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

BAKALÁŘSKÁ PRÁCE

Diskriminační analýza pro kompoziční data

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

Vypracovala:
Šárka Brodinová
ME, III. ročník

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci sepsala samostatně pod vedením pana RNDr. Karla Hrona, Ph.D. a že jsem v seznamu literatury uvedla všechny zdroje, které jsem použila při zpracování práce.

V Olomouci dne 12. dubna 2012

Poděkování

Na tomto místě bych chtěla poděkovat především svému vedoucímu bakalářské práce panu RNDr. Karlu Hronovi, Ph.D., že měl se mnou dostatek trpělivosti, aby mi pomohl dovést tuto práci ke zdárnému konci. Mé díky patří tvůrcům typografického progamu LaTeX a statistického softwaru R. Také bych ráda poděkovala své rodině, přátelům a spolužákům, kteří mě po celou dobu studia podporovali.

Obsah

Úvod

Vývoj matematické a statistické disciplíny jde v ideálním případě ruku v ruce s vývojem dalších vědních oborů. Platí to jak pro diskriminační analýzu, mnohorozměrná statistická metoda pro klasifikaci objektů, tak i u logratio analýzy kompozičních dat (kompozic), pozorování nesoucí pouze relativní informaci. Obě metody našly široké uplatnění v mnoha oblastech jako geologie, ekonomika, paleontologie a další.

Pro svou bakalářskou práci jsem si vybrala téma diskriminační analýza pro kompoziční data. První kapitola je věnována tzv. bayesovské diskriminační analýze pro standardní (nekompoziční) data. Úkolem kapitoly je podrobně seznámit čtenáře se zmíněnou problematikou tak, aby ji dobře porozumněl. Je zde rozebrán princip tvorby klasifikačního pravidla mezi dvěma (resp. více) skupinami objektů. Navíc dle charakteristik jednotlivých skupin objektů se rozlišuje lineární a kvadratická diskriminační analýza. Na závěr je popsána metoda sloužící k vyhodnocení správnosti diskriminace.

V druhé kapitole se zabývám kompozičními daty, která jsou stěžejní pro tuto práci. Mým cílem bylo, aby čtenář pochopil rozdíl mezi kompozičními a standardními daty. V této kapitole jsou uvedeny hlavní analytické nástroje, které jsou důležité pro práci s daty nesoucími relativní informaci a srozumitelnou interpretaci dosažených výsledků.

Následující kapitola popisuje aplikaci diskriminační analýzy na kompoziční data. Je zde využito poznatků z první i druhé kapitoly. Pro lepší pochopení problematiky jsou příslušná tvrzení dokázána.

Závěrečná kapitola vás seznámí s interpretací diskriminační analýzy na konkretních reálných datech. Daná problematika byla použita na vlastním příkladu z oblasti osobních financí. Pro srovnání byla na datech aplikována jak lineární diskriminace, tak i kvadratická. Výpočty byly prováděny ve statistickém softwaru R.

Cílem bakalářské práce je seznámit čtenáře, který již absolvoval základní kurz matematické statistiky, s daným tématem tak, aby pochopil jak teoretické aspekty diskriminační analýzy kompozic, tak její aplikaci na reálných datech. Práce byla vysázena v typografickém softwaru LaTeX.

1 Diskriminační analýza

1.1 Princip a hlavní cíle diskriminační analýzy

Diskriminační analýza je metoda pracující s mnohorozměrnými daty. Cílem diskriminační analýzy je zařadit různé objekty do skupin na základě jejich vlastností, které jsou popsány diskriminační funkcí nebo zařadit nový objekt do předem známé skupiny pomocí klasifikačního pravidla. Mezi oběma cíli existuje přímý vztah. Diskriminační funkci lze použít jednak ke třídění objektů, ale také k zařazení nového objektu do předem známých skupin. Třídění objektů se provádí pomocí tzv. bayesovských (pracující se známými apriorními pravděpodobnostmi) nebo fisherových (využívající projekci na směr největšího rozdílu mezi skupinami pozorování) pravidel. V tomto textu se omezíme pouze na bayesovský přístup [6, 8].

Jelikož má diskriminační analýza široké uplatnění (v medicíně, v technických oborech, v bankovnictví atd.), není těžké si uvést několik příkladů z praxe:

1. Představme si situaci, kdy jdeme k praktickému lékaři. V čekárně je několik pacientů, kteří postupně chodí do ordinace a lékař jim na základě jejich příznaků určí diagnózu a následně je pošle na specializované oddělení. Tímto způsobem lékař během dne pošle pacienta na neurologii nebo na chirurgii či na urologii atd. Postupně tvoří skupiny pacientů se stejnými příznaky na základě předešlých zkušeností. Matematicky řečeno, lékař třídí pacienty na základě vlastností (příznaků) do skupin za pomoci klasifikačního pravidla (utvořeného z předešlých zkušeností). Při příchodu každého dalšího pacienta dokáže kvalifikovaněji určit jeho (skutečnou) diagnózu.

2. Součástí výrobních procesů je kontrola jakosti výrobků. Kontrola je prováděna v několika fázích. Nejprve se provede měření kvantitativních proměnných (hmotnost, rozměry, tvrdost atd.) na výběrovém souboru výrobků. Poté tento soubor podrobíme zátěžovým zkouškám, abychom zjistili, které výrobky obstály. K tomu abychom mohli předpovědět, zda je nově vyrobený výrobek kvalitní či nikoliv, stačí pouze změřit kvantitativní proměnné. Provádíme diskriminaci na zá-

kladě vlastností (naměřených hodnot) za pomoci klasifikačního pravidla (utvořeného z předešlých naměřených hodnot) do jedné skupiny výrobků bez vady a do druhé s vadou.

1.2 Klasifikační pravidlo

Pro lepší pochopení konstrukce klasifikačního pravidla zúžíme třídění objektů do dvou tříd [2, 6].

Mějme skupinu objektů, každý z nich je charakterizován náhodným vektorem $\mathbf{X} = (X_1, \ldots, X_p)^\top$. Úkolem je rozdělit objekty do dvou tříd π_1 a π_2. Nechť soubor objektů první třídy je populace $\mathbf{x}$-ových hodnot z π_1 a obdobně hodnoty $\mathbf{x}$ druhé třídy tvoří populaci π_2. Přitom předpokládáme, že objekt patřící do π_1, (resp. π_2) se řídí spojitým rozdělením pravděpodobnosti s hustotou $f_1(\mathbf{x})$, (resp. $f_2(\mathbf{x})$). Označme Ω výběrový prostor objektů, které klasifikujeme. Z toho logicky vyplývá, že každý objekt $\mathbf{x}$ patří buď do třídy π_1, nebo do třídy π_2. Při třídění se ovšem může stát, že objekt, který ve skutečnosti patří do třídy π_1, chybně zařadíme do π_2. Přitom označme R_1 množinu všech objektů klasifikovaných do třídy π_1. Množina R_2 obsahuje všechny objekty zařazené do třídy π_2. Z čehož plyne, že sjednocením R_1 a R_2 získáme množinu všech objektů Ω, nebo-li R_1 a R_2 tvoří rozklad množiny Ω. Úkolem diskriminační analýzy je najít optimální rozklad.

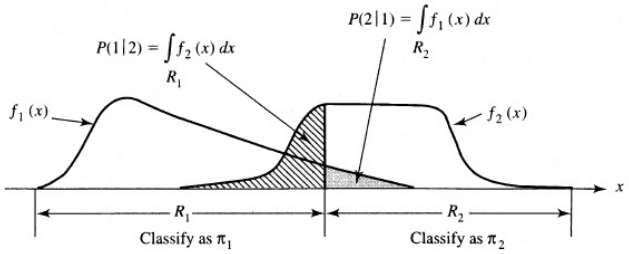
Známe-li hustoty $f_1(\mathbf{x})$ a $f_2(\mathbf{x})$, potom pravděpodobnost špatného zařazení objektů patřících do třídy π_1 je podmíněná pravděpodobnost

$$P(2|1) = P(\mathbf{X} \in R_2|\pi_1) = \int_{R_2=\Omega-R_2} f_1(\mathbf{x})d\mathbf{x},$$

nebo naopak může nastat situace, kdy o objektu patřícím do třídy π_2 rozhodneme, že patří do třídy π_1. Odpovídající podmíněná pravděpodobnost je

$$P(1|2) = P(\mathbf{X} \in R_1|\pi_2) = \int_{R_1} f_2(\mathbf{x})d\mathbf{x}.$$

Výše popsanou situaci znázorňuje obrázek 1.



Obrázek 1: Pravděpodobnosti při špatné klasifikaci [6]

Nechť p_1 je apriorní pravděpodobnost, která říká, že objekt pochází z třídy π_1, p_2 pro objekt pocházející z třídy π_2. Při splnění podmínky $p_1 + p_2 = 1$ zavádíme následující podmíněné pravděpodobnosti:

$$P\text{ (správně zařazený objekt do } \pi_1) = P(\mathbf{X} \in R_1|\pi_1)P(\pi_1) = P(1|1)p_1,$$
$$P\text{ (chybně zařazený objekt do } \pi_1) = P(\mathbf{X} \in R_1|\pi_2)P(\pi_2) = P(1|2)p_2,$$
$$P\text{ (správně zařazený objekt do } \pi_2) = P(\mathbf{X} \in R_2|\pi_2)P(\pi_2) = P(2|2)p_2,$$
$$P\text{ (chybně zařazený objekt do } \pi_2) = P(\mathbf{X} \in R_2|\pi_1)P(\pi_1) = P(2|1)p_1.$$

Špatné zařazení objektů pro nás představuje ztrátu. Logicky nulová ztráta nastane při bezchybné klasifikaci. Označme $z(1|2)$ ztrátu, kterou dostaneme, když objekt z třídy π_2 klasifikujeme do třídy π_1, a naopak $z(2|1)$. Potom vztah

$$L = z(2|1)P(2|1)p_1 + z(1|2)P(1|2)p_2 \tag{1}$$

nazveme střední (očekávaná) hodnota ztráty. Optimální rozklad množiny Ω je takový, pro který bude (1) minimální. Tomuto požadavku odpovídá R_1 jako množina takových $\mathbf{x}$, pro která platí

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{z(1|2)}{z(2|1)}\right)\left(\frac{p_2}{p_1}\right), \tag{2}$$

viz [3], str. 208 - 213.

Obdobně pro objekt z množiny R_2 platí

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{z(1|2)}{z(2|1)}\right)\left(\frac{p_2}{p_1}\right). \tag{3}$$

Z (2) a (3) lze odvodit následující speciální případy:

- jestliže $p_1 = p_2$

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{z(1|2)}{z(2|1)} \qquad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{z(1|2)}{z(2|1)}$$

- jestliže $z(1|2) = z(2|1)$

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_2}{p_1} \qquad R_2 : \frac{f_1\mathbf{x}}{f_2\mathbf{x}} < \frac{p_2}{p_1}$$

- jestliže $z(1|2) = z(2|1)$ a $p_1 = p_2$

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1 \qquad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 1$$

Poznámka 1.1. *Lze vytvořit i jiná kritéria, na jejichž základě se dá zkonstruovat klasifikační pravidlo zohledňující apriorní pravděpodobnosti. Jedno takové kritérium pro volbu R_1 a R_2 vychází z celkové pravděpodobnosti mylné klasifikace, která je dána vztahem*

$$Q = p_1 \int_{R_2} f_1(\mathbf{x}) d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x}) d\mathbf{x} = p_1 P(2|1) + p_2 P(1|2). \tag{4}$$

Můžeme vidět, že hodnota vztahu (4) je totožná s hodnotou v (1) za předpokladu, že jsou ztráty jednotkové. Navíc, jsou-li ztráty stejné, je minimalizace vztahu (1) ekvivalentní právě s minimalizací (4).

1.3 Diskriminace mezi dvěma skupinami

Nyní se omezíme na případ mnohorozměrného normálního rozdělení v obou třídách (skupinách) objektů [6, 8]. Než přistoupíme k samotné tvorbě diskriminačního pravidla, připomeňme si definici tohoto rozdělení [2].

Definice 1.1. *Nechť je dán p-složkový číselný vektor $\boldsymbol{\mu}$ a $p \times p$ pozitivně definitní číslená matice $\boldsymbol{\Sigma}$. Řekneme, že p-rozměrný náhodný vektor $\mathbf{X}$ má normální rozdělení s parametry $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$, značíme $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, jestliže pro libovolný p-složkový*

číselný vektor $\mathbf{c}$ *má náhodná veličina* $\mathbf{c}^\top\mathbf{X}$ *normální rozdělení s parametry* $\mathbf{c}^\top\boldsymbol{\mu}$ *a* $\mathbf{c}^\top\boldsymbol{\Sigma}^\top\mathbf{c}$*, t.j.* $\mathbf{c}^\top\mathbf{X} \sim N(\mathbf{c}^\top\boldsymbol{\mu},\mathbf{c}^\top\boldsymbol{\Sigma}^\top\mathbf{c})$.

Obecně v případě k skupin, funkce $f_i(\mathbf{x})$ představuje hustotu mnohorozměrného normálního rozdělení se známými parametry $\boldsymbol{\mu}_i$, což značí střední hodnotu, a varianční maticí $\boldsymbol{\Sigma}_i, i = 1, \ldots, k$. Tato funkce je dána vztahem

$$f_i(\mathbf{x}) = (2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\}, \qquad i = 1, \ldots, k. \quad (5)$$

V případě, že je populace rozdělena do dvou skupin, budeme uvažovat pouze funkce $f_1(\mathbf{x})$ a $f_2(\mathbf{x})$ se středními hodnotami $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ a variančními maticemi $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$. V následujících podkapitolách se budeme věnovat speciálním případům, kdy se varianční matice mohou a nemusí rovnat.

K ověření shodnosti variančních matic, na základě informace z náhodných výběrů z obou rozdělení, použijeme Boxův test. Testujeme hypotézu o stejných variančních maticích oproti alternativní hypotéze, že aspoň jedna matice je odlišná. K dispozici jsou výběry o rozsazích n_i s výběrovými variančními maticemi $S_i = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)^\top$, kde $\bar{\mathbf{x}}_i$ značí výběrový průměr i-tého výběru (podrobněji v kapitole 1.4). Za pomoci testovací statistiky

$$M = \frac{\prod_{i=1}^{k}|\mathbf{S}_i|^{\frac{n_i-1}{2}}}{|\bar{\mathbf{S}}|^{\frac{n-i}{2}}} \quad (6)$$

a veličiny $V = -(1-c_1)\ln M$, která má za platnosti nulové hypotézy přibližně χ^2 rozdělení o $p(p+1)(K-1)/2$ stupních volnosti, rozhodneme o platnosti nulové hypotézy. Tu na hladině α zamítáme ve prospěch alternativy, pokud hodnota veličiny V překročí příslušný α-kvantil. Přitom

$$\bar{\mathbf{S}} = \frac{\sum_{i=1}^{k}\mathbf{S}_i(n_i-1)}{n-k}, \qquad c_1 = \frac{2p^2+3p-1}{6(p+1)(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i-1} - \frac{1}{n-k}\right). \quad (7)$$

Boxova statistika je vhodná pro počet skupin menší než 5 a pro výběrové soubory mající aspoň 20 objektů.

1.3.1 Lineární diskriminační analýza (LDA)

O lineární diskriminační analýze hovoříme v situaci, kdy se populace dělí do dvou tříd, π_1 nebo π_2, a náhodný vektor $\mathbf{X} = (X_1, \ldots, X_n)^\top$ má vícerozměrné normální rozdělení se středními hodnotami $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ a shodnými variančními maticemi $\boldsymbol{\Sigma}$, tedy platí $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. Funkce hustoty je stejná jako v (5) s uvedeným rozdílem $i = 1, 2$.

Jestliže známe parametry $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ a $\boldsymbol{\Sigma}$, pak podle (2) nejmenší střední hodnota ztráty při zařazení objektů do první skupiny nastane, jestliže

$$R_1 : \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right\} \geq \left(\frac{z(1|2)}{z(2|1)}\right)\left(\frac{p_2}{p_1}\right). \tag{8}$$

Při zařazení do druhé skupiny hovoříme o nejmenší střední hodnotě ztráty podle (3), pokud

$$R_2 : \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right\} < \left(\frac{z(1|2)}{z(2|1)}\right)\left(\frac{p_2}{p_1}\right). \tag{9}$$

Nerovnosti (8) a (9) lze považovat za klasifikační pravidlo. Tvrzení vychází z následující věty.

Věta 1.1. *Nechť π_1 a π_2 jsou populace, které mají normální rozdělení se stejnými variančními maticemi $\boldsymbol{\Sigma}$. Pak klasifikační pravidlo minimalizující očekávanou hodnotu ztráty, tzv. lineární diskriminační funkce, zní:*
Nové pozorování (objekt $\mathbf{x}_0$) zařadíme do třídy π_1, když

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \ln\left(\frac{z(1|2)}{z(2|1)}\frac{p_2}{p_1}\right), \tag{10}$$

jinak klasifikujeme do třídy π_2.

Důkaz: [6], str. 93.

□

Většinou parametry $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ a $\boldsymbol{\Sigma}$ nejsou zadány, tudíž musíme pracovat s jejich odhady. Výpočet vychází z náhodných vektorů $\mathbf{X} = (X_1, \ldots, X_p)^\top$. Jestliže je v třídě π_1 přítomno n_1 měření a v třídě π_2 je k dispozici výběr o rozsahu n_2, ve výsledku pracujeme s číselnými maticemi typu $n_1 \times p$ a $n_2 \times p$,

$$\mathbf{X}_1 = \begin{pmatrix} \mathbf{x}_{11}^\top \\ \mathbf{x}_{12}^\top \\ \vdots \\ \mathbf{x}_{1n_1}^\top \end{pmatrix} \qquad \mathbf{X}_2 = \begin{pmatrix} \mathbf{x}_{21}^\top \\ \mathbf{x}_{22}^\top \\ \vdots \\ \mathbf{x}_{2n_2}^\top \end{pmatrix}.$$

Za odhady střední hodnoty $\boldsymbol{\mu}_i$ a varianční matice $\boldsymbol{\Sigma}_i$, $i = 1, 2$, považujeme výběrový průměr a výběrovou varianční matici,

$$\bar{\mathbf{x}}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} \mathbf{x}_{1j}, \qquad \mathbf{S}_1 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (\mathbf{x}_{1j} - \bar{x}_1)(\mathbf{x}_{1j} - \bar{x}_1)^\top,$$

$$\bar{\mathbf{x}}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} \mathbf{x}_{2j}, \qquad \mathbf{S}_2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (\mathbf{x}_{2j} - \bar{x}_2)(\mathbf{x}_{2j} - \bar{x}_2)^\top. \tag{11}$$

Nyní je ovšem uvažován případ, kdy se varianční matice rovnají ($\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$), tudíž provedeme spolu s $\mathbf{S}_1$ a $\mathbf{S}_2$ odhad společné výběrové varianční matice,

$$\mathbf{S} = \left(\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)} \right) \mathbf{S}_1 + \left(\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)} \right) \mathbf{S}_2. \tag{12}$$

Jelikož už známe odhady, můžeme lineární diskriminační funkci zjednodušit. Za předpokladu shodnosti ztrát (plynoucích z chybného zařazení) a rovnosti apriorních pravděpodobností, tj. $\left(\frac{z(1|2)}{z(2|1)} \frac{p_2}{p_1} \right) = 1$, dostáváme pravidlo pro zařazení objektu $\mathbf{x}_0$ do množiny R_1 jako

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \mathbf{S}^{-1} - \frac{1}{2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \geq \ln 1$$

a tedy

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \mathbf{x}_0 \geq \frac{1}{2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2), \tag{13}$$

jinak dle (10) zařadíme pozorovaný objekt do množiny R_2.

1.3.2 Kvadratická diskriminační analýza (QDA)

Jak již bylo zmíněno na začátku kapitoly o diskriminační analýze, může nastat případ, kdy u dvou vícerozměrných normálních rozdělení nejsou varianční matice shodné, tj. $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. I v tomto případě je klasifikační pravidlo postaveno na poměru hustot $f_1(\mathbf{x})/f_2(\mathbf{x})$, jako tomu bylo u lineární diskriminační analýzy. Aby tvorba diskriminačního pravidla byla snažší, je podíl zlogaritmován

$$\begin{aligned}\ln\left(\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right) &= \ln\left(\frac{|\boldsymbol{\Sigma}_2|^{1/2}}{|\boldsymbol{\Sigma}_1|^{1/2}}\right) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) \\ &= \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right) - \frac{1}{2}\mathbf{x}^\top(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_2^{-1})\mathbf{x} + (\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1})\mathbf{x} \\ &\quad - \frac{1}{2}(\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2).\end{aligned}$$

Nejmenší střední hodnota ztráty způsobená při špatném zařazení objektů do třídy π_1 nastane podle (2), jestliže

$$R_1 : -\frac{1}{2}\mathbf{x}^\top(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_1^{-1})\mathbf{x} + (\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1})\mathbf{x} - k \geq \ln\left(\frac{z(1|2)}{z(2|1)}\frac{p_2}{p_1}\right), \quad (14)$$

a pro zařazení do třídy π_2 podle (3)

$$R_2 : -\frac{1}{2}\mathbf{x}^\top(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_2^{-1})\mathbf{x} + (\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1})\mathbf{x} - k < \ln\left(\frac{z(1|2)}{z(2|1)}\frac{p_2}{p_1}\right), \quad (15)$$

kde $k = \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right) + \frac{1}{2}\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \frac{1}{2}\boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2$.

Rovnice (14) a (15) lze považovat za klasifikační pravidlo, tvrzení vychází z následující věty.

Věta 1.2. *Nechť π_1 a π_2 jsou populace, které mají normální rozdělení se středními hodnotami $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ a různými variančními maticemi $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. Pak klasifikační pravidlo minimalizující očekávanou hodnotu ztráty, tzn. kvadratická diskriminační funkce, zní:*
Nově příchozí objekt $\mathbf{x}_0$ (pozorování) zařadíme do třídy π_1, když

$$-\frac{1}{2}\mathbf{x}_0^\top(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_1^{-1})\mathbf{x}_0 + (\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1})\mathbf{x} - k \geq \ln\left(\frac{z(1|2)}{z(2|1)}\frac{p_2}{p_1}\right).$$

Jinak $\mathbf{x}_0$ *přiřadíme do* π_2.

Klasifikační pravidlo je kvadratické funkce proměnné $\mathbf{x}$, proto se hovoří o kvadratické diskriminační analýze. Za neznámé parametry se dosadí jejich odhady (viz kapitola 1.3.1).

V praxi je kvadratická diskriminační funkce potřebná, jsou-li varianční matice výrazně odlišné a rozsahy výběrů velké. Nevýhodou tohoto postupu je nutnost (alespoň přibližného) splnění předpokladu normality. Jestliže výběry jsou o malých rozsazích a rozdíly mezi variančními maticemi nejsou signifikantní, používá se lineární diskriminační funkce.

1.4 Diskriminace mezi více skupinami

V praxi se často setkáváme s diskriminací mezi více skupinami, což je potřeba zohlednit při tvorbě klasifikačního pravidla [2, 6].

Proto je tato kapitola věnována zobecnění pravidel pro diskriminaci mezi dvěma skupinami. Budeme uvažovat k počet skupin ($k > 2$), do nichž chceme zařadit pozorované objekty. Každá třída je charakterizována normálním rozdělením s hustotou

$$f_i(\mathbf{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) \right\}, \qquad i = 1, \ldots, k,$$

se střední hodnotou $\boldsymbol{\mu}_i$ a varianční maticí $\boldsymbol{\Sigma}_i$. Obdobně jako při klasifikaci do dvou skupin mohou nastat případy, kdy se varianční matice rovnají, nebo nerovnají.

Lineární diskriminační pravidlo pro i-tou třídu je

$$d_i^L(\mathbf{x}) = \boldsymbol{\mu}_i^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i, \qquad i = 1, \ldots, k. \tag{16}$$

Za $\boldsymbol{\mu}_i$ dosadíme příslušný odhad $\bar{\mathbf{x}}_i$ (výběrový průměr) a varianční matici odhadneme pomocí společné výběrové varianční matice $\mathbf{S}$,

$$\bar{\mathbf{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_{ij},$$

$$\mathbf{S} = \frac{1}{n_1 + \cdots + n_k - k}\Big((n_1 - 1)\mathbf{S}_1 + \cdots + (n_k - 1)\mathbf{S}_g\Big),$$

$$\mathbf{S}_i = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^\top.$$

Pozorovaný objekt zařadíme do třídy π_j, jestliže $d_j^L(\mathbf{x})$ je největší z hodnot $d_1^L(\mathbf{x}), \ldots, d_g^L(\mathbf{x})$.

Pravidlo pro kvadratickou diskriminační analýzu definujeme jako

$$d_i^Q(\mathbf{x}) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(\mathbf{x} - \mu_i)^\top\Sigma_i^{-1}(\mathbf{x} - \mu_i) + \ln p_i, \qquad i = 1, \ldots, g. \tag{17}$$

Opět je potřeba neznámé hodnoty nahradit jejich odhady. Princip zařazení zkoumaného objektu je stejný jako u lineární diskriminační analýzy.

1.5 Vyhodnocení správnosti diskriminace

Součástí každé analýzy je její vyhodnocení, v němž je potřeba zpětně zkontrolovat výsledky. Jedná se především o posouzení správnosti zvolené metody či nástrojů, chceme-li pravidel. U diskriminační analýzy posuzujeme kvalitu klasifikace objektů do tříd. Možností, jak posoudit kvalitu diskriminace, je mnoho. V této kapitole je zmíněna pouze křížová kontrola správnosti („jackknife“ metoda) [6, 10], která se v praxi využívá asi nejčastěji. Pro jednoduchost budeme uvažovat třídění do dvou skupin.

Vyhodnocení je založeno na sestavení tabulky (viz tabulka 1), která se skládá z původního umístění objektů a umístění provedeného pomocí diskriminace. Na vedlejší diagonále je počet chybně zařazených objektů. Pravděpodobnost chybného umístění pak představuje podíl chybně zařazených objektů a celkového počtu pozorovaných (klasifikovaných) objektů. Výsledek lze uvést i v procentech.

	původní zařazení	
zař. diskriminací	1	2
1	h_{11}	h_{12}
2	h_{21}	h_{22}

Tabulka 1: Křížová tabulka

Pravděpodobnost špatného umístění objektů je

$$\frac{h_{12} + h_{21}}{h_{11} + h_{12} + h_{21} + h_{22}}. \tag{18}$$

Pro pochopení je situace rozebrána podrobněji. Metoda spočívá v postupném vynechávání vždy jednoho objektu a jeho zařazení do třídy dle příslušného klasifikačního pravidla. Označme původní počet objektů zařazených do třídy π_1 jako n_1 (resp. n_2 pro třídu π_2). Začíná se vynecháním objektu zařazeného do třídy π_1 a následně se vytvoří klasifikační funkce pro $n_1 - 1$ a n_2 objektů. Vynechaný objekt zařadíme do třídy za pomoci nově vytvořeného klasifikačního pravidla. Takto opakujeme dokud nevynecháme postupně všechny objekty z tříd π_1 a π_2. Označme h_{21} počet špatně zařazených objektů, které pocházely z třídy π_1 (resp. h_{12}). Pravděpodobnost špatného zařazení je opravdu

$$\frac{h_{12} + h_{21}}{n_1 + n_2} = \frac{h_{12} + h_{21}}{h_{11} + h_{12} + h_{21} + h_{22}}.$$

Kvůli výpočetní náročnosti není vhodné tuto metodu použít pro velké počty objektů ve výběrech.

2 Kompoziční data

Kompoziční data (kompozice) jsou vícerozměrná data představující relativní podíly na celku. Za zakladatele prvního systematického přístupu ke statistické analýze kompozičních dat (tzv. logratio analýzy) je považován J. Aitchison, který pro tato data navrhl geometrii, dnes známou jako Aitchisonovu geometrii. Jejich vlastnostem odpovídá speciální výběrový prostor (simplex) a následně i speciální nástroje pro jejich statistickou analýzu (logratio transformace). Kompoziční data mají velké zastoupení v přírodních vědách, příkladem jsou koncentrace chemických prvků naměřených v hornině, nebo ve společenských vědách, např. měsíční výdaje domácností na různé komodity (bydlení, stravování, kultura apod.) [9, 11, 12].

Aby byl lépe pochopitelný rozdíl mezi standardními a kompozičními daty, uvedu podrobnější příklad. Představme si situaci, kdy dělník na stavbě pobírá průměrný měsíční plat 20 000 korun a manažer vysoce konkurenčně schopného podniku vydělává průměrně 40 000 korun za měsíc. Oba pánové mají sjednanou hypotéku a měsíční splátka je nastavena na 8 000 korun (absolutní hodnota splátky). Částka je pro oba stejná, ale pro každého představuje jinou (relativní) část platu. Dělník měsíčně zaplatí 40 % ze svého výdělku, ale manažer pouze 20 % (kompoziční data). Záleží, z jakého úhlu se na splátku díváme, zda-li jako na absolutní hodnoty či jako na výdajový podíl z celkového příjmu.

2.1 Vlastnosti kompozičních dat

Základní vlastnosti plynou již z definice kompozičních dat, která říká, že tato data nesou pouze relativní informaci, nikoli absolutní, jak je tomu u standardních mnohorozměrných dat.

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

Relativní informace je dána tím, že data reprezentují části celku. Z čehož

vyplývá, že mezi složkami existuje určitý vztah, pokud se hodnota jedné složky zvětší, druhá se musí automaticky zmenšit. Další vlastnosti plynou z jejich výběrového prostoru a z Aitchisonovy geometrie, navržené pro operace s těmito daty.

2.2 Výběrový prostor a Aitchisonova geometrie

Než se začne pracovat s daty, je potřeba si uvědomit, jaký je jejich výběrový prostor. Pro většinu dat je to reálný prostor s euklidovskou geometrií, proto jsou statistické metody téměř výhradně zkonstruovány pro tento typ dat. Naproti tomu kompoziční data lze vždy reprezentovat v omezeném prostoru, ve kterém jednotlivé složky kompozice nabývají hodnot od 0 do 100, nebo do libovolně zvolené konstanty κ. Omezeným prostorem máme na mysli simplex.

Definice 2.2. *Výběrový prostor kompozičních dat je simplex, definovaný jako*

$$S^D = \{\mathbf{x} = (x_1, x_2, \ldots, x_D)^\top | x_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} x_i = \kappa\}.$$

Simplex S^D je zřejmě podmnožina $\mathbb{R}^D, D \geq 2$. Pro $D = 2$ tvoří simplex úsečku a pro $D = 3$ je výběrovým prostorem trojúhelník. Výše uvedená definice charakterizuje výběrový prostor dat s konstantním součtem κ, tj. součet všech složek kompozice nám musí dát přesně zvolenou konstantu κ. Ve skutečnosti se běžně stává, že tento požadavek není splněn. Pro reprezentaci kompozice se ve výše uvedeném tvaru používá operace uzávěr, která nám převede původní kompoziční vektor s pozitivními složkami na takový, ve kterém součet všech složek bude roven zvolené konstantě.

Definice 2.3. *Pro libovolný D složkový kladný vektor $\mathbf{x} = (x_1, x_2, \ldots, x_D)^\top \in \mathbb{R}_+^D$ je uzávěr kompozice $\mathbf{x}$ definován jako*

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

Standardní euklidovská geometrie, ve které při statistické analýze běžně pracujeme, bohužel není vhodná pro kompoziční data. Proto je potřeba zavést jinou geometrii, tj. Aitchisonovu geometrii, s takovými operacemi, aby poskytovaly smysluplné informace o kompozicích podobně jako v uvedeném euklidovském prostoru. Mezi základní operace patří perturbace a mocninná transformace (vše ve spojení s uzávěrem).

Definice 2.4. *Perturbace kompozice* $\mathbf{x} \in S^D$ *kompozicí* $\mathbf{y} \in S^D$ *je kompozice*

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, x_2 y_2, \ldots, x_D, y_D)^\top.$$

Definice 2.5. *Mocninná transformace kompozice* $\mathbf{x} \in S^D$ *konstantou* $\alpha \in \mathbb{R}$ *je dána jako*

$$\alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha)^\top.$$

Skalární součin umožňuje umocnit jednotlivé složky kompozice konstantou.

Společně se zavedením operací perturbace a mocninné transformace lze zadefinovat další základní operace, jimiž jsou skalární součin, norma a vzdálenost mezi kompozicemi.

Definice 2.6. *Skalární součin dvou kompozic* $\mathbf{x}, \mathbf{y} \in S^D$ *je dán vztahem*

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

Definice 2.7. *Norma* $\mathbf{x} \in S^D$ *je definována jako*

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

Definice 2.8. *Vzdáleností mezi kompozicemi* $\mathbf{x}$, $\mathbf{y} \in S_D$ *nazveme*

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

kde $\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus ((-1) \odot \mathbf{y})$.

Uspořádaná trojice simplexu, perturbace, mocninné transformace spolu se skalárním součinem tvoří lineární euklidovský vektorový prostor.

Důkaz: [12], str. 12.

□

2.3 Logratio transformace a vztahy mezi nimi

Hlavním cílem navržení Aitchisonovy geometrie na simplexu bylo zadefinování nástrojů potřebných ke statistické analýze kompozičních dat. V tomto případě se za nástroje považují transformace, které převedou kompozici ze simplexu do reálného euklidovského prostoru [9, 11, 12].

Aby bylo možno použít běžné statistické metody (např. korelační analýzu, diskriminační analýzu nebo regresní metody) pro analýzu kompozic, navrhl J. Aitchison v knize [1] aditivní logratio (alr) a centrovanou logratio (clr) transformaci. Ovšem alr transformace nezachovává vzdálenosti (není izometrická), na proti tomu clr transformace je izometrická, ale vede k singulární varianční matici. Tyto vlastnosti vedly v [4] k zavední izometrické logratio (ilr) transformace. Základní úvahou při tvorbě ilr trasformace bylo zvolit vhodnou bázi na simplexu (vzhledem k Aitchisonově geometrii) a vyjádřit kompozici vzhledem k zvolené bázi. Nutností této báze je ortonormalita, která je základem pro izomorfismus mezi Aitchisonovou geometrií na simplexu a euklidovskou geometrií v reálném prostoru. Výsledkem ilr transformace je realný vektor, jehož složky jsou souřadnice vzhledem k (nějaké vhodně zvolené) ortonormální bázi.

Podívejme se na jednotlivé logratio transformace podrobněji. Použitím *alr transformace* pro D-složkovou kompozici $\mathbf{x} = (x_1, \ldots, x_D)$ ze simplexu S^D obdržíme $(D-1)$-rozměrný reálný vektor

$$\mathbf{y} = (y_1, \ldots, y_{D-1})^\top = alr(\mathbf{x}) = \left(\ln \frac{x_1}{x_D}, \frac{x_2}{x_D}, \ldots, \frac{x_{D-1}}{x_D} \right)^\top.$$

Tato transformace však není izometrická, tudíž nezachovává vzdálenosti. Z těchto důvodů byla navržena clr transformace.

Zobrazením D-složkové kompozice z S^D do reálného prostoru pomocí *clr transformace* dostáváme D-rozměrný vektor

$$\mathbf{w} = (w_1, \ldots, w_D)^\top = clr(\mathbf{x}) = \left(\ln \frac{x_1}{g(\mathbf{x})}, \frac{x_2}{g(\mathbf{x})}, \ldots, \frac{x_{D-1}}{g(\mathbf{x})} \right)^\top,$$

kde $g(\mathbf{x}) = \left(\prod_{i=1}^{D} x_i \right)^{1/D}$ je geometrický průměr složek kompozice. Clr transformace je izometrická, ale vede k singulární varianční matici, tj. k nulovému determinantu, který by mohl být během následného statistického zpracování dat problematický. Dále je potřeba poznamenat, že všechny clr transformovaná data leží na nadrovině $w_1 + \cdots + w_D = 0$.

Všechny vlastnosti, které jsou potřebné pro statistické zpracování kompozičních dat, splňuje *ilr transformace* a je definována pro vybranou ortonormální bázi jako

$$\mathbf{z} = (z_1, \ldots, z_{D-1})^\top = ilr(\mathbf{x}),\ z_i = \sqrt{\frac{i}{i+1}} \ln \frac{\sqrt[i]{\prod_{j=1}^{i} x_j}}{x_{i+1}},\ i = 1, \ldots, D-1.$$

Mezi výše uvedenými logratio transformacemi existují lineární vztahy [7], které lze vyjádřit jako

$$\mathbf{z} = \mathbf{U}\mathbf{w}, \qquad \mathbf{y} = \mathbf{C}\mathbf{z}, \qquad \mathbf{y} = \mathbf{F}\mathbf{w},$$

kde $\mathbf{U}$, $\mathbf{F}$ jsou matice typu $(D-1) \times D$ s plnou řádkovou hodností a navíc $\mathbf{U}\mathbf{U}^\top = \mathbf{I}_{D-1}$, z čehož vyplývá, že matice $\mathbf{C} = \mathbf{F}\mathbf{U}^\top$ je regulární, řádu $D-1$, tudíž můžeme uvažovat inverzní matici. Konkrétně řádky matice $\mathbf{U}$ tvoří vektory ortogonální báze nadroviny $w_1 + w_2 + \cdots + w_D = 0$ a $\mathbf{F} = [\mathbf{I}_{D-1}, -\mathbf{I}_{D-1}]$, kde $\mathbf{I}_{D-1}$ značí $(D-1)$-rozměrný vektor jedniček.

Následně dostaneme též

$$\mathbf{w} = \mathbf{U}^\top \mathbf{z}, \qquad \mathbf{z} = \mathbf{C}^{-1}\mathbf{y}, \qquad \mathbf{w} = \mathbf{F}^+ \mathbf{y},$$

kde $\mathbf{F}^+$ je Moore-Penroseova pseudoinverzní matice k $\mathbf{F}$, tj. matice splňující podmínky

$$\mathbf{F}\mathbf{F}^+\mathbf{F} = \mathbf{F},\ \mathbf{F}^+\mathbf{F}\mathbf{F}^+ = \mathbf{F}^+,\ (\mathbf{F}\mathbf{F}^+)^\top = \mathbf{F}\mathbf{F}^+,\ (\mathbf{F}^+\mathbf{F})^\top = \mathbf{F}^+\mathbf{F}.$$

Navíc různé ilr transformace (vzniklé odlišnou volbou ortonormální báze na simplexu) jsou vzájemně ortogonální, tedy pro $\mathbf{z}$ a $\mathbf{z}^*$ jako výsledky (různých) ilr transformací téže kompozice $\mathbf{x}$ existuje ortogonální matice $\mathbf{P}$ řádu $D-1$ ($\mathbf{P}\mathbf{P}^\top = \mathbf{P}^\top\mathbf{P} = \mathbf{I}_{D-1}$) taková, že $\mathbf{z}^* = \mathbf{P}\mathbf{z}$. Vzhledem k tomu, že právě alr a ilr transformace se při výpočtech vyskytují nejčastěji, budeme se jim věnovat i v další kapitole v kontextu diskriminační analýzy pro kompoziční data.

3 Diskriminační analýza pro kompoziční data

V následujících větách se vyšetřuje, jak se budou chovat diskriminační funkce při různých transformacích (alr, ilr) kompozičních dat [7].

Věta 3.1. *Hodnoty lineárního diskriminačního pravidla se nemění při alr nebo ilr transformaci dané kompozice (sledovaného objektu). To znamená, že nezáleží, jakou transformaci použijeme pro vyjádření kompozice v reálném prostoru.*

Důkaz: Uvažujme dvě kompozice $\mathbf{u}$ a $\mathbf{v}$, které vznikly ilr nebo alr transformací kompozice $\mathbf{x}$. Pak existuje regulární matice $\mathbf{A}$ řádu $D-1$ taková, že $\mathbf{v} = \mathbf{A}\mathbf{u}$. Použitím klasifikačního vztahu pro LDA podle (16) dostáváme

$$
\begin{aligned}
d_i^L(\mathbf{v}) &= (\mathbf{A}\boldsymbol{\mu}_i)^\top (\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)^{-1}\mathbf{A}\mathbf{u} - \frac{1}{2}(\mathbf{A}\boldsymbol{\mu}_i)^\top (\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)^{-1}\mathbf{A}\boldsymbol{\mu}_i + \ln p_i = \\
&= \boldsymbol{\mu}_i^\top \mathbf{A}^\top (\mathbf{A}^\top)^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{-1}\mathbf{A}\mathbf{u} - \frac{1}{2}\boldsymbol{\mu_i}^\top \mathbf{A}^\top (\mathbf{A}^\top)^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{-1}\mathbf{A}\boldsymbol{\mu}_i + \ln p_i = \\
&= \boldsymbol{\mu}_i\boldsymbol{\Sigma}^{-1}\mathbf{u} - \frac{1}{2}\boldsymbol{\mu}_i^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i = d_i^L(\mathbf{u}),\ i = 1, \ldots, k,
\end{aligned}
$$

kde bylo využito toho, že parametry normálního rozdělení vektoru $\mathbf{v}$ jsou rovny $\mathbf{A}\boldsymbol{\mu}_i$ a $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top$.

□

Věta 3.2. *Kvadratické diskriminační pravidlo je neměnné vzhledem k ilr transformaci. Neměnnost klasifikace je dosažena i při použití alr transformace.*

Důkaz: Tvrzení dokážeme obdobným způsobem jako u LDA podle (17); opět jsou uvažovány vektory $\mathbf{u}$ a $\mathbf{v}$, které vznikly ilr nebo alr transformací kompozice $\mathbf{x}$, tj. $\mathbf{v} = \mathbf{A}\mathbf{u}$, kde $\mathbf{A}$ je regulární matice řádu $D-1$. Potom tvrzení věty je

dokázáno s využitím vlastností determinantu,

$$
\begin{aligned}
d_i^Q(\mathbf{v}) &= -\frac{1}{2}\ln[\det(\mathbf{A}\boldsymbol{\Sigma}_i\mathbf{A}^\top)] - \frac{1}{2}(\mathbf{A}\mathbf{u}-\mathbf{A}\boldsymbol{\mu}_i)^\top(\mathbf{A}\boldsymbol{\Sigma}_i\mathbf{A}^\top)^{-1}(\mathbf{A}\mathbf{u}-\mathbf{A}\boldsymbol{\mu}_i) + \ln p_i = \\
&= -\frac{1}{2}\ln[\det(\mathbf{A})\det(\boldsymbol{\Sigma}_i)\det(\mathbf{A}^\top)] - \frac{1}{2}(\mathbf{u}-\boldsymbol{\mu}_i)^\top\mathbf{A}^\top(\mathbf{A}^\top)^{-1}\boldsymbol{\Sigma}_i^{-1}\mathbf{A}^{-1}\mathbf{A}(\mathbf{u}-\boldsymbol{\mu}_i) + \\
&+ \ln p_i = -\frac{1}{2}\ln[\det(\boldsymbol{\Sigma}_i)] - \frac{1}{2}\ln[\det(\mathbf{A}^2)] - \frac{1}{2}(\mathbf{u}-\boldsymbol{\mu}_i)^\top\boldsymbol{\Sigma}_i^{-1}(\mathbf{u}-\boldsymbol{\mu}_i) + \ln p_i = \\
&= d_i^Q(\mathbf{u}) - \frac{1}{2}\ln[\det(\mathbf{A}^2)],\ i = 1,\ldots,k.
\end{aligned}
$$

Je zřejmé, že hodnoty klasifikačního pravidla obecně nebudou stejné, tudíž je nutno uvažovat nad tím, z jakých transformací vznikly vektory $\mathbf{u}$ a $\mathbf{v}$. Jestliže vektory vznikly užitím dvou různých ilr transformací, pak matice $\mathbf{A}$ je ortogonální a $\det(\mathbf{A}) = \pm 1$, ze vztahu $\det(\mathbf{A}^2) = \det(\mathbf{A})\det(\mathbf{A}) = 1$ plyne, že hodnota klasifikačního kritéria zůstane zachována. Při použití ilr a alr transformace platí mezi vektory lineární vztah s maticí $\mathbf{A} = \mathbf{F}\mathbf{U}^\top$ (viz kapitola 2.3), jejíž determinant je různý od ± 1. Vzhledem k tomu, že výraz $\frac{1}{2}\ln[\det(\mathbf{A}^2)]$ neobsahuje index i, se výsledek klasifikace jako takové nezmění.

□

Při následných výpočtech ve čtvrté kapitole je možné se díky výše uvedeným větám omezit na užití libovolné ilr transformace se zárukou invariantních výsledků klasifikace.

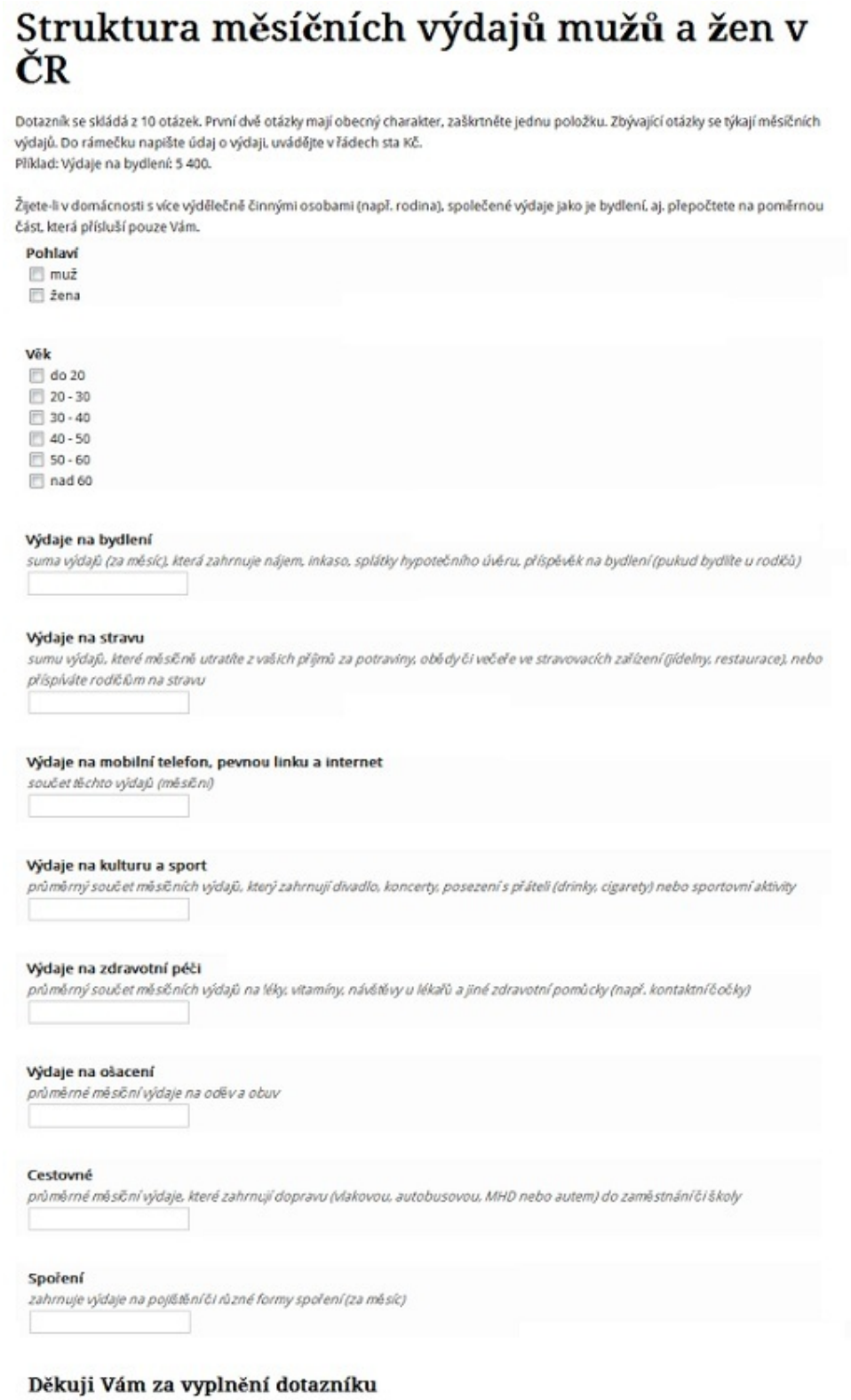
4 Praktický příklad

V této kapitole se budu zabývat praktickou interpretací diskriminační analýzy pro kompoziční data. Jelikož jsem působila v oblasti finančního poradenství a setkávala se s údaji o výdajích svých klientů, můj záměr byl propojit tuto oblast s kompozičními daty. Proto jsem si zvolila vlastní příklad, ve kterém zkoumám strukturu měsíčních výdajů mužů a žen v České republice na vybrané komodity.

Každý občan České republiky má jinou strukturu výdajů, což je dáno jednak věkem, zájmy ale také pohlavím. Jestliže zohledníme podíly jednotlivých výdajů na celkových výdajích, konkrétně, zajíma-li nás pouze relativní informace obsažená v datech, budeme pracovat s kompozičním datovým souborem.

Pomocí aplikace Google Docs byl vytvořen dotazník (viz obrázek 2), který byl následně rozeslán emailem a umístěn na sociální síti Facebook. Dotazovaní odpovídali na 10 otázek. První dvě měly obecný charakter, zbývající byly zaměřeny na měsíční výdaje. Výdaje byly členěny podle charakteru, a to na výdaje spojené s bydlením, se stravou, s komunikací, s kulturním a sportovním životem, se zdravotní péčí, s ošacením, s dopravou, ale i se spořením. Dotazovaní měli za úkol uvádět hodnoty v řádech stokorun. Vyplněná data byla automaticky zapisována do tabulky. Některá data měla nulové hodnoty. Jelikož metodika zpracování kompozic s nulovými složkami přesahuje rozsah a zaměření této práce, byl počet dotazovaných redukován na 66 (viz příloha 1), z čehož 27 bylo mužského pohlaví (označme „m“) a 39 ženského (označme „z“). Právě dle pohlaví bude prováděna následná klasifikace. Diskriminační pravidlo bylo tvořeno na 30 ženách a 20 mužích („tréningová“ data), od nichž byla data obdržena v první fázi měření, a klasifikace byla následně provedena na zbývajících objektech.

Veškeré výpočty jsou realizovány pomocí statistického softwaru R, který je volně dostupný na internetových stránkách *www.r-project.com*. Další výhodou softwaru je jednoduchost i velké množství knihoven, například knihovna *compositions*, která byla využita pro tuto práci. Pro jeho výhodné vlastnosti je R velmi rozšířený především mezi mladší generací statistiků.



Struktura měsíčních výdajů mužů a žen v ČR

Dotazník se skládá z 10 otázek. První dvě otázky mají obecný charakter, zaškrtněte jednu položku. Zbývající otázky se týkají měsíčních výdajů. Do rámečku napište údaj o výdaji, uvádějte v řádech sta Kč.
Příklad: Výdaje na bydlení: 5 400.

Žijete-li v domácnosti s více výdělečně činnými osobami (např. rodina), společené výdaje jako je bydlení, aj. přepočtete na poměrnou část, která přísluší pouze Vám.

Pohlaví
- muž
- žena

Věk
- do 20
- 20 - 30
- 30 - 40
- 40 - 50
- 50 - 60
- nad 60

Výdaje na bydlení
suma výdajů (za měsíc), která zahrnuje nájem, inkaso, splátky hypotečního úvěru, příspěvek na bydlení (pukud bydlíte u rodičů)

Výdaje na stravu
sumu výdajů, které měsíčně utratíte z vašich příjmů za potraviny, obědy či večeře ve stravovacích zařízení (jídelny, restaurace), nebo přispíváte rodičům na stravu

Výdaje na mobilní telefon, pevnou linku a internet
součet těchto výdajů (měsíční)

Výdaje na kulturu a sport
průměrný součet měsíčních výdajů, který zahrnují divadlo, koncerty, posezení s přáteli (drinky, cigarety) nebo sportovní aktivity

Výdaje na zdravotní péči
průměrný součet měsíčních výdajů na léky, vitamíny, návštěvy u lékařů a jiné zdravotní pomůcky (např. kontaktní čočky)

Výdaje na ošacení
průměrné měsíční výdaje na oděv a obuv

Cestovné
průměrné měsíční výdaje, které zahrnují dopravu (vlakovou, autobusovou, MHD nebo autem) do zaměstnání či školy

Spoření
zahrnuje výdaje na pojištění či různé formy spoření (za měsíc)

Děkuji Vám za vyplnění dotazníku

Obrázek 2: Dotazník

Postup zadávání příkazů s výsledky je uveden v přiloze 3. Veškeré výsledné hodnoty jsou zaokrouhleny na 4 desetinná místa použitím příkazu *round (object, digits)*.

Než přistoupíme k samotné diskriminační analýze je nezbytné, abychom naměřené osmisložkové kompozice převedli na data obsahující relativní informaci. Použitím následujícího příkazu, načteme knihovnu, která umožňuje pracovat s kompozičními daty.

```
> library(compositions)
```

Data načítáme z textového souboru jednotlivě pro ženy a pro muže. Při transformaci postupujeme stejným způsobem. Data byla převedena pomocí ilr transformace na sedmisložková pozorování (viz příloha 2).

Jelikož byla teoretická část věnována pouze diskriminační analýze za předpokladu normálního mnohorozměrného rozdělení uvažovaných skupin, je potřeba ověřit normalitu datového souboru mužů a žen. Pro ověření normality byl použit Anderson-Darlingův test [5, 12]. Použitím funkce *adtest* z knihovny *robCompositions* (viz příloha 3)

```
> adtest(vydajez, R = 1000, locscatt = "standard")
        A-D radius test
data:
= 1.1323, p-value = 0.067

> adtest(vydajem, R = 1000, locscatt = "standard")
        A-D radius test
data:
= 0.4219, p-value = 0.684
```

jsme na hladině 0,05 nezamítli nulovou hypotézu o normalitě datového souboru žen i mužů.

Nyní lze přistoupit k rozhodnutí, zda použít lineární nebo kvadratickou diskriminační analýzu, které je závislé na výsledku Boxova testu. Ověření je prováděno

na tréningovém souboru dat. Podle (6), (7) testujeme hypotézu o shodnosti variančních matic. Determinant výběrové varianční matice skupiny žen $|\mathbf{S}_z| = 0{,}0002$, pro skupinu mužů $|\mathbf{S}_m| = 0{,}0015$ a determinant společné výběrové varianční matice $\bar{\mathbf{S}} = 0{,}0011$. Testovací statistika $M = 8{,}1584 \cdot 10^{-12}$ a veličina

$$V = 49{,}3202,$$

jejíž hodnota nepřekračuje kvantil $v_{0,95}(36) = 51{,}00$. Tudíž na hladině 0,05 nelze zamítnout hypotézu o shodě variančních matic, proto lze pro klasifikaci použít lineární diskriminační analýzu.

4.1 LDA

Po zadání níže uvedeného příkazu, byla na tyto objekty

```
> vydaje[-train,8 ]
 [1] z z z z z z z z z m m m m m m m m
Levels: z m
```

aplikována lineární diskriminační analýza:

```
> #lda
> z <- lda(Sp ~ ., vydaje, subset = train)
> predict(z, vydaje[-train, ])$class
 [1] z z z z m z m z m m m m m z m m
Levels: z m.
```

Z výše uvedeného výsledku třídění vyplývá, že 6 žen a 6 mužů bylo klasifikováno správně, tři ženy byly zařazeny jako muž a jenom jeden muž do skupiny žen. Apriorní pravděpodobnost pro skupinu žen je 0,6 a pro skupinu mužů 0,4.

Odhady středních hodnot pro jednotlivé skupiny dostaneme jako

$$\bar{\mathbf{x}}_z = \begin{pmatrix} -0{,}3517 \\ -1{,}4065 \\ -0{,}7542 \\ -1{,}3697 \\ -0{,}2364 \\ 0{,}0076 \\ -0{,}5825 \end{pmatrix}, \qquad \bar{\mathbf{x}}_m = \begin{pmatrix} -0{,}0978 \\ -1{,}1617 \\ -0{,}3240 \\ -2{,}0068 \\ -0{,}5476 \\ -0{,}0984 \\ 0{,}2199 \end{pmatrix}.$$

Nově příchozí ilr transformovanou kompozici výdajů ženy nebo muže, tj. $\mathbf{x}_0$, zařadíme do třídy π_1 (žen), jestliže hodnota $d_z^L(\mathbf{x}_0) \geq d_m^L(\mathbf{x}_0)$, podle výběrového tvaru (16)

$$d_z^L(\mathbf{x}_0) = \begin{pmatrix} -0{,}1975 \\ -7{,}8715 \\ -1{,}2118 \\ -4{,}1906 \\ -0{,}5641 \\ -0{,}5716 \\ -0{,}5896 \end{pmatrix} \times \mathbf{x}_0 - 8{,}6226\,, \qquad d_m^L(\mathbf{x}_0) = \begin{pmatrix} 3{,}1592 \\ -0{,}5380 \\ 0{,}9644 \\ -7{,}7154 \\ -3{,}3956 \\ -2{,}5418 \\ -1{,}4127 \end{pmatrix} \times \mathbf{x}_0 - 7{,}7266.$$

Pro ilustraci uvažujme pozorovaný objekt, v tomto případě ženu s ID1, jejíž struktura výdajů je charakterizována náhodným vektorem $\mathbf{x}_0 = (2800, 2000, 600, 1600, 300, 600, 1000, 300)^\top$. Po provedení ilr transformace pomocí příslušné funkce knihovny *compositions* obdržíme souřadnice $\mathbf{x}_0 = (-0{,}2379, -1{,}1204, 0{,}0572, -1{,}4530, -0{,}5536, 0{,}0051, -1{,}1218)^\top$. Představme si, že bychom nevěděli o pohlaví pozorovaného objektu. Na základě diskriminačního pravidla jsme schopni říci, zda je pozorovaná osoba muž nebo žena. Za pomoci výše uvedených pravidel

$$d_z^L(\mathbf{x}_0) = 7{,}2340\,, \qquad d_m^L(\mathbf{x}_0) = 6{,}8414,$$

určíme pozorovanou osobu ženou, protože platí $d_z^L(\mathbf{x}_0) \geq d_m^L(\mathbf{x}_0)$.

Kapitola 1.5 byla věnována zjištění kvality klasifikace. Kontrola správnosti byla provedena pouze na tréningových datech. Výsledky vyhodnocení jsou uvedeny v následující tabulce.

	pův. zařazení	
zař.diskriminací	ž	m
ž	27	7
m	3	13

Příslušné výpočty byly provedeny ve statistickém softwaru. Příkazy jsou opět uvedeny v příloze 3. Výsledná diskriminace při postupném vypouštění objektů ze skupiny žen je posloupnost

```
z z z z m m z z z z m z z z z z z z z z z z z z z z z z z
Levels: z m
```

a mužů

```
m m m m z m z m z m z m z m m m m z z m
Levels: z m.
```

Špatně zařazených žen je 7 a chybně klasifikovaní muži jsou tři, z čehož vyplývá pravděpodobnost chybného zařazení podle (18)

$$\frac{7+3}{50} = 0{,}2.$$

Pro nízkou hodnotu pravděpodobnosti chybného zařazení je zřejmě v tomto případě lineární diskriminační analýza odpovídajícím klasifikačním nástrojem.

4.2 QDA

I když jsme na začátku kapitoly pomocí Boxova testu zjistili, že nelze zamítnout hypotézu o rovnosti variančních matic. Protože hodnota testovací statistiky byla velmi blízká příslušnému kvantilu, byla pro zajímavost na stejném datovém souboru provedena i kvadratická diskriminační analýza.

Za pomoci příkazu

```
> #qda
> v <- qda(Sp ~ ., vydaje, subset = train)
> predict(v, vydaje[-train, ])$class
```

```
 [1] z z z z z z m z m m m z m z m m
Levels: z m
```

bylo zjištěno, že dvě ženy byly klasifikovány jako muž a dva muži jako žena.

Na první pohled se zdá, že kvadratická diskriminace byla stejně uspěšná jako lineární diskriminace. Pro potvrzení takového závěru je potřeba posoudit účinnost zvolené metody pomocí pravděpodobnosti chybného zařazení. Příslušné výpočty byly provedeny ve statistickém softwaru (viz příloha 3). Metodou postupného vypouštění bylo 30 žen klasifikováno následovně

```
z z z z m m z z z z z z z z z z z z z z z m z z z z m z z
Levels: z m
```

a 20 mužů

```
m m m z m m z m z m z z z z m m m z z m
Levels: z m.
```

Pravděpodobnost špatného zařazení podle (18) je

$$\frac{4+9}{50} = 0{,}26\,,$$

jestliže jsme výše zjistili, že 4 ženy byly klasifikovány jako muž a 9 mužů bylo zařazeno do skupiny žen.

Jak je zřejmé z provedeného porovnání klasifikace na strukturu měsíčních výdajů mužů a žen, je lepší aplikovat lineární diskriminační analýzu, jelikož pravděpodobnost chybného zařazení je v porovnání s pravděpodobností chybného zařazení při použití kvadratické diskriminační analýzy menší.

Závěr

V první polovině teoretické části práce jsem shrnula poznatky týkající se diskriminační analýzy. Nejprve jsem se zabývala diskriminací mezi dvěma skupinami a poté jejímu zobecnění. Za důležité jsem považovala i popis vyhodnocení správnosti diskriminace. Druhou polovinu teoretické části jsem věnovala základním pojmům z oblasti kompozičních dat a popsala nástroje vedoucí ke statistickému zpracování uvedených dat nesoucích pouze relativní informaci. Mezi těmito nástroji (logratio transformacemi) existují lineární vztahy, které umožňují aplikaci diskriminační analýzy na kompoziční data. Speciálně lze v tomto ohledu doporučit ilr transformaci kompozic.

V praktické části jsem se zabývala aplikací uvedených postupů na výdaje mužů a žen v ČR. Hodnoty byly získány na základě vyplněných dotazníků. Bylo demonstrováno, že lineární diskriminační analýza funguje na datových souborech majících stejné varianční matice v jednotlivých skupinách. Naopak použití kvadratické diskriminační analýzy nebylo na stejném datovém souboru úspěšné, ačkoliv rozdíly mezi hodnotami pravděpodobností chybné klasifikace nebyly tak výrazné. To mohlo být způsobeno nízkým počtem dat nebo i nepřesnými údaji vyplněnými do dotazníku, protože ne každý si vede podrobné zápisy o jednotlivých výdajích.

Příloha 1

Součástí práce je CD ROM obsahující uvedená vstupní data.
Použité zkratky:

ID	identifikační číslo
P	pohlaví, žena (ž) nebo muž (m)
Bydl.	výdaje spojené s bydlením
Strava	výdaje za stravu
Tel.	výdaje spojené s účty za mobilní telefon, popř. internet
Kultur.	výdaje za kulturní nebo sportovní vyžití
Zdraví	výdaje za vitamíny, léky a jiné zdravotní pomůcky
Oděv	výdaje spojené s ošacením
Spoření	výdaje na pojištění či různé formy spoření
Cestovné	výdaje zahrunující dopravu

		Výdaje							
ID	P	Bydl.	Strava	Tel.	Kultur.	Zdraví	Oděv	Spoření	Doprav.
1	ž	2800	2000	600	1600	300	600	1000	300
2	ž	3000	4500	350	600	500	1500	5000	300
3	ž	3000	2000	500	2000	200	500	2000	200
4	ž	2200	2500	500	400	300	500	300	1000
5	ž	8000	5000	1500	1000	400	1000	2000	3000
6	ž	9000	7000	500	2500	500	1000	500	600
7	ž	6000	2000	1000	1000	1000	1000	2000	4000
8	ž	2800	2000	500	800	1000	1000	4000	300
9	ž	2600	2500	600	1500	500	1000	500	500
10	ž	4500	2500	750	400	1000	1000	1600	400
11	ž	2500	2000	500	1000	200	2500	2000	2500
12	ž	2750	1500	300	300	200	300	200	200
13	ž	15000	10000	2500	1000	500	1000	4500	1000
14	ž	3000	1000	150	1000	350	1000	650	240
15	ž	8000	5000	1000	1000	1000	2000	500	2000
16	ž	7000	2500	100	200	500	700	3000	250
17	ž	35000	15000	4000	2000	500	5000	3800	3000
18	ž	1200	2000	600	300	200	2000	1500	200
19	ž	7500	3000	1000	1500	1000	1000	1000	2000
20	ž	9100	5000	1500	5000	2000	1500	1000	2500

		Výdaje							
ID	P	Bydl.	Strava	Tel.	Kultur.	Zdraví	Oděv	Spoření	Doprav.
21	ž	4350	3500	500	300	100	1000	500	200
22	ž	1250	500	120	300	100	550	700	200
23	ž	6000	2000	600	1000	100	500	1500	500
24	ž	8500	2500	1100	4000	500	500	2000	500
25	ž	2500	500	150	200	100	200	150	500
26	ž	2800	3000	1200	500	200	500	500	200
27	ž	2900	3000	1100	500	150	500	400	150
28	ž	8700	1450	600	700	150	300	1500	900
29	ž	3000	4000	900	500	200	3000	2000	1200
30	ž	3200	4000	500	1000	200	1000	1000	500
31	ž	2800	2000	900	1000	1000	2000	2500	2500
32	ž	2000	2000	500	2000	200	500	500	160
33	ž	5500	2300	700	3000	2000	1000	1500	1000
34	ž	3000	6000	1000	800	300	1500	1500	1000
35	ž	2000	1000	350	300	200	300	1000	3000
36	ž	5000	5000	1500	100	200	500	2500	1000
37	ž	6000	2500	1000	2000	200	1200	2300	2000
38	ž	8000	3000	2000	2500	4000	2000	10000	2000
39	ž	2000	4000	500	2000	400	2000	1000	1500
1	m	6000	3500	300	500	50	500	2100	2000
2	m	2000	4000	500	1500	300	300	1000	1000
3	m	500	1000	500	3000	500	400	500	700
4	m	10000	5000	700	2500	500	1000	1000	2500
5	m	2500	2000	1500	2000	500	1000	500	1000
6	m	12000	5000	2000	3000	100	2000	2000	5000
7	m	13000	5000	2000	4000	400	3000	8000	4000
8	m	400	1500	610	500	100	500	200	2000
9	m	11000	3500	1800	1000	200	2000	1500	2000
10	m	3000	10000	9000	3000	500	8000	500	14000
11	m	3000	2000	500	500	200	300	1000	500
12	m	2000	1300	800	4000	300	500	1000	500
13	m	10500	8000	2500	1000	300	1000	3000	2000
14	m	3500	10000	500	5000	200	2000	2500	500
15	m	4500	3500	1000	2000	100	300	2000	3000
16	m	10500	4000	500	2000	100	200	1000	1500
17	m	5500	2500	500	1500	150	500	3600	3500
18	m	8000	5000	1200	3000	800	1000	100	1500

		Výdaje							
ID	P	Bydl.	Strava	Tel.	Kultur.	Zdraví	Oděv	Spoření	Doprav.
19	m	12000	8000	800	2000	500	1500	2000	500
20	m	2000	4000	1000	500	200	200	1500	1200
21	m	6200	8000	2000	2000	300	1000	3000	3000
22	m	2200	3000	3000	2500	500	1600	3000	4000
23	m	10000	3500	700	1500	200	1500	700	2500
24	m	5000	4000	800	3000	300	1000	2000	1000
25	m	3000	3000	600	300	300	300	200	1000
26	m	3000	4500	500	1000	200	1500	5000	3500
27	m	3000	2000	800	2500	1000	400	1500	1200

Příloha 2

		transformovaná kompozice						
ID	P	z_1	z_2	z_3	z_4	z_5	z_6	z_7
1	ž	-0.2379	-1.1204	0.0572	-1.4530	-0.5536	0.0051	-1.1218
2	ž	0.2867	-1.9197	-0.8907	-0.8530	0.3064	1.3736	-1.4421
3	ž	-0.2867	-1.2974	0.2831	-1.8402	-0.6660	0.7206	-1.5299
4	ž	0.0904	-1.2619	-1.0856	-1.0982	-0.4303	-0.8366	0.4017
5	ž	-0.3323	-1.1749	-1.1819	-1.7351	-0.5802	0.1513	0.5103
6	ž	-0.1777	-2.2574	-0.2024	-1.5963	-0.6706	-1.2085	-0.8761
7	ž	-0.7768	-1.0145	-0.7173	-0.5556	-0.4537	0.2583	0.8721
8	ž	-0.2379	-1.2693	-0.4905	-0.1803	-0.1472	1.1590	-1.4192
9	ž	-0.0277	-1.1812	-0.0417	-1.0150	-0.1960	-0.8073	-0.6992
10	ž	-0.4156	-1.2230	-1.4092	-0.2720	-0.2221	0.2474	-1.0825
11	ž	-0.1578	-1.2230	-0.2645	-1.6444	0.9630	0.6073	0.7347
12	ž	-0.4286	-1.5616	-1.1042	-1.2180	-0.6243	-0.9030	-0.7821
13	ž	-0.2867	-1.2974	-1.7110	-1.9453	-0.9556	0.5849	-0.9004
14	ž	-0.7768	-1.9975	0.2305	-0.7604	0.3375	-0.1136	-1.0304
15	ž	-0.3323	-1.5060	-1.0649	-0.8249	-0.0407	-1.3179	0.1554
16	ž	-0.7281	-3.0485	-1.5554	-0.3852	-0.0074	1.3411	-1.1630
17	ž	-0.5991	-1.4251	-1.6080	-2.4855	0.0726	-0.1927	-0.3880
18	ž	0.3612	-0.7745	-1.1479	-1.2518	1.0798	0.6463	-1.3251
19	ž	-0.6479	-1.2711	-0.5477	-0.7869	-0.6425	-0.5430	0.1781
20	ž	-0.4234	-1.2275	0.1747	-0.6842	-0.8213	-1.0695	-0.0691
21	ž	-0.1537	-1.6776	-1.6286	-2.2442	0.2696	-0.4139	-1.2155
22	ž	-0.6479	-1.5393	-0.2949	-1.2111	0.5674	0.7028	-0.5632
23	ž	-0.7768	-1.4315	-0.5699	-2.5009	-0.5728	0.5330	-0.5660
24	ž	-0.8653	-1.1699	0.2908	-1.6347	-1.3347	0.1554	-1.1622
25	ž	-1.1380	-1.6401	-0.9106	-1.3253	-0.4493	-0.6461	0.5667
26	ž	0.0488	-0.7200	-1.2673	-1.8012	-0.6342	-0.5360	-1.3213
27	ž	0.0240	-0.8054	-1.2523	-2.0469	-0.5722	-0.6902	-1.5152
28	ž	-1.2670	-1.4520	-0.8932	-2.0697	-1.0571	0.5966	0.0388
29	ž	0.2034	-1.1005	-1.2872	-1.8166	0.9888	0.4603	-0.0792
30	ž	0.1578	-1.6068	-0.5359	-1.8546	-0.0451	-0.0381	-0.6814
31	ž	-0.2379	-0.7893	-0.4669	-0.3617	0.3375	0.4918	0.4259
32	ž	0.0000	-1.1319	0.4002	-1.7495	-0.5920	-0.5003	-1.4992
33	ž	-0.6165	-1.3272	0.3218	-0.1134	-0.7253	-0.2376	-0.5851
34	ž	0.4901	-1.1800	-1.0276	-1.6733	0.1030	0.0870	-0.3039

		transformavaná kompozice z						
ID	P	z_1	z_2	z_3	z_4	z_5	z_6	z_7
35	ž	-0.4901	-1.1402	-0.9397	-1.0906	-0.5203	0.6749	1.6122
36	ž	0.0000	-0.9830	-3.0404	-1.7351	-0.5802	0.9997	0.0086
37	ž	-0.6190	-1.1056	-0.1815	-2.2001	-0.1607	0.4665	0.2733
38	ž	-0.6936	-0.7315	-0.3240	0.1694	-0.4944	1.0722	-0.5769
39	ž	0.4901	-1.4149	0.2001	-1.2845	0.4204	-0.2864	0.1312
1	m	-0.3811	-2.2260	-1.1316	-2.9360	-0.2953	1.0791	0.8889
2	m	0.4901	-1.4149	-0.0490	-1.4775	-1.2064	0.0951	0.0823
3	m	0.4901	-0.2830	1.3516	-0.5556	-0.6574	-0.3490	0.0125
4	m	-0.4901	-1.8883	-0.2328	-1.6199	-0.6899	-0.5830	0.3522
5	m	-0.1578	-0.3260	0.0186	-1.2255	-0.3679	-0.9526	-0.1766
6	m	-0.6190	-1.1056	-0.4306	-3.3757	-0.0215	-0.0182	0.8414
7	m	-0.6756	-1.1382	-0.2046	-2.2180	0.0284	0.9321	0.1588
8	m	0.9346	-0.1950	-0.3101	-1.6798	0.0977	-0.7657	1.4907
9	m	-0.8097	-1.0104	-1.2235	-2.3873	0.1528	-0.1372	0.1503
10	m	0.8513	0.4055	-0.6647	-2.1175	0.8021	-1.8890	1.4811
11	m	-0.2867	-1.2974	-0.9174	-1.5302	-0.8793	0.3716	-0.3266
12	m	-0.3046	-0.5723	0.9891	-1.5506	-0.7998	-0.0342	-0.6780
13	m	-0.1923	-1.0607	-1.5436	-2.2725	-0.7564	0.3778	-0.0521
14	m	0.7423	-2.0174	0.5676	-2.4394	0.1102	0.2997	-1.2459
15	m	-0.1777	-1.1255	-0.1955	-2.8309	-1.3086	0.6505	0.9426
16	m	-0.6824	-2.0918	-0.2786	-2.8953	-1.7312	0.0269	0.4026
17	m	-0.5575	-1.6360	-0.2054	-2.2186	-0.7124	1.2256	1.0350
18	m	-0.3323	-1.3571	-0.1661	-1.3109	-0.8666	-2.8642	0.0527
19	m	-0.2867	-2.0456	-0.6529	-1.7457	-0.4225	-0.0907	-1.3753
20	m	0.4901	-0.8489	-1.2006	-1.7495	-1.4285	0.6582	0.3613
21	m	0.1802	-1.0278	-0.7268	-2.2598	-0.7461	0.3866	0.3348
22	m	0.2193	0.1266	-0.0684	-1.4925	-0.1568	0.4495	0.6583
23	m	-0.7423	-1.7427	-0.5722	-2.2454	0.0060	-0.7006	0.5840
24	m	-0.1578	-1.4052	0.1510	-1.9425	-0.4870	0.2302	-0.4490
25	m	0.0000	-1.3141	-1.5295	-1.1847	-0.9673	-1.1929	0.4724
26	m	0.2867	-1.6285	-0.5512	-1.8665	0.3153	1.3812	0.8625
27	m	-0.2867	-0.9137	0.3407	-0.5556	-1.2901	0.1333	-0.0933

Příloha 3

Výsledky při zjišťování kvality diskriminace jsou pro lepší přehlednost upraveny. Součástí práce je přiložený CD ROM se scriptem obsahující jednotlivé příkazy.

```
> library(MASS)
> library(compositions)
>
> #nacteni pracovnich souboru
> vz1=read.table("vydajez1.txt")
> vz2=read.table("vydajez2.txt")
> row.names(vz2)=c("31","32","33","34","35","36","37","38","39")
> vm1=read.table("vydajem1.txt")
> vm2=read.table("vydajem2.txt")
> row.names(vm2)=c("21","22","23","24","25","26","27")
>
> #pouziti ilr transformace
> vydajez1=ilr(vz1)
> vydajem1=ilr(vm1)
> vydajez2=ilr(vz2)
> vydajem2=ilr(vm2)
>
> #overeni normality
> library(robCompositions)
> vz=data.frame(rbind(vz1,vz2))
> vm=data.frame(rbind(vm1,vm2))
> vydajez=ilr(vz)
> vydajem=ilr(vm)
> adtest(vydajez, R = 1000, locscatt = "standard")
        A-D radius test
data:
= 1.1323, p-value = 0.067
> adtest(vydajem, R = 1000, locscatt = "standard")
        A-D radius test
data:
= 0.4219, p-value = 0.684
>
> #testovani shodnosti variancnich matic - Boxova statistika
> det.a=det(cov(vydajez1))
> det.a
[1] 0.0002293314
> det.b=det(cov(vydajem1))
> det.b
```

```
[1] 0.001464614
> det.S=det((cov(vydajez1)*38+cov(vydajem1)*26)/64)
> det.S
[1] 0.001081715
> M.c=det.a^(38/2)*det.b^(26/2)
> M.j=det.S^(64/2)
> M=M.c/M.j
> M
[1] 8.158418e-12
> c=((2*16+3*8-1)/(6*9))*(1/38-1/64+1/26-1/64)
> V={-2}*(1-c)*log(M)
> V
[1] 49.3202
> qchisq(0.95,36)
[1] 50.99846
>
>
> #diskriminacni analyza
> vydajz=data.frame(cbind(vydajez),Sp=rep(c("z")))
> vydajm=data.frame(cbind(vydajem),Sp=rep(c("m")))
> vydaje=data.frame(rbind(vydajz,vydajm))
> set.seed(123)
> train= c(1:30,40:59)
> table(vydaje$Sp[train])
 z  m
30 20
> vydaje[-train,8 ]
 [1] z z z z z z z z z m m m m m m m m
Levels: z m
>
> #lda
> z <- lda(Sp ~ ., vydaje, subset = train)
> predict(z, vydaje[-train, ])$class
 [1] z z z z m z m z m m m m m z m m
Levels: z m
>
> z$prior
  z   m
0.6 0.4
>z$means
           X1        X2         X3        X4         X5           X6
z -0.35167241 -1.406499 -0.7542100 -1.369677 -0.2363951  0.007554136
m -0.09775435 -1.161735 -0.3240071 -2.006813 -0.5476148 -0.098377450
```

```
          X7
z -0.5824965
m  0.2198838
>
>
> #zarazeni  prvni zeny - hodnota d.l
> zena=vydajez1[1, ]
> cast1z=z$means[1,] %*% solve(cov(vydajez1))
> cast1z
           [,1]      [,2]      [,3]      [,4]       [,5]      [,6]       [,7]
[1,] -0.1974857 -7.871524 -1.211816 -4.190557 -0.5641156 -0.571573 -0.5896276
> cast2z=((t(z$means[1,]) %*% solve(cov(vydajez1))) %*% z$means[1,])/2 +
+   log(z$prior[1])
> cast2z
         [,1]
[1,] 8.622627
> d.lz=cast1z[1,] %*% zena - cast2z
> d.lz
         [,1]
[1,] 7.213712
>
> #vyhodnocení pro lda
> vydajev=vydaje[c(1:30,40:59),]
> vyhod.l=rep(NA,i)
> for(i in 1:50){
+ vec=1:50
+ train.l= vec[-i]
+ z <- lda(Sp ~ ., vydajev, subset = train.l)
+ print(predict(z, vydajev[-train.l,])$class)}
z z z z m m z z z z m z z z z z z z z z z z z z z z z z z
Levels: z m
m m m m z m z m z m z m z m m m m z z m
Levels: z m
>
> #qda
> v <- qda(Sp ~ ., vydaje, subset = train)
> predict(v, vydaje[-train, ])$class
 [1] z z z z z z m z m m m z m z m m
Levels: z m
>
> vyhodnoceni pro qda
> vyhod.q=rep(NA,i)
> for(i in 1:50){
```

```
+ vec=1:50
+ train.q= vec[-i]
+ v<- qda(Sp ~ ., vydajev, subset = train.q)
+ print(predict(v, vydajev[-train.q, ])$class)}
z z z z m m z z z z z z z z z z z z z z z z m z z z z m z z
Levels: z m
m m m z m m z m z m z z z z m m m z z m
Levels: z m
```

Literatura

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

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

[3] Anderson, T. W., *An introduction to multivariate statistical analysis*, 3^{rd} edition, New York: John Wiley & Sons, 2003.

[4] Egozcue, J.J., Pawlowsky-Glahn, V., *Isometric logratio transformations for compositional data analysis*, Mathematical Geology 35 (2003), 3, 279-300.

[5] Malý, O., *Testy normality pro kompoziční data*, Diplomová práce, UP Olomouc, 2010.

[6] Filzmoser, P., Multivariate Statistik, TU Wien, 2007.

[7] Filzmoser, P., Hron, K., Templ, M., *Discriminant analysis for compositional data and robust parameter estimation*, Computational Statistics (2012), DOI: 10.1007/s00180-011-0279-8.

[8] Hebák,P., *Vícerozměrné statistické metody (1)*, 2. vydání, Praha: Informatorium, 2007.

[9] Hron, K., *Elementy statistické analýzy kompozičních dat*, Informační bulletin České statistické společnosti 21 (2010), 3, 41-48.

[10] Meloun, M., Militký, J., *Kompendium statistického zpracování dat*, 2. vydání, Praha: Academia, 2006.

[11] Pawlowsky-Glahn, V., Egozcue, J.J., *Compositional data and their analysis: an introduction data analysis*, Geological Society, London, Special Publications 264 (2006), 1-10.

[12] Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., *Lecture Notes on Compositional Data Analysis*, Universitat de Girona, 2007.