ANALÝZA A KLASIFIKACE DAT

prof. Ing. Jiří Holčík, CSc.

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

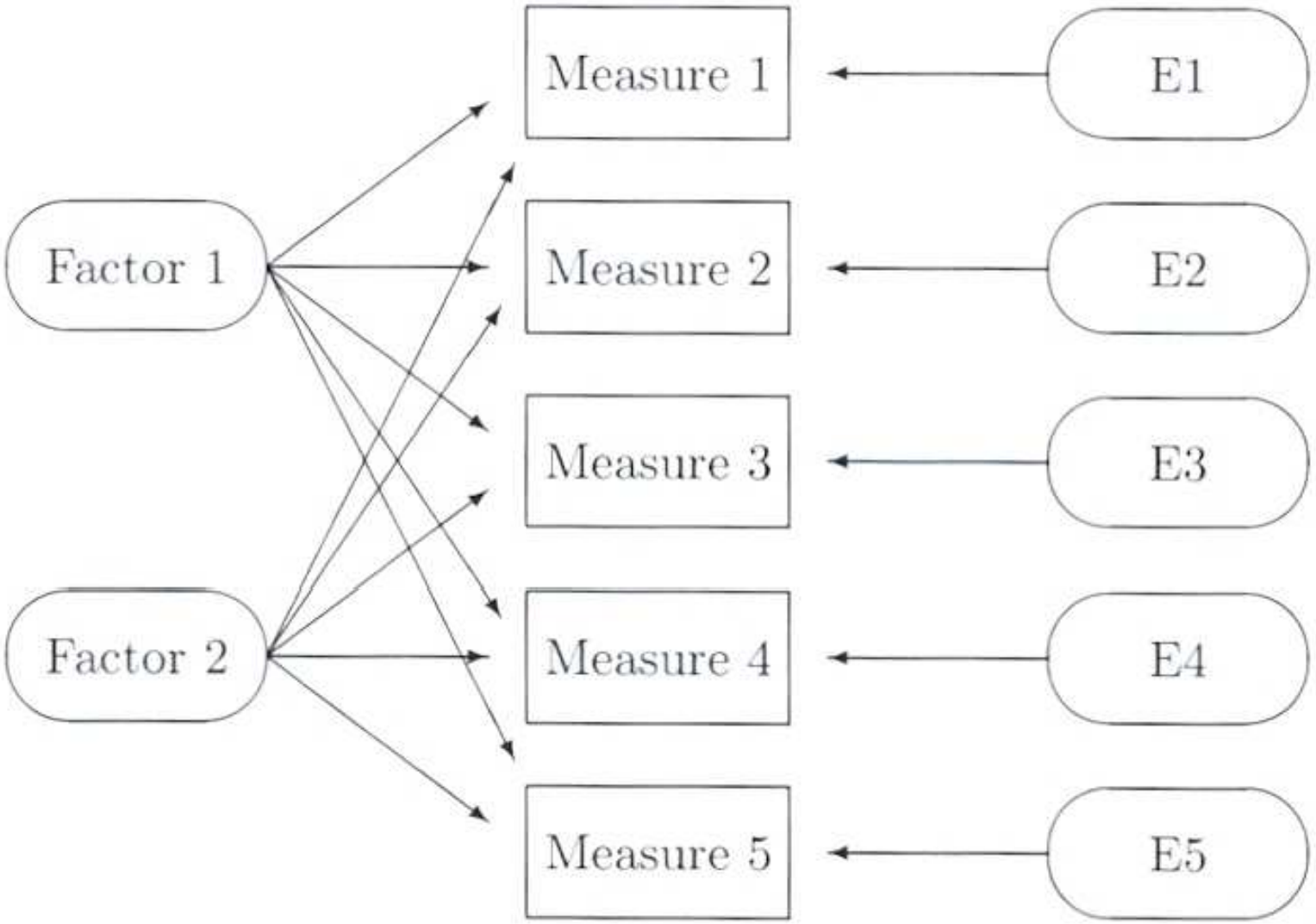
XI.
FAKTOROVÁ ANALÝZA

CO TO JE?

Faktorová analýza je (statistická) metoda používaná k popisu variability známých (pozorovaných, naměřených) proměnných pomocí nejlépe menšího (určitě ne většího) počtu skrytých (latentních) proměnných, zvaných **faktory**.

Zřejmě nediskutovanější multivariační analytická metoda. Nejvýznamnější kritika vychází ze subjektivity, která je nezbytná při interpretaci jejích výsledků.

CO TO JE?



DEFINICE

předpokládejme, že známe hodnoty m proměnných $\mathbf{x}_1, \ldots, \mathbf{x}_m$ se středními hodnotami $\mu_1, \ldots, \mu_m$. Dále předpokládejme, že pro neznámé konstanty a_{ij} a n skrytých proměnných f_j, i=1, 2, .., m a j=1,…,n platí

$$x_i - \mu_i = \lambda_{i1} f_1 + \ldots + \lambda_{in} f_n + \varepsilon_i$$

koeficienty λ_{ij} nazýváme **faktorové zátěže** (faktor loadings) i-tého obrazu u k-tého společného faktoru

ε_i jsou statisticky nezávislé **chybové členy** (**chybové, specifické faktory**) s nulovým průměrem a konečným rozptylem ($var(\varepsilon_i)=\psi_i$)

DEFINICE

$$\mathrm{cov}(\boldsymbol{\varepsilon}) = \mathrm{diag}(\psi_1, \ldots, \psi_m) = \boldsymbol{\Psi} \text{ a } E(\boldsymbol{\varepsilon}) = 0$$

maticově:

$$\mathbf{x} - \boldsymbol{\mu} = \boldsymbol{\Lambda}.\mathbf{f} + \boldsymbol{\varepsilon}$$

další předpoklady (pro F):

- ☑ **f** a **ε** jsou nekorelované (nezávislé);
- ☑ $E(\mathbf{f}) = 0$
- ☑ $\mathrm{cov}(\mathbf{f}) = \mathbf{I}$

DEFINICE

pokud cov(x) označíme **Σ**, pak za uvedených podmínek máme:

$$\text{cov}(\mathbf{x} - \boldsymbol{\mu}) = \text{cov}(\boldsymbol{\Lambda}.\mathbf{f} + \boldsymbol{\varepsilon}) \text{ nebo}$$

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}.\text{cov}(\mathbf{f}).\boldsymbol{\Lambda}^T + \text{cov}(\boldsymbol{\varepsilon}) \text{ nebo } \boldsymbol{\Sigma} = \boldsymbol{\Lambda}.\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}$$

základní faktorová věta

v praxi je matice **Σ** nahrazována výběrovou korelační **R**, resp. kovarianční maticí;

Λ je matice faktorových zátěží

$\boldsymbol{\Lambda}.\boldsymbol{\Lambda}^T$ představuje kovariační matici vektoru **Λ.f**

$\boldsymbol{\Psi} = \boldsymbol{\Gamma}^2$ matice jedinečností – kovarianční matice chybových faktorů – je diagonální, protože předpokládáme nekorelované chyby (diagonální prvky matice $\boldsymbol{\Gamma}^2$ jsou rozptyly jednotlivých sloupců zdrojové matice

DEFINICE

$$\mathbf{S}^2 = \mathbf{H}^2 + \mathbf{\Gamma}^2$$

$\mathbf{S}^2$ je diagonální matice rozptylů faktorů;

proměnlivost každého faktoru vyjádřenou sloupci zdrojové matice můžeme rozdělit do dvou složek

$\mathbf{H}^2$ – *komunalita* – představuje proměnlivost společnou všem faktorům; váha s jakou jednotlivé faktory přispívají k rozptylu odpovídající proměnné, čtverec komunality je suma faktorových zátěží faktorů

$\mathbf{\Gamma}^2$ – *jedinečnost* – část variability nevysvětlenou faktory, bývá dále rozdělena na část specificity (ta část proměnlivosti, kterou nelze vysvětlit ani chybou experiemntu, ani společnými faktory) a část nespolehlivosti (experimentální chyba při měření faktorů)

DEFINICE

faktorizace určená základní faktorovou větou nemusí existovat a pokud ano nemusí být řešení jednoznačné

pokud je **T** ortogonální matice o rozměru n x n, pak

$$(\mathbf{\Lambda T}).(\mathbf{\Lambda T})^T = \mathbf{\Lambda}.\mathbf{\Lambda}^T$$

to znamená, že pokud je **Λ** určená matice faktorových zátěží, pak **ΛT** je jí také, a i když jsou to různé matice, mohou generovat tutéž kovarianční strukturu – můžeme tedy otáčet původní řešení a hledat alternativní lepší řešení

CÍLE

základním cílem je samozřejmě redukce dat

v případě PCA promítneme původní data na vlastní vektory autokorelační (autokovarianční) matice tak, abychom získali skóre složek

$$\boldsymbol{\xi}_r = \mathbf{A}_r^T.\mathbf{x}, \text{ resp. } \boldsymbol{\xi}_r = \mathbf{A}_r^T.(\mathbf{x} - \boldsymbol{\mu})$$

(neúplná komponentní analýza)

CÍLE

základním cílem je samozřejmě redukce dat

v případě faktorové analýzy je to jinak, základní vztah mezi pozorovanými **x** a skrytými **ξ** proměnnými je

$$\mathbf{x} = \mathbf{\Lambda}.\boldsymbol{\xi} + \boldsymbol{\varepsilon}$$

(úplná komponentní analýza)

když m<n, pak není možné invertovat rovnici pro výpočet **ξ** z **x** a odtud spočítat skóre faktorů

ALGORITMUS

1. pro dané hodnoty proměnných spočítat výběrovou kovarianční matici a provést faktorovou analýzu pro určitý počet faktorů;
2. provést test souhlasu modelu s daty; pokud není, návrat na 1;
3. rotace faktorů k poskytnutí maximálních faktorových zátěží pro každou proměnnou;
4. seskupit proměnné pro každý faktor a interpretovat faktory (není třeba, když není třeba);
5. odhadnout skóre faktorů, což poskytne výklad dat s redukovaným rozměrem;

VÝBĚR FAKTORŮ

hlavní faktorová metoda (principal factor m.)

vybírá jako první ten faktor, který znamená maximální celkovou komunalitu, druhý faktor je zvolen ten, který maximálně přispěje k nárůstu komunality, … To je ekvivalentní nalezení vlastních čísel a vektorů redukované korelační matice **R***, definované jako korelační matice **R**, která má diagonální prvky nahrazené komunalitami

$$\mathbf{R}^* = \mathbf{R} - \mathbf{\Psi} = \mathbf{\Lambda}.\mathrm{cov}(\mathbf{f}).\mathbf{\Lambda}^T = \mathbf{\Lambda}.\mathbf{\Lambda}^T$$

ROTACE FAKTORŮ

- faktorový model je invariantní vůči ortogonální transformaci faktorů. Můžeme tedy nahradit matici faktorových zátěží $\boldsymbol{\Lambda}$ maticí $\boldsymbol{\Lambda T}$ bez změny aproximace modelu vzhledem ke kovarianční matici. To umožňuje určitou míru flexibility, která dovoluje otáčet s faktory a pomoci k jejich lepší interpretaci.
- Hlavním cílem rotačních technik je určit faktory tak aby naměřené proměnné měly vysoké zátěže u malého počtu faktorů a malé hodnoty zátěží na zbytku faktorů.

ROTACE FAKTORŮ

dvě základní rotační techniky:

- ☑ ortogonální rotace, která zachovává nezávislost faktorů;
- ☑ nepřímé (oblique) rotace – vytvářejí korelované faktory

ODHAD FAKTOROVÉHO SKÓRE

původní proměnné jsou popsány pomocí faktorů + chybový člen

- vícerozměrná regresní analýza

  předpokládejme, že $\mathbf{f} = \mathbf{A}^T.\mathbf{x}$; násobením zprava $\mathbf{x}^T$ a určením střední hodnoty a za předpokladu, že $E(\mathbf{x}.\mathbf{x}^T)=\mathbf{R}$ a $E(\mathbf{f}.\mathbf{x}^T)=\mathbf{\Lambda}^T$ dostáváme

$$\mathbf{\Lambda}^T = \mathbf{A}^T.\mathbf{R} \text{ nebo } \mathbf{A}^T = \mathbf{\Lambda}^T.\mathbf{R}^{-1}$$

  a z toho odhad skóre

$$\hat{\mathbf{f}} = \mathbf{\Lambda}^T\mathbf{R}^{-1}\mathbf{x}$$

KOLIK FAKTORŮ?

- v případě hlavní faktorové metody se vybírá tolik faktorů, kolik je vlastních čísel redukované korelační matice větších než jedna;
- v případě řešení pomocí metody maximální věrohodnosti – statistické postupy založené na potvrzení hypotézy, že variance všech obrazů může být vyjádřena pomocí daného počtu faktorů

FAKTOROVÁ ANALÝZA VS. PCA

FA

- neortogonální;
- skóre faktorů je třeba odhadnout;
- faktory nejsou jednoznačné;
- orientované na proměnné (faktory);
- model pro kovarianční nebo korelační matici;
- kovarianční struktura;
- neúčinná, pokud jsou proměnné nekorelované;
- odhad maximální věrohodnosti řeší problém měřítka;

PCA

- množina ortogonálních vektorů;
- skóre složek (souřadnic) lze snadno určit;
- souřadnice jsou jednoznačné;
- orientované na proměnné;
- není principiálně založeno na statistickém modelu;
- vysvětluje strukturu rozptylu (při použití disperzní matice);
- neúčinná, pokud jsou proměnné nekorelované;
- závislé na měřítku;

PŘÍKLAD

PŘÍČINY BOHATSTVÍ V ČR

1. schopnost či talent
2. štěstí
3. nepoctivost
4. pracovitost
5. dobré známosti a styky
6. narození (vstup do života)
7. hospodářský systém

MATICE ROTOVANÝCH FAKTORŮ

příčina	faktor1	faktor2
5	0.709	0.170
7	0.708	-0.212
3	0.706	-0.237
6	0.487	0.440
1	-0.160	0.706
4	-0.245	0.705
2	0.317	0.548

KLIDNÉ, SPOKOJENÉ VÁNOCE

BUJARÉ PŘIVÍTÁNÍ NOVÉHO ROKU

BEZPROBLÉMOVÉ ABSOLVOVÁNÍ ZKUŠEBNÍHO OBDOBÍ

Příprava nových učebních materiálů oboru Matematická biologie

je podporována projektem ESF

č. CZ.1.07/2.2.00/07.0318

„VÍCEOBOROVÁ INOVACE STUDIA MATEMATICKÉ BIOLOGIE"

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ