

Faktorová analýza

PSY259 – ZÁKLADY PSYCHOMETRIKY (PŘEDNÁŠKA 4)

12. 4. 2021 | ADAM ŤÁPAL

CTT vs. teorie latentních rysů (např. FA)

Klasická testová teorie:

- Položky jsou *paralelními* (zaměnitelnými) *testy* (měřítky) měřeného konstrukt
- Měřeným konstruktem je pravé skóre (*true score*) osoby v testu
- Měřený konstrukt je tedy závislý na testu (souboru položek), je jím operacionalizovaný („*Pravé skóre je to, co měříme tímto testem*“)
- **Operacionalismus:** Konstrukt (a jeho význam) nelze oddělit od metody
- **Antirealismus:** Konstrukt reálně neexistuje, je něčím, co jsme si jen vymysleli

CTT vs. teorie latentních rysů (např. FA)

Teorie latentních rysů:

- Konstrukty *reálně existují*
- Konstrukty *způsobují* reakce na stimuly / odpovědi na položky
- Konstrukty jsou *společnou příčinou* chování
(Položky v testu inteligence spolu korelují, potože správnost odpovídání na ně má společnou příčinu – inteligenci)
- **Realismus:** Konstrukty = latentní rysy existují a jsou příčinou pozorovaného chování

Latentní rys:

Schopnost rychle běžet

Skill v šachu

Jak se (třeba) projevuje?

Latentní rys:

- 1) *Jak rychle zaběhl 100m?*
- 2) *Jak rychle zaběhl 400m?*
- 3) *Jak rychle zaběhl 800m?*

Schopnost rychle běžet



- 1) *Kolikrát z 10 her porazil cvičenou opici?*
- 2) *.... okresního mistra v šachu?*
- 3) *.... Garriho Kasparova?*

Skill v šachu



Faktorová analýza

- Vysvětluje / popisuje vztahy mezi (spojitými) *manifestními* proměnnými a (spojitými) *latentními* proměnnými (rysy)
- **Manifestní proměnná (MV)** – proměnná, kterou lze přímo měřit či pozorovat
- **Latentní proměnná (LV)** – proměnná, kterou NELZE přímo měřit či pozorovat – hypotetický konstrukt. **Faktory** ve faktorové analýze jsou právě latentními proměnnými. Tedy – faktor (LV) je stále nějaká (spojitá) proměnná a různí lidé „mají“ své skóry na této proměnné (alespoň to je předpoklad 😊)

Manifestní proměnné:

Běh:

- 1) *Jak rychle zaběhl 100m?*
- 2) *Jak rychle zaběhl 400m?*
- 3) *Jak rychle zaběhl 800m?*

Šachy:

- 1) *Kolikrát z 10 her porazil cvičenou opici?*
- 2) *.... okresního mistra v šachu?*
- 3) *.... Garriho Kasparova?*

Latentní proměnné:

Schopnost rychle běžet

Skill v šachu

Měřené osoby:

Adolf

Běh: (20s, 90s, 180s)

Šachy: (3, 1, 0)

Bruno

Běh: (40s, 180s, 300s)

Šachy: (4, 2, 1)

Cecil

Běh: (50s, 190s, 320s)

Šachy: (7, 4, 3)

Faktorová analýza

- *Schopnost rychle běžet ani skill v šachu* neumíme (nemůžeme) nijak „přímo“ měřit, zbývá nám na ně *usuzovat*
- Předpokládáme, že obě latentní proměnné se *manifestují* skrze něco, co měřit nebo pozorovat můžeme – **manifestní proměnné**
- Rozdílná *schopnost rychle běžet* mezi osobami se bude manifestovat rozdílnými časy na jednotlivých tratích, ale nebude mít sama o sobě nic společného s počtem výher v šachu
- Rysy osobnosti či postoje se mohou manifestovat mírou (nesouhlasu) s tvrzeními, která by měla být pro vysokou/nízkou míru rysu typická
(„*Hrozně rád jsem ve společnosti středem pozornosti*“)

Faktorová analýza

- Faktorová analýza nám do ruky dává **matematický** nástroj (**statistický model**), který nám umožňuje vztahy mezi manifestními a latentními proměnnými studovat
- Na předchozích slidech jsme si představili základní premisu FA konceptuálně, jako takový myšlenkový experiment
- Pojdme to vzít trochu techničtěji a abstraktněji – představením modelu.
- Jak podstatu tohoto myšlenkového experimentu propojíme s reálnými daty, s něčím pozorovatelným či měřitelným?

Základní pojmy

- Jaká je typická podoba dat v případě faktorové analýzy?
- Multivariační data – data pro soubor osob, větší množství manifestních (měřených, pozorovaných) proměnných (např. skóry z testů, škál, položek...)

Datová matice:

Co sloupec, to proměnná

Co řádek, to osoba

Základní pojmy

- Jednotlivé buňky v datové matici představují skór dané osoby na dané manifestní proměnné
- Fundamentální premisa faktorové analýzy: Tyto skóry nejsou nějakými náhodnými hodnotami, ale vykazují určité systematické aspekty, kterými se můžeme zabývat

Datová matice:

Co řádek, to osoba

Co sloupec, to proměnná

Základní pojmy

Datová matice:

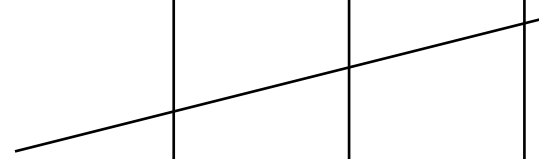
p sloupců (proměnných)

$X =$

x_{11}	x_{12}		x_{1p}
		x_{ij}	
x_{N1}	x_{N2}		x_{Np}

N řádků (osob)

Skór osoby *i* na proměnné *j*



Základní pojmy

Čeho si můžeme na těchto datech všimnout?

- Variabilita každé proměnné napříč osobami (rozptyl / SD)
- Kovariance dvou proměnných napříč osobami (kovariance / korelace)

X_{11}	X_{12}		X_{1p}
		X_{ij}	
X_{N1}	X_{N2}		X_{Np}

Základní pojmy

Korelační matice:

p manifestních proměnných

$$R = \begin{array}{cccc} 1 & r_{12} & r_{13} & \dots & r_{1p} \\ r_{21} & 1 & r_{23} & \dots & r_{2p} \\ r_{32} & r_{32} & 1 & \dots & r_{3p} \\ & & & \dots & \\ & & & r_{kj} & \\ & & & & r_{jk} \\ & & & & \\ r_{p1} & r_{p2} & r_{p3} & \dots & 1 \end{array}$$

p manifestních proměnných

Pozn.: Na obrázku je korelační matice (na diagonále jsou jedničky $r_{jj} = 1$, mimo diagonálu korelace r_{jk}), faktorová analýza ale velmi často pracuje s kovarianční maticí, kde na diagonále je rozptyl (σ_{jj}^2) a mimo diagonálu kovariance (σ_{jk}) příslušných proměnných. Typicky EFA pracuje s korelační, zatímco CFA s kovarianční maticí. Z kovarianční matice lze získat korelační matici snadno, $r_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}^2 \sigma_{kk}^2}}$. Naopak to nefunguje, protože korelační matice nenesou informaci o rozptylech.

Základní princip FA

- Korelace mezi dvěma manifestními proměnnými je způsobena tím, že tyto manifestní proměnné jsou **funkcemi** jednoho nebo více společných faktorů
- V rámci nějaké domény existuje (relativně) malé množství faktorů, které ovlivňují (relativně) velké (hypoteticky nekonečné) množství manifestních proměnných. Tím způsobují pozorovatelné korelace (kovariance) mezi těmito manifestními proměnnými
- Míra toho, jak moc ten který faktor ovlivňuje danou manifestní proměnnou, je reprezentována **faktorovým nábojem** – jakousi silou, s jakou faktor ovlivňuje manifestní proměnnou (0 = faktor MV neovlivňuje). Faktorové náboje jsou ekvivalentní **regresním koeficientům** – faktor je nezávislá proměnná (prediktor) a MV je závislá proměnná (outcome)

Model dat v FA

- Vraťme se k příkladu s během a šachy (a chvíli se tvařme, že žádné jiné latentní proměnné na světě neexistují)

$$\text{Čas } 100m_i = \lambda_B * \text{Schop. běh}_i + \lambda_\zeta * \text{Skill. šach}_i$$

- Čas, za který osoba i uběhne 100m, je lineární funkcí skóru osoby i na latentních proměnných *Schopnost běžet* a *Skill v šachu*
- λ_B a λ_ζ jsou mírou lineárního efektu těchto latentních proměnných na skór (čas) v manifestní proměnné *Běh na 100 metrů*. Jedná se o **faktorové náboje**
- Faktorové náboje nemají subscript i , nezávisí na dané osobě
- ...závisí však na MV. V tomto případě bude zřejmě platit $\lambda_\zeta = 0$

Model dat v FA

- Ovlivnily ale výkon osoby i pouze tyto latentní proměnné? Co když třeba sice dobře běhá, ale nemá rád krátké tratě (takže se moc nesnažil) a ještě k tomu mu špatně změřili čas?

$$\text{Čas } 100m_i = \lambda_B * \text{Schop. běh}_i + \lambda_\zeta * \text{Skill. šach}_i + \text{Nerad. krátké. tratě}_i + \text{Chyba}_i$$

- *Schopnost běhat* by ovlivnila i jiný výsledek člověka i , třeba v běhu na 1000 metrů – byla by v tomto případě tzv. *obecným / společným faktorem*
- Láska ke krátkým tratím i momentální chyba měření jsou v tomto případě tzv. *unikátním faktorem* – čas v běhu na 1000m neovlivní.
- Láska ke krátkým tratím je ale v tomto případě systematická – pokud by člověk i běžel 200m, projeví se a stane se v takovou chvíli obecným (společným faktorem). Takovou část unikátního faktoru nazýváme *specifickým faktorem*.

Common Factor Model

- Právě jsme si (konceptuálně) popsali tzv. Common Factor Model (L. L. Thurstone), který je modelem faktorové analýzy od 40. let 20. století do současnosti
- Dle CFM jsou manifestní proměnné funkcí dvou druhů faktorů:
 - **Obecných / společných faktorů (Common factors)**, které jsou *společné* dvěma a více MV v datové matici
 - **Unikátních faktorů (Unique factors)**, které ovlivňují pouze jednu MV. Unikátní faktory tak nevysvětlují (nezpůsobují) žádnou korelaci mezi dvěma MVs.

Common Factor Model

- Každý unikátní faktor se skládá ze dvou komponent:
 - Ze **specifického faktoru**
 - Z (náhodné) **chyby měření**

...specifický faktor reprezentuje nějaké systematické vlivy, které ovlivňují pouze jednu danou manifestní proměnnou. Chyba měření představuje náhodnou chybu.

- Pokud nemáme k dispozici žádné další informace, v modelu nelze chybu od systematického faktoru oddělit.
- Systematický faktor se ale může stát společným faktorem, jestliže nás začne zajímat nějaká další manifestní proměnná, která je jím také ovlivňována

Common Factor Model

- Rozptyl každé manifestní proměnné je rozložitelný následujícím způsobem:

Pozorovaný rozptyl = Společný rozptyl + Unikátní rozptyl

Unikátní rozptyl = Specifický rozptyl + Chybový rozptyl

→ 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

Common Factor Model

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

Průměr +

Obecné faktory

+ Unikátní faktor

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

μ_j je průměr manifestní proměnné j

Common Factor Model

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

Průměr +

Obecné faktory

+ Unikátní faktor

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

Common Factor Model

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

Common Factor Model

- Model dat slouží k vysvětlení struktury a podoby syrových dat (tedy skóru na manifestních proměnných)
- Faktorová analýza se však vlastně nezabývá strukturou a podobou syrových dat. Zabývá se vysvětlením kovariancí / korelací mezi MVs. Má to „malou“ výhodu – nepotřebujeme k tomu znát skóry osob na latentních proměnných (které stejně neznáme a znát nemůžeme – jsou nepozorované a neurčitelné [*indeterminate*])

Model kovarianční struktury

- Kovarianční struktura (tedy vysvětlení korelací / kovariancí) v Common Factor Modelu:

$$\Sigma = \Lambda\Phi\Lambda' + D_{\psi}$$

- Σ (sigma) je matice korelací / kovariancí mezi manifestními proměnnými
- Λ (lambda) je matice faktorových nábojů (apostrofov značí transpozici)
- Φ (phi / fí) je matice korelací / kovariancí mezi (obecnými) faktory. Faktory být korelované nemusí – v takovém případě lze říci, že faktory jsou tzv. *ortogonální*
- D_{ψ} (D-psi / D-psí) je matice rozptylů unikátních faktorů
- ...jak možná správně tušíte, k faktorové analýze nepotřebujete syrová data, ale korelace / kovariance mezi MVs.

O co nám tedy ve FA jde?

- Cílem 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

Příklad

Představme si, že pro vzorek jedinců máme k dispozici skóry ze 4 testů: porozumění textu (PC), slovní zásoba (VO), aritmetika (AR), matematické slovní úlohy (MPS). Z dat získáme následující korelační matici:

	PC	VO	AR	MPS
PC	1			
VO	.49	1		
AR	.14	.07	1	
MPS	.48	.42	.48	1

Příklad

Chtěli bychom identifikovat faktory, které „můžou“ za korelace mezi proměnnými, abychom těmto korelacím porozuměli. Aplikujeme metody faktorové analýzy a získáme následující matici faktorových nábojů:

	Faktor 1	Faktor 2
PC	.70	.10
VO	.70	.00
AR	.10	.70
MPS	.60	.60

Příklad

	Faktor 1	Faktor 2
PC	.70	.10
VO	.70	.00
AR	.10	.70
MPS	.60	.60

- Prvky v této matici představují sílu lineárního vztahu mezi každým faktorem a každým testem (manifestní proměnnou)
- Jaký může být význam Faktoru 1 a Faktoru 2?

Explorační a konfirmační FA

- 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!**

Na závěr

- Mějme na paměti, že FA je model – model, který reprezentuje nějakou hypotetickou strukturu uvnitř pozorovaných dat. Každý matematický model je – alespoň do nějaké míry – chybný a nedá se říct, že by perfektně a bez výhrad korespondoval s realitou
- Model, který nám sice dává smysl konceptuálně, ale vůbec neseďí na data, je (většinou) k ničemu
- Model, který skvěle sedí na data, ale nedává nám konceptuálně smysl, je (většinou) rovněž k ničemu
- Neplatí, že by jen tak jakákoli data byla vhodná pro faktorovou analýzu.