Vícerozměrná analýza dat metodou hlavních komponent a shluků

Prof. RNDr. Milan Meloun, DrSc.
Katedra analytické chemie, Chemickotechnologická fakulta, Univerzita Pardubice,
nám. Čs. Legií 565, 532 10 Pardubice,
http: //meloun.upce.cz, email: milan.meloun@upce.cz,

Souhrn: *Vícerozměrná statistická analýza je založena na latentních proměnných, které jsou lineární kombinací původních proměnných,* $y = w_1 x_1 + ... + w_m x_m$. *Zdrojová matice dat obsahuje proměnné v m sloupcích a objekty v n řádcích. Data jsou před zpracováním škálována. Cílem je nalézt shluk jako množinu podobných objektů s podobnými proměnnými. Podobnost objektů posuzujeme na základě vzdálenosti (míry) objektů v m-rozměrném prostoru: čím je vzdálenost shluků či objektů větší, tím menší je jejich podobnost. Strukturu a vazby mezi proměnnými vystihují metody snížení dimensionality, metoda hlavních komponent (PCA). Důležitou pomůckou je rozptylový diagram, který zobrazuje objekty, rozptýlené v rovině prvních dvou hlavních komponent. Graf komponentních vah porovnává vzdálenosti mezi proměnnými* x_i *a* x_j, *kde krátká vzdálenost značí silnou korelaci. Dvojný graf pak kombinuje oba předchozí grafy. Objekty lze seskupovat do shluků hierarchicky dle předem zvoleného způsobu metriky (průměrově, centroidně, nejbližším sousedem, nejvzdálenějším sousedem, medianově, mezi těžišti a průměrnou vazbou) a nehierarchicky dle uživatelem vybraných objektů. Výsledkem je vývojový strom čili dendrogram.*

ÚVOD

Vícerozměrná statistická analýza vychází z koncepce latentních proměnných (faktorů, kanonických proměnných) *y*, které jsou lineární kombinací původních proměnných *x* s vhodně volenými vazbami. Latentní proměnná *y* je kombinací *m*-tice sledovaných (měřených resp. jinak získaných) proměnných $x_1, x_2, ..., x_m$ ve tvaru $y = w_1 x_1 + ... + w_m x_m$. Jednotlivé vícerozměrné metody využívají různých způsobů stanovení vah $w_1, w_2, ..., w_m$.

Zdrojová matice má rozměr $n \times m$, *n* řádků a *m* sloupců. Před vlastní aplikací vhodné metody vícerozměrné statistické analýzy je třeba vždy provést *exploratorní (průzkumovou) analýzu dat*, která umožňuje

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

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

(a) *struktura a vazby v proměnných* nebo
(b) *struktura a vazby v objektech*:
  (1) Hledání struktury *v proměnných* v metrické škále: *faktorová analýza FA* a *analýza hlavních komponent PCA.*
  (2) Hledání struktury *v objektech* v metrické škále: *shluková analýza*.
  (3) Hledání struktury *v objektech* v metrické i v nemetrické škále: *vícerozměrné škálování.*
  (4) Hledání struktury *v objektech* v nemetrické škále: *korespondenční analýza.*
  (5) 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 proměnnými ale i mezi objekty se zabývají techniky redukce proměnných na latentní proměnné, metoda *analýzy hlavních komponent (PCA)* a *metoda*

faktorové analýzy (FA). Důležitou metodou určení vzájemných vazeb mezi proměnnými je i *kanonická korelační analýza CA*, která se používá ke zkoumání závislosti mezi dvěma skupinami proměnných, přičemž jedna ze skupin se považuje za proměnné nezávislé a druhá za skupinu proměnných závislých.

METODA HLAVNÍCH KOMPONENT PCA

Metoda hlavních komponent (PCA) je jednou 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 nebo-li 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 *hlavními komponentami*. 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.

● **Zdrojová matice dat $\boldsymbol{X}$** ($n \times m$): Zdrojová matice dat $\boldsymbol{X}$ ($n \times m$) obsahuje n objektů v řádcích a m znaků ve sloupcích. *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, a pod. 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 monografiích [22, 26, 31].

● **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.

● **Cíl metody hlavních komponent PCA:** Základním 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. 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.

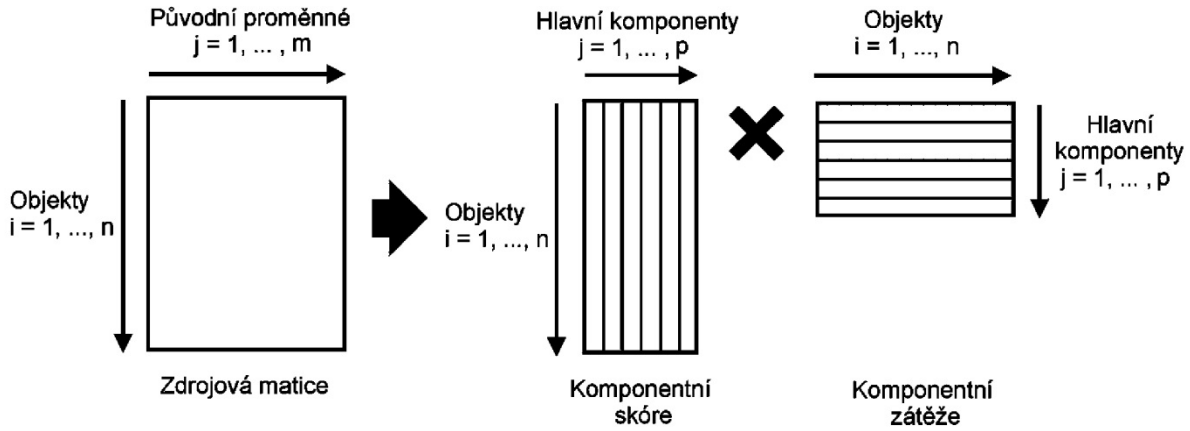
Rozdíl mezi souřadnicemi objektů v původních znacích a v hlavních komponentách čili ztráta informace projekcí do menšího počtu rozměrů se nazývá *mírou těsností proložení modelu PCA* nebo

také *chybou modelu PCA*. Na obr. 1 je tato situace schematicky znázorněna spolu s použitým označením.

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, $j = 1, ..., m$, k hlavním komponentám y_j, $j = 1, ..., k$, tvoří dominantní součásti analýzy modelu hlavních komponent PCA.

Z obr. 1 je zřejmé, že 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 [26].



Obr. 1 Rozklad zdrojové matice dat metodou hlavních komponent PCA.

● **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* $\boldsymbol{X}$.

● $\boldsymbol{X}$ **= 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, obr. 2).

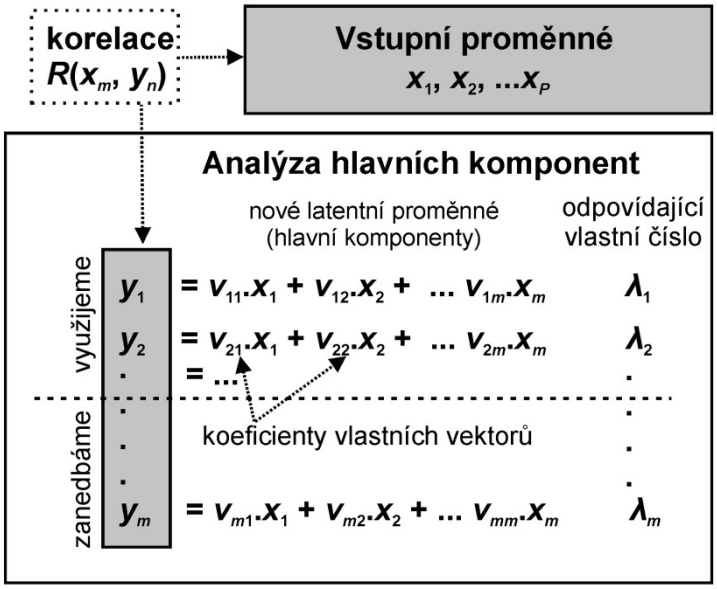
Model hlavních komponent má pak tvar $\boldsymbol{X} = \boldsymbol{T}\,\boldsymbol{P}^T + \boldsymbol{E}$ = *struktura dat* + *šum* . Zdrojovou matici dat $\boldsymbol{X}$ je třeba nejprve centrovat $x_{ik} = x_{ik} - \bar{x}_k$. V metodě hlavních komponent pracujeme za předpokladu, že zdrojová matice $\boldsymbol{X}$ je rozložena na součin matic $\boldsymbol{TP}^T$ a na matici reziduí $\boldsymbol{E}$, kde matice $\boldsymbol{T}$ je matice komponentního skóre a $\boldsymbol{P}^T$ je transponovaná matice komponentních vah, $\boldsymbol{E}$ je matice reziduí. Cílem metody hlavních komponent je místo $\boldsymbol{X}$ využívat dále jenom součin $\boldsymbol{TP}^T$ a tak oddělit šum od struktury dat $\boldsymbol{TP}^T$. 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 je možno rozepsat do tvaru

$$\mathbf{X} = \boldsymbol{t}_1\,\boldsymbol{p}_1^T + \boldsymbol{t}_2\,\boldsymbol{p}_2^T + ... + \boldsymbol{t}_A\,\boldsymbol{p}_A^T + \boldsymbol{E}$$

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 a dostane se $\boldsymbol{X} = \boldsymbol{t}_1\,\boldsymbol{p}_1^T + \boldsymbol{E}_1$.

2. Odečte se příspěvek první hlavní komponenty *PC*1 od $\boldsymbol{X}$ dle rovnice $\boldsymbol{E_1} = \boldsymbol{X} - \boldsymbol{t_1}\,\boldsymbol{p_1}^{\mathbf{T}}$.
3. Vyčíslí se $\boldsymbol{t_2}$ a $\boldsymbol{p_2}^{\mathbf{T}}$ z matice $\boldsymbol{E_1}$ a dostane se $\boldsymbol{X} = \boldsymbol{t_1}\,\boldsymbol{p_1}^{\mathbf{T}} + \boldsymbol{t_2}\,\boldsymbol{p_2}^{\mathbf{T}} + \boldsymbol{E_2}$.
4. Odečte se příspěvek druhé hlavní komponenty *PC*2 od $\boldsymbol{E_2}$ dle rovnice $\boldsymbol{E_1} = \boldsymbol{E_2} - \boldsymbol{t_2}\,\boldsymbol{p_2}^{\mathbf{T}}$.
5. Podobně se pokračuje, až se vyčíslí dostatečný počet *A* hlavních komponent.



Obr. 2 Schéma maticových výpočtů v metodě hlavních komponent PCA.

● **Komponentní váhy, zátěže - vztah mezi *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 *PC*3. 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.

Na *grafu komponentních vah* $\boldsymbol{p}$ pro *PC*1 a *PC*2 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 *PC*1 značí skóre t_{i1}. Projekce téhož objektu na druhou hlavní komponentu *PC*2 značí skóre t_{i2}, ...atd. Každý objekt má svůj soubor komponentních skóre t_{i1}, t_{i2}, ..., $t_{ip\text{-}1}$. Hodnot skóre je stejný počet jako hlavních komponent.

Matice všech skóre pro všechny objekty se nazývá *matice skóre* ***T***. Skóre pro jeden objekt v této matici tvoří jeden řádek. Sloupce v této matici jsou ortogonální. *Vektor skóre* je sloupec v matici ***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 *PC*1 proti skóre *PC*2. Je snadno k pochopení, protože jde o dva směry, podél kterých shluk objektů vykazuje největší (*PC*1) a druhé největší (*PC*2) 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 *PC*1. Obecně lze říci, že *PC*1 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 ***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* ***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 ***E***, velká hodnota A méně ponechaného šumu v matici ***E***.

Matice ***E*** obsahuje jak malá tak i velká rezidua. Vyhodnocení ***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 ***X*** je stejné, jako centrování zdrojové matice dat. Pro $A = 0$ 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 ***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 $A = 0$ bude $\boldsymbol{E_0}$ představovat 100%. *Reziduálový rozptyl* bude pak 100% a *modelem objasněný rozptyl* bude 0%. Velikost ***E*** je charakterizována čtverci reziduí čili rozptyly. Existují dva způsoby jak sčítat prvky matice ***E***: buď v řádku a obdržet rezidua každého objektu, nebo ve sloupci a obdržet rezidua každého znaku.

GRAFICKÉ DIAGNOSTIKY PCA

Graficky lze výsledek analýzy hlavních komponent zobrazit v několika grafech hlavních komponent následujícím způsobem:

(a) **Cattelův indexový graf úpatí vlastních čísel** (Scree Plot) je vlastně sloupcový diagram vlastních čísel nebo reziduálního rozptylu proti stoupající hodnotě indexu, pořadového čísla *A* (obr. 5). Zobrazuje relativní velikost jednotlivých vlastních čísel. Řada autorů ho s oblibou využívá k určení počtu *A* "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 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** (Plot Components Weights) zobrazí komponentní váhy pro první dvě hlavní komponenty (obr. 7). 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 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, některé 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. Kovariance však nemusí ještě nutně znamenat korelaci. 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 vahou v y_1 tím pádem mnohem důležitější než znaky x_j s vysokou vahou v y_2. Znaky s úhlem 0° mezi průvodiči jsou zcela pozitivně korelované, znaky s úhlem 90° jsou zcela nekorelované zatímco znaky s úhlem 180° 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** (Scatterplot) zobrazuje *komponentní skóre* čili hodnoty obyčejně prvních dvou hlavních komponent u všech objektů (obr. 8). Lze snadno nalézt shluk vzájemně podobných objektů a dále 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. 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 $m(m\text{-}1)/2$ = 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 nějvětší míru proměnlivosti v datech. Interpretace rozptylového diagramu komponentního skore lze shrnout do těchto bodů:

1. *Umístění objektů.* Objekty daleko od počátku jsou extrémy. Objekty nejblíže počátku jsou nejtypičtější.
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. To platí jen v případech, kdy se nejedná o zdánlivou nehomogenitu danou sešikmením dat a odstranitelnou transformací znaků.
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 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 biologickou podobnost.
7. *Vysvětlení místa objektu.* Umístění objektu na ploše v diagramu může být porovnáváno s komponentními vahami znaků ve dvojném grafu a pomocí znaků pak i vysvětleno.

(d) **Dvojný graf** (Biplot) kombinuje předchozí dva grafy. 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 původní znaky x_j do hlavní komponety, čili je úměrná komponentní váze. Kombinace obou grafů v jediném přináší cenné srovnání, jeden graf působí zde doplňkově vůči druhému.

METODA SHLUKŮ

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. Měly by být využity pouze takové znaky, které dostatečně rozlišují mezi objekty.

● **Identifikace odlehlých objektů:** 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.

● **Míry podobnosti:** 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. Podobnost může být měřena rozličnými způsoby, které se dají obyčejně zařadit do jedné ze tří základních skupin:

(1) *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 jedničce. V případě ordinální škály (pořadová čísla) je analogickou mírou podobností *Spearmanův korelační koeficient.*

(2) *Míry vzdálenosti:* Představují nejčastěji užívané míry, založené na prezentaci objektů v prostoru, jehož souřadnice tvoří jednotlivé znaky. 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 vzdálenost

$$d_E(x_k, x_l) = \sqrt{\sum_{j=1}^{m} (x_{kj} - x_{lj})^2}$$

představuje standardní typ vzdálenosti. Vedle Eukleidovské vzdálenosti se užívá také *čtverec Eukleidovské vzdálenosti*, který tvoří základ Wardovy metody shlukování. Často je užívaná *Manhattanská vzdálenost* zvaná také *vzdálenost městských bloků* nebo *Hammingova metrika*, definovaná vztahem

$$d_H(x_k, x_l) = \sum_{j=1}^{m} |x_{kj} - x_{lj}|$$

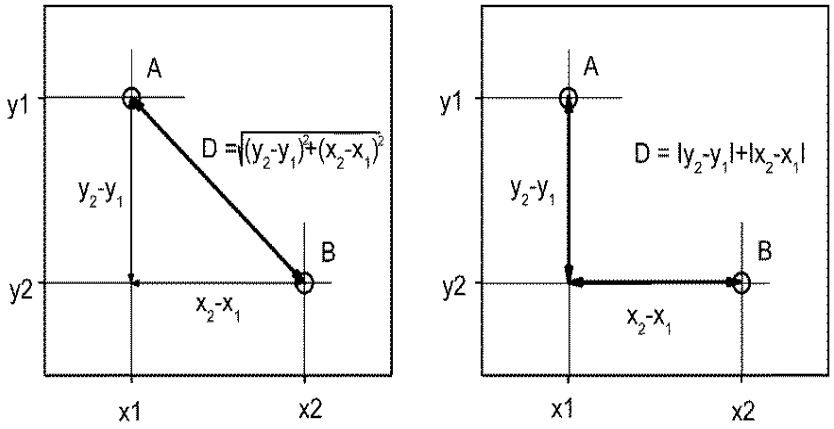
Před užitím této vzdálenosti se musíme ujistit, že znaky spolu nekorelují. Když tato podmínka není splněna, shluky jsou nesprávné. Další mírou je zobecněná *Minkovského metrika*, pro kterou platí

$$d_M(x_k, x_l) = \sqrt[z]{\sum_{j=1}^{m} |x_{kj} - x_{lj}|^z}$$

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

$$d_{CH}(x_k, x_l) = \sqrt{2 \left[1 - \frac{\sum_{j=1}^{m} x_{kj}\, x_{lj}}{\sum_{j=1}^{m} x^2_{kj} \sum_{j=1}^{m} x^2_{lj}} \right]}$$

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



Obr. 3 Nejpoužívanější míry vzdálenosti: (a) Eukleidovská *D*, (b) Manhattanovská vzdálenost *D*

Problém všech vzdálenostních měr vzniká při použití nestandardizovaných dat, které mohou způsobit rozdíly mezi shluky, díky často veliké odlišnosti jednotek měření.

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

$$d_{Ma}(x_k, x_l) = \sqrt{(x_k - x_l)^T C^{-1} (x_k - x_l)}$$

Jde vlastně o vzdálenost bodů v prostoru, jehož osy nemusí být orthogonální. Vysoce korelovaný

výběr znaků může skrytě převážit celý soubor znaků shlukování.

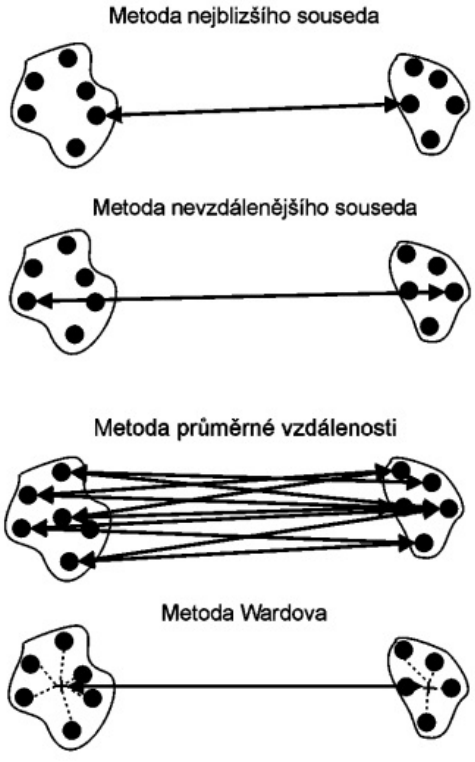
- **Standardizace dat**: Před vlastní shlukovou analýzou je třeba řešit otázku, zda je třeba data standardizovat. Musí se respektovat fakt, že většina měr vzdálenosti je velmi citlivá na měřítka (stupnice), vedoucí k různé numerické velikosti znaků. Obecně platí pravidlo, že znaky s větší mírou proměnlivosti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

(1) *Standardizování znaků*: Nejužívanější formou standardizace je *normalizace každého znaku* do svého *Z-skóre*, tj. 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.

(2) *Standardizace objektů:* Když chceme identifikovat shluky dle vzdálenosti pak standardizace není vhodná. Standardizace objektů nebo-li řádková standardizace může být však efektní ve speciálních případech.

- **Způsoby shlukování**: *Shluk* (cluster) je skupina objektů, jejichž vzdálenost je menší než vzdálenost s objekty do shluku nepatřícími. Podle způsobu shlukování se postupy dělí na *hierarchické* a *nehierarchické*. Hierarchické se dělí dále na *aglomerační* a *divizní.*

- **Hierarchické shlukovací postupy**: 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 naopak tento shluk je zařazen jako objekt. Celý postup se opakuje tak dlouho, dokud všechny objekty netvoří jeden velký shluk nebo dokud nezůstane určitý, předem zadaný počet shluků. *Divizní postup* je obrácený. Vychází se z množiny všech objektů jako jediného shluku a jeho postupným dělením získáme systém shluků, až skončíme ve stadiu jednotlivých objektů.



Obr.4 Nejčastěji užívané metriky shlukování.

Mezi metody metriky shlukování patří:

(1) *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 v prvním shluku vůči libovolnému bodu v druhém. Dva shluky jsou propojeny v libovolném stádiu nejkratší spojkou.

(2) *Metoda nejvzdálenějšího souseda:* 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.

(3) *Metoda průměrné vzdálenosti:* Kritériem vzniku shluků je průměrná vzdálenost všech objektů v jednom shluku ke všem objektům ve druhém shluku.

(4) *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.

(5) *Metoda těžiště*: 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.

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

● **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 znaků ukazuje rozlišení znaků ve shlucích. Jeho interpretace je snadná: znaky blízko sebe jsou propojeny spojovací úsečkou hodně nízko, mají malou vzdálenost čili značnou vzájemnou podobnost. Znaky propojené hodně vysoko mají malou podobnost a mezi sebou vykazují velkou vzdálenost.

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.

Míra věrohodnosti: Dendrogram lze sestrojit celou řadou technik. Prvním kritériem těsnosti proložení při volbě "nejlepšího dendrogramu", jež nejlépe odpovídá struktuře objektů a znaků mezi objekty, je *kofenetický korelační koeficient CC*. Je to Pearsonův korelační koeficient mezi skutečnou a predikovanou vzdáleností, založenou na dendrogramu.

Druhým kritériem těsnosti proložení je *kritérium delta* Δ, které měří stupeň přetvoření struktury dat spíše než stupeň podobnosti. Kritérium delta je definováno vztahem

$$\Delta_A = \left[\frac{\sum_{j<k}^{N} |d_{jk} - d_{jk}^*|^{1/A}}{\sum_{j<k}^{N} (d_{jk}^*)^{1/A}} \right]^A$$

kde A = 0.5 nebo 1, d_{ij} je vzdálenost v původní matici vzdáleností a d^*_{ij} je vzdálenost získaná z dendrogramu. Je žádoucí, aby hodnoty *delta* byly blízké nule. Řada autorů ukázala, že metoda průměrová vede obvykle k nejlepšímu dendrogramu.

Postup obecné analýzy shluků

Analýza shluků je vždy silnou analytickou pomůckou k účelům zjednodušení, průzkumu a potvrzení struktury.

1. krok: Cíle analýzy shluků: Primárním cílem je rozdělení souboru objektů do dvou nebo více skupin, tříd či shluků. Sledujeme tři cíle:

(a) *Popis systematiky:* Tradičním využitím jsou průzkumové cíle a popis systematiky - *taxonomie*, tj. empirická klasifikace objektů. Analýzou shluků se dospěje k určitým shlukům objektů, které jsou pak porovnány s jejich teoreticky odvozenou typologií.

(b) *Zjednodušení dat:* Při hledání taxonomie poskytuje analýza shluků zjednodušený pohled na objekty. Zatímco faktorová analýza se snaží nalézt strukturu znaků, analýza shluků činí totéž ale pro objekty. Na objekty se pak už nehledí jako na jeden společný soubor ale na oddělené shluky objektů, rozčleněné dle jejich vlastností.

(c) *Identifikace vztahu:* Po nalezení shluků objektů, a tím i struktury mezi objekty je snadnější odhalit vztahy mezi objekty, což by bylo mezi samotnými objekty daleko obtížnější. Shluky mohou být předmětem dalšího kvalitativního uvažování.

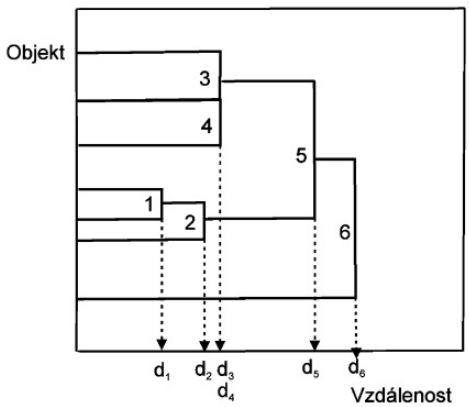
2. krok: Formulace úlohy analýzy shluků: S vybranými znaky budeme shlukovat vyšetřované objekty. Nejprve však je třeba odpovědět na tři otázky: (1) Mohou být v datech nějaké odlehlé objekty, které mohou být posléze odstraněny? (2) Jak vyjádříme podobnost objektů? (3) Měla by být data před analýzou shluků standardizována?

3. krok: Předpoklady analýzy shluků: Analýza shluků objektů není charakteru statistického testování. Požadavky normality, linearity, homoskedasticity, které jsou tolik důležité v ostatních vícerozměrných technikách nemají zde význam. Přesto existují dva kritické předpoklady: reprezentativnost a vliv multikolinearity.

4. krok: Výstavba shluků a celková těsnost proložení: Za vhodné rozlišovací kritérium je možné použít maximalizaci rozdílů mezi shluky, a toto porovnávat vůči proměnlivosti uvnitř shluků. Testování poměru roztylu mezi jednotlivými shluky vůči průměru rozptylu uvnitř shluku lze porovnat s *F* kritériem v analýze rozptylu.

Hierarchické shlukování: Tyto algoritmy se týkají konstrukce stromovité struktury shluků, dendrogramu. Existují v zásadě dva typy hierarchického shlukování, *aglomerační* a *divizní*. Grafickým zobrazením růstového stromu je diagram, také zvaný *dendrogram*.

Pět nejpoužívanějších aglomeračních algoritmů výstavby shluků jsou metoda nejbližšího souseda, metoda nejvzdálenějšího souseda, metoda průměrová, Wardova metoda, a metoda těžiště.



Obr. 5 Postupná výstavba dendrogramu.

5. krok: Interpretace shluků: Interpretace shluků se týká vyšetření každého shluku v pojmech shlukových znaků a především pojmenování shluků, které vystihuje podstatu a povahu shluků. Při zahájení interpretace si uvědomíme, že ve shlukové analýze je často užívanou mírou těžiště shluků. Uživatel se *musí* vrátit k původním datům v původních znacích a vyčíslit průměrové profily pro

původní data. Vyšetříme proto profily průměrových skóre a označíme popisným nadpisem každý shluk.

6. krok: Validace a profilování shluků: Existuje poněkud subjektivní charakter analýzy shluků stran hledání optimálního shlukového řešení.

Validace shluků: Validace čili potvrzení shluků zahrnuje pokusy zajistit, že shlukové řešení je reprezentativní v celé obecné populaci, a je proto zobecnitelné i na ostatní objekty a dále je i stabilní v čase.

Vzorová úloha 1. *Klasifikace zdrojů pitné vody* (E404 v ref. [31])

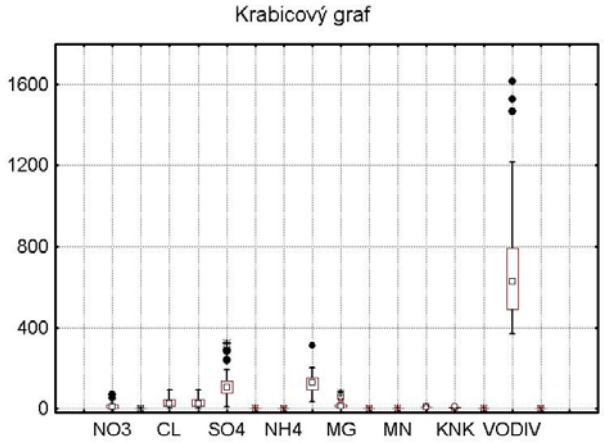
Na 62 vzorcích zdrojů pitné vody bylo stanoveno 16 proměnných kvality vody. Je třeba vyšetřit, zda krabicový graf ukazuje nutnost data standardizovat, zda lze nalézt *vybočující objekty*, resp. jejich proměnné, zda existuje korelace mezi proměnnými, zda ukazuje graf komponentních vah na korelující proměnné, zda jsou některé proměnné redundandní, zda lze odhalit v rozptylovém diagramu komponentního skóre odlehlé objekty, zda lze posoudit *podobnost objektů* shlukovou analýzou klasifikaci zdrojů.

Data: i index vzorku, **NO3** obsah dusičnanů [mg/l], **NO2** obsah dusitanů [mg/l], **Cl** obsah chloridů [mg/l], **Cl2** obsah celkového chloru [mg/l], **SO4** obsah síranů [mg/l], **PO4** obsah fosforečnanů [mg/l], **NH4** obsah amonných solí [mg/l], **Ca** obsah vápníku [mg/l], **Mg** obsah hořčíku [mg/l], **Fe** obsah železa (celkového) [mg/l], **Mn** obsah manganu [mg/l], **pH** je pH roztoku, **KNK**, **ZNK**, **Vodiv.** vodivost roztoku, **Nerozp**. nerozpuštěné látky [mg/l].

i	NO3	NO2	Cl	Cl2	SO4	PO4	NH4	Ca	Mg	Fe	Mn	pH	KNK	ZNK	Vodiv.	Nerozp.
1	2,2	0,00	6	6,00	103,50	0,03	0,02	181	17	0,016	0,050	7,08	8,10	3,40	855	0,09
2	1,9	0,01	18	18,00	127,00	0,03	0,06	161	19	0,016	0,050	7,22	6,00	1,80	706	0,08
3	1,3	0,01	1	1,00	93,50	0,03	0,04	175	14	0,016	0,050	7,16	9,30	2,10	625	0,56
4	1,5	0,02	80	80,00	291,00	0,03	0,02	310	38	0,016	0,050	7,07	9,70	5,05	1465	0,06
5	1,7	0,01	93	93,00	321,00	0,03	0,02	316	51	0,016	0,050	7,13	10,50	4,85	1526	0,03
6	11,8	0,11	51	51,00	127,00	0,03	0,02	129	24	0,016	0,100	7,21	4,70	1,35	752	0,07
7	1,6	0,05	28	28,00	281,00	0,04	0,02	179	65	1,050	0,100	7,21	8,10	2,45	1081	0,27
8	2,5	0,10	11	11,00	49,00	0,19	0,22	83	10	0,016	0,110	7,21	4,60	0,55	437	0,2
9	4,8	0,11	12	12,00	51,00	0,04	0,06	47	22	0,071	0,140	7,74	3,70	0,60	380	0,10
10	8,4	0,04	13	13,00	84,00	0,03	0,02	94	5	0,016	0,050	7,51	3,70	0,55	414	0,32
11	1,0	0,01	7	7,00	150,00	0,03	1,89	159	13	0,016	0,670	7,29	6,60	2,05	686	0,13
12	11,4	0,56	23	23,00	132,00	0,03	0,05	85	13	0,016	0,070	7,32	3,50	0,45	464	0,12
13	9,0	0,11	14	14,00	79,50	0,16	0,02	88	24	0,016	0,070	7,42	5,10	0,75	475	0,13
14	27,7	0,03	57	57,00	168,00	0,06	0,02	200	13	0,016	0,080	7,37	6,60	1,25	920	0,26
15	54,2	0,02	52	52,00	139,00	0,10	0,03	190	7	0,038	0,060	7,36	6,90	1,15	982	0,36
16	69,9	0,04	38	38,00	72,00	0,10	0,44	171	2	0,016	0,050	7,29	6,80	1,70	913	0,13
17	3,5	0,08	13	13,00	110,50	0,15	0,02	88	12	0,016	0,050	7,41	3,30	0,65	491	0,15
18	1,0	0,02	16	7,16	7,31	0,10	0,20	34	19	0,038	0,170	7,12	9,60	2,90	1614	0,19
19	1,0	0,00	87	87,00	144,00	0,03	0,31	145	45	0,053	0,050	7,62	7,80	1,10	924	0,15
20	13,2	0,01	15	15,00	50,00	0,14	0,06	94	5	0,016	0,060	7,72	4,80	0,70	491	0,17
21	1,0	0,04	45	45,00	68,00	0,08	0,28	122	3	0,018	3,660	7,19	3,40	0,85	541	0,11
22	18,4	0,02	24	24,00	127,00	0,03	0,02	137	5	0,016	0,090	7,49	5,30	1,40	610	0,10
23	12,8	0,00	15	15,00	124,00	0,03	0,02	120	6	0,016	0,050	7,59	4,50	0,55	518	0,18
24	19,9	0,03	43	43,00	187,00	0,03	0,02	183	2	0,025	0,050	7,48	5,10	1,60	882	0,11
25	33,7	0,03	9	9,00	101,00	0,03	0,02	96	6	0,016	0,050	7,65	2,70	0,75	429	0,09
26	1,3	0,05	23	23,00	243,00	0,03	0,23	161	13	0,064	1,050	7,18	2,80	1,10	686	0,12
27	1,0	0,00	16	16,00	137,00	0,03	0,02	147	7	0,016	0,260	7,51	5,70	1,30	648	0,10
28	2,1	0,03	16	16,00	130,00	0,03	0,04	155	6	0,016	0,320	7,53	5,70	1,05	660	0,10
29	1,0	0,02	7	7,00	163,00	0,03	0,78	149	9	0,025	3,050	7,31	6,10	2,20	637	0,09
30	1,0	0,00	4	4,00	72,00	0,03	0,21	88	9	0,016	0,180	7,62	4,30	0,55	414	0,11
31	1,3	0,00	57	57,00	116,00	0,04	0,13	130	58	0,066	0,050	7,48	10,90	2,05	908	0,21
32	1,0	0,00	30	30,00	48,00	0,39	0,02	168	5	0,054	0,050	8,03	7,60	1,75	672	0,09
33	1,4	0,00	46	46,00	94,00	0,03	0,02	86	54	0,066	0,050	7,79	5,50	1,10	711	0,23
34	22,6	0,00	53	53,00	58,00	0,03	0,02	105	58	0,061	0,050	7,55	6,50	1,85	729	0,11
35	21,2	0,00	45	45,00	106,00	0,03	0,02	100	72	0,064	0,050	7,72	6,80	1,40	790	0,09

36	30,2	0,00	59	59,00	238,00	0,03	0,02	154	83	0,084	0,050	7,70	6,30	1,30	1101	0,09
37	39,0	0,00	22	22,00	90,00	0,04	0,02	88	11	0,061	0,050	7,72	3,80	1,05	432	0,09
38	11,0	0,00	11	11,00	82,00	0,03	0,02	90	22	0,056	0,050	7,71	7,00	1,60	629	0,05
39	25,0	0,00	20	20,00	96,50	0,03	0,03	90	35	0,016	0,050	7,65	5,80	1,15	551	0,15
40	2,2	0,00	34	34,00	82,00	0,03	0,10	130	12	0,016	0,050	7,84	5,70	1,15	575	0,08
41	33,6	0,01	45	45,00	115,00	0,03	0,02	160	6	0,028	0,120	7,77	4,10	1,05	679	0,07
42	46,5	0,01	11	11,00	81,50	0,54	0,10	80	9	0,016	0,050	7,73	3,10	0,75	432	0,06
43	39,3	0,04	28	28,00	130,50	0,03	0,19	144	15	0,031	0,080	7,68	5,50	1,15	615	0,11
44	14,7	0,00	29	29,00	92,50	0,03	0,02	122	24	0,084	0,050	7,78	6,10	1,05	575	0,07
45	17,6	0,00	33	33,00	132,00	0,03	0,02	128	18	0,036	0,070	7,81	4,10	0,60	551	0,04
46	42,8	0,15	25	25,00	149,00	0,03	0,02	152	6	0,076	0,120	7,64	3,80	0,95	665	0,05
47	18,4	0,27	15	15,00	45,50	0,03	0,02	70	5	0,089	0,080	7,50	3,20	0,95	369	0,07
48	7,4	0,01	14	14,00	195,00	0,03	0,02	148	9	0,051	0,130	7,46	3,90	1,30	657	0,08
49	22,1	0,19	52	52,00	180,00	0,03	0,02	154	15	0,048	0,060	7,73	5,60	1,60	803	0,07
50	1,9	0,00	11	11,00	58,00	0,03	0,02	146	11	0,028	0,050	7,84	4,30	1,00	425	0,08
51	18,3	0,00	7	7,00	65,50	0,03	0,02	98	12	0,043	0,050	7,59	4,30	1,25	398	0,07
52	20,8	0,29	36	36,00	65,50	0,03	0,03	120	18	0,084	0,050	7,50	5,70	1,70	592	0,07
53	9,0	0,00	24	24,00	125,00	0,03	0,02	122	7	0,079	0,050	7,77	4,40	1,50	511	0,07
54	19,4	0,00	22	22,00	108,00	0,03	0,02	120	5	0,041	0,050	7,95	5,10	1,65	548	0,03
55	1,8	0,00	3	3,00	92,00	0,03	0,02	118	15	0,056	0,050	7,84	6,40	1,90	504	0,06
56	31,1	0,01	1	1,00	125,00	0,03	0,02	138	24	0,097	0,050	7,71	6,20	1,60	628	0,06
57	21,9	0,83	31	31,00	76,00	1,74	0,02	104	4	0,102	0,090	6,67	5,70	1,55	613	0,06
58	6,9	0,31	36	36,00	50,50	4,42	0,02	120	7	0,079	0,220	7,69	6,20	3,10	635	0,03
59	13,9	0,15	60	60,00	81,00	0,11	0,02	151	9	0,066	0,050	7,96	5,40	0,85	1217	0,66
60	7,4	0,17	56	56,00	151,00	0,18	0,02	149	19	0,016	0,180	7,61	5,60	0,85	790	0,13
61	1,0	0,07	29	29,00	68,50	0,16	0,02	81	11	0,041	0,050	7,62	4,50	0,75	441	0
62	32,8	0,01	25	25,00	115,50	0,05	0,02	102	12	0,016	0,050	7,69	2,60	0,65	436	0,05

Řešení: Na začátku každé vícerozměrné analýzy dat je třeba provést exploratorní analýzu a vyšetřit znaky co do jejich proměnlivosti a vzájemné korelace. Krabicový graf všech znaků na obr. 6 přehledně zobrazuje proměnlivost jednotlivých znaků. Je zřejmé, že data bude třeba předem standardizovat, protože se vyskytují v rozličných rozměrech i řádově se lišících hodnotách. Platí zde totiž pravidlo, že *čím větší proměnlivost znaku, tím více znak přispívá k formulaci hlavních komponent*.



Obr. 6 Krabicový graf proměnlivosti znaků.

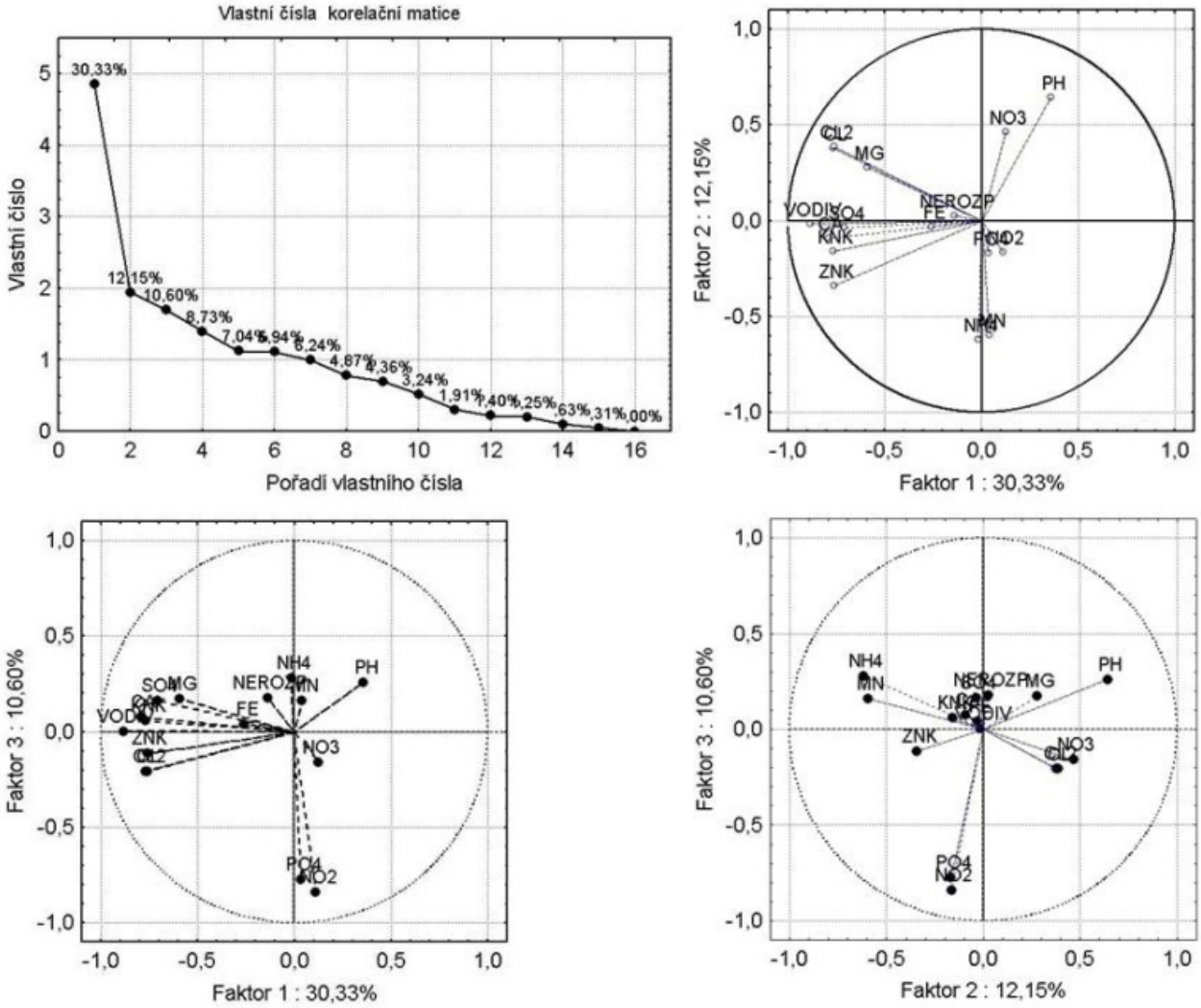
Maticový diagram ukazuje rozptylové diagramy dvojic znaků. Je zřejmé, že vysoké hodnoty korelačního koeficientu *R* vedou ke zřetelné lineární (přímkové) závislosti, zatímco nízké hodnoty *R* ukazují, že znaky nejsou v diagramu zobrazeny na přímce ale spíše v chaotickém mraků bodů.

Metoda hlavních komponent PCA

Na začátku analýzy metodou hlavních komponent je třeba určit vhodný počet hlavních komponent, který co nejlépe popisuje proměnlivost v datech.

Vyšetření indexového grafu úpatí vlastních čísel: Počet užitečných komponent je v Cattelově indexovém grafu na obr. 7a oddělen zřetelným zlomovým místem, a *x*-ová souřadnice tohoto zlomu je pak hledaná hodnota tohoto počtu a zde je rovna 2.

Vyšetření grafu komponentních vah: Vyšetření se provede porovnáním vzdáleností mezi proměnnými a dospěje se k závěru, že krátká vzdálenost mezi dvěma proměnnými znamená silnou korelaci. (Obr. 7 b,c,d)



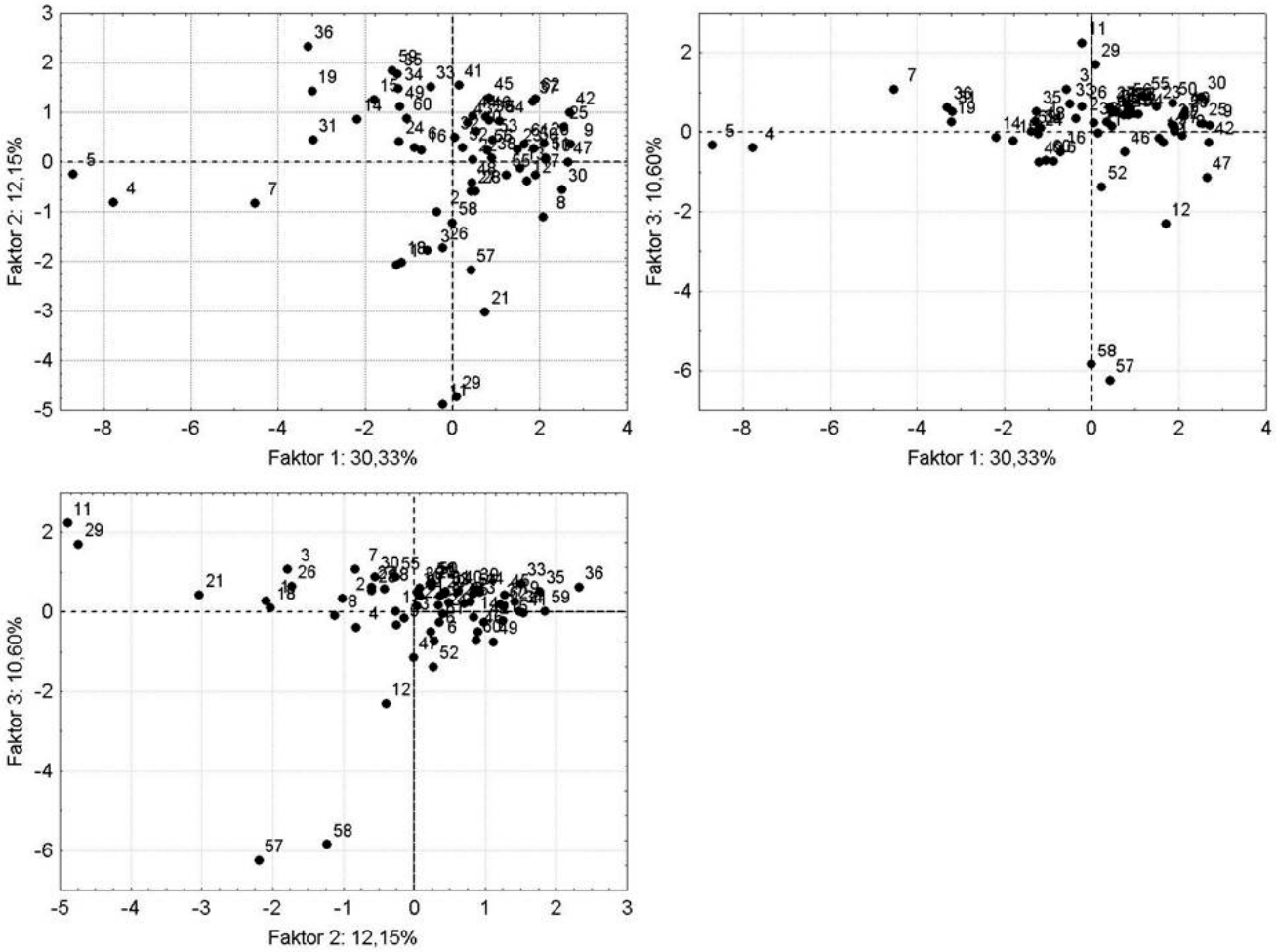
Obr. 7 (a) Cattelův indexový graf úpatí vlastních čísel pro 62 zdrojů vody a 16 znaků, (b) Graf komponentních vah 1. a 2. hlavní komponenty, (c) 1. a 3. hlavní komponenty, (d) 2. a 3. hlavní komponenty,

Grafy ukazují, jakou měrou přispívají jednotlivé původní proměnné do hlavních komponent. Lze názorně vysvětlit, jak jednotlivé původní proměnné x_j, j= 1,..., 16, přispívají do první hlavní komponenty y_1 nebo do druhé hlavní komponenty y_2. Některé původní proměnné x_j přispívají kladnou vahou, jiné zápornou. Původní proměnné x_j, j= 1,..., 16, blízko sebe a nebo proměnné x_j s malým úhlem mezi svými průvodiči proměnných a na stejné straně vůči počátku mají vysokou kladnou kovarianci, a tím také vysokou kladnou korelaci. Naopak, původní proměnné x_j daleko od sebe anebo s velikým úhlem mezi průvodiči proměnných jsou negativně korelovány. Ukazuje se užitečné, aby původní proměnné, které spolu silně korelují byly z dat odstraněny, aby se tak snížil počet původních znaků..

Vyšetření rozptylového diagramu komponentního skóre:

Nejdůležitější diagram metody hlavních komponent ukazuje celou vyšetřovanou strukturu objektů, tzn. shluky objektů, izolované objekty, odlehlé objekty, anomálie, atd. (obr. 8a, b, c). Objekty mohou být označeny textovým popisem nebo jako na obr. 5 číselně indexem. Analýzou lze dospět k těmto závěrům:

1. *Umístění objektů*: objekty daleko od počátku (4, 5, 7, 11, 12, 29, 58, 59, atd.) jsou extrémy. Objekty nejblíže počátku (22, 52, 53, 54, 45, atd.) jsou typičtější.
2. *Podobnost objektů*: objekty blízko sebe (např. 53 a 54 a 44 a 45) si jsou podobné, objekty daleko od sebe (např. 5 a 45, 4 a 30, atd) jsou si nepodobné.
3. *Objekty v shluku*: objekty umístěné zřetelně v jednom shluku (např. 12, 13, 17, 55, 23, 61, 50, 20, atd) jsou si podobné a přitom nepodobné objektům v ostatních shlucích (např. 14, 15, 24, 49, atd.). 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 (4, 5, 7) 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 (4, 5, 11, 29, 57 a 58). 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 vypovídají o existujících shlucích.

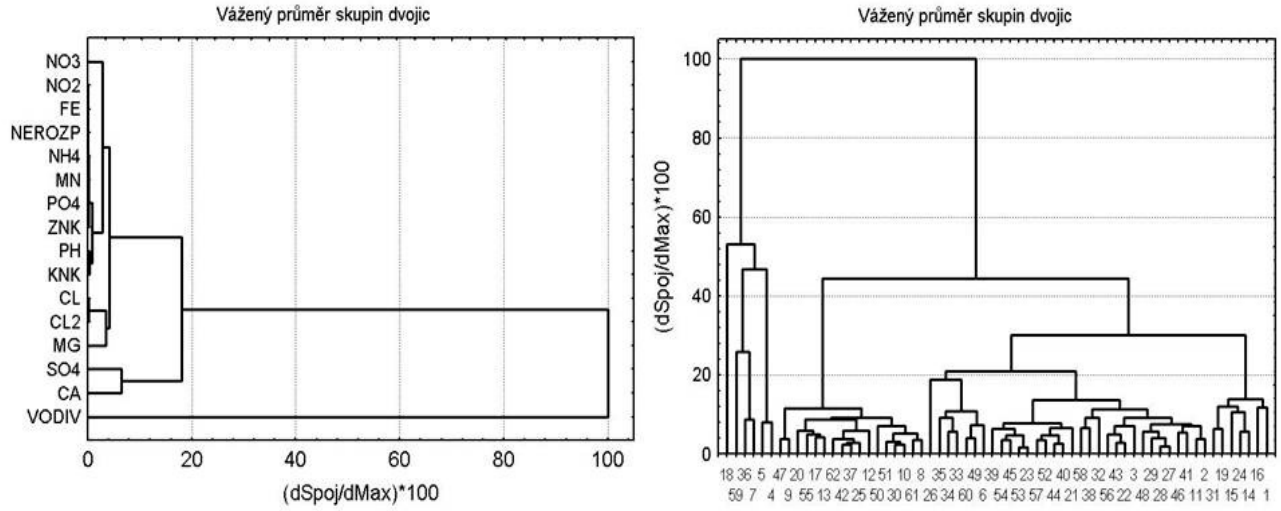


Obr. 8 (a) Rozptylový diagram komponentního skóre 1. a 2. hlavní komponenty, (b) 1. a 3. hlavní komponenty, (c) 2. a 3. hlavní komponenty.

KLASIFIKACE METODOU SHLUKů

V prvém kroku se postaví dendrogram znaků a v druhém kroku dendrogram objektů.

Dendrogram podobnosti znaků: V prvém stádiu se vyšetřuje podobnost proměnných, a tím se také odhalí jejich silná korelace. Vyšetří se, která původní proměnná (znak) je nadbytečná a kterou lze vynechat a nahradit jinou. Čím nižší je spojka dvou objektů, tím jsou si objekty podobnější.



Obr. 9 (a) Dendrogram proměnných (znaků) metodou váženého průměru skupin dvojic, (b) Dendrogram objektů (zdrojů vody) metodou váženého průměru skupin dvojic.

Metodou váženého průměru skupin dvojic nebo metodou nejbližšího souseda lze nalézt na obr 9a řadu dvojic velice si podobných znaků. Nejvíce jsou si podobné znaky v následujícím shluku: *NO2-FE-NEROZP-NH4-MN-PO4*. Dále pak ve shluku *KNK-pH-ZNK* a konečně *SO4-CA*. Poněkud ménějsou si podobné proměnné ve shluku *NO3* vůči *KNK-pH-ZNK*. Výjimečné postavení má proměnná *VODIV*, která si není podobná s žádnou jinou proměnnou. Metoda odhalila řadu shluků, dvojic podobných proměnných. Dospěla k podobným závěrům jako metoda nejbližšího souseda nebo metoda Wardova.

Dendrogram podobnosti objektů: V druhém stádiu klasifikace tvorbou shluků se vyšetřuje podobnost objektů čili zdrojů pitné vody, jež patří vlastně k nejdůležitější část klasifikační analýzy. Jde o odhalení vybočujících zdrojů, které jsou silně nepodobné ostatním, které mají anomální hodnoty znaků. Z obr. 6b je zřejmé, že zdroje pitné vody lze na základě 16 znaků rozdělit do šesti zřetelně odlišných shluků.

Závěry: Standardním využitím PCA je snížení dimenze úlohy čili redukce počtu znaků bez velké ztráty informace, a to užitím pouze prvních několika hlavních komponent. Toto 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. Dále lze vybrané hlavní komponenty využít také k testu vícerozměrné normality. Analýza hlavních komponent je rovněž součástí průzkumové analýzy dat. Snížení rozměrnosti je často využíváno při konstrukci komplexních ukazatelů jako lineárních kombinací původních znaků. Například první hlavní komponenta je vlastně vhodným ukazatelem jakosti, pokud původní znaky charakterizují její složky. Využití první hlavní komponenty jako komplexního ukazatele je běžné v oblasti ekonomie, sociologie a medicíny. 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. Oblíbené je také použití hlavních komponent v oblasti řízení jakosti.

Doporučená literatura

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

[2] Kendall M. G., Stuart A.: *The Advanced Theory of Statistics,* Vol. III. New York 1966.

[3] James W., Stein C.: *Estimation with Quadratic Loss,* Proceed. 4th Berkeley Symp. on Math. Statist., p. 361, 1961.

[4] Guanadeskian R., Kettenring J. R.: *Biometrics* **28**, 80 (1972).

[5] Campbell N. A.: *Appl. Statist.*, 29, 231 (1980).

[6] Hu J., Skrabal P., Zollinger H.: *Dyes and Pigments*, **8**, 189 (1987).

[7] Chambers J. M., Cleveland W. S., Kleiner B., Tukey P. A.: *Graphical Methods for Data Analysis*. Duxburg Press, Belmont, California 1983.

[8] Barnett V., (Edit.): *Interpreting Multivariate Data*. Wiley, Chichester 1981, kap. 6.

[9] Jolliffe I. T.: *Principal Component Analysis*. Springer Verlag, New York 1986.

[10] Barnett V., (Edit.): *Interpreting Multivariate Data*. Wiley, Chichester 1981, kap. 12.

[11] Everitt B. S.: *Graphical Techniques for Multivariate Data*. London 1978.

[12] Andrews D. F.: *Biometrics*, **28**, 125 (1972).

[13] Kulkarni S. R., Paranjape S. R.: *Commun. Statist.*, **13**, 2511 (1984).

[14] Guanadeskian R.: *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, New York 1977.

[15] Kleiner B., Hartigan J. A., *J. Amer. Statist. Assoc.*, **76**, 260 (1981).

[16] Kres H.: *Statistical Tables for Multivariate Analysis*. Springer, New York 1983.

[17] Seber G. A. F.: *Multivariate Observations*. Wiley, New York 1984.

[18] Stryjewska E., Rubel S., Henrion A., Henrion G.: *Z. Anal. Chem.*, **327**, 679 (1987).

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

[20] Johnson R.A., Wichern D.W.: *Applied Multivariate Statistical Analysis,* Prentice Hall, 1982

[21] Ajvjazin S., Bežajeva Z., Staroverov O.: *Metody vícerozměrné analýzy,* SNTL Praha 1981

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

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

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

[25] Jeffers J. N. R., *Applied Statistician*, **16**, 225 (1967).

[26] Meloun M. , Militký J., *Statistické zpracování experimentálních dat,* Plus Praha 1994, Academia Praha 2004.

[27] Martens H., Naes T., *Multivariate calibration*, Wiley (1989) Chichester.

[28] Thomas E. V., *Anal. Chem.*, **66** (1994) 795A-804A.

[29] Malinowski F., Howery D., *Factor Analysis in Chemistry*, Wiley (1980) New York.

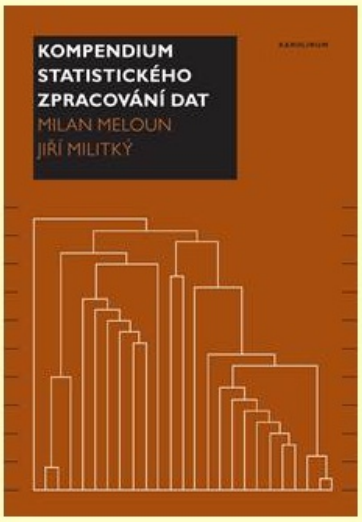
[30] Meloun M. , Militký J., *Sbírka úloh - Statistické zpracování experimentálních dat,* Univerzita Pardubice, 1996.

[31] Meloun M. , Militký J., *Kompendium statistického zpracování dat,* Academia Praha 2002, Academia Praha 2006.

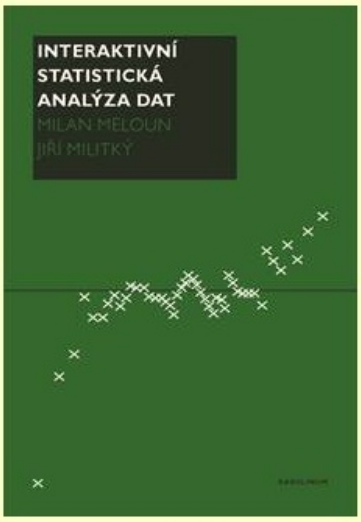
Statistické zpracování experimentálních dat

Academia Praha, Karolinum Praha

2012



2012

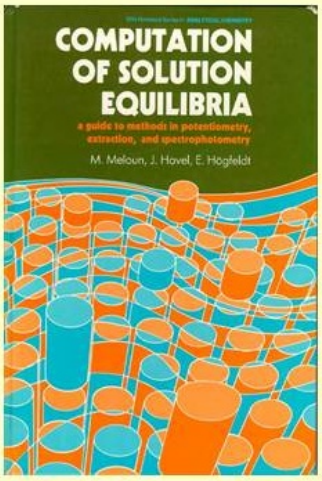


2012

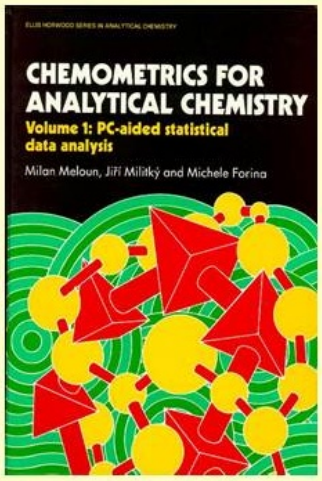
Statistické zpracování experimentálních dat

Učebnice v angličtině

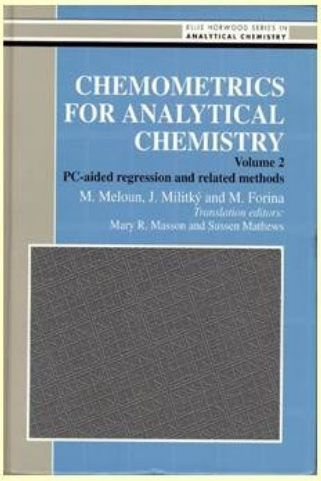
Ellis Horwood Chichester, John Wiley & Son, New York,
Woodhead Publishing WPI New Delhi



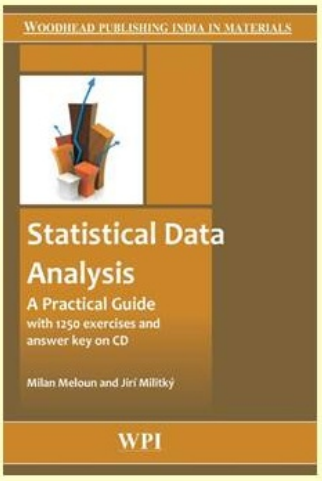
1988



1992



1994



2011