Alternativní přístup k analýze vícefaktorových dat

Kamila Fačevicová[1], Peter Filzmoser[2], Karel Hron[1]

[1] Katedra matematické analýzy a aplikací matematiky, Přírodovědecká fakulta Univerzity Palackého v Olomouci,
[2] Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology

kamila.facevicova@gmail.com

Motivace

Struktura zaměstnanosti ve 42 zemích v roce 2015. Např. pro Českou republiku (v tis.):

$$\mathbf{x} = \begin{array}{c} \\ \text{Žena} \\ \text{Muž} \end{array} \begin{pmatrix} \multicolumn{2}{c}{FT} & \multicolumn{2}{c}{PT} & \multicolumn{2}{c}{FT} & \multicolumn{2}{c}{PT} & \multicolumn{2}{c}{FT} & \multicolumn{2}{c}{PT} \\ 104.756 & & 17.128 & & 1618.415 & & 90.505 & & 317.031 & & 56.355 \\ 169.851 & & 11.165 & & 2127.849 & & 22.759 & & 467.212 & & 38.208 \end{pmatrix}$$

$$\begin{array}{cccccc} \multicolumn{2}{c}{15-24} & & \multicolumn{2}{c}{25-54} & & \multicolumn{2}{c}{55+} \end{array}$$

Zdroj: http://stats.oecd.org

Struktura zaměstnanosti v ČR

Věková struktura zaměstnanců v České republice v roce 2015:

$$\mathbf{x}^{\text{věk}} = \begin{matrix} \text{15 - 24} & \text{25 - 54} & \text{55+} \\ (0.09, & 0.62, & 0.29) \end{matrix}$$

$\Rightarrow$ **vektorová kompoziční data**

Vektorová kompoziční data

- **Kompoziční data**, standardně definovaná jako D-složkový vektor kladných hodnot $\mathbf{x} = (x_1, \ldots, x_D)'$, se vyznačují vlastností, že veškerá relevantní informace je obsažena v poměrech mezi složkami: $(5, 10)$ $(100, 105)$.
- Výběrovým prostorem je **D-rozměrný simplex** namísto celého $\mathbb{R}^D$:

$$\mathcal{S}^D = \{(x_1, \ldots, x_D)' \in \mathbf{R}^D, x_i > 0, \forall i, \sum_{i=1}^{D} x_i = \kappa\}.$$

- Data se řídí **Aitchisonovou geometrií** namísto Euclidovské:

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, \ldots, x_D y_D)' \quad \text{a} \quad \alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, \ldots, x_D^\alpha),$$

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{D} \sum_{i<j} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}.$$

Vektorová kompoziční data

- Vzhledem k relativní povaze dat není vhodné použití standardních metod.

$\Rightarrow$ Data nejprve vyjádříme v reálných souřadnicích a následně je analyzujeme pomocí standardních (klasických/robustních) analytických metod.

Vektorová kompoziční data

- Vzhledem k relativní povaze dat není vhodné použití standardních metod.

$\Rightarrow$ Data nejprve vyjádříme v reálných souřadnicích a následně je analyzujeme pomocí standardních (klasických/robustních) analytických metod.

Isometrické log-ratio (ilr) souřadnice

$$\mathrm{ilr}(\mathbf{x}) = (\langle \mathbf{x}, \mathbf{e}_1 \rangle_A, \ldots, \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_A)$$

- představují isometrický isomorfismus mezi $\mathcal{S}^D$ a $\mathbb{R}^D$

$$\mathrm{ilr}(\mathbf{x} \oplus \mathbf{y}) = \mathrm{ilr}(\mathbf{x}) + \mathrm{ilr}(\mathbf{x}), \quad \mathrm{ilr}(\alpha \odot \mathbf{x}) = \alpha \cdot \mathrm{ilr}(\mathbf{x}),$$

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \langle \mathrm{ilr}(\mathbf{x}), \mathrm{ilr}(\mathbf{x}) \rangle,$$

- neexistuje žádná standardní báze.

Vektorová kompoziční data

Pomocí **postupného binárního dělení** získáme systém $D-1$

- log-kontrastů $\xi_i = (\xi_{i1}, \ldots, \xi_{iD})$ s prvky

$$\xi_{i+} = \frac{1}{u}\sqrt{\frac{uv}{u+v}}, \quad \xi_{i-} = -\frac{1}{v}\sqrt{\frac{uv}{u+v}} \quad \text{a} \quad \xi_{i0} = 0 \quad ,$$

- ortonormálních vektorů

$$\mathbf{e}_i = \exp(\xi_i)$$

- a ilr souřadnic - **bilancí**

$$z_i = \sum_{j=1}^{D} \xi_{ij} \ln x_j = \sqrt{\frac{uv}{u+v}} \ln \frac{(x_{j_1} x_{j_2} \cdots x_{j_u})^{1/u}}{(x_{k_1} x_{k_2} \cdots x_{k_v})^{1/v}} \quad .$$

i	x_1	x_2	x_3	u	v
1	+	-	-	1	2
2		+	-	1	1

$$\xi_1 = \left(\sqrt{\tfrac{2}{3}}, -\tfrac{1}{2}\sqrt{\tfrac{2}{3}}, -\tfrac{1}{2}\sqrt{\tfrac{2}{3}}\right) \qquad z_1 = \sqrt{\tfrac{2}{3}} \ln \frac{x_1}{\sqrt{x_2 x_3}}$$

$$\xi_2 = \left(0, \sqrt{\tfrac{1}{2}}, -\sqrt{\tfrac{1}{2}}\right) \qquad z_2 = \sqrt{\tfrac{1}{2}} \ln \frac{x_2}{x_3}$$

Struktura zaměstnanosti

Věková struktura zaměstnanců v České republice v roce 2015:

$$\begin{array}{rccc} & \text{15 - 24} & \text{25 - 54} & \text{55+} \\ \mathbf{x}^{\text{věk}} = & (0.09, & 0.62, & 0.29) \end{array}$$

i	15 - 24	25 - 54	55+	u	v
1	+	-	-	1	2
2		+	-	1	1

$$z_1^{\text{věk}} = \sqrt{\tfrac{2}{3}} \ln\left(\frac{x_1}{\sqrt{x_2 x_3}}\right) = \mathbf{-1.25} \qquad (-1.53)$$

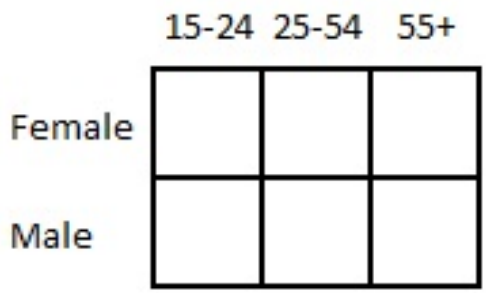
$$z_2^{\text{věk}} = \sqrt{\tfrac{1}{2}} \ln\left(\frac{x_2}{x_3}\right) = \mathbf{0.55} \qquad (0.78)$$

Struktura zaměstnanosti

Věková struktura zaměstnanců v České republice v roce 2015, podle jejich pohlaví a věku:

$$\mathbf{x} = \begin{array}{c} \\ \text{Žena} \\ \text{Muž} \end{array} \begin{array}{c} \begin{array}{ccc} 15-24 & 25-54 & 55+ \end{array} \\ \begin{pmatrix} 0.04 & 0.40 & 0.14 \\ 0.05 & 0.23 & 0.14 \end{pmatrix} \end{array}$$

$\Rightarrow$ **kompoziční tabulky**

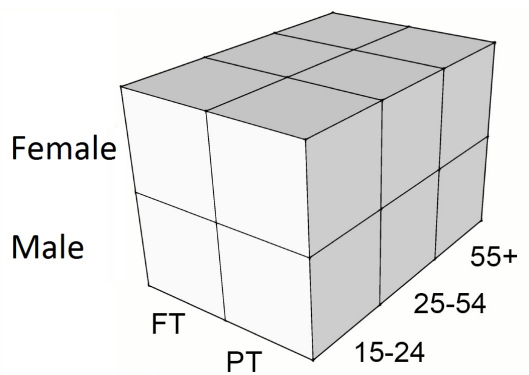


Struktura zaměstnanosti

Věková struktura zaměstnanců v České republice v roce 2015, podle jejich pohlaví, věku a typu prac. poměru:

$$\mathbf{x} = \begin{array}{c} \\ \text{Žena} \\ \text{Muž} \end{array} \begin{array}{c} \begin{array}{cc|cc|cc} \textit{FT} & \textit{PT} & \textit{FT} & \textit{PT} & \textit{FT} & \textit{PT} \end{array} \\ \left(\begin{array}{cc|cc|cc} 0.021 & 0.003 & 0.321 & 0.018 & 0.063 & 0.011 \\ 0.034 & 0.002 & 0.422 & 0.005 & 0.093 & 0.008 \end{array} \right) \\ \begin{array}{cc} 15-24 & \end{array} \quad \begin{array}{cc} 25-54 & \end{array} \quad \begin{array}{cc} 55+ & \end{array} \end{array}$$

$\Rightarrow$ **kompoziční krychle**



Kompoziční krychle

- **Kompoziční krychle**

$$\mathbf{x} = \left(\begin{array}{ccc|c|ccc} x_{111} & \cdots & x_{1J1} & & x_{11K} & \cdots & x_{1JK} \\ \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ x_{I11} & \cdots & x_{IJ1} & & x_{I1K} & \cdots & x_{IJK} \end{array} \right),$$

  kde $x_{ijk} > 0, \forall i, j, k$, představuje třífaktorové zobecnění $I \cdot J \cdot K$-složkových kompozičních dat,

- Výběrovým prostorem je $I \cdot J \cdot K$-složkový simplex

$$\mathcal{S}^{IJK} = \left\{ \mathbf{x} = (x_{111}, \ldots, x_{IJK}) | x_{ijk} > 0, \quad \forall i, j, k; \sum_{i,j,k=1}^{I,J,K} x_{ijk} = \kappa \right\}.$$

- Základní operace Aitchisonovy geometrie je potřeba modifikovat

$$\mathbf{x} \oplus \mathbf{y} = (x_{ijk} \cdot y_{ijk})_{i,j,k=1}^{I,J,K} \quad \text{a} \quad \alpha \odot \mathbf{x} = (x_{ijk}^{\alpha})_{i,j,k=1}^{I,J,K},$$

Souřadnicová reprezentace CoDa krychlí

Souřadnicový systém navržený pro vektorová kompoziční data (bilance) nerespektuje

- trojrozměrnou povahu krychlí,
- možnost jejich rozkladu na část nezávislou a části interakční

$\Rightarrow$ alternativní souřadnicový systém.

Souřadnicová reprezentace CoDa krychlí - konstrukce

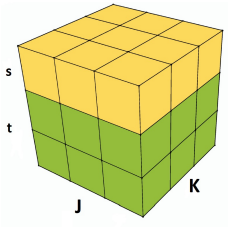
1. Definice PBD pro celé řádky, sloupce a řezy,
2. výpočet log-kontrastů ξ_i,
3. výpočet dalších log-kontrastů pomocí Hadamardova součinu dvojic a trojic log-kontrastů z různých PBD,
4. normování nových log-kontrastů,
5. výpočet $IJK - 1$ ortonormálních souřadnic s využitím vztahu

$$z_i = \sum_{j=1}^{D} \xi_{ij} \ln x_j \quad .$$

Souřadnicová reprezentace CoDa krychlí

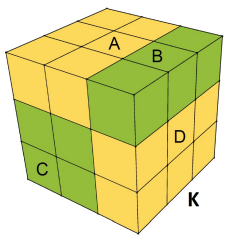
Takto získaný systém je tvořen třemi skupinami souřadnic popisujícími:

- bilance mezi úrovněmi jednotlivých faktorů



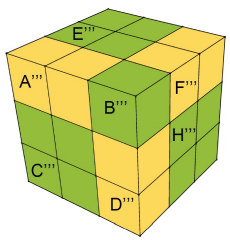
$$\text{např. } z_i^r = \sqrt{\frac{stJK}{s+t}} \ln \frac{\left[g(x_{j_1}..)\cdots g(x_{j_s}..)\right]^{1/s}}{\left[g(x_{k_1}..)\cdots g(x_{k_t}.)\right]^{1/t}}$$

- interakce mezi dvojicemi faktorů



$$\text{např. } z^{rc} = \sqrt{\frac{|A||D|}{|A|+|B|+|C|+|D|}} \ln \frac{g(x_A)g(x_D)}{g(x_B)g(x_C)}$$

- interakce mezi všemi faktory



$$\text{např. } z^{rcs} = K \ln \frac{g(x_{A'''})g(x_{D'''})g(x_{F'''})g(x_{G'''})}{g(x_{B'''})g(x_{C'''})g(x_{E'''})g(x_{H'''})}$$

Struktura zaměstnanosti

$$\mathbf{x} = \left(\begin{array}{cc|cc|cc} 0.021 & 0.003 & 0.321 & 0.018 & 0.063 & 0.011 \\ 0.034 & 0.002 & 0.422 & 0.005 & 0.093 & 0.008 \end{array}\right)$$

i	Ženy	Muži	s	t
1	+	-	1	1

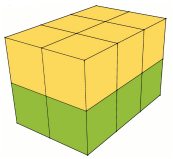
j	Full time	Part time	u	v
1	+	-	1	1

k	15 - 24	25 - 54	55+	m	n
1	+	-	-	1	2
2		+	-	1	1

Struktura zaměstnanosti

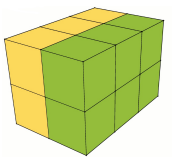
Bilance - pohlaví

$$z_1^r = \sqrt{3}\ln\left(\frac{g(x_{1..})}{g(x_{2..})}\right) = \mathbf{0.31} \qquad (0.18)$$



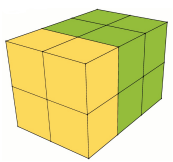
Bilance - typ

$$z_1^c = \sqrt{3}\ln\left(\frac{g(x_{.1.})}{g(x_{.2.})}\right) = \mathbf{4.67} \qquad (2.70)$$

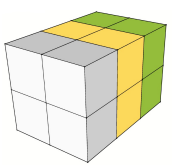


Bilance - věk

$$z_1^s = \sqrt{\frac{8}{3}}\ln\left(\frac{g(x_{..1})}{\sqrt{g(x_{..2})g(x_{..3})}}\right) = -\mathbf{2.50} \qquad (-1.53)$$



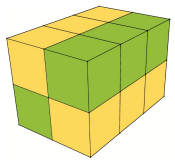
$$z_2^s = \sqrt{2}\ln\left(\frac{g(x_{..2})}{g(x_{..3})}\right) = \mathbf{1.10} \qquad (0.78)$$



Employment structure

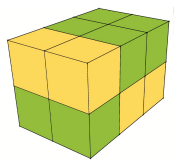
Interakce - pohlaví, typ

$$z_1^{rc} = \sqrt{\frac{3}{4}} \ln\left(\frac{g(x_{11.})g(x_{22.})}{g(x_{12.})g(x_{21.})}\right) = -\mathbf{0.97} \qquad \text{(-1.12)}$$

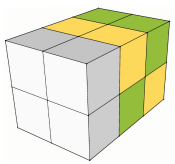


Interakce - pohlaví, věk

$$z_1^{rs} = \sqrt{\frac{2}{3}} \ln\left(\frac{g(x_{1.1})\sqrt{g(x_{2.2})g(x_{2.3})}}{g(x_{2.1})\sqrt{g(x_{1.2})g(x_{1.3})}}\right) = -\mathbf{0.25} \qquad \text{(-0.30)}$$

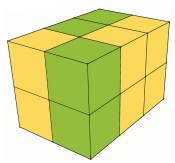


$$z_2^{rs} = \sqrt{\frac{1}{2}} \ln\left(\frac{g(x_{1.2})g(x_{2.3})}{g(x_{2.2})g(x_{1.3})}\right) = \mathbf{0.38} \qquad \text{(0.54)}$$

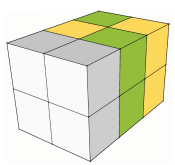


Interakce - typ, věk

$$z_1^{cs} = \sqrt{\frac{2}{3}} \ln\left(\frac{g(x_{.11})\sqrt{g(x_{.22})g(x_{.23})}}{g(x_{.21})\sqrt{g(x_{.12})g(x_{.13})}}\right) = -\mathbf{0.51} \qquad \text{(-0.63)}$$



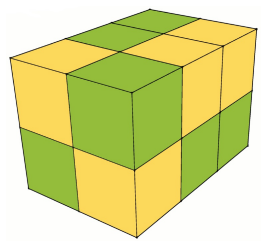
$$z_2^{cs} = \sqrt{\frac{1}{2}} \ln\left(\frac{g(x_{.12})g(x_{.23})}{g(x_{.13})g(x_{.22})}\right) = \mathbf{1.12} \qquad \text{(1.59)}$$



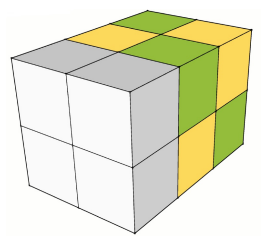
Employment structure

Plná interakce

$$z_1^{rcs} = \sqrt{\frac{1}{6}} \ln\left(\frac{x_{111}x_{221}\sqrt{x_{122}x_{123}}\sqrt{x_{212}x_{213}}}{x_{211}x_{121}\sqrt{x_{112}x_{113}}\sqrt{x_{222}x_{223}}}\right) = \mathbf{0.12} \qquad (0.30)$$

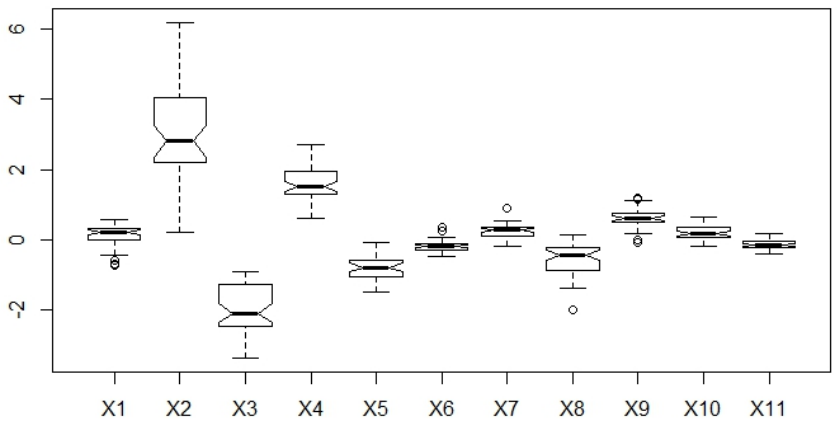


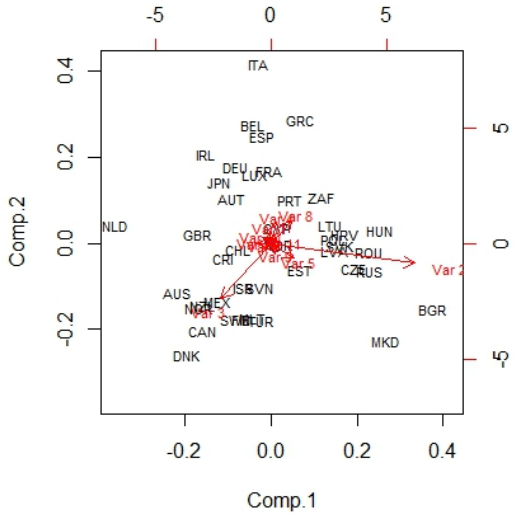
$$z_2^{rcs} = \sqrt{\frac{1}{8}} \ln\left(\frac{x_{112}x_{222}x_{123}x_{213}}{x_{122}x_{212}x_{113}x_{223}}\right) = -\mathbf{0.30} \qquad (-0.86)$$



Struktura zaměstnanosti

Celkově jsme měli k dispozici údaje o 42 zemích.





Data jsou nyní připravena k analýze pomocí standardních analytických metod.

Struktura zaměstnanosti

Při tradiční analýze souboru původních tabulek můžeme využít např. log-lineárních modelů. Tyto modely jsou však

- spíše konstruované pro analýzu jedné tabulky ne celého výběru.

Zahrneme-li do analýzy čtvrtý faktor (stát), je výsledný model

- velmi ovlivněn různými velikostmi států,
- které zastírají efekt sledovaných faktorů.

References


Agresti A (2002) *Categorical data analysis.* John Wiley & Sons, Inc., New Jersey.


Aitchison J (1986) *The statistical analysis of compositional data.* Chapman and Hall, London.


Egozcue JJ, Pawlowsky-Glahn V (2005) Groups of parts and their balances in compositional data analysis. *Math Geol* 37:795–828.


Fačevicová K, Hron K, Todorov V, Templ M (2016) Compositional tables analysis in coordinates. *Scandinavian Journal of Statistics*, 43(4): 962–977.


Fačevicová K, Hron K, Todorov V, Templ M (2016) General approach to coordinate representation of compositional tables. Under review.