

Explorační faktorová analýza

PSYb2590: Základy psychometriky | Seminář 4

28. 3. / 4. 4. 2022 | Hynek Cígler, Petr Palíšek, Edita Chvojková
Adam Ťápal (in memoriam)

Review přednášky

- **Manifestní** (pozorované) a **latentní** (skryté) proměnné, vždy spojené
- Faktory jsou latentními proměnnými, jsou **nepozorovatelné a neměřitelné**
- **Fundamentální princip FA:** Manifestní proměnné korelují právě proto, že jsou "způsobovány" jednou stejnou (nebo více stejnými) latentními proměnnými
- FA je model, který popisuje / vysvětluje korelace mezi MVs tím, že postuluje existenci společných (common) LVs - společných faktorů

Review přednášky

- Manifestní proměnné jsou **lineární funkcí** latentních proměnných
- Míra toho, jak moc která LV ovlivňuje kterou MV je reprezentována tzv. **faktorovým nábojem** (ty jsou jako regresní koeficienty)
- **Common Factor Model:**

Pozorovaný rozptyl = Společný rozptyl + Specifický rozptyl + Chybový rozptyl

$$\text{Komunalita (Communality)} = \frac{\text{Společný rozptyl}}{\text{Pozorovaný rozptyl}} = 1 - \frac{\text{Unikátní rozptyl}}{\text{Pozorovaný rozptyl}}$$

... = podíl pozorovaného rozptylu, který je způsoben obecnými (společnými) faktory

Review přednášky

$$x_{ij} = \mu_j + \lambda_{j1}z_{i1} + \lambda_{j2}z_{i2} + \cdots + \lambda_{jm}z_{im} + 1u_{ij}$$

Průměr + Obecné (společné) faktory + Unikátní faktor

x_{ij} je skór osoby i na manifestní proměnné j

μ_j je průměr manifestní proměnné j

z_{ik} je skór osoby i na obecném faktoru k

λ_{jk} je faktorový náboj manifestní proměnné j na faktoru k

u_{ij} je skór osoby i na unikátním faktoru j

Review přednášky

Rovnice modelu vypadá jako rovnice pro vícenásobnou lineární regresi

- Manifestní proměnné jsou závislými proměnnými
 - Faktory jsou nezávislými proměnnými
 - Faktorové náboje jsou regresními koeficienty
-
- Faktorový model je jako sada vícenásobných lineárních regresí, kde nezávislé proměnné jsou nepozorované a neměřené (...a nepozorovatelné a neměřitelné)
 - Všechny parciální korelace mezi jednotlivými manifestními proměnnými - ve chvíli, kdy kontrolujeme vliv obecných faktorů – jsou předpokládány za nulové
 - Jinými slovy – korelace mezi jednotlivými manifestními proměnnými jsou způsobeny pouze obecnými faktory

Review přednášky

- Cílem FA je **odhalit, pochopit a popsat** strukturu, která „způsobuje“ korelace mezi manifestními proměnnými
- Chceme tedy identifikovat (nebo ověřit) **počet a charakter** (význam) faktorů, které způsobují pozorované korelace mezi manifestními proměnnými
- Jinými slovy, chceme přijít na to, kolik obecných / společných faktorů ovlivňuje naše manifestní proměnné a **odhadnout sílu a směr (+ / -) faktorových nábojů**
- Velikost a směr faktorových nábojů nám napomáhá v určení podstaty faktoru. Význam faktoru je totiž vymezen tou podmnožinou všech manifestních proměnných, které jsou faktorem výrazně ovlivňovány

Review přednášky

- Ve světě faktorové analýzy rozlišujeme dvě situace:
- **Explorační (exploratory / unrestricted) FA:**
Nemáme žádnou (nebo jen velmi mlhavou) představu o tom, kolik faktorů a jakého charakteru je „za daty“
- **Konfirmační (confirmatory / restricted) FA:**
Máme celkem jasnou představu o tom, kolik faktorů a jakého charakteru je „za daty“
- ...teoretický model, který v obou případech používáme, je **totožný!**

Explorační faktorová analýza (EFA)

- Ještě jednou, k čemu je tedy vlastně dobrá EFA?
 - Máme data, která jsou pro FA vhodná (předpoklad existence latentních proměnných), ale nemáme (jasnou) představu o faktorové struktuře
 - Máme několik (vágních) nápadů, jak by mohla faktorová struktura vypadat
 - Teoretický model neexistuje
 - Existující teoretický model nepopisuje data dobře (a u toho nechceme skončit)
- ➔ chceme (lépe) prozkoumat (explorovat) možnou faktorovou strukturu

Explorační faktorová analýza (EFA)

- Jak to (ve zkratce) funguje?
 - 1) Stanovíme si, na jaká data model aplikujeme (jaké máme MVs?)
 - 2) Zvolíme si počet (nikoliv však charakter) společných faktorů (tohle může působit neintuitivně – vždyť to děláme proto, že nevíme! Více později 😊)
 - 3) Zvolíme si metodu odhadu parametrů modelu (více později 😊)
 - 4) Pomocí softwaru odhadneme faktorové náboje všech MVs
 - 5) Vyhodnotíme shodu modelu s daty (fit)
 - 6) Zamyslíme se, zda výsledek „působí“ přijatelně (teorie, zkušenost...) 😊

Předpoklady EFA

- Model stojí na určitých předpokladech, které jsou nutností k tomu, aby se dal odhadnout a matematicky odvodit:
 - 1) Obecné (společné) faktory a unikátní faktory jsou nezávislé a nekorelují spolu
 - 2) Unikátní faktory jsou navzájem rovněž nezávislé a nekorelují spolu
 - 3) Obecné a unikátní faktory mají z definice průměr 0
 - 4) Obecné a unikátní faktory mají z definice rozptyl 1 (a tedy i $SD = 1$)

Příklad v JASPu

- Klasický dataset Holzinger & Swineford (aka Svinibrod), 1939
- 301 dětí, skóry z 9 testů:
 - Visual Perception, Cubes, Lozenges
 - Paragraph Comprehension, Sentence Completion, Word Meaning
 - Speeded Addition, Speeded Counting, Speeded Discrimination

JASP bugs

Namísto korelační matice faktorů JASP reportuje korelační matici faktorových skóre.

- „Hey sorry, this is a bug“. <https://github.com/jasp-stats/jasp-issues/issues/1319> :-)

JASP eposkytuje některé důležité ukazatele.



hynekcigler commented now

Author 😊 ⋮

@JMBKoch @juliuspf After another year, I want to teach the same psychometric class as a year before. Unfortunately, the main bug has not been fixed yet and I am going to say my students "hey, sorry, this is a bug" :(Just to summarize:

1. **BUG:** In EFA, factor score correlation are reported instead of factor correlation (and thus even with orthogonal rotation, factors are "correlated").
2. **Serious deficiency:** residual matrix in EFA is not available.

Other problems are unpleasant, but manageable. (And thank you for the work I see you did in the related new issue linked above).

Shoda modelu s daty

- Nějaký výsledek zpravidla vždy dostaneme. Jak ale poznáme, že model popisuje data dobře?
 - 1) Rozdíl mezi pozorovanými korelacemi MVs a tzv. modelem implikovanými korelacemi MVs → **Reziduální matice** (kterou JASP neumí! 😬)
 - 2) **Test of perfect fit (χ^2):**
 - H0: Model (v populaci) naprosto přesně popisuje data
 - H1: Data nemají požadovanou faktorovou strukturu

... „žádoucí“ je tedy vysoká p-hodnota

... extrémní síla testu, extrémně senzitivní na velikost vzorku

... většinou nám stačí závěr „data mají *přibližně* požadovanou faktorovou strukturu (moc se od ní neliší)“

Shoda modelu s daty

- Nějaký výsledek vždy dostaneme. Jak ale poznáme, že model popisuje data dobře?

3) Indexy shody modelu s daty. V JASPU a JAMOVI najdeme:

TLI (Tucker-Lewis Index), jde o tzv. inkrementální index

- „Kde na kontinuu mezi nejhorším možným (0) a nejlepším možným (1) modelem se nachází náš model?“
- Typicky chceme vidět $TLI > .9$ (ale není to vytesané do kamene)

RMSEA (Root Mean Square Error of Approximation), jde o tzv. absolutní index

- Stejně jako TLI bere v úvahu komplexitu modelu
- Typicky chceme vidět $RMSEA < .08$ (ale není to vytesané do kamene)

Počet faktorů

- Počet faktorů volíme **a priori** předem
- Nemůžeme nikdy úplně znát „pravdu“ (skutečný počet společných faktorů)
...což nemalou řadu výzkumníků dost znervózňuje
- Existuje proto řada postupů, jak zvolit optimální počet faktorů, od jednoduchých rules-of-thumb po sofistikovanější postupy
- ...měli bychom je ale brát spíše jako pomocníky než jako “pravdu“. Žádná science machine, která udělá science za vás, neexistuje.
- Nejdůležitějšími kritérii jsou **shoda modelu s daty** a **interpretabilita modelu** (dá se model interpretovat a dává smysl?)

Počet faktorů

- **Kaiser-Guttmanovo** kritérium
 - Počet eigenvalues větších než 1 je *spodní hranicí* skutečného počtu faktorů
 - Pravidelně nepochopeno a zneužíváno
 - Nemá v podstatě žádnou oporu v teorii, nepoužívejte
(viz http://www.quantpsy.org/pubs/preacher_maccallum_2003.pdf)
- **Scree plot** („sutinový graf“)
 - Seřadit eigenvalues dle velikosti, zanést na graf a propojit spojnici
 - Tolik faktorů, kolikátá eigenvalue je „bodem zlomu“ na grafu
 - Subjektivní, bez dostatečné opory v teorii

Počet faktorů

- **Hornova Paralelní analýza**

- Kaiser-Guttmanovo kritérium vylepšené o zvážení výběrové chyby
- Když už se spoléhat na Kaiser-Guttmanovo pravidlo, tak jedině takto

- **Velicerovo MAP (Minimum Average Partial)**

- Iterativní procedura
- Optimální počet faktorů je takový počet, který modeluje ještě nějakou systematickou korelaci mezi MVs

Rotační indeterminace

- aka „rotační neurčitelnost“
- $\Sigma = \Lambda \Phi \Lambda' + D_\psi$
- Jako řešení hledáme matici faktorových nábojů Λ , která vyhoví rovnici výše
- Máme ale maličký, úplně malinkatý problémeček - pokud takovou matici najdeme (a pokud uvažujeme řešení s 2 a více faktory), pak existuje nekonečně mnoho dalších takových matic Λ , které jsou lineárními transformacemi té původní Λ
- Takže – pokud najdeme nějaké řešení, pak jsme jich našli nekonečně mnoho a všechny z nich jsou stejně „dobrá“

Rotační indeterminace

- To se může zdát docela divné – proč hledáme nějaké řešení, když jich existuje nekonečně mnoho stejně dobrých? A jaké si tedy máme vybrat?
- Není to tak hrozné, jak se může zdát. Tato různá řešení jsou jen transformacemi jedno druhého, jsou matematicky ekvivalentní. Jen nejsou ekvivalentní pro naše oči, a některá mohou být lépe interpretovatelná lidmi než jiná.
- Koncept jednoduché struktury „simple structure“ (Thurstone)

Rotační indeterminace

- Tohoto principu využívá tzv. **rotace**, jeden ze základních interpretačních mechanismů EFA
- Získané řešení můžeme „rotovat“ (transformovat na jiné) tak, aby se nám ulehčil proces interpretace, tedy proces **atribuce významu** jednotlivým faktorům
- Význam faktorům totiž připisujeme na základě struktury jejich **faktorových nábojů**

Rotační indeterminace

- **Ortogonalní rotace** (orthogonal) – společné faktory jsou nekorelované
 - např. Varimax, Quartimax, ...
- **Oblé rotace** (oblique) – společné faktory *mohou* korelovat
 - např. Oblimin, Simplimax, ...
- Ortogonalní rotace jsou (podle nás) spíše reliktem minulosti, protože jsou méně výpočetně náročné. Používejte oblé rotace.

Metoda odhadu parametrů

- Je jich celá řada 😊
- Nejobvyklejší metody jsou založeny na:
 - Maximum Likelihood (ML, silný předpoklad normality MVs)
 - Metodě nejmenších čtverců (Least Squares, menší předpoklad normality MVs)
 - Ordinary Least Squares (OLS)
 - Minimum Residual (Minres)
- Analýza hlavních komponent (Principal Component Analysis, PCA)
 - Nejde o FA (je to jiný model), ale SPSS to chytře vydává za metodu odhadu parametrů
 - Použití v psychologii spíše neobvyklé, ale denní chleba třeba v machine learningu

Heywood cases

- Heywoodovy případy – někdy se může stát, že v odhadnutém modelu je některý rozptylový parametr záporný.
- Rozptyl ale nemůže být záporný...je to něco jako dělit nulou. Když se to stane, někde umře koťátko nebo vesmír imploduje.
- Pokud se vám to stane, pak je váš model nejspíš příliš složitý (příliš mnoho faktorů), nebo vám zlobí nějaká položka
- ...když už jsme u toho, obecně se snažte, aby vás model nebyl složitější, než je nutno. Úspornost (parsimony) je hlavním principem modelování (jakéhokoliv)

Pár tipů

- Simple structure
- Faktor musí být identifikován alespoň 3 manifestními proměnnými
- Používejte oblé rotace
- Faktory můžete „obrátit“
- Faktorové skóry neznáme a znát nemůžeme, dají se ale odhadnout (což JASP ani JAMOVI neumí....)
- EFA tedy v tomto kurzu používejte především k ověření / exploraci faktorové struktury a na základě ověření pracujte se součtovými skóry (pokud to budete potřebovat)

Pár tipů

- Pečlivě uvádějte postup volby počtu faktorů, metodu extrakce i rotace
- U EFA je zcela akceptovatelné vyzkoušet sérii modelů, nejde o rybaření!
- Máte-li silné předpoklady o modelu, volte CFA (o ní si povíme příště)
- Alespoň $N = P * F * 5$ respondentů, kde P je počet položek a F počet faktorů
- Nepoužívejte PCA
- Nejmenší čtverce jsou vhodnou první volnou pro estimátor