MUNI FSS Katedra psychologie

Faktorová analýza

PSYb2590: Základy psychometriky | **Přednáška 4**

28. 3. 2022 | Adam Ťápal & Hynek Cígler

FA v kostce

Pokud

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

a zároveň

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

pak platí:

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

Většina vysvětlení FA je příliš redukující a zkreslující.

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

Když se na to podíváte, je to vlastně jednoduché.

- Vysvětlím na tabuli.

CTT vs. teorie latentních rysů (např. FA)

Klasická testová teorie:

- Položky jsou *paralelními* (zaměnitelnými) *testy* (měřítky) měřeného konstruktu
- 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

Pro srovnání CTT: Jak rychle zaběhl dohromady 100+400+800 = 1300m?
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 souborů?

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: (20*s*, 90*s*, 180*s*)
Šachy: (3, 1, 0)

Bruno
Běh: (40*s*, 180*s*, 300*s*)
Šachy: (4, 2, 1)

Cecil
Běh: (50*s*, 190*s*, 320*s*)
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

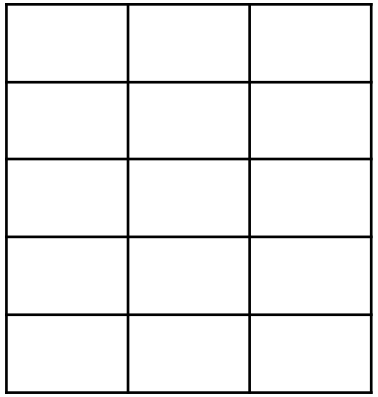
- Jaká je typická podoba dat v případě faktorové analýzy?

- 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



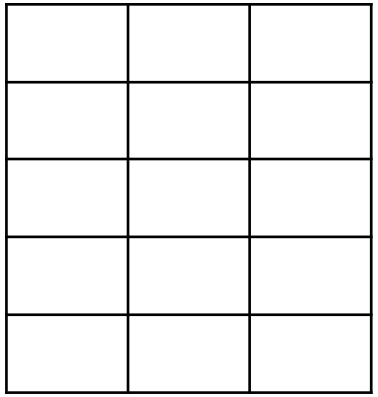
Základní pojmy

- 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

Co sloupec, to proměnná

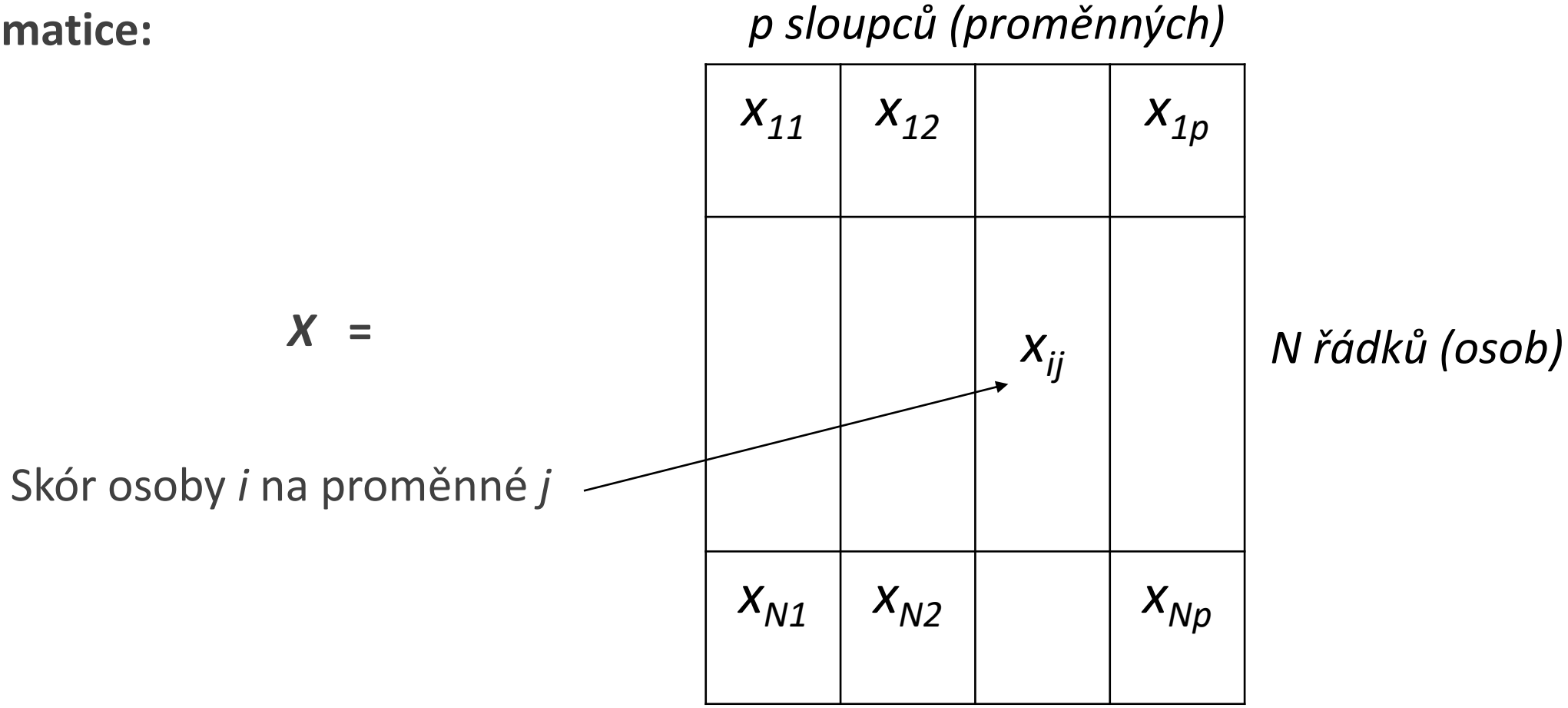
Datová matice:

Co řádek, to osoba



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 **=**

p manifestních proměnných

1	r_{12}	r_{13}			r_{1p}
r_{21}	1	r_{23}			r_{2p}
r_{32}	r_{32}	1			r_{3p}
				r_{kj}	
			r_{jk}		
r_{p1}	r_{p2}	r_{p3}			1

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 nenese 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)

$$Čas\,100m_i = \lambda_B * Schop.běh_i + \lambda_Š * Skill.šach_i$$

- 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 $\lambda_Š$ 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 $\lambda_Š = 0$

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?

$$Čas\,100m_i = \lambda_B * Schop.běh_i + \lambda_Š * Skill.šach_i + Nerad.krátké.tratě_i + Chyba_i$$

- *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 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

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

... = 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 skórů 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:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{D}_{\boldsymbol{\psi}}$$

- **Σ** (sigma) je matice korelací / kovariancí mezi manifestními proměnnými
- **Λ** (lambda) je matice faktorových nábojů (apostrof 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.

Model kovarianční struktury

Vzorec

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{D}_{\boldsymbol{\psi}}$$

lze 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:

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

- λ_{if} – náboj položky i na faktoru f
- ϕ_{fg} – korelace 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**!

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 nesedí 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.