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

DIPLOMOVÁ PRÁCE

Testy normality pro kompoziční data

Vedoucí diplomové práce:
RNDr. Karel Hron, Ph.D.
Rok odevzdání: 2010

Vypracoval:
Bc. Ondřej Malý
AME, II. ročník

Prohlášení

Prohlašuji, že jsem vytvořil tuto diplomovou práci samostatně za vedení RNDr. Karla Hrona, Ph.D. a že jsem v seznamu použité literatury uvedl všechny zdroje použité při zpracování práce.

V Olomouci dne 15. listopadu 2010.

Poděkování

Rád bych na tomto místě poděkoval vedoucímu diplomové práce RNDr. Karlu Hronovi, Ph.D. za obětavou spolupráci, trpělivost, rady, inspiraci, věcné diskuze i za čas, který mi věnoval při konzultacích.

Rovněž patří poděkování mé rodině za podporu při studiu a tvorbu potřebného zázemí.

Obsah

Úvod

Normální rozdělení má zásadní význam v teorii pravděpodobnosti a matematické statistice. Normálním rozdělením se řídí mnoho náhodných veličin a za určitých podmínek aproximuje jiná spojitá i diskrétní rozdělení. V mnohorozměrné statistické analýze hraje normální rozdělení nezastupitelnou roli.

Jen vyjímečně lze na základě zkušeností prohlásit, že náhodná veličina je normálně rozdělená, u naprosté většiny náhodných veličin je nutné normalitu ověřit. Ověření předpokladu normálního rozdělení implikuje možnost využití konkrétní teorie statistické analýzy včetně tvorby predikcí. Vhodným nástrojem k ověření normálního rozdělení jsou statistické testy.

Standardně testujeme normalitu náhodného výběru, jehož výběrovým prostorem je konečně-dimenzionální euklidovský prostor. Kompoziční data jsou speciálním typem mnohorozměrných dat se speciálním výběrovým prostorem, jejich zpracování standardními statistickými metodami může vést k zavádějícím nebo zcela chybným závěrům. Proto kompoziční data svou přirozenou povahou vyžadují zvláštní péči v podobě vhodného výběrového prostoru, matematických operací a interpretace. Teorie analýzy kompozičních dat je budována od roku 1986, jejím zakladatelem je John Aitchison.

Cílem této diplomové práce je popsat dostupné přístupy k testování normality kompozičních dat včetně jejich porovnání a demonstrace na konkrétních datech. Vedlejším cílem této práce je nalezení odpovědí na otázky, proč je vhodné zpracovávat kompoziční data odlišně od jiných statistických dat, jak taková data zpracovat a jaký má význam testování normality.

Práce je členěna do čtyř stěžejních kapitol.

V první kapitole je popsáno normální rozdělení v jeho jednorozměrné i mnohorozměrné podobě spolu s jeho vlastnostmi. K normálnímu rozdělení je přistupováno prostřednictvím charakteristické funkce.

Druhá kapitola vymezuje pojem kompoziční data jako pozorování nesoucí pouze relativní informaci včetně jejich transformací ze simplexu do euklidovského prostoru a softwarového řešení pro kompoziční data prostřednictvím knihovny

`robCompositions` softwaru R.

Třetí kapitola se věnuje normálnímu rozdělení na simplexu.

Čtvrtá kapitola analyzuje tři přístupy k testování normality kompozičních dat. Všechny tyto přístupy jsou demonstrovány na konkrétních příkladech spolu s průběžným porovnáním a interpretací výsledků testování. Poznamenejme již nyní, že k testování normálního rozdělení na simplexu jsou z hlediska dimenze dat užity marginální univariátní testy, bivariátní testy a radius test a z hlediska typu testového kritéria Anderson-Darlingův test, Cramer-von Misesův test a Watsonův test.

Práce je vysázena typografickým softwarem TEXLive, verze 2009, praktické části jsou zpracovány statistickým softwarem R.

1 Normální rozdělení

Normální rozdělení je v matematické statistice jedním z nejdůležitějších a nejužívanějších rozdělení pravděpodobností náhodné veličiny X. Důležitost normálního rozdělení je daná z věcného hlediska samotnou existencí nepřeberného množství náhodných veličin, které jsou svou povahou normálně rozdělené. Význam normálního rozdělení navíc podtrhuje fakt, že většina metod mnohorozměrné statistické analýzy závisí na předpokladu mnohorozměrného normálního rozdělení.

Normálnímu rozdělení se také říká gaussovské rozdělení dle autora Carla Friedricha Gausse a v aplikacích se můžeme setkat s termínem „rozdělení chyb“. Z normálního rozdělení jsou odvozena další užitečná rozdělení jako χ^2, Studentovo a Fisherovo rozdělení. Také součet velkého počtu nezávislých náhodných veličin je za dosti obecných předpokladů normálně rozdělen. Zřejmě lze trochu s nadsázkou tvrdit, že bez splnění předpokladu normality se v teorii pravděpodobnosti a matematické statistice nehneme z místa.

V této kapitole si shrneme základní poznatky o normálním rozdělení, a to v jeho jednorozměrné i mnohorozměrné podobě. K normálnímu rozdělení budeme přistupovat pomocí charakteristické funkce.

1.1 Charakteristická funkce

Rozdělení pravděpodobností náhodných veličin se obvykle popisuje pomocí distribuční funkce v obecném případě nebo pomocí pravděpodobnostní funkce u diskrétních náhodných veličin a hustoty u spojitých náhodných veličin. V absolutně spojitém případě je znalost hustoty ekvivalentní znalosti distribuční funkce. Je zde ovšem i jiná možnost výhodného popisu rozdělení pravděpodobností, a to prostřednictvím charakteristické funkce [14]. Charakteristická funkce existuje pro všechny náhodné veličiny.

Definice 1.1. Charakteristickou funkci ψ náhodné veličiny X definujeme jako střední hodnotu náhodné veličiny e^{itX}, tj.

$$\psi(t) = \mathsf{E}e^{itX} = \mathsf{E}\cos tX + i\mathsf{E}\sin tX,$$

kde t je reálná proměnná a i je komplexní jednotka.

V případě více náhodných veličin použijeme pro charakteristickou funkci náhodné veličiny X indexaci $\psi_X(t)$. Vlastnosti charakteristické funkce vycházejí z teorie pravděpodobnosti a komplexní analýzy. Uvedeme několik zásadních vlastností charakteristické funkce.

Pokud má náhodná veličina X distribuční funkci F, potom přímo z definice charakteristické funkce plyne vztah mezi charakteristickou funkcí a distribuční funkcí, tj.

$$\psi(t) = \int e^{itx} dF(x).$$

Jestliže distribuční funkce náhodné veličiny X je absolutně spojitá a má hustotu f, potom

$$\psi(t) = \int e^{itx} f(x) dx.$$

Věta 1.1. *Platí, že* $|\psi(t)| \leq 1$, *pro* $t = 0$ *je* $\psi(0) = 1$.

Důkaz: Protože platí nerovnosti $|e^{itX}| \leq 1$ a $|\mathsf{E}(X)| \leq \mathsf{E}(|X|)$, potom

$$|\psi(t)| \leq \mathsf{E}(|e^{itX}|) \leq 1.$$

□

Věta 1.2. $\psi(t)$ *je stejnoměrně spojitá na intervalu* $(-\infty, \infty)$.

Důkaz: Viz [14], s. 264.

□

Věta 1.3. *Jestliže existují konečné obecné momenty* $\mu_1', \ldots, \mu_n'$ *náhodné veličiny* X , *pak charakteristická funkce* ψ *má prvních* n *derivací a platí*

$$\psi^{(k)}(0) = i^k \mu_k', \ k = 1, \ldots, n,$$

a dále platí

$$\psi(t) = \sum_{k=0}^{n} \mu_k' \frac{(it)^k}{k!} + o(t^n),$$

kde $o(t^n)$ *je taková funkce, že* $\lim_{t \to 0} \frac{o(t^n)}{t^n} = 0$.

Důkaz: Viz [14], s. 266 - 267.

□

Podstata charakteristické funkce tkví v popisu rozdělení náhodných veličin. Tento poznatek si shrneme v následující větě.

Věta 1.4. *Nechť ψ je charakteristická funkce odpovídající distribuční funkci F a nechť a, b $(a < b)$ jsou body spojitosti funkce F. Pak platí vztah*

$$F(b) - F(a) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\psi(t) \frac{e^{-ita} - e^{-itb}}{2it} - \psi(-t) \frac{e^{ita} - e^{itb}}{2it} \right] dt.$$

Charakteristická funkce jednoznačně určuje distribuční funkci.

Důkaz: Viz [14], s. 272.

□

Věta 1.5. *Nechť X je náhodná veličina, $\psi_X(t)$ je její charakteristická funkce, $a, b \in \mathbb{R}$. Potom náhodná veličina $Y = a + bX$ má charakteristickou funkci*

$$\psi_Y(t) = e^{ita} \psi_X(tb).$$

Důkaz: Z vlastností střední hodnoty plyne

$$\psi_Y(t) = \mathsf{E} e^{itY} = \mathsf{E} e^{ita + itbX} = e^{ita} \mathsf{E} e^{itbX} = e^{ita} \psi_X(tb).$$

□

Věta 1.6. *Nechť X_1 a X_2 jsou nezávislé náhodné veličiny s charakteristickými funkcemi $\psi_1(t)$ a $\psi_2(t)$. Potom náhodná veličina $X = X_1 + X_2$ má charakteristickou funkci*

$$\psi_X(t) = \psi_1(t) \psi_2(t).$$

Důkaz: Za předpokladu nezávislosti je střední hodnota součinu náhodných veličin rovna součinu středních hodnot, a tedy

$$\psi_X(t) = \mathsf{E} e^{it(X_1 + X_2)} = \mathsf{E} \left(e^{itX_1} e^{itX_2} \right) = \mathsf{E} \left(e^{itX_1} \right) \mathsf{E} \left(e^{itX_2} \right) = \psi_1(t) \psi_2(t).$$

□

Definice 1.2. Nechť $\mathbf{x} = (X_1, \ldots, X_p)'$ je p-rozměrný náhodný vektor. Charakteristickou funkci ψ náhodného vektoru $\mathbf{x}$ definujeme vztahem

$$\psi(\mathbf{t}) = \psi(t_1, \ldots, t_p) = \mathsf{E}e^{i\mathbf{t}'\mathbf{x}} = \mathsf{E}e^{i\sum_{j=1}^{p} t_j X_j}.$$

Vlastnosti charakteristické funkce náhodného vektoru [14], kapitola 6.6, jsou v analogii s jednorozměrným případem. Tedy

- $|\psi(\mathbf{t})| \leq 1$ pro každý vektor $\mathbf{t} \in \mathbb{R}^p$,
- $\psi(\mathbf{0}) = 1$,
- ψ je stejnoměrně spojitá na $\mathbb{R}^p$.
- Nechť $\mathbf{b} \in \mathbb{R}^r$, $\mathbf{A} \in \mathbb{R}^{r,p}$, $\mathbf{y} = \mathbf{b} + \mathbf{A}\mathbf{x}$. Potom

$$\psi_{\mathbf{y}}(\mathbf{t}) = e^{i\mathbf{t}'\mathbf{b}}\psi_{\mathbf{x}}(\mathbf{A}'\mathbf{t}),\ \mathbf{t} \in \mathbb{R}^r.$$

- Existují-li střední hodnoty $\mathsf{E}(X_j)$, $j = 1, \ldots, p$, pak

$$\left(\frac{\partial\psi(\mathbf{t})}{\partial t_j}\right)_{\mathbf{t}=(0,\ldots,0)} = i\mathsf{E}(X_j).$$

- Existují-li střední hodnoty $\mathsf{E}(X_j X_k)$, $j, k = 1, \ldots, p$, pak

$$\left(\frac{\partial^2\psi(\mathbf{t})}{\partial t_j \partial t_k}\right)_{\mathbf{t}=(0,\ldots,0)} = -\mathsf{E}(X_j X_k).$$

- Nechť $\psi_j(t)$ je charakteristická funkce náhodné veličiny X_j. Pak

$$\psi_j(t_j) = \psi_{\mathbf{x}}(0, 0, \ldots, t_j, \ldots, 0, 0).$$

- Složky náhodného vektoru $\mathbf{x} = (X_1, \ldots, X_p)'$ jsou nezávislé právě tehdy, když

$$\psi_{\mathbf{x}}(\mathbf{t}) = \prod_{j=1}^{p} \psi_j(t_j).$$

- Nechť $\mathbf{x}$ a $\mathbf{y}$ jsou nezávislé náhodné vektory s charakteristickými funkcemi $\psi_{\mathbf{x}}(t_1,\ldots,t_p)$ a $\psi_{\mathbf{y}}(t_1,\ldots,t_p)$. Potom $\mathbf{z}=\mathbf{x}+\mathbf{y}$ má charakteristickou funkci

$$\psi_{\mathbf{z}}(\mathbf{t})=\psi_{\mathbf{x}}(\mathbf{t})\psi_{\mathbf{y}}(\mathbf{t}).$$

1.2 Jednorozměrné normální rozdělení

Jednorozměrné normální rozdělení [11] definujeme pomocí hustoty.

Definice 1.3. Řekneme, že náhodná veličina X má normální rozdělení s parametry $\mu\in\mathbb{R}$ a $\sigma^2>0$, píšeme $X\sim N(\mu,\sigma^2)$, jestliže její hustota je

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

Charakteristická funkce rozdělení $N(\mu,\sigma^2)$ je dána vztahem

$$\psi(t)=e^{\mu it}e^{-\frac{t^2\sigma^2}{2}}.$$

Tento vztah ověříme výpočtem.

Příklad: Nechť $X\sim N(\mu,\sigma^2)$. Charakteristická funkce této náhodné veličiny je

$$\psi(t)=\mathsf{E}e^{itX}=\int_{-\infty}^{\infty}e^{itx}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}e^{itx-\frac{(x-\mu)^2}{2\sigma^2}}dx=$$

$$=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}e^{uit\sigma+\mu it-\frac{u^2}{2}}\sigma du=\frac{1}{\sqrt{2\pi}}e^{\mu it}\int_{-\infty}^{\infty}e^{uit\sigma-\frac{u^2}{2}}du=$$

$$=\frac{1}{\sqrt{2\pi}}e^{\mu it}\int_{-\infty}^{\infty}e^{-\frac{1}{2}[(u-it\sigma)^2+t^2\sigma^2]}du=\frac{1}{\sqrt{2\pi}}e^{\mu it}\int_{-\infty}^{\infty}e^{-\frac{1}{2}y^2-\frac{1}{2}t^2\sigma^2}dy=$$

$$=\frac{1}{\sqrt{2\pi}}e^{\mu it}e^{-\frac{t^2\sigma^2}{2}}\int_{-\infty}^{\infty}e^{-\frac{1}{2}y^2}dy=e^{\mu it}e^{-\frac{t^2\sigma^2}{2}}.$$

Použili jsme nejprve substituci $\frac{x-\mu}{\sigma}=u$, $dx=\sigma du$, $x=\sigma u+\mu$, dále substituci $y=u-it\sigma$, $dy=du$ a z matematické analýzy víme, že $\int_{-\infty}^{\infty}e^{-\frac{1}{2}y^2}dy=\sqrt{2\pi}$.

K výpočtu momentů jednorozměrného normálního rozdělení lze kromě hustoty užít charakteristickou funkci tak, jak to říká Věta 1.3. První obecný moment, střední hodnotu, dostaneme snadným výpočtem, tj.

$$\mathsf{E}(X) = \mu_1^{'} = \frac{\psi^{(1)}(0)}{i} = \mu.$$

Výpočtem druhé derivace funkce ψ v bodě 0 dostaneme po úpravě druhý obecný moment, tj.

$$\mathsf{E}(X^2) = \mu_2^{'} = \frac{-\mu^2 - \sigma^2}{i^2} = \mu^2 + \sigma^2.$$

Zřejmě druhý centrální moment, rozptyl, bude

$$\operatorname{var} X = \mathsf{E}[X - \mathsf{E}(X)]^2 = \mathsf{E}(X^2) - [\mathsf{E}(X)]^2 = \mu^2 + \sigma^2 - \mu^2 = \sigma^2.$$

Obecně je k-tá derivace funkce ψ

$$\psi^{(k)}(t) = \frac{d^k\psi}{dt^k} = \int_{-\infty}^{\infty} \frac{d^k}{dt^k} e^{itx} f(x)dx = \int_{-\infty}^{\infty} (ix)^k e^{itx} f(x)dx.$$

Položíme-li $t = 0$, pak pro k-tý obecný moment dostáváme známý vztah

$$\mathsf{E}(X^k) = \mu_k^{'} = \int_{-\infty}^{\infty} x^k f(x)dx.$$

Víme, že pokud existuje k-tý obecný moment, pak také existuje k-tý centrální moment [11], s. 54. V případě normálního rozdělení platí [3], s. 23, že

$$\mathsf{E}[X - \mathsf{E}(X)]^{2k-1} = 0,\ k = 1, 2, \ldots,$$

$$\mathsf{E}[X - \mathsf{E}(X)]^{2k} = \frac{(2k)!}{k!}\frac{\sigma^{2k}}{2^k},\ k = 1, 2, \ldots.$$

Snadné je už dopočítat charakteristiky šikmosti a špičatosti. Koeficient šikmosti bude

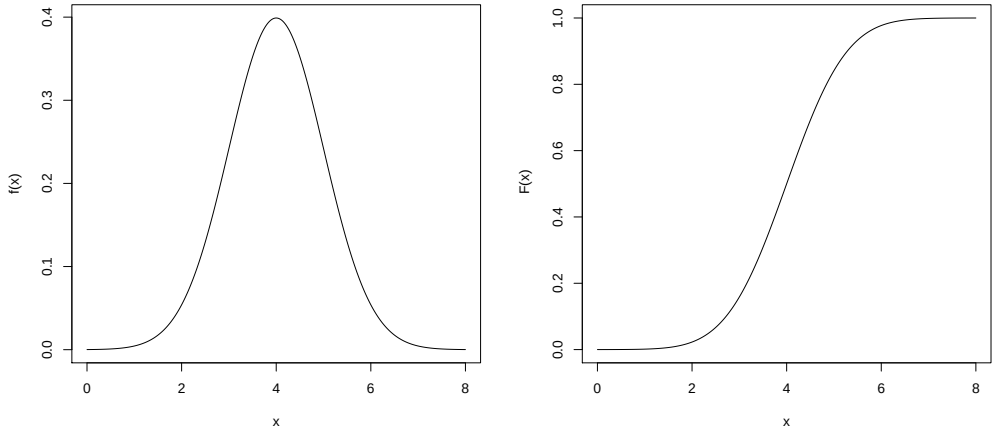
$$\alpha_3(X) = \frac{\mathsf{E}[(X - \mathsf{E}(X))^3]}{(\sqrt{\operatorname{var}(X)})^3} = 0$$

a koeficient špičatosti

$$\alpha_4(X) = \frac{\mathsf{E}[(X - \mathsf{E}(X))^4]}{(\sqrt{\mathrm{var}(X)})^4} = 3.$$

Distribuční funkce normálně rozdělené náhodné veličiny X je integrálem hustoty, tj.

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{(u-\mu)^2}{2\sigma^2}\right\} du, \ x \in \mathbb{R}.$$



Obr.: Hustota rozdělení $N(4,1)$ vlevo, distribuční funkce rozdělení $N(4,1)$ vpravo.

Hustotě rozdělení $N(\mu, \sigma^2)$ se někdy říká Gaussova křivka, má typicky zvonovitý tvar, je symetrická kolem střední hodnoty a v μ nabývá křivka svého maxima. Body $\mu - \sigma$ a $\mu + \sigma$ jsou body inflexe, μ určuje polohu křivky ve směru osy x a σ její tvar.

Hodnoty distribuční funkce $N(\mu, \sigma^2)$ hledáme pomocí statistického softwaru nebo v tabulkách s využitím normovaného normálního rozdělení, viz níže.

Nechť $Y \sim N(\mu, \sigma^2)$, potom náhodná veličina

$$X = \frac{Y - \mu}{\sigma} \ \sim N(0,1).$$

Rozdělení $N(0,1)$ nazýváme *normované normální rozdělení*, je speciálním případem normálního rozdělení s hustotou

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\},\ x \in \mathbb{R},$$

a distribuční funkcí

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{u^2}{2}\right\} du,\ x \in \mathbb{R}.$$

Substitucí $X = \frac{Y-\mu}{\sigma}$ lze distribuční funkci náhodné veličiny Y vyjádřit jako

$$F(y) = P(Y \leq y) = P\left(X \leq \frac{y-\mu}{\sigma}\right) = \Phi\left(\frac{y-\mu}{\sigma}\right).$$

Hustota rozdělení $N(0,1)$ je funkce sudá, a to výpočet hodnoty distribuční funkce v bodě ještě zjednoduší, protože platí

$$\Phi(x) = 1 - \Phi(-x),\ x > 0.$$

1.3 Mnohorozměrné normální rozdělení

Mnohorozměrné normální rozdělení [3] je esenciální složkou mnohorozměrné statistické analýzy.

Nechť $\mathbf{x} = (X_1, \ldots, X_p)'$ je p-rozměrný náhodný vektor, $p \geq 2$, jehož složky X_j jsou náhodné veličiny.

Definice 1.4. Řekneme, že náhodný vektor $\mathbf{x}$ má p-rozměrné normální rozdělení, jestliže pro každý vektor $\mathbf{c}_{(p\times 1)}$ platí, že lineární transformace $\mathbf{c}'\mathbf{x}$ má jednorozměrné normální rozdělení.

Definice převádí mnohorozměrnou problematiku zpět na jednorozměrnou, a to nám umožňuje odvolávat se na vše, co platí pro jednorozměrné normální rozdělení.

Rozdělení náhodného vektoru $\mathbf{x}_{(p\times 1)}$ je plně určeno rozdělením lineárních funkcí typu $\mathbf{c}'\mathbf{x}$, $\mathbf{c} \in \mathbb{R}^p$. To je poznatek, který ospravedlňuje definici. Známe-li totiž rozdělení každé lineární kombinace, pak známe i rozdělení náhodného

vektoru. Samozřejmě lze definovat mnohorozměrné normální rozdělení i pomocí hustoty, ale je to nepraktické řešení z hlediska dokazování matematických tvrzení.

Hustota mnohorozměrného normálního rozdělení má tvar

$$f(\mathbf{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\},$$

kde $\mathbf{x} \in \mathbb{R}^p$, $|\boldsymbol{\Sigma}|$ značí determinant matice $\boldsymbol{\Sigma}$. Matice $\boldsymbol{\Sigma}_{(p\times p)}$ je symetrická a pozitivně semidefinitní. Hodnosti matice $\boldsymbol{\Sigma}$, tj. $h(\boldsymbol{\Sigma})$, se říká *řád rozdělení*. Je-li $h(\boldsymbol{\Sigma}) = p$, resp. $h(\boldsymbol{\Sigma}) < p$, pak říkáme, že mnohorozměrné normální rozdělení je *regulární*, resp. *singulární*. Matice $\boldsymbol{\Sigma}$ musí být regulární, aby byla invertovatelná. Hustota existuje jen pro regulární rozdělení. Pro $p = 1$ se hustota zredukuje na hustotu jednorozměrného normálního rozdělení.

Nechť $\mathbf{x} = (X_1, \ldots, X_p)'$, $\mathbf{y} = (Y_1, \ldots, Y_p)'$ jsou náhodné vektory, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)' = \mathsf{E}\mathbf{x} = (\mathsf{E}X_1, \ldots, \mathsf{E}X_p)'$ je vektor středních hodnot a $\boldsymbol{\Sigma} = (\sigma_{jk})_{j,k=1}^p = \operatorname{var}\mathbf{x} = (\operatorname{cov}(X_j, X_k))$ je varianční matice.

Věta 1.7. *Nechť* $\mathbf{x}$ *má* p*-rozměrné normální rozdělení, potom existuje střední hodnota* $\mathsf{E}\mathbf{x} = \boldsymbol{\mu}$ *a varianční matice* $\operatorname{var}\mathbf{x} = \boldsymbol{\Sigma}$ *(tj. mají konečné prvky) a platí*

$$\forall \mathbf{c}_{(p\times 1)} : \mathbf{c}'\mathbf{x} \sim N(\mathbf{c}'\boldsymbol{\mu}, \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}).$$

Důkaz: Je nutné dokázat, že $\forall \mu_j < \infty$ a $\forall \sigma_{jk} < \infty$.

Zvolme tedy p-rozměrný vektor, $\mathbf{c}_j = (0, \ldots, 0, 1, 0, \ldots, 0)'$, kde na j-té pozici je jednička. Z definice mnohorozměrné normality víme, že lineární kombinace $\mathbf{c}_j'\mathbf{x}$ má jednorozměrné normální rozdělení. Protože $\mathbf{c}_j'\mathbf{x} = X_j$ má konečné momenty

$$\mathsf{E}X_j = \mu_j, \quad \operatorname{var} X_j = \sigma_{jj},$$

potom $\sigma_{jj} < \infty,\ \mu_j < \infty,\ j = 1, \ldots, p.$

Zvolme jiný p-rozměrný vektor, $\mathbf{c}_{jk} = (0, \ldots, 0, 1, 0, \ldots, 0, 1, 0, \ldots, 0)'$, kde na j-té a k-té pozici je jednička. Lineární kombinace $\mathbf{c}_{jk}'\mathbf{x}$ má jednorozměrné normální rozdělení, sčítá $X_j + X_k$, a tedy

$$\mathsf{E}(X_j + X_k) = \mu_j + \mu_k < \infty,$$

$$\mathrm{var}(X_j + X_k) = \sigma_{jj} + \sigma_{kk} + \sigma_{jk} + \sigma_{kj}.$$

Ze symetrie matice $\mathbf{\Sigma}$ platí, že $\sigma_{jk} = \sigma_{kj}$. Schwarzovou nerovností

$$\sigma_{jk} = \sigma_{kj} \leq \sqrt{\sigma_{jj}\sigma_{kk}} < \infty$$

jsme ověřili konečnost mimodiagonálních prvků matice $\mathbf{\Sigma}$. Existence je dokázána. Dopočítáme ještě momenty.

$$\mathsf{E}\mathbf{c}'\mathbf{x} = \mathbf{c}'\mathsf{E}\mathbf{x} = \mathbf{c}'\boldsymbol{\mu}, \quad \mathrm{var}\,\mathbf{c}'\mathbf{x} = \mathbf{c}'\,\mathrm{var}\,\mathbf{x}\,\mathbf{c} = \mathbf{c}'\mathbf{\Sigma}\mathbf{c}$$

□

Mnohorozměrné normální rozdělení je plně určeno parametry $\boldsymbol{\mu}_{(p\times 1)}$ a $\mathbf{\Sigma}_{(p\times p)}$. Je to zřejmé už z rozdělení lineární kombinace $\mathbf{c}'\mathbf{x}$. Používáme značení $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$.

U následujících tvrzení důkazy odcitujeme a provedeme jen ty, kde lze výhodně použít charakteristickou funkci. Charakteristická funkce rozdělení $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ [14], s. 299, je dána vztahem

$$\psi(\mathbf{t}) = \exp\left\{i\mathbf{t}'\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}'\mathbf{\Sigma}\mathbf{t}\right\}, \ \mathbf{t} \in \mathbb{R}^p,$$

kde i je komplexní jednotka. Charakteristická funkce existuje i pro singulární rozdělení. Protože není nutné invertovat matici $\mathbf{\Sigma}$, pracuje se s charakteristickou funkcí snadněji než s hustotou.

Věta 1.8. (Lineární transformace.) *Nechť* $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$. *Nechť* $\mathbf{B}_{r\times p}, \mathbf{d}_{r\times 1}$ *jsou konstantní, potom náhodný vektor*

$$\mathbf{y} = \mathbf{d} + \mathbf{B}\mathbf{x} \ \sim N_r(\mathbf{d} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\mathbf{\Sigma}\mathbf{B}').$$

Důkaz: Z vlastností charakteristické funkce náhodného vektoru plyne, že

$$\psi_{\mathbf{y}}(\mathbf{t}) = \mathsf{E}\exp\{i\mathbf{t}'(\mathbf{d} + \mathbf{B}\mathbf{x})\} = \exp(i\mathbf{t}'\mathbf{d})\mathsf{E}\exp(i\mathbf{t}'\mathbf{B}\mathbf{x}) = \exp(i\mathbf{t}'\mathbf{d})\psi_{\mathbf{x}}(\mathbf{B}'\mathbf{t}) =$$

$$= \exp(i\mathbf{t}'\mathbf{d})\psi_{\mathbf{x}}(\mathbf{u}),$$

kde $\mathbf{u} = \mathbf{B}'\mathbf{t}$. Protože $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ a má charakteristickou funkci $\psi_{\mathbf{x}}$, potom náhodný vektor $\mathbf{y}$ bude mít charakteristickou funkci r-rozměrného normálního rozdělení. Dojde tak jen ke změně parametrů normálního rozdělení.

Ze vztahu pro charakteristickou funkci mnohorozměrného normálního rozdělení bude

$$\psi_{\mathbf{y}}(\mathbf{t}) = \exp(i\mathbf{t}'\mathbf{d})\psi_{\mathbf{x}}(\mathbf{u}) = \exp(i\mathbf{t}'\mathbf{d}) \exp\left\{ i\mathbf{t}'\mathbf{B}\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}'\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}'\mathbf{t} \right\} =$$

$$= \exp\left\{ i\mathbf{t}'(\mathbf{d} + \mathbf{B}\boldsymbol{\mu}) - \frac{1}{2}\mathbf{t}'\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}'\mathbf{t} \right\} = \exp\left\{ i\mathbf{t}'[\mathsf{E}\mathbf{y}] - \frac{1}{2}\mathbf{t}'[\operatorname{var}\mathbf{y}]\mathbf{t} \right\}.$$

□

Lineární transformace zachovává normalitu. Tedy normální bude i každá lineární kombinace složek náhodného vektoru $\mathbf{x}$. Nechť

$$\mathbf{x}_{(p\times 1)} = \begin{pmatrix} {}_1\mathbf{x}_{(r\times 1)} \\ {}_2\mathbf{x}_{(s\times 1)} \end{pmatrix}, \ \boldsymbol{\mu}_{(p\times 1)} = \begin{pmatrix} {}_1\boldsymbol{\mu}_{(r\times 1)} \\ {}_2\boldsymbol{\mu}_{(s\times 1)} \end{pmatrix}, \ \boldsymbol{\Sigma}_{(p\times p)} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

kde $p = r + s$ a $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{12}$, $\boldsymbol{\Sigma}_{21}$, $\boldsymbol{\Sigma}_{22}$ jsou blokové matice po řadě typu $r \times r$, $r \times s$, $s \times r$, $s \times s$.

Věta 1.9. (O marginálním rozdělení.) *Nechť* $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *Potom náhodný vektor* ${}_1\mathbf{x} \sim N_r({}_1\boldsymbol{\mu}, \boldsymbol{\Sigma}_{11})$.

Důkaz: Nechť $\mathbf{y} = \mathbf{d} + \mathbf{B}\mathbf{x}$, nechť $\mathbf{d} = \mathbf{0}_{(r\times 1)}$ a $\mathbf{B}_{(r\times p)} = (\mathbf{I}_r, \mathbf{0}_{(r\times s)})$, kde $\mathbf{I}_r$ je jednotková matice typu $r \times r$. Dále pokračujeme obdobně jako v důkaze věty o lineární transformaci.

□

Věta 1.10. *Vektory* ${}_1\mathbf{x}, {}_2\mathbf{x}$ *jsou nezávislé tehdy a jen tehdy, když* $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}'_{21} = \mathbf{0}$.

Důkaz: Užijeme charakteristickou funkci. Víme, že ${}_1\mathbf{x}$, ${}_2\mathbf{x}$ budou nezávislé právě tehdy, když $\psi_{\mathbf{x}}(\mathbf{t}) = \psi_{{}_1\mathbf{x}}({}_1\mathbf{t})\psi_{{}_2\mathbf{x}}({}_2\mathbf{t})$. (Obdobně to platí i pro hustotu.) Kvadratickou formu upravíme podle bloků, tj.

$$\psi_{\mathbf{x}}(\mathbf{t}) = \exp\left\{ i\mathbf{t}'\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t} \right\} =$$

$$= \exp\left\{ i_1 \mathbf{t}'_1\boldsymbol{\mu} + i_2 \mathbf{t}'_2\boldsymbol{\mu} - \frac{1}{2}{}_1\mathbf{t}'\boldsymbol{\Sigma}_{11}\,{}_1\mathbf{t} - \frac{1}{2}{}_1\mathbf{t}'\boldsymbol{\Sigma}_{12}\,{}_2\mathbf{t} - \frac{1}{2}{}_2\mathbf{t}'\boldsymbol{\Sigma}_{21}\,{}_1\mathbf{t} - \frac{1}{2}{}_2\mathbf{t}'\boldsymbol{\Sigma}_{22}\,{}_2\mathbf{t} \right\} =$$

$$= \psi_{{}_1\mathbf{x}}({}_1\mathbf{t})\psi_{{}_2\mathbf{x}}({}_2\mathbf{t}),$$

pokud $\boldsymbol{\Sigma}_{12} = \mathbf{0}$.

□

K termínům nezávislost a nekorelovanost pouze poznamenejme, že z nezávislosti plyne nekorelovanost vždy, ale z nekorelovanosti plyne nezávislost pouze za splnění předpokladu normality. Nezávislost je tedy přísnější pojem.

Věta 1.11. (O hustotě.) *Nechť* $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $h(\boldsymbol{\Sigma}) = p$, *tj. regulární rozdělení. Potom existuje hustota*

$$f(\mathbf{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \right\},\ \mathbf{x} \in \mathbb{R}^p.$$

Důkaz: Viz [3], s. 65.

□

Z geometrického hlediska představuje výraz $c^2 = (\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ kvadratickou formu vzdálenosti $\mathbf{x}$ od $\boldsymbol{\mu}$. Jsou to kontury elipsoidu se středem v $\boldsymbol{\mu}$ a s poloosami tvaru $c\sqrt{\lambda_j}\mathbf{v}_j$, kde λ_j jsou vlastní čísla a $\mathbf{v}_j$ jsou ortonormální vektory matice $\boldsymbol{\Sigma}$. Hustota $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ se obvykle zakresluje pomocí elips konstantní hustoty, tj. $f(\mathbf{x}) = k$, kde k je zvolené číslo z intervalu $\langle 0; \max f(\mathbf{x}) \rangle$. Hustota $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ nabývá svého maxima pro střední hodnotu. Po dosazení a úpravě příslušné rovnice dostaneme

$$c^2 = -2\ln\left(k(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}} \right).$$

Kontury lze graficky znázornit pouze pro $p = 2$, tj. pro $f(x_1, x_2) = k$. Pokud X_1, X_2 jsou nezávislé náhodné veličiny, potom osy elipsy jsou rovnoběžné s osami souřadnic. Pokud X_1, X_2 jsou „závislé“ náhodné veličiny, potom osy elipsy jsou pootočeny a úhel pootočení závisí na rozptylech první a druhé složky a na korelačním koeficientu. Také představa hustoty $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ jako takové je omezená, proto

se nejčastěji zobrazuje hustota dvourozměrného normálního rozdělení, která má kloboukovitý tvar.

Nechť $\mathbf{x} = (X_1, X_2)'$. Potom s využitím vztahu pro obyčejný korelační koeficient ρ je

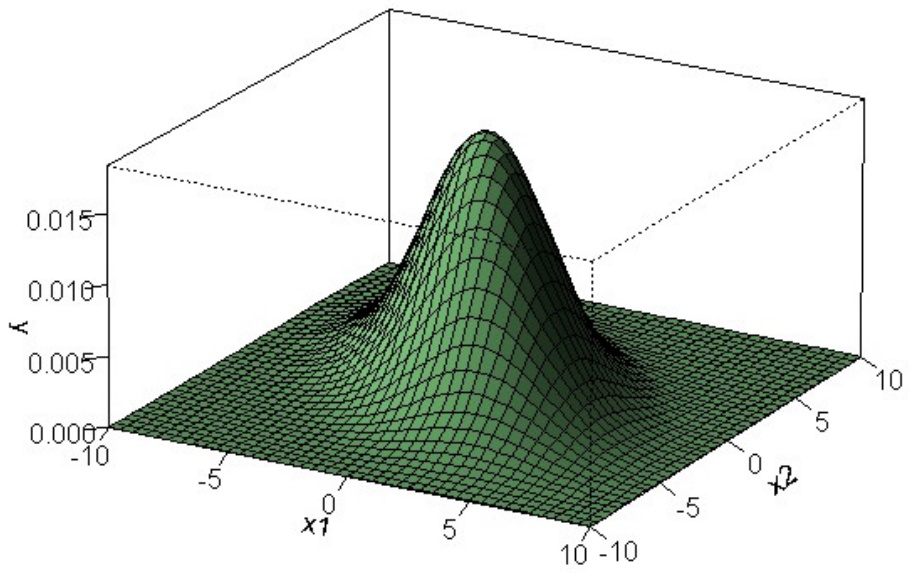
$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2, \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2, \sigma_2^2 \end{pmatrix},$$

$$\mathbf{\Sigma}^{-1} = \frac{1}{|\mathbf{\Sigma}|}\,\mathrm{adj}(\mathbf{\Sigma}) = \frac{1}{1-\rho^2}\begin{pmatrix} \frac{1}{\sigma_1^2}, \frac{-\rho}{\sigma_1\sigma_2} \\ \frac{-\rho}{\sigma_1\sigma_2}, \frac{1}{\sigma_2^2} \end{pmatrix},$$

kde $\mathrm{adj}(\mathbf{\Sigma})$ značí adjugovanou matici k matici $\mathbf{\Sigma}$. Hustota regulárního dvourozměrného normálního rozdělení je tedy ve tvaru

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times$$

$$\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right\}.$$



Obr.: Hustota dvourozměrného normálního rozdělení s parametry

$$\mu_1 = \mu_2 = 0,\ \sigma_1^2 = \sigma_2^2 = 10,\ \rho = 0.5.$$

Normální rozdělení je nezastupitelné v teorii pravděpodobnosti i v praktických statistických úlohách. V matematické statistice se také můžeme setkat se situací, kdy víme, že pozorování nepochází z normálního rozdělení. Existují cesty jak se vyhnout předpokladu normality. V tomto případě je nutné použít některou z metod nezávislých na rozdělení, například pořadové testy, nebo robustní metody, které jsou necitlivé na porušení předpokladu normality. Ta ovšem i takto zůstává stěžejní součástí většiny postupů mnohorozměrné statistické analýzy.

2 Kompoziční data

2.1 Definice a základní principy

Cílem této kapitoly je seznámení se s kompozičními daty jako matematickými objekty včetně jejich softwarového řešení, formulace hlavních principů a vlastností kompozičních dat ve spojení s jejich výběrovým prostorem [1].

Kvantitativní mnohorozměrná data můžeme z hlediska přenosu informace rozdělit na dvě hlavní skupiny. Jsou to data, která nesou absolutní informaci a data, která nesou relativní informaci, například proporce nebo procenta. Kompoziční data jsou právě speciálním typem mnohorozměrných dat, která nesou pouze relativní informaci, tedy dávají smysl jen tehdy, pokud jsou vázána k nějakému celku. Přesněji si pojem kompoziční data zformulujeme v následující definici [7].

Definice 2.1. D-složkový kompoziční vektor, nebo jednoduše kompozice, je kladný reálný vektor $\mathbf{x} = (x_1, \ldots, x_D)'$ popisující kvantitativně části nějakého celku nesoucí výhradně relativní informaci mezi složkami.

Definice tedy říká, že jedinou relevantní informaci obsahují podíly mezi složkami kompozice. Důsledkem definice kompozice je definice uzávěru.

Definice 2.2. Uzávěr kompozice $\mathbf{x} = (x_1, \ldots, x_D)'$ vzhledem ke konstantnímu součtu k je vektor

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

Definice 2.3. Subkompozice $\mathbf{x}_s$ dané kompozice $\mathbf{x}$ je vektor $(x_{i_1}, \ldots, x_{i_s})'$ představující část kompozice $\mathbf{x}$, kde indexy $i_1, \ldots, i_s$ určují vybrané složky.

V souvislosti s definicí kompozice vzniká praktický problém [4]. Složky kompozice jsou přirozeně zavedeny jako kladná reálná čísla, ale v praxi se může některá ze složek jevit jako nulová. Vyskytnout se přitom mohou dva druhy nul. Jsou to nuly strukturální, například pro měsíční výdaje domácností bude položka alkoholické nápoje u abstinentů nulová, a nuly vzniklé zaokrouhlováním, například

u koncentrace chemických prvků se může nějaký prvek vyskytnout pouze ve stopovém množství. Strukturální nuly v datovém souboru lze ošetřit tak, že vezmeme nějakou hodnotu blízkou nule nebo se snažíme pracovat se subkompozicemi. Zaokrouhlovací nuly se snažíme nahradit nějakým malým číslem, které odpovídá struktuře dat, tj. očekávanou malou hodnotou, kterou bychom změřili přesnějším přístrojem.

Protože složky dané kompozice nesou pouze relativní informaci, nese i kladný reálný násobek kompozice stejnou informaci. Definice tak umožňuje přiřadit kompozici libovolný součet $k \in \mathbb{R}^+$ jejích složek. V případě $k = 1$ nebo $k = 100$ budou složky kompozice vyjadřovat proporce nebo procentuální podíly na celku.

Kompoziční data si představíme na jednoduchém příkladě. Nápoj Piña colada se sestává z 4 cl kubánského rumu, 2 cl kokosového sirupu, 12 cl ananasového džusu a 2 cl smetany ke šlehání. Máme čtyřsložkovou kompozici $\mathbf{x} = (4, 2, 12, 2)'$ s konstantním součtem $k = 20$. Pro přípravu dvou Piña colad stačí vynásobit kompozici $\mathbf{x}$ dvěma, tedy hodnoty složek kompozice se změní, ale informace o přípravě nápoje, tj. o podílech složek v něm, je zachována. Zřejmě vynásobíme-li kompozici $\mathbf{x}$ číslem $\frac{1}{20}$, dostaneme kompozici $(0.2, 0.1, 0.6, 0.1)'$, tj. proporce jednotlivých surovin na přípravu nápoje. Důležité ale je, že se žádná ze složek kompozice nesmí vynechat, a to i při zachování proporcí ve složkách zbývajících. V kompozici je vše vzájemně provázáno jako chutě v koktejlu. Pokud bychom k přípravě nepoužili například kokosový sirup, musel by takto namíchaný koktejl nést jiné jméno. Vynechání složky popírá podstatu celku.

Definice 2.4. Říkáme, že vektory $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D_+$ jsou kompozičně ekvivalentní, jestliže existuje takové číslo $\lambda \in \mathbb{R}^+$, že platí $\mathbf{x} = \lambda\mathbf{y}$, a ekvivalentně $\mathcal{C}(\mathbf{x}) = \mathcal{C}(\mathbf{y})$.

Dalšími principy práce s kompozicemi je *permutační invariance*, tj. libovolná permutace složek kompozice nemění celkovou informaci nesenou kompozičním vektorem a *subkompoziční dominance*, tj. vzdálenost mezi každými dvěmi D-složkovými kompozicemi musí být větší nebo rovna vzdálenosti mezi dvěmi s-složkovými subkompozicemi vybranými z těchto kompozic, kde $D \geq s$.

2.2 Aitchisonova geometrie na simplexu

D-složkové kompozice přirozeně indukují odlišný výběrový prostor [13]. Je jím *D-složkový simplex* definovaný jako

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

Zvolme trojsložkovou kompozici $\mathbf{x} = (x_1, x_2, x_3)'$. Potom simplex bude rovnostranný trojúhelník s vrcholy $A = [k, 0, 0],\ B = [0, k, 0],\ C = [0, 0, k]$, s délkou strany $k\sqrt{2}$ a výškou $k\sqrt{\frac{3}{2}}$. Velikost plochy simplexu bude závislá na konkrétní volbě součtu složek kompozice.

Kompozici můžeme interpretovat jako bod na simplexu. Nicméně simplex jako výběrový prostor při pevně zvoleném k nepojme všechny kompozice. Geometricky si prostor všech kompozic, $\mathbb{R}^D_+$, můžeme představit jako prostor, který je určen vybíhajícími paprsky z počátku protínajícími vždy právě jeden bod simplexu. Každý paprsek tvoří třídu ekvivalentních kompozic. Podstatou operace uzávěru je potom přiřazení dané třídě ekvivalentních kompozic právě jednoho zástupce na simplexu. Uzávěr je tedy projekcí vektoru s kladnými složkami z $\mathbb{R}^D$ do simplexu.

V praxi se trojsložkové kompozice zobrazují pomocí ternárního diagramu. *Ternární diagram* je rovinný rovnostranný trojúhelník s vrcholy $X_1,\ X_2,\ X_3$, kompozice $\mathbf{x}$ je v něm zobrazena pomocí složek tak, že x_1 představuje vzdálenost od protilehlé strany k vrcholu X_1, x_2 je vzdálenost od protilehlé strany k vrcholu X_2 a x_3 je vzdálenost od protilehlé strany k vrcholu X_3.

Obecně pracujeme s mnohorozměrnými daty v reálném vektorovém prostoru s využitím standardní euklidovské geometrie. Umíme tedy vektory sčítat, násobit skalárem, známe neutrální prvek, tj. nulový vektor, umíme zde odvodit další vlastnosti jako je ortogonalita, pro výpočet vzdálenosti dvou vektorů používáme euklidovskou normu. Máme tedy nástroje potřebné k reprezentaci dat v rámci této geometrie. Avšak euklidovská geometrie není vhodná pro kompoziční data, například nerespektuje subkompoziční dominanci, relativní škálu kompo-

zic. Euklidovské vzdálenosti mezi kompozicemi $(30, 70)'$ a $(60, 40)'$ a kompozicemi $(10, 90)'$ a $(40, 60)'$ jsou si rovny, tj. $30\sqrt{2}$, ale přitom relativní přírůstek je v prvním případě pro první složku stoprocentní a ve druhém čtyřnásobek. Důsledkem pak může být interpretace výsledků statistického zpracování kompozic standardními metodami založenými na principech euklidovské geometrie, která může vést ke zcela nesmyslným závěrům.

Pro práci s kompozičními daty je nutné zavést vhodnou geometrii na simplexu. Hledáme analogii pro práci s kompozičními daty na simplexu k práci v euklidovském prostoru. Zavedeme nyní dvě operace na simplexu, které jsou analogické sčítání vektorů a násobení vektoru skalárem v D-dimenzionálním reálném prostoru.

Definice 2.5. Perturbace kompozice $\mathbf{x} = \mathcal{C}(x_1, \ldots, x_D)' \in S^D$ kompozicí $\mathbf{y} = \mathcal{C}(y_1, \ldots, y_D)' \in S^D$ je kompozice $\mathbf{x} \oplus \mathbf{y} \in S^D$ definovaná vztahem

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

Definice 2.6. Mocninná transformace kompozice $\mathbf{x} = \mathcal{C}(x_1, \ldots, x_D)' \in S^D$ číslem $\alpha \in \mathbb{R}$ je kompozice $\alpha \odot \mathbf{x} \in S^D$ definovaná vztahem

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

Protože operace perturbace a mocninná transformace splňují následující axiomy,

1. $(S^D, \oplus)$ tvoří komutativní grupu, tj. pro libovolné kompozice $\mathbf{x}, \mathbf{y}, \mathbf{z} \in S^D$ platí
   - komutativita: $\mathbf{x} \oplus \mathbf{y} = \mathbf{y} \oplus \mathbf{x}$,
   - asociativita: $(\mathbf{x} \oplus \mathbf{y}) \oplus \mathbf{z} = \mathbf{x} \oplus (\mathbf{y} \oplus \mathbf{z})$,
   - existuje neutrální prvek $\mathbf{n} = \mathcal{C}(1, \ldots, 1)'$ a platí $\mathbf{x} \oplus \mathbf{n} = \mathbf{x}$,
   - inverze: $\mathbf{x} \oplus \mathbf{x}^{-1} = \mathbf{n}$, kde $\mathbf{x}^{-1} = \mathcal{C}(x_1^{-1}, \ldots, x_D^{-1})'$;
2. pro libovolné kompozice $\mathbf{x}, \mathbf{y} \in S^D$ a $\alpha, \beta \in \mathbb{R}$ platí

– neutrální prvek: $1 \odot \mathbf{x} = \mathbf{x}$,

– asociativita: $\alpha \odot (\beta \odot \mathbf{x}) = (\alpha \cdot \beta) \odot \mathbf{x}$,

– distributivita: $(\alpha + \beta) \odot \mathbf{x} = (\alpha \odot \mathbf{x}) \oplus (\beta \odot \mathbf{x})$,

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

je $(S^D, \oplus, \odot)$ reálným vektorovým prostorem. Poznamenejme ještě, že ve shodě se značením v reálném vektorovém prostoru značíme kompozici $\mathbf{x} \oplus \mathbf{y}^{-1} = \mathbf{x} \ominus \mathbf{y}$.

Definujme nyní skalární součin, normu a vzdálenost na simplexu.

Definice 2.7. Aitchisonův skalární součin kompozic $\mathbf{x}, \mathbf{y} \in S^D$ definujeme vztahem

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

Definice 2.8. Aitchisonova norma kompozice $\mathbf{x} \in S^D$ je dána jako

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

Definice 2.9. Aitchisonovu vzdálenost kompozic $\mathbf{x}, \mathbf{y} \in S^D$ definujeme vztahem

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

Prostor $(S^D, \oplus, \odot)$ společně s operacemi skalární součin, norma a vzdálenost tvoří $(D-1)$-dimenzionální euklidovský vektorový prostor. V teorii kompozičních dat jej nazýváme *Aitchisonova geometrie na simplexu.*

2.3 Logratio transformace kompozičních dat

Aitchisonova geometrie na simplexu má vlastnosti euklidovské geometrie a zdá se tak pro kompoziční data nejlepší možnou volbou. Aitchisonova geometrie ovšem (až na výjimky) není vhodná pro statistickou analýzu dat tohoto typu.

Cílem logratio transformací je tak převést kompoziční data z Aitchisonovy geometrie izometricky do standardní euklidovské geometrie, abychom předešli chybným interpretacím výsledků užitých statistických metod a také proto, že práce s daty ve standardní euklidovské geometrii je snadněji interpretovatelná a „zažitější“ než práce na simplexu. Cest, jak dosáhnout tohoto cíle, je více.

Kompozici vyjádříme jako souřadnice vzhledem k bázi [5], [13]. Budeme požadovat, aby báze byla ortonormální, neboť ortonormalita báze s sebou nese výhodné vlastnosti.

Kompoziční data bohužel nelze interpretovat stejně jako v $\mathbb{R}^D$, tj. pomocí kanonické báze, protože vektory kanonické báze euklidovského prostoru $\mathbb{R}^D$ nejsou množinou generujících prvků ani bází na simplexu.

K tomu, abychom zkonstruovali ortonormální bázi na simplexu, je nutné najít vhodný generující systém. Vezměme systém kompozic $\{\mathbf{w}_1, \ldots, \mathbf{w}_D\}$, kde $\mathbf{w}_i = \mathcal{C}(1, \ldots, e, \ldots, 1)'$ pro $i = 1, \ldots, D$ a Eulerovo číslo e je na i-té pozici. Potom můžeme každou kompozici $\mathbf{x} \in S^D$ psát ve tvaru

$$\mathbf{x} = \ln x_1 \odot (e, 1, \ldots, 1)' \oplus \ldots \oplus \ln x_D \odot (1, \ldots, 1, e)'.$$

Protože koeficienty s ohledem na množinu generujících prvků nejsou určeny jednoznačně, můžeme psát ekvivalentně

$$\mathbf{x} = \ln \frac{x_1}{g(\mathbf{x})} \odot (e, 1, \ldots, 1)' \oplus \ldots \oplus \ln \frac{x_D}{g(\mathbf{x})} \odot (1, \ldots, 1, e)',$$

kde $g(\mathbf{x}) = \left(\prod_{i=1}^{D} x_i \right)^{\frac{1}{D}}$ je geometrický průměr složek dané kompozice.

Tímto se dostáváme k první transformaci zvané *centered logratio* neboli zkráceně *clr* transformace. Centered logratio transformuje kompozici $\mathbf{x} \in S^D$ na D-složkový vektor $\mathbf{y} = (y_1, \ldots, y_D)'$ euklidovského prostoru a definujeme ji vztahem

$$clr(\mathbf{x}) = \left(\ln \frac{x_1}{g(\mathbf{x})}, \ldots, \ln \frac{x_D}{g(\mathbf{x})} \right)' = \mathbf{y}.$$

Clr transformace je izometrická a symetrická ve složkách. Součet složek vektoru $\mathbf{y}$ je nula, což má za následek singularitu dat. Z algebraického hlediska je *clr* izomorfismus mezi S^D a podprostorem prostoru $\mathbb{R}^D$. Geometricky, transformovaný

vektor leží na nadrovině procházející počátkem prostoru $\mathbb{R}^D$ a ortogonální k vektoru $(1,\ldots,1)'$. *Clr* transformace se používá při konstrukci tzv. kompozičního biplotu [13], kapitola 5.4.

Důsledkem takto zavedené transformace jsou následující vztahy [13], s. 20. Pro libovolné kompozice $\mathbf{x}_1, \mathbf{x}_2 \in S^D$ a $\alpha, \beta \in \mathbb{R}$ platí

$$clr(\alpha \odot \mathbf{x}_1 \oplus \beta \odot \mathbf{x}_2) = \alpha clr(\mathbf{x}_1) + \beta clr(\mathbf{x}_2),$$

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_a = \langle clr(\mathbf{x}_1), clr(\mathbf{x}_2) \rangle\,,\ \|\mathbf{x}_1\|_a = \|clr(\mathbf{x}_1)\|\,,$$

$$d_a(\mathbf{x}_1, \mathbf{x}_2) = d(clr(\mathbf{x}_1), clr(\mathbf{x}_2)),$$

kde na pravých stranách rovností v posledních dvou řádcích jsou po řadě uvedeny euklidovský skalární součin, euklidovská norma a vzdálenost.

Vybereme-li z generujícího systému $D-1$ prvků, dostaneme bázi $\mathbf{w}_1, \ldots, \mathbf{w}_{D-1}$. Potom můžeme každou kompozici $\mathbf{x} \in S^D$ psát ve tvaru

$$\mathbf{x} = \ln \frac{x_1}{x_D} \odot (e, 1, \ldots, 1)' \oplus \ldots \oplus \ln \frac{x_{D-1}}{x_D} \odot (1, \ldots, 1, e, 1)'.$$

Tímto se dostáváme k druhé transformaci zvané *additive logratio* neboli zkráceně *alr* transformace. Additive logratio transformuje kompozici $\mathbf{x} \in S^D$ na $(D-1)$-složkový vektor $\mathbf{y} = (y_1, \ldots, y_{D-1})'$ euklidovského prostoru a definujeme ji nejčastěji vztahem

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

Alr transformací lze ovšem obecně vytvořit přesně tolik, kolik má kompozice složek. Volba dělící složky by přitom měla vycházet z výhodnosti interpretace výsledků. *Alr* transformace není symetrická ve složkách a není izometrická, přičemž druhá skutečnost má „negativní“ dopad například na výpočet vzdálenosti. *Alr* transformace se používá, mimo jiné, pro detekci odlehlých pozorování [8].

Výhodou *clr* a *alr* transformace je zdánlivě snadnější interpretace výsledných souřadnic. Problémem *clr* a *alr* transformací ovšem je, že nejsou přímo svázány s ortogonálním souřadnicovým systémem na simplexu. Tento a výše uvedené problémy řeší transformace zvaná *isometric logratio* neboli krátce *ilr* transformace,

která je založená na volbě ortonormální báze na nadrovině euklidovského prostoru dimenze D vytvořené *clr* transformací.

Volbou $D-1$ kompozic získáme bázi $\mathbf{w}_1, \ldots, \mathbf{w}_{D-1}$. Tato báze ještě obecně není ortonormální. Ortonormální báze $\mathbf{v}_1, \ldots, \mathbf{v}_{D-1} \in S^D$ vzhledem k Aitchisonově geometrii dosáhneme užitím Gram-Schmidtovy ortonormalizační metody. Ortonormálních bází je možné nalézt mnoho, jedna z možností je vzít vektory tvaru

$$\mathbf{v}_i = \mathcal{C}\left(\exp\left\{\frac{1}{\sqrt{(D-i+1)(D-i)}}\right\}, \ldots, \exp\left\{\frac{1}{\sqrt{(D-i+1)(D-i)}}\right\},\right.$$

$$\left.\exp\left\{\sqrt{\frac{D-i+1}{D-i}}\right\}, 1, \ldots, 1\right)', \ i = 1, \ldots, D-1,$$

kde poslední složky, jejichž počet je $D-i$, jsou jedničky.

Protože simplex má vlastnosti euklidovské geometrie, můžeme výhodně použít skalární součin pro nalezení ortonormálních souřadnic libovolné kompozice $\mathbf{x} \in S^D$, tj.

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

Isometric logratio transformuje kompozici $\mathbf{x} \in S^D$ při pevně zvolené bázi na $(D-1)$-složkový vektor $\mathbf{x}^* = (x_1^*, \ldots, x_{D-1}^*)'$ euklidovského prostoru a definujeme ji vztahem

$$ilr(\mathbf{x}) = (\langle \mathbf{x}, \mathbf{v}_1 \rangle_a, \ldots, \langle \mathbf{x}, \mathbf{v}_{D-1} \rangle_a)' = \mathbf{x}^*,$$

kde ortonormální souřadnice pro výše uvedenou bázi můžeme psát explicitně jako

$$x_i^* = \langle \mathbf{x}, \mathbf{v}_i \rangle_a = \sqrt{\frac{i}{i+1}} \ln \frac{\sqrt{\prod_{j=1}^{D} x_j}}{x_{i+1}}, \ i = 1, \ldots, D-1.$$

Ilr transformace má izometrické vlastnosti *clr* transformace a zároveň řeší singularitu dat. Důsledkem takto zavedené transformace jsou potom následující vztahy [13], s. 21. Pro libovolné kompozice $\mathbf{x}_1, \mathbf{x}_2 \in S^D$ a $\alpha, \beta \in \mathbb{R}$ platí

$$ilr(\alpha \odot \mathbf{x}_1 \oplus \beta \odot \mathbf{x}_2) = \alpha ilr(\mathbf{x}_1) + \beta ilr(\mathbf{x}_2),$$

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_a = \langle ilr(\mathbf{x}_1), ilr(\mathbf{x}_2) \rangle \,, \;\; \|\mathbf{x}_1\|_a = \|ilr(\mathbf{x}_1)\| \,,$$

$$d_a(\mathbf{x}_1, \mathbf{x}_2) = d(ilr(\mathbf{x}_1), ilr(\mathbf{x}_2)).$$

Ilr transformací je možné volbou ortonormální báze na simplexu vymyslet nekonečně mnoho, nevýhodou *ilr* transformace je obvykle pouze komplexní interpretace vzniklých souřadnic.

Všechny výše uvedené transformace jsou svázány lineárními vztahy [6].

2.4 Softwarové řešení pro kompoziční data

Práce s kompozičními daty by byla bez možnosti jejich softwarového zpracování neefektivní. Vhodným softwarem díky jeho veřejné licenci je pro statistickou analýzu dat software R dostupný na http://cran.r-project.org. Samotný program umožňuje používat jen základní funkce, proto je pro řešení speciálních problémů nutné tyto základní funkce doplnit o knihovny obsahující speciální příkazy nebo si naprogramovat vlastní funkce. Speciálně pro statistickou analýzu kompozičních dat je vhodné načíst knihovny `compositions` a `robCompositions`. Přitom zejména knihovna `robCompositions` [16] poskytuje užitečné nástroje k zpracování kompozičních dat včetně grafického zobrazení. `robCompositions` závisí na knihovnách `utils`, `robustbase`, `rrcov`, `car` a `MASS`, importuje z nich některé funkce, redefinuje některé funkce z `compositions` a bude pro nás z hlediska testování normality klíčová. Knihovnu `robCompositions` nainstalujeme příkazem `install.packages` a načteme příkazem `library`. Nápovědu k `robCompositions` získáme příkazem `help`.

```
> install.packages('robCompositions')
> library('robCompositions')
> help('robCompositions')
```

Nyní můžeme přistoupit k práci s kompozičními daty. Datovou matici $\mathbf{X}$ vytvoříme příkazem `matrix`, data se standardně uvádějí po sloupcích, přičemž funkce `c()` umožňuje kombinovat různé datové typy a tvoří sloupcový vektor,

parametr `nrow` udává počet řádků a `ncol` počet sloupců. Příkazem `colnames` pojmenujeme sloupce datové matice.

```
>  X = matrix(c(79.07,31.74,18.61,49.51,29.22,21.99,11.74,24.47,
5.14,15.54,57.17,52.25,77.40,10.54,46.14,16.29,32.27,40.73,49.29,
61.49,12.83,56.69,72.05,15.11,52.36,59.91,65.04,52.53,38.39,57.34,
3.81,23.73,9.13,20.34,15.97,69.18,36.20,47.41,42.74,7.63,8.10,
11.57,9.34,35.38,18.42,18.10,23.22,23.00,56.47,27.11,39.02,24.02,
13.47,69.12,37.89,14.53,31.53,11.86,7.97,30.88),nrow=20,ncol=3)
> colnames(X)=c("x1","x2","x3")
> X
         x1    x2    x3
 [1,] 79.07 12.83  8.10
 [2,] 31.74 56.69 11.57
 [3,] 18.61 72.05  9.34
 [4,] 49.51 15.11 35.38
 [5,] 29.22 52.36 18.42
 [6,] 21.99 59.91 18.10
 [7,] 11.74 65.04 23.22
 [8,] 24.47 52.53 23.00
 [9,]  5.14 38.39 56.47
[10,] 15.54 57.34 27.11
[11,] 57.17  3.81 39.02
[12,] 52.25 23.73 24.02
[13,] 77.40  9.13 13.47
[14,] 10.54 20.34 69.12
[15,] 46.14 15.97 37.89
[16,] 16.29 69.18 14.53
[17,] 32.27 36.20 31.53
[18,] 40.73 47.41 11.86
[19,] 49.29 42.74  7.97
```

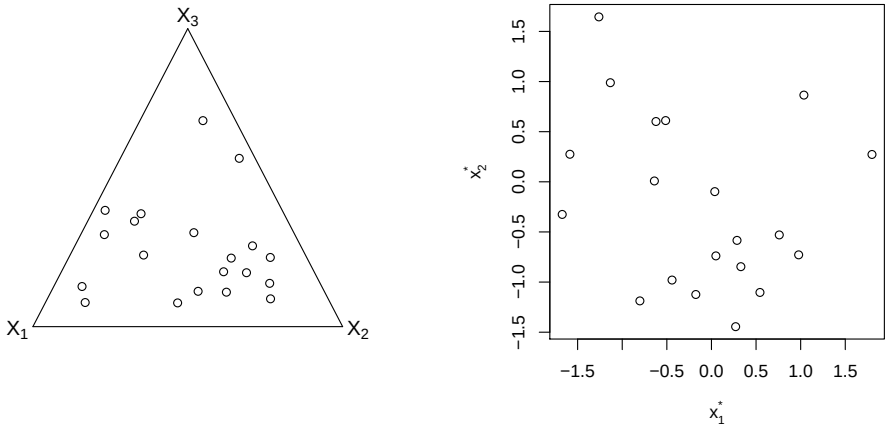
```
[20,] 61.49  7.63 30.88
```

Simulované kompozice [13], s. 10, řádky matice $\mathbf{X}$, transformujeme ze simplexu do souřadnic užitím funkce `ilr`, jejímž argumentem bude datová matice. Transformovaná matice bude mít o jeden sloupec méně, to plyne z vlastností ilr transformace. Prvky transformované matice již nemusí být nutně kladná reálná čísla.

```
> ilrX=ilr(X)
> colnames(ilrX)=c("souradnice1", "souradnice2")
> ilrX
      souradnice1  souradnice2
 [1,] -1.67260012 -0.325214048
 [2,] -0.17519651 -1.123721572
 [3,]  0.27118645 -1.444657387
 [4,] -0.62169922  0.601599316
 [5,]  0.04975451 -0.738718633
 [6,]  0.32969306 -0.846358379
 [7,]  0.97735164 -0.728311845
 [8,]  0.28658332 -0.583992574
 [9,]  1.79931367  0.272881356
[10,]  0.76018544 -0.529690718
[11,] -1.26163216  1.645045139
[12,] -0.63950172  0.008589047
[13,] -1.58642954  0.274993329
[14,]  1.03616618  0.864971717
[15,] -0.51356056  0.610922767
[16,]  0.54371505 -1.103437534
[17,]  0.03744559 -0.097665338
[18,] -0.44169414 -0.979810843
[19,] -0.80205386 -1.187551019
[20,] -1.13311458  0.988550044
```

Kompozice zobrazíme v ternárním diagramu užitím funkce `ternaryDiag` s parametry `name` pro pojmenování proměnných a `grid` s logickou hodnotou `TRUE` anebo `FALSE` pro zobrazení mřížky udávající proporcionální podíly jednotlivých složek. Funkce `expression` vytvoří objekty módu, v našem případě užitím hranatých závorek proměnnou s dolním indexem. V souřadnicích zobrazíme kompozice prostřednictvím funkce `plot` s argumenty, kterými jsou souřadnice, a s paramentry `xlab`, `ylab` pro pojmenování os, kde symbol stříška vytvoří horní index, a `type` s hodnotou `"p"` pro bodové zakreslení kompozic. Souřadnice zadáme výběrem sloupců transformované matice.

```
> ternaryDiag(X, name=c(expression(X[1],X[2],X[3])), grid=FALSE)
> ilrx1=ilrX[,1]
> ilrx2=ilrX[,2]
> plot(ilrx1, ilrx2, xlab=expression(x[1]^"*"), ylab=expression
(x[2]^"*"), type="p")
```



Obr.: Simulované kompozice v ternárním diagramu vlevo, v souřadnicích vpravo.

Knihovna `robCompositions` obsahuje, mimo jiné, funkce `alr` a `clr`, které provedou příslušnou logratio transformaci dat, funkce, které provedou inverzní

transformace, příkaz `constSum` pro operaci uzávěru a `aDist` pro výpočet Aitchisonovy vzdálenosti. Další funkce týkající se operací s kompozicemi lze například najít v knihovně `compositions` nebo nutno takové funkce naprogramovat.

3 Normální rozdělení na simplexu

Zpracování mnohorozměrných dat vyžaduje splnění předpokladu mnohorozměrného normálního rozdělení. Ne jinak je tomu i v případě kompozičních dat. V této kapitole budeme uvažovat náhodné kompozice. Nechť $\mathbf{x} = (X_1, \ldots, X_D)'$ je náhodný vektor, jehož výběrovým prostorem je simplex S^D.

Definice 3.1. Říkáme, že náhodný vektor $\mathbf{x}$ má normální rozdělení na simplexu S^D právě tehdy, když vektor ortonormálních souřadnic, $\mathbf{x}^* = ilr(\mathbf{x})$, má mnohorozměrné normální rozdělení na $\mathbb{R}^{D-1}$.

Normální rozdělení na simplexu lze definovat pomocí hustoty [12]. Postup, jak nalézt vztah pro hustotu na simplexu, se zakládá na myšlence transformovat náhodnou kompozici ze simplexu do euklidovského prostoru, zde zadefinovat hustotu transformovaného vektoru a tuto hustotu transformovat zpět na simplex.

Princip zavedení hustoty na simplexu vychází z teorie míry a transformací. Obecně jsou hustoty funkce vyjádřené s respektem k Lebesgueově míře, což je přirozená míra v euklidovském prostoru, tedy míra kompatibilní s jeho geometrickou strukturou a tudíž s absolutní mírou rozdílu. Simplex jako odlišný výběrový prostor indukuje přirozeně jinou míru.

Jestliže je nějaký výběrový prostor $E \subset \mathbb{R}^D$ úplným vektorovým prostorem se skalárním součinem, pak můžeme definovat míru λ_E kompatibilní s jeho strukturou, zvláště pak tuto míru můžeme definovat prostřednictvím Lebesgueovy míry na ortonormálních souřadnicích. Hustota f_E je definovaná na E jako Radon-Nikodýmova derivace pravděpodobnostní míry P vzhledem k míře λ_E. Míra λ_E má stejné vlastnosti v prostoru E jako Lebesgueova míra v reálném prostoru.

Přirozená míra na simplexu, λ_a, kompatibilní s jeho strukturou, je ekvivalentem k Lebesgueově míře a můžeme ji definovat užitím ortonormálních souřadnic. Míra λ_a je absolutně spojitá vzhledem k Lebesgueově míře λ v reálném prostoru a vztah mezi nimi [12], s. 496, je

$$|d\lambda_a/d\lambda| = (\sqrt{D}x_1x_2\cdots x_D)^{-1}.$$

Touto úvahou se dostáváme k vztahu pro hustotu normálního rozdělení na simplexu [12], s. 497, tj.

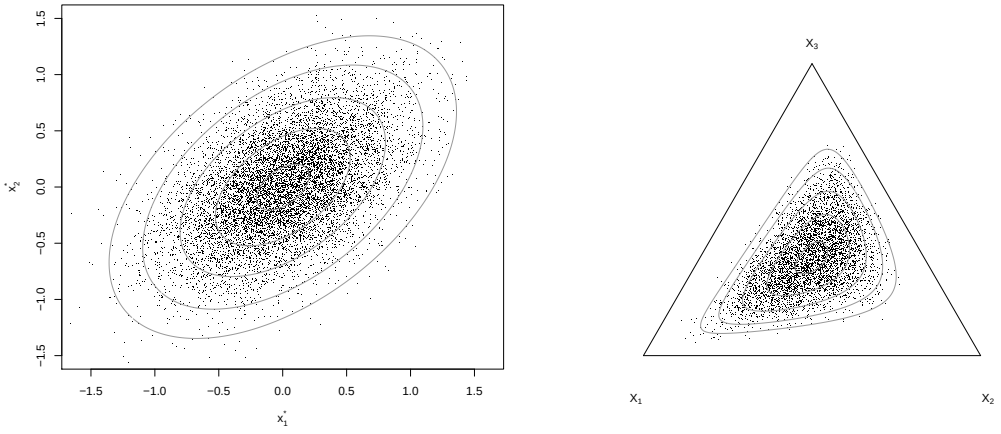
$$\frac{(2\pi)^{-\frac{(D-1)}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{\sqrt{D}x_1x_2\cdots x_D}\exp\left\{-\frac{1}{2}(ilr(\mathbf{x})-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(ilr(\mathbf{x})-\boldsymbol{\mu})\right\},\ \mathbf{x}\in S^D.$$

Hustota je vystavěna na přechodu od *alr* k *ilr* transformaci a nazývá se *logistický normální model na simplexu.* Jedná se o tzv. aditivně-logistický normální model na simplexu vyjádřený v ortonormálních souřadnicích.

Užitím algebraicko-geometrické struktury simplexu (Aitchisonovy geometrie) a uvažujeme-li přirozenou míru λ_a, definujeme normální rozdělení na simplexu prostřednictvím hustoty generických ortonormálních souřadnic [12], s. 498, tj.

$$(2\pi)^{-\frac{(D-1)}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(ilr(\mathbf{x})-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(ilr(\mathbf{x})-\boldsymbol{\mu})\right\},\ \mathbf{x}\in S^D.$$

Tato funkce je hustotou vzhledem k Lebesgueově míře v prostoru $\mathbb{R}^{D-1}$ a také vzhledem k míře λ_a v prostoru S^D.



Obr.: Normální hustota v souřadnicích vlevo a na simplexu vpravo. Data jsou náhodně generována z dvourozměrného normálního rozdělení s parametry $\mu_1 = \mu_2 = 0,\ \sigma_{11} = \sigma_{22} = 0.2,\ \sigma_{12} = \sigma_{21} = 0.1.$

S hustotou normálního rozdělení na simplexu se pracuje obdobně jako v euklidovském prostoru. Chceme-li například určit pravděpodobnost realizace náhodné

kompozice v množině $M \subset S^D$, je potřeba nejprve M transformovat a potom vypočítat příslušný $(D-1)$-rozměrný integrál.

Normální rozdělení na simplexu charakterizují, stejně jako v mnohorozměrném případě, parametry $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$. Odhadujeme je z náhodného výběru kompozic $\mathbf{x}_i = (X_{i1}, \ldots, X_{iD})'$, $i = 1, \ldots, n$, vyjádřeného v ortonormálních souřadnicích, $\mathbf{x}_i^* = (X_{i1}^*, \ldots, X_{iD-1}^*)'$, $i = 1, \ldots, n$, jako maximálně věrohodné odhady. Parametr $\boldsymbol{\mu} = \mathsf{E}[ilr(\mathbf{x})]$ odhadneme jako výběrový průměr $\hat{\boldsymbol{\mu}}$ s prvky

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}^*,\ j = 1, \ldots, D-1,$$

a parametr $\boldsymbol{\Sigma} = \mathrm{var}[ilr(\mathbf{x})]$ pomocí výběrové varianční matice $\hat{\boldsymbol{\Sigma}}$ s prvky

$$\hat{\sigma}_{jk} = \frac{1}{n} \sum_{i=1}^{n} (x_{ij}^* - \hat{\mu}_j)(x_{ik}^* - \hat{\mu}_k),\ j, k = 1, \ldots, D-1.$$

Hodnoty parametrů se liší v závislosti na volbě ortonormální báze na simplexu. Volba ortonormální báze ale nemá vliv na existenci normálního rozdělení, neboť jde jen o ortogonální rotaci transformovaných dat.

Poznamenejme ještě, že obdobou střední hodnoty na simplexu je *centrum* definované jako

$$\mathsf{E}_a[\mathbf{x}] = cen[\mathbf{x}] = \mathcal{C}(g_1, \ldots, g_D)',$$

kde $g_j = (\prod_{i=1}^{n} X_{ij})^{\frac{1}{n}}$, $j = 1, \ldots, D$. Obdobou varianční matice na simplexu je *matice rozptylů* $\mathbf{T}$ o rozměru $D \times D$ s prvky

$$t_{ij} = \mathrm{var}\left(\ln \frac{X_i}{X_j}\right).$$

4 Testování normality kompozičích dat

Ověřování normality náhodného výběru z mnohorozměrného normálního rozdělení má zásadní praktický význam. Splnění předpokladu normality implikuje využití všech možností dané teorie ke zpracování mnohorozměrných dat.

Cílem této kapitoly je analýza konkrétních přístupů k testování normality kompozičních dat včetně jejich softwarového zpracování v R a demonstrace na konkrétních příkladech.

Normalita se ověřuje statistickými testy. Obecně testujeme nulovou hypotézu, že náhodný výběr pochází z mnohorozměrného normálního rozdělení proti alternativě, že náhodný výběr pochází z nějakého jiného rozdělení, přičemž parametry rozdělení testovaná hypotéza blíže nespecifikuje.

Z normality náhodného vektoru plyne normalita příslušného marginálního rozdělení viz Věta 1.9. Naopak tvrzení obecně neplatí, platí pouze v případě nezávislosti složek náhodného vektoru. V praxi se ale ukázalo [9], že významnější odchylka od mnohorozměrné normality se projeví i na marginální normalitě, proto lze na základě těchto zkušeností usuzovat z marginální normality na mnohorozměrnou normalitu, a to s sebou nese příznivé důsledky pro její testování. Rozhodnutí o nulové hypotéze lze tedy podpořit testováním jednorozměrných nebo marginálních rozdělení, rozdělení lineárních kombinací nebo sdružených rozdělení jednotlivých párů náhodných veličin.

Účelem testování je prokázat platnost nulové hypotézy. Vhodné je provést několik různých testů z důvodu odkrytí jednotlivých zdrojů odchylek od normality. Existuje celá řada algoritmů pro testování normality [9], s. 80 - 98. Pro kompoziční data se ukázaly jako nejvhodnější [1] Anderson-Darlingův test, Cramer-von Misesův test a Watsonův test.

Anderson-Darlingův test je modifikací testu Kolmogorovova-Smirnovova [15]. Testovým kritériem Kolmogorovova-Smirnovova testu je supremum vzdálenosti mezi empirickou (skutečnou) distribuční funkcí a teoretickou (předpokládanou) plně specifikovanou distribuční funkcí, Anderson-Darlingův test uvažuje rozdíly na „chvostu“ distribuční funkce. Hodnoty empirické distribuční funkce se určují

jako kumulativní relativní četnosti ve výběru a hodnoty předpokládané distribuční funkce uspořádaných hodnot dle velikosti jsou buď dané, nebo se určí například z tabulek pro normální normované rozdělení. Cramer-von Misesův test a Watsonův test jsou modifikací testu Anderson-Darlingova. Hlavní myšlenkou pro testování p-rozměrné normality náhodného výběru $\mathbf{x}_1, \ldots, \mathbf{x}_n$ je užití toho faktu, že Mahalanobisova vzdálenost

$$(\mathbf{x} - \hat{\boldsymbol{\mu}})' \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}),$$

kde $\hat{\boldsymbol{\mu}}$ je výběrový průměr a $\hat{\boldsymbol{\Sigma}}$ je výběrová varianční matice, má χ^2 rozdělení s p stupni volnosti.

Dále nechť $D-1$ je dimenze náhodného vektoru, tj. $p = D - 1$.

Testování normality na simplexu S^D je ekvivalentní testování mnohorozměrné normality *ilr* transformovaných kompozic [13]. Testujeme hypotézu

$$H_0 : \textit{náhodný výběr pochází z normálního rozdělení na simplexu } S^D$$

proti alternativě

$$H_1 : \textit{náhodný výběr nepochází z normálního rozdělení na simplexu } S^D$$

ekvivalentně s

$$H_0 : \textit{náhodný výběr v ilr souřadnicích pochází z mnohorozměrného normálního rozdělení}$$

proti alternativě

$$H_1 : \textit{náhodný výběr v ilr souřadnicích nepochází z mnohorozměrného normálního rozdělení.}$$

Základní myšlenkou tohoto přístupu k testování normality na simplexu je vypočítat hodnoty testových kriterií pro každou souřadnici zvlášť, pro páry souřadnic a pro celou kompozici a porovnat je s kritickými hodnotami příslušných rozdělení. Zamítnutí, resp. nezamítnutí nulové hypotézy by mělo vést ve všech případech, tj. pro $D-1$ marginálních univariátních testů, $\frac{1}{2}(D-1)(D-2)$ bivariátních testů a pro $(D-1)$-dimenzionální radius test, k jednotnému rozdělení.

4.1 Marginální univariátní testy

Testujeme rozdělení každé ortonormální souřadnice $x_i^* = \langle \mathbf{x}, \mathbf{v}_i \rangle_a$ pro náhodný výběr $X_{1i}^*, \ldots, X_{ni}^*$, $i = 1, \ldots, D-1$. Testová procedura [13], s. 51, je následující.

V prvním kroce vypočítáme maximálně věrohodné odhady parametrů μ_i a σ_i^2, tj.

$$\hat{\mu}_i = \frac{1}{n}\sum_{r=1}^{n} x_{ri}^*, \quad \hat{\sigma}_i^2 = \frac{1}{n}\sum_{r=1}^{n}(x_{ri}^* - \hat{\mu}_i)^2.$$

V druhém kroce získáme, například z tabulek, hodnoty distribuční funkce normovaného normálního rozdělení pro normované hodnoty x_{ri}^*, tj.

$$\Phi\left(\frac{(x_{ri}^* - \hat{\mu}_i)^2}{\hat{\sigma}_i}\right) = z_r, \ r = 1, \ldots, n.$$

Třetím krokem je uspořádání hodnot z_r vzestupně, uspořádané hodnoty označíme jako $z_{(r)}$.

Ve čtvrtém kroce vypočítáme hodnotu testového kritéria Anderson-Darlingova "A" nebo Cramer-von Misesova "CM" nebo Watsonova "W" pro jednorozměrné rozdělení, tj.

$$A = \left(\frac{25}{n^2} - \frac{4}{n} - 1\right)\left(\frac{1}{n}\sum_{r=1}^{n}(2r-1)[\ln z_{(r)} + \ln(1 - z_{(n+1-r)})] + n\right),$$

$$CM = \left(\sum_{r=1}^{n}\left(z_{(r)} - \frac{2r-1}{2n}\right)^2 + \frac{1}{12n}\right)\left(\frac{2n+1}{2n}\right),$$

$$W = CM - \left(\frac{2n+1}{2}\right)\left(\frac{1}{n}\sum_{r=1}^{n} z_{(r)} - \frac{1}{2}\right)^2.$$

Pátým krokem je porovnání hodnoty testového kritéria s kritickou hodnotou v tabulce. Nulovou hypotézu zamítneme na hladině významnosti α ve prospěch alternativy, jestliže hodnota testového kritéria je větší než kritická hodnota.

Hladina významnosti	%	10	5	2.5	1
Anderson - Darling	A	0.656	0.787	0.918	1.092
Cramer - von Mises	CM	0.104	0.126	0.148	0.178
Watson	W	0.096	0.116	0.136	0.163

Tab.: Kritické hodnoty pro marginální univariátní testy.

4.2 Bivariátní testy

Bivariátní testy sledují chování párů *ilr* souřadnic. Pro každý pár (i, j) souřadnic, kde $i, j = 1, \ldots, D-1, i < j$, máme soubor pozorování $(x^*_{ri}, x^*_{rj}),\ r = 1, \ldots, n$. Jestliže dvojice veličin (U_i, U_j) má rozdělení $N_2(\mathbf{0}, \mathbf{I}_2)$, které nazveme v tomto kontextu kruhovým normálním rozdělením [1], potom úhel v radiánech mezi vektorem vybíhajícím z počátku k bodu (u_i, u_j) a osou u_i je rozdělen rovnoměrně na intervalu $(0, 2\pi)$.

Principem bivariátního testu je transformace dvojice veličin na kruhové normální rozdělení a pak provedení testu rozdělení úhlu podobně jako v jednorozměrném případě. Testová procedura [13], s. 53, je následující.

V prvním kroce pro každý pár (i, j) ortonormálních souřadnic, kde $i, j = 1, \ldots, D-1,\ i < j$, vypočítáme maximálně věrohodné odhady parametrů μ_i, σ_i^2, a σ_{ij}, tj.

$$\begin{pmatrix} \hat{\mu}_i \\ \hat{\mu}_j \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{r=1}^{n} x^*_{ri} \\ \frac{1}{n}\sum_{r=1}^{n} x^*_{rj} \end{pmatrix},$$

$$\begin{pmatrix} \hat{\sigma}_i^2 & \hat{\sigma}_{ij} \\ \hat{\sigma}_{ji} & \hat{\sigma}_j^2 \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{r=1}^{n}(x^*_{ri} - \hat{\mu}_i)^2 & \frac{1}{n}\sum_{r=1}^{n}(x^*_{ri} - \hat{\mu}_i)(x^*_{rj} - \hat{\mu}_j) \\ \frac{1}{n}\sum_{r=1}^{n}(x^*_{ri} - \hat{\mu}_i)(x^*_{rj} - \hat{\mu}_j) & \frac{1}{n}\sum_{r=1}^{n}(x^*_{rj} - \hat{\mu}_j)^2 \end{pmatrix}.$$

V druhém kroce pro $r = 1, \ldots, n$ vypočteme

$$u_r = \frac{1}{\sqrt{\hat{\sigma}_i^2\hat{\sigma}_j^2 - \hat{\sigma}_{ij}^2}} \left[(x^*_{ri} - \hat{\mu}_i)\hat{\sigma}_j - (x^*_{rj} - \hat{\mu}_j)\frac{\hat{\sigma}_{ij}}{\hat{\sigma}_j} \right],$$

$$v_r = \frac{(x^*_{rj} - \hat{\mu}_j)}{\hat{\sigma}_j}.$$

Třetím krokem je výpočet radiánového úhlu $\hat{\theta}$ požadovaného k rotaci osy u_r proti směru hodinových ručiček k bodu (u_r, v_r). Jestliže $\arctan(t)$ označíme funkci

arkustangens definovanou na $\mathbb{R}$ s hodnotami v intervalu $(-\frac{\pi}{2}, \frac{\pi}{2})$, pak

$$\hat{\theta} = \arctan\left(\frac{v_r}{u_r}\right) + \left(\frac{(1 - sgn(u_r))\pi}{2} + \frac{(1 + sgn(u_r))(1 - sgn(v_r))\pi}{2}\right),$$

kde sgn je znaménková funkce signum definovaná na $\mathbb{R}$, která přiřazuje záporným číslům hodnotu -1, kladným číslům hodnotu 1 a nule hodnotu 0.

Čtvrtým krokem je uspořádání hodnot $\frac{\hat{\theta}}{2\pi}$ vzestupně, uspořádané hodnoty označíme jako $z_{(r)}$.

V pátém kroce vypočítáme hodnotu testového kritéria Anderson-Darlingova "A" nebo Cramer-von Misesova "CM" nebo Watsonova "W" pro rozdělení úhlů v kruhovém normálním rozdělení, tj.

$$A = -\frac{1}{n}\sum_{r=1}^{n}(2r-1)[\ln z_{(r)} + \ln(1 - z_{(n+1-r)})] - n,$$

$$CM = \left(\sum_{r=1}^{n}\left(z_{(r)} - \frac{2r-1}{2n}\right)^2 - \frac{3.8}{12n} + \frac{0.6}{n^2}\right)\left(\frac{n+1}{n}\right),$$

$$W = \left(\sum_{r=1}^{n}\left(z_{(r)} - \frac{2r-1}{2n}\right)^2 - \frac{0.2}{12n} + \frac{0.1}{n^2} - n\left(\sum_{r=1}^{n} z_{(r)} - \frac{1}{2}\right)^2\right)\left(\frac{n+0.8}{n}\right).$$

Šestým krokem je porovnání hodnoty testového kritéria s kritickou hodnotou v tabulce. Nulovou hypotézu zamítáme na hladině významnosti α ve prospěch alternativy, jestliže hodnota testového kritéria je větší než kritická hodnota.

Hladina významnosti	%	10	5	2.5	1
Anderson - Darling	A	1.933	2.492	3.070	3.857
Cramer - von Mises	CM	0.347	0.461	0.581	0.743
Watson	W	0.152	0.187	0.221	0.267

Tab.: Kritické hodnoty pro bivariátní testy.

4.3 Radius test

Radius testem ověřujeme souhrnnou mnohorozměrnou normalitu. Za platnosti nulové hypotézy, že náhodný výběr kompozic v ilr souřadnicích, tj. $\mathbf{x}_r^*$,

$r = 1, \ldots, n$, pochází z mnohorozměrného normálního rozdělení, bude mít příslušná Mahalanobisova vzdálenost rozdělení χ^2_{D-1}. Užijeme distribuční funkci tohoto rozdělení a dostaneme hodnoty, které by měly vést k rovnoměrnému rozdělení.

Pricipem radius testu je redukce proměnných vhodnou transformací a následné univariátní testování. Testová procedura [13], s. 54, je následující.

V prvním kroce vypočítáme maximálně věrohodné odhady vektoru středních hodnot a varianční matice zcela analogicky jako v předchozím, tj. pro složky

$$\hat{\mu}_i = \frac{1}{n}\sum_{r=1}^{n} x^*_{ri},\ i = 1, \ldots, D-1,$$

$$\hat{\sigma}_{ij} = \frac{1}{n}\sum_{r=1}^{n}(x^*_{ri} - \hat{\mu}_i)(x^*_{rj} - \hat{\mu}_j),\ i, j = 1, \ldots, D-1.$$

V druhém kroce vypočítáme hodnoty Mahalanobisových vzdáleností u_r, tj.

$$u_r = (\mathbf{x}^*_r - \hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}^*_r - \hat{\boldsymbol{\mu}}),\ r = 1, \ldots, n.$$

Třetím krokem je výpočet hodnoty distribuční funkce F rozdělení χ^2_{D-1} v bodě u_r, tj.

$$z_r = F(u_r),\ r = 1, \ldots, n.$$

Ve čtvrtém kroce uspořádáme hodnoty z_r vzestupně, uspořádané hodnoty označíme jako $z_{(r)}$.

V pátém kroce vypočítáme hodnotu testového kritéria Anderson-Darlingova "A" nebo Cramer-von Misesova "CM" nebo Watsonova "W" pro radius test, tj.

$$A = -\frac{1}{n}\sum_{r=1}^{n}(2r-1)[\ln z_{(r)} + \ln(1 - z_{(n+1-r)})] - n,$$

$$CM = \left(\sum_{r=1}^{n}\left(z_{(r)} - \frac{2r-1}{2n}\right)^2 - \frac{3.8}{12n} + \frac{0.6}{n^2}\right)\left(\frac{n+1}{n}\right),$$

$$W = \left(\sum_{r=1}^{n}\left(z_{(r)} - \frac{2r-1}{2n}\right)^2 - \frac{0.2}{12n} + \frac{0.1}{n^2} - n\left(\sum_{r=1}^{n} z_{(r)} - \frac{1}{2}\right)^2\right)\left(\frac{n+0.8}{n}\right).$$

Šestým krokem je porovnání hodnoty testového kritéria s kritickou hodnotou v tabulce. Nulovou hypotézu zamítáme na hladině významnosti α ve prospěch alternativy, jestliže hodnota testového kritéria je opět větší než kritická hodnota.

Hladina významnosti	%	10	5	2.5	1
Anderson - Darling	A	1.933	2.492	3.070	3.857
Cramer - von Mises	CM	0.347	0.461	0.581	0.743
Watson	W	0.152	0.187	0.221	0.267

Tab.: Kritické hodnoty pro radius test.

4.4 Principy testování normality kompozičních dat v R

Postup ověřování platnosti nulové hypotézy v softwaru R se v některých krocích, tak jako je tomu u většiny statistických softwarů, liší od „učebnicového“ postupu. Výpočet testových statistik je v principu stejný, avšak rozhodování o platnosti nulové hypotézy se zakládá na užití p-hodnoty. p-hodnota testu je pravděpodobnost, s jakou testovací statistika nabývá hodnot více svědčících v neprospěch testované hypotézy. p-hodnota je obvyklým výstupem softwaru, udává mezní hladinu významnosti, při které bychom hypotézu ještě nezamítli. Jinými slovy, nulovou hypotézu zamítáme na hladině významnosti α právě tehdy, když p-hodnota je menší než α.

Generování p-hodnot v rámci testování normality kompozičních dat v R, v souladu s knihovnou `robCompositions` [16], se zakládá na metodě Monte Carlo. Předností této metody, nazývané také jako metoda statistických pokusů, je jednoduchost výpočtového algoritmu spočívající v provedení jednoho náhodného výběru, který se R-krát opakuje nezávisle na ostatních pokusech. Přesnost výpočtu je dána počtem pokusů, chyba klesá se zvyšujícím se počtem provedených pokusů a je úměrná číslu $\sqrt{\frac{k}{R}}$, kde konstanta k vyjadřuje povahu řešeného problému. Konkrétně R-krát náhodně vygenerujeme náhodný výběr o rozsahu n (v bivariátním případě datovou matici o rozměru $n \times 2$, v případě D-složkové kompozice matici ortonormálních souřadnic o rozměru $n \times (D-1)$) z normálního rozdělení s parametry, které byly standardně odhadnuty z dat. Z nově vygenero-

vaných dat vypočítáme hodnoty testových statistik a tyto hodnoty porovnáme s hodnotou testové statistiky vypočtené z původních dat. p-hodnotu získáme jako podíl hodnot testových statistik, které nabyly hodnot vyšších než je hodnota testové statistiky vypočtená z původních dat.

Otázkou zůstává volba počtu Monte Carlo simulací. Je-li v našem zájmu snížit chybu při výpočtu p-hodnoty desetkrát, je nutné navýšit počet simulací o jedno sto. S rostoucím počtem simulací bude narůstat počet algoritmem zpracovaných operací, a to při velkém souboru dat nemusí být zcela efektivní z hlediska výpočtové zátěže a rychlosti zpracování dat. Při volbě $R < 1$ test neproběhne a testová procedura nás vyzve k navýšení počtu simulací.

Testování neproběhne ani v případě malého souboru dat, tj. pro $n < 8$, testová procedura výpočet zastaví a informuje o problému. Určitá komplikace může z matematického hlediska vyvstat v případě radius testu, konkrétně se jedná o Mahalanobisovu vzdálenost v rámci výpočtu testového kritéria. Singulární, neinvertovatelná, varianční matice ukazuje především na nekorektní data. Singularita nenastane, pokud je soubor dat získán alespoň „předvídatelně korektním" způsobem.

Knihovna `robCompositions` poskytuje užitečné nástroje k testování normality kompozičních dat. Pro účely testování `robCompositions` nabízí uživateli funkci `ilr` transformující kompoziční data a funkce `adtest`, `adtestWrapper`, `summary.adtestWrapper`, `print.adtestWrapper`, které provedou Anderson-Darlingův test na datové matici kompozic a shrnou jeho výsledky.

Funkce `adtest` obsahuje výpočtové procedury univariátních, bivariátních testových statistik, statistiky radius testu a procedury generující p-hodnotu. Argumentem funkce `adtest` je datová matice. Parametry jsou `locscatt` s hodnotou `"standard"` pro standardní odhady parametrů rozdělení a `R`, který udává počet Monte Carlo simulací. Funkci `adtest` není nutné samostatně volat, je volaná i s parametry prostřednictvím funkce `adtestWrapper`.

Funkce `adtestWrapper` *ilr* transformuje datovou matici, zavolá funkci `adtest`, provede $D-1$ univariátních testů, poté $\frac{1}{2}(D-1)(D-2)$ bivariátních testů a ko-

nečně radius test v případě kompozic, které mají čtyři a více složek, a generuje informace o výsledcích testování. Vstupním argumentem funkce `adtestWrapper` je datová matice, parametr `alpha` udává hladinu významnosti, `R` je počet Monte Carlo simulací a parametr `robustEst` je volbou mezi standardními a robustními odhady parametrů s logickou hodnotou `FALSE` pro volbu standardních odhadů.

Funkce `summary.adtestWrapper` shrne výsledky testování, tj. vytiskne název testu, hladinu významnosti, označení testovaných proměnných v kolonce `ilrVars`, názvy dílčích testů, hodnoty odpovídajících testových statistik, p-hodnoty a konečně výsledky dílčích testů v kolonce `check` s logickou hodnotou `TRUE` anebo `FALSE`. Dílčí závěry testování jsou užitečné k odkrytí vlivu jednotlivých proměnných na zamítnutí nulové hypotézy.

Funkce `print.adtestWrapper` tiskne rozhodnutí o platnosti nulové hypotézy.

Knihovna `robCompositions` nenabízí funkce, které provedou Cramer-von Misesův a Watsonův test. Funkce `cvmtest`, `cvmtestWrapper`, `summary.cvmtestWrapper`, `print.cvmtestWrapper` a funkce `watstest`, `watstestWrapper`, `summary.watstestWrapper`, `print.watstestWrapper`, které pracují zcela analogicky včetně parametrů, odhadů p-hodnot na principu Monte-Carlo, otázky neotestování, se strukturou, jakou mají funkce Anderson-Darlingova testu, nalezneme v příloze. Tyto funkce načteme příkazem `source("cesta k hledanému R-souboru")`. Například funkci `cvmtestWrapper` načteme jako

```
> source("D:\\R-ko balicky\\cvmtest\\cvmtestWrapper.R")
```

Provedeme-li Anderson-Darlingův test na datové matici z kapitoly 2.4, tj.

```
> Normality_test= adtestWrapper(X, alpha = 0.05, R = 1000,
robustEst = FALSE)
> summary(Normality_test)

  ------------------------------------------------
  Anderson-Darling test results ( alpha = 0.05 ):
  ----------------
```

```
  ilrVars                      testName  testStat pvalue check
1       1 A-D univariate normality test 0.1530720  0.987  TRUE
2       2 A-D univariate normality test 0.4210118  0.399  TRUE
3       3  A-D bivariate normality test 0.3339600  0.687  TRUE

  ------------------------------------------------

  --> p-values and tests are obtained from standard estimates.
> Normality_test
[1] "The data follow the normal distribution on the simplex
(alpha =0.05)"
```

nebo jednoduše

```
> summary(adtestWrapper(X, alpha = 0.05, R = 1000, robustEst =
FALSE))
> print(adtestWrapper(X, alpha = 0.05, R = 1000, robustEst =
FALSE))
```

a dále Cramer-von Misesův test

```
> Normality_test= cvmtestWrapper(X, alpha = 0.05, R = 1000,
robustEst = FALSE)
> summary(Normality_test)

  ------------------------------------------------
  Cramer - von Mises test results ( alpha = 0.05 ):
  ----------------
  ilrVars                       testName   testStat pvalue check
1       1 C-vM univariate normality test 0.01872976  0.989  TRUE
2       2 C-vM univariate normality test 0.06184996  0.391  TRUE
3       3  C-vM bivariate normality test 0.04168836  0.473  TRUE
```

```
  ------------------------------------------------

  --> p-values and tests are obtained from standard estimates.
> Normality_test
[1] "The data follow the normal distribution on the simplex
(alpha =0.05)"
```

a konečně Watsonův test,

```
> Normality_test= watstestWrapper(X, alpha = 0.05, R = 1000,
robustEst = FALSE)
> summary(Normality_test)

  ------------------------------------------------
  Watson test results ( alpha = 0.05 ):
  ----------------
  ilrVars                          testName   testStat pvalue check
1       1 Watson univariate normality test 0.01862105  0.980  TRUE
2       2 Watson univariate normality test 0.05684297  0.396  TRUE
3       3  Watson bivariate normality test 0.03350805  0.396  TRUE

  ------------------------------------------------

  --> p-values and tests are obtained from standard estimates.
> Normality_test
[1] "The data follow the normal distribution on the simplex
(alpha =0.05)"
```

platnost nulové hypotézy, tj. data pocházejí z normálního rozdělení na simplexu, je na hladině významnosti 5% ověřena.

Všimněme si p-hodnot, které jsou si přibližně v konkrétních marginálních univariátních případech u všech tří testů rovny. Podstatně se liší p-hodnoty v bivariátních případech. Nabízí se otázka, jestli je nutné k ověření normality na simplexu vždy použít všechny tři testové procedury. Testy se vzájemně liší v jejich síle, tedy v přísnosti zamítání nulové hypotézy. Tato skutečnost je snadno pozorovatelná z p-hodnot, které se nejvíce blíží k hladině významnosti u Watsonova testu, nejmírnějším testem je Anderson-Darlingův test. Testy, také hladinu významnosti, je účelné volit tak, aby co nejvíce odpovídaly povaze řešeného případu. K ujištění se o normalitě je vhodné, nikoliv nezbytné, testovat data s různými volbami ilr transformace. Všechny testy mohou fungovat při odlišných volbách ortonormální báze na simplexu různě dobře, zároveň dokreslují z hlediska testování celkový pohled na konkrétní data.

4.5 Finální návrhy k testování normality na simplexu

Cílem této podkapitoly je demonstrovat dva finální přístupy k testování normality kompozičních dat včetně jejich softwarového zpracování, ukázat výhodné vlastnosti a určitá úskalí těchto přístupů, nalézt efektivní procedury k testování normality kompozičních dat a předvést tyto procedury na konkrétních datech včetně průběžného porovnání dílčích výsledků testování. Oba dále uvedené přístupy se zakládají na užití singulárního rozkladu transformované datové matice a liší se především typem použité logratio transformace.

4.5.1 Testování normality kompozičních dat užitím clr transformace a singulárního rozkladu

Standardně testujeme normalitu ilr transformovaných dat. Tento přístup k testování normality na simplexu se vrací k clr transformaci a aditivně-logistické normalitě. Hlavní myšlenka spočívá v tom, že se datová matice clr transformuje, transformovaná matice se vhodně rozloží na součin tří matic tak, že se původní informace, kterou nesou data, očistí tímto rozkladem od vzájemných vazeb mezi proměnnými. Následné testování normality se provede pouze pro matici, která

nese očištěnou informaci. Testová procedura [2] je následující.

Nechť $\mathbf{X}$ je datová matice typu $n \times D$, jejíž prvky tvoří D-složkové kompozice, tj.

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}.$$

Prvním krokem je sestavení matice $\mathbf{Y}$ typu $n \times D$ z *clr* transformované a následně centrované datové matice $\mathbf{X}$. Maticově lze tento postup zapsat jako

$$\mathbf{Y} = (\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n) \ln \mathbf{X} (\mathbf{I}_D - \frac{1}{D}\mathbf{J}_D),$$

kde $\mathbf{I}_n$ je n-rozměrná jednotková matice a $\mathbf{J}_n$ je matice typu $n \times n$, jejíž prvky jsou rovny jedné. Matice $(\mathbf{I}_D - \frac{1}{D}\mathbf{J}_D)$ provede *clr* transformaci, která transformuje data ze simplexu do euklidovského prostoru. V euklidovském prostoru jsou transformovaná data necentrovaná, centrování dat v euklidovském prostoru provede matice $(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n)$.

Druhým krokem je singulární rozklad matice $\mathbf{Y}$. *Singulárním rozkladem* [10] se rozumí rozklad původní matice na součin tří matic s výhodnými vlastnostmi, tj.

$$\mathbf{Y}_{(n\times D)} = \mathbf{Z}_{(n\times D)} \mathbf{U}_{(D\times D)} \mathbf{V}'_{(D\times D)}, \ n > D.$$

Matice $\mathbf{Z}$ a $\mathbf{V}$ jsou ortonormální, matice $\mathbf{Z}$ je složena z vlastních vektorů matice $\mathbf{Y}\mathbf{Y}'$, matice $\mathbf{V}$ je složena z vlastních vektorů matice $\mathbf{Y}'\mathbf{Y}$ a z ortonormality plyne rovnost $\mathbf{Z}'\mathbf{Z} = \mathbf{V}'\mathbf{V} = \mathbf{I}_{(D\times D)}$. Matice $\mathbf{U}$ je diagonální, na hlavní diagonále jsou uspořádány singulární hodnoty dle klesající velikosti. *Singulární hodnotou* se rozumí odmocněné nenulové vlastní číslo matice $\mathbf{Y}'\mathbf{Y}$ nebo také matice $\mathbf{Y}\mathbf{Y}'$, je jich nejvýše $D-1$. D-tá singulární hodnota je, až na numerickou chybu, rovna nule. Matice $\mathbf{Z}$ se nazývá *matice skórů*. Informaci o významnosti jednotlivých skórů, sloupců matice $\mathbf{Z}$, nám dávají singulární hodnoty, a to podle jejich vzdálenosti od nuly. Singulární hodnoty blízké nule ukazují na menší významnost skórů a vyšší přítomnost experimentálních chyb. Singulární hodnoty taktéž ukazují významnost odpovídajících *zátěží*, sloupců matice $\mathbf{V}$. Singulární rozklad lze zapsat

jako lineární kombinaci

$$\mathbf{Y} = \sum_{i=1}^{D-1} u_i \mathbf{z}_i \mathbf{v}_i'.$$

Třetím krokem je přeškálování matice $\mathbf{Z}$ z důvodu lepší interpretace [2], s. 672, konkrétně násobíme číslem $\sqrt{n}$, tj.

$$\mathbf{W} = \mathbf{Z}\sqrt{n}.$$

Na rozhodování o platnosti nulové hypotézy nemá škálování vliv.

Čtvrtým krokem je provedení testů pouze pro získanou přeškálovanou matici skórů. Normalita se testuje postupně univariátně pro první sloupec matice $\mathbf{W}$, potom univariátně pro druhý sloupec a bivariátně pro první a druhý sloupec, potom univariátně pro třetí sloupec, bivariátně pro první a třetí, druhý a třetí sloupec, radius test pro první, druhý a třetí sloupec a tak dále až do chvíle, kdy bude otestováno tolik sloupců, kolik je potřeba k ujištění se o normalitě z hlediska informace v datovém souboru vysvětlené příslušnými skóry, o jejichž významnosti nám poskytují informaci singulární hodnoty. Uvedený postup vychází z metody hlavních komponent, k této metodě blíže [13], kapitola 8.5. U konkrétního testu užijeme zároveň Anderson-Darlingovo, Cramer-von Misesovo a Watsonovo kritérium a testujeme až do 99% vysvětlené informace [2], s. 673.

Pátým krokem je rozhodnutí o nulové hypotéze. Platí ekvivalence [2], s. 673, že náhodná kompozice $\mathbf{x}$ má aditivně-logistické normální rozdělení na simplexu právě tehdy, když složky náhodného vektoru $\mathbf{w}$ (testované sloupce matice $\mathbf{W}$) jsou nekorelované, za platnosti nulové hypotézy nezávislé, a mají normální rozdělení $N(0, 1)$. Nezávislost plyne z ortogonalizačních vlastností singulárního rozkladu.

Testujeme tedy aditivně-logistickou normalitu. Pokud má náhodná kompozice $\mathbf{x}$ aditivně-logistické normální rozdělení na simplexu, pak má tato kompozice normální rozdělení na simplexu, protože aditivně-logistické normální rozdělení je jen transformací normálního rozdělení na simplexu, tedy normalita bude zachována.

4.5.2 Testování normality kompozičních dat užitím clr transformace a singulárního rozkladu v R

V praxi se samozřejmě nestává, že bychom zadávali datovou matici způsobem demonstrovaným v kapitole 2.4. Nejčastěji se datové matice načítají z datových souborů. Součástí knihovny `robCompositions` je datový soubor `skyeLavas` obsahující 23 vzorků lávy. Jedná se o trojsložkové kompozice, které tvoří sloučenina oxidů sodíku a draslíku, oxid železa a oxid hořčíku. Data jsou vyjádřena v procentech.

Datový soubor načteme příkazem `data` a příkazem `data.frame` vytvoříme datovou tabulku. Po načtení datového souboru je možné data otevřít v datovém editoru příkazem `fix`.

```
> data(skyeLavas)
> fix(skyeLavas)
> X=data.frame(skyeLavas)
> X
   sodium.potassium iron magnesium
1                52   42         6
2                52   44         4
3                47   48         5
4                45   49         6
5                40   50        10
6                37   54         9
7                27   58        15
8                27   54        19
9                23   59        18
10               22   59        19
11               21   60        19
12               25   53        22
13               24   54        22
14               22   55        23
```

15	22	56	22
16	20	58	22
17	16	62	22
18	17	57	26
19	14	54	32
20	13	55	32
21	13	52	35
22	14	47	39
23	24	56	20

Datová tabulka je v prostředí R výstupním objektem typu seznam, tuto skutečnost snadno ověříme příkazem `mode`. V našem zájmu je, aby se tabulka chovala jako datová matice, to zajistíme příkazem `as.matrix`. Diagonální matici vytvoříme příkazem `diag`. Dále provedeme *clr* transformaci užitím funkce `clr`, která je součástí `robCompositions`, a následně data vycentrujeme. Výstupem funkce `clr` je seznam, jehož první složkou je transformovaná matice dat `x.clr` a druhou složkou vektor příslušných geometrických průměrů `gm`. Přístup ke složkám seznamu jménem zajistí symbol `$`, symbol `%*%` užijeme k maticovému násobení. Bez užití `as.matrix` nastanou určité komplikace při zpracování výpočtu *clr* transformace a následujících maticových operací.

```
> mode(X)
[1] "list"
> X=as.matrix(X)
> mode(X)
[1] "numeric"
> Y=(diag(1,nrow=23)-(1/23)*matrix(1,nrow=23,ncol=23))%*%
(clr(X)$x.clr)
> Y
      sodium.potassium        iron   magnesium
 [1,]       0.92231239 -0.07569389 -0.84661850
```

```
 [2,]       1.04196075  0.09047449 -1.13243525
 [3,]       0.87117837  0.10779960 -0.97897797
 [4,]       0.77454134  0.07526698 -0.84980832
 [5,]       0.51900988 -0.04227875 -0.47673113
 [6,]       0.47650201  0.07013596 -0.54663797
 [7,]       0.07235311  0.05252708 -0.12488019
 [8,]       0.01737651 -0.07390849  0.05653198
 [9,]      -0.10101398  0.05659707  0.04441692
[10,]      -0.14867090  0.05339192  0.09527898
[11,]      -0.18528662  0.08010333  0.10518328
[12,]      -0.07656797 -0.10958405  0.18615202
[13,]      -0.11001334 -0.08351530  0.19352864
[14,]      -0.18895456 -0.05709600  0.24605056
[15,]      -0.18014347 -0.03026641  0.21040989
[16,]      -0.25538070  0.02489786  0.23048284
[17,]      -0.42637353  0.14373996  0.28263356
[18,]      -0.41361410  0.01179165  0.40182245
[19,]      -0.59424216 -0.02874762  0.62298977
[20,]      -0.64976385  0.00818780  0.64157605
[21,]      -0.66093808 -0.05907590  0.72001398
[22,]      -0.61390526 -0.18724716  0.80115242
[23,]      -0.09036583 -0.02750014  0.11786597
```

Singulární rozklad reálné datové matice provede funkce `svd`. Funkce `svd` generuje seznam nekorespondující s výše uvedenou teorií, proto složky seznamu přejmenujeme příkazem `list`.

```
> svd=list(Z=svd(Y)$u, U=svd(Y)$d, V=svd(Y)$v)
> svd
$Z
              [,1]        [,2]        [,3]
```

```
 [1,] -0.350359777 -0.36450463 -0.15364502
 [2,] -0.433874536  0.07649930  0.39474415
 [3,] -0.369714954  0.15317493 -0.30305476
 [4,] -0.324253616  0.07909772  0.29127358
 [5,] -0.197225223 -0.20423912 -0.06602124
 [6,] -0.204636964  0.11478288 -0.09353291
 [7,] -0.040130062  0.13362069 -0.04080916
 [8,]  0.009076135 -0.20797447  0.03883420
 [9,]  0.027924161  0.17395206 -0.03603845
[10,]  0.047566812  0.17304167 -0.17535946
[11,]  0.056348693  0.25328736  0.02478828
[12,]  0.054146634 -0.29122299  0.04473683
[13,]  0.061806917 -0.21329144 -0.51693411
[14,]  0.087481514 -0.12676675  0.08663137
[15,]  0.078175157 -0.05380216 -0.20789055
[16,]  0.096159815  0.11185692  0.02617454
[17,]  0.138446314  0.47023331  0.13047674
[18,]  0.161911534  0.10199692 -0.01953584
[19,]  0.242501251  0.01977323 -0.18824834
[20,]  0.256589244  0.13155281  0.09151685
[21,]  0.275580592 -0.05320302 -0.05435599
[22,]  0.284601041 -0.41670052  0.45698219
[23,]  0.041879318 -0.06116469  0.03614113

$U
[1] 3.552110e+00 4.279342e-01 1.117279e-15

$V
            [,1]       [,2]      [,3]
[1,] -0.68550605 -0.4435630 0.5773503
```

```
[2,] -0.04138379  0.8154471 0.5773503
[3,]  0.72688984 -0.3718842 0.5773503
```

Matice **U** je diagonální a poskytuje informaci o významnosti konkrétních skórů. Zřejmě první dva skóry nesou podstatnou informaci o datech, třetí singulární hodnota je (až na numerickou chybu) nulová. Matici **W** získáme přeškálováním matice **Z**.

```
> W=(svd$Z)*sqrt(23)
> W
             [,1]        [,2]        [,3]
 [1,] -1.68026646 -1.74810282 -0.73685561
 [2,] -2.08078918  0.36687777  1.89312643
 [3,] -1.77309063  0.73460117 -1.45339959
 [4,] -1.55506571  0.37933932  1.39689900
 [5,] -0.94585894 -0.97949641 -0.31662675
 [6,] -0.98140440  0.55047935 -0.44856809
 [7,] -0.19245702  0.64082231 -0.19571386
 [8,]  0.04352762 -0.99741053  0.18624229
 [9,]  0.13391957  0.83424479 -0.17283435
[10,]  0.22812242  0.82987872 -0.84099441
[11,]  0.27023884  1.21472352  0.11888042
[12,]  0.25967813 -1.39665639  0.21455030
[13,]  0.29641556 -1.02290983 -2.47912890
[14,]  0.41954660 -0.60795200  0.41546947
[15,]  0.37491488 -0.25802609 -0.99700807
[16,]  0.46116627  0.53644695  0.12552868
[17,]  0.66396520  2.25515974  0.62574444
[18,]  0.77650044  0.48916004 -0.09369062
[19,]  1.16299515  0.09482907 -0.90280731
[20,]  1.23055879  0.63090512  0.43889941
```

```
[21,]  1.32163809 -0.25515272 -0.26068217
[22,]  1.36489865 -1.99842551  2.19160960
[23,]  0.20084615 -0.29333557  0.17332677
```

Aditivně-logistickou normalitu budeme testovat na prvních dvou přeškálovaných skórech. K testování užijeme funkce `adtest`, `cvmtest` a `watstest`.

```
> adtest(W[,1], R = 1000, locscatt = "standard")

        A-D univariate normality test

data:
A = 1.0786, p-value = 0.011

> cvmtest(W[,1], R = 1000, locscatt = "standard")

        C-vM univariate normality test

data:
CM = 0.1828, p-value = 0.008

> watstest(W[,1], R = 1000, locscatt = "standard")

        Watson univariate normality test

data:
W = 0.1686, p-value = 0.013
```

Test prvního skóru vede k zamítnutí nulové hypotézy na hladině významnosti 5%. Všimněme si, že funkce `adtest`, `cvmtest` a `watstest` nevyžadují zadávání parametru `alpha`. Posouzení platnosti nulové hypotézy plyne z porovnání hladiny významnosti, kterou si zvolíme, a příslušné p-hodnoty.

Porovnáme-li p-hodnoty s hladinou významnosti 1%, nulovou hypotézu v případě testu Anderson-Darlingova a Watsonova nezamítneme a testujeme dále druhý skór. V případě testu Cramer-von Mises bychom nulovou hypotézu na hladině významnosti 1% zamítli, ale záleží na konkrétní vygenerované p-hodnotě, která je u tohoto testu velmi blízká hladině významnosti. Proto je vhodné k ujištění se o nutnosti dalšího testování, tj. zamítnutí či nezamítnutí nulové hypotézy, užít příslušný test opakovaně.

Na první pohled se může někomu jevit užití nižší hladiny významnosti, tedy „přísnějšího" testování, ve prospěch zamítání nulové hypotézy. Je tomu právě naopak, nulovou hypotézu je obtížnější zamítnout. Hladina významnosti je pravděpodobností chyby I. druhu, kdy nulovou hypotézu zamítáme, přestože platí. Volbou nižší hladiny významnosti klesá pravděpodobnost chyby I. druhu a roste pravděpodobnost chyby II. druhu, kdy nulovou hypotézu nezamítáme, přestože neplatí. Pravděpodobnost správného zamítnutí neplatné hypotézy je síla testu proti alternativě.

Při hladině významnosti 5% nemá smysl testovat další skór a můžeme přímo prohlásit závěr o neplatnosti nulové hypotézy. Při volbě hladiny významnosti 1% pokračujeme v testování druhého skóru.

```
> adtest(W[,2], R = 1000, locscatt = "standard")

        A-D univariate normality test

data:
A = 0.5325, p-value = 0.227

> cvmtest(W[,2], R = 1000, locscatt = "standard")

        C-vM univariate normality test

data:
```

```
CM = 0.087, p-value = 0.188

> watstest(W[,2], R = 1000, locscatt = "standard")

        Watson univariate normality test

data:
W = 0.0851, p-value = 0.165
```

Testy druhého skóru vypovídají ve prospěch nulové hypotézy na hladině významnosti 1%. Dále testujeme skóry bivariátně.

```
> adtest(matrix(c(W[,1],W[,2]), nrow=23, ncol=2), R = 1000,
locscatt = "standard")

        A-D bivariate normality test

data:
A = 0.3041, p-value = 0.743

> cvmtest(matrix(c(W[,1],W[,2]), nrow=23, ncol=2), R = 1000,
locscatt = "standard")

        C-vM bivariate normality test

data:
CM = 0.0361, p-value = 0.533

> watstest(matrix(c(W[,1],W[,2]), nrow=23, ncol=2), R = 1000,
locscatt = "standard")
```

```
        Watson bivariate normality test

data:
W = 0.0456, p-value = 0.184
```

Bivariátní testy potvrzují platnost nulové hypotézy na hladině významnosti 1%. Testovaná data pocházejí z aditivně-logistického normálního rozdělení.

Nyní srovnejme výše uvedený postup testování prostřednictvím funkce `adtest` s testováním téhož rozdělení prostřednictvím funkce `alnadtestWrapper`. Srovnat můžeme i s funkcemi `alncvmtestWrapper` a `alnwatstestWrapper`. Všechny tři `Wrapper`-funkce jsou zároveň s pomocnými funkcemi součástí přílohy a použijí se až na vybrané skóry z matice $\mathbf{W}$.

```
> Normality_test= alnadtestWrapper(matrix(c(W[,1],W[,2]), nrow=23,
ncol=2), alpha = 0.01, R = 1000, robustEst = FALSE)
> summary(Normality_test)

  ------------------------------------------------
  Anderson-Darling test results ( alpha = 0.01 ):
  ----------------
  Vars                       testName  testStat pvalue check
1    1 A-D univariate normality test 1.0785511  0.013  TRUE
2    2 A-D univariate normality test 0.5325345  0.235  TRUE
3    3  A-D bivariate normality test 0.3041238  0.748  TRUE

  ------------------------------------------------

  --> p-values and tests are obtained from standard estimates.
> Normality_test
[1] "The data follow the additive logistic normal distribution
on the simplex (alpha =0.01)"
```

Tímto způsobem testování skórů dospějeme ke stejnému závěru o platnosti nulové hypotézy. Funkce `alnadtestWrapper`, `alncvmtestWrapper` a `alnwatstestWrapper` užijeme k testování dvou a více skórů. Procedura testující jeden skór jako vektor nahlásí chybu a v případě, že skór vybereme jako matici rozměru $n \times 1$, užije radius test a vytiskne zcela chybný závěr.

```
> Normality_test= alnadtestWrapper(as.matrix(W[,1]), alpha = 0.05,
R = 1000, robustEst = FALSE)
> summary(Normality_test)

  ------------------------------------------------
  Anderson-Darling test results ( alpha = 0.05 ):
  ----------------
  Vars        testName  testStat pvalue check
1    1 A-D radius test 0.3869447  0.686  TRUE

  ------------------------------------------------

  --> p-values and tests are obtained from standard estimates.
```

Z praktického hlediska se zdá užití univerzálních funkcí jednodušší, nicméně otestovány budou všechny složky bez ohledu na dílčí závěry, nikoliv všechny kombinace skórů. Výhodou je naopak jednoduché zadávání testových parametrů, odkrytí jednotlivých vlivů složek na rozhodování o platnosti nulové hypotézy a přehledná tabulka.

Porovnáme-li ve výše uvedeném příkladě hodnoty testových kritérií u testů prvního skóru s kritickými hodnotami pro marginální univariátní testy, nulovou hypotézu zamítneme na hladině významnosti 1% v případě testu Cramer-von Mises i v případě Watsonova testu, což bylo vzhledem k účelu této podkapitoly záměrně opomenuto.

4.5.3 Testování normality kompozičních dat užitím ilr transformace a singulárního rozkladu

Nejprve shrneme několik důvodů proč aplikovat přístup k testování normality založený na *ilr* transformaci a singulárním rozkladu.

Zavedení *ilr* transformace odstranilo nedostatky plynoucí z užití *alr* a *clr* transformací, konkrétně k otázce vhodnosti *clr* transformace lze poukázat na její singularitu, *clr* je transformací z S^D do R^D. Tyto problémy řeší *ilr* transformace, která je ve shodě s testovanou hypotézou.

Singulární rozklad poskytuje informaci o struktuře datové matice, data očistí od vazeb, tj. kovariance mezi různými souřadnicemi budou rovny nule, a generuje skóry spolu se singulárními hodnotami. Singulární rozklad není závislý na konkrétní volbě *ilr* transformace. Rozložíme-li tedy *ilr* transformovanou matici, dostaneme při každé volbě ortonormální báze v *ilr* transformaci totožné matice.

K přístupu založenému na testování normality *clr* transformovaných dat se nabízí řada otázek. Účelem testování normality je prokázání této vlastnosti, tedy ověření nulové hapotézy, že náhodná kompozice pochází z normálního rozdělení na simplexu S^D. Ve spojení s definicí normálního rozdělení na simplexu a s ekvivalencí hypotéz bude platnost nulové hypotézy ověřena právě tehdy, když vektor ortonormálních souřadnic bude mít mnohorozměrné normální rozdělení na $\mathbb{R}^{D-1}$. Není nutné ověřovat aditivně-logistickou normalitu. Z hlediska ověřování normality jako vlastnosti kompozičních dat nejsou podstatné parametry normálního rozdělení, centrování dat je v tomto smyslu krokem navíc. Stejně tak škálování může mít účel pouze tam, kde se zpracovávají data velmi blízká nule z důvodu snadnější interpretace nebo z důvodu zamezení chyb vzniklých zaokrouhlováním. Poslední skóry obsahují rušivé elementy, nicméně i další skóry budou ovlivněny daty „neočištěnými“ od odlehlých hodnot nebo jiných artefaktů, které se vymykají převládající struktuře datového souboru. Bude tedy ovlivněn samotný singulární rozklad, kterému by mělo kompletní očištění dat od náhodných i nenáhodných vlivů předcházet.

Nechť $\mathbf{X}$ je datová matice typu $n \times D$, jejíž prvky tvoří D-složkové kompozice,

tj.

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}.$$

Testová procedura je následující.

Prvním krokem je *ilr* transformace datové matice,

$$\mathbf{X}^* = \begin{pmatrix} (\mathbf{x}_1^*)' \\ \vdots \\ (\mathbf{x}_n^*)' \end{pmatrix} = ilr(\mathbf{X}) = \begin{pmatrix} (ilr(\mathbf{x}_1))' \\ \vdots \\ (ilr(\mathbf{x}_n))' \end{pmatrix}.$$

Druhým krokem je singulární rozklad transformované datové matice,

$$\mathbf{X}^*_{(n\times(D-1))} = \mathbf{Z}_{(n\times(D-1))}\mathbf{U}_{((D-1)\times(D-1))}\mathbf{V}'_{((D-1)\times(D-1))},\ n > D-1.$$

Třetím krokem je provedení testů. Normalita se testuje, obdobně jako u clr přístupu, postupně univariátně pro první sloupec matice $\mathbf{Z}$, potom univariátně pro druhý sloupec a bivariátně pro první a druhý sloupec, potom univariátně pro třetí sloupec, bivariátně pro první a třetí, druhý a třetí sloupec, radius test pro první, druhý a třetí sloupec a tak dále až do chvíle, kdy budou otestovány všechny významné skóry. Posouzení významnosti skórů předpokládá relevantní míru subjektivního vlivu, anebo můžeme testovat 99% informace v datové matici vysvětlené příslušnými skóry.

Čtvrtým krokem je rozhodnutí o nulové hypotéze. Platí ekvivalence, že náhodná kompozice $\mathbf{x}$ má normální rozdělení na simplexu právě tehdy, když složky náhodného vektoru $\mathbf{z}$ (testované sloupce matice $\mathbf{Z}$) jsou za platnosti nulové hypotézy nezávislé a mají normální rozdělení $N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, D-1$.

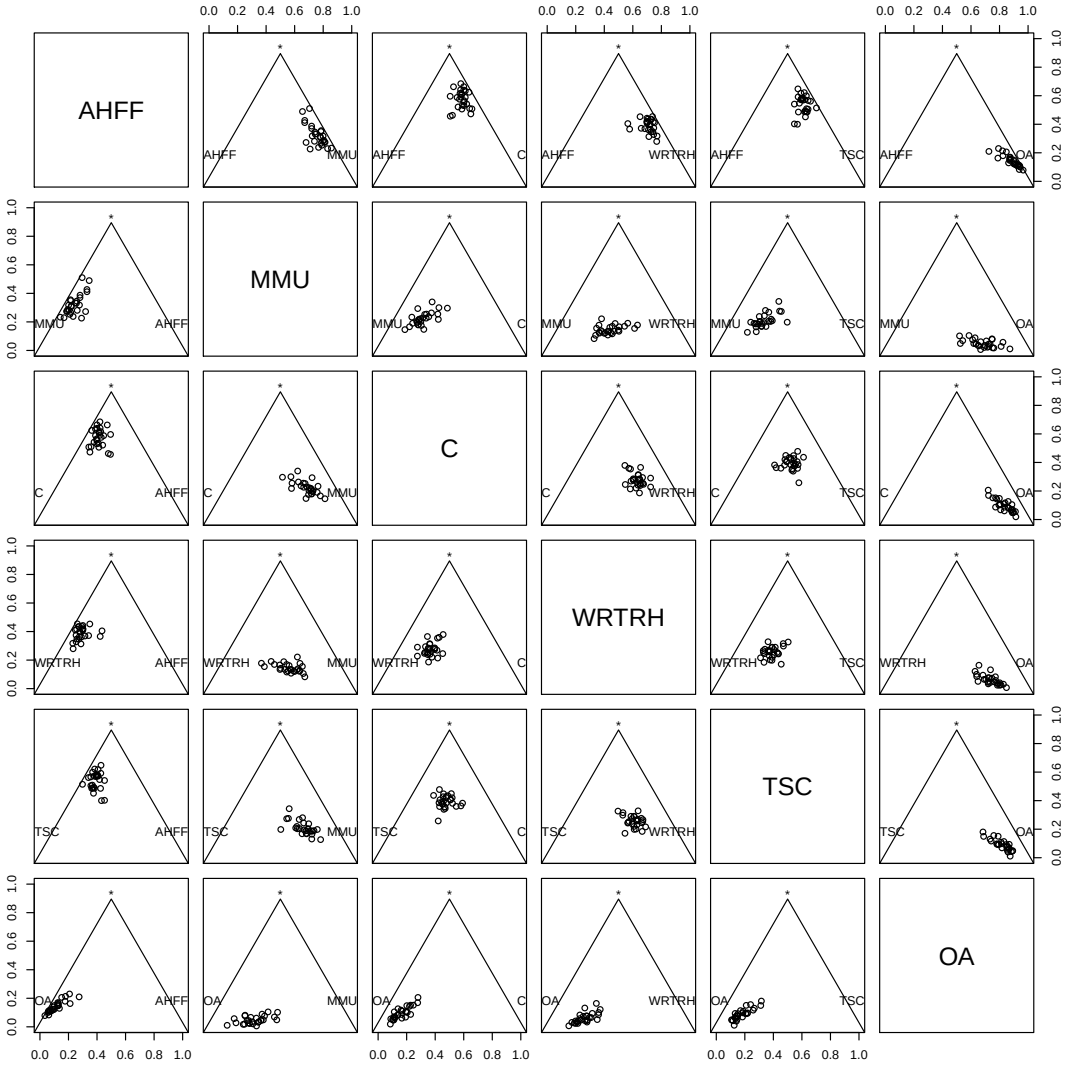
4.5.4 Testování normality kompozičních dat užitím ilr transformace a singulárního rozkladu v R

Součástí přílohy je datový soubor *GDPcomposition_EU2008*, data jsou dostupná ze stránek United Nations Statistics Division http://unstats.un.org pod heslem National Accounts, obsahující podíly složek v procentech na celkovém

hrubém domácím produktu v zemích Evropské unie za rok 2008. HDP zemí tvoří vždy šest složek, tj. AHFF (zemědělství, myslivost, lesnictví, rybolov), MMU (těžba, výroba, služby), C (stavební průmysl), WRTRH (velkoobchod, maloobchod, restaurace a hotely), TSC (doprava, skladování a spoje), OA (další aktivity).

```
> X=data.frame(read.table("GDPcomposition_EU2008.txt", header =
FALSE, row.names=c(1),col.names=c("EU","AHFF","MMU","C","WRTRH",
"TSC","OA")))
> X
```

	AHFF	MMU	C	WRTRH	TSC	OA
Austria	1.6724455	23.62040	7.243717	16.93010	6.308534	44.22480
Belgium	0.7679442	17.86381	5.311831	14.64812	8.370353	53.03794
Bulgaria	7.2935684	21.91616	8.625474	11.63737	11.910835	38.61659
Cyprus	2.0833452	10.21047	9.380986	19.84015	6.893338	51.59171
CzechRep	2.3327952	31.26360	6.294895	14.95094	10.457527	34.70024
Denmark	1.1121613	20.48297	5.791791	13.38915	8.039203	51.18472
Estonia	2.6153257	20.60632	8.355977	15.41115	10.295573	42.71565
Finland	3.0183540	24.89956	6.746108	11.70340	9.854682	43.77789
France	1.9990079	13.76618	6.682267	12.24289	6.412128	58.89753
Germany	0.8753009	25.96547	4.193032	12.07563	5.783017	51.10755
Greece	3.2907581	13.59769	6.126512	23.27585	9.898483	43.81071
Hungary	4.3057211	24.88625	4.559623	13.85435	8.329342	44.06472
Ireland	1.7317811	24.19112	9.971594	12.84698	5.258424	46.00011
Italy	2.0130278	20.84147	6.155723	14.75504	7.348548	48.88619
Latvia	3.1146258	13.79290	8.913145	18.98970	10.761646	44.42798
Lithuania	4.3989204	22.24593	9.976555	18.03437	12.748770	32.59545
Luxembourg	0.4091497	9.71299	6.195695	11.88007	9.569553	62.23255
Malta	2.5697177	17.91853	3.940834	18.14942	9.660617	47.76088
Netherlands	1.7795908	19.68344	5.778850	14.18298	6.774381	51.80076
Poland	4.5052431	23.05299	7.995867	20.03768	7.217026	37.19119
Portugal	2.3545565	17.56475	6.378360	17.39116	6.942675	49.36850
Romania	8.2509648	27.61570	8.627225	13.88047	11.532557	30.09309
Slovakia	3.4180854	28.14946	8.689013	17.97172	8.181395	33.59032
Slovenia	2.2861925	25.08994	8.929076	14.62433	7.725079	41.34538
Spain	2.7706243	17.32580	11.562637	17.68700	6.770972	43.88296
Sweden	1.5534342	22.83479	5.112606	12.31471	7.126501	51.05795
UnitedKingdom	0.9040331	17.65076	5.867503	13.84962	6.846569	54.88152



Obr.: Složky HDP v ternárních diagramech.
(Třetí souřadnice je geometrickým průměrem ostatních složek.)

Provedeme *ilr* transformaci datové matice, tj.

```
> ilrX=ilr(X)
> ilrX
           [,1]        [,2]      [,3]        [,4]      [,5]
 [1,] 2.0139928 -0.49372982 0.7274263 -0.01102165 1.3770072
 [2,] 2.6752450 -0.24175384 1.0883647  0.29682877 1.3055393
 [3,] 0.7106392 -0.35974451 0.6123354  0.49916032 0.8317227
 [4,] 1.7808009  0.40398038 0.6193726 -0.04143672 1.4232685
 [5,] 1.7610559 -0.74488705 0.8890115  0.19780429 0.8481212
 [6,] 2.3476712 -0.38185730 0.9655869  0.33920599 1.3089333
 [7,] 1.6729277 -0.25894642 0.7079584  0.25152052 1.0061079
 [8,] 1.4838192 -0.54190671 0.8083070  0.46839030 1.0544251
 [9,] 1.7339714 -0.03364543 0.7911326  0.37726199 1.5680856
[10,] 2.4713564 -0.76330405 1.1199946  0.28841899 1.5407904
[11,] 1.4033769  0.13253253 1.0917114 -0.09086519 1.0518190
[12,] 1.0892148 -0.62743573 1.1495976  0.26464514 1.1779515
[13,] 1.9684341 -0.53724160 0.3297716  0.15605606 1.5335819
[14,] 1.8132738 -0.39238831 0.9016583  0.20446211 1.3399618
[15,] 1.5053042  0.17992892 0.7364603  0.11515222 1.0025933
[16,] 1.2629548 -0.26531201 0.5834626  0.10004080 0.6637880
[17,] 3.1822711  0.35650186 0.9794061  0.58777176 1.3239101
[18,] 1.5648708 -0.25469839 1.4199137  0.13757453 1.1300623
[19,] 1.8923194 -0.36946788 0.9381888  0.22717717 1.4384368
[20,] 1.1466828 -0.42085647 0.6793551 -0.16440731 1.1593928
[21,] 1.6669032 -0.20520477 0.9047738  0.05106218 1.3870787
[22,] 0.6210460 -0.59002272 0.5817334  0.24025573 0.6782012
[23,] 1.4348661 -0.59997136 0.5827548 -0.06593161 0.9987001
[24,] 1.7758006 -0.50343627 0.5430501  0.16373183 1.1861636
[25,] 1.6014701 -0.08810281 0.3532414 -0.02102037 1.3214988
[26,] 2.0019681 -0.55317095 1.0139566  0.35729564 1.3923930
[27,] 2.5016016 -0.25860224 0.9378750  0.27451001 1.4717925
```

a singulární rozklad *ilr* transformované matice, tj.

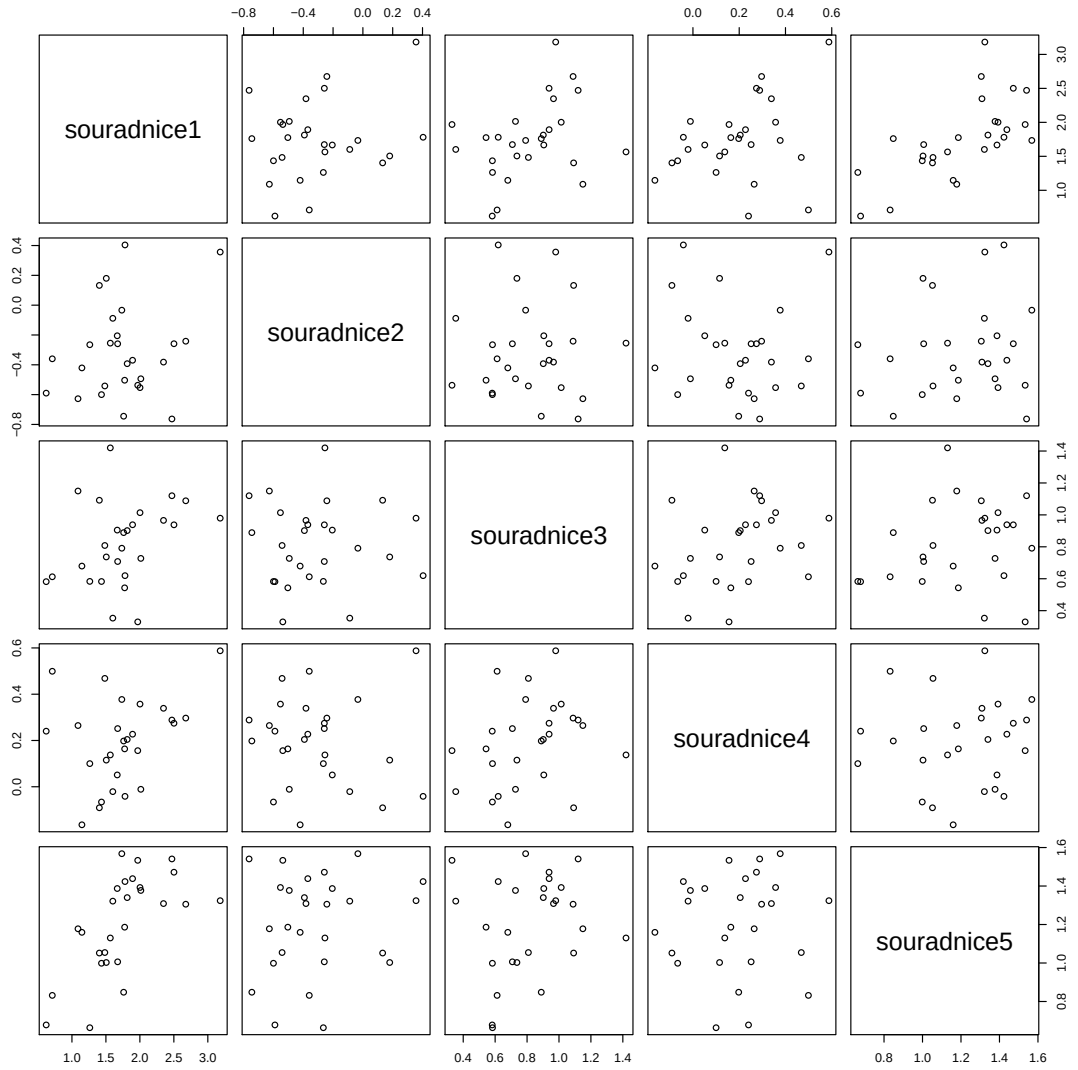
```
> svd=list(Z=svd(ilrX)$u, U=svd(ilrX)$d, V=svd(ilrX)$v)
> svd
$Z
             [,1]        [,2]        [,3]        [,4]          [,5]
 [1,] -0.20912501 -0.04423739 -0.16417998 -0.15827607  0.221981883
```

```
 [2,] -0.25732847  0.15185685 -0.05817715  0.25176199  0.150067873
 [3,] -0.10364216 -0.20723365  0.11196042  0.05555790 -0.501384350
 [4,] -0.18412192  0.33855371  0.19993160 -0.36226952 -0.005801803
 [5,] -0.17972752 -0.20229856 -0.16516894  0.27839385  0.242927868
 [6,] -0.23519818  0.03614582 -0.09398372  0.16652706  0.004174515
 [7,] -0.17116253  0.02158565 -0.04055863  0.08117218 -0.037302151
 [8,] -0.16853217 -0.17811475 -0.03736873  0.15150927 -0.274300603
 [9,] -0.19941273  0.07199106  0.14842186 -0.20457247 -0.422935991
[10,] -0.26049371 -0.16145707 -0.15458771  0.10764830  0.093797015
[11,] -0.16069155  0.08045448  0.44909206 -0.05675230  0.277122156
[12,] -0.15806327 -0.37078632  0.27881872  0.01895074 -0.111894166
[13,] -0.20318980 -0.02128545 -0.39940759 -0.32356925 -0.138145423
[14,] -0.20039274 -0.07318098  0.02902456 -0.04585415 -0.023348218
[15,] -0.15594162  0.18764644  0.20474825 -0.03689741  0.001993251
[16,] -0.12668076 -0.01437433 -0.03672714  0.10751673  0.136661853
[17,] -0.28274898  0.55155069 -0.04705690  0.38470552 -0.164508102
[18,] -0.18883492 -0.13509324  0.45150965  0.15860060  0.186350570
[19,] -0.21041549 -0.06113267  0.04896781 -0.07204473 -0.064554614
[20,] -0.14255381 -0.16730786  0.06859057 -0.29621235  0.224398300
[21,] -0.19032737 -0.01921396  0.16006231 -0.18499405  0.053019040
[22,] -0.09128356 -0.30892897  0.01906832  0.03880751 -0.144889850
[23,] -0.15365426 -0.15874997 -0.17085841 -0.10581212  0.271568504
[24,] -0.18233758 -0.05456005 -0.24262067 -0.07748662  0.004578823
[25,] -0.16624014  0.12664764 -0.11945132 -0.37226359 -0.019382764
[26,] -0.22025545 -0.13506799 -0.02027013  0.07367040 -0.106154421
[27,] -0.24916874  0.11877781 -0.08431863  0.05196138  0.017039367

$U
[1] 12.2945583  1.8146159  1.2502743  1.1751273  0.8011705

$V
            [,1]        [,2]        [,3]        [,4]       [,5]
[1,] -0.77002724  0.39403550 -0.36795818  0.26164523  0.2189580
[2,]  0.12422762  0.84421753  0.49430118 -0.09954172 -0.1327480
[3,] -0.34698275 -0.32289957  0.76972081  0.37387051  0.2075802
[4,] -0.08523425 -0.04044202 -0.04918475  0.47287045 -0.8746851
[5,] -0.51377391 -0.16165731  0.15932312 -0.74715969 -0.3553475
```

Užijeme-li jinou *ilr* transformaci, například z knihovny `compositions`,

```
> list(Z=svd(compositions::ilr(X))$u, U=svd(compositions::ilr(X))$d,
V=svd(compositions::ilr(X))$v)
```

singulární rozklad se nezmění.



Obr.: Složky HDP v souřadnicích.
(Použitá *ilr* transformace pochází z knihovny `robCompositions`.)

Pětirozměrnou normalitu testujeme na prvních čtyřech skórech, otestováno tak bude 95.3785% informace v datové matici. K testování užijeme funkci `cvmtest`. Testujeme postupně všechny kombinace vybraných skórů.

```
> cvmtest(svd$Z[,1], R = 1000, locscatt = "standard")
```

```
        C-vM univariate normality test

data:
CM = 0.0298, p-value = 0.853

> cvmtest(svd$Z[,2], R = 1000, locscatt = "standard")

        C-vM univariate normality test

data:
CM = 0.0618, p-value = 0.352

> cvmtest(matrix(c(svd$Z[,1],svd$Z[,2]), nrow=27, ncol=2), R = 1000,
locscatt = "standard")

        C-vM bivariate normality test

data:
CM = 0.0421, p-value = 0.507

> cvmtest(svd$Z[,3], R = 1000, locscatt = "standard")

        C-vM univariate normality test

data:
CM = 0.0664, p-value = 0.334

> cvmtest(matrix(c(svd$Z[,1],svd$Z[,3]), nrow=27, ncol=2), R = 1000,
locscatt = "standard")

        C-vM bivariate normality test

data:
CM = 0.0245, p-value = 0.744

> cvmtest(matrix(c(svd$Z[,2],svd$Z[,3]), nrow=27, ncol=2), R = 1000,
locscatt = "standard")

        C-vM bivariate normality test

data:
CM = 0.1224, p-value = 0.139
```

```
> cvmtest(matrix(c(svd$Z[,1],svd$Z[,2],svd$Z[,3]), nrow=27, ncol=3),
R = 1000, locscatt = "standard")

        C-vM radius test

data:
= 0.0349, p-value = 0.735

> cvmtest(svd$Z[,4], R = 1000, locscatt = "standard")

        C-vM univariate normality test

data:
CM = 0.0332, p-value = 0.801

> cvmtest(matrix(c(svd$Z[,1],svd$Z[,4]), nrow=27, ncol=2), R = 1000,
locscatt = "standard")

        C-vM bivariate normality test

data:
CM = 0.0368, p-value = 0.583

> cvmtest(matrix(c(svd$Z[,2],svd$Z[,4]), nrow=27, ncol=2), R = 1000,
locscatt = "standard")

        C-vM bivariate normality test

data:
CM = 0.0325, p-value = 0.622

> cvmtest(matrix(c(svd$Z[,3],svd$Z[,4]), nrow=27, ncol=2), R = 1000,
locscatt = "standard")

        C-vM bivariate normality test

data:
CM = 0.0637, p-value = 0.314

> cvmtest(matrix(c(svd$Z[,1],svd$Z[,2],svd$Z[,4]), nrow=27, ncol=3),
R = 1000, locscatt = "standard")

        C-vM radius test
```

```
data:
= 0.0576, p-value = 0.499

> cvmtest(matrix(c(svd$Z[,1],svd$Z[,3],svd$Z[,4]), nrow=27, ncol=3),
R = 1000, locscatt = "standard")

        C-vM radius test

data:
= 0.0361, p-value = 0.743

> cvmtest(matrix(c(svd$Z[,2],svd$Z[,3],svd$Z[,4]), nrow=27, ncol=3),
R = 1000, locscatt = "standard")

        C-vM radius test

data:
= 0.142, p-value = 0.103

> cvmtest(matrix(c(svd$Z[,1],svd$Z[,2],svd$Z[,3],svd$Z[,4]), nrow=27,
ncol=4), R = 1000, locscatt = "standard")

        C-vM radius test

data:
= 0.0697, p-value = 0.396
```

Cramer-von Mises testy potvrzují platnost nulové hypotézy na hladině významnosti 5%. Testovaná data pocházejí z normálního rozdělení na simplexu.

Doporučuje se [2] testovat skóry zároveň testem Anderson-Darlingovým a Watsonovým. Při velkém počtu testovaných skórů exponenciálně roste počet testů, a to vede k neefektivnímu testování. Účelem testování je ověření mnohorozměrné normality (radius test), každá odchylka od mnohorozměrné normality se projeví na marginální normalitě (univariátní testy), už bivariátní testy jsou dokreslujícími testy, a proto z praktického hlediska není nutné testovat všechny kombinace skórů. Srovnejme s funkcemi `scsadtestWrapper`, `scscvmtestWrapper` a `scswatstestWrapper`, které testují normalitu u vybraných skórů z matice $\mathbf{Z}$ a které jsou i s pomocnými funkcemi součástí přílohy.

```
> Normality_test= scsadtestWrapper(matrix(c(svd$Z[,1],svd$Z[,2],
svd$Z[,3],svd$Z[,4]), nrow=27, ncol=4), alpha = 0.05, R = 1000,
robustEst = FALSE)
> summary(Normality_test)
```

```
  ------------------------------------------------
  Anderson-Darling test results ( alpha = 0.05 ):
  ----------------
   Vars                       testName  testStat pvalue check
1     1 A-D univariate normality test 0.2198469  0.872  TRUE
2     2 A-D univariate normality test 0.4796890  0.299  TRUE
3     3 A-D univariate normality test 0.4696080  0.323  TRUE
4     4 A-D univariate normality test 0.2506100  0.821  TRUE
5   1 2  A-D bivariate normality test 0.4213909  0.485  TRUE
6   1 3  A-D bivariate normality test 0.2901587  0.769  TRUE
7   1 4  A-D bivariate normality test 0.4110933  0.491  TRUE
8   2 3  A-D bivariate normality test 0.7320082  0.185  TRUE
9   2 4  A-D bivariate normality test 0.3242455  0.681  TRUE
10  3 4  A-D bivariate normality test 0.4589947  0.439  TRUE
11  all               A-D radius test 0.4830687  0.480  TRUE

  ------------------------------------------------

  --> p-values and tests are obtained from standard estimates.

> Normality_test= scscvmtestWrapper(matrix(c(svd$Z[,1],svd$Z[,2],
svd$Z[,3],svd$Z[,4]), nrow=27, ncol=4), alpha = 0.05, R = 1000,
robustEst = FALSE)
> summary(Normality_test)

  ------------------------------------------------
  Cramer - von Mises test results ( alpha = 0.05 ):
  ----------------
   Vars                        testName   testStat pvalue check
1     1 C-vM univariate normality test 0.02976524  0.866  TRUE
2     2 C-vM univariate normality test 0.06180743  0.352  TRUE
3     3 C-vM univariate normality test 0.06641304  0.313  TRUE
4     4 C-vM univariate normality test 0.03324761  0.776  TRUE
5   1 2  C-vM bivariate normality test 0.04208721  0.515  TRUE
6   1 3  C-vM bivariate normality test 0.02453625  0.752  TRUE
7   1 4  C-vM bivariate normality test 0.03675613  0.603  TRUE
8   2 3  C-vM bivariate normality test 0.12240647  0.121  TRUE
9   2 4  C-vM bivariate normality test 0.03248694  0.614  TRUE
10  3 4  C-vM bivariate normality test 0.06366651  0.331  TRUE
```

```
11  all               C-vM radius test 0.06973922  0.393  TRUE

  ------------------------------------------------

  --> p-values and tests are obtained from standard estimates.

> Normality_test
[1] "The data follow the normal distribution on the simplex
(alpha =0.05)"
```

Dle testu Anderson-Darlingova a Cramer-von Misesova data pocházejí z normálního rozdělení na simplexu na hladině významnosti 5%.

```
> Normality_test= scswatstestWrapper(matrix(c(svd$Z[,1],svd$Z[,2],
svd$Z[,3],svd$Z[,4]), nrow=27, ncol=4), alpha = 0.05, R = 1000,
robustEst = FALSE)
> summary(Normality_test)

  ------------------------------------------------
  Watson test results ( alpha = 0.05 ):
  ----------------
   Vars                          testName   testStat pvalue check
1     1 Watson univariate normality test 0.02948763  0.839  TRUE
2     2 Watson univariate normality test 0.05186895  0.463  TRUE
3     3 Watson univariate normality test 0.06000805  0.352  TRUE
4     4 Watson univariate normality test 0.03197060  0.804  TRUE
5   1 2  Watson bivariate normality test 0.04770174  0.164  TRUE
6   1 3  Watson bivariate normality test 0.02963812  0.568  TRUE
7   1 4  Watson bivariate normality test 0.02796677  0.600  TRUE
8   2 3  Watson bivariate normality test 0.03493553  0.409  TRUE
9   2 4  Watson bivariate normality test 0.03306016  0.486  TRUE
10  3 4  Watson bivariate normality test 0.07277691  0.022 FALSE
11  all               Watson radius test 0.07372841  0.393  TRUE

  ------------------------------------------------

  --> p-values and tests are obtained from standard estimates.
> Normality_test
[1] "The data do not follow the normal distribution on the simplex
(alpha =0.05)"
```

Watsonův test zamítá nulovou hypotézu ve prospěch alternativy na hladině významnosti 5%. Porovnáme-li ale hodnoty vypočtených testových kritérií s kritickými hodnotami v tabulkách, dospějeme k závěru, že nulovou hypotézu nelze zamítnout ani v případě Watsonova testu.

Normalita se především ověřuje testy, ale je možné užít i grafické metody. Normalitu lze vypozorovat na základě rozložení dat v grafu nebo dle charakteristik jako je šikmost a špičatost. K tomuto účelu nám slouží například histogram, boxplot, N-P plot, Q-Q plot a P-P plot [9], s. 98 - 106. Všechny uvedené nástroje se vzájemně doplňují a dokreslují celkovou představu problému. V zásadě jsou ale testy nejužitečnějším a nejpraktičtějším nástrojem k ověřování předpokladu normality.

Závěr

Kompoziční data nesou celou svou informaci v podílech mezi složkami, standardní statistické metody takovou informaci nezachovají a zde nachází své opodstatnění logratio přístup k analýze kompozičních dat. Vhodnou geometrií, která respektuje vlastnosti kompozičních dat, je Aitchisonova geometrie zavedená na simplexu. Pochopení a interpretace výsledků práce na simplexu ovšem není snadné. Proto se kompoziční data nejčastěji transformují do euklidovského prostoru, kde můžeme pracovat ve smyslu standardní euklidovské geometrie. Jako nejlepší logratio transformace se jeví *ilr* díky její izometrii.

Protože jsou kompoziční data už svou povahou mnohorozměrnými daty, je testování normálního rozdělení na simplexu jako vlastnosti dat zcela zásadní. Způsobů, jak ověřovat platnost hypotézy o normalitě, je více a každý z nich vyžaduje určitou míru subjektivního rozhodování, tj. volba přístupu, volba testové procedury a testového kritéria, volba hladiny významnosti a počtu Monte Carlo simulací, výběr testovaných skórů, porovnání hladiny významnosti s p-hodnotou, zamítnutí či nezamítnutí nulové hypotézy, a posouzení konkrétní situace, konkrétních dat a vlastních očekávání.

Z praktického hlediska spočívá nejjednodušší varianta ověřování normality v *ilr* transformaci kompozičních dat a následném užití jednoho typu testového kritéria (Anderson-Darlingova, Cramer-von Misesova, Watsonova) na data v ortonormálních souřadnicích. Užití dalších typů kritérií má dokreslující význam. K tomu nabízí knihovna `robCompositions` `Wrapper`-funkci, která provede Anderson-Darlingův test a generuje přehlednou tabulku závěrů testování, nebo je možné užít jiná kritéria, která nabízejí `Wrapper`-funkce z přílohy. Nevýhodou tohoto postupu je, že testy v závislosti na konkrétní volbě ortonormální báze na simplexu nemusí dobře zafungovat, a proto je vhodné testovat normalitu dat pro různé volby ortonormální báze.

Přístup k testování normality založený na singulárním rozkladu a *clr* transformaci řeší některé problémy výše popsaného přístupu. Zvlášť významnou roli hraje singulární rozklad transformované matice, který data očistí od vzájemných vazeb a „vytáhne“ z nich podstatnou informaci, tím se sníží dimenze dat, tedy i počet nutných testů k ověření normality. Pokud je v našem zájmu testovat efektivně, tj. užít co nejmenší počet testů a zároveň obdržet korektní závěr, pak tento záměr komplikuje doporučení testovat všechny kombinace skórů a užít všechna testová kritéria. K testování v softwaru R použijeme funkce `adtest`, `cvmtest` a `watstest` nebo můžeme i přesto, že netestují všechny kombinace skórů, užít `aln...Wrapper`-funkce.

Přístup k testování normality založený na singulárním rozkladu a *ilr* transformaci je kompromisní variantou mezi předchozími přístupy a řeší jejich nedostatky. Singulární

rozklad se užije na ilr transformovaná data a na konkrétní volbě ortonormální báze na simplexu je nezávislý. ilr je transformací z D-složkového simplexu do euklidovského prostoru dimenze $D-1$, to znamená, že testujeme $(D-1)$-rozměrné normální rozdělení, které je úzce svázáno s normálním rozdělením na simplexu. Singulární rozklad poskytuje informaci o struktuře datové matice. Testovat můžeme všechny kombinace skórů anebo můžeme užít `scs...Wrapper`-funkce; přitom postačuje užití jednoho typu testového kritéria.

Standardně testujeme normalitu na hladině významnosti 5%, při porovnání hladiny významnosti s p-hodnotou je nutné sledovat vztah těchto hodnot. V případě, kdy jsou si hodnoty blízké, je dobré generovat p-hodnotu opakovaně. Zároveň není od věci sledovat hodnoty testových kritérií ve vztahu ke kritickým hodnotám uvedeným v tabulkách.

V této práci jsou vysvětleny všechny testové procedury k testování normality na simplexu včetně jejich zpracování v softwaru R ve spojení s knihovnou `robCompositions`. Přístupy byly demonstrovány na třech různých datových souborech. K práci přikládám datový disk, který obsahuje algoritmy funkcí prostředí R užité k testování.

Literatura

[1] AITCHISON, John. *The statistical analysis for compositional data.* London : Chapman and Hall, 1986. 416 s.

[2] AITCHISON, John; MATEU-FIGUERAS, Gloria; NG, Kai W. Characterization of Distributional Forms for Compositional Data and Associated Distributional Tests. *Mathematical geology.* 2007, 35, 6, s. 667-680. Dostupný také z WWW: <http://www.springerlink.com/content/m202gg85234n6652/>.

[3] ANDĚL, Jiří. *Matematická statistika.* 2. vydání. Praha : SNTL, 1985. 346 s.

[4] MARTÍN-FERNÁNDEZ, Josep A.; BARCELÓ-VIDAL, Carles; PAWLOWSKY-GLAHN, Vera. Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. *Mathematical Geology.* 2003, 35, 3, s. 253-278. Dostupný také z WWW: <http://www.springerlink.com/content/ku816485q4264772/>.

[5] BARCELÓ-VIDAL, Carles; MARTÍN-FERNÁNDEZ, Josep A.; PAWLOWSKY-GLAHN, Vera. Mathematical Foundations of Compositional Data Analysis. *Proceedings of IAMG'01 : the sixth annual conference of the International Association for Mathematical Geology* [online]. 2001, 20, [cit. 2010-11-20]. Dostupný z WWW: <http://ima.udg.edu/∼jamf/>.

[6] EGOZCUE, Juan José, et al. Isometric Logratio Transformations. *Mathematical Geology.* 2003, 35, 3, s. 279-300. Dostupný také z WWW: <http://www.springerlink.com/content/wx1166n56n685v82/>.

[7] EGOZCUE, Juan José. Reply to "On the Harker Variation Diagrams; ..." by J.A. Cortés. *Mathematical Geosciences.* 2009, 41, 7, s. 829-834. Dostupný také z WWW: <http://www.springerlink.com/content/x08r2624vg95710g/>.

[8] FILZMOSER, Peter; HRON, Karel. Outlier Detection for Compositional Data Using Robust Methods. *Mathematical Geosciences.* 2008, 40, 3, s. 233-248. Dostupný také z WWW: <http://www.springerlink.com/content/d662421553216861/>.

[9] HEBÁK, Petr, et al. *Vícerozměrné statistické metody (1).* 2. vydání. Praha : Informatorium, 2007. 253 s. ISBN 9788073330569.

[10] HOLÍK, Miroslav. *Cheminfo.chemi.muni.cz* [online]. 2004 [cit. 2010-11-20]. SINGULÁRNÍ ROZKLAD MATICE DAT. Dostupné z WWW: <http://cheminfo.chemi.muni.cz/ktfch/holik/Pomucky/>.

[11] KUNDEROVÁ, Pavla. *Základy pravděpodobnosti a matematické statistiky.* 1.vydání. Olomouc : Univerzita Palackého v Olomouci, 2004. 186 s. ISBN 80-244-0813-9.

[12] MATEU-FIGUERAS, Glória; PAWLOWSKY-GLAHN, Vera. A Critical Approach to Probability Laws in Geochemistry. *Mathematical Geosciences.* 2008, 40, 5, s. 489-502. Dostupný také z WWW: <http://www.springerlink.com/content/t78t0812246u6v91/>.

[13] PAWLOWSKY-GLAHN, Vera; EGOZCUE, Juan José; TOLOSANA-DELGADO, Raimon. *Lecture Notes on Compositional Data Analysis* [online]. Girona : Universitat de Girona, 28.5.2007 [cit. 2010-11-20]. Dostupné z WWW: <http://hdl.handle.net/10256/297>.

[14] RÉNYI, Alfréd. *Teorie pravděpodobnosti.* Praha : Academia, 1972. Charakteristická funkce, s. 262-314. ISBN 21-078-72.

[15] STEPHENS, Michael A. EDF Statistics for Goodness of Fit and Some Comparisons. *Journal of the American Statistical Association.* 1974, 69, 347, s. 730-737. Dostupný také z WWW: <http://www.jstor.org/stable/2286009>.

[16] TEMPL, Matthias; HRON, Karel; FILZMOSER, Peter. *Cran.r-project.org : Package robCompositions* [online]. 2010 [cit. 2010-11-20]. Dostupné z WWW: <http://cran.r-project.org/web/packages/robCompositions/index.html>.