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

BAKALÁŘSKÁ PRÁCE

Korelační analýza pro kompoziční data

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

Vypracoval:
Martin Petera
Aplikovaná statistika, III. ročník

Prohlášení

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

V Olomouci dne 31. března 2010

Poděkování

Rád bych na tomto místě poděkoval vedoucímu bakalářské práce RNDr. Karlu Hronovi, Ph.D. za obětavou spolupráci i za čas, který mi věnoval při konzultacích. Dále si zasluží poděkování můj počítač, že vydržel moje pracovní tempo, a typografický systém TEX, kterým je práce vysázena.

Obsah

Úvod

Některé problémy řešili matematikové dlouhá staletí bez uspokojivých výsledků. Mezi ně patří i práce s daty obsahujícími pouze relativní informaci (speciálně se jedná o data charakterizovaná procentuálními podíly). Přitom právě v korelační analýze, kdy vysvětlujeme vztahy mezi proměnnými, totiž může dojít k celé škále absurdních situací, jež ukazují na nutnost použití zcela odlišné koncepce při zpracování dat tohoto typu.

Až poměrně nedávno se centrem zájmu mnoha statistiků i řady odborníků z oblasti aplikací (zejména geologů) stala *kompoziční data*, jako synonymum pro výše popsaný druh pozorování, jejichž statistická pozorování se opírájí o tzv. *logratio analýzu* (log-ratio = logaritmus podílu). Zdá se, že tato metoda nabízí kýžené relevantní řešení, i když je zřejmě ještě potřeba dalšího intezivního výzkumu v této oblasti. A právě provést vás úskalími korelační analýzy pro kompoziční data si klade za cíl tato bakalářská práce.

V první kapitole jsou uvedeny základní pojmy spojené se standardní korelační analýzou. Ačkoli se jedná o pojmy známé, pro přehlednost a ucelenost celé práce je nezbytná jejich prezentace. Mimo běžného korelačního koeficientu jsou přitom zmíněny též koeficient mnohonásobné korelace a skupinový korelační koeficient.

Druhá kapitola vás seznámí se samotnými kompozičními daty, která tvoří základ celé práce. Zde jsem si vytyčil úkol nadefinovat nejen samotný pojem kompozičních dat, ale podrobně jsou popsány též vlastnosti jeho výběrového prostoru - *simplexu*.

Předposlední kapitola osvětluje samotný název této bakalářské práce. S kompozičními daty musíme totiž v případě hledání vztahů (závislostí) mezi proměnnými pracovat odlišně, než jak je zvykem v případě standardních pozorování.

V závěrečné (čtvrté) kapitole jsou podrobně popsány postupy řešení konkrétních problémů. Tato část obsahuje dva příklady, první je ilustrativní a právě na něm ukazuji, jak jsem prováděl výpočty ve zvoleném statistickém softwaru R. Druhý příklad je můj vlastní a souvisí s oblastí sdělovacích prostředků, kterou se ve svém volném čase zabývám.

1 Korelační analýza

V následující kapitole uvádíme základní pojmy standardní korelační analýzy, které budou velmi důležité pro pozdější práci s kompozičními daty. Než se ovšem dostaneme k samotným korelacím, shrneme si základní pojmy, jenž jsou nezbytně nutné k jejich pochopení ([9], str. 50).

Definice 1.1. *Nechť X je diskrétní náhodná veličina s rozdělením daným dvojicí posloupností $\{x_n\}, \{p_n\}$. Je-li*

$$\sum_n |x_n| p_n = \sum_n |x_n| \mathsf{P}(X = x_n) < \infty,$$

nazveme součet řady

$$\sum_n x_n p_n = \sum_n x_n \mathsf{P}(X = x_n)$$

střední hodnotou $\mathsf{E}(X)$ náhodné veličiny X. *Pokud není uvedená podmínka splněna, řekneme, že náhodná veličina X nemá střední hodnotu.*

Definice 1.2. *Nechť X je spojitá náhodná veličina s hustotou $f_X(x)$. Je-li*

$$\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty,$$

nazveme integrál

$$\int_{-\infty}^{\infty} x f_X(x) dx$$

střední hodnotou $\mathsf{E}(X)$ náhodné veličiny X. *Není-li podmínka splněna, řekneme, že náhodná veličina X nemá střední hodnotu.*

Definice 1.3. *Druhý centrální moment náhodné veličiny X se nazývá* rozptyl (variance, disperze) náhodné veličiny X. *Obvykle se označuje*

$$\mathrm{var}(X) = \mathsf{E}[X - \mathsf{E}(X)]^2.$$

Druhá odmocnina z rozptylu $\sqrt{\mathrm{var}(X)}$ se nazývá směrodatná (standardní, střední kvadratická) odchylka náhodné veličiny X.

Definice 1.4. Střední hodnota $\mathsf{E}(\mathbf{X})$ náhodného vektoru $\mathbf{X} = (X_1, \ldots, X_D)'$ *se definuje jako*

$$\mathsf{E}(\mathbf{X}) = (\mathsf{E}(X_1), \ldots, \mathsf{E}(X_D))',$$

jestliže příslušné střední hodnoty složek vektoru existují.

Definice 1.5. *Nechť náhodné veličin X, Y mají konečné druhé momenty.* Kovariance $\mathrm{cov}(X, Y)$ náhodných veličin X, Y *je číslo definované vztahem*

$$\mathrm{cov}(X, Y) = \mathsf{E}([X - \mathsf{E}(X)][Y - \mathsf{E}(Y)]).$$

1.1 Korelační koeficient a korelační matice

Nyní se již můžeme přesunout k samotným korelacím. Na úvod opět připomeneme definici a základní vlastnosti korelačního koeficientu dvou náhodných veličin, ze kterých budeme dále vycházet při zobecnění.

Definice 1.6. *Nechť X, Y mají konečné druhé momenty a nechť* $\mathrm{var}(X) \neq 0$, $\mathrm{var}(Y) \neq 0$. Korelační koeficient $\rho_{X,Y}$ náhodných veličin X, Y *je číslo definované vztahem*

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

Věta 1.1. ([9], str. 97) Pro kovarianci a korelační koeficient platí tyto vlastnosti:

1. $\mathrm{cov}(X,X) = \mathrm{var}(X),\ \rho_{X,X} = 1$
2. $\mathrm{cov}(X,Y) = \mathrm{cov}(Y,X),\ \rho_{X,Y} = \rho_{Y,X}$
3. Jsou-li X, Y nezávislé, je $\mathrm{cov}(X,Y) = 0$, a tedy také $\rho_{X,Y} = 0$.
4. Nechť $a, b, c, d \in \mathbb{R}$, $b \neq 0$ a $d \neq 0$, pak platí $\rho_{a+bX,c+dY} = \rho_{X,Y} \cdot \mathrm{sgn}(bd)$.
5. $|\rho_{X,Y}| \leq 1$,
   $\rho_{X,Y} = 1 \Leftrightarrow \exists\ a, b,\ b > 0$ tak, že $\mathsf{P}(Y = a + bX) = 1$,
   $\rho_{X,Y} = -1 \Leftrightarrow \exists\ a, b,\ b < 0$ tak, že $\mathsf{P}(Y = a + bX) = 1$.

Definice 1.7. *Matice*

$$\mathbf{V} = \mathrm{var}(\mathbf{X}) = (\mathrm{cov}(X_i, X_j))_{i,j=1}^{p}$$

se nazývá varianční matice náhodného vektoru.

Definice 1.8. *Matice* $\mathbf{P} = \mathrm{cor}(\mathbf{X})$*, jejíž prvky jsou korelační koeficienty veličin* X_i *a* X_j*, kde* $i, j = 1, \ldots, p$ *se nazývá* korelační matice náhodného vektoru $\mathbf{X}$.

Pro korelační matici $\mathrm{cor}(\mathbf{X})$ zřejmě platí, že je symetrická a pozitivně semidefinitní, obdobně jako je tomu u varianční matice vektoru $\mathbf{X}$. Následující lemma dává do souvislosti varianční a korelační matici.

Lemma 1.1. ([3], str. 40, Lemma 2.18) Nechť $\mathbf{X} = (X_1, \ldots, X_p)'$ je náhodný vektor s konečnými druhými momenty, jehož všechny složky mají kladné rozptyly. Označme $\mathbf{V} = \mathrm{var}(\mathbf{X})$, $\mathbf{P} = \mathrm{cor}(\mathbf{X})$, $\sigma_i = \sqrt{\mathrm{var}(X_i)}$ pro $i = 1, \ldots, p$ a

$$\mathbf{D} = \mathrm{Diag}\{\sigma_1, \ldots, \sigma_p\} = \begin{pmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ & \vdots & & \\ 0 & 0 & \ldots & \sigma_p \end{pmatrix}.$$

Pak $\mathbf{P} = \mathbf{D}^{-1}\mathbf{V}\mathbf{D}$.

Nyní se blíže zaměříme na výběrový korelační koeficient. V praxi totiž není v našich silách pracovat s celým rozdělením pravděpodobnosti, proto se tu uplatňuje náhodný výběr z nějakého dvojrozměrného rozdělení.

Před samotnou definicí označme $\bar{X}, S_X^2$ jako charakteristiky prvního výběru $X_1, \ldots, X_n$. Obdobně učiníme i pro druhý výběr $Y_1, \ldots, Y_n$ $(\bar{Y}, S_Y^2)$.

Definice 1.9. *Mějme náhodný výběr o rozsahu n z dvourozměrného rozdělení náhodného vektoru $(X, Y)'$,*

$$(X_1, Y_1)', \ldots, (X_n, Y_n)'.$$

Potom výběrovou kovarianci definujeme jako

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}).$$

Definice 1.10. *Jsou-li výběrové rozptyly nenulové, potom jako* výběrový korelační koeficient *označujeme*

$$r_{X,Y} = \frac{S_{XY}}{\sqrt{S_X^2 S_Y^2}}.$$

Vlastnosti výběrového korelačního koeficientu kopírují vlastnosti jeho teoretické obdoby. Pro vícerozměrné rozdělení slouží *výběrová korelační matice*, opět s obdobnými vlastnostmi jako $\mathrm{cor}(\mathbf{X})$.

Mějme $\mathbf{X}_1, \ldots, \mathbf{X}_n$ náhodný výběr o rozsahu n z nějakého p-rozměrného rozdělení. *Výběrovou varianční matici* definujeme

$$\mathbf{S} = (S_{ij}) = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})',$$

kde $\bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i$ představuje výběrový průměr. Abychom mohli stanovit vztah pro *výběrovou korelační matici*, musíme uvažovat situaci, kdy všechny diagonální prvky matice $\mathbf{S}$ jsou kladné. Pak

$$\mathbf{R} = (r_{ij}) = \left(\frac{S_{ij}}{\sqrt{S_{ii} S_{jj}}} \right)_{i,j=1}^{p}.$$

1.2 Koeficient mnohonásobné korelace a skupinový korelační koeficient

Často se ovšem vyskytují i situace, kdy chceme popsat sílu lineárního vztahu mezi náhodnou veličinou Y a náhodným vektorem $\mathbf{X}$ pouze jedním číslem. K tomu

slouží *koeficient mnohonásobné korelace* $\rho_{Y,\mathbf{X}}$. Ještě před tím, než však tento nadefinujeme, je třeba znát bližší informace k pojmu *lineární náhrada* veličiny Y náhodným vektorem $\mathbf{X}$. Jinak řečeno, seznámíme se se vztahem pro regresi Y na $X_1, \ldots, X_p$ (složky náhodného vektoru $\mathbf{X}$); úkolem je nahradit náhodnou veličinu Y lineární funkcí $\widehat{Y}$ veličin $X_1, \ldots, X_p$,

$$\widehat{Y} = a + b_1 X_1 + \cdots + b_p X_p = a + \mathbf{b}'\mathbf{X}$$

tak, aby $\mathsf{E}(Y - \widehat{Y})^2$ byla minimální. K řešení tohoto úkolu nám pomůže následující věta:

Věta 1.2. *Nechť je varianční matice* $\mathrm{var}(\mathbf{X})$ *regulární. Pak platí*

$$\mathsf{E}(Y - a - \mathbf{b}'\mathbf{X})^2 \leq \mathrm{var}(Y) - \mathrm{cov}(Y, \mathbf{X})[\mathrm{var}(\mathbf{X})]^{-1}\mathrm{cov}(\mathbf{X}, Y)$$

a rovnosti je dosaženo tehdy a jen tehdy, když

$$\mathbf{b} = [\mathrm{var}(\mathbf{X})]^{-1}\mathrm{cov}(\mathbf{X}, Y), \;\; a = \mathsf{E}(Y) - \mathbf{b}'\mathsf{E}(\mathbf{X}),$$

kde

$$\mathrm{cov}(\mathbf{X}, Y) = (\mathrm{cov}(X_1, Y), \ldots, \mathrm{cov}(X_p, Y))' = [\mathrm{cov}(Y, \mathbf{X})]'.$$

Důkaz: Viz [3], str. 38, Věta 2.15.

□

Zobecněním (obyčejného) korelačního koeficientu mezi dvěma náhodnými veličinami je právě *koeficient mnohonásobné korelace*. Tentokrát tedy hledáme vztah mezi náhodnou veličinou a skupinou náhodných veličin.

Definice 1.11. *Koeficient mnohonásobné kolerace* $\rho_{Y,\mathbf{X}}$ *je (obyčejný) korelační koeficient mezi* Y *a její optimální lineární náhradou (náhodnou veličinou)* $\widehat{Y} = a + \mathbf{b}'\mathbf{X}$*, kde regresní koeficienty jsou uvedeny ve Větě 1.2. Tedy*

$$\rho_{Y,\mathbf{X}} = \begin{cases} \rho_{Y,a+\mathbf{b}'\mathbf{X}}, & \mathbf{b} \neq \mathbf{0}; \\ 0, & \mathbf{b} = \mathbf{0}. \end{cases}$$

Z této definice je zřejmé, že $\rho_{Y,\mathbf{X}}$ může nabývat pouze nezáporných hodnot.

Věta 1.3. *Označme* $\mathbf{P} = \mathrm{cor}(\mathbf{X})$*. Potom* koeficient mnohonásobné korelace $\rho_{Y,\mathbf{X}}$ *lze vyjádřit ve výpočetně vhodném tvaru*

$$\rho^2_{Y,\mathbf{X}} = \mathrm{cor}(Y, \mathbf{X})\mathbf{P}^{-1}\mathrm{cor}(\mathbf{X}, Y).$$

Přitom $\mathrm{cor}(\mathbf{X}, Y) = (\rho_{X_1,Y}, \ldots, \rho_{X_p,Y})', \;\; \mathrm{cor}(Y, \mathbf{X}) = [\mathrm{cor}(\mathbf{X}, Y)]'$.

Analogicky jako u korelačního koeficientu zavedeme pojem *výběrového koeficientu mnohonásobné korelace*.

Definice 1.12. *Nechť*

$$(Y_1, \mathbf{X}_1)', \ldots, (Y_n, \mathbf{X}_n)'$$

je náhodný výběr z rozdělení, které má náhodný vektor $(Y, \mathbf{X}')'$*, a nechť výběrová korelační matice* $\mathbf{R}$ *vektoru* $\mathbf{X}$ *je regulární. Dále nechť* $\widehat{\mathrm{cor}}(Y, \mathbf{X})$ *a* $\widehat{\mathrm{cor}}(\mathbf{X}, Y)$ *obsahují příslušné výběrové korelační koeficienty. Pak*

$$r_{Y,\mathbf{X}} = \sqrt{\widehat{\mathrm{cor}}(Y, \mathbf{X})\mathbf{R}^{-1}\widehat{\mathrm{cor}}(\mathbf{X}, Y)}$$

se nazývá výběrový koeficient mnohonásobné korelace.

Než postoupíme dále, uvedeme si následující pomocnou větu:

Lemma 1.2. ([7], str. 188, Theorem 13.3.8) Nechť

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$$

je taková bloková matice, že bloky $\mathbf{A}, \mathbf{D}$ jsou čtvercové. Je-li $\mathbf{D}$ regulární, pak pro její determinant platí

$$\begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{vmatrix} = |\mathbf{D}| \cdot |\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}|.$$

Je-li $\mathbf{A}$ regulární, pak

$$\begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{vmatrix} = |\mathbf{A}| \cdot |\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}|.$$

Podívejme se na korelační koeficient dvou veličin X a Y z trochu jiného pohledu ([4], str. 307). Ten totiž úzce souvisí s tzv. Hadamardovou nerovností. Pro determinant matice

$$\mathbf{W} = \mathrm{var}\left[(X, Y)'\right] = \begin{pmatrix} \mathrm{var}(X), & \mathrm{cov}(X, Y) \\ \mathrm{cov}(Y, X), & \mathrm{var}(Y) \end{pmatrix}$$

platí

$$|\mathbf{W}| \leq \mathrm{var}(X) \cdot \mathrm{var}(Y).$$

Podíl levé a pravé strany nerovnosti přitom činí

$$\frac{|\mathbf{W}|}{\mathrm{var}(X) \cdot \mathrm{var}(Y)} = 1 - \rho^2_{X,Y}.$$

Čím více se tedy blíží podíl obou stran v nerovnosti jedné, tím je menší lineární závislost mezi zkoumanými veličinami. Jestliže se naopak levá strana nerovnosti rovná nule, je korelační koeficient v absolutní hodnotě roven jedné.

Pro $\mathbf{X} = (X_1, \ldots, X_p)'$ zaveďme matici

$$\mathbf{W} = \mathrm{var}\left[(Y, \mathbf{X}')'\right] = \begin{pmatrix} \mathrm{var}(Y) & \mathrm{cov}(Y, \mathbf{X}) \\ \mathrm{cov}(\mathbf{X}, Y) & \mathrm{var}(\mathbf{X}) \end{pmatrix}.$$

Z Hadamardovy nerovnosti víme, že platí následující nerovnost,

$$|\mathbf{W}| \leq |\mathrm{var}(\mathbf{X})| \cdot \mathrm{var}(Y).$$

Současně z Lemmatu 1.2

$$|\mathbf{W}| = |\mathrm{var}(\mathbf{X})| \cdot [\mathrm{var}(Y) - \mathrm{cov}(Y, \mathbf{X}) \cdot (\mathrm{var}(\mathbf{X}))^{-1} \cdot \mathrm{cov}(\mathbf{X}, Y)].$$

Podíl levé a pravé strany nerovnosti činí

$$\frac{|\mathbf{W}|}{|\mathrm{var}(\mathbf{X})| \cdot \mathrm{var}(Y)} = 1 - \rho^2_{Y,\mathbf{X}}.$$

Tímto krokem jsme obdrželi koeficient mnohonásobné korelace.

Situaci můžeme dále zobecnit, konkrétně tedy budeme hledat vztah mezi dvěma skupinami náhodných veličin. To vše provedeme na základě analogie, následně zavedeme další pojem - *skupinový korelační koeficient.*

Matice $\mathbf{W}$ tentokrát označuje $\mathrm{var}\left[(\mathbf{X}', \mathbf{Y}')'\right]$ a její struktura je analogická jako v případě koeficientu mnohonásobné korelace ($\mathbf{Y} = (Y_1, \ldots, Y_q)'$),

$$\mathbf{W} = \mathrm{var}[(\mathbf{X}', \mathbf{Y}')'] = \begin{pmatrix} \mathrm{var}(\mathbf{X}) & \mathrm{cov}(\mathbf{X}, \mathbf{Y}) \\ \mathrm{cov}(\mathbf{Y}, \mathbf{X}) & \mathrm{var}(\mathbf{Y}) \end{pmatrix},$$

kde

$$\mathrm{cov}(\mathbf{X}, \mathbf{Y}) = \begin{pmatrix} \mathrm{cov}(X_1, Y_1) & \mathrm{cov}(X_1, Y_2) & \ldots & \mathrm{cov}(X_1, Y_q) \\ \mathrm{cov}(X_2, Y_1) & \mathrm{cov}(X_2, Y_2) & \ldots & \mathrm{cov}(X_2, Y_q) \\ \vdots & & & \\ \mathrm{cov}(X_p, Y_1) & \mathrm{cov}(X_p, Y_2) & \ldots & \mathrm{cov}(X_p, Y_q) \end{pmatrix}, \ \mathrm{cov}(\mathbf{Y}, \mathbf{X}) = [\mathrm{cov}(\mathbf{X}, \mathbf{Y})]'.$$

Poznámka: Počet složek náhodného vektoru $\mathbf{X}$ se nemusí rovnat počtu složek náhodného vektoru $\mathbf{Y}$.

Z Hadamardovy nerovnosti a díky analogii k předchozímu dostáváme novou charakteristiku,

$$\frac{|\mathbf{W}|}{|\mathrm{var}(\mathbf{X})| \cdot |\mathrm{var}(\mathbf{Y})|} = 1 - \rho^2_{\mathbf{X},\mathbf{Y}}.$$

Definice 1.13. *Číslo* $\rho_{\mathbf{X},\mathbf{Y}}$ *nazveme* skupinovým korelačním koeficientem *vektorů* $\mathbf{X}$ *a* $\mathbf{Y}$.

Z nerovnosti $|\mathbf{W}| \leq |\mathrm{var}(\mathbf{X})| \cdot |\mathrm{var}(\mathbf{Y})|$ vyplývá, že $\rho^2_{\mathbf{X},\mathbf{Y}} \leq 1$. Skupinový korelační koeficient $\rho_{\mathbf{X},\mathbf{Y}}$ nám tedy určuje sílu vztahu mezi dvěma skupinami náhodných veličin.

Je přitom zřejmé, že obdobné vztahy platí též pro výběrové verze uvedených koeficientů.

2 Kompoziční data

O kompozičních datech jsme se krátce zmínili již v úvodu této bakalářské práce. Věnujme se jim nyní podrobněji. Čerpáno bylo zejména z [12] a [13].

Definice 2.1. *Sloupcový vektor* $\mathbf{x} = (x_1, x_2, \ldots, x_D)'$ *se nazývá* D*-složková kompozice, jestliže všechny jeho složky jsou* kladná čísla *a nesou pouze* relativní informaci.

Touto *relativní informací* máme na mysli, že nikoli absolutní hodnoty, ale pouze podíly mezi složkami jsou pro nás relevantní. Jinak řečeno, vektor $\mathbf{x} = (x_1, x_2, \ldots, x_D)'$ a $a\mathbf{x} = (ax_1, ax_2, \ldots, ax_D)'$ nám dávají stejnou informaci $(a > 0)$. Možným způsobem, jak následně zjednodušit použití a interpretaci kompozičních dat, je tedy možnost reprezentovat je jako kladné vektory se součtem složek rovným kladné konstantě κ. Při volbě $\kappa = 1$ nebo 100 je takto kompozice $\mathbf{x}$ (bez ztráty informace) reprezentovaná ve formě proporcí nebo procentuálních podílů.

Tato úvaha nás vede k následujícím definicím:

Definice 2.2. *Výběrový prostor kompozičních dat je simplex, podmnožina dimenze* $D - 1$ *reálného prostoru* $\mathbb{R}^D$*, definovaná jako*

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

Definice 2.3. *Pro každou* D*-složkovou kompozici*

$$\mathbf{z} = (z_1, z_2, \ldots, x_D)' \in \mathbb{R}_+^D$$

($z_i > 0, \forall i = 1, 2, \ldots, D$*) je* uzávěrem *rozuměn vektor*

$$\mathcal{C}(\mathbf{z}) = \left[\frac{\kappa \cdot z_1}{\sum_{i=1}^{D} z_i}, \frac{\kappa \cdot z_2}{\sum_{i=1}^{D} z_i}, \cdots, \frac{\kappa \cdot z_D}{\sum_{i=1}^{D} z_i} \right]'.$$

Definice 2.4. *Pro danou kompozici* $\mathbf{x}$ *obdržíme* podkompozici $\mathbf{x}_s$ *(obsahující* s *částí,* $s < D$*) aplikací operace uzávěru na podvektor* $(x_{i_1}, x_{i_2}, \ldots, x_{i_s})'$ *vektoru* $\mathbf{x}$*. Pro indexy* $i_1, \ldots, i_s$*, které určují vybrané složky kompozice* $\mathbf{x}$*, přitom platí* $1 \leq i_1 < \ldots < i_s \leq D$*.*

Na úvod ještě poznamenejme, že od této chvíle budeme značit případné náhodné objekty i jejich realizace malými písmeny. Dle kontextu též bude zřejmé, zda pracujeme s teoretickými či výběrovými charakteristikami těchto objektů.

2.1 Aitchisonova geometrie

V reálném prostoru jsme zvyklí provádět operace s vektory (sčítání, násobení skalárem, zjisťování jejich vzájemné vzdálenosti, ortogonality a podobně), vše založené na standardní vektorové algebře a euklidovské metrice. Přestože je simplex speciálním výběrovým prostorem, tak i zde chceme nějaké takovéto operace používat. Musíme je však nejprve upravit vzhledem k charakteru dat, například euklidovská metrika by zde totiž mohla činit velké potíže.

Pochopitelně nebudeme zavádět zcela nové pojmy. Raději zvolíme smysluplnou cestu, která povede k analogickým operacím jako na reálném prostoru.

Definice 2.5. *Perturbací kompozic* $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ *rozumíme kompozici*

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

Definice 2.6. *Mocninná transformace kompozice* $\mathbf{x} \in \mathcal{S}^D$ *konstantou* $\alpha \in \mathbb{R}$ *je definována jako*

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

Připomeňme si následující obecnou definici:

Definice 2.7. *Čtveřici* $\mathcal{V} = (\mathbf{V}, +, \mathbf{T}, \cdot)$ *nazýváme vektorový prostor nad tělesem* $\boldsymbol{T}$*, jestliže*

1. $(\mathbf{V}, +)$ *je abelovská grupa;*
2. $\mathbf{T}$ je číselné těleso;
3. operace $\cdot$ splňuje požadavek $\mathbf{T} \cdot \mathbf{V} \to \mathbf{V}$;
4. $\forall \mathbf{u}, \mathbf{v} \in \mathbf{V}$ a současně $\forall c, d \in \mathbf{T}$:
   - $c \cdot (\mathbf{u} + \mathbf{v}) = c \cdot \mathbf{u} + c \cdot \mathbf{v}$,
   - $(c + d) \cdot \mathbf{u} = c \cdot \mathbf{u} + d \cdot \mathbf{u}$,
   - $(c \cdot d) \cdot \mathbf{u} = c \cdot (d \cdot \mathbf{u})$,
   - $1 \cdot \mathbf{u} = \mathbf{u}$.

Věta 2.1. *Simplex společně s perturbací a mocninnou transformací* $(\mathcal{S}^D, \oplus, \odot)$ *tvoří reálný vektorový prostor.*

Důkaz: Jedním z předpokladů, aby výše uvedená trojice byla skutečně vektorým prostorem, je splnění následující podmínky: $(\mathcal{S}^D, \oplus)$ je *abelovská komutativní grupa*, což znamená:

1. Splňuje komutativní zákon: $\mathbf{x} \oplus \mathbf{y} = \mathbf{y} \oplus \mathbf{x}$.

   Vezmeme-li v úvahu, že při operaci perturbace pracujeme (až na uzávěr) pouze s násobením složek vektorů, jedná se v tomto případě o „obyčejné“ násobení, pro něž komutativní zákon evidentně platí.

2. Asociativní zákon: $(\mathbf{x} \oplus \mathbf{y}) \oplus \mathbf{z} = \mathbf{x} \oplus (\mathbf{y} \oplus \mathbf{z})$.

   Dokážeme obdobně jako výše.

3. Má neutrální prvek:

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

   Dokažme nyní, že $\mathbf{n}$ je skutečně neutrálním prvkem. Provedeme perturbaci

$$\mathbf{n} \oplus \mathbf{x} = \mathcal{C}[1 \cdot x_1, 1 \cdot x_2, \ldots, 1 \cdot x_D] = \mathcal{C}[x_1, x_2, \ldots, x_D] = \mathbf{x}.$$

   Navíc je jasné, že $\mathbf{n}$ je jediným neutrálním prvkem.

4. Posledním z předpokladů splnění podmínky abelovské grupy je přítomnost inverzního prvku $\mathbf{x}$, který označíme $\mathbf{x}^{-1} = \mathcal{C}[x_1^{-1}, x_2^{-1}, \ldots, x_D^{-1}]$; pokud je $\mathbf{x}^{-1}$ inverzním prvkem, dostáváme $\mathbf{x} \oplus \mathbf{x}^{-1} = \mathbf{n}$. A skutečně,

$$\mathbf{x} \oplus \mathbf{x}^{-1} = \mathcal{C}\left[x_1 \cdot x_1^{-1}, x_2 \cdot x_2^{-1}, \ldots, x_D \cdot x_D^{-1}\right] = \mathcal{C}[1, 1, \ldots, 1] = \mathbf{n}.$$

   V tomto kontextu zmiňme, že zapisujeme $\mathbf{x} \oplus \mathbf{y}^{-1} = \mathbf{x} \oplus ((-1) \odot \mathbf{y}) = \mathbf{x} \ominus \mathbf{y}$.

Nyní jsme dokázali, že $(\mathcal{S}^D, \oplus)$ je komutativní abelovská grupa. Stále ovšem nevíme, zda námi definovaná trojice je skutečně vektorovým prostorem. Ověříme tedy, zda pro každé $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ a všechny $\alpha, \beta \in \mathbb{R}$ platí následující čtyři vztahy:

1. $\alpha \odot (\beta \odot \mathbf{x}) = (\alpha \cdot \beta) \odot \mathbf{x}$:

$$\begin{aligned}
\alpha \odot (\beta \odot \mathbf{x}) &= \alpha \odot (\mathcal{C}[(x_1)^\beta, (x_2)^\beta, \ldots, (x_D)^\beta]) &&= \\
&= \mathcal{C}\left[\left((x_1)^\beta\right)^\alpha, \left((x_2)^\beta\right)^\alpha, \ldots, \left((x_D)^\beta\right)^\alpha\right] &&= \\
&= \mathcal{C}[(x_1)^{\alpha\cdot\beta}, (x_2)^{\alpha\cdot\beta}, \ldots, (x_D)^{\alpha\cdot\beta}] &&= (\alpha \cdot \beta) \odot \mathbf{x};
\end{aligned}$$

2. $\alpha \odot (\mathbf{x} \oplus \mathbf{y}) = (\alpha \odot \mathbf{x}) \oplus (\alpha \odot \mathbf{y})$:

$$\alpha \odot (\mathbf{x} \oplus \mathbf{y}) = \mathcal{C}[(x_1 y_1)^\alpha, (x_2 y_2)^\alpha, \ldots, (x_D y_D)^\alpha] =$$

$$= \mathcal{C}[x_1^\alpha y_1^\alpha, x_2^\alpha y_2^\alpha, \ldots, x_D^\alpha y_D^\alpha] = (\alpha \odot \mathbf{x}) \oplus (\alpha \odot \mathbf{y});$$

3. $(\alpha + \beta) \odot \mathbf{x} = (\alpha \odot \mathbf{x}) \oplus (\beta \odot \mathbf{x})$:

   Ukážeme obdobně jako v předchozím případě;

4. $1 \odot \mathbf{x} = \mathbf{x}$:

$$1 \odot \mathbf{x} = \mathcal{C}[x_1^1, x_2^1, \ldots, x_D^1] = \mathcal{C}[x_1, x_2, \ldots, x_D] = \mathbf{x}.$$

Protože trojice $(\mathcal{S}^D, \oplus, \odot)$ splnila všechny požadované podmínky, jedná se o reálný vektorový prostor.

□

Definice 2.8. *Skalární součin dvou kompozic* $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ *lze definovat následovně:*

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

Definice 2.9. *Normu vektoru* $\mathbf{x} \in \mathcal{S}^D$ *definujeme*

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

Definice 2.10. *Vzdálenost mezi* $\mathbf{x}$ *a* $\mathbf{y} \in \mathcal{S}^D$ *získáme jako*

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

O této vzdálenosti hovoříme jako o tzv. *Aitchisonově vzdálenosti*, ve skutečnosti se jedná o upravenou euklidovskou vzdálenost pro všechny přípustné kombinace logaritmů podílů složek kompozic $\mathbf{x}$ a $\mathbf{y}$. Rizikům použití standardní euklidovské vzdálenosti na kompoziční data se věnuje několik zdrojů, zmiňme například [1]. Analogicky hovoříme též o *Aitchisonově normě* a *Aitchisonově skalárnímu součinu*. Společně s vektorovým prostorem na simplexu pak mluvíme o *Aitchisonově geometrii*.

2.2 Vyjádření v souřadnicích

Z předchozí kapitoly je zřejmé, že aplikace standardních statistických metod, založených na vlastnostech euklidovské geometrie, na „neupravená“ kompoziční data může vést k nesmyslným výsledkům (viz též dále). Proto byly navrženy tzv. *logratio transformace*, zobrazující kompozice z $\mathcal{S}^D$ do $(D-1)$-rozměrného, resp. D-rozměrného reálného prostoru, kde je již použití standardních metod pro jejich statistické zpracování možné. Volba dané transformace a možnosti interpretace zobrazených dat ovšem úzce souvisí s vyjádřením kompozic v souřadnicích, jak si nyní uvedeme.

Ještě relativně nedávno byla problematika vyjádření v souřadnicích kompozičních dat považována za nedůležitou a téměř opomíjena. To se odráželo už při zavádění logratio transformací v knize [2] - aditivní logratio (alr) a centrované

logratio (clr) transformace. Metrické vlastnosti kompozic *alr* transformace nezachovává, *clr* sice ano, ale vede k singulární varianční matici. Vyjádření v souřadnicích na simplexu vede tedy k novému přístupu (někdy též označovánému jako *izometrická logratio (ilr) transformace*).

Na simplexu $\mathcal{S}^D$ jsou kompozice vyjádřitelné pomocí kanonické báze $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_D\}$ prostoru $\mathbb{R}^D$. Každý vektor $\mathbf{x} \in \mathbb{R}^D_+$ lze zapsat jako

$$\mathbf{x} = x_1[1, 0, \ldots, 0]' + x_2[0, 1, 0, \ldots, 0]' + \cdots + x_D[0, \ldots, 0, 1]' = \sum_{i=1}^{D} x_i \cdot \mathbf{e}_i.$$

Nastává ale zásadní problém, výše zmíněná kanonická báze totiž není ani generujícím systémem, ani bází vzhledem k Aitchisonově geometrii na simplexu $\mathcal{S}^D$.

Proto vyvstává důležitá otázka, jak vytvořit vhodnou bázi na simplexu. Vezměme generující systém,

$$\mathbf{w}_i = \mathcal{C}(\exp(\mathbf{e}_i)) = \mathcal{C}[1, 1, \ldots, e, \ldots, 1], \;\; i = 1, 2, \ldots, D.$$

Nyní vyjádřeme kompozici $\mathbf{x}$:

$$\mathbf{x} = \bigoplus_{i=1}^{D} \ln \frac{x_i}{g(\mathbf{x})} \odot \mathbf{w}_i = \ln \frac{x_1}{g(\mathbf{x})} \odot [e, 1, \ldots, 1]' \oplus \cdots \oplus \ln \frac{x_D}{g(\mathbf{x})} \odot [1, 1, \ldots, e]',$$

kde

$$g(\mathbf{x}) = \left(\prod_{i=1}^{D} x_i \right)^{\frac{1}{D}} = \exp\left(\frac{1}{D} \sum_{i=1}^{D} \ln x_i \right)$$

je geometrický průměr složek kompozice $\mathbf{x}$.

Na pravé straně výše uvedené rovnosti pro kompozici nalézáme právě známou centrovanou logratio transformaci. Označme vyjádření kompozice pomocí *clr* koeficientů jako

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

Inverzní transformace, která nám poskytne koeficienty v kanonické bázi reálného prostoru, je potom ve tvaru

$$\mathrm{clr}^{-1}(\boldsymbol{\xi}) = \mathcal{C}[\exp(\xi_1), \exp(\xi_2), \ldots, \exp(\xi_D)].$$

Jak už bylo uvedeno výše, *clr* transformace sice zachovává vzdálenosti, ale varianční matice transformace $\boldsymbol{\xi}$ je singulární. Formální definici clr koeficientů můžeme vyjádřit následovně:

Definice 2.11. *Pro kompozici* $\mathbf{x} \in \mathcal{S}^D$ *jsou clr koeficienty složky vektoru* $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_D)'$*, jediného vektoru splňujícího*

$$\mathbf{x} = \mathrm{clr}^{-1}(\boldsymbol{\xi}) = \mathcal{C}(\exp(\boldsymbol{\xi})), \ \sum_{i=1}^{D} \xi_i = 0.$$

Potom i*-tý clr koeficient je*

$$\xi_i = \ln \frac{x_i}{g(\mathbf{x})}.$$

Označme $d(\cdot,\cdot)$, $||\cdot||$ a $\langle \cdot, \cdot \rangle$ standardní eukleidovskou vzdálenost, normu a skalární součin. Potom můžeme vyslovit následující větu:

Věta 2.2. ([13], str. 20, Property 4.1.) Uvažujme $\mathbf{x}_k \in \mathcal{S}^D$, $k = 1, 2$, a reálné konstanty α, β, potom platí:

1. $\mathrm{clr}(\alpha \odot \mathbf{x}_1 \oplus \beta \odot \mathbf{x}_2) = \alpha \cdot \mathrm{clr}(\mathbf{x}_1) + \beta \cdot \mathrm{clr}(\mathbf{x}_2)$,
2. $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_a = \langle \mathrm{clr}(\mathbf{x}_1), \mathrm{clr}(\mathbf{x}_2) \rangle$,
3. $||\mathbf{x}_1||_a = ||\mathrm{clr}(\mathbf{x}_1)||, \ d_a(\mathbf{x}_1, \mathbf{x}_2) = d(\mathrm{clr}(\mathbf{x}_1), \mathrm{clr}(\mathbf{x}_2))$.

2.3 Ortonormální souřadnice

Jako klíčový se (též vzhledem k již uvedenému) jeví požadavek na ortonormalitu báze na simplexu.

Uvažujme ortonomální bázi $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}$ na simplexu $\mathcal{S}^D$ a dále uvažujme matici $\boldsymbol{\Psi}_{(D-1),D}$, jejíž řádky tvoří $\mathrm{clr}(\mathbf{e}_i)$. Pro tuto ortonormální bázi platí $\langle \mathbf{e}_i, \mathbf{e}_j \rangle_a = \delta_{ij}$, přičemž δ_{ij} je tzv. Kroneckerovo delta, které je rovno nule pro $i \neq j$ a jedné pro $i = j$.

Jakmile jsme zvolili bázi, kompozici $\mathbf{x} \in \mathcal{S}^D$ lze vyjádřit následovně,

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

kde $\mathbf{x}^* = (x_1^*, \ldots, x_{D-1}^*)'$ je vektor souřadnic $\mathbf{x}$ vzhledem k vybrané bázi.

Předpis, který umožňuje vyjádřit kompozici $\mathbf{x}$ v souřadnicích $\mathbf{x}^*$, zobrazuje ji tak z $\mathcal{S}^D$ do $\mathbb{R}^{D-1}$, bývá někdy označován jako izometrická transformace.

Následující rovnost vyjadřuje vztah mezi souřadnicemi $\mathbf{x}^*$ a clr transformací:

$$\mathbf{x}^* = \mathrm{ilr}(\mathbf{x}) = \boldsymbol{\Psi} \cdot \mathrm{clr}(\mathbf{x}).$$

Inverzi *ilr* transformace (opětovné získání kompozice z jejích souřadnic) odvodíme z předchozího vztahu a využitím inverzní clr transformace,

$$\mathrm{clr}(\mathbf{x}) = \boldsymbol{\Psi}'\mathbf{x}^*, \ \mathbf{x} = \mathcal{C}(\exp(\boldsymbol{\Psi}'\mathbf{x}^*)),$$

je tvořena třemi kroky:

1. zkonstruovat clr matici báze $\mathbf{\Psi}$,

2. vypočítat maticový součin $\mathbf{\Psi}'\mathbf{x}^*$,

3. aplikovat inverzi clr k získání $\mathbf{x}$.

Existuje několik způsobů k definování ortonormální báze na simplexu. Hlavním kritériem volby báze je možnost interpretace kompozice v souřadnicích. Zejména se přitom zaměříme na báze vzniklé pomocí tzv. *sekvenčního binárního dělení*, protože jsou takto vzniklé souřadnice (nazývané v tomto kontextu též bilance nebo rovnováhy) snadno interpretovatelné z hlediska skupin složek kompozice. Jim se budeme věnovat v další podkapitole.

2.4 Bilance

Budeme-li následně chtít zjistit vztahy mezi jednotlivými zkoumanými skupinami složek kompozice pomocí korelační analýzy, poslouží nám právě bilance. Pro lepší vysvětlení si situaci ilustrujeme na vytvořeném ilustrativním příkladě, který ukazuje přiřazení ortonormálních souřadnic (bilancí) kompozici vybraných televizních stanic, vysílajících na území České republiky.

Jelikož jsme zvolili šest různých televizí (složky kompozice), informaci o vztazích mezi nimi nám popíše pět bilancí. Postup při sestavení sekvenčního binárního dělení není nijak složitý. Popišme jej v několika srozumitelných krocích.

V prvním bodu zapíšeme tabulku obsahující jednotlivé zkoumané televizní subjekty:

komerční				**veřejnoprávní**	
terestrické		kabelové			
Nova	Prima	Spektrum	Disney Channel	ČT1	ČT2
x_1	x_2	x_3	x_4	x_5	x_6

Nyní už máme přehledně uspořádané názvy a můžeme pokračovat další fází. V ní se věnujeme vytváření tabulky bilancí, jedná vlastně se o posloupnost dělení složek kompozic vždy na dvě skupiny. V prvním kroku jsou všechny složky rozděleny na dvě podskupiny, což provedeme i na našem příkladu. První čtyři složky představují komerční televizní stanice, zbylé dvě složky pak představují kanály veřejnoprávní České televize. Toto rozdělení zaznačíme do tabulky, první skupině přiřadíme znaménko $+$, druhé potom pro odlišení znaménko opačné $-$. Postup je samozřejmě možné obrátit a přiřadit odlišná znaménka, toto by mělo následně pouze vliv na interpretaci hodnot bilance. Jednalo by se přitom o ortogonální transformaci nově reprezentovaných složek kompozice vzhledem k původně vytvořeným bilancím.

V dalších krocích členíme jednotlivé skupiny do menších podskupin. První skupinou komerčních stanic jsou terestrické, tedy šířené volně vzduchem. Nesou označení x_1, x_2, druhá polovina komerčních je potom označena jako kabelové, tedy placené programy - x_3, x_4. Jelikož jsme získali dvě další podskupiny, chceme je opět odlišit. Proto první znovu přidělíme znaménko $+$, druhé pak $-$.

V posledním kroku zůstaly tři skupiny složek - x_1, x_2; x_3, x_4; x_5, x_6. I tentokrát přidělíme každému prvku znaménko $+$, respektive $-$.

Celý postup shrnuje následující tabulka. Jsou v ní uspořádány všechny stanovené stanice (označeny x_i), zahrnuty jsou i bilance. Těch je pět (tedy o jednu méně než počet subjektů).

bilance	x_1	x_2	x_3	x_4	x_5	x_6	r	s
z_1	$+$	$+$	$+$	$+$	$-$	$-$	4	2
z_2	$+$	$+$	$-$	$-$			2	2
z_3	$+$	$-$					1	1
z_4			$+$	$-$			1	1
z_5					$+$	$-$	1	1

Na konci každého řádku tabulky jsou taktéž dva sloupce nesoucí značení r a s. Jejich význam je zřejmý; písmeno r udává řádkový součet kladných znamének, s poté značí řádkový součet záporných znamének.

Pro konečné vytvoření bilancí musíme dosadit do obecného vzorce pro tvorbu bilancí. Ten má následující tvar:

$$b = \sqrt{\frac{rs}{r+s}} \ln \frac{(\Pi_+ x_j)^{\frac{1}{r}}}{(\Pi_- x_k)^{\frac{1}{s}}} \text{ pro } i = 1, \ldots, D-1.$$

Výše uvedené znamená, že pro i-tou bilanci bude mít každá složka váhu

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

pro složky nezahrnuté do dělení. Výsledná bilance tak vyjdřuje rovnováhu mezi „plusovými“ a „mínusovými“ složkami.

V případě, že požadujeme konkrétní tvar námi vytvořených bilancí, dosadíme do výše uvedeného obecného vzorce součty kladných a záporných znamének u příslušných bilancí. Zvolíme-li například první bilanci (z_1), potom $r = 4$ a $s = 2$, dostaneme tak první ze vztahů v následující tabulce:

z_1	z_2	z_3	z_4	z_5
$\sqrt{\frac{8}{6}} \ln \frac{(x_1x_2x_3x_4)^{1/4}}{(x_5x_6)^{1/2}}$	$\ln \frac{(x_1x_2)^{1/2}}{(x_3x_4)^{1/2}}$	$\frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2}$	$\frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2}$	$\frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2}$

Je přitom zřejmé, že např. bilance z_2, z_3, z_4 vyčerpávají veškerou informaci o skupině komerčních stanic, z_5 potom o stanicích veřejnoprávních. Nakonec

uveďme i tvar matice $\boldsymbol{\Psi}$ ze začátku předchozí kapitoly,

$$\boldsymbol{\Psi} = \begin{pmatrix} \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 0 & 0 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}.$$

2.5 Matice rozptylů

Bilance, tak jak jsme je zavedli v předchozí kapitole, nám následně umožní použít běžnou korelační analýzu pro skupiny složek náhodné kompozice $\mathbf{x} = (x_1, \ldots, x_D)'$. Nicméně, vyjádřením všech možných dvousložkových podkompozic v souřadnicích je možné najít míru podobnou korelační pro dvojice složek x_i a x_j, $1 \leq i, j \leq D, i \neq j$. Výsledná bilance je v tomto případě ve tvaru $\frac{1}{\sqrt{2}} \ln \frac{x_i}{x_j}$, je to tudíž náhodná veličina, přitom vztah mezi mezi složkami x_i a x_j budeme popisovat pomocí rozptylu. Takto obdržíme

$$t_{ij} = \operatorname{var}\left(\frac{1}{\sqrt{2}} \ln \frac{x_i}{x_j}\right),$$

což vede k následující definici:

Definice 2.12. Maticí rozptylů *náhodné kompozice* $\mathbf{x}$ *nazveme čtvercovou matici*

$$\mathbf{T} = \begin{pmatrix} t_{11} & t_{12} & \ldots & t_{1D} \\ t_{21} & t_{22} & \ldots & t_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ t_{D1} & t_{D2} & \ldots & t_{DD} \end{pmatrix}, \; t_{ij} = \operatorname{var}\left(\frac{1}{\sqrt{2}} \ln \frac{x_i}{x_j}\right).$$

Z definice je zřejmé, že tato matice je symetrická a na diagonále má samé nuly, míra t_{ij} zřejmě nezávisí na měřítku (škále) dat. Hodnota t_{ij} blízká nule vyjadřuje téměř konstantní podíl x_i/x_j, což značí silný vztah mezi i-tou a j-tou složkou. S velkými hodnotami x_i lze takto rovněž očekávat velké hodnoty x_j. Nulové hodnoty t_{ij} (tedy podíl x_i/x_j je konstantní) pak značí, že složka x_i je fixní proporcí složky x_j, jedná se o nejsilnější formu závislosti mezi dvěma složkami.

Naopak pro složky s velkými rozdíly mezi příslušnými hodnotami podílu míra poroste. Obecně míra vzroste též v případě, že data splňují nepřímou úměru, což znamená, že s růstem x_i klesá x_j. Z tohoto důvodu se t_{ij} nechová jako obvyklá kovariance, resp. korelační koeficient.

Stejně tak normovaná verze $\exp(-t_{ij})$ neudává výsledky očekávané od korelační analýzy; je ovšem vhodnější, protože transformuje obdržené hodnoty t_{ij} do intervalu $\langle 0, 1\rangle$. Tuto kapitolu zakončíme následující definicí:

Definice 2.13. *Míra celkové variability kompozice se nazývá* celkový rozptyl *a je dán vztahem*

$$\text{totvar}[\mathbf{x}] = \frac{1}{D}\sum_{i=1}^{D}\sum_{j=1}^{D}\text{var}\left(\frac{1}{\sqrt{2}}\ln\frac{x_i}{x_j}\right) = \frac{1}{D}\sum_{i=1}^{D}\sum_{j=1}^{D}t_{ij}.$$

Ani celkový rozptyl nezávisí na kontantě κ z výběrového prostoru $\mathcal{S}^D$. Matice rozptylů $\mathbf{T}$ navíc vysvětluje, jak je celkový rozptyl rozdělen mezi jednotlivé podíly složek (lépe řečeno - mezi logaritmy podílů - logratios).

3 Korelační analýza pro kompoziční data

Dostáváme se do stěžejní části bakalářské práce. Jak už bylo několikrát řečeno, kompoziční data musí být zpracována zcela odlišným způsobem, než je tomu v případě standardních pozorování (nesoucích absolutní informaci). Problémy, které mohou nastat při jejich korelační analýze, objasníme v následující kapitole, v níž bude zahrnut i ilustrativní příklad.

3.1 Problémy standardního přístupu

Korelační analýza patří mezi oblíbené statistické metody. Jednoduše totiž určí, zda jsou dvě či více proměnných na sobě závislé. Vztah je potom určen jedním číslem, informace obsažená v tomto čísle ve skutečnosti zastupuje n pozorovnání (ve výběrovém případě). Je ovšem na místě si opět uvědomit, že kompoziční data nesou pouze relativní informaci (ta je obsažena v podílech mezi jednotlivými složkami). Úskalí použití korelační analýzy pro kompoziční data ukazuje následující příklad, který byl převzat z [11].

Příklad 3.1. Dva vědci A a B zkoumají složení tří vzorků půdy, které se skládají z rostlinné, živočišné, neživé a vodní části. V této podobě vzorků také naměřil příslušné koncentrace složek vědec A. Druhý vědec přišel s mírně rozdílným přístupem. Pro něj není voda ve vzorku důležitá, a proto půdu před zahájením dalších prací u své části vzorků nejprve vysušil. Až poté, co tuto proceduru provedl, zkoumal koncentrace zbylých složek.

Naměřená data obou vědců byla následně pro lepší interpretaci převedena na konstantní součet složek (zde roven jedné, jedná se tak o proporce) a přenesena do následující tabulky:

vědec A	x_1	x_2	x_3	x_4		vědec B	x_1^+	x_2^+	x_3^+
1	0,1	0,2	0,1	0,6		1	0,25	0,50	0,25
2	0,2	0,1	0,2	0,5		2	0,40	0,20	0,40
3	0,3	0,3	0,1	0,3		3	0,43	0,43	0,14

Je zřejmé, že proporce prvních tří složek (vyjádřené v podílech) si opravdu odpovídají, v případě prvního vzorku totiž například

$$\frac{x_1}{x_2} = \frac{x_1^+}{x_2^+} = \frac{1}{2}.$$

A právě nyní se ukáže, kde se skrývá problém. Srovnáme-li obě výběrové korelační matice

$$\mathbf{R_x} = \begin{pmatrix} 1{,}00 & & & \\ 0{,}50 & 1{,}00 & & \\ 0{,}00 & -0{,}87 & 1{,}00 & \\ -0{,}98 & -0{,}65 & 0{,}19 & 1{,}00 \end{pmatrix}, \quad \mathbf{R_{x^+}} = \begin{pmatrix} 1{,}00 & & \\ -0{,}57 & 1{,}00 & \\ -0{,}05 & -0{,}79 & 1{,}00 \end{pmatrix},$$

dojdeme ke zjištění, že i když oba vědci zkoumali shodné vzorky, došli k protichůdným závěrům. Například hodnota korelačního koeficientu mezi rostlinnou a živočišnou složkou vyšla vědci A rovna 0,5, zatímco vědec B získal pro ty samé složky hodnotu $-0{,}57$. Obdržené hodnoty korelačních koeficietnů mezi jednotlivými složkami tedy zjevně postrádají smysl a následně i možnost jakékoli interpretace.

Tohoto problému si již v minulosti všimnul Karl Pearson v článku [13]. Pokud nejsou všechny složky kompozic k dispozici, korelace mezi nimi závisí na použitých podkompozicích. Hlavní roli ve znehodnocení výsledků přitom hrála možnost přeškálovat data na konstantní součet, v případě kompozic přitom bez ztráty informace.

Speciálně potom, uvažujeme-li kompozici (x_1, x_2) s konstantním součtem složek $x_1 + x_2 = 1$, dojdeme navíc ke vztahu

$$\mathrm{cov}(x_1, x_2) = \mathrm{cov}(x_1, 1 - x_1) = -\mathrm{cov}(x_1, x_1) = -\mathrm{var}(x_1),$$

tedy kovariance složek se od příslušných rozptylů liší pouze znaménkem. Zřejmě totiž také

$$\mathrm{var}(x_2) = \mathrm{var}(1 - x_1) = \mathrm{var}(x_1).$$

I v tomto případě, když uvažujeme pevný součet složek náhodné kompozice, tedy korelace ztrácí svou funkci charakteristiky závislosti mezi proměnnými. Obecně takto hovoříme o tzv. negativním biasu, kovariance jsou totiž pro $\mathbf{x} = (x_1, \ldots, x_D)'$, $\sum_{i=1}^{D} x_i = 1$, vázané podmínkou

$$\mathrm{cov}(x_1, x_2) + \mathrm{cov}(x_1, x_3) + \cdots + \mathrm{cov}(x_1, x_D) = -\mathrm{var}(x_1).$$

Tento efekt je ostatně pozorovatelný i na korelačních maticích v předchozím příkladu.

3.2 Korelační míra

Ačkoli jsme se věnovali korelacím již ve druhé kapitole, nyní se na ně podíváme z jiného pohledu. Tentokrát je totiž použijeme jako míry závislosti mezi disjunktními skupinami složek náhodné kompozice, z nichž každá je vyjádřena pomocí jedné nebo více souřadnic (bilancí) $z_i,\ i = 1, \ldots, D-1$.

Korelační koeficient tedy nyní vyjadřuje sílu vztahu mezi podíly složek (vyjádřené pomocí bilancí) v obou skupinách. Jak je níže uvedeno, mohou ovšem nastat potíže při interpretaci. Zřejmě budeme uvažovat tři případy:

1. V prvním (triviálním) případě nás zajímá míra lineární závislosti mezi dvěma bilancemi z_i a z_j (pro $1 \leq i, j \leq D-1$), které zastupují dvě dvousložkové podkompozice. Proto použijeme korelační koeficient ρ_{z_i,z_j}. Jeho hodnoty se pohybují v intervalu $\langle -1, 1 \rangle$.

2. Druhým případem je zkoumání korelace mezi jedinou bilancí a skupinou bilancí, která zastupuje vícesložkovou podkompozici. V této situaci uvažujeme koeficient mnohonásobné korelace $\rho_{z_i,\mathbf{z}_k}$ mezi náhodnou veličinou z_i a skupinou náhodných proměnných $\mathbf{z}_k$. Hodnota tohoto koeficientu se realizuje v intervalu $\langle 0, 1\rangle$. Komplikace nastávají právě na tomto místě, obtížnost interpretace výsledku se objevuje proto, že určujeme vztah mezi jednou náhodnou veličinou a skupinou náhodných veličin (bilancí). Abychom mohli vypočítat výsledek, je úkolem určit lineární náhradu této skupiny, což při analýze výsledku ztěžuje situaci.

3. V posledním případě nás zajímá korelační míra mezi dvěma skupinami bilancí. Uvažovat tedy budeme skupinový korelační koeficient $\rho_{\mathbf{z}_k,\mathbf{z}_l}$ pro dva náhodné vektory $\mathbf{z}_k$ a $\mathbf{z}_l$. V tomto kroku musíme interpretovat vztah mezi dvěma lineárnímu náhradami. Při určování lineární náhrady přitom ztrácíme část informace, kterou nám skupina dává. Jestliže takto nastává problém u standardních dat, u kompozičních dat je daná situace ještě obtížnější.

Nyní je uvedena věta, jejíž znalost je klíčová pro možnost relevantního zpracování dat a odpovídající interpretace.

Věta 3.1. *Korelační koeficient, koeficient mnohonásobné korelace a skupinový korelační koeficient jsou invariantní vzhledem k výběru sekvenčního binárního dělení pro reprezentaci daných skupin složek kompozice.*

Důkaz: (dle [6], str. 14) Označme $\mathbf{z}_1$ a $\mathbf{z}_2$ náhodné vektory reprezentující dvě disjunktní skupiny složek (bez vzájemného průniku) pomocí g_1, resp. g_2 bilancí. Jejich varianční matice potom označme $\mathbf{\Sigma}_{\mathbf{z}_1} = \mathrm{var}(\mathbf{z}_1)$ a $\mathbf{\Sigma}_{\mathbf{z}_2} = \mathrm{var}(\mathbf{z}_2)$. Dále nechť $\mathbf{P}_i$, kde $i = 1, 2$, jsou ortonormální matice řádu g_i, tedy $\mathbf{P}_i\mathbf{P}_i' = \mathbf{P}_i'\mathbf{P}_i = \mathbf{I}$, kde $\mathbf{I}$ je jednotková matice.

Ortogonální transformace vektoru $\mathbf{z}_i$, jež odpovídá odlišné reprezentaci skupiny složek pomocí bilancí (dané sekvenčním binárním dělením), je dána vztahem

$$\mathbf{w}_i = \mathbf{P}_i\mathbf{z}_i.$$

Nechť $\mathbf{\Sigma}_{\mathbf{w}_i} = \mathrm{var}(\mathbf{w}_i)$, vypočítejme determinant této varianční matice,

$$|\mathbf{\Sigma}_{\mathbf{w}_i}| = |\mathbf{P}_i\mathbf{\Sigma}_{\mathbf{z}_i}\mathbf{P}_i'| = |\mathbf{P}_i||\mathbf{\Sigma}_{\mathbf{z}_i}||\mathbf{P}_i'| = |\mathbf{\Sigma}_{\mathbf{z}_i}|.$$

Uvedené rovnosti jsme obdrželi ze skutečnosti, že determinanty jednotlivých matic jsou čísly, platí mezi nimi komutativní zákon a mohli jsme tak využít vztahu $\mathbf{P}_i\mathbf{P}_i' = \mathbf{P}_i'\mathbf{P}_i = \mathbf{I}$ a zároveň poznatku, že $|\mathbf{P}_i| = \pm 1$ (vyplývající z ortogonality matice).

Dále budeme uvažovat náhodné vektory $\mathbf{z} = (\mathbf{z}_1', \mathbf{z}_2')'$ a $\mathbf{w} = (\mathbf{w}_1', \mathbf{w}_2')'$ a jejich varianční matice $\mathbf{\Sigma_z} = \mathrm{var}(\mathbf{z})$ a $\mathbf{\Sigma_w} = \mathrm{var}(\mathbf{w})$. Z postupu sekvenčního binárního dělení je jasné, že varianční matice $\mathbf{w}$ je rovna

$$\mathbf{\Sigma_w} = \begin{pmatrix} \mathbf{P}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_2 \end{pmatrix} \mathbf{\Sigma_z} \begin{pmatrix} \mathbf{P}_1' & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_2' \end{pmatrix},$$

přičemž $\mathbf{0}$ značí matici nul příslušných rozměrů. Protože matice

$$\begin{pmatrix} \mathbf{P}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_2 \end{pmatrix}$$

je ortogonální, dostáváme opět rovnost $|\mathbf{\Sigma_w}| = |\mathbf{\Sigma_z}|$. Všechny hodnoty potřebné k výpočtu skupinového korelačního koeficientu tedy zůstávají při odlišné volbě sekvenčního binárního dělení nezměněny.

Speciálně potom z tohoto obecného důkazu obdržíme též analogie pro korelační koeficient, resp. koeficient mnohonásobné korelace.

□

4 Praktické početní příklady

Aplikaci výše uvedené teorie provedeme na dvou příkladech. První je ilustrativní a byl převzat z [6] s tím, že všechny výpočty byly znovu podrobně provedeny. Druhý příklad je vlastní a nebyl dosud nikde publikován.

Výpočetní část práce byla provedena pomocí volně dostupného statistického softwaru R [14], jenž je nápomocen především těm, kteří nemají finanční možnosti k zakoupení drahých komerčních softwarů. Kromě této nesporné výhody nabízí R i možnost spolupodílet se na vytváření dalších knihovem a nástrojů pro zpracování statistických dat a je rovněž poměrně uživatelsky přívětivý. Dnes je však v hojné míře využíván i ve státní a soukromé sféře, kde je plnohodnotnou náhradou již zmíněných soukromých softwarů. Při práci se softwarem jsme využili zejména knihovnu `compositions` [5].

4.1 Ilustrativní příklad - Geologická data

Na prvním příkladu budeme mimo jiné demonstrovat též použití statistického softwaru R. Budeme přitom prezentovat zejména postup, samotnou interpretací výsledků se ještě zabývat nebudeme, blíže viz [6]. V tomto příkladu pracujeme s daty pojmenovanými jako *Kola data*. Ta obsahují informaci o koncetracích více než 50 chemických prvků v 600 vzorcích půdy získaných na poloostrově Kola a dostupných v knihovně softwaru R `StatDA`.

Poloostrov Kola [8] se nachází na severu Ruska, konkrétně v tzv. Murmanské oblasti, sousedí s Barentsovým mořem na severu a s Bílým mořem na výhodě a jihu. Poslední doba ledová zapříčinila odkytí extrémně bohatého naleziště různých rud a minerálů (včetně apatitu, hlidníku, železné rudy a dalších). V letech 1993-1998 byl uskutečněn geologický průzkum Finska, Norska a Ruska. Uvedené vzorky byly odebrány v pěti různých vrstvách. My se zaměříme na svrchní vrstvu půdy, tzv. O-horizon. Pro jeho načtení v softwaru R jsou potřeba následující příkazy:

```
> library(StatDA)
> data(ohorizon)
```

Pokročme k samotnému řešení zadaného problému. Ještě před početním řešením příkladu je nutné uspořádat informace do přehledných tabulek.

V následující tabulce jsou vybrány chemické prvky reprezentující příslušné efekty:

Skupina	**Prvky**
Znečištění (P)	Co, Cu, Ni
Splaveniny (S)	Mg, Na, S
Kontaminace (C)	As, Bi, Cd, Sb
Mineralizace (M)	Ag, Pb
Bioproduktivita (B)	As, Bi, Cd, Sb, Ag, Pb

V tabulce se tedy nachází 12 prvků, jež popíšeme 11 bilancemi. K popisu první skupiny P potřebujeme dvě bilance, stejně jak i v případě S, o jednu bilanci se rozšíří skupina C a pro poslední skupinu M vyžadujeme jedinou bilanci. Skupiny C a M pak reprezentují dohromady skupinu B. Zbývající bilance budou použity na „propojení“ jednotlivých skupin.

bilance	Co	Cu	Ni	Mg	Na	S	As	Bi	Cd	Sb	Ag	Pb
z_1	+	+	+	+	+	+	−	−	−	−	−	−
z_2	+	+	+	−	−	−						
z_3	+	+	−									
z_4	+	−										
z_5				+	+	−						
z_6				+	−							
z_7							+	+	+	+	−	−
z_8							+	+	−	−		
z_9							+	−				
z_{10}									+	−		
z_{11}											+	−

Bilance jsou reprezentovány následovně: pro znečištění z_3, z_4, splaveniny z_5, z_6, kontaminace z_8, z_9, z_{10} a v neposlední řadě mineralizace z_{11}. Skupiny C a M jsou spojeny ve skupinu B pomocí bilance z_7.

Jiná možnost volby bilancí reprezentujících dané skupiny složek je pak uvedena v následující tabulce:

bilance	Co	Cu	Ni	Mg	Na	S	As	Bi	Cd	Sb	Ag	Pb
z_1	+	+	+	−	−	−	−	−	−	−	−	−
z_2	+	+	−									
z_3	+	−										
z_4				+	+	+	−	−	−	−	−	−
z_5				+	+	−						
z_6				+	−							
z_7							+	+	+	+	−	−
z_8							+	+	+	−		
z_9							+	+	−			
z_{10}							+	−				
z_{11}											+	−

V další fázi přikročme k práci se softwarem R. Při zpracování kompozičních dat je nezbytné nahrát knihovnu `compositions`. Tuto operaci provedeme pomocí příkazu

```
> library(compositions)
```

Všechna vybraná pozorování nyní spojíme do jedné matice dat, to vše pomocí příkazu

```
> data_kolaset=ohorizon[,(c("Co","Cu","Ni","Mg","Na","S",
"As","Bi","Cd","Sb","Ag","Pb"),ncol=12)]
```

Postupme k vytvoření kompozičních dat (jako třídy `acomp`):

```
> comps<-acomp(data_kolaset)
```

Pro vytvoření sekvenčního binárního dělení dle výše uvedené (první) tabulky musíme přistoupit k následujícímu:

```
> signary=cbind(
c(1,1,1,1,1,1,-1,-1,-1,-1,-1,-1),c(1,1,1,-1,-1,-1,0,0,0,0,0,0),
c(1,1,-1,0,0,0,0,0,0,0,0,0),c(1,-1,0,0,0,0,0,0,0,0,0,0),
c(0,0,0,1,1,-1,0,0,0,0,0,0),c(0,0,0,1,-1,0,0,0,0,0,0,0),
c(0,0,0,0,0,0,1,1,1,1,-1,-1),c(0,0,0,0,0,0,1,1,-1,-1,0,0),
c(0,0,0,0,0,0,1,-1,0,0,0,0),c(0,0,0,0,0,0,0,0,0,1,-1,0,0),
c(0,0,0,0,0,0,0,0,0,0,1,-1))
```

Přitom dělení složek do jednotlivých skupin odpovídá vzájemným vztahům mezi sobě blízkými složkami. V závěru první fáze ještě aplikujeme *ilr* transformaci, tedy vytvoříme příslušné souřadnice, následně vypočteme varianční matici souřadnic:

```
> base=gsi.buildilrBase(signary)
> z=ilr(comps,V=base)
> S=var(z)
```

Výsledky korelací jednotlivých skupin složek pomocí bilancí jsou uvedeny v následující tabulce. Výsledky se (s využitím Věty 3.1) nezmění ani v případě, že použijeme jiné sekvenční binární dělení.

	skup.kor.koef.		
skupina	P	S	B
P	–	0,244	0,590
S	–	–	0,460
B	–	–	–

V závěru tohoto příkladu ještě uveďme příkazy k výpočtu skupinového korelačního koeficientu mezi skupinou P a S, zbylé byly vypočteny analogicky.

```
> K=S[3:4,3:4]
> L=S[5:6,5:6]
> sigma=S[3:6,3:6]
> rhoPS=sqrt(1-det(sigma)/(det(K)*det(L)))
```

První a druhý řádek vytvoří varianční matice pro vektor $(z_3, z_4)'$, resp. $(z_5, z_6)'$. Stejně vytvoříme i varianční matici pro $(z_3, z_4, z_5, z_6)'$. V posledním řádku je uveden postup pro vypočtení skupinového korelačního koeficientu. Jen pro úplnost dodejme, že příkaz `det` je určen k vypočtení determinantu matice.

4.2 Vlastní příklad - Poslechovost rádií v krajích

Druhý příklad pochází z mediální oblasti. Tentokrát však nezamíříme do televizního světa, nýbrž budeme zkoumat rozhlasový éter. Nejprve krátce osvětlíme, jak se data poslechovosti v tomto případě získávají.

Průzkum poslechovosti provádí společnosti STEM/MARK a MEDIAN u populace ČR ve věkové kategorii 12-79 let. Dotazování probíhá sedm dní v týdnu (8-21 hodin). Výstupy poslechovosti jsou zveřejňovány čtyřikrát ročně. Jeden tento výstup označujeme jako *vlnu poslechovosti*, v každé vlně získáváme data za uplynulý půlrok.

Pravidelné výsledky RadioProjektu jsou určeny především pro potřeby komerčních (ale i veřejnoprávních) subjektů. Získat tato data není takový problém jako v případě sledovanosti televizí, v tomto případě si sami provozovatelé rozhodují, zda výsledky zveřejní či nikoliv. Data poslechovosti prezentují především zpravodajské servery zabývající se mediální problematikou, v České republice se jí nejšířeji věnuje server `RadioTV.cz` [15], na jehož webové prezentaci jsou dostupná data z uplynulých dvou let u všech rádií (konkrétně na `radiotv.cz/poslechovost`) a zároveň jsou přístupná podrobná data ve zpravodajském archivu.

V tabulce jsou uvedena data týdenní poslechovosti. Týdenní poslechovost udává počet lidí, kteří alespoň jednou týdně ve zkoumaném období stanici naladili. Data tohoto měření poslechovosti RadioProjektu pochází z období 1. dubna - 3. září 2008.

	veřejnoprávní		komerční			
			rádia pro mladé		střední věk	
kraj	ČRo 1	ČRo 2	E2	Kiss	F1	Hitrádio
Praha	220	100	259	72	152	110
Středočeský	133	104	203	111	160	31
Jihočeský	67	30	105	54	104	163
Plzeňský	64	31	99	58	85	10
Karlovarský	29	14	83	2	47	82
Ústecký	71	39	140	51	134	94
Liberecký	73	40	98	12	74	33
Královéhradecký	61	47	99	10	87	98
Pardubický	61	37	103	40	87	14
Vysočina	49	34	74	30	90	160
Jihomoravský	147	78	177	166	221	20
Olomoucký	67	33	104	20	139	63
Moravskoslezský	164	60	148	146	252	194
Zlínský	103	34	47	65	118	69
průměr	93,50	48,64	124,21	59,79	125,00	81,50

Bez ztráty informace si nyní výše uvedená data uvedeme v procentuálních podílech. Důvodem je lepší možnost následné interpretace.

	veřejnoprávní		komerční			
			rádia pro mladé		střední věk	
kraj	ČRo 1	ČRo 2	E2	Kiss	F1	Hitrádio
Praha	24,10	10,95	28,37	7,89	16,65	12,05
Středočeský	17,92	14,02	27,36	14,96	21,56	4,18
Jihočeský	12,81	5,74	20,08	10,33	19,89	31,17
Plzeňský	18,44	8,93	28,53	16,71	24,50	2,88
Karlovarský	11,28	5,45	32,30	0,78	18,29	31,91
Ústecký	13,42	7,37	26,47	9,64	25,33	17,77
Liberecký	22,12	12,12	29,70	3,64	22,42	10,00
Královéhradecký	15,17	11,69	24,63	2,49	21,64	24,38
Pardubický	17,84	10,82	30,12	11,70	25,44	4,09
Vysočina	11,21	7,78	16,93	6,86	20,59	36,61
Jihomoravský	18,17	9,64	21,88	20,52	27,32	2,47
Olomoucký	15,73	7,75	24,41	4,69	32,63	14,79
Moravskoslezský	17,01	6,22	15,35	15,15	26,14	20,12
Zlínský	23,62	7,80	10,78	14,91	27,06	15,83

Abychom mohli následně přistoupit ke korelační analýze, provedeme opět nejprve sekvenční binární dělení. Rozklad vychází z předchozí tabulky. První dvě rádia jsou veřejnoprávní, nejsou tedy primárně financována z výnosů reklamy jako komerční subjekty, základním zdrojem příjmů jsou koncesionářské poplatky. Druhá skupina pak zahrnuje komerční subjekty, u nichž naopak reklamní příjmy představují stěžejní část rozpočtu. Komerční rádia jsou zde rozdělena na rádia pro mladou generaci a střední věkovou skupinu. Označení jednotlivých skupin vychází ze [10], podívejme se na ně podrobněji.

První z nich je ve stručnosti označena jako skupina rádií pro mladé (zde Evropa 2 a Kiss Radia). Označit tuto skupinu stanic bychom mohli jako Contemporary Hits Radio (CHR), cílovou skupinou jsou lidé ve věku 14 - 24 let. Ačkoli oficiálně tuzemská rádia tento přídomek nemají, právě Evropu 2 a Kiss Radia takto označit můžeme. Jedná se o agresivní hudební formát zaměřený na aktuální hitparádové hity. Playlist je velmi úzký a rotace ve vysílání vysoké. Skladby zůstávají v playlistu obvykle 3-6 měsíců, jen výjimečně déle.

Druhá skupina (Frekvence 1, Hitrádia) míří na střední generaci, zpravidla na skupinu posluchačů ve věku 25 - 50 let, tento formát nese označení Adult Contemporary (AC). Jedná se o nejmasovější nenáročný formát hrající popovou hudbu, mírně se mohou objevit náznaky softrocku či lehké taneční hudby. Playlisty jsou postaveny na největších hitech z posledních dvou až tří dekád.

Poslední skupinu tvoří veřejnoprávní stanice (Český rozhlas 1 - Radiožurnál a Český rozhlas 2 - Praha). Zde nelze přímo hovořit o konkrétním formátu, Ra-

diožurnál primárně sice hraje taktéž formát AC, přesto ve struktuře posluchačů obě stanice míří spíše na starší generaci. Největší podíl posluchačů u druhého programu České rozhlasu tvoří lidé od 50 do 79 let.

Popis dělení jako obvykle shrnuje tabulka

bilance	x_1	x_2	x_3	x_4	x_5	x_6
z_1	+	+	−	−	−	−
z_2	+	−				
z_3			+	+	−	−
z_4			+	−		
z_5					+	−

Nyní přichází ke slovu statistický software `R`. Data do softwaru načteme ve formě textového souboru `poslechovost`. Dělení a bázi vytvoříme pomocí následujících příkazů, které byly popsány v minulém příkladu.

```
> signary=cbind(
c(1,1,-1,-1,-1,-1),c(1,-1,0,0,0,0),c(0,0,1,1,-1,-1),
c(0,0,1,-1,0,0),c(0,0,0,0,1,-1))
>base=gsi.buildilrBase(signary)
>z=ilr(comps,V=base)
```

Po provedení výše uvedených příkazů můžeme pracovat s transformovanými daty získanými dle požadavků. V tabulce jsou uvedeny absolutní hodnoty korelačních koeficientů, jež má smysl uvažovat. Uvedené výsledky jsou zaokrouhleny na tři desetinná místa.

$\rho_{z_2,z_4} = 0{,}537$	$\rho_{z_2,z_5} = 0{,}045$	$\rho_{z_4,z_5} = 0{,}140$	$\rho_{z_2,(z_3,z_4,z_5)} = 0{,}744$

Přistupme k interpretaci získaných výsledků. Korelační koeficient $\rho_{z_2,z_4} = 0{,}537$ (tedy závislost mezi veřejnoprávními rádii a rádii pro mladé) značí střední závislost mezi oběma skupinami, jedná se o druhou nejsilnější závislost ze všech zkoumaných. Výsledek je poměrně překvapivý, protože bychom neočekávali, že nejsilnější závislost existuje mezi dvěma (na první pohled) nejodlišnějšími skupinami, konkrétně tedy rádii veřejnoprávními a rádii určeným mladším ročníkům (veřejnoprávní rozhlas poslouchají zpravidla lidé od věku 40 let). Posluchačskou obec Českého rozhlasu tvoří zejména vysokoškolsky vzdělaní lidé a také starší generace, zatímco stanice Evropa 2 a Kiss rádia svým zaměřením cílí na velmi mladé dospívající jedince.

V případě zkoumání závislosti mezi $\rho_{z_2,z_5} = 0{,}045$, dojdeme k zjištění, že mezi skupinami reprezentujícími veřejnoprávní rádia a rádia určená střední generaci

neexistuje prakticky žádná závislost. To vše i přes fakt, že tyto dvě skupiny mají k sobě svým zaměřením blíže než výše uvedené skupiny. Zde pozorujeme nejnižší míru závislosti.

Korelační koeficient ρ_{z_4,z_5} je roven pouze 0,140, tudíž mezi rádii pro mladou a střední generaci je jen nepatrná lineární závislost.

Z uvažovaných vztahů se v posledním kroku zaměřme ještě na vztah mezi bilancí z_2 (veřejnoprávní stanice) a skupinou bilancí z_3, z_4, z_5 (komerční stanice). Vychází opět celkem překvapivý výsledek, tento korelační koeficient nám totiž vyjadřuje nejsilnější míru závislosti mezi všemi uvažovanými skupinami. Obě skupiny existují vedle sebe a doplňují poptávku posluchačů.

Podívejme se nyní, jak dopadne interpretace, zaměříme-li se na matici rozptylů,

$$\mathbf{T} = \begin{pmatrix} 0 & 0{,}031 & 0{,}086 & 0{,}428 & 0{,}258 & 0{,}578 \\ 0{,}031 & 0 & 0{,}059 & 0{,}557 & 0{,}293 & 0{,}624 \\ 0{,}086 & 0{,}059 & 0 & 0{,}669 & 0{,}172 & 0{,}562 \\ 0{,}428 & 0{,}557 & 0{,}669 & 0 & 0{,}945 & 1{,}180 \\ 0{,}258 & 0{,}293 & 0{,}172 & 0{,}945 & 0 & 0{,}466 \\ 0{,}578 & 0{,}624 & 0{,}562 & 1{,}180 & 0{,}466 & 0 \end{pmatrix}.$$

Pro naše zkoumání bude však vhodná spíše „normovaná“ matice s hodnotami v intervalu $\langle 0, 1 \rangle$,

$$\mathrm{e}^{-\mathbf{T}} = \begin{pmatrix} 1 & 0{,}969 & 0{,}918 & 0{,}652 & 0{,}772 & 0{,}561 \\ 0{,}969 & 1 & 0{,}943 & 0{,}573 & 0{,}746 & 0{,}536 \\ 0{,}918 & 0{,}943 & 1 & 0{,}512 & 0{,}842 & 0{,}570 \\ 0{,}652 & 0{,}573 & 0{,}512 & 1 & 0{,}389 & 0{,}307 \\ 0{,}772 & 0{,}746 & 0{,}842 & 0{,}389 & 1 & 0{,}627 \\ 0{,}561 & 0{,}536 & 0{,}570 & 0{,}307 & 0{,}627 & 1 \end{pmatrix}.$$

Všechny výše uvedené hodnoty reprezentují variabilitu podílů mezi dvěma složkami. V případě, že se hodnota blíží jedné, můžeme hovořit o jejich stabilitě. Výsledky můžeme interpretovat zejména z regionálního hlediska. Velkou část výsledných dat lze hodnotit tak, že podíly napříč všemi kraji v České republice jsou spíše nestabilní. Jelikož se v případě koeficientů $\exp(-t_{ij})$ jedná o hodnoty exponenciální funkce, o stabilitě můžeme hovořit snad jen u hodnot velmi blízkých jedné, řekněme od hranice 0,8.

Nejnižší stabilitu podílů složek jednoznačně vykazují dvojice (Kiss, F1) a (Kiss, Hitrádio). Z výsledků tedy vyplývá, že hodnoty napříč jednotlivými regiony jsou silně rozkolísané, nejvíce mezi všemi sledovanými dvojicemi subjektů.

O těsné závislosti lze hovořit v případě podílů složek ČRo 1 - Radiožurnál a ČRo 2 - Praha, o něco slabší vztah existuje mezi druhým veřejnoprávním okruhem a rádiem Evropa 2, respektive mezi dvojicí ČRo 1 - Radiožurnál a Evropa 2. Sledujeme-li hodnoty napříč regiony, skutečně napozorujeme poměrně stabilní hodnoty podílů v porovnání s ostatními.

Data však nemusíme interpretovat jen jako porovnání mezi dvěma subjekty. Jednotlivé vektory hodnot v řádku (sloupci) u příslušných rozhlasových stanic vypovídají o stabilitě podílu daného rádia na trhu. V tomto ohledu pozorujeme nestabilnější podíl v případě Radiožurnálu. Při pohledu na vstupní data poměrně překvapivě stabilní podíl na rozhlasovém trhu vykazuje též ČRo 2 - Praha a Evropa 2. I zde vidíme hodnoty variability jednotlivých podílů složek blízké jedné, navíc se všechny příslušné hodnoty v řádku liší jen velmi málo (ovšem se dvěma výjimkami). O vyrovnaném (ale již méně stabilním) podílu můžeme hovořit ještě u rádia Frekvence 1.

Celkově lze říci, že stabilnější hodnoty podílů na trhu vykazují obecně celoplošné stanice. Nakonec podíváme-li se blíže na síťová rádia (stanice Kiss a Hitrádia), je stabilita spíše menší. Nizké hodnoty prvků matice $e^{-\mathbf{T}}$ v řádku u rádií Kiss, resp. Hitrádií, mohou prezentovat rozkolísanou a nestabilní pozici, stejné závěry lze učinit i tehdy, pokud sledujeme dlouhodobější vývoj poslechovosti.

Z příkladu se nám jeví, že bližší našim očekáváním i současným možnostem interpretace jsou hodnoty obdržené z matice rozptylů. Problém ovšem může být i z důvodu potřeby komplexní interpretace bilancí, které vstupují do jednotlivých korelačních koeficientů. Je zřejmé, že v této oblasti je ještě třeba dalšího intenzivního výzkumu.

Závěr

Alternativní přístup ke korelační analýze, resp. analýze vztahů mezi proměnnými vůbec, se v případě kompozičních dat jeví jako rozumný přístup pro práci s pozorováními tohoto typu. Je však třeba mít na paměti, že obor zabývající se kompozičními daty se stále teprve rozvíjí a v případě některých interpretací tak existují občasné spory i mezi největšími odborníky.

Prakticky nejobtížnějším úkolem se stalo počítání příkladů a především, v případě toho druhého, následná interpretace. Na druhou stranu se samotným softwarem R pracuji relativně často, proto jsem s ním velké komplikace neměl. Když mi byly poskytnuty materiály ke knihovně, vztahující se ke kompozičním datům, celkem snadno jsem si její znalost osvojil a nečinilo mi další potíže s ní pracovat.

Tvorba této práce mne naučila mnohému. Kromě samotných kompozičních dat jsem byl donucen studovat ze zahraniční literatury (především anglické), což je neocenitelný krok do budoucna. Navíc, jak se říká, nejlépe si člověk znalosti vštípí v případě, že informace sám aktivně vyhledává a pracuje s nimi.

Reference

[1] Aitchison, J. a kol., *Logratio Analysis and Composition Distance*, Mathematical Geology, **3** (32), 273 (2000).

[2] Aitchison, J., *The Statistical Analysis Of Compositional Data*, London: Chapman & Hall, 1986.

[3] Anděl, J., *Základy matematické statistiky*, 2. vydání, Praha: MATFYZPRESS, 2007.

[4] Anděl, J., *Matematická statistika*, 2. vydání, Praha: SNTL/ALFA, 1985.

[5] van der Boogaart, K. G., *Using the R package "compositions"* [online], dostupné z http://www.stat.boogaart.de/compositions/UsingCompositions.pdf, [citováno 11.10.2009].

[6] Filzmoser, P., Hron, K., *Correlation analysis for compositional data* [online], dostupné z http://www.statistik.tuwien.ac.at/forschung/SM/SM-2008-2complete.pdf, [citováno 5.4.2009].

[7] Harville, David A., *Matric algebra from statistician's perspective*, 1. vydání, New York: Springer, 1997.

[8] Kola Peninsula [online], dostupné z http://en.wikipedia.org/wiki/Kola_Peninsula, [citováno 11.10.2009].

[9] Kunderová, P., *Základy pravděpodobnosti a matematické statistiky*, 1. vydání, Univerzita Palackého v Olomouci, 2004.

[10] *Malý lexikon rozhlasových formátů* [online], dostupné z http://www.radiotv.cz/p_radio/r_program/mal-lexikon-rozhlasovch-formt/, [citováno 20.3.2009].

[11] Pawlowsky-Glahn, V., *The Aitchison geometry of the simplex and the statistical analysis of compositional data* [online], dostupné z http://www.statistik.tuwien.ac.at/public/filz/Pawlowsky2009Vienna.pdf, [citováno 7.1.2010].

[12] Pawlowsky-Glahn, V., Egozcue J.J., *Groups of parts and their balances in compositional data analysis*, Mathematical Geology, **7** (37), 796-801 (2005).

[13] Pawlowsky-Glahn, V., Egozcue, J.J., Tolasana-Delgado R., *Lecture Notes on Compositional Data Analysis*, The University of Girona, 2007 [online], dostupné z http://www.sediment.uni-goettingen.de/staff/tolosana/extra/CoDa.pdf [citováno 5.4.2008].

[14] *The R Project for Statistical Computing* [online], dostupné z http://www.r-project.org/ [citováno 5.4.2008].

[15] *RadioProjekt II. - III.Q 2008: Několik rádií dosáhlo dlouhodobá maxima* [online], dostupné z http://www.radiotv.cz/p_radio/r_obchod/radioprojekt-ii-iii-q-2008-nkolik-rdi-doshlo-dlouhodob-maxima/, [citováno 20.3.2009].