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

DIPLOMOVÁ PRÁCE

Kompoziční analýza dvourozměrných diskrétních pravděpodobností

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

Vypracovala:
Lenka Štefková
AME, II. ročník

Prohlášení

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

V Olomouci dne 8. dubna 2010

Poděkování

Na tomto místě bych chtěla poděkovat především svému vedoucímu diplomové práce panu RNDr.Karlu Hronovi Ph.D. za obětavě věnovaný čas a podporu při psaní práce. Můj dík patří také mé rodině a přátelům za to, že mi připravili podmínky, které vznik této práce umožnily.

Obsah

Úvod

V mé diplomové práci jsem se zaměřila na problém analýzy kontingenčních tabulek s využitím poznatků o kompozičních datech jako pozorováních nesoucích pouze relativní informaci. Výchozím bodem úvah je skutečnost, že na kontingenční tabulku (resp. příslušnou tabulku pravděpodobností) lze pohlížet jako na kompozici a pro řešení souvisejících problémů je možno použít myšlenky kompoziční analýzy postavené na Aitchisonově geometrii na simplexu.

Těžiště práce tvoří zpracování stěžejního článku *Compositional analysis of bivariate discrete probabilities* [4], který pojednává o kompoziční analýze obecné kontingenční tabulky. Pro lepší názornost a snažší orientaci čtenáře jsem se v této práci omezila na případ tabulky čtyřpolní. Tento přístup mi mimo jiné umožnil odvození některých tvrzení, která jsou v článku uvedena bez důkazu a měl by vést k větší přehlednosti nově zaváděné terminologie.

Práce je členěna do třech hlavních kapitol. V první kapitole má čtenář možnost seznámit se s kompozičními daty a se souvisejícími pojmy, se kterými pracuji v dalších kapitolách. Druhá kapitola přestavuje kontingenční tabulku a nejpoužívanější přístupy k její analýze. Slouží zejména pro uvedení do kontextu a srovnání s kapitolou třetí, která se zabývá již zmíněnou kompoziční analýzou kontingenčních tabulek. V závěrečné čtvrté kapitole jsou na didaktickém příkladu aplikovány získané poznatky z kompoziční analýzy na rozbor konkrétní čtyřpolní kontingenční tabulky.

1. Kompoziční data

V této kapitole se seznámíme s důležitými poznatky o kompozičních datech, která byla poprvé uceleně popsána Johnem Aitchisonem v polovině osmdesátých let 20. století v knize [1]. Z ní jakož i ze zdrojů [5], [6] a [8] jsem také při tvorbě následujícího textu nejvíce čerpala.

Definice 1. *Řádkový vektor* $\mathbf{x} = (x_1, \ldots, x_D)$ *nazýváme D-složkovou kompozicí, jestliže všechny složky tohoto vektoru jsou kladná reálná čísla nesoucí pouze relativní informaci.*

Kompoziční data jsou tedy definována jako vícerozměrná data, u kterých nezáleží na absolutní velikosti složek, důležité jsou pouze podíly mezi složkami vektoru. To znamená, že kompozice $(x_1, \ldots, x_D)$ a $(ax_1, \ldots, ax_D)$, pro $a > 0$, nám sdělují totožnou informaci.

Uvedená vlastnost přináší zjednodušení práce s kompozicemi, neboť můžeme všechny kompozice reprezentovat jako kladné vektory se součtem složek rovným konstantě k, aniž by došlo ke ztrátě informace. Hovoříme pak o operaci uzávěru kompozice.

Definice 2. *Uzávěr kompozice* $\mathbf{x} = (x_1, \ldots, x_D)$ *na konstantu* k *označíme* $C(\mathbf{x})$ *a definujeme vztahem*

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

Definice 3. *Výběrovým prostorem kompozičních dat je D-složkový simplex,*

$$S^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D),\ x_i > 0,\ i = 1, 2, \ldots, D,\ \sum_{i=1}^{D} x_i = k \right\}.$$

Simplex je tedy množina všech kompozičních dat se součtem složek rovným k.

Často se naše pozornost zaměřuje jen na některé složky kompozice, které jsou pro nás relevantní. Mluvíme pak o tzv. podkompozici, která obsahuje pouze vybrané složky původní kompozice.

Definice 4. *Mějme kompozici* $\mathbf{x} = C(x_1, \ldots, x_D)$ *v prostoru* S^D *a množinu indexů* $R = \{i_1, \ldots, i_r\}$. *R-podkompozici* $\mathbf{x}$ *reprezentuje následující vektor v* S^r

$$sub(\mathbf{x}; R) = C(x_{i_1}, x_{i_2}, \ldots, x_{i_r}).$$

Indexy $i_1, \ldots, i_r$ značí, které složky vybereme do podkompozice (ne nutně prvních r) a na vektor těchto složek pak aplikujeme operaci uzávěru.

1.1. Aitchisonova geometrie na simplexu

Pokud chceme pracovat s kompozičními daty na simplexu analogicky jako s běžnými mnohorozměrnými pozorováními v reálném vektorovém prostoru, musíme zavést vhodnou geometrii, která nám toto umožní.

Definice 5. *Perturbace kompozice* $\mathbf{x} = C(x_1, \ldots, x_D) \in S^D$ *s kompozicí* $\mathbf{y} = C(y_1, \ldots, y_D) \in S^D$ *je definována jako*

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

Výsledek perturbace, ani následně zavedené mocninné transformace, nezávisí na konstantách uzavírajících vektory $\mathbf{x}$ a $\mathbf{y}$, protože v obou definicích je zahrnuta operace uzávěru.

Perturbace splňuje následující vlastnosti:

1. komutativitu: $\mathbf{x} \oplus \mathbf{y} = \mathbf{y} \oplus \mathbf{x}$,
2. asociativitu: $(\mathbf{x} \oplus \mathbf{y}) \oplus \mathbf{z} = \mathbf{x} \oplus (\mathbf{y} \oplus \mathbf{z})$,
3. existenci neutrálního prvku: $\mathbf{e} = C(1, 1, \ldots, 1)$,
4. existenci inverzního prvku: $\mathbf{x}^{-1} = C(x_1^{-1}, x_2^{-1}, \ldots, x_D^{-1})$, kde $\mathbf{x} \oplus \mathbf{x}^{-1} = \mathbf{e}$.

Struktura $(S^D, \oplus)$ proto tvoří komutativní grupu. Perturbaci $\mathbf{x}$ s inverzním prvkem k $\mathbf{y}$ značíme symbolem $\ominus$: $\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus \mathbf{y}^{-1}$; zřejmě $\mathbf{x} \ominus \mathbf{x} = \mathbf{e}$.

Definice 6. *Mocninná transformace kompozice* $\mathbf{x} = C(x_1, \ldots, x_D) \in S^D$ *konstantou* $\alpha \in \mathbb{R}$ *je dána vztahem*

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

Pro $\alpha, \beta \in \mathbb{R}$ a $\mathbf{x}, \mathbf{y} \in S^D$ splňuje mocninná transformace vlastnosti:

1. asociativitu: $\alpha \odot (\beta \odot \mathbf{x}) = (\alpha \cdot \beta) \odot \mathbf{x}$,
2. distributivitu zleva: $\alpha \odot (\mathbf{x} \oplus \mathbf{y}) = (\alpha \odot \mathbf{x}) \oplus (\alpha \odot \mathbf{y})$,
3. distributivitu zprava: $(\alpha + \beta) \odot \mathbf{x} = (\alpha \odot \mathbf{x}) \oplus (\beta \odot \mathbf{x})$,
4. existenci neutrálního prvku 1, neboť $1 \odot \mathbf{x} = \mathbf{x}$.

Vidíme tedy, že simplex spolu s perturbací a mocninnou transformací tvoří vektorový prostor $(S^D, \oplus, \odot)$.

Vektorová struktura simplexu nám umožňuje definovat pojem perturbační závislost a nezávislost.

Definice 7. *Kompozice $\mathbf{x}_1, \ldots, \mathbf{x}_m \in S^D$ se nazývají perturbačně závislé, jestliže existuje alespoň jedno nenulové číslo $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$ takové, že platí*

$$(\alpha_1 \odot \mathbf{x}_1) \oplus \cdots \oplus (\alpha_m \odot \mathbf{x}_m) = \mathbf{e},$$

kde $\mathbf{e} = C(1, \ldots, 1)$.

V opačném případě se kompozice $\mathbf{x}_1, \ldots, \mathbf{x}_m$ nazývají perturbačně nezávislé. Na simplexu S^D je maximální počet perturbačně nezávislých kompozic $D - 1$. S^D je tedy vektorový prostor dimenze $D - 1$.

Nechť $\mathbf{e}_1, \ldots, \mathbf{e}_{D-1} \in S^D$ jsou perturbačně nezávislé a tvoří tedy bázi S^D. Každou kompozici $\mathbf{x} \in S^D$ pak můžeme vyjádřit v následujícím tvaru pomocí bázových kompozic,

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

kde $\alpha_i \in \mathbb{R},\ i = 1, \ldots, D-1$.

Definice 8. *Koeficienty $\alpha_i,\ i = 1, \ldots, D-1$, ze vztahu (1) nazýváme souřadnice $\mathbf{x}$ vzhledem ke zvolené bázi.*

Máme-li danou bázi, jsou souřadnice $\mathbf{x}$ určeny jednoznačně. Proto můžeme reprezentovat kompozici jejími souřadnicemi vzhledem k dané bázi.

Již víme, že struktura $(S^D, \oplus, \odot)$ je vektorovým prostorem. Nyní nadefinujeme skalární součin a z něj odvozenou normu a vzdálenost. Dá se ukázat, že $(S^D, \oplus, \odot)$ je $(D-1)$-dimenzionálním prostorem Hilbertovým.

Definice 9. *Skalární součin kompozic* $\mathbf{x}, \mathbf{y} \in S^D$ *je definován jako*

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \sum_{i=1}^{D} (\ln x_i \cdot \ln y_i) - \frac{1}{D} \left(\sum_{j=1}^{D} \ln x_j \right) \cdot \left(\sum_{k=1}^{D} \ln y_k \right).$$

Ekvivalentně můžeme psát

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

kde $g(\mathbf{x}) = \sqrt[D]{x_1 \cdot \ldots \cdot x_D}$.

Takto definovaný skalární součin splňuje vlastnosti:

1. $\langle \mathbf{x}, \mathbf{x} \rangle_a \geq 0,\ \langle \mathbf{x}, \mathbf{x} \rangle_a = 0 \Leftrightarrow \mathbf{x} = \mathbf{e}$,
2. komutativitu $\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle \mathbf{y}, \mathbf{x} \rangle_a$,
3. distributivitu 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. linearitu vzhledem k násobení skalárem $\langle c \cdot \mathbf{x}, \mathbf{y} \rangle_a = c \cdot \langle \mathbf{x}, \mathbf{y} \rangle_a$.

Definice 10. *Norma kompozice* $\mathbf{x} \in S^D$ *je dána*

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

Definice 11. *Vzdálenost kompozic* $\mathbf{x}, \mathbf{y} \in S^D$ *definujeme*

$$d_a(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} \ominus \mathbf{y}||_a.$$

Strukturu $(S^D, \oplus, \odot)$ spolu se zavedenými pojmy skalárního součinu, normy a vzdálenosti nazýváme souhrnně *Aitchisonovou geometrií na simplexu* a hovoříme o Aitchisonově skalárním součinu, normě a vzdálenosti.

Vektorový prostor se skalárním součinem splňujícím výše uvedené vlastnosti 1.–4., a tedy i struktura $(S^D, \oplus, \odot, \langle \cdot, \cdot \rangle_a)$ je Hilbertovým prostorem. To znamená, že simplex S^D je plně ekvivalentní reálnému prostoru $\mathbb{R}^{D-1}$.

1.2. Logratio transformace kompozic

Pro praktické použití vyjádříme kompozice pomocí koeficientů vzhledem ke generujícímu systému (clr transformace) nebo přímo pomocí souřadnic vzhledem k bázi na simplexu (alr, ilr transformace) a pracujeme s nimi jako se standardními vektory v reálném vektorovém prostoru. Každá ze tří zmíněných transformací má odlišné vlastnosti, které určují jejich užití.

Clr transformace

Definice 12. *Centrovanou logratio (clr) transformací* [1] *kompozice* $\mathbf{x} \in S^D$ *rozumíme vektor*

$$clr(\mathbf{x}) = \left(\ln \frac{x_1}{g(\mathbf{x})}, \ln \frac{x_2}{g(\mathbf{x})}, \ldots, \ln \frac{x_D}{g(\mathbf{x})} \right).$$

Složky tohoto vektoru nazýváme clr koeficienty kompozice $\mathbf{x}$.

Věta 1. *Nechť* $\mathbf{x}, \mathbf{y} \in S^D$ *a* $\alpha, \beta \in \mathbb{R}$, *pak platí*

1. $clr(\alpha \odot \mathbf{x} \oplus \beta \odot \mathbf{y}) = \alpha \cdot clr(\mathbf{x}) + \beta \cdot clr(\mathbf{y})$,
2. $\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle clr(\mathbf{x}), clr(\mathbf{y}) \rangle$,
3. $||\mathbf{x}||_a = ||clr(\mathbf{x})||, \quad d_a(\mathbf{x}, \mathbf{y}) = d(clr(\mathbf{x}), clr(\mathbf{y}))$.

Clr transformace tedy zachovává vzdálenosti a umožňuje nahradit perturbaci a mocninnou transformaci operacemi sčítání vektorů a násobení vektoru skalárem (je izometrií ze simplexu S^D do reálného prostoru $\mathbb{R}^D$). Výhodou je též symetrie

vzhledem ke složkám, ale použití nástrojů pro statistickou analýzu komplikuje singularita dat (součet koeficientů transformované kompozice je roven nule).

Alr transformace

Definice 13. *Aditivní logratio (alr) transformace* [1] *kompozice* $\mathbf{x} \in S^D$ *je vektor*

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

Alr transformací získáme souřadnice vzhledem k bázi na simplexu, neboť libovolná kompozice $\mathbf{x} \in S^D$ může být zapsána v následujícím tvaru,

$$\mathbf{x} = (a_1 \odot \mathbf{e}_1) \oplus (a_2 \odot \mathbf{e}_2) \oplus \cdots \oplus (a_{D-1} \odot \mathbf{e}_{D-1}),$$

kde $a_i = \ln \frac{x_i}{x_D},\ i = 1, \ldots, D-1$, a vektory $\mathbf{e}_i = C(1, 1, \ldots, e, \ldots, 1),$ (e na i-té pozici) tvoří bázi na simplexu S^D.

Alr transformace představuje zobrazení z S^D do $\mathbb{R}^{D-1}$ a převádí operace na simplexu na obyčejné sčítání vektorů a násobení vektoru skalárem (stejně jako clr transformace). Zásadním problémem je ovšem fakt, že tato transformace není izometrií ze simplexu s Aitchisonovou metrikou do reálného prostoru s eukleidovskou metrikou (nezachovává vzdálenosti). Alr transformace je asymetrická vzhledem ke složkám, neboť záleží na volbě složky ve jmenovateli (v našem případě x_D).

Ilr transformace

Vyjádřit kompozice pomocí souřadnic vzhledem k ortonormální bázi na simplexu umožňuje tzv. ilr transformace.

Mějme ortonormální bázi na simplexu $\{\mathbf{e}_1, \ldots, \mathbf{e}_{D-1}\}$. Kompozici $\mathbf{x} \in S^D$ je možno vyjádřit následovně,

$$\mathbf{x} = (x_1^* \odot \mathbf{e}_1) \oplus (x_2^* \odot \mathbf{e}_2) \oplus \ldots (x_{D-1}^* \odot \mathbf{e}_{D-1}), \quad x_i^* = \langle \mathbf{x}, \mathbf{e}_i \rangle_a .$$

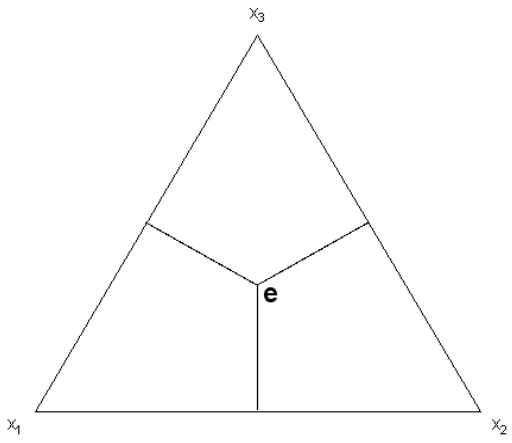
Potom $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_{D-1}^*)$ je vektor souřadnic $\mathbf{x}$ vzhledem k vybrané bázi. Funkci $ilr : S^D \to \mathbb{R}^{D-1}$, přiřazující souřadnice $\mathbf{x}^*$ k $\mathbf{x}$, nazýváme izometrická logratio (ilr) transformace (viz [6]).

Ilr transformace je izometrie z S^D do $\mathbb{R}^{D-1}$ a splňuje tedy všechny vlastnosti věty 1. Blíže se jí budeme zabývat i v dalších podkapitolách.

1.3. Ternární diagram a ortonormální báze

Existují různé možnosti grafického zobrazení kompozičních dat. Dvousložkové kompozice zobrazíme jako body v intervalu $(0, k)$. Pro trojsložkové kompozice používáme ternární, pro čtyřsložkové potom tetrahedrální diagram. Zobrazujeme tak vlastně vždy reprezentace kompozic (s předepsaným součtem složek k) na simplexu.

Kompozice $\mathbf{x} = (x_1, x_2, x_3)$, zobrazená v ternárním diagramu, je bodem uvnitř rovnostranného trojúhelníku, přičemž vzdálenosti tohoto bodu od jednotlivých stran představují hodnoty jednotlivých složek. V následujícím obrázku je zakreslena kompozice $\mathbf{e}$ (neutrální prvek), jejíž složky jsou si rovny, a leží tedy v těžišti daného trojúhelníku.



Obrázek 1: Reprezentace trojsložkové kompozice $\mathbf{e} = C(1, 1, 1)$.

Jak již bylo zmíněno, S^D je vektorový prostor dimenze $D-1$. Báze v S^D musí tedy obsahovat $D-1$ perturbačně nezávislých vektorů, ozn. $\mathbf{e}_1, \ldots, \mathbf{e}_{D-1} \in S^D$. Pokud navíc tyto vektory splňují následující podmínky,

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

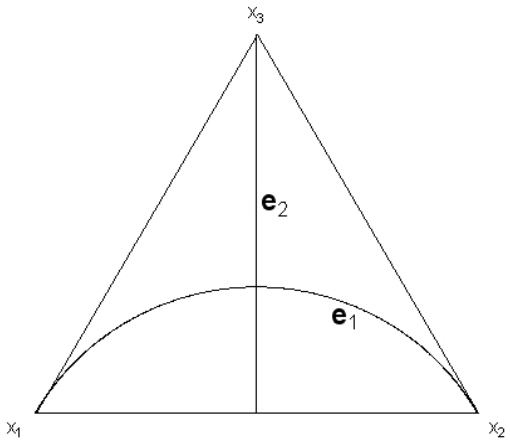
pak tvoří ortonormální bázi v S^D.

Příkladem ortonormální báze v S^3 jsou kompozice

$$\mathbf{e}_1 = C\left(\exp\left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0\right)\right),$$

$$\mathbf{e}_2 = C\left(\exp\left(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}}\right)\right).$$

V ternárním diagramu by pak směry dané těmito dvěma bázovými vektory (vzhledem k Aitchisonově geometrii) vypadaly následovně:



Obrázek 2: Směry dané bázovými vektory na simplexu, zakreslené v ternárním diagramu.

1.4. Postupný binární rozklad a bilance

Vytvoření ortonormální báze na simplexu není tak intuitivní jako je tomu v reálném vektorovém prostoru (na simplexu neexistuje kanonická báze). Abychom získali ortonormální bázi na simplexu, používáme různé postupy. V této kapitole se zaměříme na jeden z nich, a to na postupný binární rozklad (angl. sequential binary partition), který je veden snahou o co nejlepší názornost a interpretovatelnost získaných výsledků.

Při postupném binárním rozkladu kompozičního vektoru postupujeme po krocích následujícím způsobem. V prvním kroku rozdělíme složky dané kompozice do

dvou skupin. Obě získané skupiny v dalším kroku rozdělíme do dvou podskupin a tak dále, dokud se všechny skupiny nerozdělí do prvků (neboli jednotlivých složek kompozice). Počet kroků nutný k provedení postupného binárního rozkladu kompozice $\mathbf{x} \in S^D$ je $D-1$ a je přímo spojen se získáním $D-1$ ortonormálních bázových kompozic na S^D.

krok	x_1	x_2	x_3	x_4	r	s
1	$+1$	$+1$	-1	-1	2	2
2	$+1$	-1	0	0	1	1
3	0	0	$+1$	-1	1	1

Tabulka 1: Schéma znázorňující postupný binární rozklad kompozice.

Tabulka 1 zobrazuje postupný proces dělení čtyřsložkové kompozice. V každém kroku rozkládáme skupinu z předchozí úrovně do dvou nových podskupin, prvkům jedné z nich přiřazujeme symbol +1, prvkům z druhé pak -1. Symbol 0 znamená, že se složka na dané úrovni rozkladu neúčastní.

Postupným binárním rozkladem získáme ortonormální kompozice v tomto tvaru,

$$\mathbf{e}_i = C(exp(a_{i1}, a_{i2}, \ldots, a_{iD-1})),\ i = 1, \ldots, D-1,$$

kde koeficienty a_{ij} nabývají hodnot v závislosti na zvoleném postupu rozkladu. Předpokládejme, že v i-tém kroku je skupina $r+s$ složek rozdělena do skupin o r složkách (symbol +1) a s složkách (symbol -1). Koeficienty a_{ij} jsou pak ve tvaru

$$a_+ = \sqrt{\frac{s}{r(r+s)}},\ a_- = -\sqrt{\frac{r}{s(r+s)}},\ a_0 = 0,$$

a_+ pro složky první skupiny, a_- pro složky druhé skupiny, a_0 odpovídá složkám nefigurujícím v rozkladu.

Souřadnice kompozice $\mathbf{x}$ vzhledem k bázi, získané postupným binárním dělením, mají tvar

$$x_i^* = \sqrt{\frac{rs}{r+s}} \ln \frac{\left(\prod_+ x_j\right)^{1/r}}{\left(\prod_- x_k\right)^{1/s}}.$$

Souřadnice získané tímto postupem se nazývají *bilance*, vektory výsledné ortonormální báze nazýváme *bilanční prvky*.

Pro postupný binární rozklad podle tabulky 1 dostáváme následující bilance,

$$x_1^* = \ln \frac{\sqrt{x_1 x_2}}{\sqrt{x_3 x_4}}, \quad x_2^* = \frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2}, \quad x_3^* = \frac{1}{\sqrt{2}} \ln \frac{x_3}{x_4},$$

a bilanční prvky

$$\begin{aligned}
\mathbf{e}_1 &= C \exp\left(\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}\right), \\
\mathbf{e}_2 &= C \exp\left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0, 0\right), \\
\mathbf{e}_3 &= C \exp\left(0, 0, \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right).
\end{aligned}$$

Význam bilancí ozřejmí následující úvaha. Mějme kompozici z S^D, jejíchž q složek ($q \leq D$) jsme vybrali do skupiny Q. Dále předpokládejme, že skupina Q je rozdělena do dvou podskupin R a S obsahujících po řadě r a s složek, $q = r + s$. Hodnoty podílů mezi složkami skupiny Q sdělují kompoziční informaci o podkompozici Q. Podíly složek v rámci podkompozic R a S zachycují pouze část informace o kompozici Q, schází nám vztah mezi R a S. Pro zúplnění informace musíme přidat podíly jedné složky z R a druhé z S. Právě tuto dodatkovou informaci představuje *bilance mezi skupinami R a S*.

1.5. Projekce na podkompozice

Jak bylo uvedeno v definici 4, představuje podkompozice vektor skládající se z vybraných složek původní kompozice. Postup, kterým získáme žádanou podkompozici, je jednoduchý: z kompozice v S^D vybereme r složek, které určí novou kompozici popsanou množinou indexů R vybraných složek. Na tuto kompozici použijeme operaci uzávěru. Znamená to, že se přesuneme z S^D na S^r, dojde ke zmenšení dimenze z D na r. Redukujeme-li kompozici na $R-$podkompozici, je naším záměrem zbavit se nadbytečné informace o podílech mezi složkami, které nás dále nezajímají.

Je-li dána kompozice $\mathbf{x} \in S^D$, můžeme se tázat, jak najít podkompozici $\mathbf{x}_R \in S^D$, která by obsahovala pouze informaci $R-$podkompozice, a jejíž Aitchisonova vzdálenost $d_a(\mathbf{x}, \mathbf{x}_R)$ by byla minimální. Jednoduchým výpočtem dostáváme

$$\mathbf{x}_R = (x_1, x_2, \ldots, x_r, a, a, \ldots, a), \quad a = \left(\prod_{i=1}^{r} x_i \right)^{\frac{1}{r}},$$

kde předpokládáme bez újmy na obecnosti $R = \{1, 2, \ldots, r\}$.

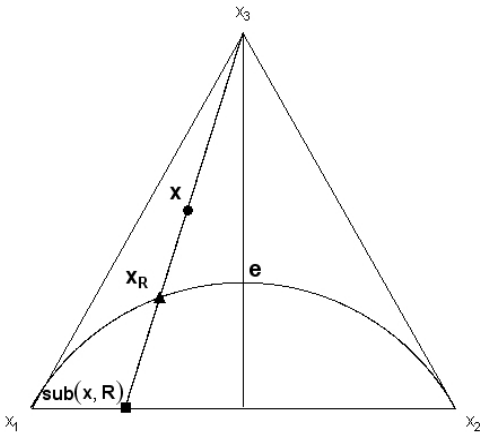
Za povšimnutí stojí poznatek, že

$$||\mathbf{x}_R||_a = ||sub(\mathbf{x}; R)||_a,$$

přičemž tyto normy jsou počítány na odlišných simplexech, jmenovitě S^D a S^r. To znamená, že $R-$podkompozice v S^r může být reprezentována pomocí $\mathbf{x}_R$, nacházející se stále v S^D, a tedy $\mathbf{x}_R$ je kompozice přidružená k $R-$podkompozici.

Tato dvojí reprezentace podkompozice je názorně zobrazena na obrázku 3.

Z geometrického hlediska představuje $\mathbf{x}_R$ ortogonální projekci $\mathbf{x}$ do podprostoru určeného podkompozicí. Tato operace může být provedena jednoduše užitím odpovídajících souřadnic. Pro nalezení vhodných souřadnic se užije postupný binární rozklad $R-$podkompozice. Například postupný binární rozklad z tabulky 1 obsahuje rozklad skupiny složek $R = \{3, 4\}$, který se provádí v kroku 3. Souřadnici x_3^* nazýváme přidruženou k $R-$podkompozici $sub(\mathbf{x}; \{3, 4\})$.



Obrázek 3: Původní kompozice $\mathbf{x} \in S^3$, podkompozice $sub(\mathbf{x}; \{1, 2\}) \in S^2$, ortogonální projekce $\mathbf{x}_R$.

2. Kontingenční tabulky

Sledujeme-li u statistických jednotek dva znaky, které jsou buď diskrétní a nabývají konečně mnoha hodnot nebo umožňují jednoznačné zařazení hodnot do kategorií, je obvyklé výsledky sledování zaznamenat do tvaru kontingenční tabulky, kdy jednotky roztřídíme podle variant znaků do $r \cdot c$ tříd, r je počet řádků a c počet sloupců tabulky. V této kapitole jsem čerpala zejména z [2],[7] a [10].

2.1. Základní pojmy

Mějme náhodný vektor (X, Y), který má diskrétní rozdělení, přičemž veličina X nabývá hodnot $1, \ldots, r$ a veličina Y hodnot $1, \ldots, c$. Označme

$$p_{ij} = P(X = i, Y = j), \quad p_{i.} = \sum_j p_{ij}, \quad p_{.j} = \sum_i p_{ij}.$$

Předpokládejme, že se uskutečnil výběr o rozsahu n z tohoto rozdělení. Nechť n_{ij} je počet těch případů, kdy se ve výběru vyskytla dvojice (i, j). Matici (n_{ij}) nazýváme *kontingenční tabulka* (viz tabulka 2). Podobně jako pro pravděpodobnosti píšeme i pro empirické četnosti

$$n_{i.} = \sum_j n_{ij}, \quad n_{.j} = \sum_i n_{ij}.$$

Zřejmě platí

$$n = \sum_i n_{i.} = \sum_j n_{.j} = \sum_i \sum_j n_{ij}.$$

Číslům $p_{i.}$ a $p_{.j}$ se říká *marginální (řádkové a sloupcové) pravděpodobnosti* a hodnotám $n_{i.}$ a $n_{.j}$ *marginální četnosti*.

Jak již bylo řečeno, ke vzniku takovéto dvourozměrné kontingenční tabulky dochází, když na statistických jednotkách sledujeme dva znaky. Tyto znaky mohou být buď kvalitativní nebo kvantitativní, rozhodující je schopnost zařadit jejich hodnoty jednoznačně do tříd.

Častou úlohou při rozboru dvourozměrných kontingenčních tabulek je otestování nezávislosti znaků X a Y.

	Y			
X	1	$\dots$	c	$\sum$
1	n_{11}	$\dots$	n_{1c}	$n_{1.}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
r	n_{r1}	$\dots$	n_{rc}	$n_{r.}$
$\sum$	$n_{.1}$	$\dots$	$n_{.c}$	n

Tabulka 2: Kontingenční tabulka typu $r \times c$.

2.2. Test nezávislosti

Uspořádání četností v kontingenční tabulce může naznačovat závislost obou znaků ve výběrovém souboru tím, že představuje rozdílná podmíněná rozdělení četností. Testem nezávislosti ověřujeme, zda takové uspořádání nemohlo vzniknout pouhou náhodou.

Věta 2. *Veličiny X a Y jsou nezávislé tehdy a jen tehdy, platí-li $p_{ij} = p_{i.}p_{.j}$ pro všechny dvojice (i, j).*

Proto *hypotézu nezávislosti* H_0 můžeme psát ve tvaru

$$H_0 \colon p_{ij} = p_{i.}p_{.j}, \quad i = 1, \dots, r; \; j = 1, \dots, c.$$

Pro odhady marginálních pravděpodobností lze odvodit vztahy (viz [2], str. 281)

$$\hat{p}_{i.} = \frac{n_{i.}}{n}, \quad i = 1, \dots, r,$$

$$\hat{p}_{.j} = \frac{n_{.j}}{n} \quad j = 1, \dots, c.$$

Veličina

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} \tag{2}$$

má při platnosti nulové hypotézy asymptoticky rozdělení χ^2 s počtem stupňů volnosti $(r-1)(c-1)$.

Hypotézu H_0 o nezávislosti veličin X a Y zamítáme na hladině významnosti α v případě, že realizace testovací statistiky překročí hodnotu příslušného $(1-\alpha)$-kvantilu, tedy $\chi^2 \geq \chi^2_{(r-1)(c-1)}(1-\alpha)$. Aby došlo ke shodě s limitním rozdělením, klade se požadavek, aby všechny teoretické četnosti $\frac{n_{i.}n_{.j}}{n}$ (někdy se uvádí alespoň 80% všech četností) byly větší než 5. Není-li tato podmínka splněna, obvykle se spojí některé řádky nebo sloupce a při testu se s nimi zachází jako s třídou jedinou.

2.3. Test χ^2 ve čtyřpolních tabulkách

Je-li $r = c = 2$, dostáváme čtyřpolní kontingenční tabulku

n_{11}	n_{12}	$n_{1.}$
n_{21}	n_{22}	$n_{2.}$
$n_{.1}$	$n_{.2}$	n

Věta 3. *Ve čtyřpolní kontingenční tabulce platí*

$$\chi^2 = n\frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}. \tag{3}$$

Důkaz. Užitím vztahu

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

a dosazením $n_{i2} = n_{i.} - n_{i1},\ n_{.2} = n - n_{.1}$ postupně dostáváme

$$\begin{aligned}
\chi^2 &= n\sum_{i=1}^{r}\left[\frac{\left(n_{i1} - \frac{n_{i.}n_{.1}}{n}\right)^2}{n_{i.}n_{.1}} + \frac{\left(n_{i2} - \frac{n_{i.}n_{.2}}{n}\right)^2}{n_{i.}n_{.2}}\right] \\
&= \frac{1}{n}\sum_{i=1}^{r}\left[\frac{(nn_{i1} - n_{i.}n_{.1})^2}{n_{i.}n_{.1}} + \frac{(n(n_{i.} - n_{i1}) - n_{i.}(n - n_{.1}))^2}{n_{i.}n_{.2}}\right] \\
&= \frac{1}{n}\sum_{i=1}^{r}(nn_{i1} - n_{i.}n_{.1})^2\left[\frac{1}{n_{i.}n_{.1}} + \frac{1}{n_{i.}n_{.2}}\right] = \frac{n^2}{n_{.1}n_{.2}}\sum_{i=1}^{r}n_{i.}\left(\frac{n_{i1}}{n_{i.}} - \frac{n_{.1}}{n}\right)^2 \\
&= \frac{n^2}{n_{.1}n_{.2}}\left[n_{1.}\left(\frac{n_{11}}{n_{1.}} - \frac{n_{.1}}{n}\right)^2 + n_{2.}\left(\frac{n_{21}}{n_{2.}} - \frac{n_{.1}}{n}\right)^2\right].
\end{aligned}$$

Ověříme-li, že

$$\left(\frac{n_{11}}{n_{1.}} - \frac{n_{.1}}{n}\right)^2 = \frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}^2 n^2},$$
$$\left(\frac{n_{21}}{n_{2.}} - \frac{n_{.1}}{n}\right)^2 = \frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{2.}^2 n^2},$$

pak již snadnou úpravou dostaneme vztah (3).

Při testu nezávislosti pro čtyřpolní tabulku vypočteme tedy χ^2 podle vztahu (3). Pokud vyjde $\chi^2 \geq \chi_1^2(1-\alpha)$, zamítáme hypotézu o nezávislosti.

2.4. Míry závislosti dvou proměnných

Statistika χ^2 pro test nezávislosti v dvojrozměrných tabulkách neměří stupeň závislosti mezi sledovanými znaky. V literatuře (např. [3], [7] nebo [10]) je popsána řada různých měr závislosti, které umožňují měřit intenzitu závislosti. Uveďme například koeficient ϕ,

$$\phi = \sqrt{\frac{\chi^2}{n}},$$

který je ve speciálním případě čtyřpolní tabulky roven

$$\phi = \frac{|n_{11}n_{22} - n_{12}n_{21}|}{\sqrt{n_{1.}n_{2.}n_{.1}n_{.2}}}.$$

Koeficient ϕ je vždy číslo kladné a roste se vzrůstající závislostí mezi X a Y. Pro měření intenzity závislosti však tato charakteristika není nejvhodnější, protože se její hodnoty nepohybují v pevném intervalu.

Dále se používá Pearsonův koeficient kontingence,

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}},$$

jehož hodnoty se vzrůstající závislostí rostou od nuly do jedné, přičemž hodnoty jedna nemůže tento koeficient nikdy dosáhnout.

Cramerovo V definované

$$V = \sqrt{\frac{\chi^2}{n(m-1)}}, \quad \text{kde} \quad m = min\{r, c\},$$

nabývá hodnot mezi 0 a 1, přitom vyšší hodnota koeficientu odpovídá těsnější závislosti znaků.

Ve všech třech případech je χ^2 hodnota testovací statistiky dané vztahem (2).

Dále se zaměříme na tzv. *poměr šancí* ve čtyřpolní tabulce. Ukazatel

$$b = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

se nazývá *poměr šancí* (odds-ratio, cross-product ratio). Protože $\frac{n_{ij}}{n}$ je odhadem pravděpodobnosti p_{ij}, je b vlastně odhadem *teoretického poměru šancí*

$$\beta = \frac{p_{11}p_{22}}{p_{12}p_{21}}.$$

Snadno lze dokázat, že teoretický poměr šancí β má následující vlastnosti

1. je invariantní vůči násobení řádků a sloupců tabulky pravděpodobností kladnými konstantami, pokud součet pravděpodobností zůstává jedna,
2. je invariantní vůči transpozici tabulky,
3. $\beta = 1$ právě tehdy, když jsou znaky nezávislé ($p_{ij} = p_{i.}p_{.j}, \ \forall i, j$).

Poměr šancí lze snadno interpretovat, což si ukážeme na názorném příkladě, který byl převzat z [9], str. 90.

Příklad Vyšetřuje se průběh onemocnění v závislosti na tom, zda byl pacient očkován, nebo ne. Data jsou v tabulce 3.

Šance očkovaných pacientů na lehký průběh nemoci je 20 : 4, bez očkování je šance na lehký průběh nemoci 8 : 22. Protože poměr šancí

$$b = \frac{20:4}{8:22} = \frac{55}{4}$$

	průběh nemoci	
pacient	lehký	těžký
očkován	20	4
neočkován	8	22

Tabulka 3: Údaje o pacientech.

je významně větší než 1, bude zřejmě výhodnější dát se očkovat.

Závislost znaků bude tedy tím větší, čím více se bude β vzdalovat od jedné. Přitom nesmíme zapomenout, že $0 \leq \beta \leq \infty$.

Nesymetrie hodnot β kolem bodu 1 vedla k zavedení *logaritmické interakce* d a *teoretické logaritmické interakce* δ, které se definují jako

$$d = \ln b, \quad \delta = \ln \beta.$$

Nyní si ukažme, jak souvisí pojem logaritmické interakce s pojmem interakce v logaritmicko-lineárním modelu.

2.5. Logaritmicko-lineární model pro čtyřpolní tabulky

Jedním z důležitých nástrojů statistického vyhodnocování je logaritmicko-lineární model, který se dá zobecnit i na tabulky vyšších dimenzí. Teorie této podkapitoly je převzata z [2].

V našem případě se omezíme na čtyřpolní kontingenční tabulku. Odpovídající tabulku pravděpodobností můžeme parametrizovat následujícím způsobem,

$$\begin{aligned}
\ln p_{11} &= \mu + \alpha_1 + \beta_1 + \lambda_{11}, \\
\ln p_{12} &= \mu + \alpha_1 + \beta_2 + \lambda_{12}, \\
\ln p_{21} &= \mu + \alpha_2 + \beta_1 + \lambda_{21}, \\
\ln p_{22} &= \mu + \alpha_2 + \beta_2 + \lambda_{22},
\end{aligned}$$

kde parametry $\alpha_i,\ \beta_j$ a λ_{ij} splňují tzv. *podmínky identifikovatelnosti*

$$\begin{aligned}\alpha_1 + \alpha_2 = 0, \quad \beta_1 + \beta_2 = 0,\\ \lambda_{11} + \lambda_{12} = 0, \quad \lambda_{21} + \lambda_{22} = 0,\\ \lambda_{11} + \lambda_{21} = 0, \quad \lambda_{12} + \lambda_{22} = 0.\end{aligned}$$

Podobně jako v analýze rozptylu mají parametry α_i, β_j význam efektů příslušného řádku či sloupce, parametry λ_{ij} mají význam interakcí mezi znaky.

Z podmínek identifikovatelnosti plyne

$$\begin{aligned}\alpha_2 = -\alpha_1, \quad \beta_2 = -\beta_1,\\ \lambda_{12} = -\lambda_{11}, \quad \lambda_{21} = -\lambda_{11}, \quad \lambda_{22} = \lambda_{11}.\end{aligned}$$

Proto můžeme psát

$$\begin{aligned}\ln p_{11} &= \mu + \alpha_1 + \beta_1 + \lambda_{11},\\ \ln p_{12} &= \mu + \alpha_1 - \beta_1 - \lambda_{11},\\ \ln p_{21} &= \mu - \alpha_1 + \beta_1 - \lambda_{11},\\ \ln p_{22} &= \mu - \alpha_1 - \beta_1 + \lambda_{11}.\end{aligned}$$

Po úpravě dostáváme logaritmickou interakci δ rovnu

$$\delta = \ln \frac{p_{11}p_{22}}{p_{12}p_{21}} = 4\lambda_{11}.$$

Proto interakce v logaritmicko-lineárním modelu λ_{11} je až na koeficient $1/4$ rovna logaritmické interakci δ.

V případě kontingenčních tabulek se však nepoužívají symboly $\mu,\ \alpha_i,\ \beta_j$ a λ_{ij} a také se místo modelu pro $\ln p_{ij}$ používá model pro $\ln m_{ij}$, kde $m_{ij} = np_{ij}$ jsou *očekávané četnosti*. Protože $\ln m_{ij} = \ln n + \ln p_{ij}$, přechod k m_{ij} ovlivní pouze parametr dosud označovaný jako μ.

V obecné dvojrozměrné kontingenční tabulce typu $r \times c$ používáme obvykle značení

$$\ln m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}, \quad i = 1, \ldots, r, \quad j = 1, \ldots, c,$$

za splnění podmínek

$$\sum_{i=1}^{r} u_{1(i)} = \sum_{j=1}^{c} u_{2(j)} = \sum_{i=1}^{r} u_{12(ij)} = \sum_{j=1}^{c} u_{12(ij)} = 0.$$

Přitom $u_{1(i)}, u_{2(j)}$ nazýváme *hlavní efekty*, $u_{12(ij)}$ *interakce prvního řádu.*

3. Kompoziční analýza pro kontingenční tabulky

Pokud v kontingenční tabulce nahradíme pozorované četnosti teoretickými pravděpodobnostmi, dostaneme matici pravděpodobností vyjadřující sdružené rozdělení pravděpodobností dvou diskrétních náhodných veličin.

	Y			
X	1	...	c	$\sum$
1	p_{11}	...	p_{1c}	$p_{1.}$
...	...	...	...	...
r	p_{r1}	...	p_{rc}	$p_{r.}$
$\sum$	$p_{.1}$	...	$p_{.c}$	1

Tabulka 4: Matice pravděpodobností $r \times c$.

Prvky této matice jsou kladná čísla, jejich součet je přirozeně 1, stejně jako součet marginálních řádkových a sloupcových pravděpodobností. Této vlastnosti je možno využít a interpretovat matici pravděpodobností jako $(r \cdot c)$-složkovou kompozici a pro další analýzu uplatnit poznatky o Aitchisonově geometrii na simplexu.

V dalším vycházíme z teorie a značení užitém v článku [4], ve většině případů se však omezíme pouze na případ matice rozměru 2×2 (čtyřpolní tabulka).

3.1. Značení a základní operace

Matici dvourozměrných diskrétních pravděpodobností budeme dále nazývat zkráceně DDP matice nebo jen DDP.

Uvažujme matici typu $r \times c$ s kladnými prvky. Tento typ matic značíme $\mathbf{x}, \mathbf{y}, \mathbf{z}$, jejich prvky pak např. x_{ij}. Kladné násobky těchto matic jsou považovány za ekvivalentní, třída ekvivalentních matic je reprezentována jedinou DDP maticí. Množinu DDP matic typu 2×2 představuje čtyřsložkový simplex, značíme S^4.

Vektor pravděpodobností v řádku (resp. ve sloupci) je podkompozice v S^2, značíme

$$\mathrm{row}_i[\mathbf{x}] = C(x_{i1}, x_{i2}), \quad \text{resp.} \quad \mathrm{col}_j[\mathbf{x}] = C(x_{1j}, x_{2j}), \ i, j = 1, 2.$$

Každý vektor vyjmutý z původní matice se stane automaticky řádkovým vektorem.

Vektor marginálních sloupcových pravděpodobností (angl. marginal row of $\mathbf{x}$) značíme

$$\mathrm{mrgr}[\mathbf{x}] = C(x_{11} + x_{21}, x_{12} + x_{22}),$$

je kompozicí v S^2. Podobně vektor marginálních řádkových pravděpodobností (angl. marginal column of $\mathbf{x}$) je

$$\mathrm{mrgc}[\mathbf{x}] = C(x_{11} + x_{12}, x_{21} + x_{22}).$$

Operátor, jehož užitím získáme z původní matice řádkový vektor, tj. mrgc, mrgr, row_i, col_j, je následován argumentem v hranatých závorkách. Když bude argument uváděn v závorkách kulatých, znamená to, že výsledkem operace je DDP matice v S^4 (viz např. rovnice (4)).

Standardní operace na simplexu, perturbaci a mocninnou transformaci, použijeme nyní na DDP matice. Prvky perturbace

$$\mathbf{z} = \mathbf{x} \oplus \mathbf{y}, \quad \text{jsou} \quad z_{ij} = \frac{k x_{ij} y_{ij}}{\sum_{r,c} x_{rc} y_{rc}}, \ k > 0.$$

Mocninná transformace reálným číslem $a \in \mathbb{R}$,

$$\mathbf{z} = a \odot \mathbf{x}, \quad \text{má prvky} \quad z_{ij} = \frac{k x_{ij}^a}{\sum_{r,c} x_{rc}^a}.$$

Jak bylo dokázáno, simplex spolu s těmito operacemi tvoří vektorový prostor. Navíc po přidání definice skalárního součinu a odpovídající normy a vzdálenosti dostáváme strukturu Hilbertova prostoru.

V kapitole 1.1 byly zadefinovány pojmy Aitchisonův skalární součin, norma a vzdálenost. Pro DDP matice má tento skalární součin odpovídající tvar

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \sum_{i,j} (\ln x_{ij} \cdot \ln y_{ij}) - \frac{1}{rc} \left(\sum_{i,j} \ln x_{ij} \right) \cdot \left(\sum_{i,j} \ln y_{ij} \right).$$

Pro připomenutí uveďme též pojmy normy a vzdálenosti,

$$||\mathbf{x}||_a = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_a}, \quad d_a(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} \ominus \mathbf{y}||_a.$$

3.2. Řádkové a sloupcové podprostory

Euklidovská struktura prostoru S^{rc} umožňuje definovat podprostory a ortogonální projekce DDP matice. Pak i-tý řádek DDP matice $\mathbf{x} \in S^4$ je podkompozice $\mathrm{row}_i[\mathbf{x}] \in S^2$. Nicméně tato podkompozice může být reprezentována ortogonální projekcí $\mathbf{x}$ do podprostoru S^4 určeného $\mathrm{row}_i[\mathbf{x}]$. Takový podprostor nazveme řádkový podprostor a označíme $S^4(\mathrm{row}_i)$, jeho dimenze je rovna jedné.

Dříve než vytvoříme daný podprostor a projekci, musíme uvažovat ortonormální bázi v S^2 (protože $\mathrm{row}_i[\mathbf{x}] \in S^2$), kterou tvoří kompozice

$$\mathbf{e} = C \exp\left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right), \quad \mathrm{clr}(\mathbf{e}) = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right).$$

Ortonormální báze v $S^4(\mathrm{row}_i)$ mají tvar

$$\mathbf{E}_1 = C \exp\begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 0 & 0 \end{pmatrix} \text{ v } S^4(\mathrm{row}_1)$$

a

$$\mathbf{E}_2 = C \exp\begin{pmatrix} 0 & 0 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \text{ v } S^4(\mathrm{row}_2),$$

kde operátor exp použijeme postupně na všechny složky kompozice.

Matice $\mathbf{E}_1$, resp. $\mathbf{E}_2$ jsou tedy báze podprostorů určených výběrem prvního, resp. druhého řádku matice $\mathbf{x}$.

Ortogonální projekci $\mathbf{x}$ na podprostor $S^4(\mathrm{row}_i)$ značíme $\mathrm{row}_i(\mathbf{x})$.

Uvědomme si rozdíl mezi značením $\mathrm{row}_i[\mathbf{x}] \in S^2$ pro řádkový vektor a $\mathrm{row}_i(\mathbf{x}) \in S^4$ pro matici DDP. Platí

$$\mathrm{row}_i(\mathbf{x}) = \langle \mathbf{x}, \mathbf{E}_i \rangle_a \odot \mathbf{E}_i, \ i = 1, 2. \tag{4}$$

Provedeme rozvoj pro vyjádření matice projekce,

$$\mathbf{x} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}, \quad \mathbf{E}_1 = C \begin{pmatrix} e^{\frac{1}{\sqrt{2}}} & e^{-\frac{1}{\sqrt{2}}} \\ 1 & 1 \end{pmatrix}, \quad g(\mathbf{E}_1) = 1.$$

Obdržíme

$$\langle \mathbf{x}, \mathbf{E}_1 \rangle_a = \ln \frac{x_{11}}{g(\mathbf{x})} \ln e^{\frac{1}{\sqrt{2}}} + \ln \frac{x_{12}}{g(\mathbf{x})} \ln e^{-\frac{1}{\sqrt{2}}} + \ln \frac{x_{21}}{g(\mathbf{x})} \ln 1 + \ln \frac{x_{22}}{g(\mathbf{x})} \ln 1 =$$

$$= \frac{1}{\sqrt{2}} \ln \frac{x_{11}}{g(\mathbf{x})} - \frac{1}{\sqrt{2}} \ln \frac{x_{12}}{g(\mathbf{x})} = \frac{1}{\sqrt{2}} \ln \frac{\frac{x_{11}}{g(\mathbf{x})}}{\frac{x_{12}}{g(\mathbf{x})}} = \frac{1}{\sqrt{2}} \ln \frac{x_{11}}{x_{12}},$$

$$\langle \mathbf{x}, \mathbf{E}_1 \rangle_a \odot \mathbf{E}_1 = C \begin{pmatrix} e^{\frac{1}{\sqrt{2}}\left(\frac{1}{\sqrt{2}} \ln \frac{x_{11}}{x_{12}}\right)} & e^{-\frac{1}{\sqrt{2}}\left(\frac{1}{\sqrt{2}} \ln \frac{x_{11}}{x_{12}}\right)} \\ 1 & 1 \end{pmatrix} = C \begin{pmatrix} \sqrt{\frac{x_{11}}{x_{12}}} & \sqrt{\frac{x_{12}}{x_{11}}} \\ 1 & 1 \end{pmatrix} =$$

$$= C \begin{pmatrix} x_{11} & x_{12} \\ \sqrt{x_{11}x_{12}} & \sqrt{x_{11}x_{12}} \end{pmatrix}.$$

Označíme-li $g(\mathrm{row}_i[\mathbf{x}])$ jako průměr i-tého řádku, můžeme pak psát

$$\mathrm{row}_1(\mathbf{x}) = C \begin{pmatrix} x_{11} & x_{12} \\ g(\mathrm{row}_1[\mathbf{x}]) & g(\mathrm{row}_1[\mathbf{x}]) \end{pmatrix}.$$

Analogicky dostaneme též

$$\mathrm{row}_2(\mathbf{x}) = C \begin{pmatrix} g(\mathrm{row}_2[\mathbf{x}]) & g(\mathrm{row}_2[\mathbf{x}]) \\ x_{21} & x_{22} \end{pmatrix}.$$

Přitom $\mathrm{row}_i(\mathbf{x})$ má k původní matici $\mathbf{x}$ nejblíže vzhledem k Aitchisonově vzdálenosti ze všech DDP matic shodujícími se v i-tém řádku.

Dále se dá dokázat ortogonalita $\mathrm{row}_1(\mathbf{x})$ a $\mathrm{row}_2(\mathbf{x})$. Užitím Aitchisonova skalárního součinu dostaneme

$$\langle \mathrm{row}_1(\mathbf{x}),\ \mathrm{row}_2(\mathbf{x}) \rangle_a = \ln \frac{x_{11}}{\sqrt{x_{11}x_{12}}} \ln \frac{\sqrt{x_{21}x_{22}}}{\sqrt{x_{21}x_{22}}} + \ln \frac{x_{12}}{\sqrt{x_{11}x_{12}}} \ln \frac{\sqrt{x_{21}x_{22}}}{\sqrt{x_{21}x_{22}}} +$$

$$+ \ln \frac{\sqrt{x_{11}x_{12}}}{\sqrt{x_{11}x_{12}}} \ln \frac{x_{21}}{\sqrt{x_{21}x_{22}}} + \ln \frac{\sqrt{x_{11}x_{12}}}{\sqrt{x_{11}x_{12}}} \ln \frac{x_{21}}{\sqrt{x_{21}x_{22}}} = 0.$$

Analogický postup pro sloupcové vektory vede k vytvoření dvou vzájemně ortogonálních podprostorů dimenze 1 určených prvním a druhým sloupcem, které jsou obdobně nazvány sloupcové podprostory a značeny $S^4(\mathrm{col}_j)$,

$$\mathrm{col}_1(\mathbf{x}) = C \begin{pmatrix} x_{11} & g(\mathrm{col}_1[\mathbf{x}]) \\ x_{21} & g(\mathrm{col}_1[\mathbf{x}]) \end{pmatrix},$$

$$\mathrm{col}_2(\mathbf{x}) = C \begin{pmatrix} g(\mathrm{col}_2[\mathbf{x}]) & x_{12} \\ g(\mathrm{col}_2[\mathbf{x}]) & x_{22} \end{pmatrix}.$$

Přitom řádkové a sloupcové podprostory nejsou navzájem ortogonální, jak dokazuje následující příklad.

Příklad Uvažujme DDP matici rozměru 2×2,

$$\mathbf{x} = \begin{pmatrix} 0.1 & 0.4 \\ 0.2 & 0.3 \end{pmatrix}, \ \mathrm{mrgr}[\mathbf{x}] = (0.3, 0.7), \ \mathrm{mrgc}[\mathbf{x}] = (0.5, 0.5).$$

$||\mathbf{x}||_a^2 = 1.084$. Vyjádříme projekce $\mathbf{x}$ na první řádek a první sloupec

$$\mathrm{row}_1(\mathbf{x}) = C \begin{pmatrix} 0.1 & 0.4 \\ 0.2 & 0.2 \end{pmatrix}, \ ||\mathrm{row}_1(\mathbf{x})||_a^2 = 0.961,$$

$$\mathrm{col}_1(\mathbf{x}) = C \begin{pmatrix} 0.1 & 0.141 \\ 0.2 & 0.141 \end{pmatrix}, \ ||\mathrm{col}_1(\mathbf{x})||_a^2 = 0.24.$$

Skalární součin obou projekcí je roven $\langle \mathrm{row}_1(\mathbf{x}), \mathrm{col}_1(\mathbf{x}) \rangle_a = 0.238$, je tedy nenulový a projekce na řádek a sloupec nejsou ortogonální. Svírají úhel přibližně $60°$.

3.3. Geometrické marginální řádkové a sloupcové podprostory

Dále se budeme zabývat otázkou nalezení ortogonálního doplňku ke všem řádkovým podprostorům a ke všem sloupcovým podprostorům.

Protože dimenze prostoru S^4 je rovna 3 a dimenze řádkových (resp. sloupcových) podprostorů $S^4(\mathrm{row}_i(\mathbf{x}))$, resp. $S^4(\mathrm{col}_j(\mathbf{x}))$ jsou rovny jedné, je dimenze ortogonálního doplňku k obou řádkovým (resp. sloupcovým) podprostorům také rovna jedné,

$$\dim S^4 - 2 \dim S^4(\mathrm{row}_i(\mathbf{x})) = \dim S^4 - 2 \dim S^4(\mathrm{col}_j(\mathbf{x})) = 3 - 2 = 1.$$

Prostory ortogonálních doplňků značíme $S^4(\mathrm{row}^\perp)$, resp. $S^4(\mathrm{col}^\perp)$. Báze $S^4(\mathrm{row}^\perp)$ je tvaru

$$\mathbf{F} = C \exp \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix},$$

přičemž vektor $\left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)$ tvoří ortonormální bázi v $\mathbb{R}^1$.

Ortogonalitu ověříme provedením skalárního součinu $\langle \mathbf{F}, \mathbf{E}_i \rangle_a = 0,\ i = 1, 2.$

$$\mathbf{E}_1 = C \begin{pmatrix} e^{\frac{1}{\sqrt{2}}} & e^{-\frac{1}{\sqrt{2}}} \\ 1 & 1 \end{pmatrix}, \quad g(\mathbf{E}_1) = 1, \quad \mathbf{F} = C \begin{pmatrix} e^{\frac{1}{\sqrt{2}}} & e^{\frac{1}{\sqrt{2}}} \\ e^{-\frac{1}{\sqrt{2}}} & e^{-\frac{1}{\sqrt{2}}} \end{pmatrix}, \quad g(\mathbf{F}) = 1.$$

$$\langle \mathbf{F}, \mathbf{E}_1 \rangle_a = \ln e^{\frac{1}{\sqrt{2}}} \ln e^{\frac{1}{\sqrt{2}}} + \ln e^{\frac{1}{\sqrt{2}}} \ln e^{-\frac{1}{\sqrt{2}}} + \ln e^{-\frac{1}{\sqrt{2}}} \ln 1 + \ln e^{-\frac{1}{\sqrt{2}}} \ln 1 = \frac{1}{2} - \frac{1}{2} = 0.$$

Pro $\langle \mathbf{F}, \mathbf{E}_2 \rangle_a$ je výpočet analogický.

Vezmeme-li v úvahu, že norma matice $\mathbf{F}$ není jednotková, neboť $||\mathbf{F}||_a^2 = 2$, (není tedy ortonormální bází prostoru $S^4(\mathrm{row}^\perp)$), můžeme projekci $\mathbf{x}$ na prostor $S^4(\mathrm{row}^\perp)$ vyjádřit rozvojem

$$\mathrm{row}^\perp(\mathbf{x}) = \frac{1}{2} \odot \left(\langle \mathbf{x}, \mathbf{F} \rangle_a \odot \mathbf{F} \right).$$

Výpočet probíhá následovně. Ze zadaných prvků

$$\mathbf{x} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}, \quad \mathbf{F} = C \begin{pmatrix} e^{\frac{1}{\sqrt{2}}} & e^{\frac{1}{\sqrt{2}}} \\ e^{-\frac{1}{\sqrt{2}}} & e^{-\frac{1}{\sqrt{2}}} \end{pmatrix}$$

postupně obdržíme

$$\begin{aligned} \langle \mathbf{x}, \mathbf{F} \rangle_a &= \ln \frac{x_{11}}{g(\mathbf{x})} \ln e^{\frac{1}{\sqrt{2}}} + \ln \frac{x_{12}}{g(\mathbf{x})} \ln e^{\frac{1}{\sqrt{2}}} + \ln \frac{x_{21}}{g(\mathbf{x})} \ln e^{-\frac{1}{\sqrt{2}}} + \ln \frac{x_{22}}{g(\mathbf{x})} \ln e^{-\frac{1}{\sqrt{2}}} = \\ &= \frac{1}{\sqrt{2}} \ln \left(\frac{x_{11}}{g(\mathbf{x})} \cdot \frac{x_{12}}{g(\mathbf{x})} \cdot \frac{g(\mathbf{x})}{x_{21}} \cdot \frac{g(\mathbf{x})}{x_{22}} \right) = \frac{1}{\sqrt{2}} \ln \frac{x_{11} x_{12}}{x_{21} x_{22}}, \end{aligned}$$

$$\langle \mathbf{x}, \mathbf{F} \rangle_a \odot \mathbf{F} = C \begin{pmatrix} e^{\frac{1}{\sqrt{2}} \left(\frac{1}{\sqrt{2}} \ln \left(\frac{x_{11} x_{12}}{x_{21} x_{22}} \right) \right)} & e^{\frac{1}{\sqrt{2}} \left(\frac{1}{\sqrt{2}} \ln \left(\frac{x_{11} x_{12}}{x_{21} x_{22}} \right) \right)} \\ e^{-\frac{1}{\sqrt{2}} \left(\frac{1}{\sqrt{2}} \ln \left(\frac{x_{11} x_{12}}{x_{21} x_{22}} \right) \right)} & e^{-\frac{1}{\sqrt{2}} \left(\frac{1}{\sqrt{2}} \ln \left(\frac{x_{11} x_{12}}{x_{21} x_{22}} \right) \right)} \end{pmatrix} = C \begin{pmatrix} \sqrt{\frac{x_{11} x_{12}}{x_{21} x_{22}}} & \sqrt{\frac{x_{11} x_{12}}{x_{21} x_{22}}} \\ \sqrt{\frac{x_{21} x_{22}}{x_{11} x_{12}}} & \sqrt{\frac{x_{21} x_{22}}{x_{11} x_{12}}} \end{pmatrix},$$

$$\frac{1}{2} \odot (\langle \mathbf{x}, \mathbf{F} \rangle_a \odot \mathbf{F}) = C \begin{pmatrix} \sqrt[4]{\frac{x_{11}x_{12}}{x_{21}x_{22}}} & \sqrt[4]{\frac{x_{11}x_{12}}{x_{21}x_{22}}} \\ \sqrt[4]{\frac{x_{21}x_{22}}{x_{11}x_{12}}} & \sqrt[4]{\frac{x_{21}x_{22}}{x_{11}x_{12}}} \end{pmatrix} = C \begin{pmatrix} \sqrt{x_{11}x_{12}} & \sqrt{x_{11}x_{12}} \\ \sqrt{x_{21}x_{22}} & \sqrt{x_{21}x_{22}} \end{pmatrix},$$

a tedy

$$\mathrm{row}^{\perp}(\mathbf{x}) = C \begin{pmatrix} g(\mathrm{row}_1[\mathbf{x}]) & g(\mathrm{row}_1[\mathbf{x}]) \\ g(\mathrm{row}_2[\mathbf{x}]) & g(\mathrm{row}_2[\mathbf{x}]) \end{pmatrix},$$

kde prvky původní matice $\mathbf{x}$ jsou nahrazeny geometrickými průměry příslušného řádku. Kompozici složenou z geometrických průměrů řádků nazvěme geometrický marginální sloupec (angl. geometric marginal column) a označme

$$\mathrm{gmrgc}[\mathbf{x}] = C(g(\mathrm{row}_1[\mathbf{x}]), g(\mathrm{row}_2[\mathbf{x}])).$$

Slovo marginální se v označení daného vektoru objevuje z toho důvodu, že pokud zaměníme geometrické průměry za aritmetické, dostaneme tradiční vektor marginálních pravděpodobností $\mathrm{mrgc}[\mathbf{x}]$ (vektor řádkových součtů). Obdobně obdržíme

$$\mathrm{col}^{\perp}(\mathbf{x}) = C \begin{pmatrix} g(\mathrm{col}_1[\mathbf{x}]) & g(\mathrm{col}_2[\mathbf{x}]) \\ g(\mathrm{col}_1[\mathbf{x}]) & g(\mathrm{col}_2[\mathbf{x}]) \end{pmatrix}$$

a

$$\mathrm{gmrgr}[\mathbf{x}] = C(g(\mathrm{col}_1[\mathbf{x}]), g(\mathrm{col}_2[\mathbf{x}])),$$

nazýváme geometrický marginální řádek (angl. geometric marginal row). Projekce $\mathrm{row}^{\perp}(\mathbf{x})$, $\mathrm{col}^{\perp}(\mathbf{x})$ nazvěme geometrické marginální sloupcové (řádkové) DDP matice.

Ortogonalita projekcí $\mathrm{row}_i(\mathbf{x})$ a $\mathrm{row}^{\perp}(\mathbf{x})$ umožňuje ortogonální dekompozici $\mathbf{x}$,

$$\begin{aligned} \mathbf{x} &= \mathrm{row}^{\perp}(\mathbf{x}) \oplus \big(\mathrm{row}_1(\mathbf{x}) \oplus \mathrm{row}_2(\mathbf{x})\big) = \\ &= \Big(\langle \mathbf{x}, \mathbf{F} \rangle_a \odot \mathbf{F}\Big) \oplus \Big(\big(\langle \mathbf{x}, \mathbf{E}_1 \rangle_a \odot \mathbf{E}_1\big) \oplus \big(\langle \mathbf{x}, \mathbf{E}_2 \rangle_a \odot \mathbf{E}_2\big)\Big). \end{aligned}$$

Následujícím postupem provedeme ověření uvedené dekompozice:

$$\mathrm{row}_1(\mathbf{x}) = C \begin{pmatrix} x_{11} & x_{12} \\ \sqrt{x_{11}x_{12}} & \sqrt{x_{11}x_{12}} \end{pmatrix}, \quad \mathrm{row}_2(\mathbf{x}) = C \begin{pmatrix} \sqrt{x_{21}x_{22}} & \sqrt{x_{21}x_{22}} \\ x_{21} & x_{22} \end{pmatrix},$$

$$\mathrm{row}_1(\mathbf{x}) \oplus \mathrm{row}_2(\mathbf{x}) = C \begin{pmatrix} x_{11}\sqrt{x_{21}x_{22}} & x_{12}\sqrt{x_{21}x_{22}} \\ x_{21}\sqrt{x_{11}x_{12}} & x_{22}\sqrt{x_{11}x_{12}} \end{pmatrix},$$

$$\mathrm{row}^{\perp}(\mathbf{x}) = C \begin{pmatrix} \sqrt{x_{11}x_{12}} & \sqrt{x_{11}x_{12}} \\ \sqrt{x_{21}x_{22}} & \sqrt{x_{21}x_{22}} \end{pmatrix},$$

$$(\mathrm{row}_1(\mathbf{x}) \oplus \mathrm{row}_2(\mathbf{x})) \oplus \mathrm{row}^{\perp}(\mathbf{x}) = C \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} = \mathbf{x}.$$

Pro sloupcové projekce dostáváme dekompozici

$$\mathbf{x} = \mathrm{col}^{\perp}(\mathbf{x}) \oplus \big(\mathrm{col}_1(\mathbf{x}) \oplus \mathrm{col}_2(\mathbf{x})\big).$$

Podprostory $S^4(\mathrm{row}^{\perp})$ a $S^4(\mathrm{col}^{\perp})$ jsou vzájemně ortogonální. Abychom toto tvrzení dokázali, vezměme clr transformace dvou matic DDP: matice $\mathbf{x}$ s identickými řádky a matice $\mathbf{y}$ s identickými sloupci,

$$\mathrm{clr}(\mathbf{x}) = \begin{pmatrix} \xi_1 & \xi_2 \\ \xi_1 & \xi_2 \end{pmatrix}, \quad \mathrm{clr}(\mathbf{y}) = \begin{pmatrix} \eta_1 & \eta_1 \\ \eta_2 & \eta_2 \end{pmatrix}.$$

Součet složek transformované kompozice je roven nule, proto je i skalární součin obou matic nulový,

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle \mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{y}) \rangle = \xi_1\eta_1 + \xi_2\eta_1 + \xi_1\eta_2 + \xi_2\eta_2 = 0.$$

Důležitou vlastností geometrických marginálních pravděpodobností je linearita vzhledem k perturbaci. Nechť $\mathbf{x}$ a $\mathbf{y}$ jsou DDP v S^{rc}. Pak platí

$$\mathrm{gmrgr}[\mathbf{x} \oplus \mathbf{y}] = \mathrm{gmrgr}[\mathbf{x}] \oplus \mathrm{gmrgr}[\mathbf{y}], \quad \mathrm{gmrgc}[\mathbf{x} \oplus \mathbf{y}] = \mathrm{gmrgc}[\mathbf{x}] \oplus \mathrm{gmrgc}[\mathbf{y}].$$

Tato vlastnost přitom zřejmě neplatí pro klasické marginální pravděpodobnosti $(\mathrm{mrgr}, \mathrm{mrgc})$.

3.4. Podprostor nezávislých matic

Při rozboru dvourozměrných kontingenčních tabulek je důležité nejen identifikovat vztahy mezi řádky a sloupci, ale také odhalit tu část, která žádné vztahy nezahrnuje, nazvěme ji nezávislou částí.

Z teorie pravděpodobnosti víme, že prvky matice dvojrozměrných diskrétních pravděpodobností nezávislých znaků vzniknou součinem řádkových a sloupcových marginálních pravděpodobností. Situace pro nezávislou matici pak vypadá takto,

$$p_{ij} = p_{i.}p_{.j}, \qquad \begin{array}{|cc|c|} \hline & & \\ \hline p_{11} & p_{12} & p_{1.} \\ p_{21} & p_{22} & p_{2.} \\ \hline p_{.1} & p_{.2} & \\ \hline \end{array}$$

Na simplexu je takováto matice výsledkem perturbace dvou DDP matic, jejichž řádky resp. sloupce jsou identické,

$$\begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} p_{1.}p_{.1} & p_{1.}p_{.2} \\ p_{2.}p_{.1} & p_{2.}p_{.2} \end{pmatrix} = \begin{pmatrix} p_{1.} & p_{1.} \\ p_{2.} & p_{2.} \end{pmatrix} \oplus \begin{pmatrix} p_{.1} & p_{.2} \\ p_{.1} & p_{.2} \end{pmatrix}.$$

Množinu všech matic vzniklých touto perturbací (s geometrickými marginálními pravděpodobnostmi) nazvěme podprostor nezávislých matic a označme S^4_{ind}, symbolicky můžeme psát $S^4_{ind} = S^4(\mathrm{row}^\perp) \oplus S^4(\mathrm{col}^\perp)$.

Protože dimenze podprostorů $S^4(\mathrm{row}^\perp)$ a $S^4(\mathrm{col}^\perp)$ je rovna jedné a jsou vzájemně ortogonální, jejich perturbací vznikne podprostor dimenze 2.

Ortogonální projekce DDP matice $\mathbf{x}$ na prostor S^4_{ind} pak představuje tu část matice $\mathbf{x}$, která obsahuje pravděpodobnosti dvou nezávislých náhodných veličin. Zbývající část pak obsahuje interakce. Nezávislou část spočítáme snadno jako perturbaci geometrické marginální sloupcové a řádkové DDP matice,

$$\mathbf{x}_{ind} = \mathrm{row}^\perp(\mathbf{x}) \oplus \mathrm{col}^\perp(\mathbf{x}).$$

Perturbaci rozepíšeme, abychom zjistili výpočetní tvar nezávislé matice:

$$\mathrm{row}^\perp(\mathbf{x}) = C \begin{pmatrix} \sqrt{x_{11}x_{12}} & \sqrt{x_{11}x_{12}} \\ \sqrt{x_{21}x_{22}} & \sqrt{x_{21}x_{22}} \end{pmatrix}, \quad \mathrm{col}^\perp(\mathbf{x}) = C \begin{pmatrix} \sqrt{x_{11}x_{21}} & \sqrt{x_{12}x_{22}} \\ \sqrt{x_{11}x_{21}} & \sqrt{x_{12}x_{22}} \end{pmatrix},$$

$$\mathbf{x}_{ind} = \mathrm{row}^\perp(\mathbf{x}) \oplus \mathrm{col}^\perp(\mathbf{x}) = C \begin{pmatrix} x_{11}\sqrt{x_{12}x_{21}} & x_{12}\sqrt{x_{11}x_{22}} \\ x_{21}\sqrt{x_{11}x_{22}} & x_{22}\sqrt{x_{12}x_{21}} \end{pmatrix}.$$

Kdybychom pro výpočet použili aritmetické marginální pravděpodobnosti, výsledek by byl odlišný, $\mathbf{x}_{ind} \neq \mathrm{mrgc}(\mathbf{x}) \oplus \mathrm{mrgr}(\mathbf{x})$, protože

$$\mathrm{row}^\perp(\mathbf{x}) \neq \mathrm{mrgc}(\mathbf{x}) = (\mathrm{mrgc}[\mathbf{x}]', \mathrm{mrgc}[\mathbf{x}]'),$$

$$\mathrm{col}^{\perp}(\mathbf{x}) \neq \mathrm{mrgr}(\mathbf{x}) = (\mathrm{mrgr}[\mathbf{x}]', \mathrm{mrgr}[\mathbf{x}]')'.$$

Rovnost $\mathbf{x}_{ind} = \mathrm{mrgc}(\mathbf{x}) \oplus \mathrm{mrgr}(\mathbf{x})$ nastává v případě, že matice $\mathbf{x}$ je tvořena pouze nezávislou částí, je to tedy matice pravděpodobností dvou nezávislých veličin, žádné řádkové ani sloupcové efekty se neprojeví. Pak platí

$$\begin{aligned}\mathrm{mrgc}(\mathbf{x}) = \mathrm{mrgc}(\mathbf{x}_{ind}) &= C \begin{pmatrix} x_{11}\sqrt{x_{12}x_{21}} + x_{12}\sqrt{x_{11}x_{22}} & x_{11}\sqrt{x_{12}x_{21}} + x_{12}\sqrt{x_{11}x_{22}} \\ x_{21}\sqrt{x_{11}x_{22}} + x_{22}\sqrt{x_{12}x_{21}} & x_{21}\sqrt{x_{11}x_{22}} + x_{22}\sqrt{x_{12}x_{21}} \end{pmatrix} = \\ &= C \begin{pmatrix} \sqrt{x_{11}x_{12}}(\sqrt{x_{11}x_{21}} + \sqrt{x_{12}x_{22}}) & \sqrt{x_{11}x_{12}}(\sqrt{x_{11}x_{21}} + \sqrt{x_{12}x_{22}}) \\ \sqrt{x_{21}x_{22}}(\sqrt{x_{21}x_{11}} + \sqrt{x_{22}x_{12}}) & \sqrt{x_{21}x_{22}}(\sqrt{x_{21}x_{11}} + \sqrt{x_{22}x_{12}}) \end{pmatrix} = \\ &= C \begin{pmatrix} g(\mathrm{row}_1[\mathbf{x}]) & g(\mathrm{row}_1[\mathbf{x}]) \\ g(\mathrm{row}_2[\mathbf{x}]) & g(\mathrm{row}_2[\mathbf{x}]) \end{pmatrix}.\end{aligned}$$

Analogicky by výpočet proběhl i pro marginální řádkovou matici $\mathrm{mrgr}(\mathbf{x})$ a v případě $\mathbf{x} = \mathbf{x}_{ind}$ tedy platí

$$\mathrm{row}^{\perp}(\mathbf{x}) = \mathrm{mrgc}(\mathbf{x}), \quad \mathrm{col}^{\perp}(\mathbf{x}) = \mathrm{mrgr}(\mathbf{x}).$$

V této situaci je projekce matice $\mathbf{x}$ na řádek a na sloupec nulová a projekce na nezávislý podprostor odpovídá původní matici, která je rovna součinu marginálních pravděpodobností.

3.5. Analýza dvourozměrných diskrétních pravděpodobností

Podobně jako v teorii logaritmicko-lineárních modelů je naším cílem rozložit tabulku DDP do tvaru perturbace tabulek, které je možno interpretovat. Jednou z typických dílčích tabulek je nezávislá tabulka.

Prvním krokem bude tedy ortogonální rozklad DDP v S^4 ve tvaru

$$\mathbf{x} = \mathbf{x}_{ind} \oplus \mathbf{x}_{int},$$

kde

$$\mathbf{x}_{ind} = \mathrm{col}^{\perp}(\mathbf{x}) \oplus \mathrm{row}^{\perp}(\mathbf{x}).$$

Matice $\mathbf{x}_{int}$ představuje tu část $\mathbf{x}$, která není vysvětlena pomocí $\mathbf{x}_{ind}$, obsahuje tedy informaci o vazbách mezi řádky a sloupci. Nazvěme ji proto interakční

maticí,

$$\mathbf{x}_{int} = \mathbf{x} \ominus \mathbf{x}_{ind} = C \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} \ominus C \begin{pmatrix} x_{11}\sqrt{x_{12}x_{21}} & x_{12}\sqrt{x_{11}x_{22}} \\ x_{21}\sqrt{x_{11}x_{22}} & x_{22}\sqrt{x_{12}x_{21}} \end{pmatrix} =$$

$$= C \begin{pmatrix} \frac{1}{\sqrt{x_{12}x_{21}}} & \frac{1}{\sqrt{x_{11}x_{22}}} \\ \frac{1}{\sqrt{x_{11}x_{22}}} & \frac{1}{\sqrt{x_{12}x_{21}}} \end{pmatrix}.$$

Dále nás bude zajímat vzdálenost $\mathbf{x}$ od projekce $\mathbf{x}_{ind}$. Vyhovující mírou vzdálenosti se ukazuje druhá mocnina Aitchisonovy vzdálenosti,

$$\Delta^2(\mathbf{x}) = ||\mathbf{x} \ominus \mathbf{x}_{ind}||_a^2 = ||\mathbf{x}_{int}||_a^2 = ||\mathbf{x}||_a^2 - ||\mathbf{x}_{ind}||_a^2.$$

Protože hodnota normy závisí na rozměru $\mathbf{x}$, bude vhodnější použít k výpočtu relativní míru závislosti,

$$R_\Delta^2(\mathbf{x}) = \frac{\Delta^2(\mathbf{x})}{||\mathbf{x}||_a^2}, \quad 0 \leq R_\Delta^2 \leq 1.$$

Pro $R_\Delta^2(\mathbf{x}) = 1$ se $\mathbf{x}$ skládá pouze z interakcí (je interakční maticí), zatímco $R_\Delta^2(\mathbf{x}) = 0$ znamená, že je $\mathbf{x}$ nezávislá matice. Je zřejmé, že $R_\Delta^2(\mathbf{x}_{int}) = 1$ a $R_\Delta^2(\mathbf{x}_{ind}) = 0$. R_Δ^2 tedy vyčísluje, do jaké míry se matice $\mathbf{x}$ liší od nezávislé matice, neboli jak důležitou úlohu hrají efekty a interakce.

Další oblastí zájmu jsou příspěvky jednotlivých řádků a sloupců. V obecném případě však tyto nejsou ortogonální. V kapitole 3.3 jsme uvedli dvě ortogonální dekompozice: pomocí projekcí na řádky (vyjadřuje příspěvky řádků) a geometrické marginální sloupcové matice,

$$\mathbf{x} = \mathrm{row}^{\perp}(\mathbf{x}) \oplus \big(\mathrm{row}_1(\mathbf{x}) \oplus \mathrm{row}_2(\mathbf{x})\big) \tag{5}$$

a analogicky pomocí projekcí na sloupce a geometrické marginální řádkové matice,

$$\mathbf{x} = \mathrm{col}^{\perp}(\mathbf{x}) \oplus \big(\mathrm{col}_1(\mathbf{x}) \oplus \mathrm{col}_2(\mathbf{x})\big). \tag{6}$$

Nezávislý podprostor není ortogonální k žádné z částí těchto dekompozic.

Užitím dekompozic (5) a (6) pro $\mathbf{x}_{int}$ získáme následující rozklad

$$\begin{aligned}\mathbf{x} &= \mathbf{x}_{ind} \oplus \Big(\mathrm{row}_1(\mathbf{x}_{int}) \oplus \mathrm{row}_2(\mathbf{x}_{int})\Big) \\ &= \mathbf{x}_{ind} \oplus \Big(\mathrm{col}_1(\mathbf{x}_{int}) \oplus \mathrm{col}_2(\mathbf{x}_{int})\Big).\end{aligned}$$

Příspěvek i-tého řádku k druhé mocnině interakční matice $\mathbf{x}_{int}$ je takto roven $||\mathrm{row}_i(\mathbf{x}_{int})||_a^2$ a analogicky $||\mathrm{col}_j(\mathbf{x}_{int})||_a^2$ pro příspěvek j-tého sloupce.

Zbývá ještě vyšetřit interakce mezi řádky a sloupci. Ortogonální dekompozice pro tuto situaci není možná. Interakce mezi řádky a sloupci matice $\mathbf{y} = \mathbf{x}_{int}$ lze definovat více ekvivalentními způsoby.

Prvním z možných způsobů je zavedení pojmu *křížový kontrast* (angl. cross-contrast) pro bilanci mezi prvkem na pozici (i, j) a všemi ostatními prvky i-tého řádku a j-tého sloupce, označme $I_{cross}(i, j)$. Obecně pro matici $\mathbf{y}$ typu $r \times c$ platí

$$I_{cross}(i,j) = \sqrt{\frac{r+c-2}{r+c-1}} \ln \frac{y_{ij}}{\left(\prod_{i\neq k=1}^{r} y_{kj} \prod_{j\neq l=1}^{c} y_{il}\right)^{1/(r+c-2)}},$$

speciálně pro čtyřpolní tabulku

$$I_{cross}(1,1) = \sqrt{\frac{2}{3}} \ln \frac{y_{11}}{\sqrt{y_{12}y_{21}}},$$

$$I_{cross}(1,2) = \sqrt{\frac{2}{3}} \ln \frac{y_{12}}{\sqrt{y_{11}y_{22}}},$$

$$I_{cross}(2,1) = \sqrt{\frac{2}{3}} \ln \frac{y_{21}}{\sqrt{y_{11}y_{22}}},$$

$$I_{cross}(2,2) = \sqrt{\frac{2}{3}} \ln \frac{y_{22}}{\sqrt{y_{12}y_{21}}}.$$

Tyto bilance nejsou ortogonální, ale platí, že součet bilancí je nulový a součet druhých mocnin bilancí je úměrný druhé mocnině normy $\mathbf{x}_{int}$,

$$\sum_{i=1}^{r}\sum_{j=1}^{c} I_{cross}(i,j) = 0,$$

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\left(I_{cross}(i,j)\right)^2 = \frac{(r+c)^2}{(r+c-1)(r+c-2)} \cdot ||\mathbf{x}_{int}||_a^2. \tag{7}$$

Druhým způsobem, jak definovat interakce řádků a sloupců je zavedení pojmu *složková interakce* (angl. cell-interaction). Složková interakce je definována jako bilance mezi složkou na pozici (i,j) a všemi ostatními prvky DDP matice,

$$I_{cell}(i,j) = \sqrt{\frac{rc-1}{rc}} \ln \frac{y_{ij}}{\left(\prod_{(k,l)\neq(i,j)} y_{kl}\right)^{1/(rc-1)}}.$$

Konkrétně pro čtyřpolní tabulku dostáváme

$$I_{cell}(1,1) = \sqrt{\frac{3}{4}} \ln \frac{y_{11}}{\sqrt[3]{y_{12}y_{21}y_{22}}},$$

$$I_{cell}(1,2) = \sqrt{\frac{3}{4}} \ln \frac{y_{12}}{\sqrt[3]{y_{11}y_{21}y_{22}}},$$

$$I_{cell}(2,1) = \sqrt{\frac{3}{4}} \ln \frac{y_{21}}{\sqrt[3]{y_{11}y_{12}y_{22}}},$$

$$I_{cell}(2,2) = \sqrt{\frac{3}{4}} \ln \frac{y_{22}}{\sqrt[3]{y_{11}y_{12}y_{21}}}.$$

Znaménko $I_{cell}(i,j)$, resp. $I_{cross}(i,j)$, udává, zda se jedná o interakci konstruktivní nebo destruktivní. Konstruktivní interakce znamená, že pravděpodobnost výskytu kombinace dané řádkové a sloupcové hodnoty v původní tabulce $\mathbf{x}$ je vyšší než v tabulce nezávislé $\mathbf{x}_{ind}$, pro destruktivní interakci je naopak pravděpodobnost v původní tabulce nižší oproti pravděpodobnosti v tabulce nezávislé.

Stejně jako pro křížový kontrast platí i pro složkovou interakci, že součet druhých mocnin bilancí je úměrný druhé mocnině normy interakční matice. Platí

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\left(I_{cell}(i,j)\right)^2 = \frac{rc}{rc-1} \cdot ||\mathbf{x}_{int}||_a^2. \tag{8}$$

Z uvedeného vyplývá, že křížový kontrast i složková interakce poskytují ekvivalentní informaci o interakci mezi řádkem a sloupcem.

Rozklady druhé mocniny normy $||\mathbf{x}_{int}||_a^2$ ze vztahu (7) a (8) odpovídají neortogonálnímu perturbačnímu rozkladu interakční matice. Původní DDP matici pak můžeme rozložit následujícím způsobem,

$$\mathbf{x} = \mathbf{x}_{ind} \oplus \Big(\bigoplus_{i=1}^{r} \bigoplus_{j=1}^{c} \big(a \cdot I_{cross}(i,j)\big) \odot \mathbf{c}_{ij} \Big), \quad a = \frac{(r+c-2)(r+c-1)}{(r+c)^2}, \tag{9}$$

kde DDP matice křížových interakcí jsou rovny

$$\mathbf{c}_{ij} = C \exp \begin{pmatrix} 0 & \dots & -B & \dots & 0 \\ 0 & \dots & -B & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ -B & \dots & A & \dots & -B \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & -B & \dots & 0 \end{pmatrix},$$

$$A = (r+c-2)^{1/2}(r+c-1)^{-1/2}, \quad B = ((r+c-2)(r+c-1))^{-1/2}$$

a nenulový řádek je i-tý a sloupec j-tý; $||\mathbf{c}_{ij}||_a = 1$.

Pro čtyřpolní tabulku můžeme psát

$$\mathbf{x} = \mathbf{x}_{ind} \oplus \Big(\bigoplus_{i=1}^{2} \bigoplus_{j=1}^{2} \Big(\frac{3}{8} I_{cross}(i,j) \Big) \odot \mathbf{c}_{ij} \Big), \tag{10}$$

$$A = \frac{\sqrt{6}}{3}, \quad B = \frac{\sqrt{6}}{6} \quad \text{a např.} \quad \mathbf{c}_{22} = C \exp \begin{pmatrix} 0 & -\frac{\sqrt{6}}{6} \\ -\frac{\sqrt{6}}{6} & \frac{\sqrt{6}}{3} \end{pmatrix}.$$

Uvedený vztah (10) pro rozklad čtyřpolní matice $\mathbf{x}$ ověříme zpětně postupnými výpočty.

Pro výpočet bilancí $I_{cross}(i,j)$ vycházíme z interakční matice, proto si pro usnadnění zápisu symbolicky označme její prvky a zopakujme si, jak vypadá její explicitní tvar,

$$\mathbf{x}_{int} = C \begin{pmatrix} x_{int11} & x_{int12} \\ x_{int21} & x_{int22} \end{pmatrix} = C \begin{pmatrix} \frac{1}{\sqrt{x_{12}x_{21}}} & \frac{1}{\sqrt{x_{11}x_{22}}} \\ \frac{1}{\sqrt{x_{11}x_{22}}} & \frac{1}{\sqrt{x_{12}x_{21}}} \end{pmatrix}.$$

Označíme první prvek perturbace $\bigoplus_{i=1}^{2}\bigoplus_{j=1}^{2}\left(\frac{3}{8}I_{cross}(i,j)\right)\odot\mathbf{c}_{ij}$ jako matici $\mathbf{I}$,

$$\left(\frac{3}{8}\cdot I_{cross}(1,1)\right)\odot\mathbf{c}_{11} = \begin{pmatrix} \mathrm{I}_{11} & \mathrm{I}_{12} \\ \mathrm{I}_{21} & \mathrm{I}_{22} \end{pmatrix} = \mathbf{I}.$$

Pro explicitní vyjádření matice $\mathbf{I}$ potřebujeme matici $\mathbf{c}_{11}$, která je tvaru

$$\mathbf{c}_{11} = C\exp\begin{pmatrix} \frac{\sqrt{6}}{3} & -\frac{\sqrt{6}}{6} \\ -\frac{\sqrt{6}}{6} & 0 \end{pmatrix}$$

a bilanci $I_{cross}(1,1)$,

$$I_{cross}(1,1) = \frac{\sqrt{6}}{3}\ln\frac{x_{int11}}{\sqrt{x_{int12}x_{int21}}} = \frac{\sqrt{6}}{3}\ln\frac{\sqrt{x_{11}x_{22}}}{\sqrt{x_{12}x_{21}}}.$$

Po provedení mocninné transformace získáme prvky matice $\mathbf{I}$:

$$\mathrm{I}_{11} = e^{\frac{\sqrt{6}}{3}\frac{3}{8}\frac{\sqrt{6}}{3}\ln\frac{\sqrt{x_{11}x_{22}}}{\sqrt{x_{12}x_{21}}}} = \sqrt[8]{\frac{x_{11}x_{22}}{x_{12}x_{21}}},$$

$$\mathrm{I}_{12} = e^{-\frac{\sqrt{6}}{6}\frac{3}{8}\frac{\sqrt{6}}{3}\ln\frac{\sqrt{x_{11}x_{22}}}{\sqrt{x_{12}x_{21}}}} = \sqrt[16]{\frac{x_{12}x_{21}}{x_{11}x_{22}}},$$

$$\mathrm{I}_{21} = \mathrm{I}_{12}, \quad \mathrm{I}_{22} = e^{0} = 1.$$

Jejich uspořádáním do matice $\mathbf{I}$ dostáváme první prvek perturbace,

$$\mathbf{I} = C\begin{pmatrix} \sqrt[8]{\frac{x_{11}x_{22}}{x_{12}x_{21}}} & \sqrt[16]{\frac{x_{12}x_{21}}{x_{11}x_{22}}} \\ \sqrt[16]{\frac{x_{12}x_{21}}{x_{11}x_{22}}} & 1 \end{pmatrix}.$$

Analogicky postupujeme i pro ostaní prvky perturbace, které označíme **II**, **III** a **IV**. Kroky vedoucí k výpočtu matice **II** rozepíšeme stejným způsobem jako pro matici **I**, zbývající prvky **III** a **IV** uvedeme pouze ve výsledném tvaru.

$$\left(\frac{3}{8}\cdot I_{cross}(1,2)\right)\odot \mathbf{c}_{12} = \begin{pmatrix} \mathrm{II}_{11} & \mathrm{II}_{12} \\ \mathrm{II}_{21} & \mathrm{II}_{22} \end{pmatrix} = \mathbf{II},$$

$$\mathbf{c}_{12} = C\exp\begin{pmatrix} -\frac{\sqrt{6}}{6} & \frac{\sqrt{6}}{3} \\ 0 & -\frac{\sqrt{6}}{6} \end{pmatrix},$$

$$I_{cross}(1,2) = \frac{\sqrt{6}}{3}\ln\frac{x_{int12}}{\sqrt{x_{int11}x_{int22}}} = \frac{\sqrt{6}}{3}\ln\frac{\sqrt{x_{12}x_{21}}}{\sqrt{x_{11}x_{22}}},$$

$$\mathrm{II}_{11} = e^{-\frac{\sqrt{6}}{6}\frac{3}{8}\frac{\sqrt{6}}{3}\ln\frac{\sqrt{x_{12}x_{21}}}{\sqrt{x_{11}x_{22}}}} = \sqrt[16]{\frac{x_{11}x_{22}}{x_{12}x_{21}}}, \quad \mathrm{II}_{12} = e^{\frac{\sqrt{6}}{3}\frac{3}{8}\frac{\sqrt{6}}{3}\ln\frac{\sqrt{x_{12}x_{21}}}{\sqrt{x_{11}x_{22}}}} = \sqrt[8]{\frac{x_{12}x_{21}}{x_{11}x_{22}}},$$

$$\mathrm{II}_{21} = e^0 = 1, \quad \mathrm{II}_{22} = \mathrm{II}_{11},$$

$$\mathbf{II} = C\begin{pmatrix} \sqrt[16]{\frac{x_{11}x_{22}}{x_{12}x_{21}}} & \sqrt[8]{\frac{x_{12}x_{21}}{x_{11}x_{22}}} \\ 1 & \sqrt[16]{\frac{x_{11}x_{22}}{x_{12}x_{21}}} \end{pmatrix}.$$

Pro matice **III** i **IV** tedy analogickým postupem obdržíme

$$\mathbf{III} = C \begin{pmatrix} \sqrt[16]{\frac{x_{11}x_{22}}{x_{12}x_{21}}} & 1 \\ \sqrt[8]{\frac{x_{12}x_{21}}{x_{11}x_{22}}} & \sqrt[16]{\frac{x_{11}x_{22}}{x_{12}x_{21}}} \end{pmatrix},$$

$$\mathbf{IV} = C \begin{pmatrix} 1 & \sqrt[16]{\frac{x_{12}x_{21}}{x_{11}x_{22}}} \\ \sqrt[16]{\frac{x_{12}x_{21}}{x_{11}x_{22}}} & \sqrt[4]{\frac{x_{11}x_{22}}{x_{12}x_{21}}} \end{pmatrix}.$$

Na závěr provedeme perturbaci

$$\bigoplus_{i=1}^{2}\bigoplus_{j=1}^{2}\left(\frac{3}{8}\cdot I_{cross}(i,j)\right)\odot \mathbf{c}_{ij} = \mathbf{I}\oplus\mathbf{II}\oplus\mathbf{III}\oplus\mathbf{IV} = C\begin{pmatrix} \sqrt[4]{\frac{x_{11}x_{22}}{x_{12}x_{21}}} & \sqrt[4]{\frac{x_{12}x_{21}}{x_{11}x_{22}}} \\ \sqrt[4]{\frac{x_{12}x_{21}}{x_{11}x_{22}}} & \sqrt[4]{\frac{x_{11}x_{22}}{x_{12}x_{21}}} \end{pmatrix}.$$

Z předchozích znalostí (kapitola 3.4) víme, že

$$\mathbf{x}_{ind} = C\begin{pmatrix} x_{11}\sqrt{x_{12}x_{21}} & x_{12}\sqrt{x_{11}x_{22}} \\ x_{21}\sqrt{x_{11}x_{22}} & x_{22}\sqrt{x_{12}x_{21}} \end{pmatrix},$$

a proto konečnou perturbací získáme původní matici $\mathbf{x}$,

$$\mathbf{x}_{ind} \oplus \left(\bigoplus_{i=1}^{2}\bigoplus_{j=1}^{2}\left(\frac{3}{8}\cdot I_{cross}(i,j)\right)\odot \mathbf{c}_{ij}\right) =$$

$$= C\begin{pmatrix} x_{11}\sqrt[4]{x_{11}x_{12}x_{21}x_{22}} & x_{12}\sqrt[4]{x_{11}x_{12}x_{21}x_{22}} \\ x_{21}\sqrt[4]{x_{11}x_{12}x_{21}x_{22}} & x_{22}\sqrt[4]{x_{11}x_{12}x_{21}x_{22}} \end{pmatrix} = \mathbf{x}.$$

4. Příklad na ortogonální dekompozici

Předchozí teoretické úvahy budeme demonstrovat na konkrétních datech. Zadání následujícího příkladu převzato z [2], str. 287.

Příklad V Anglii zkoumali, zda u zločinců hmotnost souvisí s jejich rozumovými schopnostmi. Získaná data jsou v následující tabulce,

Rozumové schopnosti (Rs)	Hmotnost (Hm) Do 150 liber	Nad 150 liber	Celkem
Normální (N)	272	124	396
Snížené (S)	82	15	97
Celkem	354	139	493

Tabulka 5: Údaje o zločincích.

Prvky příslušné matice pravděpodobností ztotožníme s odhady $\hat{p}_{ij} = \frac{n_{ij}}{n}$, $i, j = 1, 2$.

Rs	Hm < 150	> 150	mrgc	gmrgc
N	0.552	0.252	0.804	0.840
S	0.166	0.030	0.196	0.160
mrgr	0.718	0.282	1	
gmrgr	0.776	0.224		$g(\mathbf{x}) = 0.163$

Tabulka 6: $\mathbf{x}$ – matice pravděpodobností, $||\mathbf{x}||_a^2 = 4.543$.

Nejprve se pokusíme rozložit výchozí matici $\mathbf{x}$ na nezávislou a interakční matici získané přístupem přes aritmetické marginální pravděpodobnosti. Prvky nezávislé matice $\mathbf{x}_{ind}$ vzniknou součinem $p_{i.}p_{.j}$, interakční matice pak vznikne odečtením $\mathbf{x} \ominus \mathbf{x}_{ind}$. Výsledek výpočtu vidíme v tabulkách 7 a 8.

Původní matice $\mathbf{x}$ vykazuje druhou mocninu normy rovnu 4.543, nezávislá tabulka, vzniklá násobením řádkových a sloupcových marginálních pravděpodobností, má druhou mocninu normy rovnu 2.876 a pro interakční matici jsme vypočí-

tali $||\mathbf{x}_{int}||_a^2 = 0.353$. Skutečnost, že $4.543 > 2.876 + 0.353$, dokládá neortogonalitu takovéto dekompozice.

	Hm		
Rs	< 150	> 150	mrgc
N	0.577	0.227	0.804
S	0.141	0.055	0.196
mrgr	0.718	0.282	1

Tabulka 7: Nezávislá tabulka (aritm.).

	Hm		
Rs	< 150	> 150	mrgc
N	0.252	0.292	0.544
S	0.310	0.146	0.456
mrgr	0.562	0.438	1

Tabulka 8: Interakční tabulka (aritm.).

V následujících dvou tabulkách (tabulka 9 a tabulka 10) jsou zaneseny výsledky projekce na ortogonální doplňky řádků a sloupců, jde o geometrickou marginální sloupcovou a řádkovou matici $(\text{row}^{\perp}(\mathbf{x}),\ \text{col}^{\perp}(\mathbf{x}))$, jejichž pertrubací, jak víme, vznikne projekce na podprostor nezávislých matic, tedy matice $\mathbf{x}_{ind}$. Symbolickým zápisem

$$\mathbf{x}_{ind} = \text{row}^{\perp}(\mathbf{x}) \oplus \text{col}^{\perp}(\mathbf{x}).$$

	Hm			
Rs	< 150	> 150	mrgc	gmrgc
N	0.420	0.420	0.840	0.840
S	0.080	0.080	0.160	0.160
mrgr	0.500	0.500	1	
gmrgr	0.500	0.500		

Tabulka 9: $\text{row}^{\perp}(\mathbf{x})$ – geometrická marginální sloupcová matice.

Vypočteme-li normu vzniklé nezávislé matice, $||\mathbf{x}_{ind}||_a^2 = 4.327$ a porovnáme-li ji s normou matice původní, $||\mathbf{x}||_a^2 = 4.543$, zjistíme, že na příslušnou interakční

Rs	Hm < 150	Hm > 150	mrgc	gmrgc
N	0.388	0.112	0.500	0.500
S	0.388	0.112	0.500	0.500
mrgr	0.776	0.224	1	
gmrgr	0.776	0.224		

Tabulka 10: $\mathrm{col}^{\perp}(\mathbf{x})$ – geometrická marginální řádková matice.

matici zbývá 0.216. Z toho můžeme usoudit, že mezi veličinami reprezentujícími rozumové schopnosti zločinců a jejich váhou sice vztah existuje, ale je spíše slabý. Potvrdí nám to i výpočet míry závislosti,

$$R^2_{\Delta} = 0.216/4.543 = 0.047.$$

Rs	Hm < 150	Hm > 150	mrgc	gmrgc
N	0.652	0.188	0.840	0.840
S	0.124	0.036	0.160	0.160
mrgr	0.776	0.224	1	
gmrgr	0.776	0.224		$g(\mathbf{x}_{ind}) = 0.153$

Tabulka 11: $\mathbf{x}_{ind}$ – nezávislá tabulka, $||\mathbf{x}_{ind}||^2_a = 4.327$.

Rs	Hm < 150	Hm > 150	mrgc	gmrgc
N	0.194	0.306	0.500	0.500
S	0.306	0.194	0.500	0.500
mrgr	0.500	0.500	1	
gmrgr	0.500	0.500		$g(\mathbf{x}_{int}) = 0.244$

Tabulka 12: $\mathbf{x}_{int}$ – interakční tabulka, $||\mathbf{x}_{int}||^2_a = 0.216$.

Pro srovnání otestujme tabulku $\mathbf{x}$ testem nezávislosti a zjistěme hodnotu Pearsonova koeficientu, Cramerova V a poměru šancí.

Při testu nezávislosti vypočteme χ^2 podle vztahu (3), vyjde $\chi^2 = 9.67$. Protože $\chi^2 \geq \chi^2_1(0.95) = 3.84$, zamítneme na hladině 0.05 hypotézu, že rozumové

schopnosti nemají vliv na hmotnost zločinců.

Poměr šancí je roven

$$b = 0.401 \neq 1,$$

což indukuje spíše slabou závislost znaků ($d = -0.916$).

Pro Pearsonův koeficient dostaneme

$$C = 0.139,$$

a pro Cramerovo V

$$V = 0.140.$$

Hodnoty obou koeficientů jsou téměř shodné a v intervalu $(0, 1)$ leží blízko nulové hranice, čímž potvrzují nízkou míru závislosti veličin.

Druhá mocnina normy interakční matice může být dále rozložena dle vztahu (7), resp. (8) z kapitoly 3.5. Zaneseme-li hodnoty interakcí (7) na jednotku normy,

$$\frac{(I_{cross}(i,j))^2}{||\mathbf{x}_{int}||_a^2} \cdot \frac{(r+c-1)(r+c-2)}{(r+c)^2}$$

do tabulky, kde pro náš případ

$$I_{cross}(1,1) = \sqrt{\frac{2}{3}} \ln \frac{y_{11}}{\sqrt{y_{12}y_{21}}} = -0.378,$$

$$I_{cross}(1,2) = \sqrt{\frac{2}{3}} \ln \frac{y_{12}}{\sqrt{y_{11}y_{22}}} = 0.378,$$

$$I_{cross}(2,1) = \sqrt{\frac{2}{3}} \ln \frac{y_{21}}{\sqrt{y_{11}y_{22}}} = 0.378,$$

$$I_{cross}(2,2) = \sqrt{\frac{2}{3}} \ln \frac{y_{22}}{\sqrt{y_{12}y_{21}}} = -0.378,$$

máme k dispozici další hodnotný nástroj k analýze závislostí vztahů řádků a sloupců. Zachováme-li navíc znaménko příslušné bilance $I_{cross}(i,j)$, získáme informaci, jak velký vliv má příslušný řádek na sloupec a zda je tato interakce konstruktivní nebo destruktivní.

	Hm	
Rs	< 150	> 150
N	-0.25	0.25
S	0.25	-0.25

Tabulka 13: Interakce na jednotku normy se znaménkem příslušné bilance.

Čtyřpolní interakční tabulka má na hlavní a na vedlejší diagonále stejné prvky, proto jsou si i všechny křížové kontrasty čtyřpolní tabulky až na znaménko rovny. Interakce na jednotku normy nám proto neposkytují prostor pro další interpretaci (jak je vidět též z tabulky 13). Kladná hodnota příslušné interakce znamená, že daná kombinace hodnot znaků převažuje nad ostatními kombinacemi v „křížovém schématu“.

Pro vysvětlení vazeb můžeme použít přímo interakční tabulku, ze které je patrné, že normální rozumové schopnosti se častěji vyskytují v kombinaci s vyšší váhou než naopak. Stejnou informaci nám z klasické analýzy poskytuje i poměr šancí. Protože jeho hodnota je menší než jedna, je patrně větší „šance“ zločinců s normálními rozumovými schopnostmi na vyšší hmotnost než na hmotnost nižší.

V obecném případě interakce na jednotku normy představují poslední článek informace, kterou poskytuje kompoziční analýza aplikovaná na kontingenční tabulky. Zastupují sílu interakce (konstruktivní nebo destruktivní) daného řádku a sloupce. Pokud je interakce kladná, pravděpodobnost výskytu dané kombinace hodnot dvou statistických znaků se proti případu nezávislých veličin zvyšuje, pokud je záporná, tak snižuje.

Závěr

Snahou této práce bylo představit čtenáři nový pohled na řešení problémů spojených s kontingenčními tabulkami. Protože tento nový úhel pohledu vychází z teorie kompozičních dat, jejíž principy nejsou ještě široce známy, věnovala jsem příslušný prostor pro seznámení čtenáře s touto problematikou. Neopomněla jsem uvést ani srovnání nového postupu s již zavedeným přístupem k analýze kontingenčních tabulek zejména pomocí testování nezávislosti a sestavování logaritmicko-lineárních modelů.

Doufám, že se mi podařilo zpracovat danou tematiku srozumitelným a přehledným způsobem, který dovolí čtenáři pochopit danou oblast bez větších nesnází.

Největší osobní přínos této práce pro mě znamenala možnost seznámit se s oblastí kompozičních dat, která mi do této doby byla neznámým pojmem. Bylo vzrušující nahlédnout do tak mladého oboru statistiky, který v současné době prochází bouřlivým rozvojem.

Literatura

[1] Aitchison, J.: The statistical analysis of compositional data, Monographs on statistics and applied probability. Chapman & Hall, London, 1986.

[2] Anděl J.: Základy matematické statistiky. Matfyzpress, Praha, 2005.

[3] Bakytová H. a kol.: Základy statistiky. Vydavatelstvo technickej a ekonomickej literatúry, Bratislava, 1975.

[4] Egozcue, J., J., Diaz-Barrero, J., L., Pawlowsky-Glahn, V.: Compositional analasis of bivariate discrete probabilities. CoDaWork'08, Girona, 2008.

[5] Egozcue, J., J., Pawlowsky-Glahn,V.: Simplicial geometry for compositional data. In: Buccianti A., Mateau-Figueras G., Pawlowsky-Glahn V. (eds.): Compositional data analasis in the Geosciences: From theory to practise. The Geological Society of London, 2006.

[6] Egozcue, J., J., Pawlowsky-Glahn,V., Mateu-Figueras, G., Barceló-Vidal, C.: Isometric Logratio Transformations for Compositional Data Analysis, Mathematical Geology, 2003, **35**, 3, 145-158.

[7] Jarošová E., Pecáková I.: Příklady z předmětu statistiky B. VŠE, Praha, 2000.

[8] Pawlowsky-Glahn,V., Egozcue,J.,J., Tolosana-Delgado, R.: Lecture notes on compositional data analasis, 2007 [online],
dostupné z http://handl.net/10256/297, [citováno 15.11. 2009].

[9] Prášková Z.: Kontingenční tabulky. Univerzita Karlova, Praha, 1985.

[10] Zvára K.: Biostatistika. Nakladatelství Karolinum, Praha, 2004.