Exploratorní faktorová analýza (EFA) v JASP

K tématu existují tři videa, v nichž je problematika vysvětlená velmi podrobně. Z toho nejpodstatnějšího v nich vznikl následující text a na základě doporučení také PDF s ukázkou výpočtu provedeném na datech, které máte k dispozici v Moodle v této kapitole.

Petr Soukup seminář - 1 o EFA (základní vysvětlení principu a smyslu faktorování):

https://www.youtube.com/watch?v=OZNEEDuu3B0

Petr Soukup seminář - 2 o EFA (ukázka výpočtu s jedním faktorem na základě dat, která jsou k dispozici v databázi dat v JASP. Jde o původní data od Ch. Spearmana, na nichž extrahoval obecný faktor inteligence (g-factor) z dílčích otázek v IQ testu:

https://www.youtube.com/watch?v=D7f2fchTBRQ

Petr Soukup seminář - 3 o EFA (výpočet vícefaktorové analýzy v JASP):

https://www.youtube.com/watch?v=IUYrkOdSexE

Exploratorní faktorovou analýzu (dále již EFA) používáme, máme-li větší počet položek (otázek), které měří určitý jev z různého pohledu nebo v různých kontextech. Principem EFA je to, že zkoumáme, zda některé položky spolu navzájem souvisí a zda by se zachycením těchto souvislostí dala data zjednodušit tak, abychom nemuseli porovnávat různé podsoubory respondentů s každou otázkou z dotazníku (testu) zvlášť. Můžeme je pak porovnávat se shluky proměnných, které navzájem korelují, tj. měří podobný aspekt sledovaného jevu. Ptáme se tedy: „***Mají některé z položek dotazníku/testu něco společného?***“ - Ze 30-položkového dotazníku můžeme pomocí EFA získat např. 2-5 faktorů, které zahrnují položky s nejvyššími vzájemnými korelacemi.

- U získaných faktorů pak můžeme hledat rozdíly mezi podsoubory respondentů.
- EFA se často používá i tehdy, když nemůžeme něco měřit napřímo. Měříme tzv. *latentní proměnné*. Např. Spearmanův obecný faktor inteligence (tzv. g-faktor) vznikl právě na základě faktorové analýzy. Pomocí EFA byl také např. na základě jednotlivých otázek ověřen konstrukt extraverze a introverze.

Platí, že každá položka je sycená každým extrahovaným faktorem (protože téměř každá položka s každou koreluje, byť třeba jen minimálně). Podobně jako máme korelační koeficienty od -1 do +1, získáváme v EFA ***faktorové zátěže***, které se také pohybují ve stejném rozmezí. Faktorová zátěž (či náboj) u dané položky vyjadřuje těsnost vztahu k příslušnému faktoru. Čím vyšší je absolutní číslo této zátěže, tím silněji položka s faktorem souvisí. [1]

[1] Faktor (= nezávisle proměnná) je vždy SPOJITÁ/KARDINÁLNÍ PROMĚNNÁ. JASP však neumí ukládat u respondentů skóry pro každý faktor (tzv. faktorové skóry). **Můžeme si však v Excelu vytvořit průměrové škály položek, které jsou nejvíc sycené konkrétními faktory** (tj. které mají u daného faktoru nejvyšší náboje).

Teď něco nepříliš intuitivního: ***Faktor je nezávisle proměnná, položky jsou závisle proměnné***. Znamená to, že položky jsou vysvětlené něčím „schovaným“, abstraktním (podobně jako Spearman vysvětloval správně zodpovězené položky v IQ testu obecnou inteligencí, která je předchůdcem a příčinou zdařilého řešení těchto položek a lze ji tušit v pozadí).[2]

TYP PROMĚNNÝCH PRO EFA

Proměnné musí být v JASP nastavené jako kardinální/spojité proměnné (se znakem „oranžového pravítka“ u jejich názvu), jinak je EFA v JASP nevezme do výpočtu.

Ideálně vychází EFA z kardinálních/spojitých proměnných s normálním rozdělením. Pak by se pro odhad faktorových zátěží (=lambd) – ideální by zde byla pro faktorový výpočet „metoda maximální věrohodnosti“ (= maximal likelihood, ML). To je však ve společenských vědách vzácností. Většinou máme spíš ordinální/pořadové proměnné (někdy dokonce jen dvoustupňové, dichotomní). Tam se metoda maximální věrohodnosti použít nedá.

POSTUP PŘI EFA

Pro exploratorní analýzu musím mít minimálně 3 položky na jeden faktor. Počet faktorů tedy nesmí být vyšší, než je trojnásobný počet faktorovaných položek.

Technik, jak použít výsledky z EFA, je vícero. Většinou používáme ***řešení s rotací***, které vede k lépe interpretovatelným faktorům.

- Nejprve musíme **zkontrolovat data**, zda nic nechybí, zda tam nejsou nereálné proměnné (v Descriptives kontrolujeme průměry či mediány a minima + maxima)
- Poté provedeme kontrolu, **zda jsou naše proměnné nějak vzájemně provázané** – prostřednictvím ***korelačních koeficientů***. Pokud by mezi nimi nebyly příliš vysoké korelace, nemá EFA smysl.
- Dále provedeme výpočetní texty, které ověřují ***vztahů*** mezi všemi faktorovanými proměnnými komplexně. Je to trochu podobné jako u korelací, ale jde o ***globální*** výpočet provázaností mezi proměnnými – abychom nemuseli prozkoumávat desítky dílčích korelací (jde o ***KMO*** a ***Bartlettův test*** sféricity) v JASP.
- Musíme **určit, v jaké podobě budeme mít vstupní data** – může jít o původní data, nebo o korelační matici. Pro kardinální data můžeme vycházet z Pearsonovy korelační matice, ale ne u krátkých ordinálních stupnic (3-4 stupně), ani u dichotomických stupnic. O pětistupňových Likertových škál už ale můžeme vycházet z Pearsonovy korelační matice.

[2] Nepodstatná teorie: Rovnice pro EFA jsou násobky faktorových zátěží s hodnotou konkrétního faktoru, jejich součet a přičtení chyby. Položka1 = Lambda1 * F1 + Lambda2 * F2 + … + e.
U ***Principal component analysis (PCA)*** by to bylo naopak. Tam jsou položky nezávisle proměnnými a naopak faktory jsou proměnnou závislou. Zmizela by i chybová komponenta, protože u nezávisle proměnných není žádná chybovost.
Zatímco EFA se snaží vysvětlit souvislosti (korelace) mezi proměnnými, cílem PCA je vysvětlit variabilitu (rozptyly), tj. šíři informací proměnných. Obojí se však při interpretaci částečně překrývá.

- **Určení odhadovací techniky** pro určení počtu faktorů
- **Určení počtu faktorů**
- **Výpočet nerotovaného řešení**, na jeho základě ***vyřadit položky***, které se nepozdávají (podle určitých indikátorů), protože mají slabé faktorové náboje nebo vysoké u více faktorů.
- **Zpravidla následně provedeme stejné řešení rotované šikmé** (metoda oblimin) a kontrolujeme, jak spolu faktory korelují.
- Pokud je tato korelace vysoká, zůstáváme u tohoto ***šikmého řešení***. Je-li relativně nízká, volíme pro rotaci ***kolmé řešení*** (metoda varimax).
- Sledujeme, které položky jsou nejsilněji syceny jednotlivými faktory. Podle jejich názvu hledáme zobecňující **názvy pro každý z těchto faktorů**.
- následně provedeme pomocí reliability **výpočet vnitřní konsistence** pro shluky položek, které mají nevyšší náboje u určitého faktoru. (výpočet reliability provedeme u položek s nejvyššími náboji pro každý z extrahovaných faktorů. Každý faktor by měl mít své vlastní položky s nejvyššími náboji. Ty položky, které mají slabé náboje u všech faktorů, vynechváme.) Cronbachova alfa by měla být u každého z těchto shluků alespoň 0,7 nebo vyšší. Zároveň by žádná položka neměla tuto výši výrazně snižovat (viz funkce „if item is deleted“).
- Poté z takto zkontrolovaných shluků položek, které mají vysoké náboje u daného faktoru, **vypočítáme v Excelu průměrovou škálu** (pro každý faktor jednu).
- S těmito průměrovými škálami můžeme dál počítat jako s novými spojitými/kardinálními proměnnými. Můžeme např. sledovat rozdíly v nich u mužů a žen (třeba pomocí t-testu), nebo je korelovat s dalšími (třeba s věkem respondentů) apod.

TROCHA TEORIE - VSTUPNÍ DATA (NEPOVINNÉ)

Pro kardinální data a data s alespoň pětistupňovou stupnicí lze použít Pearsonovu korelační matici. Jsou-li data normálně rozdělená, lze využít metodu maximální věrohodnosti (maximum likelihood). **Nejsou-li úplně normálně rozdělené** (anebo mám **kratší ordinální stupnici**), používám PCA či PA (PCA=PA), resp. v JASP jde o PA, tj. **Principal component analysis** – metoda hlavních komponent. Ta je sice zaměřená na zjišťování rozptylu, ale dle Soukupa se dá použít, i když někteří statistici jsou striktně proti míchání těchto metod. V JASP ji najdeme v samostatné záložce „**principal component analysis**“ a Exploratory Factor analysis má v JASP samostatné zadání.

Při kratších ordinálních proměnných (do 4 stupňů včetně) se nedoporučuje Pearsonova korelační matice, protože bychom zbytečně dostali horší řešení. Používá se ***polychorický***

koeficient a u dichotomických proměnných ***tetrachorický koeficient***. Asi to ale neumí JASP, spíš software „R“. U pětistupňových škál už lze použít Pearsonovu korelační matici.[3]

Metoda maximální věrohodnosti (maximum likelihood) hledá odhad maximalizovat, ale v sociálních vědách se skoro nedá použít.

Metoda hlavních komponent (PCA) – maximalizace vysvětleného rozptylu proměnných. Vhodná i pro ordinální stupnice. Nesměřuje na vysvětlení pro vztahy. Vychází se z korelační matice (diagonála sama se sebou).

TEĎ UŽ POVINNÉ INFO:

Většina autorů preferuje Principal Axis (=**metoda hlavních os**). Tu používáme, pokud nemáme dlouhé kardinální proměnné s normálním rozdělením, které by byly počítatelné metodou maximum likelihood.[4]

- **Metoda hlavních os** je v nabídce „Exploratory factor analysis“.[5]

VOLBA POČTU FAKTORŮ

Existuje několik metod, které nám pomáhají určit, kolik faktorů máme při extrakci zvolit:

(1) Kaiserovo pravidlo – beru počet faktorů, které mají ***hodnotu vlastních čísel*** (*eigenvalues*) korelační matice nad 0. Doporučuje se brát aspoň tak silné faktory, jaké jsou výchozí proměnné.

(2) Sutinový test (*scree plot*) – Jde o metodu R. Cattella: hledáme zlom v grafu a vezmeme o 1 faktor méně, než je zlom (tj. počet faktorů nad zlomem).

(3) Vysvětlený rozptyl – sledujeme vysvětlené procento rozptylu. Jak velké by mělo být? Jaká je minimální doporučená mez? Doporučuje se 50-80 % (PJ: ale často bývá tato hodnota jen kolem 40 %).

(4) Alternativně lze použít **paralelní analýzu** – z čísel nad 1 u sutinového grafu a z nasimulovaných náhodných dat (95. kvantil) – z nich se nakreslí jejich jednotková čísla (v JASP se zadá volbou varianty „Paralel analysis“). Tam, kde se obě čáry protnou, tak nad nimi by příslušný počet faktorů měl být přijatý.

---

[33] Pro určité typy dichotomních proměnných typu „ano-ne“ (např. nemoc-zdraví, žije-zemřel apod.) se exploratorní faktorová analýza nehodí, protože nevycházejí z kontinua, které by bylo rozstříhané na ordinální kategorie. Lepší je zde ***analýza latentních tříd***. Jinak se ale v dotaznících předpokládá, že respondent většinou necítí úplně stoprocentní „ano“ nebo „ne“, takže tam to kontinuum (tj. spojitou proměnnou) předpokládáme a pak můžeme faktorovat.

[4] U metody hlavních os se v diagonálách (úhlopříčkách) v korelační matici místo jedniček (=korelace položek se sebou samými) dosazují komunality, tj. míry vysvětleného rozptylu faktorovou analýzou. Tím dostaneme realističtější řešení.

[5] Dále zde najdeme metody založené na principu nejmenších čtverců (podobně jako lineární regrese, viz za týden): „Ordinary least squares“, „Weighted least squares“ (větší důraz na proměnné s vysokými korelacemi s mnoha proměnnými), „generalized least squares“ apod.

Petr Soukup otevírá z nabídky v JASP (open => data library => 6.factor) originální původní data od Spearmana, z nichž určoval latentní proměnnou (obecný faktor inteligence, g-factor), která sytí dílčí projevy inteligence v dílčích úlohách v IQ testech. Budeme zde počítat z dat očištěných od rušivých vlivů daných různým věkem dětských respondentů (s názvem začínajícím „Residuals"). V názvech proměnných musíme mít pro Pearsonův korelační koeficient hodnoty označené jako kardinální (=obrázek „oranžového pravítka" u názvu proměnné).

V nabídce pod „reggresion" volíme „correlate" a díváme se, jak vzájemně koreluje 7 sledovaných proměnných. Přesuneme je do okénka „variables". Největší korelace se prokázaly mezi jazyky a matematikou (kolem 0,7-0,8). Je tam shluk korelujících proměnných, které naznačují minimálně jeden shluk (faktor) společně korelujících proměnných.

PROCEDURA FAKTOROVÁNÍ

Proměnné změníme na „kardinální proměnné" (oranžové měřítko). Poté fakturujeme: „Factor" => „Exporatory factor analysis", do kolonky „Variables" přenášíme proměnné, které chceme faktorovat.

JASP nabízí možnosti – volba počtu faktorů

- pomocí tzv. paralelní analýzy (viz výše)
- pomocí Kaiserova pravidla („Eigenvalues")
- „manual" – umožňuje, abychom rovnou napsali, kolik těch faktorů chceme.

V Output Options níže nejprve řešíme, zda jsou proměnné rozumně provázané jako celek. Potřebujeme výpočty: „*KMO*" a „*Bartlett*". Získáváme ***Kaiser-Meyer-Oiken celkové kritérium*** **(Overall MSA)** a pak tyto hodnoty pro jednotlivé proměnné.

- Doporučení autorů těchto testů je, aby hodnota celkového výpočtu KMO i u jednotlivých položek v KMO byla ***vyšší než 0,5***. Pokud je to nižší hodnota, nemůžeme provádět EFA, nebo můžeme vynechat proměnné, které to nesplňují. Proměnné s nižší hodnotnou s ostatními nekorelují, tak je můžeme vynechat.
- Zároveň by měl být ***Bartlettův test statisticky významný*** (tj. $p < 0,05$).

Nahlížíme také do tabulky Loading Factor: jde o co největší hodnotu ve sloupci „Factor 1" a nejnižší hodnotu v „***Uniqueness***" (čím jedinečnější je položka, tím méně souvisí s ostatními). Na základě výše uvedených kritérií se můžeme zbavit dvou proměnných (ve Spearmanových datech je to položka: „Light" a „Weight"). Pokud je procento v Uniqueness nad 0,9, je to položka k vyřazení.

Podobně také, je-li faktorový náboj pod 0,3, doporučuje se s položkou nepracovat.

Po vynechání nevhodných položek se znovu přepočítává vše. Zvýší se KMO.

Z hlediska počtu faktorů si zde můžeme říct dopředu i o „Scree plot". Kaiserovo pravidlo hodnot (Eigenvalues) nad 0 by podle Scree plotu doporučovalo všechny hodnoty nad nulu.

Paralelní analýza - první zkřížení čar = doporučuje 1 faktor. Cattelovo pravidlo (o 1 faktor méně, než je zlom) = také doporučuje 1 faktor. (Proto se méně preferuje kritérium Eigenvalue, ale naopak **se velmi preferuje Paralelní analýza a Cattelův sutinový test**.)

Pro úpravu dat ve faktorové matici v „**Output options**“ si na posuvném měřítku můžu nastavit, jak nízké faktorové náboje už nebudou ve faktorové matici znázorněny. Tím se stává faktorová matice přehlednější a snadnější pro porozumění výsledků a jejich interpretaci.

Můžeme si říct o ***Path diagram***, který ukazuje faktorové zátěže v grafické podobě s barevným znázorněním vztahů vůči faktoru. – To je skvělé pro popis a interpretaci.

Pokud nám vyjde faktor, který sytí jen 2 položky, tak nemá smysl. V takovém případě je vhodné odstranit jednu z nich, nebo obě. – Je zřejmé, že faktorování je hodně na výzkumníkovi, který se v různých situacích musí nějak rozhodnout. Správných řešení může být víc, vždy jde i o úhel pohledu. (PJ: Roli také hraje smysluplnost a interpretovatelnost faktorů, která může být důležitější než statisticky perfektně správné, ale nesrozumitelné řešení.)

TYPY ROTACÍ: ROTACE KOLMÁ/ORTOGONÁLNÍ (VARIMAX) ŠIKMÁ (OBLIMIN)

Velká část jevů, které zkoumáme v EFA, je nějak provázaných. Pak je nepříliš moudré volit pravoúhlou rotaci, která vnucuje faktorům vzájemnou nezávislost. Tím se data deformují. Vypadá to pak, že faktory jsou navzájem nezávislé a že lidé hodnotí realitu v polohách „buď-anebo“. Ale oni mohli hodnotit jen něco o trochu silněji než něco jiného (pak to znamená, že se ty faktory vzájemně nevylučují, ale jen shlukují vzájemně si bližší položky oproti těm méně podobným).

Rotaci provádíme vždy, pokud původní (nerotované) řešení nedává dobrý smysl. U PCA a u hlavních os (Principal Axis) vlastně vždy. Avšak při jednom faktoru není třeba nic rotovat.

Vždy začínáme s šikmou rotací (oblimin). To je ta obecnější verze. Teprve když zjistíme, že faktory nejsou příliš provázané, pak teprve volíme kolmou rotaci. To zjistíme tak, že při šikmé rotaci („**Oblique rotation**“) si řekneme v nabídce „Option output“ o korelaci mezi faktory „factor correlations“. Velikost korelace nám napoví, jak moc spolu faktory souvisí. Je ta korelace mezi faktory šikmá velká, nebo malá: Když je r=0,6, tak je neudržitelná představa toho, že by tyto faktory mohly být nezávislé. Když si však nejsme jistí, jestli máme zvolit šikmou, nebo kolmou rotaci, vypočítáme obě. – Jsou-li obě řešení podobná, můžeme pracovat s kolmou. Pokud se liší, volíme rotaci šikmou.

Volíme-li kolmou rotaci, tak volíme Varimax – poskytuje dobré řešení. Jde o metodu, v níž se hledá řešení, ve kterém mají položky v jednom z faktorů co nejvyšší náboj a v těch ostatních co nejnižší (tzv. čisté řešení).

Z path diagram vidíme provazby mezi položkami a faktory i mezi faktory. To nám pomáhá výsledky interpretovat.