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

DIPLOMOVÁ PRÁCE

Bilance v analýze kompozičních dat

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

Vypracovala:
Bc. Klára Hrůzová
AME, II. ročník

Prohlášení

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

V Olomouci dne 26. března 2012

Poděkování

Ráda bych na tomto místě poděkovala vedoucímu diplomové práce RNDr. Karlu Hronovi, Ph.D. za obětavou spolupráci i za čas, který mi věnoval při konzultacích.

Obsah

Úvod

Práce se zabývá speciálním typem mnohorozměrných dat, zvaných kompoziční data. Kompoziční data jsou specifická tím, že nesou pouze relativní informaci, proto nás zajímají poměry mezi složkami kompozičního vektoru. Pro práci s tímto typem dat byla zavedena Aitchisonova geometrie na simplexu. Tato geometrie je odlišná od euklidovské na reálném prostoru, a bylo by tedy potřeba uzpůsobit jí i statistické metody. Tento postup se ovšem zdá být poněkud složitý, z toho důvodu se začaly využívat transformace kompozičních dat do reálného prostoru. V minulosti byly hojně používané alr a clr transformace, které ovšem měly teoretické nedostatky, a proto se zavedla ilr transformace. Ilr transformace převádí kompoziční vektor o D složkách do reálného prostoru dimenze $D-1$ pomocí vyjádření kompozice v souřadnicích vzhledem k ortonormálním bázím na simplexu. Tato práce ukazuje jeden z postupů, jak najít takovouto ortonormální bázi a jí odpovídající souřadnice, v tomto případě zvané bilance. Na takto transformovaná data už je možné použít standardní statistické metody, ale je potřeba dát si pozor na interpretaci výsledků. Dále si zde ukážeme, jak je možné snížit dimenzi simplexu pomocí projekce na podkompozice. A v neposlední řadě představíme program pro práci s kompozičními daty, CoDaPack, a tvorbu bilančního dendrogramu.

Práce je rozdělena do šesti kapitol, které jsou dále děleny na podkapitoly. V první kapitole jsou představena kompoziční data - základní definice a vlastnosti. Ve druhé kapitole si přiblížíme Aitchisonovu geometrii na simplexu. Ve třetí kapitole jsou uvedeny možnosti transformace kompozičních dat. Zatímco čtvrtá kapitola se přímo týká hlavního tématu této diplomové práce - tzn. objasníme si konstrukci a interpretaci bilancí. Pátá kapitola je věnována kovarianční struktuře bilancí, tzn. jak spočítáme rozptyl, kovarianci a co tyto charakteristiky o bilancích vypovídají. V šesté kapitole se seznámíme s programem CoDaPack, který se používá při analýze kompozičních dat a předvedeme pomocí něj vytvoření

bilančního dendrogramu. Ten následně aplikujeme na reálný příklad (chemické složení minerálních vod).

1 Kompoziční data

Pojem kompoziční data (kompozice) není prozatím příliš rozšířený navzdory tomu, že se s tímto typem mnohorozměrných dat setkáváme v každodenním životě. Data jsou zvláštní tím, že jedinou relevantní informaci nám poskytuje podíl jednotlivých proměnných. Nejčastěji se s kompozičními daty setkáváme v přírodních vědách jako je geologie, geografie, chemie, biologie a podobně. Jako příklad ze života můžeme uvést podíly jednotlivých každodenních výdajů k rozpočtu domácnosti, poměr politických stran ve Sněmovně, rozdělení obyvatelstva např. podle národnosti, víry, věku. V ekonomii můžeme zmínit například úrokové sazby, složení investičního portfolia, kapitálovou přiměřenost a různé ekonomické indexy. Nyní si tedy představíme základní pojmy týkající se kompozic.

Tato kapitola vychází z literatury [1], [9], [10].

Definice 1.1. *Kompoziční vektor o D složkách, $\mathbf{x} = [x_1, x_2, \ldots, x_D]$, je definován jako vektor s kladnými složkami, kde jediná relevantní informace je obsažena v podílech mezi jednotlivými složkami.*

Tvrzením, že všechna relevantní informace je obsažena v poměrech, dojdeme k závěru, že pokud a je kladné reálné číslo, pak $[x_1, \ldots, x_D]$ a $[ax_1, \ldots, ax_D]$ vyjadřují stejnou informaci, a tudíž jsou ekvivalentní. Množina kompozic je tedy tvořena třídami ekvivalentních kompozičních vektorů.

Způsob, jak zjednodušit užití a interpretaci kompozic, je reprezentovat je v uzavřené formě, tzn. jako kladný vektor, jehož složky mají předepsaný součet, κ. Obvyklými hodnotami konstanty κ jsou 1 pro proporce, 100 pro procenta, atd. Ovšem z definice existují i kompoziční data, která jsou reprezentována v neuzavřené formě, což můžeme jednoduše „napravit“.

Definice 1.2. *Pro každý vektor složený z D kladných reálných čísel*

$$\mathbf{x} = [x_1, x_2, \ldots, x_D] \in \mathbb{R}_+^D$$

je uzávěr definovaný takto:

$$\mathcal{C}(\mathbf{x}) = \left[\frac{\kappa \cdot x_1}{\sum_{i=1}^{D} x_i}, \frac{\kappa \cdot x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa \cdot x_D}{\sum_{i=1}^{D} x_i} \right].$$

Množina kladných reálných vektorů uzavřených na konstantu κ se nazývá simplex o D složkách. Můžeme ho tedy považovat za přirozený výběrový prostor kompozičních dat. Formálně simplex definujeme takto:

Definice 1.3. *D-složkový simplex je podmnožina $\mathbb{R}^D$,*

$$\mathcal{S}^D = \left\{ \mathbf{x} = [x_1, x_2, \ldots, x_D] \,|x_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} x_i = \kappa \right\}.$$

Často se pozornost zaměřuje na skupinu složek kompozice. Poměry složek ve skupině jsou považovány za relevantní, zatímco poměry zahrnující některé složky mimo skupinu jsou ignorovány. To odpovídá definici subkompozice (podkompozice), která zahrnuje pouze složky ve skupině.

Definice 1.4. *Mějme kompozici $\mathbf{x}$, pak podkompozici $\mathbf{x}_s$ o s složkách získáme aplikací operace uzávěru na podvektor $[x_{i_1}, x_{i_2}, \ldots, x_{i_s}]$ z $\mathbf{x}$. Přitom podindexy $i_1, \ldots, i_s$ nám říkají, které složky byly do podkompozice vybrány (ne nutně prvních s složek).*

Dále nechť množinu s indexů označíme S. Potom lze uvedenou podkompozici $\mathbf{x}_s$ také označit jako sub$(\mathbf{x}; S)$.

1.1 Principy analýzy kompozičních dat

Abychom mohli na kompoziční data aplikovat standardní statistické metody, musí tyto metody splňovat tři podmínky: škálová neměnnost, neměnnost při změně pořadí a podkompoziční soudržnost.

1.1.1 Škálová neměnnost

Nejdůležitější vlastností kompozičních dat je, že nesou pouze relativní informaci. Pro analýzu užíváme nejčastěji hodnoty složek kompozice po operaci uzávěru. Uzávěr je projekce jakéhokoliv bodu v kladné části reálného prostoru dimenze D na simplex. Tuto situaci si lze intuitivně představit v případě trojrozměrného reálného prostoru ($D = 3$). Tehdy jsou všechny body na polopřímce začínající v počátku zobrazeny na ten samý bod v $\mathcal{S}^D$. Navíc pokud změníme jednotky původních dat, jednoduše vynásobíme všechny odpovídající kompoziční složky konstantou změny jednotek, posuneme je dál po jejich polopřímce do průniku s dalším rovnostranným trojúhelníkem (trojsložkovým simplexem), rovnoběžným s tím původním.

Definice 1.5. *Dva vektory o D kladných reálných složkách $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^D$ jsou kompozičně ekvivalentní, jestliže existuje kladná konstanta $\lambda \in \mathbb{R}^+$ taková, že $\mathbf{x} = \lambda \cdot \mathbf{y}$ a tedy $\mathcal{C}(\mathbf{x}) = \mathcal{C}(\mathbf{y})$.*

Přitom nám samozřejmě záleží na tom, abychom totožné výsledky statistické analýzy dostali nezávisle na konstantě λ, což nazýváme škálová neměnnost.

Definice 1.6. *Funkce $f(\cdot)$ je škálově neměnná, jestliže pro každou kladnou reálnou konstantu $\lambda \in \mathbb{R}^+$ a pro jakoukoliv kompozici $\mathbf{x} \in \mathcal{S}^D$, funkce splňuje rovnost $f(\lambda\mathbf{x}) = f(\mathbf{x})$, tzn. že dostaneme ty samé výsledky pro každý kompozičně ekvivalentní vektor.*

1.1.2 Neměnnost při změně pořadí

Zjednodušeně můžeme říci, že funkce je neměnná při změně pořadí složek kompozice, jestliže i po změně pořadí jednotlivých složek kompozice dostaneme stejnou funkční hodnotu.

1.1.3 Podkompoziční soudržnost

Podkompozice by se měly chovat jako ortogonální projekce ve standardní reálné analýze. Můžeme si to názorně představit na situaci, kdy délka úsečky v rovině po projekci je vždy menší nebo rovna délce úsečky původní.

Z tohoto principu vyplývá několik důležitých vlastností jako je např. podkompoziční dominance (vzdálenost měřená mezi dvěma kompozicemi musí být větší nebo rovna vzdálenosti mezi jakýmikoliv jim odpovídajícími podkompozicemi), nebo dále, že jestliže vynecháme složky, které neobsahují pro nás důležitou informaci, výsledky analýzy týkající se zbytku složek kompozice se nezmění.

2 Aitchisonova geometrie

Abychom s kompozičními daty mohli pracovat, potřebujeme si nejprve nadefinovat základní operace jako je sčítání, násobení skalárem, skalární součin apod. Vzhledem k povaze dat není možné použít pro jejich geometrickou reprezentaci standardní reálný prostor. Proto zavádíme tzv. Aitchisonovu geometrii na simplexu, tedy na výběrovém prostoru kompozičních dat, jak již bylo řečeno v předchozí kapitole.

Kapitola opět vychází z literatury [1], [9], [10].

Definice 2.1. *Perturbací kompozice* $\mathbf{x} \in \mathcal{S}^D$ *kompozicí* $\mathbf{y} \in \mathcal{S}^D$ *nazveme kompozici definovanou takto:*

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

Věta 2.1. *$(\mathcal{S}^D, \oplus)$ je komutativní grupa, tzn. pro $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{S}^D$ platí:*

1. *komutativita:* $\mathbf{x} \oplus \mathbf{y} = \mathbf{y} \oplus \mathbf{x}$;
2. *asociativita:* $(\mathbf{x} \oplus \mathbf{y}) \oplus \mathbf{z} = \mathbf{x} \oplus (\mathbf{y} \oplus \mathbf{z})$;
3. *neutrální prvek:*

$$\mathbf{n} = \mathcal{C}\left[1, 1, \ldots, 1\right] = \left[\frac{1}{D}, \frac{1}{D}, \ldots, \frac{1}{D}\right];$$

   kde $\mathbf{n}$ *je střed simplexu a je jediný.*

4. *inverze:* $\mathbf{x}^{-1} = \mathcal{C}\left[x_1^{-1}, x_2^{-1}, \ldots, x_D^{-1}\right]$; *tedy* $\mathbf{x} \oplus \mathbf{x}^{-1} = \mathbf{n}$.

Definice 2.2. *Mocninná transformace kompozice* $\mathbf{x} \in \mathcal{S}^D$ *konstantou* $\alpha \in \mathbb{R}$ *je kompozice definována vztahem:*

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

Věta 2.2. *Mocninná transformace má stejné vlastnosti jako vnější součin v euklidovském prostoru:*

1. *asociativita:* $\alpha \odot (\beta \odot \mathbf{x}) = (\alpha \odot \beta) \odot \mathbf{x}$*;*
2. *distributivní vlastnost 1:* $\alpha \odot (\mathbf{x} \oplus \mathbf{y}) = (\alpha \odot \mathbf{x}) \oplus (\alpha \odot \mathbf{y})$*;*
3. *distributivní vlastnost 2:* $(\alpha + \beta) \odot \mathbf{x} = (\alpha \odot \mathbf{x}) \oplus (\beta \odot \mathbf{x})$*;*
4. *neutrální prvek:* $1 \odot \mathbf{x} = \mathbf{x}$*; neutrální prvek je jediný.*

Definováním těchto dvou operací se tedy simplex, $(\mathcal{S}^D, \oplus, \odot)$, stává vektorovým prostorem.

Poznámka 2.1. *Operace uzávěru vyruší veškeré konstanty, proto konstanta* κ *není při početních operacích z matematického pohledu důležitá, což lze vyjádřit následovně:*

$$\mathbf{x} \oplus (\alpha \odot \mathbf{z}) = \mathbf{x} \oplus (\alpha \odot \mathcal{C}(\mathbf{z})).$$

Je nutné ovšem poznamenat, že konstanta uzávěru je velice důležitá pro správnou interpretaci výsledků.

Definice 2.3. *Aitchisonův skalární součin kompozic* $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ *definujeme takto:*

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

Věta 2.3. *Aitchisonův skalární součin splňuje standardní vlastnosti:*

1. *pozitivita:* $\langle \mathbf{x}, \mathbf{x} \rangle_a > 0$ *pokud* $\mathbf{x} \neq \mathbf{n}$*;*
2. *komutativita:* $\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle \mathbf{y}, \mathbf{x} \rangle_a$*;*
3. *distributivita vzhledem k perturbaci:* $\langle \mathbf{x} \oplus \mathbf{z}, \mathbf{y} \rangle_a = \langle \mathbf{x}, \mathbf{y} \rangle_a + \langle \mathbf{z}, \mathbf{y} \rangle_a$*;*
4. *linearita vzhledem k násobení skalárem:* $\langle c \odot \mathbf{x}, \mathbf{y} \rangle_a = c \cdot \langle \mathbf{x}, \mathbf{y} \rangle_a$.

Poznámka 2.2. *Díky těmto vlastnostem má simplex $\mathcal{S}^D$ strukturu euklidovského prostoru dimenze $D-1$. Tedy $\mathcal{S}^D$ je ekvivalentní s $\mathbb{R}^{D-1}$.*

Definice 2.4. *Aitchisonova norma $\mathbf{x} \in \mathcal{S}^D$ je dána následujícím vztahem,*

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

Definice 2.5. *Aitchisonova vzdálenost mezi $\mathbf{x}$ a $\mathbf{y} \in \mathcal{S}^D$ je definována následovně,*

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

Věta 2.4. *Hlavními vlastnostmi Aitchisonovy vzdálenosti jsou:*

1. *nezávisí na konstantě uzávěru;*
2. *neměnnost po perturbaci: nechť $\mathbf{p} \in \mathcal{S}^D$, pak*

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

3. *přeškálování umocněním: mějme c reálné číslo, pak*

$$d_a(c \odot \mathbf{x}, c \odot \mathbf{y}) = |c| \cdot d_a(\mathbf{x}, \mathbf{y});$$

4. *neměnnost po permutaci složek kompozice;*
5. *garantuje podkompoziční dominanci: nechť R je množina r indexů, $1 < r < D$, pak*

$$d_a(sub(\mathbf{x}; R), sub(\mathbf{y}; R)) \leq d_a(\mathbf{x}, \mathbf{y}).$$

3 Reprezentace kompozičních dat v souřadnicích

Abychom mohli pro analýzu kompozičních dat aplikovat běžné statistické metody, založené na předpokladu standardní euklidovské geometrie v reálném prostoru, využíváme transformace kompozic, které jsou již z podstaty dat založené na poměrech jejich složek. Mezi původní transformace, zavedené již na počátku 80. let 20. století zakladatelem log-ratio analýzy kompozičních dat Johnem Aitchisonem (log-ratio = logaritmus podílu), patří aditivní log-ratio (alr) a centrované log-ratio (clr) transformace ze simplexu do reálného prostoru. Mezi nimi je rozdíl v použití, protože alr-transformace není izometrická (nezachovává vzdálenost), používala se proto k modelování dat, clr transformace vede zase k singulární varianční matici, proto se při statistické analýze používala pro techniky založené na metrice.

Nedávno se začaly používat i izometrické log-ratio (ilr) transformace, díky nimž můžeme přejít ze simplexu do kartézské soustavy souřadnic (izometrie mezi $\mathcal{S}^D$ a $\mathbb{R}^{D-1}$).

Kapitola vznikla za použití zdrojů [1], [2], [9], [10], [12].

Nyní se zaměříme nejprve obecně na báze a generující systémy na simplexu a pak i na jednotlivé transformace.

Struktura vektorového prostoru $\mathcal{S}^D$ nám dovoluje použít koncept lineární závislosti a nezávislosti. Množina m kompozic v $\mathcal{S}^D$, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$, a m skalárů, $\alpha_1, \alpha_2, \ldots, \alpha_m$, je spojena v perturbačně-lineární kombinaci tak, že

$$\mathbf{x} = (\alpha_1 \odot \mathbf{x}_1) \oplus (\alpha_2 \odot \mathbf{x}_2) \oplus \cdots \oplus (\alpha_m \odot \mathbf{x}_m) = \bigoplus_{i=1}^{m} (\alpha_i \odot \mathbf{x}_i),$$

což je perturbačně-mocninná verze tradiční lineární kombinace v reálném vektorovém prostoru. Množina $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ se nazývá perturbačně závislá, pokud se perturbačně-lineární kombinace rovná neutrálnímu prvku, $\mathbf{x} = \mathbf{n}$, pro nějaké $\alpha_i \neq 0, i = 1, \ldots, m$.

Naopak, pokud identická rovnost $\mathbf{x} = \mathbf{n}$ vyjadřuje, že všechny skaláry $\alpha_i = 0$, pak množina $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ se nazývá perturbačně nezávislá.

V $\mathcal{S}^D$ je maximální počet perturbačně nezávislých kompozic $D-1$, tedy $\mathcal{S}^D$ je vektorový prostor dimenze $D-1$. Pokud $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}$ jsou perturbačně nezávislé, tvoří bázi $\mathcal{S}^D$. Což znamená, že každá kompozice $\mathbf{x} \in \mathcal{S}^D$ může být vyjádřena jako perturbačně-lineární kombinace

$$\mathbf{x} = (\alpha_1 \odot \mathbf{e}_1) \oplus (\alpha_2 \odot \mathbf{e}_2) \oplus \cdots \oplus (\alpha_{D-1} \odot \mathbf{e}_{D-1}) = \bigoplus_{i=1}^{D-1} (\alpha_i \odot \mathbf{e}_i),$$

pro nějaké koeficienty α_i, které představují souřadnice vzhledem k dané bázi.

3.1 Ortonormální báze

Jak již bylo zmíněno, $\mathcal{S}^D$ je vektorový prostor dimenze $D-1$, tedy $D-1$ perturbačně nezávislých vektorů v $\mathcal{S}^D$ tvoří bázi. Pokud kompozice $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}$ v $\mathcal{S}^D$ splňuje

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

tvoří ortonormální bázi $\mathcal{S}^D$.

3.2 Ortonormální souřadnice

Euklidovská prostorová struktura simplexu zaručuje existenci ortonormálních bází, které mohou být jednoduše získány aplikací Gram-Schmidtova ortonormalizačního procesu na jakoukoliv bázi. Přitom bázi můžeme získat z jakéhokoliv generujícího systému. Báze takto získaná však bude jen jednou z mnoha ortonormálních bází, které můžeme nadefinovat na euklidovském prostoru. Navíc není zřejmé, jak určit, která báze je nejvhodnější pro řešení daného problému.

Jestliže si tedy vybereme ortonormální bázi, můžeme kompozici $\mathbf{x} \in \mathcal{S}^D$ vyjádřit jako

$$\mathbf{x} = \bigoplus_{i=1}^{D-1} x_i^* \odot \mathbf{e}_i, x_i^* = \langle \mathbf{x}, \mathbf{e}_i \rangle_a ,$$

kde $\mathbf{x}^* = \left[x_1^*, \ldots, x_{D-1}^*\right]$ je vektor souřadnic $\mathbf{x}$ vzhledem k vybrané bázi.

Poznámka 3.1. *Souřadnice x_i^* je ortogonální projekce kompozice* $\mathbf{x}$ *na směr definovaný jednotkovou kompozicí* $\mathbf{e}_i$.

Funkce ilr: $\mathcal{S}^D \to \mathbb{R}^{D-1}$ přiřazující souřadnici $\mathbf{x}^*$ kompozici $\mathbf{x}$ se nazývá izometrická log-ratio transformace, což je izometrický izomorfismus vektorového prostoru. Pro jednoduchost se tato funkce někdy označuje jako h. Z výše uvedeného je zřejmé, že ilr transformace izometricky zobrazuje Aitchisonovu geometrii na simplexu do reálného prostoru se standardní euklidovskou geometrií.

Věta 3.1. *Uvažujme* $\mathbf{x}_k \in \mathcal{S}^D, k = 1, 2$ *a reálné konstanty* α, β, *pak*

$$h(\alpha \odot \mathbf{x}_1 \oplus \beta \odot \mathbf{x}_2) = \alpha \cdot h(\mathbf{x}_1) + \beta \cdot h(\mathbf{x}_2) = \alpha \cdot \mathbf{x}_1^* + \beta \cdot \mathbf{x}_2^*;$$

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_a = \langle h(\mathbf{x}_1), h(\mathbf{x}_2) \rangle = \langle \mathbf{x}_1^*, \mathbf{x}_2^* \rangle ;$$

$$\|\mathbf{x}_1\|_a = \|h(\mathbf{x}_1)\| = \|\mathbf{x}_1^*\| ;$$

$$d_a(\mathbf{x}_1, \mathbf{x}_2) = d(h(\mathbf{x}_1), h(\mathbf{x}_2)) = d(\mathbf{x}_1^*, \mathbf{x}_2^*).$$

Je několik způsobů, jak definovat ortonormální bázi na simplexu. Hlavním kritériem pro výběr takovéto báze je dobrá interpretovatelnost reprezentace kompozice v souřadnicích. Připomeňme si v této souvislosti, že při analýze hlavních komponent je ortonormální báze vybrána tak, aby první souřadnice (hlavní komponenta) reprezentovala směr maximální variability. Konkrétní případy, které si zaslouží pozornost, jsou báze související s postupným binárním dělením kompozičního vektoru, o kterých se zmíníme v následující kapitole. Vzhledem k interpretaci odpovídajících souřadnic tyto nejčastěji nazýváme bilance.

Konkrétním příkladem ortonormální báze na simplexu je $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$, kde

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

s příslušnými souřadnicemi

$$x_i^* = \sqrt{\frac{i}{i+1}} \ln\left(\frac{(x_1 x_2 \cdots x_i)^{1/i}}{x_{i+1}}\right), i = 1, 2, \ldots, D-1.$$

3.3 Centrovaná log-ratio transformace

Výsledkem clr transformace nejsou souřadnice vzhledem k bázi, ale vzhledem ke generujícímu systému na simplexu. Přesto se clr transformace kompozic využívá, a to pro její jednoduchost a některé její výhodné vlastnosti.

Koeficienty clr transformace kompozice $\mathbf{x}$ z $\mathcal{S}^D$ do $\mathbb{R}^D$ jsou definovány následovně,

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

kde $g(\cdot)$ značí geometrický průměr. Z definice můžeme vidět, že $\sum_{i=1}^{D} \xi_i = 0$, tedy množina clr koeficientů tvoří nadrovinu v $\mathbb{R}^D$.

Z koeficientů clr transformace můžeme kompozici získat zpět pomocí inverzní transformace:

$$\mathbf{x} = [x_1, x_2, \ldots, x_D] = \mathcal{C}\left[\exp \xi_1, \exp \xi_2, \ldots, \exp \xi_D\right].$$

Mezi hlavní vlastnosti clr koeficientů patří:

1. Převádí perturbaci a mocninnou transformaci kompozic na běžný součet vektorů a násobení vektoru skalárem.

2. Standardní euklidovská vzdálenost mezi vektory clr koeficientů je rovna Aitchisonově vzdálenosti odpovídajících kompozic, což platí také pro skalární součin a normu.

Naopak hlavní nevýhoda clr transformace plyne z výše uvedené podmínky na nulový součet clr koeficientů, totiž varianční matice clr transformovaných kompozic bude vždy singulární. Přitom podmínka regularity varianční matice je nedílnou součástí většiny statistických metod, regresní analýzou počínaje a korelační analýzou konče.

3.4 Aditivní log-ratio souřadnice

Výsledkem aditivní log-ratio transformace kompozice $\mathbf{x} = [x_1, \ldots, x_D] \in \mathcal{S}^D$ je reálný vektor

$$\mathbf{a} = [a_1, a_2, \ldots, a_{D-1}] = \left[\ln \frac{x_1}{x_D}, \ln \frac{x_2}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D}\right].$$

Speciální roli zde hraje složka x_D, která dělí všechny ostatní složky kompozice. Pozice této složky je ovšem volitelná, pro tento účel může být vybrána jakákoli složka kompozice.

Zpětnou (inverzní) transformací můžeme získat původní kompozici:

$$\mathbf{x} = \mathcal{C}\left[\exp a_1, \exp a_2, \ldots, \exp a_{D-1}, 1\right].$$

Podobně jako clr a ilr souřadnice i alr souřadnice transformují operace perturbace a mocninné transformace na simplexu na obvyklý reálný součet a násobení skalárem. Avšak metrické vlastnosti se při alr transformaci bohužel nezachovávají. Proto se více používají ostatní zmíněné log-ratio transformace.

3.5 Práce v ortonormálních souřadnicích

Princip práce v souřadnicích popsaných výše spočívá v tom, že všechny standardní statistické metody mohou být aplikovány na souřadnice kompozic vzhledem k (nějaké) ortonormální bázi.

Nejjednoduššími případy tohoto principu jsou základní operace na simplexu, jako je perturbace a mocninná transformace, o čemž jsme se zmínili již dříve z ryze geometrického hlediska. Souřadnice vzhledem k ortonormální bázi na simplexu umožňují místo nich použít standardní operace. Perturbace kompozic v $\mathcal{S}^D$ je tak ekvivalentní sčítání vektorů v reálném prostoru, mocninná transformace v $\mathcal{S}^D$ je ekvivalentní násobení skalárem. Tedy pokud zvážíme vektor souřadnic $h(\mathbf{x}) = \mathbf{x}^* \in \mathbb{R}^{D-1}$ kompozičního vektoru $\mathbf{x} \in \mathcal{S}^D$ vzhledem k libovolné ortonormální bázi, dostaneme

$$h(\mathbf{x} \oplus \mathbf{y}) = h(\mathbf{x}) + h(\mathbf{y}) = \mathbf{x}^* + \mathbf{y}^*, h(\alpha \odot \mathbf{x}) = \alpha \cdot h(\mathbf{x}) = \alpha \cdot \mathbf{x}^*,$$

můžeme tedy přemýšlet o perturbaci kompozice jako by měla stejné vlastnosti na simplexu jako součet vektorů v reálném prostoru a mocninná transformace kompozice stejné vlastnosti jako násobení vektoru skalárem.

Navíc

$$d_a(\mathbf{x}, \mathbf{y}) = d(h(\mathbf{x}), h(\mathbf{y})) = d(\mathbf{x}^*, \mathbf{y}^*),$$

kde d je obvyklá euklidovská vzdálenost. To znamená, že pokud analyzujeme kompoziční data, výsledky získané užitím kompozic a Aitchisonovy geometrie jsou stejné jako ty získané za použití souřadnic a euklidovské geometrie.

Pokud bychom ovšem na tyto operace použili alr souřadnice, stejné výsledky nedostaneme, protože jsme použili souřadnice vzhledem k bázi, která není ortonormální. Použitím clr souřadnic získáme správné výsledky, avšak operace odpovídají prostoru $\mathbb{R}^D$.

Pro shrnutí, princip práce v souřadnicích můžeme jednoduše použít na simplexu, a to tak, že místo s kompozicemi pracujeme s jejich vyjádřením v souřadnicích, aplikujeme standardní statistické metody a, pokud je to nutné pro interpretaci dosažených výsledků, vrátíme se zpět na simplex použitím zpětné transformace. Ačkoliv je tento postup jednoduchý, v některých případech je nutné dát si pozor právě na správnou interpretaci výsledků v souřadnicích, protože jsou tyto vyjádřeny v podobě logaritmů podílů skupin složek kompozic.

3.6 Odejít či zůstat na simplexu

Jedním ze způsobů, jak statisticky analyzovat kompoziční data, je transformovat kompozice do reálného prostoru užitím vhodné log-ratio transformace a potom aplikovat standardní statistické metody; po obdržení výsledků se vrátit zpět do simplexu užitím inverzní log-ratio transformace a výsledky interpretovat. Jediným rozdílem mezi log-ratio přístupem, který byl v této podobě navržený již Johnem Aitchisonem v 80. letech minulého století, a principem práce v souřadnicích je doporučené použití ortonormálních (ilr) souřadnic ve druhém přístupu. Avšak oba přístupy jsou v podstatě stejné, protože problémy kvůli singulární varianční matici v případě clr transformace nebo nerespektování Aitchisonovy metriky

(alr transformace) mohou být překonány použitím adekvátních algebraických technik. Cílem této kapitoly je srovnat oba přístupy s ohledem na reprezentaci pravděpodobnostních modelů a odpovídajících očekávání.

Log-ratio přístup byl dlouho považován za transformační techniku podobnou log-normálnímu modelu: log transformací náhodné veličiny s log-normálním rozdělením dostaneme normálně rozdělenou náhodnou veličinu. Obě náhodné proměnné předpokládají $\mathbb{R}$ jako výběrový prostor, ačkoliv u první z nich je tento omezen kladnými hodnotami. Pozornost nebyla věnována faktu, že $\mathbb{R}_+$ je euklidovský prostor s vlastními operacemi, metrikou a příslušnou mírou. Tento typ transformační úvahy je spojen s tradičním Aitchisonovým log-ratio přístupem. Alternativou je zvážit $\mathbb{R}_+$ jako prostor odlišný od $\mathbb{R}$ a pak věnovat speciální pozornost příslušné míře (referenční míra). V $\mathbb{R}$ Lebesgueova míra označuje délku $\lambda(a,b) = \|b-a\|$ intervalu (a,b), zatímco míra odpovídající $\mathbb{R}_+$ je $\lambda_+(a,b) = \|\ln b - \ln a\|$.

Pokud předpokládáme, že log-normální náhodná proměnná existuje v $\mathbb{R}$, přecházíme z $\mathbb{R}_+$ do $\mathbb{R}$, tedy vlastně odcházíme z $\mathbb{R}_+$. Jestliže uvažujeme, že náhodná proměnná má výběrový prostor $\mathbb{R}_+$ a její logaritmus je právě její souřadnicí, pak řekneme, že zůstáváme v $\mathbb{R}_+$.

Podobný koncept může být použit pro náhodné kompozice. Výběrovým prostorem je $\mathcal{S}^D$ s Aitchisonovou geometrií. To umožňuje souřadnicovou reprezentaci náhodné kompozice (přístup zůstat). Alternativně můžeme transformovat náhodnou kompozici na reálný náhodný vektor (přístup odejít). Jedním z hlavních důsledků k přístupu zůstat je nutnost užití přirozené referenční míry. Přirozenou referenční mírou v $\mathcal{S}^D$ je Aitchisonova míra. Aitchisonova míra v $\mathcal{S}^D$ je jednoduše definovaná užitím ortonormálních souřadnic. Mějme rovnoběžnostěn V v prostoru ilr souřadnic $\mathbb{R}^{D-1}$. Lebesgueovu míru tohoto útvaru, $\lambda(V)$, spočítáme jako součin délek jeho hran. Potom $U = \mathrm{ilr}^{-1}(V) \subseteq \mathcal{S}^D$ má Aitchisonovu míru $\lambda_a(U) = \lambda(V)$.

Princip práce v souřadnicích spolu s přístupem zůstat může být použit k definování parametrických modelů na simplexu, například užitím funkce hustoty ortonormálních souřadnic. V prostoru souřadnic je tato hustota klasickou husto-

tou, jakou je hustota vzhledem k Lebesgueově míře, ale na simplexu je to hustota s ohledem na Aitchisonovu míru. V plném log-ratio přístupu (přístup odejít) je náhodná kompozice transformována do reálného prostoru souřadnic (typicky alr souřadnice). Zpětná transformace z log-ratio souřadnic do simplexu je vnímána jako jednoduchá transformace reálného vektoru souřadnic do reálného vektoru na simplexu (tj. náhodné kompozice). K získání hustoty náhodné kompozice je požadován Jakobián k zajištění Lebesgueovy referenční míry, předpokládané v případě souřadnic.

Standardní statistické metody byly tradičně vyvíjeny za předpokladu euklidovské geometrie v reálném prostoru. Jestliže má výběrový prostor jinou algebraicko-geometrickou strukturu (např. Aitchisonovu geometrii v případě simplexu), pak použití standardních statistických metod není vhodné, můžeme tak dojít k matoucím, či dokonce chybným závěrům.

Teorie práce v souřadnicích na simplexu je obecně stejná jako práce pomocí log-ratio přístupu. Nicméně, když pracujeme v souřadnicích, zdůrazňujeme ortonormální souřadnice, jako jsou ilr souřadnice, navíc analytik si může dovolit vybrat nejlépe interpretovatelné souřadnice pro daný problém. V log-ratio přístupu jsou dostupné všechny log-ratio transformace (alr, clr a ilr), proto si analytik zdánlivě může vybrat nejlépe vyhovující transformaci pro každou konkrétní analýzu, pokud vezmeme v potaz jejich výhody a omezení. Avšak alr a clr vektory nejsou souřadnice vzhledem k ortonormální bázi, a tudíž má každý z nich jistá omezení. Navíc princip práce v souřadnicích je přirozeně spojen se strategií zůstat na simplexu, která používá celou strukturu Aitchisonovy geometrie na simplexu včetně Aitchisonovy míry, například k definování pravděpodobnostních hustot a momentů.

4 Skupiny složek a jejich bilance

Kapitola byla vypracována s využitím zdrojů [8], [9].

4.1 Úvod

Analýza kompozičních dat je založená na poměrech jejich složek, což je základ pro hlubší porozumění tomuto typu dat. Aitchisonova geometrie na simplexu je založena na specifických operacích jako je perturbace a mocninná transformace a definicích Aitchisonovy vzdálenosti, normy a skalárního součinu.

Statistická analýza kompozičních dat tak obvykle vyžaduje interpretaci výsledků pomocí podílů a logaritmů podílů (log-ratios), které jsou mnohem obtížnější na interpretaci než reálné vektory ve standardní mnohorozměrné analýze. Pro zjednodušení analýzy můžeme složky seřadit tak, aby mohly být rozděleny do dvou a více podmnožin, které jsou nějakým způsobem lépe interpretovatelné. Analytik se potom může zajímat o studium složek kompozice ze dvou hledisek:

1. vztah nebo lépe bilance (rovnováha) mezi skupinami složek kompozice (meziskupinová analýza),

2. chování složek ve skupině (vnitroskupinová analýza).

Skupiny složek mohou být vnímány jako podkompozice, nebo jako skupina uvnitř celé kompozice. Podkompoziční analýza je určena pro práci se složkami ve skupině a vztahy vzhledem k ostatním skupinám jsou z analýzy vyloučeny. Pro studium vztahů mezi podkompozicemi byl v literatuře [1] zaveden koncept amalgamace neboli sloučení několika složek kompozice do nové složky. Později se ale ukázalo, že nelineární charakter amalgamace vzhledem k Aitchisonově geometrii vede k problémům. Amalgamace složek uvnitř každé skupiny může být relevantní, pokud má jasný, dobře definovaný smysl a slouží například ke studiu variability nově vzniklé kompozice. Když srovnáme analýzu amalgamovaných složek kompozice s kompoziční analýzou původních složek, očekávali bychom, že obě analýzy budou zaměnitelné a jejich interpretace si bude odpovídat. Bohužel

většinou je pravdou pravý opak tohoto očekávání. Amalgamace nezachovává Aitchisonovu vzdálenost na simplexu a vzdálenost amalgamovaných kompozic tak mají složité, nemonotónní chování vzhledem ke vzdálenostem původních kompozic. V následujícím příkladu si ukážeme jeden z důvodů zavedení nového postupu pro snížení dimenze kompozice.

Příklad 4.1. *Mějme třísložkovou náhodnou kompozici* $\mathbf{x}_j = [x_{j1}, x_{j2}, x_{j3}], j = 1, 2, \ldots, J$*, její střed spočítáme jako geometrický průměr složek kompozice*

$$cen(\mathbf{x}) = \mathcal{C}\left[\left(\prod_{j=1}^{J} x_{j1}\right)^{1/J}, \left(\prod_{j=1}^{J} x_{j2}\right)^{1/J}, \left(\prod_{j=1}^{J} x_{j3}\right)^{1/J}\right].$$

Nyní na kompozici provedeme amalgamaci do dvou složek,

$$\mathcal{C}\left[\left(\prod_{j=1}^{J} x_{j1}\right)^{1/J} + \left(\prod_{j=1}^{J} x_{j2}\right)^{1/J}, \left(\prod_{j=1}^{J} x_{j3}\right)^{1/J}\right]. \tag{1}$$

Očekávali bychom, že střed nově vzniklé kompozice $\mathcal{C}[x_{j1}+x_{j2}, x_{j3}], j = 1, 2, \ldots, J$ *bude vypadat stejně jako (1). Níže si ovšem můžeme všimnout, že střed nové kompozice je odlišný:*

$$\mathcal{C}\left[\left(\prod_{j=1}^{J} (x_{j1} + x_{j2})\right)^{1/J}, \left(\prod_{j=1}^{J} x_{j3}\right)^{1/J}\right].$$

Navzdory nechtěnému chování je operace amalgamace složek kompozice v praxi často užívaná, protože je jednoduchá a očividně intuitivní způsob sloučení složek kompozice slouží především k získání nižší dimenze kompozičních dat. Pokud se ovšem zajímáme o analýzu jak celé kompozice, tak reprezentace nižší dimenze, potřebujeme alternativní a správný způsob analýzy seskupování složek uvnitř kompozice. Hlavním požadavkem je snadná interpretovatelnost a kompatibilita s Aitchisonovou geometrií na simplexu.

4.2 Ortonormální báze postupného binárního dělení

4.2.1 Postupné binární dělení

Jak již bylo zmíněno, kompoziční vektory o D složkách jsou často děleny do skupin prezentující jistý vztah. Mějme nyní kompozici $[x_1, \ldots, x_n]$ na simplexu o D složkách. Každé rozdělení složek kompozice do skupin může být vnímáno jako mezikrok při postupném binárním dělení složek kompozice. V prvním kroku binárního dělení získáme dvě skupiny složek, v druhém kroku rozdělíme jednu skupinu z prvního pořadí na dvě skupiny; postup opakujeme tak dlouho, dokud v každé skupině nebude pouze jedna složka. Počet kroků, než postupné binární dělení složek dospěje ke konci, je $D-1$. Postupné binární dělení můžeme vyjádřit v tabulce, kdy v každém kroku označíme složky první skupiny symbolem '+', složky druhé skupiny '−' a složky, jichž se daný krok dělení netýká, jako 0. Ukažme si použití tohoto postupu na následujícím příkladu:

Příklad 4.2. *Mějme kompozici složenou ze šesti složek* $[x_1, x_2, x_3, x_4, x_5, x_6]$*, kde každá složka kompozice zastupuje chemický prvek obsažený v minerální vodě Mattoni. Postupně se jedná o složky* Mg^{2+}*,* Ca^{2+}*,* Na^{+}*,* Cl^{-}*,* SO_4^{-}*,* HCO_3^{-}*.*

V prvním kroku od sebe oddělíme kationty a anionty, ve druhém kroku oddělíme prvky hořčík a vápník od sodíku z důvodu postavení v periodické tabulce prvků (Mg a Ca jsou kovy alkalických zemin). Následně oddělíme složky ve zbylé dvojici kationtů. Ve skupině aniontů nejprve oddělíme bezkyslíkatou sůl od solí kyslíkatých a následně i složky kyslíkatých solí - SO_4^{-}*,* HCO_3^{-}*. Výsledek je zobrazen v následující tabulce:*

Krok	x_1	x_2	x_3	x_4	x_5	x_6
1	+	+	+	−	−	−
2	+	+	−	0	0	0
3	+	−	0	0	0	0
4	0	0	0	+	−	−
5	0	0	0	0	+	−

4.2.2 Ortonormální báze dělení

V dalším se budeme snažit dát do souvislosti ortonormální báze $\mathcal{S}^D$ s postupným binárním dělením. Z každého postupného binárního dělení můžeme totiž získat ortonormální bázi na simplexu s odpovídajícími souřadnicemi, zvanými bilance mezi skupinami prvků. Tyto bilance (vlastně ilr souřadnice) můžeme následně použít ve vnitro- i meziskupinové analýze. Předpokládejme, že v l-tém kroku binárního dělení, $l = 1, \ldots, D-1$, oddělujeme složky $x_{k+1}, \ldots, x_{k+r}$ (r složek) od $x_{k+r+1}, \ldots, x_{k+r+s}$ (s složek). Navíc můžeme uvažovat i zbývající složky z předchozích dělení. Ty jsou reprezentovány k složkami $x_1, \ldots, x_k$, resp. j složkami $x_{k+r+s+1}, \ldots, x_D$. To znamená, že $D = k+r+s+j$, $l \leq D-r-s+1$ a že k a j mohou být 0. Jednotkový vektor ortonormální báze na simplexu vzniklý v l-tém kroku binárního dělení nazýváme bilanční element, který definujeme takto:

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

kde

$$a = \frac{\sqrt{s}}{\sqrt{r(r+s)}}, b = \frac{-\sqrt{r}}{\sqrt{s(r+s)}}. \tag{2}$$

Pro každé binární dělení tento vztah jednoznačně definuje ortonormální bázi, avšak znaménka odpovídající složkám v tabulce mohou být změněna k získání bází, které se liší jen v orientaci. Také různá binární dělení mohou souviset s ortogonálními bázemi, které se pouze liší pořadím jednotlivých kompozic v bázi nebo permutací složek v rámci těchto kompozic.

Příklad 4.3. *Z binárního dělení z předchozího příkladu vypočítáme ortonormální bázi. Máme pět kroků binárního dělení, tzn. že dostaneme pět vektorů ortonormální báze, které vypočítáme dle vztahu (2).*

$$\mathbf{e}_1 = \left[e^{\frac{1}{\sqrt{6}}}, e^{\frac{1}{\sqrt{6}}}, e^{\frac{1}{\sqrt{6}}}, e^{-\frac{1}{\sqrt{6}}}, e^{-\frac{1}{\sqrt{6}}}, e^{-\frac{1}{\sqrt{6}}}\right],$$

$$\mathbf{e}_2 = \left[e^{\frac{1}{\sqrt{6}}}, e^{\frac{1}{\sqrt{6}}}, e^{-\frac{\sqrt{2}}{\sqrt{3}}}, 1, 1, 1\right],$$

$$\mathbf{e}_3 = \left[e^{\frac{1}{\sqrt{2}}}, e^{-\frac{1}{\sqrt{2}}}, 1, 1, 1, 1\right],$$
$$\mathbf{e}_4 = \left[1, 1, 1, e^{\frac{\sqrt{2}}{\sqrt{3}}}, e^{-\frac{1}{\sqrt{6}}}, e^{-\frac{1}{\sqrt{6}}}\right],$$
$$\mathbf{e}_5 = \left[1, 1, 1, 1, e^{\frac{1}{\sqrt{2}}}, e^{-\frac{1}{\sqrt{2}}}\right].$$

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

$$x_l^* = \ln \left[\frac{(x_{k+1} \cdots x_{k+r})^{\sqrt{s/(r(r+s))}}}{(x_{k+r+1} \cdots x_{k+r+s})^{\sqrt{r/(s(r+s))}}} \right] = \sqrt{\frac{rs}{r+s}} \ln \frac{\left(\prod_{i=1}^{r} x_i\right)^{1/r}}{\left(\prod_{i=r+1}^{r+s} x_i\right)^{1/s}}.$$

Konstrukce x_l^* intuitivně ukazuje, proč se x_l^* označuje jako bilance mezi skupinami složek $x_{k+1}, \ldots, x_{k+r}$ a $x_{k+r+1}, \ldots, x_{k+r+s}$, a proč $\mathbf{e}_l$ je nazýván bilančním elementem pro dvě skupiny složek kompozice. Podívejme se, v jakém budou tvaru pro postupné binární dělení kompozice z Příkladu 4.2.

Příklad 4.4. *Pomocí báze získáme souřadnice, zvané bilance, díky nímž se můžeme pohybovat v reálném prostoru dimenze $D-1$.*

$$x_1^* = \ln \left[\frac{(x_1 x_2 x_3)^{\frac{1}{\sqrt{6}}}}{(x_4 x_5 x_6)^{\frac{1}{\sqrt{6}}}} \right] = \sqrt{\frac{9}{6}} \ln \frac{(x_1 x_2 x_3)^{1/3}}{(x_4 x_5 x_6)^{1/3}},$$
$$x_2^* = \ln \left[\frac{(x_1 x_2)^{\frac{1}{\sqrt{6}}}}{(x_3)^{\frac{\sqrt{2}}{\sqrt{3}}}} \right] = \sqrt{\frac{2}{3}} \ln \frac{(x_1 x_2)^{1/2}}{x_3},$$
$$x_3^* = \ln \left[\frac{(x_1)^{\frac{1}{\sqrt{2}}}}{(x_2)^{\frac{1}{\sqrt{2}}}} \right] == \sqrt{\frac{1}{2}} \ln \frac{x_1}{x_2},$$
$$x_4^* = \ln \left[\frac{(x_4)^{\frac{\sqrt{2}}{\sqrt{3}}}}{(x_5 x_6)^{\frac{1}{\sqrt{6}}}} \right] = \sqrt{\frac{2}{3}} \ln \frac{x_4}{(x_5 x_6)^{1/2}},$$
$$x_5^* = \ln \left[\frac{(x_5)^{\frac{1}{\sqrt{2}}}}{(x_6)^{\frac{1}{\sqrt{2}}}} \right] = \sqrt{\frac{1}{2}} \ln \frac{x_5}{x_6}.$$

4.3 Projekce na podkompozici

Jak víme, podkompozice je kompozice vytvořená z vybraných složek původní kompozice. Proces, jak získat podkompozici, je jednoduchý: máme kompozici v $\mathcal{S}^D$, vybereme r složek (popsané množinou indexů R) a výslednou R-podkompozici získáme tak, že aplikujeme operaci uzávěru odpovídající reprezentaci podkompozice na simplexu $\mathcal{S}^r$. Výslednou podkompozici v tomto kontextu označujeme $sub(\mathbf{x}; R)$. Indexy složek nezahrnutých v R označíme jako $\overline{R}$. Je důležité poznamenat, že R-podkompozice obsahuje pouze informace o podílech mezi složkami, jejichž indexy jsou v R, ostatní podíly jsou ignorovány.

Mějme kompozici $\mathbf{x} \in \mathcal{S}^D$ a hledejme kompozici $\mathbf{x}_R \in \mathcal{S}^D$, která obsahuje pouze informace R-podkompozice a navíc Aitchisonova vzdálenost $d_a(\mathbf{x}, \mathbf{x}_R)$ je minimální. Tyto vlastnosti splňuje

$$\mathbf{x}_R = [x_1, x_2, \ldots, x_r, a, \ldots, a], \tag{3}$$

kde

$$a = \left(\prod_{i=1}^{r} x_i \right)^{\frac{1}{r}},$$

přičemž jsme pro jednoduchost zvolili $R = \{1, 2, \ldots, r\}$. Zajímavým faktem je, že $\|\mathbf{x}_R\|_a = \|sub(\mathbf{x}; R)\|_a$, kde normu počítáme na různých výběrových prostorech ($\mathcal{S}^D$ a $\mathcal{S}^r$). To znamená, že R-podkompozice v $\mathcal{S}^r$ může být reprezentována pomocí $\mathbf{x}_R$, která je stále v $\mathcal{S}^D$, a proto $\mathbf{x}_R$ je kompozice související s R-podkompozicí.

Z geometrického úhlu pohledu je $\mathbf{x}_R$ ortogonální projekcí $\mathbf{x}$ na podprostor definovaný podkompozicí. Tato operace může být jednoduše představena pomocí odpovídajících souřadnic. Podrobněji se tomuto tématu budeme věnovat v dalších kapitolách.

4.3.1 Vnitroskupinová analýza: podkompozice

Ortonormální báze odpovídající postupnému binárnímu dělení může být použita na definování ortonormální báze („podbáze“) související se skupinami složek

kompozice. Nechť $\mathbf{e}_i, i = 1, \ldots, D-1$, je ortonormální báze binárního dělení a $x_{k+1}, \ldots, x_{k+r}$ je skupina složek získaná v l-tém kroku binárního dělení. Soustředíme se na r prvkovou podkompozici (R-podkompozici) definovanou množinou indexů $R = \{k+1, k+2, \ldots, k+r\}$

$$sub(\mathbf{x}; R) = \mathcal{C}\left[x_{k+1}, \ldots, x_{k+r}\right].$$

Prvek původní báze $\mathbf{e}_j$ je v podbázi (tj. bázi odpovídající podkompozici), pokud $sub(\mathbf{e}_j; R) \neq \mathbf{n}_r$, kde $\mathbf{n}_r = [1/r, 1/r, \ldots, 1/r]$ je neutrální prvek v $\mathcal{S}^r$. Prvky báze, kde $sub(\mathbf{e}_j; R) = \mathbf{n}_r$ nesouvisí se skupinou složek odpovídajících množině indexů R (R-skupinu složek), protože nás neinformují o její vnitřní struktuře.

Bázové kompozice získané z postupného binárního dělení do l-tého kroku včetně nejsou spojeny s R-skupinou. Pouze $r-1$ zbývajících kompozic $\mathbf{e}_{l+1}, \ldots, \mathbf{e}_{D-1}$ jsou spojeny s R-skupinou: ty složky, jejichž indexy $k+1, \ldots, k+r$ nejsou stejné.

Podbáze R-skupiny generuje podprostor dimenze $r-1$, $\mathcal{S}^D(R) \subset \mathcal{S}^D$. Hlavní charakteristikou tohoto podprostoru je, že ortogonální projekce kompozice z $\mathcal{S}^D$ do $\mathcal{S}^D(R)$ R-podkompozici neovlivní. Jinak řečeno, podíly složek v R-podkompozici mohou být studovány přímo v $\mathcal{S}^D(R)$ po projekci, která sníží dimenzi z $D-1$ na $r-1$.

Jsou dva možné způsoby, jak studovat R-podkompozici nějakého datového souboru: buď vybrat R-podkompozici z dat a potom ji analyzovat, nebo nejprve získat souřadnice vzhledem k bázi a potom vybrat ty souřadnice, které odpovídají R-podbázi.

Ve druhém případě nechť $x_i^*, i \in R^*$ jsou souřadnice vzhledem k R-podbázi, kde R^* je množina $r-1$ indexů odpovídajících souřadnic. Je třeba poznamenat, že indexy mohou být v libovolném pořadí. Projekci dat do $\mathcal{S}^D(R)$ získáme jako $\bigoplus_{i \in R^*}(x_i^* \odot \mathbf{e}_i)$, což je stále kompozice o D složkách, ačkoliv je v prostoru dimenze $r-1$. Abychom získali efektivní snížení dimenze, musíme ji reprezentovat v $\mathcal{S}^r$. Abychom to mohli učinit, musíme najít reprezentační bázi $\mathbf{h}_i \in \mathcal{S}^r$ tak, že pro libovolné $\mathbf{x} \in \mathcal{S}^D$,

$$sub(\mathbf{x}; R) = \bigoplus_{i \in R^*}(x_i^* \odot \mathbf{h}_i).$$

Vhodnou reprezentační bází je $\mathbf{h}_i = sub(\mathbf{e}_i; R)$, kde $\mathbf{e}_i$ jsou prvky podbáze spojené s R-skupinou. Výhodou tohoto přístupu je, že po získání podkompozice nepotřebujeme znovu počítat souřadnice, jako jsou právě ty odpovídající prvkům podbáze souvisejících s R-skupinou.

Příklad 4.5. *Teorii si ukážeme opět na příkladu s minerální vodou Mattoni. Podíváme se, jak vypadá podkompozice kationtů - Mg^{2+}, Ca^{2+}, Na^{+}. Vybrali jsme si tedy 3 indexy, tzn. máme množinu indexů $R = \{1, 2, 3\}$. Souřadnice odpovídající této skupině jsou $x_2^* = \sqrt{\frac{2}{3}} \ln \frac{(x_1 x_2)^{1/2}}{x_3}$, $x_3^* = \sqrt{\frac{1}{2}} \ln \frac{x_1}{x_2}$. Vybrané reprezentační podbáze jsou $\mathbf{h}_2 = sub(\mathbf{e}_2, R) = \left[e^{1/\sqrt{6}}, e^{1/\sqrt{6}}, e^{-\sqrt{2}/\sqrt{3}}\right]$, $\mathbf{h}_3 = sub(\mathbf{e}_3, R) = \left[e^{1/\sqrt{2}}, e^{-1/\sqrt{2}}, 1\right]$, pro $R = \{1, 2, 3\}$. Víme, že $sub(\mathbf{x}; R) = \bigoplus_{i \in R^*} (x_i^* \odot \mathbf{h}_i)$. Pro náš konkrétní příklad máme*

$$(x_2^* \odot \mathbf{h}_2) = \mathcal{C}\left[(x_1 x_2)^{1/2}, (x_1 x_2)^{1/2}, x_3\right],$$

$$(x_3^* \odot \mathbf{h}_3) = \mathcal{C}\left[x_1, x_2, (x_1 x_2)^{1/2}\right].$$

Potom tedy

$$sub(\mathbf{x}, R) = \mathcal{C}\left[x_1, x_2, x_3\right].$$

4.3.2 Podkompozice jako ortogonální projekce

Pro lepší porozumění bilančních souřadnic a podprostoru, který souvisí se skupinou složek kompozice, je zapotřebí formálního popisu situace. Hlavním pojmem je podprostor spojený se skupinou složek, který nás odkazuje na projekce motivované vnitroskupinovou analýzou.

Nechť R je neprázdná množina indexů z množiny $1, 2, \ldots, D$ a r je počet indexů v R, tzn. $1 \leq r = Card(R) \leq D - 1$.

Definice 4.1. *Kompozice $\mathbf{x} \in \mathcal{S}^D$ je spojená s R-skupinou, jestliže $sub(\mathbf{x}; R) \neq \mathbf{n}_r$, $sub(\mathbf{x}; \overline{R}) = \mathbf{n}_{D-r}$, kde $R \cup \overline{R} = 1, 2, \ldots, D$, $R \cap \overline{R} = \emptyset$ a $\mathbf{n}_r$, $\mathbf{n}_{D-r}$ jsou*

neutrální prvky $\mathcal{S}^r$ a $\mathcal{S}^{D-r}$. Množinu kompozic spojených s R-skupinou, doplněná neutrálním prvkem $\mathcal{S}^D$, $\mathbf{n}_D$, označíme $\mathcal{S}^D(R)$.

Věta 4.1. *$\mathcal{S}^D(R)$ je $(r-1)$-dimenzionálním podprostorem $\mathcal{S}^D$.*

Důkaz: Pokud se $r = 1$, pak $n_r = [1]$ a pro všechny kompozice $\mathbf{x} \in \mathcal{S}^D$ platí $sub(\mathbf{x}; R) = \mathbf{n}_r$, což je ve sporu s Definicí 4.1. Jedinou složkou v $\mathcal{S}^D(R)$ je pak $\mathbf{n}_D$, je to tedy degenerovaný podprostor nulové dimenze. Pro $2 \leq r \leq D-1$ je $\mathcal{S}^D(R)$ uzavřený na operace perturbace a mocninná transformace, tudíž je to podprostor $\mathcal{S}^D$. Nyní nechť $\mathbf{e}_i, i = 1, 2, \ldots, r-1$, tvoří ortonormální bázi v $\mathcal{S}^r$ a předpokládejme, že R-složky umístíme na prvních r pozic v kompozici. Pro získání ortonormálních kompozic o D složkách doplníme kompoziční vektor o $D-r$ stejných (jednoznačně určených) konstant a_i tak, že získáme $\mathbf{z}_i = [e_{i1}, e_{i2}, \ldots, e_{ir}, a_i, \ldots, a_i]$, $\|\mathbf{z}_i\|_a = 1$. Kompozice $\mathbf{z}_i$ jsou ortonormální a patří do $\mathcal{S}^D(R)$. Dimenze $\mathcal{S}^D(R)$ je tedy větší nebo rovna $r-1$. Kdyby tato dimenze byla $s_1 \geq r$, pak dimenze $\mathcal{S}^D(\overline{R})$ by byla $s_2 \geq D-r$, ale pak $s_1 + s_2 \geq D > D-1$. Protože $\mathcal{S}^D(\overline{R})$ a $\mathcal{S}^D(R)$ jsou ortogonální, pak součet jejich dimenzí by měl být menší nebo roven $D-1$. Proto dimenze $\mathcal{S}^D(R)$ je $r-1$.

□

Definice 4.2. *$\mathcal{S}^D(R)$ se nazývá podprostor spojený s R-skupinou složek kompozice.*

Věta 4.2. *Nechť R definuje skupinu r složek, $2 \leq r \leq D-1$ získaných v postupném binárním dělení o počtu kroků l, $l < D-1$, a nechť $\mathbf{e}_i$, $i = 1, 2, \ldots, D-1$, je odpovídající ortonormální báze. Nechť $x_1^*, x_2^*, \ldots, x_{D-1}^*$ jsou souřadnice $\mathbf{x} \in \mathcal{S}^D$ vzhledem k této bázi. Pak platí:*

1. *báze obsahuje $r-1$ prvků, které tvoří podbázi pro $\mathcal{S}^D(R)$, jsou to $\mathbf{e}_j$, $j \in R^*$,*
2. *ortogonální projekce $\mathbf{x}$ na $\mathcal{S}^D(R)$ má souřadnice x_j^* pro $j \in R^*$ a 0 jinde,*
3. *$\mathbf{h}_j = sub(\mathbf{e}_j; R)$, $j \in R^*$ tvoří ortonormální bázi $\mathcal{S}^r$,*
4. *souřadnice $sub(\mathbf{x}; R)$ vzhledem k bázi $\mathbf{h}_j$, $j \in R^*$, jsou x_j^*.*

Důkaz: Víme, že prvky báze, které vychází z postupného binárního dělení až do l-tého kroku, s R-skupinou nesouvisí. Abychom získali dělení R-skupiny do jednosložkových podskupin, potřebujeme $r-1$ kroků postupného binárního dělení R-skupiny, z tohoto dělení získáme podbázi v $\mathcal{S}^D(R)$, tedy $\mathbf{e}_j, j \in R^*$. K těmto $r-1$ prvkům podbáze získáme souřadnice $x_j^*, j \in R^*$. Nyní bychom potřebovali snížit dimenzi z $\mathcal{S}^D(R)$ na $\mathcal{S}^r$. K tomu nám poslouží reprezentační báze $\mathbf{h}_j$, kterou vybereme z již získané ortonormální báze, dimenze této báze však bude pouze r. I přesto se stále jedná o ortonormální bázi, tentokrát však prostoru $\mathcal{S}^r$. Souřadnice podkompozice vzhledem k této reprezentační bázi jsou ty samé souřadnice x_j^*.

□

Příklad 4.6. *Mějme $R = \{1, 2, 3\}$ skupinu o $r = 3$ složkách získaných ve druhém kroku postupného binárního dělení v Příkladu 4.2. Ortonormální báze a souřadnice získané v tomto postupném binárním dělení jsou uvedeny v Příkladu 4.3 a 4.4. Pro tuto skupinu máme bázi obsahující dva prvky $\mathbf{e}_2$, $\mathbf{e}_3$ a těmto prvkům odpovídají souřadnice x_2^*, x_3^*. Reprezentační báze odpovídající této podkompozici obsahuje dva vektory:*

$$\mathbf{h}_2 = sub(\mathbf{e}_2; R) = \left[e^{1/\sqrt{6}}, e^{1/\sqrt{6}}, e^{-\sqrt{2}/\sqrt{3}}\right],$$

$$\mathbf{h}_3 = sub(\mathbf{e}_3, R) = \left[e^{1/\sqrt{2}}, e^{-1/\sqrt{2}}, 1\right].$$

Souřadnice odpovídající těmto ortonormálním bázím jsou stále x_2^*, x_3^*.

Důsledek 4.1. *Ortogonální projekce kompozice na podprostor R-skupiny složek neovlivní hodnoty složek R-podkompozice.*

Příslušná ortonormální projekce kompozice $\mathbf{x}$ na R-podkompozici, $\mathbf{x}_R$, bude (po případném přeindexování složek) přesně ve tvaru (3). Tato projekce přitom eliminuje všechnu informaci, která není vztažená k R-podkompozici, a tedy můžeme identifikovat projekci s příslušnou R-podkompozicí. Tato úvaha je přesněji zformulována v následující větě, kde ortonormální báze nutně nesouvisí s postupným binárním dělením.

Věta 4.3. *Nechť* $\{\mathbf{e}_j, j \in R^*\}$, *je množina* $r-1$ *ortonormálních kompozic v* $\mathcal{S}^D$. *Tato množina je ortonormální bází* $\mathcal{S}^D(R)$ *tehdy a jen tehdy, jestliže* $\forall \mathbf{x} \in \mathcal{S}^D$ *platí* $sub(\mathbf{x}; R) = sub(\bigoplus_{j \in R^*}(x_j^* \odot \mathbf{e}_j); R)$, *kde* $x_j^* = \langle \mathbf{x}; \mathbf{e}_j \rangle_a$.

Důkaz: Předpokládejme, že $\mathbf{e}_j, j \in R^*$ tvoří ortonormální bázi $\mathcal{S}^D(R)$ a doplňme je na ortonormální bázi v $\mathcal{S}^D$ $D-r$ kompozicemi $\mathbf{e}_k, k \in \overline{R}^*$. Tyto kompozice splňují $sub(\mathbf{e}_k; R) = \mathbf{n}_D$, protože dimenze $\mathcal{S}^D(R)$ je $r-1$. Tvrzení věty pak získáme tak, že vezmeme R-podkompozici z reprezentace kompozice $\mathbf{x}$.

Nyní předpokládejme, že věta platí. Potom $\mathbf{e}_j, j \in R^*$ je $r-1$ ortonormálních kompozic, které tvoří ortonormální bázi $\mathcal{S}^D$ s $D-r$ prvky $\mathbf{e}_k, k \in \overline{R}^*$. Každá kompozice $\mathbf{x} \in \mathcal{S}^D$ je vyjádřena jako $\mathbf{x} = \bigoplus_i (x_i^* \odot \mathbf{e}_i)$. Toto platí specielně pro $\mathbf{e}_k, k \in \overline{R}^*$ a vztah v tvrzení věty se redukuje na $sub(\mathbf{e}_k; R) = \mathbf{n}_r$. Z toho vyplývá, že $\mathbf{e}_k, k \in \overline{R}^*$ neleží v $\mathcal{S}^D(R)$ a tvoří ortonormální bázi ortogonálního doplňku $\mathcal{S}^D(R)$. Jestliže předpokládáme, že $\mathbf{e}_j, j \in R^*$ jsou ortogonální k $\mathbf{e}_k, k \in \overline{R}^*$, pak ty první leží v podprostoru $\mathcal{S}^D(R)$.

□

Důsledek 4.2. *Podíly a logaritmy podílů prvků v R-podkompozici jsou stejné jako poměry a log-ratio R-prvků v ortogonální projekci na* $\mathcal{S}^D(R)$.

4.4 Meziskupinová analýza: bilance

K reprezentaci vztahu mezi skupinami složek používáme odpovídající souřadnice - bilance. To obecně způsobuje snížení dimenze - je méně skupin než složek - a tudíž potřebujeme reprezentaci na simplexu s nižší dimenzí. Ačkoliv se tato situace podobá vnitroskupinové analýze, vyskytují se v tomto případě navíc normalizační konstanty, a to z toho důvodu, že je obecně v každé skupině různý počet složek.

Předpokládejme, že máme bilance mezi $l+1$ skupinami, které tvoří dělení celé množiny D složek (např. postupné binární dělení). Nechť $\mathbf{e}_i, i = 1, \ldots, l$, jsou prvky báze odpovídající postupnému binárnímu dělení až do požadovaného kroku. Odpovídající souřadnice $x_i^*, i = 1, \ldots, l$, jsou bilance mezi skupinami složek a zahrnují všechny informace o vztazích mezi skupinami.

Explicitní vyjádření ortogonální projekce kompozice na bilanční element, $x_i^* \odot \mathbf{e}_i$, pro jakékoliv $i \leq l$, nám umožní lepší pohled na to, co bilance reprezentují. Předpokládejme, že v i-tém kroku oddělujeme dvě skupiny složek tvořené r a s složkami (viz. kapitola 4.2.2). Potom

$$x_i^* \odot \mathbf{e}_i = \mathcal{C}\left[\underbrace{\left(\prod_{i=k+1}^{k+r+s} x_i\right)^{\frac{1}{r+s}}}_{k \text{ opakovaných složek}}, \underbrace{\left(\prod_{i=k+1}^{k+r} x_i\right)^{\frac{1}{r}}}_{r \text{ opakovaných složek}}, \underbrace{\left(\prod_{i=k+r+1}^{k+r+s} x_i\right)^{\frac{1}{s}}}_{s \text{ opakovaných složek}}, \underbrace{\left(\prod_{i=k+1}^{k+r+s} x_i\right)^{\frac{1}{r+s}}}_{j \text{ opakovaných složek}}\right],$$

kde $k+r+s+j = D$. Musíme si uvědomit, že každá původní složka ve skupině je nahrazena geometrickým průměrem složek zahrnutých v této skupině. Mimo skupinu je každá složka nahrazena geometrickým průměrem všech složek zahrnutých v obou skupinách. Bilanční prvky $\mathbf{e}_i, i = 1, \ldots, l$, tvoří ortogonální bázi podprostoru a projekce $\mathbf{x}$ na tento podprostor je kompozice o D složkách, ve které jsou uvažovány pouze vztahy mezi složkami odpovídajícími různým skupinám a informace o vnitroskupinových vztazích byly eliminovány. Vyjádření této projekce je

$$\bigoplus_{i=1}^{l}(x_i^* \odot \mathbf{e}_i) = \mathcal{C}\left[\underbrace{\left(\prod_{j=1}^{r_1} x_j\right)^{\frac{1}{r_1}}}_{r_1 \text{ opakovaných složek}}, \underbrace{\left(\prod_{j=1}^{r_2} x_{r_1+j}\right)^{\frac{1}{r_2}}}_{r_2 \text{ opakovaných složek}}, \cdots, \underbrace{\left(\prod_{j=1}^{r_l} x_{n+1-j}\right)^{\frac{1}{r_{l+1}}}}_{r_{l+1} \text{ opakovaných složek}}\right],$$

kde $r_1, \ldots, r_{l+1}$ jsou počty složek každé z $l+1$ skupin získaných v l-tém kroku binárního dělení.

Bilanční souřadnice $x_i^*, i = 1, 2, \ldots, l$, můžeme reprezentovat v $l+1$-složkovém simplexu o dimenzi l, s $l+1$ prvky a dimenze l. Abychom to mohli udělat, potřebujeme vybrat ortonormální bázi a přiřadit každé bilanční souřadnici $x_i^*, i = 1, 2, \ldots, l$, vektor reprezentační báze, $\mathbf{h}_i$, stejně jako pro podkompozice. Pak bilance mezi skupinami mohou být reprezentovány v $\mathcal{S}^{l+1}$ jako $\bigoplus_{i=1}^{l}(x_i^* \odot \mathbf{h}_i)$. Ačkoliv je výběr reprezentační báze libovolný, většinou ji konstruujeme následovně: V l-tém kroku postupného binárního dělení máme požadované skupiny složek.

Každá skupina může být brána jako jedna složka, pak tedy postupné binární dělení do l-tého kroku ztotožníme s postupným binárním dělením $l+1$ složek; odpovídající ortonormální báze $\mathcal{S}^{l+1}$ tedy může být brána jako reprezentační báze $\mathbf{h}_i, i = 1, 2, \ldots, l$.

Příklad 4.7. *Mějme dáno postupné binární dělení z Příkladu 4.2. Zájem je soustředěn na krok $l = 2$, kde rozdělujeme skupinu kationtů. Dle předchozí teorie si tedy vybereme bázi*

$$\mathbf{e}_1 = \left[e^{\frac{1}{\sqrt{6}}}, e^{\frac{1}{\sqrt{6}}}, e^{\frac{1}{\sqrt{6}}}, e^{-\frac{1}{\sqrt{6}}}, e^{-\frac{1}{\sqrt{6}}}, e^{-\frac{1}{\sqrt{6}}}\right]$$

a

$$\mathbf{e}_2 = \left[e^{\frac{1}{\sqrt{6}}}, e^{\frac{1}{\sqrt{6}}}, e^{-\frac{\sqrt{2}}{\sqrt{3}}}, 1, 1, 1\right],$$

odpovídající souřadnice jsou $x_1^* = \sqrt{\frac{9}{6}} \ln \frac{(x_1x_2x_3)^{1/3}}{(x_4x_5x_6)^{1/3}}$ *a* $x_2^* = \sqrt{\frac{2}{3}} \ln \frac{(x_1x_2)^{1/2}}{x_3}$. *Ortogonální projekce kompozice na bilanční element vypadá následovně:*

$$(x_1^* \odot \mathbf{e}_1) = \mathcal{C}\left[(x_1x_2x_3)^{1/3}, (x_1x_2x_3)^{1/3}, (x_1x_2x_3)^{1/3}, (x_4x_5x_6)^{1/3}, (x_4x_5x_6)^{1/3}, (x_4x_5x_6)^{1/3}\right],$$

$$(x_2^* \odot \mathbf{e}_2) = \mathcal{C}\left[(x_1x_2x_3)^{1/3}, (x_1x_2x_3)^{1/3}, (x_1x_2x_3)^{1/3}, (x_1x_2)^{1/2}, (x_1x_2)^{1/2}, x_3\right].$$

A potom projekci, ve které jsou vnitroskupinové vztahy eliminovány, získáme jednoduše:

$$\bigoplus_{i=1}^{2} (x_i^* \odot \mathbf{e}_i) = \mathcal{C}\left[(x_1x_2)^{1/2}, (x_1x_2)^{1/2}, x_3, (x_4x_5x_6)^{1/3}, (x_4x_5x_6)^{1/3}, (x_4x_5x_6)^{1/3}\right].$$

4.5 Podkompoziční a bilanční dominance pro vzdálenosti

Důležitou vlastností analýzy skupin složek je zachování požadovaných vlastností vzdálenosti na simplexu. Bylo ukázáno, že Aitchisonova vzdálenost v $\mathcal{S}^D$ mezi dvěma kompozicemi $\mathbf{x}$ a $\mathbf{y}$ může být vyjádřena jako obvyklá euklidovská vzdálenost díky souřadnicím vzhledem k ortonormální bázi,

$$d_a^2(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{D-1} (x_i^* - y_i^*)^2.$$

Předpokládejme, že nás zajímá $l+1$ skupin definovaných v l-tém kroku dělení. Prvky báze dané postupným binárním dělením mohou být rozděleny do skupin:

1. obsahující bilanční prvky mezi skupinami daného dělení, např. $\mathbf{e}_i, i = 1, \ldots, l$,

2. obsahující podbáze příslušné skupinám složek v dělení; nechť $\mathbf{e}_j, j \in R_k^*$, je podbáze, která souvisí s R_k-skupinou r_k složek. Počet indexů v R_k^* je pak $r_k - 1$, kde $\sum_{j=1}^{l+1}(r_k - 1) = D - l - 1$ a $\bigcup_{k=1}^{l+1} R_k = \{l+1, l+2, \ldots, D\}$.

Nyní druhou mocninu vzdálenosti mezi kompozicemi můžeme rozdělit na výrazy meziskupinových bilancí a vnitroskupinových souřadnic,

$$d_a^2(\mathbf{x}, \mathbf{y}) = \underbrace{\sum_{i=1}^{l}(x_i^* - y_i^*)^2}_{meziskupinová\ vzdálenost} + \sum_{k=1}^{l+1} \underbrace{\sum_{j \in R_k^*}(x_j^* - y_j^*)^2}_{vnitro-k-skupinová\ vzdálenost}.$$

Součet výrazů pro $i = 1, \ldots, l$, je příspěvek meziskupinových bilancí ke čtverci vzdálenosti. Součet výrazů pro $j \in R_k^*$ je vnitroskupinový příspěvek jednotlivých R_k-skupin; každý dílčí součet představuje Aitchisonovu vzdálenost mezi $\mathbf{x}$ a $\mathbf{y}$ měřenou v R_k-podkompozici.

Prvním závěrem je, že podkompoziční přínos ke čtverci vzdálenosti je dominován vzdáleností v $\mathcal{S}^D$,

$$d_a^2(\mathbf{x}, \mathbf{y}) \geq \sum_{j \in R_k^*}(x_j^* - y_j^*).$$

Druhým závěrem je, že meziskupinový přínos k druhé mocnině vzdálenosti je rovněž dominován vzdáleností v $\mathcal{S}^D$,

$$d_a^2(\mathbf{x}, \mathbf{y}) \geq \sum_{i=1}^{l}(x_i^* - y_i^*).$$

Tyto dvě výše uvedené vlastnosti jsou velmi důležité pro bezespornou statistickou analýzu kompozičních dat při práci se skupinami složek, ať už prostřednictvím bilancí mezi skupinami složek nebo podkompozic. Navíc všechny výrazy

jsou čtvercové vzdálenosti měřené podél ortogonálních směrů - směrů prvků ortonormální báze - a proto by dekompozice měla být interpretována jako Pythagorova věta: čtvercové vzdálenosti jsou získány sečtením čtvercových vzdáleností odpovídajících ortogonálních příspěvků.

5 Kovarianční struktura bilancí

Tato kapitola vychází z literatury [3].

Doposud jsme získané bilance interpretovali především jako rovnováhy mezi dvěma skupinami složek kompozice. Vzhledem k tomu, že veškerá informace v kompozici je obsažena v podílech mezi jejími složkami, nabízí se otázka, zda podíly (resp. logaritmy podílů) hrají nějakou roli i při interpretaci bilancí. Následující věty, týkající se rozkladu kovarianční struktury bilancí na lineární kombinace rozptylů logaritmů podílů, jednu takovouto alternativní interpretaci umožňují. Tato vychází z hlavní myšlenky metody hlavních komponent, kdy jsou rozptyly proměnných ztotožněny s informací nesenou danými proměnnými. Než této myšlenky využijeme k interpretaci bilancí, řekněme si, že pod pojmem log-kontrast dané D-složkové kompozice $\mathbf{x} = [x_1, \ldots, x_D]$ rozumíme výraz

$$\sum_{i=1}^{D} a_i' \ln x_i = \mathbf{a}(\ln \mathbf{x})',$$

kde $\sum_{i=1}^{D} a_i = 0$. Specielně je pak i každá bilance log-kontrastem.

Kovarianční struktura log-kontrastů je popsána následující větou, která je dále používána ve větách o rozptylech a kovariancích bilancí.

Věta 5.1. *Rozptyly a kovariance pro log-kontrasty* $\mathbf{a}(\ln \mathbf{x})'$ *a* $\mathbf{b}(\ln \mathbf{x})'$ *D-prvkové kompozice* $\mathbf{x}$ *jsou*

$$\operatorname{var}\left[\mathbf{a}(\ln \mathbf{x})'\right] = -\frac{1}{2}\mathbf{a}\mathbf{T}\mathbf{a}',$$

$$\operatorname{cov}\left[\mathbf{a}(\ln \mathbf{x})', \mathbf{b}(\ln \mathbf{x})'\right] = -\frac{1}{2}\mathbf{a}\mathbf{T}\mathbf{b}',$$

kde $\mathbf{T}$ *je varianční matice kompozice* $\mathbf{x}$ *definovaná předpisem*

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

Věta 5.2. *Zvolme libovolnou bilanci x^* mezi dvěma disjunktními skupinami složek kompozice tak, že $i_1, i_2, \ldots, i_r$ jsou indexy r složek první skupiny a $j_1, j_2, \ldots, j_s$ jsou indexy s složek druhé skupiny. Pak rozptyl x^* je*

$$\mathrm{var}(x^*) = \frac{1}{r+s}\sum_{p=1}^{r}\sum_{q=1}^{s}\mathrm{var}\left(\ln\frac{x_{i_p}}{x_{j_q}}\right) - \frac{s}{2r(r+s)}\sum_{p=1}^{r}\sum_{q=1}^{r}\mathrm{var}\left(\ln\frac{x_{i_p}}{x_{i_q}}\right) -$$

$$-\frac{r}{2s(r+s)}\sum_{p=1}^{s}\sum_{q=1}^{s}\mathrm{var}\left(\ln\frac{x_{j_p}}{x_{j_q}}\right).$$

Jak již bylo zmíněno výše, rozptyl bilance x^* je vlastně lineární kombinací rozptylů logaritmů podílů. Přitom rozptyly s kladným znaménkem vlastně charakterizují logaritmy podílů, které jsou touto bilancí vysvětleny. Můžeme tak říci, že bilance x^* vysvětluje všechny logaritmy podílů mezi složkami jedné a druhé skupiny. Ilustrujme si to na konkrétním příkladu:

Příklad 5.1. *Mějme opět postupné binární dělení z Příkladu 4.2. Jestliže se zajímáme o rozptyl bilance x_1^*, dostaneme*

$$\mathrm{var}(x_1^*) = \frac{1}{6}\left[\mathrm{var}\left(\ln\frac{x_1}{x_4}\right) + \mathrm{var}\left(\ln\frac{x_1}{x_5}\right) + \mathrm{var}\left(\ln\frac{x_1}{x_6}\right) + \mathrm{var}\left(\ln\frac{x_2}{x_4}\right) +\right.$$

$$\left.+\mathrm{var}\left(\ln\frac{x_2}{x_5}\right) + \mathrm{var}\left(\ln\frac{x_2}{x_6}\right) + \mathrm{var}\left(\ln\frac{x_3}{x_4}\right) + \mathrm{var}\left(\ln\frac{x_3}{x_5}\right) + \mathrm{var}\left(\ln\frac{x_3}{x_6}\right)\right] -$$

$$-\frac{1}{12}\left[\mathrm{var}\left(\ln\frac{x_1}{x_2}\right) + \mathrm{var}\left(\ln\frac{x_1}{x_3}\right) + \mathrm{var}\left(\ln\frac{x_2}{x_1}\right) + \mathrm{var}\left(\ln\frac{x_2}{x_3}\right) + \mathrm{var}\left(\ln\frac{x_3}{x_1}\right) +\right.$$

$$\left.+\mathrm{var}\left(\ln\frac{x_3}{x_2}\right)\right] - \frac{1}{12}\left[\mathrm{var}\left(\ln\frac{x_4}{x_5}\right) + \mathrm{var}\left(\ln\frac{x_4}{x_6}\right) + \mathrm{var}\left(\ln\frac{x_5}{x_4}\right) + \mathrm{var}\left(\ln\frac{x_5}{x_6}\right) +\right.$$

$$\left.+\mathrm{var}\left(\ln\frac{x_6}{x_4}\right) + \mathrm{var}\left(\ln\frac{x_6}{x_5}\right)\right].$$

Pro úplnost se podívejme také na strukturu kovariance mezi dvěma bilancemi:

Věta 5.3. *Nechť jsou dány bilance $x_k^*, k = 1, 2$, mezi dvěma skupinami tak, že $i_1^k, i_2^k, \ldots, i_{r_k}^k$ jsou indexy r_k složek první skupiny a $j_1^k, j_2^k, \ldots, j_{s_k}^k$ jsou indexy s_k složek druhé skupiny. Pak kovariance mezi x_1^* a x_2^* je*

$$\operatorname{cov}(x_1^*, x_2^*) = \frac{C}{2r_1s_2}\sum_{p=1}^{r_1}\sum_{q=1}^{s_2}\operatorname{var}\left(\ln\frac{x_{i_p^1}}{x_{j_q^2}}\right) + \frac{C}{2r_2s_1}\sum_{p=1}^{r_2}\sum_{q=1}^{s_1}\operatorname{var}\left(\ln\frac{x_{i_p^2}}{x_{j_q^1}}\right) -$$

$$-\frac{C}{2r_1r_2}\sum_{p=1}^{r_1}\sum_{q=1}^{r_2} var\left(\ln\frac{x_{i_p^1}}{x_{i_q^2}}\right) - \frac{C}{2s_1s_2}\sum_{p=1}^{s_1}\sum_{q=1}^{s_2}\operatorname{var}\left(\ln\frac{x_{j_p^1}}{x_{j_q^2}}\right),$$

kde koeficient C je definován takto

$$C = \sqrt{\frac{r_1r_2s_1s_2}{(r_1+s_1)(r_2+s_2)}}.$$

Poznámka 5.1. *Nechť $I_k^+ = \left\{i_1^k, i_2^k, \ldots, i_{r_k}^k\right\}$ je množina indexů r_k složek první skupiny a $J_k^- = \left\{j_1^k, j_2^k, \ldots, j_{s_k}^k\right\}$ je množina indexů s_k složek ze druhé skupiny. Potom rozptyl bilance x_k^* mezi těmito dvěma skupinami můžeme přepsat jako*

$$\operatorname{var}(x_k^*) = \frac{1}{\left|I_k^+\right| + \left|J_k^-\right|}\sum_{p\in I_k^+}\sum_{q\in J_k^-}\operatorname{var}\left(\ln\frac{x_p}{x_q}\right) - \frac{\left|J_k^-\right|}{2\left|I_k^+\right|\left(\left|I_k^+\right| + \left|J_k^-\right|\right)}\sum_{p\in I_k^+}\sum_{q\in I_k^+}\operatorname{var}\left(\ln\frac{x_p}{x_q}\right) -$$

$$-\frac{\left|I_k^+\right|}{2\left|J_k^-\right|\left(\left|I_k^+\right| + \left|J_k^-\right|\right)}\sum_{p\in J_k^-}\sum_{q\in J_k^-}\operatorname{var}\left(\ln\frac{x_p}{x_q}\right),$$

kde symbol $\left|I_k^+\right|$ označuje kardinalitu množiny I_k^+. Podobně kovariance mezi bilancemi x_1^* a x_2^* může být přepsána jako

$$\operatorname{cov}(x_1^*, x_2^*) = \frac{C}{2\left|I_1^+\right|\left|J_2^-\right|}\sum_{p\in I_1^+}\sum_{q\in J_2^-}\operatorname{var}\left(\ln\frac{x_p}{x_q}\right) + \frac{C}{2\left|J_1^-\right|\left|I_2^+\right|}\sum_{p\in I_2^+}\sum_{q\in J_1^-}\operatorname{var}\left(\ln\frac{x_p}{x_q}\right) -$$

$$-\frac{C}{2\left|I_1^+\right|\left|I_2^+\right|}\sum_{p\in I_1^+}\sum_{q\in I_2^+}\operatorname{var}\left(\ln\frac{x_p}{x_q}\right) - \frac{C}{2\left|J_1^-\right|\left|J_2^-\right|}\sum_{p\in J_1^-}\sum_{q\in J_2^-}\operatorname{var}\left(\ln\frac{x_p}{x_q}\right),$$

kde

$$C = \sqrt{\frac{\left|I_1^+\right|\left|J_1^-\right|\left|I_2^+\right|\left|J_2^-\right|}{(\left|I_1^+\right|+\left|J_1^-\right|)(\left|I_2^+\right|+\left|J_2^-\right|)}}.$$

Poznámka 5.2. *Pokud zaměníme ve vzorcích z předchozích vět* $\operatorname{var}\left(\ln\frac{x_p}{x_q}\right)$ *za* $\operatorname{var}\left(\frac{1}{\sqrt{2}}\ln\frac{x_p}{x_q}\right)$ *(s přizpůsobením příslušných konstant), získáme rozptyly odpovídajících bilancí jako lineární kombinace rozptylů podkompozic* (x_p, x_q).

Jestliže bilance zvolíme dle vztahu

$$x_i^* = \sqrt{\frac{D-i}{D-i+1}}\ln\frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}, i = 1, \ldots, D-1, \tag{4}$$

kovarianční struktura se nám zjednoduší. Tato volba bilancí je v praxi velmi oblíbená, protože x_1^* obsahuje veškerou relativní informaci složky x_1 vzhledem k ostatním složkám kompozice (tj. vysvětluje všechny podíly složky x_1 a ostatních složek).

Potom tedy

$$\operatorname{var}(x_i^*) = \frac{1}{D-i+1}\sum_{p=i+1}^{D}\operatorname{var}\left(\ln\frac{x_i}{x_p}\right) - \frac{1}{2(D-i)(D-i+1)}\sum_{p=i+1}^{D}\sum_{q=i+1}^{D}\operatorname{var}\left(\ln\frac{x_p}{x_q}\right)$$

a

$$\operatorname{cov}(x_i^*, x_j^*) = \frac{C}{2(D-j)}\sum_{p=j+1}^{D}\operatorname{var}\left(\ln\frac{x_i}{x_p}\right) + \frac{C}{2(D-i)}\sum_{p=i+1}^{D}\operatorname{var}\left(\ln\frac{x_j}{x_p}\right) -$$

$$-\frac{C}{2}\operatorname{var}\left(\ln\frac{x_i}{x_j}\right) - \frac{C}{2(D-i)(D-j)}\sum_{p=i+1}^{D}\sum_{q=j+1}^{D}\operatorname{var}\left(\ln\frac{x_p}{x_q}\right),$$

kde

$$C = \sqrt{\frac{(D-i)(D-j)}{(D-i+1)(D-j+1)}}.$$

5.1 Kovarianční struktura třísložkové kompozice

Statistická analýza třísložkové kompozice je poměrně častá, protože jí odpovídají dvě ortonormální souřadnice (bilance), jejichž hodnoty můžeme zobrazit v rovinném grafu a udělat si tak dobrou představu o struktuře datového souboru.

Jak jsme viděli výše, obecné binární dělení vede k poměrně složité interpretaci kovarianční struktury souřadnic. Pokud však spočítáme bilance dle vzorce (4), dostaneme jednodušší vyjádření a pro $D = 3$ získáme následující souřadnice,

$$x_1^* = \sqrt{\frac{2}{3}} \ln \frac{x_1}{\sqrt{x_2 x_3}}, x_2^* = \frac{1}{\sqrt{2}} \ln \frac{x_2}{x_3}$$

s

$$\mathrm{var}(x_1^*) = \frac{1}{3}\mathrm{var}\left(\ln \frac{x_1}{x_2}\right) + \frac{1}{3}\mathrm{var}\left(\ln \frac{x_1}{x_3}\right) - \frac{1}{6}\mathrm{var}\left(\ln \frac{x_2}{x_3}\right), \mathrm{var}(x_2^*) = \frac{1}{2}\mathrm{var}\left(\ln \frac{x_2}{x_3}\right)$$

a

$$\mathrm{cov}(x_1^*, x_2^*) = \frac{1}{2\sqrt{3}}\mathrm{var}\left(\ln \frac{x_1}{x_3}\right) - \frac{1}{2\sqrt{3}}\mathrm{var}\left(\ln \frac{x_1}{x_2}\right).$$

Obecně může být rozptyl $\mathrm{var}\left(\ln \frac{x_i}{x_j}\right), i = 1, \ldots, D$ použit jako míra variability poměru mezi x_i a x_j: pokud je rozptyl nulový (nebo téměř nulový), potom je poměr x_i/x_j konstantní, nebo alespoň téměř konstantní. Tedy pro $\mathrm{cov}(x_1^*, x_2^*) = 0$ je variabilita poměrů x_1/x_3 a x_1/x_2 stejná. Na druhou stranu může být nulová kovariance také použita k závěru, že oba poměry mají stejný podíl na výsledném rozptylu x_1^*. Pokud je $\mathrm{cov}(x_1^*, x_2^*) > 0$, resp. $\mathrm{cov}(x_1^*, x_2^*) < 0$, ukáže se dominantní role x_1/x_3, resp. x_1/x_2.

Analogických závěrů bychom dosáhli také použitím příslušného korelačního koeficientu,

$$\rho_{x_1^*, x_2^*} = \frac{\mathrm{cov}(x_1^*, x_2^*)}{\sqrt{\mathrm{var}(x_1^*) \cdot \mathrm{var}(x_2^*)}}.$$

Předchozí úvaha může být použita k získání obecných doporučení pro interpretaci bilancí v případě třísložkové kompozice. Podíly mezi složkami x_1, x_2, x_3 jsou vysvětleny dvěma bilancemi, jejichž pozice očividně není stejná. Zmíněný

výběr bilancí odpovídá důležité roli poměru x_2/x_3, jak je znázorněno bilancí x_2^*. Na druhou stranu první bilance obsahuje veškerou relativní informaci o x_1. Dominantní roli jednoho z poměrů x_1/x_2 a x_1/x_3 pro výslednou hodnotu rozptylu x_1^* můžeme odvodit ze znaménka odpovídající kovariance a korelačního koeficientu. Navíc nízké rozptyly obou bilancí (ale zejména té první) a přibližně nulová kovariance znamená konstantní podíl x_1 v kompozici, tedy konstantní podíly x_1/x_2 a x_1/x_3.

Provedené teoretické úvahy z této i předchozích kapitol využijeme dále při řešení příkladu s reálnými daty pomocí softwaru CoDaPack.

6 Bilanční dendrogram a používání programu CoDaPack

V předchozích kapitolách jsme zjistili, že statistická analýza kompozičních dat založená na Aitchisonově geometrii simplexu vyžaduje vytvoření odpovídající báze k reprezentaci dat. Jednoduchý a intuitivní způsob konstrukce této báze je postupné binární dělení kompozičního vektoru. Dělení spolu se statistickými charakteristikami souřadnic (bilancí) mohou být reprezentovány v grafu typu dendrogram. V této kapitole bude představen program pro práci s kompozičními daty zvaný CoDaPack a ukážeme, jak s jeho pomocí statisticky analyzovat kompozice a v neposlední řadě, jak sestrojit a interpretovat bilanční dendrogram.

Kapitola byla vytvořena za použití zdrojů [7], [11].

6.1 Program CoDaPack

Tento program pro analýzu kompozičních dat je volně dostupný na internetu (na adrese: http://ima.udg.edu/codapack/) a uživatelsky přívětivý. CoDaPack se skládá ze tří menu (Data, Statistics a Graphs) pro Microsoft Excel za použití programovacího jazyka Visual Basic. Používání balíčku je jednoduché a intuitivní, po instalaci se každá část zobrazuje jako menu nebo podmenu v Microsoft Excel. Jesliže aktivujeme nějaký proces, otevře se nové okno a uživatel je vyzván ke vložení proměnných a dotázán na konkrétní možnosti vzhledem k vytvoření grafických nebo numerických výsledků. Numerické výsledky se zobrazí na stejném listu, zatímco grafické výsledky se zobrazí v novém okně.

Analyzovaná data by měla být uložena na listu CoDaPack.xls, na který jsou jednotlivé procedury navázány. Musí být seřazeny do matice, jejíž sloupce odpovídají složkám a řádky pozorováním. Maximální velikost je 200 složek a 6000 pozorování. První řádek je ponechán pro označení (pojmenování) složek kompozičního vektoru, případně může zůstat prázdný.

V prvním menu najdeme transformace mezi simplexem a reálným prostorem, a naopak, jako je clr, alr a ilr transformace a jejich inverze, dále pak operace

amalgamace a základní algebraické operace na simplexu, mimo jiné perturbace, mocninná transformace a operace uzávěru pro podkompozici. Druhé menu nabízí výpočet základních statistických charakteristik a test logistického normálního rozdělení (viz kapitola 7.3). A konečně třetí menu aktivuje grafické procesy: známé ternární diagramy, grafy pro alr, clr a ilr transformovaná data, a dále biplot a bilanční dendrogram.

6.2 Bilanční dendrogram

Bilanční dendrogram byl vytvořen pro zobrazování postupného binárního dělení, bilancí, dekompozice celkového rozptylu (rovného součtu rozptylů jednotlivých bilancí) a dalších jednorozměrných charakteristik.

Bilanční dendrogram je sestaven z následujících částí:

- Označení složek ve spodní části grafu.
- Horizontální úsečky, které spojují složky nebo skupiny složek. Tyto horizontální úsečky jsou užity pro reprezentaci popisných statistik bilancí; každá z nich reprezentuje stejně dlouhý interval.
- Vertikální spojení, jejichž délka je určena proporcí celkového rozptylu vysvětlované bilance reprezentované horizontální přímkou v nižším konci dané spojnice.

Bilanční dendrogram umožňuje zobrazení:

- Postupného binárního dělení pomocí vertikálních a horizontálních spojnic v dendrogramu.
- Výběrového průměru odpovídajícího každé bilanci, což je bod, kde končí vertikální spojnice.
- Proporce výběrového celkového rozptylu odpovídající každé bilanci, reprezentované délkou vertikální spojnice. Součet délek všech vertikálních spojnic reprezentuje celkový rozptyl a dlouhá spojnice v určité části dendro-

gramu znamená velkou variabilitu ve výběru (prostřednictvím rozptylu odpovídající bilance). To odpovídá tomu, že tato bilance vysvětluje velkou část celkového rozptylu.

- Souhrnných statistik každé bilance, reprezentované box-plotem 0,05, 0,25, 0,50, 0,75 a 0,95-percentilů na každé horizontální úsečce.

Nyní si tento obecný postup demonstrujeme na příkladu, který nám zároveň poslouží jako návod pro vytvoření bilančního dendrogramu.

Data příkladu tvoří 10 různých minerálních vod a obsah minerálů v nich (v mg/l). Tabulka byla vytvořena pouze z kationtů a aniontů, které byly obsaženy v každé z uvedených minerálek (tím pádem jsme z analýzy vyloučili aniont chloridu, vyskytující se v předchozích příkladech s minerální vodou Mattoni):

	A	B	C	D	E	F
1		Mg^{2+}	Ca^{2+}	Na^{+}	HCO_3^{-}	SO_4^{2-}
2	Mattoni	25	84,5	69,9	52,8	40
3	Magnesia	170	37,4	6,17	970	11,1
4	Toma Natura	6,63	30,1	1,06	104	25
5	Rajec	19	87,4	18,3	321	20,3
6	Bonaqua	39,5	61,3	1,6	381	21,2
7	Dobrá voda	7,65	5,31	13,3	105	1,74
8	Korunní	30	78,3	98,2	547	57,3
9	Hanácká kyselka	68	275	251	1454	2
10	Poděbradka	69,4	172	495	1437	78,5
11	Ondrášovka	21,4	200	30	75	13,6

Nyní si ve zkratce uvedeme, jak jsou pro nás jednotlivé minerály přínosné (shrnuto ze zdrojů [5], [6]).

Ca^{2+} - kationt vápenatý, má velmi důležitou roli při zvyšování mechanické odolnosti tkání a dodává tkáním, zejména kostem a zubům, jejich tvrdost a mechanickou odolnost.

Mg^{2+} - kationt hořečnatý, nepostradatelný prvek pro získávání energie na všechny reakce, které souvisí s přeměnami energie v lidském těle. Funkce svalů, přenos vzruchu a funkce nervové tkáně, tvorbu tuků a bílkovin v těle, reakce

důležité pro ochranu před toxickými projevy.

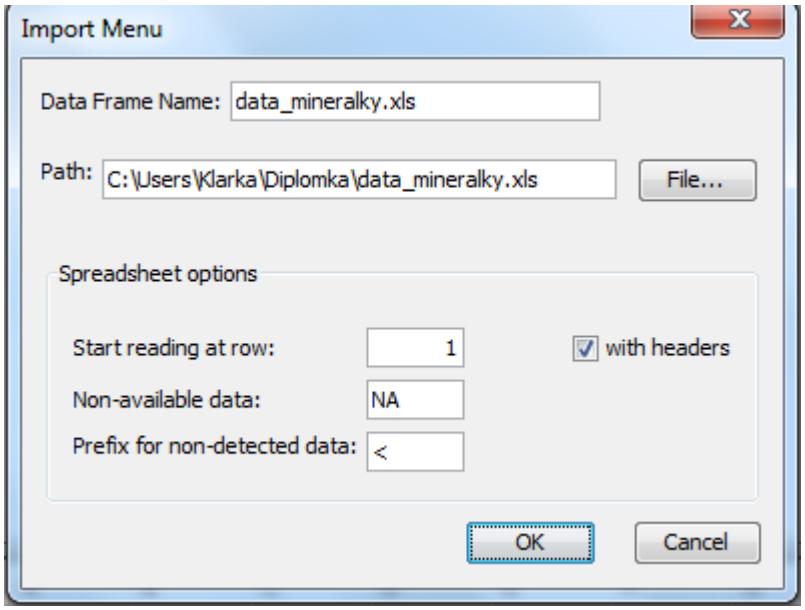
Na^+ - kationt sodný, je důležitý pro udržení tzv. osmotického tlaku, který zabraňuje přílišnému pronikání vody do buněk a iontové síly tělních tekutin. Při nedostatku sodíku se dostavují svalové křeče, bolesti hlavy a průjmy.

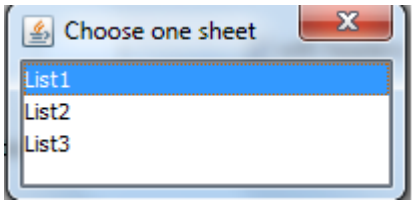
HCO_3^- - hydrogenuhličitan, podporuje vylučování žaludeční šťávy a usnadňuje tak trávení. Kromě toho napomáhá vstřebávání dalších minerálních látek i některých léků, například antibiotik.

SO_4^{2-} - síran, má silně projímavé účinky, proto je vhodný pro pročištění.

6.2.1 Postup pro vytvoření bilančního dendrogramu

Nejprve v programu CoDaPack musíme načíst data; jak již bylo uvedeno, je přitom nutné, aby byla data uložena jako excelovský dokument. V hlavním menu vybereme nabídku File → Import Data a najdeme potřebný soubor. Vyskočí nám tabulka, ve které si vybereme, na kterém řádku začneme data načítat, jestli chceme nechat popisky složek, jak se označí prázdná místa atd. A vybereme si list, na kterém chceme data zobrazit.





Po načtení dat se můžeme nejprve podívat na charakteristiky dat. Tím jsou myšleny průměry a rozptyly logaritmů podílů složek kompozice (ty druhé vlastně představují prvky symetrické varianční matice), uspořádány jako dolní a horní diagonála čtvercové tabulky. K tomu jednoduše použijeme menu Statistics → Compositional statistics summary. Výsledkem je ovšem nejprve tabulka obsahující geometrické průměry (ozn. Center) jednotlivých minerálních prvků, minimum (ozn. 0), maximum (ozn. 100) a jednotlivé kvartily (ozn. 25, 50, 75).

Statistics

	Center	0	25	50	75	100
Mg2+	0.0674	0.0308	0.0370	0.0575	0.0783	0.1423
Ca2+	0.1548	0.0313	0.0764	0.1341	0.1876	0.5882
Na+	0.0551	0.0032	0.0064	0.1000	0.1224	0.2568
HCO3-	0.6863	0.1940	0.6235	0.6888	0.7551	0.8119
SO42-	0.0363	0.0010	0.0131	0.0420	0.0707	0.1499

Samotná avizovaná tabulka je uvedena až dále v tomto menu.

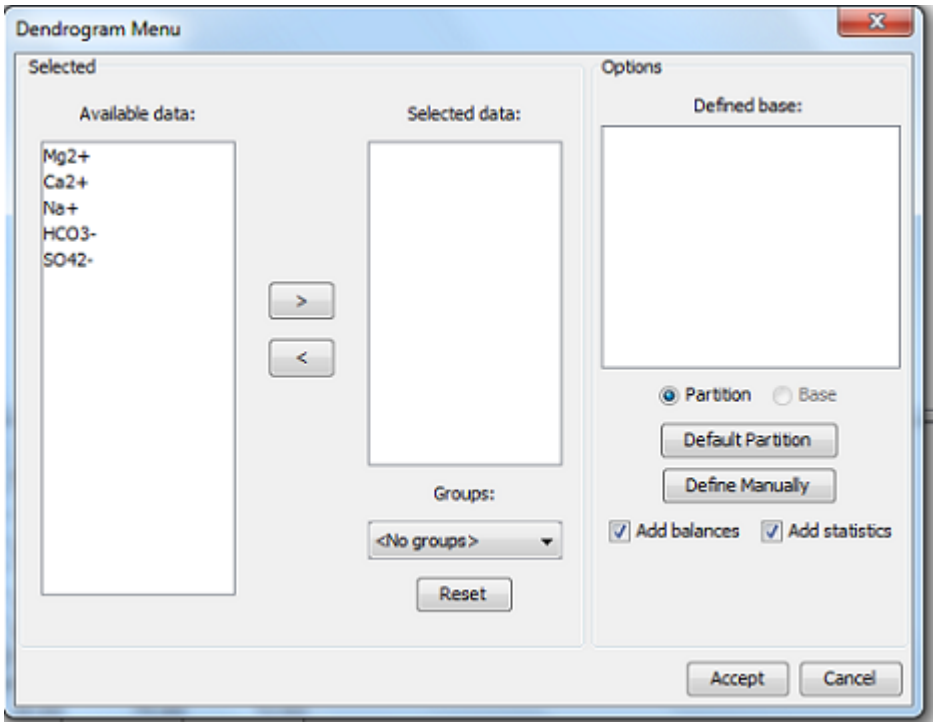
Variation array:

Variance ln(Xi/Xj)

Xi\Xj	Mg2+	Ca2+	Na+	HCO3-	SO42-	clr variances
Mg2+		1.0495	3.2681	0.5876	2.1272	0.7032
Ca2+	0.8316		2.5157	1.6147	1.7746	0.6955
Na+	-0.2018	-1.0334		3.3509	4.7398	1.3874
HCO3-	2.3205	1.4889	2.5223		2.8407	0.8394
SO42-	-0.6178	-1.4494	-0.4160	-2.9383		1.1482
Mean ln(Xi/Xj)						4.7738 **Total Variance**

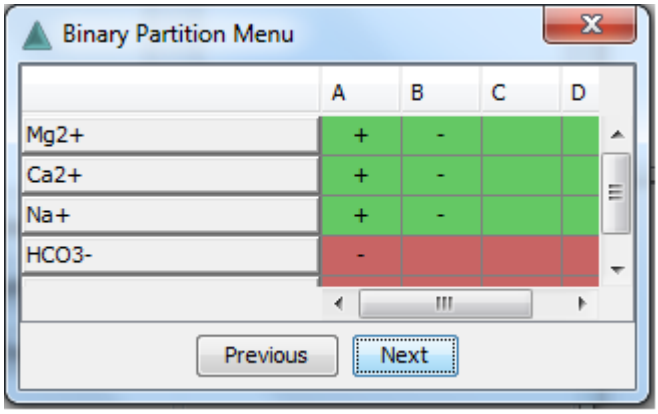
Pro vytvoření dendrogramu v menu Graphs vybereme Balance Dendrogram.

V menu, které se nám objeví, si můžeme sami určit, které složky chceme do dendrogramu zahrnout a také si sami nadefinujeme postupné binární dělení.



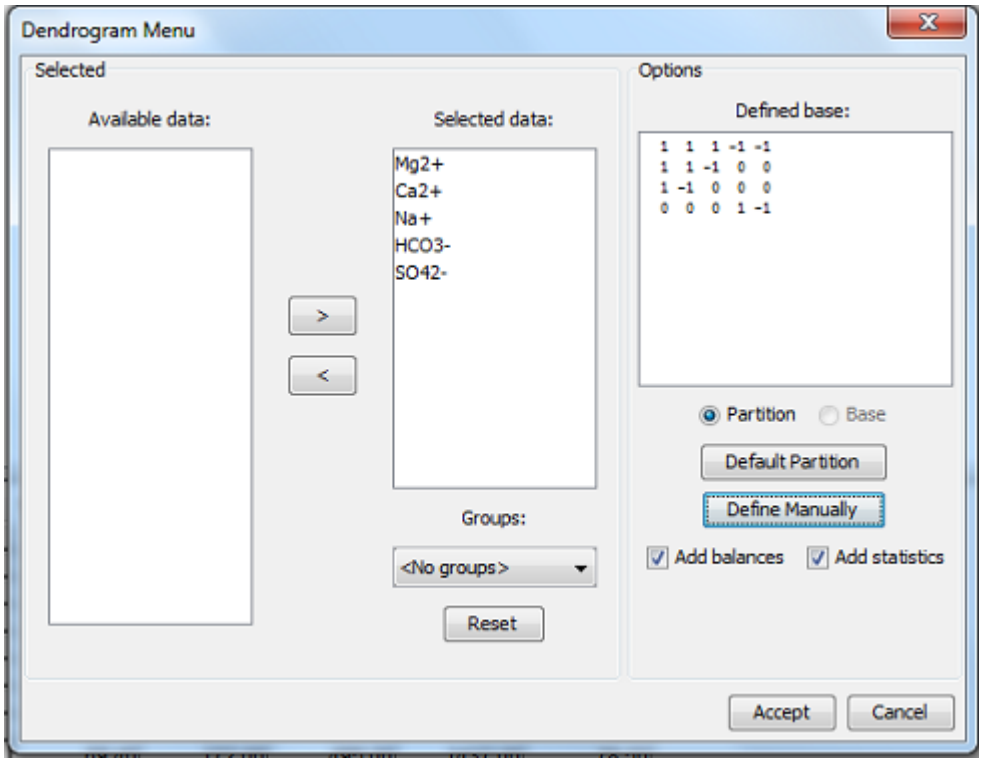
V zobrazené tabulce označíme složku, kterou chceme vybrat a klikneme na šipku směřující do vedlejší kolonky. To učiníme pro každou složku zvlášť. Po načtení všech požadovaných složek, které má dendrogram obsahovat, můžeme konečně určit postupné binární dělení. Máme možnost si vybrat, buďto si ho nadefinujeme sami (tlačítko Define Manually) nebo ho necháme provést programem tak, že se oddělí v každém kroku dělení vždy první složka (tlačítko Default Partition). Pro naše potřeby si postupné binární dělení nadefinujeme sami. Klikneme na tlačítko Define Manually. Objeví se tabulka (barevně rozlišená), ve které je na začátku u každého prvku mínus, které změníme na plus pouhým kliknutím do potřebného rámečku. Ve chvíli kdy máme pomocí plus vytvořenou první skupinu, klikneme na Next. Zeleně se zabarví skupina označená plusy v původním kroku, tato skupina bude dále dělena. Ve chvíli kdy je první skupina rozdělena až na jednotlivé složky, zabarví se červeně a zeleně se naopak zabarví skupina v prvním pořadí označena mínusy. Opět měníme znaménka a klikáme na Next, dokud není skupina rozdělena na jednotlivé složky. Po rozdělení klikneme na Next,

tabulka zbělá, zkontrolujeme zda je postupné binární dělení dle našich představ (v případě, že ne, stačí kliknout na Previous a zadání změnit) a klikneme na Accept.

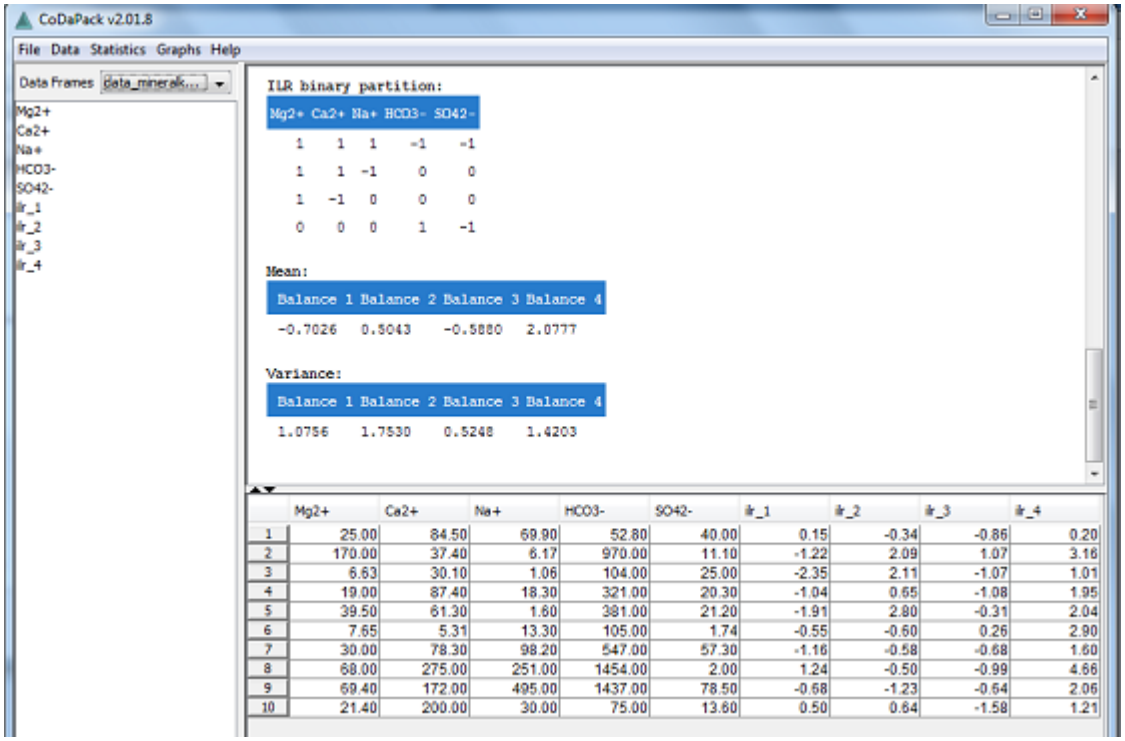


V postupném binárním dělení jsme tedy nejprve oddělili kationty od aniontů a pokračovali jsme dělením kationtů. Nejprve jsme oddělili Na^{+} od Mg^{2+} a Ca^{2+}, důvodem je to, že kationt sodný patří mezi alkalické kovy a zbylé dva kationty mezi kovy alkalických zemin. Ve třetím pořadí tedy jen oddělíme kationt hořečnatý od kationtu vápenatého a v posledním pořadí od sebe oddělíme anionty.

Po nadefinování postupného binárního dělení máme menu Dendrogramu doplněné o potřebná data a můžeme si nechat dendrogram vytvořit pouhým kliknutím na Accept.

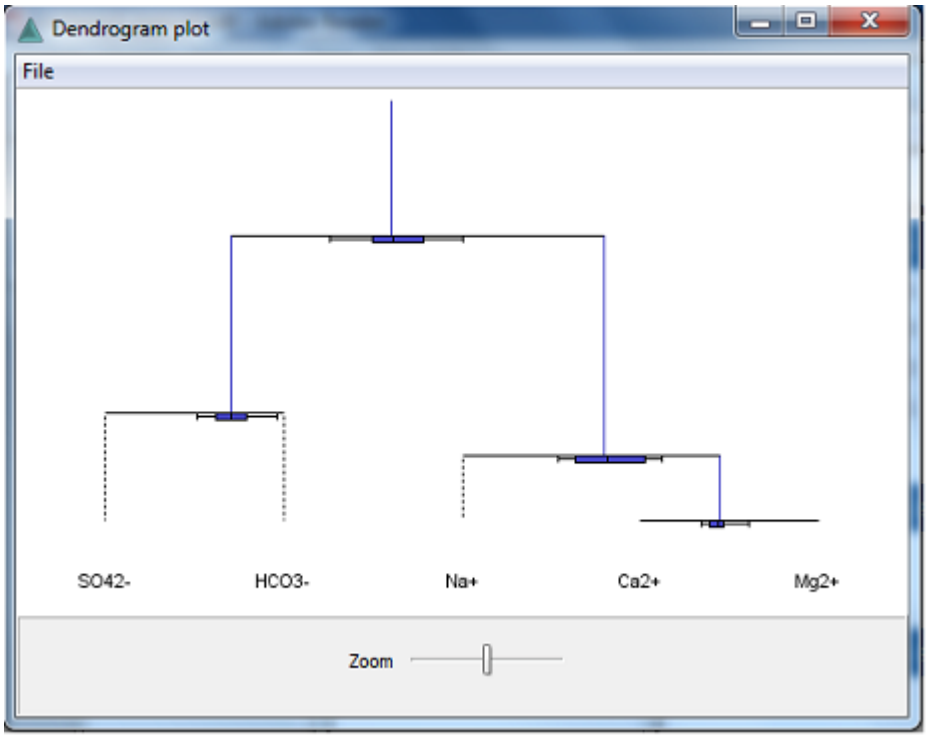


Výsledkem je jednak zobrazený dendrogram, jednak tabulka obsahující výběrové průměry jednotlivých bilancí, rozptyly a vypočtené hodnoty bilancí.



	Mg2+	Ca2+	Na+	HCO3-	SO42-	ilr_1	ilr_2	ilr_3	ilr_4
1	25.00	84.50	69.90	52.80	40.00	0.15	-0.34	-0.86	0.20
2	170.00	37.40	6.17	970.00	11.10	-1.22	2.09	1.07	3.16
3	6.63	30.10	1.06	104.00	25.00	-2.35	2.11	-1.07	1.01
4	19.00	87.40	18.30	321.00	20.30	-1.04	0.65	-1.08	1.95
5	39.50	61.30	1.60	381.00	21.20	-1.91	2.80	-0.31	2.04
6	7.65	5.31	13.30	105.00	1.74	-0.55	-0.60	0.26	2.90
7	30.00	78.30	98.20	547.00	57.30	-1.16	-0.58	-0.68	1.60
8	68.00	275.00	251.00	1454.00	2.00	1.24	-0.50	-0.99	4.66
9	69.40	172.00	495.00	1437.00	78.50	-0.68	-1.23	-0.64	2.06
10	21.40	200.00	30.00	75.00	13.60	0.50	0.64	-1.58	1.21

Na tomto obrázku můžeme vidět postupné binární dělení, dále průměry jednotlivých bilancí a také jejich rozptyly. Už zde si můžeme všimnout, že největší rozptyl je u druhé bilance, viditelné je to také u dendrogramu. Navíc se nám v tabulce dat objevily ilr souřadnice.

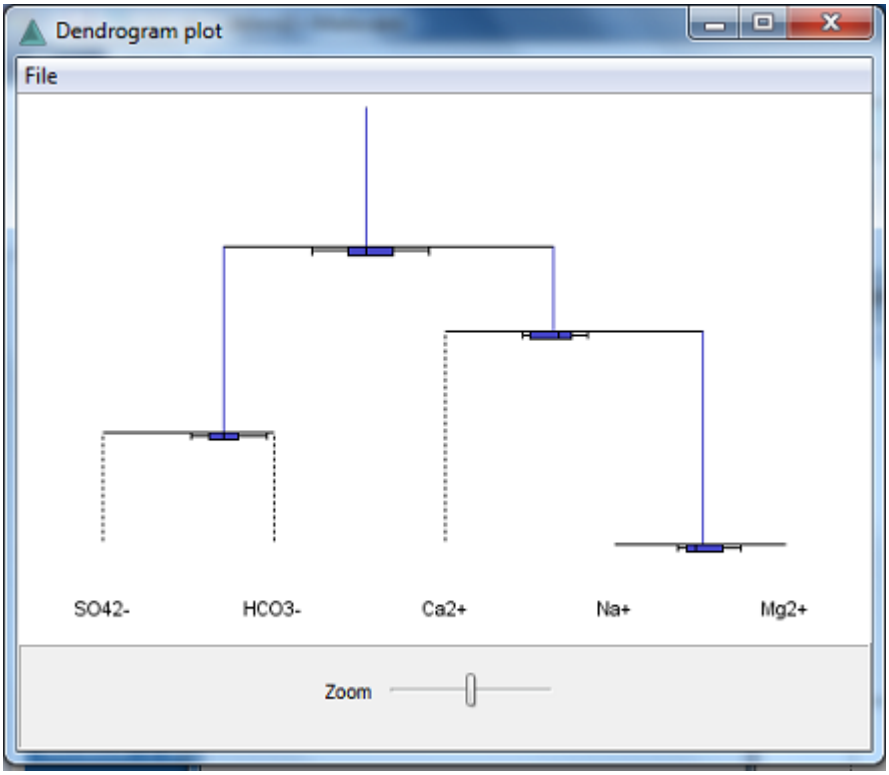


Obrázek 1: Dendrogram č.1

Z obrázku 1 je patrné, že největší variabilita je u druhé bilance, ačkoliv bychom spíše intuitivně očekávali, že největší variabilitu bude obsahovat první bilance. Předpoklad plyne z toho, že oddělujeme kationty od aniontů. Největší rozdíl v datech je ovšem mezi Na^{+} na jedné straně a Mg^{2+} a Ca^{2+} na straně druhé. Druhý největší rozptyl má pak bilance aniontů. Rozptyl je pravděpodobně ovlivněn rozdílností minerálních pramenů. Každý pramen je totiž bohatší na různé prvky. Je samozřejmé, že k větší vypovídací hodnotě provedené analýzy by také prospělo zvětšení rozsahu výběru.

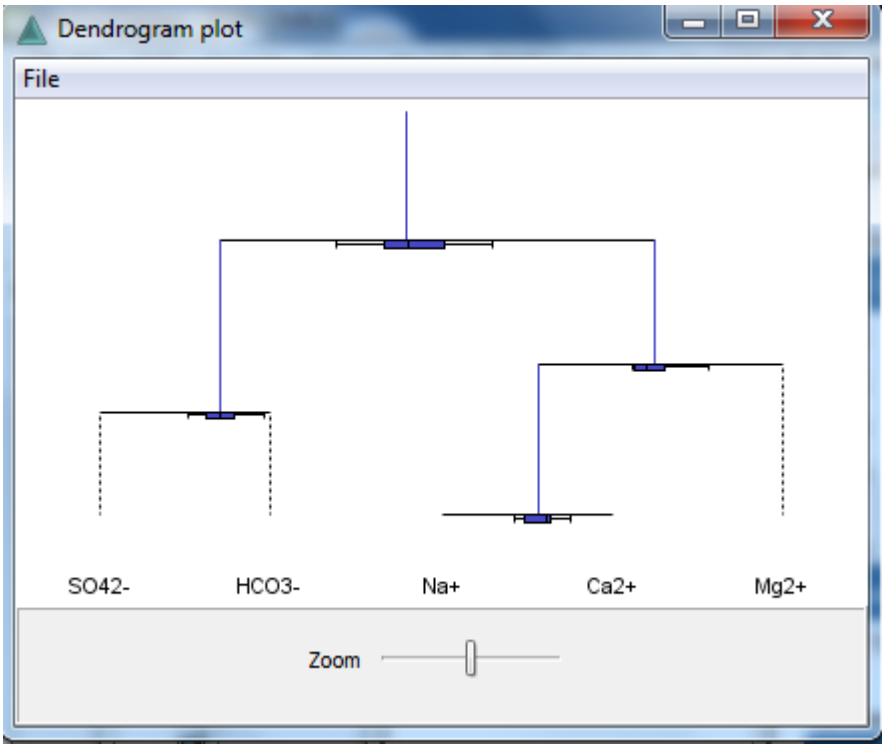
Pro ukázku jsme vytvořili ještě další dendrogramy. První dva se liší pouze v tom, který prvek kationtů byl oddělen jako první. Nejprve jsme tedy oddělili Ca^{2+} od Na^{+} a Mg^{2+} (obrázek 2). Z dendrogramu je patrné, že největší rozptyl se přesunul z druhé bilance ke třetí bilanci, čili největší variabilita je mezi Na^{+} a Mg^{2+}.

Ve druhém dendrogramu (obrázek 3) byl jako první oddělen kationt hořečnatý od ostatních. Toto postupné binární dělení se zdá být nejméně přehledné. Všechny rozptyly jsou víceméně stejné a dendrogram nám neukazuje, jak se jednotlivá data liší.



Obrázek 2: Dendrogram č.2

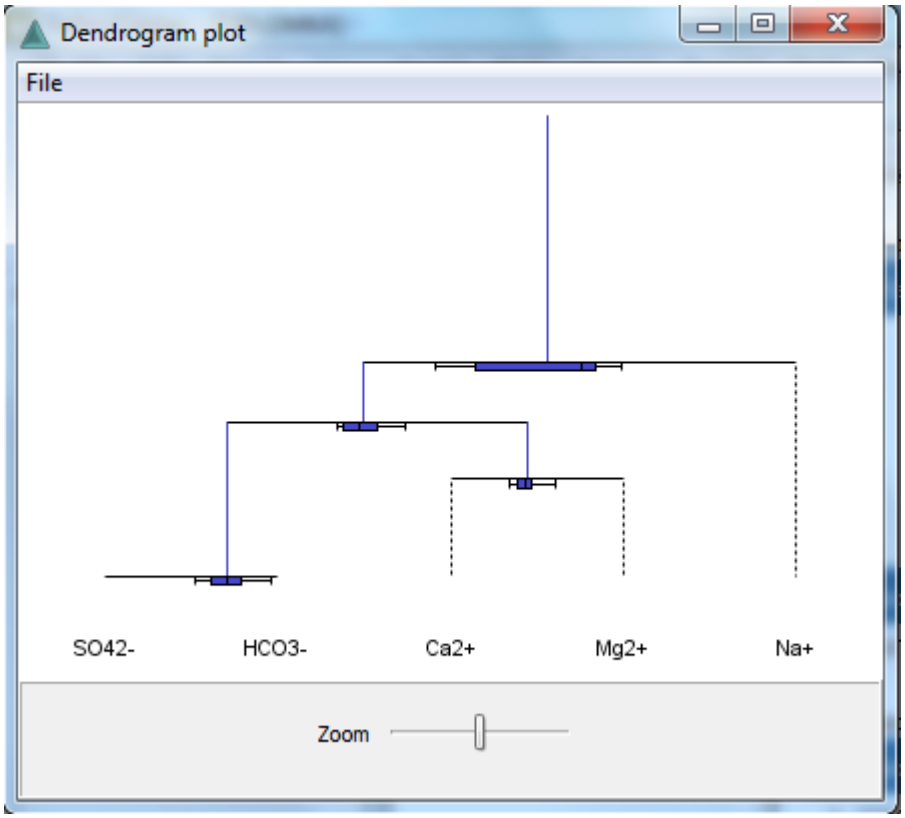
Z dendrogramů vyplývá zejména to, že největší variabilita je mezi prvky Na^{+} a Mg^{2+}, zatímco nejmenší variabilita je mezi prvky Mg^{2+} a Ca^{2+}. To je pravděpodobně způsobeno chemickými vazbami mezi uvedenými prvky s praktickým důsledkem, že pro spravné vstřebání hořčíku do těla je zapotřebí vápník, bez vápníku by se hořčík bez účinku vyloučil. Poznamenejme přitom, že úsudek



Obrázek 3: Dendrogram č.3

o variabilitě jednotlivých logaritmů podílů u skupiny kationtů bychom si byli schopni udělat již z jedné volby postupného binárního dělení užitím interpretace kovarianční struktury bilancí, jak byla popsána v předchozí kapitole.

Vzhledem k výše uvedené rozdílnosti mezi Na^+ a dalšími kationty uvedeme ještě dendrogram, ve kterém byl v prvním kroku oddělen kationt sodný od ostatních, následuje ve druhém kroku oddělení zbývajících kationtů od aniontů a potom samozřejmě zvlášť oddělujeme kationty a anionty (obrázek 4). Výsledkem je dendrogram bilance mezi kationtem sodným a ostatními složkami kompozice, která má zároveň největší rozptyl. Můžeme si všimnout velmi stabilní bilance mezi kationtem hořečnatým a vápenatým, což svědčí o výše uvedeném tvrzení o chemických vazbách mezi těmito prvky. Druhý největší rozptyl, tak jako i v předešlých dendrogramech, má bilance aniontů, což vypovídá o rozdílnosti obsahu těchto prvků v jednotlivých minerálkách.



Obrázek 4: Dendrogram č.4

Jak si můžeme všimnout, dendrogram je velmi dobrým nástrojem pro interpretaci postupného binárního dělení, jednotlivých bilancí a jejich rozptylů. Vzhledem k jednoduchosti zobrazení dendrogramu si můžeme vyzkoušet všechny možnosti dělení a vybrat si tu nejlépe odpovídající potřebám analytika a interpretaci analýzy.

Závěr

V práci jsme se zabývali tím, jak kompoziční data převést ze simplexu do reálného prostoru za použití ortonormálních bazí, abychom pak na data mohli aplikovat standardní statistické metody. Pro nalezení ortonormální báze bylo použito postupné binární dělení. Postupné binární dělení nám pomohlo i při tvorbě souřadnic, podkompozic a bilancí. Hlavní důraz byl v této práci kladen na konstrukci a interpretaci bilancí, k čemuž nám posléze pomohl i bilanční dendrogram. Snahou bylo veškerou teorii ukázat na jednoduchém příkladu, což se, jak doufám, podařilo.

Díky této práci jsem se seznámila s novým odvětvím statistické analýzy, začala jsem se orientovat v Aitchisonově geometrii a nakonec jsem pochopila i bilance a to, co obnášejí a znamenají pro analýzu kompozičních dat. Naučila jsem se pracovat s programem CoDaPack a zobrazovat bilance pomocí bilančního dendrogramu, což je mnohem názornější, než se dívat na nicneříkající čísla. Nejtěžší na všem bylo zorientovat se v úplně novém prostoru a geometrii a následně pochopit, co bilance znázorňují. Překážkou pro celkové pochopení byly nedostatečné znalosti z oblasti lineární algebry a samotné nastudování a pochopení teorie psané anglickým jazykem. Cením si zejména nově nabytých znalostí, které využiji v dalším studiu, a rovněž velké trpělivosti a pomoci svého vedoucího.

Literatura

[1] Aitchison, J., *The Statistical Analysis of Compositional Data*. Chapman and Hall, London, New York, 1986.

[2] Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., *Isometric Logratio Transformations for Compositional Data Analysis*. Mathematical Geology 35, 2003, 279-300.

[3] Fišerová, E., Hron, K., *On interpretation of orthonormal coordinates for compositional data*. Mathematical Geosciences 43, 2011, 455-468.

[4] manuál programu CoDaPack [online], dostupné z: http://ima.udg.edu/codapack/assets/codapack-manual.pdf [citováno 15.1.2012]

[5] minerální prvky [online], dostupné z: http://www.pijtezdravouvodu.cz/o-vode/jakou-vodu-pit/ [citováno 10.1.2012]

[6] minerální prvky [online], dostupné z: http://cs.wikipedia.org/ [citováno 10.1.2012]

[7] Gram-Schmidtova ortogonalizace [online], dostupné z: http://nptel.iitm.ac.in/courses/Webcourse-contents/IIT-KANPUR/mathematics-2/node49.html [citováno 15.1.2012]

[8] Pawlowsky-Glahn, V., Egozcue, J. J., *Groups of Parts and Their Balances in Compositional Data Analysis*. Mathematical Geology 37, 2005, 795-828.

[9] Pawlowsky-Glahn, V., Egozcue, J. J., *Simplicial geometry for compositional data*. In: Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V., *Compositional Data Analysis in the Geosciences From Theory to Practice*. Geological Society, London, 2006, 145-159.

[10] Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, R., *Lecture Notes on Compositional Data Analysis* [online], dostupné z: http://dugi-doc.udg.edu/bitstream/10256/297/1/CoDa-book.pdf [citováno 15.8.2011]

[11] Pawlowsky-Glahn, V., Mateu-Figueras, G., *A Critical Approach to Probability Laws in Geochemistry*. Mathematical Geosciences 40, 2008, 489-502.

[12] Pawlowsky-Glahn, V., Mateu-Figueras, G., Egozcue, J. J., *The principal of working on coordinates*. In: *Compositional Data Analysis. Theory and Aplications*. Wiley, Chichester, 2011, 33-43.

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

7 Přílohy

7.1 Gram-Schmidtova ortogonalizace

Tato věta byla převzata ze zdroje [7].

Věta 7.1. *Nechť $\mathcal{V}$ je prostor se skalárním součinem. Předpokládejme dále, že $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ je množina lineárně nezávislých vektorů z $\mathcal{V}$. Pak existuje množina vektorů $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ z $\mathcal{V}$, která splňuje:*

1. $\|\mathbf{v}_i\| = 1$ *pro* $i = 1, 2, \ldots, n$
2. $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0$ *pro* $1 \leq i, j \leq n, i \neq j$
3. *Množiny vektorů* $\mathbf{v}_1, \ldots, \mathbf{v}_i$ *a* $\mathbf{u}_1, \ldots, \mathbf{u}_i$ *generují stejné podprostory pro* $i = 1, 2, \ldots, n$.

Věta popisuje vlastnosti nezávislých vektorů v libovolném vektorovém prostoru se skalárním součinem, včetně Aitchisonovy geometrie. Pokud tedy místo euklidovské geometrie použijeme Aitchisonovu, dostaneme právě ortonormální bázi simplexu.

7.2 Logaritmicko-normální rozdělení

Náhodný vektor $\mathbf{x}$ má logaritmicko-normální (log-normální) rozdělení s parametry $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$, jestliže má hustotu pravděpodobnosti

$$f(\mathbf{x}) = (2\pi)^{-(D-1)/2} \left|\boldsymbol{\Sigma}\right|^{-1/2} \exp\left[(-\frac{1}{2}(\ln \mathbf{x} - \boldsymbol{\mu}) \left|\boldsymbol{\Sigma}\right|^{-1} (\ln \mathbf{x} - \boldsymbol{\mu})\right], \mathbf{x} \in \mathbb{R}_+.$$

7.3 Logistické normální rozdělení

Logistické normální rozdělení na simplexu definoval John Aitchison v knize [1], a to následovně: transformujeme náhodnou kompozici ze simplexu do reálného prostoru, zde definujeme funkci hustoty transformovaných dat a nakonec se vrátíme zpět do simplexu užitím transformace proměnných. Výsledkem je funkce hustoty pro výchozí náhodnou kompozici vzhledem k Lebesgueově míře.

$$f(\mathbf{x}) = \frac{(2\pi)^{-(D-1)/2} |\boldsymbol{\Sigma}|^{-1/2}}{\sqrt{D} x_1 x_2 \cdots x_D} \exp\left[-\frac{1}{2} \left(\mathrm{ilr}(\mathbf{x}) - \boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1} \left(\mathrm{ilr}(\mathbf{x}) - \boldsymbol{\mu}\right)\right],$$

kde $\mathbf{x}$ je kompozice, $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$ jsou parametry rozdělení.

7.4 Normální rozdělení na simplexu

Užitím algebraicko-geometrické struktury na simplexu a Aitchisonovy míry λ_a, definujeme normální rozdělení na $\mathcal{S}^D$ prostřednictvím funkce hustoty ortonormálních souřadnic jako

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

Vidíme, že je to obvyklá hustota normálního rozdělení aplikovaná na ilr-souřadnice. Tudíž je to hustota vzhledem k Lebesgueově míře v prostoru souřadnic $\mathbb{R}^{D-1}$. Tedy je to hustota na $\mathcal{S}^D$ vzhledem k míře λ_a.

Výše uvedená rozdělení byla převzata z literatury [11].