančnou maticou). Možný je však aj iný prah.

URČENIE POČTU FAKTOROV

- Určenie pomocou **rozdelenia súboru na polovice.** FA sa uskutoční v každej polovici nezávisle a ponechajú sa iba tie faktory, ktoré majú veľmi podobné saturácie v obidvoch častiach.
- Určenie pomocou **testov významnosti.** V prípade metódy PCA sa dá testovať významnosť jednotlivých vlastných čísel; ponechajú sa iba faktory s významnými vlastnými číslami. Ak je rozsah výberu veľký, vedie táto metóda k zbytočne veľkému počtu faktorov, ktoré často vysvetľujú iba malú časť variability.

ROTÁCIE

- Metóda **varimax** maximalizuje priemerný rozptyl vysvetlený spoločnými faktormi, t.j. priemer V_j. Minimalizuje tak počet premenných s vysokými saturáciami u každého faktora. To uľahčuje interpretáciu faktorov.
- Metóda **quartimax** sa snaží maximalizovať saturáciu nejakého faktora danou premennou. Minimalizuje tak počet faktorov potrebných k vysvetleniu pozorovaných premenných. Má tendenciu vytvárať jeden všeobecný faktor s všetkými saturáciami vysokými.
- Metóda **equamax** sa snaží dosiahnuť jednoduchú štruktúru matice L vzhľadom na riadky i stĺpce. Je kombináciou varimaxu a quartimaxu.

ROTÁCIE

- Metóda **oblimin** robí šikmé (neortogonálne) rotácie. Má parameter $\delta \leq 0{,}8$, ktorý je mierou šikmosti rotácie. Kladné hodnoty sa vo všeobecnosti neodporúčajú, keďže vedú k silnej závislosti až nerozlíšiteľnosti faktorov. Preto je základné nastavenie $\delta = 0$ (metóda **quartimin**). Veľké záporné hodnoty vedú takmer k ortogonalite faktorov.
- Metóda **promax** tiež robí šikmé rotácie. Má parameter $\kappa \geq 1$. Čím je κ väčšia, tým sú faktory viac korelované a štruktúra saturácií jednoduchšia. Odporúčaná hodnota je $\kappa = 4$ ako kompromis medzi jednoduchosťou a korelovanosťou. Je výpočtovo jednoduchšia ako oblimin, takže sa odporúča pre rozsiahle dáta.

VÝPOČET FAKTOROVÝCH SKÓROV

- **Regresná metóda** je najbežnejšia. Skóry majú strednú hodnotu 0 a rozptyl rovný štvorcu korelácie medzi odhadnutými skórmi a skutočnými hodnotami. Skóry môžu byť korelované, aj keď sú faktory ortogonálne.
- **Bartlettove skóry** minimalizujú vážený súčet štvorcov rezíduí, pričom váhy sú prevrátenou hodnotou príslušnej jedinečnosti. Majú strednú hodnotu 0.
- **Anderson-Rubinova metóda** je modifikáciou Bartlettovej metódy, ktorá zaručuje ortogonalitu (nekorelovanosť) odhadnutých faktorov. Skóry majú strednú hodnotu 0 a rozptyl 1.

DIAGNOSTIKA

- **p-hodnoty** (významnosti) jednotlivých korelačných koeficientov. Ak korelácie nie sú významné, model FA je nevhodný.
- **Bartlettov test sférickosti** testuje hypotézu $R = I$. V takom prípade je model FA je nevhodný.
- **Sutinový graf** (scree plot) zobrazuje rozptyl vysvetlený každým faktorom. Používa sa na určenie počtu faktorov, ktoré treba extrahovať. Typicky má strmú časť u významných faktorov a plochú časť u zvyšných (sutina).
- **Kaiser-Meyer-Olkinov koeficient** (KMO measure) je číslo z intervalu $(0; 1)$, ktoré je mierou vhodnosti výberu premenných. Malo by mať hodnotu najmenej $0{,}5$. Malé hodnoty koeficientu ukazujú, že sú malé parciálne korelácie dvojíc premenných s vylúčením vplyvu všetkých ostatných a že teda model FA je nevhodný.

DIAGNOSTIKA

- **Reprodukovaná korelačná matica** je odhadom korelačnej matice na základe extrahovaných faktorov a ich záťaží. Na diagonále sú odhady komunalít. **Reziduálna matica** je rozdiel reprodukovanej a pozorovanej korelačnej matice. Cieľom je mať reziduálnu maticu blízku nule.
- **Korigovaná korelačná matica** (anti-image correlation matrix) obsahuje mínus parciálne korelácie dvojíc premenných s vylúčením vplyvu všetkých ostatných, **korigovaná variančná matica** obsahuje mínus parciálne kovariancie dvojíc premenných. Diagonálny prvok korigovanej korelačnej matice je mierou vhodnosti príslušnej premennej do modelu FA. Dobrý model FA má všetky mimodiagonálne prvky malé.

POTVRDZOVACIA FA

- Používa model štruktúrnych rovníc (SEM, sw balíky AMOS alebo LISREL).
- Predpokladá, že výskumník už pozná počet faktorov i ktoré pozorované premenné ich saturujú.
- Testovanie modelu FA vyžaduje nasledujúci postup:
  - Z modelu sa odstránia všetky priame šípky spájajúce latentné premenné;
  - pridajú sa oblúkové šípky predstavujúce koreláciu/kovarianciu medzi všetkými pármi latentných premenných;
  - ponechajú sa všetky priame šípky spájajúce latentné premenné s ich saturujúcimi indikátorovými premennými;
  - ponechajú sa všetky priame šípky od chybových členov k príslušným premenným.
- Takto vytvorený model sa potom vyhodnotí pomocou štandardných kritérií a mier vhodnosti (goodness of fit).