

Komparační faktorová analýza profilů (Model a metoda LINDA-K)

JAN ŘEHÁK

Ústav pro filozofii a sociologii ČSAV, Praha

IVANA LOUČKOVÁ

katedra aplikované sociologie
Univerzita Palackého, Olomouc

Komparační analýza patří k základním a nejčastějším postupům při práci s daty, přesto je však metodika pro některé typy datových uspořádání stále ještě nedostatečně rozvinuta. V této stati se zabýváme aplikací kanonického rozkladu matice pro vyhledávání struktury profilových komparačních tabulek a pro identifikaci faktorů, které působí na rozdílnosti mezi profily. Metoda patří do třídy modelů LINDA (lineární dekompoziční analýza) a vzhledem k jejímu komparačnímu využití ji nazýváme LINDA-K.

Metoda poskytuje různé interpretace numerických výsledků podle obsahového modelu, s nímž k analýze přistupujeme:

- vyhledávání faktorů, které působí systematické rozdíly mezi profily;
- určení komparační báze profilů, z níž jsou skládány různými kombinacemi jednotlivé profily;
- grafická korespondenční mapa vyjadřující pomocí vzdáleností bodů stupeň typičnosti položek profilu pro porovnávané objekty, resp. stupeň specifického ovlivnění položky objektem vzhledem ke vlivu na jiné položky a zobrazení podobností objektů a položek;
- uspořádání položek a objektů a jejich kvantifikace.

V článku je metoda podrobně komentována na příkladě ilustrujícím základní komparační úlohu, je popsán matematický model a je zařazen příklad, který ukazuje možnost kvantifikačního využití postupu. LINDA-K je příbuzná metodě LINDA-A (viz Řehák, Loučková [1984]) a vychází ze stejného matematického základu kanonické dekompozice obdélníkových matic. V dodatku je shrnut doplněný a opravený algoritmus výpočtů. Poznamenejme, že existuje též příbuznost s metodou DISTAN (viz Řehák, Loučková [1983]), která vyhledává vztahy mezi profily, zatímco LINDA-K vychází z interakčních profilů, které jsou očištěny odečtením řádkových a sloupcových efektů.

LINDA-K patří mezi explorační postupy, které odpovídají na otázky: „Jaké vztahy v datech lze nalézt?“, „Existuje nějaká interpretovatelná struktura datových vztahů, která odhaluje vazby přímo nezjistitelné?“. Je to formální model, jehož použitelnost závisí na interpretabilitě číselných výsledků a možnostech využít i dalších poznatků při hledání jejich významů.

1. Popis úlohy a charakteristika dat

Sociologické objekty statisticky komparujeme posouzením odlišností údajů, které je charakterizují a které vznikají realizací (měřením) určitého zvoleného souboru proměnných. Proto je statistická komparace prováděna vždy jen vzhledem k těm vlastnostem, které jsou ve zpracovávaných datech odráženy a zastoupeny. Soubor proměnných tvoří vektor (označíme jej $\mathbf{X} = (X_1, X_2, \dots, X_S)$) a soubor hodnot

zjištěných u r -tého objektu tvoří *datový vektor* (označíme jej $\mathbf{x}_r = (x_{r1}, x_{r2}, \dots, x_{rS})$). Jestliže všechny proměnné mají stejnou stupnici, tj. můžeme-li porovnávat jejich číselné hodnoty navzájem, nazveme $\mathbf{x}_r$ *datovým profilem r -tého objektu*. Typické uspořádání dat je znázorněno v tabulce 1, kterou nazýváme *profilovou* nebo *komparační tabulkou*.

Tabulka 1. *Profilová tabulka pro 3 objekty a 4 proměnné*
($R = 3, S = 4$)

objekt r	proměnná X_s				řádkový průměr	řádkový efekt
	1	2	3	4		
1	x_{11}	x_{12}	x_{13}	x_{14}	$x_{1.}$	$y_{1.}$
2	x_{21}	x_{22}	x_{23}	x_{24}	$x_{2.}$	$y_{2.}$
3	x_{31}	x_{32}	x_{33}	x_{34}	$x_{3.}$	$y_{3.}$
sloupcový průměr	$x_{.1}$	$x_{.2}$	$x_{.3}$	$x_{.4}$	celkový průměr $x_{..}$	
sloupcový efekt	$y_{.1}$	$y_{.2}$	$y_{.3}$	$y_{.4}$		

Přímá komparace profilů zahrnuje rozdílnosti mezi čísly tak, jak jsou dány v tabulce. Ty však mají různé zdroje. Jsou způsobeny jednak rovnoměrným posunutím v řádcích, tedy vlivem objektů, jednak posunutím charakteristickým pro jednotlivé profilové položky, a nakonec specifickými vlivy projevujícími se u různých dvojic objektu a položky různě. Pokud působí jen vlivy prvních dvou uvedených typů, lze vyjádřit model *marginálním aditivním vztahem*:

$$(1) \quad x_{rs} = \text{vliv } r\text{-tého objektu} + \text{posunutí na škále } s\text{-té položky}$$

Struktura dat marginálního aditivního modelu se v praxi vyskytuje velmi řídce. Častější je model:

$$(2) \quad x_{rs} = \text{vliv } r\text{-tého objektu} + \text{posunutí } s\text{-té položky} + \text{specifický posun } s\text{-té položky u } r\text{-tého objektu}$$

Vlivy objektů se projeví rozdílem řádkových průměrů nebo *řádkovými efekty (efekty objektů)*:

$$(3) \quad y_{r.} = x_{r.} - \bar{x}, \quad x_{r.} = \sum_s x_{rs}/S, \quad \bar{x} = \sum_r \sum_s x_{rs}/RS$$

Obdobně celkové posunutí na škále položky se projeví sloupcovými průměry nebo *sloupcovými efekty (efekty položek)*:

$$(4) \quad y_{.s} = x_{.s} - \bar{x}, \quad x_{.s} = \sum_r x_{rs}/R$$

Rovnici (1) lze napsat:

$$(5) \quad x_{rs} = \bar{x} + y_{r.} + y_{.s}$$

a rovnici (2) jako:

$$(6) \quad x_{rs} = \bar{x} + y_r + y_s + y_{rs} \\ = \text{celkový průměr} + \text{řádkový efekt} + \text{sloupcový efekt} + \text{interakční člen}$$

Členy y_{rs} znamenají odchylky od modelu marginální aditivity ((1) resp. (5)). Nazývají se proto *odchylkové*, *zbytkové* nebo *reziduální*. Vzhledem k tomu, že charakterizují vzájemný vztah mezi položkou a objektem, nazývají se též *interakční (vztahové) členy* nebo *efekty*. Vektor těchto efektů pro r -tý objekt, $y_r = (y_{r1}, y_{r2}, \dots, y_{rs})$, se nazývá r -tým *interakčním profilem* nebo *reziduálním profilem*.

Interakční profily mohou vznikat pouhou náhodou, chybou měření, nepatrnými vlivy, které v interpretaci nebereme v úvahu. V tom případě přijímáme model marginální aditivity (1). Interakční profily mohou ale být vytvářeny také systematicky nějakými latentními faktory⁽¹⁾, které působí s různou intenzitou na různé členy y_{rs} , na jejich velikost i směr, přičemž se skládají:

$$(7) \quad y_{rs} = \text{příspěvek faktoru 1} + \text{příspěvek faktoru 2} + \dots$$

Tyto faktory působí diferenciaci objektů vzhledem k položkám, proto je nazýváme *diferenčními faktory*.

Smyslem metody LINDA-K je nalezení takových faktorů, určení jejich vlivu a jejich obsahová interpretace. Spojením rovnic (6) a (7) a využitím výsledků metody chceme sestavit obsahový model charakterizující vznik struktury dat profilové tabulky. Hodnocení faktorů však může být značně rozdílné podle povahy dat.

Mohou to být:

- příčiny korelované s významem objektů, působící rozdíly v profilech;
- vlastnosti projevující se prostřednictvím položek profilu (nebo vlastností profilem indikované či jeho pomocí operacionalizované), které jsou příčinou diferencovanosti objektů;
- proces homogenizace některých objektů a heterogenizace jiných;
- zprostředkující proměnné v procesu vzniku profilů ap.

Úlohy praktické analýzy dat, které mají za cíl určit faktory, mohou mít různé *cíle*:

- a) *prosté zjednodušení* a přehledný popis dat, redukce informací;
- b) *rozklad složité struktury* na jednoduché interpretovatelné *složky*;
- c) úvahy o *příčinách* variability profilů, kauzální analýza vztahů mezi objekty a vlastnostmi, které jsou reprezentovány v profilech;
- d) využití modelových rovnic pro *predikce*, *rozhodování* či *vyhlazení* dat o nepodstatné složce chybového charakteru.

Podle povahy dat se setkáváme s několika typy profilových tabulek:

- A) *Profily průměrů*. Hodnoty x_{rs} vznikají jako průměry hodnot jednotlivých jednotek r -tého souboru (skupiny) nebo jako průměr opakovaných měření téže vlastnosti u jedné jednotky. Při komparaci odhlížíme od vnitřní heterogenity skupin (o níž v profilové tabulce nemáme ani informaci), všechny závěry se týkají agregovaných dat.
- B) *Individuální profily*. Hodnoty x_{rs} vznikají jako výsledek jednoho měření proměnné X_s u r -tého objektu. Individuální data neobsahují žádnou vnitřní variabilitu, jsou však zpravidla zatížena daleko vyšší náhodnou chybou než data agre-

(1) Latentní příčinou nazýváme takový vliv, jehož přímé zjišťování nebylo v datech provedeno, ale nepřímo se *projevuje svými důsledky*. V našem případě působí na hodnoty interakčních efektů.

govaná, a proto značná část interakční variability bývá ztotožněna s chybovou variací.

- C) *Profily procentních hodnot nebo četností.* Speciálním případem A) jsou procentní nebo četnostní údaje, x_{rs} je procento nebo četnost s -tého jevu ze zkoumaného souboru jevů u r -tého objektu (procento vzniká jako stonásobek průměru indikační proměnné: 1 = jev nastal, 0 = jev nenastal).
- D) *Profily rozložení četností.* x_r reprezentuje rozložení četností nějakého kategorizovaného znaku s S hodnotami, nebo nějaké funkce četností, jako jsou koncentrace polí, kumulativní četnosti, rezidua od hypotézy symetrie či dobré shody ap.
- E) *Indikační profily.* Profilové hodnoty x_{rs} mají indikační roli: pro každý z S jevů označují, zda nastal (= 1), nebo nenastal (= 0). Jde o speciální případ B).
- F) *Pořadové profily.* Hodnoty x_{rs} jsou hodnotami pořadí pro jeden z případů:
 - a) pro každý řádek (tj. hodnoty 1, 2, ..., S , příp. spojená pořadí)*
 - b) pro každý sloupec (tj. hodnoty 1, 2, ..., R , příp. spojená pořadí);
 - c) pro celou tabulku (tj. hodnoty 1, 2, ..., RS , příp. spojená pořadí).

Uvedené typy profilů jsou nejčastější, každý z nich má své zvláštní metodologické vlastnosti, které se projeví i v interpretaci a v nakládání s profilovou tabulkou (např. při testování existence nenáhodných interakčních hodnot, tj. testu platnosti modelu marginální aditivity).

2. Faktorová komparace profilů: příklad využití metody

Metodu LINDA-K ukážeme v této části na konkrétních datech výzkumu, v němž byly šetřeny příčiny zhoršování manželských vztahů. Příklad má ukázat možnosti, které interpretačně před výzkumníkem vyvstanou při aplikaci metody, a způsob, jakým je nová odkrytá informace prezentována. Ukázka má ovšem jen ilustrační charakter, neboť výzkumné závěry nelze dělat z této jedné tabulky a z dat vytržených z dalších souvislostí. Přesto věříme, že příklad má dost názorných rysů pro obecnější pochopení modelu a jeho výsledků.

Tabulka 2 obsahuje profily průměrných skóre, jimiž respondenti ve čtyřech typech manželství (*harmonické, stabilizované, narušené, rozvrácené*) označovali (na stupnici 1–7), v jaké míře se uplatnily jednotlivé důvody při rozhodnutí o vstupu do manželství.

Před interpretací údajů je nutno si uvědomit, že výpověď o důvodech je založena na skutečnosti, ta však je zkreslena subjektivními hledisky v závislosti na stavu manželství (pocit rozladění, trpkosti, současné disharmonie), na časovém odstupu, vývoji osobnosti a současných životních okolností i na struktuře skupin.

Podrobná analýza vztahů tabulky en face je velmi obtížná (a tato obtížnost roste s rozměry tabulky), ne-li nemožná. Systematické posuny v řádcích i ve sloupcích znemožňují určit přímo, které položky vykazují typické hodnoty pro které typy. Sloupcové⁽²⁾ efekty řadí položky podle stupně uváděné intenzity důvodů, znaménko dělí důvody na část s nadprůměrným ohodnocením (+) a část s podprůměrným postavením na škále (–). Nejvýše je postaven „citový vztah“ (1.56), nejnižší intenzitu má důvod nástupu na základní vojenskou službu (–.52). Řádkové efekty ukazují na celkovou úroveň uvádění důvodů v jednotlivých typech: celková intenzita je vysoká u prvních dvou typů a poté postupně klesá (.37, .40, –.22, –.54). To značí,

(2) Sloupcové průměry zde vznikají jako *proste průměry průměrů*. Charakterizují tak střední skóre položky, který je očištěn od vlivu velikostí výběrů ve skupinách, odhlíží od velikosti skupin. Tyto sloupcové průměry se liší numericky od průměrných skóre položek, spočtených pro celý spojený soubor dat.

Tabulka 2. Důvody vstupu do manželství — průměrné hodnoty vyjadřují, jak silně byly jednotlivé důvody uváděny ve čtyřech typech manželství (stupnice intenzity: 1 = „u nás se neprojevovalo“, 7 = „byl to velmi silný důvod“)

typ manželství	důvody vstupu do manželství									řádkové průměry	řádkové efekty
	A citový vztah	B majetková situace partnera	C společenské postavení partnera	D těho- tenství	E možnost získat byt	F naléhání partnera	G naléhání rodičů	H osobnost partnera	I základní vojenská služba		
1. harmonické	5.338	1.946	1.792	3.434	2.303	1.609	1.946	3.296	1.946	2.623	0.366
2. stabilizované	4.595	1.946	2.079	3.807	2.639	2.303	1.792	3.136	1.609	2.656	0.399
3. narušené	3.258	1.946	1.609	2.398	1.792	1.792	1.792	1.946	1.792	2.036	—0.221
4. rozvrácené	2.079	1.609	1.609	2.079	1.609	1.609	1.609	1.609	1.609	1.714	—0.543
sloupcové průměry	3.818	1.862	1.773	2.930	2.086	1.828	1.785	2.497	1.739	celkový průměr 2.257	
* sloupcové efekty	1.560	—0.396	—0.485	0.672	—0.172	—0.429	—0.473	0.239	—0.518		

Zdroj dat: J. Vítek a kol.: Výzkumné šetření manželských dvojic. Olomouc, UP 1982.

Poznámka: Data jsou získána ze šetření speciální populace, a proto nemohou být zobecněna na populaci ČSSR — jsou uváděna jen pro ilustraci metody.

že v horších typech se postupně snižuje přiznání důležitosti důvodů v celku. Interpretace marginálních efektů má svoji samostatnou důležitost, ale neříká nic o vztazích mezi objekty a položkami (i když v souvislosti s další analýzou může vnést do interpretace důležitou informaci pro vysvětlení nových poznatků). V dalším postupu chceme odhalit, do jaké míry jsou pro jednotlivé typy manželství charakteristické jednotlivé důvody či jejich absence. Model pro takovou analýzu můžeme zapsat podle (2):

- (8) stupeň uvádění důvodu = stupeň charakteristický pro celý soubor
 + odchylka daná typem manželství
 + odchylka daná obsahem položky
 + specifická odchylka konkrétního důvodu pro daný typ manželství

Odtud je vidět, jak se spočítají interakční členy (viz též (6)):

$$(9) \quad y_{rs} = x_{rs} - x_r - x_s + \bar{x} \\ = x_{rs} - y_r - y_s - \bar{x}$$

Interakční členy jsou uvedeny v tab. 3. V závorce jsou tam také uvedeny hodnoty profilů při hypotéze marginální aditivní, tj. při absenci nenáhodných vychýlení interakčních členů, a tedy při neexistenci specifických projevů položek pro typy. Získáme je podle vzorce:

$$(10) \quad \text{očekávaný průměr} = \text{celkový průměr} + \text{řádkový efekt} + \text{sloupcový efekt} \\ \hat{x}_{rs} = \bar{x} + y_r + y_s$$

Testování významnosti interakcí, tj. přijetí alternativní hypotézy k (10) se provádí různými technikami podle povahy dat, většinou je však obtížné, a tak model obvykle zavrhneme, mají-li odchylky nějakou *interpretovatelnou logiku*.⁽³⁾

Vysoké hodnoty reziduí s kladným znaménkem ukazují na charakteristické položky pro danou skupinu, záporné vysoké hodnoty ukazují na typickou absenci položky. Tak například „*citový vztah*“ je výrazný pro *harmonický* typ, zatímco chybí u *rozevřeného* typu.⁽⁴⁾

V interakční tabulce mají vztahy mezi členy specifickou interpretaci. Například v 1. sloupci jsou kategorie podle hodnot uspořádány, mezi kvalitou manželství a důvodem „*citový vztah*“ je zřetelný trend (dokonce lineárního charakteru, postupně rozdíl reziduí jsou .78, .72, .86, tedy přibližně rovnoměrné). Podobná závislost je vidět u „*osobnosti partnera*“, opačná u „*společenské situace partnera*“ a „*naléhání partnera*“. U „*majetkové situace partnera*“ vidíme trend, který nerozlišuje pořadí, ale dvě skupiny. U jiných důvodů jsou vztahy hodnot reziduí jiného typu. Podobně můžeme hodnotit řádky co do pořadí hodnot, jejich znamének, výrazných odchylek a podobně.

Tabulka tak přináší zcela novou informaci o roli vstupu do manželství v procesu jeho dalšího vývoje. Je tu i další informace: v marginální analýze se ukázalo, že „*základní vojenská služba*“ je důvod, který je uváděn celkem s nejmenší intenzitou, v interakcích má však závažnou roli, vyskytuje se jako silný faktor uváděný v rozvrácených a narušených dvojicích v kontrastu se stabilizovanými. Nejde tedy o častý důvod, ale tam, kde nastává, působí intenzivně nebo existuje tendence svěst zhoršování vztahu na tento důvod. (Přitom je nutno si uvědomit, že název důvodu je jen

(3) Poznamenejme k tomu, že celý postup je svou povahou především *explorační*, závisí tedy na postupu interpretace a na tom, jak stavíme otázky pro další analýzu. Výsledky metody vedou většinou, kromě silných a průkazných faktorů, k formulování hypotéz a dalších výzkumných otázek.

(4) Interakční členy jsou měřeny na stejné stupnici jako původní hodnoty. Např. 0.5 značí vychýlení o půl stupně původní stupnice znaku.

Tabulka 3. Tvary profilů
 Tabulka reziduí pro profily tabulky 2 obsahuje hodnoty $y(j, k) = x(j, k) - x(j, *) - x(*, k) + x(*, *)$
 V závorce jsou uvedeny predikční hodnoty $\bar{x}(j, k) = x(j, *) + x(*, k) - x(*, *)$

typ manželství	důvody vstupu do manželství								
	A citový vztah	B majetková situace partnera	C společenská situace partnera	D těhotenství	E možnost získat byt	F naléhání partnera	G naléhání rodičů	H osobnost partnera	I základní vojenská služba
1. harmonické	1.154 (4.18)	-0.282 (2.23)	-0.347 (2.14)	0.139 (3.30)	-0.149 (2.45)	-0.585 (2.19)	-0.205 (2.15)	0.433 (2.86)	-0.159 (2.11)
2. stabilizované	0.379 (4.22)	-0.315 (2.26)	-0.092 (2.17)	0.478 (3.33)	0.155 (2.48)	0.075 (2.23)	-0.392 (2.18)	0.240 (2.90)	-0.529 (2.14)
3. narušené	-0.338 (3.60)	0.305 (1.64)	0.058 (1.55)	-0.310 (2.71)	-0.073 (1.86)	0.185 (1.61)	0.228 (1.56)	-0.329 (2.28)	0.274 (1.52)
4. rozvrácené	-1.195 (3.27)	0.291 (1.32)	0.380 (1.23)	-0.307 (2.39)	0.067 (1.54)	0.325 (1.28)	0.368 (1.24)	-0.344 (1.95)	0.414 (1.20)

nositelem určité komplexní situace, skrytých podstatných důvodů a nepravých korelací). Interakční hodnoty charakterizují skupiny vzhledem k jednotlivým položkám a stávají se tak škálovými hodnotami a charakteristikami intenzity působení položky na skupinu nebo (podle významu) intenzity projevu skupiny v položce. Každý sloupec i každý řádek proto tvoří samostatnou škálu vlivů, která charakterizuje skupinu nebo položku v jejich zvláštnostech oproti ostatním.

Pohled na interakční tabulku odhaluje působení řádkových či sloupcových vstupů jako samostatných faktorů. Podobná pořadí a první interpretace hodnot v řádcích a sloupcích však nabízejí hypotézu, že existují nějaké společné příčiny, které působí ve více sloupcích a řádcích, a to s různou intenzitou i směrem a které se vzájemně skládají podle rovnice (7). Abychom mohli takové společné latentní faktory identifikovat, předpokládáme u nich platnost některých vlastností (jejich matematická formulace bude provedena v další části).⁽⁵⁾

- faktory působí na sobě *nezávisle a jejich vlivy se sčítají*;
- každý faktor působí *s různou intenzitou a směrem na jednotlivé řádky a na jednotlivé sloupce*;
- každý faktor působí *v jednotlivých polích jako součin jeho řádkové a sloupcové intenzity*, tedy úměrně vlivu na řádek i vlivu na sloupec;
- každý faktor má *nulovou očekávanou hodnotu*.

Faktory tak ovlivňují řádky a sloupce a jejich prostřednictvím úměrně i interakční příspěvek. V tabulce 4 uvádíme pro ilustraci nenulová charakteristická čísla a charakteristické vektory matic YY' a $Y'Y$. Charakteristická čísla jsou hodnotami faktorových rozptylů, složky k nim příslušných charakteristických vektorů jsou hodnoty standardizovaných faktorů pro řádky a sloupce.

Faktorové koeficienty, které ukazují číselně vyjádřený vliv každého faktoru na řádky i sloupce, jsou uvedeny v tabulce 5 (vznikají z tabulky 4b, v níž postupně násobíme sloupce $\lambda_m^{\frac{1}{2}}$).

Interpretace faktorů, tj. latentních, přímo nezjišťovaných proměnných, které se projevují souhrnem interakčních efektů a reprezentují jejich společnou variabilitu, se provádí:

- a) *na základě významu a obsahu řádků a sloupců tabulky a možných mechanismů jejich vzájemného vztahu* (směr kauzálního působení, druh ovlivňování, existence latentního procesu, který působí přechod objektů do jiných typů, řadícího sloupce či řádky do vývojové řady, seskupování objektů do homogenních skupin, seskupování proměnných do bloků sumovaných indexů atp.);
- b) *podle hodnot faktorových koeficientů* (viz tab. 5): vysoké hodnoty označují silný vliv, resp. projev faktoru pro řádek nebo sloupec (či jeho prostřednictvím), znaménka koeficientů určují souběžnost či protichůdnost směru působení;
- c) *podle hodnot součinů řádkových a sloupcových koeficientů faktoru pro každé pole*: ty vyjadřují konkrétní hodnotu a směr příspěvku faktoru v každém poli, tj. jeho podíl na vzniku každého interakčního členu, kladnou nebo zápornou souvztažnost řádku a sloupce danou faktorem;
- d) *na základě nekorelovanosti odhalených faktorů*, které působí nezávisle na sobě, a tudíž jsou nezávislé i obsahově, nepřekrývají se ve své skladbě nezávislých drobných příčin;

(5) Verbální popis faktorů je možno upřesnit v případě, že je předem přijat směr a způsob kauzálního působení. Formulace „faktor působí na“ může být podle situace nahrazena výrazem „řádkové (sloupcové) vlastnosti působí na faktor“, „faktor působí prostřednictvím řádků (sloupců)“, „faktor působí souběžně s řádky (sloupci)“, „faktor se projevuje v řádcích (sloupcích)“ ap.

Tabulka 4. Charakteristická čísla a vektory matic YY^T a Y^TY

a) Charakteristická čísla a odvozené koeficienty determinace a diferenciacce

faktor	charakteristické číslo λ	λ	koeficient determinace faktoru	kumulativní koeficient determinace	koeficient diferenciacce škály
1	5.3091	2.3042	.879	.879	.679
2	0.6465	0.8041	.107	.986	.237
3	0.0818	0,2859	.014	1.000	.084
součet	6.0374	3.3942	1.000	—	1.000

b) Charakteristické vektory matice YY^T — normalizované faktorové koeficienty pro řádky

řádky	vektor 1	vektor 2	vektor 3
1. harmonické	.612	—,519	—,326
2. stabilizované	.336	.784	.149
3. narušené	—,296	—,333	.743
4. rozvrácené	—,652	.068	—,566

c) Charakteristické vektory matice Y^TY — normalizované faktorové koeficienty pro sloupce

sloupce	vektor 1	vektor 2	vektor 3
A citový vztah	.743	—,338	.370
B majetková situace partnera	—,242	—,227	.374
C společenská situace partnera	—,221	.143	—,255
D těhotenství	.233	.479	—,107
E možnost získat byt	—,027	.283	—,071
F naléhání partnera	—,260	.402	.543
G naléhání rodičů	—,245	—,313	—,107
H osobnost partnera	.290	.061	—,543
I základní vojenská služba	—,272	—,491	—,202

- Poznámky: 1. Nenulová charakteristická čísla matic YY^T a Y^TY jsou stejná.
 2. Počet nenulových charakteristických čísel je vždy menší než menší z rozměru tabulky Y .
 3. Charakteristické vektory jsou normalizovány tak, aby součet čtverců koeficientů (jak pro řádky-viz tab. b), tak pro sloupce — viz tab. c) byl roven jedné.

e) na základě korelace faktoru a interakčních odchylek — čím vyšší je koeficient determinace faktoru, tím silněji je jím interakce ovlivňována a tím důležitější, ale i výraznější faktor je; při nepatrných koeficientech determinace faktory do interpretace nezahrnujeme a shrnujeme je do náhodných vlivů.

O problémovosti kauzální interpretace faktorů identifikovaných v tab. 5 jsme se již zmínili. Je velmi obtížné apriori odhadnout, zda v datech převládá výpověď objektivní, či zda se prosadí současný pohled respondenta (subjektivní zkreslení), či jaký je jejich vzájemný podíl. Předpokládáme, že jsou oba smíšeny, že faktory jsou vytvářeny skutečnými důvody, že jsou však zkresleny (zesíleny nebo zeslabeny) současným stavem manželského svazku. Proto k modelu přistoupíme jako k vyjádření vzájemné souvztažnosti a korelovanosti typů a položek.

Tabulka 5. Faktorové řešení metodou hlavních komponent
(Tabulky obsahují faktorové koeficienty)

typy manželství	řádkové efekty	faktorové koeficienty pro řádky		
		faktor 1	faktor 2	faktor 3
1. harmonické	0.37	0.929	—0.466	—0.174
2. stabilizované	0.40	0.511	0.703	0.080
3. narušené	—0.22	—0.450	—0.299	0.397
4. rozvrácené	—0.54	—0.990	0.061	—0.303

důvody uzavření sňatku	sloupcové efekty	faktorové koeficienty pro sloupce		
		faktor 1	faktor 2	faktor 3
A citový vztah	1.56	1.128	—0.303	0.198
B majetková situace partnera	—0.40	—0.368	—0.203	0.200
C společenská situace partnera	—0.48	—0.335	0.128	—0.136
D těhotenství	0.67	0.354	0.430	—0.057
E možnost získat byt	—0.17	—0.040	0.254	—0.038
F naléhání partnera	—0.43	—0.394	0.361	0.290
G naléhání rodičů	—0.47	—0.372	—0.281	—0.057
H osobnost partnera	0.24	0.440	0.055	—0.291
I základní vojenská služba	—0.52	—0.412	—0.440	—0.108

koeficient determinace (v %)	87.9	10.7	1.4
kumulární koeficient determinace (v %)	87.9	98.6	100.0
koeficienty diferenciací škál (v %)	67.9	23.7	8.4

Do vztahového modelu přijímáme takové faktory, které mají jednak dostatečně vysoký podíl na vysvětlení interakční variability a které jsme schopni interpretovat, tj. ty, jejichž koeficienty nabízejí smysluplné vysvětlení.

Faktor 1: Podle koeficientů sloupcového vlivu sdružuje s velmi silným podílem „citový vztah“ (1.29), dále ve stejném směru „těhotenství“ (.35) a „osobnost partnera“ (.44), zatímco v opačném směru proti těmto důvodům stojí „majetková (— .37) a společenská (— .34) situace partnera“, „naléhání partnera (— .39) a rodičů“ (— .37) a „základní vojenská služba“ (— .41). Současně řadí faktor skupiny podle stupně harmoničnosti a rozvrácenosti (.93, .51, —.45, —.99). Působení faktoru je velmi silné: vysvětluje 88 % variability interakčních členů. Položky profilu se podle koeficientů dělí na dvě skupiny: s harmonickým manželstvím korespondují zvýšený stupeň vzájemné náklonnosti partnerů, vyjádření aktivního podílu respondenta. Pro narušení je specifické hledání důvodů mimo respondenta vnější tlak (svádění viny na někoho či něco jiného). Kvalita manželského vztahu diagnostikovaná ze zcela jiných symptomů je tu silně korelována s kvalitou vztahu manželů a jejich vzájemné náklonnosti projevených ve výpovědích o důvodech vstupu do manželství. Výsledek tohoto silného faktoru⁽⁶⁾ je trojí:

(6) Připomeňme však, že data jsou pouze ilustrační s pouhým náznakem interpretace. Pro praxi a teoretické závěry by analýza byla značně složitější, zahrnovala by úvahy o vlivech dalších proměnných: délka trvání manželství, věk respondenta a jeho pohlaví, věk při vstupu do man-

- a) vede k hypotéze, že proces zhoršování manželského vztahu je částečně predikovatelný (ve statistickém smyslu⁷⁾, už na samém počátku, přičemž citový vztah a vzájemná náklonnost jsou pro vznik a trvání harmonie rozhodující, majetkové a společenské aspekty manželství neudržují;
- b) validizuje použité diagnostické pravidlo pro určování typu manželství — obsahová validita;
- c) umožňuje konstrukci součtového indexu, kterým můžeme na základě výpovědi o důvodech diagnostikovat stav manželství (a ve spojení s dalšími fakty popřípadě i index složitější, který by predikoval trvání a úspěšnost manželství).

Součtový index, který odráží v důvodech současný stupeň vztahu. lze napsat pomocí faktorových vah položek (označíme je zde jako f_s):

$$(11) \quad T = \sum_s^S f_s I_s,$$

kde I_s je indikátor výskytu důvodu nebo nějaký skór jeho intenzity. Pomocí tohoto indexu můžeme také posuzovat vliv různých kombinací důvodů, například současný výskyt důvodů A, D, G má skór: $1.13 + .35 - .37 = 1.11$, současný výskyt B, C, D má skór: $-.37 - .34 + .35 = -.36$ ap. Tím můžeme získat i určitou představu o kompenzacích různých položek (v některých případech i relativně přesná empirická rozhodovací a predikční pravidla). Hodnoty porovnáváme jednak mezi sebou, jednak vzhledem k nule a k dosažitelným hranicím (v našem případě jsou to 1.92 a -1.92). Ve většině sociologických analýz však musíme dávat pozor na absolutizaci takových numerických výsledků.

U faktoru se vyskytuje zajímavá okolnost: znaménka řádkových i sloupcových koeficientů si odpovídají se znaménky marginálních efektů a i hodnoty si v hrubých relacích odpovídají. To naznačuje hypotézu, že v marginálních efektech je už skryta značná (ne-li převážná) část subjektivního zkreslení objektivitu výpovědi o skutečných důvodech vstupu do manželství, a tak je tento vliv již z interakcí z části odečten a interakce obsahují objektivnější obsah než původní profily.⁽⁸⁾

Tabulka 6 obsahuje příspěvky faktoru 1 k jednotlivým interakčním hodnotám v polích⁽⁹⁾ a zbytkové, ještě nevysvětlené hodnoty interakcí, které budou tvořeny dalšími faktory.

Faktor 2 působí nezávisle na prvním faktoru a odlišuje stabilizovaný typ od obou sousedních. To ukazuje na specifickou roli typu mezi ostatními a jeho odchýlení od čistého uspořádání. U tohoto typu manželství se (v části jeho typové třídy) projevují

želství, shoda manželů při výpovědích, věková a sociální struktura páru, počet dětí a popřípadě i osobnostní charakteristiky, hodnotové hierarchie a další sociologické a psychologické proměnné. Zároveň je třeba při interpretaci počítat s tím, že se některé položky mohou, vzhledem k souběžnosti působení a projevů a k nepravým korelacím, sloučit do jednoho faktoru empiricky, aniž by k sobě obsahově patřily. Informace v tabulce je omezená a nemusí umožnit identifikaci faktorů v čisté podobě. Souběžně působící (i obsahově odlišné) faktory tvoří ve výsledcích metody empirické koreláty, které nejsou matematicky separovatelné. (Toto omezení je společné pro všechny modely faktorové analýzy, latentních struktur, škálování ap.). O to větší význam má využití vnější informace, spojování výsledků různých analýz, analýzy na dílčích souborech a komparace různých šetření, která používají různé soubory proměnných, otázek, operacionalizací.

(7) To znamená, že nelze určit přesný vývoj, ale statisticky pro určitý komplex okolností lze očekávat typický (v modálním či průměrovém smyslu) výsledek.

(8) V případě, že by se ukázala praktická využitelnost faktorů pro jakékoli diagnostické či predikční cíle, vysoká závažnost hypotézy by vyžadovala prověření a další upřesnění modelu.

(9) Ve sloupci ALFA jsou uvedeny řádkové hodnoty standardizovaného faktoru, v řádku BETA sloupcové hodnoty standardizovaného faktoru, $\sqrt{\lambda_1}$ je konstanta úměrnosti ve vztahu: $\text{příspěvek faktoru v poli} = \sqrt{\lambda_1} \times ALFA \times BETA$. $\lambda_1 = 5.3091$ je součet čtverců všech příspěvků faktoru 1 v interakční tabulce.

Tabulka 6. Příspěvky faktoru 1 k hodnotám polí tabulky Y
(V závorce je hodnota rezidua po odečtení vlivu 1. faktoru)

typ manželství	důvody uzavření sňatku									ALFA	podíl řádku na působení faktoru 1 (v %)
	A citový vztah	B majetková situace partnera	C společenská situace partnera	D těho- tenství	E možnost získat byt	F naléhání partnera	G naléhání rodičů	H osobnost partnera	I základní vojenská služba		
1. harmonické	1.047 (0.11)	-0.324 (0.06)	-0.311 (-0.04)	0.329 (-0.19)	-0.038 (-0.11)	-0.366 (-0.22)	-0.345 (0.14)	0.408 (0.03)	-0.383 (0.22)	0.612	37.4
2. stabilizované	0.576 (-0.20)	-0.188 (-0.13)	-0.171 (0.08)	0.181 (0.30)	-0.021 (0.18)	-0.201 (0.28)	-0.190 (-0.20)	0.224 (0.02)	-0.211 (-0.32)	0.336	11.3
3. narušené	-0.507 (0.17)	0.165 (0.14)	0.151 (-0.09)	-0.159 (-0.15)	0.018 (-0.09)	0.177 (0.01)	0.167 (0.06)	0.198 (-0.13)	0.185 (0.09)	-0.296	8.8
4. rozvrácené	-1.116 (-0.08)	0.364 (-0.07)	0.331 (0.05)	-0.350 (0.04)	0.040 (0.03)	0.390 (-0.07)	0.368 (0.00)	-0.435 (0.09)	0.408 (0.01)	-0.652	42.5
BETA	0.743	-0.242	-0.221	0.233	-0.027	-0.260	-0.245	0.290	-0.272	součet řádky = 100.0 $\sqrt{\lambda_1} = 2.3042$	
podíl sloupce na působení faktoru 1 (v %)	55.2	5.9	4.9	5.4	0.1	6.8	6.0	8.4	7.4	součet sloupce = 100.1	

Poznámky: 1. Hodnoty reziduí v závorkách budou vysvětlovány dalšími faktory (tj. druhým a třetím)

2. Ve sloupci ALFA jsou hodnoty charakteristického vektoru 1 z tab. 3b

3. V řádku BETA jsou hodnoty charakteristického vektoru z tab. 3c

4. Příspěvek v poli vzniká jako součin $ALFA \times BETA \times \sqrt{\lambda}$ ($1.047 \doteq 0.612 \times 0.743 \times 2.3042$)

5. Hodnoty reziduí vznikly jako $Y - ALFA \times BETA \times \sqrt{\lambda}$ ($0.11 = 1.154 - 1.047$)

dodatečné faktory stabilizace, které odpovídají racionálnímu pohledu (děti, byt) a potlačení či negaci původních emocí *a/nebo* se jedná o dvojice s dominantním partnerem respondenta. Faktor tak lze chápat jako *vliv racionálního postoje*. Vliv faktoru na rezidua je vidět v tabulce 7. V závorce jsou uvedeny zbytky interakcí po odečtení vlivu prvních dvou faktorů, které budou vysvětleny v našem případě už jen třetím faktorem (více jich pro tabulku 4×9 nalézt nelze). Podíl na vysvětlení interakcí je 11 %.

Faktor 3 má koeficient determinace 1.4 %, tedy velmi nízký, a ani jeho koeficienty nenaznačují výraznou odchylku, která — ač s malou vahou — by mohla indikovat nějaký důležitý vliv.

Podíly faktorů jsou sumarizovány v tab. 8, z níž je vidět, že 20.5 % celkové variability údajů vysvětlují řádkové efekty, 57.9 % je způsobeno hierarchizací položek profilu a 21.6 % připadá na interakční variabilitu. Ta je čerpána z 87.9 % prvním faktorem (jehož podíl na celkové varianci je tak 10.0 %), 10.7 % druhým faktorem (2.3 % z celku) a zbytek připadá na neinterpretovaný faktor třetí.⁽¹⁰⁾

Řešení úlohy není jednoznačné. *Kanonický rozklad* (tzv. *hlavní řešení*) vede na jeden z možných faktorových rozkladů, který má tu vlastnost, že první faktor vyčerpává co největší možný stupeň variability, druhý opět maximálním způsobem zbytek atd. Další možná řešení hledáme transformací nazývanou *rotaci*. Pokoušíme se nalézt takové faktory, které by se co nejvíce přimykaly k jednotlivým objektům a vytvářely homogenní, ale co nejkontrastněji odlišené skupiny objektů, s tím, že každý objekt je ovlivňován co nejmenším počtem faktorů. Jinou možností je hledat faktory rozdělující položky na skupiny, které se pokud možno nepřekrývají a jejichž působení na objekty je co nejshodnější; každá položka reprezentuje pokud možno jen jeden, či co nejmenší počet faktorů.⁽¹¹⁾ Taková řešení umožňují často zřetelnější a sociologicky zajímavější interpretace. V tabulce 8 je uvedeno rotované řešení metodou Varimax, v němž byla hledána prostá struktura řádkových koeficientů, tedy takové faktory, které co nejzřetelněji rozlišují mezi typy manželství.

Transformační matice $\mathbf{T} = ||t_{ij}||$ pro rotaci je:

$$\mathbf{T} = \begin{vmatrix} .8555 & .5177 \\ -.5177 & .8555 \end{vmatrix}$$

V porovnání s hlavním řešením se 1. faktor v uspořádání objektů neliší, skupina „harmonické“ je však výrazně vzdálena od ostatních a i druhý extrém se od svého souseda liší podstatně více, zatímco obě střední kategorie se k sobě relativně přiblížily. Tuto změnu působí položky „*citový vztah*“ a osobnost na jedné straně a tři důvody spojené s partnerem na druhé straně. Faktor má tak dva položkové protipóly: vzájemnost a vyzvednutí osobních předností partnera proti vnějškovým jednostranným okolnostem spojeným však na rozdíl od nerotovaného řešení pouze s partnerem.

Druhý faktor se změnil podstatně: staví nyní do kontrastu skupinu „stabilizované“ nikoli proti sousedům, ale proti oběma horším typům. Působí tu obdobně jako kladný vliv v prvním faktoru, doplněný o důvod těhotenství; proti stojí majetek, základní

(10) V některých případech vznikají řádkové *a/nebo* sloupcové efekty uměle, či nemají vhodnou interpretaci. Tehdy znamenají pouhou *korekci*, standardizaci dat, odečítají nežádoucí vlivy a do tabulky je nezahrnujeme. Aby byla tabulka informativní a měla smysl, do celkové variability přijmeme jen ty efekty, které můžeme interpretovat věcně vzhledem k problému.

(11) Těchto vlastností dosahujeme *rotací k tzv. prosté struktuře řádkových* (v prvním případě) *nebo sloupcových* (v druhém případě) faktorových koeficientů.

Tabulka 7. Příspěvky faktoru 2 k hodnotám polí tabulky Y
(V závorce je uvedena hodnota po odečtení vlivu 1. a 2. faktoru)

typ manželství	důvody uzavření sňatku									ALFA	podíl řádku na působení faktoru 2 (v %)
	A citový vztah	B majetková situace partnera	C společenská situace partnera	D těho- tenství	E možnost získat byt	F naléhání partnera	G naléhání rodičů	H osobnost partnera	I základní vojenská služba		
1. harmonické	0.141 (-0.03)	0.095 (-0.03)	-0.060 (0.02)	-0.200 (0.01)	-0.118 (-0.01)	-0.168 (-0.05)	0.131 (0.01)	-0.026 (0.05)	0.205 (0.02)	-0.519	27.0
2. stabilizované	-0.213 (0.02)	-0.143 (0.02)	0.090 (-0.01)	0.302 (0.00)	0.178 (0.00)	0.254 (0.02)	-0.197 (0.00)	0.039 (-0.02)	-0.309 (-0.01)	0.784	61.5
3. narušené	0.090 (0.08)	0.061 (0.08)	-0.038 (-0.05)	-0.128 (-0.02)	-0.076 (-0.02)	-0.108 (0.12)	0.084 (-0.02)	-0.016 (-0.12)	0.131 (-0.04)	-0.333	11.1
4. rozvrácené	-0.019 (-0.06)	-0.012 (-0.06)	0.008 (0.04)	0.026 (0.02)	0.016 (0.01)	0.022 (-0.09)	-0.017 (0.02)	0.003 (0.09)	-0.027 (0.03)	0.068	0.5
BETA	-0.338	-0.227	0.143	0.479	0.283	0.402	-0.313	0.061	-0.491	součet řádky = 100.1 $\sqrt{2} = 0.80$	
podíl sloupce na působení faktoru 2 (v %)	11.4	5.1	2.0	23.0	8.0	16.2	9.8	0.4	24.1	součet sloupce = 100.0	

Poznámky: 1. Hodnoty v závorkách budou vysvětlovány dalšími faktory, v tomto případě už jen posledním, třetím faktorem. Ukazují proto i příspěvek třetího faktoru.
2. Viz poznámky u tab. 5.

Obr. Korespondenční mapa typologie manželství a důvodů uzavření sňatku (vodorovná osa = 1. faktor, svislá osa = 2. faktor)

TYPOLOGIE MANŽELSTVÍ 1 harmonické 2 stabilizované 3 narušené 4 rozvrácené E F	② D H A ①
C B G ③ I ④	DŮVODY UZAVŘENÍ SŇATKU A citový vztah B majetková a finanční situace partnera C společenské postavení partnera D těhotenství E možnost dostat byt F naléhání partnera G naléhání rodičů H osobní vlastnosti partnera I nástup do základní vojenské služby

vojenská služba a naléhání rodičů.⁽¹²⁾ V převaze se tu jeví důvody, které mají společného jmenovatele v uspěchaném rozhodnutí.

Rotované faktory mají koeficienty determinace $R_1^{*2} = .60$, $R_2^{*2} = .24$ a příspěvek společného působení je $\Delta_{12} = .15$; koeficient determinace obou faktorů současně je

$$R_{1,2}^{*2} = .60 + .24 + .15 = .99$$

Korelace obou rotovaných faktorů je $r_{12}^* = .19$.

Korelace mezi rotovanými a nerotovanými faktory $r(F_i, F_j^*)$:

	F_1^*	F_2^*
F_1	.8867	.5120
F_2	.1133	.4880

(Výpočty byly provedeny podle vzorců v dodatku.)

Důležitým aspektem metody LINDA-K je možnost grafického zobrazení řádků a sloupců pomocí korespondenční mapy. Tato mapa reprezentuje společný prostor

(12) V tabulkách faktorových koeficientů si můžeme všimnout dvojí možné interpretace nulových hodnot. U objektů: střed škály, vřazení objektu do pořadí na neutrální, průměrnou pozici, nebo absenci vlivu faktoru na objekt, objekt je mimo sféru faktorového vlivu. U položek: vynechání položky z faktoru, její nulový podíl na faktoru nebo role průměrné položky s neutrálním vlivem.

Tabulka 8. Podíly variability pro skupiny manželství, složky profilu důvodů, interakce a jejich faktory — shrnutí

zdroj variability	součet čtverců	podíl na celkové variabilitě	podíl faktorů na reziduích
celková variabilita údajů tabulky	27.9345	100 %	—
typy manželství	5.7357	20.5 %	100 %
profil důvodů	16.1620	57.9 %	
interakce typů a profilu	6.0369	21.6 %	
faktor 1	5.3091	19.0 %	87.9 %
faktor 2	0.6465	2.3 %	10.7 %
faktor 3	0.0818	0.3 %	1.4 %

- Poznámky: 1. Celková variabilita tabulky se počítá jako součet čtverců všech hodnot (tab. 1) od jejich průměru (= 2.257)
2. Řádkové rozdíly (typy manželství) charakterizuje součet čtverců řádkových efektů (obdobně sloupce)
3. Reziduální variabilita je součet čtverců interakčních členů (z tab. 2)
4. Součet čtverců pro jednotlivé faktory se rovná jejich charakteristickým číslům

objektů a položek, které jsou znázorněny body a jejich vzdálenosti ukazují na vztahy mezi nimi. Na obrázku 1 jsou znázorněny body objektů i body položek tak, že faktorové koeficienty slouží jako souřadnice, vodorovná osa značí faktor 1, svislá faktor 2. Vztahy posuzujeme podle vzdáleností bodů. Blízké body značí podobnost objektů, nebo podobnost působení položek na objekty, nebo typičnost položky pro objekt. Velká vzdálenost vyjadřuje nepodobnost, protichůdnost působení či typickou absenci položky pro objekt. V obrázku je vidět, že umístění bodů je výrazně plošné, že 1. škála nestačí k vysvětlení vztahů mezi skupinami, ale že je třeba uvažovat výrazný odklon typu „stabilizované“. Pro „harmonické“ je zcela typický důvod A = citový vztah, který působí oddělení bodu A od ostatních. Pro „stabilizované“ je nejtypičtější bodem důvod D = těhotenství, pro „narušené“ jsou to B, G, H. „Rozvrácené“ svůj jednoznačně typický důvod nemá. Důvod H je v pozici mezi 1 a 2 a tudíž se jeho působení přibližně rovnoměrně štěpí, E má polohu středu a tak zhruba stejný vliv na všechny položky. Průměty bodů na osy poskytují jednorozměrné škály.

Veškerá interpretace je ovšem vázána na ty dva faktory, které jsou v obrázku reprezentovány. V našem případě je to — až na zanedbatelný třetí faktor — plná korespondence profilové tabulky. Je-li však faktorů více, umožňuje tento způsob posuzovat pouze vlivy každých dvou faktorů zvláště (připraví se obrázek pro každou dvojici faktorů).⁽¹³⁾

Interpretace obrázku umožňuje rychlý celostní pohled na vliv dvojice faktorů. Vyčteme z něj jak podobnosti, tak korespondence. Hodnotíme ovšem nejen dvojice bodů, ale i jejich seskupení, tvary naznačených křivek a to, co v obrázku chybí (např. typické položky pro některý objekt). Rozmístění bodů v obrázku je určeno souřadnicemi obou os (jednotlivých škál). Váha každé osy pro určení polohy může být charakterizována koeficientem diferenciace škály, který zavádíme jako podíl součtu čtverců jejich hodnot k celkovému součtu čtverců vzdáleností bodů od počátku. Koeficient diferenciace se uvádí zpravidla vzhledem ke K-rozměrné škále, kterou uvažujeme, i když obrázky vyjadřují pouze dílčí dvourozměrné škály.

(13) Je též možno použít některých vícerozměrných způsobů grafického zobrazení — ty však autorům nejsou zatím k dispozici.

3. Popis matematického modelu

Formalizaci modelu provedeme stručně ve třech možných variantách: faktorový model, model komparační báze a model korespondenční analýzy. V další části bude na příkladě uvedeno ještě použití modelu kvantifikace.

A) Dekompoziční vlastnost matice $\mathbf{Y}$ společná všem modelům

Matice reziduí (viz rovnice (9)) o rozměrech $R \times S$, v níž řádky odpovídají objektům a sloupce profilovým položkám (složkám), může být vyjádřena ve formě kanonického rozkladu:

$$(12) \quad \mathbf{Y} = \mathbf{ADB}' = \mathbf{PQ}' = \sum_m^M d_m \mathbf{A}_m \mathbf{B}_m' = \sum_m^M \mathbf{P}_m \mathbf{Q}_m'.$$

kde $\mathbf{A} = \|a_{rm}\|$ je matice $R \times M$, jejíž sloupce $\mathbf{A}_m$ jsou charakteristické vektory matice $\mathbf{Y}\mathbf{Y}'$;

$\mathbf{B} = \|b_{sm}\|$ je matice $S \times M$, jejíž sloupce $\mathbf{B}_m$ jsou charakteristické vektory matice $\mathbf{Y}'\mathbf{Y}$;

$\mathbf{D} = \text{diag}\{d_m\}$ je diagonální matice, jejíž prvky jsou odmocniny nenulových charakteristických čísel matice $\mathbf{Y}\mathbf{Y}'$ resp. $\mathbf{Y}'\mathbf{Y}$;

M = počet nenulových charakteristických čísel matice $\mathbf{Y}\mathbf{Y}'$ resp. $\mathbf{Y}'\mathbf{Y}$;

$$\mathbf{P} = \mathbf{AD}^{\frac{1}{2}}, \quad \mathbf{Q} = \mathbf{BD}^{\frac{1}{2}}, \quad p_{rm} = \sqrt{d_m} a_{rm}, \quad q_{sm} = \sqrt{d_m} b_{sm}$$

Vyjádřeno pomocí součtů

$$(13) \quad y_{rs} = \sum_m^M d_m a_{rm} b_{sm} = \sum_m^M p_{rm} q_{sm} \quad \text{pro všechna } (r, s)$$

Dekompozice matice má dále tento význam: vezmeme-li postupně dvojice matic $(\mathbf{A}^k, \mathbf{B}^k)$, $k = 1, 2, \dots, M$, kde matice $\mathbf{A}^k, \mathbf{B}^k$ jsou tvořeny vždy prvními k sloupci matic $\mathbf{A}$ a $\mathbf{B}$ a $\mathbf{D}^k$ je vytvořena prvními k řádky a sloupci matice $\mathbf{D}$, platí, že součiny $\mathbf{A}^k \mathbf{D}^k \mathbf{B}^k$ jsou *postupně nejlepších aproximací* matice $\mathbf{Y}$ pomocí matic hodnot 1, 2, ..., $k, \dots, M$ ve smyslu metody nejmenších čtverců, tj.

$$\sum_r^R \sum_s^S (y_{rs} - \mathbf{a}_r^k \mathbf{D}^k \mathbf{b}_s^k)^2 = \min, \quad \mathbf{a}_r^k = r\text{-tý řádek } \mathbf{A}^k, \quad \mathbf{b}_s^k = s\text{-tý řádek } \mathbf{B}^k$$

(Eckart-Youngova věta).

Tyto vlastnosti jsou základem všech modelů. Existuje několik jejich různých vyjádření, která mají odlišné interpretace pro datové matice. (A též odlišné interpretace pro různé typy dat uspořádané v matici $\mathbf{Y}$).

B) Model diferenčních faktorů

vychází z toho, že po odečtení řádkového vlivu (faktoru objektů vyjádřeného pomocí řádkových efektů) a po odečtení sloupcového vlivu (faktoru položek vyjádřeného pomocí sloupcových efektů) z profilové tabulky (matice $\mathbf{X}$, viz (6)) se rozkládají rezidua y_{rs} (uspořádaná do matice $\mathbf{Y}$) podle rovnice (13).

Předpokládáme existenci M interakčních faktorů $\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_M$ které jsou navzájem nezávislé a mají nulovou střední hodnotu. Tyto faktory však působí jako součin nezávislých náhodných veličin, $\mathbf{F}_m = \mathbf{U}_m \mathbf{W}_m$. $\mathbf{U}_m$ působí prostřednictvím řádků (ovlivňuje objekty), $\mathbf{W}_m$ působí prostřednictvím sloupců (vliv položek). Vznik interakčního efektu je řízen rovnicí:

$$(14) \quad y_{rs} = \sum_m^M \mathbf{F}_m(r, s) = \sum_m^M \mathbf{U}_m(r) \cdot \mathbf{W}_m(s); \quad \mathbf{Y} = \sum \mathbf{F}_m = \sum \mathbf{U}_m \mathbf{W}_m,$$

kde $F_m(r, s)$ je hodnota m -tého faktoru v poli (r, s) $U_m(r)$ je hodnota vlivu na r -tý objekt, $W_m(s)$ je hodnota působení s -té položky.

Faktory F_m můžeme nazvat též *diferenciálními faktory*.

Rovnice (14) je analogická ke vztahu (13), a tak prvky matic $\mathbf{P}$, $\mathbf{Q}$ můžeme chápat jako odhady hodnot $U_m(r)$, $W_m(s)$ a jejich součiny jako odhady $F_m(r, s)$. Prvky matic $\mathbf{A}$, $\mathbf{B}$ v tom případě charakterizují hodnoty standardizovaných veličin $\bar{U}_m$, $\bar{W}_m$. Rozklad (12) má tyto vlastnosti (viz Gantmacher 1966 a Gollob 1968a, b):

$$\sum_r^R a_{rm} = \sum_s^S b_{sm} = \sum_r^R \sum_s^S a_{rm} b_{sm} = 0, \quad \sum_r^R a_{rm}^2 = \sum_s^S b_{sm}^2 = 1 \quad \text{pro všechna } m$$

$$\sum_r^R a_{rm} a_{rk} = \sum_s^S b_{sm} b_{sk} = \sum_r^R \sum_s^S (a_{rm} b_{sm}) (a_{rk} b_{sk}) = 0 \quad \text{pro všechna } k \neq m$$

Proto platí:

$$(15) \quad \mathbf{E} F_m = \mathbf{E} U_m = \mathbf{E} W_m = 0 \quad \text{pro } m = 1, 2, \dots, M$$

$$\text{var } F_m = d_m^2, \quad \text{var } U_m = \text{var } W_m = d_m, \quad \text{pro } m = 1, 2, \dots, M$$

Faktory jsou *nekorelované*:

$$(16) \quad \text{cov}(F_m, F_{m'}) = \text{cov}(U_m, U_{m'}) = \text{cov}(W_m, W_{m'}) = 0 \quad \text{pro } m \neq m'$$

Koeficient determinace faktoru F_m je dán vzorcem:

$$(17) \quad R^2(\mathbf{Y}, F_m) = R_m^2 = \frac{d_m^2}{\sum_m d_m^2} = \frac{(\sum_r p_{rm}^2)^2}{\sum_m (\sum_r p_{rm}^2)^2} = \frac{(\sum_s q_{sm}^2)^2}{\sum_m (\sum_s q_{sm}^2)^2}$$

Koeficient determinace několika faktorů $F_k + F_m + \dots$ je určen jako součet příslušných $R_k^2 + R_m^2 + \dots$.

Kanonický rozklad poskytuje tzv. *hlavní řešení*, které má tu vlastnost (plynoucí z Eckart-Youngovy věty), že první faktor vysvětluje maximální možný podíl variability interakcí y_{rs}^2 , tj. *vede k nejvyššímu možnému R_1^2* . Obdobně druhý faktor maximalizuje podíl vysvětlení zbytku interakcí po odečtení vlivu prvního faktoru atd. pro další faktory. Hlavní řešení má tak *vlastnost postupné maximalizace koeficientů determinace*.

Faktory, které mají nízký, zanedbatelný koeficient determinace nebo které neumíme interpretovat, vypouštíme z modelu a rovnicí (14) zapisujeme jako vztah:

$$(18) \quad \mathbf{Y} = \tilde{\mathbf{Y}} + \varepsilon,$$

$$\text{kde } \tilde{\mathbf{Y}} = \sum_m^K \mathbf{F}_m \quad (K < M) \quad \text{a} \quad \varepsilon = \sum_{m=K+1}^M \mathbf{F}_m = \text{náhodné vlivy}$$

Rozklad matice $\mathbf{Y}$ pomocí charakteristických čísel a vektorů není jediný. Existují i jiná řešení, která sice nemaximalizují determinace, ale která mají jiné výhodné interpretovatelné vlastnosti. Hledáme je tak, že hlavní řešení (resp. ty jeho faktory, které jsme nevyloučili) transformujeme tzv. *rotací*, tj. *orthonormální lineární transformací*.⁽¹⁴⁾

(14) Běžně se používá rotace metodou VARIMAX. Ta vede k takovému souboru koeficientů, které vykazují co největší rozdíly ve sloupcích a co nejvíce nulových hodnot v řádcích matic.

Kritériem pro nové řešení je buď

- *maximální rozlišení objektů* každým jednotlivým faktorem, nebo
- *maximální diferencované působení položek* na objekty
- *kompromisní řešení* mezi oběma předchozími, které co nejvíce rozlišuje objekty při co nejčistším působení položek.

Rotační transformace je vyjádřena pomocí transformační matice $\mathbf{T}$ (s rozměry $M \times M$) $\mathbf{P}^* = \mathbf{P}\mathbf{T}$ a $\mathbf{Q}^* = \mathbf{Q}\mathbf{T}$. Tím je získáno nové řešení.⁽¹⁵⁾

Veličiny $\mathbf{U}_m^*$, $\mathbf{W}_m^*$ však nyní působí složitěji a faktory $\mathbf{F}_m^* = \mathbf{U}_m^* \mathbf{W}_m^*$ jsou korelované. Platí rovnice obdobná k (14):

$$(19) \quad y_{rs} = \sum \mathbf{F}_m^*(r, s) = \sum \mathbf{U}_m^*(r) \cdot \mathbf{W}_m^*(s).$$

Koeficienty determinace faktorů $\mathbf{F}_m^*$:

$$(20) \quad R^2(\mathbf{Y}, \mathbf{F}_m^*) = R_m^{*2} = \frac{[\sum d_{kt}^2 t_{km}^2]^2}{d_m^2} = \frac{\sum_r \sum_s p_{rm}^{*2} q_{sm}^{*2}}{\sum_r \sum_s y_{rs}^2}$$

Vzhledem ke korelovanosti faktorů *neplatí součtová vlastnost pro koeficient determinace* současného vlivu několika faktorů. Ten je vyšší a spočítá se podle vzorců:

$$(21) \quad R_{m,k}^2 = \frac{\frac{K}{\sum_i} [d_{it}^2 t_{im}^2]^2 + \frac{K}{\sum_i} [d_{it}^2 t_{ik}^2]^2 + 2 [\sum_i d_{it}^2 t_{im} t_{ik}]^2}{\sum_m^M d_m^2} =$$

$$= \frac{\sum_r \sum_s (p_{rm}^* q_{sm}^* + p_{rk}^* q_{sk}^*)^2}{\sum_r \sum_s y_{rs}^2}$$

$$(22) \quad R_\alpha^2 = \frac{\sum_{m \in \alpha} \sum_{k \in \alpha} [\sum_{i=1}^K d_{it}^2 t_{ik} t_{im}]^2}{\sum_m^M d_m^2} = \frac{\sum_r \sum_s (\sum_{m \in \alpha} p_{rm}^* q_{sm}^*)^2}{\sum_r \sum_s y_{rs}^2},$$

kde α je množina indexů určujících podmiňující faktory.

C) Model komparační báze

Vektory $\mathbf{Y}_r = (y_{r1}, y_{r2}, \dots, y_{rs})$ lze vyjádřit jako *lineární kombinace báze vektorového prostoru*. Tu vytvářejí vektory $\mathbf{V}_m$, $m = 1, 2, \dots, M$, které mají vlastnost lineární nezávislosti a jednotkové délky ($M \leq \min(R, S) - 1$). Kanonická dekompozice dává jedno řešení takové báze:

$$(23) \quad \mathbf{Y}_r = \sum_m^M c_{rm} \mathbf{B}_m, \quad \text{kde } c_{rm} = a_{rm} d_m,$$

přičemž $\mathbf{B}_1$ má pro určení $\mathbf{Y}_r$ největší podíl, $\mathbf{B}_2$ druhý největší atd.

(15) Vzorec zahrnuje i případ, v němž rotujeme jen K vybraných faktorů $\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_K$. V matici $\mathbf{T}$ o rozměrech $M \times M$ dosadíme za $t_{ij} = 1$ pro $i = j$ a $= 0$ naopak pro $i > K, j > K$. Tím se faktory $\mathbf{F}_{K+1}, \dots, \mathbf{F}_M$ nemění.

Při postupném zvětšování počtu vektorů $\mathbf{B}_m$ ve výrazu

$$(24) \quad \tilde{\mathbf{Y}}_r = \sum_m^K c_{rm} \mathbf{B}_m$$

dostáváme stále lepší aproximaci vektorů $\mathbf{Y}_r$.⁽¹⁶⁾ Navíc je aproximace matice $\mathbf{Y}$ nejlepší možná (ve smyslu součtu čtverců vzdáleností $\sum_r \|\tilde{\mathbf{Y}}_r - \mathbf{Y}_r\|^2$ pomocí daného počtu vektorů. ($\|\mathbf{a}\|^2$ je čtverec normy.)

Rotací báze nebo její části ($\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_K$), $K \leq M$ dostáváme nové řešení, které můžeme volit tak, aby vektory nové báze co nejvíce odlišily jednotlivé složky, nebo tak, aby koeficienty c_{rm} byly co nejkontrastnější pro každý interakční profil $\mathbf{Y}_r$.

Vztah (24) a analogický vztah pro rotované vektory ukazuje, jak vynecháním nepodstatných bazových složek lze provést *vyrovnaní dat*. Kromě tohoto použití umožňuje komparační báze predikci profilu pro nový objekt, jehož vztah k objektům profilové tabulky je znám a je matematicky vyjádřitelný. Analytický význam této varianty modelu spočívá v pokusu interpretovat složky báze jako nezávislé obsahové složky profilových vlastností a odhad jejich vah ve skladbě pro každý jednotlivý objekt. Složky báze lze tak považovat za *mnohorozměrný faktor* a váhy c_{rm} za charakteristiky vztahu mezi faktorem a objektem.

D) Model korespondenční mapy

Rozklad (12) a rovnice (13) ukazují, že rezidua y_{rs} vznikají jako *skalární součiny* r -tého řádku $\mathbf{p}_r$ matice $\mathbf{P}$ a s -tého řádku $\mathbf{q}_s$ matice $\mathbf{Q}$ (resp. při aproximaci půjde o řádky matic $\mathbf{P}^k$ a $\mathbf{Q}^k$).

$$(25) \quad y_{rs} = \langle \mathbf{p}_r, \mathbf{q}_s \rangle \quad y_{rs} = \langle \mathbf{p}_r^k, \mathbf{q}_s^k \rangle$$

To naznačuje možnost vyjádřit vztahy mezi řádky a mezi sloupci profilové tabulky geometricky. V K -rozměrném⁽¹⁷⁾ eukleidovském prostoru znázorníme objekty i položky pomocí bodů o souřadnicích $\mathbf{p}_r^k$ a $\mathbf{q}_s^k$. Tím dostaneme tři druhy vztahů:

- a) blízkost objektových bodů znamená jejich podobnost vzhledem k vlivu faktorů;
- b) blízkost položkových bodů značí podobný vliv na objekty;
- c) blízkost položkových a objektových bodů značí typičnost položky pro objekt.

Body takto určené mají těžiště v počátku (a to jak body objektů, tak i body položek) a jejich rozptýlení lze charakterizovat pomocí součtu čtverců vzdáleností. Podíl m -té škály (dimenze prostoru) na tomto rozptýlení je dán *koeficientem diferenciací* m -té škály vzhledem ke K -rozměrné škále K faktorů:

$$(26) \quad v_m^2 = \frac{\sum_r p_{rm}^2}{\sum_m \sum_r p_{rm}^2} = \frac{\sum_s q_{sm}^2}{\sum_m \sum_s q_{sm}^2} = \frac{d_m}{\sum_m d_m}$$

Při rotaci se vzdálenosti bodů mezi sebou ani od počátku nemění (rotace znamená natáčení os kolem počátku), a proto koeficient diferenciací m -té rotované škály vzhledem k mnohorozměrné škále K faktorů spočítáme obdobně:

$$(27) \quad v_m^{*2} = \frac{\sum_r p_{rm}^{*2}}{\sum_m \sum_r p_{rm}^{*2}} = \frac{\sum_s q_{sm}^{*2}}{\sum_m \sum_s q_{sm}^{*2}}$$

(16) Pro $K = M$ dostáváme plnou reprodukci všech profilů.

(17) Pro $K = 1$ je to přímka, pro $K = 2$ rovina, $K = 3$ třírozměrný prostor.

E) Aplikační poznámky k modelu

V některých typech dat je pro analýzu výhodné nejprve údaje v profilové tabulce přetransformovat tak, aby data přirozeně vyhovovala součtovému modelu. Dvě nejčastější transformace jsou:

a) *logaritmus*

$$(28) \quad Z = \ln(X - A) \quad \text{resp.} \quad Z = \ln(X - A + k)$$

kde k je malá kladná konstanta pro případ, že by nastalo $X = A$ a A je dolní hranice hodnot proměnné X .

Tato transformace se používá u dat, u nichž je přirozené porovnávat různé hodnoty poměrově (tj. klademe otázku „kolikrát“ spíše než „o kolik“). Po logaritmování se poměry a násobky změny na rozdíly a součty. Je obvyklá u četností.

b) *logit*

$$(29) \quad Z = \text{logit}(A, B, k) = \ln \left(\frac{X - A + k}{B + k - X} \right),$$

kde X je proměnná z intervalu $\langle A, B \rangle$ a k je malá kladná konstanta pro případ, že $X = A$ nebo $X = B$.

Logitová transformace bývá výhodná u proměnných, které nabývají hodnot z uzavřeného intervalu a jejichž význam vede k tomu, že význam stejných rozdílů se liší, jsou-li odvozeny ze středu škály či z jejích krajů.

Kromě těchto transformací existuje celá řada dalších, například odmocninové transformace či speciální transformace používané pro některá rozložení a typy dat.

V popisu modelu jsme používali v souvislostech s řádkovými a sloupcovými faktorními koeficienty **A** a **B** výrazy „položky ovlivňují objekty“ a „objekty se liší, diferencují“. V souladu s poznámkami o směrech kauzálního působení, různého významu dat, či možné existence latentních procesů ap., které byly uvedeny v předchozích částech, by bylo možno podle významu obou vstupů do profilové tabulky formulovat model jako „objekty ovlivňují profily“, „položky jsou ovlivňovány“ nebo „proces souběžně vytváří hodnotu položek a projevuje se s danou intenzitou u objektu“ atd. atd. Interpretace je tak závislá na datech a obsahu tabulky.

4. Postup kvantifikace metodou LINDA-K — příklad

Základní využití metody LINDA-K nacházíme v komparaci profilů. Lze ji však využít i pro kvantifikace profilových položek při hledání vah jejich důležitosti v celkovém rámci profilu. Postup ukážeme na praktickém ilustračním příkladě, v němž je cílem kvantifikovat kategorie znaku, jehož typ je nejasný. Za profily bereme rozložení relativních četností.

Ve výzkumu návštěvníků zoologických zahrad byly položeny otázky na výši postihu za překročení patnácti různých bodů návštěvního řádu. Pro každý přestupek byla stanovena stupnice **T** devíti různých postihů: **I** = „není přestupek“, **II** = „vykázat ze ZOO“, **III** = „do 20 Kčs“, **IV** = „do 50 Kčs“, **V** = „do 100 Kčs“, **VI** = „do 200 Kčs“, **VII** = „do 500 Kčs“, **VIII** = „do 1000 Kčs“, **IX** = „vice než 1000 Kčs“. Každý z 15 různých přestupků byl respondenty hodnocen na této stupnici. Statistická analýza jednotlivých distribucí je však komplikována tím, že není jasný typ znaků. Kategorie **II**, o které nic nevíme, není vřazena do ordinální škály ostatních kategorií. Proto hodnocení nemůže být provedeno metodami pro ordinální data ani pro kardinální data (není jasné ani uspořádání ani významová kvantifikace).

Tabulka 9. Rotované řešení — rotace k prosté struktuře řádkových faktorových koeficientů, matice sloupcových faktorových koeficientů je přizpůsobena: rotovány 2 faktory hlavního řešení tabulky 4

typ manželství	řádkové efekty	faktor 1	faktor 2
1. harmonické	0.37	1.036	0.082
2. stabilizované	0.40	0.073	0.866
3. narušené	—0.22	—0.230	—0.488
4. rozvrácené	—0.54	—0.878	—0.460

důvody uzavření sňatku	sloupcové efekty	faktor 1	faktor 2
A citový vztah	1.56	1.122	0.325
B majetková situace partnera	—0.40	—0.210	—0.364
C společenská situace partnera	—0.49	—0.352	—0.064
D těhotenství	0.67	0.080	0.551
E možnost získat byt	—0.17	—0.166	0.196
F naléhání partnera	—0.43	—0.525	0.105
G naléhání rodičů	—0.47	—0.173	—0.433
H osobnost partnera	0.24	0.348	0.275
I základní vojenská služba	—0.52	—0.125	—0.590
		1.90	1.21

koeficient diferenciací škály (v %)	61.2	38.8
koeficient determinace (v %)	59.9	24.1
společně (v %)	98.6	

Poznámka: 1. Příspěvek společného působení faktorů je 14.6 %

Nominální chápání znaku vede jasně ke ztrátě informace. (Místo, na které byla vřazena II = „vykázat ze ZOO“, bylo určeno předběžně logikou výše vstupního a finanční ztráty způsobené vykazáním, tedy řazení vyšlo z předpokladu, že jde o nejnižší postih.)

Hlavním úkolem je řazení přestupků podle závažnosti. Jaké však zvolit kritérium u znaku, jehož typ nelze považovat za ordinální ani kardinální? Pomocí metody LINDA-K se pokusíme provést vřazení kategorie „vyvést ze ZOO“ do stupnice ostatních kategorií a nalézt přibližný finanční ekvivalent, který je jí zkoumaným souborem přisuzován. To provedeme pomocí modelu, který tu jen stručně naznačíme.

Tabulka profilů obsahuje podmíněná rozložení odpovědí na 15 otázek. Objektem je tu přestupek a profilovými položkami jsou kategorie postihu. Tabulku označíme $\mathbf{X} = \|f_{s/r}\|$. Je-li vektor $\mathbf{b}$ kvantifikací znaku (b_s = hodnota s -té kategorie), jsou řádkové průměry (tj. průměrné hodnocení přestupků) vyjádřeny složkami vektoru $\mathbf{Xb}$.

Pro dekompozici (12) platí odvozeně:

$$(30) \quad \mathbf{YQ} = \mathbf{PD} \quad \mathbf{YQ}_m = d_m \mathbf{P}_m, \quad m = 1, 2, \dots, M$$

Tabulka 10. Faktorový rozklad profilů rozložení četnosti pro odhad výše trestu u různých přestupků v ZOO Praha (hlavní řešení, uvedeny pouze tři první faktory s 98.2 % vyčerpané reziduální heterogenity)

výše trestu	diferenciační faktory		
	F1	F2	F3
1. není přestupek	1.75	—1.45	.39
2. vykázat ze ZOO	—1.19	—1.39	—1.21
3. do 20 Kčs	1.33	.40	—1.02
4. do 50 Kčs	.88	.85	—1.15
5. do 100 Kčs	.20	.74	.05
6. do 200 Kčs	—1.35	.49	.20
7. do 500 Kčs	—1.72	.25	.41
8. do 1000 Kčs	—1.27	—1.18	.44
9. více než 1000 Kčs	—1.62	—1.72	—1.11
koeficient determinace ($100 r_m^2$)	80.4 %	14.8 %	3.0 %
kumulativní koeficient determinace (v %)	80.4 %	95.2 %	98.2 %
přestupek			
	F1	F2	F3
A dráždění zvířat větve	.11	.15	—1.53
B házení smrtících krmiv	—1.46	—1.41	—1.45
C krmení přinesenou potravou	.48	—1.19	.59
D krmení ulámanými větve	.59	—1.24	.73
E ničení informačních tabulek	—1.09	.84	.12
F pohazování odpadků po zemi	.91	.52	—1.18
G vstup do prostoru zvířat	—1.23	.11	.03
H ubližování zvířatům	—1.06	—1.07	.04
I rušení hlukem	.83	—1.17	—1.14
J ničení zařízení	—1.45	1.00	.11
K vypouštění nebezpečných zvířat na svobodu	—1.11	—1.80	.09
L vypouštění zvířat (ne nebezpečných)	—1.69	—1.55	.33
M fotografování bez poplatku	1.66	—1.71	—1.08
N ničení stromů, keřů, květin	—1.17	.83	.03
O vodění psů, koček a jiných zvířat do ZOO	.68	—1.31	—1.69
koeficient determinace ($100 r_m^2$)	80.4 %	14.8 %	3.0 %
kumulativní koeficient determinace	80.4 %	95.2 %	98.2 %

Poznámky: 1. Zdroj dat: Výzkum návštěvníků ZOO v ČSSR (S. Plieka, J. Řehák, 1979).
2. Data pocházejí z jednoho z dílčích souborů výzkumu.
3. Před aplikací LINDA-K byly relativní četnosti distribucí logaritmovány: $y = \ln f$.

Navíc má první sloupec $\mathbf{Q}_1$ tu vlastnost, že zajišťuje maximální varianci hodnot $\|p_{r1}\|$ vzhledem k omezením, která klademe na dekompozici, a zároveň platí, že složky vektoru $\mathbf{Xb}$ a $\mathbf{Yb}$ jsou od sebe konstantně rozdílné.

Při podmínce $\sum b_s = 0$ můžeme za kvantifikaci přijmout první hlavní faktor $\mathbf{Q}_1$ řešení LINDA-K. Vektor $\mathbf{P}_1$ zároveň řadí řádky podle průměrů a zajišťuje jejich

nejlepší rozlišitelnost. Faktor můžeme interpretovat jako *hodnotící kritérium, které každé polože připisuje určitou postihovou hodnotu a každému přestupku určitý stupeň postihu*.

Hlavní řešení⁽¹⁸⁾ je uvedeno v tab. 10.

Faktorová dekompozice profilů odhalila osm faktorů, z nichž však uvádíme jen první tři ($R_1^2 + R_2^2 + R_3^2 = .982$). Pro řešení bude ale podstatný první hlavní faktor. Koeficienty kategorií tohoto faktoru jsou seřazeny podle výše peněžního trestu, u kategorie „*vykázat ze ZOO*“ však efekt neodpovídá jejímu zařazení ve stupnici. Je vřazena ordinálně mezi „*do 100 Kčs*“ a „*do 200 Kčs*“, takže je v hodnocení návštěvníků ekvivalentní přibližně pokutě 150–180 Kčs. To samo o sobě je již důležitý sociologický poznatek, který staví postih do souvislosti s peněžními částkami, a tak odhaluje relace, které nelze stanovit přímo. To pochopitelně může mít vliv na další analýzu dat tím, že *prohlásíme znaky postihu jednotlivých přestupků za ordinální nebo za kardinální* (přiřadíme každé kategorii peněžní hodnotu).

Jestliže přijmeme $\mathbf{Q}_1$ za kvantifikační vektor, ukazují hodnoty složek vektoru $\mathbf{P}_1$ relace mezi průměrnými postihy (k vlastním průměrům bychom přišli lineární transformací). Skóry přiřazené prvním hlavním faktorem jsou opačně orientovány, a proto pro výhodnější interpretaci u nich změníme znaménko, aby vyšší kladná hodnota znamenala vyšší postih.

Kategorie „*vykázat ze ZOO*“ se týká také slabý faktor 3. Přes malou determinaci (3.0 %) ukazuje odchylku pro tento postih, a to u přestupků zjevně nebezpečných vystavovaným exponátům (dráždění, jedy, zavlečení infekce). Můžeme vyslovit hypotézu, že u části respondentů tak je vykázání ze ZOO spojeno nejen s postihem, ale také s obranou exponátů, popřípadě, že šlo o návštěvníky s odborným nebo zájmovým zázemím, kteří si nejen uvědomují nutnost pořádku, ale také nebezpečí pro samotné exponáty.

Podobných případů lze v analytické praxi nalézt značné množství. Jsou to buď znaky s nejasným určením typu, u nichž se ptáme, zda lze předpokládat jejich ordinální nebo kardinální charakter, nebo u nichž chceme určit jejich obecný typ pomocí skóre určených vícerozměrnými vzdálenostmi řádků matice $\mathbf{Q}$. Ve výše uvedeném příkladě uživatele jistě napadne určitá analogie s Thurstonovou metodou následných intervalů a zlepšené možnosti dekompozičního postupu. Pro kontingenční tabulky lze obdobnou kvantifikaci provést pomocí kanonické analýzy (viz Lauber, Linhart [1978]. Nishisato [1980]).

Závěr

Modely, které vedou na využití metody LINDA-K jsou značně rozmanité. Nejčastějším aplikačním postupem je využití faktorové interpretace a grafického zobrazení geometrických vztahů (korespondenční mapa). Profily údajů komparujeme v sociologické práci velmi často, a proto metoda, která umožňuje vniknout do vzájemných skrytých vztahů v datech, je užitečným prostředkem.

Ze zkušeností, jichž autoři nabyli v průběhu experimentování s metodou, se ukazuje, že výsledky metody v jednoduché formě sumarizují závěry, jež je jinak nutno dělat z únavných a zdoluhavých analýz podrobných třídění dat. Navíc klasický přístup neumožňuje odhalit některé málo výrazné, ale důležité vztahy, které se objevují ve faktorech s nízkou determinací.

Explorační metoda LINDA-K netestuje *statistické* hypotézy, ale umožňuje je formulovat. Výsledky metody u silných faktorů jsou často průkazné, zatímco

(18) Původní profilovou tabulku (distribucí pro úsporu místa neuvádíme. Je to soubor distribucí 1. stupně třídění pro baterii otázek na postih jednotlivých přestupků. V příkladu byly profily před zpracováním logaritmičky transformovány.

faktory, které se prosazují s nižší intenzitou, vedou často k formulaci důležitých hypotéz a podněcují sociologickou imaginaci.

Úspěšnost aplikace závisí na možnostech interpretace číselných výsledků. V mnoha případech se ukazuje, že výsledky nedávají smysluplný závěr, že nejsou výzkumníkem vysvětlitelné, nebo že nepřinášejí nic nového. To bývá způsobeno nepravými korelacemi, heterogenitou skupin i celého souboru, nevyvážeností skladby celého souboru, nevhodnou volbou profilových položek a formulací otázek, podle nichž získáváme data, nebo chybnou či nepřímou operacionalizací. Někdy je též zapotřebí data transformovat, aby jejich obsahové vlastnosti odpovídaly aditivním vztahům modelu. Proto je třeba k závěrům přistupovat opatrně. Jestliže data neodpovídají modelu, pak číselné výsledky mají jen formální charakter bez reálného významu.

Metoda je v různých variantách a nezávislých větvích ve světové sociologii velmi populární, především v analýze kategorizovaných dat (viz např. Nishisato [1980] a tam uvedená podrobná bibliografie). Zaslouží si však aplikace i pro obecnější datové vstupy.

Dodatek

Algoritmus výpočtů faktorových koeficientů

V této části je podán opravený a modifikovaný algoritmus LINDA pro různé varianty modelu (včetně metody LINDA-A, viz Řehák, Loučková [1984], kterou tímto opravujeme).

krok 1. Vstup obdélníkové matice profilů $\mathbf{X} = \|x_{rs}\|$ o rozměrech $R \times S$

krok 2. Výpočet reziduí $\mathbf{Y} = \|y_{rs}\|$ o rozměrech $R \times S$

$$y_{rs} = x_{rs} - x_{r.} - x_{.s} + x_{..}, \quad x_{r.} = \frac{1}{S} \sum_s x_{rs}$$

$$x_{.s} = \frac{1}{R} \sum_r x_{rs}, \quad x_{..} = \frac{1}{RS} \sum_r \sum_s x_{rs}$$

a jejich pomocí součty čtverců:

$$SS_T = \sum_r \sum_s (x_{rs} - x_{..})^2$$

$$SS_R = S \sum_r (x_{r.} - x_{..})^2$$

$$SS_S = R \sum_s (x_{.s} - x_{..})^2$$

$$SS_I = \sum_r \sum_s y_{rs}^2 = \sum_r \sum_s (x_{rs} - x_{r.} - x_{.s} + x_{..})^2$$

$$\text{Podíl řádkových efektů} = \frac{SS_R}{SS_T}$$

$$\text{Podíl sloupcových efektů} = \frac{SS_S}{SS_T}$$

$$\text{Podíl interakčních efektů} = \frac{SS_I}{SS_T}$$

krok 3. Je-li $R \leq S$: výpočet matice $\mathbf{Y}\mathbf{Y}'$, jejich nenulových charakteristických čísel $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_M > 0$ a k nim příslušných charakteristických vektorů $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_M$ (o rozměrech $R \times 1$). Označme $\mathbf{\Lambda} = \text{diag} \{\lambda_i\}$, $\mathbf{D} = \text{diag} \sqrt{\lambda_i} = \mathbf{\Lambda}^{\frac{1}{2}}$, $\mathbf{D}^{\frac{1}{2}} = \text{diag} \{\sqrt{d_i}\}$, $d_i = \sqrt{\lambda_i}$

krok 4. Výpočet $\mathbf{P} = \mathbf{A}\mathbf{D}^{\frac{1}{2}}$, $\mathbf{A}$ je matice ze sloupců $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_M$ o rozměru $R \times M$

krok 5. Výpočet $\mathbf{Q} = \mathbf{Y}'\mathbf{P}\mathbf{D}^{-1}$

krok 6. Výpočet koeficientů determinace

$$r_m^2 = \frac{\lambda_m}{\lambda_1 + \lambda_2 + \dots + \lambda_M} = \frac{\frac{R}{\sum_r (\sum_r p_{rm}^2)^2}}{\sum_m (\sum_r p_{rm}^2)^2} = \frac{\frac{S}{\sum_s (q_{sm}^2)^2}}{\sum_m (\sum_s q_{sm}^2)^2}$$

krok 7. Výpočty faktorových příspěvků k interakcím, které poskytují matice $\mathbf{P}_1\mathbf{Q}_1', \mathbf{P}_2\mathbf{Q}_2', \dots, \mathbf{P}_M\mathbf{Q}_M'$ jejich kumulace pro prvních k faktorů a výpočet zbytkových členů $\mathbf{Y} - \mathbf{P}_1\mathbf{Q}_1' - \mathbf{P}_2\mathbf{Q}_2' - \dots - \mathbf{P}_k\mathbf{Q}_k', \quad k = 1, 2, \dots, M-1$

krok 8. Výpočet koeficientů diferenciace v_m^2 :

$$v_m^2 = \frac{d_m}{\sum d_m} = \frac{\sum_r p_{rm}^2}{\sum_m \sum_r p_{rm}^2} = \frac{\sum_s q_{sm}^2}{\sum_m \sum_s q_{sm}^2}$$

Součet podle m odpovídá škálám, které uvažujeme a k jejichž celku výpočet vztahujeme.

krok 9. Sestrojení korespondenčního grafu

Je-li $R > S$, nahradí se kroky 3—5 postupem:

krok 3' Výpočet nenulových charakteristických čísel matice $\mathbf{Y}'\mathbf{Y}$, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_M$ a k nim příslušných charakteristických vektorů $\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_M$ ($\mathbf{B} = \|\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_M\|$)

krok 4' $\mathbf{Q} = \mathbf{B}\mathbf{D}^{\frac{1}{2}}$

krok 5' $\mathbf{P} = \mathbf{Y}\mathbf{Q}\mathbf{D}^{-1}$

Algoritmus rotace

krok 1: Volba rotační strategie (tj. matice pro rotaci) a volba K faktorů, které budou rotovány.

krok 2: Aplikace metody rotace (např. VARIMAX) pro vybranou matici, zjištění transformační matice $\mathbf{T}$ a přizpůsobení druhé matice

I: najde se $\bar{\mathbf{P}}^* = \bar{\mathbf{P}}\mathbf{T}_p$ a přizpůsobí se $\bar{\mathbf{Q}}^* = \bar{\mathbf{Q}}\mathbf{T}_p$

II: najde se $\bar{\mathbf{Q}}^* = \bar{\mathbf{Q}}\mathbf{T}_q$ a přizpůsobí se $\bar{\mathbf{P}}^* = \bar{\mathbf{P}}\mathbf{T}_q$

III: najde se $\bar{\mathbf{C}}^* = \bar{\mathbf{C}}\mathbf{T}$

kde $\bar{\mathbf{P}}, \bar{\mathbf{Q}}$ jsou matice vytvořené těmi sloupci $\mathbf{P}, \mathbf{Q}$, které odpovídají vybraným faktorům,

$\bar{\mathbf{C}}$ je bloková matice $\bar{\mathbf{C}} = \left\| \begin{array}{c} \mathbf{P} \\ \mathbf{Q} \end{array} \right\|$

krok 3: Výpočet příspěvků jednotlivých faktorů $\mathbf{P}_1^*\mathbf{Q}_1^*, \mathbf{P}_2^*\mathbf{Q}_2^*, \dots$ a příspěvků skupin faktorů $(\mathbf{F}_1^*, \mathbf{F}_2^*, \dots, \mathbf{F}_m^*)$, $m = 1, 2, \dots, K-1$

$$\sum_{k=1}^m \mathbf{P}_k^* \mathbf{Q}_k^*$$

krok 4: Výpočet koeficientů determinace pro rotované faktory

$$r_m^{*2} = \frac{\sum_s \sum_r p_{rm}^{*2} q_{sm}^{*2}}{\sum_r \sum_s y_{rs}^2} = \frac{\frac{K}{k} [\sum_k d_k t_{km}^2]^2}{\sum_k d_k^2}$$

a výpočet kumulativních koeficientů determinace

$$R_m^{*2} = \frac{\sum_r \sum_s [\sum_{k=1}^m p_{rk}^{*2} q_{sk}^{*2}]^2}{\sum_r \sum_s y_{rs}^2} = \frac{\sum_k \sum_l [\sum_i d_i t_{ik} t_{il}]^2}{\sum_k d_k^2}$$

pro $m = 1, 2, \dots, K$

krok 5: Výpočet koeficientů diferenciace škál:

$$v_m^{*2} = \frac{\sum_r p_{rm}^{*2}}{\sum_k \sum_r p_{rk}^{*2}} = \frac{\sum_s q_{sm}^{*2}}{\sum_k \sum_s q_{sk}^{*2}}$$

krok 6: Sestrojení korespondenční mapy z bodů pro objekty:

$$p_r^* = (p_{r1}^*, p_{r2}^*, \dots, p_{rK}^*) \quad r = 1, 2, \dots, R$$

a z bodů pro položky profilu:

$$q_s^* = (q_{s1}^*, q_{s2}^*, \dots, q_{sK}^*)$$

krok 7: Výpočet korelací rotovaných faktorů

$$r(\mathbf{F}_j^*, \mathbf{F}_k^*) = \frac{[\sum_{m=1}^K d_m t_{mj} t_{mk}]^2}{(\sum_m d_m t_{mj}^2) (\sum_m d_m t_{mk}^2)}$$

krok 8: Výpočet korelací rotovaných a nerotovaných faktorů

$$r(\mathbf{F}_i, \mathbf{F}_j^*) = \frac{d_i t_{ij}^2}{\sum_m d_m t_{mj}^2}$$

Literatura

- Anderson T. W.: *An Introduction to Multivariate Statistical Analysis*. New York, Wiley 1958.
- Ajvazjan S. A., Běžajeva Z. I., Starověrov O. V.: *Klassifikacija mnogorozměrnych nabljudenij*. Moskva, Statistika 1974.
- Benzécri J. P.: *Statistical Analysis as a Tool to Make Patterns Emerge from Data*. In: *Methodologies of Pattern Recognition* (S. Waternabe, ed.) New York, Academic Press 1969.
- Gantmacher F. G.: *Teoriya matric*. Moskva, Nauka 1966.
- Gollob H. F.: *A Statistical Model Which Combines Features of Factor Analytic and Analysis of Variance Techniques*. Psychometrika 33, 1968a, s. 73—115.
- Gollob H. F.: *Confounding of Sources of Variation in Factor — Analytic Techniques*. Psychological Bulletin 70, 1968b, s. 330—344.
- Guttman L.: *The Quantification of a Class of Attributes: a Theory and Methods of Scale Construction*, in: *The Prediction Adjustment* (P. Horst, ed.), New York: Social Science Research Council 1941, s. 251—364.

- Guttman L.: *Metricizing Rank-Ordered or Unordered Data for Linear Factor Analysis*. Sankhya 21, 1959, s. 257—268.
- Harmann H. H.: *Modern Factor Analysis*. Chicago, The University of Chicago Press 1967.
- Hill M. O.: *Correspondence Analysis: A Neglected Multivariate Method*. Applied Statistics 23, 1974, s. 340—354.
- Kasalová H., Řehák J.: *Diferenciální tendence v recepci umění*. Sociologický časopis 1984, 20, č. 1, s. 35—37.
- Kildišev G. S., Abolencev Ju. I.: *Mnogorozměrnýje grupirovki*, Moskva, Statistika 1978.
- Loučková I., Řehák J.: *LINDA — Soubor programů pro analýzu reziduálních odchylek v drojroz-
měrných uspořádaných dat*, in: *Současný vývoj programů pro potřeby sociologie*. Střín. Sekce
metod a technik ČSSS při ČSAV 1979, s. 24—44.
- Nishisato S.: *Analysis of Categorical Data: Dual Scaling and its Applications*. Toronto, University
of Toronto Press 1980.
- Řehák J., Loučková I.: *Korespondenční analýza*, kapt. in: *Škádování a kvantifikace* (red. J. Řehák).
Mikulov, ČSSS, 1981.
- Řehák J., Loučková I.: *Faktorová analýza kontingenčních tabulek*, Robust. Praha, JČMF 1982.
- Řehák J., Loučková I.: *Klasické mnohorozměrné škádování (Aplikace metody DISTAN)*. Socio-
logický časopis 1983, 19, č. 5, s. 535—554.
- Řehák J., Loučková I.: *Faktorová analýza v kontingenčních tabulkách (Model LINDA-A —
— popis a užití)*. Sociologický časopis 1984, 20, č. 2, s. 174—193.
- Řehák J., Řeháková B.: *Základní charakteristiky proměnných s konečným počtem hodnot a dis-
tanční analýza jejich rozložení*. Sociologický časopis 1979, 15, č. 2, s. 214—233.
- Řehák J., Řeháková B.: *Analýza kategorizovaných dat v sociologii*. Praha, ACADEMIA 1984
(v tisku).
- Überla K.: *Faktorenanalyse*, Springer Verlag, Z. Berlin 1977, též Iberla K.: *Faktornyj analiz*.
Moskva, Statistika 1980.
- Williams J. S.: *Canonical Analysis: A Factor Analytic Method of Comparing Finite Discrete
Distribution Functions*. Journal of the American Statistical Association 73, 1978, s. 781—786.

Резюме

**Я. Ржегак - И. Лоучкова: Сопоставительный факторный анализ профилей
(Модель и метод LINDA-K)**

Метод линейного декомпозиционного анализа, основанный на теореме Эккарта-Янга (здесь называемый LINDA-K), позволяет решать несколько задач сопоставления объектов, характеризующихся векторами данных, упорядоченных в таблице профилей. Это четыре возможные интерпретации:

- a) факторный анализ профилей ведет к выявлению причин систематических различий между профилями,
- b) сопоставительный базис для таблицы профилей предоставляет декомпозицию посредством совокупности линейно независимых векторов, которые складывают и сглаживают первоначальные данные,
- в) графическая карта соответствий показывает степень типичности признаков профилей для отдельных объектов, степень специфической ассоциации объектов и признаков, степень схождения объектов применительно к независимым компонентам вектора признаков и совпадение в действии признаков на объекты,
- г) упорядочение признаков и объектов и их квантификация выявляют основную нумерическую информацию квантификационной модели.

После описания проблемы и данных, для которых применим характеризующий метод (часть I), приводится практический пример, подробно иллюстрирующий сопоставительный метод и применение модели (часть 2). Математическая модель описана в третьей части, а затем в четвертой части приводится следующий пример, показывающий возможность применения метода для количественного выражения. В дополнительном разделе коротко описывается алгоритм математического вычисления.

Широкая практическая верификация методики на исследовательских данных позволяет рекомендовать ее для более частого применения в социологическом анализе данных. Метод носит эксплорационный характер и относится к классу методов анализа соответствий и дуального шкалирования.

Summary

J. Rehák — I. Loučková: Comparative Factor Analysis of Profiles (LINDA-K Model and Method)

The method of linear decompositional analysis based on Eckart-Young theorem (model called here LINDA-K) allows for solving several tasks in the comparative method. For objects characterized by data vectors arranged in profile tables, four ways of interpretation are presented:

- a) *factor analysis* of profile tables leads to an identification of the sources of systematic differences among profiles;
- b) *comparative base* for profile table yields the decomposition of the table into a linearly independent set of vectors that compose and smooth the original data;
- c) graphical *correspondence map* exhibits the degree of typicalness of profile items for individual objects and/or the degree of specific association of the objects to the items, the similarities of objects with respect to the independent components of item vector, and the congruences in the item influence;
- d) *ordering* of items and objects and their *quantification* reveals the basic numerical properties of quantification model.

After the description of the problem and the data for which the method is applicable (part 1), a practical example illustrating the comparative method and the use of the model is commented in detail (part 2), the mathematical model (including rotation and its properties) is described (part 3), and another example is given that shows the quantification possibilities of the method (part 4). In Appendix, the algorithm for mathematical solution of the problem is sketched.

The extensive practical experimentation with the method on real research data suggests its wider use in sociological data analysis. The procedure is of an exploratory nature and belongs to the family of correspondence analysis and dual scaling methods.