Úvod do statistické analýzy kompozičních dat

Mgr. Adéla Vrtková

Katedra aplikované matematiky
Fakulta elektrotechniky a informatiky
Vysoká škola báňská - Technická univerzita Ostrava

Statistický seminář, 27. září 2016

Kompoziční data

- D-složková kompozice - popisuje části celku, nese relativní informaci

$$\mathbf{x} = (x_1, ..., x_D)', x_i > 0, i = 1, ..., D$$

- simplex

$$S^D = \{\mathbf{x} = (x_1, ..., x_D)', x_i > 0, i = 1, ..., D, \sum_{i=1}^{D} x_i = \kappa\}$$

- příklad - koncentrace prvků v horninách (mg/l, mg/g), volební hlasy v jednotlivých obvodech, výdaje domácností, metabolomická data

Předpoklady relevantní analýzy

- invariance na změnu měřítka
  - uzávěr

$$C(\mathbf{x}) = \left(\frac{\kappa x_1}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa x_D}{\sum_{i=1}^{D} x_i} \right)'$$

- invariance na permutaci
  - nezávislost analýzy na pořadí složek kompozic
- podkompoziční soudržnost
  - invariance na změnu měřítka pro libovolnou podkompozici
  - souvislost se zachováním geometrických vlastností, které charakterizují např. eukleidovskou geometrii

Aitchisonova geometrie na simplexu

Operace, které respektují vlastnosti kompozic a navíc vedou ke struktuře eukleidovského prostoru. Nechť $\mathbf{x}, \mathbf{y} \in S^D$, $\alpha \in \mathbb{R}$

- perturbace

$$\mathbf{x} \oplus \mathbf{y} = C(x_1 y_1, ..., x_D y_D)$$

- mocninná transformace

$$\alpha \odot \mathbf{x} = C(x_1^\alpha, ..., x_D^\alpha)$$

- Aitchisonův skalární součin

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

Aitchisonova geometrie na simplexu

- norma kompozice

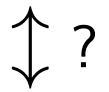
$$\|\mathbf{x}\|_a = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_a} = \sqrt{\frac{1}{D} \sum_{i<j} \left(\ln \frac{x_i}{x_j} \right)^2}$$

- Aitchisonova vzdálenost

$$\mathrm{d}_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_a = \sqrt{\frac{1}{D} \sum_{i<j} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2},$$

kde $\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus [(-1) \odot \mathbf{y}]$

Aitchisonova geometrie na simplexu



reálný prostor s eukleidovskou geometrií

Práce v souřadnicích

Uvažujeme určitou bázi a kompozice vyjádříme v souřadnicích vzhledem k této bázi.

- alr souřadnice - zobrazí kompozice ze simplexu do $\mathbb{R}^{D-1}$

$$\mathsf{alr}(\mathbf{x}) = \ln\left(\frac{x_1}{x_D}, \frac{x_2}{x_D}, ..., \frac{x_{D-1}}{x_D}\right)$$

- clr souřadnice - zobrazí kompozice do $\mathbb{R}^{D}$

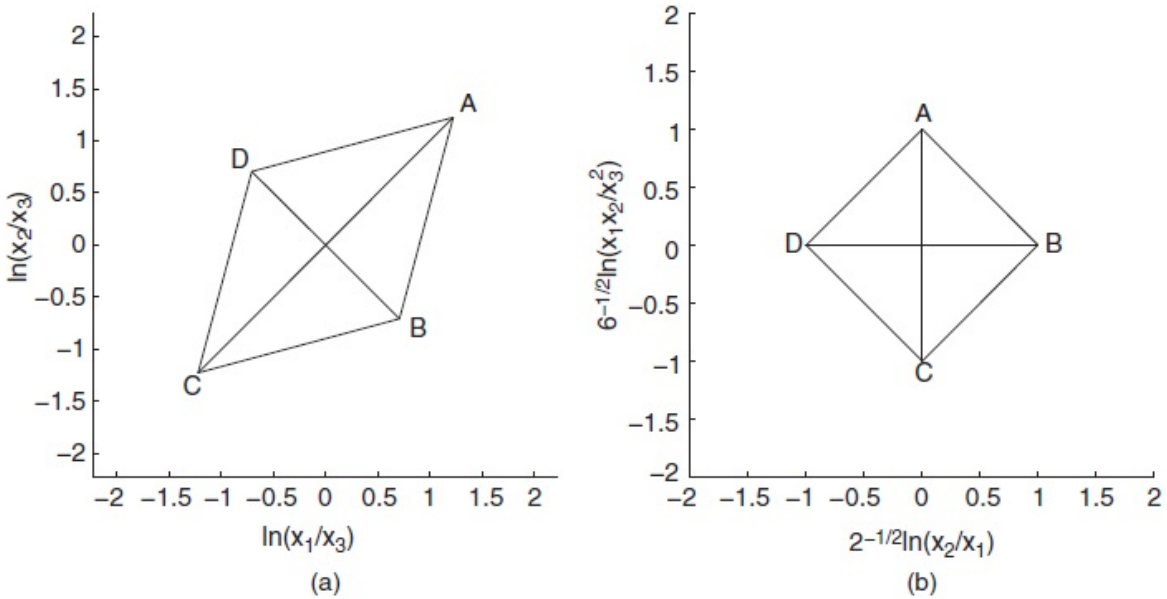
$$\mathsf{clr}(\mathbf{x}) = \ln\left[\frac{x_1}{\mathrm{g}(\mathbf{x})}, ..., \frac{x_D}{\mathrm{g}(\mathbf{x})}\right], \ \mathrm{g}(\mathbf{x}) = \left(\prod_{i=1}^{D} x_i\right)^{1/D}$$

Práce v souřadnicích

- ilr souřadnice

$$\text{ilr}(\mathbf{x}) = \left(\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, ..., \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a\right),$$

kde $\{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_{D-1}\}$ je ortonormální báze na simplexu



- ilr souřadnice

$$\mathbf{x} = (x_1, ..., x_D)' \rightarrow \mathbf{x}^{(l)} = (x_l, x_1, ..., x_{l-1}, x_{l+1}, ..., x_D)'$$

$$\rightarrow \mathbf{x}^{(l)} = (x_1^{(l)}, ..., x_l^{(l)}, x_{l+1}^{(l)}, ...x_D^{(l)})'$$

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}} \ \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}}, \quad i = 1, ..., D-1$$

- první souřadnice $z_1^{(l)}$ obsahuje veškerou relativní informaci, která se týká x_l ve vztahu k ostatním složkám, zbývající souřadnice $z_2^{(l)}, ..., z_{D-1}^{(l)}$ pak vysvětlují zbytek informace

Výskyt nul v kompozičních datech

- nuly vzniklé zaokrouhlením (rounded zeros)
- strukturní nuly (structural zeros)
- diskrétní nuly (count zeros)

Metody pro nuly vzniklé zaokrouhlením

- neparametrické nahrazení

$$xr_{ij} = \begin{cases} \delta_{ij}, & x_{ij} = 0, \\ x_{ij}\left(1 - \dfrac{\sum\limits_{k|x_{ik}=0} \delta_{ik}}{\kappa_i}\right), & \text{jinak} \end{cases}$$

```
> LPdata[c(1:3,91:93),]
     Cr  B   P   V   Cu    Ti   Ni  Y  Sr La  Ce  Ba  Li     K  Rb
1  33.3 23 393  47  5.3  3715  9.1 24  38  9 180  48  76 16617  53
2  64.7 35 978  83  9.1  4215 19.8 21  93 19 259  93  95 16527  64
3  30.4 23 433  42  3.8  3305 16.6 22  59 14 240  75  80 12209  53
91 11.2 14 105  35  0.0   884  0.0  3   5  0  37  10 224 30511 167
92  4.2  4  39  21  0.0  2005  0.0 12  33 10 139  45  86 20106  83
93 22.1 11 186  59  6.4  2525 15.5 10  38  7 197  74 236 33909 195

> DL
 Cr   B    P   V  Cu     Ti  Ni   Y  Sr  La   Ce   Ba   Li     K   Rb
3.6 3.0 37.0 9.0 2.0 549.0 6.3 3.0 5.0 1.0 18.0 10.0 22.0 159.0 15.0
```

```
XN <- multRepl(X,label=0,DL,delta=2/3)
```

```
> round(XN[c(1:3,91:93),],2)
     Cr  B   P   V    Cu    Ti   Ni  Y  Sr    La  Ce  Ba  Li     K  Rb
1  33.3 23 393  47  5.30  3715  9.1 24  38  9.00 180  48  76 16617  53
2  64.7 35 978  83  9.10  4215 19.8 21  93 19.00 259  93  95 16527  64
3  30.4 23 433  42  3.80  3305 16.6 22  59 14.00 240  75  80 12209  53
91 11.2 14 105  35  1.33   884  4.2  3   5  0.67  37  10 224 30511 167
92  4.2  4  39  21  1.33  2005  4.2 12  33 10.00 139  45  86 20106  83
93 22.1 11 186  59  6.40  2525 15.5 10  38  7.00 197  74 236 33909 195
```

- ilr-EM algoritmus
  - první fáze
    1. nahradit všechny nuly 65% detekčního limitu
    2. seřadit sloupce (složky) sestupně podle počtu nul
  - druhá fáze (pro $l = 1, ..., D$)
    3. vyjádřit kompozice v ilr souřadnicích $\rightarrow \mathbf{Z}^{(l)} = [\mathbf{z}_1^{(l)}, \mathbf{Z}_{-1}^{(l)}]$
    4. (robustní) lineární regrese $\mathbf{z}_1^{(l)}$ na $\mathbf{Z}_{-1}^{(l)} \rightarrow \widehat{\boldsymbol{\beta}}^{(l)}$
    5. přepočet nul v $\mathbf{z}_1^{(l)}$ pomocí parametrů z lineární regrese
    6. převést zpět na simplex
  - třetí fáze
    7. uspořádat kompozice podle původního pořadí

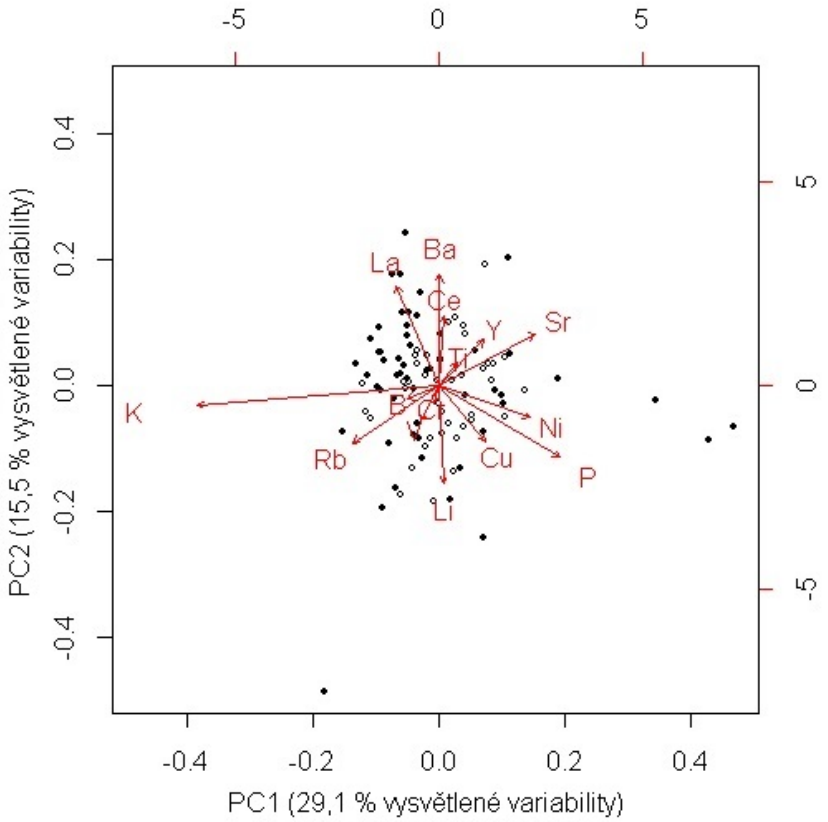
- ilr-EM algoritmus
  - nahrazení nul pomocí tzv. „censored regression“

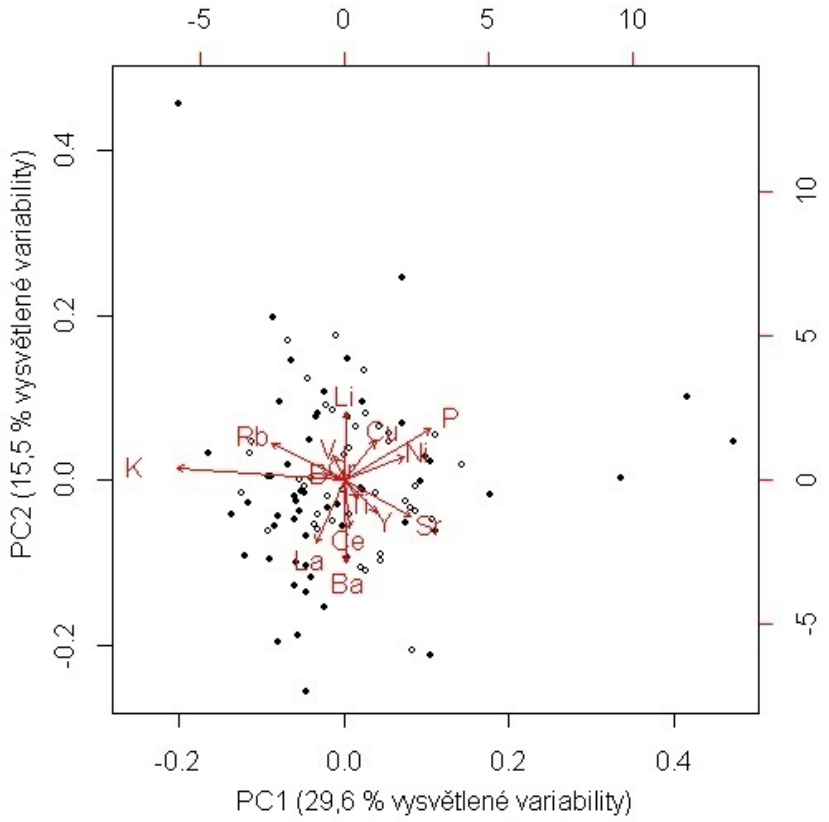
$$\hat{z}_{i1}^{(l)} = \mathbf{z}_{i,-1}^{(l)'}\hat{\boldsymbol{\beta}}^{(l)} - \hat{\sigma}^{(l)} \frac{\phi\left(\frac{\psi_{i1}^{(l)} - \mathbf{z}_{i,-1}^{(l)'}\hat{\boldsymbol{\beta}}^{(l)}}{\hat{\sigma}^{(l)}}\right)}{\Phi\left(\frac{\psi_{i1}^{(l)} - \mathbf{z}_{i,-1}^{(l)'}\hat{\boldsymbol{\beta}}^{(l)}}{\hat{\sigma}^{(l)}}\right)}$$

```
> fitI <- impRZilr(as.matrix(LPdata),dl=DL,method="lm")
> XI = fitI$x
```

```
> round(XI[c(1:3,91:93),],2)
     Cr  B   P  V   Cu   Ti    Ni  Y Sr   La  Ce Ba  Li     K  Rb
1  33.3 23 393 47 5.30 3715  9.10 24 38  9.0 180 48  76 16617  53
2  64.7 35 978 83 9.10 4215 19.80 21 93 19.0 259 93  95 16527  64
3  30.4 23 433 42 3.80 3305 16.60 22 59 14.0 240 75  80 12209  53
91 11.2 14 105 35 1.45  884  3.29  3  5  0.9  37 10 224 30511 167
92  4.2  4  39 21 1.11 2005  3.24 12 33 10.0 139 45  86 20106  83
93 22.1 11 186 59 6.40 2525 15.50 10 38  7.0 197 74 236 33909 195
```

Graf: Clr-biplot vycházející z matice s nahrazenými nulami pomocí neparametrického nahrazení (vlevo) a ilr-EM algoritmu (vpravo)





Metody pro strukturní nuly - detekce odlehlých hodnot

- dvoustupňový algoritmus

1. označit nuly jako chybějící hodnoty, nahradit vhodným algoritmem, vyjádřit matici v ilr souřadnicích
2. určit robustní odhady střední hodnoty a varianční matice pro nenulové složky
3. vypočítat Mahalanobisovy vzdálenosti

$$\mathrm{MD}(\tilde{\mathbf{z}}_i^*) = \left[(\tilde{\mathbf{z}}_i^* - \tilde{\mathbf{t}}_i^*)' \tilde{\mathbf{C}}_i^{*-1} (\tilde{\mathbf{z}}_i^* - \tilde{\mathbf{t}}_i^*)\right]^{\frac{1}{2}}$$

4. identifikovat odlehlá pozorování podle MD a příslušného kvantilu χ^2-rozdělení
5. detekce vzhledem ke struktuře strukturních nul

- dvoustupňový algoritmus
  - $\mathbf{x} = (x_1, ..., x_D)'$
  - $\tilde{\mathbf{x}} = (0, ..., 0, x_{j_1}, ..., x_{j_K})'$, $\tilde{\mathbf{x}} = \tilde{\mathbf{P}}\mathbf{x}$
  - $\mathbf{z} = \text{ilr}(\mathbf{x})$
  - $\tilde{\mathbf{z}} = \mathbf{Q}'\mathbf{z}$
  - odhady $\mathbf{t}$ a $\mathbf{C}$
  - $\tilde{\mathbf{t}} = \mathbf{Q}'\mathbf{t}$ a $\tilde{\mathbf{C}} = \mathbf{Q}'\mathbf{C}\mathbf{Q}$
  - výběr složek odpovídajícím nenulovým složkám $\longrightarrow \tilde{\mathbf{z}}^*$, $\tilde{\mathbf{t}}^*$, $\tilde{\mathbf{C}}^*$

Knihovny pro kompoziční data v softwaru R

- compositions
- zCompositions
- robCompositions

Literatura

- AITCHISON, J. The statistical analysis of compositional data. 1986.
- AITCHISON, J., BARCELÓ-VIDAL, C., MARTÍN-FERNÁNDEZ, J. A, PAWLOWSKY-GLAHN, V. Logratio analysis and compositional distance. *Mathematical Geology*, 2000, 32.3: 271-275.
- AITCHISON, J., GREENACRE, M. Biplots of compositional data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2002, 51.4: 375-392.
- FRY, J. M., FRY, T. R. L., MCLAREN, K. R. Compositional data analysis and zeros in micro data. *Applied Economics*, 2000, 32.8: 953-959.
- HRON, K. Elementy statistické analýzy kompozičních dat. *Informační bulletin České statistické společnosti*, 2010, 21.3: 41-48.
- MARTÍN-FERNÁNDEZ, J. A., HRON, K., TEMPL, M., FILZMOSER, P., PALAREA-ALBALADEJO, J. Model-based replacement of rounded zeros in compositional data: classical and robust approaches. *Computational Statistics & Data Analysis*, 2012, 56.9: 2688-2704.

- MONTI, G. S., HRON, K., TEMPL, M., FILZMOSER, P. Covariance-Based Outlier Detection for Compositional Data with Structural Zeros: Application to Italian Survey of Household Income and Wealth Data. *Advances in Latent Variables-Methods, Models and Applications*. 2013.
- PALAREA-ALBALADEJO, J., MARTÍN-FERNÁNDEZ, J. A. Values below detection limit in compositional chemical data. *Analytica chimica acta*, 2013, 764: 32-43.
- PALAREA-ALBALADEJO, J., MARTÍN-FERNÁNDEZ, J. A., BUCCIANTI, A. Compositional methods for estimating elemental concentrations below the limit of detection in practice using R. *Journal of Geochemical Exploration*, 2014, 141: 71-77.
- PAWLOWSKY-GLAHN, V., BUCCIANTI, A. *Compositional data analysis: Theory and applications*. John Wiley & Sons, 2011.
- PAWLOWSKY-GLAHN, V., EGOZCUE, J. J., TOLOSANA DELGADO, R. Lecture notes on compositional data analysis. 2007.
- TEMPL, M., HRON, K., FILZMOSER, P., MONTI, G. S. Outlier detection in compositional data with structural zeros. *ODAM 2013*, 2013, 61.
- TEMPL, M., HRON, K., FILZMOSER, P. Exploratory tools for outlier detection in compositional data with structural zeros. *Odesláno*, 2015.