DIRICHLETOVO ROZDĚLENÍ VZHLEDEM K AITCHISONOVĚ MÍŘE NA SIMPLEXU

Petra Kynčlová
kynclova.petra@gmail.com
Department of Statistics and Probability Theory,
Vienna University of Technology, Austria

Většina standardních statistických metod předpokládá, že zkoumaná data pocházejí z reálného prostoru s euklidovskou geometrií. Geometrická struktura simplexu, výběrového prostoru kompozičních dat, je přitom odlišná a je charakterizována tzv. Aitchisonovou geometrií. Z tohoto důvodu se pro simplexový výběrový prostor zavádí alternativní, relativní míra. Tato míra se označuje jako Aitchisonova míra a je zavedena pomocí transformace Lebesgueovy míry z prostoru ortonormálních souřadnic na simplex. Jako vhodný nástroj pro parametrické modelování kompozičních dat se tradičně uvádí Dirichletovo rozdělení, jelikož předpokládá simplex jako výběrový prostor. Jeho hustota je ovšem typicky vyjádřena vzhledem k Lebesgueově míře. Cílem příspěvku je popsat vlastnosti a číselné charakteristiky Dirichletova rozdělení na simplexu vzhledem k Aitchisonově míře, resp. vzhledem k Lebesgueově míře v prostoru ortonormálních souřadnic, a důsledky volby parametrů na tvar Dirichletova rozdělení.

KOMPOZIČNÍ DATA

- D-složkovou kompozicí rozumíme kladný reálný vektor $\mathbf{x} = (x_1, \ldots, x_D)'$, jehož složky nesou výhradně relativní informaci.
- Kompozice můžeme reprezentovat jako data s konstantním součtem, tj.

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

- Výběrový prostor reprezentací kompozic při zvoleném κ je **simplex**,

$$\mathcal{S}^D = \left\{ (x_1, x_2, \ldots, x_D)' : x_1 > 0, x_2 > 0, \ldots, x_D > 0; \sum_{i=1}^{D} x_i = \kappa \right\}.$$

- Při práci s kompozicemi se používá **Aitchisonova geometrie na simplexu** [1], která je reprezentována operacemi *perturbace*, *mocnění* a *Aitchisonův skalární součin*

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}\left(x_1 y_1, x_2 y_2, ..., x_D y_D\right)',$$

$$\alpha \odot \mathbf{x} = \mathcal{C}\left(x_1^{\alpha}, x_2^{\alpha}, ..., x_D^{\alpha}\right)',$$

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

- Zavedení Aitchisonovy geometrie na simplexu zaručuje existenci ortonormální báze $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$. **Izometrická logratio (ilr) transformace** kompozice $\mathbf{x}$ představuje souřadnice jakékoliv kompozice $\mathbf{x}$ vzhledem k ortonormální bázi $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$ [2], tj

$$\mathrm{ilr}(\mathbf{x}) = \left(\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, \langle \mathbf{x}, \mathbf{e}_2 \rangle_a, \ldots, \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a \right)'.$$

Jednou konkrétní volbou ortonormální báze dostaneme ilr souřadnice

$$\mathrm{ilr}(\mathbf{x}) = \mathbf{z} = (z_1, \ldots, z_{D-1})', \;\; z_i = \sqrt{\frac{i}{i+1}} \ln \frac{\sqrt[i]{\prod_{i=1}^{j} x_j}}{x_{i+1}}.$$

ČÍSELNÉ CHARAKTERISTIKY NA SIMPLEXU

- **Střed kompozice**, střední hodnota náhodné kompozice, je definována pomocí geometrické interpretace střední hodnoty náhodného vektoru [4]. Představuje kompozici $\mathrm{cen}(\mathbf{x})$, která minimalizuje výraz $\mathrm{E}[d_a^2(\mathbf{x}, \mathrm{cen}(\mathbf{x}))]$, tj.

$$\mathrm{cen}(\mathbf{x}) = \mathcal{C}(\exp(\mathrm{E}[\ln \mathbf{x}])).$$

Konkrétně pro izometrickou logratio transformaci platí

$$\mathrm{ilr}(\mathrm{cen}[\mathbf{x}]) = \mathrm{E}[\mathrm{ilr}(\mathbf{x})].$$

- **Metrický rozptyl** udává variabilitu náhodné kompozice jako střední hodnotu čtvercové Aitchisonovy vzdálenosti kompozice od jejího středu, tj.

$$\mathrm{Mvar}[\mathbf{x}] = \mathrm{E}[d_a^2(\mathbf{x}, \mathrm{cen}[\mathbf{x}])].$$

Při použití izometrické logratio transformace pro metrický rozptyl platí

$$\mathrm{Mvar}[\mathbf{x}] = \mathrm{E}[d_e^2(\mathrm{ilr}(\mathbf{x}), \mathrm{ilr}(\mathrm{cen}[\mathbf{x}]))].$$

AITCHISONOVA MÍRA

- Hustoty rozdělení pravděpodobnosti pro data z reálného prostoru s euklidovskou geometrií jsou vyjádřeny vzhledem k Lebesgueově pravděpodobnostní míře. Geometrická struktura výběrového prostoru však může být v některých případech odlišná a je tedy nutné pracovat s jinou mírou než právě s Lebesgueovou.
- Nechť je dán vektorový prostor E, na kterém je zaveden skalární součin. Zde můžeme zavést pravděpodobnostní míru λ_E, jež bude se strukturou prostoru E kompatibilní, a to prostřednictvím Lebesgueovy míry na ortonormálních souřadnicích. Funkce hustoty f_E, která je definována na E, je pak dána jako Radon-Nikodýmova derivace pravděpodobnostní míry P vzhledem k míře λ_E. Míra λ_E má v prostoru E stejné vlastnosti jako Lebesueova míra v reálném prostoru (tedy v prostoru ortonormálních souřadnic) [3].
- Stejným způsobem je zavedena i **Aitchisonova míra** λ_a, která odpovídá geometrické struktuře simplexu. Míra λ_a je relativní a je absolutně spojitá vzhledem k Lebesgueově míře λ. Vztah mezi mírami λ_a a λ je dán pomocí Jakobiánu

$$\frac{d\lambda_a}{d\lambda} = \frac{1}{\sqrt{D} x_1 \cdots x_D}.$$

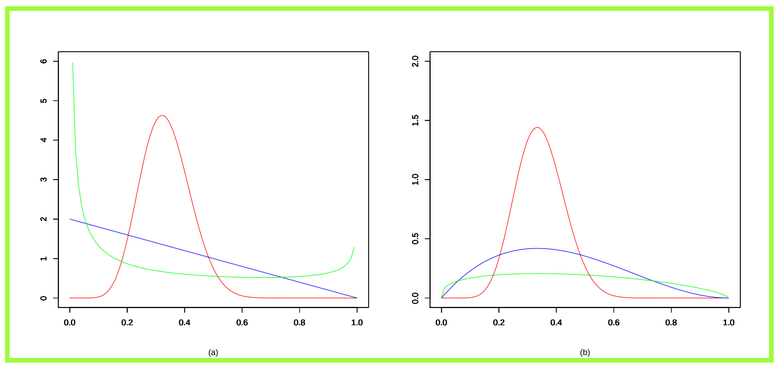
DIRICHLETOVO ROZDĚLENÍ NA SIMPLEXU

- Náhodný vektor $\mathbf{X} \in \mathcal{S}^D$ má D-rozměrné Dirichletovo rozdělení (vzhledem k Lebesgueově míře na simplexu) s parametrem $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)' \in \mathbb{R}_+^D$, jestliže jeho hustota pravděpodobnosti má tvar

$$f(\mathbf{x}) = \frac{dP}{d\lambda}(\mathbf{x}) = \frac{\Gamma(\alpha_+)}{\prod_{i=1}^{D} \Gamma(\alpha_i)} \prod_{i=1}^{D} x_i^{\alpha_i - 1},$$

kde λ je Lebesgueova míra, $\alpha_+ = \sum_{i=1}^{D} \alpha_i$, a Γ je gamma funkce. Značíme $\mathbf{X} \sim \mathcal{D}^D(\boldsymbol{\alpha})$ [4].
Zároveň můžeme hustotu Dirichletova rozdělení vyjádřit vzhledem k Aitchisonově míře λ_a ve tvaru

$$f_a(\mathbf{x}) = \frac{dP}{d\lambda_a}(\mathbf{x}) = \frac{\Gamma(\alpha_+)\sqrt{D}}{\prod_{i=1}^{D} \Gamma(\alpha_i)} \prod_{i=1}^{D} x_i^{\alpha_i}.$$



Obrázek 1. Hustoty Dirichletova rozdělení vzhledem (a) k Lebesgueově míře λ a (b) k Aitchisonově míře λ_a s parametry $\boldsymbol{\alpha} = (10, 20)'$ (——), $\boldsymbol{\alpha} = (1, 2)'$ (——) a $\boldsymbol{\alpha} = (1/3, 2/3)'$ (——).

- Obrázek 1 znázorňuje hustoty Dirichletova rozdělení pro případ $D = 2$. Vzhledem k Aitchisonově míře λ_a dostáváme vždy unimodální funkci. Pro hustotu vzhledem k Lebesgueově míře λ to neplatí. Pro případ, že jsou všechny složky $\alpha_i < 1$, má funkce vertikální asymptoty v 0 a 1, speciálně pro $\boldsymbol{\alpha} = (1, 1)'$ je hustota konstantní.
- Odlišnosti jsou patrné i při výpočtu charakteristik Dirichletova rozdělení.
  - Modus a střední hodnota vzhledem k Lebesgueově míře

$$\mathrm{modus}(\mathbf{X}) = \left(\frac{\alpha_1 - 1}{\alpha_+ - D}, \ldots, \frac{\alpha_D - 1}{\alpha_+ - D} \right)',$$

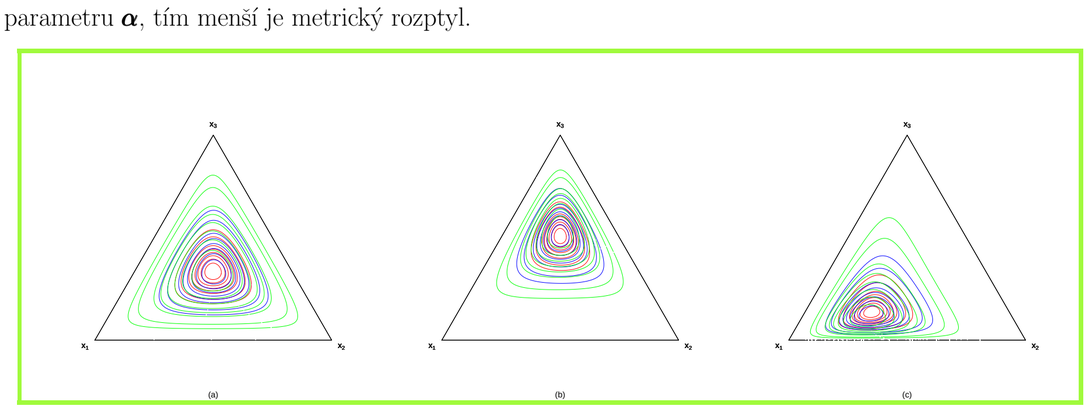
$$\mathrm{E}(\mathbf{X}) = \left(\frac{\alpha_1}{\alpha_+}, \ldots, \frac{\alpha_D}{\alpha_+} \right)'.$$

  - Modus a střední hodnota vzhledem k Aitchisonově míře

$$\mathrm{modus}_a(\mathbf{X}) = \left(\frac{\alpha_1}{\alpha_+}, \ldots, \frac{\alpha_D}{\alpha_+} \right)',$$

$$\mathrm{E}(\mathbf{X})_a = \mathcal{C}\left(e^{\psi(\alpha_1)}, \ldots, e^{\psi(\alpha_D)} \right)'.$$

- Analogické chování platí i v případě $D = 3$ - hustoty Dirichletova rozdělení vzhledem k Aitchisonově míře λ_a na simplexu jsou vždy unimodální (i v $D = 2$ jsme měli simplex). V případě, že volíme parametr $\boldsymbol{\alpha}$ v rámci třídy ekvivalentních kompozic, pak mají hustoty vždy stejný modus. Vzhledem k Lebesgueově míře ani jedno tvrzení neplatí. Unimodalita zde nastává pouze v případě, že jsou všechny složky kompozice $\boldsymbol{\alpha}$ větší než jedna, ale modus stejný není. Pro metrický rozdíl obecně platí, že čím větší jsou hodnoty složek parametru $\boldsymbol{\alpha}$, tím menší je metrický rozptyl.



Obrázek 2. Hustoty Dirichletova rozdělení vzhledem k Aitchisonově míře λ_a na simplexu s parametry
(a) $\boldsymbol{\alpha} = (10, 10, 10)'$ (——), $\boldsymbol{\alpha} = (5, 5, 5)'$ (——), $\boldsymbol{\alpha} = (2, 2, 2)'$ (——),
(b) $\boldsymbol{\alpha} = (10, 10, 20)'$ (——), $\boldsymbol{\alpha} = (5, 5, 10)'$ (——), $\boldsymbol{\alpha} = (2.5, 2.5, 5)'$ (——),
(c) $\boldsymbol{\alpha} = (20, 10, 5)'$ (——), $\boldsymbol{\alpha} = (10, 5, 2.5)'$ (——), $\boldsymbol{\alpha} = (5, 2.5, 1.25)'$ (——).

Literatura

[1] J. Aitchison (1986). *The Statistical Analysis of Compositional Data.* Chapman & Hall, London.
[2] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, C. Barceló-Vidal (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3), 279–300.
[3] G. Mateu-Figueras, V. Pawlowsky-Glahn, J.J. Egozcue (2011). The principle of working on coordinates. In: V. Pawlowsky-Glahn, A. Buccianti, eds. *Compositional Data Analysis: Theory and Applications*, pp. 31–42, Wiley, Chichester.
[4] G.S. Monti, G. Mateu-Figueras, V. Pawlowsky-Glahn, J.J. Egozcue (2011). The shifted-scaled Dirichlet distribution in the simplex, In: J.J. Egozcue, R. Tolosana-Delgado, M.I. Ortego, eds. *Compositional Data Analysis Workshop – CoDaWork'11, Proceedings*, International Center for Numerical Methods in Engineering (CIMNE), Barcelona.