UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Korelační analýza kompozičních dat v souřadnicích

Katedra matematické analýzy a aplikací matematiky
Vedoucí diplomové práce: **doc. RNDr. Karel Hron, Ph.D.**
Vypracovala: **Klára Juráňová**
Studijní program: N1103 Aplikovaná matematika
Studijní obor Aplikace matematiky v ekonomii
Forma studia: prezenční
Rok odevzdání: 2019

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Klára Juráňová

Název práce: Korelační analýza kompozičních dat v souřadnicích

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: doc. RNDr. Karel Hron, Ph.D.

Rok obhajoby práce: 2019

Abstrakt: Cílem práce je představit a porovnat různé přístupy ke korelační analýze kompozičních dat. První kapitola shrnuje základní myšlenky kompozičních dat společně s problémy, které se objevují při použití standardní korelační analýzy. Druhá kapitola ukazuje různé metody založené na převedení dat na nové souřadnice. Ve třetí části jsou metody diskutovány a použity na datech týkajících se fyzické aktivity dětí během dne.

Klíčová slova: korelační analýza, kompoziční data, souřadnice, symetrické pivotové souřadnice, b-asociace

Počet stran: 60

Počet příloh: 0

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Klára Juráňová

Title: Correlation analysis of compositional data in coordinates

Type of thesis: Master's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: doc. RNDr. Karel Hron, Ph.D.

The year of presentation: 2019

Abstract: The goal of the diploma thesis is to introduce and compare different approaches to the correlation analysis of compositional data. The first chapter sums up concepts of compositional data together with problems occurring when the standard correlation analysis is used. The second chapter shows different methods based on transforming data into new coordinates. In the third part, methods are discussed and used on data about the physical activity of children during the day.

Key words: correlation analysis, compositional data, coordinates, symmetric balances, b-association

Number of pages: 60

Number of appendices: 0

Language: Czech

Prohlášení

Prohlašuji, že jsem bakalářskou práci zpracovala samostatně pod vedením doc. RNDr. Karla Hrona, Ph.D. a všechny použité zdroje jsem uvedla v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Poděkování

Ráda bych na tomto místě poděkovala vedoucímu své diplomové práce panu doc. RNDr. Karlu Hronovi, Ph.D. za odborné vedení práce, cenné rady a čas strávený při konzultacích. Rovněž bych chtěla poděkovat doc. Mgr. Alešovi Gábovi, Ph.D. z Fakulty tělesné kultury UP za poskytnutí dat pro praktický příklad.

Úvod

V mnohorozměrné statistické analýze dat pracujeme s pozorováními více náhodných proměnných zároveň. Nabízí se tedy možnost studovat vztahy mezi proměnnými, a objevit tak skrytou strukturu dat. Analýza závislostí mezi proměnnými má uplatnění v nejrůznějších oborech lidské činnosti. Asi nejznámější charakteristikou vztahu dvou veličin je Pearsonův korelační koeficient vyjadřující míru těsnosti lineárního vztahu mezi dvěma veličinami. Avšak už sám Karl Pearson zmiňuje některé nevýhody tohoto koeficientu v určitých situacích.

Jedním z případů, kdy s jednoduchou aplikací korelačního koeficientu na pozorovaná data neuspějeme, je práce s kompozičními daty. Kompozičními nazýváme data obsahující relativní informaci. Snadno si je představíme jako procenta či proporce z daného celku. Zvětšíme-li proporci jedné složky mnohorozměrného pozorování, činíme tak na úkor ostatních složek. Korelace mezi složkami se tímto zkresluje, neboť je tlačena do záporných hodnot.

Proto v korelační analýze kompozičních dat převádíme data na speciální souřadnice ve tvaru logaritmů podílů jejich složek. Pokud transformaci zkonstruujeme speciálním způsobem, získáme dobře interpretovatelné souřadnice (tzv. *bilance*) s výhodnými vlastnostmi. Práce dále uvádí dva přístupy k měření asociace mezi složkami kompozičních dat. Prvním je použití *symetrických pivotových souřadnic*, výpočet klasického korelačního koeficientu a odpovídající interpretace hodnot. Druhý přístup přináší nové míry asociace mezi složkami vycházející z lineárního modelu mezi bilancemi.

Třetí kapitola demonstruje použití metod na reálném datovém souboru. Data se týkají rozložení míry fyzické aktivity u dětí během dne a byla naměřena na Fakultě tělesné kultury UP. Analýza odhaluje vztahy mezi složkami kompozičního datového souboru a zároveň ukazuje výhody a nevýhody jednotlivých metod, což je hlavním cílem diplomové práce.

1 Kompoziční data

Na začátku studia korelační analýzy kompozičních dat je třeba se dobře seznámit s kompozičními daty jako takovými. K tomu mohou posloužit například knihy [3], [17], [18], na kterých je úvodní kapitola postavena.

Kompoziční data jsou mnohorozměrná pozorování, jejichž složky popisují jednotlivé části celku. Již z definice 1.1 [18] vyplývá, že předmětem zájmu je relativní informace obsažená v datech. Podstatné jsou tedy podíly mezi složkami kompozičních dat a samotná hodnota složek je irelevantní. Jako příklad kompozičních souborů dat lze uvést rozložení volného času lidí mezi jednotlivé aktivity či složení živin v jejich stravě.

Definice 1.1 *Kompozicí nazveme vektor D kladných reálných složek $\boldsymbol{x} = [x_1, \ldots, x_D]$, který nese pouze relativní informaci.*

Někdy se v definici uvádí rovněž podmínka na konstantní součet složek κ. Ta ovšem nemusí být splněna, protože třeba v uvedených příkladech mají jednotliví lidé různé množství volného času nebo zkonzumují různé množství potravy. Nicméně lze vždy kompoziční data do takového tvaru převést tzv. *uzavřením*. Často bývá konstantní součet roven 1 nebo 100, pak hovoříme o *proporcích* a *procentech*. V případě uzavřených kompozic je jejich výběrovým prostorem simplex.

Definice 1.2 *D-složkovým simplexem nazveme množinu*

$$\mathbb{S}^D := \left\{ \boldsymbol{x} = [x_1, \cdots, x_D] : x_i > 0, \sum_{i=1}^{D} x_i = \kappa \right\}.$$

Z podstaty kompozičních dat vycházejí tři základní principy pro jejich analýzu. Jelikož je informace obsažená v kompozicích pouze relativní povahy, tak se vynásobením libovolnou kladnou konstantou nezmění. Kompoziční data jsou tedy *invariantní vůči změně měřítka*. Dále nezáleží na pořadí složek v kompozici, což se nazývá *invariance vůči permutaci*. Nakonec zbývá zásada *podkompoziční soudržnosti*. Podkompozicí chápeme kompozici, která je podvektorem jiné

kompozice. Jestliže je splněna podkompoziční soudržnost, potom nedosáhneme protichůdných závěrů analýzy týkající se složek společných pro kompozici i její podkompozici, ať už pracujeme s celou kompozicí nebo právě pouze s její podkompozicí.

1.1 Aitchisonova geometrie

Protože operace standardního euklidovského vektorového prostoru nesplňují uvedené principy analýzy kompozičních dat, byly zavedeny speciální operace pro kompoziční data [18]:

Definice 1.3 *Nechť jsou dány D-složkové kompozice* $\boldsymbol{x} = [x_1, \ldots, x_D]$, $\boldsymbol{y} = [y_1, \ldots, y_D]$ *a* $\alpha \in \mathbb{R}$. *Operaci perturbace, mocninné transformace, a dále Aitchisonův skalární součin definujeme v tomto pořadí jako*

$$\boldsymbol{x} \oplus \boldsymbol{y} = \mathcal{C}[x_1 \cdot y_1, \ldots, x_D \cdot y_D],$$

$$\alpha \odot \boldsymbol{x} = \mathcal{C}[x_1^\alpha, \ldots, x_D^\alpha],$$

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_A = \frac{1}{D} \sum_{i>j}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j},$$

kde $\mathcal{C}$ značí operaci uzavření. Jmenované operace dohromady tvoří euklidovskou geometrii na simplexu, která nese také přívlastek Aitchisonova. Aitchisonův skalární součin implikuje i příslušnou Aitchisonovu normu $\|\boldsymbol{x}\|_A = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_A}$ a vzdálenost $d_A(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} \ominus \boldsymbol{y}\|_A$.

1.2 Falešná korelace

John Aitchison ve známé knize [1] nebyl první, kdo upozornil na nutnost přistupovat ke kompozicím speciálně. Již Karl Pearson uvedl ve svém článku z roku 1897 [19] fenomén, který nazval *falešnou korelací* (z anglického *spurious correlation*) a který je ve svém důsledku popsán níže.

Korelace vyjadřuje míru lineárního vztahu mezi dvěma náhodnými veličinami a nejčastěji ji měříme pomocí *korelačního koeficientu*, který je definován následovně jako normovaná kovariance na interval $\langle -1, 1\rangle$ či kovariance mezi normovanými veličinami s nulovou střední hodnotou a jednotkovým rozptylem [14].

Definice 1.4 *Nechť náhodné veličiny X a Y mají konečné druhé momenty, střední hodnoty $E(X)$, $E(Y)$ a rozptyly* $\mathrm{var}(X) \neq 0$, $\mathrm{var}(Y) \neq 0$. *Korelační koeficient $\rho_{X,Y}$ náhodných veličin X a Y je číslo definované vztahem*

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X) \cdot \mathrm{var}(Y)}} = \mathrm{cov}\left(\frac{X - E(X)}{\sqrt{\mathrm{var}(X)}}, \frac{Y - E(Y)}{\sqrt{\mathrm{var}(Y)}}\right).$$

V praxi pak tuto charakteristiku odhadujeme použitím *výběrového korelačního koeficientu* mezi statistickými znaky X a Y. Korelační koeficient je symetrický a normovaný na interval $\langle -1, 1\rangle$. Čím více se hodnota koeficientu blíží k 1, popř. -1, tím je lineární vztah mezi znaky těsnější. Hodnota 0 značí, že mezi znaky není žádná lineární závislost (nevylučuje to však závislost jiného charakteru).

Pearson [19] se zabýval korelací mezi dvěma podíly a zmínil příklad, kdy vezmeme tři náhodná po dvou nekorelovaná čísla (veličiny) x, y a z. Pokud pro každou takovou trojici vytvoříme podíly x/y a z/y, nalezneme mezi těmito zlomky korelaci, ačkoliv původní čísla byla nekorelovaná. Bylo by tedy chybou z nalezené korelace usuzovat závěry.

Pro lepší představu si demonstrujeme problém na reálných datech, která nazveme `activity`, protože se týkají rozložení pohybové aktivity během dne u dětí od 7 do 13 let. Jejich podrobný popis lze nalézt v podkapitole 3.1. Kompozice obsahuje čtyři složky: sedavá aktivita (`sedentary`), dále pak lehká (`light`), středně náročná (`moderate`) a náročná (`vigorous`) pohybová aktivita. Původní jednotkou jsou minuty za den. Převedeme data tak, aby se součet složek rovnal jedné. Provedeme tedy operaci uzávěru a spočteme standardní korelační koeficient pro každou dvojici proměnných, čímž získáme *korelační matici* z tabulky 1.1.

	sedentary	light	moderate	vigorous
sedentary	1.000	-0.935	-0.502	-0.187
light	-0.935	1.000	0.200	-0.089
moderate	-0.502	0.200	1.000	0.401
vigorous	-0.187	-0.089	0.401	1.000

Tabulka 1.1: Výběrová korelační matice kompozice obsahující všechny čtyři složky.

Pokud se rozhodneme vyloučit sedavou aktivitu, vytvoříme podkompozici obsahující zbývající tři složky. Tu opět uzavřeme, aby byl součet složek 1, dostaneme se ke korelační matici v tabulce 1.2.

	light	moderate	vigorous
light	1.000	-0.874	-0.765
moderate	-0.874	1.000	0.356
vigorous	-0.765	0.356	1.000

Tabulka 1.2: Výběrová korelační matice kompozice obsahující pouze tři složky.

Vidíme, že v druhém případě dostáváme pro stejné dvojice složek naprosto odlišné hodnoty. Korelace tedy nezávisí pouze na dvojici uvažovaných složek, pro které ji počítáme, ale také na všech ostatních složkách v kompozici. To se vylučuje s definicí korelačního koeficientu, protože ten hovoří pouze o vztahu dvou statistických znaků.

Na Pearsonovy myšlenky navázal až v roce 1960 geolog Felix Chayes [5], který upozornil, že pokud nezávislé složky uzavřeme na konstantní součet, dostaneme mezi nimi zápornou korelaci, tzv. *negativní vychýlení*. Protože pro D-složkovou kompozici $[x_1, \ldots, x_D]$ se součtem složek $x_1 + \cdots + x_D = 1$ platí

$$\mathrm{cov}(x_1, x_1 + \cdots + x_D) = 0,$$

lze odvodit vztah

$$\mathrm{cov}(x_1, x_2) + \cdots + \mathrm{cov}(x_1, x_D) = -\mathrm{var}(x_1).$$

Pravá strana je kromě případu konstantní první složky vždy záporná. Tudíž alespoň jedna z kovariancí na levé straně musí být záporná. Proto se korelace

nemohou volně pohybovat na intervalu $\langle -1, 1\rangle$ jako u mnohorozměrných dat bez omezení [2]. Dalším problémem omezení na konstantní součet je singularita varianční matice, neboť některé mnohorozměrné statistické metody používají její inverzi.

1.3 Logpodílové souřadnice

Z předchozích poznatků plyne, že k měření korelace pro kompoziční data nemůžeme použít absolutní hodnoty jejich původních složek. Aitchison [1] proto navrhl vyjádřit kompoziční data ve tvaru logaritmů podílů složek, tzv. *logpodílových souřadnicích*. Veškerá informace je totiž obsažena v podílech složek kompozic a logaritmus zobrazuje a symetrizuje kladné podíly na celou reálnou osu, aby případně bylo možno použít metody založené na normálním rozdělení dat.

1.3.1 Alr souřadnice a clr koeficienty

John Aitchison nejprve přišel s *aditivními logpodílovými (alr) souřadnicemi*:

$$\mathrm{alr}(\boldsymbol{x}) = (\ln(x_i/x_D))_{i=1,\dots,D-1},$$

které ovšem nezachovávají vzdálenosti, tedy nejsou izometrické se zavedenou Aitchisonovou geometrií (viz podkapitola 1.1). Podmínku izometrie už ale splňují *centrované logpodílové (clr) koeficienty*:

$$\mathrm{clr}(\boldsymbol{x}) = \left(\ln \frac{x_i}{g(\boldsymbol{x})}\right)_{i=1,\dots,D}, \quad \text{kde } g(\boldsymbol{x}) = \sqrt[D]{x_1 \cdot x_2 \cdots x_D}.$$

Clr koeficienty ale vedou k singulární varianční matici, což je problém, pokud je třeba získat její inverzi.

1.3.2 Ilr souřadnice a bilance

Kvůli nedostatkům výše uvedených souřadnicových systémů byly zkonstruovány tzv. *izometrické logpodílové (ilr) souřadnice*. Jak již název napovídá, jedná

se o izometrické zobrazení mezi euklidovskými vektorovými prostory na $\mathbb{S}^D$ a $\mathbb{R}^{D-1}$, které navíc vede ke vzniku ortonormálních souřadnic (vzhledem k Aitchisonově geometrii). Většinou se kvůli dobré interpretaci využívá speciální případ ilr souřadnic získaných pomocí *postupného binárního dělení*, které se nazývají *bilance*.

Během postupného binárního dělení rozdělujeme v každém kroku složky do dvou nepřekrývajících se podskupin, než dostaneme v každé podskupině pouze jednu složku. Pro D-složkovou kompozici tak získáme celkem $D-1$ bilancí. V k-tém kroku tak dělení definuje bilanci mezi dvěma podskupinami složek, kde $i_1, i_2, \cdots, i_r$ je r složek první podskupiny označené znaménkem + a $j_1, j_2, \cdots, j_s$ s složek druhé podskupiny značené pomocí $-$. Bilance je tak definovaná jako normovaný logaritmus podílu geometrického průměru složek jedné podskupiny ku geometrickému průměru podskupiny druhé. Vyjadřuje míru celkové relativní významnosti jedné skupiny složek vzhledem ke skupině druhé.

$$b_k = \sqrt{\frac{rs}{r+s}} \ln \frac{(x_{i_1} x_{i_2} \cdots x_{i_r})^{1/r}}{(x_{j_1} x_{j_2} \cdots x_{j_s})^{1/s}} = \ln \frac{(x_{i_1} x_{i_2} \cdots x_{i_r})^{a_+}}{(x_{j_1} x_{j_2} \cdots x_{j_s})^{a_-}}.$$

To znamená, že pro k-tou bilanci dostanou složky váhu buď

$$a_+ = \frac{1}{r}\sqrt{\frac{rs}{r+s}}, \quad a_- = \frac{1}{s}\sqrt{\frac{rs}{r+s}} \text{ nebo } a_0 = 0,$$

konkrétně a_+ pro složky v čitateli, a_- pro ty ve jmenovateli a a_0 pro všechny, které nebyly zahrnuty do dělení. Bilanci potom získáme jako

$$b_k = \sum_{i=1}^{D} a_{ik} \ln x_i, \tag{1.1}$$

kde se a_{ik} rovná a_+ (resp. $-a_-$), pokud v k-tém kroku postupného binárního dělení byla i-tá složka označena znaménkem + (resp. $-$). Jestliže v k-tém kroku i-tá složka nefiguruje, použije se váha a_0. Výsledné bilance můžeme uspořádat do vektoru $\text{ilr}(\boldsymbol{x}) = [b_1, \ldots, b_{D-1}]$.

Bilance je lineární kombinací logaritmů složek a pro součet koeficientů platí

$$r \cdot \frac{1}{r}\sqrt{\frac{rs}{r+s}} - s \cdot \frac{1}{s}\sqrt{\frac{rs}{r+s}} = 0,$$

je tedy speciálním případem tzv. *logkontrastu* (ostatně jako i další uvedené log-podílové souřadnice). Matice obsahující složky a_{ik} pak tvoří ortonormální bázi pro $i = 1, \ldots, D$ a $k = 1, \ldots, D-1$ (v reálném prostoru potažmo na simplexu).

Koncept bilancí hraje v korelační analýze kompozičních dat významnou roli, protože na něm stojí oba přístupy prezentované v kapitole 2. Proto si ukážeme celý proces sestrojení bilancí pro reálná data `activity`. Možností pro rozdělení složek do skupin je samozřejmě více, uvedeme pouze jednu vzorovou možnost. V prvním kroku sledujeme poměr sedavé složky ke všem pohybovým složkám. Příslušná bilance z_1 vyjadřuje míru relativní významnosti sedavé aktivity (složka x_1) vzhledem k průměrné pohybové aktivitě (složky x_2, x_3 a x_4)

$$z_1 = \sqrt{\frac{3}{4}} \ln \frac{x_1}{\sqrt[3]{x_2 x_3 x_4}}.$$

V dalších krocích pak rozdělujeme zbylé tři složky, než zůstane v každé skupině pouze jedna a interpretujeme analogicky. Dělení složek zapisujeme do tabulky 1.3. Ostatní bilance mají potom tvar

$$z_2 = \sqrt{\frac{2}{3}} \ln \frac{x_2}{\sqrt{x_3 x_4}}, \qquad z_3 = \sqrt{\frac{1}{2}} \ln \frac{x_3}{x_4}.$$

	Bilance		
Složka	z_1	z_2	z_3
`sedentary`	+		
`light`	−	+	
`moderate`	−	−	+
`vigorous`	−	−	−

Tabulka 1.3: Postupné binární dělení pro datový soubor `activity`.

Odpovídající ortonormální báze, zapsaná ve formě matice $\boldsymbol{V}$, vypadá následovně

$$\boldsymbol{V} = \begin{pmatrix} \sqrt{\frac{3}{4}} & 0 & 0 \\ -\frac{1}{3}\sqrt{\frac{3}{4}} & \sqrt{\frac{2}{3}} & 0 \\ -\frac{1}{3}\sqrt{\frac{3}{4}} & -\frac{1}{2}\sqrt{\frac{2}{3}} & \sqrt{\frac{1}{2}} \\ -\frac{1}{3}\sqrt{\frac{3}{4}} & -\frac{1}{2}\sqrt{\frac{2}{3}} & -\sqrt{\frac{1}{2}} \end{pmatrix}$$

a pro kompozici $\boldsymbol{x} = [x_1, \ldots, x_D]$ (uvažujeme řádkový vektor) platí

$$\mathrm{clr}(\boldsymbol{x}) \cdot \boldsymbol{V} =: \mathrm{ilr}(\boldsymbol{x}).$$

1.3.3 Vlastnosti souřadnic

Všechny tři typy logpodílových souřadnic převádí perturbaci a mocninnou transformaci na simplexu na součet a násobení skalárem v reálném prostoru. Clr koeficienty a ilr souřadnice navíc zachovávají skalární součin, jsou tedy izometrické. U ilr souřadnic nevzniká ani problém se singularitou varianční matice. Lze proto na začátku analýzy kompoziční data převést na souřadnice vzhledem k ortonormální bázi, což odpovídá ilr souřadnicím, a provést požadovanou analýzu. Poté je případně možné výsledky zobrazit zpět na kompozice a interpretovat výsledky.

Pro výpočet korelačního koeficientu ve smyslu původních kompozičních složek nejsou ale clr koeficienty a ilr souřadnice obecně vhodné. U clr koeficientů totiž narazíme na problém negativního vychýlení, protože platí, že součet clr koeficientů je konstantní (konkrétně nulový). Vektor ilr souřadnic D-složkové kompozice rovněž není ve svém obecném tvaru vyhovující, protože má pouze $D - 1$ prvků, a proto není možné jednoznačně přiřadit souřadnici ke konkrétní složce kompozice. V následující kapitole tudíž sestrojíme sofistikovanější ilr souřadnice, které jsou určené speciálně pro výpočet korelačního koeficientu, tzv. *symetrické pivotové souřadnice*. Dále pak představíme i jiné způsoby měření lineární asociace mezi složkami kompozičních dat, opět založené na bilancích.

2 Míry lineární asociace

Druhá kapitola nejdříve uvede standardní Aitchisonův přístup pro měření asociace mezi dvěma složkami kompozice pomocí *variační matice*. Poté představíme přístupy pro určení míry lineární asociace mezi složkami, a to s využitím symetrických pivotových souřadnic [15] a také pomocí *míry b-asociace* [7]. Hlavním zdrojem informací jsou články [7], [15] a knihy [3], [17] a [18].

2.1 Variační matice

Protože výpočet standardní korelační matice nedává pro původní složky kompozičních dat uspokojivé výsledky (viz podkapitola 1.2), přišel John Aitchison v [1] s vlastním způsobem měření asociace mezi dvěma složkami, která odpovídá míře proporcionality mezi složkami. Ta je vyjádřena pomocí rozptylu logaritmu podílu příslušných složek. Pokud spočteme takovou charakteristiku pro všechny kombinace složek, získáme *variační matici*.

Definice 2.1 *Variační maticí (náhodné) kompozice* $\boldsymbol{x} = [x_1, \ldots, x_D]$ *z* $\mathbb{S}^D$ *nazveme matici* D^2 *prvků, definovaných jako*

$$t_{ij} = \mathrm{var}\left(\ln \frac{x_i}{x_j}\right)$$

a odhadovaných z výběru kompozic $\boldsymbol{x}_n = [x_{n1}, \ldots, x_{nD}], \quad n = 1, \ldots, N$ *hodnotou*

$$\widehat{t}_{ij} = \left[\frac{1}{N-1}\sum_{n=1}^{N} \ln^2 \frac{x_{ni}}{x_{nj}}\right] - \frac{N}{N-1}\ln^2 \frac{\overline{x}_i}{\overline{x}_j},$$

kde N *je počet pozorování.*

Míra proporcionality mezi složkami je větší, jestliže logaritmus podílu má tendenci být konstantní. V případě vysoké míry proporcionality se tudíž hodnoty příslušného prvku matice blíží 0. Logaritmus podílu můžeme normovat tak,

aby tvořil bilanci dvousložkové kompozice $[x_i, x_j]$ a počítáme klasický výběrový rozptyl pro tyto bilance.

$$t^*_{ij} = \mathrm{var}\left(\frac{1}{\sqrt{2}} \ln \frac{x_i}{x_j}\right).$$

Z vlastnosti rozptylu vyplývá, že $t^*_{ij} = \frac{1}{2} t_{ij}$. Obě verze matice jsou symetrické, jelikož pro veličiny A, B a C s kladným oborem hodnot platí, že $\ln(A/B) = -\ln(B/A)$ a $\mathrm{var}(-C) = \mathrm{var}(C)$. Na diagonále obou matic se nachází nuly.

Pokud chceme prvky variační matice normovat na interval $(0, 1\rangle$, vypočteme její prvky následovně [4],

$$\tau_{ij} = \exp\left(-\mathrm{var}\left(\frac{1}{\sqrt{2}} \ln \frac{x_i}{x_j}\right)\right) \tag{2.1}$$

V tomto případě značí vysokou proporcionalitu složek hodnoty blížící se k 1. Nicméně se stále jedná pouze o normované hodnoty variační matice, a ne o korelační míru. Měření proporcionality totiž neumožňuje rozlišovat mezi pozitivní a negativní asociací jako u korelačního koeficientu. Dalším problémem je absence statistických metod pro testování významnosti prvků variační matice.

2.2 Symetrické pivotové souřadnice

Nedostatky variační matice vedly ke snaze měřit sílu asociace mezi složkami kompozice pomocí korelačního koeficientu mezi určitými ortonormálními ilr souřadnicemi vzhledem k Aitchisonově geometrii. Jelikož se jedná o ortonormální souřadnice, mohou být použity korelační míry založené na euklidovské geometrii, jako třeba právě Pearsonův korelační koeficient. Příkladem takových souřadnic jsou symetrické pivotové souřadnice [15], které zachycují veškerou informaci ve formě agregovaných logaritmů podílů obou kompozičních složek, jejichž vztah sledujeme. Pojďme se nyní blíže zaměřit na jejich konstrukci.

2.2.1 Konstrukce symetrických pivotových souřadnic

Abychom mohli počítat korelační koeficient, potřebujeme sestavit bilance, které vyjadřují dominanci konkrétní složky vůči ostatním složkám. Pokud vezmeme postupné binární dělení vedoucí k bilancím [9]

$$\boldsymbol{z} = [z_1, \ldots, z_{D-1}], \qquad z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}, \tag{2.2}$$

kde $i = 1, \ldots, D-1$, pak bilance z_1 vyjadřuje veškerou relativní informaci složky x_1 vůči zbývajícím složkám. Lze ji tedy interpretovat z hlediska dominance složky x_1 vůči průměrnému chování zbývajících složcek (neboť bereme jejich geometrický průměr). U následujících bilancí už ale chybí složka x_1, a tedy i potřebná interpretace bilancí. Pro účely měření vztahu mezi dvěma složkami nás konkrétně zajímá bilance vyjadřující míru dominance složky x_2 v kompozici. Příslušný souřadnicový systém $\boldsymbol{z}^* = [z_1^*, \ldots, z_{D-1}^*]$ vznikne permutací složek $\boldsymbol{z} = [z_1, \ldots, z_{D-1}]$ ve vztahu (2.2). Proto je třeba uvažovat uvedené dva systémy souřadnic $\boldsymbol{z}$ a $\boldsymbol{z}^*$, respektive jejich první dvě souřadnice

$$z_1 = \sqrt{\frac{D-1}{D}} \ln \frac{x_1}{\sqrt[D-1]{\prod_{i=2}^{D} x_i}}, \quad z_2 = \sqrt{\frac{D-2}{D-1}} \ln \frac{x_2}{\sqrt[D-2]{\prod_{i=3}^{D} x_i}}$$

a

$$z_1^* = \sqrt{\frac{D-1}{D}} \ln \frac{x_2}{\sqrt[D-1]{x_1 \prod_{i=3}^{D} x_i}}, \quad z_2^* = \sqrt{\frac{D-2}{D-1}} \ln \frac{x_1}{\sqrt[D-2]{\prod_{i=3}^{D} x_i}}.$$

První dvě souřadnice z obou systémů úplně popisují podkompozici $[x_1, x_2]$ v rámci dané kompozice. Vezměme nyní části ortonormálních bází odpovídající příslušným dvěma bilancím z obou souřadnicových systémů

$$\underset{D\times 2}{\boldsymbol{V}_{\boldsymbol{z}}} = \begin{pmatrix} \sqrt{\frac{D-1}{D}} & 0 \\ -\frac{1}{D-1}\sqrt{\frac{D-1}{D}} & \sqrt{\frac{D-2}{D-1}} \\ -\frac{1}{D-1}\sqrt{\frac{D-1}{D}} & \frac{1}{D-2}\sqrt{\frac{D-2}{D-1}} \\ \vdots & \vdots \\ -\frac{1}{D-1}\sqrt{\frac{D-1}{D}} & \frac{1}{D-2}\sqrt{\frac{D-2}{D-1}} \end{pmatrix},$$

$$\underset{D\times 2}{\boldsymbol{V}_{\boldsymbol{z}^*}} = \begin{pmatrix} -\frac{1}{D-1}\sqrt{\frac{D-1}{D}} & \sqrt{\frac{D-2}{D-1}} \\ \sqrt{\frac{D-1}{D}} & 0 \\ -\frac{1}{D-1}\sqrt{\frac{D-1}{D}} & \frac{1}{D-2}\sqrt{\frac{D-2}{D-1}} \\ \vdots & \vdots \\ -\frac{1}{D-1}\sqrt{\frac{D-1}{D}} & \frac{1}{D-2}\sqrt{\frac{D-2}{D-1}} \end{pmatrix},$$

kde matice $\boldsymbol{V}_{\boldsymbol{z}^*}$ vznikne výměnou prvního a druhého řádku matice $\boldsymbol{V}_{\boldsymbol{z}}$. Lze odvodit [15], že mezi prvními dvěma souřadnicemi $\boldsymbol{z}$ a $\boldsymbol{z}^*$ platí vztah

$$[z_1^*, z_2^*] = [z_1, z_2]\, \boldsymbol{V}_{\boldsymbol{z}}{}^T \boldsymbol{V}_{\boldsymbol{z}^*},$$

kde ortogonální matice $\boldsymbol{V}_{\boldsymbol{z}}{}^T \boldsymbol{V}_{\boldsymbol{z}^*}$ má tvar

$$\boldsymbol{V}_{\boldsymbol{z}}{}^T \boldsymbol{V}_{\boldsymbol{z}^*} = \begin{pmatrix} -\frac{1}{D-1} & \sqrt{\frac{D-2}{D}}\frac{D}{D-1} \\ \sqrt{\frac{D-2}{D}}\frac{D}{D-1} & \frac{1}{D-1} \end{pmatrix}.$$

Podle vztahu (1.1) pro bilance z kapitoly 1.3.2 platí

$$z_1 = \sum_{i=1}^{D} v_{i1} \ln x_i, \qquad z_2 = \sum_{i=1}^{D} v_{i2} \ln x_i,$$

$$z_1^* = \sum_{i=1}^{D} v_{i1}^* \ln x_i, \qquad z_2^* = \sum_{i=1}^{D} v_{i2}^* \ln x_i.$$

Tyto bilance použijeme k sestrojení nových symetrických pivotových souřadnic. Veškerou relativní informaci týkající se složky x_1 zachycují vektory $\boldsymbol{v}_1$ a $\boldsymbol{v}_2^*$, tedy první sloupec (řádkový vektor) matice $\boldsymbol{V_z}$ a druhý sloupec matice $\boldsymbol{V_{z^*}}$. Proto nová symetrická pivotová souřadnice z_1^s odpovídá směru osy úhlu mezi $\boldsymbol{v}_1$ a $\boldsymbol{v}_2^*$. Podobně i druhá souřadnice z_2^s odpovídá směru osy úhlu mezi vektory $\boldsymbol{v}_2$ a $\boldsymbol{v}_1^*$. Novou souřadnici z_1^s tedy spočteme jako

$$z_1^s = \frac{1}{\|\boldsymbol{v}_1 + \boldsymbol{v}_2^*\|} \ln \boldsymbol{x}(\boldsymbol{v}_1 + \boldsymbol{v}_2^*)^T,$$

kde

$$\boldsymbol{v}_1 + \boldsymbol{v}_2^* = \left[\frac{D-1+\sqrt{D(D-2)}}{\sqrt{D(D-1)}}, -\frac{1}{\sqrt{D(D-1)}}, \right.$$
$$\left. -\frac{\sqrt{D-2}+\sqrt{D}}{\sqrt{D(D-1)(D-2)}}, \ldots, -\frac{\sqrt{D-2}+\sqrt{D}}{\sqrt{D(D-1)(D-2)}} \right]$$

a norma $\|\boldsymbol{v}_1 + \boldsymbol{v}_2^*\| =$

$$\sqrt{\frac{\left((D-1)+\sqrt{D(D-2)}\right)^2}{D(D-1)} + \frac{1}{D(D-1)} + (D-2)\cdot\frac{\left(\sqrt{D-2}+\sqrt{D}\right)^2}{D(D-1)(D-2)}} =$$
$$\sqrt{\frac{2\cdot(D-1+\sqrt{D(D-2)})}{(D-1)}}.$$

Z předchozích vztahů lze odvodit koeficienty pro logkontrast příslušný symetrické pivotové souřadnici z_1^s

$$\frac{\boldsymbol{v}_1+\boldsymbol{v}_2^*}{\|\boldsymbol{v}_1+\boldsymbol{v}_2^*\|}=\left[\frac{\sqrt{D-1+\sqrt{D(D-2)}}}{\sqrt{2D}},-\frac{1}{\sqrt{2D(D-1+\sqrt{D(D-2)})}},\right.$$
$$-\frac{\sqrt{D-2}+\sqrt{D}}{\sqrt{2D(D-2)(D-1+\sqrt{D(D-2)})}},\dots,$$
$$\left.-\frac{\sqrt{D-2}+\sqrt{D}}{\sqrt{2D(D-2)(D-1+\sqrt{D(D-2)})}}\right].$$

Výsledné symetrické pivotové souřadnice pak vypadají následovně (při konstrukci z_2^s postupujeme analogicky)

$$z_1^s=\sqrt{\frac{D-1+\sqrt{D(D-2)}}{2D}}\ln\frac{x_1}{x_2^{\frac{1}{D-1+\sqrt{D(D-2)}}}(x_3x_4\cdots x_D)^{\frac{\sqrt{D-2}+\sqrt{D}}{\sqrt{D-2}(D-1+\sqrt{D(D-2)})}}},$$

$$z_2^s=\sqrt{\frac{D-1+\sqrt{D(D-2)}}{2D}}\ln\frac{x_2}{x_1^{\frac{1}{D-1+\sqrt{D(D-2)}}}(x_3x_4\cdots x_D)^{\frac{\sqrt{D-2}+\sqrt{D}}{\sqrt{D-2}(D-1+\sqrt{D(D-2)})}}}.$$

Souřadnice z_1^s a z_2^s vyjadřují symetricky dominanci složek x_1 a x_2 vzhledem k dalším složkám. Souřadnice $z_1^s, z_2^s, z_3, \dots, z_{D-1}$, resp. $z_1^s, z_2^s, z_3^*, \dots, z_{D-1}^*$ potom tvoří ortonormální souřadnice kompozice $\boldsymbol{x}$. Složitý tvar mocnitelů ve jmenovateli slouží k normování souřadnic, aby bylo dosaženo ortonormality. Odlišný koeficient u složky x_2, resp. x_1 ve jmenovateli zase umožňuje dosažení symetrie souřadnic. Když upravíme poměr obou exponentů ve jmenovateli

$$\frac{\frac{1}{D-1+\sqrt{D(D-2)}}}{\frac{\sqrt{D-2}+\sqrt{D}}{\sqrt{D-2}(D-1+\sqrt{D(D-2)})}}=\frac{\sqrt{D-2}}{\sqrt{D-2}+\sqrt{D}},$$

tak vidíme, že s rostoucím počtem složek se poměr blíží 1/2 ve prospěch zbývajících složek. S rostoucím počtem složek rovněž klesá efekt různých exponentů

ve jmenovateli a souřadnice z_1^s a z_2^s se blíží (až na konstantu) prvním dvěma clr koeficientům

$$\ln \frac{x_1}{\sqrt[D]{x_1 \cdot x_2 \cdots x_D}}, \quad \ln \frac{x_2}{\sqrt[D]{x_1 \cdot x_2 \cdots x_D}} \quad \text{pro} \quad D \to \infty.$$

2.2.2 Korelační analýza se symetrickými pivotovými souřadnicemi

Symetrické pivotové souřadnice byly sestrojeny především za účelem určení korelace, v tomto případě mezi souřadnicemi, které vyjadřují vždy jednu složku kompozice vzhledem k ostatním složkám. K určení korelace mezi souřadnicemi můžeme použít např. klasický Pearsonův korelační koeficient (viz definice 1.4)

$$\rho(z_1^s, z_2^s) = \frac{\mathrm{cov}(z_1^s, z_2^s)}{\sqrt{\mathrm{var}(z_1^s) \cdot \mathrm{var}(z_2^s)}}.$$

Výhodou použití je jeho velká rozšířenost ve statistice. Na rozdíl od variační matice lze rozlišit pozitivní a negativní asociaci mezi bilancemi. Navíc můžeme použít standardní testování významnosti. Nicméně je důležité si uvědomit, že se nejedná o korelaci mezi původními složkami kompozice, ale k nim přiřazenými souřadnicemi. Hodnota korelačního koeficientu je ovlivněna i ostatními složkami kompozice, nejen dvěma sledovanými složkami, což ovšem koresponduje s vlastnostmi kompozičních dat.

Kladná hodnota korelačního koeficientu znamená, že dominance sledovaných složek nad průměrným zástupcem ostatních složek roste souběžně. Naopak záporná hodnota značí, že bilance se pohybují opačně. Nulová hodnota odpovídá nekorelovanosti bilancí. Protože každá bilance obsahuje jak složku x_1, tak složku x_2, pak pokud vzroste hodnota z_1^s, musí nutně hodnota z_2^s poklesnout. Složka x_2 v souřadnici z_1^s, resp. x_1v z_2^s je však pouze jednou z $D-1$ složek ve jmenovateli, a proto je efekt mnohem menší.

Pokud spočteme korelační koeficient mezi bilancemi odpovídajícími každé dvojici složek v kompozici $\boldsymbol{x} = [x_1, \ldots, x_D]$, získáme *kompoziční korelační matici*

$\boldsymbol{R}_C(\boldsymbol{x})$ o rozměrech $D \times D$. Matice je symetrická, na diagonále má jedničky. Navíc, pokud kompozici $\boldsymbol{x}$ přeškálujeme a posuneme, tzn. perturbujeme $\boldsymbol{x}$ kompozicí $\boldsymbol{b} = [b_1, \ldots, b_D]$ a mocninně transformujeme reálnou konstantou a (a dostaneme dle definice 1.3 kompozici $a \odot \boldsymbol{x} \oplus \boldsymbol{b} = [x_1^a b_1, \ldots, x_D^a b_D]$), tak se kompoziční korelační matice nezmění [18]

$$\boldsymbol{R}_C(a \odot \boldsymbol{x} \oplus \boldsymbol{b}) = \boldsymbol{R}_C(\boldsymbol{x}).$$

Existují ale ještě další metody založené na souřadnicích $\boldsymbol{z}$ a $\boldsymbol{z}^*$ z podkapitoly 2.2.1. Například vypočítat korelační koeficient pro dvojice souřadnic z_1, z_2 a z_1^*, z_2^* a poté vzít průměrnou hodnotu koeficientů

$$\rho_{ave}(\boldsymbol{z}, \boldsymbol{z^*}) = \frac{\rho(z_1, z_2) + \rho(z_1^*, z_2^*)}{2}.$$

Další možností je spočítat varianční matice $\boldsymbol{\Sigma}_{\boldsymbol{z}}$ a $\boldsymbol{\Sigma}_{\boldsymbol{z}^*}$ odpovídající prvním dvěma souřadnicím ze systémů $\boldsymbol{z}$ i $\boldsymbol{z}^*$ zvlášť a následně spočítat společnou varianční matici $\boldsymbol{\Sigma}_p(\boldsymbol{z}, \boldsymbol{z}^*)$ (tzv. *pooled covariance matrix*) jako průměr matic $\boldsymbol{\Sigma}_{\boldsymbol{z}}$ a $\boldsymbol{\Sigma}_{\boldsymbol{z}^*}$, přičemž

$$\boldsymbol{\Sigma}_{\boldsymbol{z}} = \begin{pmatrix} \mathrm{var}(z_1) & \mathrm{cov}(z_1, z_2) \\ \mathrm{cov}(z_2, z_1) & \mathrm{var}(z_2) \end{pmatrix} \quad , \quad \boldsymbol{\Sigma}_{\boldsymbol{z}^*} = \begin{pmatrix} \mathrm{var}(z_1^*) & \mathrm{cov}(z_1^*, z_2^*) \\ \mathrm{cov}(z_2^*, z_1^*) & \mathrm{var}(z_2^*) \end{pmatrix}$$

a

$$\boldsymbol{\Sigma}_p(\boldsymbol{z}, \boldsymbol{z}^*) = \frac{\boldsymbol{\Sigma}_{\boldsymbol{z}} + \boldsymbol{\Sigma}_{\boldsymbol{z}^*}}{2} = \begin{pmatrix} \boldsymbol{\Sigma}_{p_{11}} & \boldsymbol{\Sigma}_{p_{12}} \\ \boldsymbol{\Sigma}_{p_{21}} & \boldsymbol{\Sigma}_{p_{22}} \end{pmatrix}.$$

Využitím prvků této matice pak získáme výsledný korelační koeficient

$$\rho_{pool}(\boldsymbol{z}, \boldsymbol{z}^*) = \frac{\boldsymbol{\Sigma}_{p_{12}}}{\sqrt{\boldsymbol{\Sigma}_{p_{11}} \boldsymbol{\Sigma}_{p_{22}}}}.$$

2.3 Míry b-asociace

V další části se budeme zabývat měřením asociace mezi složkami kompozice definované na základě bilancí konstantních napříč výběrem. Tento vztah nazý-

váme *b-asociací*. V praxi se přesná b-asociace příliš nevyskytuje, proto se používají míry stupně b-asociace, které lze i statisticky testovat.

2.3.1 Přesná lineární asociace proměnných

V mnohorozměrné analýze reálných dat řekneme, že je skupina g proměnných s indexy z množiny G přesně korelovaná se skupinou h proměnných s indexy z H, pokud platí vztah

$$\sum_{i \in G} \alpha_i x_i = \beta_0 + \sum_{j \in H} \beta_j x_j.$$

Tedy všechny body z výběru leží na jedné nadrovině reálného prostoru. Pokud každá skupina obsahuje pouze jednu složku, dostaneme jednoduchý lineární model $x_1 = (\beta_0/\alpha_1) + (\beta_1/\alpha_1)x_2$.

Pro kompoziční datový soubor $\boldsymbol{X}_{n \times D}$ (náhodný výběr z rozdělení kompozice $[x_1, \ldots, x_D]$) kde řádky jsou D-složkové kompozice $\boldsymbol{x}_i$, vypadá jednorozměrné lineární omezení na simplexu následovně

$$\sum_{i=1}^{D} \alpha_i \ln x_i = k, \qquad \sum_{i=1}^{D} \alpha_i = 0.$$

Tzn. existuje logkontrast, který je konstantní napříč výběrem. Jestliže tento vztah platí, pak řekneme, že existuje kompoziční asociace mezi skupinou složek s kladnými, resp. zápornými hodnotami koeficientů α, kterou nazýváme *c-asociací*. Podmínka na koeficienty α_i zajistí, že logkontrast je invariantní vůči změně měřítka D-složkové kompozice $[x_1, \ldots, x_D]$. Pokud výběr splňuje $d < D - 1$ nezávislých jednorozměrných omezení, pak leží všechny body výběru na nadrovině dimenze $D - 1 - d$.

Pro větší počet složek se c-asociace interpretuje mnohem obtížněji. V takovém případě potřebujeme najít pouze složky, které hrají v logkontrastu významnou roli (tedy absolutní hodnota příslušných α je velká). Dále hledáme co největší počet složek, jejichž α je blízké nule, abychom je mohli v asociaci zanedbat. Problém se komplikuje, pokud zároveň vyžadujeme nízkou variabilitu logkontrastů.

Neboť když se omezíme jen na významné příspěvky složek, nezaručíme tím nízkou variabilitu takto upravených logkontrastů.

Existují speciální logkontrasty, jejichž interpretace je značně jednodušší. Jedná se o bilance představené v podkapitole 1.3.2. Asociace mezi bilancemi nazýváme *b-asociacemi*. Nechť je dán výběr z D-složkové kompozice $\boldsymbol{X}$ o rozsahu n. Složky kompozice rozdělíme do tří nepřekrývajících se skupin $G = \{x_1, x_2, \ldots, x_g\}$, $H = \{x_{g+1}, \ldots, x_{g+h}\}$ a $R = \{x_{g+h+1}, \ldots, x_D\}$. Uvažujme bilanci

$$B(G/H) = \ln \frac{g(G)}{g(H)}, \tag{2.3}$$

kde $g(\cdot)$ značí geometrický průměr argumentů. Bilance $B(G/H)$ je logkontrast se třemi různými α koeficienty: $1/g$ pro skupinu G, $1/h$ pro skupinu H a 0 pro skupinu R. Skupiny G a H jsou b-asociované, pokud je bilance $B(G/H)$ konstantní napříč výběrem, tedy $\mathrm{var}[B(G/H)] = 0$, kde $\mathrm{var}(\cdot)$ označuje výběrový rozptyl. Přitom b-asociaci skupin G a H lze snadno interpretovat jako proporcionalitu mezi geometrickými průměry složek v každé skupině. Pokud je bilance konstantní, pak $\ln g(G) = k_1 + \ln g(H)$ pro reálnou konstantu k_1. Jestliže na rovnici aplikujeme exponenciální funkci, dostaneme tvar $g(G) = k_2 \cdot g(H)$, kde $k_2 = \exp(k_1)$ je kladná konstanta.

Pro případ kdy $G = \{x_1\}$ a $H = \{x_2\}$ má bilance tvar jednoduchého logaritmu podílu, b-asociace je potom vyjádřena jednou z ekvivalentních podmínek

$$B(\{x_1\}/\{x_2\}) = \ln \frac{x_1}{x_2} = k, \qquad \mathrm{var}\left[B(\{x_1\}/\{x_2\})\right] = 0.$$

2.3.2 Přibližná b-asociace

Protože se přesná b-asociace v praxi nevyskytuje, je třeba zavést míry stupně b-asociace. Prvním přístupem je Aitchisonova variační matice (viz podkapitola 2.1), která je mírou (nedostatku) b-asociace. Pokud chceme posoudit významnost b--asociace, musíme hodnoty variační matice normovat. Nejprve zavedeme celkový rozptyl

$$\mathrm{totVar}(\boldsymbol{X}) = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\mathrm{var}\left[\ln\frac{x_i}{x_j}\right].$$

Členy variační matice pak normujeme součtem všech složek variační matice následovně

$$t_{ij}^{norm} = \frac{\mathrm{var}\left[\ln x_i/x_j\right]}{2D\cdot\mathrm{totVar}\left(\boldsymbol{X}\right)}.$$

Výsledné hodnoty ale závisí na počtu složek v kompozici, proto je třeba použít jiný přístup. Pokud by složky kompozice byly naprosto neproporcionální, její variační matice by měla hodnoty rovnoměrně rozložené mezi všechny svoje nediagonální členy. Hodnota celkového rozptylu by byla $2D\cdot\mathrm{totVar}(\boldsymbol{X})/(D\cdot(D-1))$, kterou lze použít k normování [6]

$$T_{ij}^{norm} = \frac{D(D-1)\mathrm{var}(\ln(x_i/x_j))}{2D\cdot\mathrm{totVar}\left(\boldsymbol{X}\right)} = \frac{(D-1)\mathrm{var}(\ln(x_i/x_j))}{2\cdot\mathrm{totVar}\left(\boldsymbol{X}\right)}.$$

Pokud $T_{ij}^{norm} < 1$, pak lze usuzovat na asociaci mezi příslušnými složkami kompozice. Ze zkušenosti ale vyplývá, že až hodnoty pod 0,2 značí skutečnou b-asociaci [7].

Měření míry b-asociace prostřednictvím lineárního modelu

V podkapitole 2.3.1 jsme uvedli, že b-asociace mezi skupinami složek G a H se dá vyjádřit lineárním modelem $g(G) \cong k\cdot g(H)$, resp. pokud vztah logaritmujeme, dostaneme tvar

$$\ln g(G) \cong k_1 + \beta_1 \ln g(H), \quad k_1 = \ln k.$$

V případě b-asociace je směrnice β_1 přibližně rovna jedné. Model není invariantní vůči změně měřítka, proto je potřeba ho převést. Jestliže se kompozice skládá ze

tří skupin složek G, H a R, můžeme například od každé strany rovnice odečíst $\ln g(R)$ a dostaneme lineární model

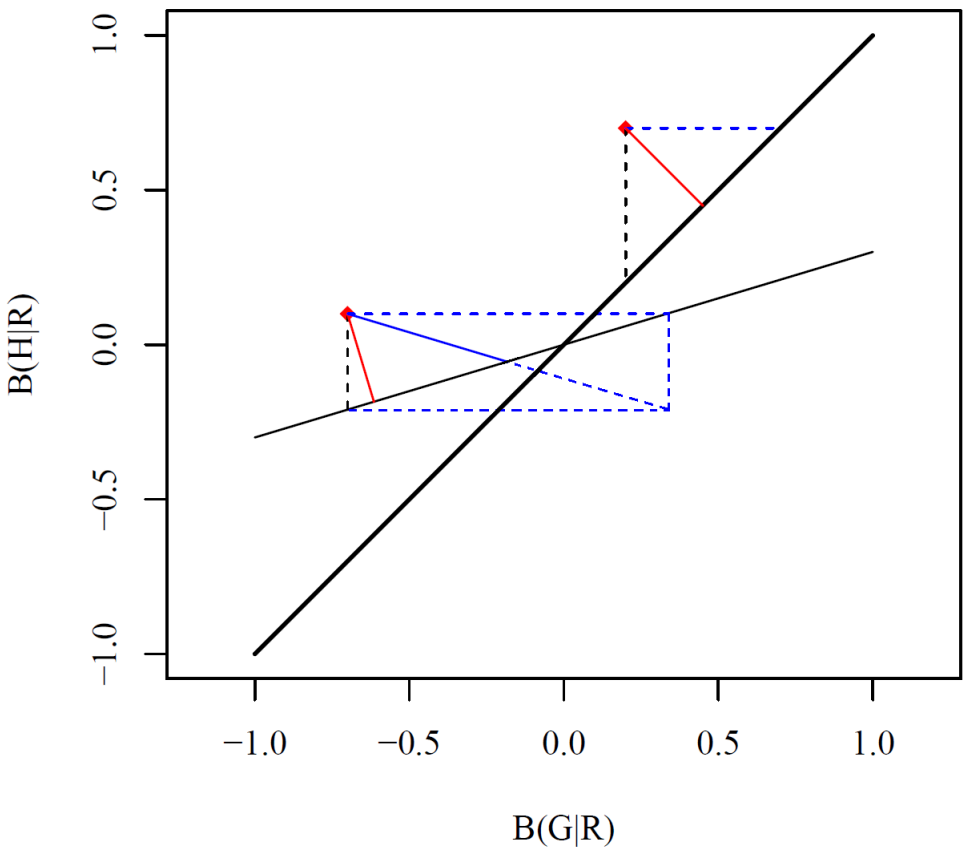
$$\ln \frac{g(G)}{g(R)} = \beta_0 + \beta_1 \ln \frac{g(H)}{g(R)} + \epsilon, \tag{2.4}$$

kde ϵ značí náhodnou chybu a skupina složek R referenční kompozici. Bilanci $\ln g(G)/g(R)$ budeme, jak již bylo zmíněno dříve, dále označovat $B(G/R)$ a obdobně bude vypadat i bilance obsahující skupinu H, tzn. $B(H/R)$ (B jako bilance). Pokud je β_1 přibližně rovno jedné, platí proporcionalita skupin složek pro malé hodnoty náhodné chyby. Její velikost je relativní vzhledem k bilancím $B(G/R)$ a $B(H/R)$, které jsou závislé na referenční kompozici.

Protože obě bilance hrají v modelu symetrickou roli, nemůžeme parametry odhadovat metodou nejmenších čtverců (*ordinary least squares*, zkráceně OLS). Pro tyto případy je třeba použít ortogonální regresi, konkrétně odhad MA (z anglického *major axis*, tedy hlavní osa) a SMA (z anglického *standardized major axis*, tedy standardizovaná hlavní osa). Cílem totiž není predikovat veličinu Y na základě X jako u klasické regrese, ale popsat vztah mezi dvěma proměnnými. Odhadnutá regresní přímka vyjadřuje dvourozměrná data pomocí jediné dimenze, dospějeme k redukci počtu rozměrů. Přístupy se liší tvarem reziduí, které v regresi minimalizujeme. Při použití klasické metody nejmenších čtverců minimalizujeme součet čtverců hodnot reziduí z modelu (2.4).

V obrázku 2.1, který je převzatý z článku [7] a následně upraven, vidíme dvě regresní přímky se směrnicí 1 (tučná) a 0,3 (tenká). Černou přerušovanou čárou jsou vyznačena rezidua metody nejmenších čtverců, rovnoběžná s vertikální osou. MA rezidua vedeme kolmo od bodu na příslušnou regresní přímku (červeně), jedná se o nejkratší vzdálenost mezi bodem a přímkou. Metoda MA odpovídá první hlavní komponentě vypočtené z varianční matice pro centrovaná data.

Odhad SMA získáme jako odhad MA použitý na standardizovaná data. SMA a MA rezidua se kryjí, pokud je směrnice rovna jedné. Obecně je najdeme jako půlku úhlopříčky obdélníku, jenž má vrchol v bodě, jehož reziduum hledáme a jehož strany získáme projekcemi bodu na regresní přímku, které jsou rovnoběžné



Obrázek 2.1: Rezidua různých odhadů regresního modelu.

s horizontální a vertikální osou. SMA reziduum je vyznačeno plnou modrou čarou a čárkovaně je naznačený i příslušný obdélník. Úhlopříčka je úměrná obsahu trojúhelníku, který je tvořen pozorováním a regresní přímkou. Lze proto říci, že SMA odhad minimalizuje plochu těchto trojúhelníků přes všechna pozorování. Metoda SMA je ekvivalentní hledání první hlavní komponenty při použití korelační matice [20]. Je nutno dodat, že ortogonální regrese předpokládá, že souřadnicové osy jsou na sebe kolmé. Osy v podobě bilancí $B(G/R)$ a $B(H/R)$ však ortogonální nejsou. Naštěstí se tím odhady parametrů modelu nezmění, ale mohou se mírně změnit výpočty reziduí a jejich součtu čtverců.

Pomocí charakteristik zavedených bilancí definujeme míry (nedostatku) b--asociace. První z nich je ϕ statistika [16]

$$\phi(B(G/R), B(H/R)) = \frac{\mathrm{var}(B(G/R) - B(H/R))}{\mathrm{var}(B(G/R))}. \tag{2.5}$$

Označme si podíl $g(G)/g(R)$ jako X a $g(H)/g(R)$ jako Y a upravme výraz $\mathrm{var}(\ln(X/Y))$ pomocí vlastností rozptylu náhodné veličiny a dalšími ekvivalentními úpravami

$$\begin{aligned}
\mathrm{var}(\ln(X/Y)) &= \mathrm{var}(\ln X) + \mathrm{var}(\ln Y) - 2 \cdot \mathrm{cov}(\ln X, \ln Y) \\
&= \mathrm{var}(\ln X)\left(1 + \frac{\mathrm{var}(\ln Y)}{\mathrm{var}(\ln X)} - 2\sqrt{\frac{\mathrm{var}(\ln Y)}{\mathrm{var}(\ln X)}}\frac{\mathrm{cov}(\ln X, \ln Y)}{\sqrt{\mathrm{var}(\ln X)\mathrm{var}(\ln Y)}}\right) \\
&= \mathrm{var}(\ln X)(1 + \beta^2 - 2\beta\,|r_{\ln X, \ln Y}|),
\end{aligned}$$

kde β je SMA odhad až na znaménko

$$\beta = \sqrt{\frac{\mathrm{var}(\ln Y)}{\mathrm{var}(\ln X)}}$$

a $r_{\ln X, \ln Y}$ značí korelační koeficient mezi $\ln X$ a $\ln Y$. Při výpočtu statistiky ϕ z dat pak použijeme místo teoretických hodnot rozptylu a korelačního koeficientu výběrové (značíme stříškou) a za X s Y dosadíme zpět příslušné výrazy. Z předchozího odvození platí, že

$$\phi(B(G/R), B(H/R)) = 1 + \hat{\beta}^2 - 2\hat{\beta}\left|\hat{r}_{B(G/R),\,B(H/R)}\right|.$$

Z tvaru charakteristiky je zřejmé, že přesné b-asociace je dosaženo, když $\hat{\beta} = 1$ a $\hat{r}_{B(G/R),\,B(H/R)} = 1$. Každé odchýlení od těchto hodnot zvětšuje hodnoty $\phi(B(G/R), B(H/R))$, tudíž hovoříme o míře nedostatku b-asociace. ϕ statistika postrádá symetrii, kterou přirozeně od míry asociace mezi dvěma veličinami očekáváme. Proto zavedeme další míru asociace ρ

$$\rho(B(G/R), B(H/R)) = \frac{2\mathrm{cov}(B(G/R), B(H/R))}{\mathrm{var}(B(G/R)) + \mathrm{var}(B(H/R))}. \tag{2.6}$$

Opět provedeme substituci $g(G)/g(R)$ za X a $g(H)/g(R)$ za Y

$$\begin{aligned}\frac{2\mathrm{cov}(\ln X,\ln Y)}{\mathrm{var}(\ln X)+\mathrm{var}(\ln Y)} &= 2\cdot\frac{\mathrm{cov}(\ln X,\ln Y)}{\sqrt{\mathrm{var}(\ln X)\mathrm{var}(\ln Y)}}\cdot\frac{\sqrt{\mathrm{var}(\ln X)\mathrm{var}(\ln Y)}}{\mathrm{var}(\ln X)+\mathrm{var}(\ln Y)}\\ &= \frac{2r_{\ln X,\ln Y}}{\frac{\mathrm{var}(\ln Y)}{\sqrt{\mathrm{var}(\ln X)\mathrm{var}(\ln Y)}}+\frac{\mathrm{var}(\ln X)}{\sqrt{\mathrm{var}(\ln X)\mathrm{var}(\ln Y)}}}\\ &= \frac{2r_{\ln X,\ln Y}}{\sqrt{\frac{\mathrm{var}(\ln Y)}{\mathrm{var}(\ln X)}}+\sqrt{\frac{\mathrm{var}(\ln X)}{\mathrm{var}(\ln Y)}}} = \frac{2r_{\ln X,\ln Y}}{\beta+\frac{1}{\beta}},\end{aligned}$$

kde $r_{\ln X,\ln Y}$ a β je definováno stejně jako u ϕ statistiky. Po dosazení odhadů parametrů dostáváme míru b-asociace

$$\rho(B(G/R),B(H/R)) = \frac{2\hat{r}_{B(G/R),B(H/R)}}{\hat{\beta}+\frac{1}{\hat{\beta}}},$$

jejíž hodnoty se pohybují na intervalu $\langle -1,1\rangle$. Hodnota 1 odpovídá přesné proporcionalitě, naopak -1 přesnou reciprocitu. Je ovšem třeba pamatovat, že obě zmíněné statistiky jsou definované vzhledem k dané refenční (pod)kompozici R.

2.3.3 Testování b-asociace

Pokud chceme testovat b-asociaci, nelze jednoduše použít nulovou hypotézu

$$H_0: \quad B(G/H) = k,$$

protože nepřipouští žádnou variabilitu testovací statistiky. Každá odchylka od nulové hypotézy by totiž vedla k jejímu zamítnutí. Proto je třeba zadefinovat nulovou hypotézu alternativně.

Testování na jednotkovou směrnici

Pokud uvažujeme model (2.4), lze hypotézu o b-asociaci skupin složek G a H naformulovat následovně

$$H_0': \quad \beta_1 = 1.$$

H_0' neimplikuje H_0, protože může platit i v případě velkých reziduí modelu. Je tedy slabší než H_0. Testování H_0' provedeme pomocí MA nebo SMA odhadu pro β_1 [20]

$$\hat{\beta}_{MA} = \frac{(s_{yy}^2 - s_{xx}^2) + \sqrt{(s_{yy}^2 - s_{xx}^2)^2 + s_{xy}^2}}{2s_{xy}}, \quad \hat{\beta}_{SMA} = \text{sign}(s_{xy})\frac{s_{yy}}{s_{xx}},$$

kde indexy x a y značí osy $B(G/R)$ a $B(H/R)$ a s je výběrová kovariance mezi veličinami určenými indexy. Oba odhady budou téměř shodovat, když je směrnice přímky blízko jedné (viz obrázek 2.1). Pro MA i SMA lze parametr β_0 odhadnout jako $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$, kde pruhem značíme příslušné výběrové průměry.

Testovací statistika je založena na korelaci mezi vyrovnanými hodnotami a rezidui za předpokladu, že H_0' platí, tzn. $\beta_1 = 1$. Dosadíme tedy za β_1 do vztahu pro rezidua $\epsilon' = y - \hat{\beta}_0 - \hat{\beta}_1 \cdot x$ a pro vyrovnané hodnoty

$$f'_{MA} = \hat{\beta}_1 \cdot (y - \hat{\beta}_0) + x, \quad f'_{SMA} = (y - \hat{\beta}_0) + \hat{\beta}_1 \cdot x.$$

Samotná testovací statistika má tvar

$$F = (n-2)\frac{r_{\epsilon' f'}^2}{1 - r_{\epsilon' f'}^2}.$$

Za platnosti nulové hypotézy H_0' a zároveň splnění podmínek na nezávislost, normalitu a homoskedasticitu chyb ϵ ve výběru má F rozdělení o 1 a $n-2$ stupni volnosti. Jestliže platí nulová hypotéza, tak rezidua obsahují už jen náhodnou informaci, a tudíž nemohou být signifikantně korelována s vyrovnanými hodnotami.

Problémem uvedeného testování je velká citlivost F-rozdělení na nedodržení normality reziduí. Pokud předpoklad neplatí, poté F-statistika sice stále měří míru odchylky od jednotkové směrnice, ale příslušná p-hodnota může být neinformativní (vychýlená). Jako další nevýhodu zmiňme, že ve skutečnosti testujeme, jestli se směrnice regresní přímky $\beta_1 = \pm 1$, ne jenom $\beta_1 = 1$. Proto je třeba ručně

ověřit, jestli nezamítnutí nulové hypotézy nenastalo kvůli hodnotám blízkým -1, což může být při realizaci mnoha testů pracné. Řešením může být automatické zamítnutí nulové hypotézy v případě záporného odhadu $\hat{\beta}_1$.

Regresní test

Další test předpokládá zavedení referenční kompozice. Složky kompozice opět rozdělujeme do skupin G, H a R, například v pořadí (G, H, R), můžeme však použít i jiné podkompozice jako referenční. Uvažujme model

$$B(G/H) = \gamma_0 + \gamma_1 B(G, H/R) + \sum_{j=2}^{r} \gamma_j B_j + e, \tag{2.7}$$

kde e značí náhodnou chybu, $B(G, H/R)$ je bilance zřetězených skupin G a H nad R. Za B_j volíme bilance pro podkompozici R, které mohou být ortogonální. Standardní regresní F-test testující, že $\gamma_j = 0$ pro $j = 1, \ldots, r$ odpovídá nulové hypotéze

$$H_0'' : \quad B(G/H) = \gamma_0 + e.$$

Stejně jako H_0', i H_0'' se liší od H_0. Nulovou hypotézu z regresního testu nezamítáme ani pro velká rezidua v modelu, pokud se nedají predikovat z referenční kompozice. To znamená, že test je silnější pro větší počet prediktorů v modelu. V případě velké kompozice obsahující desítky až stovky proměnných je ovšem test až příliš přísný. Model lze zjednodušit a uvažovat pouze jeden člen $B(G/H) = \gamma_0 + \gamma_1 B(G, H/R) + e$. Původní model (2.7) užijeme jen v případě silného vztahu mezi G a H a chceme-li být přísní v redukci chyby druhého druhu. Zjednodušený model dává stejné p-hodnoty jako test jednotkové směrnice, nicméně zjednodušený regresní test nerozlišuje znaménko β_1. Opět lze hodnoty F-statistiky brát jako směrodatné, i pokud nebyly splněny předpoklady na rezidua. Na p-hodnoty je ovšem nutné pohlížet kriticky, zvlášť když máme k dispozici kompozice s velkým počtem složek.

3 Korelační analýza reálných dat

3.1 Data

Ve třetí kapitole provedeme korelační analýzu datového souboru pocházejícího ze studie na Fakultě tělesné kultury Univerzity Palackého v Olomouci. Podrobné informace o studii a sběru dat nalezneme v článcích [10], [11]. Kompoziční datový soubor, který budeme označovat názvem `activity`, obsahuje data o rozložení fyzické aktivity během dne. Skládá se celkem ze čtyř složek: sedavá (`sedentary`) aktivita, lehká (`light`), středně náročná (`moderate`) a náročná (`vigorous`) pohybová aktivita v minutách za den. Data byla sbírána pomocí akcelerometru připevněného k pasu účastníků výzkumu v cpm (*counts per minute*). Hranice jednotlivých kategorií byly určeny dle tabulky 3.1.

sedentary	do 100 cpm
light	od 100 do 2296 cpm
moderate	od 2296 do 4012 cpm
vigorous	od 4012 cpm

Tabulka 3.1: Hranice pro rozdělení času mezi jednotlivé typy aktivit.

Nicméně je třeba zdůraznit, že nelze najít dokonalé hranice, a proto vždy nemusí odpovídat zařazení aktivity realitě. To je způsobeno i problematickou jednotkou cpm (*counts per minute*), která je v současnosti nahrazována jinými jednotkami.

Může se zdát trošku zavádějící, že první kategorii označujeme jako „sedavou“, neboť do ní zahrnujeme kromě sezení i ležení a stání. Dále však budeme tyto činnosti označovat jako `sedentary` (tedy sedavé). Pod lehkou náročnou aktivitou si můžeme představit například chůzi či domácí práce. Obecně je ale přechod mezi sedavou a lehkou aktivitou špatně detekovatelný. Středně náročnou aktivitu reprezentuje lehčí forma sportování a náročnou aktivitu už účastníci popisují jako činnost, při které se zpotí a zadýchají. Měření probíhalo na dětech od 7 do 12 let z náhodně vybraných základních škol na místech s různým počtem

obyvatel ve východní oblasti České republiky. Fyzická aktivita byla monitorována ve dvou obdobích během roku 2013 a 2014. Měření probíhalo po celý týden ve všední i víkendové dny a hodnoty byly následně zprůměrovány. Bohužel tedy nelze sledovat rozdílnou strukturu mezi školními dny a víkendem, přestože režim může být často velmi odlišný. Pro lepší představu si zobrazíme první čtyři z 426 řádků kompozičního datového souboru `activity` (tabulka 3.2).

sedentary	light	moderate	vigorous
349,60	217,00	24,80	14,40
338,20	250,20	13,60	8,00
329,20	250,60	36,60	2,80
243,75	326,50	42,50	7,25

Tabulka 3.2: Prvních pět řádků datového souboru `activity`, jednotlivé aktivity jsou uvedeny v minutách za den.

Abychom získali přehled o populaci, na které byla data měřena, zobrazíme si ještě souhrn charakteristik některých doplňujících proměnných (tabulka 3.3).

	věk	výška	váha	procento tělesného tuku	BMI
minimum	7	117,5	20,5	4,0	12,9
1. kvartil	9	136,5	29,5	12,1	15,7
medián	10	142,5	35,4	16.2	17,0
průměr	9,8	142,9	36,4	17,8	17,6
3. kvartil	11	149,5	41,3	22,6	19,0
maximum	12,8	175,0	76,1	43,1	30,0

Tabulka 3.3: Souhrn charakteristik doplňujících proměnných týkající se pozorovaných dětí.

3.2 Explorační analýza dat

Protože provádíme kompoziční analýzu dat, převedeme si nejdříve data na konstantní součet, neboť nás zajímají podíly složek na celku. Tabulka 3.2 se tedy změní na tabulku 3.4, která představuje proporce dne strávené jednotlivými aktivitami.

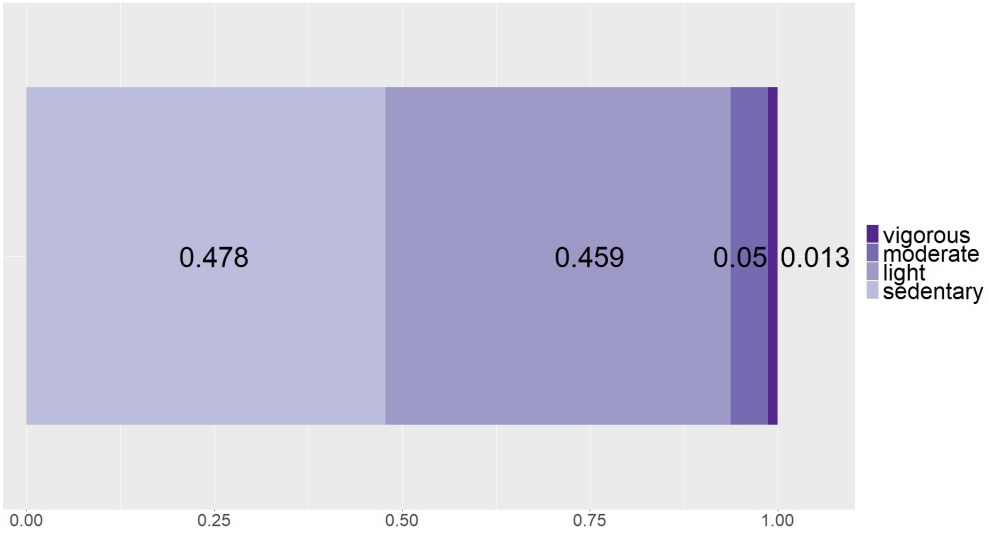
sedentary	light	moderate	vigorous
0,577	0,358	0,041	0,024
0,554	0,410	0,022	0,013
0,532	0,405	0,059	0,005
0,393	0,527	0,069	0,012

Tabulka 3.4: Prvních pět řádků datového souboru `activity` uzavřeného na konstantní součet.

Dále budeme pozorovat, kolem kterých hodnot se složky kompozice pohybují, a to pomocí kompozičního průměru (tzv. centra). Spočítáme ho jako aritmetický průměr clr koeficientů, následně převedený zpět na kompozici

$$\overline{\boldsymbol{x}} = clr^{-1}\left(\frac{1}{N}\sum_{n=1}^{N} clr(\boldsymbol{x}_n)\right).$$

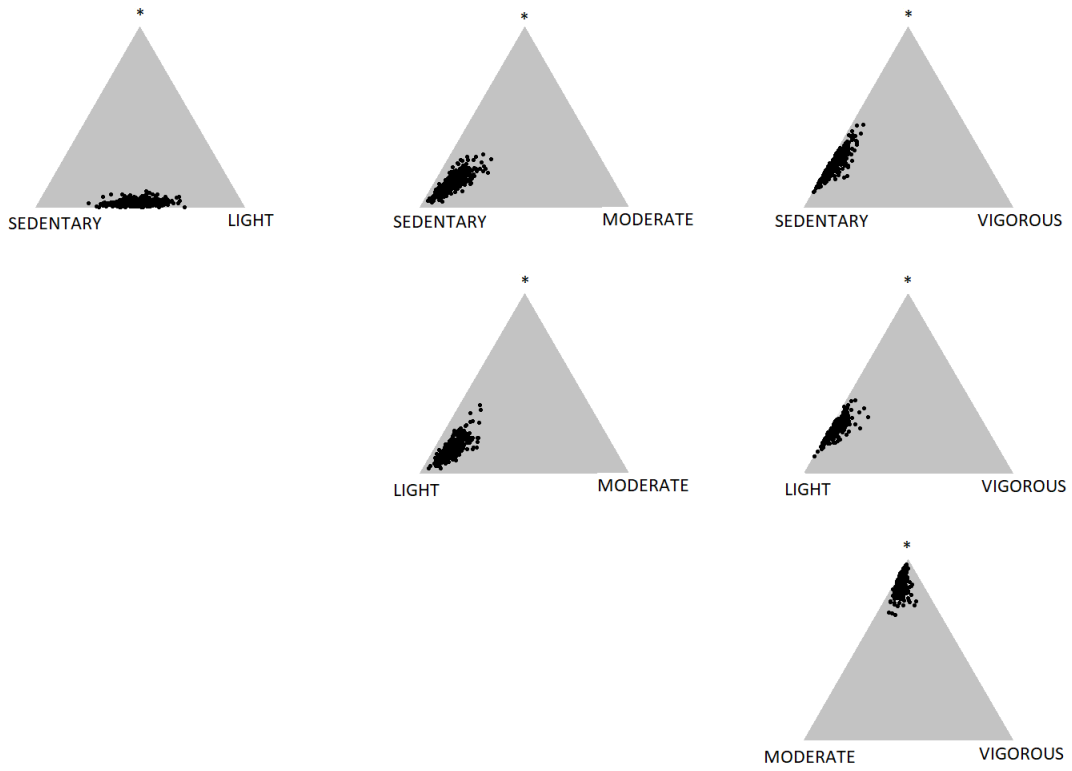
Hodnoty kompozičního průměru si zobrazíme pomocí sloupcového grafu (viz obrázek 3.1). Průměrné dítě stráví necelých 48 % dne sezením a téměř 46 % dne lehkou pohybovou aktivitou. Zmíněné složky tedy dominují průměrné kompozici. Středně náročná a náročná aktivita zaberou v průměru už jen 6,3 %.



Obrázek 3.1: Kompoziční průměr datového souboru `activities`.

Podrobnější představu o poloze a tvaru datového mraku si lze učinit, když si data vykreslíme do matice ternárních diagramů. Jedná se o obdobu matice bo-

dových grafů pro dvojice proměnných, které ovšem nejsou pro kompoziční data vhodné [13]. Vytvoříme ternární diagram pro každou dvojici proměnných (obrázek 3.2). Jako třetí složku pak zvolíme geometrický průměr všech zbývajících složek a uzavřeme na konstantní součet. Kompoziční datový soubor následně zakreslíme do trojsložkového simplexu, což je trojúhelník, který lze snadno vizualizovat. Velikosti složek jednotlivých pozorování potom získáme kolmými projekcemi bodu na jednotlivé strany trojúhelníku (označovaného v tomto kontextu právě jako ternární diagram). V grafech se opět projevuje dominance složek odpovídající sedavé a lehce náročné aktivitě. Například v grafu na pozici $(1, 1)$ všechny data leží u strany trojúhelníka mezi sedavou a lehce náročnou aktivitou. Nicméně pokud zprůměrujeme sedavou a lehkou aktivitu (graf na posledním řádku), můžeme sledovat i variabilitu v poměru ke složkám středně náročné a náročné aktivity.



Obrázek 3.2: Horní diagonála matice ternárních diagramů pro datový soubor `activities`.

Jako další nástroj k prozkoumání dat použijeme analýzu hlavních komponent. Protože ale analyzujeme kompoziční data, aplikujeme ji na clr koeficienty centrovaného datového souboru $\boldsymbol{X}$ dle knihy [18], který si označíme jako $\boldsymbol{Z}$. Provedeme singulární rozklad matice $\boldsymbol{Z}$, který získáme pomocí matice vlastních vektorů $\boldsymbol{U}$ matice $\boldsymbol{Z}\boldsymbol{Z}^T$, matice vlastních vektorů $\boldsymbol{V}$ matice $\boldsymbol{Z}^T\boldsymbol{Z}$ a diagonální matice $\boldsymbol{\Lambda}^{1/2}_{D\times D}$ singulárních hodnot matice $\boldsymbol{Z}\boldsymbol{Z}^T$ (singulární hodnota matice se rovná odmocnině z příslušného vlastního čísla matice)

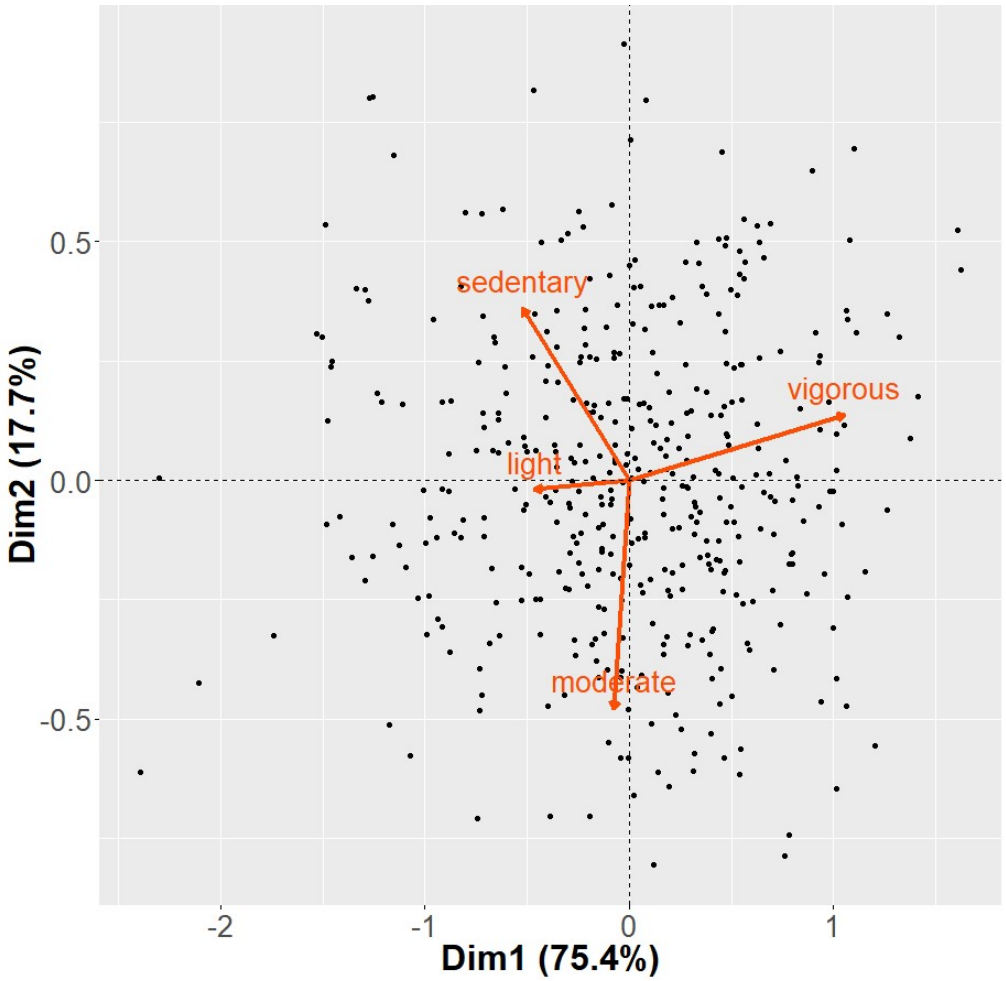
$$\boldsymbol{Z}_{n\times D} = \boldsymbol{U}_{n\times D}\boldsymbol{\Lambda}^{1/2}_{D\times D}\boldsymbol{V}^T_{D\times D}.$$

Matice $\boldsymbol{U}$ obsahuje standardizované souřadnice kompozice. Matici $\boldsymbol{V}$ označujeme jako *matici zátěží*, kde sloupce jsou clr koeficienty nové ortonormální báze na simplexu. Jednotlivá vlastní čísla odpovídají rozptylu komponent, přičemž komponenty jsou seřazeny od největšího rozptylu po nejmenší. Poslední vlastní číslo je vždy nulové, protože složky clr koeficientů dávají součet 0, a proto je hodnost matice $\mathrm{clr}(\boldsymbol{X})$ nejvýše $D-1$. Matice $\boldsymbol{U}\boldsymbol{\Lambda}^{1/2}$ obsahuje centrované ilr souřadnice vzhledem k bázi $\boldsymbol{V}$ a nazývá se *matice skórů*.

Protože analýza hlavních komponent slouží k redukci dimenze, můžeme komponenty vysvětlující pouze malou část variability zanedbat a pracovat pouze s těmi významnými. Pokud uvažujeme jen dvě nejvýznamnější komponenty, lze je zakreslit do dvourozměrného *biplotu* (obrázek 3.3), ve kterém je zachyceno

$$\frac{\lambda_1+\lambda_2}{\sum_{i=1}^{r}\lambda_i}$$

procent variability dat, kde λ_i značí vlastní číslo odpovídající i-té komponentě. Biplot na obrázku 3.3 reprezentuje datový soubor `activity` dobře, protože jsme pomocí prvních dvou komponent schopni vysvětlit 93,1% variability dat. V grafu jsou zakreslené jak aproximace jednotlivých pozorování (body) pomocí skórů prvních dvou komponent, tak clr koeficienty příslušných složek prostřednictvím zátěží ve formě šipek. Skóry a zátěže jsou ale před zakreslením do biplotu normovány diagonální maticí singulárních hodnot.



Obrázek 3.3: Biplot pro clr koeficienty datového souboru `activity`.

Kompoziční biplot je třeba interpretovat jinak než ten běžný [12]. Šipky se protínají v centru kompozičního datového souboru, který je v počátku souřadnicové osy, neboť jsme data centrovali. Délky jednotlivých šipek odpovídají směrodatným odchylkám jednotlivých clr koeficientů, interpretovatelných ve smyslu dominance jednotlivých složek vůči průměrnému chování ostatních složek v kompozici. Vidíme, že výrazně menší směrodatnou odchylku má clr koeficient příslušný lehké fyzické aktivitě, ostatní koeficienty mají směrodatnou odchylku podobnou. Označme si $\boldsymbol{b}_j$ dvojici souřadnic vektoru zátěží odpovídajícího j-tému clr koeficientu kompozice. Délka spojnice mezi vektory $\boldsymbol{b}_j$ a $\boldsymbol{b}_k$ se přibližně rovná

rozptylu bilance $\ln(x_j/x_k)$. Jelikož jedna z definic b-asociace uvádí právě rozptyl bilance, lze ji v biplotu hledat jako krátkou spojnici mezi vrcholy daných složek. V grafu 3.3 tedy najdeme nejsilnější b-asociaci mezi složkami sedavé a lehce náročné aktivity a poté mezi složkami lehké a středně náročné činnosti.

Posledním nástrojem, který použijeme pro explorační analýzu, bude CoDa dendrogram, který vizualizuje jednotlivé bilance z postupného binárního dělení popsaného v části 1.3.2. Protože je konstrukce postupného binárního dělení čistě na rozhodnutí statistika, získáme i různé verze CoDa dendrogramu. Graf zobrazuje hodnoty bilancí na vertikálních osách spojující složky pomocí boxplotů. Hodnoty osy jsou implicitně nastavené na interval $(-4, 4)$. Délky čar vycházející z jednotlivých boxplotů vyjadřují, jakou část variability dat vysvětluje daná bilance.

Z prvního CoDa dendrogramu na obrázku 3.4 a popsaného tabulkou 3.5 vyčteme, že největší podíl variability vysvětlí bilance tří nejlehčích aktivit oproti náročné fyzické aktivitě. Boxplot posunutý ke kladným hodnotám pak potvrzuje malou proporci náročné fyzické aktivity na celku. Z pohledu b-asociace nás zajímají malé rozptyly bilancí. Příkladem může být bilance mezi sedavou a lehkou aktivitou.

	Bilance		
Složka	z_1	z_2	z_3
`sedentary`	+	+	+
`light`	+	+	−
`moderate`	+	−	
`vigorous`	−		

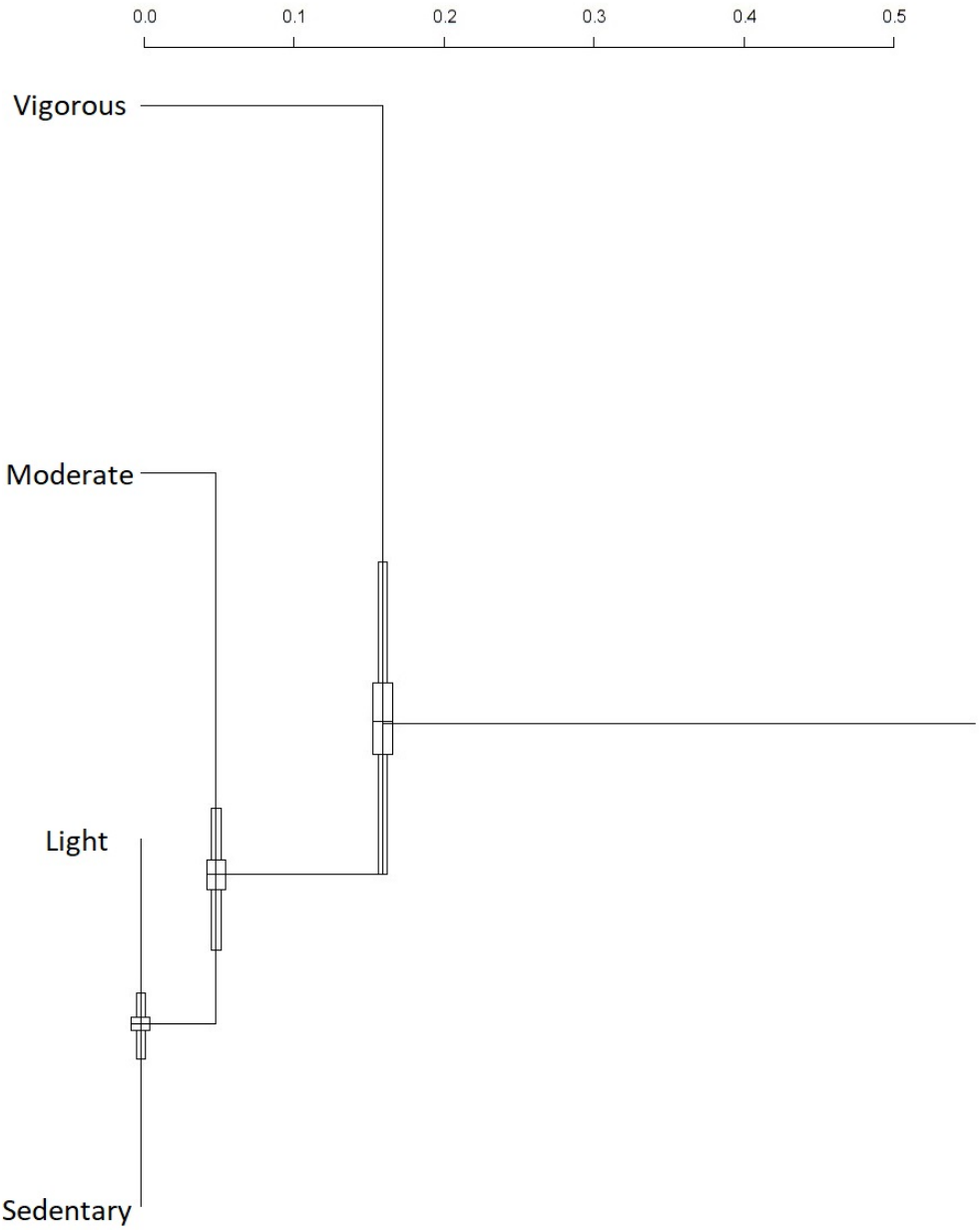
Tabulka 3.5: Postupné binární dělení pro datový soubor `activity` odpovídající obrázku 3.4.

Obdobně lze v druhém CoDa dendrogramu popsaného tabulkou 3.6 na obrázku 3.5 sledovat malou variabilitu bilance mezi složkami `Moderate` a `Light`. Největší část variability tentokrát vysvětluje bilance průměru složek `Moderate` a `Light` oproti složce `Vigorous`, což je obdobné zjištění jako v předchozím dendro-

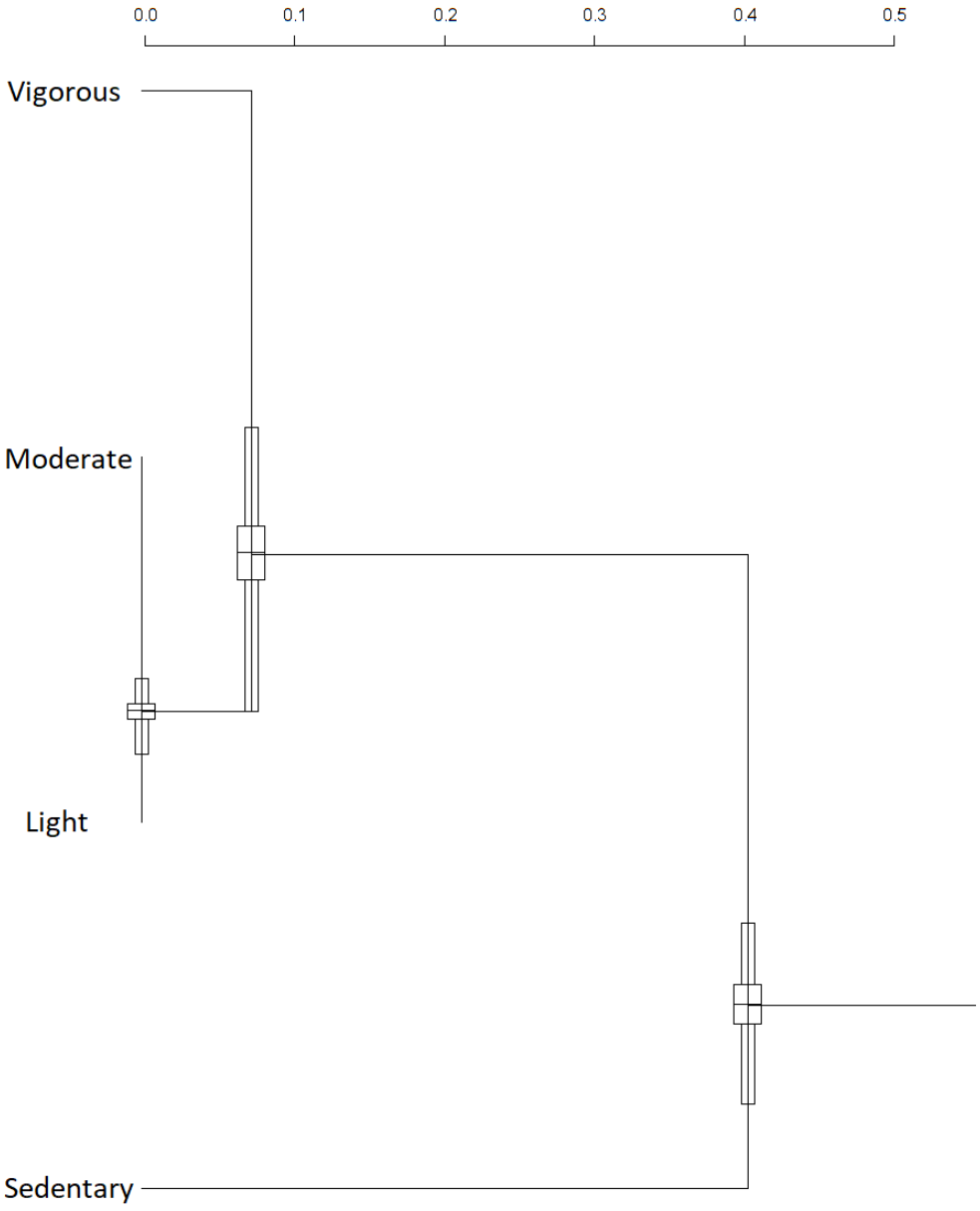
gramu. V obou případech jsou tedy rozdíly mezi dětmi ukryté hlavně v poměru k této složce.

	Bilance		
Složka	z_1	z_2	z_3
`sedentary`	+		
`light`	−	+	+
`moderate`	−	+	−
`vigorous`	−	−	

Tabulka 3.6: Postupné binární dělení pro datový soubor `activity` odpovídající obrázku 3.5.



Obrázek 3.4: CoDa dendrogram postupného binárního dělení datového souboru `activity` z tabulky 3.5.



Obrázek 3.5: CoDa dendrogram jiného postupného binárního dělení datového souboru `activity` z tabulky 3.6.

3.3 Variační matice

Nyní spočteme a budeme interpretovat variační matici (tabulka 3.7) pro datový soubor `activity`. V knihovně `compositions` softwaru R k tomu slouží funkce `variation`.

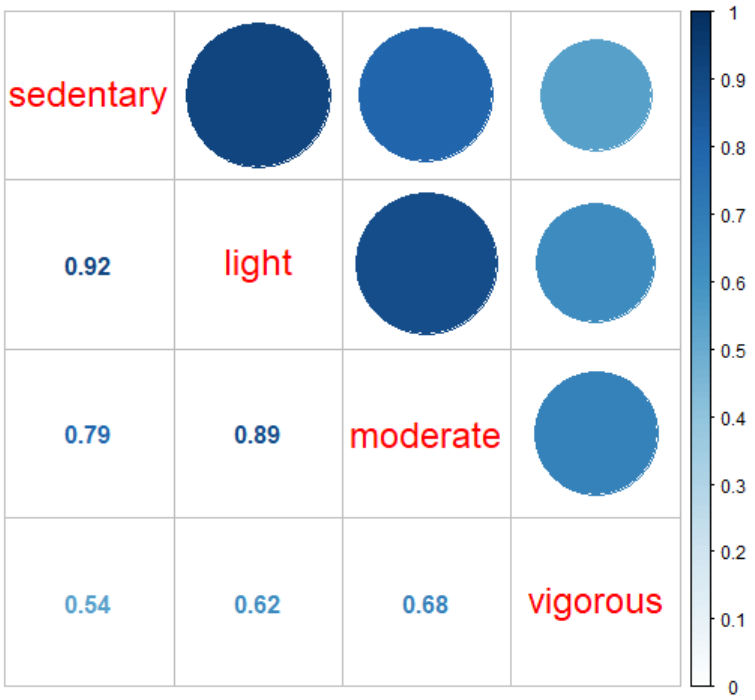
	sedentary	light	moderate	vigorous
sedentary	0.000	0.101	0.237	0.675
light	0.101	0.000	0.146	0.627
moderate	0.237	0.146	0.000	0.439
vigorous	0.675	0.627	0.439	0.000

Tabulka 3.7: Variační matice pro datový soubor `activity`.

Tabulka 3.7 ukazuje, že nejvíce proporcionální jsou sedavá a lehce náročná aktivita. Naopak největší rozdíly jsou v poměru mezi sedavou a náročnou pohybovou aktivitou. Nyní hodnoty normujeme pomocí exponenciální funkce podle vztahu (2.1) a vizualizujeme pomocí knihovny `corrplot` (obrázek 3.6). Zatímco na dolním trojúhelníku matice jsou konkrétní hodnoty prvků variační matice, tak v horním trojúhelníku jsou kruhy, jejichž velikost je úměrná těmto hodnotám. Vše je doplněno odpovídající barvou ze škály na pravé straně grafu. Hodnoty se z definice pohybují pouze mezi 0 a 1.

Ačkoliv se hodnoty změnily, jejich interpretace zůstala stejná jako u předchozí matice. Tentokrát se jedná o míru proporcionality, a ne o míru nedostatku proporcionality jako u původní Aitchisonovy verze. Například vysoká hodnota koeficientu mezi sedavou a lehce náročnou aktivitou značí jejich vysokou proporcionalitu.

Z obrázku 3.6 lze usoudit, že variační matice není vhodným nástrojem pro hledání asociace. Pro všechny dvojice proměnných je hodnota téměř stejná a lze těžko poznat, který vztah je skutečně významný. Navíc nezobrazuje informaci o negativních asociacích.



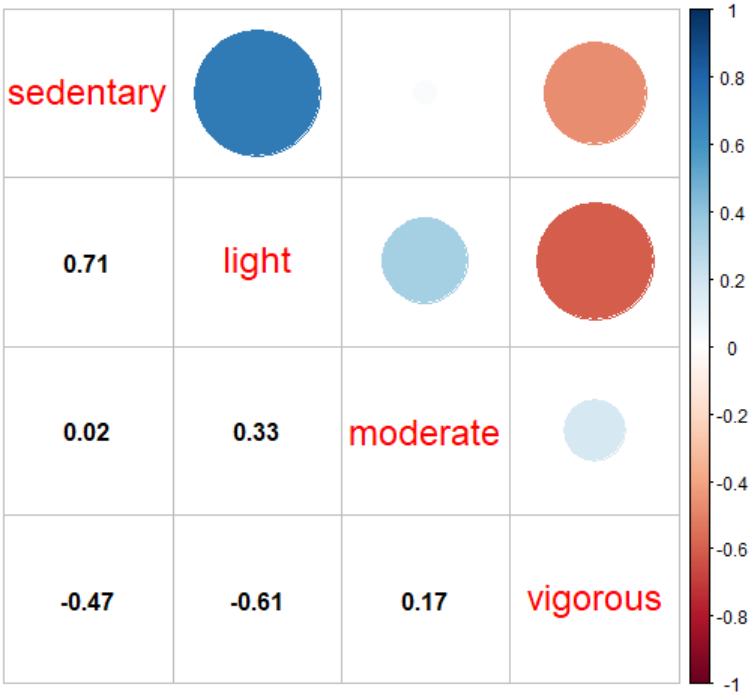
Obrázek 3.6: Normovaná variační matice pro datový soubor `activity`.

3.4 Korelace mezi symetrickými pivotovými souřadnicemi

Na rozdíl od varianční matice, kompoziční korelační matice mezi symetrickými pivotovými souřadnicemi pro datový soubor `activity` na obrázku 3.7 ukazuje mnohem bohatější strukturu vztahů. Hodnoty lze jednoduše získat užitím příkazu `corCoDa` z knihovny `robCompositions`. Jednotlivé symetrické pivotové souřadnice vyjadřují dominanci složky kompozice vzhledem k ostatním složkám. Korelace potom počítáme mezi dominancemi dvojic složek a lze je zobrazit do korelační matice.

Nejsilnější kladný vztah je mezi symetrickou pivotovou souřadnicí složek sedavé a lehce náročné aktivity. To může poukazovat na jev, že na jednu stranu existují méně sportovně zaměřené děti, které tráví hodně času dvěma nejlehčími aktivitami a naopak sportovně zaměřené děti, které náročnější složky upředno-

stňují na úkor sedavé a lehce náročné aktivity. Také to může souviset s tím, že hranice mezi sedavou a lehkou aktivitou je nejasná, a proto mohou být tyto aktivity snadno zaměněny.

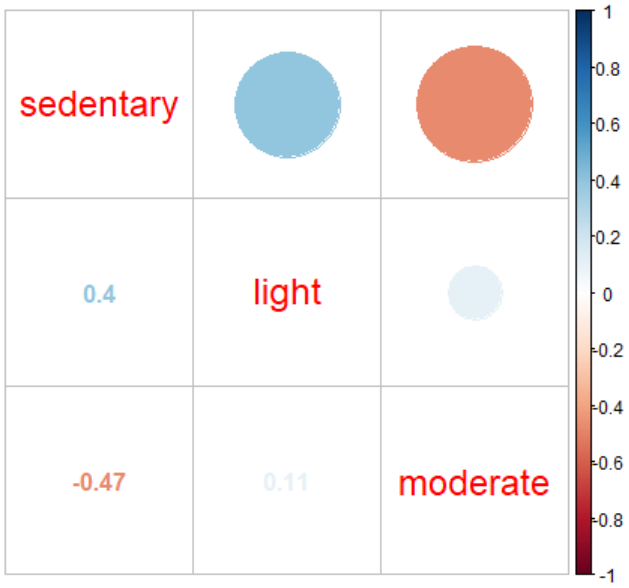


Obrázek 3.7: Korelační matice pro symetrické pivotové souřadnice datového souboru `activity`.

Tentokrát můžeme sledovat i negativní vztahy, nejsilněji mezi bilancemi lehké a náročné aktivity a podobně i mezi bilancemi sedavé a náročné aktivity. I tato skutečnosti nejspíš odpovídá naší intuitivní představě, že pokud je podíl lehké či sedavé aktivity vyšší, tak dominance náročné aktivity bude klesat. Míra negativní korelace je menší pro bilance odpovídající sedavé a náročné složce, než pokud vezmeme místo sedavé složku lehce náročnou. To může být ovlivněno faktem, že všechny děti musí určitou část dne prosedět ve škole, a poté vztah nemá takovou sílu.

Abychom si lépe uvědomili, že se jedná o korelaci mezi bilancemi v kontextu celé kompozice, vykreslíme si ještě graf pro podkompozici prvních tří složek (ob-

rázek 3.8). Ačkoliv korelace mezi bilancemi vyjadřujícími dominanci sedavé a středně náročné složky v rámci celé kompozice byla téměř nulová, tak se situace výrazně změní, pokud uvažujeme pouze první tři složky. Při vyloučení náročné složky totiž středně náročná aktivita reprezentuje nejnáročnější aktivitu v kompozici, proto je korelace mezi dominancemi odpovídajícími sedavé a středně náročné složce záporná. Pokud bychom například do výzkumu zapojili i proporci dne strávenou spánkem, získali bychom opět odlišné vztahy mezi symetrickými pivotovými souřadnicemi. Nemůžeme tedy počítané charakteristiky ztotožňovat s korelací mezi dvěma samotnými složkami, ale je nutné se zamyslet nad všemi složkami současně.



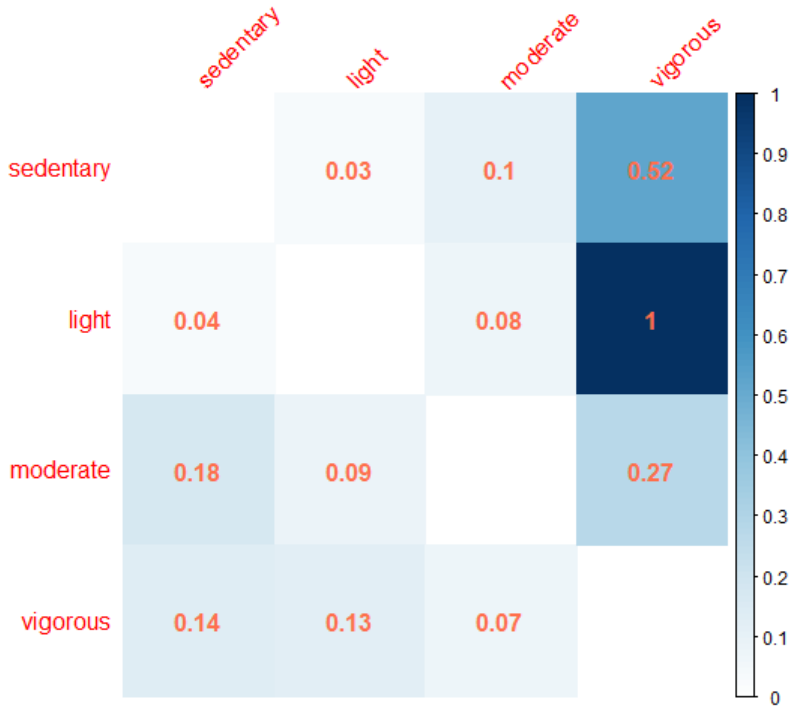
Obrázek 3.8: Korelační matice pro symetrické pivotové souřadnice podkompozice prvních tří složek datového souboru `activity`.

3.5 Míry b-asociace

Druhým uvedeným přístupem pro měření vztahu mezi složkami kompozičních dat je b-asociace. Sledujeme proporcionalitu mezi složkami na základě modelu (2.4).

3.5.1 ϕ statistika

První ze zavedených měr je ϕ statistika dle vztahu (2.5), která je definovaná pomocí bilancí $B(G/R)$ a $B(H/R)$. Za G a H volíme jednotlivé dvojice složek a za R zbývající složky. K určení ϕ statistiky pak už stačí jen výpočet příslušných výběrových rozptylů. Protože hodnoty rozptylu nejsou shora nijak omezené, normujeme hodnoty statistiky největší hodnotou z matice ϕ statistik. Tím získáme matici na obrázku 3.9.



Obrázek 3.9: ϕ statistika normovaná největší hodnotou pro datový soubor `activity`.

K interpretaci je třeba dodat, že se jedná o míru nedostatku b-asociace. Čím je hodnota menší, tím je proporcionalita mezi složkami silnější. Přesné b-asociace bychom dosáhli pro hodnotu 0. Největším problémem charakteristiky je absence

symetrie, která je patrná například pro dvojice `light-vigorous` a `vigorous-light`. Zatímco pro první variantu bychom b-asociaci vyloučili, při obrácení pořadí už to není tak jisté. I přesto jsme schopni z matice vyčíst silnou b-asociaci mezi složkami sedavé a lehké aktivity a lehké a středně náročné fyzické aktivity. Obdobně jako u variační matice nerozlišujeme mezi kladnou a zápornou asociací, což ostatně vyplývá z definice ϕ statistiky pomocí rozptylu. Zajímavé je, že nejvyšší hodnoty v horním trojúhelníku matice dostáváme pro dvojice složek, které odpovídají negativní korelaci mezi symetrickými pivotovými souřadnicemi.

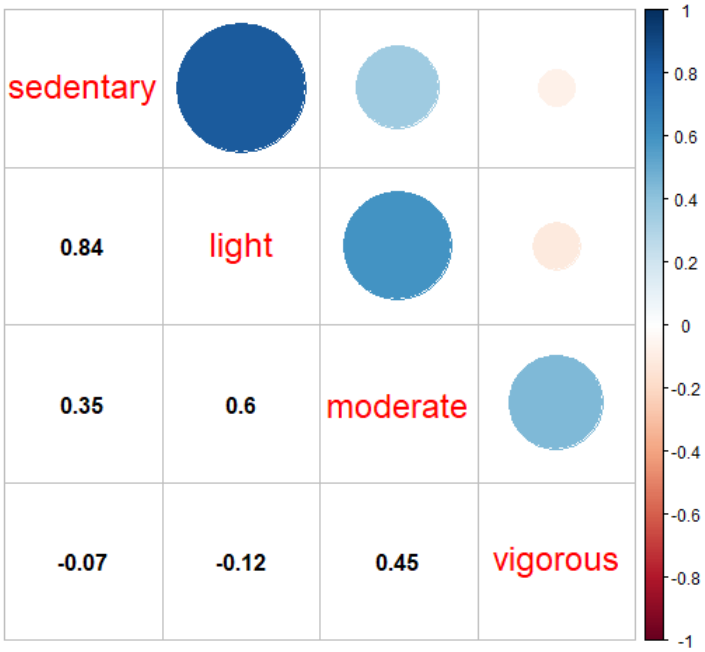
3.5.2 ρ statistika

Některé problémy ϕ statistiky řeší ρ statistika ze vztahu (2.6), která je symetrickou mírou b-asociace. Pokud zvolíme za G a H vždy dvojici složek kompozice a za R zbytek složek, pak získáme matici z obrázku 3.10. Hodnoty charakteristiky se pohybují od -1 do 1, nicméně interpretace krajních hodnot je jiná než u korelace. Zatímco 1 znamená přesnou proporcionalitu bilancí, tak -1 značí jejich reciprocitu, tzn. že hodnoty jsou vzájemně převrácené.

Při pohledu na obrázek 3.10 vidíme již mnoha způsoby nalezený vztah mezi sedavou a lehkou fyzickou aktivitou. Druhou nejsilnější asociaci nalezneme mezi lehkou a středně náročnou aktivitou, což se rovněž neliší od předchozích přístupů. Výrazným rozdílem oproti přístupu se symetrickými pivotovými souřadnicemi jsou dvojice s negativní hodnotou ρ statistiky. Tentokrát je síla negativní asociace mnohem menší. Nicméně je potřeba si uvědomit, že každá z metod měří něco jiného. Zatímco pro ρ statistiku rozumíme negativní hodnotou míru reciprocity, u předchozího přístupu značila záporná hodnota negativní lineární vztah mezi dominancemi složek ku zbytku složek.

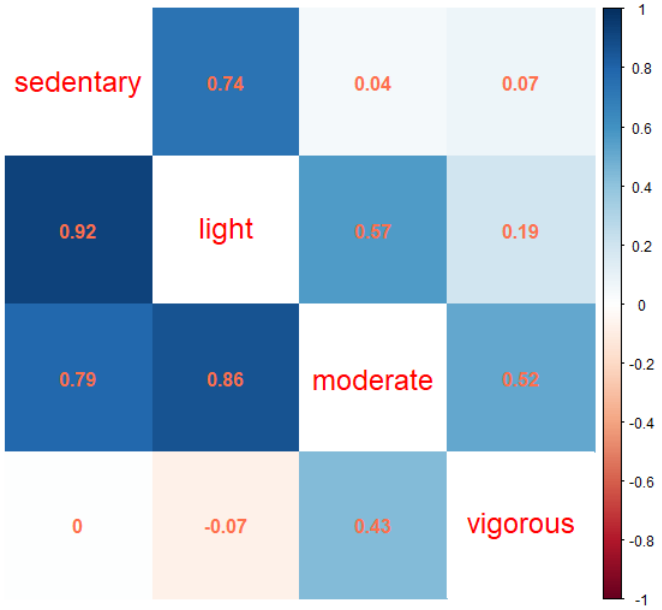
Při využívání ρ statistiky k analýze kompozičních dat je nutné mít na mysli, že hodnoty závisí na volbě referenční kompozice R. Doposud jsme jako referenční používali množinu všech zbývajících složek. Zajímá nás ale, jaký bude rozdíl, když zvolíme jako referenční nejprve první ze zbývajících složek a následně druhou v pořadí. Abychom mohli porovnat kontrast mezi variantami, zobrazíme si

jednotlivé možnosti do horního a dolního trojúhelníku matice ρ statistik (obrázek 3.11). Následně porovnáme, zda je matice přibližně symetrická či nikoliv.



Obrázek 3.10: ρ statistika pro datový soubor `activity`.

Vidíme, že například pro asociaci mezi sedavou a středně náročnou aktivitou získáme velmi odlišné výsledky. Pokud se podíváme i na předchozí hodnoty z obrázků 3.7, 3.8 a 3.10, zjistíme, že vztah vždy závisí na tvaru příslušné bilance (resp. na ostatních složkách v nich vystupujících). Opět jsme došli k závěru, že informaci o vztahu mezi dvěma složkami nelze oddělit od ostatních složek.



Obrázek 3.11: Matice ρ statistik pro různé referenční kompozice v dolním a horním trojúhelníku matice.

3.6 Testování

Nakonec bychom chtěli ověřit, zda jsou nalezené míry b-asociace statisticky signifikantní. Jinak řečeno, zda-li nejsou pouze náhodné.

3.6.1 Testování na jednotkovou směrnici

Budeme testovat nulovou hypotézu, že směrnice tečny lineárního modelu (2.4) se rovná jedné. Protože je test definován na základě ortogonální regrese, najdeme v knihovně `smatr` funkci `sma`, která má v sobě test implementován. Stačí zadat

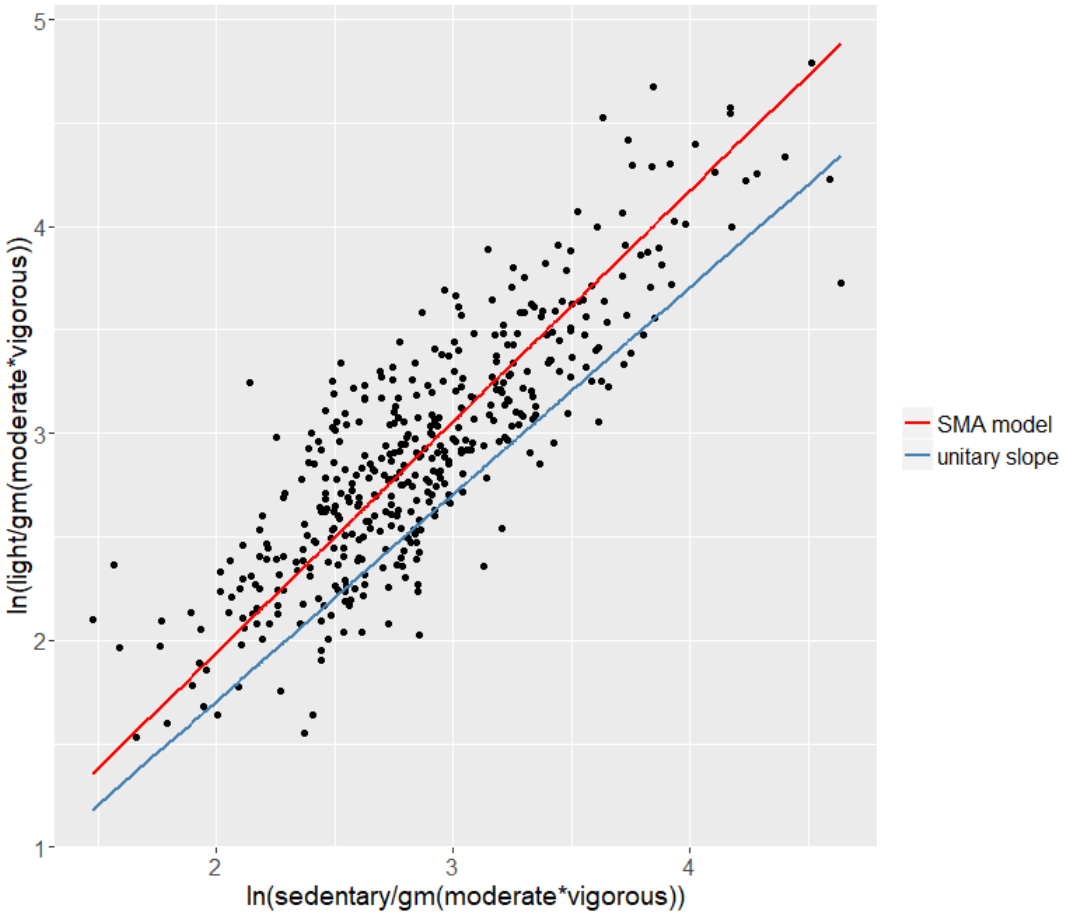
`> sma(y~x,method=c("SMA"),slope.test=1)`, kde za y a x zvolíme požadované bilance $B(H/R)$ a $B(G/R)$ dle vztahu (2.3). Za referenční kompozici bereme všechny zbývající složky kompozice mimo H a G. Hypotézu o jednotkové směrnici zamítáme, pokud je p-hodnota menší než 0,05 (bereme v úvahu 95 % hladinu spolehlivosti).

Když vezmeme jako referenční kompozici pro každou dvojici složek všechny zbývající složky, dostaneme pro datový soubor `activities` p-hodnoty z obrázku 3.12.

Obrázek 3.12: p-hodnoty testu na jednotkovou směrnici pro datový soubor `activities`.

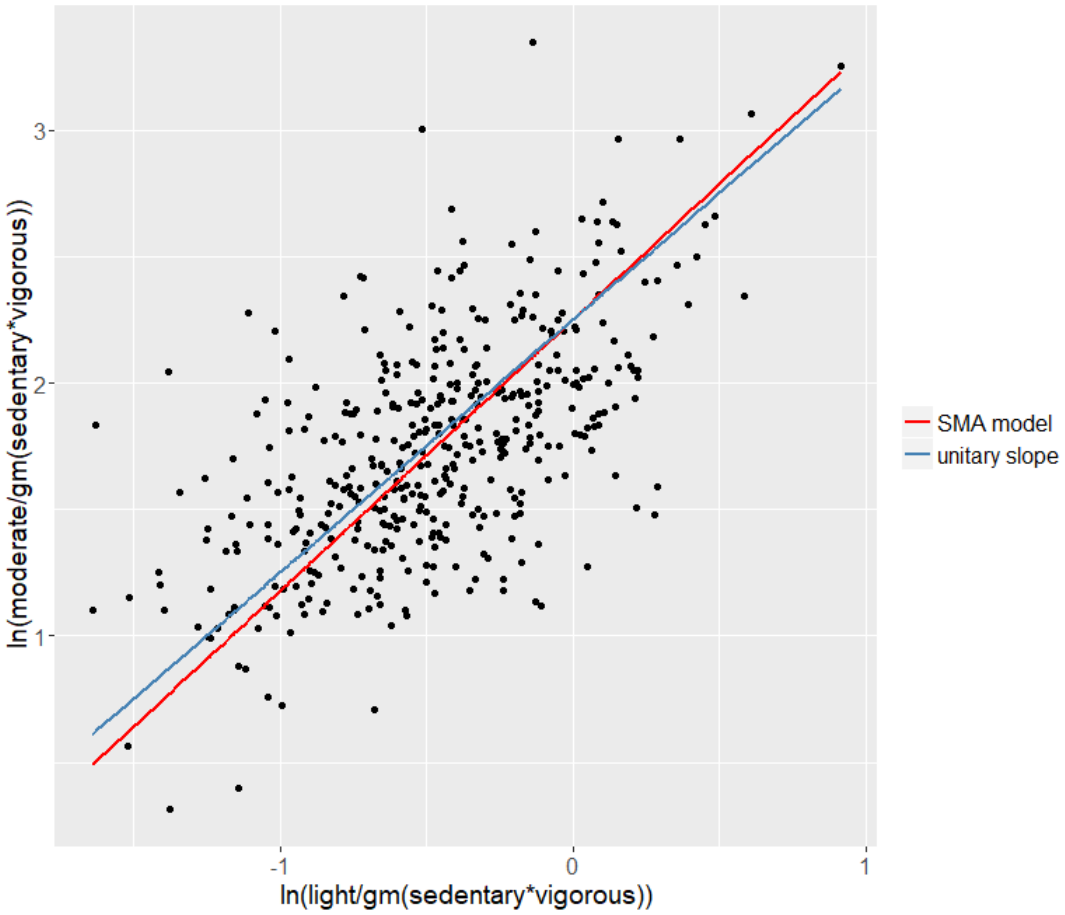
Hypotézu o jednotkové směrnici nezamítáme pouze pro bilance odpovídající dvojici složek lehké a středně náročné fyzické zátěže. Možná v rozporu s očekáváním zamítáme nulovou hypotézu pro první dvojici složek sedavé a lehce náročné aktivity. Předchozí výsledky totiž naznačovaly, že se mezi těmito bilancemi vyskytuje silná míra b-asociace. Je nutno si ale uvědomit, že nulová hypotéza odpovídá jednotkové směrnici, a ne přímo definici b-asociace. Abychom lépe pochopili, na jakém základě je test založen, zobrazíme si odhadnutou regresní přímku z or-

togonální regrese společně s daty a zároveň si vykreslíme i přímku se stejným absolutním členem, ale s jednotkovou směrnicí (obrázek 3.13).



Obrázek 3.13: SMA odhad regresní přímky a přímka s jednotkovou směrnicí pro bilance příslušné k sedavé a lehce náročné složce datového souboru `activity`.

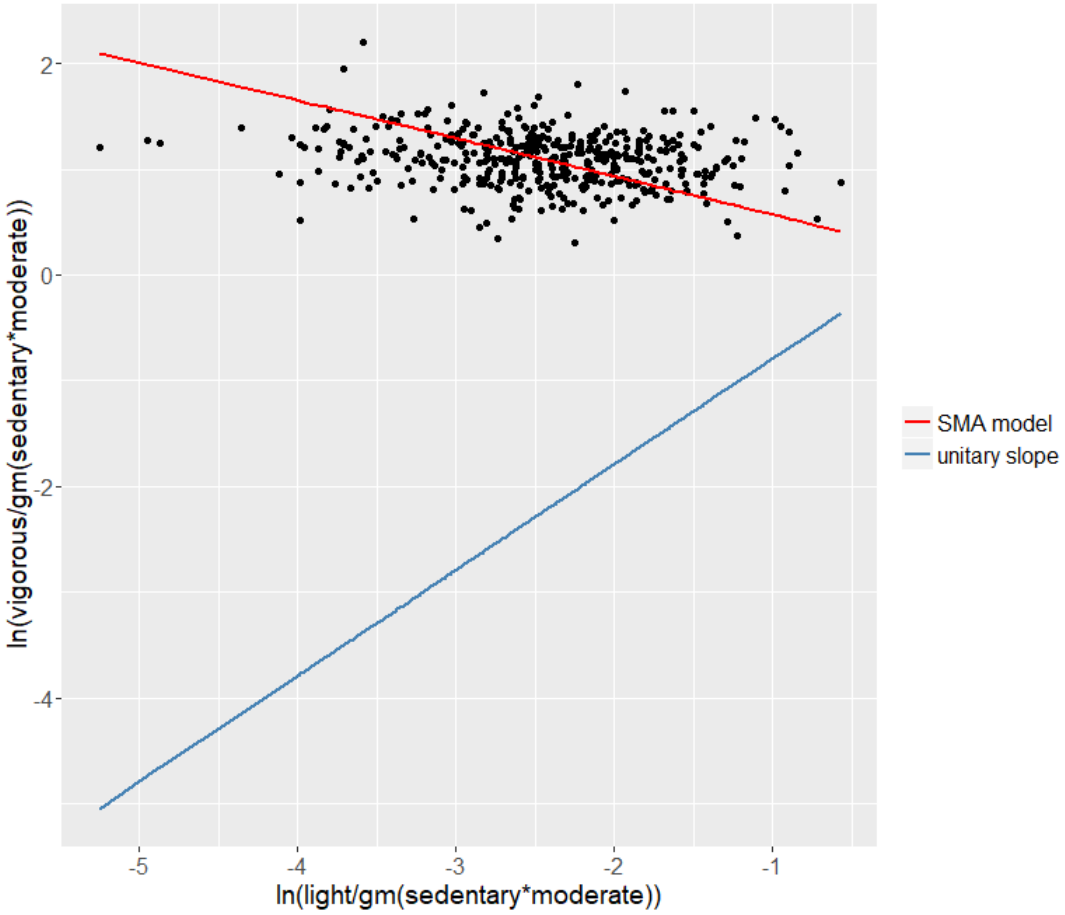
Srovnejme obrázek 3.13 s obrázkem 3.14 pro dvojici bilancí, pro kterou jsme nulovou hypotézu nezamítli. Vidíme, že ve druhém případě se přímky téměř překrývají. Tím se vysvětluje, proč testování ukazuje výsledky v nesouladu s ρ statistikou. V prvním obrázku 3.13 se zdá být vztah mezi bilancemi silnější, což potvrzuje i koeficient determinace 71 %. Nicméně test se rozhoduje na základě



Obrázek 3.14: SMA odhad regresní přímky a přímka s jednotkovou směrnicí pro bilance příslušné k lehce a středně náročné složce datového souboru `activity`.

odhadu parametru β_1 modelu a už neřeší, že druhý model z obrázku 3.14 vysvětluje jen 36 % variability dat.

Podobně je tomu i pro negativní SMA odhady regresního modelu, které nejsou testem vůbec reflektovány (nulová hypotéza sleduje pouze hodnotu $+1$, a ne -1). Proto je pro dvojice se zápornou ρ statistikou p-hodnota nulová. Příkladem může být vztah dvojice bilancí odpovídajících lehce náročné a náročné složce kompozice `activity` na obrázku 3.15, z kterého je zřejmé, že směrnice přímek mají opačná znaménka.



Obrázek 3.15: SMA odhad regresní přímky a přímka s jednotkovou směrnicí pro bilance příslušné k lehce náročné a náročné složce datového souboru `activity`.

3.6.2 Regresní test

Pro regresní test užijeme nejjednodušší model

$$B(G/H) = \gamma_0 + \gamma_1 B(G, H/R) + e,$$

kde za referenční kompozici R volíme vždy všechny zbývající složky kompozice a testujeme nulovou hypotézu $H_0'' : B(G/H) = \gamma_0 + e$ pomocí F-statistiky klasického regresního modelu. Nicméně pro všechny dvojice bilancí docházíme k téměř nulovým p-hodnotám, tudíž všechny nulové hypotézy zamítáme. Regresní test je pro data příliš striktní a nedává proto příliš mnoho informace. Navíc získáváme nesymetrické výsledky pro dvojice bilancí.

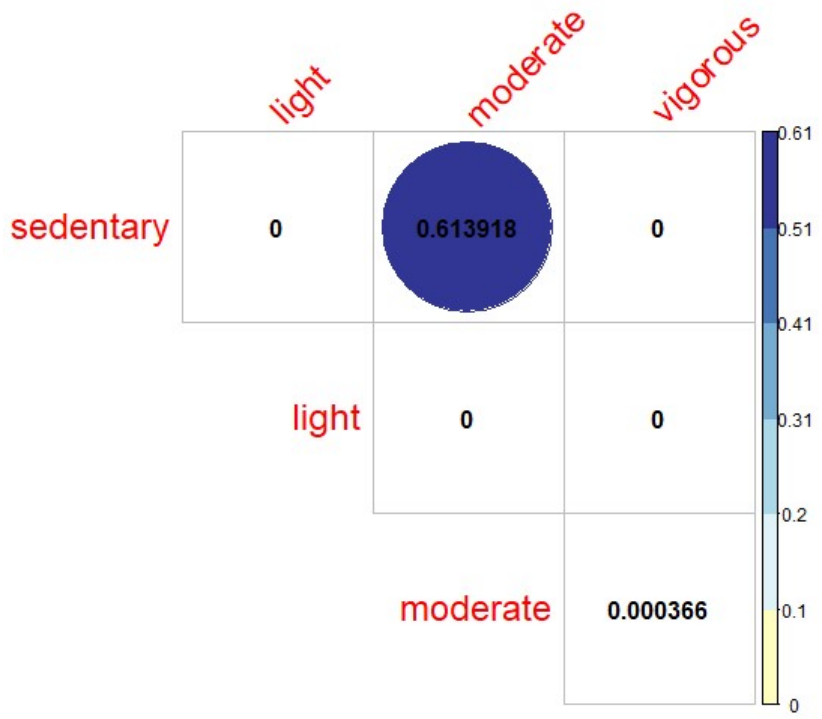
3.6.3 Test korelace mezi symetrickými pivotovými souřadnicemi

Pro úplnost si ještě otestujeme korelaci mezi symetrickými pivotovými souřadnicemi. Tentokrát budeme testovat nulovou hypotézu o nulovosti korelace mezi dvěma veličinami. Využijeme k tomu klasický test, který je implementován do funkce `cor.test`. Nejprve však vypočítáme hodnoty symetrických pivotových souřadnic využitím funkce `pivotCoord` z knihovny `robCompositions`, kde za metodu zvolíme variantu `"symm"`. Funkce pak počítá hodnoty symetrických pivotových souřadnic, které mají v čitateli postupně první a druhou složku datového souboru, který byl vstupem. Proto musíme datový soubor permutovat, abychom získali výsledky testů pro všechny dvojice složek (obrázek 3.16). Hypotézu nezamítáme pro dvojici bilancí odpovídající sedavé a středně náročné složce, což je v souladu s naším očekáváním po shlédnutí hodnot korelací mezi symetrickými pivotovými bilancemi na obrázku 3.7.

3.7 Diskuze

Ačkoliv variační matice z kapitol 2.1 a 3.3 řeší problémy standardního korelačního koeficientu aplikovaného na složky původní kompozice, přináší s sebou problémy nové. Zatímco interpretace korelačního koeficientu je celkem intuitivní, pochopit význam variační matice už nemusí být tak snadné. Navíc vůbec nerozlišuje znaménka asociací.

Korelace mezi zavedenými symetrickými pivotovými souřadnicemi se opírá o známý korelační koeficient, přičemž ale v konstrukci souřadnic vystupují i ostatní složky kompozice. Pracujeme tedy s více složkami současně, což je pro kompoziční data přirozené, musíme se s tím však naučit operovat. Jako nevýhodu metody se symetrickými pivotovými souřadnicemi lze uvést složitou konstrukci a komplikovaný tvar souřadnic. Na druhou stranu takto dosáhneme požadovaných



Obrázek 3.16: p-hodnoty testů o nulovosti korelačních koeficientů mezi dvojicemi symetrických pivotových souřadnic datového souboru `activity`.

vlastností bilancí a konečné výsledky metody už vypadají elegantně. Navíc díky funkci v knihovně `robCompositions` je použití velmi snadné.

Z obrázků 3.7 a 3.8 by se mohlo na první pohled zdát, že korelace mezi symetrickými pivotovými souřadnicemi nedodržuje podkompoziční soudržnost, když jsme dospěli k jiným výsledkům při použití celé kompozice a poté podkompozice. Je nutné si ale uvědomit, že pro každý případ jsme sestrojili jiné symetrické pivotové souřadnice. Poté už jsou odlišné výsledky jasnější.

Obdobně si lze „hrát“ se zapojováním různých složek kompozice i u měření míry b-asociace. Tím nejlépe pochopíme strukturu vztahů mezi složkami kompozičních dat. Ukazuje se, že pro měření míry b-asociace je lepší ρ statistika, protože ϕ statistika postrádá požadovanou vlastnost symetrie. Jelikož je ρ statistika zavedena jinak než korelace mezi symetrickými pivotovými souřadnicemi, provedená analýza naznačuje, že míra b-asociace ukazuje nižší míru negativních vztahů. Ale

i přesto se zjištěné závěry z obou metod nevylučují, naopak jsou do určité míry analogické.

Nakonec jsme se zabývali testováním b-asociací navrženým v článku [7]. Přitom jsme implicitně předpokládali normalitu jednotlivých vstupních veličin. Došli jsme k podobným závěrům jako autoři, a to že testování b-asociací nedává uspokojivé výsledky. Upravená nulová hypotéza testu na jednotkovou směrnici neodpovídá výsledkům získaným výpočtem měr b-asociace. Regresní test je příliš striktní, obzvláště při vyšším počtu pozorování, a zamítá všechny nulové hypotézy. Tudíž pomocí něho nejsme schopni nijak rozlišit míru asociace mezi skupinami kompozičních složek, reprezentovanými pomocí bilancí. Oba testy nejsou citlivé na velká rezidua modelů, což jsme demonstrovali na obrázku 3.14, kde nezamítáme hypotézu o jednotkové směrnici, možná protože dosahujeme pouze nízkého koeficientu determinace. Testování korelace mezi symetrickými pivotovými souřadnicemi funguje podle očekávání a je v souladu s vypočítanými korelacemi.

Závěr

Diplomová práce představuje možnosti korelační analýzy kompozičních dat. První část rekapituluje základní poznatky v oblasti kompozičních dat a poukazuje na nutnost speciálního přístupu k analýze míry asociací mezi složkami kompozic. Jako výchozí bod uvádí variační matici, která využívá logpodílového pohledu na kompoziční data.

Dále pak práce seznamuje se sofistikovanější metodou založenou na symetrických pivotových souřadnicích, které jsou speciálně určeny pro výpočet korelace. Takto lze získat již mnohem zajímavější výsledky. I druhý přístup využívající b-asociace je postaven na bilancích, ale používá jiné míry asociace než předchozí metoda. Výsledky však mají podobný charakter jako u symetrických pivotových souřadnic. Navržené testování b-asociací dává rozporuplné závěry, a tudíž na jejich základě není možné provést relevantní rozhodnutí.

Práce potvrzuje, že s kompozicí je vždy potřeba pracovat jako s celkem a získané transformace původních složek vhodným způsobem interpretovat. Problematika korelační analýzy kompozičních dat je velmi komplexní a je potřebné ji dále rozvíjet.

Literatura

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

[2] Aitchison, J.: *A concise guide to compositional data analysis.* University of Glasgow, 2005.

[3] van den Boogaart, K. G., Tolosana-Delgado, R.: *Analyzing compositional data with R.* Springer, Heidelberg, 2013.

[4] Buccianti A., Pawlowsky-Glahn, V.: *New perspectives on water chemistry and compositional data analysis.* Math Geol, 37(7), 703–727, 2005.

[5] Chayes, F.: *On correlation between variables of constant sum.* Journal of Geophysical Research, 65(12), 4185-4193, 1960.

[6] Egozcue, J. J., Lovell, D., Pawlowsky-Glahn, V.: *Testing compositional association.* Proceedings of the 5th workshop on compositional data analysis -CoDaWork, 2013.

[7] Egozcue, J. J., Pawlowsky-Glahn, V., Gloor, G. B.: *Linear association in compositional data analysis.* Austrian Journal of Statistics, 47(1), 3–31, 2018.

[8] Evenson, K. R., Catellier, D. J., Gill, K., Ondrak, K.S., McMurray, R.G.: *Calibration of two objective measures of physical activity for children.* J Sports Sci, 26, 1557–65, 2008.

[9] Filzmoser, P., Hron, K., Reimann, C.: *Univariate statistical analysis of environmental (compositional) data: Problems and possibilities.* Sci Total Environ, 407(23), 6100–6108, 2009.

[10] Gába, A., Mitáš, J., Jakubec, L.: *Associations between accelerometer-measured physical activity and body fatness in school-aged children.* Environmental Health and Preventive Medicine, 22(1), 2017.

[11] Gába, A., Dygrýn, J., Mitáš, J., Jakubec, L., Frömel, K.: *Effect of Accelerometer Cut-Off Points on the Recommended Level of Physical Activity for Obesity Prevention in Children.* PLoS ONE, 11(10), 2017.

[12] Gardlo, A.: *Diplomová práce: Kompoziční biplot.* Univerzita Palackého v Olomouci, 2012.

[13] Juráňová, K.: *Bakalářská práce: Statistická analýza kompozičních dat pomocí knihovny „compositions“ softwaru R.* Univerzita Palackého v Olomouci, 2017.

[14] Hron, K., Kunderová, P.: *Základy počtu pravděpodobnosti a metod matematické statistiky.* Univerzita Palackého v Olomouci, 2013.

[15] Kynčlová, P., Hron, K., Filzmoser, P.: *Correlation between compositional parts based on symmetric balances.* Math Geosci, 49(6), 777–796, 2017.

[16] Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S., Bähler, J.: *Proportionality: A valid alternative to correlation for relative data.* PLoS Comput Biol, 11(3), 2015. Dostupné z `https://journals.plos.org/ploscompbiol/article/file?id=10.1371/journal.pcbi.1004075&type=printable`.

[17] Pawlowsky-Glahn, V., Egozcue, J. J. and Tolosana-Delgado, R.: *Lecture notes on compositional data analysis.* Universitat de Girona, 2007 [online].[cit. 2016-10-23]. Dostupné z `http://dugi-doc.udg.edu/bitstream/handle/10256/297/CoDa-book.pdf?sequence=1`.

[18] Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, R.: *Modeling and analysis of compositional data.* Wiley, Chichester, 2015.

[19] Pearson, K.: *Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs.* Proceedings of the Royal Society of London, LX, 489–502, 1897.

[20] Warton, D. I., Wright, I. J., Falster, D. S., Westoby, M.: *Bivariate line-fitting methods for allometry.* Biological Reviews, 81, 259–291, 2006.