MDA-metody klasifikace a shlukování

Prof. RNDr. Milan Meloun, DrSc.
Katedra analytické chemie, Univerzita Pardubice,
532 10 Pardubice, Česká republika,
http://meloun.upce.cz
Email: milan.meloun@upce.cz

Klasifikační metody struktury ve znacích a objektech

Hledáním struktury a vzájemných vazeb v objektech se zabývají klasifikační metody vícerozměrné statistické analýzy MDA. *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. Významnou roli při hledání struktury a vazeb mezi objekty na základě jejich podobnosti tvoří také *vícerozměrné škálování MDS.*

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*.

To tedy znamená

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 se závisle proměnné 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).*

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).

Obecný postup vícerozměrné analýzy dat MDA

K dosažení úspěšné analýzy vícerozměrných dat a odhalení všech skrytých vnitřních vazeb včetně klasifikace objektů do tříd je třeba vycházet z následujícího typového postupu:

1. ***Standardizace:*** vícerozměrné analýze obvykle předchází standardizace znaků ve sloupcích.

2. ***Odhady parametrů polohy, rozptýlení, tvaru a intenzita vztahu mezi znaky:*** vyčíslení výběrové střední hodnoty každého znaku, proměnné, odhad kovarianční matice ***S*** a její normované podoby - korelační matice ***R.*** Matice ***R*** obsahuje Pearsonovy párové korelační koeficienty ρ_{ij}, které se je třeba diskutovat. Užitečný je zde především diagram korelační matice.

3. ***Exploratorní analýza dat EDA:*** (a) posouzení podobnosti objektů pomocí vizuálních rozptylových diagramů typu casement plot, draftsman plot, dále symbolových a profilových grafů (hvězdičky, sluníčka, obličeje, křivky, stromy), (b) nalezení vybočujících objektů nebo vybočujících znaků či proměnných mnohdy nevhodných pro další analýzu.

4. ***Určení vhodného počtu latentních proměnných:*** matice ***S*** nebo ***R*** se rozloží na vlastní čísla a vlastní vektory. Z Cattelova indexového grafu úpatí vlastních čísel (Scree plot) se určí vhodný počet latentních proměnných (při zobrazení v rovině se obvykle dává přednost prvním dvěma latentním proměnným), které ještě dostatečně popisují proměnlivost v datech.

5. ***Určení struktury v proměnných (PCA a FA):*** hledání struktury a vzájemných vazeb (korelace) proměnných se provede v grafu komponentních vah (v software zvané Plot of components weights, loadings). Hledání struktury v objektech a třídění objektů do shluků se provede v rozptylovém diagramu komponentního skóre (v software zvané Plot of principal components).

6. ***Určení struktury a vzájemných vazeb v objektech:*** klasifikační postupy zařadí v diskriminační analýze analyzovaný objekt do jednoho z již existujících a předem zadaných shluků. Neutříděnou skupinu objektů lze uspořádat do shluků a výsledek třídění zobrazit dendrogramem v analýze shluků. V hierarchickém postupu je třeba k vytvoření shluků vybrat vzdálenost mezi objekty (Eukleidovskou, Manhattanovskou, Mahalanobisovu) a jednu z možných metod: průměrovou, centroidní, nejbližšího souseda, nejvzdálenějšího souseda, mediánovou, Wardovu.

7. ***Soulad nalezené struktury objektů a vazeb v dendrogramu a PCA (či FA) grafech:*** je třeba vyšetřit a komentovat nalezenou strukturu a vazby *jednotlivých znaků,* nalezenou jednak v PCA (či FA) a jednak v dendrogramu podobnosti proměnných analýzou vzniklých shluků. Dále je třeba komentovat také strukturu a vazby *klasifikovaných objektů,* nalezenou v PCA a v dendrogramu podobnosti objektů.

8. ***Soulad nalezené struktury objektů a vazeb diskriminační analýzou:*** je třeba vyšetřit a komentovat nalezenou strukturu a vazby *jednotlivých proměnných* do tříd objektů také metodou diskriminační analýzy a následného využití Fisherovy lineární diskriminační funkce ke klasifikaci neznámých vzorků.

Z 11 nejpoužívanějších metod vícerozměrné statistické analýzy MDA bylo vybráno pět následujících počítačových metod v různým software, které nejlépe odpovídají požadovanému zadání analýzy a odpovídajícím vícerozměrným datům.

1. Exploratorní analýza dat EDA

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.

PŘÍKLAD 2.2 ***Průzkumová analýza zdrojové matice dat demografického souboru Lidé***

Vyšetřete grafickými diagnostikami průzkumové analýzy vícerozměrných dat, které ze 12 znaků demografického souboru dat *Lidé jsou nejvýhodnější k charakterizaci osob a které znaky mají největší míru rozptýlení. Matice obsahuje data pro n = 32 osob a m = 12 znaků, kde 16 osob bylo vybráno ze Skandinávie (kód A) a 16 osob ze Středomoří (kód B), 16 osob jsou muži (kód M) a 16 osob jsou ženy (kód F).*

Data: *Znaky obsahují u každé osoby výšku [cm], hmotnost [kg], délku vlasů [krátká: −1, dlouhá: +1], velikost boty [evropský standard], věk [roky], příjem [Euro], spotřeba piva [litry na rok], spotřeba vína [litry na rok], pohlaví [muž: −1, žena: +1], schopnost plavat [naměřený čas na uplavání 500 m], původ [A: −1 Skandinávie, B: +1 Středomoří], inteligenční kvocient IQ [evropský standardizovaný test IQ]. Mezi znaky jsou tři dichotomické, binární proměnné, a to pohlaví, délka vlasů a původ a ostatních 9 znaků nabývá kvantitativních hodnot.*

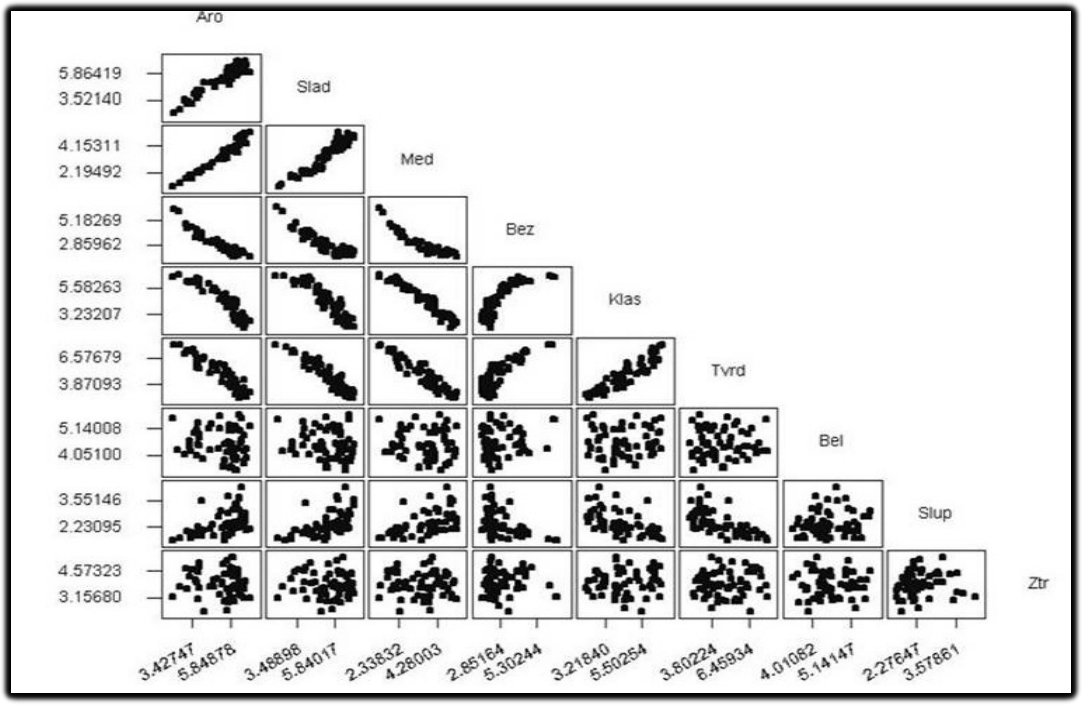
Znaky x1 až x 12

Objekty (n=1 až 32)

Osoba	Výška	Hmotnost	Vlasy	Boty	Věk	Příjem	Pivo	Víno	Sex
	Plavání	Původ	IQ						
MA	198	92	−1	48	48	45000	420	115	−1
	98	−1	100						
MA	184	84	−1	44	33	33000	350	102	−1
	92	−1	130						
MA	183	83	−1	44	37	34000	320	98	−1
	91	−1	127						
FA	166	47	−1	36	32	28000	270	78	1
	75	−1	112						

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 užitečnou mírou vnitřního lineárního vztahu mezi dvěma proměnnými x_1 a x_2. *Personův korelační koeficient r* se odhaduje podle

$$r = \frac{\sum_{i=1}^{n}(x_{1i}-\overline{x}_1)(x_{2i}-\overline{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{1i}-\overline{x}_1)^2\sum_{i=1}^{n}(x_{2i}-\overline{x}_2)^2}}$$

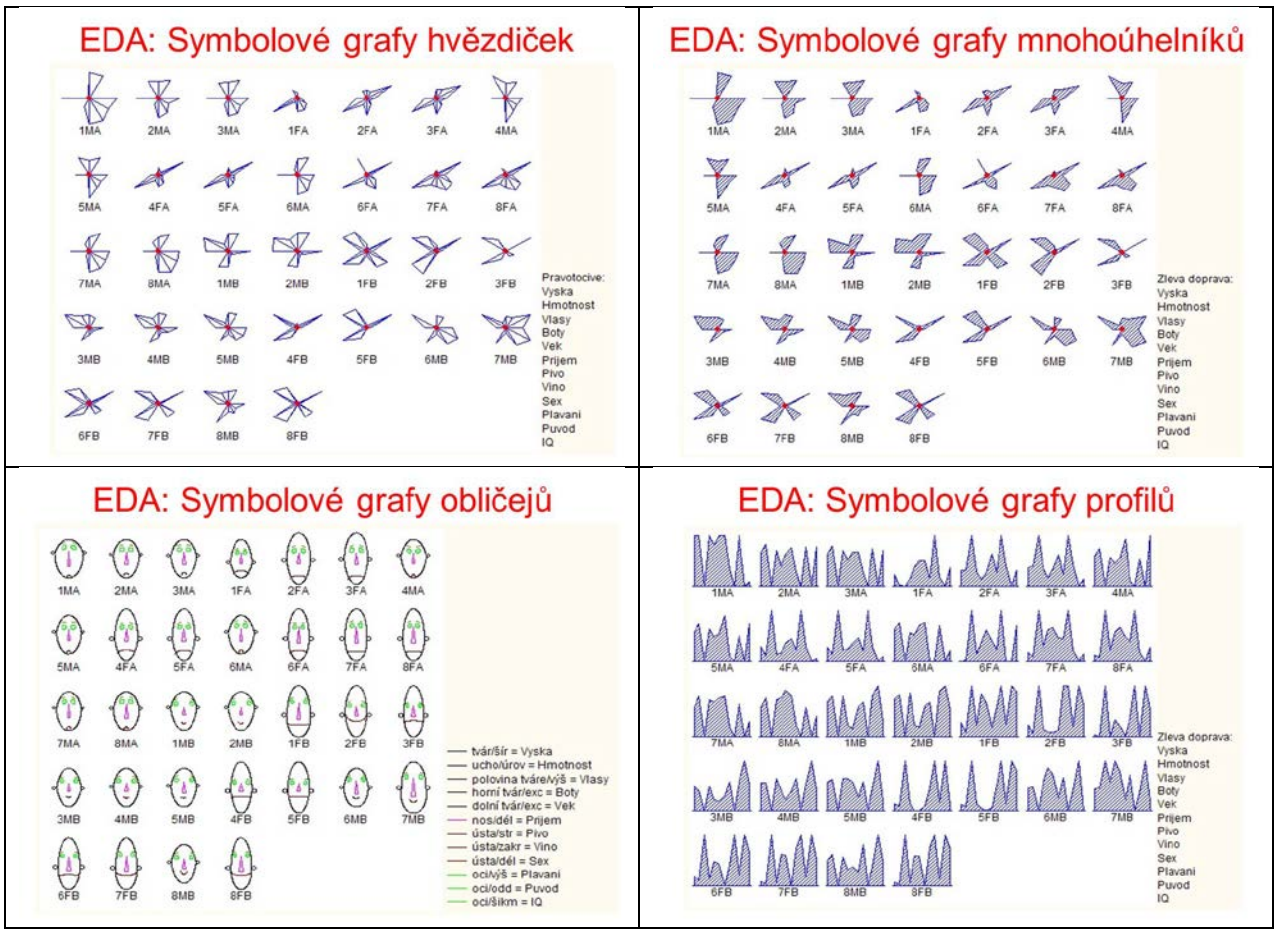


Koeficient determinace D = r^2 popisuje podíl celkového rozptylu, který lze objasnit lineárním vztahem. Koeficient determinace 100*D* [%] vyjádřený v procentech často zvaný *regresní rabat* je běžnou mírou k vystižení stupně korelace a vystihuje procento bodů, které odpovídají lineárnímu regresnímu modelu.

Úvodní exploratorní (průzkumová) analýza vícerozměrných dat je stejně jako u jednorozměrných dat založena na použití grafických diagnostik k zobrazení objektů vícerozměrného prostoru. Pro případ, kdy jsou jednotlivé znaky ve sloupcích matice $\boldsymbol{X}$ málo korelované postačují *rozptylové diagramy* pro jednotlivé kombinace složek vektoru $\boldsymbol{x}$ a pro nekorelované pak sloupce matice $\boldsymbol{X}$.

Rychlé posouzení podobnosti mezi jednotlivými objekty čili řádky datové matice usnadňují především ***symbolové grafy***. Jednotlivé znaky jsou v nich "kódovány" s ohledem na jejich konkrétní hodnoty do určitých geometrických tvarů, *symbolů*. Každému objektu $\boldsymbol{x}_i$ tak odpovídá jistý obrazec zvaný *symbol*. Vlastnosti dat se posuzují s ohledem na vizuální rozdíly mezi symboly. Tím lze do jednoho grafu zahrnout *více proměnných* x_j , $j = 1, ..., m$. Prvním krokem před vlastním zobrazením do symbolů je obvykle standardizace. Mezi základní typy zobrazovaných symbolů patří ***profily, polygony, tváře, křivky*** a ***stromy***.

Polygony jsou vlastně profily v polárních souřadnicích, kdy každý znak objektu $\boldsymbol{x}_i^{\mathrm{T}}$, $i = 1, ..., n$, odpovídá délce paprsku vycházejícího ze společného středu. Paprsky dělí kružnici ekvidistantně, znaky jsou standardizovány do intervalu [0, 1]. Mezi polygony patří ***graf slunečních paprsků*** a ***hvězdicový graf,*** který vypadá na první pohled jako předchozí graf sluníček. Sestává z paprsků, reprezentujících relativní hodnoty znaků u jednotlivých objektů, které se pro každý objekt spojují v jednom centrálním bodě. Stejně směřující paprsky u různých objektů se liší svojí délkou. *Nejkratší paprsek* indikuje, že u objektu nabývá příslušný znak nejmenší hodnoty z celého výběru.



Podobně *nejdelší paprsek* informuje o nejvyšší hodnotě příslušného znaku. Délky ostatních paprsků se pohybují podle relativní velikosti hodnot znaku u příslušného objektu mezi těmito dvěma krajními mezemi. Podobně vypovídají i ***mnohoúhelníky, tváře*** a ***křivky***.

2. Analýza hlavních komponent PCA

Podstata metody PCA

Metoda hlavních komponent (PCA) je jednou z nejstarších a nejvíce používaných metod vícerozměrné analýzy. Nyla zavedena Pearsonem v roce 1901 a nezávisle Hotellingem v roce 1933. Cílem je zjednodušení popisu skupiny vzájemně lineárně závislých neboli 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 orthogonální 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.

Standardním využitím PCA je snížení dimenze úlohy čili redukce počtu znaků bez velké ztráty informace, a to užitím pouze prvních několika hlavních komponent. Snížení dimenze úlohy se netýká počtu původních znaků. Je výhodné především pro možnost zobrazení vícerozměrných dat. Předpokládá se, že nevyužité hlavní komponenty obsahují malé množství informace, protože jejich rozptyl je příliš malý. Tato metoda je atraktivní především z důvodu, že hlavní komponenty jsou nekorelované. Namísto vyšetřování velkého počtu původních znaků s komplexními vnitřními vazbami analyzuje uživatel pouze malý počet nekorelovaných hlavních komponent.

Zaměření metody PCA charakterizují její složky. První dvě respektive první tři hlavní komponenty se využívají především jako techniky zobrazení vícerozměrných dat v projekci do roviny nebo do prostoru. Výhodou je, že tato projekce zachovává vzdálenosti a úhly mezi jednotlivými objekty. V řadě případů jsou hlavní komponenty pouze jednou z fází komplexnější analýzy.

Zdrojová matice dat $\boldsymbol{X}$ ($n \times m$) obsahuje n objektů a m znaků. *Objekty* jsou pozorování, vzorky, experimenty, měření, atd., zatímco *znaky* či proměnné jsou druhy signálu měření, měřená veličina, ale také 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 závislý na charakteru každé úlohy. Velikou výhodou metody PCA je to, že lze použít 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 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.

Cílem PCA je transformace původních znaků x_i, $j = 1, ..., m$, 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. 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.*

Grafické diagnostiky metody hlavních komponent

Graficky lze výsledek analýzy hlavních komponent zobrazit v několika diagnostických grafech:

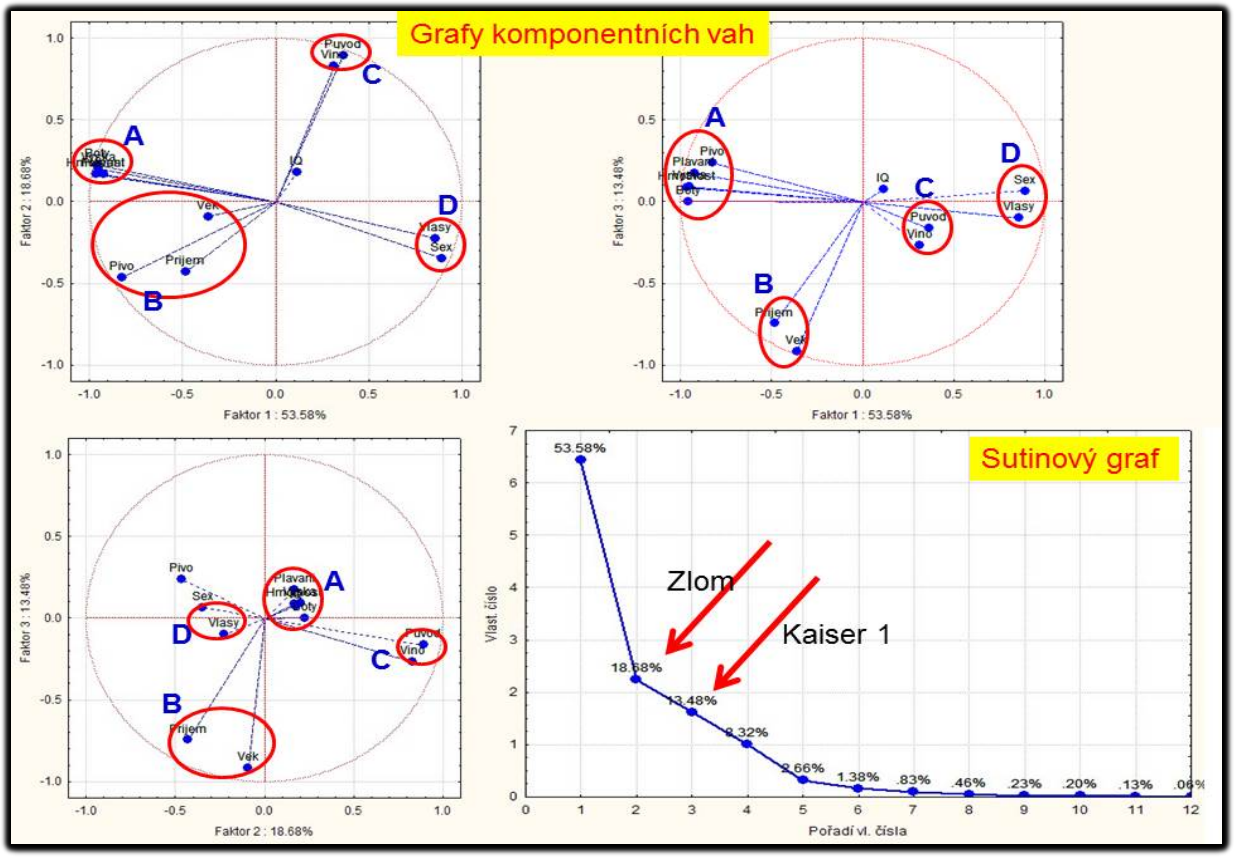
a) Cattelův indexový graf úpatí vlastních čísel (v software zvaný Scree Plot a česky v software sutinový graf). Je sloupcovým diagramem vlastních čísel nebo reziduálního rozptylu pro stoupající hodnoty indexu, pořadového čísla. Řada autorů ho s oblibou využívá k určení počtu „užitečných“ hlavních komponent. Cattel vysvětluje *scree* jako úpatí mořského útesu čili zlomové místo mezi kolmou stěnou a vodorovným dnem. Vybrané „užitečné“ hlavní komponenty (nebo také faktory) pak tvoří kolmou stěnu útesu a „neužitečné“ hlavní komponenty (nebo faktory) představují vodorovné mořské dno. Užitečné komponenty jsou tak odděleny zřetelným zlomovým místem, úpatím a souřadnice x tohoto zlomu je hledaná hodnota indexu. Jiným, hrubším kritériem je Kaiserovo pravidlo, podle kterého využíváme ty hlavní komponenty, jejichž vlastní číslo je větší než jedna. Pravidlo vychází z myšlenky, že není třeba uvažovat komponenty, jejichž rozptyl je menší než jednotkový rozptyl každého normovaného znaku. Graf úpatí se však jeví objektivnějším a praktičtějším.

b) Graf komponentních vah, zátěží (v software zvaný Plot Components Weights). Zobrazí komponentní váhy například pro první dvě hlavní komponenty. V tomto grafu se porovnávají vzdálenosti mezi znaky. Krátká vzdálenost mezi dvěma znaky znamená silnou korelaci. Lze nalézt i shluk podobných znaků, jež spolu korelují. Tento graf můžeme považovat za most či korelaci mezi znaky a hlavními komponentami, protože ukazuje, jakou měrou přispívají jednotlivé znaky do hlavních komponent. Někdy se podaří hlavní komponenty y_1, y_2, ... pojmenovat, vysvětlit a přidělit jim fyzikální, chemický nebo biologický význam. Pak lze názorně vysvětlit, jak jednotlivé znaky x_j, $j = 1,..., m$, přispívají do první hlavní komponenty y_1 nebo do druhé hlavní komponenty y_2. Některé znaky x_j přispívají kladnou vahou, jiné zápornou. Bývá zajímavé sledovat kovarianci znaků x_j v prostorovém 3D grafu komponentních vah y_1, y_2 a y_3. Jsou-li znaky x_j, $j = 1,..., m$, blízko sebe v prostorovém shluku, jde o silnou pozitivní kovarianci. Výklad grafu komponentních vah lze obecně shrnout do následujících bodů:

1. *Důležitost znaků* x_j, $j = 1,..., m$: znaky x_j s vysokou mírou proměnlivosti v datech objektů 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 proměnlivost znaků: jestliže například y_1 objasňuje 70 % proměnlivosti a y_2 jenom 5 % (přečteno z indexového grafu úpatí vlastních čísel), jsou znaky x_j, $j = 1,..., m$, s vysokou váhou v y_1 tím pádem mnohem důležitější než znaky x_j s vysokou vahou v y_2. Znaky s úhlem 0 stupňů mezi dvěma průvodiči jsou zcela pozitivně korelované, znaky s úhlem 90 stupňů jsou zcela nekorelované, zatímco znaky s úhlem 180 stupňů jsou negativně korelované.

2. *Korelace a kovariance:* znaky x_j, $j = 1,..., m$, 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, znaky x_j daleko od sebe, anebo s velikým úhlem mezi průvodiči znaků, jsou negativně korelovány.



c) **Rozptylový diagram komponentního skóre** (v software zvaný Scatterplot). Zobrazuje *komponentní skóre* čili hodnoty obyčejně prvních dvou hlavních komponent u všech objektů. Lze snadno nalézt shluk vzájemně podobných objektů a dále také objekty odlehlé a silně odlišné od ostatních. Diagram komponentního skóre však může být prostorový ve třech hlavních komponentách a v rovinném grafu se pak sleduje pouze jeho průmět do roviny. Tento diagram se užívá k identifikaci odlehlých objektů, identifikaci trendů, tříd, shluků objektů, k objasnění podobnosti objektů atd. Je často nemožné analyzovat všechny diagramy, protože jich je velmi mnoho: 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 vybíráme diagramy y_1 vs. (y_2), y_1 vs. (y_3), y_1 vs. (y_4) atd. Držíme se první hlavní komponenty y_1, protože objasňuje největší míru proměnlivosti v datech. Interpretace rozptylového diagramu komponentního skóre lze shrnout do těchto bodů:

1. *Umístění objektů*. Objekty daleko od počátku jsou extrémy.

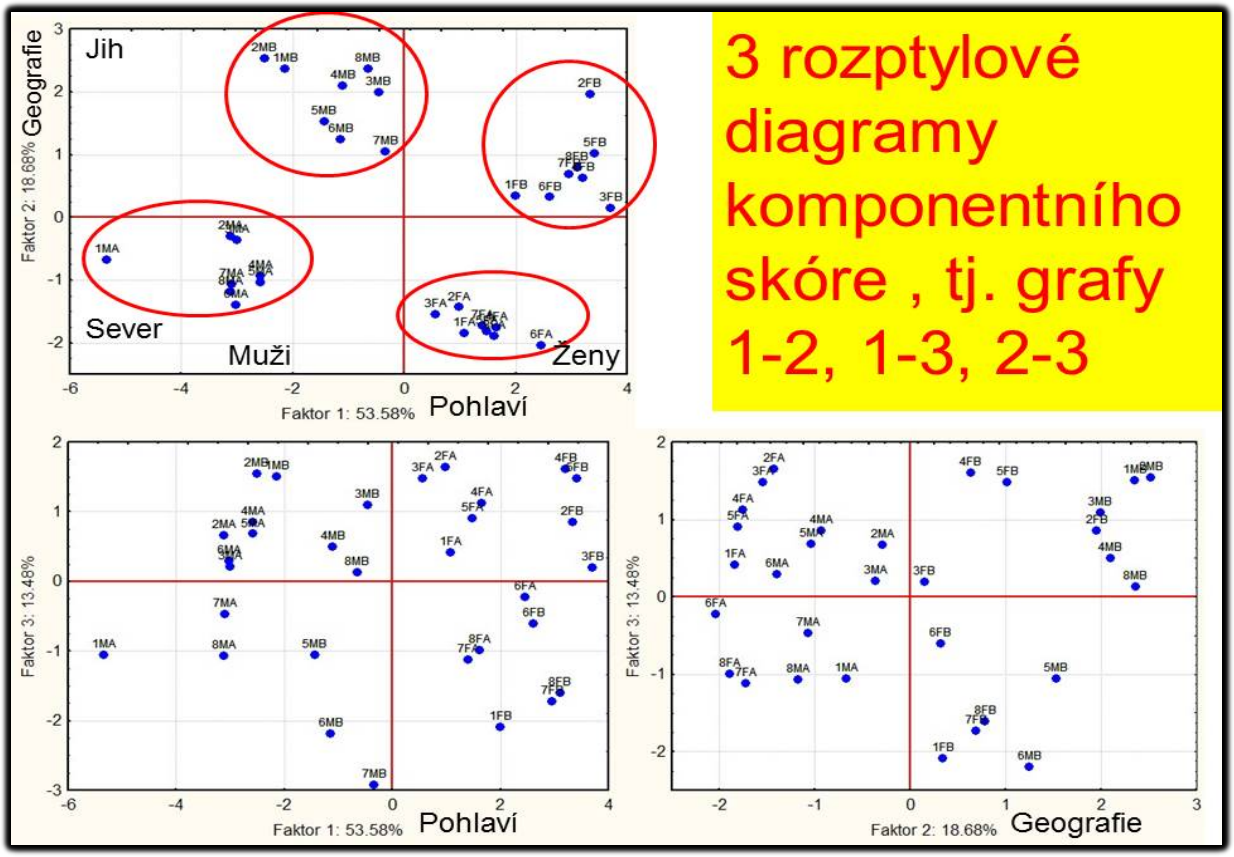
2. *Podobnost objektů*. Objekty blízko sebe si jsou podobné, objekty daleko od sebe jsou si nepodobné.

3. *Objekty 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.

4. *Osamělé objekty*. Izolované objekty mohou být odlehlé objekty, které jsou silně nepodobné ostatním objektům.

5. *Odlehlé objekty*. 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 roztřídí po celé ploše diagramu a teprve pak vypovídají o existujících shlucích.



6. *Pojmenování objektů.* Výstižná jména objektů slouží k hledání hlubších souvislostí mezi objekty a mezi pojmenovanými hlavními komponentami. Snadno obkroužíme shluky podobných objektů nebo nakreslením spojky mezi objekty vystihneme jejich fyzikální či chemickou nebo biologickou podobnost.

7. *Vysvětlení místa objektu.* Umístění objektu na ploše v diagramu může být porovnáváno s komponentními váhami znaků ve dvojném grafu a pomocí znaků pak i vysvětleno.

3. Faktorová analýza FA

Podstata metody FA

Faktorová analýza je vícerozměrná technika k vyšetření vnitřních souvislostí a vztahů a odhalení základní struktury zdrojové matice dat. Týká se analýzy struktury vnitřních vztahů mezi velkým počtem původních znaků pomocí 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 zdrojové matice 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 lze 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.

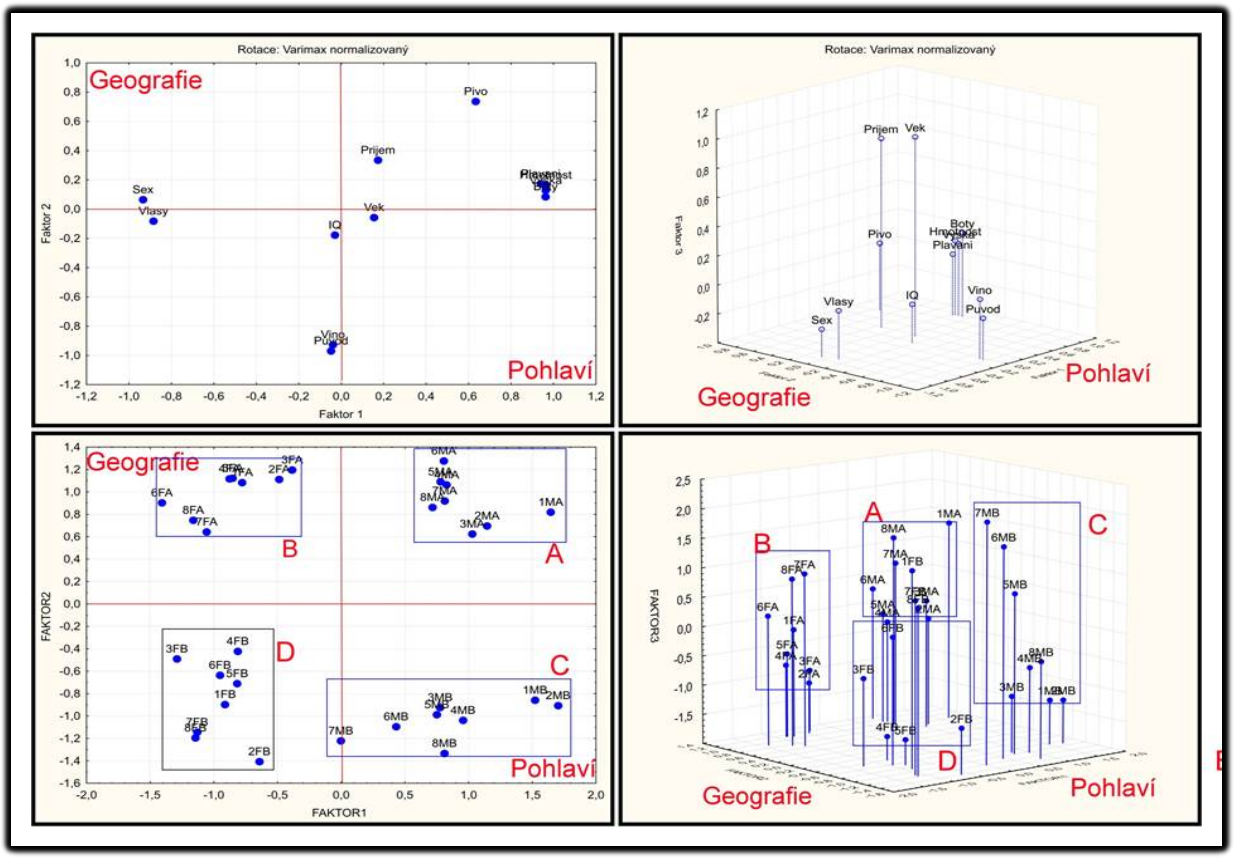
Existuje řada různých faktorových modelů objasňujících u stejných dat korelační strukturu původních znaků. Ne každý model je však stejně informativní. Pro dobrou interpretaci faktorů F se většinou požaduje co nejjednodušší struktura. Vůbec nejjednodušší je taková struktura, kde má každý znak nenulovou zátěž pouze pro jeden faktor. 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, $i = 1, ..., m$, 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 $x_1, ..., x_m$ a p společných faktorů $F_1, ..., F_p$. *Varimaxová rotace* maximalizuje rozptyl zátěží v každém sloupci faktorové matice. 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. Otočené faktory jsou v ideálním případě zvoleny tak, že některé zátěže dosahují vysokých hodnot, blízkých ±1, a zbývající pak velmi nízkých, téměř nulových. Je vhodné, aby každý znak dosahoval vysoké zátěže pouze u jediného faktoru, čili byl *faktorově čistý.*

Grafické pomůcky 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.

a) 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.

b) 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. Uvažujeme-li k-tý objekt, je skóre pro j-tý faktor F_{jk} vyčísleno vztahem, 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 udává celkový počet všech původních znaků.



Diagnostikování metodou FA

Uvedeme stručný postup analýzy vnitřní struktury dat, který obsahuje čtyři základní kroky:

1. Vyčíslí se korelační matice všech znaků a vyznačí se znaky, které nejsou v korelaci s ostatními. Vyčíslí se vhodnost faktorového modelu. Uživatel se musí rozhodnout, co bude dělat s přípustnou neúplnou vstupní maticí dat.
2. Stanoví se potřebné faktory. Uživatel se musí ujistit, jak dobře faktorový model prokládá data.
3. Rotací faktorů se docílí jejich lepší interpretace.
4. Vyčíslí se faktorové skóre pro každý objekt a naleznou se shluky podobných objektů.

Zvyšováním počtu znaků se zvyšuje možnost, že znaky nebudou nekorelované. Uživatel by měl vědět, jak jsou znaky vnitřně svázány, aby tak lépe vysvětlil svá experimentální data. Je-li počet znaků příliš veliký, nebo je-li třeba vystihnout data menším počtem znaků, faktorová analýza poslouží vytvořením nových znaků při zachování původní povahy a charakteru dat. Všechny znaky jsou uvažovány souběžně, každý je vztažen ke všem ostatním. Jde o lineární kombinaci všech původních znaků.

Na každý faktor se lze dívat jako na závisle proměnnou, která je funkcí celého souboru ostatních původních znaků.

a) Nejprve identifikujeme s t r u k t u r u zestručněním dat, a to vyšetřením korelací mezi znaky (ve sloupcích) nebo korelací mezi objekty (v řádcích). Je-li cílem zestručnění popisných charakteristik, aplikuje se faktorová analýza na korelační matici znaků a analýza se nazývá *R-faktorová analýza.* Faktorová analýza se může aplikovat také na korelační matici objektů, a pak se nazývá *Q-faktorová analýza.* Zhušťuje velký počet objektů do různě velkých shluků. Výhodnější bývá užít analýzu shluků.

b) Pak následuje *redukce dat.* Faktorová analýza může z velkého počtu původních znaků, vedle identifikace reprezentativní skupiny souboru některých znaků, vytvořit naprosto novou skupinu znaků o daleko menším počtu a tou nahradit původní znaky. Cílem je zachovat povahu a charakter původních znaků při redukci jejich počtu za účelem zjednodušení vícerozměrné analýzy dat. Příspěvky každého znaku do faktoru, zvaného *zátěž*, je totiž vše, co je potřebné k této analýze.

c) Užití faktorové analýzy s ostatními technikami vícerozměrné analýzy dat poskytuje přímý pohled do vnitřních vztahů mezi znaky a objekty. Zestručnění popisu dat poskytuje uživateli jasné vysvětlení, jak velký význam má ten který znak, s kterými vlastnostmi je spojen a kolik znaků má vliv na vlastní analýzu.

d) Redukce dat a jejich sumarizace může být provedena buď s původní skupinou znaků, nebo se znaky vytvořenými novou analýzou. Když bude do analýzy zahrnut nekriticky velký počet všech původních znaků, je malá šance na dobrý výsledek.

4. Analýza shluků CLU

Podstata metody 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. 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.

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:

1) *popis systematiky*, jenž je tradičním využitím shlukové analýzy pro průzkumové cíle a taxonomii, což je empirická klasifikace objektů,

2) *zjednodušení dat*, kdy analýza shluků poskytuje při hledání taxonomie zjednodušený pohled na objekty,

3) *identifikaci vztahu*, kdy po nalezení shluků objektů, a tím i struktury mezi objekty, je snadnější odhalit vztahy mezi objekty.

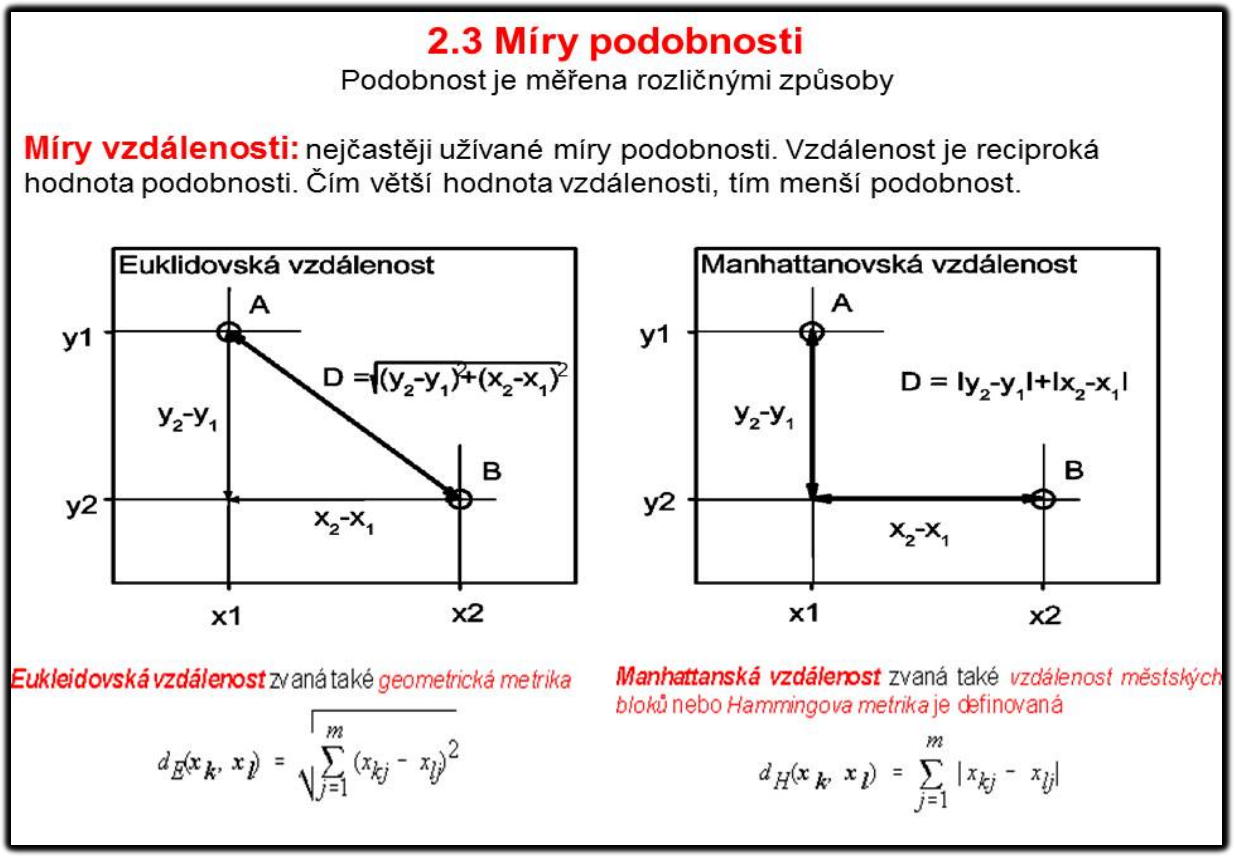
Cíle shlukové analýzy nelze oddělit od hledání a volby vhodných znaků k charakterizování shlukovaných objektů. Nalezené shluky vystihují strukturu dat pouze s ohledem na vybrané znaky. Volba znaků musí být provedena na základě teoretických, pojmových a praktických hledisek. Vlastní shluková analýza neobsahuje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

Při odhalování struktury objektů je shluková analýza velmi citlivá na přítomnost *nevýznamných znaků.* Je citlivá také na přítomnost *odlehlých objektů,* které se silně odlišují ode všech ostatních objektů. Odlehlé objekty mohou představovat buď skutečně odchýlené, patologické objekty, které nepředstavují 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 těžkopádnými při velkém počtu objektů nebo znaků. Vybočující objekty je obvykle třeba z dat odstranit, i když si musíme být vědomi, že jejich odstraněním se mnohdy zbortí aktuální struktura. Vypouštění objektů by proto mělo být velmi uvážlivé.

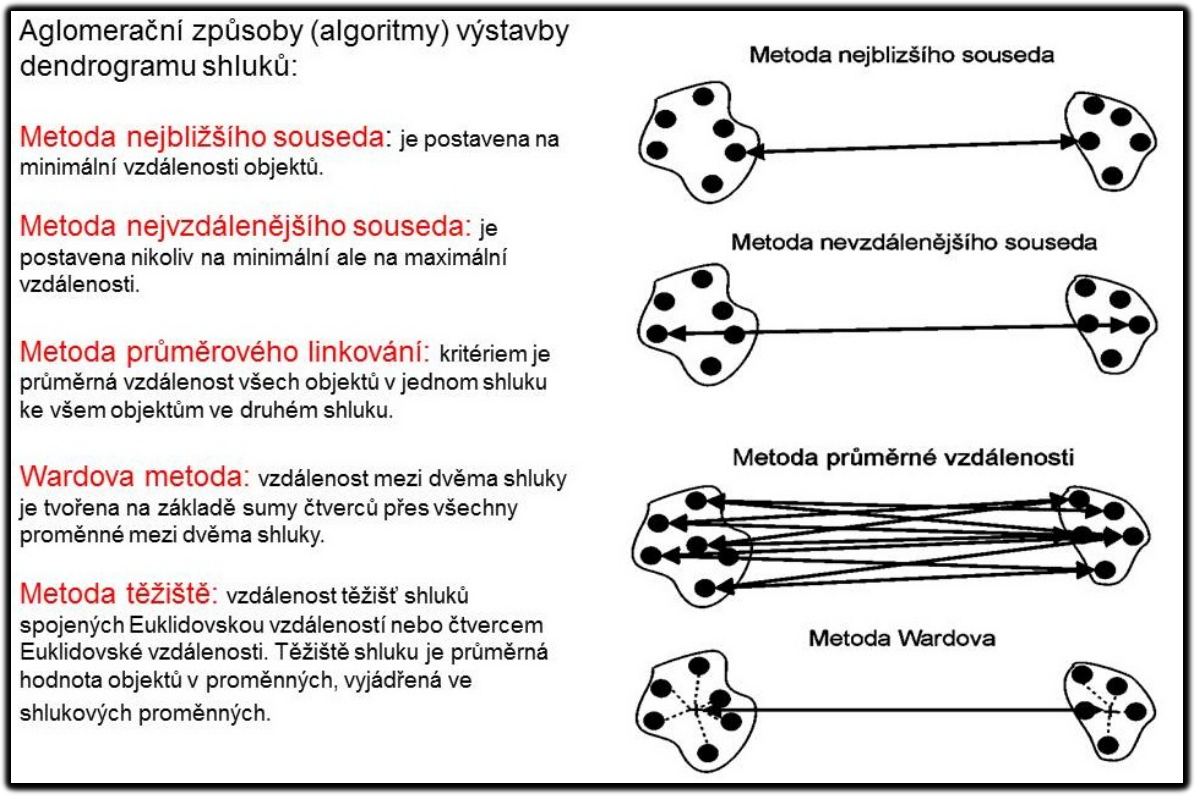
Míry podobnosti

Myšlenka podobnosti objektů je v analýze shluků myšlenkou 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.



a) **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. Nejpoužívanější míry vzdálenosti: a) eukleidovská *D*, b) manhattanská vzdálenost *D*.



b) 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 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*. Před užitím této vzdálenosti se musíme ujistit, že znaky spolu nekorelují. Když tato podmínka není splněna, shluky jsou nesprávné. 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*. Jde vlastně o vzdálenost bodů v prostoru, jehož osy nemusí být ortogonální. Vysoce korelovaný výběr znaků může skrytě převážit celý soubor znaků shlukování.

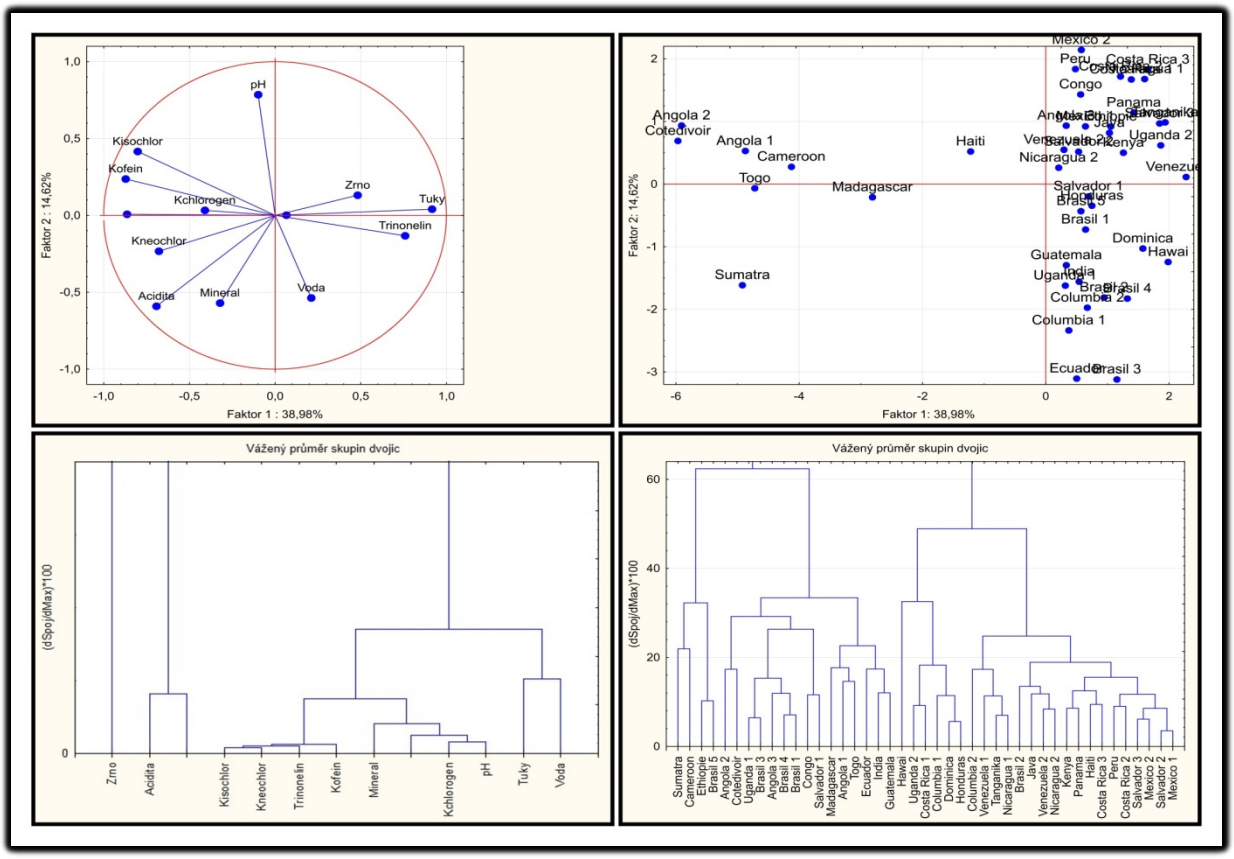
PŘÍKLAD 9.13 ***Klasifikace vlastností rozličných druhů kávy***

Byl získán výběr 43 vzorků kávy, pocházejících ze 30 zemí. U každého druhu kávy byly změřeny jeho chemické a fyzikální vlastnosti. Splňují data požadavky na homogenitu a je možné indikovat dvě či více rozličných kategorií? Vytvořte dendrogram klasifikovaných druhů kávy.

○ ***Data:*** Soubor dat *Kava* obsahuje 2 druhy kávy, Robusta a Arabica ve 43 vzorcích ze 30 zemí a popsaných 13 fyzikálně-chemickými znaky:

i značí index kávy,
Objekt značí původ kávy,
Voda značí obsah vody x_1,
Zrno značí hmotnost zrn x_2,
Extrakt značí extrakt x_3,
pH značí hodnotu pH x_4,
Acidita značí hodnotu volné acidity x_5,
Mineral značí obsah minerálů x_6,
Tuky značí obsah tuků x_7,
Kofein značí obsah kofeinu x_8,
Trinonelin značí obsah trinonelinu x_9,
Kchlorogen značí obsah kyseliny chlorogenikové x_{10},
Kneochlor značí obsah kyseliny neochlorogenikové x_{11},
Kisochlor značí obsah kyseliny isochlorogenikové x_{12},
Sumakys značí sumu kyselin chlorogenikových x_{13}.

i	*Objekt*	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
1	Mexico 1	8.9	156.6	33.5	5.8	32.7	3.8	15.2	1.1	1.0	5.4	0.4	0.8	6.6
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
43	Hawai	9.7	191.2	35.1	5.6	34.6	4.2	14.2	1.1	0.9	0.7	0.5	0.3	6.5



Hierarchické shlukování

Analýzou shluků budeme sledovat a vyšetřovat jednak podobnost objektů, analyzovanou pomocí *dendrogramu objektů*, a jednak podobnost proměnných analyzovanou pomocí *dendrogramu proměnných.*

Dendrogram, diagram shluků nebo také nazývaný vývojový strom se objeví v případě zadání hodnot původních proměnných. Výsledkem je zobrazení hodnot ve dvojrozměrném prostoru, kde osy tvoří zadané proměnné. V některých programech se objeví také „obkroužení“ objektů v jednotlivých shlucích.

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

Dendrogram podobnosti proměnných. Odhaluje nejčastěji dvojice či trojice (obecně *m*-tice) proměnných, které jsou si velmi podobné a silně spolu korelují. Odhaluje proměnné, které jsou ve společném shluku, které jsou si tím pádem značně podobné a které jsou také vzájemně nahraditelné. To má značný význam při plánování experimentu a respektování úsporných ekonomických kritérií. Některé vlastnosti (či proměnné, znaky) není třeba vůbec měřit, protože jsou snadno nahraditelné jinými a nepřispívají do celku velkou vypovídací schopností.

5. Diskriminační analýza DA

Podstata 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.

Příklad 4.24 ***Botanické třídění kosatců diskriminační analýzou***

Ve Fisherově úloze o rozměrech okvětních lístků u 150 kosatců se analyzují květy tří základních tříd: (1) *Iris setosa*, (2) *Iris versicolor*, (3) *Iris virginica*. Květy kosatců jsou popsány čtyřmi znaky: délkou kališních lístků v mm anglicky *lsepal* a jejich šířkou *wsepal*, dále délkou korunních plátků v mm *lpetal* a jejich šířkou *wpetal*.

Každý objekt je popsán $p = 4$ znaky, a to *SepalLength*, *SepalWidth*, *PetalLength*, *PetalWidth*.

Data: Data jsou v pořadí proměnných *E418a, E418b, E418c, E418d:*

5.1	3.5	1.4	0.2	5.4	3.9	1.3	0.4	5.2	4.1	1.5	0.1	5.3	3.7	1.5	0.2
4.9	3.0	1.4	0.2	5.1	3.5	1.4	0.3	5.5	4.2	1.4	0.2	5.0	3.3	1.4	0.2
4.7	3.2	1.3	0.2	5.7	3.8	1.7	0.3	4.9	3.1	1.5	0.2	7.0	3.2	4.7	1.4
4.6	3.1	1.5	0.2	5.1	3.8	1.5	0.3	5.0	3.2	1.2	0.2	6.4	3.2	4.5	1.5
5.0	3.6	1.4	0.2	5.4	3.4	1.7	0.2	5.5	3.5	1.3	0.2	6.9	3.1	4.9	1.5
5.4	3.9	1.7	0.4	5.1	3.7	1.5	0.4	4.9	3.6	1.4	0.1	5.5	2.3	4.0	1.3
4.6	3.4	1.4	0.3	4.6	3.6	1.0	0.2	4.4	3.0	1.3	0.2	6.5	2.8	4.6	1.5
5.0	3.4	1.5	0.2	5.1	3.3	1.7	0.5	5.1	3.4	1.5	0.2	5.7	2.8	4.5	1.3
4.4	2.9	1.4	0.2	4.8	3.4	1.9	0.2	5.0	3.5	1.3	0.3	6.3	3.3	4.7	1.6
4.9	3.1	1.5	0.1	5.0	3.0	1.6	0.2	4.5	2.3	1.3	0.3	4.9	2.4	3.3	1.0
5.4	3.7	1.5	0.2	5.0	3.4	1.6	0.4	4.4	3.2	1.3	0.2	6.6	2.9	4.6	1.3
4.8	3.4	1.6	0.2	5.2	3.5	1.5	0.2	5.0	3.5	1.6	0.6	5.2	2.7	3.9	1.4
4.8	3.0	1.4	0.1	5.2	3.4	1.4	0.2	5.1	3.8	1.9	0.4	5.0	2.0	3.5	1.0
4.3	3.0	1.1	0.1	4.7	3.2	1.6	0.2	4.8	3.0	1.4	0.3	5.9	3.0	4.2	1.5
5.8	4.0	1.2	0.2	4.8	3.1	1.6	0.2	5.1	3.8	1.6	0.2	6.0	2.2	4.0	1.0
5.7	4.4	1.5	0.4	5.4	3.4	1.5	0.4	4.6	3.2	1.4	0.2	6.1	2.9	4.7	1.4

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. 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.

Volba znaků, diskriminátorů

Otázkou je, zda volba znaků ***x***, tj. *diskriminátorů*, je schopna zajistit dostatečně přesné zařazení objektů do tříd, tj. *diskriminaci*. Principem volby diskriminátorů je zajištění dostatečné separability tříd a volba takových diskriminátorů, které vedou k maximalizaci nějaké míry. Jindy se volí postup, který začne se všemi původními diskriminátory a postupně se vypouštějí takové, které vedou k nedostatečné separaci.

Při výběru „účinných“ diskriminátorů jde o analogii s vícenásobnou regresní analýzou. V diskriminační analýze se místo testování, zda se hodnota čtverce vícenásobného korelačního koeficientu R^2 změní přidáním nebo odebráním proměnné, testuje, zda se změní hodnota Mahalanobisovy vzdálenosti D_M^2. Užívají se popsaná testační kritéria tak, že pro testační kritérium F se, pokud přidáváme proměnné, za hladinu statistické významnosti α dosazuje hodnota 0.15. Bohužel není známá vhodná hodnota pro α při použití F-testu, odebíráme-li proměnné, a proto je v literatuře obvykle doporučována hodnota α = 0.30. Všeobecně užívaným algoritmem je *krokový výběr diskriminátorů*, jehož principy jsou známé z lineární regrese. Postup kombinuje jak přidávání diskriminátorů, tak i jejich odstraňování. V krokové metodě má první diskriminátor, zahrnutý do modelu ve výběrovém kritériu, největší přijatelnou hodnotu. Po zavedení prvního diskriminátoru je hodnota kritéria přepočítána pro všechny diskriminátory v modelu a diskriminátor s největší přijatelnou hodnotou *zaváděcího kritéria* je zaveden do modelu jako další. V tomto okamžiku je diskriminátor, který byl zaveden do modelu jako první, znovu přepočten, zda splňuje také *odstraňovací kritérium.* Jestliže ano, je z modelu odstraněn. Vybírání diskriminátorů se ukončí, když žádné další diskriminátory nesplňují zaváděcí nebo odstraňovací kritérium.

Kritéria pro vybírání diskriminátorů

Existuje několik rozhodovacích kritérií k vybírání diskriminátorů. U Wilkova kritéria λ platí že, když diskriminátor v diskriminační funkci poskytuje nejmenší hodnotu Wilkova kritéria λ, je tento diskriminátor zahrnut do modelu. K zavedení nebo odstranění diskriminátoru je dovolen jeden krok. Maximální počet kroků k vybírání diskriminátorů je roven dvojnásobku jejich počtu.

Podobně jako ve vícenásobné regresi, když jsou některé nezávisle proměnné lineárními kombinacemi ostatních nezávisle proměnných, není možné očekávat jediné řešení. Aby se předešlo výpočetním problémům, je před zavedením diskriminátoru do modelu stanovena jeho *tolerance.* Tolerance je mírou stupně lineární asociace mezi diskriminátory. Pro i-tý diskriminátor platí $1 - R_i^2$, kde R_i^2 je čtverec vícenásobného korelačního koeficientu, když je uvažován i-tý diskriminátor za závisle proměnnou a když je uvažována regresní rovnice mezi tímto i-tým diskriminátorem a ostatními diskriminátory. Malé hodnoty tolerance indikují, že i-tý diskriminátor je většinou lineární kombinací ostatních diskriminátorů. Diskriminátory s tolerancí menší než 0.001 nejsou do modelu zařazeny.

Významnost změny Wilkova kritéria λ po zavedení diskriminátoru do modelu nebo odstranění z modelu je založena na testačním kritériu F. Aktuální hodnota testačního kritéria F nebo vypočtená statistická významnost α slouží jako kritérium pro zavedení diskriminátoru do modelu nebo k jeho odstranění. Obě kritéria však nemusí být ekvivalentní, protože pevné hodnoty kvantilu F mají rozdílnou pravděpodobnost v závislosti na počtu diskriminátorů v modelu. Aktuální vypočtená statistická významnost, spojená s kvantilem F při zavedení a s kvantilem F při odstranění, není obyčejně vypočtena z rozdělení F, protože je zde vyšetřeno mnoho diskriminátorů a jsou vybrány největší a nejmenší hodnota F. Skutečnou hladinu významnosti α je obtížné vyčíslit, protože závisí na mnoha faktorech včetně uvažováné korelace mezi diskriminátory.

Dříve než začne pracovat krokový algoritmus, jsou na začátku, tj. v nultém kroku, jak tolerance, tak i minimum tolerance položeny rovné 1, protože v modelu dosud nejsou diskriminátory. Vedle Wilkova kritéria λ se pro statistickou významnost každého diskriminátoru užívá také *F*-test. V každém kroku je ten diskriminátor, který způsobuje nejmenší hodnotu Wilkova kritéria λ, zařazen do modelu.

1.7 Automatická volba účinných znaků

Úkolem v diskriminační analýze je výběr znaků. Z velké skupiny možných znaků je třeba vybrat menší výběr těch nejlepších, maximálně 8 proměnných.

Krok	Činnost v kroku	Diskrimin.	% změny lambda	v *F*-test	Spočtená hladina α	Wilkovo lambda
0	Žádné					1.000000
1	Zavedeno	PetalLength	94.14	1180.16	0.000000	0.058628
2	Zavedeno	SepalWidth	37.09	43.04	0.000000	0.036884
3	Zavedeno	PetalWidth	32.29	34.57	0.000000	0.024976
4	Zavedeno	SepalLength	6.15	4.72	0.010329	0.023439
..	.	.....			...	.....

Detail ve 4. kroku automatického volby znaku:

Status	Diskrimin.	% změny lambda	v *F*-test	Spočtená hladina α	R^2 na ostatních *X*
Uvnitř	SepalLength	6.15	4.72	0.010329	0.858612
Uvnitř	SepalWidth	23.35	21.94	0.000000	0.524007
Uvnitř	PetalLength	33.08	35.59	0.000000	0.968012
Uvnitř	PetalWidth	25.70	24.90	0.000000	0.937850

Celkové Wilkovo λ = 0.023439

Automatický výběr znaků se plní krokově - nejprve se nalezne nejlepší znak a potom druhý nejlepší. Postupný krokový proces přidávání nejlepšího zbývajícího znaku s ověřením, zda by jeden aktivní znak mohl být odebrán, pokračuje, dokud je ještě nějaký znak k dispozici. *% změny v lambda* značí procento snížení v hodnotě lambda, které je výsledkem tohoto kroku. Wilkovo kritérium λ je analogické (1 - R^2) ve vícenásobné regresi. Vhodnější model sníží Wilkovo kritérium λ .

Vedle Wilkova kritéria λ existují ještě další kritéra:

Mahalanobisova vzdálenost $D_{1,2}^2$ je zobecněnou mírou vzdálenosti mezi dvěma třídami 1 a 2. Protože je Mahalanobisova vzdálenost $D_{1,2}^2$ kritériem pro volbu diskriminátorů, je toto kritérium všech párů tříd vyčísleno jako první. Ten diskriminátor, který měl největší hodnotu D_M^2 pro dvě od začátku nejtěsnější třídy, čili které měly nejmenší hodnotu $D_{1,2}^2$, je zařazen do modelu. Testační kritérium *F* mezi třídami, kdy provádíme test nulové hypotézy H_0: „dva vektory středních hodnot tříd objektů jsou stejné“, může být postaveno na Mahalanobisově vzdálenosti $D_{1,2}^2$. Toto kritérium může být také použito k výběru diskriminátorů. V každém kroku je zařazen do modelu diskriminátor, který vykazuje největší hodnotu kritéria *F*. Protože je Mahalanobisova vzdálenost vážená velikostí výběru, výsledky ze dvou tříd se mohou lišit, když je užito k výběru diskriminátorů kritérium *F*.

Po zavedení dalších diskriminátorů do modelu jsou sledovány změny hodnoty testačního kritéria *F*-změny. Proces přidávání diskriminátorů do modelu buď pokračuje, nebo je zastaven terminačním kritériem.

Jiný užitečný test k vyšetření, zda jeden přidaný diskriminátor zlepší zařazení objektů, využívá toho, že do souboru založeném na diskriminátorech $x_1, x_2, ..., x_p$ přiřadíme další diskriminátor x_{p+1} s Mahalanobisovou vzdáleností D^2_p. Obecně můžeme testovat, zda přidaný diskriminátor x_{p+1} významně zvětší hodnotu Mahalanobisovy vzdálenosti D_{p+1}^2, čili testujeme nulovou hypotézu H_0: $D_{p+1}^2 = D_p^2$. Zobecnění posledního testu dovoluje prověřit příspěvek několika přidaných *q* diskriminátorů současně. Když budeme začínat s diskriminátory $x_1, x_2, ..., x_p$, můžeme testovat, zda

další diskriminátory x_{p+1}, x_{p+2}, ..., x_{p+q} zlepší predikci. Budeme testovat nulovou hypotézu H_0: $D^2_{p+q} = D^2_p$.

Interpretace výsledků diskriminační analýzy

a) Exploratorní analýza zdrojové matice dat. Průzkumová analýza dat vyšetří pomocí histogramu a krabicového grafu symetrii rozdělení jednotlivých diskriminátorů a odhalí odlehlé hodnoty a porovnat proměnlivost diskriminátorů u rozličných objektů.

b) Výpočet bodových odhadů parametrů polohy a rozptýlení diskriminátorů. Nejprve zobrazíme přehled popisných statistik všech diskriminátorů. Z výsledků je vidět, zda v datech existují nějaké chybějící údaje čili „díry“. Jsou uvedeny směrodatné odchylky každého diskriminátoru, a to v každé třídě objektů. Diskriminační analýza je postavena na předpokladu, že kovarianční matice jsou stejné pro každou třídu. Tak lze posoudit, zda jsou směrodatné odchylky ve třídách zhruba stejné.

c) Korelační matice v číselné podobě. Korelační koeficienty umožňují vyšetřit celkovou korelaci, to znamená korelaci v každém páru diskriminátorů zdrojové matice.

1) *Celkové korelace* (pod diagonálou) a *kovariance* (nad diagonálou): bývají uvedeny korelace a kovariance vytvořené, když jsou ignorovány smíšené diskriminátory.

2) *Mezitřídní korelace* (pod diagonálou) a *kovariance* (nad diagonálou): výstup programu obsahuje korelace a kovariance vytvořené za použití průměrů místo jednotlivých objektů.

3) *Vnitrotřídní korelace* (pod diagonálou) a *kovariance* (nad diagonálou): jsou uvedeny korelace a kovariance, vytvořené z dat, ve kterých byly třídní průměry odečteny.

4) *Maticový diagram znaků objektů*. Maticový diagram diskriminátorů objektů zobrazuje hodnoty Pearsonových korelačních koeficientů a ukazuje největší a nejmenší korelaci dvojice znaků. Korelační matice proto indikuje jakousi zřetelnou strukturu v datech. Když by se v tomto diagramu body nerozdělily do několika mráčků, byla by malá naděje, že diskriminace bude vůbec možná.

2. Diskriminační funkce.

2.1 Odhady diskriminačních koeficientů $a_0, a_1, \ldots, a_p$ diskriminační funkce Z_k pro třídy Setosa G_1, Versicolor G_2, Virginica G_3.

Znak	Setosa G_1	Versicolor G_2	Virginica G_3
Absolutní člen	-85.20985	-71.754	-103.2697
SepalLength	2.354417	1.569821	1.244585
SepalWidth	2.358787	0.707251	0.3685279
PetalLength	-1.643064	0.5211451	1.276654
PetalWidth	-1.739841	0.6434229	2.107911

Tabulka obsahuje celkem tři diskriminační funkce, jednu pro každou třídu ve sloupci. Tato diskriminační funkce je určena k zařazování nových, dosud nezařazených objektů do tříd čili ke klasifikaci, a proto se v této formě nazývá *klasifikační funkce.*

Klasifikační funkce *pro první třídu* G_1: -85.20985+2.35442*SepalLength +2.35879*SepalWidth-1.64306*PetalLength-1.73984*PetalWidth.

Tyto tři funkce mohou být použity k zařazení nových kosatců do navržených tříd.

d) Výstavba diskriminační funkce a vyšetření jednotlivých diskriminátorů. Prvotním cílem diskriminační analýzy je klasifikace objektů dle zadaných diskriminátorů nazývaná výstavbou modelu vhodné Fisherovy lineární diskriminační funkce. Princip spočívá v postupném

přidávání (nebo odebírání) jednotlivých diskriminátorů do diskriminační funkce a sledování vlivu dotyčného diskriminátoru na diskriminaci objektů do tříd.

1) **Výstavba při přidávání (či odstraňování) dotyčného diskriminátoru z (nebo do) Fisherovy lineární diskriminační funkce:** Při každém kroku výstavby diskriminačního modelu jde o přidání znaku do diskriminační funkce. Sledují se následující rozhodčí testační kritéria, zda přidávaný diskriminátor způsobil významné zlepšení diskriminace objektů do tříd, což by se mělo projevit ve zlepšené hodnotě dotyčného testačního kritéria. Program tak vybere diskriminátor s největší hodnotou kritéria F. Při postupném odebírání jednotlivých diskriminátorů program vybere ten diskriminátor k vyřazení, který vykazuje nejmenší hodnotu kritéria F. Po každém kroku se vyčíslí také vícenásobný korelační koeficient R a jeho čtverec R^2 zvaný koeficient determinace, a to pro každý diskriminátor korelující se všemi dosud zadanými znaky. Kritérium tolerance se vypočte jako 1 - R^2 a značí míru nadbytečnosti dotyčného diskriminátoru. Když například vykáže právě přidaný diskriminátor hodnotu tolerance 0.01, znamená to, že tento diskriminátor vykazuje 99% nadbytečnost vůči všem dosud zavedeným diskriminátorům. V takovém případě, kdy jeden nebo i více diskriminátorů by vykázaly příliš velikou nadbytečnost, by nebylo totiž vůbec možné diskriminační analýzu objektů provést. Je proto vhodné zařadit v datech kritérium nevyhovujícího diskriminátoru, vykazujícího toleranci například 0.01.

2) **Ovlivnění diskriminace každým samotným diskriminátorem**. Je uveden vliv jednotlivých diskriminátorů na výsledky diskriminační analýzy.

a) **Odhady parametrů b_0, b_1, ..., b_p lineární Fisherovy lineární diskriminační funkce pro jednotlivé třídy**: Jsou uvedeny odhady neznámých parametrů b_0, b_1, ..., b_p Fisherovy lineární diskriminační funkce. Tyto parametry jsou nazývány diskriminačními koeficienty. Technika předpokládá, že diskriminátory v každé třídě objektů vykazují vícerozměrné normální rozdělení se shodnými variančně-kovariančními maticemi ve třídách. Technika je dostatečně robustní i při nesplnění těchto předpokladů.

b) **Odhady parametrů b_0, b_1, ..., b_p lineárního regresního modelu pro jednotlivé třídy**. Regresní parametry b_0, b_1, ..., b_p lineárního regresního modelu pro každou třídu objektů jsou vyčísleny postupem: (1) Vytvoří se indikátorové proměnné, jedna je pro každý druh objektů. Každý diskriminátor je položen roven jedné. (2) Proloží se vícenásobnou regresí nezávisle proměnných každá ze tříd objektů. (3) Obdržíme odhady regresních parametrů. Těmito regresními parametry budou predikované hodnoty ležet mezi nulou a jednou. Určení, ke které třídě objektů jedinec patří, se provede tak, že se vybere třída s nejvyšším skóre.

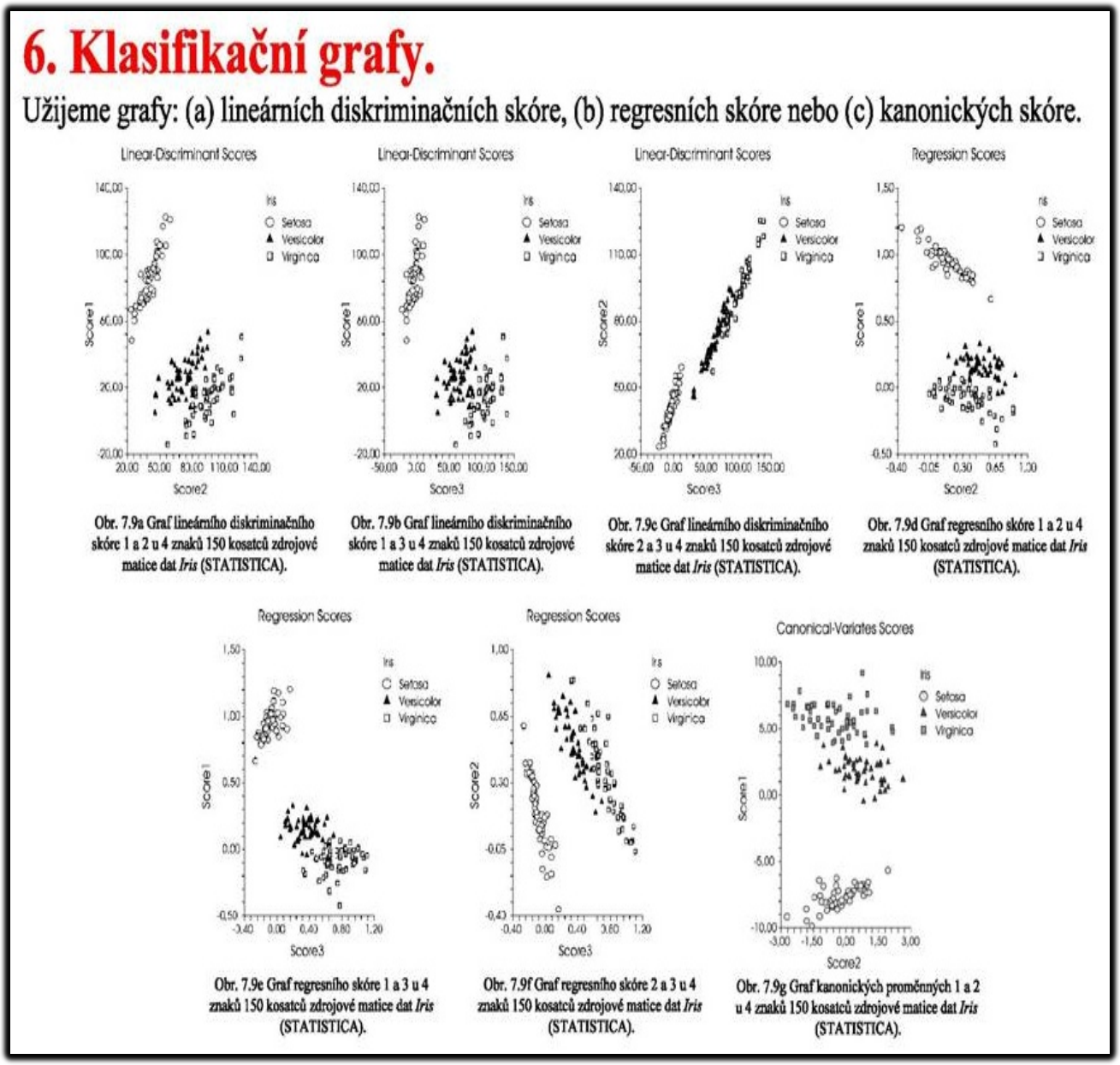
Klasifikace objektů do tříd

a) **Klasifikační funkce při diskriminaci do tříd**: Klasifikační funkce by se neměly zaměňovat s diskriminační funkcí. Klasifikační funkce jsou vyčísleny pro každou třídu objektů a mohou být použity přímo ke klasifikování objektů. Objekt bude patřit do té třídy, ve které dosáhne nejvyšší hodnoty vyčísleného klasifikačního skóre.

b) **Diskriminace objektů do jednotlivých tříd a celkově.** Tabelárně je ukázáno, jak navržené diskriminační funkce klasifikují objekty v datech. Bylo-li dosaženo perfektní klasifikace, obdržíme v matici mimo diagonálu nuly. Řádky v tabulce představují skutečné a zadávané třídy objektů, zatímco sloupce představují predikované třídy objektů.

c) **Přehled chybně zařazených objektů v řádcích při diskriminaci do jednotlivých tříd**. V řádku se u každého chybně klasifikovaného objektu nachází název známé a název zadávané třídy objektů a dále název predikované třídy objektů. Následuje 100krát zvětšená hodnota

pravděpodobnosti vyjádřená v procentech, že objekt se nachází v dané třídě objektů. Procento pravděpodobnosti se jeví totiž názornější než normovaný odhad v rozmezí 0 a 1. Hodnota blízko 100 % ukazuje, že objekt patří do dotyčné třídy. Při užití lineární diskriminační techniky se vyčíslí pravděpodobnosti $P(i)$, že tento řádek patří do i-té třídy: nechť f_i, $i = 1, ..., K$, je hodnota lineární diskriminační funkce a $\max(f_k)$ je maximální skóre ze všech tříd.



d) **Zařazení objektů klasifikací pomocí diskriminační funkce do jednotlivých tříd**. Pro každý objekt obsahuje tabulka vždy skutečnou, čili známou třídu objektů, predikovanou třídu objektů a procento pravděpodobnosti zařazení do dotyčné třídy objektů.

Výklad grafů diskriminace všech objektů do tříd.

V grafické interpretaci řady software se nabízí několik zobrazení. Jsou to

(a) grafy lineárních diskriminačních skóre,

(b) grafy regresních skóre nebo

(c) grafy kanonických skóre.

Na základě diagramů těchto tří druhů skóre lze pak snáze vytvořit svou vlastní interpretaci diskriminace objektů. Diagramy poskytnou vizuální vysvětlení, jak diskriminační funkce klasifikují objekty v datech. Diagram obvykle ukazuje hodnoty prvního a druhého kanonického skóre.

3. Klasifikace objektů.

3.1 Zařazení (klasifikace) objektů klasifikační funkcí do tříd tvoří klasifikační matici kosatců ve třídách Setosa G_1, Versicolor G_2, Virginica G_3:

	Nalezená predikovaná třída			
Daná třída	Setosa G_1	Versicolor G_2	Virginica G_3	Celkově $G_1+G_2+G_3$
Setosa	50	0	0	50
Versicolor	0	34	16	50
Virginica	0	7	43	50
Celkově	50	41	59	150

Redukce v klasifikační správnosti v důsledku proměnných $X = 77.0$ %.

Klasifikační matice ukazuje, jak diskriminační funkce zařazují objekty do tří tříd. Bylo-li dosaženo perfektního zařazení, obdržíme v matici mimo diagonálu nuly. *Redukce v klasifikační správnosti* obsahuje procento snížení v zařazovací správnosti, dosažené diskriminačními funkcemi vůči očekávanému zařazení dle původní závisle proměnné.

Z grafu bývá patrné klasifikační pravidlo: první kanonická funkce postačuje k diskriminování mezi druhy objektů, protože třídy objektů mohou být snadno odděleny vertikální osou. Existuje však také software (například náročnější na obsluhu objektově programovaný software S-Plus), který umožňuje 3D zobrazení s rotací podél os v prostoru. Potom by bylo vytvoření a rozlišení tříd objektů ještě názornější.

3.2 Přehled chybně zařazených objektů v řádcích do tříd Setosa G_1, Versicolor G_2, Virginica G_3 dle klasifikačních funkcí proti původní závisle proměnné.

			Procento zařazení do jednotlivé třídy		
Objekt	**Daná třída**	**Nalezená třída**	**Setosa G_1**	**Versicolor G_2**	**Virginica G_3**
5	Virginica	Versicolo	-1.8	58.6	43.1
9	Versicolo	Virginica	10.3	20.2	69.5
22	Versicolo	Virginica	18.8	22.6	58.6
28	Versicolo	Virginica	22.1	35.5	42.4
29	Versicolo	Virginica	22.1	27.4	50.6
38	Versicolo	Virginica	10.6	38.3	51.1
45	Virginica	Versicolo	-31.4	66.4	65.0
57	Virginica	Versicolo	-18.6	83.9	34.7
62	Versicolo	Virginica	24.4	34.0	41.6
66	Versicolo	Virginica	11.9	37.9	50.2
70	Versicolo	Virginica	12.1	41.5	46.3
78	Virginica	Versicolo	-7.3	58.4	48.9
91	Virginica	Versicolo	-16.1	83.8	32.3
95	Versicolo	Virginica	23.7	14.3	62.0
106	Versicolo	Virginica	20.7	30.7	48.7
111	Virginica	Versicolo	-21.4	63.8	57.6
112	Virginica	Versicolo	-23.9	71.8	52.1
114	Versicolo	Virginica	17.1	35.6	47.2
117	Versicolo	Virginica	22.1	38.9	39.0
130	Versicolo	Virginica	30.9	32.4	36.8
131	Versicolo	Virginica	14.0	39.6	46.4
142	Versicolo	Virginica	21.4	38.6	40.0
148	Versicolo	Virginica	6.8	36.8	56.4

V řádku je hodnota pravděpodobnosti (v procentech), že objekt se nachází v dané třídě kosatců. Hodnota blízko 100 % značí, že objekt patří do dotyčné třídy.

LITERATURA:

1. Siotani M., Hayakawa T., Fujikoshi Y.: *Modern Multivariate Statistical Analysis*, A Graduate Course and Handbook. American Science Press, Columbia 1985.

2. Seber G. A. F.: *Multivariate Observations*. Wiley, New York 1984.

3. Meloun M., Militký J. , Forina M.: *Chemometrics for Analytical Chemistry, Volume 1. PC-Aided Statistical Data Analysis*, Ellis Horwood, Chichester 1992.

4. Brereton R. G. *Multivariate Pattern Recognition in Chemometrics, Illustrated by Case Studies*, Elsevier 1992,

5. Krzanowski W. J.: *Principles of Multivariate Analysis, A User's Perspective*, Oxford Science Publications 1988,

6. Meloun M., Militký J., *Kompendium statistického zpracování experimentálních dat*, Academia Praha, 2002, Karolinum Praha 2012 (4. vydání).

7. Meloun, M., Militký J.: *Interaktivní statistická analýza dat*, Karolinum Praha 2012 (4. vydání).

8. Meloun, M., Militký, J., Hill, M: *Statistická analýza vícerozměrných dat v příkladech,* Academia 2012 (2. vydání)

9. Hebák, P., Hustopecký, J.: *Vícerozměrné statistické metody s aplikacemi,* SNTL Praha 1987.

10. Meloun M., Militký J.: *Statistical Data Analysis, A Practical Guide with 1250 Exercises and Answer Key on CD*, Woodhead Publishing India in Materials, PVT LTD, New Delhi, Cambridge2011.