Bayesovská exploratorní faktorová analýza: srovnání s klasickým přístupem

JAKUB MAZANEC[1], EVA DVOŘÁKOVÁ[1]

Abstrakt: *V článku se věnujeme srovnání klasické a bayesovské exploratorní faktorové analýzy. Použili jsme náhodně generovaná testovací data s různými velikostmi vzorku, strukturou faktorových nábojů, škálami odpovědí a velikostmi náhodné chyby. Provedli jsme oba druhy exploratorní faktorové analýzy a na základě srovnání výsledků a celkové zkušenosti z provedených analýz můžeme uzavřít, že u klasické a bayesovské exploratorní faktorové analýzy není rozdíl v přesnosti odhadu faktorových nábojů. Bayesovský přístup poskytuje více informací, které jsou intuitivněji interpretovatelné, ale na druhou stranu námi použitý model má i nevýhody, jako je nutnost ruční úpravy zdrojových kódů a časová náročnost výpočtů. Bayesovský přístup je tak podle nás vhodnější spíše pro konfirmatorní faktorovou analýzu.*

Klíčová slova: *exploratorní faktorová analýza; bayesovská statistika*

Faktorová analýza je v psychometrice oblíbenou statistickou technikou, používanou na odhalení struktury dat; díky faktorové analýze je možné ověřit, do jaké míry konstrukty reprezentují původní proměnné (Henson & Roberts, 2006). V dnešní době už její použití není omezeno technickými možnostmi, což je také důvod, proč se používá rutinně a proč se dočkala řady metodologických inovací (Lopes & West, 2004). Má však i své nevýhody, které vesměs vycházejí z toho, že vyžaduje od výzkumníka řadu arbitrárních rozhodnutí.

Klasická exploratorní faktorová analýza se skládá ze čtyř kroků. Jedná se o 1) výběr dimenzí faktorového modelu, poté o 2) alokování naměřených hodnot faktorům, 3) odhad faktorových nábojů a 4) vyřazení položek, které sytí více faktorů (Yong & Pearce, 2013, Conti et al., 2014). Všechny tyto kroky jsou předmětem ad hoc úsudků: jaký zvolit počet faktorů v kroku 1, jaké zvolit metody extrakce a rotace faktorů v kroku 3, či jak interpretovat získané odhady v kroku 4? Na rozdíl od klasické, frekventistické faktorové analýzy se ta bayesovská dokáže s jistou arbitrárností v jednotlivých krocích vypořádat tím, že ji zahrne do samotného modelu (např. Conti et al., 2014); navíc má řadu dalších výhod bayesovského přístupu (dle Kruschke, 2014):

1) Lze získat kompletní a intuitivně interpretovatelné informace o pravděpodobnostním rozdělení hodnot všech parametrů.

[1] Katedra psychologie, Fakulta sociálních studií MU, Joštova 10, 602 00, Brno

Došlo do redakce: 27. 1. 2017
Schváleno k publikaci: 7. 5. 2017

2) Snadno se zachází s chybějícími hodnotami, např. pokud testovaná osoba vynechá některé položky; inference je možná i bez použití *listwise* či *pairwise* vynechávání záznamů.
3) Je snadné se odchýlit od obvyklých předpokladů modelu, např. použití normálního rozdělení, či nezávislosti subjektů, a místo toho zvolit třeba zešikmené rozdělení, anebo přidávat do modelu různé další závislosti.
4) Lze snadno vytvářet hierarchické modely, např. korelovat faktorové skóry napříč různými geografickými jednotkami či v čase.

Stojí ovšem použití bayesovské faktorové analýzy za vynaložené náklady, jako je nutnost učit se ovládat komplikovanější software a vypořádat se s výpočetní náročností *Markov chain Monte Carlo* (dále MCMC) samplovacích metod? Protože v psychometrice se faktorová analýza používá při rozhodování, jaké položky zařadit do testů, rozhodli jsme se zaměřit při srovnávání klasického a bayesovského přístupu především na přesnost odhadu faktorových nábojů. S použitím simulovaných dat jsme zjišťovali, jak se oba přístupy liší ve schopnosti odhalit simulovanou faktorovou strukturu, jak se odhady faktorových nábojů liší od jejich *true* hodnot, zdali 95% intervaly spolehlivosti a věrohodnosti zahrnují *true* hodnoty, a kolik to všechno zabere času.

1 BAYESOVSKÝ MODEL

Bayesovský model se v základu neliší od toho klasického. Mějme množinu J měření na intervalové škále uspořádaných do vektoru $\boldsymbol{y}_{(i)} = \left(y_{(i,1)}, \dots, y_{(i,J)}\right)^\top$ pro osobu $i, i = 1, \dots I$; a matici $\boldsymbol{Y} = \left(\boldsymbol{y}_{(1)}, \dots, \boldsymbol{y}_{(I)}\right)^\top$ pro celý vzorek. Faktorový model vztahuje vektor pozorovaných měření $\boldsymbol{y}_{(i)}$ k množině K latentních faktorů:

$$\boldsymbol{y}_{(i)} = \boldsymbol{\Lambda}\boldsymbol{\theta}_{(i)} + \boldsymbol{\varepsilon}_{(i)}$$

$\boldsymbol{\Lambda}$ je $J \times K$ matice faktorových nábojů; $\boldsymbol{\theta}_{(i)}$ je $K \times 1$ vektor faktorových skórů pro osobu i; jedinečnosti jsou $\boldsymbol{\varepsilon}_{(i)} \sim \mathcal{N}_J(0, \boldsymbol{U})$, kde $\boldsymbol{U}$ je diagonální matice rozptylů jedinečností. Určíme $\boldsymbol{\theta}_{(i)} \sim \mathcal{N}_K(0, \boldsymbol{I}_K)$, čímž identifikujeme jednotku faktorových nábojů. Marginalizací přes faktory dostaneme $\mathrm{Cov}\left(\mathbf{y}_{(i)}\right) = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{U} = \boldsymbol{Q} + \boldsymbol{U}$, s komunalitou $\boldsymbol{Q}$ a jedinečností $\boldsymbol{U}$ (Lee, 2007).

Limitací tohoto modelu je, že nemá jedno řešení. Ortogonální rotace faktorových nábojů a skórů vedou ke stejným hodnotám $\boldsymbol{Y}$: pro jakoukoliv $K \times K$ ortogonální rotační matici $\boldsymbol{P}^\top$, pokud $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\boldsymbol{P}^\top$ a $\boldsymbol{\theta}^* = \boldsymbol{P}\boldsymbol{\theta}$, pak $\boldsymbol{\Lambda}^*\boldsymbol{\theta}^* = \boldsymbol{\Lambda}\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{\theta} = \boldsymbol{\Lambda}\boldsymbol{\theta}$. *Likelihood* tedy neidentifikuje faktorové náboje, identifikována je pouze matice komunalit $\boldsymbol{Q}$: pro každé $J \times K$ matice faktorových nábojů $\boldsymbol{\Lambda}$ a $\boldsymbol{\Lambda}^*$, které mají plnou sloupcovou hodnost, kde $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\boldsymbol{P}^\top$ pro nějakou ortogonální rotační matici, $\boldsymbol{Q}^* = \boldsymbol{\Lambda}^*\boldsymbol{\Lambda}^{*\top}$ a zároveň se rovná $\boldsymbol{Q} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$ (Lockwood, Savitsky & McCaffrey, 2015).

Ve frekventistickém pojetí je absence identifikace modelu vyřešena arbitrárním výběrem nějaké $\boldsymbol{\Lambda}$ tak, aby odpovídala *maximal likelihood* odhadu komunalit: $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top = \widehat{\boldsymbol{Q}}_{\mathrm{MLE}}$ a pak rotováním $\boldsymbol{\Lambda}$ tak, aby splnila určitá kritéria pro interpretovatelnost: cílem je mít „jednoduchou strukturu“, např. aby každá položka byla co nejvíce sycena právě jedním faktorem, a zbylými co nejméně[2]. To splňuje varimax rotace (Kaiser, 1958); faktorové náboje, které poskytuje ovšem nejsou unikátní – existuje $2^K K!$ matic faktorových nábojů,

[2] Strukturu je ale samozřejmě definovat i pomocí jiných kritérií.

které splňují varimax kritérium a liší se pouze pořadím a znaménky sloupců. To již ovšem tolik nevadí – výzkumník si může v získané matici prohodit sloupce a změnit znaménka tak, aby se mu výsledky dobře interpretovaly.

V bayesovském pojetí je neexistence unikátního řešení problém, protože posteriorní rozdělení je díky tomu multimodální a obtížně se z něj sampluje. Do modelu lze přidat restrikce, které umožní jeho identifikaci; obvykle se použije takové, že horní trojúhelníková část matice faktorových nábojů je omezena na 0, zatímco diagonální prvky matice jsou omezeny na kladné hodnoty (dle Geweke & Zhou, 1996):

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{(1,1)} & 0 & \cdots & 0 & 0 \\ \lambda_{(2,1)} & \lambda_{(2,2)} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \lambda_{(J-1,1)} & \lambda_{(J-1,2)} & \cdots & \lambda_{(J-1,K-1)} & 0 \\ \lambda_{(J,1)} & \lambda_{(J,2)} & \cdots & \lambda_{(J,K-1)} & \lambda_{(J,K)} \end{bmatrix}$$

kde $\lambda_{(k,k)} \geq 0, k = 1, \dots, K$.

To ovšem do exploratorní analýzy vnáší další prvek arbitrárnosti – restrikce faktorových nábojů jsou zvoleny čistě kvůli identifikaci modelu, a nikoliv pro opodstatněnou intepretaci; řešení je díky tomu závislé na pořadí položek (Lopes & West, 2004).

Lockwood, Savitsky & McCaffrey (2015) přišli s alternativním způsobem, jak tento problém vyřešit. Jejich postup se sestává ze tří kroků: samplování z modelu, který neomezuje faktorové náboje, rotace takto získaných nábojů s pomocí varimax kritéria, a reorientace rotovaných nábojů tak, aby odpovídaly vždy pouze jednomu z $2^K K!$ neunikátních řešení. Následující popis naší analýzy je tedy adaptací tohoto postupu.

Prvním krokem je bayesovský odhad s pomocí výše popsaného modelu s použitím standardizovaných dat – řádky matice dat $\boldsymbol{Y}$, které představují odpovědi osob na jednotlivé položky, jsou centrovány (tj. od jednotlivých hodnot je odečten průměr řádku) a škálovány (tj. centrované hodnoty jsou vyděleny směrodatnou odchylkou řádku).

Využívá se toho, že komunalita, na rozdíl od faktorových nábojů, identifikovaná je: každé $\boldsymbol{Q}$ definuje nekonečnou množinu matic $\boldsymbol{\Lambda}$ tak, že $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top = \boldsymbol{Q}$. Takže pokud je pro komunality určeno vhodné priorní rozdělení, lze pak vytvořit pravidlo pro výběr faktorových nábojů asociovaných s danou komunalitou. Toto priorní rozdělení musí být invariantní vůči permutacím položek. Toho je dosaženo tím, že 1) neklademe žádná výše popsaná omezení na faktorové náboje, a 2) pro faktorové náboje zvolíme priorní rozdělení, která jsou vzájemně zaměnitelná, tzn. po permutaci řádků matice $\boldsymbol{\Lambda}$ bude priorní rozdělení pro $\boldsymbol{Q} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$ stejné. Spolu s tím, že zvolíme zaměnitelná priorní rozdělení i pro jedinečnosti, dosáhneme priorního rozdělení invariantního vůči permutaci i pro $\mathrm{Cov}(\mathbf{y}_{(i)}) = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{U} = \boldsymbol{Q} + \boldsymbol{U}$.

My používáme následující priorní rozdělení:

$$\lambda_{(j,k)} \sim \mathcal{C}(0,1)$$

$$u_{(j,j)} \sim \mathcal{HC}(\mu, \gamma)$$

$$\mu \sim \mathcal{HC}(0,1)$$

$$\gamma \sim \mathcal{HC}(0,1)$$

Tedy pro faktorové náboje je použito Cauchyho rozdělení, a pro diagonální prvky matice rozptylů jedinečností je použito poloviční Cauchyho rozdělení s hyperparametry.

Protože nejsou identifikovány faktorové náboje, ale pouze komunalita, jednotlivé MCMC vzorky faktorových nábojů proto nemusí splňovat varimax kritérium. Druhým krokem je proto varimax rotace všech vzorků. Pro kandidátskou matici $\boldsymbol{\Lambda}$, varimax vede k matici $\boldsymbol{\Lambda}_{\mathrm{V}} = \boldsymbol{\Lambda}\mathrm{R}_{\mathrm{V}}(\boldsymbol{\Lambda})$, kde

$$\mathrm{R}_{\mathrm{V}}(\boldsymbol{\Lambda}) = \arg\max_{R} \sum_{k=1}^{K} \left(\frac{1}{J} \sum_{j=1}^{J} (\boldsymbol{\Lambda}\boldsymbol{R})_{(j,k)}^{4} - \left(\frac{1}{J} \sum_{j=1}^{J} (\boldsymbol{\Lambda}\boldsymbol{R})_{(j,k)}^{2} \right)^{2} \right)$$

a $(\boldsymbol{\Lambda}\boldsymbol{R})_{(j,k)}$ označuje j a k element matice $\boldsymbol{\Lambda}\boldsymbol{R}$. Finální varimaxové faktorové náboje $\boldsymbol{\Lambda}_{\mathrm{V}}$ jsou specifické k matici komunalit $\boldsymbol{Q}$ v tom, že pokud $\boldsymbol{\Lambda}$ a $\boldsymbol{\Lambda}^{*}$ splňují $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\top} = \boldsymbol{\Lambda}^{*}\boldsymbol{\Lambda}^{*\top} = \boldsymbol{Q}$ pak $\boldsymbol{\Lambda}\mathrm{R}_{\mathrm{V}}(\boldsymbol{\Lambda}) = \boldsymbol{\Lambda}^{*}\mathrm{R}_{\mathrm{V}}(\boldsymbol{\Lambda}^{*})$ až do třídy ekvivalence $2^{K}K!$ matic, které se liší 2^{K} změnami znamének u sloupce a $K!$ permutacemi sloupců. Jinými slovy, pro každou matici $\boldsymbol{Q}$ existuje $2^{K}K!$ matic faktorových nábojů, které splňují varimax kritérium a liší se pouze pořadím a znaménky sloupců.

Pro každé $\boldsymbol{\Lambda}_{b}$, kde $b = 1, \ldots, B$, kde B je celkový počet MCMC vzorků, mějme rotační matici $\mathrm{R}_{\mathrm{V}}(\boldsymbol{\Lambda}_{\boldsymbol{b}})$, takže $\boldsymbol{\Lambda}_{\mathrm{V}b} := \boldsymbol{\Lambda}_{b}\mathrm{R}_{\mathrm{V}}(\boldsymbol{\Lambda}_{b})$ splňuje varimax kritérium. Každá matice faktorových nábojů $\boldsymbol{\Lambda}_{\mathrm{V}b}$ je jednou z oněch $2^{K}K!$ orientací (tj. prohození znamének sloupců a permutací sloupců); napříč vzorky se navíc mohou orientace lišit. Rotací jednotlivých samplů faktorových nábojů jsme nicméně redukovali výběr z nekonečna možných matic faktorových nábojů odpovídajících Q.

Třetím krokem je reorientace všech $\boldsymbol{\Lambda}_{\mathrm{V}b}$ vzorků na společnou orientaci. Když vezmeme dva vzorky $\boldsymbol{\Lambda}_{\mathrm{V}b}$ a $\boldsymbol{\Lambda}_{\mathrm{V}b}$,kde $b \neq b'$, tak matice $\boldsymbol{\Lambda}_{\mathrm{V}b}$, je buď orientována stejně jako $\boldsymbol{\Lambda}_{\mathrm{V}b}$, nebo má jednu ze zbývajících $2^{K}K! - 1$ orientací. Reorientace $\boldsymbol{\Lambda}_{\mathrm{V}b}$, ji tudíž učiní podobnější $\boldsymbol{\Lambda}_{\mathrm{V}b}$.

Pro každou matici $\boldsymbol{\Lambda}_{\mathrm{V}b}$ vytvoříme všech $L = 2^{K}K!$ možných reorientací tak, že u $\boldsymbol{\Lambda}_{\mathrm{V}b}$ permutujeme pořadí sloupců a vynásobíme je buď 1, nebo −1. Z těchto L matic vybereme tu, která má nejmenší eukleidovskou vzdálenost d od referenční matice $\boldsymbol{\Lambda}_{\mathrm{V}}^{*}$:

$$d = \sqrt{\sum_{j=1}^{J} \sum_{k=1}^{K} \left(\lambda_{\mathrm{V}(j,k)}^{*} - \lambda_{\mathrm{V}bl(j,k)} \right)^{2}}$$

Referenční matice $\boldsymbol{\Lambda}_{\mathrm{V}}^{*}$ je nejdříve vybrána náhodně ze všech $\boldsymbol{\Lambda}_{\mathrm{V}}$. Jakmile je provedena reorientace všech vzorků, je vypočítána nová $\boldsymbol{\Lambda}_{\mathrm{V}}^{*}$ jako průměr ze všech reorientovaných $\boldsymbol{\Lambda}_{\mathrm{V}}$. Reorientace s pomocí eukleidovské vzdálenosti je pak opakována, až do konvergence. Reorientované matice faktorových nábojů označujeme $\boldsymbol{\Lambda}_{\mathrm{F}}$; z množiny vzorků $\{\boldsymbol{\Lambda}_{\mathrm{F}b}\}$ lze pak spočítat průměry a intervaly věrohodnosti pro jednotlivé faktorové náboje. Vzorky lze rovněž ještě libovolně reorientovat, aby bylo dosaženo žádoucí intepretace[3].

[3] My tak činíme, abychom mohli $\boldsymbol{\Lambda}_{\mathrm{F}}$ porovnat s *true* hodnotami.

2 TESTOVACÍ DATA

Výše popsaný model bayesovské exploratorní faktorové analýzy (dále BEFA), implementovaný pomocí programu Stan (Stan Development Team, 2016) v *R* (R Core Team, 2016) jsme porovnávali s klasickou faktorovou analýzou (dále CEFA) s metodou extrakce *maximum likelihood* a varimax rotací, počítanou pomocí balíčku *psych* (Revelle, 2016). Zdrojové kódy, stejně jako kompletní výsledky, jsou k dispozici v git repozitáři https://github.com/jakubmazanec/bayesian-efa.

Pro porovnání obou metod jsme vytvořili několik sad testovacích dat, náhodně generovaných pomocí následujícího modelu:

$$y_{(i,j)} = \left(\sum_{k=1}^{K} \lambda_{(j,k)}\theta_{(i,k)}\right) + \varepsilon_{(i,j)}$$

$$\theta_{(i,j)} \sim \mathcal{N}(0,1)$$

$$\varepsilon_{(i,j)} \sim \mathcal{N}\left(0, \sigma_{\varepsilon_{(j)}}\right)$$

$$\sigma_{\varepsilon_{(j)}} \sim \text{Unif}(0, e_{\max})$$

Počet faktorů *K* byl vždy roven 2, počet položek *J* byl vždy roven 20. Velikost vzorku *I* byla 100, 250, nebo 500 osob (tyto varianty dále označujeme jako *small*, *medium*, *large*). Hodnota $e_{\max}$ nabývala hodnoty 0,5, nebo 1 (neoznačená varianta, respektive varianta *noisy*). Matice faktorových nábojů $\boldsymbol{\Lambda}$ měla tři možné varianty, viz tabulku 1. Vygenerované hodnoty odpovědí $\boldsymbol{Y}$ buď byly nezměněny (varianta *int*), nebo byly transformovány (varianta *ord*) následující funkcí[4]:

$$f(y) = \begin{cases} 1, y \leq -2{,}5 \\ 2, y > -2{,}5 \wedge y \leq -1{,}5 \\ 3, y > -1{,}5 \wedge y \leq -0{,}5 \\ 4, y > -0{,}5 \wedge y \leq 0{,}5 \\ 5, y > 0{,}5 \wedge y \leq 1{,}5 \\ 6, y > 1{,}5 \wedge y \leq 2{,}5 \\ 7, y > 2{,}5 \end{cases}$$

Tabulka 1 – Hodnoty $\boldsymbol{\Lambda}$ pro generování testovacích datových sad

	Varianta 1		Varianta 2		Varianta 3	
i	$k = 1$	$k = 2$	$k = 1$	$k = 2$	$k = 1$	$k = 2$
1	0,90	0,00	0,60	0,00	0,60	0,00
2	0,90	0,00	0,60	0,30	−0,60	0,60

[4] Tímto simulujeme pořadovou proměnnou s přibližně normálním rozdělením; častým předpokladem (či přáním) při tvorbě psychologických testů a dotazníků je, že se takto projevují např. několikabodové položky s póly „zcela nesouhlasím“ na jedné a „zcela souhlasím“ na druhé straně.

3	0,90	0,00	0,55	0,25	0,60	−0,10
4	0,90	0,00	0,50	0,15	−0,60	0,10
5	0,80	0,10	0,45	0,20	0,00	0,50
6	0,80	0,10	0,45	0,15	0,50	−0,50
7	0,80	0,10	0,40	0,05	−0,10	0,50
8	0,70	0,20	0,40	0,05	0,10	−0,50
9	0,70	0,20	0,40	0,05	0,40	0,00
10	0,60	0,30	0,30	0,10	−0,40	0,40
11	0,30	0,60	0,30	0,60	0,40	−0,10
12	0,20	0,70	0,25	0,60	−0,40	0,10
13	0,20	0,70	0,30	0,55	0,00	0,30
14	0,10	0,80	0,15	0,50	0,30	−0,30
15	0,10	0,80	0,15	0,45	−0,10	0,30
16	0,10	0,80	0,10	0,45	0,10	−0,30
17	0,00	0,90	0,10	0,40	0,20	0,00
18	0,00	0,90	0,05	0,40	−0,20	0,20
19	0,00	0,90	0,10	0,30	0,20	−0,10
20	0,00	0,90	0,05	0,25	−0,20	0,00

Kombinací všech variant vzniklo celkem 36 datových sad.

3 VÝSLEDKY

Po vygenerování testovacích dat jsme na každé sadě spočítali jak CEFA, tak BEFA. Výsledky CEFA pro jednotlivé datové sady, tj. *p*-hodnoty, RMSEA a CFI, obsahuje tabulka č. 2. Pro účely našeho srovnání jsou však podstatnější odhady hodnot faktorových nábojů[5]. Kompletní výsledky jsou v příloze, zde se soustředíme pouze na souhrnné statistiky, uvedené v tabulce č. 3.

Sloupec „odchylka“ porovnává *true* hodnoty a bodové odhady: absolutní hodnoty rozdílu těchto dvou hodnot jsou pro každou datovou sadu zprůměrovány. Čím menší je tedy výsledné číslo, tím přesněji byly faktorové náboje odhadnuty. Při celkovém pohledu se tedy zdá, že se odhady CEFA a BEFA neliší; ani při bližším zkoumání a porovnávání jednotlivých odhadů odpovídajících faktorových nábojů jsme neodhalili žádné systematické rozdíly. Celkový průměr odchylek napříč všemi datovými sadami je pro CEFA 0,135 a pro BEFA 0,145.

[5] Kvůli tomuto zaměření jsme pro BEFA žádné indexy vhodnosti neimplementovali.

Sloupec „správných odhadů“ udává procentuální úspěšnost „trefení se“ intervalovým odhadem tak, aby obsahoval *true* hodnotu faktorového náboje. BEFA má obvykle vyšší počet, neboť 95% intervaly věrohodnosti jsou širší než 95% intervaly spolehlivosti u CEFA vytvořené pomocí bootstrappingu[6].

To ukazuje právě sloupec „šířka intervalu“: šířky těchto 95% intervalů jsou pro každou datovou sadu zprůměrovány. Celkový průměr šířek intervalů napříč všemi datovými sadami je pro CEFA 0,143, ale pro BEFA 0,227 – je možné, že použitá varianta bootstrappingu šířku intervalů podceňuje. Nicméně vzhledem k tomu, že bayesovské intervaly věrohodnosti lze interpretovat jako interval, který na základě daných dat s 95% pravděpodobností obsahuje *true* hodnotu, jsou pro praktické použití vhodnější – poskytují lepší informaci o tom, zdali má vybraná položka skutečně nulový či nenulový faktorový náboj.

Sloupec „délka analýzy“ udává délku analýzy dané datové sady v minutách. Zatímco u CEFA se délka pohybuje vždy pod 1 minutu (včetně bootstrappingu), u BEFA je průměrná délka 17,9 minuty, se směrodatnou odchylkou 21,3 minuty; u větších vzorků není výjimkou, že analýza může trvat více než hodinu.

Celkově nízké hodnoty „správných odhadů“ – celkový průměr napříč datovými sadami je 42,2 % pro CEFA a 53,6% pro BEFA – však ukazují, že pro přesný odhad velikost faktorových nábojů jsou zapotřebí větší vzorky, než které jsme použili. V některých případech je i u největších vzorků průměrná odchylka větší než 0,2 – vzhledem k tomu, s jak malými faktorovými náboji se v psychometrice někdy pracuje, je zřejmě, že je třeba interpretovat výsledky explorativní faktorové analýzy velmi opatrně. Vliv má jednoduchost struktury: varianta 1 matice $\boldsymbol{\Lambda}$, která obsahuje položky, které jsou vždy syceny především právě jedním faktorem, vede k mnohem lepším odhadům, než varianty 2 a 3 (což je vidět jak z menších průměrných odchylek, tak z většího poměru správných odhadů). Varianty 2 a 3 jsou přitom podle našeho názoru reálnější, neboť hledat jednodimenzionální položky je těžké. Matoucí rovněž může být zjištění, že *noisy* varianty dosahují menších odchylek; tento zdánlivý paradox má jednoduché vysvětlení: pokud jsou komunality $\boldsymbol{Q}$ nízké a zároveň jsou nízké i rozptyly chyb $\boldsymbol{U}$, jsou odhady faktorových nábojů nadhodnocené[7].

[6] S pomocí algoritmu z *R* balíčku *psych* (Revelle, 2016): z původního simulovaného vzorku respondentů byl náhodně s nahrazením vybrán stejný počet respondentů, na kterém byla provedena faktorová analýza s použitím původní rotace. Tento postup byl 1000krát opakován pro získání dostatečného počtu *bootstrap* vzorků faktorových nábojů.

[7] Detailnější vysvětlení tohoto jevu spolu se simulacemi je součástí přílohy.

Tabulka 2 – Výsledky CEFA

Datová sada	Vysvětlený rozptyl	$p^{\dagger}$	$\chi^{2\ddagger}$	RMSEA	CFI
int large 1	0,88	**0,00**	373,36	0,06	0,99
int large 1 *noisy*	0,71	0,08	175,70	0,02	1,00
int large 2	0,74	**0,00**	233,41	0,03	0,99
int large 2 *noisy*	0,58	0,09	175,10	0,02	1,00
int large 3	0,79	**0,03**	186,22	0,02	1,00
int large 3 *noisy*	0,51	0,49	150,92	0,01	1,00
int medium 1	0,87	0,74	139,49	0,00	1,00
int medium 1 *noisy*	0,72	0,57	147,40	0,01	1,00
int medium 2	0,80	**0,00**	376,52	0,08	0,98
int medium 2 *noisy*	0,53	0,78	137,31	0,00	1,00
int medium 3	0,61	0,52	149,33	0,01	1,00
int medium 3 *noisy*	0,58	0,57	147,16	0,01	1,00
int small 1	0,93	**0,03**	185,89	0,06	0,99
int small 1 *noisy*	0,70	0,58	146,69	0,03	1,00
int small 2	0,74	**0,02**	187,91	0,06	0,98
int small 2 *noisy*	0,59	0,05	180,15	0,06	0,99
int small 3	0,69	**0,04**	182,20	0,06	0,99
int small 3 *noisy*	0,44	0,73	139,71	0,01	1,01
ord large 1	0,79	**0,02**	187,52	0,02	1,00
ord large 1 *noisy*	0,65	0,09	174,81	0,02	1,00
ord large 2	0,62	**0,00**	307,40	0,05	0,98
ord large 2 *noisy*	0,34	0,63	144,51	0,00	1,00
ord large 3	0,52	**0,00**	249,71	0,04	0,98
ord large 3 *noisy*	0,45	**0,00**	445,38	0,06	0,94
ord medium 1	0,78	**0,00**	242,66	0,05	0,99
ord medium 1 *noisy*	0,61	0,69	141,90	0,00	1,00
ord medium 2	0,57	**0,00**	242,19	0,05	0,97
ord medium 2 *noisy*	0,44	0,16	168,16	0,02	0,99
ord medium 3	0,52	0,16	168,26	0,02	0,99

ord medium 3 *noisy*	0,37	0,07	177,16	0,03	0,98
ord small 1	0,77	**0,04**	182,71	0,06	0,99
ord small 1 *noisy*	0,74	0,73	140,01	0,01	1,01
ord small 2	0,54	0,20	165,25	0,05	0,99
ord small 2 *noisy*	0,49	0,79	136,92	0,00	1,01
ord small 3	0,51	**0,01**	199,45	0,07	0,96
ord small 3 *noisy*	0,39	0,17	167,67	0,05	0,97

† Tučně vyznačené jsou hodnoty $p < 0{,}05$.
‡ $df = 151$

Tabulka 3 – Srovnání CEFA a BEFA

	Odchylka*		Šířka intervalu†		Správných odhadů‡		Délka analýzy††	
Datová sada	CEFA	BEFA	CEFA	BEFA	CEFA	BEFA	CEFA	BEFA
int large 1	0,08	0,08	0,05	0,12	23 %	**53 %**	0,47	91,61
int large 1 *noisy*	0,07	0,07	0,07	0,13	40 %	**63 %**	0,53	51,37
int large 2	0,21	0,22	0,07	0,12	**20 %**	**20 %**	0,73	55,86
int large 2 *noisy*	0,17	0,17	0,10	0,14	33 %	**38 %**	0,67	52,53
int large 3	0,27	0,27	0,07	0,12	**13 %**	**13 %**	0,62	62,46
int large 3 *noisy*	0,15	0,16	0,10	0,15	43 %	**50 %**	0,57	43,22
int medium 1	0,08	0,09	0,07	0,16	45 %	**63 %**	0,51	16,75
int medium 1 *noisy*	0,04	0,05	0,10	0,18	65 %	**90 %**	0,36	8,73
int medium 2	0,25	0,25	0,09	0,18	8 %	**15 %**	0,70	25,18
int medium 2 *noisy*	0,15	0,15	0,15	0,21	**50 %**	48 %	0,55	13,53
int medium 3	0,19	0,20	0,13	0,20	28 %	**33 %**	0,34	8,73
int medium 3 *noisy*	0,18	0,19	0,14	0,22	43 %	**53 %**	0,69	24,87
int small 1	0,09	0,12	0,08	0,25	38 %	**70 %**	0,58	11,37
int small 1 *noisy*	0,08	0,09	0,17	0,31	68 %	**85 %**	0,47	2,52
int small 2	0,23	0,25	0,16	0,30	18 %	**28 %**	0,43	5,30
int small 2 *noisy*	0,18	0,20	0,21	0,33	45 %	**58 %**	0,63	18,58
int small 3	0,23	0,25	0,18	0,32	30 %	**38 %**	0,66	4,49
int small 3 *noisy*	0,15	0,17	0,27	0,37	60 %	**68 %**	0,29	2,24
ord large 1	0,05	0,05	0,06	0,13	48 %	**65 %**	0,31	15,01
ord large 1 *noise*	0,07	0,07	0,08	0,13	43 %	**63 %**	0,37	28,19
ord large 2	0,18	0,19	0,10	0,14	8 %	**25 %**	0,33	16,19
ord large 2 *noisy*	0,08	0,08	0,15	0,18	**60 %**	**60 %**	0,32	7,44
ord large 3	0,15	0,16	0,11	0,16	23 %	**35 %**	0,35	24,64
ord large 3 *noisy*	0,12	0,13	0,13	0,16	**55 %**	53 %	0,33	16,64
ord medium 1	0,06	0,07	0,09	0,18	53 %	**78 %**	0,30	2,99

ord medium 1 *noisy*	0,08	0,08	0,13	0,20	48 %	**63 %**	0,36	2,99
ord medium 2	0,16	0,17	0,15	0,20	20 %	**30 %**	0,33	7,13
ord medium 2 *noisy*	0,11	0,12	0,17	0,22	58 %	**65 %**	0,29	6,78
ord medium 3	0,15	0,16	0,15	0,22	43 %	**45 %**	0,30	2,12
ord medium 3 *noisy*	0,12	0,14	0,20	0,25	**55 %**	**55 %**	0,24	3,67
ord small 1	0,07	0,08	0,15	0,29	65 %	**75 %**	0,47	2,23
ord small 1 *noisy*	0,07	0,09	0,16	0,30	60 %	**85 %**	0,34	2,71
ord small 2	0,15	0,17	0,26	0,36	50 %	**55 %**	0,26	1,56
ord small 2 *noisy*	0,15	0,17	0,27	0,37	50 %	**65 %**	0,31	1,58
ord small 3	0,15	0,17	0,26	0,40	48 %	**58 %**	0,33	1,72
ord small 3 *noisy*	0,12	0,13	0,32	0,44	70 %	**78 %**	0,25	1,47

* Průměr absolutních hodnot rozdílů mezi *true* hodnotami a bodovými odhady faktorových nábojů; menší je lepší.
† Průměr šířek intervalů spolehlivosti, respektive intervalů věrohodnosti odhadů faktorových nábojů.
‡ Počet (v procentech) intervalů spolehlivosti, respektive intervalů věrohodnosti odhadů faktorových nábojů, které zahrnují *true* hodnotu; větší je lepší.
†† Délka trvání analýzy v minutách; menší je lepší.

4 ZÁVĚR

Bayesovský model pro exploratorní faktorovou analýzu tedy sice neposkytuje přesnější odhady faktorových nábojů, ale stále těží z obecných výhod bayesovského přístupu: poskytuje více informací, intuitivněji interpretovatelných.

Naše srovnání má jednu nevýhodu – pro každou datovou sadu jsme vygenerovali data pouze jedenkrát; ideální by však bylo opakovat generování a analýzu mnohokrát a popisné statistiky odchylek a správnosti odhadů zobrazit jako rozdělení hodnot – tak bychom se vyvarovali případů, kdy jsou některá data, souhrou více a větších náhodných chyb a složitější struktury *true* faktorových nábojů, hůře použitelná pro odhad (jako je tomu např. u datové sady *int large* 2). To by však vyžadovalo příliš času a výpočetního výkonu, který jsme neměli k dispozici[8]. Bayesovské modely jsou stále poněkud časově náročné; délka trvání MCMC samplování závisí na velikosti dat i počtu parametrů modelu, ale neexistuje tu lineární vztah; často má velký vliv i to, jak *sampler* prochází parametrový prostor, což je z podstaty jednak náhodné, a rovněž závisející na konkrétních hodnotách dat.

[8] Čtenář s dostatkem volného času však může – a my to doporučujeme – měnit *random seed* ve zdrojovém kódu a analýzy opakovat.

Drobné nevýhody má i samotná implementace modelu, která zatím není připravena pro praktické použití (neboť vyžaduje ruční úpravy zdrojového kódů, např. pro volbu jiného priorního rozdělení parametrů či jiné rotace faktorových nábojů) a neumí se vyrovnat s chybějícími hodnotami (což pro srovnávací využití nebylo nutné).

Vzhledem k výše uvedenému proto bayesovský přístup k faktorové analýze doporučujeme používat prozatím spíše pro konfirmatorní analýzu.

5 PŘÍLOHY

5.1 ZDROJOVÉ KÓDY

Zdrojové kódy a kompletní výsledky jsou k dispozici v git repozitáři https://github.com/jakubmazanec/bayesian-efa.

Je nutné mít nainstalovány *R* balíčky *psych* (který se instaluje standardním způsobem v *R* příkazem `install.packages("psych")`) a *rstan* (podrobný návod je k dispozici jak pro Windows: https://github.com/stan-dev/rstan/wiki/Installing-RStan-on-Windows, tak pro Linux či OS X: https://github.com/stan-dev/rstan/wiki/Installing-RStan-on-Mac-or-Linux).

Protože analýzy jsou časově náročné, doporučujeme nevolat funkci `testData` vícekrát najednou.

5.2 DODATEČNÉ SIMULACE CEFA

Abychom demonstrovali, jakým způsobem spolu souvisí faktorová struktura, velikost rozptylů chyb a velikost vzorku, provedli jsme jednoduchou simulaci CEFA. Vygenerovali jsme data podle dvoufaktorového modelu: 20 položek, z nich polovina byla sycena pouze prvním faktorem, zatímco druhá polovina pouze druhým faktorem. Nenulové faktorové náboje měly všechny stejnou hodnotu, stejně jako rozptyly chyb.

Variovali jsme velikost vzorku (100, 200, 400, 800, 1600 a 3200), velikost faktorových nábojů (0,1–0,9, po desetinách) a velikosti rozptylů chyb (0,1–1,0, po desetinách). Hodnotícím kritériem přesnosti odhadu byla opět průměrná odchylka od *true* hodnot, tj. průměr absolutních hodnot rozdílů mezi odhadem a *true* hodnotou faktorového náboje. Celkem tak vzniklo 540 variant dat, na kterých jsme spočítali CEFA s metodou extrakce *maximum likelihood* a varimax rotací.

Výsledky ukázaly, že mezi všemi variovanými proměnnými, tj. velikostí vzorku, velikostí faktorových nábojů a velikostí rozptylů chyb je interakce. U vysokých faktorových nábojů a nízkých rozptylů chyb jsou odchylky pochopitelně nízké; u vysokých faktorových nábojů a vysokých rozptylů chyb jsou odchylky větší. Ovšem u kombinace nízkých faktorových nábojů a nízkých rozptylů chyb jsou odchylky ještě větší, a dokonce větší než u kombinace nízkých faktorových nábojů a vysokých rozptylů chyb. Proč?

Podívejme se na odhady faktorových nábojů pro dvě varianty[9]: 1) velikost vzorku 100, velikost faktorových nábojů 0,3 a velikost rozptylů chyb 0,1 a 2) velikost vzorku 100, velikost faktorových nábojů 0,3 a velikost rozptylů chyb 0,7. Zatímco u první varianty je průměrná odchylka 0,23, u druhé varianty je to 0,11. Jak je vidět z následující tabulky, u první varianty jsou odhady faktorových nábojů relativně konzistentní, ale velmi nadhodnocené, a proto sice mnohem méně konzistentní, ale zato nenadhodnocené, odhady v druhé variantě dosahují menších odchylek:

			$\sigma_\varepsilon = 0{,}1$		$\sigma_\varepsilon = 0{,}7$	
	Odhadnuté hodnoty $\boldsymbol{\Lambda}$		*True* hodnoty $\boldsymbol{\Lambda}$		*True* hodnoty $\boldsymbol{\Lambda}$	
i	$k = 1$	$k = 2$	$k = 1$	$k = 2$	$k = 1$	$k = 2$
1	0,30	0,00	0,42	0,04	0,13	0,03
2	0,30	0,00	0,48	0,10	0,04	0,08
3	0,30	0,00	0,45	0,09	0,15	0,18
4	0,30	0,00	0,40	0,05	0,12	0,13
5	0,30	0,00	0,44	0,09	0,08	0,08
6	0,30	0,00	0,43	0,00	0,00	0,01
7	0,30	0,00	0,46	0,03	0,26	0,06
8	0,30	0,00	0,38	0,04	0,01	0,12
9	0,30	0,00	0,38	0,09	0,15	0,07
10	0,30	0,00	0,40	0,11	0,21	0,06
11	0,00	0,30	0,02	0,35	0,07	0,04
12	0,00	0,30	0,01	0,44	0,03	0,10
13	0,00	0,30	0,02	0,28	0,10	0,06
14	0,00	0,30	0,01	0,41	0,25	0,11
15	0,00	0,30	0,04	0,48	0,22	0,11
16	0,00	0,30	0,17	0,33	0,13	0,03
17	0,00	0,30	0,02	0,48	0,12	0,21
18	0,00	0,30	0,01	0,39	0,02	0,28
19	0,00	0,30	0,02	0,49	0,13	0,06
20	0,00	0,30	0,11	0,41	0,21	0,03

[9] Kompletní tabulka s odchylkami pro všechny varianty se nachází v git repozitáři.

U větších vzorků je tento efekt ještě výraznější – i v datech obsahujících hodně náhodného šumu lze slabou faktorovou strukturu odhadnout velmi přesně, neboť ve velkém množství vzorků se náhodné chyby vždy „vyruší“, bez ohledu na jejich velikost; naopak pokud je chybový rozptyl malý, jsou malé *true* hodnoty faktorových nábojů odhadnuty nadhodnocené, podobně jako u menšího vzorku.

6 POUŽITÉ ZDROJE

Conti, G., Frühwirth-Schnatter, S., Heckman, J. J., & Piatek, R. (2014). Bayesian exploratory factor analysis. *Journal of Econometrics*, *183*(1), 31-57.

Geweke, J., & Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *Review of Financial Studies*, *9*(2), 557-587.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research common errors and some comment on improved practice. *Educational and Psychological measurement*, *66*(3), 393-416.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*(3), 187-200.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Lee, S. Y. (2007). *Structural equation modeling: A Bayesian approach*. John Wiley & Sons.

Lockwood, J. R., Savitsky, T. D., & McCaffrey, D. F. (2015). Inferring constructs of effective teaching from classroom observations: An application of Bayesian exploratory factor analysis without restrictions. *The Annals of Applied Statistics, 9*(3), 1484-1509.

Lopes, H. F., & West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, 41-67.

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Revelle, W. (2016). *Psych: Procedures for Personality and Psychological Research*. Northwestern University, Evanston, Illinois, USA. https://CRAN.R-project.org/package=psych

Stan Development Team (2016). *RStan: the R interface to Stan. R package version 2.14.1*. http://mc-stan.org

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology, 9*(2), 79-94.

Mazanec & Dvořáková (2017): Bayesian exploratory factor analysis: A comparison with the classical approach

Abstract: *In this paper we compare the classic and Bayesian exploratory factor analysis. We have used randomly generated data with different sample sizes, factor loading structures, response scales, and random error sizes. We have conducted both types of exploratory factor analysis and we can conclude, on the basis of comparison of our results and overall experience from the analyses, that the classic and Bayesian exploratory factor analyses do not differ in their accuracy in estimating factor loadings. Bayesian approach provides more information which can be interpreted more intuitively, but on the other hand, the model we have used has certain disadvantage: the necessity to manually edit the source codes, and the time-consuming calculations. Therefore, we consider Bayesian approach to be more suitable for the confirmatory factor analysis.*

Keywords: *exploratory factor analysis; Bayesian statistics*