Úvod do strukturního modelování

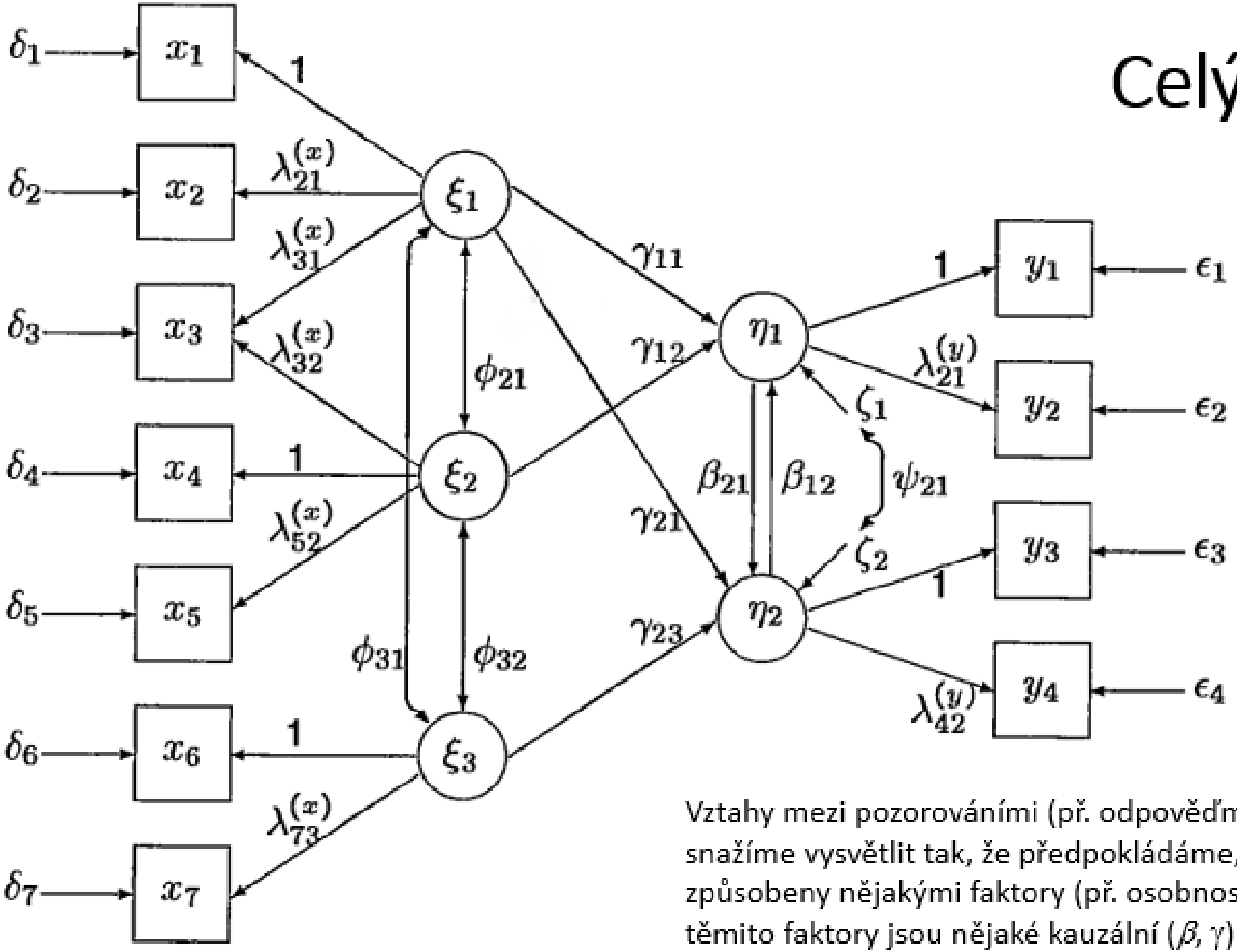
PSY028_E – Statistická analýza dat v psychologii

Blok 2 – Faktorová analýza

Program

- 1. 09:00 – 12:00 : Uvedení do faktorové analýzy
- 2. 12:00 – 13:00 : Přestávka
- 3. 13:00 – 15:00 : Faktorová analýza v R / *lavaan*

Celý SEM model



Vztahy mezi pozorováními (př. odpověďmi na položkami – *x*, *y*) se snažíme vysvětlit tak, že předpokládáme, že tyto odpovědi jsou způsobeny nějakými faktory (př. osobnostními – ξ, η) a, že mezi těmito faktory jsou nějaké kauzální (β, γ) či nekauzální (ϕ, ψ) vztahy.

Dva základní pojmy

- **Manifestní proměnná (MV)** – proměnná, kterou lze přímo měřit či pozorovat

- **Latentní proměnná (LV)** – proměnná, kterou *nelze* přímo měřit či pozorovat – hypotetický konstrukt. **Faktory** v rámci FA jsou právě latentními proměnnými. Tedy – faktor je stále nějaká proměnná a různí lidé „mají“ své skóry na této proměnné (základní předpoklad)

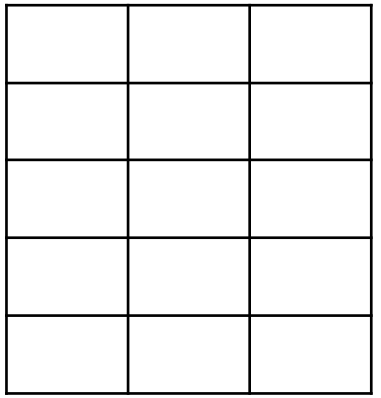
Základní principy FA

- 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...)

Co sloupec, to proměnná

Datová matice:



Co řádek, to osoba

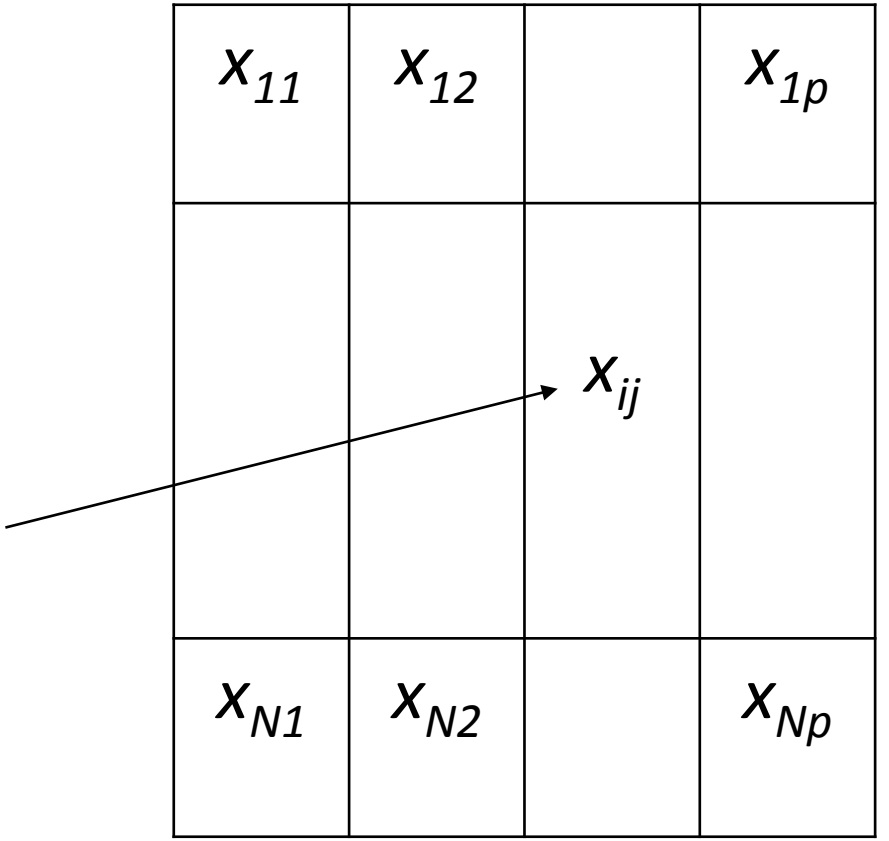
Základní principy FA

Datová matice:

p sloupců (proměnných)

X **=** *N řádků (osob)*

Skór osoby *i* na proměnné *j*



R:

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

- Cílem faktorové analýzy je **odhalit** a **pochopit** strukturu, která „způsobuje“ korelace mezi manifestními proměnnými – a to pomocí **faktorů**.

- Základní princip – v rámci domény existuje (relativně) malé množství faktorů, které ovlivňují (relativně) velké množství manifestních proměnných a tím způsobují mezi těmito manifestními proměnnými korelace (kovariance)

- Korelace mezi dvěma manifestními proměnnými je způsobena tím, že tyto manifestní proměnné jsou funkcemi jednoho nebo více společných faktorů

- To, jak moc ten který faktor ovlivňuje danou manifestní proměnnou, je reprezentováno **faktorovými náboji** (*factor loadings*).

- Hodnoty těchto faktorových nábojů představují sílu lineárního vztahu mezi faktorem a manifestní proměnnou. Faktorové náboje jsou ekvivalentní regresním koeficientům – faktor je nezávislá proměnná a MV je závislá proměnná

- Celkový obraz faktorových nábojů nám napomáhá v interpretaci podstaty faktoru

- Rozptyl každé manifestní proměnné je rozložitelný následujícím způsobem na několik základních komponent:

$$\text{Pozorovaný rozptyl} = \text{Obecný rozptyl} + \text{Unikátní rozptyl}$$

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

...podíl pozorovaného rozptylu, který je způsoben obecnými faktory (takové R^2)

Model dat ve faktorové analýze

Jak vlastně vypadá?

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

Průměr + **Obecné faktory** **+ Unikátní faktor**

Kde:

- x_{ij} **je skóre osoby *i* na manifestní proměnné *j***
- μ_j **je průměr manifestní proměnné *j***
- z_{ik} **je skóre osoby *i* na obecném faktoru *k***
- λ_{jk} **je faktorový náboj manifestní proměnné *j* na obecném faktoru *k***
- u_{ij} **je skóre osoby *i* na unikátním faktoru *j***

Model dat ve faktorové analýze

- 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)

- 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**!

...v kurzu se budeme věnovat pouze **konfirmační FA**

Model dat ve faktorové analýze

- **Vstupujeme do světa, kde už nám takový zápis pro dobré porozumění přestává stačit, a je potřeba začít s maticovou algebrou:**

$$\boldsymbol{x = \mu + \Lambda z + u}$$

- **Model dat slouží k vysvětlení struktury a podoby syrových dat (tedy skórů na manifestních proměnných)**

- **Faktorová analýza se ale vlastně nezabývá strukturou a podobou syrových dat. Zabývá se vysvětlením kovariancí / korelací mezi manifestními proměnnými. Má to „malou“ výhodu – nepotřebujeme k tomu znát skóry osob na latentních proměnných, které stejně neznáme a znát nemůžeme – jsou nepozorovatelné a neurčitelné.**

Kovarianční struktura

- **Kovarianční struktura v maticovém zápisu:**

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

- $\boldsymbol{\Sigma}$ **je matice korelací / kovariancí mezi manifestními proměnnými**
- $\boldsymbol{\Lambda}$ **je matice faktorových nábojů**
- $\boldsymbol{\Phi}$ **je matice korelací / kovariancí mezi (obecnými) faktory. Faktory (obecné) být korelované nemusí – v takovém případě lze říci, že faktory jsou tzv. ortogonální.**
- $\boldsymbol{D}_{\psi}$ **je matice rozptylů unikátních faktorů**
- **...jak možná správně tušíte, k faktorové analýze syrová data nepotřebujete. Jako vstup postačí korelační / kovarianční matice MV**

Kovarianční struktura

- **Model kovarianční struktury**

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

...je pořád jen model. Pokud se nejedná o *právě identifikovaný* model, matice korelací / kovariancí nebude vysvětlena perfektně.

Identifikace

- Vše, co bylo včera řečeno i identifikaci, nadále platí – počet odhadovaných parametrů nemůže být větší, než počet „kousků informace“
- To však nutně nestačí k tomu, aby byl model identifikovaný

- Rozptyl faktoru je nutno nějak určit – 3 způsoby

(Omezit rozptyl faktoru, omezit 1 náboj přímo, omezit 1 náboj nepřímo)

- Umožnit unikátní řešení

CFA model

- Matice $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$ a $\boldsymbol{D_{\psi}}$ obsahují parametry modelu
- Hypotéza ohledně počtu a povaze faktorů je přímo „přeložena“ do modelu prostřednictvím prvků těchto tří matic
- Parametry modelu můžeme rozdělit do tří skupin:
  - Volně odhadované (free parameters)
  - Omezené na jednu hodnotu (fixed parameters)
  - Omezené vztahem s dalšími parametry (constrained parameters)

CFA model

- Tato omezení jsou zdaleka nejčastěji jedna a tatáž – předem určujete, které parametry (faktorové náboje, korelace mezi faktory) nabývají hodnoty 0.

- Máte jasno ohledně jak **počtu**, tak **charakteru** faktorů
  -- vaše hypotéza se týká toho, kolik tušíte faktorů a jaké proměnné by měl ten který faktor ovlivňovat, a jaký je vztah jednotlivých faktorů mezi sebou

CFA model

- **Představme si situaci, kdy máte šest manifestních proměnných (x_1 až x_6) a dva faktory (z_1 a z_2).**

- **Vaše hypotéza zní:**
  - **Faktorové náboje prvních tří manifestních proměnných na faktoru z_1 mají významnou velikost (jsou nenulové) a náboje dalších tří manifestních proměnných jsou v podstatě nulové.**
  - **Faktorové náboje prvních tří manifestních proměnných na faktoru z_2 jsou v podstatě nulové a náboje dalších tří manifestních proměnných mají významnou velikost (jsou nenulové).**
  - **Faktory z_1 a z_2 spolu korelují.**

CFA model

- Nejdříve si představíme modelové matice. Máme $p = 6$ manifestních a $m = 2$ latentních proměnných.
- Z toho víme, že $\boldsymbol{\Lambda}$ má 6 řádků a 2 sloupce, $\boldsymbol{\Phi}$ má velikost 2 x 2 a $\boldsymbol{D}_{\boldsymbol{\psi}}$ má velikost 6 x 6
- Pojďme tedy matice zkonstruovat a zaplnit je volně odhadovanými i omezenými parametry

CFA model

CFA model

CFA model

- Kolik máme stupňů volnosti?

- Nejdřív pojďme spočítat volně odhadované parametry:

  6 faktorových nábojů + 6 reziduálních rozptylů + 1 korelace mezi faktory = 13 parametrů k odhadnutí

- Naše data, matice korelací mezi pozorovanými proměnnými, je 6 x 6 korelační matice (může být i kovarianční, ale...), která obsahuje [6 * (6-1)]/2 = 15 ne-redundantních prvků – „kousků informace“
- Počet stupňů volnosti je tedy 15 – 13 = 2

CFA model

- Když odhadnu model:
  - Vypadají odhady parametrů v pořádku? Dostal jsem nějaká varování?
  - Mají parametry přípustné hodnoty?
  - Dávají mi hodnoty odhadnutých parametrů smysl?
  - Jak vypadají směrodatné chyby odhadu parametrů?

  - Jak model sedí na data?

Shoda modelu s daty

- Dobrá shoda s daty ještě neznamená, že váš model je „nejlepší“ nebo „správný“

- Reziduální korelační / kovarianční matice
- Chí-kvadrát modelu, $\chi^2_M = (N-1)F_{ML}$ se stupni volnosti jako má model
- Podíl χ^2_M k počtu stupňů volnosti (χ^2/df ratio) – více konvencí
- RMSEA, TLI, CFI, SRMR – indexy fitu (inkrementální, absolutní, reziduální)
- Indexy založené na teorii informace – AIC a BIC

RMSEA

- **RMSEA = Root Mean Square Error of Approximation**
- **Steiger & Lind, 1980; Browne & Cudeck, 1992**

$$\text{RMSEA} = \sqrt{\frac{F_0}{df}}, \text{ kde}$$

...*df* je počet stupňů volnosti modelu

... $F_0 = \hat{F} - \frac{df}{N-1}$, kde $\hat{F}$ je hodnota diskrepanční funkce

RMSEA

- Browne & Cudeck, 1992 o hodnotách RMSEA:

- < .05 -- close fit
- .05 - .08 -- good fit
- .08 - .10 -- acceptable fit
- > .10 -- unacceptable fit

- Konfidenční interval RMSEA je důležitější, než bodový odhad

TLI

- **Tucker-Lewis Index:**

$$\frac{(\chi_0^2/df_0) - (\chi_m^2/df_m)}{(\chi_0^2/df_0) - 1}$$

- **Srovnání odhadnutého modelu (*m*) s nulovým modelem (*0*)**
- **Vysoce korelovaný s CFI, TLI je přísnější**
- **Doporučené hodnoty: >.95 excellent, >.90 good**

Informační kritéria

- Založeny na *deviance* – funkci hodnoty diskrepanční funkce
- Deviance = -2*log-likelihood
- Kombinují shodu modelu s daty (deviance) s komplexitou modelu (počtem parametrů)
- Akaike's Information Criterion (AIC):

$$AIC = 2k + Deviance$$

- Bayesian (Schwarz) Information Criterion (BIC):

$$BIC = ln(n)\, k + Deviance$$