

Faktorová analýza

PSYb2590: Základy psychometriky | Přednáška 4

27. 3. 2023 | Hynek Cígler (& Adam Ťápal in memoriam)

FA v kostce

Pokud

$$\mathbf{R}_k^* = \mathbf{\Delta}_k (\mathbf{\Lambda}_k \mathbf{\Phi}_k \mathbf{\Lambda}'_k + \mathbf{I}) \mathbf{\Delta}_k = \mathbf{\Delta}_k \mathbf{\Lambda}_k \mathbf{\Phi}_k \mathbf{\Lambda}'_k \mathbf{\Delta}_k + \mathbf{\Delta}_k^2$$

a zároveň

$$\mathbf{\Sigma}_k^* = \begin{vmatrix} \mathbf{\Phi}_k + \mathbf{I} & \mathbf{\Phi}_k \mathbf{\Lambda}'_{sk} \\ \mathbf{\Lambda}_{sk} \mathbf{\Phi}_k & \mathbf{\Lambda}_{sk} \mathbf{\Phi}_k \mathbf{\Lambda}'_{sk} + \mathbf{I} \end{vmatrix}$$

pak platí:

$\rho_{X\tilde{X}}$

$$= \frac{\sum_{j=1}^J \sum_{j'=1}^J [\sum_{c=1}^{C-1} \sum_{c'=1}^{C-1} \Phi_2(\tau_{V_{jc}}, \tau_{V_{j'c'}}) - (\sum_{c=1}^{C-1} \Phi_1(\tau_{V_{jc}})) (\sum_{c=1}^{C-1} \Phi_1(\tau_{V_{j'c'}}))] \sum_{k=1}^K \sum_{k'=1}^K \lambda_{V_j^* F_k} \lambda_{V_{j'}^* F_{k'}} \rho_{F_k F_{k'}}}{\sum_{j=1}^J \sum_{j'=1}^J [\sum_{c=1}^{C-1} \sum_{c'=1}^{C-1} \Phi_2(\tau_{V_{jc}}, \tau_{V_{j'c'}}) - (\sum_{c=1}^{C-1} \Phi_1(\tau_{V_{jc}})) (\sum_{c=1}^{C-1} \Phi_1(\tau_{V_{j'c'}}))]}$$

Metaforické pochopení FA nestačí.

Je nezbytně nutné skutečně rozumět analytickému principu.

Uvedené vzorce je bezpodmínečně nutné chápat, znát a umět použít.

Jsou jednoduché; zbytek hodiny se je pokusíme interpretovat.



CTT vs. teorie latentních rysů (např. FA)

Klasická testová teorie:

- Položky jsou *paralelními* (zaměnitelnými) *testy* (měřítky) měřeného konstruktů.
- 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 „*vymyšlen*“ pro popis pozorování.

CTT vs. teorie latentních rysů (např. FA)

Teorie latentních rysů:

- Konstrukty *reálně existují*.
- Konstrukty kauzálně *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?

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

Latentní rys:

Schopnost rychle běžet

Pro srovnání CTT: Jak rychle zaběhl dohromady
 $100+400+800 = 1300\text{m?}$

Alternativně CTT: Jak rychle zaběhne průměrný závod
vylosovaný z domény běžných závodů?

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

Skill v šachu

Pro srovnání CTT: Kolikrát z 10+10+10 her porazil
cvičenou opici + okresního mistra + Kasparova?

Alternativně CTT: Kolikrát z deseti her porazí
průměrného soupeře vylosovaného z domény běžných
soupeřů?

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

Pojďme 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

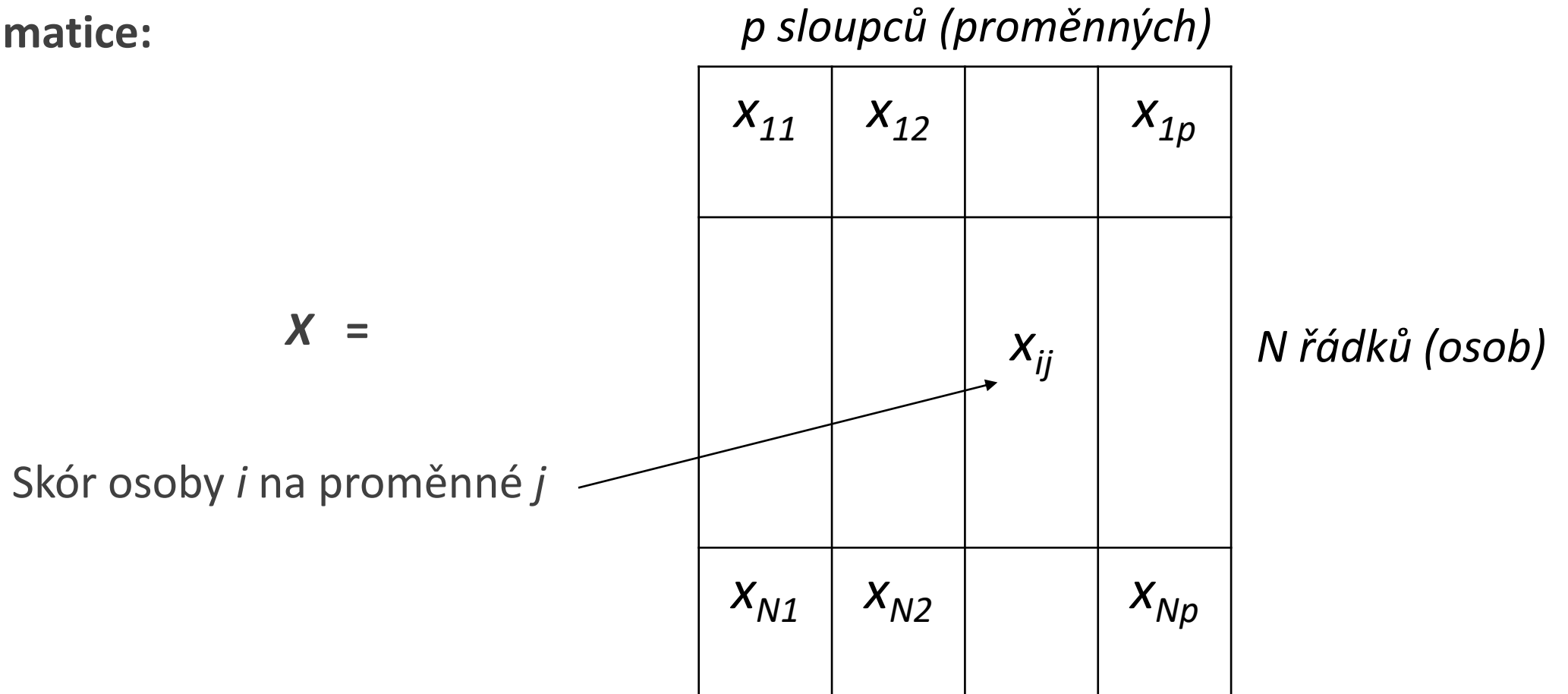
Datová matice:

Co řádek, to osoba

Co sloupec, to proměnná

Základní pojmy

Datová matice:



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:

$$R = \begin{array}{ccccccc} & \text{\textit{1}} & r_{12} & r_{13} & \dots & \dots & \dots & r_{1p} \\ r_{21} & & \text{\textbf{1}} & r_{23} & & & & r_{2p} \\ r_{32} & & r_{32} & \text{\textbf{1}} & & & & r_{3p} \\ \vdots & & & & \ddots & r_{kj} & & \\ \vdots & & & & r_{jk} & \ddots & & \\ \vdots & & & & & & \ddots & \\ r_{p1} & r_{p2} & r_{p3} & & & & & \text{\textbf{1}} \end{array} \begin{array}{l} p \text{ manifestních proměnných} \\ \\ \\ \\ \\ \\ p \text{ manifestních} \\ \text{proměnných} \end{array}$$

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.

Off-topic: 2 druhy analýz

Pokud chceme analyzovat nějaký dataset pomocí faktorové analýzy (aj.), máme v zásadě dvě možnosti:

Limited-information approach:

- Nevyužijeme *všechna* data, ale data si (1.) zjednodušíme a pak (2.) analyzujeme tato zjednodušená data.
- Tento přístup má nějaké omezení (předpoklady) a mnoho výhod (analytická jednoduchost).
- Typicky: EFA i CFA, které pracují právě s kovariační maticí (tedy bivariačními statistikami položek).

Full-information approach:

- Pro analýzu využijeme všechna data.
- Tento postup má méně omezení, občas není potřebný, je statisticky náročnější, ale má řadu výhod.
- Typicky: tzv. item-factor analysis (teorie odpovědi na položku), modelování nelineárních vztahů, ale třeba i tzv. FIML práce s chybějícími daty v CFA (částečně, stále pracuje s kovarianční maticí).

Základní princip a předpoklady 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
 - Existovaly dřívější modely faktorové analýzy, jako např. analýza tetrád aj.
 - Existují i jiné příbuzné modely, jako např. analýza hlavních komponent (PCA). Neplést!
- 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 specifického (systematického) faktoru oddělit.
- Specifický (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

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

... = 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 (mí; mean) je průměr manifestní proměnné j

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

λ_{jk} (lambda; loading) 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 jen a pouze obecnými faktory.

- Porušení tohoto předpokladu vede k neshodě modelu s daty.

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.

- V případě jediné dimenze je kovariance položek a, b rovna $\sigma_{ab} = \lambda_a \lambda_b \phi$, kde ϕ je rozptyl faktoru.
- V případě korelační (nikoli kovarianční) matice a pokud $\phi = 1$ (typicky EFA) platí $r_{ab} = \lambda_a \lambda_b$.

Má to tedy „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 (a případně reziduálních kovariancí).
- ...jak možná správně tušíte, k faktorové analýze nepotřebujete syrová data, ale korelace / kovariance mezi MVs.

Model kovarianční struktury

Vzorec

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

Ize rozepsat do rovnice pro každý pár dvou položek (případně pro jedinou položku, pokud $i = j$).

Kovariance σ_{ij}^2 proměnných i, j (případně rozptyl jediné proměnné i , pokud $i = j$) je v případě přítomnosti dvou faktorů f a g roven:

$$\sigma_{ij}^2 = \lambda_{if}\lambda_{jf}\phi_{ff} + \lambda_{ig}\lambda_{jg}\phi_{gg} + \lambda_{if}\lambda_{jg}\phi_{fg} + \theta_{ij}$$

- λ_{if} – náboj položky i na faktoru f .
- ϕ_{ff}, ϕ_{gg} – rozptyly faktorů f, g .
- ϕ_{fg} – kovariance faktorů f, g .
- θ_{ij} – reziduální kovariance položek i, j (typicky 0).

V případě F faktorů:

$$\sigma_{ij}^2 = \theta_{ij} + \sum_{f=1}^F \sum_{g=f}^F \lambda_{if}\lambda_{jg}\phi_{fg}$$

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

porozumění textu (PC)
slovní zásoba (VO)
aritmetika (AR)
matematické slovní úlohy (MPS)

Příklad

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

porozumění textu (PC)
slovní zásoba (VO)
aritmetika (AR)
matematické slovní úlohy (MPS)

- 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!**

Explorační a konfirmační FA

Ve světě faktorové analýzy rozlišujeme dvě situace.

Explorační (exploratory/unrestricted) FA: EFA

- Nemáme žádnou (nebo jen mlhavou) představu o tom, kolik faktorů a jakého charakteru je „za daty“.
- Cílem je prozkoumat data a navrhnout možnou teoretickou strukturu, která by data „vysvětlila“.

Konfirmační (confirmatory/restricted) FA: CFA

- Máme celkem jasnou představu o tom, kolik faktorů a jakého charakteru je „za daty“.
- Cílem je otestovat teoretickou strukturu oproti těmto datům.

Ovšem teoretický model, který v obou případech používáme, je totožný!

- I CFA lze použít „exploračně“ (například srovnat několik různých modelů, modely různě upravovat a snažit se je „napasovat na data“ a podobně).

Explorační faktorová analýza (EFA)

Long story short:

- Vybereme data, na něž EFA aplikujeme.
- Zvolíme si počet faktorů. Můžeme použít nějakou z vhodných metod.
- Zvolíme si metodu odhadu.
- Pomocí SW realizujeme vlastní analýzu. To zahrnuje:
 - Odhad korelační matice (co s chybějícími daty?).
 - Odhad faktorových nábojů s využitím této matice.
- Vyhodnotíme shodu modelu s daty.
- Zvolíme si metodu rotace a rotaci pomocí SW provedeme (více později).
- Vyhodnotíme, zda výsledek „působí“ přijatelně (teorie, zkušenost...).

EFA: předpoklady

Obecné (společné) faktory a unikátní faktory jsou nezávislé a nekorelují spolu.

Unikátní faktory jsou navzájem rovněž nezávislé a nekorelují spolu.

Obecné a unikátní faktory mají z definice průměr 0.

Obecné a unikátní faktory mají z definice rozptyl 1 (a tedy i $SD = 1$).

Bez těchto předpokladů není model identifikovaný a není jej možné odhadnout.

EFA: počet faktorů

Počet faktorů může být odvozen teoreticky na základě očekávání. Občas je ale očekávání příliš vágní, a proto lze využít vhodnou metodu odhadu.

Klíčové slovo: *eigenvalue* (vlastní hodnota).

- Spojeno s vysvětleným rozptylem. Jde o vyjádření množství vysvětleného rozptylu jedním faktorem.
- Každému faktoru je přiřazena určitá hodnota eigenvalue.
- Protože lze extrahovat maximálně tolik faktorů, kolik je položek, maximální suma eigenvalue dané korelační matice je rovna počtu položek.
- 1 eigenvalue lze tedy vnímat jako variabilita dat vysvětlená 1 průměrnou položkou.
- *Eigenvalue/(počet položek)* reprezentuje u (nekorelovaného) řešení tzv. „vysvětlený rozptyl“.

Dva druhy eigenvalue:

- PCA: eigenvalue („před extrakcí“) je navýšené o chybu měření. Většina pravidel (včetně popisu výše) pracuje s touto variantou.
- EFA: eigenvalue („po extrakci“) je očištěné o chybu měření.

EFA: počet faktorů

Kaiserovo kritérium (Kaiserovo-Guttmanovo).

- Extrahujeme všechny faktory, jejichž eigenvalue (před extrakcí) > 1 .
- Nepochopeno a zneužíváno, slabá opora v teorii – jen silně orientační! Typicky nadhodnocuje.

Vizuální inspekce (Cattellova) scree-plotu.

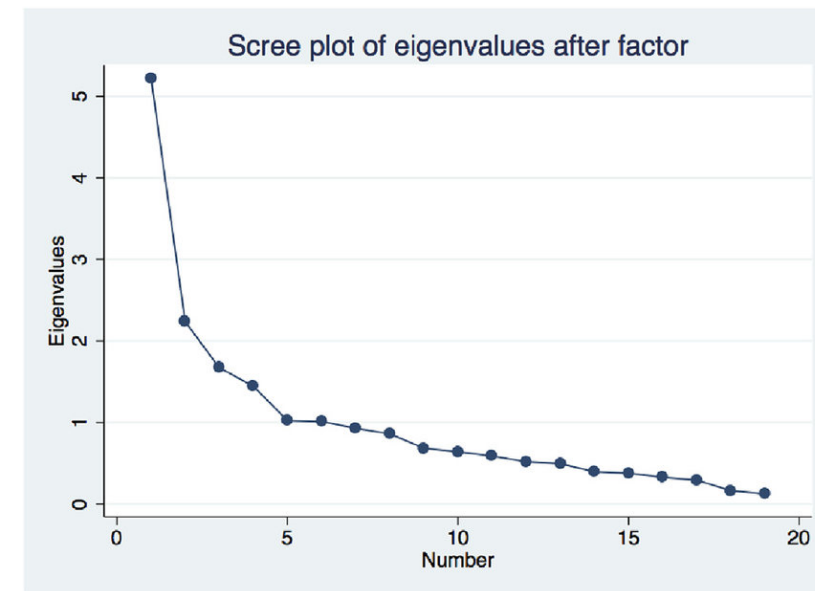
- Eigenvalue se seřadí podle velikosti a vynesou do grafu.
- Identifikuje se bod zlomu a vezmou se všechny faktory s vyšším (a nebo stejným?) eigenvalue.
- Subjektivní, problematické, ale dobrý nástroj pro exploraci.

(Hornova) paralelní analýza.

- Kombinuje výhody obojího (Kaiserovo kritérium + vizuální inspekce).
- Očišťuje Kaiserovo kritérium o výběrovou chybu.

Velicerovo MAP (Minimum Average Partial).

- Optimální počet faktorů je takový počet, který ještě modeluje nějakou systematickou korelaci MV.



EFA: rotace a rotační indeterminace

Rotační indeterminace = rotační neurčitost.

$$\text{Vzorec FA: } \Sigma = \Lambda\Phi\Lambda' + D_{\psi}$$

Jako řešení hledáme matici Λ , která co nejlépe popíše data.

Drobný problém (při 2 a více faktorech): existuje nekonečně mnoho matic faktorových nábojů Λ , které jsou lineárními transformacemi původní matice.

Pokud tedy najdeme nějaké řešení, našli jsme nekonečně mnoho **stejně dobrých řešení**.

Klíčová otázka: *Které z nich je to „správné“?*

EFA: rotace a rotační indeterminace

Koncept jednoduché struktury (Thurstone; **simple structure**).

Toho využívají tzv. rotace: najdou řešení, které splní nějaké kritérium jednoduché struktury.

- A díky tomu se nám výsledek *snáze interpretuje*.

Žádná rotace: první faktor vysvětlí maximum rozptylu, zbylé faktory popisují „odlišnosti“.

Rotace: snaží zjednodušit faktorovou strukturu. Velké množství rotací, obecně:

Ortogonální rotace: faktory spolu nekorelují.

Šikmé rotace: faktory spolu korelují.

TLDR: Pokud nevíte, používejte šikmé rotace 😊

EFA: 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

EFA: 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)

EFA: 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)

EFA: 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.
- Alespoň $N = P * F * 5$ respondentů, kde P je počet položek a F počet faktorů
- Nepoužívejte PCA
- Nejmenší čtverce jsou vhodnou první volbou pro estimátor

EFA: Pár tipů

- 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.

Konfirmační faktorová analýza (CFA)

CFA testuje teoretický model oproti datům.

Testování spočívá v tom, že *omezíme* nějaké parametry, čím dostaneme *restriktivnější* model oproti EFA.

- Omezení: parametry jsou nastavené do nějaké rozmezí (např. >0 , $<0,05$ atp.).
- Fixování: parametry jsou nastavené na nějakou konkrétní hodnotu (typicky 0).

Základní princip: některé faktorové náboje jsou nastaveny na 0. Tím je přímo „vytvořena“ jednoduchá struktura.

Tato omezení (zpravidla) vedou k jednoznačné identifikaci faktorů, není nutná rotace.

Konfirmační faktorová analýza (CFA)

Long story short:

- Vyjasním si faktorovou strukturu, „zvolím“ si jednoduchou strukturu.
- Zvolím estimátor (jiné možnosti než u EFA).
- Pomocí SW odhadnu model, případně modely (lze navzájem srovnat).
- Pokud model neseďí na data, zvážím pár úprav.
- Pokud model nelze zlepšit, zamítám jej (a mohu nalézt lepší např. pomocí EFA).

Posouzení shody s daty

1. Reziduální kovarianční matice (CFA i EFA).

- Které predikované kovariance se liší od pozorovaných kovariancí? Lokální shoda...

2. Chí-kvadrát test (CFA i EFA).

- Signifikance odlišnosti pozorované a predikované kovarianční matice. Enormní síla testu.

3. Indexy dobré/přibližné shody (CFA i EFA).

- Jak věcně významná je tato odlišnost?

4. Modifikační indexy (typicky jen CFA).

- Které omezené parametry by šlo uvolnit a tím zlepšit shodu s daty?

Ale o tom všem na semináři...