EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Počítačová analýza vícerozměrných dat v oborech přírodních, technických a společenských věd

Učební texty ke kurzu

Autoři:

Prof. RNDr. Milan Meloun, DrSc. (Univerzita Pardubice, Pardubice)

Datum:

20.-24. června 2011

Centrum pro rozvoj výzkumu pokročilých řídicích a senzorických technologií CZ.1.07/2.3.00/09.0031

TENTO STUDIJNÍ MATERIÁL JE SPOLUFINANCOVÁN EVROPSKÝM SOCIÁLNÍM FONDEM A STÁTNÍM ROZPOČTEM ČESKÉ REPUBLIKY

OBSAH

1. POČÍTAČOVÁ ANALÝZA VÍCEROZMĚRNÝCH DAT V PŘÍKLADECH V OBORECH PŘÍRODNÍCH, TECHNICKÝCH A SPOLEČENSKÝCH VĚD

Počítačově orientovaná statistická analýza vícerozměrných dat je populárně a značně nematematicky vysvětlena na 50 obsáhlých praktických příkladech. Použité metody umožňují extrahovat v datech ukrytou informaci a odhalovat vnitřní vazby a strukturu vlastností znaků a objektů datové matice. Každá z 11 kapitol se týká vždy jedné metody vícerozměrné analýzy a popisuje cíl spolu se zaměřením, podrobným pracovním postupem a interpretací výsledků. Kniha se soustřeďuje na v praxi nejpoužívanější metody statistické analýzy vícerozměrných dat jako jsou průzkumová analýza dat s prvotní úpravou dat, která je následována metodou hlavních komponent, faktorovou analýzou, diskriminační analýzou, kanonickou korelací, logistickou regresí, analýzou shluků, vícerozměrným škálováním a korespondenční analýzou. Důraz je kladen zejména na formulaci úlohy a na interpretaci výsledků. Výklad metod je usnadněn několika sty obrázky s grafy a diagramy, ve kterých je diagnostikování vnitřních vztahů zvláště názorné a vhodné k logické interpretaci. Příklady jsou voleny především z oborů přírodních věd, zejména z biochemické a klinické praxe, dále z analytické chemie ale také ze sociologie a psychologie. V řadě užívaných počítačových programů dominuje statistický software STATISTICA. Kniha poslouží jako učebnice začátečníkům a studentům přírodních, technických i společenských věd, kteří začínají s vícerozměrnou analýzou pomocí počítače. Užitečnost této praktické příručky umocňuje přiložené CD s programem STATISTICA a s databází vstupních dat ode všech 50 uvedených příkladů včetně jejich počítačových výstupů.

---

2. CHARAKTER VÍCEROZMĚRNÝCH DAT

2.1. Nepřímá pozorování a korelace

● **Příroda je vícerozměrná**. Většina technologických ale i ostatních dat zpracovávaných statistickými metodami je vícerozměrná. Chceme-li studovat jeden jediný jev, musíme si uvědomit, že je ovlivněn řadou faktorů. Počasí, např. závisí na ovlivňujících faktorech jako je vítr, tlak vzduchu, teplota, rosný bod, které označíme znaky. Obecně se m-tice znaků sleduje pro n objektů, kde bývá obvyklé, že $n \gg m$. Je tedy vždy výhodné použít co nejmenší (postačující) počet znaků m. Při konstrukci modelů na základě experimentů lze v řadě případů řadu znaků buď zkonstantnit nebo znáhodnit. V limitním případě pak tvoří data náhodný výběr z dobře definovaného souboru. Také v případě vícerozměrných dat se často využívá předpokladu náhodného výběru, který umožňuje konstrukci intervalů resp. oblastí spolehlivosti, testů hypotéz a dalších statistických závěrů o chování souboru.

Typické pro neexperimentální data, tj. data získaná na základě pasivního pozorování nebo sběru informací, je vágnost v definici souboru resp. přesněji skupiny, pro kterou se hledají skryté souvislosti. To vede ke stavu, kdy statistické závěry jsou pouze formální a primární je pouze deskripce vazeb respektive vztahů.

(a) Analýza experimentálních dat je zaměřena na redukci rozměrnosti již ve fázi konstrukce výběru tak, aby bylo možné zkoumat obecně nelineární vztahy mezi znaky.
(b) Pro analýzu neexperimentálních dat je typické, že redukce rozměrnosti se provádí až v rámci statistické analýzy a předpokládají se lineární vztahy mezi znaky.

● **Přímá a nepřímá pozorování**: Pokud jsou data přímo požadovaným výsledkem, jde o přímá pozorování. To je případ, kdy přístroj monitoruje studovaný jev přímo. Je patrné, že i jednorozměrné výsledky mohou být tvořeny na základě kombinace více znaků. Přímá měření poskytuje poměrně málo měřících metod. Typické je například měření délky měřítkem. Měření teploty standardním teploměrem je však vlastně měření délky třeba rtuťového

sloupce a pro přepočet na teplotu je třeba převést naměřenou délku na teplotu kalibrací. Jde o příklad klasického nepřímého pozorování. V řadě případů je výsledek funkční kombinací více přímých a nepřímých pozorování, například koncentrace vyjádřená jako podíl hmotnosti a objemu.

● **Data musí obsahovat užitečnou informaci**: Důležitým předpokladem použití analýzy dat je podmínka, aby data obsahovala požadovanou informaci. Když chceme například nalézt koncentraci sloučeniny v roztoku, musí měření roztoku monitorovat tuto sloučeninu. Žádná statistická metoda nemůže pomoci, když data neobsahují dostatečný podíl informace o studované vlastnosti či jevu. Objem informace v datech závisí na způsobu definování problému, zda byla provedena dostatečná pozorování, měření či potřebné experimenty. Experimentátor by měl naměřit relevantní data, která budou dostatečně vypovídat o dané vlastnosti.

2.2. Zdrojová matice dat

U vícerozměrných dat se v řadě případů předpokládá vztah mezi souborem naměřených znaků a požadovaným výsledkem neboli vlastností. Matematicky budeme výslednou vlastnost $\boldsymbol{Y}$ chápat jako závisle proměnnou, která je funkcí několika nezávisle proměnných $\boldsymbol{X}$. Výsledek obyčejně představuje náročné, ekonomicky drahé měření, zatímco proměnné $\boldsymbol{X}$ představuje snadno dostupné a ekonomicky ne náročná měření. Takto definovaná úloha je označována jako vícerozměrná kalibrace. V řadě případů jsou vlastnosti vystiženy q-ticí proměnných či znaků. Pro n-tici objektů je pak k dispozici matice $\boldsymbol{Y}$ $(n \times q)$ a výchozí matice dat $\boldsymbol{Z}$ $(n \times m)$. Platí, že $n \gg m > q$. Maticově lze pak vyjádřit vstupní data buď náhodný výběr, nebo neexperimentální data ve tvaru

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_j^T \\ \vdots \\ \boldsymbol{y}_n^T \end{pmatrix} = \begin{pmatrix} y_{1,1} & \dots & y_{1,i} & \dots & y_{1,q} \\ \vdots & & \vdots & & \vdots \\ y_{j,1} & \dots & y_{j,i} & \dots & y_{j,q} \\ \vdots & & \vdots & & \vdots \\ y_{n,1} & \dots & y_{n,i} & \dots & y_{n,q} \end{pmatrix}, \boldsymbol{Z} = \begin{pmatrix} \boldsymbol{z}_1^T \\ \vdots \\ \boldsymbol{z}_j^T \\ \vdots \\ \boldsymbol{z}_n^T \end{pmatrix}$$
$$= \begin{pmatrix} z_{1,1} & \dots & z_{1,i} & \dots & z_{1,m} \\ \vdots & & \vdots & & \vdots \\ z_{j,1} & \dots & z_{j,i} & \dots & z_{j,m} \\ \vdots & & \vdots & & \vdots \\ z_{n,1} & \dots & z_{n,i} & \dots & z_{n,m} \end{pmatrix},$$

Řádky matice Z představují objekty (vzorky, jedince, pacienty), na kterých se provádí zkoumání. Sloupce matice $\boldsymbol{Y}$ a $\boldsymbol{Z}$ představují sledované znaky, respektive charakteristiky objektů, které se na objektech zkoumají. Při dalším strukturování se matice Y dělí do několika skupin vysvětlovaných proměnných $\boldsymbol{Y}_1$ $(n \times q1)$, $\boldsymbol{Y}_2$ $(n \times q2)$, ..., $\boldsymbol{Y}_0$ $(n \times q0)$, kde $q = \sum_{i=1}^{o} q_i$. Pokud se pro statistickou analýzu použije matice $\boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{Z})$ jde o tzv. nestrukturovaná data. Strukturování vícerozměrných dat souvisí úzce s tím, jaké problémy jsou pomocí vícerozměrných statistických metod řešeny.

Vlastní metody vícerozměrné statistické analýzy závisí na škále, ve které jsou data měřena. Podle množství informace obsažených v jednotlivých škálách jsou na nich definovány různé typy operací:

a) *Nominální škála* má zavedenu pouze operaci rovnosti (=) nebo nerovnosti (≠). Rozlišeny jsou pouze různé stavy. Ze statistického hlediska jde o kvalitativní proměnné, které mohou být buď binární (definující přítomnost 1 nebo nepřítomnost 0 nějakého znaku), nebo vícestavové (kódované obyčejně čísly 0, 1, 2, ...). Příkladem vícestavové náhodné veličiny je typ katalyzátoru, druh přístroje, barva objektu (pokud se vyjadřuje subjektivně) atd. Vícestavové kvalitativní proměnné se obyčejně převádějí na umělé binární proměnné [1].

b) *Ordinální škála* je škála, kde k operaci rovnosti a nerovnosti přistupují ještě operace typu menší (<) nebo větší (>). Tento typ škály se často vyskytuje, když jsou znaky hodnoceny subjektivně a lze provést logické uspořádání do stupnice od nejhoršího k nejlepšímu. Příkladem jsou stupnice stálostí na světle

a senzorická hodnocení vzhledu, omaku, vůně atd. Ze statistického hlediska jde o *semikvantitativní znaky*, kde jsou sice kategorie uspořádány, ale nemají mezi sebou měřitelné rozdíly.

c) *Kardinální škála* je škála, v níž je zavedena metrika (vzdálenost), takže lze provádět matematické operace jako je sčítání, odčítání, násobení a dělení atd. Kardinální škála se ještě člení na *intervalovou škálu a poměrovou škálu*. V intervalové škále lze provádět také sčítání a odčítání. Není zde však zaveden přirozený nulový bod. V poměrové škále je možné vyjádřit i poměr mezi objekty (dělení), tj. je zaveden přirozený počátek. Ze statistického hlediska jde buď o *diskrétní kvantitativní znaky* (kategorizované do disjunktních kategorií), nebo *kvantitativní znaky spojité*.

Platí, že vyšší typ škály v sobě zahrnuje operace všech nižších typů a dá se převést (při ztrátě informací) na nižší typ škály. V některých případech je však tento převod z různých důvodů výhodný. Například pořadová transformace dat před vícerozměrnou statistickou analýzou „přirozeně “ odstraní problém s vybočujícími hodnotami. Na druhé straně mohou vznikat potíže v případech, kdy jsou použité metody založeny na jiných předpokladech (normalitě atd.). Z metrologického hlediska je pochopitelně žádoucí pracovat s daty v poměrové škále, která obsahuje maximum informací.

2.3. Druhy dat

V současné době existuje celá řada metod vícerozměrné statistické analýzy, které jsou často modifikovány pro speciální účely (taxonometrie, ekologie, časové řady, chemometrie, biometrie, psychometrie) a v nichž se využívá speciálních zvláštností dat (například jde o signály, prostorově závislá data, respektive časové řady, kde je další proměnnou čas, prostorové souřadnice, atd.) S ohledem na orientaci v tom, které metody se pro dané účely hodí, je výhodné využít dělení na strukturovaná a nestrukturovaná data.

2.4. Odhady parametrů polohy, rozptýlení a tvaru

Standardně se zde vychází z předpokladu, že matice X představuje náhodný výběr vektorové náhodné veličiny. Lze však analogicky zpracovat i výsledky neexperimentálních "pozorovacích" studií (observational study), kde však již mohou vznikat problémy speciálně v případě konstrukce testů hypotéz. Statistické chování vícerozměrných náhodných veličin je detailně popsané v řadě monografií a jejich přehled je uveden v učebnicích [20] a [21].

Z vícerozměrného výběru objektů velikosti n, definovaného n-ticí m-rozměrných vektorů $\boldsymbol{x}_i = \left(x_{i_1}, \ldots, x_{i_m}\right)^T$, $i = 1, \ldots, n$, je možné určit vektor *odhadů středních hodnot* $\widehat{\boldsymbol{\mu}}$ podle vztahu

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \tag{1}$$

Rozptyl je základní mírou rozptýlení znaků. Pro případ jedné proměnné x je definován vztahem

$$\boldsymbol{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^2 \tag{2}$$

Je běžné vyjadřovat tuto míru rozptýlení ve stejných jednotkách jako jsou měřené veličiny, a to známou s*měrodatnou odchylkou* $\boldsymbol{s}$, což je druhá odmocnina z rozptylu.

Kovariance $c(x_1, x_2)$ mezi dvěma proměnnými x_1 a x_2 je mírou jejich lineární závislosti. Velká absolutní hodnota kovariance indikuje silnou lineární vazbu mezi dvěma proměnnými. Malá hodnota kovariance znamená, že při změně x_1 se příliš nezmění x_2. Kovariance je mírou, která závisí na použitých jednotkách proměnných. Limitní (maximální) hodnota kovariance je rovna odmocnině z rozptylů $s_{x_1}^2$ a $s_{x_2}^2$. Tedy max $c(x_1, x_2) = \sqrt{s_{x_1}^2 s_{x_2}^2}$. Pozitivní kovariance znamená přímou vazbu mezi x_1 a x2 , tj. Při změně $x_1$se změní x_2 ve stejném smyslu, například růst x_1 je doprovázen růstem x_2. Negativní kovariance znamená nepřímou vazbu mezi x_1a x_2, tj. při změně x1 se změní x_2v opačném smyslu, například růst x_1 je doprovázen poklesem x_2. Nulová kovariance

znamená nekorelovanost, tj. lineární nezávislost. Ještě stále však může být mezi x_1a x_2 speciální typ nelineární závislosti.

Pro odhad kovarianční matice $\boldsymbol{S}^0$ platí rovnice

$$\boldsymbol{S}^0 = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^T \tag{3}$$

Diagonální prvky S_{ii}^0 jsou rozptyly $S_{x_i}^2$ a mimodiagonální prvky jsou kovariance $S_{ij}^0 = c(x_i, x_j)$. Podobně jako u jednorozměrných dat se používá *výběrová korigovaná kovarianční matice*

$$\boldsymbol{S} = \frac{n}{n-1}\boldsymbol{S}^0 \tag{4}$$

která je nevychýleným odhadem kovarianční matice souboru C.

Korelace mezi dvěma proměnnými x_1 a x_2 je praktičtější mírou lineárního vztahu, protože jde o standardizovanou kovarianci a tedy bezrozměrnou míru. Standardizace se provádí podělením součinem směrodatných odchylek. Jeví se nejužitečnější mírou vnitřního lineárního vztahu mezi dvěma proměnnými x_1a x_2. *Personův korelační koeficient* r je pak definován vztahem

$$r = \sqrt{\frac{\sum_{i=1}^{n}(x_{1i}-\overline{x_1})(x_{2i}-\overline{x_2})}{\sum_{i=1}^{n}(x_{1i}-\overline{x_1})^2 \sum_{i=1}^{n}(x_{2i}-\overline{x_2})^2}} \tag{5}$$

Koeficient determinace $D = r^2$ popisuje podíl celkového rozptylu, který lze objasnit tímto lineárním vztahem. Korelace rovná 0.0 značí, že mezi dvěma proměnnými není lineární vztah. Korelace rovná 1.0 značí, že mezi dvěma proměnnými existuje pozitivní lineární vztah. Korelace rovná -1.0 značí, že mezi dvěma proměnnými existuje negativní lineární vztah. Koeficient determinace $100D$ [%] vyjádřený v procentech je běžnou mírou k vystižení korelace, protože nezávisí na znaménku korelačního koeficientu.

● **Kauzalita versus korelace**: Korelace je statistický pojem pro vyjádření míry lineárního vztahu a jde o čistě pojmovou míru. Příčina a důsledek se týkají deterministických závislostí. Je třeba vždy rozebrat všechna možná vysvětlení příčinných souvislostí, aby bylo možné z korelace určit příčinu a důsledek.

Statistické ročenky o demografii ukazují, že například počet narozených děti na vesnicích ve Skandinávii koreluje s počtem čápů vyskytujících se v tomto kraji s korelačním koeficientem $r \approx 0.75$. Přesto nelze přítomnost čápů v tomto kraji brát jako příčinu narozených dětí. Uvedený příklad selhání příčinného výkladu na základě popisné statistiky ukazuje, že je třeba vždy určité opatrnosti ve výkladu korelace.

● **Struktury ukryté v datech**: Přirozeně se nalezne vždy nějaká korelace mezi sloupci matice $\boldsymbol{X}$. V řadě případů jde o současný vliv *několika* rozličných znaků. To znamená, že jeden znak je lineární kombinací ostatních znaků. Pokud jde o strukturovaná data a výsledek y závisí na jediném znaku a kovariance $c(y, x_j)$ je dostatečně vysoká, jde o tzv. "selektivní znak". V případě vektoru vstupních veličin existuje více úrovní selektivity.

Data obsahují často znaky, které mohou být irrelevantní k výsledku y. Tyto znaky se pak zařazují mezi chyby. Instrumentální šum a ostatní náhodné chyby budou vždy přítomny v datech, ale mohou zde být i jiné jevy, které se odehrají v průběhu měření a které jsou těsně spojeny s vyšetřovaným problémem.

Stanovujeme, například, analýzou spekter koncentraci látky A, která je ve směsi s látkami B a C, obr. 1.1. Naměřená spektra budou obsahovat vedle absorpčního pásu látky A také pásy látek B a C, které bychom potřebovali z dat spíše odstranit. Signál látek B a C zde bude v roli šumu. Jde vždy o to si uvědomit, co je cílem stanovení čili co je vytýčený model a co do modelu nepatří, co představuje rušivý šum. Vícerozměrná pozorování se proto běžně modelují jako součet dvou složek*: struktura a šum*.

3. PŘEDÚPRAVA VÍCEROZMĚRNÝCH DAT

Předúprava dat je důležitý krok v řadě technik vícerozměrné analýzy dat. V některých případech musí být provedena před vlastním použitím dané metody. V jiných případech tvoří procedura předúpravy součást metody, takže na vstupu mohou být i původní data [21].

3.1. Formy standardizace dat

Standardizace dat znamená odstranění závislosti na jednotkách a na parametru polohy respektive i rozptýlení. Po provedené standardizaci můžeme pomocí vah přiřadit znakům různou důležitost. Standardizace tvoří často první krok v předúpravě vícerozměrných dat. Obecný termín *škálování* vystihuje, že operace se týká jak jednotek veličin, tak i počátku stupnice. Škálování může být použito na znaky, na objekty nebo na obojí. Škálování by mělo zahrnout:

a) posun centra souřadného systému,
b) protažení nebo zkrácení měřítka na osách.

Po posunu centra do nuly tj. centrování sloupců se vzdálenost mezi dvěma objekty nezmění. To však neplatí při změně měřítka. Znaky před škálováním v prostoru objektů dobře oddělené mohou být po škálování totožné. Z tohoto hlediska jsou některé škálovací techniky dat daleko od reality. Uvedeme nejběžnější škálovací techniky, ve kterých bude yij představovat pro i-tý objekt transformací upravený čili j-tý *škálovaný znak* odpovídající původnímu prvku x_{ij}.

● **Sloupcové centrování**. Novým centrem stupnice znaku v j-tém sloupci je nula. Původním centrem byl průměr prvků znaku $\bar{x_j}$. Sloupcově centrovaná data y_{ij} vzniknou dle vztahu $y_{ij} = x_{ij} - \bar{x_j}$, kde $\bar{x}_j$ je průměr prvků j-tého sloupce vyčíslený dle vztahu

$$\bar{x}_j = \sum_{i=1}^{n} \frac{x_{ij}}{n}. \tag{6}$$

● **Sloupcová standardizace**. Prvky znaku původních dat v j-tém sloupci x_{ij} jsou děleny odpovídající směrodatnou odchylkou tedy $y_{ij} = x_{ij}/s_j$, kde s_j je

směrodatná odchylka počítaná z hodnoty prvků j-tého sloupce vyčíslená dle vztahu

$$s_j = \sqrt{\frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}{n-1}}. \quad (7)$$

● **Autoškálování**. Je názvem běžně užívaným k označování kombinace sloupcového centrování a sloupcové standardizace. Jde vlastně o tzv. studentizaci

$$y_{ij} = (x_{ij} - \bar{x}_j)/s_j, \quad (8)$$

která je analogická Z-transformaci pro velké výběry, kdy předpokládáme, že známe μ_j a σ_j

$$y_{ij} = \frac{x_{ij}-\mu_j}{\sigma_j}. \quad (9)$$

Autoškálování užívá odhadů jak střední hodnoty, tak i směrodatné odchylky.

● **Škálování sloupcovým rozsahem**. Znaky jsou škálovány, aby bylo získáno minimum každého znaku rovné 0 a maximum rovné 1 dle vztahu

$$y_{ij} = \frac{x_{ij}-\min_j x_{ij}}{\max_j x_{ij}-\min_j x_{ij}}. \quad (10)$$

● **Řádkové centrování**. Znaky jsou škálovány dle vzorce .$y_{ij} = x_{ij} - \bar{x}_i$

● **Řádková standardizace**. Znaky jsou škálovány dle vzorce $y_{ij} = x_{ij}/s_i$.

● **Celkové centrování**. Znaky jsou škálovány dle vzorce $y_{ij} = x_{ij} - \bar{x}$, kde $\bar{x}$ je celkový průměr vyčíslený pro celou zdrojovou matici dat rozměru $n \times m$.

● **Celková standardizace**. Znaky jsou škálovány dle vzorce $y_{ij} = x_{ij}/s$, kde s je směrodatná odchylka od průměru pro všechny prvky zdrojové matice rozměru $n \times m$.

● **Dvojité centrování**. Znaky jsou škálovány nejdříve sloupcovým centrováním a následně řádkovým centrováním.

● **Řádkové profily**. Znaky jsou škálovány dle vzorce $y_{ij} = x_{ij}/(\bar{x}_i m)$. Tento poněkud zvláštní případ škálování se užívá hodně v chemii, kdy je znak relativní a je vyjádřen v procentech. Součet řádku je pak 1.

● **Sloupcové profily**. Znaky jsou škálovány dle vzorce $y_{ij} = x_{ij}/(\bar{x}_j n)$.

3.2. Průzkumová analýza vícerozměrných dat

3.2.1. *Zobrazení vícerozměrných dat*

Pro účely průzkumové analýzy vícerozměrných dat se používá různých technik, umožňujících jejich grafické zobrazení ve dvourozměrném souřadnicovém systému a u novějších výpočetních prostředků i v třírozměrném souřadnicovém systému. Toto zobrazení umožňuje:

a) identifikovat vektory $\boldsymbol{x}_i$ nebo jejich složek, které se jeví jako vybočující,
b) indikovat různé struktury v datech jako jsou shluky, jež ukazují na heterogenitu použitého výběru nebo přítomnost různých dílčích výběrů s odlišným chováním.

Na základě těchto informací a výsledků testů normality (popř. grafických ekvivalentů těchto testů) pak může být před vlastní statistickou analýzou provedena řada různých korekcí vedoucích k odstranění nehomogenity výběru a přiblížení se k vícerozměrné normalitě.

Většina používaných technik pro zobrazení vícerozměrných dat se dá zařadit do jedné ze dvou základních skupin, kterými jsou *zobecněné rozptylové diagramy a symbolové grafy*.

Pro základní případ dvojice náhodných znaků ($m = 2$) lze konstruovat rozptylové grafy, které umožňují sledovat statistické zvláštnosti dat. Ke konstrukci výběrového rozdělení nebo jeho porovnání s rozděleními teoretickými je možné použít i histogramy, neparametrické odhady hustoty a jiné grafy. Problémy však nastávají u vícerozměrných dat pro $m > 2$, kdy je třeba buď volit několik různých grafů, či vhodným způsobem provést transformace na dvoudimenzionální data.

Nelze obecně říci, která metoda zobrazení vícerozměrných dat je nejlepší. Závisí to na počtu znaků m vektorů $\boldsymbol{x}_i$, $i = 1, \ldots, n$, počtu měření a specifických zvláštnostech dat.

4. METODY K ODHALENÍ STRUKTURY VE ZNACÍCH A OBJEKTECH

Zdrojová matice dat $\boldsymbol{X}$ má rozměr $n \times m$. Před vlastní aplikací vhodné metody vícerozměrné statistické analýzy je třeba vždy provést průzkumovou (exploratorní) analýzu dat, která umožňuje:

a) posoudit *podobnost objektů* pomocí rozptylových a symbolových grafů,
b) nalézt *vybočující objekty*, resp. jejich znaky,
c) stanovit, zda lze použít předpoklad *lineárních vazeb*,
d) *ověřit předpoklady o datech* (normalitu, nekorelovanost, homogenitu).

Jednotlivé techniky k určení vzájemných vazeb se dále dělí podle toho, zda hledají

1. *strukturu a vazby ve znacích* nebo

2. *strukturu a vazby v objektech*, což je:

a) hledání struktury ve znacích v metrické škále − faktorová analýza FA, analýza hlavních komponent PCA a shluková analýza.
b) hledání struktury v objektech v metrické škále − *shluková analýza*.
c) hledání struktury v objektech v metrické i v nemetrické škále − *vícerozměrné škálování*.
d) hledání struktury v objektech v nemetrické škále − *korespondenční analýza*.

Většina metod vícerozměrné statistické analýzy umožňuje zpracování *lineárních vícerozměrných modelů*, kde závisle proměnné se uvažují jako lineární kombinace nezávisle proměnných, resp. vazby mezi proměnnými jsou lineární. V řadě případů se také uvažuje normalita metrických proměnných.

Určením struktury a vzájemných vazeb mezi znaky ale i mezi objekty se zabývají techniky redukce znaků na latentní proměnné, *metoda analýzy hlavních komponent (PCA) a metoda faktorové analýzy (FA)*. Důležitou metodou určení vzájemných vazeb mezi znaky je i *kanonická korelační analýza CA*, která se používá ke zkoumání závislosti mezi dvěma skupinami znaků, přičemž jedna ze skupin se považuje za proměnné nezávislé a druhá za skupinu proměnných závislých.

● *Maticový diagram*: diagram ukazuje rozptylové diagramy jednotlivých dvojic znaků. Je zřejmé, že vysoké hodnoty korelačního koeficientu vedou ke zřetelné lineární závislosti (přímka), zatímco nízké hodnoty ukazují, že osoby neleží v grafu zobrazeny na přímce ale leží spíše v chaotickém mraku bodů.

● *Směry maximálního rozptylu* v rozptylovém diagramu: Intuitivně tušíme, že by například na obr. 3.13 bylo nejlepší proložit objekty novou souřadnou osu tak, aby byla totožná s nejtěsněji prokládanou přímkou. Z toho plyne, že vhodný počet potřebných os není 3 ale pouze 1. Je to proto, že znaky x_1, x_2 a x_3 spolu silně korelují a objekty leží na prostorové přímce. Nejtěsněji objekty prokládaná přímka, a tím pádem i nová souřadnicová osa zvaná *PC1* leží ve směru *největšího rozptylu objektů*. Souřadnicová osa *PC1* není obecně přitom paralelní s žádnou z původních os znaků x_1, x_2 a x_3.

● Novou souřadnou osu *PC1* nazveme v následujících kapitolách první hlavní komponentou. Bude ležet ve *směru maximálního rozptylu* zdrojové matice dat $\boldsymbol{X}$. S touto osou je spjata jakási nová latentní proměnná, jejíž význam zatím neznáme. Metoda hlavních komponent PCA nám poskytne nejenom tuto první komponentu ale také další hlavní komponenty a je jenom na nás, jak je budeme interpretovat, co znamenají v analýze a co popisují. Když vezmeme příklad ze spektrometrie, kde počet znaků m je roven počtu vlnových délek, při kterých měříme absorbanci a kdy m může dosahovat stovek až tisíců, potom vynášení dosavadního 2D-rozptylového grafu pro vybrané dvojice vlnových délek je pro velký počet dvojic dat poněkud nereálné. Zde pomůže metoda hlavních komponent PCA odhalit skrytou strukturu v datech, a to díky svým vlastnostem. Obvykle jedna nebo dvě, maximálně však tři, nové souřadnicové osy *PC1, PC2* a *PC3* vystihnou dohromady největší podíl proměnlivosti dat.

● *První hlavní komponenta proložením nejmenšími čtverci*: Účelem je proložit přímku v prostoru x_1, x_2 a x_3 umístěnými objekty. Když z každého i-tého objektu spustíme k této přímce kolmici, obdržíme vzdálenost e_i zvanou *reziduum*. Nejtěsněji proloženou přímku pak dostaneme, když suma čtverců reziduí bude dosahovat své minimální velikosti dle vztahu

$$RSS = \sum_{i=1}^{n} e_i^2 = minimum. \qquad (11)$$

Existují tedy dvě kritéria ke konstrukci první hlavní komponenty: 1. Směr největšího rozptylu, 2. metoda nejmenších čtverců. Ukazuje se, že obě metody vedou ke stejnému výsledku.

● *Rozšíření k dalším hlavním komponentám*: Může nastat případ, že první hlavní komponenta *PC1* nestačí dostatečně popsat rozptýlení objektů. Druhá hlavní komponenta *PC2* je ortogonální vůči první (je na ni kolmá) a je umístěna do směru druhého největšího rozptýlení objektů v rovině. Podobně, vytvoří-li objekty v prostoru přibližný elipsoid, budou do prostorových os elipsoidu umístěny tři osy vystihující největší rozptýlení objektů čili osy tří hlavních komponent *PC1*, *PC2* a *PC3*. Nové tři osy nekorelují, jsou vzájemně ortogonální tj. na sebe kolmé. Metoda hlavních komponent PCA umožňuje tuto geometrickou projekci zobecnit na libovolný počet m znaků.

○ *Závěr*: Vyšetřili jsme matici dat v 2D- a 3D-rozptylových diagramech a získali řadu závěrů, např. čím vyšší jsou osoby, tím jsou těžší, nebo schopnost plavat koreluje s výškou a hmotností osob. Především bylo zjištěno, že existují minimálně dva shluky, jeden pro muže a jeden pro ženy. Pro některé znaky se ukázaly také shluky odpovídající původu osob. Některé korelace jsou ale typicky falešné, například plavání proti výšce. Diskrétní znaky jdou vyšetřit velmi obtížně a zobrazit je nelze. Docházíme k názoru, že i přes veškerou snadnost a názornost 2D- a 3D-rozptylových diagramů zobrazení všech 9 znaků je časově náročné a pracné. Budeme proto hledat výhodnější postup především v metodě hlavních komponent. Také interpretace není zdaleka tak snadná zejména pro jiné případy. Poněkud lepší situace je v případech, kdy data tvoří vícerozměrný náhodný výběr.

5. ANALÝZA HLAVNÍCH KOMPONENT (PCA)

5.1. Zaměření metody PCA

Metoda hlavních komponent (PCA) je jedna z nejstarších a nejvíce používaných metod vícerozměrné analýzy. Poprvé byla zavedena Pearsonem již v roce 1901 a nezávisle Hotellingem v roce 1933. Cílem analýzy hlavních komponent je především zjednodušení popisu skupiny vzájemně lineárně závislých čili korelovaných znaků čili rozklad zdrojové matice dat do *matice strukturní* a do *matice šumové*. V analýze hlavních komponent nejsou znaky děleny na závisle a nezávisle proměnné jako v regresi. Techniku lze popsat jako metodu lineární transformace původních znaků na nové, nekorelované proměnné, nazvané *hlavní komponenty*. Každá hlavní komponenta představuje lineární kombinací původních znaků. Základní charakteristikou každé hlavní komponenty je její míra variability čili *rozptyl*. Hlavní komponenty jsou seřazeny dle důležitosti, tj. dle klesajícího rozptylu, od největšího k nejmenšímu. Většina informace o variabilitě původních dat je přitom soustředěna do první komponenty a nejméně informace je obsaženo v poslední komponentě. Platí pravidlo, že má-li nějaký původní znak malý či dokonce žádný rozptyl, není schopen přispívat k rozlišení mezi objekty.

5.2. Podstata metody PCA

● **Zdrojová matice dat $\boldsymbol{X}$** $(n \times m)$: Zdrojová matice dat $\boldsymbol{X}$ $(n \times m)$ obsahuje n objektů a m znaků. Objekty jsou pozorování, vzorky, experimenty, měření, pacienti, rostliny, atd., zatímco znaky či proměnné jsou druhy signálu měření, měřená veličina, vlastnosti (sladký, kyselý, hořký, slaný, cholerický, atd.), barva, apod. Důležitá je zde skutečnost, že každý znak je znám pro všech n objektů. Správná skladba zdrojové matice $\boldsymbol{X}$ čili volba, které znaky použít a které objekty zařadit je delikátní úkol silně odvislý od charakteru každé úlohy. Velikou výhodou metody PCA je použití jakéhokoliv počtu proměnných ve zdrojové matici $\boldsymbol{X}$ k vícerozměrné charakterizaci. Cílem každé vícerozměrné analýzy je zpracovat data tak, aby se zřetelně indikoval model a tak odkryl skrytý jev. Myšlenka sledování rozptylu je velice důležitá, protože je vlastně základním

předpokladem vícerozměrné analýzy dat, že “nalezené směry maximálního rozptylu” jsou více či méně spjaty s těmito skrytými jevy. Matematické pozadí metody je detailně popsáno v monografii [21].

● **Zobrazení objektů v m rozměrech**: Zdrojová matice dat $\boldsymbol{X}$ $(n \times m)$ s m sloupci znaků a n řádky objektů může být zobrazena v m-rozměrném eukleidovském prostoru, tj. ortogonálním systému souřadnic rozměru m. Osy proměnných jsou *ortogonální* a mají *společný počátek*, ale mají *různé měrné jednotky*. Prostorem proměnných rozměru m se nazývá m-rozměrný systém, jehož rozměr m se rovná hodnosti zdrojové matice, což matematicky značí počet m nezávislých základních vektorů zdrojové matice a statisticky představuje počet m nezávislých zdrojů proměnlivosti zdrojové matice dat. Vícerozměrná analýza dat slouží ke snížení rozměrnosti a tedy k určení efektivní dimensionality. Na začátku však vždy vycházíme z m rozměrů. Je snahou využívat 1D-, 2D- a 3D-rozměrný prostor, i když vícerozměrný prostor než 3D- je rovněž možný, ale nedá se jednoduše zobrazit.

5.3. Cíl metody hlavních komponent PCA

Základním cílem PCA je transformace původních znaků x_i, $j = 1, \ldots, m$, do menšího počtu latentních proměnných y_j. Tyto latentní proměnné mají vhodnější vlastnosti: je jich výrazně méně, vystihují téměř celou *proměnlivost původních znaků* a jsou vzájemně nekorelované. Latentní proměnné jsou nazvány *hlavními komponentami* a jsou to lineární kombinace původních proměnných: *první hlavní komponenta* y_1 popisuje největší část proměnlivosti čili rozptylu původních dat, *druhá hlavní komponenta* y_2 zase největší část rozptylu neobsaženého v y_1 atd. Matematicky řečeno, první hlavní komponenta je takovou lineární kombinací vstupních znaků, která pokrývá největší rozptyl mezi všemi ostatními lineárními kombinacemi.

Rozdíl mezi souřadnicemi objektů v původních znacích a v hlavních komponentách čili ztráta informace projekcí do menšího počtu rozměrů se nazývá *mírou těsností proložení modelu* PCA nebo také *chybou modelu* PCA.

I při velkém počtu původních znaků m může být k velmi malé, běžně 2 až 5. Volba počtu užitých komponent k vede k *modelu hlavních komponent* PCA. Vysvětlení užitých hlavních komponent, jejich pojmenování a vysvětlení vztahu původních znaků x_j, $j = 1,\ldots,m$, k hlavním komponentám y_j, $j = 1,\ldots,k$, tvoří dominantní součásti analýzy modelu hlavních komponent PCA.

Zdrojová centrovaná matice $\boldsymbol{X}_C$ se rozkládá na matici komponentních skóre $\boldsymbol{T}$ rozměru $n \times k$ a matici komponentních zátěží $\boldsymbol{P}_T$ rozměru $k \times m$.

Model PCA odpovídá *aproximaci* zdrojové matice dat $\boldsymbol{X}$, který užijeme místo původní zdrojové matice dat $\boldsymbol{X}$. Aproximace má řadu výhod v interpretaci dat. Nejde zde pouze o změnu systému souřadnic, ale především o nalezení a vypuštění šumu. PCA má proto dvojí cíl: transformace do nového systému os a snížení rozměrnosti úlohy užitím několika prvních hlavních komponent, které vystihují strukturu v datech. Problémem zůstává, kolik hlavních komponent je nutno použít. Statistická analýza metody je detailně popsána v monografii [21].

● **Maximální počet hlavních komponent**: Existuje horní mez počtu hlavních komponent, které mohou být odvozeny ze zdrojové matice dat $\boldsymbol{X}$. Největší počet hlavních komponent se buď rovná číslu $n - 1$ nebo m v závislosti na tom, které z těchto dvou čísel je menší. Je-li $\boldsymbol{X}$ složena například $n = 40$ spekter měřených při $m = 2000$ vlnových délek, bude maximální počet hlavních komponent 39. Počet efektivních hlavních komponent se rovná *hodnosti zdrojové matice* $\boldsymbol{X}$.

● **X = Struktura + šum**: Všechny hlavní komponenty jsou vzájemně ortogonální a souvisí postupně se snižující hodnotu rozptylu objektů. Poslední hlavní komponenta souvisí s nejmenším rozptylem, tj. *stochastickým rozptylem*. Vyšší hlavní komponenty (obvykle vyšší než 3) se často týkají pouze šumu. Dochází k rozdělení původní zdrojové matice $\boldsymbol{X}$ na *část struktury* (první hlavní komponenty) a *část šumu* (ostatní zbývající hlavní komponenty). Model hlavních komponent má pak tvar

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\boldsymbol{T}} + \boldsymbol{E} = struktura\ dat + šum \qquad (12)$$

Zdrojovou matici dat X je třeba nejprve centrovat $x_{ik} = x_{ik}$ - . V metodě hlavních komponent pracujeme za předpokladu, že zdrojová matice $\boldsymbol{X}$ je rozložena na součin matic TPT a na matici reziduí $\boldsymbol{E}$, kde matice $\boldsymbol{T}$ je matice komponentního skóre a PT je transponovaná matice komponentních vah, E je matice reziduí. Cílem metody hlavních komponent je místo $\boldsymbol{X}$ využívat dále jenom součin TPT a tak oddělit šum od struktury dat TPT. Matici $\boldsymbol{E}$ se nazývá matice reziduí, která není objasněna modelem hlavních komponent PCA. Matice $\boldsymbol{E}$ souvisí s "těsností proložení" a ukazuje, jak dobře jsou objekty proloženy modelem hlavních komponent.

Rovnici (12) je možno rozepsat do tvaru

$$\boldsymbol{X} = \boldsymbol{t_1 p_1^T} + \boldsymbol{t_2 p_2^T} + \cdots + \boldsymbol{t_A p_A^T} + \boldsymbol{E} \tag{13}$$

Každý sčítanec $\boldsymbol{t_1}\ \boldsymbol{p_1^T}$ je matice rozměru $n \times m$, která má hodnost 1. Postupný výpočet hlavních komponent se rozkládá do těchto kroků:

1. Vyčíslí se $\boldsymbol{t_1}$ a $\boldsymbol{p_1^T}$ z matice $\boldsymbol{X}$ například s využitím SVD [20,21] a dostane se $\boldsymbol{X} = \boldsymbol{t_1}\,\boldsymbol{p_1^T} + \boldsymbol{E_1}$.

2. Odečte se příspěvek první hlavní komponenty PC1 od $\boldsymbol{X}$ dle rovnice $\boldsymbol{E_1} = \boldsymbol{X} - \boldsymbol{t_1}\,\boldsymbol{p_1^T}$.

3. Vyčíslí se $\boldsymbol{t_2}$ a $\boldsymbol{p_2^T}$ z matice $\boldsymbol{E_1}$ a dostane se $\boldsymbol{X} = \boldsymbol{t_1}\,\boldsymbol{p_1^T} + \boldsymbol{t_2}\,\boldsymbol{p_2^T} + \boldsymbol{E_2}$.

4. Odečte se příspěvek druhé hlavní komponenty *PC2* od $\boldsymbol{E_2}$ dle rovnice $\boldsymbol{E_1} = \boldsymbol{E_2} - \boldsymbol{t_2}\,\boldsymbol{p_2^T}$.

5. Podobně se pokračuje, až se vyčíslí dostatečný počet A hlavních komponent.

● **Střed modelu**: Hlavní komponenty mají společný počátek, který *odpovídá průměrnému objektu* čili těžišti celého shluku objektů. Postup se nazývá *centrování*. Tento bod je pouhou abstrakcí stejně jako latentní proměnné hlavní komponenty.

● **Komponentní váhy**, **zátěže - vztah mezi $\boldsymbol{X}$ a *PC***: Hlavní komponenty *PC* jsou vhodně škálované vektory v prostoru znaků. Kterákoliv hlavní komponenta

představuje *lineární kombinaci všech* m vektorů v prostoru znaků, tj. jednotkové vektory podél každé osy původního znaku v m rozměrném prostoru. Lineární kombinace v každé hlavní komponentě bude obsahovat m koeficientů p_{ka}, kde k je index m-tého znaku a a je index směru hlavní komponenty. Například p_{23} znamená koeficient pro druhý znak v lineární kombinaci, která vytvoří PC3. Tyto koeficienty se nazývají komponentní váhy. Váhy pro všechny hlavní komponenty tvoří matici $\boldsymbol{P}$. Tato matice je vlastně transformační maticí, která převádí původní znaky zdrojové matice $\boldsymbol{X}$ do nových latentních proměnných, tj. hlavních komponent. *Vektory vah* čili sloupce v matici $\boldsymbol{P}$ jsou ortogonální.

Váhy informují o vztahu mezi původními m znaky a hlavními komponentami. Tvoří tak most mezi prostorem původních znaků a prostorem hlavních komponent. Váhy souvisejí se směrovými kosiny každé hlavní komponenty vzhledem k systému os původních znaků.

Na *grafu komponentních vah* p pro PC1 a PC2 jsou místo objektů jejich znaky. Tak lze vyšetřovat závislosti a podobnosti mezi znaky. Tento graf rovněž ukazuje, jak každý původní znak přispívá do každé hlavní komponenty. Je třeba si uvědomit, že hlavní komponenty představují lineární kombinace jednotkových vektorů původních znaků. Komponentní váhy představují koeficienty v těchto lineárních kombinacích. Každý původní znak může přispívat do více než jedné hlavní komponenty. Na x-ové ose je patrné jak jednotlivé původní znaky přispívají do první hlavní komponenty. Analogicky na y-ové ose je vidět jak jednotlivé znaky přispívají do druhé hlavní komponenty. Některé znaky zobrazují *kladnou váhu*, pak jde o kladné koeficienty v lineárních kombinacích, zatímco jiné znaky zobrazují *zápornou váhu*. Jsou-li v grafu komponentních vah znaky blízko sebe, znamená to, že spolu *silně korelují*. Jsou-li naopak daleko od sebe, nekorelují.

● **Komponentní skóre - souřadnice objektů v prostoru hlavních komponent:**

Souřadnice každého objektu na osách hlavních komponent nazýváme *skóre*. Projekce i-tého objektu na první hlavní komponentu *PC1* značí skóre ti1. Projekce téhož objektu na druhou hlavní komponentu *PC2* značí skóre t_{i2}, atd.

Každý objekt má svůj soubor komponentních skóre $t_{i1}, t_{i2}, \ldots, t_{ip-1}$. Hodnot skóre je stejný počet jako hlavních komponent.

Matice všech skóre pro všechny objekty se nazývá *matice skóre* $\boldsymbol{T}$. Skóre pro jeden objekt v této matici tvoří jeden řádek. Sloupce v této matici jsou ortogonální. *Vektor skóre* je sloupec v matici $\boldsymbol{T}$ a obsahuje skóre pro jednu hlavní komponentu.

Jedním z nejdůležitějších grafů metody hlavních komponent je *graf komponentního skóre*. Jde o zobrazení dvou skórových vektorů vynesených v systému kartézských os jeden proti druhému. Skórové vektory zde představují znázornění objektů na hlavních komponentách. Vynesení skórových vektorů odpovídá vynesení objektů v prostoru hlavních komponent. Nejužívanějším grafem ve vícerozměrné analýze dat je vektor skóre *PC1* proti skóre *PC2*. Je snadno k pochopení, protože jde o dva směry, podél kterých shluk objektů vykazuje největší (*PC1*) a druhé největší (*PC2*) rozptýlení.

Graf komponentního skóre t_1 proti t_2 se obvykle vyšetřuje jako první. Do tohoto grafu lze vynášet libovolný pár hlavních komponent. Otázkou však zůstává, které hlavní komponenty jsou ty nejdůležitější. Je-li počet znaků m menší než počet objektů n, lze vynést $m(m-1)/2$ možných 2D-grafů komponentního skóre. Počet možných grafů se tak stává nekontrolovatelným a není možné vyšetřovat všechny grafy. Uvedeme si pravidlo k volbě grafů komponentního skóre:

1. Na x-ové ose užijeme vždy stejnou hlavní komponentu (obvykle první) u všech grafů komponentního skóre: t_1 proti t_2, t_1 proti t_3, t_1 proti t_4, t_1 proti t_5, ... atd., takže budeme vyšetřovat ostatní hlavní komponenty proti stále stejné, první. To nejlépe pomůže získat požadovaný přehled o hledané struktuře dat.

2. Užijeme tu hlavní komponentu, která vykazuje největší hodnotu rozptýlení pro vyšetřovanou úlohu a vyneseme ji na x-ovou osu. U mnoha úloh se ukáže, že jde o *PC1*. Obecně lze říci, že *PC1* popisuje největší strukturní změnu v libovolném souboru dat. Korelace však není totožná s kauzalitou.

Obě pravidla jsou velmi obecná a mají řadu výjimek. Existuje také řada úloh, kdy konkrétní informace je skryta v řadě ostatních hlavních komponent. Konkrétní informace pak lze odhalit v jedné či několika prostředních hlavních komponentách.

● **Rezidua objektů**: Metoda hlavních komponent PCA přináší mnoho výhod v analýze zdrojové matice dat $\boldsymbol{X}$ při nahrazení m původních znaků objektů skóry hlavních komponent. Velikost projekční vzdálenosti e_i představuje určitou "ztrátu informace" vlivem aproximace původního souboru dat. Vzdálenost e_i se nazývá *reziduum* a všechna rezidua objektů jsou obsažena v *matici reziduí* $\boldsymbol{E}$. Velikosti reziduí jsou v přímé návaznosti na využitý počet hlavních komponent A stejně jako na podíl odhalené struktury v datech. "Velká rezidua" ukazují, že proložení objekty není nejlepší a model nedostatečně popisuje data. "Malá rezidua" značí dobrý model. Existuje mnoho statistických kritérií k posouzení těsnosti proložení technikou statistické analýzy reziduí. Platí obecné pravidlo: malá hodnota A čili málo využitelných hlavních komponent značí hodně zbývajícího šumu v matici $\boldsymbol{E}$, velká hodnota A méně ponechaného šumu v matici $\boldsymbol{E}$.

Matice $\boldsymbol{E}$ obsahuje jak malá tak i velká rezidua. Vyhodnocení $\boldsymbol{E}$ je vždy relativní vůči celkovému rozptylu. Na základě průměru každého znaku se určí počátek dat, společný všem hlavním komponentám (0, 0, ..., 0), který se *nazývá nulová hlavní komponenta*, a který vlastně popisuje průměrný objekt.

Odečtením průměrného objektu od zdrojové matice $\boldsymbol{X}$ je stejné, jako centrování zdrojové matice dat. Pro $\boldsymbol{A} = \boldsymbol{0}$ je matice reziduí $\boldsymbol{E}_0$ a bude rovna centrované matici $\boldsymbol{X}_C$. Matice $\boldsymbol{E}_0$ hraje důležitou referenční roli při kvantitativním posouzení relativní velikosti $\boldsymbol{E}$.

Rezidua se budou měnit při přidávání dalších hlavních komponent. Index i u písmene $\boldsymbol{E}_i$ bude vždy značit počet hlavních komponent vyčíslených v modelu hlavních komponent PCA. Výsledná rezidua se porovnávají s rezidui matice $\boldsymbol{E}_0$. Pro $\boldsymbol{A} = 0$ bude $\boldsymbol{E}_0$ představovat 100%. *Reziduálový rozptyl* bude pak 100% a *modelem objasněný rozptyl* bude 0%. Velikost $\boldsymbol{E}$ je charakterizována čtverci reziduí čili rozptyly. Existují dva způsoby jak sčítat prvky matice $\boldsymbol{E}$: buď v řádku a

obdržet rezidua každého objektu, nebo ve sloupci a obdržet rezidua každého znaku.

Rezidua objektu: čtverec reziduí i-tého objektu e_i^2 je dán vzorcem

$$e_i^2 = \sum_{k=1}^{p} e_{ik}^2 \tag{14}$$

Reziduální rozptyl i-tého objektu je pak roven e_i^2/p. Druhá odmocnina tohoto rozptylu bude představovat vzdálenost mezi i-tým objektem a modelovou nadrovinou použitých A hlavních komponent, vyjádřenou v prostoru původních znaků. Reziduum objektu je mírou vzdálenosti mezi objektem v prostoru znaků a modelovou projekcí objektu do prostoru hlavních komponent. Čím je tato vzdálenost menší, tím je těsnější modelová projekce původního objektu. Jinými slovy, řádky v $\boldsymbol{E}$ indikují, jak dobře model prokládá objekt za použití právě A hlavních komponent.

Celkový čtverec reziduí všech objektů: Modelem hlavních komponent se prokládají *všechny objekty* současně. Je třeba definovat *celkový čtverec reziduí všech objektů* vztahem

$$e_{tot}^2 = \sum_{i=1}^{n} e_i^2 \tag{15}$$

Celkový rozptyl reziduí je pak roven $e_{tot}^2/(p\ n)$.

5.4. Grafické pomůcky analýzy hlavních komponent

Výsledek analýzy hlavních komponent lze zobrazit v několika různých typech grafů.

5.4.1. *Cattelův indexový graf úpatí vlastních čísel (Scree Plot)*

Je vlastně sloupcovým diagramem vlastních čísel $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$ zdrojové matice dat $\boldsymbol{X}$ v závislosti na indexu i (obr. 4.2a). Zobrazuje relativní velikost jednotlivých vlastních čísel. Řada autorů ho s oblibou využívá k určení počtu k „významných“ hlavních komponent. Úpatí je zlomové místo mezi kolmou stěnou a vodorovným dnem. Nevýznamné hlavní komponenty (nebo faktory) představují vodorovné dno. Významné komponenty jsou tak odděleny

zřetelným zlomovým místem, úpatím a hodnota indexu i tohoto zlomu udává počet významných komponent. Alternativou je graf úpatí pro $\ln \lambda_j$. Určení počtu využitelných hlavních komponent je jedním z důležitých úkolů v analýze hlavních komponent a uplatňuje se například ve spektroskopii při hledání počtu světlo-absorbujících částic analýzou hodnosti absorbanční matice. Protože jsou hlavní komponenty seřazeny dle klesajícího rozptylu, můžeme vybrat prvních několik hlavních komponent k jako představitele m-tice původních znaků. Existuje také celá řada jednoduchých pravidel, které vyhovují praktickému použití:

a) Výběr pouze k-tice komponent, které objasňují zadané procento P variability původních znaků. Obyčejně se volí $P = 70$ až 90 %. To znamená, že se vybírá k takové, že platí

$$100 \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \geq P \tag{16}$$

b) Vyloučení těch komponent, kterým odpovídají vlastní čísla menší než průměrná hodnota $(\sum_{i=1}^{m} \lambda_i)/m$. Protože $\sum_{i=1}^{m} \lambda_i = \mathrm{tr}\, \boldsymbol{S}$, je průměrná hodnota vlastních čísel ekvivalentní průměrné hodnotě rozptylu původních znaků. Vybírají se tedy ty komponenty, jejichž variabilita je nadprůměrná. Pokud se provádí PCA pro korelační matici $\boldsymbol{R}$, je $\mathrm{tr}\, \boldsymbol{R} = m$ a průměrná hodnota je rovna jedné. To vede k jednoduchému pravidlu vyloučit hlavní komponenty, pro které jsou vlastní vektory menší než 1. Na základě simulací bylo zjištěno, že je lépe vypouštět komponenty, pro které jsou vlastní čísla menší než 0.7. Při použití tohoto pravidla pro korelační matici stačí provést normalizaci $\lambda_j / \sum_{i=1}^{m} \lambda_i$. Je možné také využít efektivního rozptylu $D_e = 1 - (\det \boldsymbol{R})^{1/m}$. Počet významných hlavních komponent k je pak roven

$$k = m(0.8 - 0.5 D_e) \tag{17}$$

Toto kritérium odpovídá výběru hlavních komponent objasňujících 90 % variability [6].

c) Pravidlo „zlomené hole“ Cattelova indexového grafu úpatí vychází z faktu, že pokud je jednotkový úsek náhodně rozdělen na m částí, je očekávaná délka k-té části rovna

$$g_k = \frac{1}{m}\sum_{i=k}^{m}\frac{1}{i} \tag{18}$$

Do PCA se pak zahrnují ty komponenty, pro které platí, že $\lambda_k > g_k$.

5.4.2. *Graf komponentních vah, zátěží (Plot Components Weights)*

Zobrazuje komponentní váhy čili zátěže pro první dvě hlavní komponenty (obr. 4.30. Každý bod v grafu odpovídá jednomu znaku a v grafu se porovnávají vzdálenosti mezi znaky. Krátká vzdálenost mezi dvěma znaky znamená silnou korelaci. Lze nalézt i shluk podobných znaků, jež spolu korelují. Graf můžeme považovat za most mezi původními znaky a hlavními komponentami, protože ukazuje, jakou měrou přispívají jednotlivé původní znaky do hlavních komponent. Někdy se podaří hlavní komponenty pojmenovat, vysvětlit a přidělit jim fyzikální, chemický nebo biologický význam. Pak lze názorně vysvětlit, jak jednotlivé původní znaky $x_j, j = 1,\ldots,m$, přispívají do první hlavní komponenty nebo do druhé hlavní komponenty. Některé původní znaky x_j přispívají kladnou vahou, jiné zápornou. Bývá zajímavé sledovat kovarianci původních znaků x_j v prostorovém 3D-grafu prvních tří hlavních komponent. Jsou-li znaky x_j, $j = 1,\ldots,m$, blízko sebe v prostorovém shluku, jde o silnou pozitivní kovarianci. Kovariance však nemusí ještě nutně znamenat korelaci. Interpretaci grafu komponentních vah lze obecně shrnout do těchto bodů:

a) *Důležitost původních znaků* x_j, $j = 1,\ldots,m$. Znaky x_j s vysokou mírou proměnlivosti v objektech mají vysoké hodnoty komponentní váhy. Ve 2D-diagramu prvních dvou hlavních komponent pak leží hodně daleko od počátku. Znaky s malou důležitostí leží blízko počátku. Když určíme *důležitost znaků*, určíme tím také variabilitu znaků: jestliže například první hlavní komponenta y_1 objasňuje 70 % proměnlivosti a druhá hlavní komponenta y_2 jenom 5 % (určeno z indexového grafu úpatí vlastních čísel), jsou původní znaky xj, $j = 1,\ldots,m$, s vysokou váhou v y_1 tím pádem mnohem důležitější než znaky x_j

s vysokou váhou v y_2. Znaky s úhlem 0° mezi průvodiči grafu jsou silně pozitivně korelované, znaky s úhlem 90° jsou zcela nekorelované, zatímco znaky s úhlem 180° jsou negativně korelované.

b) *Korelace a kovariance*. Původní znaky x_j, $j = 1, \ldots, m$, jsou blízko sebe, anebo znaky x_j s malým úhlem mezi svými průvodiči znaků a na stejné straně vůči počátku mají vysokou kladnou kovarianci a vysokou kladnou korelaci. Naopak, původní znaky x_j daleko od sebe, anebo s velikým úhlem mezi průvodiči znaků, jsou negativně korelovány.

c) *Spektroskopická data*. Ve spektroskopických datech je graf komponentních vah často nejvhodnější. I zde platí pravidlo, že vysoké komponentní váhy představují vysokou důležitost znaků x_jj (například vlnových délek).

5.4.3. *Rozptylový diagram komponentního skóre (Scatterplot)*

Zobrazuje komponentní skóre $\boldsymbol{t}_1$, $\boldsymbol{t}_2$ obyčejně pro první dvě hlavní komponenty u všech objektů (obr. 4.4). Body v tomto grafu jsou t_{1i}, t_{2i}, $i = 1, \ldots, n$. Dokonalé rozptýlení objektů v rovině obou hlavních komponent ukazuje na dokonalé rozlišení objektů. Lze snadno nalézt shluk vzájemně podobných objektů a dále objekty odlehlé a silně odlišné od ostatních objektů. Diagram komponentního skóre však může být i ve třech či více hlavních komponentách a v rovinném grafu se pak sleduje pouze jeho průmět do roviny. Tento diagram se užívá k identifikaci odlehlých objektů, identifikaci trendů, identifikaci tříd, shluků objektů, k objasnění podobnosti objektů atd. Nelze analyzovat všechny možné rozptylové diagramy, protože jich je velmi mnoho. Například pro $m = 10$ znaků, existuje $m(m-1)/2 = 45$ diagramů, pro $m = 11$ pak 55 diagramů, pro $m = 12$ pak 66 diagramů atd. Obvykle se vybírají diagramy závislosti y_1 na y_2, y_1 na y_3, y_1 na y_4 atd. Pro identifikaci odlehlých bodů se někdy doporučuje graf pro poslední dvě hlavní komponenty, $\boldsymbol{t}_k$, $\boldsymbol{t}_{k-1}$. Výklad rozptylového diagramu komponentního skóre se týká

a) *Umístění objektů*. Objekty daleko od počátku, respektive centra jsou extrémy. Objekty nejblíže počátku jsou nejtypičtější.

b) *Podobnosti objektů*. Objekty blízko sebe si jsou podobné, objekty daleko od sebe jsou si nepodobné.
c) *Objektů v shluku*. Objekty umístěné zřetelně v jednom shluku jsou si podobné a přitom nepodobné objektům v ostatních shlucích. Dobře oddělené shluky prozrazují, že lze nalézt vlastní model pro samotný shluk. Jsou-li shluky blízko sebe, znamená to značnou podobnost objektů.
d) *Osamělých objektů*. Izolované objekty mohou být odlehlými objekty, které jsou silně nepodobné ostatním objektům. To platí, pokud se nejedná o zdánlivou nehomogenitu danou zešikmením dat a odstranitelnou transformací znaků.
e) *Odlehlých objektů.* V ideálním případě bývají objekty rozptýlené po celé ploše diagramu. V opačném případě je něco špatného v modelu, obyčejně je přítomen silně odlehlý objekt. Odlehlé objekty jsou totiž schopny zbortit celý diagram, ve srovnání se silně vybočujícím objektem jsou ostatní objekty nakumulovány do jediného úzkého shluku. Po odstranění vybočujícího objektu se ostatní objekty buď rozptýlí po celé ploše diagramu, anebo vypovídají o existujících shlucích.
f) *Pojmenováním objektů.* Výstižná jména objektů slouží k hledání hlubších souvislostí mezi objekty a mezi pojmenovanými hlavními komponentami. Snadno lze ohraničit shluky podobných objektů nebo nakreslením spojky mezi objekty vystihnout jejich fyzikální či biologický vztah.
g) *Vysvětlením místa objektu*. Umístění objektu v diagramu může být porovnáváno s komponentními váhami původních znaků ve dvojném grafu a pomocí původních znaků pak i vysvětleno.

5.4.4. *Dvojný graf (Biplot)*

Kombinuje předchozí dva grafy (obr. 4.6). Úhel mezi průvodiči dvou znaků x_j a x_k je nepřímo úměrný velikosti korelace mezi těmito dvěma znaky. Čím je menší úhel, tím je větší korelace. Každý průvodič má své souřadnice na první a na druhé hlavní komponentě. Délka této souřadnice je úměrná příspěvku původního znaku x_j do hlavní komponenty, čili je úměrná komponentní váze. Kombinace obou grafů do jediného společného přináší cenné srovnání, jeden graf zde působí jako doplněk vůči druhému. Když se ve dvojném grafu nachází

objekt v blízkosti určitého znaku x_j, znamená to, že tento objekt „obsahuje“ velký podíl právě tohoto znaku a je s ním v interakci. Interakce znaků a objektů umožňuje také vysvětlit umístění objektů vpravo od nuly na ose y_1 (či vlevo od nuly) pomocí pozice znaků v tomto grafu, popř. umístění nahoře od nuly (či dole od nuly) na ose y_2.

Modifikací je 2D-graf s osami odpovídajícími prvním dvěma hlavním komponentám. Body zde indikují objekty a vektory jsou projekce znaků. Jednotlivá skóre $\boldsymbol{t}_1$ a $\boldsymbol{t}_2$ obsahují souřadnice bodů, které se obvykle označují čísly objektů, stejně jako v rozptylovém grafu komponentních skóre. Vektory zde obsahují souřadnice délek projekcí p_{ij} i-tého bodu do j-té hlavní komponenty. Tyto projekce jsou vyjádřeny v 2D-podprostoru n rozměrného prostoru pro znaky, když každý znak má n složek. Podobně jsou komponentní skóre vynesena v 2D podprostoru m rozměrného prostoru objektů [7]. Jako měřítko na osách se volí měřítko odpovídající komponentním skóre, takže body leží uvnitř elipsy, ohraničené křivkou

$$\frac{p_{i1}^2}{\hat{\lambda}_1} + \frac{p_{i2}^2}{\hat{\lambda}_2} = 1 \tag{19}$$

Interpretace tohoto grafu je shodná s interpretací dvojného grafu.

5.4.5. *Graf reziduí jednotlivých objektů*

Rozptyl reziduí jednoho i-tého objektu představuje vzdálenost mezi tímto objektem a modelem (obr. 4.5). Je vhodné zobrazovat tuto veličinu pro všechny objekty v jednom společném grafu a odhalit tak odlehlé objekty či jiné anomálie objektů. Toto zobrazení je zvláště výhodné, když potřebujeme porovnat rezidua jednotlivých objektů mezi sebou. Dosahuje-li reziduum určitého objektu na y-nové ose proti ostatním objektům abnormálně velké hodnoty, půjde pravděpodobně o vybočující objekt.

5.4.6. *Graf celkového reziduálového rozptylu všech objektů*

Výstavba modelu hlavních komponent se provádí postupným přidáváním hlavních komponent. Začíná se s centrovanou maticí dat $\boldsymbol{X}_C$, označenou jako

celkový rozptyl reziduí $\boldsymbol{E}_0$. Pak se vyčíslí $\boldsymbol{t}_1$ a $\boldsymbol{p}_1$ a odečte se $\boldsymbol{E}_0$ a dostane se $\boldsymbol{E}_1$. Matice $\boldsymbol{E}_1$ poskytne novou hodnotu celkového rozptylu reziduí $e^2_{tot,1}$ která musí být menší v porovnání s předešlou hodnotou $e^2_{tot,0}$. Postupným přidáváním dalších hlavních komponent se hodnota celkového rozptylu reziduí e2tot,i bude zmenšovat tak, že nová hodnota $e^2_{tot,i+1}$ bude vždy menší než předešlá. Vynesením hodnoty $e^2_{tot,i}$proti počtu použitých hlavních komponent obdržíme *indexový graf úpatí celkového reziduálového rozptylu všech objektů* (obr. 4.2b). Sestupná křivka tohoto grafu úpatí se bude blížit k nule, když počet hlavních komponent A se bude rovnat největšímu číslu, tj. hodnoty $\min(n, p)$. Graf celkového reziduálového rozptylu je obdobou grafu úpatí vlastních čísel a nalezení zlomu na sestupné křivce čili úpatí analogicky vystihuje optimální počet A využitelných hlavních komponent.

6. FAKTOROVÁ ANALÝZA (FA)

6.1. Zaměření metody FA

Faktorová analýza je vícerozměrná technika k vyšetření vnitřních souvislostí a vztahů čili korelace a odhalení základní struktury zdrojové matice dat. Týká se analýzy struktury vnitřních vztahů mezi velkým počtem původních znaků využitím souboru menšího počtu latentních proměnných, zvaných *faktory*. Nejprve jsou identifikovány faktory, a pak je každému faktoru přidělen obsahový, obvykle fyzikální, význam, pomocí kterého je každý původní znak vysvětlen vybraným faktorem. Jde o dva primární cíle faktorové analýzy, a to jednak sumarizaci a jednak redukci dat. V *sumarizaci dat* využívá faktorová analýza faktorů tak, aby data vysvětlila a usnadnila jejich pochopení daleko menším počtem latentních proměnných, než je počet původních znaků. *Redukce dat* je dosaženo vyčíslením skóre pro každý faktor a následnou náhradou původních znaků novými latentními proměnnými – faktory.

Podobně jako metoda hlavních komponent patří faktorová analýza mezi metody snížení dimenze čili redukce počtu původních znaků. Ve faktorové analýze předpokládáme, že každý vstupující znak můžeme vyjádřit jako lineární kombinaci nevelkého počtu společných skrytých faktorů a jediného specifického faktoru. Na rozdíl od PCA se ve faktorové analýze snažíme vysvětlit závislost znaků. K nevýhodám metody patří zejména nutnost zvolit *počet společných faktorů* ještě před prováděním vlastní analýzy.

Demonstrujme si nejdříve myšlenku faktorové analýzy na příkladu dvou znaků x_1, x_2, které jsou oba ovlivněny latentním, společným faktorem F a navíc je každý z obou znaků ovlivněn samostatně faktorem e_1, popř. e_2. Odpovídající model má pak tvar

$$x_i = a_i F + b_i e_i, \qquad i = 1{,}2. \tag{20}$$

Konstanty úměrnosti a_i, b_i se nazývají *faktorové zátěže*. Platí, že faktory e_i a faktor F jsou vzájemně nekorelované, tedy

$$\mathrm{cov}(e_1\ e_2) = 0,\ \mathrm{cov}(e_1\ F) = 0, \qquad i = 1{,}2. \tag{21}$$

Navíc jsou x_i, e_i, F vyjádřené ve standardizovaném tvaru, tj. odečtené od průměru a dělené směrodatnou odchylkou. Rozptyl i-tého znaku bude

$$\sigma^2(x_i) = a_i^2 + b_i^2 = 1. \tag{22}$$

Podíl rozptylu, připadající na společný faktor F, je a_i^2 a označuje se jako *komunalita* a podíl rozptylu, připadající na samostatný faktor e_i, je b_i^2 a označuje se jako *specifita*. Pro kovarianci respektive korelaci mezi x_i a F platí, že

$$\mathrm{cov}(x_i, F) = a_i \tag{23}$$

a pro kovarianci (korelaci) mezi x1 a x2 platí, že

$$\mathrm{cov}(x_1, x_2) = a_1 a_2. \tag{24}$$

Známe-li faktorové zátěže, můžeme snadno určit korelace mezi původními znaky. Na druhé straně však znalost korelací mezi původními znaky nemusí vést k jednoznačnému určení

faktorových zátěží. Platí, že různé typy modelů, rozšiřujících jednoduchý model (20) na více původních znaků, poskytují stejné korelace. Lze definovat dvě krajní situace:

1. Korelační struktura původních znaků je tvořena faktorovými modely se stejným počtem společných faktorů ale s různými faktorovými zátěžemi.
2. Korelační struktura původních znaků je tvořena faktorovými modely s různým počtem společných faktorů.

Existuje řada různých faktorových modelů objasňujících u stejných dat korelační strukturu původních znaků. Ne každý model je však stejně informativní. Pro dobrou interpretaci faktorů F se většinou požaduje co nejjednodušší struktura. Vůbec nejjednodušší je taková struktura, má-li každý znak nenulovou zátěž pouze pro jeden faktor. Struktura vyjádřená rovnicí (20) je proto nejjednodušší. Je snahou, aby každý společný faktor definoval rozdílné skupiny vzájemně korelovaných původních znaků. Taková procedura se označuje jako *rotace faktorů*. V souřadném systému, jehož osy představují jednotlivé společné

faktory, které jsou nekorelované, tvoří znaky x_i, $i = 1, \ldots, m$, shluky a účelem rotace je potočit souřadný systém tak, aby byly jednotlivé shluky co nejblíže nových os. Rotace, kdy se zachovává pravý úhel mezi společnými faktory, se označuje jako *ortogonální rotace*. Pokud se jednotlivé osy pootáčejí nezávisle na sobě, není již zachován pravý úhel a jde o *zobecněnou* (oblique) *rotaci*. Obecný faktorový model má celkem m znaků $x_1, \ldots, x_m$ a p společných faktorů $F_1, \ldots, F_p$. Jednotlivé faktorové zátěže se pak uspořádají do tzv. *faktorové matice* rozměru $m \times p$, jejíž prvek a_{ij} odpovídá zátěži j-tého společného faktoru pro znak x_i. Uveďme, že komunalita původního znaku x_j je dána součtem

$$\sigma^2(x_i) = \sum_{j=1}^{p} a_{ij}^2 = 1. \tag{25}$$

Jednotlivé metody rotace pak souvisejí s rozptyly čtverců zátěží. *Varimaxová rotace* maximalizuje rozptyl zátěží v každém sloupci faktorové matice. Maximalizuje se výraz

$$C_j = \frac{1}{m} \sum_{i=1}^{m} \left(a_{ij}^2 + \bar{a}_j^2\right)^2, \quad j = 1, \ldots, p, \tag{26}$$

kde $\bar{a}_j$ je aritmetický průměr v j-tém sloupci. Hodnota C_j se blíží k nule, pokud jsou zátěže ve sloupci stejné, a blíží se k maximu, pokud jedna zátěž nabývá hodnot blízkých jedné.

Alternativou je *quartimaxová rotace*, kdy se maximalizuje rozptyl zátěží v každém řádku, tj. výraz

$$d_i = \frac{1}{p} \sum_{j=1}^{p} \left(a_{ij}^2 + \bar{a}_j^2\right)^2, \quad j = 1, \ldots, m, \tag{27}$$

kde $\bar{a}_j$ je aritmetický průměr v j-tém řádku.

Podle toho, které znaky se vzájemně ovlivňují, rozeznáváme tři základní skupiny společných faktorů:

a) Obecné společné faktory, které mají vysoké zátěže pro všechny znaky.
b) Skupinové faktory, které mají vysoké zátěže pro skupinu znaků. Pro dva znaky se takové faktory označují jako *doublety*.

c) Specifické obecné faktory, které mají vysokou zátěž pouze pro jeden znak.

Jednotlivé společné faktory se pak interpretují podle toho, pro které znaky mají vysoké zátěže.

Faktorová analýza se obyčejně dělí do dvou kategorií. Uvedené úvahy se týkaly modelu *klasické faktorové analýzy FA*, který lze vyjádřit obecně vztahem

$$x_i = \sum_{j=1}^{m} a_{ij} F_j + b_i e_i. \tag{28}$$

Druhým modelem je vlastně *analýza hlavních komponent PCA*, kdy se nepoužívají žádné předpoklady o struktuře znaků, a společnými, latentními faktory jsou hlavní komponenty $y_1, \ldots, y_p$. Tento model je pak ve tvaru

$$x_i = \sum_{j=1}^{m} b_{ij} y_j. \tag{29}$$

kde b_{ij} jsou zátěže.

Základní rozdíl mezi klasickou faktorovou analýzou FA a metodou hlavních komponent PCA je ve specifikaci korelační matice $\boldsymbol{R}$ původních znaků rozměru $m \times m$. Pokud jsou na diagonále této matice hodnoty 1 (korelace znaku samotného se sebou je 1), vede faktorová analýza této matice na analýzu hlavních komponent.

Pokud se však na diagonále korelační matice dosadí odhady komunalit, odstraní se jednoznačnost každého znaku a analyzuje se pouze jejich část zahrnutá ve faktorovém modelu. Komunality lze však určit až po faktorové analýze, a proto se používají různé odhady:

a) největší párová korelace znaku s ostatními,
b) průměrná párová korelace znaku s ostatními (averoid),
c) vícenásobný korelační koeficient,
d) komunalita pro původní faktorovou analýzu, kdy jsou na diagonále hodnoty jedna.

Komunality je možné iterativně zpřesňovat. Faktorová analýza je vlastně typem regrese, kdy všechny znaky $x_1, x_2, \ldots, x_m$ jsou uvažovány jako závisle proměnné

a nezávisle proměnné jsou latentní faktory $F_1, \ldots, F_p$. Je také možné uvažovat model

$$F_j = \sum_{i=1}^{m} a_{ij} x_j. \tag{30}$$

Veličiny F_j se pak nazývají faktorová skóre a používají se místo původních znaků v řadě vícerozměrných statistických metod.

6.2. Podstata metody faktorové analýzy FA

Vyjdeme z matice $\boldsymbol{X}_R$ rozměru $n \times m$ standardizovaných dat $x_{R_{ij}} = (x_{ij} - \bar{x}_j)/s_j$, kde x_{ij} jsou prvky původní zdrojové matice dat $\boldsymbol{X}$. Protože se standardizace provádí téměř vždy, vynecháme v tomto odstavci symbol R. Matice $\boldsymbol{X} = \boldsymbol{X}_R$ obsahuje m sloupců znaků $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ a n řádků objektů. Všechny znaky mají vzhledem ke standardizaci střední hodnotu 0 a rozptyly rovné 1 a navíc jejich kovariance jsou vlastně párové korelační koeficienty. Předpokládejme, že $x_i^T = (x_1, x_2, \ldots, x_m)^T$ je jeden obecný i-tý objekt pro dané znaky s korelační maticí $\boldsymbol{R}$. Tento objekt je v modelu faktorové analýzy (FA) vyjádřen vztahy

$$\begin{aligned} x_1 &= a_{11}F_1 + a_{12}F_2 + \cdots + a_{1p}F_p + e_1 \\ x_2 &= a_{21}F_1 + a_{22}F_2 + \cdots + a_{2p}F_p + e_2 \\ &\vdots \\ x_m &= a_{m1}F_1 + a_{m2}F_2 + \cdots + a_{mp}F_p + e_m \end{aligned} \tag{31}$$

kde $F_1, F_2, \ldots, F_p$ je p vybraných společných faktorů, které vyvolávají korelace mezi m původními znaky. Faktory mají nulovou střední hodnotu a jednotkový rozptyl, $e_1, e_2, \ldots, e_m$ *jsou specifické (chybové) faktory*, které přispívají pouze k rozptylu jednotlivých znaků. Koeficienty aij nazýváme *faktorové zátěže* i-tého znaku na j-tém společném faktoru F_j, tyto koeficienty představují prvky matice faktorových zátěží.

V maticovém tvaru pro n-tici objektů lze regresní model FA vyjádřit ve tvaru

$$\boldsymbol{X} = \boldsymbol{F}\boldsymbol{A}^T + \boldsymbol{E}^* \tag{32}$$

v němž matice $\boldsymbol{X}$ je rozměru $n \times m$, $\boldsymbol{F}$ má rozměr $n \times p$, $\boldsymbol{A}$ má rozměr $m \times p$ a $\boldsymbol{E}^*$ je rozměru $n \times m$. Zde matice $\boldsymbol{F}$ má řádky, které obsahují p-tici faktorů nahrazujících m-tici znaků, a $\boldsymbol{E}^*$ je matice chybových členů, specifik. Faktorová analýza je tedy podobná PCA.
U hlavních komponent je každá hlavní komponenta vyjádřena jako lineární kombinace původních znaků. Počet hlavních komponent je roven počtu původních znaků, i když se všechny hlavní komponenty využívají jen zřídka. U faktorové analýzy je však již předem počet faktorů menší, než je počet původních znaků. V ideálním případě by měl být počet faktorů předem znám, v praxi to nemusí být ale splněno. Data sama umožňují tento počet určit.

Faktorový model rozdělí každý původní znak x_i na dvě části, rozdělí však i rozptyl každého znaku na dvě části. Pokud je x_i standardizovaná, je její rozptyl $s_i^2 = 1$. Obecně je rozptyl s_i^2 pro i-tý znak $x_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m$ ve tvaru

$$s_i^2 = \sum_{j=1}^{p} a_{ij}^2 + \sum_{j\neq k}^{p}\sum_{j\neq k}^{p} a_{ij}a_{ik}M_{jk} + N_i, \quad i = 1, \dots, m. \tag{33}$$

Pokud jsou faktory ortogonální, je $\boldsymbol{M} = \boldsymbol{E}$ a všechny mimodiagonální prvky jsou rovny 0. Pak je rozptyl i-tého znaku ve tvaru

$$s_i^2 = \sum_{j=1}^{p} a_{ij}^2 + N_i = h_i^2 + u_i^2. \tag{34}$$

Ve specializované literatuře se provádí ještě další rozklad s_i^2, [9] a podrobný popis metody je detailně popsán v monografii [21].

Komunalita h_i^2 je část rozptylu příslušející použitým faktorům (objasněná část rozptylu).

Jedinečnost u_i^2 je část rozptylu příslušející chybovému členu (zbytková variabilita). Často se dělí na specifitu u_{Si}^2, kterou nelze nijak vysvětlit, a *nespolehlivost* u_{Ni}^2, která představuje experimentální chybu měření.

Základní úlohou FA je pak určit matici faktorových zátěží $\boldsymbol{A}$ a hodnoty komunalit h_i^2, $i = 1, \dots, m$. Existuje mnoho metod, jak je vyčíslit. Nalezenému řešení se říká počáteční faktorová extrakce. Po vyčíslení prvních odhadů faktorů je

faktorová matice pootočena a odpovídající faktory jsou vyčísleny za účelem zlepšení vysvětlení původních znaků. Řada metod FA vyžaduje, aby počet společných faktorů p byl předem znám. Není-li tomu tak, je třeba číslo p určit některou z metod používaných v PCA, například pomocí grafu úpatí vlastních čísel. Protože numerické výsledky jsou silně závislé na zvoleném počtu p, mnoho uživatelů provádí analýzu s několika rozličnými hodnotami p.

● **Faktorová rotace**. Jejím hlavním cílem je odvodit z dat snadno vysvětlitelné a pojmenovatelné společné faktory. Počáteční odhady faktorů však bývají často obtížně vysvětlitelné. Je to tím, že většina faktorů je korelována s více znaky. Protože jedním z cílů faktorové analýzy je právě identifikace smysluplných faktorů, které zestručňují popis původních dat, ukazuje se, že *rotace (otočení) faktorů* je důležitou transformací původních faktorů, které už budou snadněji vysvětlitelné. Otočené faktory jsou v ideálním případě zvoleny tak, že některé zátěže dosahují vysokých hodnot blízko ±1 a zbývající pak velmi nízkých, téměř nulových. Je vhodné, aby každý znak dosahoval vysoké zátěže pouze u jediného faktoru, čili byl *faktorově čistý*.

Cílem je transformovat původní obtížně vysvětlitelnou matici s nepojmenovanými faktory do nové, otočené podoby, ve které lze faktory už snadněji pojmenovat a ve které jsou původní znaky převážně faktorově čisté. Matici lze zobrazit v *grafu faktorových zátěží*. Otočí-li se souřadné osy faktorů F_1 a F_2 o 90°, nazýváme toto otočení *ortogonální rotace*. Otočí-li se souřadné osy faktorů o jiný úhel než 90°, nazýváme toto otočení *neortogonální rotace*. Účelem otočení faktorů je získání ***jednodušší struktury***, která se snadněji vysvětlí a znaky budou spíše faktorově čisté. To znamená, že se požaduje, aby každý faktor měl nenulové zátěže pouze pro několik původních znaků. To pomůže při vysvětlování a pojmenování faktorů. Také by bylo vhodné, aby měl každý původní znak nenulové zátěže pouze pro několik málo faktorů, nejlépe pro jediný. To pak dovolí, aby se faktory odlišovaly jeden od druhého. Má-li několik faktorů vysoké zátěže na témže znaku, je obtížné rozhodnout, v čem se faktory vlastně liší.

Existují jednoduchá strukturní pravidla týkající se matice faktorových zátěží $\boldsymbol{A}$. Tato pravidla umožňují zjednodušení modelu FA výběrem vhodných nulových prvků v matici $\boldsymbol{A}$:

- Každý řádek matice $\boldsymbol{A}$ by měl obsahovat alespoň jeden nulový prvek.
- Pokud se uvažuje p faktorů, měl by každý sloupec matice $\boldsymbol{A}$ obsahovat nejméně p nulových prvků.
- Pro každou dvojici sloupců matice $\boldsymbol{A}$ mají být některé znaky nevýznamné v jednom sloupci $a_{ij} \approx 0$ a významné v druhém.
- Pro každou dvojici sloupců má být pouze malý podíl znaků významný v obou sloupcích.

Tato pravidla umožňují zjednodušení modelu FA pomocí výběru vhodných nulových prvků v matici $\boldsymbol{A}$. Otočením faktorů se neovlivní těsnost proložení faktorového řešení. Tedy ačkoliv se faktorová matice změní, komunality a procento vysvětleného rozptylu z celkového se nezmění. Otočení předistribuuje vysvětlený rozptyl pro jednotlivé faktory. Různé metody otočení faktorů mohou vést k identifikaci poněkud rozdílných faktorů, neboť řada algoritmů užívá pro ortogonální rotaci *metodu varimax*, která minimalizuje počet znaků, jež vykazují vysokou zátěž faktoru. Nové osy faktorů jsou vybrány tak, že procházejí shluky znaků. Když jsou faktorové zátěže vyčísleny podělením každé zátěže komunalitou dotyčného znaku, nazývá se tento výpočet *Kaiserova normalizace*. Výpočet vede k rovnosti působení znaků s proměnlivými komunalitami. Když neužijeme tuto metodu, znaky s vyššími komunalitami silně ovlivní konečné řešení. Další je *metoda quartimax*, která minimalizuje počet faktorů potřebných k popisu původních znaků, a konečně *metoda equimax*, která je kombinací obou předchozích. Přehled základních typů rotací je uveden v Dartonově práci [10].

Ortogonální otočení vede k faktorům, které jsou nekorelované. Ačkoliv je toto výhodná vlastnost, jsou případy, kdy korelace mezi faktory dovolují zjednodušit faktorovou matici. Neortogonální rotace vede k podstaně smysluplnějším faktorům. Konečně vlivy v přírodě jsou také vesměs korelované. Neortogonální rotace zachovává komunality znaků tak, jak to činí ortogonální rotace.

6.3. Grafické pomůcky faktorové analýzy FA

Grafické pomůcky jsou daleko názornější než numerické hodnoty faktorových zátěží zapsané v tabulce. Uvedeme proto dvě užitečné grafické diagnostiky, které usnadní odhalení vnitřní struktury v datech.

● **Graf faktorových zátěží**. Vyšetření úspěšnosti ortogonálního otočení faktorů nabízí 2D nebo 3D graf původních znaků, když na souřadných osách jsou jednotlivé faktory. Jestliže bylo otočením dosaženo jednoduché struktury, objeví se v grafu shluky znaků. Znaky při konci souřadnicové osy mají vysokou faktorovou zátěž pouze na přiřazené ose. Tyto faktory se označují jako faktorově *čisté původní znaky*. Původní znaky blízko počátku (0, 0) mají malé zátěže obou faktorů. Původní znaky, které nejsou blízké některé souřadnici, jsou vysvětleny oběma faktory a označují se jako *znaky faktorově nečisté*. Když bylo dosaženo jednodušší struktury, mělo by být málo znaků faktorově nečistých. K identifikování faktorů je také třeba vytvořit skupiny znaků, které mají vysokou faktorovou zátěž pro stejné faktory. Pak lze v tomto grafu snadno určit i shluky znaků.

● **Graf faktorového skóre**. Jelikož je cílem faktorové analýzy redukovat veliký počet původních znaků na menší počet faktorů, je rovněž užitečné vyčíslit faktorové skóre pro každý objekt. Graf faktorového skóre pro vybraný pár faktorů je užitečný k odhalení výjimečných hodnot objektů, především odlehlých pozorování. Pro k-tý objekt je skóre pro j-tý faktor F_{jk} vyčísleno vztahem $\hat{F}_{ik} = \sum_{i=1}^{m} w_{ji} x_{ik}$, kde x_{ik} je standardizovaná hodnota i-tého znaku pro k-tý objekt a w_{ji} je koeficient faktorového skóre pro j-tý faktor a i-tý znak, m je celkový počet všech původních znaků. Existuje řada sofistikovaných metod k vyčíslení koeficientů faktorového skóre. Každá metoda má své vlastnosti a výsledky metod jsou pak i rozličné. Především užívané *regresní postupy* kombinují vnitřní korelace mezi znaky x_i a faktorovými zátěžemi, čímž vzniknou veličiny zvané *koeficienty faktorového skóre*. Tyto koeficienty se užijí lineárním způsobem.

7. KANONICKÁ KORELAČNÍ ANALÝZA CCA

7.1. Zaměření metody CCA

Kanonická korelační analýza byla navržena v roce 1935 Hotellingem v souvislosti s hledá-ním lineární kombinace jedné skupiny znaků $\boldsymbol{x} = (x_1, \ldots, x_q)$, která nejlépe koreluje s lineární kombinací druhé skupiny znaků $\boldsymbol{y} = (y_1, \ldots, y_p)$. Vychází z předpokladu společného rozdělení obou skupin znaků. Podobně jako u PCA a FA se hledají lineární kombinace znaků obou skupin, tj. hypotetických kanonických proměnných, které vedou k maximálním vzájemným korelacím. Tyto kanonické proměnné tvoří nový souřadnicový systém vzájemně ortogonálních složek. Přesněji řečeno, jde o krokový proces, podobně jako v PCA, kdy se v prvním kroku hledá lineární kombinace $\boldsymbol{x}$ a lineární kombinace $\boldsymbol{y}$, jejichž korelace je maximální. Tyto lineární kombinace tvoří první složky souřadnicových systémů kanonických proměnných pro $\boldsymbol{x}$ a $\boldsymbol{y}$. V dalších krocích se hledají další lineární kombinace $\boldsymbol{x}$ a $\boldsymbol{y}$, tj. kanonické proměnné takové, které mají maximální vzájemnou korelaci a přitom jsou nekorelované s kanonickými proměnnými (složkami obou nových souřadnicových systémů) nalezených v předchozích krocích. Přímé využití této metody je při snižování dimenze, kdy jsou skupiny původních znaků veliké a účelem je nalézt malý počet kanonických proměnných (lineární kombinace původních znaků), které postihují v maximální míře korelace mezi původními skupinami znaků.

Kanonická korelační analýza úzce souvisí s chováním vícenásobného korelačního koeficientu R mezi jednou náhodnou veličinou a lineární kombinací jiných proměnných. Tento koeficient nabývá maxima pro případ, kdy jsou koeficienty lineární kombinace přímo koeficienty regresními. Interpretace kanonických proměnných je ještě problematičtější než u faktorové analýzy. Slouží proto v řadě případů jako předzpracování dat pro další analýzu (například diskriminační).

Kanonická korelační analýza se často využije v situacích, ve kterých se tvoří regresní modely a kde existuje více než jedna závisle proměnná. Zvláště je

užitečná v situacích, kdy závisle proměnné jsou vnitřně korelovány, takže nemá cenu je vyhodnocovat odděleně, protože by se zanedbala jejich vzájemná vnitřní korelace. Užitečnou vlastností kanonické korelace je možnost ověření nezávislosti mezi skupinami znaků $\boldsymbol{x}$ **a** $\boldsymbol{y}$.

7.2. Podstata metody CCA

Předpokládejme, že chceme studovat vztah mezi skupinou znaků $x = (x_1, x_2, \ldots, x_p)^T$ a skupinou znaků $y = (y_1, y_2, \ldots, y_q)^T$. Bez újmy na obecnosti lze předpokládat, že všechny znaky jsou centrované. V kanonické korelační analýze se tvoří kanonické proměnné

$$\boldsymbol{U}_1 = a_1\boldsymbol{y}_1 + a_2\boldsymbol{y}_2 + \cdots + a_q\boldsymbol{y}_q \tag{35}$$

ze znaků $\boldsymbol{y}$ a kanonické proměnné

$$\boldsymbol{V}_1 = b_1\boldsymbol{x}_1 + b_2\boldsymbol{x}_2 + \cdots + b_q\boldsymbol{x}_q \tag{36}$$

ze znaků $\boldsymbol{x}$. V každé skupině znaků se vyhledávají koeficienty a_i a b_i tak, aby pro kanonické proměnné $\boldsymbol{U}_1$ a $\boldsymbol{V}_1$ vykazovaly maximální *párový korelační koeficient* R. V druhém kroku se pak hledají další lineární kombinace čili kanonické proměnné $\boldsymbol{U}_2$ a $\boldsymbol{V}_2$, které mají druhý největší korelační koeficient ρ_2 za podmínky, že $\boldsymbol{U}_2$ a $\boldsymbol{V}_2$ jsou nekorelované s prvními kanonickými proměnnými $\boldsymbol{U}_1$ a $\boldsymbol{V}_1$.

Korelace R mezi $\boldsymbol{U}_1$ a $\boldsymbol{V}_1$ se nazývá *první kanonická korelace*. První kanonická korelace je tudíž největší možná korelace mezi lineárními kombinacemi znaků $\boldsymbol{y}$ a lineárními kombinacemi znaků $\boldsymbol{x}$. První kanonická korelace představuje jistou analogii k vícenásobnému korelačnímu koeficientu ve vícenásobné lineární regresi. Rozdíl je v tom, že u kanonické korelace existuje několik znaků y a je nutné navíc hledat optimální lineární kombinaci mezi nimi. *Druhá kanonická proměnná* $\boldsymbol{V}_2$ je lineární kombinací $\boldsymbol{x}$ a jí odpovídající druhá kanonická proměnná $\boldsymbol{U}_2$ je lineární kombinací $\boldsymbol{y}$. Koeficienty těchto lineárních kombinací jsou vybrány tak, že současně platí:

1. $\boldsymbol{V}_2$ je nekorelované s $\boldsymbol{V}_1$ a $\boldsymbol{U}_1$.

2. $\boldsymbol{U}_2$ je nekorelované s $\boldsymbol{V}_1$ a $\boldsymbol{U}_1$.

Kanonické proměnné $\boldsymbol{V}_2$ a $\boldsymbol{U}_2$ mají maximální možnou korelaci R (podmíněnou splněním podmínek 1 a 2), která se nazývá *druhá kanonická korelace* a jež bude nutně menší nebo rovna první kanonické korelaci.

V dalších krocích se určují kanonické proměnné $\boldsymbol{U}_3$ a $\boldsymbol{V}_3$, $\boldsymbol{U}_4$ a $\boldsymbol{V}_4$ atd. Maximální počet kanonických korelací a jim odpovídajících dvojic kanonických proměnných je roven menšímu ze dvou čísel p a q, když p je počet znaků $\boldsymbol{x}$ a q je počet znaků $\boldsymbol{y}$.

Kanonické proměnné je nutno také interpretovat, protože stejně jako ve faktorové analýze pracujeme i zde s hypotetickými proměnnými, které je často obtížné fyzikálně vysvětlit. Důležitost každé kanonické proměnné se vyhodnocuje ze dvou hledisek:

a) Určujeme intenzitu vztahu mezi kanonickou proměnnou $\boldsymbol{U}$ a původním znakem y nebo kanonickou proměnnou $\boldsymbol{V}$ a znaky $\boldsymbol{x}$.

b) Vyjadřujeme také intenzitu vztahu mezi oběma kanonickými proměnnými $\boldsymbol{V}$ a $\boldsymbol{U}$.

Jelikož kanonická korelační analýza předpokládá pouze *lineární závislost* mezi proměnnými, třeba vyšetřit grafy každého páru proměnných a prověřit linearitu a odlehlé body [21].

Výběrové kanonické proměnné je možné získat z datových matic $\boldsymbol{X}$ rozměru $(n \times p)$ a $\boldsymbol{Y}$ rozměru $(n \times q)$, uvažujme $p \leq q$. Společná výběrová kovarianční matice je pak pro případ, že jsou všechny znaky centrované (v odchylkách od průměrů jednotlivých znaků), rovna

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{xx} & \boldsymbol{S}_{xy} \\ \boldsymbol{S}_{yx} & \boldsymbol{S}_{yy} \end{pmatrix}. \tag{37}$$

Kovarianční matice znaků $\boldsymbol{x}$ je rovna

$$\boldsymbol{S}_{xx} = \frac{1}{n-1} \boldsymbol{X}^T \boldsymbol{X} \tag{38}$$

a znaků $\boldsymbol{y}$ je rovna

$$\boldsymbol{S}_{yy} = \frac{1}{n-1}\boldsymbol{X}^T\boldsymbol{X}. \tag{39}$$

Výběrová matice kovariancí znaků $\boldsymbol{x}$ a znaků $\boldsymbol{y}$ je rovna

$$\boldsymbol{S}_{xy} = \frac{1}{n-1}\boldsymbol{X}^T\boldsymbol{Y}. \tag{40}$$

Matice $S_{yx} = S_{xy}^T$. Rozkladem matice

$$(\boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{xy}\boldsymbol{S}_{yy}^{-1}\boldsymbol{S}_{yx}) = \boldsymbol{V}\boldsymbol{\lambda}\boldsymbol{V}^T \tag{41}$$

na vlastní čísla $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ jako diagonální prvky diagonální matice $\boldsymbol{\lambda}$ a vlastní vektory $\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_p$ jako sloupce ortogonální matice $\boldsymbol{V}$ jsou určeny kanonické korelační koeficienty a koeficienty $\boldsymbol{b}_i = \boldsymbol{V}_i$. Koeficienty $\boldsymbol{a}_i$ se pak určí ze vztahu

$$\boldsymbol{a}_i = \frac{1}{\sqrt{\lambda_i}}\boldsymbol{S}_{yy}^{-1}\boldsymbol{S}_{xy}\boldsymbol{b}_i. \tag{42}$$

Kanonické proměnné pro výběrová data jsou reprezentovány *kanonickými skóre*

$$\boldsymbol{V} = \boldsymbol{X}\boldsymbol{B} \quad a \quad \boldsymbol{U} = \boldsymbol{Y}\boldsymbol{A}, \tag{43}$$

kde $\boldsymbol{B}$ obsahuje jako sloupce vlastní vektory $\boldsymbol{V}_i = \boldsymbol{b}_i$ a $\boldsymbol{A}$ obsahuje jako sloupce koeficienty $\boldsymbol{a}_i$. Pro interpretaci kanonických proměnných se počítají také matice zátěží podle vztahů

$$\boldsymbol{L}_x = \boldsymbol{S}_{xx}\boldsymbol{B} \quad a \quad \boldsymbol{L}_y = \boldsymbol{S}_{yy}\boldsymbol{A}, \tag{44}$$

Bývá zvykem pracovat s normovanými znaky (centrovanými a dělenými směrodatnou odchylkou), pak se v uvedených vztazích použijí místo kovariančních matic $\boldsymbol{S}$ matice korelační $\boldsymbol{R}$. Obyčejně se také používá pouze malý počet kanonických proměnných, a to menší než $\min(p, q)$.

Pro normované znaky jsou $\boldsymbol{L}_x$ a $\boldsymbol{L}_y$ přímo korelačními koeficienty mezi znaky $\boldsymbol{x}$ a kanonickými proměnnými $\boldsymbol{V}$, respektive mezi $\boldsymbol{y}$ a $\boldsymbol{U}$. Pro centrované proměnné se získávají korelační matice vynásobením, respektive, kde $\boldsymbol{D}$ jsou

diagonální matice obsahující na diagonále směrodatné odchylky jednotlivých znaků.

7.2.1. *Test významnosti kanonických korelací*

Většinou se určují koeficienty pro všechny kanonické proměnné, dále hodnoty kanonických korelací a hodnoty kanonických proměnných pro všechny prvky výběru, nazýváme je *kanonická skóre*. Standardní je test nulové hypotézy H_0: „k nejmenších kanonických korelací souboru je rovno nule“. Dva testy jsou vhodné: Bartlettův χ^2-test a aproximativní F-test. Oba testy byly odvozeny za předpokladu, že $\boldsymbol{x}$ a $\boldsymbol{y}$ pocházejí z vícerozměrného normálního rozdělení. Velká hodnota χ^2 a velká hodnota F jsou indikátorem, že ne všechny korelace mezi proměnými v souboru jsou nulové. V datech s více znaky mohou být tyto testy užitečným průvodcem k výběru počtu statisticky významných kanonických korelací. Výsledky testu vyšetřujeme proto, abychom určili, ve kterém kroku lze zbývající kanonické korelace považovat za nulové.

7.2.2. *Vysvětlení kanonických proměnných*

Důležitým výstupem jsou korelace mezi kanonickými proměnnými a původními znaky, ze kterých byly kanonické proměnné odvozeny. Tyto korelace vysvětlují kanonické proměnné, když některé ze znaků ať už ze skupiny x, nebo ze skupiny y jsou navzájem silně vnitřně korelované. Tyto korelace se nazývají *zátěže kanonických proměnných* nebo *koeficienty kanonické struktury*. Protože kanonické zátěže mohou být vysvětlovány jako korelace mezi každým původním znakem a kanonickou proměnnou, jsou užitečné v pochopení vztahu mezi znaky a kanonickými proměnnými. Je-li skupina znaků v jedné kanonické proměnné nekorelovaná, kanonické zátěže jsou rovny standardizovaným koeficientům kanonické proměnné. Jsou-li některé ze znaků silně vnitřně korelovány, pak zátěže a koeficienty jsou zcela rozličné. To je právě případ vysvětlení kanonických zátěží, který je poněkud snazší než koeficientů kanonické proměnné. Předpokládejme například, že existují dva znaky x, které jsou vysoce kladně korelovány a že každý je pozitivně korelován s kanonickou proměnnou. Pak je možné, že jeden koeficient kanonické proměnné bude

kladný a jeden záporný, zatímco kanonické zátěže jsou obě kladné, jak by se dalo očekávat.

7.2.3. *Analýza redundance*

Průměrná hodnota čtverců kanonických zátěží pro první kanonickou proměnnou $\boldsymbol{V_1}$ poskytuje podíl rozptylu znaků, jež je vysvětlen první kanonickou proměnnou. Obdobně to platí i pro $\boldsymbol{U_1}$ a znaky $\boldsymbol{y}$. Někdy je podíl rozptylu takto objasněného malý, i když jde o vysokou kanonickou korelaci. To se může stát v důsledku jednoho či dvou znaků, které mají hlavní vliv na kanonickou proměnnou. Těmto výpočtům se říká *analýza redundance* (nadbytečnosti). Redundanční index je mírou průměrného podílu rozptylu ve skupině znaků $\boldsymbol{y}$, který nebyl zahrnut do souboru $\boldsymbol{V}$. Je analogický čtverci vícenásobné korelace ve vícenásobné lineární regresi. Je také možné získat podíl rozptylu ve skupině znaků $\boldsymbol{x}$, jež je objasněn všemi proměnnými $\boldsymbol{U}$.

7.2.4. *Grafické pomůcky*

Mezi užitečné grafické diagnostiky patří *graf závislosti skóre kanonických proměnných* $\boldsymbol{U_i}$ na $\boldsymbol{V}_i$. Stupeň rozptýlení bodů po ploše je důležitým ukazatelem. Graf je užitečný v odkrývání neobvyklých případů ve výběrech, jako jsou odlehlé body. Graf závislosti $\boldsymbol{U}_1$ na $\boldsymbol{V}_1$ může vést ve zdánlivě nelineární roztylový diagram nebo eliptický tvar typický pro dvojrozměrné normální rozdělení.

8. DISKRIMINAČNÍ ANALÝZA (DA)

Hledáním struktury a vzájemných vazeb v objektech se zabývají klasifikační metody vícerozměrné statistické analýzy. *Klasifikační metody* jsou postupy, pomocí kterých se jeden objekt zařadí do existující třídy (*diskriminační analýza DA*), nebo pomocí nichž lze neuspořádanou skupinu objektů uspořádat do několika vnitřně sourodých tříd či shluků (*analýza shluků CLU*). Postup klasifikace je založen na určitých předpokladech o vlastnostech klasifikovaných objektů, například, když rozdělení náhodného vektoru charakterizujícího objekty je normální, pak hovoříme o *parametrických klasifikačních metodách*. Není-li klasifikace založena na znalostech rozdělení náhodného vektoru, mluvíme o *neparametrických klasifikačních metodách*. Významnou roli při hledání struktury a vazeb mezi objekty na základě jejich podobnosti tvoří také *vícerozměrné škálování MDS*.

Obecně patří tyto úlohy do kategorie statistického učení (statistical learning), kdy se na základě výstupů buď metrických (kardinální), nebo nemetrických (ordinální, nomimální) konstruuje predikce, která je založena na skupině znaků (vstupní data). Pro vybrané výstupy a znaky jsou k dispozici tzv. *trénovací data*, kde je pro každý objekt určen jak výstup (y), tak i hodnoty všech znaků ($\boldsymbol{x}$). Na základě těchto dat se sestavuje predikční model $y = f(\boldsymbol{x})$, který umožňuje předpovědět výstup pro nový objekt charakterizovaný znaky $\boldsymbol{x}_0$. Je zřejmé, že takto definovaná úloha statistického učení souvisí velmi úzce s regresní analýzou. Protože se na základě *trénovacích dat* „učí“ predikční model předvídat y, označuje se tento postup jako učení s učitelem (supervised learning). Při učení bez učitele (unsupervised learning) jsou k dispozici určité znaky pro objekty a nikoliv výstupy. Úlohou je pak pouze stanovit organizaci dat (shluky). Základy statistického učení jsou popsány například v knize [13].

8.1. Zaměření metody DA

Klasická klasifikační diskriminační analýza, zavedená Ronaldem Fisherem v roce 1936, patří mezi metody zkoumání vztahu mezi skupinou p nezávislých znaků, zvaných *diskriminátory* (sloupců zdrojové matice), a jednou kvalitativní závisle

proměnnou − výstupem. Výstupem je v nejjednodušším případě binární proměnná y, nabývající hodnotu 0 pro případ, že objekt je v první třídě, respektive hodnotu 1 pro případ, že objekt je ve druhé třídě. O třídách je známé, že jsou zřetelně odlišené a každý objekt patří do jedné z nich. Účelem může být také identifikace, které znaky přispívají do procesu klasifikace. Ve vstupních datech trénovací skupiny jsou svými hodnotami diskriminátorů a výstupů všechny objekty zařazené do tříd. Účelem je nalézt predikční model umožňující zařadit nové objekty do tříd.

Řešení této úlohy využitím logistické regrese je jednoduché. U konstrukce modelů v logistické regresi se omezíme na standardní pravděpodobnostní přístup vycházející z předpokladu, že ve všech skupinách mají znaky stejné rozdělení (normální), lišící se pouze některými parametry. Symbol π_1 označuje *apriorní pravděpodobnost* příslušnosti objektu do první skupiny a $\pi_2 = 1 - \pi_1$ je pak pravděpodobnost příslušnosti objektu do druhé skupiny. Pomocí Bayesovy věty pak můžeme obecně určit *aposteriorní pravděpodobnost* příslušnosti k j-té skupině ($j = 1, 2$)

$$P(G = j|x) = \frac{f_j(x)\pi_j}{\sum_{i=1}^{2} f_i(x)\pi_i} \quad , \quad j = 1,2. \tag{45}$$

Tento zápis se dá snadno rozšířit i pro případ více kategorií. Pokud je $P(G = j|x)$ známé, je k dispozici rozhodovací pravidlo, na jehož základě můžeme klasifikovat nový objekt do té skupiny, kde je tato aposteriorní pravděpodobnost vyšší. Je zřejmé, že existuje vždy pravděpodobnost chybné klasifikace, nesprávného zařazení do tříd. Dá se ukázat, že tato pravděpodobnost je minimalizována pro případ, že jako rozhodovací pravidlo byla vybrána aposteriorní pravděpodobnost $P(G = j \mid x)$. Z rovnice (7.6) je zřejmé, že objekt zařadíme do první skupiny, pokud $\pi_1 f_1(x) > \pi_2 f_2(x)$, protože součet ve jmenovateli je pouze normalizační konstantou. Pro zařazení objektu do první skupiny musí proto platit, že

$$\frac{f_1(x)}{f_2(x)} > \pi_2/\pi_1. \tag{46}$$

Je zřejmé, že konkrétní diskriminační pravidla budou záviset na tom, které parametry obou rozdělení se liší. Jak bude ukázáno dále, vede nejjednodušší případ, kdy se $f_1(x)$ a $f_2(x)$ liší pouze středními hodnotami, na tzv. *lineární diskriminační analýzu* (LDA) a případ, kdy se rozdělení liší i kovariančními maticemi, na *kvadratickou diskriminační analýzu* (QDA).

8.2. Zařazovací pravidla DA

Před formálním odvozením diskriminačních pravidel na základě rovnice (7.6) si ukažme intuitivní postup. Pro zjednodušení uvažujme, že účelem je klasifikace do jedné ze dvou tříd (A, B) a že klasifikace se provádí na základě jednoho znaku x s normálním rozdělením. Ve třídě A jde o rozdělení $N(\mu_A, \sigma_A^2)$ a ve třídě B jde o rozdělení $N(\mu_B, \sigma_B^2)$. Nový objekt nechť má hodnotu x. Je logické vybrat tu třídu, pro kterou je x blíže ke střední hodnotě dané třídy. Můžeme tedy určit prahový bod $C = (\mu_A + \mu_B)/2$. Pro případy, kdy $x < C$ se pak objekt zařadí do třídy (A) a pro $x \geq C$ do druhé třídy (B). Tato pravděpodobnost je pro obě kategorie stejná. Pokud by se toto pravidlo použilo i pro případ nestejných rozptylů obou rozdělení, kdy například $\sigma_A^2 < \sigma_B^2$, došlo by k situaci, že pravděpodobnost nesprávné klasifikace pro třídu A by vyšla větší, než pro třídu B. Bude tedy třeba penalizovat C s ohledem na nestejný rozptyl. Je zřejmá analogie s t-testem porovnání dvou středních hodnot pro nestejné rozptyly. Pro případ dvou diskriminátorů $(x_1\ x_2)$ je výhodné zobrazovat v prostoru $x_1\ x_2$ elipsy konstantní hustoty (například pro pravděpodobnost 0.95). Záleží na tom, zda jsou kovarianční matice $\boldsymbol{C_A}$ a $\boldsymbol{C}_B$ shodné či nikoliv.

Pro shodné kovarianční matice a stejné apriorní pravděpodobnosti zařazení do kategorií ($\pi_1 = \pi_2$) bude dělicí čarou pro obě třídy přímka, určená rovnicí $a_1\ x_1 + a_2\ x_2 = C$. Tato situace se označuje jako *lineární diskriminace* (LDA). Pro různé kovarianční matice $\boldsymbol{C}_A \approx \boldsymbol{C}_B$ bude dělicí čára definovaná polynomem druhého stupně, což je *kvadratická diskriminace* (QDA). Pomocí diskriminační funkce se převede dvourozměrný problém zařazování do tříd na problém jednorozměrný, tj. výpočet $Z_i = a_1 x_{1i} + a_2 x_{2i}$ a porovnání s C.

8.3. Lineární (LDA) a kvadratická (QDA) diskriminační funkce

Při formálním odvození tvaru diskriminačních funkcí je možné vyjít z rovnice (45) pro aposteriorní pravděpodobnost příslušnosti k j-té skupině ($j = 1, 2$) a hledat její maximum. Podle typu hustot pravděpodobnosti pro znaky $f_1(x)$ a $f_2(x)$ se tak liší jednotlivé diskrimi-načních metody v tom, jak jsou specifikovány dělicí oblasti a jaký mají tvar:

- pro normální rozdělení, lišící se jen středními hodnotami tříd, dostáváme lineární diskriminační analýzu (LDA),
- pro normální rozdělení, lišící se jak středními hodnotami, tak i kovariančními maticemi tříd dostáváme kvadratickou diskriminační analýzu (QDA),
- pro případ směsí normálních rozdělení vycházejí nelineární diskriminační funkce,
- pro případ neparametrických hustot rozdělení znaků ve třídách obdržíme flexibilní diskriminační funkce,
- naivní Bayesův přístup spočívá v představě, že každá hustota pravděpodobnosti ve třídě je získána jako součin marginálních hustot znaků (znaky jsou zde považovány za podmíněně nezávislé).

Pro případ vícerozměrného Gaussova rozdělení v i-té třídě má odpovídající hustota pravděpodobnosti tvar

$$f_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{m/2}\det(\boldsymbol{C}_i)^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T C_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right), \tag{47}$$

kde $m = 2$ je počet znaků [21].

1. Lineární diskriminační funkce LDA

LDA vychází z předpokladu, že $\boldsymbol{C_1} = \boldsymbol{C}_2 = \boldsymbol{C}$. Pak pro dvě třídy 1, 2 máme logaritmus poměru aposteriorních pravděpodobností

$$\ln\frac{P(G=1|\boldsymbol{x})}{P(G=2|\boldsymbol{x})} = \ln\left(\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} + \ln\frac{\pi_1}{\pi_2}\right) =$$

$$= \ln\frac{\pi_1}{\pi_2} - 0.5(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T V^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \boldsymbol{x}^T\boldsymbol{C}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu}_2). \tag{48}$$

Vzhledem k proměnné $\boldsymbol{x} = (x_1, x_2)^T$ jde o lineární funkci, takže dělicí funkcí bude přímka. Tato dělicí funkce je množinou bodů x, pro které platí $P(G = 1 \mid \boldsymbol{x}) = P(G = 0 \mid \boldsymbol{x})$. Pro případ více znaků jsou dělicí čáry lineárními úseky procházejícími průsečíky jednotlivých elips konstantní hustoty pro jednotlivé třídy.

Pro případ, že $\boldsymbol{C} = \sigma^2 \boldsymbol{E}$, tj. jednotlivé znaky jsou ortogonální, přecházejí elipsy v kružnice. Dělicí čáry procházejí středem spojnice mezi body středních hodnot a jsou na tyto spojnice kolmé.

Z uvedené rovnice pro logaritmus podílu aposteriorních hustot můžeme snadno určit pravidlo pro zařazení do první skupiny ve tvaru

$$\boldsymbol{a}^T \boldsymbol{x} + \boldsymbol{b} > 0, \tag{49}$$

kde

$$\boldsymbol{a}^T = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{C}^{-1}, \tag{50}$$

je vektor koeficientů u lineárního členu a absolutní člen je ve tvaru

$$\boldsymbol{b} = -0.5 \boldsymbol{a}^T (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \ln \frac{\pi_1}{\pi_2}, \tag{51}$$

Lineární diskriminační funkce je pak $\boldsymbol{L}(\boldsymbol{x}) = \boldsymbol{a}^T \boldsymbol{x}$ Rozdělení této funkce je normální a závisí na Mahalanobisově vzdálenosti mezi středními hodnotami v obou skupinách

$$D_M^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{C}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{52}$$

Pro obě skupiny pak platí, že $E(L(\boldsymbol{x})) = D_M^2/2$ *a* $D(L(x)) = D_M^2$. Dělicí rovina (přímka pro $m = 2$) je pak určena rovnicí

$$\boldsymbol{L}(\boldsymbol{x}) + \boldsymbol{b} = \boldsymbol{a}^T \boldsymbol{x} + \boldsymbol{b} = 0. \tag{53}$$

Znalost rozdělení statistiky $\boldsymbol{L}(\boldsymbol{x})$ umožňuje stanovit chybu nesprávné klasifikace ω ze vztahu

$$\omega = \pi_1 \varphi\left(\frac{\ln\frac{\pi_1}{\pi_2} - \frac{D_M^2}{2}}{D_M}\right) + \pi_2 \varphi\left(\frac{-\ln\frac{\pi_2}{\pi_1} - \frac{D_M^2}{2}}{D_M}\right), \tag{54}$$

kde $\varphi(x)$ je distribuční funkce normovaného normálního rozdělení. Pro případ neinforma-tivní apriorní pravděpodobnosti $\pi_1 = \pi_2 = 0.5$ vyjde jednoduchý vztah $\omega = \varphi(-D_M/2)$. Chyba nesprávné klasifikace do první třídy je stejná jako chyba klasifikace do druhé třídy. Při použití Mahalanobisovy vzdálenosti u reálných dat se doporučuje použít korekci

$$\frac{n-m-3}{n-2} D^2 - \frac{nm}{n_1 n_2}. \tag{55}$$

Místo lineární diskriminační funkce je možné použít lineární diskriminační kritérium, které nevyžaduje stanovení koeficientů $\boldsymbol{a}$, $\boldsymbol{b}$. Toto kritérium $Lk_j(\boldsymbol{x})$ se počítá pro každou j-tou třídu zvlášť. Při klasifikaci objektů se pak objekt zařazuje do l-té třídy, pro kterou vyjde $Lk_l(\boldsymbol{x}) = \max\{Lk_j(\boldsymbol{x})\}$. Pro LDA má lineární diskriminační kritérium tvar

$$Lk_j(\boldsymbol{x}) = \boldsymbol{x^T C^{-1} \mu_j} - 0.5 \boldsymbol{\mu}_j^T \boldsymbol{C}^{-1} \boldsymbol{\mu}_j + \ln \pi_j. \tag{56}$$

Ukažme si praktické použití LDA na případu dvou tříd, tedy $m = 2$. Vychází se ze známých vstupních matic $\boldsymbol{X}_1$ rozměru $n_1 \times m$ pro třídu 1 a $\boldsymbol{X}_2$ rozměru $n_2 \times m$ pro třídu 2. Jednotlivé objekty v matici $\boldsymbol{X}$ všech dat se tedy zařadí do tříd podle výstupu G. Standardním postupem se pro jednotlivé třídy vyčíslí výběrové průměry a a společná kovarianční matice

$$\boldsymbol{S} = \frac{(n_1-1)\boldsymbol{S}_1 + (n_2-1)\boldsymbol{S}_2}{n_1+n_2-2}. \tag{57}$$

Vhodným způsobem se určí apriorní pravděpodobnosti. Nejjednodušší je předpoklad $\pi_1 = \pi_2 = 0.5$. Pokud je výběr informativní, tj. reprezentativní, a byl pořízen jako celek a pak teprve rozdělen do skupin, je možné použít relativních četností

$$\hat{\pi}_1 = \frac{n_1}{n_1+n_2} \quad a \quad \hat{\pi}_2 = \frac{n_2}{n_1+n_2}\ . \tag{58}$$

Pokud lze přijmout předpoklad normality, určí se koeficienty Fisherovy lineární diskriminační funkce z odhadů

$$\widehat{\boldsymbol{a}} = (\bar{x}_1 - \bar{x}_2)\boldsymbol{S} \ . \tag{59}$$

a

$$\widehat{\boldsymbol{b}} = -0.5\boldsymbol{a}^T(\bar{x}_1 - \bar{x}_2) - \ln\frac{\widehat{\pi}_2}{\widehat{\pi}_1} \ . \tag{60}$$

Při zařazování nových objektů s hodnotami znaků $\boldsymbol{x}_0$ se pak uvažuje pravidlo, že do první skupiny je podle použitého pravidla objekt klasifikován, pokud platí $\widehat{\boldsymbol{a}}^T\boldsymbol{x}_0 + \widehat{\boldsymbol{b}} \geq \boldsymbol{0}$. V opačném případě je klasifikován do druhé skupiny. Využitím Mahalanobisovy vzdálenosti lze odhadnout i chybu klasifikace.

Koeficienty diskriminační funkce jsou také nazývány *koeficienty kanonické diskriminační funkce*, protože jsou totožné s těmi, které vyčíslíme kanonickou korelační analýzou, když budeme uvažovat maximálně korelované lineární kombinace závisle proměnné jak tříd, tak i jejich diskriminátorů.

Jiný soubor koeficientů je nazván *koeficienty Fisherovy lineární diskriminační funkce* neboli klasifikační koeficienty, které mohou být užity přímo k zařazování nových dosud nezařazených objektů do tříd čili ke klasifikaci. Soubor takových koeficientů je vyčíslen pro každou třídu a objekt je zařazen do té třídy, pro kterou dosáhne největší hodnoty diskriminačního skóre. Jsou-li použity obě kanonické diskriminační funkce, jsou klasifikační výsledky stejné.

9. LOGISTICKÁ REGRESE (LR)

9.1. Zaměření metody LR

Logistická regrese byla navržena v 60-tých letech [14] jako alternativní postup k metodě nejmenších čtverců pro případ, že vysvětlovaná (závisle) proměnná je binární. V minulosti se týkala většina úloh aplikace logistické regrese oblasti medicíny a epidemiologie. Vysvětlovaná (závisle) proměnná představuje přítomnost nebo nepřítomnost choroby. Příkladem mohou být výpočty rizika vzniku srdeční choroby nebo rakoviny jako funkce osobních a chování se týkajících charakteristik (věk, váha, krevní tlak, hladina cholesterolu a kouření). Logistická regrese může třeba v jedné úloze odhalit, které charakteristiky mohou predikovat těhotenství mladistvých, kdy lze predikci provést na základě rozličných proměnných, jako je bodové hodnocení mladistvého průměrem známek v základní škole, náboženstvím nebo velikostí rodiny. V průmyslu mohou být hodnoceny výrobní jednotky buď jako úspěšné, nebo neúspěšné. K tomu je sledována řada charakteristik a vyšetří se, které charakteristiky budou účinně předvídat komerční úspěch. Logistická regrese úzce souvisí s diskriminační analýzou a analýzou směsi normálních rozdělení [15]. Je alternativní metodou klasifikace, když nejsou splněny předpoklady vícerozměrného normálního modelu. Může se aplikovat na libovolnou kombinaci diskrétních nebo spojitých proměnných. Vyžaduje však znalost obou, jak závisle proměnné, tak i nezávisle proměnných u analyzovaného výběru. Výsledný model pak může být využit k budoucímu klasifikování, když jsou uživateli dostupné pouze vysvětlující, nezávisle proměnné.

Logistická regrese se liší od lineární regrese v tom, že predikuje pravděpodobnost dané události, která se buď stala, nebo nestala. Vypočtená pravděpodobnost je tedy buď rovna 0, nebo 1. Aby se vytvořila tato vazebná podmínka, užívá logistická regrese tzv. *logitovou transformaci*, která vede na sigmoidální vztah mezi závisle proměnnou y a vektorem nezávisle proměnných x. Při velmi nízkých hodnotách nezávisle proměnné se pravdě-podobnost proměnné y blíží k nule, zatímco při vysokých hodnotách nezávisle proměnné se blíží k jedné, obr. 8.1. Rozdíl mezi logistickou a lineární regresí spočívá v tom,

že logistická regrese používá *kategorickou vysvětlovanou*, *závisle proměnnou* zatímco lineární regrese užívá pouze *spojitou vysvětlovanou, závisle proměnnou*. Centrální roli zde hraje logitová transformace, která vychází z tzv. *poměru šancí či naděje*. Dle typu vysvětlující proměnné se rozlišují:

a) *Binární logistická regrese*, která se týká binární závisle proměnné či znaku, nabývající pouze dvou možných hodnot, jako je například přítomnost a nepřítomnost, muž a žena. Vektor vysvětlujících, nezávislých proměnných či znaků může obsahovat jednu či více proměnných či znaků, a to spojitých zvaných *prediktory* anebo kategorických zvaných *faktory*.
b) *Ordinální logistická regrese*, která se týká ordinální závisle proměnné či znaku, nabývající tři a více možných stavů přirozeného charakteru, třeba stoupající síly jako je silný nesouhlas, nesouhlas, souhlas, silný souhlas. Vektor vysvětlujících nezávisle proměnných či znaků může obsahovat jak prediktory, tak i faktory.
c) *Nominální logistická regrese*, která se týká nominální závisle proměnné, znaku o více než třech úrovních různých stavů, mezi kterými je definována pouze odlišnost. Vektor vysvětlujících nezávisle proměnných může obsahovat jak prediktory, tak i faktory.

9.2. Logistický regresní model

V logistické regresi potřebujeme vědět, zda se událost stala nebo nestala. Jde proto o dichotomickou hodnotu závisle proměnné, ze které se odhaduje pravděpodobnost, že se událost stala či nestala. Je-li predikovaná pravděpodobnost větší než 0.50, pak se událost stala, je-li menší než 0.50, pak se nestala. Název logistická regrese je odvozen od logitové transformace závisle proměnné. Je-li tato transformace užita, logistická regrese a její koeficienty nabývají poněkud jiného významu než v regresi metrické proměnné. Postup, vyčíslující logistické koeficienty, porovnává pravděpodobnost události odehrané $L_{(1)}$ vůči pravděpodobnosti události neodehrané $L_{(0)} = 1 - L_{(1)}$ využitím *pravděpodobnostního poměru* $L_{(1)}/L_{(0)}$, ve kterém je pravděpodobnost $L_{(1)}$ vyjádřena logistickou funkcí

$$L_{(1)} = \frac{1}{1+e^{c-z}} \ . \tag{61}$$

Pravděpodobnostní poměr, také nazývaný *„poměr vyhlídek, šancí“*, lze vyjádřit vztahem

$$\frac{L_{(1)}}{L_{(0)}} = \exp a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_p x_p \ . \tag{62}$$

kde odhadované koeficienty $a_0, a_1, a_2, \ldots, a_p$ jsou míry změny poměru pravděpodobností, který je lineární funkcí diskriminační funkce o p původních nezávisle proměnných

$$Z = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_p x_p \ . \tag{63}$$

Po dosazení, zlogaritmování výrazu (61) a úpravě vyjde

$$C - Z = \ln\left(\frac{L_{(0)}}{L_{(1)}}\right). \tag{64}$$

V souladu s klasifikačními postupy (statistické učení s učitelem) je $L_{(0)} = P(G = 1 \mid x)$ a $L_{(1)} = P(G = 1 \mid x) = 1 - P(G = 1 \mid x)$. Po úpravách rovnice (61) vyjde

$$\ln\left(\frac{L_{(1)}}{L_{(0)}}\right) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p. \tag{65}$$

kde $b_0 = -C + a_0$, $b_i = a_i$ pro $i = 1, \ldots, p$. Je patrná podobnost s vícenásobnou lineární regresní funkcí. Vyčíslení poměru pravděpodobností je běžně užívaná technika vysvětlení pravděpodobnosti. Ve sportu například řekneme, že tým má šanci 3:1. Toto tvrzení říká, že favorizovaný tým má pravděpodobnost vítězství $\frac{3}{3+1} = \frac{3}{4} = 0.75$. Platí tedy pravděpodobnostní poměr $\frac{L_{(1)}}{L_{(0)}} = \frac{0.75}{1-0.75} = 3$.

Logistický model lze snadno rozšířit na případ K tříd, kde se předpokládá, že aposteriorní pravděpodobnost $P(G = j \mid x)$ zařazení do j-té kategorie (vektor obsahující nuly a jedničku pouze u j-tého prvku) je podmíněna hodnotou $\boldsymbol{x}$ vektoru znaků (vysvětlujících proměnných). Odpovídající lineární regresní model má tvar

$$\ln \frac{P(G=1|\boldsymbol{x})}{P(G=K|\boldsymbol{x})} = b_{1,0} + \boldsymbol{b}_1^T \boldsymbol{x}. \tag{66}$$

$$\ln \frac{P(G=2|\boldsymbol{x})}{P(G=K|\boldsymbol{x})} = b_{2,0} + \boldsymbol{b}_2^T \boldsymbol{x}. \tag{67}$$

$$\vdots$$

$$\ln \frac{P(G=K-1|\boldsymbol{x})}{P(G=K|\boldsymbol{x})} = b_{K-1,0} + \boldsymbol{b}_{K-1}^T \boldsymbol{x}. \tag{68}$$

Tento model představuje tedy $K - 1$ logitových transformací s omezením, že součet pravděpodobností $\sum_{i=1}^{n} P(G = i|\boldsymbol{x}) = 1$. Po zpětné transformaci pak vychází

$$P(G = j|\boldsymbol{x}) = \frac{\exp(b_{j,0} + \boldsymbol{b}_j^T \boldsymbol{x})}{1 + \sum_{l=1}^{K-1} \exp(b_{l,0} + \boldsymbol{b}_l^T \boldsymbol{x})}. \tag{69}$$

a

$$P(G = K|\boldsymbol{x}) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp(b_{l,0} + \boldsymbol{b}_l^T \boldsymbol{x})}. \tag{70}$$

V dalším výkladu označíme pravděpodobnost $P(G = K|\,\boldsymbol{x}) = p_k(\boldsymbol{x}, \boldsymbol{b})$, aby se zvýraznilo, že jde o funkci regresních parametrů $\boldsymbol{b} = [b_{1,0}, \boldsymbol{b}_1, b_{2,0}, \boldsymbol{b}_2, \ldots, b_{K-1,0}, \boldsymbol{b}_{K-1}]$. Je zřejmé, že pro $K = 2$ přechází tento model na standardní logistický model pro binární vysvětlovanou proměnnou $y = G$.

1. Odhady parametrů

Pro odhad parametrů logistických modelů se standardně používá metoda maximální věrohodnosti. Pro obecný model se předpokládá multinomické rozdělení a pro standardní model se dvěma třídami rozdělení binomické. Přítomnost v první třídě $y = 1$ je pro $G = 1$ a nepřítomnost v první třídě $y = 0$ je pro $G = 2$ (přítomnost ve druhé třídě). Výchozí data jsou ve formě vektoru $\boldsymbol{y}$ rozměru $n \times 1$ a matice $\boldsymbol{X}$ rozměru $n \times m$. Pro i-tý objekt má y_i hodnotu buď 0, nebo 1 a x_i^T je i-tý řádek matice $\boldsymbol{X}$. Označme $p(\boldsymbol{x}, \boldsymbol{b}) = p_1(\boldsymbol{x}, \boldsymbol{b})$ a $1 - p(\boldsymbol{x}, \boldsymbol{b}) = p_2(\boldsymbol{x}, \boldsymbol{b})$. Pak za předpokladu binomického rozdělení y lze zapsat logaritmus věrohodnostní funkce ve tvaru

$$\mathrm{Ln}\, L(\boldsymbol{b}) = \sum_{i=1}^{n}\{y_i \ln p(\boldsymbol{x}_i, \boldsymbol{b}) + (1 - y_i) \ln 1 - p(\boldsymbol{x}_i, \boldsymbol{b})\} =$$

$$= \sum_{i=1}^{n}\{y_i \boldsymbol{b}^T \boldsymbol{x}_i - \ln(1 + \exp \boldsymbol{b}^T \boldsymbol{x}_i)\}, \quad (71)$$

kde $\boldsymbol{b}^T = \{b_0, b_1\}$ a předpokládá se, že první sloupec matice $\boldsymbol{X}$ obsahuje pouze jedničky (absolutní člen). Pro maximalizaci $\ln L(\boldsymbol{b})$ se využívá toho, že první derivace v maximu jsou rovny nule,

$$J = \frac{d \ln L(\boldsymbol{b})}{d\, \boldsymbol{b}} = \sum_{i=1}^{n} x_i (y_i - p(\boldsymbol{x}_i, \boldsymbol{b})). \quad (72)$$

Jde o soustavu $m + 1$ nelineárních rovnic vzhledem k $\boldsymbol{b}$.

2. Interpretace regresních koeficientů

Základní předpoklad v logistické regresi říká, že logaritmus pravděpodobnostního poměru $\ln(L_{(1)}/L_{(0)})$ je lineární funkcí nezávisle proměnných. Žádné předpoklady o rozdělení nezávisle proměnných x zde neexistují. Hlavní výhodou této metody je fakt, že x mohou být jak diskrétní (*faktory*) tak i spojité veličiny (*prediktory*). Logistický model je zapsán ve tvaru (8.11) a nazývá se *vícenásobný logistický regresní model* a koeficienty b_i mohou být interpretovány jako regresní koeficienty. Jsou vyjádřeny v logaritmech, takže bude nutné je přetransformovat zpět a pak bude jejich relativní vliv na pravděpodobnosti možné vysvětlit snadněji. Pro logaritmus pravděpodobnostního poměru $\ln(L_{(1)}/L_{(0)})$ se užívá termín *logit*, popř. *logit transformace pravděpodobnosti*. Vícenásobný logistický regresní model lze upravit při dosazení za $L_{(1)} = (1 - L_{(0)})$ do ekvivalentního tvaru

$$L_{(0)} = \frac{1}{1+\exp[-(b_0+b_1x_1+b_2x_2+\cdots+b_px_p)]}. \quad (73)$$

Kladné znaménko koeficientu b_i zvyšuje pravděpodobnost $L_{(1)}$ a záporné ji snižuje. Je-li bi kladné, funkce exp je větší než 1 a pravděpodobnostní poměr $(L_{(1)}/L_{(0)})$ se bude zvyšovat. Zvýšení se objeví, když predikovaná pravděpodobnost odehrané události $L_{(1)}$ se zvýší a predikovaná pravděpodobnost neodehrané události $L_{(0)}$ se sníží. Proto má model vyšší predikovanou pravděpodobnost odehrané události $L_{(1)}$. Podobně, je-li b_i

záporné, je funkce exp menší než 1 a pravděpodobnost se sníží. Pro koeficient rovný nule vede funkce $\exp 0$ k hodnotě 1 čili k žádné změně pravděpodobnosti.

Všimněme si, že $L_{(1)}$ se mění od 0 do 1 a pravděpodobnostní poměr $L_{(1)}/L_{(0)}$ se mění od 0 do $\propto$. Je-li $L_{(1)}$= 0.5, je poměr $L_{(1)}/L_{(0)}$ roven 1. Ve stupnici pravděpodobnostního poměru odpovídají hodnoty od 0 do 1 hodnotám pravděpodobnosti $L_{(1)}$ od 0 do 0.5. Na druhé straně hodnoty pravděpodobnosti $L_{(1)}$ od 0.5 do 1 vedou k poměru $L_{(1)}/L_{(0)}$ od 1 do ∞. Po zlogaritmování poměru $L_{(1)}/L_{(0)}$ bude tato asymetrie odstraněna, neboť pro $L_{(1)}$ = 0 je $\ln L_{(1)}/L_{(0)} = \infty$, pro $L_{(1)}$ = 0.5 je $\ln L_{(1)}/L_{(0)} = 0$, pro $L_{(1)}$ = 1.0 je $\ln L_{(1)}/L_{(0)} = -\infty$.

3. Test významnosti regresních koeficientů

Logistická regrese umožňuje ověřit hypotézu, že regresní koeficient v logistickém regresním modelu se liší od nuly. Nula zde značí, že pravděpodobnostní poměr $L_{(1)}/L_{(0)}$ se nemění a pravděpodobnost tím pádem není ovlivněna. Test významnosti každého regresního koeficientu je analogický jako u lineární regrese a můžeme proto rovněž použít Studentův t-test k vyšetření statistické významnosti jednotlivých regresních koeficientů. Pro velké výběry se užívá u logistické regrese *Waldovo testační kritérium* $W_{a,i} = \left(b_i/s(b_i)\right)^2$, které vyčísluje statistickou významnost nulové hypotézy pro jednotlivé odhady regresních koeficientů stejně jako ve vícenásobné regresi. Waldova statistika $W_{a,i}$ má χ^2-rozdělení s 1 stupněm volnosti a představuje čtverec poměru odhadu regresního koeficientu a jeho směrodatné odchylky. Pro kategorizované proměnné má $W_{a,i}$ o 1 stupeň volnosti méně, než je počet kategorií. Waldova statistika W_a má jednu nežádoucí vlastnost. Když je absolutní hodnota regresního koeficientu b_i veliká a odhad jeho směrodatné odchylky $s(b_i)$ je také veliký, je výsledkem příliš malá hodnota testačního kritéria $W_{a,i}$, která vede k selhání zamítnutí nulové hypotézy, že regresní koeficient je nulový. Proto, je-li regresní koeficient veliký, by se nemělo užívat

Waldova kritéria. Místo toho se měl použít logistický model s touto proměnnou a bez této proměnné a vyšetřit změnu rovnic v logitovém kritériu.

4. Parciální korelace

Podobně jako v lineární regresi tak i v logistické regresi je obtížné určit příspěvek jednotlivých proměnných. Příspěvek každé proměnné závisí na ostatních proměnných v logistickém modelu. Statistika, která se užívá k vyšetření parciální korelace mezi závisle proměnnou a každou nezávisle proměnnou se nazývá *korelační koeficient* R_i, jehož hodnoty leží v intervalu od −1 do +1. Kladné hodnoty ukazují, že když roste hodnota R_i, zvyšuje se pravděpodobnost objektu v „události“ $L_{(1)}$. Je-li R_i záporné, je tomu naopak. Malé hodnoty R_i ukazují, že proměnná má malý vliv na model. Korelační koeficient R_i se vyčíslí vztahem

$$R_i = \pm\sqrt{\frac{W_{a,i}-2df}{-2\ln L_{(0)}}}, \tag{74}$$

kde df značí počet stupňů volnosti a týká se počtu odhadovaných parametrů. Ve jmenovateli je záporný dvojnásobek logaritmu pravděpodobnosti základního logistického modelu, který neobsahuje žádné proměnné kromě absolutního členu (úseku). Je-li Waldova statistika $W_{a,i}$ menší než $2df$, je $R_i = 0$.

10. ANALÝZA SHLUKŮ (CLU)

Analýza shluků (Cluster analysis, CLU) patří mezi metody, které se zabývají vyšetřováním podobnosti *vícerozměrných objektů* (tj. objektů, u nichž je změřeno větší množství proměnných) a jejich klasifikací do tříd čili shluků. Hodí se zejména tam, kde objekty projevují přirozenou tendenci se seskupovat. Navzdory starému přísloví, že opaky se přitahují, v přírodě se ukazuje, že platí spíše pravidlo, že podobné věci se sjednocují. Nejenom ptáci podobného peří ale také ostatní živočichové, kteří sdílejí podobné či stejné vlastnosti mají tendenci se seskupovat, shlukovat. V biologii se proto užívá shluková analýza ke klasifikování živočichů a rostlin. Tato klasifikace se nazývá numerická taxonomie.

V medicíně analýza shluků identifikuje nemoci a jejich stadia. I když analýza shluků a diskriminační analýza klasifikují objekty do kategorií, diskriminační analýza vyžaduje znát předem příslušnost objektů do tříd, aby se odvodilo klasifikační pravidlo. Když chceme například rozlišit mezi několika nemocemi, musíme předem znát diagnózy pacientů, jež nám poslouží ke klasifikaci. Potom na základě příslušnosti do jednotlivých tříd odvozuje diskriminační analýza pravidlo k umístění dosud nediagnostikovaných nezařazených pacientů. V analýze shluků je neznámá příslušnost do tříd všech objektů. Dokonce i počet tříd či shluků je neznámý.

Lze formulovat tři hlavní cíle analýzy shluků:

- *popis systematiky*, je tradičním využitím shlukové analýzy pro průzkumové cíle a taxonomii, což je empirická klasifikace objektů,
- *zjednodušení dat*, kdy analýza shluků poskytuje při hledání taxonomie zjednodušený pohled na objekty,
- *identifikaci vztahu*, kdy po nalezení shluků objektů, a tím i struktury mezi objekty, je snadnější odhalit vztahy mezi objekty.

Cíle shlukové analýzy nelze oddělit od hledání a volby vhodných znaků k charakterizování shlukovaných objektů. Nalezené shluky vystihují strukturu dat pouze s ohledem na vybrané znaky. Volba znaků musí být provedena na základě teoretických, pojmových a praktických hledisek. Vlastní shluková

analýza neobsahuje techniku k rozlišení významných a nevýznamných znaků. Provede pouze odlišení shluků. Nesprávné zařazení znaků vede k zahrnutí i odlehlých objektů, které mohou mít rušivý vliv na výsledky analýzy. Měly by být využity pouze takové znaky, které dostatečně rozlišují mezi objekty.

1. Vliv odlehlých objektů

Při odhalování struktury objektů je shluková analýza velmi citlivá na přítomnost nevýznamných znaků. Je ovšem citlivá také na přítomnost *odlehlých objektů*, které se silně odlišují ode všech ostatních objektů. Odlehlé objekty mohou představovat buď skutečně odchýlené, patologické objekty, které nejsou představiteli analyzované populace, nebo chybný výběr objektu z populace, který způsobí nevhodné zastoupení původní populace. V obou případech odchýlené objekty zbortí strukturu dat a způsobí, že nalezené shluky nebudou odrážet skutečnou strukturu analyzované populace. Nejsnadnějším způsobem předběžného odhalení odlehlých objektů je *profilový diagram znaků*. Je možné použít také další způsoby znázornění vícerozměrných dat, které umožňují indikaci vybočujících objektů. Postupy průzkumové analýzy vícerozměrných dat se stávají těžkopádné při velkém počtu objektů nebo znaků. Vybočující objekty je obvykle třeba z dat odstranit, i když si musíme býti vědomi, že jejich odstraněním se mnohdy zbortí aktuální struktura. Vypouštění objektů by proto mělo být velmi uvážlivé.

2. Míry podobnosti

Myšlenka podobnosti objektů je v analýze shluků základní. Podobnost mezi objekty je užita jako kritérium tvorby shluků objektů. Nejdříve se stanovují znaky určující podobnost, které se dále kombinují do podobnostních měr. Tímto způsobem pak může být objekt porovnán s jiným objektem. Analýza shluků vytváří shluky podobných objektů. Meziobjektová podobnost může být měřena rozličnými způsoby, které se dají obyčejně zařadit do jedné ze tří základních skupin, a to míry korelace, míry vzdálenosti a míry asociace. Každá z nich představuje zvláštní pohled na podobnost, která je závislá na objektech a na typu dat. Korelační a vzdálenostní míry jsou míry metrických dat, zatímco asociační míry jsou určeny spíše pro nemetrická data.

● **Korelační míry**. Základní mírou podobnosti dvou objektů či znaků x_i a x_j, vyjádřených v kardinální škále může být *Pearsonův párový korelační koeficient* r. Objekty jsou si tím podobnější, čím je jejich párový korelační koeficient větší a bližší jedné. V případě ordinální škály je analogickou mírou podobností *Spearmanův korelační koeficient*. Obyčejně se vychází z transponované matice dat $\boldsymbol{X}^T$, kdy sloupce představují objekty a řádky pak znaky. Korelační koeficienty mezi dvěma sloupci matice $\boldsymbol{X}^T$ představují korelaci mezi dvojicí objektů. Tomu odpovídá podobnost jejich profilů v profilovém diagramu. Vysoká korelace prozrazuje vysokou „podobnost" a nízká korelace pak „nepodobnost" profilů. Korelační míry představují podobnost odpovídajících si „vzorů" posuzovanou přes všechny znaky. Korelační míry se však užívají zřídka, protože v praktické analýze je kladen důraz více na velikosti objektů než na tvar jejich profilových křivek.

● **Míry vzdálenosti**. Představují nejčastěji užívané míry založené na prezentaci objektů v prostoru, jehož souřadnice tvoří jednotlivé znaky. Pokud tyto míry splňují požadavky symetrie $d(x,y) = d(y,x)$ a trojúhelníkovou nerovnost $d(x,y) \leq d(x,z) + d(y,z)$, jde o tzv. *metriky*. Při porovnání na profilovém diagramu je zřejmé, že vzdálenosti se zaměřují na velikost hodnot a vyhledání křivek v profilovém diagramu, které jsou blízko sebe, i když u jednotlivých znaků velmi odlišného tvaru. Použití korelace vede na jiné shluky než použití vzdálenostních měr. Nejčastější vzdálenostní mírou je *eukleidovská vzdálenost* zvaná také *geometrická metrika*, která představuje délku přepony pravoúhlého trojúhelníka a její výpočet je založen na Pythagorově větě. Platí, že

$$d_E(\boldsymbol{x}_k, \boldsymbol{x}_l) = \sqrt{\sum_{j=1}^{m} \left(x_{kj} - x_{lj}\right)^2} \tag{75}$$

představuje standardní typ vzdálenosti. Vedle eukleidovské vzdálenosti se užívá také *čtverec eukleidovské vzdálenosti*, který tvoří základ Wardovy metody shlukování. Existuje několik dalších vzdálenostních měr. Často je užívaná *manhattanská vzdálenost* zvaná také *vzdálenost městských bloků* nebo *Hammingova metrika*, definovaná vztahem

$$d_H(\boldsymbol{x}_k, \boldsymbol{x}_l) = \sum_{j=1}^{m} \left| x_{kj} - x_{lj} \right|. \tag{76}$$

Před užitím této vzdálenosti se musíme ujistit, že znaky spolu nekorelují. Když tato podmínka není splněna, shluky jsou nesprávné. Další mírou je zobecněná *Minkovského metrika*, pro kterou platí

$$d_M(\boldsymbol{x}_k, \boldsymbol{x}_l) = \sqrt[z]{\sum_{j=1}^{m} |x_{kj} - x_{lj}|^z}, \tag{77}$$

kde pro $z = 1$ jde o Hammingovu metriku a pro $z = 2$ o Eukleidovu. Čím je z větší, tím více je zdůrazňován rozdíl mezi vzdálenými objekty. V některých případech se používá také *tětivová vzdálenost* (anglicky chord distance), definovaná vztahem

$$d_{CH}(\boldsymbol{x}_k, \boldsymbol{x}_l) = \sqrt{2\left(1 - \frac{\sum_{j=1}^{m} x_{kj} x_{lj}}{\sum_{j=1}^{m} x_{kj}^2 \sum_{j=1}^{m} x_{lj}^2}\right)}. \tag{78}$$

V případě třech znaků je tětivová vzdálenost přímou vzdáleností dvou bodů na povrchu koule s jednotkovým poloměrem a počátkem v těžišti.

Problém všech vzdálenostních měr vzniká při použití nestandardizovaných dat, které mohou způsobit rozdíly mezi shluky, a to odlišností jednotek měření. Shluky různých vzdálenostních měr se budou lišit, největší rozptýlení mezi shluky bude u čtverce eukleidovské vzdálenosti. Pořadí podobností se významně změní se změnou měřítka nebo změnou jednotek jednoho ze znaků. Škálování znaků má veliký dopad na konečné řešení. Standardizace znaků by se měla vždy využít, pokud to je jenom možné.

Všechny dosud uvedené metriky neuvažují závislost mezi znaky. Zahrneme-li do vztahu pro vzdálenost i vazby mezi znaky, vyjádřené kovarianční maticí C, dostaneme statistickou míru, kterou nazýváme *Mahalanobisova metrika*

$$d_{Ma}(\boldsymbol{x}_k, \boldsymbol{x}_l) = \sqrt{(\boldsymbol{x}_k - \boldsymbol{x}_l)^T \boldsymbol{C}^{-1} (\boldsymbol{x}_k - \boldsymbol{x}_l)}, \tag{79}$$

Jde vlastně o vzdálenost bodů v prostoru, jehož osy nemusí být ortogonální. Vysoce korelovaný výběr znaků může skrytě převážit celý soubor znaků shlukování.

● **Rozdíly mezi Eukleidovou a Mahalanobisovou vzdáleností u dvou znaků.** Ukažme rozdíly mezi Eukleidovou a Mahalanobisovou vzdáleností dvou znaků x_1 a x_2. Vektor průměrů je v tomto případě $\overline{\boldsymbol{x}} = (\overline{\boldsymbol{x}}_1, \overline{\boldsymbol{x}}_2)$ a pro kovarianční matici platí že

$$\boldsymbol{S} = \begin{pmatrix} s_1^2 & rs_1s_2 \\ rs_1s_2 & s_2^2 \end{pmatrix}, \tag{80}$$

kde r je párový korelační koeficient a s_1 a s_2 jsou výběrové směrodatné odchylky. Inverzí této matice dostaneme

$$\boldsymbol{S^{-1}} = \frac{1}{s_1^2 s_2^2 (1-r)} \begin{pmatrix} s_2^2 & -rs_1s_2 \\ -rs_1s_2 & s_1^2 \end{pmatrix}, \tag{81}$$

Eukleidovská vzdálenost i-tého objektu od těžiště x je definována vztahem

$$d_E(\boldsymbol{x}_j, \overline{\boldsymbol{x}}) = \sqrt{(x_{i1} - \bar{x}_1)^2 + (x_{i2} - \bar{x}_2)^2} \tag{82}$$

a Mahalanobisova vzdálenost i-tého objektu od těžiště x je rovna

$$d_{Ma}(\boldsymbol{x}_j, \overline{\boldsymbol{x}}) = \sqrt{(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})} =$$

$$= \sqrt{\left(\frac{x_{i1} - \bar{x}_1}{s_1}\right)^2 + \left(\frac{x_{i2} - \bar{x}_2}{s_2\sqrt{1-r^2}} - \frac{r(x_{i1} - \bar{x}_1)}{s_1\sqrt{1-r^2}}\right)^2}. \tag{83}$$

Porovnáním těchto dvou výrazů je zřejmé, že díky vzájemné korelaci, vyjádřené korelačním koeficientem r, se do Mahalanobisovy vzdálenosti promítne jen část rozdílu $x_{i2} - \bar{x}_2$, zmenšená o podíl první proměnné, který je úměrný korelaci. Je patrné, že pro nekorelovaná data s jednotkovými rozptyly bude $d_E = d_{Ma}$. Uveďme, že vztah $d_E(x, \bar{x}) = konst$ definuje kružnici, kdežto vztah $d_{Ma}(x, \bar{x}) = konst$ definuje elipsu. Zobecněná Mahalanobisova vzdálenost vyjadřuje vzdálenost mezi objekty podobně jako koeficient determinace R^2 v regresní analýze.

Při volbě vhodné míry vzdálenosti bychom měli brát do úvahy fakt, *že různé míry vzdálenosti nebo změna měřítka znaků vedou k různým shlukům*. Doporučuje se proto užívat několik měr a porovnat výsledky s teoretickými

nebo dosud známými tvary shluků. Když jsou znaky vzájemně korelovány, jeví se Mahalanobisova vzdálenost jako nejlepší, protože zařazuje korelace do výpočtu a váží všechny znaky stejně. ❒

● **Míry asociace**. Míry asociace, resp. podobnosti (binární) se používají k porovnání objektů, pokud jsou jejich znaky nemetrického charakteru (binární proměnné). Uveďme příklad, kdy respondent odpověděl na řadu otázek odpovědí *ano* nebo *ne*. Míra asociace pak vyjadřuje stupeň souhlasu každého páru respondentů. Nejjednodušší mírou asociace bude procento souhlasu, kdy oba respondenti na danou otázku odpověděli *ano* nebo *ne*. Rozšíření tohoto jednoduchého „souhlasného koeficientu“ je podstatou míry asociace k vyhodnocování více kategorií nominálních nebo ordinálních znaků. Přehled různých typů koeficientů asociace lze nalézt v pracích [2] a [4]. Uvedeme příklad, kdy budeme předpokládat, že sledujeme asociaci mezi dvěma objekty O_i a O_j. Možné binární odezvy typu 0-1 je pak možné zapsat do tzv. kontingenční tabulky.

Kontingenční tabulka

	Objekt O_i 1		
Objekt O_j	0	1	0
	1	a	b
	0	c	d

V této tabulce jsou shrnuty všechny možné kombinace počtu znaků pro dva objekty:

a udává počet znaků, kde mají oba objekty O_j a $O_i$hodnotu 1 a jde o tzv. *pozitivní shodu*,

b představuje počet znaků, kde má objekt O_j hodnotu 1 a objekt $O_i$hodnotu 0.

c vyjadřuje počet znaků, kde má objekt $O_j$hodnotu 0 a objekt $O_i$hodnotu 1.

d značí počet znaků, kde mají oba objekty O_ja $O_i$hodnotu 0 a jde o tzv. *negativní shodu.*

Míry asociace pak vyjadřují relativní podíly počtu znaků s ohledem na to, zda má smysl uvažovat negativní shodu, nebo zda má nulová hodnota znaku u porovnávaných objektů stejnou příčinu. Mezi základní koeficienty podobnosti patří:

a) *Sokalův−Michenerův koeficient asociace (čili koeficient jednoduché shody)*

$$S_{SM} = \frac{a+d}{a+b+c+d}, \tag{84}$$

b) *Russelův−Raoův koeficient asociace*

$$S_{RR} = \frac{d}{a+b+c+d}, \tag{85}$$

c) *Jaccardův koeficient*, jenž se liší od předešlého Russelova−Raova koeficientu asociace jen tím, že má v čitateli počet shodných znaků a,

d) *Hamannův koeficient asociace*

$$S_H = \frac{a+b-c-d}{a+b+c+d}, \tag{86}$$

a také lze konstruovat následující obdobu nazývanou

e) *korelační koeficient*

$$r_B = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}, \tag{87}$$

f) *Rogersův a Tanimotův koeficient asociace*, definovaný vztahem

$$S_{RT} = \frac{a+d}{a+2b+2c+d}, \tag{88}$$

g) *Sörensenův koeficient asociace*, který je definován vztahem

$$S_S = \frac{2a}{2a+b+c} \tag{89}$$

h) Specifický případ pro binární data představují tzv. *genetické vzdálenosti* používané při studiu délkového polymorfismu fragmentů DNA, které ukážeme na příkladu 9.1.

V některých případech se pracuje se smíšenými daty, tj. jisté znaky jsou nominální nebo binární a jisté znaky jsou kardinální.

i) *Growerův koeficient podobnosti* je definován vztahem

$$S_{Gij} = \frac{\sum_{k=1}^{m} w_{ijk} S_{ijk}}{\sum_{k=1}^{m} w_{ijk}} \tag{90}$$

kde i, j jsou indexy dvou objektů charakterizovaných znaky $k = 1, \ldots, m$. Váha znaku w_{ijk} může nabýt hodnoty 1 nebo 0 podle toho, zda lze srovnávat hodnoty znaku k u objektů i a j a Sijk označuje skóre (hodnotu) pro k-tý znak. Pro všechny škály měření uvažujme jen případy, kdy jsou hodnoty znaku x_{ik} a x_{jk} známé. Pro nominální data se provádí libovolné číslování začínající od 1, tedy $w_{ijk} = 1$. Pokud nejsou hodnoty nějakého znaku známé, je automaticky $w_{ijk} = 0$. Platí, že

1. pro *nominální znaky* je $S_{ijk} = 0$ pro $x_{ik} \approx x_{jk}$ a $S_{ijk} = 1$ pro $x_{ik} = x_{jk}$.

2. pro *binární znaky* je $S_{ijk} = w_{ijk} = 1$ pro případ, že $x_{ik} = x_{jk} = 1$, resp. $x_{ik} = x_{jk} = 0$. Pokud je $x_{ik} \approx x_{jk}$, je $S_{ijk} = 0$ a $w_{ijk} = 1$. V případě, že $x_{ik} = x_{jk} = 0$ a negativní shoda nemá smysl, je $S_{ijk} = w_{ijk} = 0$.

3. pro *kardinální znaky* se počítá

$$S_{ijk} = 1 - \frac{|x_{ij} - x_{jk}|}{R_k} \tag{91}$$

kde R_k je rozpětí čili rozdíl mezi maximální a minimální hodnotou znaku u všech objektů. Jde o manhattanskou metriku.

Lze také definovat vzdálenost pro smíšená data [16]. Na základě měr podobnosti objektů se konstruují míry podobnosti mezi objekty a třídami a míry podobnosti mezi třídami.

3. Standardizace dat

Před vlastní shlukovou analýzou je třeba řešit otázku, zda je třeba *data standardizovat*. Musí se respektovat fakt, že většina měr vzdálenosti je velmi citlivá na měřítka (stupnice), vedoucí k různé numerické velikosti znaků. Obecně platí pravidlo, že znaky s větší mírou proměnlivosti čili větší směrodatnou odchylkou mají větší vliv na míru podobnosti. Standardizace se týká jednak standardizace znaků a jednak standardizace objektů.

· *Standardizování znaků*. Nejužívanější formou standardizace je normalizace každého znaku do svého Z-skóre, a to odečtením průměru a dělením směrodatnou odchylkou. Tato standardizace je známa pod názvem *normovací Z-funkce*. Tato transformace eliminuje rozdíly v měřítku, mnohdy i řádově se lišících znaků. Výhody standardizace znaků jsou následující:

a) Znaky lze v jednotném měřítku, kde je průměrná hodnota 0 a směrodatná odchylka 1, vzájemně porovnávat snadněji. Kladné hodnoty jsou nad průměrem a záporné hodnoty jsou pod průměrem,

b) Se změnou měřítka nedojde k rozdílu mezi standardizovanými znaky. I když změníme například u znaku čas jednotky z minut na sekundy, standardizované hodnoty budou stále stejné.

Standardizace znaků eliminuje efekty rozdílů stupnic mezi rozličnými znaky, efekty jednotek u jednoho a téhož znaku. Uživatel by však přesto měl být opatrný při standardizaci a neměl by standardizovat paušálně. Existuje-li nějaký „přirozený“ vztah, obsahující škálování znaků, pak se standardizace nejeví jako vhodná. Rozhodnutí zda standardizovat znaky musí být založeno na nějakém empirickém či pojmovém důvodu.

◆ *Standardizace objektů*. Nabízí se otázka, zda lze standardizovat jenom znaky nebo také objekty? Když chceme identifikovat shluky dle vzdálenosti, pak

standardizace není vhodná. Standardizace objektů nebo-li řádková standardizace může být však efektní ve speciálních případech. V některých případech je výhodné provádět také vhodnou transformaci dat vedoucí k odstranění nekonstantnosti rozptylu, popř. zešikmení rozdělení dat.

4. Způsoby shlukování

Shluk (cluster) je skupina objektů, jejichž vzdálenost (nepodobnost) je menší než vzdálenost, resp. nepodobnost, kterou mají objekty do shluku nepatřící. Podle způsobu shlukování se postupy dělí na *hierarchické a nehierarchické shlukování*. Hierarchické se dělí dále na *aglomerační a divizní shlukování.*

● **Hierarchické shlukovací postupy**. Jsou založeny na hierarchickém uspořádání objektů a jejich shluků. Graficky se hierarchicky uspořádané shluky zobrazují formou *vývojového stromu* nebo *dendrogramu*. U *aglomeračního shlukování* se dva objekty, jejichž vzdálenost je nejmenší, spojí do prvního shluku a vypočte se nová matice vzdáleností, v níž jsou vynechány objekty z prvního shluku, a tento shluk je pak zařazen jako objekt. Celý postup se opakuje tak dlouho, dokud všechny objekty netvoří jeden velký shluk, nebo dokud nezůstane určitý, předem zadaný počet shluků. Postup *divizního shlukování* je obrácený. Vychází se z množiny všech objektů jako jediného shluku a jeho postupným dělením získáme systém shluků, až skončíme ve stadiu jednotlivých objektů. Výhodou hierarchických metod je nepotřebnost informace o optimálním počtu shluků v procesu shlukování; tento počet se určuje až dodatečně. Při shlukování vznikají pouze dva základní problémy, prvním je způsob vyjádření podobnosti mezi objekty a druhým je volba vhodné shlukovací procedury. Volba vhodné shlukovací procedury souvisí se zvoleným způsobem vyjádření metriky. Mezi metody metriky shlukování patří:

● *Metoda nejbližšího souseda*. Postup je postaven na minimální vzdálenosti. Naleznou se dva objekty oddělené nejkratší vzdáleností a umístí se do shluku. Další shluk je vytvořen přidáním třetího nejbližšího objektu. Proces se opakuje, až jsou všechny objekty v jednom společném shluku. Vzdálenost mezi dvěma shluky je definována jako nejkratší vzdálenost libovolného bodu ve shluku vůči libovolnému bodu ve shluku jiném. Dva shluky jsou propojeny v libovolném

stadiu nejkratší spojkou. Častou nevýhodou metody nejbližšího souseda je řetězový efekt, kdy se spojují shluky, jejichž dva objekty jsou sice nejbližší, ale vzhledem k většině ostatních objektů nejde o nejbližší shluky.

● *Metoda nejvzdálenějšího souseda*. Jde o metodu podobnou předchozí kromě toho, že kritérium je postaveno nikoliv na minimální, ale na maximální vzdálenosti. Nejdelší vzdálenost mezi objekty v každém shluku představuje nejmenší kouli, která obklopuje všechny objekty v obou shlucích. Metoda se také nazývá *metodou úplného propojení*, protože všechny objekty ve shluku jsou propojeny každý s každým při maximální vzdálenosti čili minimální podobnosti. Můžeme říci, že podobnost uvnitř shluku je rovna průměru shluku. Obě míry vystihují pouze jeden aspekt. Nejkratší vzdálenost postihuje pouze jednoduchý pár nejtěsnějších objektů a nejvzdálenější postihuje také jediný pár, ale pár dvou extrémů.

● *Metoda průměrné vzdálenosti*. Kritériem vzniku shluků je průměrná vzdálenost všech objektů v jednom shluku ke všem objektům ve druhém shluku. Takové techniky nezávisí na extrémních hodnotách, jako je tomu u nejbližšího souseda nebo u nejvzdálenějšího souseda, ale vznik shluku závisí na všech objektech shluku a ne jenom na jediném páru dvou extrémních objektů.

● *Wardova metoda*. Principem není optimalizace vzdáleností mezi shluky ale minimalizace heterogenity shluků podle kritéria minima přírustku vnitroskupinového součtu čtverců odchylek objektů od těžiště shluků. V každém kroku se pro všechny dvojice odchylek spočítá přírůstek součtu čtverců odchylek, vzniklý jejich sloučením a pak se spojí ty shluky, kterým odpovídá minimální hodnota tohoto přírůstku. V případě, že shluk tvoří k objektů, které jsou charakterizovány m znaky, je k dispozici matice $k \times m$ s prvky x_{ij} (hodnota j-tého znaku pro k-tý objekt). Vnitroshluková variabilita VSS je pak dána vztahem

$$VSS = \sum_{j=1}^{m} \sum_{i=1}^{k} \left(x_{ij} - \bar{x}_j\right)^2, \tag{92}$$

kde $\bar{x}_j = \frac{1}{k}\sum_{i=1}^{k} x_{ij}$. Přidáváním dalších shluků s k_1 objekty se zvětší počet řádků výchozí matice na $k + k_1$ a VSS se počítá pro větší počet objektů. Pokud se začíná od jednoprvkových shluků, bude výchozí $VSS = 0$. Tento postup má tendenci kombinovat shluky s malým počtem objektů.

● *Metoda těžiště*. U této metody jde o vzdálenost dvou těžišť shluků vyjádřených eukleidovskou vzdáleností nebo čtvercem eukleidovské vzdálenosti. Těžiště shluku má souřadnice odpovídající průměrným hodnotám objektů pro jednotlivé znaky. Po každém kroku shlukování se počítá nové těžiště. Poloha těžiště shluku poněkud migruje tak, jak se připojují nové objekty a vznikají větší shluky. Mohou se objevit také zmatečné shluky, když vzdálenost mezi těžišti jednoho páru je menší než vzdálenost mezi těžišti jiného páru, utvořeného v předešlém kroku. Výhodou této metody je menší ovlivnění odlehlými body, než je tomu u ostatních hierarchických metod.

● *Metoda mediánová*. Jde o jisté vylepšení metody těžiště, neboť se snaží odstranit rozdílné významnosti, které metoda těžiště dává různě velkým shlukům.

10.1. Dendrogramy hierarchického shlukování

Analýzou shluků je možné hodnotit jednak podobnost objektů, analyzovanou pomocí dendrogramu objektů, a jednak podobnost znaků analyzovanou pomocí dendrogramu znaků.

Dendrogram shluků (vývojový strom) se konstruuje pouze v případě, kdy je k dispozici matice původních znaků.

Dendrogram podobnosti objektů je standardní výstup hierarchických shlukovacích metod, ze kterého je patrná struktura objektů ve shlucích.

Dendrogram podobnosti znaků odhaluje nejčastěji dvojice či trojice (obecně m-tice) znaků, které jsou si velmi podobné a silně spolu korelují. Znaky, které jsou ve společném shluku, si jsou značně podobné a jsou také vzájemně nahraditelné. To má značný význam při plánování experimentu a respektování

úsporných ekonomických kritérií. Některé vlastnosti či znaky není třeba vůbec měřit, protože jsou snadno nahraditelné jinými a nepřispívají do celku velkou vypovídací schopností.

11. MAPOVÁNÍ OBJEKTŮ VÍCEROZMĚRNÝM ŠKÁLOVÁNÍM (MDS)

Vícerozměrné škálování objektů (MultiDimensional Scaling, MDS) je technika vytvoření subjektivní mapy relativního umístění objektů v rovině dvojrozměrného grafu, a to na základě vzdáleností či podobností mezi objekty, tzv. *matice proximity* (blízkosti). Cílem MDS je detekovat základní souřadnice, které dovolují objasnit pozorované podobnosti či vzdálenosti mezi vyšetřovanými objekty. Zatímco ve faktorové analýze jsou podobnosti mezi objekty vyjádřeny v korelační matici, v MDS lze vedle korelační matice analyzovat jakýkoliv druh podobnostní či vzdálenostní matice.

11.1. Zaměření metody MDS

Vícerozměrné škálování je založeno na vzájemném porovnávání objektů. Proto mapa objektů může obsahovat jeden, dva, tři a zřídka i více rozměrů nazvaných *souřadnic*. Představme si matici vzdáleností mezi městy. MDS analýzou chceme vytvořit novou matici poloh měst čili mapu, kde polohu každého města budeme vyjadřovat dvěma souřadnicemi, sever-jih na y-nové ose a západ-východ na x-ové ose. Prvním úkolem bude určení počtu vhodných souřadnic k vytvoření mapy měst, obvykle zvolíme 2. Druhým úkolem bude přepočítat zadané vzdálenosti mezi městy na souřadnice polohy každého města na mapě. Podobně jako ve faktorové analýze bývá zde i třetí úkol otočení souřadnicových os tak, aby co nejlépe vystihoval polohy měst a aby se poloha dalace nejlépe vysvětlit.

Každý objekt je popsán svými dimenzemi zvanými také znaky, a to *znaky subjektivními* (znaky vnímané člověkem) a *znaky objektivními* (znaky měřitelné fyzikálně), viz Hair a kol., [11].

Zatímco subjektivní znaky jsou nemetrické, postavené na názoru respondenta (například hezký, ošklivý, laciný, drahý, kvalitní, nekvalitní), objektivní znaky jsou metrické veličiny přístroji měřitelné. Mezi objektivními a subjektivními znaky objektu jsou určité rozdíly, jimiž jsou:

a) Individuální rozdíly, kdy znaky chápané člověkem nemusí být v souhlasu s objektivními znaky. Očekává se, že objekty se budou lišit v subjektivních znacích, ale musí se připustit, že rozdíly mohou být i v objektivních znacích. Objekty se mohou lišit také v důležitosti přikládané každému jednotlivému znaku.

b) Vzájemná závislost vyjadřující, že objektivní a subjektivní znaky se mohou ovlivňovat jeden s druhým za vzniku neočekávaných výsledků, například jemný nápoj se může subjektivně jevit sladší než jiný, protože má ovocnou vůni, i když objektivně oba obsahují stejné množství cukru. Doporučuje se proto nejprve pochopit subjektivní znaky, a pak je teprve dávat do souvislostí s objektivními. Dalším rozborem by se mělo určit, které vlastnosti předvídají polohu objektu v subjektivních i objektivních znacích. Měli bychom být opatrni při vysvětlování znaků. Tento proces se jeví spíše uměním než vědou, musíme se totiž bránit ovlivnění subjektivním názorem kvalitativního, i když objektivního znaku.

Užitečnost metody vícerozměrného škálování MDS spočívá v tom, že můžeme analyzovat libovolnou matici vzdáleností (tj. nepodobností) a podobností. Podobnosti mohou vyjadřovat hodnocení objektů člověkem, procento souhlasu mezi rozhodčími, počet selhání subjektu na podnět, atd.

11.2. Podstata metody MDS

Matice proximity čili vstupní data mohou být trojího typu a mohou obsahovat:

- vzdálenosti mezi objekty d_{ij} čili nepodobnosti,
- podobnost mezi objekty S_{ij},
- hodnoty proměnných (sloupce) pro jednotlivé objekty (řádky) X_{ij}.

Mohou být také definovány maticemi $\boldsymbol{D}$, $\boldsymbol{S}$, $\boldsymbol{X}$ sestavenými z těchto hodnot, jak je běžné například v softwarech SPSS [12] nebo NCSS2000 [8].

Vzdálenost (nepodobnost) d_{ij} představuje vzdálenost mezi objekty, může být měřena přímo, jako například vzdálenost dvou měst. Vícerozměrné škálování objektů užívá vzdálenost v datech přímo a matice vzdáleností mezi objekty D je symetrická.

Podobnost S_{ij} vyjadřuje, jak blízko se nacházejí dva objekty. Vícerozměrné škálování objektů MDS umožňuje načíst míru podobnosti pro každý pár objektů. Matice podobností objektů $\boldsymbol{S}$ je opět symetrická. Podobnost objektů lze konvertovat do vzdálenosti vzorcem

$$d_{ij} = \sqrt{S_{ii} + S_{jj} - 2S_{ij}}, \tag{93}$$

kde d_{ij} představuje vzdálenost i-tého a j-tého objektu a S_{ij} jejich podobnost.

Hodnoty znaků x_{ij} představující proměnné pro jednotlivé objekty představují spíše standardní míry. Z nich se vypočte nejprve korelační matice objektů $\boldsymbol{R}$ a potom matice eukleidovských vzdáleností či matice Mahalanobisových vzdáleností objektů $\boldsymbol{D}$.

Vícerozměrné škálování objektů neužívá na rozdíl od ostatních vícerozměrných technik náhodné proměnné. Vytváří si svou „náhodnou proměnnou“, tj. subjektivní *souřadnici* vhodnou k vzájemnému porovnání objektů. Tyto souřadnice jsou odvozovány z celkových měr podobnosti mezi objekty. Jde o volbu závisle proměnné (čili podobnost mezi objekty) a k ní pak přiřazujeme nezávisle proměnné (tj. subjektivní souřadnice). Vícerozměrné škálování objektů má výhodu v redukování vlivu člověka tím, že nepožaduje přesné vymezení užívaných proměnných při vzájemném porovnávání objektů, jako je to vyžadováno třeba v analýze shluků. Nevýhodou však je, že uživatel si není jist, které proměnné respondent k porovnání objektů vůbec použil.

Technika vícerozměrného škálování objektů vyčíslí *metrické klasické řešení* (CMDS) nebo nemetrické řešení (NNMDS) a vychází buď přímo z experimentálních hodnot $\boldsymbol{X}$, z korelační matice $\boldsymbol{R}$, nebo z matice podobností $\boldsymbol{S}$ či matice vzdáleností $\boldsymbol{D}$. Pro n objektů existuje $n(n-1)/2$ podobností, resp. vzdáleností mezi páry rozličných objektů. Tyto podobnosti objektů pak tvoří vstupní data. Když podobnosti objektů nemohou být kvantifikovány, jako například podobnost mezi barvami, užijeme za vstupní data pořadová čísla podobností objektů. Vzdálenost mezi dvěma objekty je eukleidovská, počítaná na základě Pythagorovy věty. I když vynášíme vzdálenosti do dvojrozměrného grafu roviny, může být d_{ij} vyčísleno na základě většího počtu proměnných

$m \geq 2$. S růstem počtu objektů však roste i počet souřadnic, takže pro tři objekty je to dvojrozměrná rovina, pro čtyři objekty pak trojrozměrný prostor atd.

Předpokládejme dále, že podobnosti objektů S lze vzestupně uspořádat do řady

$$S_{i_1,j_1} \leq S_{i_2,j_2} \leq S_{i_3,j_3} \leq \cdots \leq S_{i_m,j_m}, \tag{94}$$

kde S_{i_1,j_1} značí nejmenší podobnost ze všech m podobností. Index indikuje pár objektů nejméně si podobných a obsahuje proto objekty o pořadovém čísle 1. Chceme nalézt q-rozměrnou sestavu n objektů tak, že vzdálenosti mezi párem objektů souhlasí s pořadovými čísly uvedených podobností. *Řada pořadových čísel* objektů

$$d_{i_1,j_1}^{(q)} \leq d_{i_2,j_2}^{(q)} \leq d_{i_3,j_3}^{(q)} \leq \cdots \leq d_{i_m,j_m}^{(q)}, \tag{95}$$

popisuje sestupné řazení vzdáleností párů objektů od největší vzdálenosti objektů v páru do nejmenší. Když budeme používat tato pořadová čísla, pak vlastní vzdálenosti mezi objekty budou pro nás nedůležité. Dále se zaměřme na dva parametry, a to na posouzení *kritéria maximální věrohodnosti* a stanovení *počtu souřadnic.*

Jak těsně prokládá model vícerozměrného škálování objektů MDS daná experimentální data lze posoudit *testem těsnosti proložení* využitím statistické míry *stress*. Tato míra je založena na rozdílu mezi skutečnou vzdáleností dvou objektů d_{ij} a modelem predikovanou čili vypočtenou vzdáleností dle vzorce

$$stress = \sqrt{\frac{\sum_{k=1}^{m}\left(d_{ij}-d_{ij,vyp}\right)^2}{\sum_{k=1}^{m} d_{ij}^2}}, \tag{96}$$

kde $d_{ij,vyp}$ je vypočtená vzdálenost objektů, založená na modelu MDS. Vypočtená hodnota závisí především na počtu užitých souřadnic a metrickém či nemetrickém algoritmu. Je-li hodnota stress blízká nule, jeví se proložení vícerozměrným škálováním objektů jako nejlepší. Obecně platí pravidlo, čím

menší je hodnota stress, tím těsnějšího proložení mezi vypočtenými a zadanými souřadnicemi objektů bylo dosaženo.

Pro q souřadnic není možné nalézt sestavu bodů, jejichž párové vzdálenosti jsou monotónně vztaženy k původním podobnostem objektů. Kruskal [22] proto navrhl *míru důležitosti* $stress(q)$, podle které je geometrický souhlas modelu MDS s daty co nejlepší. Míra stress je definována vzorcem

$$stress(q) = \sqrt{\frac{\sum_{i<j}^{m}\sum_{j}^{m}\left(d_{ij}^{(q)} - d_{ij,vyp}^{(q)}\right)^2}{\sum_{i<j}^{m}\sum_{j}^{m}\left[d_{ij}^{(q)}\right]^2}}, \tag{97}$$

kde $d_{ij,vyp}^{(q)}$ jsou *vypočtená vzdálenostní pořadová čísla* objektů, která jsou vypočtena dle monotónní funkce podobností objektů. Nejde tedy o vzdálenosti objektů, vyjádřené ve vzdálenostních jednotkách, ale o pořadí. Cílem je nalézt zobrazení objektů tak, aby body v q souřadnicích byly objekty organizovány tak, aby $stress$ dosáhl svého minima. Kruskal [22] navrhl užívat míru důležitosti $stress(q)$ dle následujícího pravidla: je-li $stress(q)$ na hodnotě 20 % je těsnost proložení nedostatečná, je-li 10 % je dostatečná, je-li 5 % je dobrá, je-li 2.5 % je výtečná a konečně je-li 0 % je proložení daty úplně perfektní.

Druhou mírou shody objektů, zavedenou Takanem [5] je *přednostní kritérium* S_{stress}. Pro daný počet souřadnic q je tato míra dána vztahem

$$S_{stress} = \sqrt{\frac{\sum_{i<j}^{m}\sum_{j}\left(d_{ij}^{2} - d_{ij,vyp}^{2}\right)^2}{\sum_{i<j}^{m}\sum_{j} d_{ij}^{4}}}, \tag{98}$$

Hodnota přednostního kritéria S_{stress} je definována v intervalu 0 až 1. Hodnoty menší než 0.1 se týkají dobré prezentace objektů body nalezeného uspořádání.

Důležitým úkolem ve vícerozměrném škálování objektů je určit počet potřebných souřadnic v modelu MDS. Každá souřadnice představuje určitou latentní proměnnou. Cílem vícerozměrného škálování objektů je udržet počet souřadnic na co možná nejmenší hodnotě. Obvykle volíme k zobrazení dvoj- maximálně trojrozměrný prostor. Vychází-li vyšší počet souřadnic, není technika vícerozměrného škálování objektů k analýze dotyčných dat vhodná.

Počet souřadnic se volí na základě co nejmenší hodnoty kritéria $stress$. Když se objekty umístí v q souřadnicích, jejich vektory souřadnic rozměru $q \times 1$ mohou být považovány za vícerozměrná pozorování. K zobrazovacím účelům je vhodné prezentovat tento q-rozměrný rozptylový graf v souřadnicovém systému hlavních komponent.

Někteří autoři si pomáhají Cattelovým indexovým grafem úpatí relativní velikosti hodnot $stress$, která jsou vyčíslována pro rostoucí počet souřadnic. Postup a interpretace jsou pak stejné jako u grafu úpatí vlastních čísel v metodě hlavních komponent PCA nebo faktorové analýzy FA. Míru stress zapíšeme jako funkci počtu q souřadnic geometrického zobrazení objektů. Pro každé q může být získáno uspořádání vedoucí k minimální hodnotě $stress$. Jak se zvyšuje q, bude se $stress$ zmenšovat, až dosáhne pro $q = n - 1$ toleranci experimentálních chyb. Od $q = 1$ se vynáší indexový graf úpatí $stress = f(q)$. Na křivce tohoto grafu se hledá zlom a tomuto zlomu odpovídající hodnota nejlepšího počtu souřadnic q.

● **Klasická metrická metoda MDS**. Je dána matice vzdáleností objektů $\boldsymbol{D}$, která vystihuje mezi-objektové vzdálenosti objektů $\boldsymbol{X}$ v prostoru spíše nižšího rozměru. Jednotlivé kroky klasické metody vícerozměrného škálování objektů MDS jsou následující:

1. Z $\boldsymbol{D}$ se vypočte $\boldsymbol{A} = \{-0.5 d_{ij}^2\}$.

2. Z $\boldsymbol{A}$ se vypočte $\boldsymbol{B} = \{a_{ij} - a_{i.} - a_{.j} + a_{..}\}$, kde $a_{i.}$ je průměr všech a_{ij} přes j a $a_{..}$ je celkový průměr.

3. Nalezne se m největších vlastních čísel $\lambda_1 > \lambda_2 > \ldots > \lambda_m$ matice $\boldsymbol{B}$ a odpovídající vlastní vektory $\boldsymbol{L} = \boldsymbol{L}_{(1)}, \boldsymbol{L}_{(2)}, \ldots, \boldsymbol{L}_{(m)}$, které jsou normovány, takže $\boldsymbol{L}_{(i)}^T \boldsymbol{L}_{(i)} = \lambda_i$. Předpokládáme, že m je voleno tak, že vlastní hodnoty jsou relativně velké a kladné.

4. Souřadnicemi objektů jsou řádky matice $\boldsymbol{L}$.

Klasické řešení odpovídá metodě nejmenších čtverců tím, že nejlepší odhad $\boldsymbol{L}$ minimalizuje sumu čtverců vzdáleností mezi skutečnými prvky matice $\boldsymbol{D}$, tj. d_{ij},

a jejich dle MDS modelu vypočtenými hodnotami vzdáleností $d_{ij,vyp}$. Předpokládejme, že hodnoty skutečných vzdáleností objektů d_{ij} jsou zatíženy náhodnou chybou ε_{ij} dle vzorce $d_{ij} = \delta_{ij} + \varepsilon_{ij}$, kde ε_{ij} představuje kombinaci náhodných chyb z měření, distorze vzdáleností, když model MDS zcela neodpovídá konfiguraci navržených m vzdáleností. Model závislosti mezi vypočtenou vzdáleností dvou objektů $d_{ij,vyp}$ je definován vztahem $d_{ij} = \beta_0 + \beta_1\delta_{ij} + \varepsilon_{ij}$. Nalezením nejlepších odhadů b_0 pro β_0 a b_1 pro β_1 obdržíme odhad vypočtené vzdálenosti $d_{ij,vyp} = b_0 + b_1\delta_{ij}$. Optimalizační procedura vychází z účelové funkce

$$U = \sum_{i<j}^{n}\left(d_{ij} - d_{ij,vyp}\right)^2 \approx min. \tag{99}$$

Aby byla zajištěna úplná invariantnost vůči transformaci proměnných, dává se přednost užití modifikované účelové funkci U_{mod} dle vztahu

$$U_{mod} = \frac{\sum_{i<j}^{n}\left(d_{ij} - d_{ij,vyp}\right)^2}{\sum_{i<j}^{n} d_{ij}{}^2}. \tag{100}$$

Často se užívá její druhá odmocnina zvaná $stress = \sqrt{U_{mod}}$. Je výhodné hledat optimální počet souřadnic dimenzí, který vezmeme k vyčíslení vyčíslené vzdálenosti $d_{ij,vyp}$ pomocí minimální hodnoty veličiny $stress$. Platí zde pravidlo, že pro $stress$ menší než 0.05 je těsnost proložení ještě přijatelná a pro $stress$ menší než 0.01 je těsnost proložení výtečná.

● **Nemetrická MDS**. V dosavadním postupu se předpokládalo, že vzdálenosti objektů jsou vyčísleny metricky. Jsou však situace, kdy jedna hodnota nestačí k vystižení skutečnosti; například při porovnávání barev na stupnici může být jedna barva zářivější než druhá a tento fakt nikterak neovlivní polohu barvy na stupnici. Vypočtené vzdálenosti objektů jsou vyčíslovány monotónní regresí, kdy skutečné vzdálenosti objektů jsou uspořádány vzestupně do řady pořadových čísel objektů

$$d_{i1,j1} \leq d_{i2,j2} \leq \cdots \leq d_{iN,jN} \tag{101}$$

kde $N = n(n-1)/2$ a jsou odhadovány tak, aby splnily podmínku *slabé monotónnosti* (WM)

$$d_{i1,j1,vyp} \leq d_{i2,j2,vyp} \leq \cdots \leq d_{iN,jN,vyp} \tag{102}$$

nebo podmínku *silné monotónnosti* (SM)

$$d_{i1,j1,vyp} < d_{i2,j2,vyp} < \cdots < d_{iN,jN,vyp} \tag{103}$$

Prvním krokem k získání počátečních odhadů predikovaných vzdáleností objektů $d_{ij,vyp}$ bývá vždy metrické vyčíslení. Pak následuje nemetrický přístup monotónní regrese. *Cattelův indexový graf úpatí veličiny* $stress$ je užitečnou pomůckou i u nemetrické metody. Hledá se jednak zlom na tomto grafu a jednak se vyšetřuje, kdy veličina $stress$ nabude hodnot menších než 0.05, resp. 0.01. Takový počet souřadnic q, se pak jeví jako optimální. Obdobně, jako metrická metoda CMDS, ústí i nemetrická NNMDS ve vícerozměrnou škálovací mapu objektů, na které se sleduje roztřídění vyšetřovaných objektů.

Porovnejme nyní vícerozměrné škálování objektů s ostatními technikami, což je založeno na přístupu k definování struktury:

Faktorová analýza třídí proměnné, které definují dotyčné rozměry v původním souboru proměnných. Proměnné, které silně korelují, jsou pak zařazeny spolu.

Analýza shluků objektů třídí objekty podle jejich profilu v souboru proměnných, ve kterých jsou objekty v těsné blízkosti zařazovány spolu.

Vícerozměrné škálování objektů se liší od analýzy shluků objektů ve dvou klíčových bodech:

- Řešení může být získáno pro každý objekt.
- Neužívá se proměnných ale souřadnic v diagramu subjektivní mapy objektů.

Ve vícerozměrném škálování objektů je objekt jednotkou analýzy matice proximit. Představme si, například v průzkumu trhu, že máme podle 10 kritérií vzájemně posoudit šest automobilů. Každé auto budeme porovnávat s každým

a bodovat vzájemnou podovnost tak, že 1 značí auta si zcela podobná, stejná a 10 auta zcela nepodobná, různá. Bodová hodnota podobnosti páru bude tvořit jeden prvek matice proximit podobnosti 6 aut. MDS mapa zobrazí rozptylový diagram, který zobrazí rozdíly v autech při uvažování desatera jejich vlastností. Analýza poskytne vyhodnocení všech uvažovaných objektů, takže řešení se získá pro každý objekt, což není v analýze shluků nebo ve faktorové analýze. Pozornost není proto soustředěna na samotné objekty, ale na to, jak respondent chápe objekty, jak je subjektivně hodnotí. Struktura je definována v subjektivních souřadnicích při porovnání názorů několika respondentů. Jsou-li definovány subjektivní názory, může se provést relativní porovnání mezi objekty.

12. KORESPONDENČNÍ ANALÝZA (CA)

Analýza je grafická metoda k zobrazení skryté vnitřní závislosti, asociace v tabulce četností [17] až [19]. Soustředíme se na dvojrozměrnou tabulku četností zvanou *kontigenční tabulka*, která obsahuje n řádkových kategorií a m sloupcových kategorií. Diagram korespondenční analýzy – *subjektivní mapa* – obsahuje proto dvě skupiny bodů: skupinu n bodů odpovídajících řádkových kategorií a skupinu m bodů odpovídajících sloupcových kategorií.

12.1. Zaměření metody CA

Korespondenční analýza je kompoziční technika, protože subjektivní mapa je založena na asociaci mezi souborem objektů v řádcích a souborem popisných znaků ve sloupcích, zadaných člověkem. Polohy bodů pak přímo vyjadřují asociaci. I když je podobná faktorové analýze, přesahuje ji. Její přímou aplikací je zobrazování korespondence kategorií proměnných, znaků, které jsou měřeny v nominální stupnici. Tato korespondence je základem vytváření subjektivní mapy.

Řádkové body, které jsou těsně u sebe, indikují řádky, které mají podobné profily v celém řádku. *Sloupcové body*, které jsou blízko u sebe, indikují sloupce s podobnými profily směrem dolů přes všechny řádky. Konečně řádkové body, které jsou v těsné blízkosti ke sloupcovým bodům, představují kombinace, které se objeví častěji, než by se očekávalo u nezávislého modelu, ve kterém řádkové kategorie nejsou vztaženy ke sloupcovým.

Běžný výstup z korespondenční analýzy obsahuje „nejlepší“ dvojrozměrné zobrazení dat, ve kterém jsou souřadnice zobrazených bodů, a dále míru *inercie* vyjadřující množství informace zobrazené v každé dimenzi. Korespondenční analýza přináší řadu výhod:

První výhodou je subjektivní mapa, která zobrazuje tabulku vícerozměrných kategorických proměnných, jako jsou třeba znaky výrobků ve vztahu k výrobcům. Přístup umožňuje buď analyzovat existující odezvy, nebo shromáždit odezvy u nejméně omezeného typu měření nominální nebo kategorické

úrovně. Respondent potřebuje například u souboru objektů hodnotit *ano* a *ne* pro celou řadu znaků. Odpovědi jsou shromážděny v tabulce a analyzovány. Ostatní vícerozměrné techniky jako faktorová analýza vyžadují u každého objektu numerickou hodnotu každého znaku.

Druhou výhodou je zobrazení vedle vztahu mezi řádky a sloupci také vztahů mezi kategoriemi buď řádků, nebo sloupců. Jsou-li například sloupci znaky v těsné blízkosti, budou mít u všech objektů vesměs podobné profily. Tento způsob třídění znaků je velmi podobný faktorové analýze a metodě hlavních komponent.

Třetí výhodou je poskytnutí společného obrazu řádkových a sloupcových kategorií ve stejném počtu dimenzí. Modifikace některých programů dovoluje porovnání vnitřních bodů, ve kterém je relativní blízkost vztažena k vyšší asociaci mezi oddělenými body {1, 2}.

U těchto srovnání je umožněno současné vyšetření řádkových a sloupcových kategorií. Analýza umožní identifikovat třídy objektů, charakterizovaných znaků, které jsou v těsné blízkosti.

Vedle výhod přináší korespondenční analýza také nevýhody a omezení. Jde o popisnou techniku, která se nehodí ke statistickému testování hypotéz. Korespondenční analýza se uplatní v rámci exploratorní analýzy dat. Týká se snížení dimenze, i když nemá proceduru k určení vhodného počtu dimenzí.

12.2. Podstata metody CA

Korespondenční analýza vychází z kontingenčních tabulek s n řádky a m sloupci. Označme $\boldsymbol{U}$ matici $n \times m$ s prvky U_{ij}, které odpovídají prvkům kontingenční tabulky. Zaveďme řádkové N_{j+} a sloupcové N_{+i} součty a celkový součet N_T vztahy

$$N_{j+} = \sum_{i=1}^{m} U_{ij}, \tag{104}$$

$$N_{+i} = \sum_{j=1}^{n} U_{ij}, \tag{105}$$

$$N_T = \sum_{j=1}^{n} N_{j+} + \sum_{i=1}^{m} N_{+i} \, . \tag{106}$$

Dále lze definovat marginální relativní četnosti řádků $r_j = N_{j+}/N_T$ a relativní četnosti sloupců $c_i = N_{+i}/N_T$. Pokud se zavede matice četností $\boldsymbol{P}$ s prvky $P_{ij} = U_{ij}/N_T$, pak $\boldsymbol{P} = \boldsymbol{U}/N_T$ a χ^2- statistika pro test nulové hypotézy, že neexistuje asociace mezi sloupci a řádky, má tvar

$$\chi^2 = N_T \sum_{j=1}^{n} \sum_{i=1}^{m} \left(p_{ij} - r_i c_j\right)^2 / r_i c_j \, . \tag{107}$$

Veličina $t = \chi^2/N_T$ se označuje jako *Pearsonova střední kvadratická kontingence*. Homogenita v kontingenční tabulce je charakterizována malou hodnotou t a heterogenita velkou hodnotou t. Veličina t se dá vyjádřit ve tvaru

$$t = \sum_{i=1}^{n} r_i \sum_{j=1}^{m} [\left(\frac{p_{ij}}{r_i} - c_j\right)^2 / c_j], \tag{108}$$

což odpovídá vážené eukleidovské vzdálenosti mezi vektorem relativních četností a průměrným řádkovým profilem.

Řádkový profil matice $\boldsymbol{U}$ je definován jako vektor $\boldsymbol{p}_j = (p_{j1}, \ldots, p_{jm})$ pro $j = 1, \ldots, n$ s prvky $p_{ij} = U_{ij}/N_j$+ a průměrný řádkový profil $\boldsymbol{p}^T = (p_1, \ldots, p_m)^T$ má pak složky $p_j = N_{+i}/N_T$. Jako míru toho, jak se řádkové profily liší od průměrných řádkových profilů se používá χ^2-vzdálenost

$$d_{C_j}^2 = \left(\boldsymbol{p}_j - \overline{\boldsymbol{p}}\right)^T \boldsymbol{D}_p^{-1} \left(\boldsymbol{p}_j - \overline{\boldsymbol{p}}\right), \tag{109}$$

kde $\boldsymbol{D}_p^{-1}$ je diagonální matice s prvky $1/\overline{\boldsymbol{p}_i}$ na diagonále. Pearsonova statistika χ^2 je pak ve tvaru

$$\chi^2 = \sum_{j=1}^{n} N_{j+} d_{C_j}^2 \, . \tag{110}$$

Označme dále $\boldsymbol{r} = \boldsymbol{PI}$ a $\boldsymbol{c} = \boldsymbol{P^T I}$, kde $\boldsymbol{I}$ jsou vektory ze samých jedniček, a zaveďme diagonální matice $\boldsymbol{D_r}$ obsahující na diagonále prvky vektoru $\boldsymbol{r}$ a $\boldsymbol{D_c}$ obsahující na diagonále prvky vektoru $\boldsymbol{c}$. Matice $\boldsymbol{J}$

$$\boldsymbol{J} = \boldsymbol{D_r^{-\frac{1}{2}}}(\boldsymbol{P} - \boldsymbol{rc}^T)\boldsymbol{D}_C^{-1/2} \tag{111}$$

má prvky úměrné standardizovaným reziduím kontingenční tabulky $\boldsymbol{U}$, tj. odmocniny členů, ze kterých se tvoří statistika χ^2. Pomocí techniky SVD se obecně obdélníková matice $\boldsymbol{E}$ rozkládá na tři matice $\boldsymbol{E} = \boldsymbol{USV^T}$, kde $\boldsymbol{S}$ je matice singulárních čísel a $\boldsymbol{U}$ a $\boldsymbol{V}$ jsou levé a pravé vlastní vektory. Složky řádkových profilů kontingenční tabulky $\boldsymbol{f}_i$ jsou pak řádky matice

$$\boldsymbol{F} = \boldsymbol{D_r^{-\frac{1}{2}}US}. \tag{112}$$

Složky sloupcových profilů kontingenční tabulky $\boldsymbol{g}_i$ jsou pak řádky matice

$$\boldsymbol{G} = \boldsymbol{D_c^{-\frac{1}{2}}VS}. \tag{113}$$

Dvojice řádkových a sloupcových složek $\boldsymbol{f}_i$, $\boldsymbol{g}_i$ jsou prvky ortogonálního rozkladu reziduí (odchylek od modelu nezávislosti sloupců a řádků, kde závislost se chápe ve smyslu variace), které jsou uspořádány sestupně podle důležitosti. Tento rozklad se označuje jako korespondenční analýza. Je zřejmé, že $\boldsymbol{f}_i$ a $\boldsymbol{g}_i$ jsou nekorelované a mají nulové střední hodnoty. Jsou však vzájemně spjaty vazbami

$$\boldsymbol{G} = \boldsymbol{D_c^{-\frac{1}{2}}P^TFS^{-1}}. \tag{114}$$

$$\boldsymbol{F} = \boldsymbol{D_r^{-\frac{1}{2}}PGS^{-1}}. \tag{115}$$

Singulární vlastní čísla, tj. diagonální prvky matice $\boldsymbol{S}$, se označují jako *hlavní setrvačnosti.* Pro zobrazení obou profilů se využívá dvojný graf (biplot) stejně jako u hlavních komponent. Hlavním cílem je identifikovat zdroje heterogenity v kontingenčních tabulkách. Korespondenční analýza umožňuje tedy rozklad statistiky χ^2 tak, aby bylo možné posoudit struktury v matici $\boldsymbol{N}$.

SEZNAM POUŽITÉ LITERATURY

[1] K. Marhold, J. Suda: *Statistické zpracování mnohorozměrných dat v taxonometrii*. Nakladatelství Karolinum, Praha 2002.

[2] A. Raveh: *Amer. Statist*. **39,** 39 (1985).

[3] J. Muruzahal, A. Muñoz: *J. Comput. and Graph. Statist*. **6,** 355 (1977).

[4] G. S. Mudholkar, M. S. Trivedi, T. C. Lin: *Technometrics*. **24**, 139 (1982).

[5] J. E. Jackson: *A User's Guide to Principal Components*. Wiley, New York 1991.

[6] D. Peña, J. Rodrigues: *J. Multivariate Anal*. **85**, 361 (2003).

[7] G. M. Arnold, A. J. Collins: *Appl. Statist*. **42**, 381 (1993).

[8] J. L. Hintze: *NCSS2000, Statistical System for Windows*. Kaysville, Utah USA, 1999.

[9] H. M. Harman: *Modern Factor Analysis*, Univ. of Chicago Press, 1976.

[10] R. A. Darton: *The Statistician* 29, 167 (1980).

[11] J. F. Hair, R. E. Anderson, R. L. Tatham, W. C. Black: *Multivariate Data Analysis*. Prentice Hall, Upper Saddle River, New Jersey 07458, USA, 1998.

[12] *SPSS for Windows*, Base System User's Guide, Chigago, USA, 1993.

[13] T. Hostie, R. Tibshirani, J. Friedman: *The elements of statistical learning*. Springer, New York 2001.

[14] C. J. Peng, H. T. So: *Understanding Statistics* 1, 31 (2002).

[15] S. K. Sapra: *Amer. Statist*. **45**, 365 (1991).

[16] K. Marhold, J. Suda: *Statistické zpracování dat v taxonomii*. Nakladatelství Karolinum, Praha 2002.

[17] D. J. Carroll, P. E. Green, C. M. Schaffer: *J. of Marketing Research* **23,** 271−280 (1986).

[18] M. J. Greenacre: *Theory an Applications of Correspondence Analysis*. Academic Press, London 1984.

[19] L. Lebart, A. Morineau, K. M. Warwick: *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. Wiley, New York, 1984.

[20] M. Meloun, J. Militký: *Kompendium statistického zpracování dat,* Academia Praha 2002.

[21] M. Meloun, J. Militký: *Statistická analýza experimentálních dat,* Academia Praha 2004.

[22] J. B. Kruskal, F. J. Carmone: How to use MDSCAL, version 5-M and other useful information. Unpublished results, Bell Laboratories, 1967.

Centrum pro rozvoj výzkumu pokročilých řídicích a senzorických technologií
CZ.1.07/2.3.00/09.0031

Ústav automatizace a měřicí techniky
VUT v Brně
Kolejní 2906/4
612 00 Brno
Česká Republika

http://www.crr.vutbr.cz

info@crr.vutbr.cz