Výběr proměnných v kompozičních datech

S. Donevska E. Fišerová P. Filzmoser K. Hron

ROBUST 2014

Kompoziční data

Kompoziční data (CoDa) = kvantitativní popis časti nějakého celku, tudíž data nesoucí pouze **relativní informaci**.

- Simplex s Aitchisonovou geometrií = výběrový prostor CoDa,

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

- Aitchisonova geometrie není úplně vhodná pro provádění standardních statistických metod.
  - $\Rightarrow$ Potřeba najít vhodnou reprezentaci CoDa v reálném prostoru.
- Navrženy transformace logaritmu podílů (logratio): aditivní logratio (alr) transformace, centrovaná logratio (clr) transformace a i izometrická logratio (ilr) transformace.
- CoDa se reprezentuji pomocí podílu složek.

Clr transformace

Clr transformace je izometrické zobrazeni mezi $\mathcal{S}^D$ a nadrovinou v $\mathbb{R}^D$,

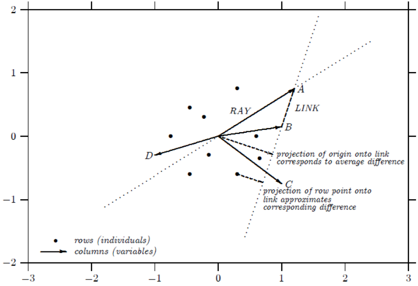
$$\mathbf{y} = clr(\mathbf{x}) = (y_1, y_2, ..., y_D)' = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, ..., \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right)' . \quad (1)$$

- Nevýhody clr proměnných:
  - nejsou souřadnice vzhledem k bázi na simplexu,
  - vedou k singulární varianční matici ($y_1 + \cdots + y_D = 0$),
  - nejsou subkompozičně soudržné.
- Výhody clr proměnných:
  - Převádí operace perturbace a mocninné transformace CoDa do obyčejného součtu a násobení skalárem clr koeficientů,
  - Euklidovská vzdálenost mezi vektory clr proměnných = Aitchisonova vzdálenost jejich příslušných kompozic. Platí i pro skalární součin a normu.

Interpretace CoDa biplotu

Kompoziční biplot je získaný jako standardní biplot pro clr transformovaná data.

1. Vzdálenost mezi vrcholy šipek jsou úměrné směrodatné odchylce logaritmu podílu příslušných složek.
2. Šipky představují clr proměnné.
3. Délka šipky je úměrná směrodatné odchylce clr proměnné.
4. Kosinus úhlu mezi dvěma vrcholy šípek dává hodnotu korelačního koeficientu mezi odpovídajícími logaritmy podílu.



Miry variability pro CoDa

Základní míra variability náhodné kompozice $\mathbf{x} = (x_1, \ldots, x_D)'$ je variační matice definovaná jako

$$\mathbf{T} = \left\{ \mathrm{var}\left(\ln \frac{x_i}{x_j} \right) \right\}_{i,j=1}^{D}.$$

- $\mathbf{T}$ je symetrická matice, která má na hlavní diagonále 0.
- Prvky matice $\mathbf{T}$ popisují variabilitu logaritmu podílu složek x_i a x_j.

(Normovaný) součet prvků variační matice $\mathbf{T}$ se nazývá celková variance,

$$\mathrm{totvar}(\mathbf{x}) = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \mathrm{var}\left(\ln \frac{x_i}{x_j} \right),$$

vyjadřuje celkovou variabilitu v kompozičním datovém souboru.

Miry variability

- Platí, že $\mathrm{totvar}(\mathbf{x}) = \sum_{i=1}^{D} \mathrm{var}(y_i)$, kde

$$\mathrm{var}(y_i) = \frac{D-1}{D^2} \sum_{j=1}^{D} \mathrm{var}\left(\ln \frac{x_j}{x_i}\right) - \frac{1}{2D^2} \sum_{\substack{j=1 \\ j \neq i}}^{D} \sum_{\substack{l=1 \\ l \neq i}}^{D} \mathrm{var}\left(\ln \frac{x_j}{x_l}\right),$$

$\Rightarrow$ Existuje celkem silný vztah mezi $\mathrm{var}(y_i)$ a součtem prvků v i-tém řádku (sloupce) příslušné variační matici $\mathbf{T}$.

Věta

Nechť jsou dány clr proměnné y_i a y_j, $i \neq j$, $i, j = 1, \ldots, D$. Potom, $\mathrm{var}(y_i) \geq \mathrm{var}(y_j)$, právě tehdy, když

$$\sum_{p=1}^{D} \mathrm{var}\left(\ln \frac{x_i}{x_p}\right) \geq \sum_{p=1}^{D} \mathrm{var}\left(\ln \frac{x_j}{x_p}\right).$$

Navržená kroková procedura

Uvažujme kompozici $\mathbf{x} = (x_1, \ldots, x_D)'$, takovou, že

$$\mathrm{var}(y_1) \geq \cdots \geq \mathrm{var}(y_D) \tag{2}$$

$$\Leftrightarrow$$

$$\sum_{p=1}^{D} \mathrm{var}\left(\ln \frac{x_1}{x_p}\right) \geq \sum_{p=1}^{D} \mathrm{var}\left(\ln \frac{x_2}{x_p}\right) \geq \cdots \geq \sum_{p=1}^{D} \mathrm{var}\left(\ln \frac{x_D}{x_p}\right). \tag{3}$$

Algoritmus:

1. Vynecháme složku x_D, jejíž rozptyl odpovídající clr proměnné je nejmenší.
   Uvažujeme tedy podkomozici $\mathbf{x}_1 = (x_1, \ldots, x_{D-1})'$ a provedeme clr transformaci na $\mathbf{x}_1$.
   Vypočítáme rozptyly clr transformovaných proměnných z $\mathbf{x}_1$.
2. Opakujeme krok 1.
3. STOP nejpozději po $D - 2$ krocích.

Navržená kroková procedura

- Bude dodrženo pořadí variancí clr proměnných pří přechodu z D-složkové k $D-1$- složkové podkompozici?

$\Rightarrow$ Ano, ale jen za splnění předpokladu

$$\mathrm{var}\left(\ln\frac{x_1}{x_D}\right) \geq \mathrm{var}\left(\ln\frac{x_2}{x_D}\right) \geq \cdots \geq \mathrm{var}\left(\ln\frac{x_{D-1}}{x_D}\right).$$

- Kdy máme skončit s výběrem kompozičních složek?

$\Rightarrow$ Poté co použijeme stop kritérium, které porovnává celkovou varianci podkompozici $\mathbf{x}_i$, získanou v i-tém kroku algoritmu $i = 1, \ldots, D-2$, s celkovou variancí podkompozici $\mathbf{x}_{i-1}$ z předchozího kroku.

Navržená kroková procedura – STOP kritérium

H_0 : $\text{totvar}(\mathbf{x}_i) = \text{totvar}(\mathbf{x}_{i-1})$ v.s. H_A : $\text{totvar}(\mathbf{x}_i) < \text{totvar}(\mathbf{x}_{i-1})$

- Za předpokladu platnosti nulové hypotézy

$$U_i^+ = \frac{\widehat{\mathrm{totvar}}(\mathbf{x}_i) - \mathrm{totvar}(\mathbf{x}_{i-1})}{\sqrt{\frac{2}{n-1}\mathrm{tr}\left(\widehat{\boldsymbol{\Sigma}}_i^2\right)}} \sim \mathrm{N}(0,1), \ \text{pro } i \to \infty,$$

kde $\widehat{\boldsymbol{\Sigma}}_i$ je výběrová varianční matice kompozice $\mathbf{x}_i$ v (libovolných) ilr souřadnicích.

- H_0 se zamítá na hladině významnosti α jestliže

$$u_i^+ \in \mathrm{W} = (-\infty, u_\alpha\rangle,$$

kde u_i^+ je realizace U_i^+ a u_α je α- kvantil normovaného normálního rozdělení.

- V každém kroku spočítáme U_i^+ proceduru ukončíme když $u_i^+ \in \mathrm{W}$.

Přiklad – Data Kola

Datový soubor Kola je výsledkem velkého geochemického projektu, který byl prováděný od 1992 do 1998 Geologickým průzkumným ústavem Finska a Norska, Centrální Kola expedicí, Rusko.

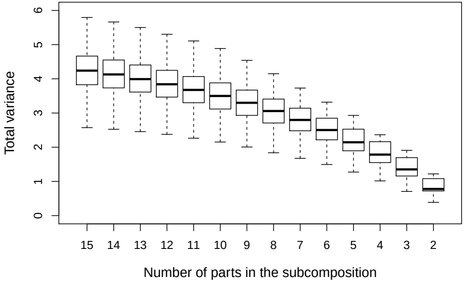
- Zkoumaná oblast: 188 000 km^2 na poloostrově Kola v Severní Evropě.
- Celkem je odebráno cca 600 vzorků ze 4 různých půdních horizontů (horizont O, horizont A, horizont B, horizont C).
- Analyzovaný jsou více než 50 chemických prvků v půdních horizontech.
- Data jsou dostupná v balíčku `StatDA` softwaru R (R Development Core Team, 2012).

Příklad – Data Kola – I. Experiment

- Náhodně vybereme 15 proměnných z 30 prvků z horizontu O.
- Algoritmus je aplikovaný dokud nedocílíme 2-složkovou podkompozici.
- V každém kroku je vypočítaná celková variance.
- Celá procedura je opakovaná 1000 krát.

Příklad – Data Kola – I. Experiment

- Náhodně vybereme 15 proměnných z 30 prvků z horizontu O.
- Algoritmus je aplikovaný dokud nedocílíme 2-složkovou podkompozici.
- V každém kroku je vypočítaná celková variance.
- Celá procedura je opakovaná 1000 krát.



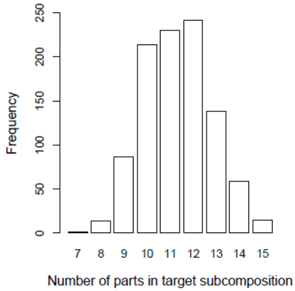
Obr: Celková variance podkompozic obdržená z postupné krokové procedury.

Příklad – Data Kola – II. experiment

- Znovu vybíráme náhodně 15 proměnných z 30 prvků z horizontu O.
- Postupná procedura je aplikovaná dokud testová statistika nenavrhne zastavení procesu.
- Celá procedura je opakovaná 1000 krát.

Příklad – Data Kola – II. experiment

- Znovu vybíráme náhodně 15 proměnných z 30 prvků z horizontu O.
- Postupná procedura je aplikovaná dokud testová statistika nenavrhne zastavení procesu.
- Celá procedura je opakovaná 1000 krát.



Obr: Sloupcový graf počtů složek výsledných podkompozic z postupné krokové procedury užitím STOP kritéria.

Příklad – Data Kola – II. experiment

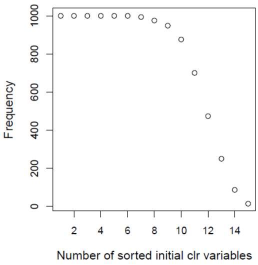
Obsahuje výsledná kompozice složky, které měly vysokou clr varianci v původní D-složkové kompozici?

- Složky ze všech 1000 počátečních podkompozic jsou vybrané dle jejích klesajících hodnot clr variancí.
- Počítáme, jak často jsou top k clr proměnné zahrnuty do výsledné kompozice, kde $k = 1, \ldots, 15$.

Příklad – Data Kola – II. experiment

Obsahuje výsledná kompozice složky, které měly vysokou clr varianci v původní D-složkové kompozici?

- Složky ze všech 1000 počátečních podkompozic jsou vybrané dle jejích klesajících hodnot clr variancí.
- Počítáme, jak často jsou top k clr proměnné zahrnuty do výsledné kompozice, kde $k = 1, \ldots, 15$.



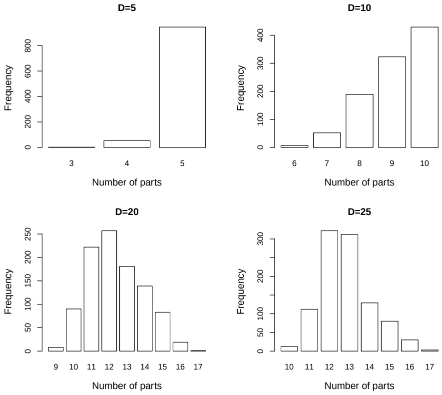
Obr: Clr proměnné z původní kompozice uspořádané dle klesajících variancí, vs. počet, kolikrát příslušné kompoziční složky byly zahrnuty ve výsledné podkompozicí.

Příklad – Data Kola – III. experiment

- Použijeme stejné simulační nastavení jako před tím, ale vybereme 5, 10, 20 a 25 - složkové počáteční kompozice z datového souboru Kola.
- Cela procedura je opakovaná 1000 krát.

Příklad – Data Kola – III. experiment

- Použijeme stejné simulační nastavení jako před tím, ale vybereme 5, 10, 20 a 25 - složkové počáteční kompozice z datového souboru Kola.
- Cela procedura je opakovaná 1000 krát.



Obr: Sloupcové grafy počtů složek výsledné podkompozice z krokové procedury užitím STOP – kritéria.

Příklad – Baltický průzkum půdy

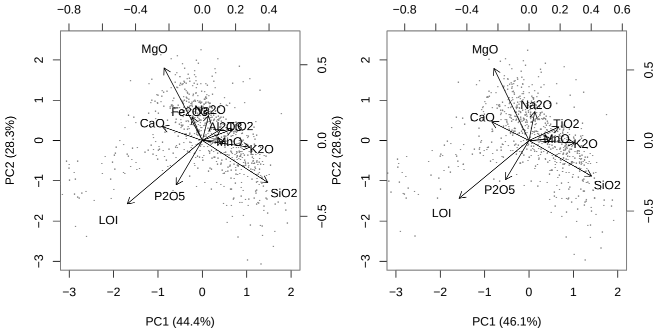
BSS datový soubor (Reimann et al., 2003), je ziskany z rozsáhlého geochemického projektu který, je realizovaný v Severní Evropě v oblasti rozsahu cca 1 800 000 km^2.

- 769 vzorků zemědělské půdy.
- Vzorky pocházejí ze dvou vrstev: horní vrstva (0-25cm) a spodní vrstva (50-75 cm).
- U všech vzorku je analyzovaná koncentrace z více než 40 chemických sloučenin.
- Používáme významné sloučeniny (Al_2O_3, Fe_2O_3, K_2O, MgO, MnO, CaO, TiO_2, Na_2O, P_2O_5 a SiO_2), plus LOI (Loss on ignition) z horní vrstvé tj., 11-ti složkovou CoDa.
- Datový soubor z horní a dolní vrstvy půdy je přístupny v baličku `mvoutlier` v R.

Příklad – Baltický průzkum půdy

Vliv postupné procedury je zobrazen pomocí CoDa biplotu na BSS datech.

- Pří použití testového kritéria pro $\alpha = 0.05$ jsme získali 9-složkovou podkompozici (Al_2O_3 a Fe_2O_3 jsou vyloučené).



Obr: Biplot pro BSS data pro všechny významné sloučeniny (vlevo) a po odstranění Al_2O_3 a Fe_2O_3 (vpravo).

Literatura

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

Egozcue JJ (2009) Reply to ”On the Harker Variation Diagrams; . . . ” by J. A. Cortés. Math Geosci 41:829–834.

Filzmoser P, Hron K, Reimann C (2012) Interpretation of multivariate outliers for compositional data. Computers & Geosciences 39: 77–85.

Hron K, Filzmoser P, Donevska S, Fišerová E (2013) Covariance based variable selection for compositional data. Mathematical geosciences 45:487–498.

Hron K, Kubáček L (2011) Statistical properties of the total variation estimator for compositional data. Metrika 74: 221–230.