BILANCE A BILANČNÍ DENDROGRAM KOMPOZIČNÍCH DAT

KLÁRA HRŮZOVÁ

Přírodovědecká fakulta Univerzity Palackého v Olomouci
Katedra matematické analýzy a aplikací matematiky

KOMPOZIČNÍ DATA

Kompoziční data jsou mnohorozměrná data nesoucí pouze relativní informaci. Zajímají nás tedy pouze podíly mezi složkami kompozičního vektoru. Výběrovým prostorem kompozičních dat je simplex, na němž byla zavedena Aitchisonova geometrie, která se liší od standardní euklidovské geometrie [1].

Základními operacemi Aitchisonovy geometrie jsou:

Perturbace: $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}\left[x_1 y_1, x_2 y_2, \ldots, x_D y_D\right]$;

Mocninná transformace: $\alpha \odot \mathbf{x} = \mathcal{C}\left[x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha\right]$;

Skalární součin: $\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}$;

Norma: $\|\mathbf{x}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j}\right)^2}$;

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

Zde $\mathbf{x} = [x_1, x_2, \ldots, x_D] \in \mathcal{S}^D$, $\mathbf{y} = [y_1, y_2, \ldots, y_D] \in \mathcal{S}^D$ jsou kompoziční vektory na D složkovém simplexu a $\alpha \in \mathbb{R}$ je kladná reálná konstanta, $C[\cdot]$ označuje operaci uzávěru a dolní index a označuje operaci Aitchisonovy geometrie.

Transformace kompozičních dat

Vzhledem ke geometrickým vlastnostem kompozičních dat a odlišnosti Aitchisonovy geometrie od euklidovské, nemůžeme použít standardní statistické metody pro jejich zpracování přímo na simplexu. Nejjednodušším řešením tohoto problému je transformovat data ze simplexu do reálného prostoru a takto transformovaná data analyzovat. Pokud je to nutné, pomocí inverzní transformace se můžeme vrátit zpět do simplexu [6].
Máme tři typy transformací založené na logaritmovaném podílu složek kompozice:

1. Centrované log-ratio transformace

$$\mathrm{clr}(\mathbf{x}) = \left[\ln \frac{x_1}{g(\mathbf{x})}, \ln \frac{x_2}{g(\mathbf{x})}, \ldots, \ln \frac{x_D}{g(\mathbf{x})}\right] = [\xi_1, \xi_2, \ldots, \xi_D]\,;$$

2. Aditivní log-ratio transformace

$$\mathrm{alr}(\mathbf{x}) = [a_1, a_2, \ldots, a_{D-1}] = \left[\ln \frac{x_1}{x_D}, \ln \frac{x_2}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D}\right];$$

3. Izometrické log-ratio transformace

$$\mathrm{ilr}(\mathbf{x}) = [z_1, z_2, \ldots, z_{D-1}]\,,\ z_i = \langle \mathbf{x}, \mathbf{e}_i \rangle_a\,,$$

kompozici tedy můžeme vyjádřit následovně: $\mathbf{x} = \bigoplus_{i=1}^{D-1} z_i \odot \mathbf{e}_i$, kde $\mathbf{e}_i$ je vektor ortonormální báze na simplexu a z_i je souřadnice vzhledem k ortonormální bázi.

BILANCE KOMPOZICE

Nyní se blíže zaměříme na ilr transformaci, která využívá ortonormální báze simplexu. $\mathcal{S}^D$ je vektorový prostor dimenze $D-1$. Pokud kompozice $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}$ v $\mathcal{S}^D$ splňuje

$$\|\mathbf{e}_i\|_a^2 = \langle \mathbf{e}_i, \mathbf{e}_i \rangle_a = 1, \langle \mathbf{e}_i, \mathbf{e}_j \rangle_a = 0, i, j = 1, 2, \ldots, D-1, i \neq j,$$

tvoří ortonormální bázi $\mathcal{S}^D$. V našem případě získáme ortonormální bázi (resp. souřadnice vzhledem k této bázi) použitím postupného binárního dělení na kompoziční vektor [5]. Tato metoda se zdá být nejvhodnější zejména z důvodu interpretace dosažených výsledků. Postupujeme tak, že v prvním kroku binárního dělení rozdělíme kompozici na dvě skupiny složek, v druhém kroku rozdělíme jednu skupinu z prvního pořadí na dvě skupiny; postup opakujeme tak dlouho, dokud v každé skupině nebude pouze jedna složka. Počet kroků, než postupné binární dělení složek dospěje ke konci, je $D-1$. Postupné binární dělení můžeme vyjádřit v tabulce, kdy v každém kroku označíme složky první skupiny symbolem '+', složky druhé skupiny '−' a složky, jichž se daný krok dělení netýká, jako 0.

Příklad SBP:
Mějme kompozici $\mathbf{x} = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8]$, jedná se (po řadě) o výdaje na bydlení, stravu, telefon, kulturu, zdraví, oděv, spoření a dopravu. Postupné binární dělení tohoto kompozičního vektoru může vypadat následovně:

Krok	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	+	+	−	−	+	−	−	+
2	+	+	0	0	−	0	0	−
3	+	−	0	0	0	0	0	0
4	0	0	0	0	+	0	0	−
5	0	0	−	−	0	−	+	0
6	0	0	−	+	0	−	0	0
7	0	0	+	0	0	−	0	0

Z každého kroku postupného binárního dělení získáme vektor ortonormální báze pomocí vzorce:

$$a = \frac{\sqrt{s}}{\sqrt{r(r+s)}}, b = \frac{-\sqrt{r}}{\sqrt{s(r+s)}},$$

kde r a s označují počet prvků v dané skupině. Jednotkový vektor ortonormální báze získaný v i-tém kroku pak vypadá následovně:

$$\mathbf{e}_i = \mathcal{C}\left[\exp\left(\underbrace{0, 0, \ldots, 0}_{k\ složek}, \underbrace{a, a, \ldots, a}_{r\ složek}, \underbrace{b, b, \ldots, b}_{s\ složek}, \underbrace{0, 0, \ldots, 0}_{j\ složek}\right)\right].$$

Projekce kompozice $\mathbf{x} \in \mathcal{S}^D$ na získané jednotkové kompoziční vektory ortonormální báze jsou počítány užitím skalárního součinu mezi danou kompozicí a odpovídajícím bázovým elementem, $z_i = \langle \mathbf{x}, \mathbf{e}_i \rangle_a$. Jsou to vlastně souřadnice $\mathbf{x}$ vzhledem k bilančnímu elementu $\mathbf{e}_i$, $i = 1, 2, \ldots, D-1$:

$$z_i = \ln \left[\frac{(x_{k+1} \cdots x_{k+r})^{\sqrt{s/(r(r+s))}}}{(x_{k+r+1} \cdots x_{k+r+s})^{\sqrt{r/(s(r+s))}}}\right] = \sqrt{\frac{rs}{r+s}} \ln \frac{\left(\prod_{j=1}^{r} x_j\right)^{1/r}}{\left(\prod_{j=r+1}^{r+s} x_j\right)^{1/s}}. \qquad (1)$$

Ze vztahu (1) je zřejmé, proč se tyto souřadnice nazývají bilance, jedná se o poměr dvou skupin složek kompozice. Pro interpretaci statistické analýzy jsou bilance vhodnější než jakékoliv jiné souřadnice, protože v postupném binárním dělení lze lehce vyčíst, jakým způsobem jsme k bilancím dospěli.

ILUSTRATIVNÍ PŘÍKLAD

Pro ukázku aplikace bilancí při analýze kompozičních dat jsme vybrali měsíční výdaje 27 mužů a 39 žen, které byly rozděleny do osmi složek: výdaje na bydlení, stravu, telefon, kulturu, zdraví, oděv, spoření a dopravu (data byla převzata z [2], získána pomocí dotazníku). Data tohoto typu byla použita i v [3], ovšem byla analyzována pomocí standardních statistických metod, které však na kompoziční data (pokud se nejedná o data transformovaná do reálného prostoru) nelze aplikovat. Ukážeme jednu z možností, jak tato data transformovat do reálného prostoru.
Pomocí speciální volby bilancí z [4] můžeme získat představu o rozdílu ve výdajích mezi muži a ženami. Bilance tedy spočítáme pomocí následující formule:

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

pro $i = 1, \ldots, D-1$, $l = 1, \ldots, D$. Zde l označuje, o kterou složku v pořadí se jedná a i označuje krok v postupném binárním dělení. Pro naši potřebu bude stačit $z_1^{(l)}$, tedy pro každou složku vypočítáme bilanci dané složky vůči všem ostatním (v 1. kroku postupného binárního dělení oddělujeme l-tou složku v pořadí od všech ostatních).
Bilance jsou shrnuty do tabulky, která obsahuje jejich průměrné hodnoty (*směrodatné odchylky*):

Bilance	Bydlení	Strava	Telefon	Kultura	Zdraví	Oděv	Spoření	Doprava
Muži	1,22 (*0,76*)	1,072 (*0,46*)	-0,37 (*0,56*)	0,19 (*0,65*)	-1,73 (*0,78*)	-0,53 (*0,61*)	-0,07 (*0,93*)	0,23 (*0,70*)
Ženy	1,45 (*0,52*)	0,98 (*0,51*)	-0,53 (*0,52*)	-0,24 (*0,74*)	-1,15 (*0,66*)	-0,18 (*0,54*)	0,14 (*0,73*)	-0,46 (*0,79*)

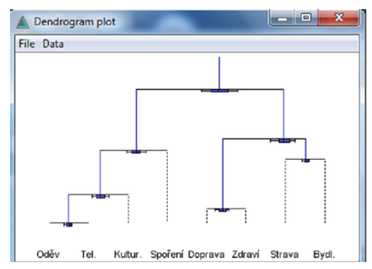
Získané výsledky nám ukazují poměr konkrétního výdaje v domácnosti vůči všem ostatním. Můžeme si všimnout některých zajímavostí:

- relativní náklady na bydlení dominují všem ostatním výdajům;
- strava se řadí mezi druhé nejvyšší výdaje, ženy do stravy neinvestují tolik jako muži;
- mezi nejnižší relativní výdaje patří náklady na zdraví, můžeme se domnívat, že anketu vyplňovali mladí lidé, kteří na zdraví ještě tolik nemyslí a mají nižší náklady na léky, ženy však do zdraví investují o něco více než muži;
- nejvýraznějšího rozdílu ve výdajích mezi muži a ženami si můžeme všimnout u položek dopravy a kultury, muži do těchto dvou položek vkládají více finančních prostředků;
- další zajímavostí je i to, že ženy přes veškeré předsudky spoří více než muži, oproti tomu se potvrdilo nadšení žen pro nakupování oblečení.

S interpretací by se dalo lépe pracovat, pokud bychom měli k dispozici věk, vzdělání a druh zaměstnání respondentů. Navíc musíme vzít do úvahy různé hodnoty směrodatných odchylek jednotlivých bilancí.

Bilanční dendrogram

Bilanční dendrogram je nástroj pro zobrazení postupného binárního dělení, bilancí, dekompozice celkového rozptylu (rovného součtu rozptylů jednotlivých bilancí) a dalších jednorozměrných charakteristik [7]. Lze snadno vytvořit pomocí programu CoDaPack, který je volně dostupný na internetové adrese http://ima.udg.edu/codapack/.



Data byla rozdělena v postupném binárním dělení (viz Příklad SBP): v prvním kroku byly odděleny výdaje nezbytné (bydlení, strava, zdraví, doprava) od zbytných (telefon, kultura, oděv, spoření), ve druhém kroku pak byly odděleny nezbytné výdaje, které jsou životně důležité (tzn. bydlení a strava) od zbylých. V třetím a čtvrtém kroku od sebe oddělíme všechny ostatní nezbytné výdaje. V pátém kroku oddělíme spoření od ostatních výdajů kvůli důležitosti do budoucna. V šestém kroku oddělíme kulturu a nakonec zbylé dva - oděv a telefon.

Ze získaného bilančního dendrogramu je zřejmé, že největší rozptyl je u bilance dopravy a zdraví. Není tedy souvislost mezi těmito dvěma výdaji; někteří respondenti investují do dopravy, někteří mohou být naopak nemocní, a proto musí nakupovat léky a častěji navštěvovat lékaře. Nejmenší rozptyl je naopak u stravy a bydlení, to znamená, že podíl vydajů na stravu a bydlení je v zásadě stabilní, zároveň se jedná o dvě největší položky v rozpočtu jakékoliv domácnosti. Dále je zajímavé si všimnout rozptylu první bilance, kde poměřujeme výdaje zbytné a nezbytné, očekávali bychom asi menší stabilitu těchto dvou skupin složek.

Použité zdroje

[1] Aitchison, J., *The Statistical analysis of compositional data*. Chapman and Hall, London, New York, 1986.

[2] Brodinová, Š., *Diskriminační analýza pro kompoziční data*. Univerzita Palackého, Olomouc, 2012.

[3] Deaton, A., *Looking for boy-girl discrimination in household expenditure data*. The World Bank Economic Review 3, 1989, 1-15.

[4] Filzmoser, P., Hron, K., Reimann, C., *Interpretation of multivariate outliers for compositional data*. Computers & Geosciences 39, 2012, 77-85.

[5] Pawlowsky-Glahn, V., Egozcue, J. J., *Groups of parts and their balances in compositional data analysis*. Mathematical Geology 37, 2005, 795-828.

[6] Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, R., *Lecture notes on compositional data analysis* [online], dostupné z: http://dugi-doc.udg.edu/bitstream/10256/297/1/CoDa-book.pdf

[7] Thió-Henestrosa, S., Egozcue, J. J., Pawlowsky-Glahn, V., Kovács, L. Ó., Kovács, G. P.: *Balance-dendrogram. A new routine of CoDaPack*. Computers & Geosciences 34, 2008, 1682-1696.