PSY117/454 Statistická analýza dat v psychologii – přednáška 12

Úvod do multivariačních technik

Shluková analýza 1

- Cílem shlukové analýzy je kategorizovat objekty
  - typicky respondenty (za objekty lze považovat po transpozici datové matice i proměnné)
  - kategorizujeme
- Objekty kategorizujeme podle jejich hodnot ve vstupních proměnných
  - např. kategorizujeme lidi podle *věku* a *pohlaví*
    - v takovém případě bychom měli získat 4 kategorie (shluky) – chlapce, muže, dívky a ženy
- Objekty jsou kategorizovány na základě **podobnosti**
  - existují různé **ukazatele podobnosti** (např. vzdálenost v *n*-rozměrném prostoru, kde *n* = počet vstupních proměnných)
  - maximem podobnosti je „stejnost" - identita

AJ: cluster analysis, similarity indices (measures),

Shluková analýza 2 – princip metody

- Analýza se skládá z neustálého opakování následujícího kroku:
- Najdi 2 nejpodobnější objekty a vytvoř z nich shluk
  - „vytvoř shluk" = utvoř z nich skupinu, kt. bude dále vystupovat jako pomyslný průměrný objekt vytvořený ze 2 původních objektů
  - po několika opakováních již budou shlukovány shluky vytvořené v předchozích krocích (do větších shluků)
- Postup končí, když jsou všechny případy v jednom velkém shluku
- Takto vzniká hierarchická struktura připomínající strom
  - na jedné straně *n* objektů, které se postupným slučováním nakonec slučují až do jednoho velkého shluku
  - grafickou podobou této struktury je **dendrogram**
- Na základě výsledků se rozhodujeme, které shluky jsou smysluplné a které ne
  - pro posouzení používáme běžná kategorizační pravidla, tj. chceme aby objekty ve shluku si byly velmi (kvalitativně) podobné a co nejvíce se odlišovaly od objektů v ostatních shlucích

Shluková analýza 3 – body rozhodování

1. Jaký ukazatel podobnosti objektů využít?
   - musí odpovídat úrovni měření dat
2. Jak definovat podobnost shluků (cluster method)
   - např. vzdálenost „průměrů“(středů) shluků (centroid), nebo vzdálenost nejbližších prvků…
   - často se používá tzv. Wardova metoda
3. Kolik shluků chceme?
   - jen málo formálních pravidel a i ta jsou přibližná
   - rozhodnutí je dáno „smysluplností“ shluků
4. Jaká data potřebujeme
   - na velikosti vzorku příliš nezáleží (záleží na účelu klasifikace)
   - počet vstupních proměnných držíme na minimu
   - vstupní proměnné by spolu ideálně neměly moc souviset (korelovat)

Shluková analýza 4 - použití

- obvykle k exploraci dat (*Do jakých skupin by se dali roztřídit?*)
- vede k empirické typologii
- obvykle se její výsledky dají jen obtížně zobecňovat (opatrnost!)
- jen zřídka poskytuje teoreticky využitelné výsledky

Mnohonásobná lineární regrese

- = vícerozměrná, multivariační… = tj. jedna závislá (predikovaná) a více nezávislých (prediktorů)
- cílem je stále statistická predikce
  - s více informacemi (proměnnými) dokážeme obvykle predikovat lépe
- prvky, kt. komplikují jednoduché rozšíření z jednoho na více prediktorů
  - korelace mezi prediktory (**multikolinearita**) – ve vyšší míře snižuje reliabilitu odhadů, v nižší pouze komplikuje interpretaci (srovnávání vlivu(predikční síly) prediktorů)
  - srovnávání síly prediktorů – k tomu vytvořen **standardizovaný regresní koeficient** β (vzniká standardizací reg. koeficientu b)
  - pořadí, v němž budou prediktory do regrese vstupovat
    - s jistým zjednodušením platí „kdo dřív přijde … ten má možnost vysvětlit více rozptylu“
    - lze to považovat za výhodu a vkládat prediktory do regrese postupně – **stepwise** (nebo postupně odebírat) a sledovat, jak se mění výsledky regrese (zejm. vysvětlený rozptyl)

Faktorová analýza (exploratorní) 1

- Účel:
  - Redukce většího množství proměnných na menší, snáze uchopitelnou matici
  - Je-li cílem např. zjednodušit výzkumný design, je možné pomocí FA identifikovat trsy korelujících proměnných a případně je sloučit do jediné škály.
- Předpoklady:
  - Proměnné měřené min. na intervalové škále
  - Rozložení proměnných symetrické, blízké normálnímu
  - Velikost vzorku N > 20 * počet položek
  - Věcně významné korelace v matici

Faktorová analýza (exploratorní) 2

- Proměnné, které spolu vysoce korelují, pravděpodobně měří totéž.
- Mohou tedy být nahrazeny jedinou proměnnou, která je jejich lineární kombinací – faktorem.
- Jde o analýzu korelační matice.
- Faktorový náboj (Fx, Fy) je korelace původní proměnné s daným faktorem.
- Komunalita $h^2 = FxP1^2 * FyP1^2$ je faktorový rozptyl položky, podíl rozptylu položky vyčerpaný daným faktorovým řešením.
- U přijatelného FA řešení neklesají h^2 pod 0,7.
- „Dobrá struktura" je požadavek na jasnost faktorové matice. Každá položka by měla vysoko (více než -+0,7) skorovat v právě jednom faktoru, každý faktor by měl obsahovat dva nebo více vysokých faktorových nábojů.

Korelační matice	P1	P2	P3	P4
P1	1	-,14	,74	,08
P2	-,14	1	,19	,59
P3	,74	,19	1	,17
P4	,08	,59	,17	1

Faktorová matice	F1	F2	h^2
P1	,73	-,60	,89
P2	,47	,77	,81
P3	,87	-,35	,87
P4	,58	,66	,77

Strukturální modelování
(Structural Equation Modelling, SEM)

- Strukturální modelování je souborné označení skupiny statistických procedur, užívaných k testování hypotéz o určité podobě kovariančních matic proměnných, tedy lineárních vztahů mezi proměnnými.
- Hypotézou pro test je v tomto případě určitý model, pro něž jsou výzkumníkem definovány hodnoty kovariancí. Testována je shoda tohoto modelu s daty.
- SEM se nejčastěji využívá pro dva účely: *konfirmatorní faktorové analýzy* a *(mnohorozměrné) lineární regrese.* Existují i jiné účely použití.

- *Konfirmatorní faktorová analýza* testuje hypotézu o určité faktorové struktuře. Tedy je dána množina manifestních (měřených) proměnných (položek), a míra jejich očekávané kovariance s latentními proměnnými (faktory).
  - „Manifestní proměnná"– měřená položka, proměnná v datech.
  - „Latentní proměnná" – konstrukt vyššího řádu, determinující manifestace.
  - Testována je hypotéza o faktorových nábojích MP v LP, tedy o determinaci skupiny manifestních proměnných určitým faktorem či faktory.`

- ***Lisrel*** *je počítačový program pro strukturální modelování, který se o zasloužil o značné rozšíření strukturního modelování v sociálních vědách. Někdy je jeho název používán i jako označení pro strukturální modelování jako metodu.*

Structural Equation Modelling (SEM)

- *(Mnohorozměrná) lineární regrese* v rámci strukturálního modelování umožňuje testování hypotéz o lineárních vztazích mezi latentními proměnnými.
- Je testována hypotéza o nějaké kovarianční matici exogenních (nezávisle) proměnných a endogenních (závisle) proměnných. Exogenní i endogenní proměnné jsou koncipovány jako latentní (viz FA).
- Testování hypotéz o maticích probíhá pomocí různých metod testu dobré shody, založených obvykle na kriteriu χ^2.
- Předpoklady dat:
  - Na každý parametr (korelaci, rozptyl proměnné) minimálně 5 N. (To odpovídá cca 20 N na manifestní proměnnou.)
  - Plná datová matice bez missing values.
  - *Existence hypotézy o vztazích v datech. SEM je vždy konfirmatorní metodou analýzy dat. Nelze užít exploratorně.*

Strukturální model - příklad