UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

DIPLOMOVÁ PRÁCE

Rozdělení na simplexu

Vedoucí diplomové práce:
RNDr. Karel Hron, Ph.D.
Rok odevzdání: 2012

Vypracovala:
Bc. Petra Kynčlová
AME, II. ročník

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením RNDr. Karla Hrona, Ph.D. s použitím uvedené literatury.

V Olomouci dne 21. března 2012

Poděkování

Na tomto místě bych chtěla hlavně poděkovat svému vedoucímu diplomové práce RNDr. Karlu Hronovi, Ph.D., za jeho nesmírnou trpělivost, užitečné rady a čas, který mi věnoval při konzultacích. Poděkování patří rovněž všem, kteří mě během mého studia podporovali.

Obsah

1. Úvod

V mé diplomové práci jsem se zabývala typy rozdělení pravděpodobnosti na simplexu, který představuje výběrový prostor pro kompoziční data. Tento prostor je specifický, jelikož kompoziční data popisují relativní příspěvky částí na daném celku, bez jehož znalosti tato data ztrácejí svůj význam. Z tohoto důvodu musí být statistický přístup ke kompozičním datům odlišný od klasické statistické analýzy.

Cílem mé práce bylo především zjistit, zda jsou modely Dirichletova rozdělení, posunutého Dirichletova rozdělení a posunutého škálovaného Dirichletova rozdělení vhodné pro práci s kompozičními daty jako alternativa k tzv. normálnímu rozdělení na simplexu. Do současné doby nebyly tyto přístupy komplexně popsány a nebyla zavedena jednotná terminologie. K rozdělením na simplexu jsem přistupovala pomocí funkce hustoty, a to vzhledem k Aitchisonově míře na simplexu a vzhledem k Lebesgueově míře v prostoru ortonormálních souřadnic. Součástí práce bylo sjednocení terminologie a vyhodnocení praktické využitelnosti Dirichletova rozdělení pro popis kompozičních dat.

Práce je členěna do tří kapitol. V první části jsem se zabývala popisem užívaných spojitých mnohorozměrných rozdělení. Zvláštní pozornost byla věnována normálnímu a Dirichletovu rozdělení, z jejichž principů vychází i formy těchto rozdělení definovaných na simplexu. Druhá část je zaměřena na popis kompozičních dat a specifikaci Aitchisonovy geometrie na simplexu, na kterém jsou tato data definována. Zavěrečná kapitola je věnována typům rozdělení pravděpodobnosti na simplexu. Podstatná část kapitoly popisuje chování Dirichletova rozdělení, posunutého Dirichletova rozdělení a posunutého škálovaného Dirichletova rozdělení na simplexu.

Práce byla vysázena použitím typografického softwaru TEXLive. Grafy a obrázky vznikly ve statistickém softwaru R.

2. Užívaná mnohorozměrná rozdělení

Existuje mnoho mnohorozměrných modelů rozdělení pravděpodobnosti používaných pro popis rozdělení náhodného vektoru, resp. rozdělení příslušného náhodného výběru. Mezi nejpoužívanější patří normální rozdělení. Dále se ve spojitém případě využívá také Studentovo mnohorozměrné rozdělení nebo (pro diskrétní vektory) multinomické rozdělení [6].

Tato diplomová práce se zabývá rozdělením pravděpodobnosti na simplexu, který tvoří přirozený výběrový prostor pro kompoziční data. Z tohoto důvodu se následující kapitola věnuje popisu dvou významných mnohorozměrných rozdělení, normálním a Dirichletovým, které se k popisu rozdělení kompozičních dat mohou teoreticky využívat.

2.1. Normální rozdělení

Normální rozdělení patří v jednorozměrném případě mezi nejpoužívanější a nejdůležitější rozdělení pravděpodobnosti náhodné veličiny X [3,6]. Někdy se také normální rozdělení označuje jako Gaussovo rozdělení nebo zákon chyb. Široká škála využití normálního rozdělení plyne z faktu, že většina náhodných dějů, které se vyskytují v přírodě nebo ve společnosti, se modeluje právě normálním rozdělením. Jako příklad uveďme třeba výšku jedinců v populaci nebo chyby měření. Důležitost normálního rozdělení je vidět i z toho, že jsou z něj odvozena jiná často používaná rozdělení jako je χ^2, Studentovo či Fisherovo rozdělení [3,6]. V praxi bývá též normální rozdělení používáno jako aproximace jiných pravděpodobnostních rozdělení spojitého i diskrétního typu [3,6].

V následující kapitole jsou shrnuty základní poznatky o jednorozměrné i mnohorozměrné formě normálního rozdělení.

2.1.1. Jednorozměrné normální rozdělení

Jednorozměrné normální rozdělení je spojité rozdělení pravděpodobnosti náhodné veličiny X, které je charakterizováno pomocí dvou parametrů. První

parametr odpovídá střední hodnotě $\mu \in \mathbb{R}$ a druhý rozptylu $\sigma^2 \in \mathbb{R}_0^+$. Říkáme, že náhodná veličina X má jednorozměrné normální rozdělení s parametry μ a σ^2, jestliže její hustota má tvar

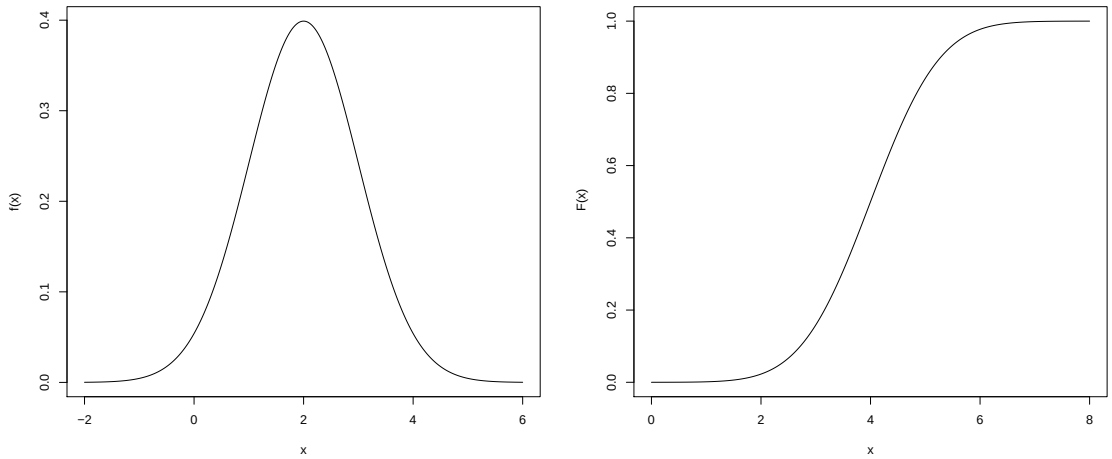
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{ - \frac{(x-\mu)^2}{2\sigma^2} \Big\}, \qquad x \in \mathbb{R}.$$

Má-li náhodná veličina X jednorozměrné normální rozdělení, symbolicky tuto skutečnost zapisujeme $X \sim \mathcal{N}(\mu, \sigma^2)$, kde $\mu = \mathrm{E}X$ a $\sigma^2 = \mathrm{var}X$ [3].

Hustota pravděpodobnosti normálního rozdělení je symetrická kolem své střední hodnoty, kde funkce nabývá svého maxima. Střední hodnota jako parametr polohy nám udává, kde se budou hodnoty nejčastěji pohybovat, budeme-li náhodně opakovat pokus řídící se normálním rozdělením. Naopak rozptyl představuje parametr variability a značí, v jak úzkém okolí střední hodnoty se naměřené hodnoty vyskytují. Body inflexe funkce jsou dány jako $\mu - \sigma$ a $\mu + \sigma$ (obrázek 1).

Distribuční funkci normálního rozdělení získáme, když zintegrujeme funkci hustoty, tj.

$$F(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \exp\Big\{ - \frac{(t-\mu)^2}{2\sigma^2} \Big\} dt, \qquad x \in \mathbb{R}.$$



Obrázek č. 1: Funkce hustoty $f(x)$ a distribuční funkce $F(x)$ jednorozměrného normálního rozdělení s parametry $\mu = 2$ a $\sigma^2 = 1$.

Speciálním typem jednorozměrného normálního rozdělení je tzv. normované normální rozdělení. Jestliže $Y \sim \mathcal{N}(\mu, \sigma^2)$, pak náhodná veličina

$$X = \frac{Y - \mu}{\sigma}$$

má normované normální rozdělení s parametry $\mu = 0$ a $\sigma^2 = 1$, značíme $X \sim \mathcal{N}(0, 1)$. Dosazením hodnot parametrů získáme funkci hustoty

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\Big\{ -\frac{x^2}{2} \Big\}, \qquad x \in \mathbb{R},$$

a distribuční funkci

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\Big\{ -\frac{t^2}{2} \Big\} dt, \qquad x \in \mathbb{R}.$$

Normované normální rozdělení se ve statistice používá jako časté rozdělení výběrových funkcí, užívaných ke konstrukci intervalových odhadů nebo k testování parametrických hypotéz [3].

2.1.2. Mnohorozměrné normální rozdělení

Mezi nejpoužívanější spojitá rozdělení v oblasti vícerozměrné matematické statistiky patří mnohorozměrné normální rozdělení, též označované jako mnohorozměrné gaussovské rozdělení [3].

Mnohorozměrné normální rozdělení náhodného vektoru $\mathbf{X}$ je zobecněnou formou jednorozměrného normálního rozdělení náhodné veličiny X. Nechť $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ je p-rozměrný náhodný vektor, kde $p \geq 2$ a jehož složky jsou náhodné veličiny.

K definici mnohorozměrného normálního rozdělení můžeme využít toho, že dokážeme určit rozdělení náhodného vektoru, pokud známe rozdělení každé jeho lineární kombinace.

Definice 2.1. *Nechť je dán náhodný vektor* $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$*, dále nechť* $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ *je číselný vektor středních hodnot a* $\boldsymbol{\Sigma} = (\sigma_{ij}) \in \mathbb{R}^{p \times p}$ *je symetrická pozitivně definitní matice. Vektor* $\mathbf{X}$ *má p-rozměrné normální rozdělení*

s parametry $\boldsymbol{\mu}$ *a* $\boldsymbol{\Sigma}$*, jestliže pro libovolný vektor* $\mathbf{c} \in \mathbb{R}^p$ *platí, že* $\mathbf{c}'\mathbf{X}$ *má jednorozměrné normální rozdělení s parametry* $(\mathbf{c}'\boldsymbol{\mu}, \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c})$*.*

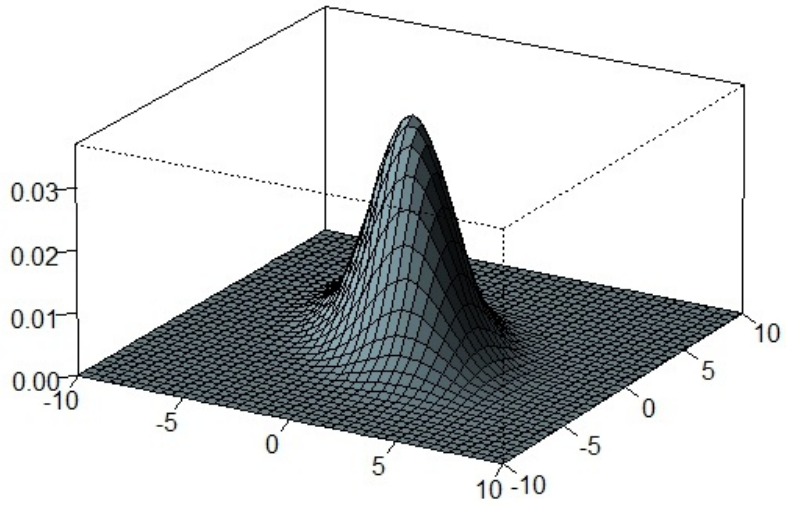
Hodnost matice $h(\boldsymbol{\Sigma})$ se nazývá řád rozdělení. V případě, že je matice $\boldsymbol{\Sigma}$ regulární, tj. její determinant je nenulový, hovoříme o regulárním p-rozměrném normálním rozdělení. V případě, že je determinant matice $\boldsymbol{\Sigma}$ nulový, pak mluvíme o singulárním p-rozměrném normálním rozdělení. Podobně jako u jednorozměrného normálního rozdělení můžeme symbolicky zapsat, že $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Poznemenejme, že $\boldsymbol{\mu}$ je zároveň střední hodnotou a $\boldsymbol{\Sigma}$ varianční maticí tohoto rozdělení [3].

Jiný způsob, kterým můžeme definovat vícerozměrné normální rozdělení, je pomocí funkce hustoty [3]. Řekneme, že náhodný vektor $\mathbf{x}$ má p-rozměrné normální rozdělení, má-li jeho hustota tvar

$$f_{\mathbf{x}}(\mathbf{x}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\Big[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big].$$

kde $\mathbf{X} \in \mathbb{R}^p$, $\boldsymbol{\mu} \in \mathbb{R}^p$ je vektor středních hodnot a $|\boldsymbol{\Sigma}|$ je determinant varianční matice $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$. Na obrázku 2 je znázorněna hustota dvourozměrného normálního rozdělení, k jejímuž vykreslení bylo využito softwaru R (zdrojové kódy jsou k dispozici na přiloženém CD).

Funkce hustoty existuje pouze v případě, že je p-rozměrné normální rozdělení regulární. Pro singulární mnohorozměrné rozdělení funkce hustoty vzhledem k Lebesgueově míře neexistuje [3,6]. Můžeme ji ale vyjádřit na nadrovině dimenze $m < p$, kde $m = h(\boldsymbol{\Sigma})$ představuje řád rozdělení. Další možností přístupu k vícerozměrnému normálnímu rozdělení je pracovat s tzv. charakteristickou funkcí, která existuje i pro singulární rozdělení [6].



Obrázek č. 2: Funkce hustoty dvourozměrného normálního rozdělení s parametry $\mu_1 = 0, \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 25, \rho = 0.5$.

Dále předpokládejme, že $\mathbf{X} = (X_1, \ldots, X_p)'$ je náhodný vektor, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)' = \mathrm{E}\mathbf{X} = (\mathrm{E}X_1, \ldots, \mathrm{E}X_p)'$ je vektor středních hodnot a $\boldsymbol{\Sigma} = \mathrm{var}\mathbf{X} = \mathrm{cov}(X_i, X_j)$ je varianční matice.

Věta 2.1. *Nechť je dán vektor* $\mathbf{a} \in \mathbb{R}^{m\times 1}$*, matice* $\mathbf{B} \in \mathbb{R}^{m\times p}$ *a dále platí, že náhodný vektor* $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ *má normální rozdělení* $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*. Pak pro náhodný vektor* $\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X}$ *platí*

$$\mathbf{Y} \sim \mathcal{N}_m(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}').$$

Důkaz. viz [3], str. 64 □

Nechť má náhodný vektor $\mathbf{X}$ n-rozměrné normální rozdělení $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ a k je celé číslo, pro které platí $1 \leq k < p$. Položme

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11}\ \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21}\ \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

kde $\mathbf{X}_1 \in \mathbb{R}^{k\times 1}$, $\boldsymbol{\mu}_1 \in \mathbb{R}^{k\times 1}$ a $\boldsymbol{\Sigma}_{11} \in \mathbb{R}^{k\times k}$. Pak vektor $\mathbf{X}_1$ složený z prvních k složek vektoru $\mathbf{X}$ má marginální rozdělení $\mathcal{N}_k(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$, jelikož víme, že lineární transformace zachovává normalitu [3].

Dále uvažujme náhodný vektor $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$, který je složen nejméně ze dvou složek. Označme $\mathbf{X}_1 = (X_1, X_2, \ldots, X_k)'$ a $\mathbf{X}_2 = (X_{k+1}, X_{k+2}, \ldots, X_p)'$, kde $1 \leq k < p$.

Věta 2.2. *Nechť* $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2')' \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *Jestliže platí* $\operatorname{cov}(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{0}$, *pak jsou náhodné vektory* $\mathbf{X}_1$ *a* $\mathbf{X}_2$ *vzájemně nezávislé.*

Důkaz. viz [3], str. 66 □

Označme vektor $\mathbf{X}^0 = (X_1^0, \ldots, X_r^0)'$, kde $X_1^0, \ldots, X_r^0$ jsou nezávislé náhodné veličiny s normovaným normálním rozdělením $\mathcal{N}_1(0, 1)$, tedy náhodný vektor $\mathbf{X}^0$ má r-rozměrné normální rozdělení $\mathcal{N}_r(\mathbf{0}, \mathbf{I}_r)$.

Věta 2.3. *Nechť* $\mathbf{X} = (X_1, X_2, \ldots, X_p)' \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *Nechť* $h(\boldsymbol{\Sigma}) = r \geq 1$ *a nechť* $\mathbf{B} \in \mathbb{R}^{p\times r}$, *kde* $h(\mathbf{B}) = r$ *a platí* $\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}'$. *Pak*

$$\boldsymbol{\mu} + \mathbf{B}\mathbf{X}^0 \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Důkaz. viz [3], str. 65 □

Mnohorozměrné normální rozdělení hraje v matematické statistice zásadní roli. Předpoklad normality je součástí většiny statistických metod a postupů. Proto je nutné předpoklad normálního rozdělení náhodného výběru otestovat. K tomu využíváme například v jednorozměrném případě chí kvadrát test dobré shody, Kolmogorov-Smirnovův test [3,6], v mnohorozměrném případě lze použít třeba Anderson-Darlingův test [6].

2.2. Dirichletovo rozdělení

Dalším typem spojitého mnohorozměrného rozdělení pravděpodobnosti je Dirichletovo rozdělení [6]. Dirichletovo rozdělení je odvozeno pomocí gamma a beta rozdělení.

2.2.1. Gamma rozdělení

Gamma rozdělení je rozdělení pravděpodobnosti, které je jednoznačně určeno dvěma parametry. Prvním z nich je parametr měřítka α a druhým je parametr tvaru β. Gamma rozdělení patří do rodiny exponenciálních rozdělení.

Gamma rozdělení definujeme pomocí tzv. gamma funkce, která je dána jako

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$$

pro všechna $\alpha > 0$ a nabývá pouze kladných hodnot. Zřejmě bude-li $\alpha = 1$, pak

$$\Gamma(1) = \int_0^\infty e^{-y} dy = 1.$$

Jestliže je $\alpha > 1$, pak použitím metody per partes dostaneme vztah

$$\Gamma(\alpha) = (\alpha - 1) \int_0^\infty y^{\alpha-2} e^{-y} dy = (\alpha - 1)\Gamma(\alpha - 1).$$

Pro přirozená čísla $\alpha > 1$ tedy platí

$$\Gamma(\alpha) = (\alpha - 1)(\alpha - 2) \cdots 3 \cdot 2 \cdot 1 \cdot \Gamma(1) = (\alpha - 1)!.$$

Provedeme-li substituci $y = \frac{x}{\beta}$, kde $\beta > 0$, můžeme gamma funkci $\Gamma(\alpha)$ vyjádřit ve tvaru

$$\Gamma(\alpha) = \int_0^\infty \left(\frac{x}{\beta}\right)^{\alpha-1} e^{-x/\beta} \left(\frac{1}{\beta}\right) dx.$$

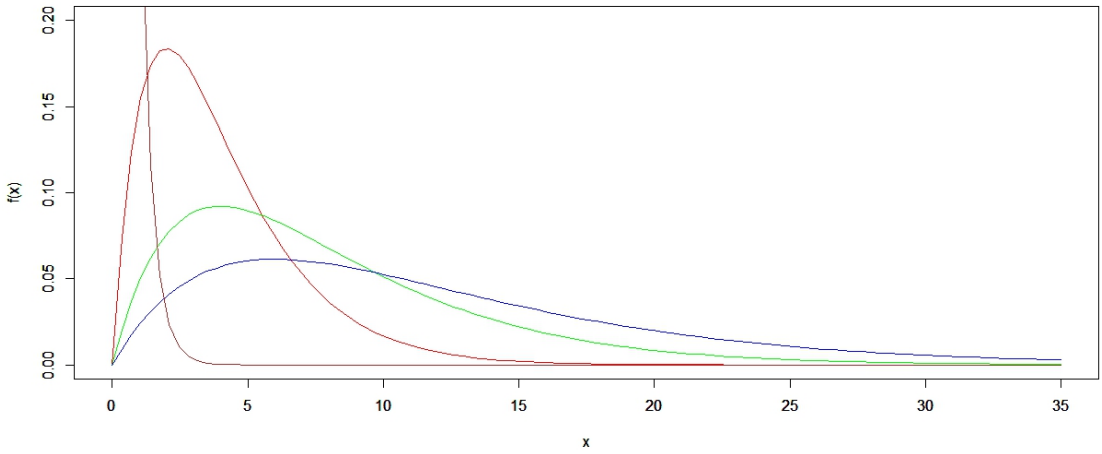
Pomocí odpovídajících ekvivalentních úprav poté dostaneme

$$1 = \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} dx.$$

Za předpokladu, že $\alpha > 0$, $\beta > 0$ a $\Gamma(\alpha) > 0$, je funkce

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} & 0 < x < \infty \\ 0 & \text{jinak.} \end{cases}$$

hustotou rozdělení pravděpodobnosti spojité náhodné veličiny X. O takovéto náhodné veličině říkáme, že má gamma rozdělení s parametry α a β. Symbolicky značíme $X \sim \Gamma(\alpha, \beta)$.



Obrázek č. 3: Funkce hustoty gamma rozdělení s parametry $\alpha = 2$, $\beta = 2$ (červená křivka), $\alpha = 2$, $\beta = 4$ (zelená křivka), $\alpha = 2$, $\beta = 6$ (modrá křivka) a $\alpha = 0.5$, $\beta = 0.5$ (hnědá křivka).

Funkce hustoty gamma rozdělení je asymetrická. Její průběh v závislosti na volbě parametrů α a β je znázorněn na obrázku 3. Gamma rozdělení se používá například v teorii front nebo v teorii spolehlivosti.

2.2.2. Beta rozdělení

Dalším důležitým rozdělením potřebným k definici Dirichletova rozdělení je beta rozdělení [6]. Beta rozdělení je spojité rozdělení pravděpodobnosti definované na intervalu $(0, 1)$. Je jednoznačně určeno dvěma kladnými parametry, které oba určují jeho tvar.

Beta rozdělení odvodíme z dvojice nezávislých gamma náhodných proměnných. Nechť tedy X_1 a X_2 jsou dvě nezávislé náhodné veličiny, které mají gamma

rozdělení a sdruženou funkci hustoty pravděpodobnosti ve tvaru

$$h(x_1, x_2) = \begin{cases} \frac{1}{\Gamma(\alpha)\Gamma(\beta)} x_1^{\alpha-1} x_2^{\beta-1} e^{-x_1-x_2} & 0 < x_1 < \infty, \quad 0 < x_2 < \infty, \\ 0 & \text{jinak,} \end{cases}$$

kde $\alpha > 0, \beta > 0$. Položme

$$Y_1 = X_1 + X_2,$$

$$Y_2 = \frac{X_1}{X_1 + X_2}.$$

Pomocí transformace náhodného vektoru (viz příloha) lze dokázat, že Y_1 a Y_2 jsou nezávislé.

Uvažujme tedy dvojnásobný integrál

$$\iint_{\mathcal{A}} h(x_1, x_2) dx_1 dx_2,$$

kde integrujeme přes množinu $\mathcal{A}$, která vymezena prvním kvadrantem roviny (x_1, x_2) bez bodů, které leží na souřadnicových osách, tj.

$$\mathcal{A} = \{(x_1, x_2) : 0 < x_i < \infty, i = 1, 2\}.$$

Funkce

$$y_1 = u_1(x_1, x_2) = x_1 + x_2,$$

$$y_2 = u_2(x_1, x_2) = \frac{x_1}{x_1 + x_2},$$

můžeme vyjádřit inverzně jako

$$x_1 = y_1 y_2,$$

$$x_2 = y_1(1 - y_2).$$

Odtud spočítáme jakobián

$$J = \begin{vmatrix} y_2 & y_1 \\ 1 - y_2 & -y_1 \end{vmatrix} = -y_1 \neq 0.$$

Tato transformace představuje zobrazení z $\mathcal{A}$ do množiny $\mathcal{B}$, kde

$$\mathcal{B} = \{(y_1, y_2) : 0 < y_1 < \infty, 0 < y_2 < 1\},$$

tj. do roviny (y_1, y_2). Sdružená funkce hustoty pravděpodobnosti pro náhodné veličiny Y_1 a Y_2 má tedy tvar

$$g(y_1, y_2) = y_1 \frac{1}{\Gamma(\alpha)\Gamma(\beta)} (y_1 y_2)^{\alpha-1} [y_1(1-y_2)]^{\beta-1} e^{-y_1},$$

tedy

$$g(y_1, y_2) = \begin{cases} \frac{y_2^{\alpha-1}(1-y_2)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} y_1^{\alpha+\beta-1} e^{-y_1} & 0 < y_1 < \infty, 0 < y_2 < 1, \\ 0 & \text{jinak.} \end{cases}$$

Marginální funkce hustoty pro veličinu Y_2 je funkce

$$g_2(y_2) = \frac{y_2^{\alpha-1}(1-y_2)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} \int_0^\infty y_1^{\alpha+\beta-1} e^{-y_1} dy_1,$$

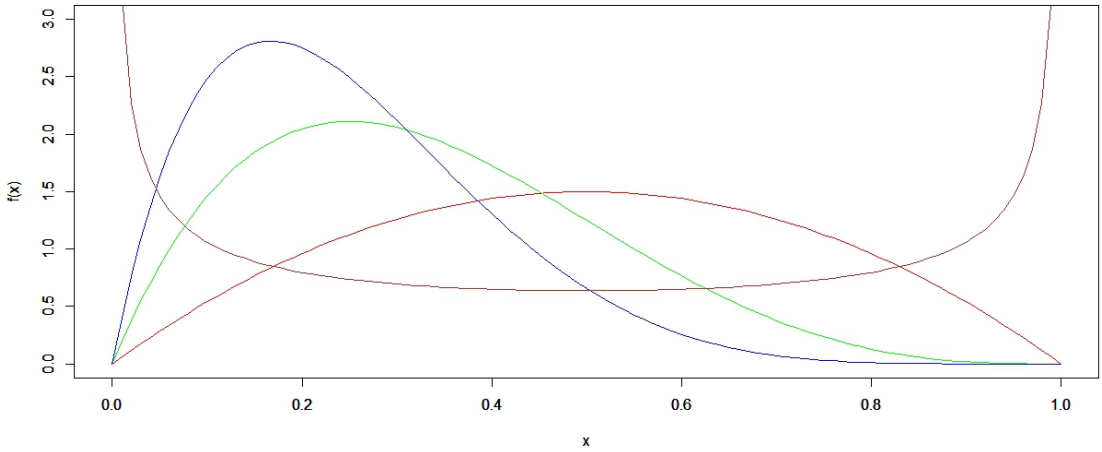
neboli

$$g_2(y_2) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y_2^{\alpha-1}(1-y_2)^{\beta-1} & 0 < y_2 < 1, \\ 0 & \text{jinak.} \end{cases}$$

Tato funkce je pak hustotou pravděpodobnosti beta rozdělení s parametry α a β. Veličiny Y_1 a Y_2 jsou nezávislé, protože platí $g(y_1, y_2) = g_1(y_1)g_2(y_2)$, přitom hustota Y_1 má tvar

$$g_1(y_1) = \begin{cases} \frac{1}{\Gamma(\alpha+\beta)} y_1^{\alpha+\beta-1} e^{-y_1} & 0 < y_1 < \infty, \\ 0 & \text{jinak.} \end{cases}$$

Jedná se o hustotu gamma rozdělení s hodnotami parametrů $\alpha + \beta$ a 1.



Obrázek č. 4: Funkce hustoty beta rozdělení s parametry $\alpha = 2$, $\beta = 2$ (červená křivka), $\alpha = 2$, $\beta = 4$ (zelená křivka), $\alpha = 2$, $\beta = 6$ (modrá křivka) a $\alpha = 0.5$, $\beta = 0.5$ (hnědá křivka).

Průběh funkce hustoty beta rozdělení je znázorněn na obrázku 4. Parametry jsou určeny stejně jako u gamma rozdělení.

2.2.3. Dirichletovo rozdělení

Dirichletovo rozdělení pravděpodobnosti patří mezi spojitá vícerozměrná rozdělení pravděpodobnosti [6]. Obecně se jedná o mnohorozměrné zobecnění beta rozdělení. Dirichletovo rozdělení bývá velmi často používáno jako apriorní rozdělení pravděpodobnosti v bayesovské statistice.

Dirichletovo rozdělení je stejně jako beta rozdělení odvozeno pomocí transformace gamma náhodných veličin (viz příloha).

Nechť $Y_1, Y_2, \ldots, Y_{p+1}$ jsou nezávislé náhodné veličiny, které mají gamma rozdělení s parametrem $\beta = 1$. Sdruženou funkci hustoty pravděpodobnosti můžeme zapsat ve tvaru

$$h(y_1, y_2, \ldots, y_{p+1}) = \begin{cases} \prod_{i=1}^{p+1} \frac{1}{\Gamma(\alpha_i)} y_i^{\alpha_i - 1} e^{-y_i} & 0 < y_i < \infty \\ 0 & \text{jinak.} \end{cases}$$

Nechť

$$X_i = \frac{Y_i}{Y_1 + Y_2 + \cdots + Y_{p+1}}, \qquad i = 1, 2, \ldots, p,$$

$$X_{p+1} = Y_1 + Y_2 + \cdots + Y_{p+1}$$

označují nové náhodné veličiny, kterých je celkem $p + 1$. Použijeme-li stejnou transformaci jako u beta rozdělení, pak máme zobrazení z množiny

$$\mathcal{A} = \{(y_1, \ldots, y_{p+1}) : 0 < y_i < \infty, i = 1, \ldots, p + 1\}$$

do množiny

$$\mathcal{B} = \{(x_1, \ldots, x_p, x_{p+1}) : 0 < x_i, i = 1, \ldots, p, x_1 + x_2 + \cdots + x_p < 1, 0 < x_{p+1} < \infty\}.$$

Odpovídající inverzní funkce mají tvar

$$\begin{aligned} y_1 &= x_1 x_{p+1}, \\ &\vdots \\ y_p &= x_p x_{p+1}, \\ y_{p+1} &= x_{p+1}(1 - x_1 - \cdots - x_p). \end{aligned}$$

Nyní můžeme sestavit Jakobián

$$J = \begin{vmatrix} x_{p+1} & 0 & \cdots & 0 & x_1 \\ 0 & x_{p+1} & \cdots & 0 & x_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & x_{p+1} & x_p \\ -x_{p+1} & -x_{p+1} & \cdots & -x_{p+1} & (1 - x_1 - \cdots - x_p) \end{vmatrix} = x_{p+1}^p,$$

tedy ze vztahu

$$\int \cdots \int_{\mathcal{A}} h(y_1, y_2, \ldots, y_{p+1}) dy_1 dy_2 \cdots dy_{p+1}$$

$$= \int \cdots \int_{\mathcal{B}} h[x_1 x_{p+1}, x_2 x_{p+1}, \ldots, x_{p+1}(1 - x_1 - \cdots - x_p)]|J| dx_1 dx_2 \cdots dx_{p+1}.$$

dostáváme simultánní funkci hustoty veličin $X_1, \ldots, X_p, X_{p+1}$ jako

$$g(x_1, \ldots, x_p, x_{p+1}) = \frac{x_{p+1}^{\alpha_1+\cdots+\alpha_{p+1}-1} x_1^{\alpha_1-1} \cdots x_p^{\alpha_p-1}(1 - x_1 - \cdots - x_p)^{\alpha_{p+1}-1} e^{-x_{p+1}}}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_p)\Gamma(\alpha_{p+1})},$$

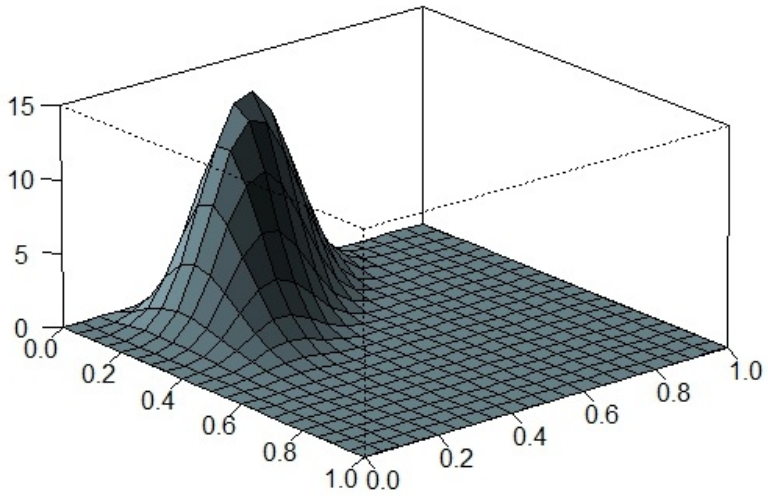
za předpokladu, že $(x_1, \ldots, x_p, x_{p+1})$ je z množiny $\mathcal{B}$. V ostatních případech je hodnota funkce hustoty nulová.

Sdruženou hustotu pravděpodobnosti pro veličiny $X_1, \ldots, X_p$ pak obdržíme ve tvaru

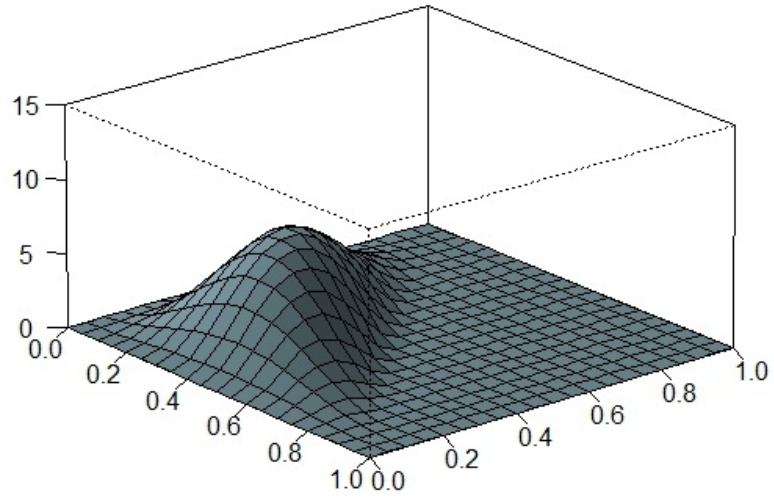
$$g(x_1, \ldots, x_p) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_{p+1})}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_{p+1})} x_1^{\alpha_1-1} \cdots x_p^{\alpha_p-1}(1 - x_1 - \cdots - x_p)^{\alpha_{p+1}-1} \quad (1)$$

pro $x_i > 0$, kde $i = 1, \ldots, p$ a $x_1 + \cdots + x_p < 1$. V opačném případě je funkce $g(x_1, \ldots, x_p)$ nulová. O náhodných veličinách $X_1, \ldots, X_p$, které mají funkci hustoty (1), říkáme, že mají Dirichletovo rozdělení pravděpodobnosti s parametrem $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{p+1})'$. Pro speciální případ $p = 1$ odpovídá Dirichletova funkce hustoty předpisu hustoty pravděpodobnosti pro beta rozdělení.

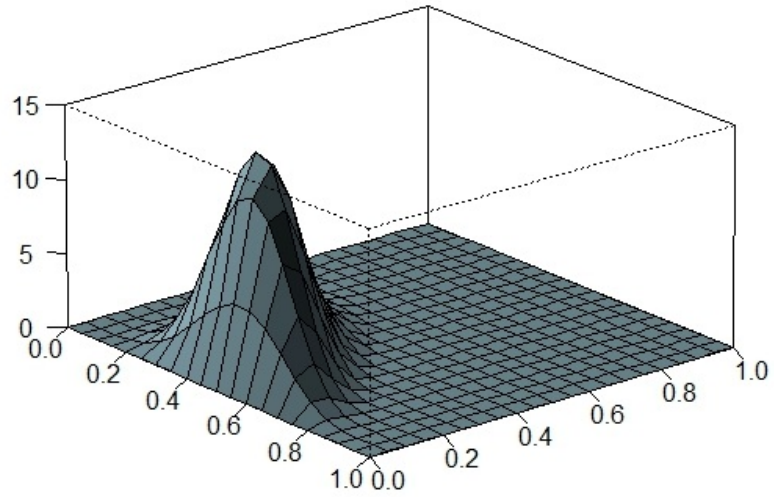
Z tvaru simultánní funkce hustoty pro $X_1, \ldots, X_p, X_{p+1}$ lze dále vyvodit, že náhodná veličina X_{p+1} má gamma rozdělení s parametry $\alpha_1 + \cdots + \alpha_p + \alpha_{p+1}$ a $\beta = 1$ a navíc je veličina X_{p+1} nezávislá na $X_1, \ldots, X_p$.



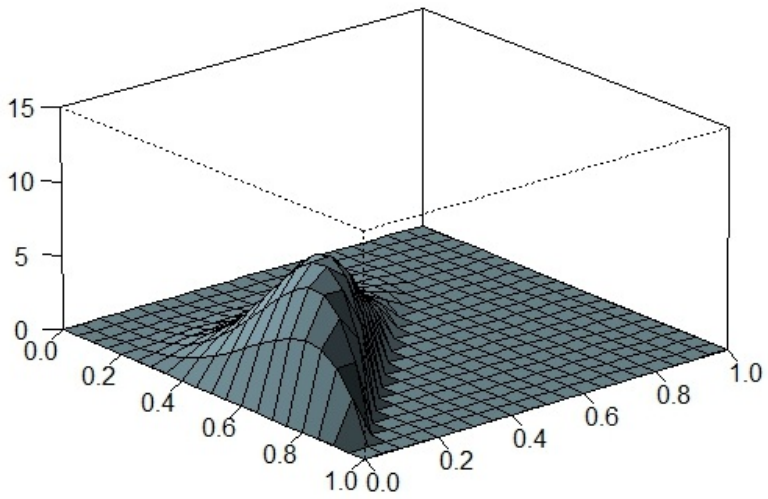
Obrázek č. 5: Funkce hustoty Dirichletova rozdělení s parametrem $\boldsymbol{\alpha} = (3, 6, 7)'$.



Obrázek č. 6: Funkce hustoty Dirichletova rozdělení s parametrem $\boldsymbol{\alpha} = (3, 3, 3)'$.



Obrázek č. 7: Funkce hustoty Dirichletova rozdělení s parametrem $\boldsymbol{\alpha} = (7, 3, 6)'$.



Obrázek č. 8: Funkce hustoty Dirichletova rozdělení s parametrem $\boldsymbol{\alpha} = (5, 2, 2)'$.

Obrázky 5 - 8 ukazují, jak se hustota Dirichletova rozdělení mění při různé volbě parametru $\boldsymbol{\alpha}$.

2.2.4. Číselné charakteristiky Dirichletova rozdělení

Pravděpodobnostní chování náhodné veličiny je ve spojitém případě popsáno její distribuční funkcí a její hustotou. Takový popis bývá někdy komplikovaný, a proto se k popisu rozdělení náhodných veličin často používají číselné charakteristiky, nejčastěji střední hodnota a rozptyl [3,6]. Stejně je tomu i u Dirichletova rozdělení. Předtím, než si je uvedeme, si ovšem vyjádříme Dirichletovo rozdělení ve tvaru, který se často užívá v aplikacích.

Definice 2.2. *Náhodný vektor* $\mathbf{X} = (X_1, \ldots, X_D)'$ *má Dirichletovo rozdělení s parametrem* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)'$*, jestliže má hustotu pravděpodobnosti*

$$g(x_1, \ldots, x_D) = \frac{\Gamma(\alpha_+)}{\prod_{i=1}^{D} \Gamma(\alpha_i)} \prod_{i=1}^{D} x_i^{\alpha_i - 1}$$

pro $x_i > 0$*,* $i = 1, \ldots, D$*,* $x_1 + \cdots + x_D = 1$*, a* $g(x_1, \ldots, x_D) = 0$ *jinak.*

Je zřejmé, že je tato definice ekvivalentní s (1) pro $D = p + 1$.

Nechť tedy $\mathbf{X} = (X_1, X_2, \ldots, X_k)'$ je náhodný vektor, který má Dirichletovo rozdělení s parametrem $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_D)'$ a označme $\alpha_+ = \sum_{i=1}^{D} \alpha_i$. Střední hodnota pro Dirichletovo rozdělení je dána ve tvaru

$$\mathrm{E}(\mathbf{X}) = \left(\frac{\alpha_1}{\alpha_+}, \frac{\alpha_2}{\alpha_+}, \ldots, \frac{\alpha_D}{\alpha_+}\right)',$$

tj. vektorem středních hodnot jednotlivých náhodných veličin X_i, pro které platí

$$\mathrm{E}(X_i) = \frac{\alpha_i}{\alpha_+}, \qquad i = 1, \ldots, D.$$

Modus náhodného vektoru $\mathbf{X}$ je vyjádřen obdobně

$$\mathrm{modus}(\mathbf{X}) = \left(\frac{\alpha_1 - 1}{\alpha_+ - D}, \frac{\alpha_2 - 1}{\alpha_+ - D}, \ldots, \frac{\alpha_D - 1}{\alpha_+ - D}\right)'.$$

Momenty druhého řádu, rozptyl a kovariance, jsou definovány následovně

$$\mathrm{var}(X_i) = \frac{\alpha_i(\alpha_+ - \alpha_i)}{{\alpha_+}^2(\alpha_+ + 1)} = \frac{\mathrm{E}(X_i)(1 - \mathrm{E}(X_i))}{(\alpha_+ + 1)},$$

$$\mathrm{cov}(X_i, X_j) = -\frac{\alpha_i \alpha_j}{(\alpha_+)^2(\alpha_+ + 1)} = -\frac{\mathrm{E}(X_i)\mathrm{E}(X_j)}{(\alpha_+ + 1)}.$$

Ze vztahu určujícího kovariance je patrné, že korelační koeficienty u Dirichletova rozdělení budou vždy nekladné.

Důležitou roli při interpretaci charakteristik hraje parametr α_+; čím více vzroste jeho hodnota, tím více bude rozdělení koncentrováno kolem své očekávané hodnoty.

2.2.5. Vlastnosti Dirichletova rozdělení

Dirichletovo rozdělení je z praktického hlediska relativně málo používaným typem vícerozměrného rozdělení pravděpodobnosti, i když se standardně uvádí jako jediné známé rozdělení náhodného vektoru s konstantním součtem složek.

Důvodem je předpoklad nezávislosti jednotlivých gamma rozdělených náhodných veličin, užitých při jeho konstrukci [1]. Již v definici Dirichletova rozdělení je přitom jedna náhodná proměnná vyjádřena pomocí ostatních. Tudíž je logické očekávat alespoň slabou formu závislosti mezi náhodnými veličinami.

Další slabinou použití Dirichletova rozdělení je předpoklad, že vzájemná kovariance dvou náhodných veličin je z definice záporná. Z tohoto důvodu se tak toto rozdělení může zdát vhodné pouze pro data, která vykazují negativní korelační strukturu (!).

Přesto Dirichletovo rozdělení nabízí i některé výhodné matematické vlastnosti. Mezi ně patří například vlastnosti, že podmíněné a marginální rozdělení k Dirichletovu je opět Dirichletovo rozdělení [12].

Dirichletovo rozdělení se proto prakticky používá především v bayesovské statistice. Jedná se o moderní odvětví matematické statistiky, které je založeno na práci s podmíněnými pravděpodobnostmi. Principem bayesovské statistiky je, že pravděpodobnost výchozí hypotézy je postupně zpřesňována na základě předchozích zkušeností či výsledků náhodného pokusu. Její matematický aparát je postaven na Bayesově větě, podle které získala i svůj název. Dirichletovo rozdělení je zde používáno jako apriorní rozdělení pravděpodobnosti např. při testování životnosti [12].

Hlavní příčinou "selhání" Dirichletova rozdělení se ovšem zdá nevhodnost použití standardní Lebesgueovy míry pro modelování dat s konstantním součtem (tzv. kompozičních dat). Alternativní přístup k zavedení Dirichletova rozdělení, který by mohl pomoci uvedený nedostatek eliminovat, je rozebrán ve čtvrté kapitole této práce.

3. Kompoziční data

Ve statistice se pod pojmem kompoziční data rozumí data nesoucí pouze relativní informaci, například proporce či procentuální části celku. Od ostatních dat se liší především tím, že se vztahují k danému celku a bez jeho znalosti ztrácejí svůj význam [1,14].

3.1. Definice a základní vlastnosti

Definice 3.1. *D-složkovým kompozičním vektorem, nebo také kompozicí, rozumíme kladný reálný vektor* $\mathbf{x} = (x_1, \ldots, x_D)'$*, jehož složky nesou výhradně relativní informaci.*

Kompoziční data jsou tedy charakteristická tím, že informace, kterou nesou, je obsažena v podílech mezi složkami. Dalším specifickým znakem těchto dat je možnost reprezentovat je jako data s konstantním součtem.

Definice 3.2. *Uzávěrem kompozice* $\mathbf{x} = (x_1, \ldots, x_D)' \in \mathbb{R}_+^D$ *vzhledem ke konstantnímu součtu* k *nazveme vektor*

$$\mathcal{C}(\mathbf{x}) = \left(\frac{kx_1}{\sum_{i=1}^{D} x_i}, \ldots, \frac{kx_D}{\sum_{i=1}^{D} x_i} \right)'.$$

Povšimněme si, že definice 3.2. představuje pouze reprezentaci kompozičních dat (pro $k = 1$ dostaneme případ z předchozí kapitoly). Jejich zavedení pomocí (obecnější) definice 3.1. umožňuje rozvinout ucelený přístup k jejich stochastickému modelování.

Jelikož je charakter kompozičních dat odlišný od standardních mnohorozměrných pozorování, je přirozené, že i jejich výběrový prostor bude specifický.

Definice 3.3. *Výběrový prostor kompozičních dat je simplex, který je definován vztahem*

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

Uvažujeme-li, že vektor $\mathbf{x} = (x_1, \ldots, x_D)'$ představuje proporcionální části daného celku, tj. $\sum_{i=1}^{D} x_i = 1$, pak výběrovým prostorem těchto dat bude jednotkový simplex.

V praxi často nemusíme mít k dispozici všechny složky kompozice, proto často pracujeme pouze s jejich podkompozicemi.

Definice 3.4. *Podkompozicí* $\mathbf{x}_s$ *dané kompozice* $\mathbf{x}$ *nazveme vektor* $(x_{i_1}, \ldots, x_{i_s})'$, *který představuje určitou část kompozice* $\mathbf{x}$. *Indexy* $i_1, \ldots, i_s$, $1 \leq i_1 < \cdots < i_s \leq D$, *určují, které složky byly do podkompozice vybrány.*

U složek kompozice se předpokládá, že nabývají pouze kladných reálných hodnot. V praxi se ovšem může stát, že některé složky mohou být nulové [7]. Vyskytnout se mohou dva typy nul. Jedním případem jsou nuly vzniklé zaokrouhlováním, máme-li (skutečné) hodnoty složek velmi blízké nule. Druhým typem jsou tzv. strukturální nuly (například výdaje za cigarety ve spotřebním koši nekuřáků).

3.2. Aitchisonova geometrie na simplexu

Při práci s mnohorozměrnými daty jsme zvyklí pracovat v reálném vektorovém prostoru, na kterém je definovaná standardní euklidovská geometrie. To znamená, že v tomto prostoru existují operace sčítání vektorů a násobení vektoru skalárem, které nám práci s vektory umožňují. Chceme-li v takovémto prostoru spočítat vzdálenost dvou vektorů, využíváme k tomu euklidovskou metriku.

Euklidovská geometrie není ovšem vhodná pro kompoziční data [14]. Z toho důvodu je nutné zavést jinou geometrii, která by vedla ke správným výsledkům při statistickém zpracování kompozičních dat. Je tedy nutné zadefinovat operace, které by byly analogické ke sčítání a násobení skalárem v reálném prostoru [14].

Definice 3.5. *Perturbací kompozice* $\mathbf{x} = \mathcal{C}(x_1, \ldots, x_D)' \in \mathcal{S}^D$ *kompozicí* $\mathbf{y} = \mathcal{C}(y_1, \ldots, y_D)' \in \mathcal{S}^D$ *nazveme kompozici* $\mathbf{x} \oplus \mathbf{y} \in \mathcal{S}^D$, *která je definována*

vztahem

$$\mathbf{x} \oplus \mathbf{y} = \frac{k \cdot (x_1 y_1, x_2 y_2, \ldots, x_D y_D)}{x_1 y_1 + x_2 y_2 + \cdots + x_D y_D} = \mathcal{C}(x_1 y_1, x_2 y_2, \ldots, x_D y_D)'.$$

Operace perturbace na simplexu představuje analogii sčítání dvou vektorů v reálném prostoru.

Definice 3.6. *Mocninnou transformací kompozice* $\mathbf{x} = \mathcal{C}(x_1, \ldots, x_D)' \in \mathcal{S}^D$ *reálným číslem* $\alpha \in \mathbb{R}$ *nazveme kompozici* $\alpha \odot \mathbf{x} \in \mathcal{S}^D$*, která je dána předpisem*

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

Mocninná transformace na simplexu je obdobou násobení vektoru skalárem v reálném prostoru.

Simplexový prostor společně s operacemi perturbace a mocninná transformace tvoří reálný vektorový prostor $(\mathcal{S}^D, \oplus, \odot)$, a proto tyto operace splňují stejné vlastnosti, které platí pro operace sčítání a násobení skalárem v prostoru reálném.

Pro libovolné kompozice $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{S}^D$ vzhledem k operaci perturbace platí:

i) komutativita: $\mathbf{x} \oplus \mathbf{y} = \mathbf{y} \oplus \mathbf{x}$;

ii) asociativita: $(\mathbf{x} \oplus \mathbf{y}) \oplus \mathbf{z} = \mathbf{x} \oplus (\mathbf{y} \oplus \mathbf{z})$;

iii) existuje neutrální prvek: $\mathbf{n} = \mathcal{C}(1, 1, \ldots, 1)' = \left(\frac{1}{D}, \frac{1}{D}, \ldots, \frac{1}{D}\right)'$, kde $\mathbf{n}$ představuje těžiště daného simplexu;

iv) existuje inverzní prvek $\mathbf{x}^{-1}$ ke kompozici $\mathbf{x}$: $\mathbf{x}^{-1} = \mathcal{C}\left(\frac{1}{x_1}, \frac{1}{x_2}, \ldots, \frac{1}{x_D}\right)'$, pro který platí $\mathbf{x} \oplus \mathbf{x}^{-1} = \mathbf{n}$, ekvivalentně $\mathbf{x} \ominus \mathbf{x} = \mathbf{n}$.

Podobně vzhledem k operaci mocninná transformace pro libovolné kompozice $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ a konstanty $\alpha, \beta \in \mathbb{R}$ platí:

i) asociativita: $\alpha \odot (\beta \odot \mathbf{x}) = (\alpha \cdot \beta) \odot \mathbf{x}$;

ii) distributivita: $\alpha \odot (\mathbf{x} \oplus \mathbf{y}) = (\alpha \odot \mathbf{x}) \oplus (\alpha \odot \mathbf{y})$;

iii) distributivita: $(\alpha \oplus \beta) \odot \mathbf{x} = (\alpha \odot \mathbf{x}) \oplus (\beta \odot \mathbf{y})$;

iv) existuje neutrální prvek: $1 \odot \mathbf{x} = \mathbf{x}$.

Definice 3.7. *Aitchisonův skalární součin kompozic* $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ *je dán vztahem*

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

kde $g(\mathbf{x}) = \prod_{i=1}^{D} (x_i)^{\frac{1}{D}}$*, resp.* $g(\mathbf{y}) = \prod_{i=1}^{D} (y_i)^{\frac{1}{D}}$ *představuje geometrický průměr složek kompozice* $\mathbf{x}$*, resp.* $\mathbf{y}$*.*

Zavedením výše uvedených operací a skalárního součinu můžeme tvrdit, že prostor $(\mathcal{S}^D, \oplus, \odot)$ tvoří $(D-1)$-rozměrný euklidovský lineární vektorový prostor. V oblasti kompozičních dat hovoříme v souvislosti s tímto prostorem jako o Aitchisonově geometrii na simplexu [14].

S existencí Aitchisonova skalárního součinu jsou zavedeny pojmy Aitchisonovy normy a vzdálenosti.

Definice 3.8. *Aitchisonova norma kompozice* $\mathbf{x} \in \mathcal{S}^D$ *je definována vztahem*

$$\| \mathbf{x} \|_a = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_a}.$$

Definice 3.9. *Aitchisonovu vzdálenost mezi kompozicí* $\mathbf{x}$ *a kompozicí* $\mathbf{y} \in \mathcal{S}^D$ *definujeme předpisem*

$$d_a(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{D} \sum_{i<j} \Big(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \Big)^2}.$$

Aitchisonova vzdálenost má standardní vlastnosti jako vzdálenost euklidovská. Je invariantní vůči permutaci

$$d_a(\mathbf{x}, \mathbf{y}) = d_a(\mathbf{p} \oplus \mathbf{x}, \mathbf{p} \oplus \mathbf{y}),$$

invariantní na změnu škály

$$d_a(\alpha \odot \mathbf{x}, \alpha \odot \mathbf{y}) = |\alpha| d_a(\mathbf{x}, \mathbf{y})$$

a nezávisí na pořadí složek kompozice, tj.

$$d_a(\mathbf{P}\mathbf{x}, \mathbf{P}\mathbf{y}) = d_a(\mathbf{x}, \mathbf{y}),$$

kde $\mathbf{P}$ je libovolná permutační matice.

3.3. Logratio transformace kompozičních dat

Aitchisonova geometrie na simplexu je vhodnou volbou typu geometrie pro práci s kompozičními daty. Problém ale nastává při samotné interpretaci výsledků [14]. Z toho důvodu se snažíme data z Aitchisonovy geometrie převést do standardní euklidovské geometrie, ve které je ostatně zkonstruována drtivá většina mnohorozměrných statistických metod. K tomu využíváme tzv. logratio transformace [4,14].

D-složkové kompozice obvykle vyjadřujeme ve tvaru kanonické báze $\{\vec{\mathbf{e}_1}, \vec{\mathbf{e}_2}, \ldots, \vec{\mathbf{e}_D}\}$ prostoru $\mathbb{R}^D$ použitím operací sčítání a násobení skalárem. Libovolnou kompozici $\mathbf{x} \in \mathcal{S}^D$ můžeme zapsat ve tvaru

$$\mathbf{x} = (x_1, x_2, \ldots, x_D)' = x_1(1, 0, \ldots, 0)' + x_2(0, 1, 0, \ldots, 0)' + \cdots + x_D(0, \ldots, 0, 1)'$$

neboli

$$\mathbf{x} = \sum_{i=1}^{D} x_i \cdot \vec{\mathbf{e}_i}.$$

Kanonická báze prostoru $\mathbb{R}^D$ ovšem není bází prostoru $\mathcal{S}^D$ a toto vyjádření není lineární kombinací na simplexu vzhledem k jeho vektorové struktuře. Nicméně víme, že prostor $\mathcal{S}^D$ je vektorový prostor dimenze $D-1$, což nám existenci báze zaručuje.

Nejprve je potřeba zadefinovat množinu, která prostor $\mathcal{S}^D$ generuje, tu označíme jako

$$B^* = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_D\},$$

kde

$$\mathbf{w}_i = \mathcal{C}(\exp(\vec{\mathbf{e}}_i))' = \mathcal{C}(1, 1, \ldots, e, \ldots, 1)', \quad i = 1, \ldots, D,$$

kde e značí Eulerovo číslo a tvoří i-tou složku kompozice $\mathbf{w}_i$.

Libovolnou kompozici $\mathbf{x} \in \mathcal{S}^D$ lze tedy vyjádřit ve tvaru

$$\mathbf{x} = (\ln x_1 \odot \mathbf{w}_1) \oplus (\ln x_2 \odot \mathbf{w}_2) \oplus \cdots \oplus (\ln x_D \odot \mathbf{w}_D) = \bigoplus_{i=1}^{D} \ln x_i \odot \mathbf{w}_i,$$

kde symbol $\bigoplus$ představuje opakovanou perturbaci.

Analogicky můžeme použít k vyjádření kompozice $\mathbf{x}$ kvůli nejednoznačnosti koeficientů vzhledem ke generujícímu systému i vztahu

$$\mathbf{x} = \left(\ln \frac{x_1}{g(\mathbf{x})} \odot \mathbf{w}_1\right) \oplus \left(\ln \frac{x_2}{g(\mathbf{x})} \odot \mathbf{w}_2\right) \oplus \cdots \oplus \left(\ln \frac{x_D}{g(\mathbf{x})} \odot \mathbf{w}_D\right) = \bigoplus_{i=1}^{D} \ln \frac{x_i}{g(\mathbf{x})} \odot \mathbf{w}_i,$$

kde $g(\mathbf{x}) = \prod_1^D (x_i)^{\frac{1}{D}} = \exp\left(\frac{1}{D} \sum_{i=1}^D \ln x_i\right)$ je opět geometrický průměr kompozice $\mathbf{x}$.

Máme-li danou kompozici $\mathbf{x} \in \mathcal{S}^D$, možným vyjádřením vektoru koeficientů (souřadnic) vzhledem ke generující množině B^* je tak

$$\mathrm{clr}(\mathbf{x}) = (c_1, c_2, \ldots, c_D)' = \left(\ln \frac{x_1}{g(\mathbf{x})}, \ln \frac{x_2}{g(\mathbf{x})}, \ldots, \ln \frac{x_D}{g(\mathbf{x})}\right)',$$

čímž jsme získáli tzv. centrovanou logratio (clr) transformaci kompozice $\mathbf{x}$ [1]. Tato transformace je sice izometrická a symetrická ve složkách, ale součet jejích složek je roven nule, což vede k singularitě příslušné varianční matice.

Dimenze prostoru $\mathcal{S}^D$ je $D - 1$, můžeme tedy vypustit libovolnou kompozici z množiny B^*, abychom obdrželi bázi prostoru $\mathcal{S}^D$. Vypustíme-li poslední kompozici $\mathbf{w}_D$, dostaneme novou bázi ve tvaru

$$B = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{D-1}\}.$$

Nyní můžeme každou kompozici $\mathbf{x} \in \mathcal{S}^D$ vyjádřit jako lineární kombinaci

$$\mathbf{x} = \left(\ln \frac{x_1}{x_D} \odot \mathbf{w}_1\right) \oplus \left(\ln \frac{x_2}{x_D} \odot \mathbf{w}_2\right) \oplus \cdots \oplus \left(\ln \frac{x_{D-1}}{x_D} \odot \mathbf{w}_{D-1}\right) = \bigoplus_{i=1}^{D-1} \ln \frac{x_i}{x_D} \odot \mathbf{w}_i.$$

Souřadnice kompozice $\mathbf{x}$ vzhledem k bázi B tvoří aditivní logratio (alr) transformaci [1].

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

Můžeme ale také vypustit z báze B^* libovolnou jinou kompozici $\mathbf{w}_i$ pro $i = 1, \ldots, D-1$, což znamená, že jsme schopni vytvořit tolik různých alr transformací, kolik máme složek kompozice. Alr transformace není izometrická ani symetrická ve složkách.

Zavedení skalárního součinu a normy v prostoru $\mathcal{S}^D$ nám zaručuje existenci ortonormální báze. Množina $B = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{D-1}\}$ však ortonormální bází není. Pomocí Gram-Schmidtovy ortonormalizační metody jsme z ní schopni obdržet bázi $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$, která už ortonormální bude. Kompozici $\mathbf{x}$ jsme nyní schopni vyjádřit ve tvaru lineární kombinace

$$\mathbf{x} = (\langle \mathbf{x}, \mathbf{e}_1 \rangle_a \odot \mathbf{e}_1) \oplus (\langle \mathbf{x}, \mathbf{e}_2 \rangle_a \odot \mathbf{e}_2) \oplus \cdots \oplus (\langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a \odot \mathbf{e}_{D-1}).$$

Souřadnice jakékoliv kompozice $\mathbf{x}$ vzhledem k ortonormální bázi $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$ tvoří izometrickou logratio (ilr) transformaci kompozice $\mathbf{x}$ [4], tj.

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

Stejně jako v reálném prostoru, i na $\mathcal{S}^D$ existuje nekonečné množství ortonormálních bází, což znamená, že i ilr transformací jsme schopni vytvořit nekonečně mnoho. Ilr transformace je izometrická stejně jako alr, ale narozdíl od ní nemá problém se singularitou příslušné varianční matice. Jednou konkrétní volbou ortonormální báze dostaneme ilr souřadnice [5]

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

Inverzní ilr transformací vyjádříme ilr souřadnice zpět na simplexu tak, že $\mathbf{x} = \text{ilr}^{-1}(\mathbf{z})$, konkrétně

$$x_i = \exp\left(\sum_{j=i}^{D} \frac{z_j}{\sqrt{j(j+1)}} - \sqrt{\frac{i-1}{i}} z_{i-1}\right),$$

kde $z_0 = z_D = 0$ pro $i = 1, \ldots, D$.

Máme-li souřadnice vzhledem k ortonormální bázi, můžeme na ně použít standardní používané metody. Operace perturbace $\oplus$ a mocninné transformace $\odot$ jsou ekvivalentní ke klasickým operacím součtu a násobení skalárem použitých na souřadnicích vzhledem k libovolné bázi, která nemusí být nutně ortonormální. V případě, že máme souřadnice vzhledem k ortonormální bázi, lze na ně aplikovat standardní skalární součin i euklidovskou vzdálenost v prostoru $\mathbb{R}^{D-1}$.

Vztahy mezi jednotlivými transformacemi alr, clr a ilr můžeme interpretovat jako změnu báze nebo generujícího systému, které souvisejí se souřadnicemi $\text{alr}(\mathbf{x})$, $\text{clr}(\mathbf{x})$ a $\text{ilr}(\mathbf{x})$ [9]. Platí

$$\begin{array}{ll} \text{alr}(\mathbf{x}) = \mathbf{F}\text{clr}(\mathbf{x}); & \text{clr}(\mathbf{x}) = \mathbf{F}^*\text{alr}(\mathbf{x}); \\ \text{clr}(\mathbf{x}) = \mathbf{U}\text{ilr}(\mathbf{x}); & \text{ilr}(\mathbf{x}) = \mathbf{U}'\text{clr}(\mathbf{x}); \\ \text{alr}(\mathbf{x}) = \mathbf{F}\mathbf{U}\text{ilr}(\mathbf{x}); & \text{ilr}(\mathbf{x}) = \mathbf{U}'\mathbf{F}^*\text{alr}(\mathbf{x}). \end{array}$$

Matice $\mathbf{U} \in \mathbb{R}^{D\times(D-1)}$ je matice, jejíž sloupce tvoří vektory $\text{clr}(\mathbf{e}_i)$ pro $i = 1, \ldots, D-1$. Matice $\mathbf{F} \in \mathbb{R}^{(D-1)\times D}$ a $\mathbf{F}^* \in \mathbb{R}^{D\times(D-1)}$ vypadají následovně:

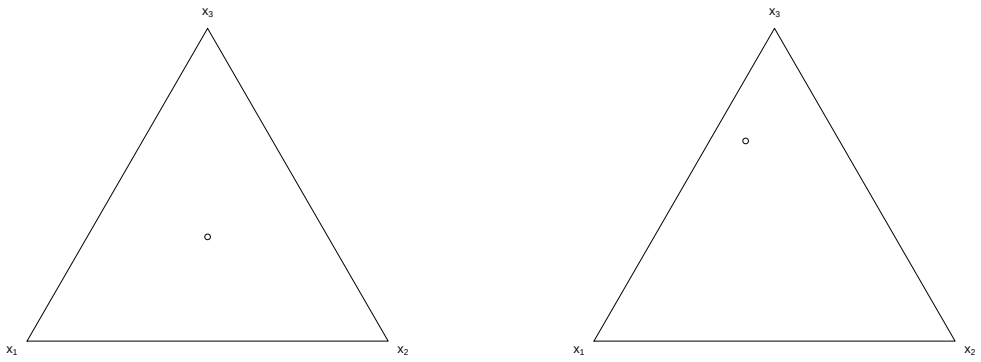
$$\mathbf{F} = \begin{bmatrix} 1 & 0 & \ldots & 0 & 1 \\ 0 & 1 & \ldots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 1 \end{bmatrix}, \qquad \mathbf{F}^* = \frac{1}{D}\begin{bmatrix} D-1 & -1 & \ldots & -1 \\ -1 & D-1 & \ldots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \ldots & D-1 \\ -1 & -1 & \ldots & -1 \end{bmatrix}.$$

3.4. Grafické zobrazení kompozičních dat

Výběrovým prostorem kompozičních dat je simplex, na kterém je zavedena Aitchisonova geometrie. Z toho je evidentní, že kompoziční data nelze zobrazovat

standardním způsobem, jelikož se pohybujeme v geometrii s odlišnými vlastnostmi [14].

U grafického zobrazení kompozičních dat je zvykem, že dvojsložkové kompozice (reprezentované pomocí operace uzávěru jako data s konstantním součtem k) zobrazujeme jako body na intervalu $(0, k)$ a trojsložkové kompozice $\mathbf{x} = (x_1, x_2, x_3)'$, $x_1 + x_2 + x_3 = k$ (nejčastěji $k = 1$) vykreslujeme do tzv. ternárního diagramu. Ternární diagram je rovinný rovnostranný trojúhelník s vrcholy X_1, X_2, X_3. Každá kompozice je zobrazena jako bod, který se nachází uvnitř trojúhelníku, a hodnoty jednotlivých složek představují vzdálenosti tohoto bodu od jednotlivých stran (obrázek 9).



Obrázek č. 9: Ternární diagram pro kompoziční data. Vlevo zakreslena kompozice $\mathbf{x} = \mathcal{C}(1, 1, 1)'$, vpravo kompozice $\mathbf{y} = \mathcal{C}(0.26, 0.1, 0.64)'$.

4. Rozdělení na simplexu

4.1. Úvod

V matematické statistice existuje široká škála statistických modelů. Cílem každé analýzy dat je najít takový model, který by co nejpřesněji odpovídal zkoumaným datům a řešenému problému. Většina statistických metod přitom předpokládá, že naše data jsou realizace reálného náhodného vektoru, tj. že pocházejí z reálného prostoru s euklidovskou geometrií. V tom případě (uvažujeme-li spojitý náhodný vektor) jsou hustoty rozdělení pravděpodobnosti vyjádřeny vzhledem k Lebesgueově pravděpodobnostní míře. V některých případech se ovšem geometrická struktura daného výběrového prostoru může lišit a bude tedy nutné pracovat s jinou mírou než s Lebesgueovou. Jako příklad prostoru s jinou geometrií uveďme reálný prostor $\mathbb{R}_+$ nebo simplex $\mathcal{S}^D$ [13].

Uvažujme náhodnou veličinu či vektor, který má omezený výběrový prostor $E \subset \mathbb{R}^D$, pak by aplikace metod používaných v reálném prostoru mohla vést k absurdním a zkresleným výsledkům, např. nepravé korelaci mezi jednotlivými proporcemi (viz kapitola 2.2.4.).

Jestliže víme, že E je vektorový prostor, na kterém je zaveden skalární součin, můžeme zde zavést pravděpodobnostní míru λ_E, jež bude s danou strukturou prostoru kompatibilní, prostřednictvím Lebesgueovy míry na ortonormálních souřadnicích. Funkce hustoty f_E, která je definována na E, je dána jako Radon–Nikodýmova derivace pravděpodobnostní míry P vzhledem k míře λ_E. Míra λ_E má v prostoru E stejné vlastnosti jako Lebesgueova míra v reálném prostoru.

Hlavní problém nastává v okamžiku, kdy bychom chtěli spočítat pravděpodobnost náhodného jevu A, tj.

$$P(A) = \int_A f_E(\mathbf{x}) d\lambda_E(\mathbf{x}),$$

jelikož se nejedná o standardní integrál, který bychom byli zvyklí běžně řešit. Tuto komplikaci je možné obejít tak, že budeme pracovat v souřadnicích, protože

vlastnosti prostoru souřadnic lze přenést do prostoru E. Například jsme nyní schopni z funkce hustoty f_E na prostoru E získat funkci f, hustotu v souřadnicích, a odtud spočítat pravděpodobnost libovolného jevu $A \subseteq E$ jako

$$P(A) = \int_V f(\mathbf{v}) d\lambda(\mathbf{v}).$$

Přitom V a $\mathbf{v}$ jsou reprezentace A a $\mathbf{x}$ v ortonormálních souřadnicích a λ je Lebesgueova míra v prostoru souřadnic.

4.1.1. Geometrická struktura prostoru $\mathbb{R}_+$

Jednoduchým příkladem, jak se může geometrická struktura prostoru lišit, je výběrový prostor $\mathbb{R}_+$, tj. kladná reálná přímka. Pokud se podíváme na geometrii prostoru $\mathbb{R}^p$, víme, že je na něm zaveden skalární součin a euklidovská vzdálenost kompatibilní s operacemi sčítání vektorů a násobení vektoru skalárem [13]. Otázkou ovšem zůstává, zda je tato geometrie vhodná i pro prostor $\mathbb{R}_+$ [8].

Uvažujme, že dva dny po sobě měříme obsah oxidu siřičitého SO_2 v ovzduší na dvou různých místech. Dostáváme dvě dvojice vzorků; $\{25, 50\}$ a $\{200, 225\}$, kde jsou naměřené hodnoty udávány v $\mu g/m^3$. Vidíme, že absolutní rozdíl mezi hodnotami je stejný, konkrétně 25. V prvním případě ale můžeme říci, že byl naměřen dvojnásobek SO_2, zatímco v druhém případě bylo oxidu uhličitého v ovzduší sice mnohonásobně více, ale změna byla nepatrná.

Při tomto či jiném podobném měření předpokládáme, že měření rozdílu je relativní. Tento předpoklad ovšem není kompatibilní se známou operací sčítání, jelikož není invariantní vůči posunutí. Navíc, pokud bychom přičetli ke kladnému číslu jiné kladné či záporné reálné číslo, nebo jej násobili libovolným reálným číslem, mohlo by dojít k situaci, že se výsledky budou nacházet mimo prostor $\mathbb{R}_+$.

Nechť máme dva prvky $x, y \in \mathbb{R}_+$. Analogii ke sčítání v $\mathbb{R}$ zde představuje standardní součin

$$x \oplus y = x \cdot y$$

a analogií k násobení skalárem je zde operace

$$\alpha \odot x = x^{\alpha}$$

pro každé $\alpha \in \mathbb{R}$. Dále pak definujeme skalární součin, normu a relativní míru rozdílu neboli vzdálenost

$$\langle x, y \rangle_{+} = \ln x \cdot \ln y,$$

$$\| x \|_{+} = |\ln x|,$$

$$d_{+}(x, y) = |\ln x - \ln y|.$$

Prostor $\mathbb{R}_{+}$ je jednorozměrný, a proto má pouze dvě ortonormální báze, jednotkový vektor e a jeho inverzní prvek vzhledem k operaci $\oplus$, tedy e^{-1}. Každý prvek $x \in \mathbb{R}_{+}$ můžeme vyjádřit ve tvaru

$$x = \ln x \odot e = e^{\ln x},$$

což znamená, že $\ln x$ je souřadnicí prvku x vzhledem k ortonormální bázi e. Pro daný interval $(a, b) \subset \mathbb{R}_{+}$ lze definovat míru prostoru $\mathbb{R}_{+}$ jako

$$\lambda_{+}(a, b) = \lambda(\ln a, \ln b) = |\ln b - \ln a|,$$

a odtud plyne, že Jakobián je $d\lambda_{+}/d\lambda = 1/x$.

V praxi vyjadřujeme libovolný vektor $x \in \mathbb{R}_{+}$ většinou jako $x = x \cdot 1$. Problém ovšem je, že 1 je sice báze prostoru $\mathbb{R}$, ale není bází v $\mathbb{R}_{+}$ (má nulovou normu). Lze ovšem využít kanonické báze $\mathbb{R}_{+}$ a psát, že $x = e^{\ln x}$, kde souřadnice vzhledem k bázi je právě $\ln x$. Práce s ortonormálními bázemi nám tedy umožňuje použít standardní statistickou analýzu v souřadnicích. Je zřejmé, že výše zavedené operace (které evokují perturbaci a mocninnou transformaci pro kompozice) jsou ekvivalentní k operacím sčítání a násobení skalárem na souřadnicích, neboli

$$x \oplus y = x \cdot y = e^{\ln x} e^{\ln y} = e^{\ln x + \ln y}, \qquad \alpha \odot x = x^{\alpha} = e^{\ln x^{\alpha}} = e^{\alpha \ln x}.$$

Stejně tak můžeme aplikovat standardní skalární součin a euklidovskou vzdálenost v $\mathbb{R}$ na souřadnice vzhledem ke kanonické bázi v $\mathbb{R}_{+}$, tj.

$$\langle x, y \rangle_{+} = \ln x \cdot \ln y = \langle \ln x, \ln y \rangle_{e},$$

$$d_+(x, y) = |\ln x - \ln y| = d_e(\ln x, \ln y),$$

kde index e zde značí, že se jedná právě o standardní operace v $\mathbb{R}$.

4.2. Aitchisonova míra na simplexu

Při práci s kompozičními daty používáme prostor, jehož přirozenou mírou není míra Lebesgueova. Z tohoto důvodu byla definována alternativní míra, která je simplexovému prostoru vlastní, a byla nazvána Aitchisonova míra λ_a [15]. Tato míra je relativní a odpovídá geometrické struktuře na simplexu. Jednoduchým způsobem, jak tuto míru zavést, je převést Lebesgueovu míru v prostoru ortonormálních souřadnic na simplex [12].

Jak jsme zmínili již v kapitolách 2.2.5 a 3.1, v praxi se často uvažují pod pojmem kompoziční data kladné vektory s konstantním součtem rovným jedné. Pak předpokládáme, že máme danou kompozici $\mathbf{x} = (x_1, \ldots, x_D)' \in \mathcal{S}^D$, kterou můžeme reprezentovat pomocí jiného vektoru, například $\mathbf{x}_- = (x_1, x_2, \ldots, x_{D-1})'$. Vektor $\mathbf{x}_-$ pochází z $(D-1)$-rozměrného prostoru, jelikož složku x_D z kompozice $\mathbf{x}$ je možné vyjádřit ve tvaru

$$x_D = 1 - \sum_{i=1}^{D-1} x_i.$$

Vektor $\mathbf{x}_-$ tak považujeme za prvek prostoru $\mathbb{R}^{D-1}$, kde pracujeme s Lebesgueovou mírou λ.

My ovšem dále s výhodou využijeme obecnější definice kompozic a Aitchisonovy geometrie na simplexu. Nechť jsou dány kompozice $\mathbf{e}_i$ pro $i = 1, 2, \ldots, D-1$, které tvoří ortonormální bázi prostoru $\mathcal{S}^D$, a kompozice $\mathbf{x} \in \mathcal{S}^D$, pak souřadnice této kompozice vzhledem k uvedené bázi označíme $\mathrm{ilr}_i(\mathbf{x})$, kde $i = 1, 2, \ldots, D-1$. V prostoru souřadnic $\mathbb{R}^{D-1}$ máme dán vícerozměrný rovnoběžnostěn R, který je určen dvěma body, $\mathbf{a} = (a_1, a_2, \ldots, a_{D-1})'$ a $\mathbf{b} = (b_1, b_2, \ldots, b_{D-1})'$. Jeho Lebesgueova míra je pak vyjádřena jako součin délek hran ve směru jednotlivých

souřadnic, tj.

$$\lambda_{D-1}(R) = \prod_{i=1}^{D-1} |b_i - a_i|,$$

kde Lebesgueovu míru v $\mathbb{R}^{D-1}$ značíme dolním indexem $D-1$, abychom ji odlišili od předchozí Lebesgueovy míry λ. Rozdíl mezi mírami λ a λ_{D-1} není dán prostorem, na kterém jsou definovány, ale způsobem, jakým jsou prvky $\mathbb{R}^{D-1}$ interpretovány. Pro míru λ je to $D-1$ složek kompozice $\mathbf{x}_-$ a pro míru λ_{D-1} jsou to ilr souřadnice kompozice $\mathbf{x}$. Použitím inverzní transformace ilr^{-1} je možné definovat Aitchisonovu míru množiny $S = \text{ilr}^{-1}(R) \subset \mathcal{S}^D$, kde $R \subset \mathbb{R}^{D-1}$, jako

$$\lambda_a(S) = \lambda_a(\text{ilr}^{-1}(R)) = \lambda_{D-1}(R).$$

Aitchisonova míra je absolutně spojitá vzhledem k míře Lebesgueově.

Uvažujme pravděpodobnostní míru P definovanou na simplexu. Jestliže je P absolutně spojitá vzhledem k míře λ a λ_a, pak Radon-Nikodýmovou derivací P vzhledem k oběma mírám dostaneme funkci hustoty pravděpodobnosti, tj. $\frac{dP}{d\lambda}$, resp. $\frac{dP}{d\lambda_a}$. Jejich integrály přes měřitelnou množinu $S \subset \mathcal{S}^D$ odpovídají pravděpodobnostem

$$P(S) = \int_S \frac{dP}{d\lambda} d\lambda = \int_S \frac{dP}{d\lambda_a} d\lambda_a.$$

Vztah mezi oběma hustotami pravděpodobnosti je dán následovně

$$\frac{dP}{d\lambda} = \frac{dP}{d\lambda_a} \cdot \frac{d\lambda_a}{d\lambda},$$

kde $\frac{d\lambda_a}{d\lambda}$ je Jakobián, který popisuje vztah mezi Lebesgueovou mírou pro složky kompozice na simplexu a mírou Aitchisonovou. Tento Jakobián je dán vztahem

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

Poznamenejme, že stejným způsobem se dá definovat Lebesgueova míra libovolného euklidovského prostoru.

4.2.1. Střed kompozice a její variabilita

Pracujeme-li se souborem dat, která pocházejí z nějakého rozdělení pravděpodobnosti, zajímají nás také jejich číselné charakteristiky. V případě kompozičních dat jsou to zejména střed kompozice a metrický rozptyl.

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

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

nebo ekvivalentně

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

S využitím předchozí rovnosti a vzájemných vztahů, které platí pro logratio transformace získáme následující vztahy

$$\mathrm{alr}(\mathrm{cen}[\mathbf{x}]) = \mathrm{E}[\mathrm{alr}(\mathbf{x})],$$

$$\mathrm{clr}(\mathrm{cen}[\mathbf{x}]) = \mathrm{E}[\mathrm{clr}(\mathbf{x})],$$

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

To znamená, že jsme schopni spočítat střední hodnoty $\mathrm{E}[\mathrm{alr}(\mathbf{x})]$, $\mathrm{E}[\mathrm{clr}(\mathbf{x})]$ a $\mathrm{E}[\mathrm{ilr}(\mathbf{x})]$ použitím standardní definice a následně skrze odpovídající lineární kombinace a aplikací exponentu získat kompozici $\mathrm{cen}(\mathbf{x})$. Můžeme tak tvrdit, že použití standardní statistické metodiky na souřadnice vzhledem k bázi je ekvivalentní s prací přímo na kompozicích. Obecně lze říci, že k určení středu kompozice $\mathrm{cen}(\mathbf{x})$ platí rovnost

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

pro libovolný izomorfismus h. Jinými slovy, můžeme použít koeficienty alr, clr nebo ilr bez rozdílu pro výsledek na simplexu.

Podíváme-li se i na celkový rozptyl náhodného vektoru z hlediska jeho geometrické interpretace, jedná se o střední hodnotu čtvercové euklidovské vzdálenosti

od jeho očekávané hodnoty. Variabilita náhodné kompozice se nazývá metrický rozptyl a je tedy definován jako

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

Vzhledem k tomu, že metrický rozptyl definujeme pomocí euklidovské vzdálenosti, můžeme jej vyjádřit pouze pomocí souřadnic ilr a clr, jelikož báze B alr souřadnic není ortonormální. Aitchisonova vzdálenost mezi $\mathbf{x}$ a $\mathrm{cen}(\mathbf{x})$ a euklidovská vzdálenost mezi příslušnými alr souřadnicemi si nejsou rovny. Pro metrický rozptyl tak platí

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

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

Obecně opět můžeme zapsat jako

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

pro každé h představující izometrické zobrazení, a jak víme, alr transformace izometrická není.

Metrický rozptyl náhodné kompozice se dá zadefinovat i jiným způsobem [12]. Předpokládejme existenci dvou log-kontrastů (standardních kontrastů logaritmovaného vektoru) náhodné kompozice $\mathbf{x}$ s koeficienty $z_1^{(j)}, z_2^{(j)}, \ldots, z_D^{(j)}$ takových, že

$$Z^{(j)}(\mathbf{x}) = \sum_{i=1}^{D} z_i^{(j)} \log x_i, \quad \sum_{i=1}^{D} z_i^{(j)} = 0, \quad j = 1, 2.$$

Variabilita náhodné kompozice $\mathbf{x}$ je popsána jako bilineární forma, která přiřazuje kovarianci každé dvojici log-kontrastů $Z^{(1)}(\mathbf{x}), Z^{(2)}(\mathbf{x})$. Bilineární forma může být vyjádřena různými způsoby, které jsou ovšem vzájemně provázány, záleží na reprezentaci $Z^{(1)}(\mathbf{x}), Z^{(2)}(\mathbf{x})$. Například lze log-kontrasty vyjádřit ve tvaru

$$Z^{(j)}(\mathbf{x}) = \sum_{i=1}^{D} z_i^{(j)} \log \frac{x_i}{g_m(\mathbf{x})} = \sum_{i=1}^{D} z_i^{(j)} \mathrm{clr}_i(\mathbf{x}), \quad j = 1, 2,$$

odtud pro $\mathbf{z}^{(1)} = (\mathbf{z}_1^{(1)}, \ldots, z_D^{(1)})'$ a $\mathbf{z}^{(2)} = (\mathbf{z}_1^{(2)}, \ldots, z_D^{(2)})'$

$$\mathrm{cov}(Z^{(1)}, Z^{(2)}) = [\mathbf{z}^{(1)}]\boldsymbol{\Gamma}\mathbf{z}^{(2)}, \quad [\boldsymbol{\Gamma}]_{ij} = \mathrm{cov}(\mathrm{clr}_i(\mathbf{x}), \mathrm{clr}_j(\mathbf{x})), \quad i, j = 1, 2, \ldots, D.$$

Log-kontrasty můžeme rovněž vyjádřit jako lineární kombinace $D-1$ souřadnic $\mathrm{ilr}(\mathbf{x})$ náhodné kompozice $\mathbf{x}$ vzhledem k ortonormální bázi na simplexu $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}$. Potom dostáváme varianční matici $\boldsymbol{\Upsilon} = \mathrm{var}(\mathrm{ilr}(\mathbf{x})) \in \mathbb{R}^{(D-1)\times(D-1)}$, která představuje druhý moment kompozice $\mathbf{x}$. Analogicky lze odvodit matici $\boldsymbol{\Sigma} = \mathrm{var}(\mathrm{alr}(\mathbf{x}))$ stejného řádu jako matice $\boldsymbol{\Upsilon}$, kde jsou log-kontrasty vyjádřeny pomocí lineární kombinace alr souřadnic. Vztahy mezi variančními matice $\boldsymbol{\Sigma}, \boldsymbol{\Gamma}$ a $\boldsymbol{\Upsilon}$ jsou následující

$$\boldsymbol{\Gamma} = \mathbf{F}^*\boldsymbol{\Sigma}\mathbf{F}'^*,$$

$$\boldsymbol{\Upsilon} = (\mathbf{U}'\mathbf{F}^*)\boldsymbol{\Sigma}(\mathbf{U}'\mathbf{F}^*)' = \mathbf{U}'\boldsymbol{\Gamma}\mathbf{U},$$

kde matice $\mathbf{U} \in \mathbb{R}^{D\times(D-1)}$ a $\mathbf{F}^* \in \mathbb{R}^{D\times(D-1)}$ byly definovány v kapitole 3.3.

Metrický rozptyl můžeme následně spočítat také jako

$$\mathrm{Mvar}[\mathbf{x}] = \mathrm{trace}(\boldsymbol{\Gamma}) = \mathrm{trace}(\boldsymbol{\Upsilon}).$$

Jelikož alr transformace není izometrická, platí $\mathrm{trace}(\boldsymbol{\Gamma}) \neq \mathrm{trace}(\boldsymbol{\Sigma})$, tedy

$$\mathrm{trace}(\boldsymbol{\Gamma}) = \mathrm{trace}(\boldsymbol{\Sigma}) - D^{-1}\mathbf{1}'_{D-1}\boldsymbol{\Sigma}\mathbf{1}_{D-1},$$

kde $\mathbf{1}_{D-1}$ představuje sloupcový vektor jedniček o $D-1$ složkách.

4.3. Normální rozdělení na simplexu

Následující kapitola se věnuje problematice mnohorozměrného normálního rozdělení na simplexu. Základní myšlenkou je transformovat náhodnou kompozici ze simplexu do reálného euklidovského prostoru. V tomto prostoru pak standardními metodami definujeme hustotu transformovaného vektoru, kterou poté převedeme zpět na simplex [9].

Definice 4.1. *Náhodný vektor* $\mathbf{x}$ *má normální rozdělení na simplexu* $\mathcal{S}^D$ *právě tehdy, když vektor ortonormálních souřadnic,* $\mathrm{ilr}(\mathbf{x})$*, má mnohorozměrné normální rozdělení na* $\mathbb{R}^{D-1}$*.*

Normální rozdělení na simplexu lze definovat pomocí funkce hustoty [9].

Definice 4.2. *Řekneme, že náhodná kompozice má normální rozdělení na $\mathcal{S}^D$ s parametry $\boldsymbol{\xi}$ a $\boldsymbol{\Upsilon}$, jestliže funkce hustoty souřadnic vzhledem k ortonormální bázi $\mathcal{S}^D$ má tvar*

$$f^*_{\mathbf{x}}(\mathbf{x}) = (2\pi)^{-(D-1)/2}|\boldsymbol{\Upsilon}|^{-1/2}\exp\Big[-\frac{1}{2}(\mathrm{ilr}(\mathbf{x})-\boldsymbol{\xi})'\boldsymbol{\Upsilon}^{-1}(\mathrm{ilr}(\mathbf{x})-\boldsymbol{\xi})\Big]. \qquad (2)$$

Zapisujeme $\mathbf{x} \sim N^D_{\mathcal{S}}(\boldsymbol{\xi}, \boldsymbol{\Upsilon})$. Dolní index $\mathcal{S}$ nám říká, že se jedná o model definovaný na simplexu, a horní index D udává počet složek kompozice. Parametry $\boldsymbol{\xi}$ a $\boldsymbol{\Upsilon}$ odkazují na vektor středních hodnot a varianční matici vektoru souřadnic $\mathrm{ilr}(\mathbf{x})$.

Hustota $f^*_{\mathbf{x}}(\mathbf{x})$ opravdu odpovídá hustotě pravděpodobnosti normálně rozděleného náhodného vektoru v $\mathbb{R}^{D-1}$.

$f^*_{\mathbf{x}}(\mathbf{x})$ je hustota souřadnic $\mathbf{x}$ vzhledem k ortonormální bázi na $\mathcal{S}^D$ a proto se jedná o Radon–Nikodýmovu derivaci vzhledem k Lebesgueově míře v $\mathbb{R}^{D-1}$, který představuje prostor souřadnic. Tato znalost nám umožňuje spočítat pravděpodobnost libovolného jevu $A \subset \mathcal{S}^D$ pomocí obyčejného integrálu jako

$$P(A) = \int_{A^*} (2\pi)^{-(D-1)/2}|\boldsymbol{\Upsilon}|^{-1/2}\exp\Big[-\frac{1}{2}(\mathrm{ilr}(\mathbf{x})-\boldsymbol{\xi})'\boldsymbol{\Upsilon}^{-1}(\mathrm{ilr}(\mathbf{x})-\boldsymbol{\xi})\Big]d\lambda_{D-1}(\mathrm{ilr}(\mathbf{x})),$$

kde A^* představuje souřadnice A vzhledem k dané ortonormální bázi a λ_{D-1} je Lebesgueova míra v prostoru $\mathbb{R}^{D-1}$, tedy pro $\mathrm{ilr}(\mathbf{x}) = \mathbf{z} = (z_1, \ldots, z_{D-1})'$ obdržíme

$$P(A) = \int_{A^*} (2\pi)^{-(D-1)/2}|\boldsymbol{\Upsilon}|^{-1/2}\exp\Big[-\frac{1}{2}(\mathbf{z}-\boldsymbol{\xi})'\boldsymbol{\Upsilon}^{-1}(\mathbf{z}-\boldsymbol{\xi})\Big]dz_1\cdots dz_{D-1}.$$

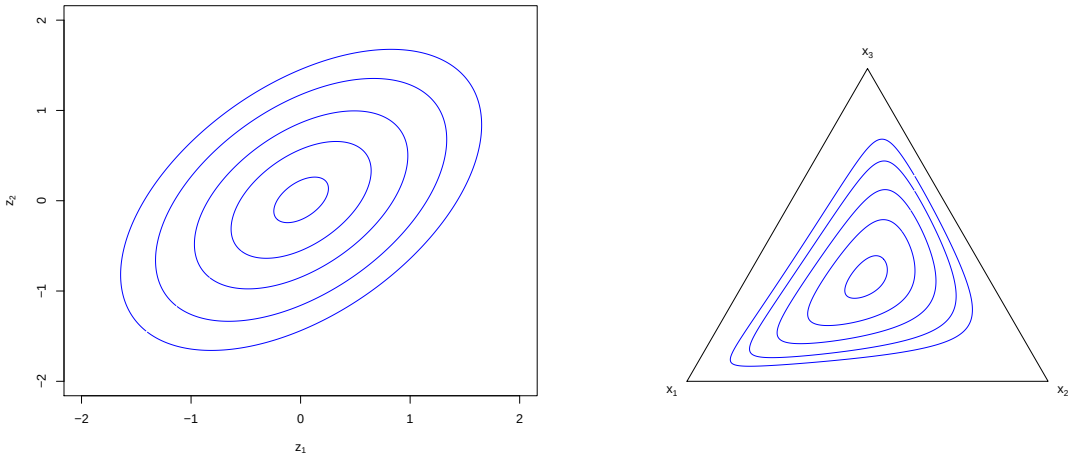
Funkce hustoty pravděpodobnosti tzv. aditivně logistického modelu [1] odpovídá Radon–Nikodýmově derivaci vzhledem k Lebesgueově míře λ v $\mathbb{R}^{D-1}$. Pak lze tedy pravděpodobnost libovolného jevu $A \subset \mathcal{S}^D$ spočítat standardním integrálem

$$P(A) = \int_{A} \frac{D^{-1/2}(\prod_{i=1}^{D} x_i)^{-1}}{(2\pi)^{-(D-1)/2}|\boldsymbol{\Upsilon}|^{-1/2}}\exp\Big[-\frac{1}{2}(\mathrm{ilr}(\mathbf{x})-\boldsymbol{\xi})'\boldsymbol{\Upsilon}^{-1}(\mathrm{ilr}(\mathbf{x})-\boldsymbol{\xi})\Big]dx_1\cdots dx_{D-1}.$$

Přestože je interpretace hustoty v tomto případě jiná, významy parametrů $\boldsymbol{\xi}$ a $\boldsymbol{\Upsilon}$ zůstávají stejné. Tuto hustotu jsme vzhledem k diskuzi v úvodu kapitoly 4.2 obdrželi s využitím Jakobiánu $\frac{d\lambda_a}{d\lambda} = D^{-1/2}(\prod_{i=1}^{D} x_i)^{-1}$.

Pravděpodobnost jevu $A \subset \mathcal{S}^D$ je stejná nezávisle na tom, který model použijeme, zda normální na $\mathcal{S}^D$ či logistický normální. Oba tyto modely jsou tedy ekvivalentní na $\mathcal{S}^D$ z hlediska výpočtu pravděpodobnosti. Naopak model normality na $\mathcal{S}^D$ a model normality aditivně logistické nejsou ekvivalentní vzhledem ke geometrii daného prostoru.

Obrázek 10 ukazuje chování vrstevnic hustoty normálního rozdělení na $\mathcal{S}^3$, jak v prostoru ortonormálních souřadnic, tak na simplexu.



Obrázek č. 10: Hustota normálního rozdělení na simplexu s parametry $\boldsymbol{\xi} = (0, 0)'$, $\Upsilon_{11} = \Upsilon_{22} = 0.3$ a $\Upsilon_{12} = \Upsilon_{21} = 0.15$. Vlevo vzhledem k Lebesgueově míře v prostoru ilr souřadnic $\mathbb{R}^2$, vpravo vzhledem k Aitchisonově míře na simplexu.

Věta 4.1. *Nechť* $\mathbf{x} \sim \mathcal{N}_{\mathcal{S}}^{D}(\boldsymbol{\xi}, \boldsymbol{\Upsilon})$ *je D-složková náhodná kompozice. Nechť* $\mathbf{a} \in \mathcal{S}^D$ *je kompozice konstant a* $b \in \mathbb{R}$ *je skalár. Pak D-složková kompozice* $\mathbf{x}^* = \mathbf{a} \oplus (b \odot \mathbf{x})$ *má rozdělení* $\mathcal{N}_{\mathcal{S}}^{D}(\mathrm{ilr}(\mathbf{a}) + b\boldsymbol{\xi}, b^2\boldsymbol{\Upsilon})$.

Důkaz. Ilr souřadnice kompozice $\mathbf{x}^*$ získáme pomocí lineární transformace ilr sou-

řadnic kompozice $\mathbf{x}$, protože $\mathrm{ilr}(\mathbf{x}^*) = \mathrm{ilr}(\mathbf{a}) + b \cdot \mathrm{ilr}(\mathbf{x})$. Můžeme pracovat s funkcí hustoty ilr koeficientů $\mathbf{x}$ jako s hustotou v reálném prostoru. Tedy v případě, že se jedná o funkci hustoty v reálném prostoru, můžeme použít větu o lineární transformaci (věta 2.1) a dostaneme tak hustotu náhodného vektoru $\mathrm{ilr}(\mathbf{x}^*)$. Střední hodnotu a varianční matici vektoru $\mathrm{ilr}(\mathbf{x}^*)$ získáme standardním výpočtem

$$\begin{aligned} \mathrm{E}[\mathrm{ilr}(\mathbf{x}^*)] &= \mathrm{E}[\mathrm{ilr}(\mathbf{a}) + b\mathrm{ilr}(\mathbf{x})] = \mathrm{ilr}(\mathbf{a}) + b\mathrm{E}[\mathrm{ilr}(\mathbf{x})] = \mathrm{ilr}(\mathbf{a}) + b\boldsymbol{\xi}, \\ \mathrm{var}[\mathrm{ilr}(\mathbf{x}^*)] &= \mathrm{var}[\mathrm{ilr}(\mathbf{a}) + b\mathrm{ilr}(\mathbf{x})] = b^2\mathrm{var}[\mathrm{ilr}(\mathbf{x})] = b^2\boldsymbol{\Upsilon}. \end{aligned}$$

Tedy $\mathbf{x}^* \sim \mathcal{N}_{\mathcal{S}}^D(\mathrm{ilr}(\mathbf{a}) + b\boldsymbol{\xi}, b^2\boldsymbol{\Upsilon})$. □

Věta 4.2. *Nechť máme náhodnou kompozici* $\mathbf{x} \sim \mathcal{N}_{\mathcal{S}}^D(\boldsymbol{\xi}, \boldsymbol{\Upsilon})$ *a vektor konstant* $\mathbf{a} \in \mathcal{S}^D$. *Pak* $f^*_{\mathbf{a}\oplus\mathbf{x}}(\mathbf{a} \oplus \mathbf{x}) = f^*_{\mathbf{x}}(\mathbf{x})$, *kde* $f^*_{\mathbf{a}\oplus\mathbf{x}}$ *a* $f^*_{\mathbf{x}}$ *značí hustoty pravděpodobnosti náhodných kompozic* $\mathbf{x}$ *a* $\mathbf{a} \oplus \mathbf{x}$.

Důkaz. Z předchozí věty víme, že $\mathbf{a} \oplus \mathbf{x} \sim \mathcal{N}_{\mathcal{S}}^D(\mathrm{ilr}(\mathbf{a}) + \boldsymbol{\xi}, \boldsymbol{\Upsilon})$. Proto

$$\begin{aligned} &f^*_{\mathbf{a}\oplus\mathbf{x}}(\mathbf{a} \oplus \mathbf{x}) = (2\pi)^{-(D-1)/2}|\boldsymbol{\Upsilon}|^{-1/2} \\ &\times \exp\Big[-\frac{1}{2}(\mathrm{ilr}(\mathbf{a} \oplus \mathbf{x}) - (\mathrm{ilr}(\mathbf{a}) + \boldsymbol{\xi}))'\boldsymbol{\Upsilon}^{-1}(\mathrm{ilr}(\mathbf{a} \oplus \mathbf{x}) - (\mathrm{ilr}(\mathbf{a}) + \boldsymbol{\xi}))\Big] \\ &= (2\pi)^{-(D-1)/2}|\boldsymbol{\Upsilon}|^{-1/2} \\ &\times \exp\Big[-\frac{1}{2}(\mathrm{ilr}(\mathbf{a}) + \mathrm{ilr}(\mathbf{x}) - (\mathrm{ilr}(\mathbf{a}) + \boldsymbol{\xi}))'\boldsymbol{\Upsilon}^{-1}(\mathrm{ilr}(\mathbf{a}) + \mathrm{ilr}(\mathbf{x}) - (\mathrm{ilr}(\mathbf{a}) + \boldsymbol{\xi}))\Big] \\ &= (2\pi)^{-(D-1)/2}|\boldsymbol{\Upsilon}|^{-1/2} \exp\Big[-\frac{1}{2}(\mathrm{ilr}(\mathbf{x}) - \boldsymbol{\xi})'\boldsymbol{\Upsilon}^{-1}(\mathrm{ilr}(\mathbf{x}) - \boldsymbol{\xi})\Big] = f^*_{\mathbf{x}}(\mathbf{x}). \end{aligned}$$

□

Věta 4.3. *Nechť* $\mathbf{x} \sim \mathcal{N}_{\mathcal{S}}^D(\boldsymbol{\xi}, \boldsymbol{\Upsilon})$ *je* D*-složková náhodná kompozice. Nechť* $\mathbf{x}_P = \mathbf{P}\mathbf{x}$ *je kompozice, jejíž složky jsou uspořádány permutační maticí* $\mathbf{P}$. *Potom* $\mathbf{x}_P$ *má rozdělení* $\mathcal{N}_{\mathcal{S}}^D(\boldsymbol{\xi}_P, \boldsymbol{\Upsilon}_P)$, *a platí*

$$\boldsymbol{\xi}_P = \mathbf{U}'\mathbf{P}\mathbf{U}\boldsymbol{\xi} \qquad \boldsymbol{\Upsilon}_P = (\mathbf{U}'\mathbf{P}\mathbf{U})\boldsymbol{\Upsilon}(\mathbf{U}'\mathbf{P}\mathbf{U})',$$

kde sloupce matice $\mathbf{U} \in \mathbb{R}^{D\times(D-1)}$ *jsou vektory* $\mathrm{clr}(\mathbf{e}_i)$, $i = 1, 2, \ldots, D-1$.

Důkaz. Abychom zjistili, jak vypadá rozdělení náhodné kompozice $\mathbf{x}_P$, potřebujeme znát maticový vztah mezi ilr souřadnicemi obou kompozic $\mathbf{x}$ a $\mathbf{x}_P$. Pracujeme-li s clr koeficienty, používáme vztah $\mathrm{clr}(\mathbf{x}_P) = \mathbf{P}\mathrm{clr}(\mathbf{x})$. S využitím znalosti vztahů mezi logratio transformacemi získáme rovnost $\mathrm{ilr}(\mathbf{x}_P) = (\mathbf{U}'\mathbf{P}\mathbf{U})\mathrm{ilr}(\mathbf{x})$. Jestliže má vektor $\mathrm{ilr}(\mathbf{x})$ normální rozdělení, můžeme opět aplikovat větu 2.1 o lineární transformaci normálního rozdělení v reálném prostoru, abychom dokázali, že náhodná kompozice $\mathbf{x}_P$ má normální rozdělení $\mathcal{N}_{\mathcal{S}}^{D}(\mathbf{U}'\mathbf{P}\mathbf{U}\boldsymbol{\xi}, (\mathbf{U}'\mathbf{P}\mathbf{U})\boldsymbol{\Upsilon}(\mathbf{U}'\mathbf{P}\mathbf{U})')$. □

Jako důsledek zmíníme větu o rozdělení podkompozice, jejíž důkaz je uveden v [9].

Věta 4.4. *Nechť* $\mathbf{x} \sim \mathcal{N}_{\mathcal{S}}^{D}(\boldsymbol{\xi}, \boldsymbol{\Upsilon})$ *je D-složková náhodná kompozice. Nechť* $\mathbf{s} = \mathcal{C}(\mathbf{S}\mathbf{x})$ *je C-složková podkompozice, kterou jsme získali ze selekční matice* $\mathbf{S} \in \mathbb{R}^{C\times D}$. *Přitom matice* $\mathbf{S}$ *má C prvků rovných jedné (jeden v každém řádku a nejvýš jeden v každém sloupci) a zbývající prvky rovny nule. Pak* $\mathbf{s}$ *má rozdělení* $\mathcal{N}_{\mathcal{S}}^{C}(\boldsymbol{\xi}_S, \boldsymbol{\Upsilon}_S)$, *a platí*

$$\boldsymbol{\xi}_S = \mathbf{U}^{*\prime}\mathbf{S}\mathbf{U}\boldsymbol{\xi} \qquad \boldsymbol{\Upsilon}_S = (\mathbf{U}^{*\prime}\mathbf{S}\mathbf{U})\boldsymbol{\Upsilon}(\mathbf{U}^{*\prime}\mathbf{S}\mathbf{U})',$$

kde $\mathbf{U} \in \mathbb{R}^{D\times(D-1)}$ *se sloupci tvořenými vektory clr koeficientů ortonormální báze* $\mathcal{S}^D$, *a* $\mathbf{U}^* \in \mathbb{R}^{C\times(C-1)}$ *se sloupci tvořenými vektory clr koeficientů odpovídající ortonormální báze* $\mathcal{S}^C$.

Věta 4.5. *Nechť* $\mathbf{x} \sim \mathcal{N}_{\mathcal{S}}^{D}(\boldsymbol{\xi}, \boldsymbol{\Upsilon})$ *je D-složková náhodná kompozice a nechť máme* $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_{D-1})'$. *Pak* $\mathrm{cen}(\mathbf{x}) = (\xi_1 \odot \mathbf{e}_1) \oplus (\xi_2 \odot \mathbf{e}_2) \oplus \cdots \oplus (\xi_{D-1} \odot \mathbf{e}_{D-1})$, *kde* $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$ *je ortonormální báze* $\mathcal{S}^D$.

Důkaz. Střední hodnota každého náhodného vektoru je prvek daného prostoru. Použijeme-li standardní definici očekávané hodnoty na souřadnice kompozice $\mathbf{x}$ vzhledem k ortonormální bázi $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$ s využitím hustoty (2), dostaneme souřadnice kompozice $\mathrm{cen}(\mathbf{x})$ vzhledem k uvažované ortonormální bázi.

Aplikací vztahů z kapitoly 4.2.1 získáme $\mathrm{ilr}(\mathrm{cen}(\mathbf{x})) = \mathrm{E}[\mathrm{ilr}(\mathbf{x})] = \boldsymbol{\xi}$. Kompozici $\mathrm{cen}(\mathbf{x})$ pak obdržíme jako lineární kombinaci $(\xi_1 \odot \mathbf{e}_1) \oplus (\xi_2 \odot \mathbf{e}_2) \oplus \cdots \oplus (\xi_{D-1} \odot \mathbf{e}_{D-1})$. □

Věta 4.6. *Nechť* $\mathbf{x} \sim \mathcal{N}_{\mathcal{S}}^{D}(\boldsymbol{\xi}, \boldsymbol{\Upsilon})$ *je D-složková náhodná kompozice. Pak metrický rozptyl* $\mathrm{Mvar}[\mathbf{x}] = \mathrm{trace}(\boldsymbol{\Upsilon})$.

Důkaz. Metrický rozptyl je definován jako $\mathrm{Mvar}[\mathbf{x}] = \mathrm{E}[d_a^2(\mathbf{x}, \mathrm{cen}[\mathbf{x}])]$. Aitchisonova vzdálenost d_a mezi dvěma kompozicemi je stejná jako euklidovská vzdálenost d_e mezi odpovídajícími souřadnicemi vzhledem k ortonormální bázi. Můžeme tedy psát $\mathrm{Mvar}[\mathbf{x}] = \mathrm{E}[d_e^2(\mathrm{ilr}(\mathbf{x}), \mathrm{E}[\mathrm{ilr}(\mathbf{x})])]$. Tato hodnota je rovna stopě matice $\mathrm{var}(\mathrm{ilr}(\mathbf{x}))$, a tedy použitím varianční matice normálního rozdělení v reálném prostoru dostáváme vztah $\mathrm{Mvar}[\mathbf{x}] = \mathrm{trace}(\boldsymbol{\Upsilon})$. □

4.4. Dirichletovo rozdělení na simplexu

Dirichletovo rozdělení se na první pohled jeví jako vhodný nástroj pro práci s kompozičními daty. Předpokládá simplex jako výběrový prostor, neboli součet složek kompozice je roven 1. Začneme-li se tímto rozdělením zabývat trochu více, zjistíme, že skutečnost je jiná.

Jak již bylo zmíněno v kapitole 1.2, konstrukce Dirichletova rozdělení vyžaduje nezávislost příslušných kompozičních složek. Tato vlastnost ovšem není použitelná z hlediska popisu kompozic dle definice 3.1, kdy je nezávislost jednotlivých složek nemyslitelná. Předpoklad nezávislosti, který platí pro každou složku kompozice u Dirichletova rozdělení vzhledem k Lebesgueově míře, je ovšem i v tomto standardním případě velmi silný a v praktické situaci téměř neaplikovatelný. Cílem je tedy najít obecnější třídu Dirichletova rozdělení, navíc vzhledem k Aitchisonově míře na simplexu, kdy již předpoklad nezávislosti složek kompozice není klíčový.

4.4.1. Dirichletovo rozdělení na simplexu

Připomeňme si definici Dirichletova rozdělení z kapitoly 2.2.4 [10,11,12].

Definice 4.3. *Náhodný vektor* $\mathbf{X} \in \mathcal{S}^D$ *má D-rozměrné Dirichletovo rozdělení s parametrem* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)' \in \mathbb{R}_+^D$*, jestliže jeho hustota pravděpodobnosti má tvar*

$$f(\mathbf{x}) = \frac{dP}{d\lambda}(\mathbf{x}) = \frac{\Gamma(\alpha_+)}{\prod_{i=1}^D \Gamma(\alpha_i)} \prod_{i=1}^D x_i^{\alpha_i - 1}, \tag{3}$$

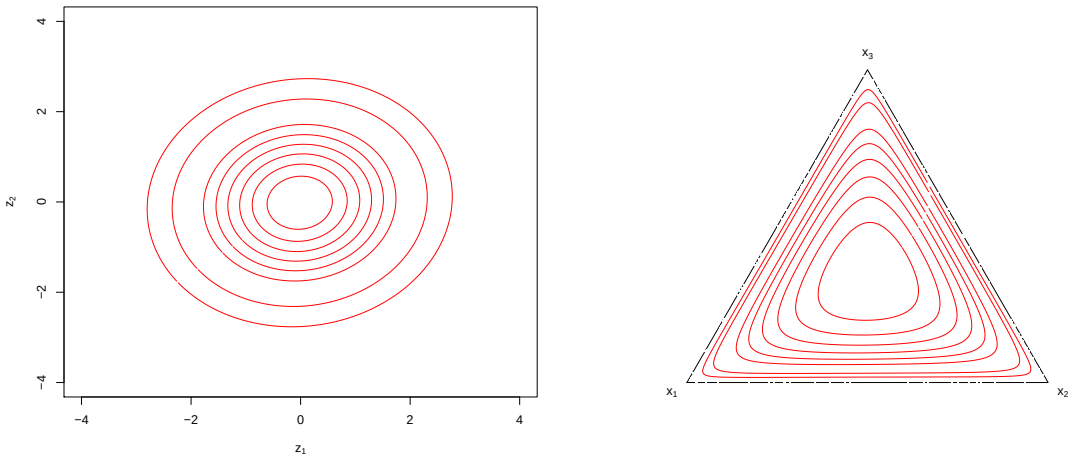
kde P je Dirichletova pravděpodobnostní míra, $\alpha_+ = \sum_{i=1}^D \alpha_i$*, a* Γ *je gamma funkce. Značíme* $\mathbf{X} \sim \mathcal{D}^D(\boldsymbol{\alpha})$*.*

Rovnice (3) představuje funkci hustoty jako Radon–Nikodýmovu derivaci vzhledem k Lebesgueově míře prostoru o $D - 1$ složkách kompozice. Když změníme míru a vyjádříme hustotu vzhledem k Aitchisonově míře λ_a, dostaneme funkci hustoty Dirichletova rozdělení ve tvaru

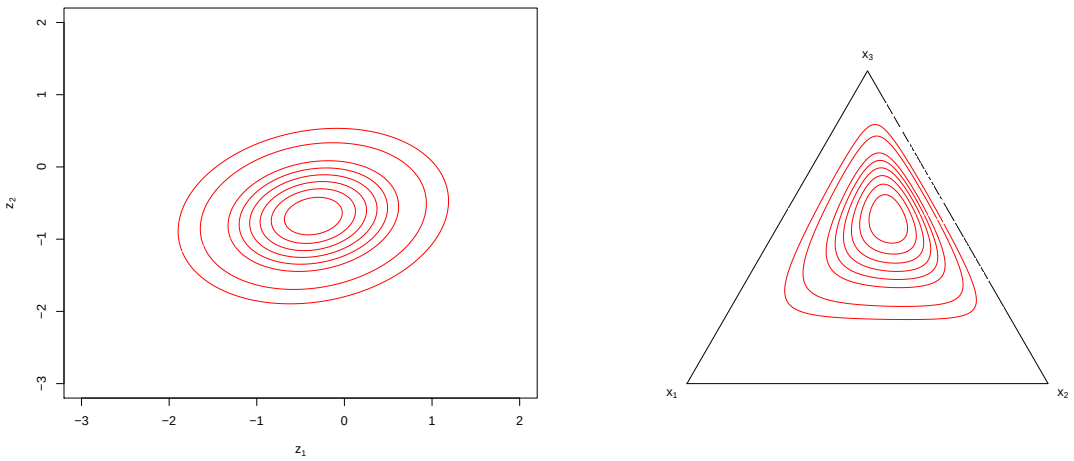
$$f_a(\mathbf{x}) = \frac{dP}{d\lambda_a}(\mathbf{x}) = \frac{\Gamma(\alpha_+)\sqrt{D}}{\prod_{i=1}^D \Gamma(\alpha_i)} \prod_{i=1}^D x_i^{\alpha_i}.$$

Explicitní vyjádření hustoty Dirichletova rozdělení vzhledem k Lebesgueově míře v prostoru ilr souřadnic je značně komplikované, tudíž i interpretace jednotlivých parametrů $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)'$ je téměř nemožná.

Obrázky 11 a 12 znázorňují vrstevnice funkce hustoty Dirichletova rozdělení v prostoru ortonormálních souřadnic a na simplexu s různou volbou parametru $\boldsymbol{\alpha}$ (zdrojové kódy k softwaru R jsou k dispozici na přiloženém CD).



Obrázek č. 11: Funkce hustoty Dirichletova rozdělení s parametrem $\boldsymbol{\alpha} = (1, 1, 1)'$. Vlevo vzhledem k Lebesgueově míře v prostoru ilr souřadnic $\mathbb{R}^2$, vpravo vzhledem k Aitchisonově míře na simplexu.



Obrázek č. 12: Funkce hustoty Dirichletova rozdělení s parametrem $\boldsymbol{\alpha} = (2, 3, 5)'$. Vlevo vzhledem k Lebesgueově míře v prostoru ilr souřadnic $\mathbb{R}^2$, vpravo vzhledem k Aitchisonově míře na simplexu.

Pokud zaměníme dvě libovolné náhodné veličiny X_i a X_j pro $i, j = 1, \ldots, D$, $i \neq j$ a zároveň zaměníme i odpovídající parametry α_i a α_j, pak funkce hustoty zůstává stejná. O hustotě Dirichletova rozdělení tedy můžeme tvrdit, že je abso-

lutně permutačně symetrická. Následující věta opět přímo navazuje na kapitolu 2.2.4.

Věta 4.7. *Nechť* $\mathbf{X} \sim \mathcal{D}^D(\boldsymbol{\alpha})$*, pak modus a střední hodnota Dirichletova rozdělení vzhledem k míře* λ_a *mají následující tvar:*

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

$$\mathrm{E}(\mathbf{X})_a = \mathcal{C}\left(e^{\psi(\alpha_1)}, \ldots, e^{\psi(\alpha_D)}\right)', \tag{4}$$

kde $\psi(t) = \frac{\partial ln\Gamma(t)}{\partial t}$ *je digamma funkce a* $\mathcal{C}$ *je operátor uzávěru.*

Zatímco určení modu kompozice $\mathbf{X}$ je velmi podobné a výpočetně snadné vzhledem k míře Lebesgueově i Aitchisonově, u střední hodnoty je to mnohem komplikovanější. Střední hodnota vzhledem k Lebesgueově míře je stejná jako modus vzhledem k míře λ_a. Nejjednodušším způsobem, jak určit očekávanou hodnotu kompozice $\mathbf{X}$ vzhledem k Aitchisonově míře spočívá ve vyjádření funkce hustoty Dirichletova rozdělení pomocí souřadnic vzhledem k ortonormální bázi a následně aplikovat standardní definici střední hodnoty na vektor $\operatorname{ilr}(\mathbf{x})$ [10]. Výsledkem jsou souřadnice kompozice $\mathrm{E}_a(\mathbf{X})$ vzhledem k dané ortonormální bázi a použitím vztahů z kapitoly 4.2.1 se dostaneme k (4).

Ke zjištění variability, např. k výpočtu metrického rozptylu, je nutné pracovat přímo na souřadnicích, tedy s $(D-1)$-složkovými vektory, jelikož metrický rozptyl není prvkem simplexu [12]. Je to pouze numericky určená hodnota, která vyjadřuje míru celkové disperze kompozice. Dosud nebyl zjištěn žádný způsob, kterým bychom získali explicitní vyjádření, aniž by bylo nutné volit ortonormální bázi pro reprezentaci kompozice.

Věta 4.8. *Předpokládejme, že máme náhodný vektor* $\mathbf{X} \sim \mathcal{D}^D(\boldsymbol{\alpha})$ *definovaný na jednotkovém simplexu* $\mathcal{S}^D$*, pak metrický rozptyl* $\mathbf{X}$ *je*

$$\operatorname{Mvar}(\mathbf{X}) = \frac{D-1}{D}(\psi'(\alpha_1) + \cdots + \psi'(\alpha_D)),$$

kde $\psi'(t)$ *pro* $t > 0$ *je trigamma funkce.*

4.4.2. Posunuté Dirichetovo rozdělení na simplexu

Posunuté Dirichletovo rozdělení získáme tak, že použijeme operaci perturbace na náhodnou kompozici, která má funkci hustoty pravděpodobnosti Dirichletova rozdělení [11,12].

Definice 4.4. *Náhodný vektor $\mathbf{X} \in \mathcal{S}^D$ má posunuté Dirichletovo rozdělení s parametry $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)' \in \mathbb{R}_+^D$ a $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_D)' \in \mathbb{R}_+^D$, jestliže jeho funkce hustoty pravděpodobnosti má tvar*

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

kde P je Dirichletova pravděpodobnostní míra, $\alpha_+ = \sum_{i=1}^D \alpha_i$, a Γ je gamma funkce. Značíme $\mathbf{X} \sim \mathcal{SD}^D(\boldsymbol{\alpha}, \boldsymbol{\beta})$.

Počet parametrů je $2D$. Jestliže položíme parametr $\boldsymbol{\beta}$ roven $(1, \ldots, 1)'$, $\mathcal{C}(1, \ldots, 1)'$ nebo $\mathcal{C}(\beta, \ldots, \beta)'$, obdržíme klasický Dirichletův model. Stejným způsobem jako u Dirichletova rozdělení můžeme vyjádřit hustotu posunutého Dirichletova rozdělení vzhledem k míře λ_a,

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

Věta 4.9. *Posunuté Dirichletovo rozdělení vznikne normováním vektoru celkem D nezávislých náhodných veličin $W_i \sim \Gamma(\alpha_i, \beta_i), i = 1, \ldots, D$, tj. jestliže $\mathbf{X} = \mathcal{C}(\mathbf{W})$, pak $\mathbf{X} \sim \mathcal{SD}^D(\boldsymbol{\alpha}, \boldsymbol{\beta})$.*

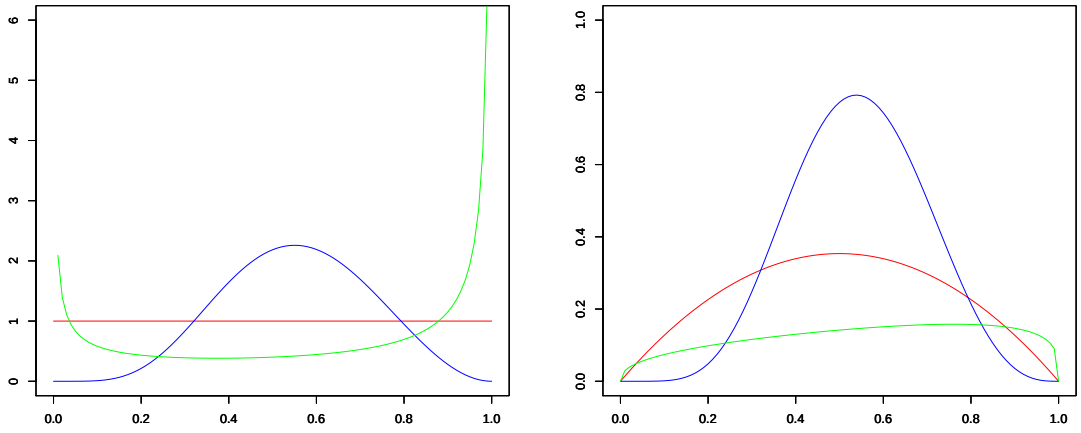
Máme-li náhodnou kompozici $\mathbf{X} \sim \mathcal{D}^D$ definovanou na $\mathcal{S}^D$ a kompozici $\mathbf{p} \in \mathcal{S}^D$, pak náhodná kompozice $\widetilde{\mathbf{X}} = \mathbf{p} \oplus \mathbf{X}$ má rozdělení $\mathcal{SD}^D(\boldsymbol{\alpha}, \boldsymbol{\beta} = \mathbf{p}^{-1})$. Parametr $\boldsymbol{\beta}$ je v tomto případě prvek prostoru $\mathcal{S}^D$, tj. kompozice, a tedy počet parametrů v modelu je $2D - 1$, navíc, proporcionální reprezentace $\boldsymbol{\beta}$ vedou ke stejnému rozdělení. Vzhledem ke geometrické struktuře na simplexu je posunuté Dirichletovo rozdělení posunem Dirichletova rozdělení v rámci simplexu. Můžeme tedy říci, že tato dvě rozdělení patří do stejné třídy rozdělení.

Pokud je $D = 2$, budou mít odpovídající funkce hustoty tvar

$$f_s(\mathbf{x}) = \frac{dP}{d\lambda}(\mathbf{x}) = \frac{1}{B(\alpha_1, \alpha_2)} \frac{\beta_1^{\alpha_1} x^{\alpha_1 - 1} \beta_2^{\alpha_2} (1-x)^{\alpha_2 - 1}}{(\beta_1 x + \beta_2 (1-x))^{\alpha_1 + \alpha_2}},$$

$$f_{sa}(\mathbf{x}) = \frac{dP}{d\lambda_a}(\mathbf{x}) = \frac{\sqrt{2}}{B(\alpha_1, \alpha_2)} \frac{(\beta_1 x)^{\alpha_1} (\beta_2 (1-x))^{\alpha_2}}{(\beta_1 x + \beta_2 (1-x))^{\alpha_1 + \alpha_2}}.$$

Víme, že pro $D = 2$ se jedná o beta rozdělení, a proto i v tomto případě mluvíme o posunutém beta rozdělení.



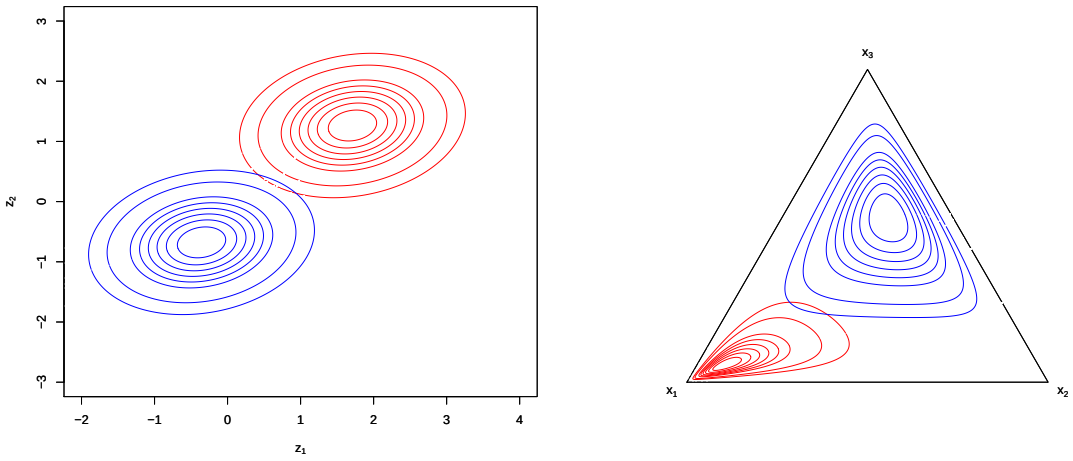
Obrázek č. 13: Funkce hustoty posunutého Dirichletova rozdělení pro $D = 2$. Na levém obrázku vzhledem k Lebesgueově míře λ, na pravém obrázku vzhledem k Aitchisonově míře λ_a. Pro parametry $\boldsymbol{\alpha} = (1,1)', \boldsymbol{\beta} = (1,1)'$ červená křivka, $\boldsymbol{\alpha} = (7,3)', \boldsymbol{\beta} = (0.6, 0.3)'$ modrá křivka, pro $\boldsymbol{\alpha} = (0.4, 0.25)'$, $\boldsymbol{\beta} = (0.4, 0.8)'$ zelená křivka.

Na obrázku 13 je znázorněna hustota vzhledem k Lebesgueově míře λ na intervalu $\langle 0, 1 \rangle$ a hustota vzhledem k Aitchisonově míře λ_a na $\mathcal{S}^2$. V případě hustoty vzhledem k míře λ_a dostáváme vždy unimodální funkci. Pro posunuté beta rozdělení vzhledem k míře λ tato vlastnost neplatí. Z obrázku je patrné, že pro $\boldsymbol{\alpha} = (1,1)'$ a $\boldsymbol{\beta} = (1,1)'$ je funkce hustoty konstantní a pro $\boldsymbol{\alpha} = (0.4, 0.25)'$ a $\boldsymbol{\beta} = (0.4, 0.8)'$ obdržíme funkci hustoty, která má v bodech 0 a 1 vertikální

asymptoty. Analogicky se funkce hustoty posunutého Dirichletova rozdělení chová i pro $D > 2$.

Věnujme se situaci $D = 3$. Na obrázku 14 jsou modrou barvou znázorněny vrstevnice posunutého Dirichletova rozdělení s parametry $\boldsymbol{\alpha} = (2, 3, 5)'$ a $\boldsymbol{\beta} = (1, 1, 1)'$. Charakter parametrů nám poukazuje na to, že se jedná o Dirichletovo rozdělení. Použijeme-li operaci perturbace na toto rozdělení, získáme posunuté Dirichletovo rozdělení. Vrstevnice hustoty vzniklé perturbací kompozicí $\mathbf{p} = (0.93, 0.05, 0.02)'$ jsou vykresleny červenou barvou. V pravé části obrázku je ternární diagram, který hustoty vyjadřuje vzhledem k Aitchisonově míře λ_a v prostoru $\mathcal{S}^3$, vlevo jsou ty stejné vrstevnice vyobrazeny v prostoru souřadnic vzhledem k ortonormální bázi.

Z prvního obrázku je vidět, že použití operace perturbace na hustotu Dirichletova rozdělení ve skutečnosti představuje posun původní neperturbované hustoty na simplexu. Tato vlastnost plyne z invariance hustoty vzhledem k perturbaci.



Obrázek č. 14: Funkce hustoty posunutého Dirichletova rozdělení pro $D = 3$. Vlevo vzhledem k Lebesgueově míře v prostoru ilr souřadnic $\mathbb{R}^2$, vpravo vzhledem k Aitchisonově míře na simplexu. Pro parametry $\boldsymbol{\alpha} = (2, 3, 5)'$ a $\boldsymbol{\beta} = (1/0.93, 1/0.05, 1/0.02)'$ červené křivky, pro parametry $\boldsymbol{\alpha} = (2, 3, 5)'$ a $\boldsymbol{\beta} = (1, 1, 1)'$ modré křivky.

Ze znalosti této vlastnosti můžeme pomocí perturbace jednodušše určit i modus a střední hodnotu pro posunuté Dirichletovo rozdělení.

Věta 4.10. *Modus a střední hodnotu* $\widetilde{\mathbf{X}} \sim \mathcal{SD}^D(\boldsymbol{\alpha}, \boldsymbol{\beta})$ *vzhledem k míře* λ_a *vyjádříme jako*

$$\mathrm{modus}_a(\widetilde{\mathbf{X}}) = (\ominus\boldsymbol{\beta}) \oplus \mathrm{modus}_a(\mathbf{X}), \tag{5}$$

$$\mathrm{E}_a(\widetilde{\mathbf{X}}) = (\ominus\boldsymbol{\beta}) \oplus \mathrm{E}_a(\mathbf{X}), \tag{6}$$

kde $\mathbf{X} \sim \mathcal{D}^D(\boldsymbol{\alpha})$ *a* $\ominus$ *je inverzní operace k perturbaci.*

Pro kompozici $\widetilde{\mathbf{X}} \sim \mathcal{SD}^D(\boldsymbol{\alpha}, \boldsymbol{\beta})$ neexistuje žádné explicitní vyjádření pro modus($\widetilde{\mathbf{X}}$) a E($\widetilde{\mathbf{X}}$) vzhledem k Lebesgueově míře λ v reálném prostoru. Pokud bychom tyto charakteristiky chtěli spočítat, museli bychom použít numerickou integraci.

Z rovnic (5) a (6) plyne, že vektor parametrů $\boldsymbol{\beta}$ se podílí na umístění kompozice $\widetilde{\mathbf{X}}$ a nikoli na měřítku. Pokud je navíc parametr $\boldsymbol{\alpha}$ vektorem konstant, tj. $\boldsymbol{\alpha} = (\alpha, \ldots, \alpha)'$, tak modus i střední hodnota vzhledem k Aitchisonově míře splývá s neutrálním prvkem simplexu $\mathbf{n}$.

Věta 4.11. *Metrický rozptyl pro kompozici* $\widetilde{\mathbf{X}} \sim \mathcal{SD}^D(\boldsymbol{\alpha}, \boldsymbol{\beta})$ *se shoduje s metrickým rozptylem pro kompozici* $\mathbf{X} \sim \mathcal{D}^D(\boldsymbol{\alpha})$.

Metrické rozptyly jsou si rovny, jelikož platí

$$d_a(\mathbf{x}, \mathbf{y}) = d_a(\mathbf{p} \oplus \mathbf{x}, \mathbf{p} \oplus \mathbf{y}),$$

což znamená, že číselná charakteristika metrického rozptylu je invariantní vůči perturbaci.

4.4.3. Posunuté škálované Dirichletovo rozdělení

Tato kapitola se zabývá rozdělením náhodné kompozice, kterou obdržíme použitím operace perturbace a mocninné transformace na náhodnou kompozici s Dirichletovým rozdělením [12]. Jinak řečeno, budeme se zabývat tvarem hustoty pro kompozici $\widetilde{\mathbf{X}} = \mathbf{p} \oplus (a \odot \mathbf{X})$, kde $\mathbf{p} \in \mathcal{S}^D, a \in \mathbb{R}_+$ a $\mathbf{X} \sim \mathcal{D}^D(\boldsymbol{\alpha})$.

Definice 4.5. *Náhodný vektor* $\mathbf{X} \in \mathcal{S}^D$ *má posunuté škálované Dirichletovo rozdělení s parametry* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)' \in \mathbb{R}_+^D$, $\mathbf{p} = (p_1, \ldots, p_D)' \in \mathcal{S}^D$ *a* $a \in \mathbb{R}_+$, *jestliže jeho funkce hustoty pravděpodobnosti má tvar*

$$f_{ps}(\mathbf{x}) = \frac{dP}{d\lambda}(\mathbf{x}) = \frac{\Gamma(\alpha_+)}{\prod_{i=1}^D \Gamma(\alpha_i)} \frac{1}{a^{D-1}} \frac{\prod_{i=1}^D {p_i}^{-(\alpha_i/a)} {x_i}^{(\alpha_i/a)-1}}{\left(\sum_{i=1}^D (x_i/p_i)^{(1/a)}\right)^{\alpha_+}},$$

kde P *je Dirichletova pravděpodobnostní míra,* $\alpha_+ = \sum_{i=1}^D \alpha_i$, *a* Γ *je gamma funkce. Značíme* $\mathbf{X} \sim p\mathcal{SD}^D(\boldsymbol{\alpha}, \mathbf{p}, a)$.

Počet parametrů je $2D$. Pro $a = 1$ se jedná o model škálovaného Dirichletova rozdělení s parametry $\boldsymbol{\alpha}$ a $\ominus\mathbf{p}$. Jestliže $a = 1$ a vektor $\mathbf{p} = \mathcal{C}(1, \ldots, 1)'$ nebo $\mathbf{p} = \mathcal{C}(p, \ldots, p)'$ pro nějakou konstantu p, pak dostaneme klasický Dirichletův model, protože parametry odpovídají neutrálním prvkům vzhledem k provedeným operacím.

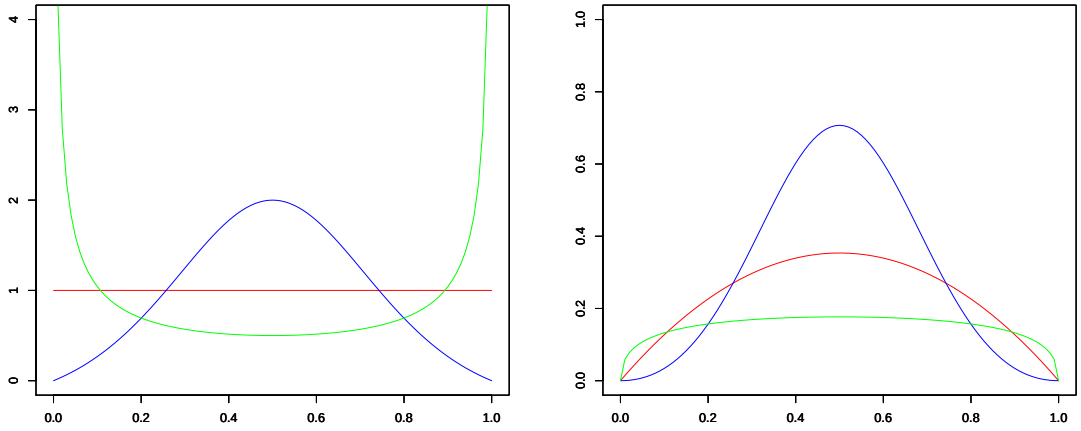
Stejně jako v předchozích případech můžeme vyjádřit funkci hustoty vzhledem k Aitchisonově míře λ_a,

$$f_{ps}(\mathbf{x}) = \frac{dP_{ps}}{d\lambda}(\mathbf{x}) = \frac{\sqrt{D}\Gamma(\alpha_+)}{\prod_{i=1}^D \Gamma(\alpha_i)} \frac{1}{a^{D-1}} \frac{\prod_{i=1}^D (x_i/p_i)^{(\alpha_i/a)}}{\left(\sum_{i=1}^D (x_i/p_i)^{(1/a)}\right)^{\alpha_+}}.$$

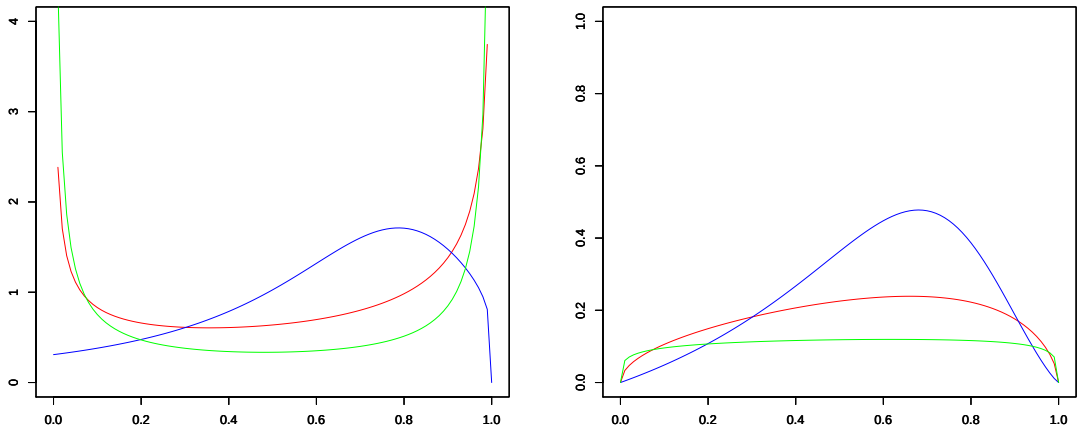
Na obrázcích 15 a 16 jsou vykresleny hustoty posunutého škálovaného Dirichletova rozdělení pro $D = 2$ vzhledem k Lebesguově míře λ na intervalu $\langle 0, 1\rangle$ a Aitchisonově míře λ_a na $\mathcal{S}^2$. Stejně jako u posunutého Dirichletova rozdělení platí, že při práci s hustotou vzhledem k Aitchisonově míře λ_a je tato funkce vždy unimodální. V tomto případě hraje roli parametru měřítka parametr a, který určuje jak bude rozdělení koncentrováno kolem střední hodnoty, tj. čím větší bude parametr a, tím více bude rozdělení koncentrováno kolem očekávané hodnoty. Podíváme-li se ovšem, jak se chová funkce hustoty vzhledem k míře Lebesgueově, vidíme, že při vysokých hodnotách parametru a má funkce vertikální asymptoty v bodech 0 a 1.

Na obrázku 15 jsou vyobrazeny hustoty vzhledem k mírám λ a λ_a při parametrech $\boldsymbol{\alpha} = (1, 1)'$ a $\mathbf{p} = (1, 1)'$. Z levého obrázku je patrné, že pro $a \in (0, 1)$

je funkce hustoty unimodální a pro $a = 1$ je konstatní, zatímco pro $a > 1$ začíná mít funkce konkávní tvar. Při pohledu na pravý obrázek je jasné, že funkce hustoty vzhledem k Aitchisonově míře se s různou volbou parametru a chová úplně opačně. Na obrázku 16 lze pozorovat analogické vlastnosti v případě různé volby parametrů $\boldsymbol{\alpha}$ a $\mathbf{p}$.



Obrázek č. 15: Funkce hustoty posunutého škálovaného rozdělení pro $D = 2$ s parametry $\boldsymbol{\alpha} = (1, 1)'$ a $\mathbf{p} = (1, 1)'$. Na levém obrázku vzhledem k Lebesgueově míře λ, na pravém obrázku vzhledem k Aitchisonově míře λ_a. Pro $a = 0.5$ modrá křivka, $a = 1$ červená křivka a $a = 2$ zelená křivka.



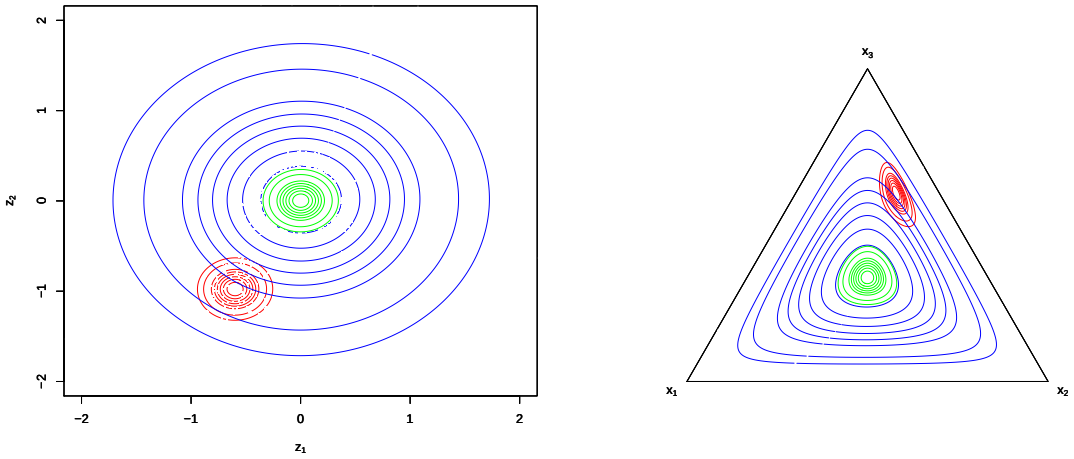
Obrázek č. 16: Funkce hustoty posunutého škálovaného rozdělení pro $D = 2$ s parametry $\boldsymbol{\alpha} = (0.5, 0.6)'$ a $\mathbf{p} = (0.7, 0.3)'$. Na levém obrázku vzhledem k Lebesgueově míře λ, na pravém obrázku vzhledem k Aitchisonově míře λ_a. Pro $a = 0.5$ modrá křivka, $a = 1$ červená křivka a $a = 2$ zelená křivka.

Podobné chování hustoty posunutého škálovaného Dirichletova rozdělení můžeme sledovat i pro $D = 3$. Na obrázku 17 jsou vykresleny vrstevnice hustot pro tři různá nastavení parametrů. Na pravém obrázku je znázorněn ternární diagram a hustota vzhledem k Aitchisonově míře λ_a, zatímco na levém obrázku máme hustotu vzhledem k Lebesgueově míře, vykreslenou v prostoru ilr souřadnic, tj. v $\mathbb{R}^2$. Modré křivky odpovídají posunutému škálovanému Dirichletovu rozdělení s parametry $\boldsymbol{\alpha} = (2, 2, 2)'$, $\mathbf{p} = \mathcal{C}(1, 1, 1)'$ a $a = 1$. Hustota s takto nastavenými parametry ve skutečnosti představuje hustotu Dirichletova rozdělení, které má střední hodnotu i modus přímo ve středu (težišti) ternárního diagramu.

Zelené křivky představují vrstevnice hustoty se stejnými parametry $\boldsymbol{\alpha}$ a $\mathbf{p}$ jako v předchozím případě, parametr a byl ale zmenšen na $a = 0.2$. Jinak řečeno, na původní náhodnou kompozici s Dirichletovým rozdělením jsme použili operaci mocninné transformace. Z obrázku 17 je zřejmé, že parametr a udává míru disperze, tj. koncentraci hodnot kolem střední hodnoty.

Poslední hustota, vykreslená červenou barvou, představuje posunuté škálo-

vané Dirichletovo rozdělení s parametry $\boldsymbol{\alpha} = (2, 2, 2)'$, $\mathbf{p} = \mathcal{C}(0.15, 0.05, 0.02)'$ a $a = 0.2$. Jelikož hodnota parametru a je stejná, míra disperze se od předchozího případu nemění. Hustotu označenou červenou barvou jsme obdrželi použitím operace perturbace na hustotu barvy zelené. Podíváme-li se na levý obrázek, je patrné, že v prostoru souřadnic je červená hustota opravdu pouze posunutím hustoty zelené.



Obrázek č. 17: Hustota posunutého škálovaného Dirichletova rozdělení pro $D = 3$ s parametrem $\boldsymbol{\alpha} = (2, 2, 2)'$. Vlevo vzhledem k Lebesgueově míře v prostoru ilr souřadnic $\mathbb{R}^2$, vpravo vzhledem k Aitchisonově míře na simplexu.

Věta 4.12. *Modus a střední hodnotu* $\widetilde{\mathbf{X}} \sim p\mathcal{SD}^D(\boldsymbol{\alpha}, \mathbf{p}, a)$ *vzhledem k míře* λ_a *vyjádříme jako*

$$\mathrm{modus}_a(\widetilde{\mathbf{X}}) = \mathbf{p} \oplus (a \odot \mathrm{modus}_a(\mathbf{X})),$$

$$\mathrm{E}_a(\widetilde{\mathbf{X}}) = \mathbf{p} \oplus (a \odot \mathrm{E}_a(\mathbf{X})),$$

kde $\mathbf{X} \sim \mathcal{D}^D(\boldsymbol{\alpha})$.

Ke zjištění modu a střední hodnoty kompozice $\widetilde{\mathbf{X}} \sim p\mathcal{SD}^D(\boldsymbol{\alpha}, \mathbf{p}, a)$ vzhledem k Lebesgueově míře λ v prostoru o $D - 1$ složkách kompozice je opět zapotřebí použít numerickou integraci.

Věta 4.13. *Pro metrický rozptyl náhodné kompozice* $\widetilde{\mathbf{X}} \sim p\mathcal{SD}^D(\boldsymbol{\alpha}, \mathbf{p}, a)$ *platí*

$$\mathrm{Mvar}(a \odot (\mathbf{p} \oplus \mathbf{X})) = \mathrm{Mvar}(\widetilde{\mathbf{X}}) = a^2 \mathrm{Mvar}(\mathbf{X}),$$

kde $\mathbf{X} \sim \mathcal{D}^D(\boldsymbol{\alpha})$.

Z uvedené definice střední hodnoty (vlastně centra) a modu posunutého škálovaného Dirichletova rozdělení je patrné, že parametr $\mathbf{p}$ určuje posunutí rozdělení, tudíž se jedná o parametr polohy. Druhý parametr a určuje koncentraci rozdělení kolem střední hodnoty. Tento parametr je tedy parametrem měřítka. Čím je parametr a menší, tím je rozdělení více koncentrováno kolem svého centra.

Propojenost Dirichletova rozdělení, posunutého Dirichletova rozdělení a posunutého škálovaného Dirichletova rozdělení je dána konkrétní volbou hodnot parametrů. Pokud v modelu posunutého škálovaného Dirichletova rozdělení nastavíme parametr $a = 1$, dostaneme model posunutého Dirichletova rozdělení. V případě, že nastavíme parametry $a = 1$ a $\mathbf{p} = \mathcal{C}(1, 1, 1)'$, obdržíme klasický model Dirichletova rozdělení. Z této vlastnosti můžeme usuzovat, že všechny tři modely rozdělení pravděpodobnosti pocházejí ze stejné rodiny rozdělení.

5. Závěr

V současné době existuje řada statistických přístupů ke zpracování kompozičních dat. Tato data se od ostatních liší tím, že nesou pouze relativní informaci, přísluší danému celku a bez jeho znalosti ztrácejí svůj význam. Z tohoto důvodu použité pravděpodobnostní a statistické techniky vykazují určité charakteristické rysy, přičemž důraz je kladen zejména na odpovídající geometrickou strukturu výběrového prostoru a na něm zavedenou pravděpodobnostní míru.

Jedním z přístupů k parametrickému modelování kompozičních dat je vedle normálního rozdělení na simplexu využití Dirichletova rozdělení. Toto rozdělení je specifické svým výběrovým prostorem, kterým je simplex. Z tohoto důvodu se jeví jako vhodný nástroj pro práci s kompozičními daty. Ukazuje se však, že konstrukce Dirichletovo rozdělení vyžaduje splnění podmínek, které se v praxi jeví jako obtížně dosažitelné. Jedním z cílů diplomové práce tak bylo studovat alternativní vyjádření Dirichletova rozdělení vzhledem k Aitchisonově míře na simplexu a zkoumat jeho možná zobecnění.

Při psaní této práce jsem získala nové znalosti z oblasti kompozičních dat a prohloubila své vědomosti z mnohorozměrné statistické analýzy a teorie míry. Nejtěžší pro mě bylo seznámit se s novou problematikou a vyrovnat se se zaváděním české terminologie.

Doufám, že se mi podařilo vytvořit ucelený a srozumitelný pohled na rozdělení pravděpodobnosti na simplexu, který by byl přínosný i pro další zájemce o danou problematiku.

Příloha

Transformace náhodného vektoru

Uvažujme integrál

$$\int \dots \int_{\mathcal{A}} h(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_n$$

na množině $\mathcal{A}$, která je podmnožinou n-rozměrného prostoru $\mathcal{S}$. Dále nechť

$$y_1 = u_1(x_1, x_2, \dots, x_n),$$

$$y_2 = u_2(x_1, x_2, \dots, x_n),$$

$$\vdots$$

$$y_n = u_n(x_1, x_2, \dots, x_n),$$

a jejich inverzní funkce

$$x_1 = w_1(y_1, y_2, \dots, y_n),$$

$$x_2 = w_2(y_1, y_2, \dots, y_n),$$

$$\vdots$$

$$x_n = w_n(y_1, y_2, \dots, y_n)$$

představují prosté zobrazení, které zobrazuje množinu $\mathcal{A}$ z $\mathcal{S}$ do množiny $\mathcal{B}$ z $\mathcal{T}$, tj. do prostoru souřadnic $(y_1, y_2, \dots, y_n)$. Nechť jsou první parciální derivace inverzních funkcí spojité a nechť je Jakobián

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \dots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \dots & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \dots & \frac{\partial x_n}{\partial y_n} \end{vmatrix}$$

nenulový v prostoru $\mathcal{T}$. Pak

$$\int \dots \int_{\mathcal{A}} h(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_n$$

$$= \int \dots \int_{\mathcal{B}} h[w_1(y_1, \dots, y_n), w_2(y_1, \dots, y_n), \dots, w_n(y_1, \dots, y_n)]|J| dy_1 dy_2 \cdots y_n.$$

Jsou-li splněny všechny výše uvedené podmínky, můžeme určit sdruženou hustotu pravděpodobnosti pro n funkcí n náhodných veličin. Pak tedy sdružená hustota náhodných veličin $Y_1 = u_1(X_1, \dots, X_n), \dots, Y_n = u_n(X_1, \dots, X_n)$, kde veličiny $X_1, \dots, X_n$ mají hustotu $h(x_1, \dots, x_n)$, je dána ve tvaru

$$g(y_1, y_2, \dots, y_n) = |J| \cdot h[w_1(y_1, \dots, y_n), w_2(y_1, \dots, y_n), \dots, w_n(y_1, \dots, y_n)],$$

kde $(y_1, y_2, \dots, y_n) \in \mathcal{T}$, a v ostatních případech je nulová.

Literatura

[1] Aitchison, J.: *The statistical analysis of compositional data*, London: Chapman and Hall, 1986.

[2] Aitchison J.: *The one-hour course in compositional data analysis or compositional data analysis in simple*, In: Pawlowsky-Glahn, V.: The third annual conference of the International Association for Mathematical Geology – IAMG'97, Proceedings, International Center for Numerical Methods in Engineering (CIMNE), Barcelona, 1997.

[3] Anděl, J., *Základy matematické statistiky*, 3. vydání. Praha: Matfyzpress, 2011.

[4] Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., *Isometric logratio transformations for compositional data analysis*, Mathematical Geology, 35, 3, 279-300, 2003.

[5] Filzmoser, P., Hron, K., *Outlier detection for compositional data using robust methods*, Mathematical Geosciences, 40, 3, 233-248, 2007.

[6] Hogg, R. V., McKean, J. W., Craig, A. T., *Introduction to Mathematical Statistics*, 6. vydání. New Jersey: Pearson Education, Inc., 2005.

[7] Martín-Fernández, J. A., Barceló-Vidal, C., Pawlowsky-Glahn, V., *Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation*, Mathematical Geology, 35, 3, 253-278, 2003.

[8] Mateu-Figueras, G., Pawlowsky-Glahn, V., *A critical approach to probability laws in geochemistry*, Mathematical Geosciences, 40, 5, 489-502, 2008.

[9] Mateu-Figueras, G., Pawlowsky-Glahn, V., Barceló-Vidal, C., *Distributions on the simplex*, In: Thió-Henestrosa, S., Martín-Fernández, J.A.: Compositional Data Analysis Workshop – CoDaWork'03, Proceedings, Universitat de Girona, 2003.

[10] Mateu-Figueras, G., Pawlowsky-Glahn, V., *The Dirichlet distribution with respect to the Aitchison measure on the simplex - a first approach*, In: Mateu-Figueras, G., Barceló-Vidal, C.: Compositional Data Analysis Workshop – CoDaWork'05, Proceedings, Universitat de Girona, 2005.

[11] Monti, G.S., Mateu-Figueras, G., Pawlowsky-Glahn, V., *Notes on the scaled Dirichlet distribution*, In: Pawlowsky-Glahn, V., Buccianti, A.: Compositional Data Analysis: Theory and Applications, Chichester: John Wiley & Sons, 128-138, 2011.

[12] Monti, G.S.,Mateu-Figueras, G.,Pawlowsky-Glahn, V., Egozcue, J.J., *The shifted-scaled Dirichlet distribution in the simplex*, In: Egozcue, J.J., Tolosana-Delgado, R., Ortego, M.I.: Compositional Data Analysis Workshop – CoDaWork'11, Proceedings, International Center for Numerical Methods in Engineering (CIMNE), Barcelona, 2011.

[13] Pawlowsky-Glahn, V., Egozcue, J.J., *Geometric approach to statistical analysis on the simplex*, Stochastic Enviromental Research and Risk Assessment, 15, 384-398, 2001.

[14] Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., *Lecture notes on compositional data analysis* [online]. Girona: Universitat de Girona, 28.5.2007 [citováno 11.3.2012]. Dostupné z WWW: <http://hdl.handle.net/10256/297>.

[15] Pawlowsky-Glahn, V., *Statistical modelling on coordinates*, In: Thió-Henestrosa, S., Martín-Fernández, J.A.: Compositional Data Analysis Workshop – CoDaWork'03, Proceedings, Universitat de Girona, 2003.