John Tukey, in talking about data analysis, has made the useful distinction between exploratory and confirmatory study. In confirmatory work, we know, a priori, or we think we know, a great deal about what will go on and we are prepared to state rather sharp hypotheses about how the data will act. And our purpose is to adjudicate, in depth, these rather specific ideas. In exploratory work, on the other hand, we know very little ahead of time; the best we can do is take observations on a whole pot-full of random variables which we suspect may be relevant, and then see what happens. We're engaged, as it were, in a fishing expedition.

Kaiser, H.F. (1970). A second generation little jiffy. *Psychometrika* **35**, 401–415. https://doi.org/10.1007/BF02291817

MUNI FSS Katedra psychologie

Faktorová analýza

PSYb2590: Základy psychometriky | **Přednáška 5**

19. 3. 2024 | Petr Palíšek, Petra Hubatková & Hynek Cígler (& Adam Ťápal)

Obsah

CFA

- Restrikce a identifikace
- Shoda modelu s daty (globální, lokální, vizualizace, stupně volnosti)
- Modifikační indexy (a tunění modelu)
- Reportování

EFA

- Specifikace
- Rotace
- Metody odhadu počtu faktorů
- Tipy

Ostatní

CFA

“konfirmační” faktorová analýza – ale lépe “**restricted**”

Předem se specifikuje očekávaná latentní struktura

Specifikace modelu: *fixování*, *omezování*, nebo *uvolňování* prvků v maticích (lambda, phi, d psi)

Např.:

- faktorový náboj prvního faktoru na první položku je 0,3 (lambda11 = 0,3)
- Korelace prvního a druhého faktoru je nulová (phi12 = 0 = phi21)
- Residální kovariance mezi první a druhou MV je volně odhadovaná (Dpsi12 = Dpsi21 = ?)

Restrikce v CFA

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ \lambda_2 & 0 \\ \lambda_3 & 0 \\ \lambda_4 & 0 \\ \lambda_5 & 0 \\ 0 & \lambda_6 \\ 0 & \lambda_7 \\ 0 & \lambda_8 \\ 0 & \lambda_9 \\ 0 & \lambda_{10} \end{bmatrix} \qquad \boldsymbol{\Phi} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{D_\Psi} = \begin{bmatrix} \xi_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \xi_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \xi_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \xi_4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \xi_5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \xi_6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \xi_7 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \xi_8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \xi_9 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \xi_{10} \end{bmatrix}$$

$$\boldsymbol{\Lambda'} = \begin{bmatrix} \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 & \lambda_5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda_6 & \lambda_7 & \lambda_8 & \lambda_9 & \lambda_{10} \end{bmatrix}$$

Identifikace

Specifikace modelu v CFA musí vést k právě jednomu možnému řešení, tj. tzv. i**dentifikaci modelu**

Neidentifikovaný model neumožňuje najít právě jedno řešení

Analogie situace se soustavou rovnic s více neznámými než rovnicemi

Příklady příčin neidentifikace:

- chybí škála pro latentní proměnné
- 0 stupňů volnosti
- méně než 3 MVs na LV

Identifikace

Specifikace modelu v CFA musí vést k právě jednomu možnému řešení, tj. tzv. i**dentifikaci modelu**

Neidentifikovaný model neumožňuje najít právě jedno řešení

Analogie situace se soustavou rovnic s více neznámými než rovnicemi

Příklady příčin neidentifikace:

- chybí škála pro latentní proměnné
- 0 stupňů volnosti
- **méně než 3 MVs na LV**

Identifikace

U jednoduchých modelů (např. >10 položek, <3 faktorů, bez residuálních kovariancí, bez crossloadingů) není s identifikací problém, škálování zajistí JASP

Uživatel specifikováním modelu implicitně stanovuje spoustu restrikcí, protože přiřazením položky k faktoru zároveň říká, že jsou faktorové náboje ostatních LVs nulové

Shoda modelu s daty

Jedním z cílů FA je zjistit, jestli specifikace modelu odpovídá pozorováním

Proto je potřeba umět posoudit, jak dobře model sedí na data

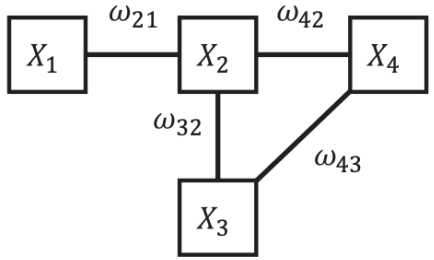
Pokud sedí dobře, je tzv. dobře specifikovaný (**well-specified**) a získáváme podporu pro to, že specifikace modelu odpovídá data-generujícímu procesu implikovanému teorií

ALE! Jen na základě posouzení shody modelu z daty nejde dovodit, jak vypadá data-generující process, protože vždy **existuje spousta ekvivalentních modelů, které na data sedí stejně dobře, ačkoliv mají jinou specifikaci**

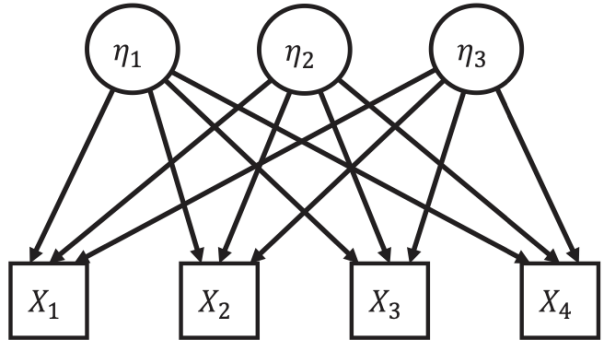
Uvažování proto musí být taženo teorií, ne statistikou (Borsboom: “statistics is a science of nothing”)

https://quantitudepod.org/s2e24-the-equivalent-models-problem/

A



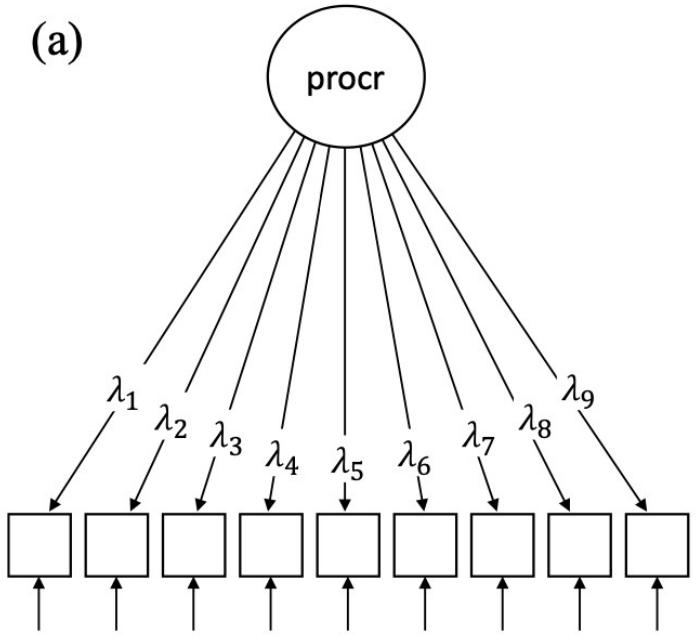
B



$$\boldsymbol{\Omega} = \begin{bmatrix} 0 & \omega_{12} & 0 & 0 \\ \omega_{21} & 0 & \omega_{23} & \omega_{24} \\ 0 & \omega_{32} & 0 & \omega_{34} \\ 0 & \omega_{42} & \omega_{43} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0.30 & 0 & 0 \\ 0.30 & 0 & 0.20 & 0.25 \\ 0 & 0.20 & 0 & 0.35 \\ 0 & 0.25 & 0.35 & 0 \end{bmatrix} \qquad \boldsymbol{A} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & 0 \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & 0 \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & 0 \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & 0 \end{bmatrix} = \begin{bmatrix} 0.407 & -0.780 & 0.113 & 0 \\ 0.796 & -0.372 & -0.126 & 0 \\ 0.739 & 0.423 & 0.249 & 0 \\ 0.782 & 0386 & -0.166 & 0 \end{bmatrix}$$

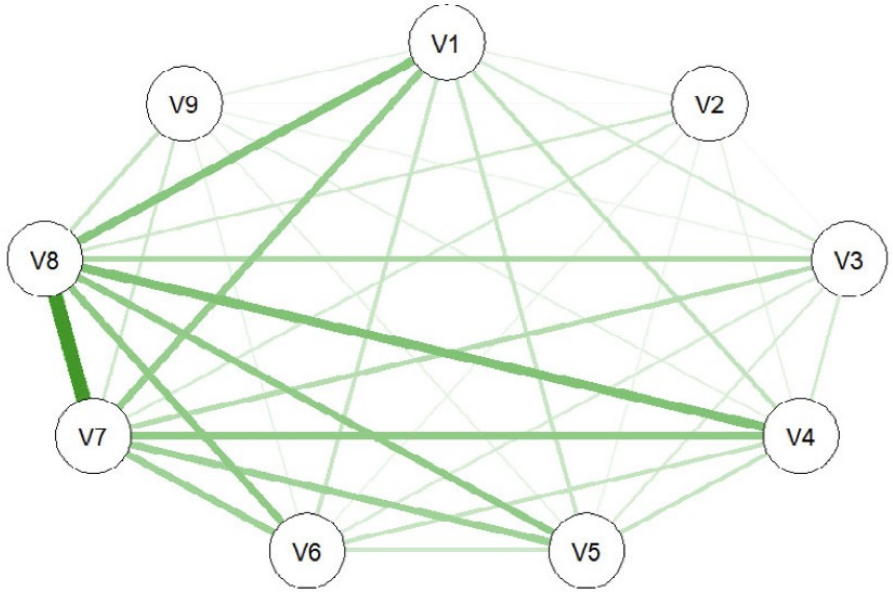
Riet van Bork, Mijke Rhemtulla, Lourens J. Waldorp, Joost Kruis, Shirin Rezvanifar & Denny Borsboom (2021) Latent Variable Models and Networks: Statistical Equivalence and Testability, *Multivariate Behavioral Research*, 56:2, 175-198, DOI: 10.1080/00273171.2019.1672515

(a)



λ_1 = 0.618
λ_2 = 0.300
λ_3 = 0.488
λ_4 = 0.640
λ_5 = 0.589
λ_6 = 0.590
λ_7 = 0.800
λ_8 = 0.830
λ_9 = 0.344

(b)



Shoda modelu s daty

Stupně volnosti

Model je (obvykle) zjednodušením reality, jinak by nemělo smysl jej tvořit

Např. máme-li 8 MVs a jednoduchý dvoufaktorový model, tak namísto celkem k(k+1)/2 = 8*9/2 = 36 kovariancí a rozptylů odhadujeme jen:

- Faktorový náboj pro každou MV (8 nábojů)
- Korelaci mezi faktory (1 korelace)
- Residuální rozptyl každé MV (8 residuálních rozptylů)

Z 56 kusů informace jsme se tak dostali na model o 8 + 1 + 8 = 17 parametrech, což je zjednodušení o 36-17 = 19. Takový model by tedy měl právě **19 stupňů volnosti.**

Shoda modelu s daty

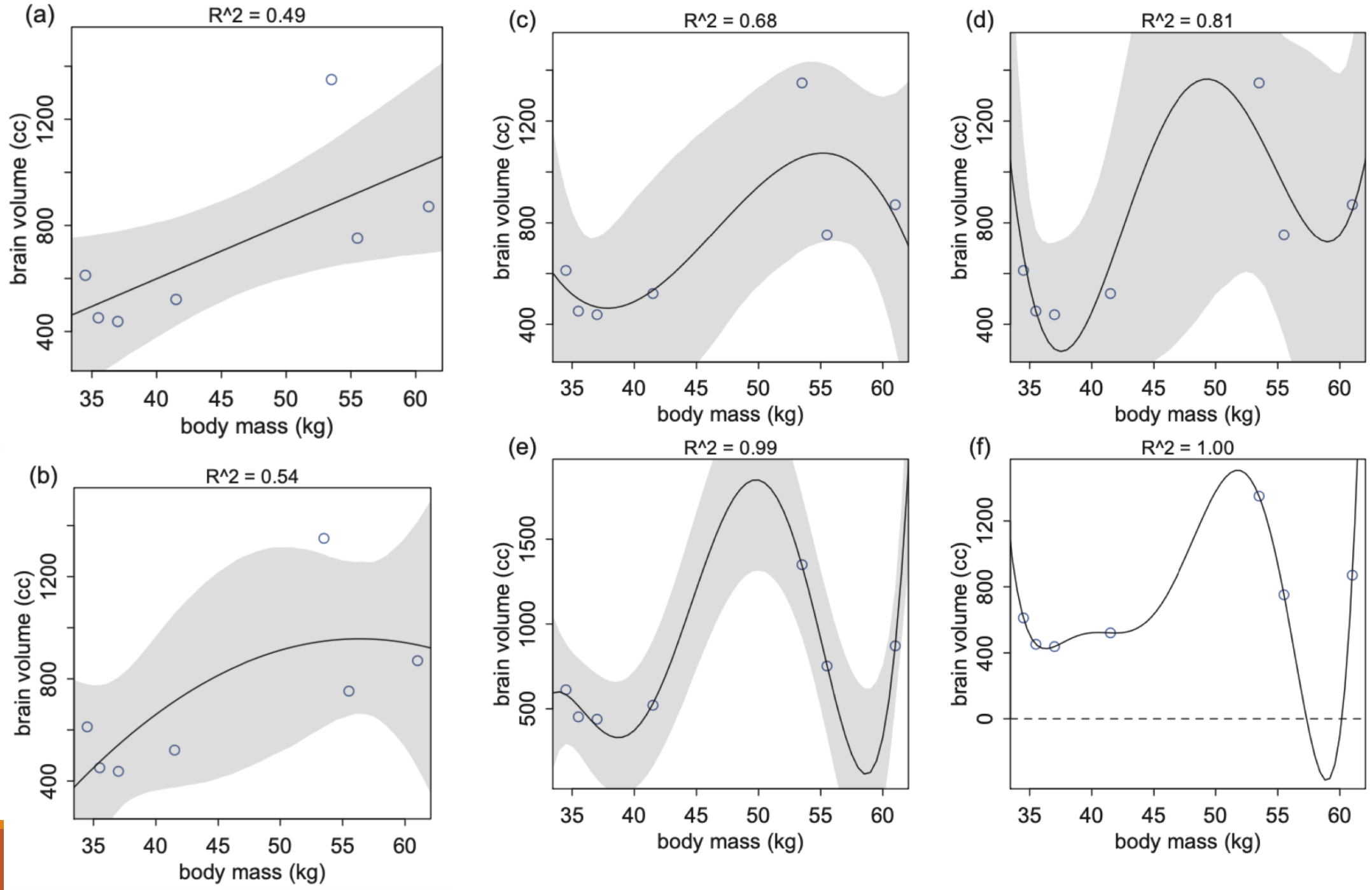
Stupně volnosti (df) tak vyjadřují, jak jednoduchý (*parsimonický*) model je oproti pozorováním

Logika je stejná i mimo FA: **za každý parametr platíte kusem informace**

Např.:

V případě rozptylu df = N-1, protože si pro odhadnutí rozptylu kupujete průměr

V případě regrese df = N-k, protože si kupujete průsečík a regresní koeficienty


(a)
R^2 = 0.49
brain volume (cc)
400
800
1200
35
40
45
50
55
60
body mass (kg)
(b)
R^2 = 0.54
brain volume (cc)
400
800
1200
35
40
45
50
55
60
body mass (kg)
(c)
R^2 = 0.68
brain volume (cc)
400
800
1200
35
40
45
50
55
60
body mass (kg)
(d)
R^2 = 0.81
brain volume (cc)
400
800
1200
35
40
45
50
55
60
body mass (kg)
(e)
R^2 = 0.99
brain volume (cc)
500
1000
1500
35
40
45
50
55
60
body mass (kg)
(f)
R^2 = 1.00
brain volume (cc)
0
400
800
1200
35
40
45
50
55
60
body mass (kg)

Shoda modelu s daty

Víme tedy, že pokud máme mít df > 0, tak pomocí $\Lambda\phi\Lambda' + D_\Psi$ nikdy pozorovanou korelační matici S nezrekonstruujeme přesně. Není to bug, ale feature – zjednodušení od modelu chceme!

Shoda modelu s daty

Víme tedy, že pokud máme mít df > 0, tak pomocí $\Lambda\phi\Lambda' + D_\Psi$ nikdy pozorovanou korelační matici S nezrekonstruujeme přesně. Není to bug, ale feature – zjednodušení od modelu chceme!

Platí ale, že pokud:

- **vezmeme pozorovanou korelační matici S a**
- **pomocí odhadnutých parametrů v Λ, ϕ a D_Ψ dopočítáme korelační matici $\hat{\Sigma} = \Lambda\phi\Lambda' + D_\Psi$**

Vznikne mezi pozorovanou a modelem implikovanou korelační maticí rozdíl:

$$S - \hat{\Sigma}$$

Shoda modelu s daty

$$\boldsymbol{C = S - \hat{\Sigma}}$$

C ... residuální matice (neplést s maticí residuálních kovariancí D_Ψ)

Prvky matice C obsahují jednotlivé rozdíly mezi pozorovanou a modelem implikovanou korelací mezi MVs

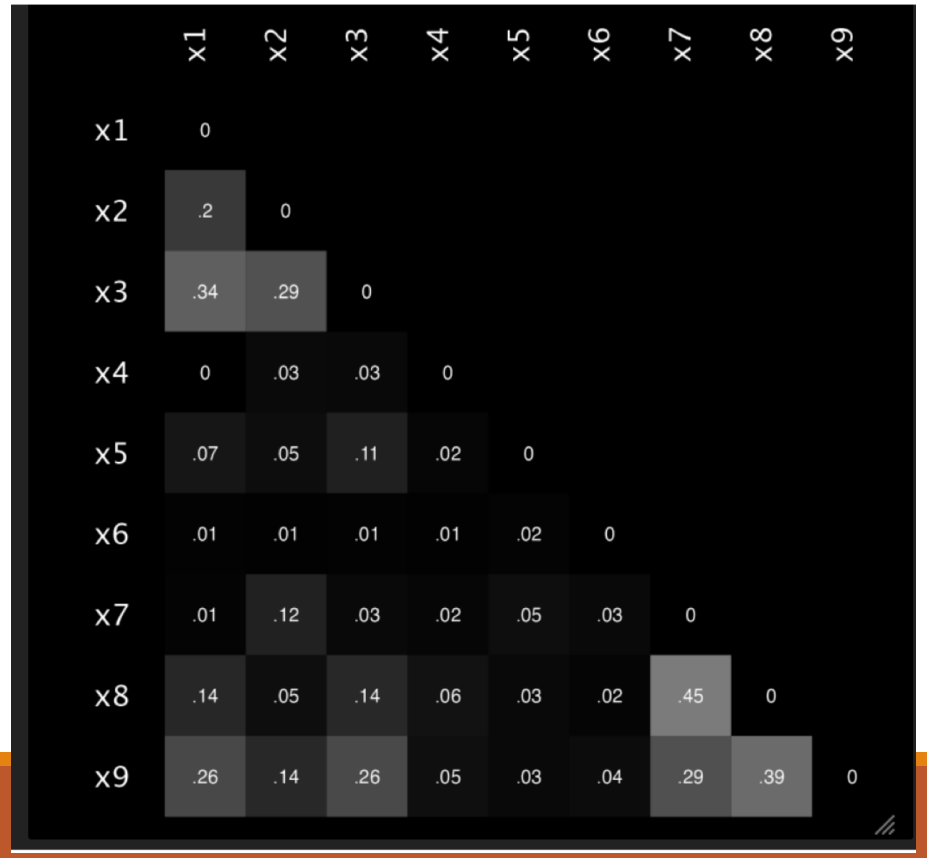
Čím méně df, tím menší rozdíl je, protože se model víc a víc přizpůsobuje datům.

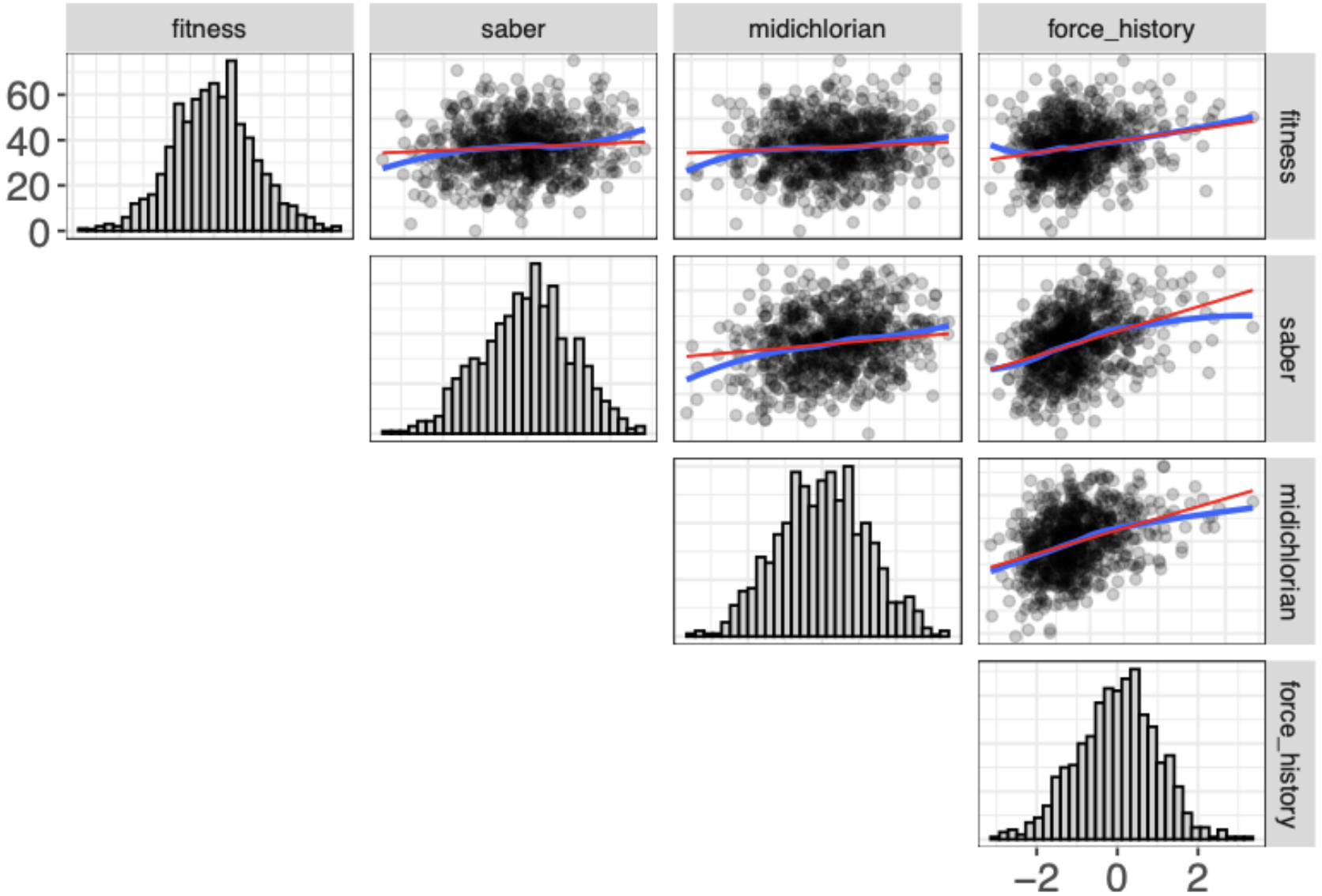
df = 0 => *plně saturovaný model* a **C** má mimo diagonálu samé 0 => vzniká *perfect fit*

Shoda modelu s daty

Residuální matici jde interpretovat přímo, v takovém případě jde o **local fit assessment**

Jednoduše se podíváme, vztahy mezi kterými MVs jsou modelem špatně vystižené





https://osf.io/preprints/psyarxiv/qm7kj

Figure 7. Scatterplot matrix showing the model-implied fit (red) and loess lines (blue) between four simulated indicator variables. These four variables are the indicators for the force latent variable. The diagonals show the histograms of the ICRs.

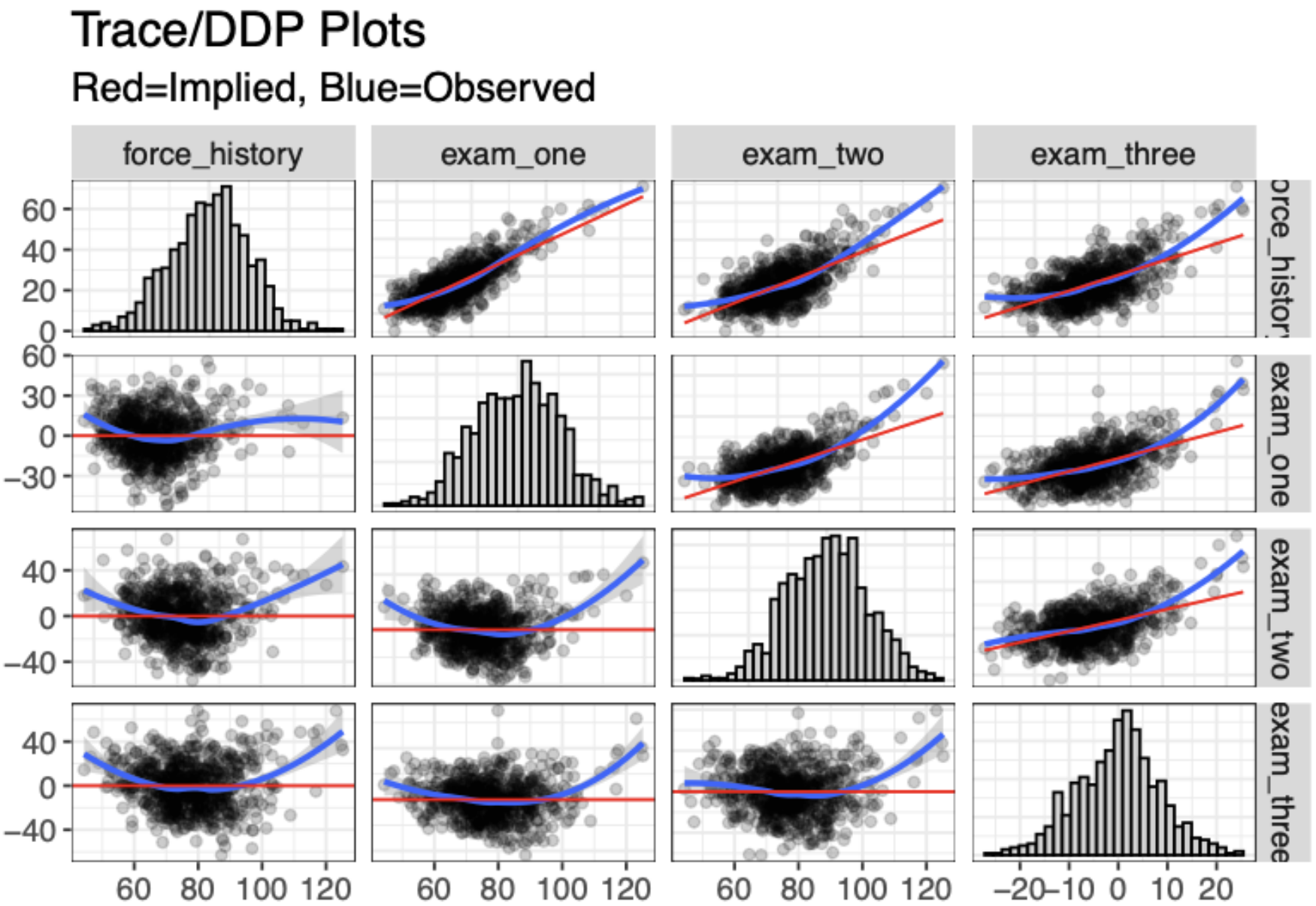


Figure 9. The upper triangle of this plot is the same as the plot shown in Figure 8. However, the lower triangle adds the disturbance-dependence plots.

Shoda modelu s daty

Residuální matice jde také shrnout do jednoho čísla, které následně interpretujeme, tomu se říká **global fit assessment**

Takovým souhrnům se říká **fit indices**, ”ukazatele shody modelu s daty”

Existuje jich celá řada, každý má své výhody / nevýhody a jejich interpretace je poměrně obtížná, např.: RMSEA, SRMR, TLI

Výzkumníci se často spoléhají na pravidla doporučená Hu a Bentlerem (1999), což je ale chyba podobná $p < .05$

Ideální je buď interpretovat všechny informace o modelu společně a s porozuměním, nebo počkat ještě pár let na vymakání: https://www.dynamicfit.app/connect/

Pro potřeby tohoto kurzu stačí používat tento návod: https://davidakenny.net/cm/fit.htm

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

SRMR

Standardized Root Mean Squared Residual

$$SRMR = \sqrt{\frac{\sum(Expected - Observed)^2}{Number\ of\ elements}}$$

Výsledkem je průměrná velikost prvku v residuální matici (tj. průměrná vzdálenost mezi odhadnutou a pozorovanou korelací)

Čím menší, tím lepší. Tradiční cut-off: < 0,08

Výhody: Velmi snadno interpretovatelné, nezávislé na jiných vlastnostech modelu

Nevýhody: U malých N méně spolehlivé kvůli výběrové chybě

Chí-kvadrát

Vzdálenost (diskrepanci) $S - \hat{\Sigma}$ jde vyjádřit i pomocí statistiky F, která má po vynásobení N-1 přibližně chí-kvadrát rozložení: $F(N-1) \sim \chi^2(df)$

U správně specifikovaného modelu v průměru očekáváme "diskrepanci" $\chi^2(df) = df$ (vyšší naznačuje špatně specifikovaný model)

Tuto hodnotu lze testovat přímo pomocí *test of perfect fit*, který testuje H0, že mezi S a $\hat{\Sigma}$ není rozdíl => chceme proto vysokou p-hodnotu

Test of perfect fit má nicméně tendenci zamítat všechny modely odhadnutné na solidních vzorcích, proto **má vypovídací hodnotu, jen s vysokou p-hodnotou (tehdy totiž model musí sedět fakt dobře).**

RMSEA

Root Mean Squared Error of Approximation

$$RMSEA = \sqrt{\frac{\chi^2 - df}{[df(N-1)]}}$$

Čím menší, tím lepší. Tradiční cut-off: < 0,05 (někdy 0,08)

Výhody: ”effect size” pro test of perfect fit, velmi rozšířené, má intervaly spolehlivosti, bere v potaz N a df

Nevýhody: méně srozumitelné

TLI

Tucker-Lewis Index

$$TLI = \frac{\frac{\chi^2_{null}}{df_{null}} - \frac{\chi^2}{df}}{\frac{\chi^2_{null}}{df_{null}} - 1}$$

Odpověď na otázku: **V kolika procentech cesty mezi nejhorším a nejlepším možným modelem se nachází můj model?**

Čím vyšší, tím lepší. Tradiční cut-off: > 0,90

Výhody: srozumitelný, bere v potaz kontext daného modelu

Nevýhody: silně závislý na správném nastavení škály od “nejhoršího” k “nejlepšímu” – když je nejhorší (nulový) model moc dobrý, tak podhodnocuje

Hodnocení shody modelu s daty

1. Jsou odhady parametrů plausibilní (korelace od -1 do 1, rozptyly > 0)? Mám aspoň df = 1?

- NE: stala se někde chyba a nemá smysl jít dál.

2. Má test of perfect fit vysokou p-hodnotu?

- ANO: wow!
- NE: nevadí

3. Jak vypadá SRMR?

4. Jak vypadá RMSEA a TLI?

5. Jak vypadá residuální matice? Není někde výrazný ústřel?

Pro účely tohoto kurzu stačí kroky 3 a 4 vyhodnocovat dle tradičních cut-offů.

Když model nesedí

Pokud máme podezření na špatnou specifikaci lze:

1. Model **porovnat s konkurenčním modelem** (třeba je špatný, ale pořád nejlepší)
2. **Využít modifikačních indexů**, tj. odhadnuté “doporučení” pro změnu restrikcí tak, aby se model přiblížil datům – nutno používat velmi opatrně a tunění vždy obsahově odůvodnit (pokud nejde o přiznanou exploraci)
3. **Přejít do EFA.** Nelze ale konfiromovat a explorovat na těch stejných datech. Nepodložené úpravy znamenají exploraci, a musí tak být i reportovány.

Reportování CFA

Co nesmí chybět

1. Metoda odhadu (typicky ML)
2. Specifikace (např. slovní popis)
3. Samotné odhady parametrů (obsah matic lambda, phi, d-psi)
4. Chí-kvadrát, df, p-hodnota; SRMR; RMSEA a CI pro RMSEA; TLI.
   Užitečné může být i uvádění chí-kvadrátu a df nulového (baseline) modelu, který nabízí JASP.
5. Residuální matice (klidně do přílohy; v praxi se skoro nevyskytuje, ale je to škoda)

EFA

"Unrestricted" FA

Specifikuje se jen počet faktorů a metoda rotace

Jinak vždy platí, že každá LV způsobuje každou MV (proto unrestricted)

V EFA nejsou residuální kovariance (D_Ψ má mimo diagonálu vždy samé 0)

Typickým cílem je pomocí explorace najít a popsat latentní strukturu za daty. Nejde tedy jen o hru s čísly, ale o spojení obsahových úsudků o povaze LVs a jejich vztahu k MVs s empirickými výsledky.

EFA končí, když je nejen specifikován počet faktorů, ale jsou i pojmenovány / popsány

Rotace

Pro EFA s >2 faktory vzniká **rotational indeterminancy = existuje nekonečně mnoho stejně dobrých řešení** (model není identifikovaný)

Lze si tedy vybrat řešení, které je nejlépe **interpretovatelné**, obvykle ve smyslu **simple structure:**

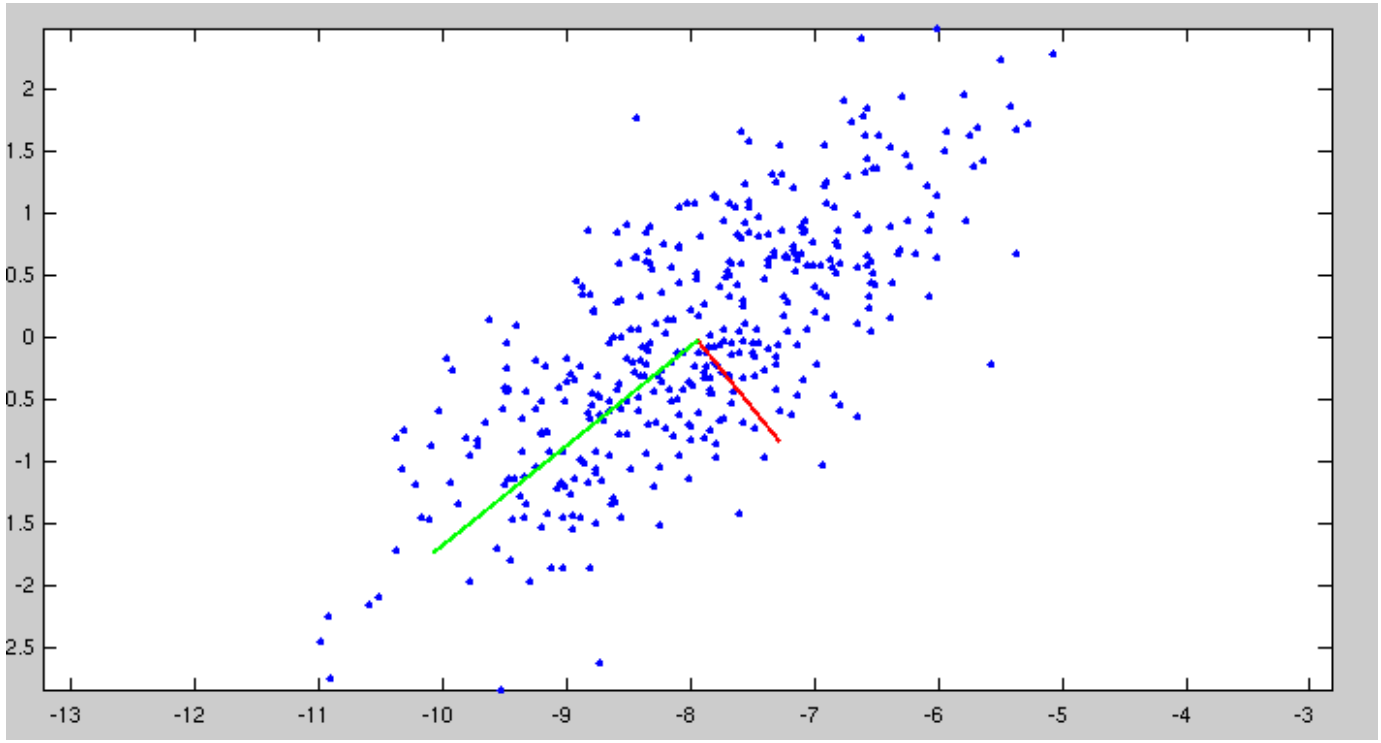
- Aby každá MV byla způsobována 1 LV (tj. minimum crossloadingů)

Rotace existují **ortogonální** (faktory nemohou korelovat), nebo **šikmé** (faktory mohou korelovat)

Jednotlivé rotace se jmenují dle kritéria, pomocí něhož hledají konkrétní řešení (např. varimax)

Eigenvalues

Složitější koncept, ale stačí eigenvalues brát jako **vyjádření rozptylu v korelační matici**



Metody odhadu počtu faktorů

Historicky:

Kaiserovo pravidlo = extrahujte tolik faktorů, kolik jich má eigenvalue > 1

Vizuální věštění ze scree plotu = hledání bodu, kde se scree plot láme

Moderně:

Hornova paralelní analýza = Kaiserovo pravidlo očištěné o výběrovou chybu

Empirické metody je ale třeba kombinovat s obsahovou úvahou (interpretovatelností LVs)

EFA: Tipy

Reportujte:

- estimátor,
- postup odhadu počtu faktorů,
- volbu rotace,
- lambda a phi matici
- % vysvětleného rozptylu (či jiné ukazatele shody s daty)

Je žádoucí zkusit víc řešení!

Pozor na Heywoodovy případy (např. komunalita > 1)

Výborně sedící model, co nedává smysl, je k ničemu.

Smysluplný model, co nesedí na data, je taky k ničemu.

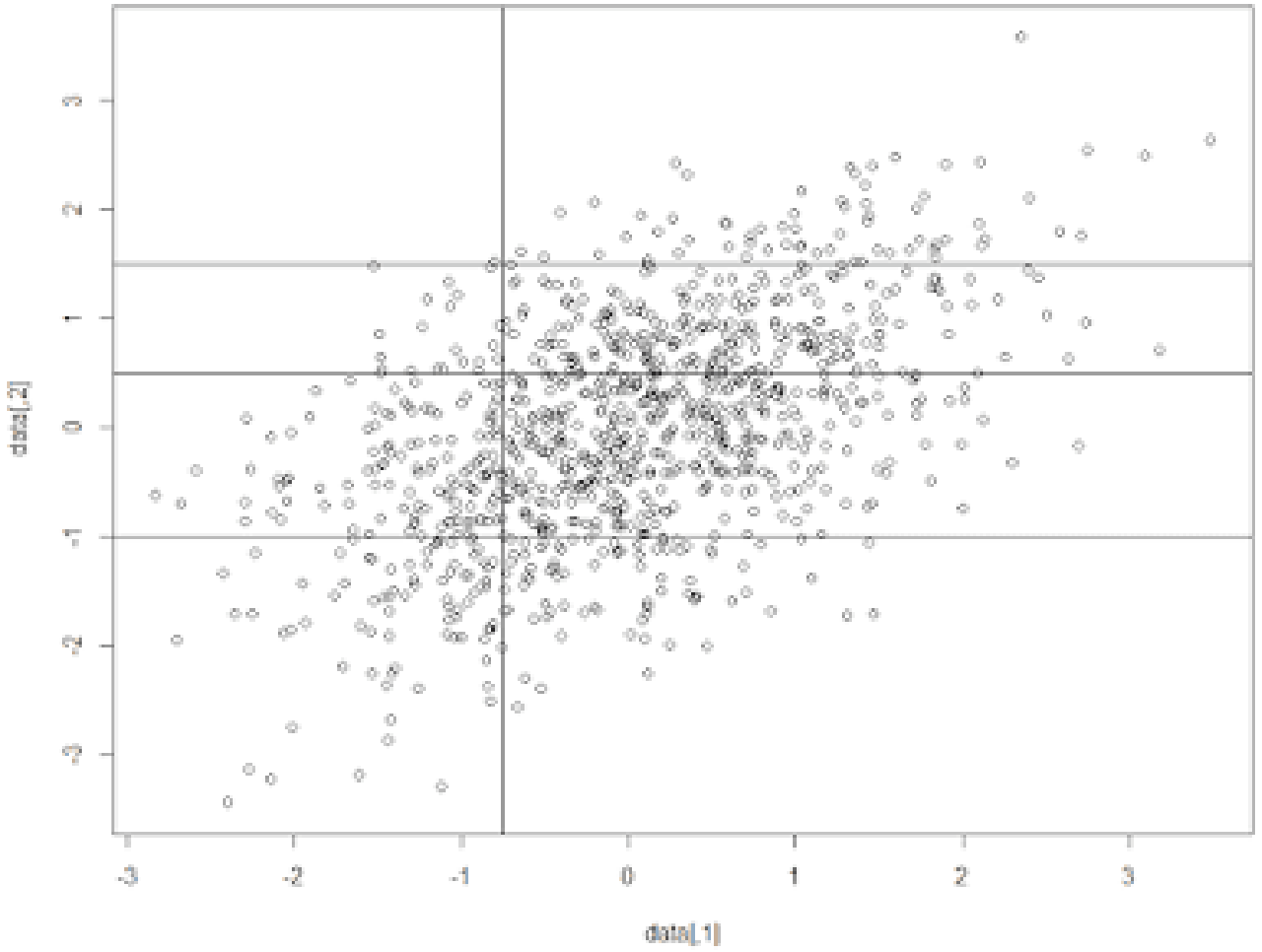
Ordinální FA

Původní FA očekává spojité MVs, v psychologii ale typicky máme Likertovu škálu = ordinální MVs

Při dostatečném počtu bodů je možné použít FA a spojitost aproximovat

Namísto Pearsonových korelací ale lze použít i **polychorické korelace**

Používá se estimátor DWLS, který ale nadhodnocuje shodu s daty, proto se musí korigovat pomocí **WLSMV**



PCA

Principal Components Analysis (analýza hlavních komponent)

Příbuzný EFA (autorem je Hotelling ve 20. letech 20. století)

Technicky velmi podobná FA, ale nepředpokládá kauzální vliv LVs na MVs, jde jen o *data reduction technique*

Rozdíl se promítá do absence konceptu komunality/unicity

Ačkoliv by PCA a EFA často přinesly podobné výsledky, **je důležité je nezaměňovat**

SEM

Structural Equation Modelling (strukturní modelování)

FA je speciálním případem SEM, ve kterém LVs způsobují MVs, zatímco spolu vzájemně mohou korelovat

SEM umožňuje přidat kauzální cesty mezi MVs / LVs a společně modelovat:

1. **measurement part** = “skládat” LVs z MVs jako v FA
2. **structural part** = kauzální cesty mezi LVs / MVs

IRT

Item Response Theory (Teorie odpovědi na položku)

Modely z rodiny IRT mají skoro stejnou logiku jako FA

Pokud je FA lineární regrese, pak se o tradičních IRT modelech dá uvažovat jako o logistické regresi