Statistická analýza vícerozměrných dat

Prof. RNDr. Milan Meloun, DrSc.,
Katedra analytické chemie, Univerzita Pardubice, 532 10 Pardubice
a
Prof. Ing. Jiří Militký, CSc.,
Katedra textilních materiálů, Technická univerzita Liberec, 461 17 Liberec

Souhrn: *Vícerozměrná statistická analýza je založena na latentních proměnných, které jsou lineární kombinací původních proměnných,* $y = w_1 x_1 + \ldots + w_m x_m$. *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 (vzdálenost Euklidovská, Manhattanská, Minkovského a Mahalanobisova), párového korelačního koeficientu a koeficientu asociace (Sokalův-Michelenerův, Russelův-Raoův a Hamanův): čím je vzdálenost shluků či objektů větší, tím menší je jejich podobnost. K rychlému posouzení podobnosti slouží grafy exploratorní analýzy vícerozměrných dat: profily, polygony, tváře, křivky, stromy, sluníčka a hvězdičky. Strukturu a vazby mezi proměnnými vystihují metody snížení dimensionality, metoda hlavních komponent (PCA) a metoda faktorové analýzy. 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-představitelů. Výsledkem je dendrogram.*

V technické praxi se vedle informací, obsažených v náhodném skaláru ξ, vyskytují i vícerozměrné informace, obsažené v náhodném vektoru ξ s m složkami $\xi_1, \ldots, \xi_m$. Příklady vícerozměrných informací jsou

a) vyjádření vlastností produktů jako jsou potraviny, oleje, slitiny, atd. pomocí řady různých analytických metod,

b) hodnocení spekter pomocí poloh a ploch absorpčních pásů sloužící k charakterizaci a identifikaci chemických sloučenin,

c) sledování složení surovin, produktů, odpadů, v závislosti na čase nebo na místě výskytu,

d) regulace jakosti na základě různých procesních proměnných,

e) stanovení charakteristiky produktu na základě měření souvisejících proměnných, např. spekter (vícerozměrná kalibrace).

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

$y = w_1x_1 + w_2x_2 + ... + w_mx_m$. Jednotlivé vícerozměrné metody využívají různých způsobů stanovení vah $w_1, w_2, ..., w_m$.

Zdrojová matice, tj. matice výchozích dat (popisující např. řadu aut) obsahuje **proměnné** v m sloupcích (např. obsah motoru, výkon, spotřeba paliva, hmotnost vozu, zrychlení, výška, šířka, délka, atd.) a **objekty** v n řádcích (např. auta různých výrobců), na nichž jsou tyto proměnné (vlastnosti) měřeny. Protože měřené proměnné mají různé jednotky, a často se řádově liší, bývá zdrojová matice před zpracováním ještě upravována, *škálována*, a to buď (a) *centrováním*, kdy se od prvků sloupce odečte jejich sloupcový aritmetický průměr, nebo (b) *standardizací* čili *normováním*, kdy se prvky centrovaných sloupců vydělí svou sloupcovou směrodatnou odchylkou.

Statistická analýza je založena na předpokladu, že hodnoty x_{ij} tvoří *náhodný výběr.* Tento výběr je tvořen n-ticí vektorů $\boldsymbol{x}_i^T = (x_{i,1}, ..., x_{i,m})$, které lze chápat jako řádky zdrojové matice nebo souřadnice n bodů v m-rozměrném prostoru proměnných. Tento výběr lze pak vyjádřit maticí rozměru $(n \times m)$

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_i^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} = \begin{bmatrix} X_{1,1} & \cdots & X_{1,j} & \cdots & X_{1,m} \\ \vdots & & \vdots & & \vdots \\ X_{i,1} & \cdots & X_{i,j} & \cdots & X_{i,m} \\ \vdots & & \vdots & & \vdots \\ X_{n,1} & \cdots & X_{n,j} & \cdots & X_{n,m} \end{bmatrix}$$

Řádek zdrojové matice čili i-tý vektor $\boldsymbol{x}_i^T = (x_{i,1}, ..., x_{i,m})$ nazýváme *objektem* (např. auto určitého typu a modelu) a můžeme ho chápat jako jeden bod v m-rozměrném prostoru. Tento objekt je charakterizován *proměnnými*, a to buď *kvantitativními,* metrickými, tj. číselnými hodnotami, nebo proměnnými *kvalitativními*, nemetrickými.

Metrické proměnné se vyskytují ve čtyřech škálách:

(a) *Proměnné v absolutní škále* mají na škále přirozený počátek a jediné měřítko, např. obsah uhlíku v %, rychlostní konstanta.

(b) *Proměnné v poměrové škále* mají zachován podíl hodnot charakteristik $c = x_2/x_1$, např. vztah vůči standardní sloučenině, vztah vůči jevu s definovaným nulovým počátkem, parametr σ v Hammettově rovnici.

(c) *Proměnné v intervalové škále* mají zachován podíl rozdílů $c = x_2 - x_1$. Jedná se o poměrovou škálu s přirozeným počátkem pro obě srovnávané hodnoty, např. Poměr

absorbancí indikátoru, vztažený na absorbanci nulové linie.

(d) *Proměnné v rozdílové škále* jsou vztahovány k různému počátku, např. Hodnoty časových škál, stáří, atd.

Nemetrické proměnné se vyskytují ve dvou škálách:

(a) *Proměnné v ordinální škále* mají svou hodnotu danou pořadím v neklesající posloupnosti proměnných dle nějakého kritéria, např. Počet atomů chloru v molekule, žebříček umístění, pořadové číslo.

(b) *Proměnné v nominální škále* jsou nejméně informativní. Obsahují kód, např. barvu kódem 1 až 16, rodinný stav (svobodný 1, ženatý 2, rozvedený 3, vdovec 4).

(c) *Proměnné v alternativní (binární) škále* vyjadřují rovnost či nerovnost vůči nějakému kritériu. Mají binární charakter, relaci můžeme popsat dvojicí 1 (ano), 0 (ne).

Třídu nebo ***shluk*** chápeme jako množinu objektů se společnými nebo alespoň blízkými proměnnými, znaky (např. auta typu BMW). Blízkost či podobnost objektů posuzujeme na základě *míry blízkosti* či *vzdálenosti objektů* v m-rozměrném prostoru proměnných. Vyjádření vzdálenosti objektů pro *kvantitativní* proměnné jsou

Euklidova metrika čili *geometrická vzdálenost* je nejjednodušší typ vzdálenosti, definovaný vztahem

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

Hammingova metrika čili *Manhattanská vzdálenost*, definovaná vztahem

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

Zobecněná Minkowskiho metrika, definovaná vztahem

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

kde pro $n = 1$ jde o Hammingovu metriku a pro $n = 2$ o Euklidovu. Čím je n větší, tím více je zdůrazňována vzdálenost mezi blízkými objekty. Všechny tyto metriky předpokládají nezávislost mezi proměnnými. Zahrneme-li však do vztahu pro vzdálenost i vazby mezi proměnnými, vyjádřené kovarianční maticí C dostaneme novou míru, zvanou *Mahalanobisova metrika*

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

Ta se společně s Euklidovou metrikou nejvíce používá v praxi. Ve všech uvedených případech jsou si dva objekty tím bližší, čím je jejich vzdálenost menší.

Mírou podobnosti dvou objektů či proměnných x_i a x_j může být *párový korelační koeficient r*. Objekty jsou si tím podobnější, čím je párový korelační koeficient větší. V

případě ordinální škály je analogickou mírou podobností *Spearmanův korelační koeficient.* Podobnost binárních nebo nominálních proměnných vyjadřují různé koeficienty asociace. Označíme-li počet případů negativní shody typu 0-0 písmenem a, počet případů s neshodou typu 1-0 písmenem b, počet případů s neshodou typu 0-1 písmenem c a počet případů s pozitivní shodou typu 1-1 písmenem d, dojdeme ke následujícím vzorcům koeficientů podobnosti:

(a) *Sokalův-Michelenerův koeficient asociace*

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

(b) *Russelův-Raoův koeficient asociace*

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

(c) *Hamannův koeficient asociace*

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

a také lze konstruovat *obdobu korelačního koeficientu*

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

Míra podobnosti mezi objekty, charakterizovanými různými typy proměnných se vypočte jako vážený průměr jednotlivých měr podobnosti.

Na základě měr podobnosti objektů se konstruují míry podobnosti mezi objekty a třídami a míry podobnosti mezi třídami. Jako nejčastější míra podobnosti se používá vzdálenost tříd $d(x_k, x_l)$. Analogicky zde užijeme způsobů vyjádření vzdálenosti objektů, protože objekt můžeme chápat jako třídu o jednom objektu. Čím větší je vzdálenost, tím menší je podobnost:

(a) *Vzdálenost nejbližšího souseda:* nejbližší jsou ty třídy, které mají nejmenší vzdálenost mezi dvěma nejbližšími objekty dvou pozorovaných tříd.

(b) *Vzdálenost nejvzdálenějšího souseda:* nejbližší jsou ty třídy, které mají nejmenší vzdálenost mezi dvěma nejvzdálenějšími objekty.

(c) *Vzdálenost mezi těžišti tříd:* nejbližší jsou ty třídy, které mají nejmenší vzdálenost mezi svými těžišti.

(d) *Vzdálenost průměrné vazby:* nejbližší jsou ty třídy, které mají nejmenší průměrnou vzdálenost mezi všemi objekty jedné a všemi objekty druhé třídy.

1. Postup analýzy vícerozměrných dat

Postup analýzy vícerozměrných dat záleží na typu dat a na druhu požadované informace

jež se z dat má získat. Data by však měla být již shromaždována s ohledem na získání požadovaných informací.

Typ dat

Otázky: Před vlastní analýzou je proto třeba zodpovědět tři základní otázky:

(1) Je možné rozdělit vyšetřované proměnné na *závislé* a *nezávislé*?

(2) Kolik proměnných se uvažuje jako *závisle proměnných*?

(3) V jaké škále jsou jednotlivé proměnné měřeny, tj. *kardinální* čili číselné, *ordinální* čili pořadové, nebo *nominální* čili znakové. Kardinální škála se označuje jako *metrická* a ostatní dvě, ordinální a nominální jako škály *nemetrické*.

Odpovědi:

(1) Pokud je odpověď na první otázku kladná, volí se techniky pro stanovení *vztahu* mezi závisle proměnnými a vhodnou kombinací nezávisle proměnných.

(2) Pokud je odpověď na první otázku záporná, volí se techniky pro stanovení *vzájemných vazeb* tj. provádí se simultánní analýza všech proměnných.

Typ informace

Jednotlivé techniky pro *stanovení závislosti* se dále dělí podle počtu závisle proměnných a podle typu či škály měření. Klasická vícerozměrná regrese je případem jedné závisle proměnné v metrické škále. Schematicky lze vztahy mezi jednotlivými technikami analýzy vícerozměrné závislosti zapsat ve formě těchto přiřazení:

(a) **Kanonická korelace**:

$$\underset{\text{(metrická, nemetrická)}}{y_1 + y_2 + \ldots + y_m} \Leftarrow \underset{\text{(metrická, nemetrická)}}{x_1 + x_2 + \ldots + x_m}$$

(b) **Vícerozměrná analýza rozptylu**:

$$\underset{\text{(metrická)}}{y_1 + y_2 + \ldots + y_m} \Leftarrow \underset{\text{(nemetrická)}}{x_1 + x_2 + \ldots + x_m}$$

(c) **Analýza rozptylu**: $\underset{\text{(metrická)}}{y_1} \Leftarrow \underset{\text{(nemetrická)}}{x_1 + x_2 + \ldots + x_m}$

(d) **Diskriminační analýza**:

$$\underset{\text{(nemetrická)}}{y_1} \Leftarrow \underset{\text{(metrická)}}{x_1 + x_2 + \ldots + x_m}$$

(e) **Vícerozměrná regrese a kalibrace**:

$$\underset{\text{(metrická)}}{y_1} \Leftarrow \underset{\text{(metrická, nemetrická)}}{x_1 + x_2 + \ldots + x_m}$$

(f) **"Conjoint" analýza**:

$$\underset{\text{(metrická, nemetrická)}}{y_1} \Leftarrow \underset{\text{(nemetrická)}}{x_1 + x_2 + \ldots + x_m}$$

(g) **Strukturní rovnice:**

$$
\begin{array}{ccc}
y_1 & \Leftarrow & x_{11} + x_{12} + \ldots + x_{1m} \\
y_2 & \Leftarrow & x_{21} + x_{22} + \ldots + x_{2m} \\
 & & \ldots\ldots \\
y_n & \Leftarrow & x_{n1} + x_{n2} + \ldots + x_{nm} \\
\text{(metrická)} & & \text{(metrická, nemetrická)}
\end{array}
$$

Uvedené schema umožňuje výběr konkrétní analýzu dat s ohledem na cíl analýzy a počet a typ závisle resp. nezávisle proměnných.

2. Určení struktury a vazeb mezi proměnnými a nebo mezi objekty

Zdrojová matice má rozměr $n \times m$. Data v nemetrické škále lze kódovat s využitím i umělých (*dummy*) proměnných, nabývajících např. hodnot 1 (přítomnost nominálního znaku) nebo 0 (nepřítomnost nominálního znaku). To umožňuje "rozšíření" faktorové a shlukové analýzy o data v nemetrické škále. Před vlastní vícerozměrnou statistickou analýzou je třeba provést *exploratorní (průzkumovou) analýzu dat*, která umožňuje

(a) posoudit *podobnost objektů* pomocí rozptylových diagramů 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 pro stanovení vzájemných vazeb se dále dělí podle toho, zda se hledají struktury v proměnných nebo v objektech:

(1) Hledání struktury *v proměnných* v metrické škále: *faktorová analýza* a *analýza hlavních komponent.*
(2) Hledání struktury *v objektech* v metrické škále: *shluková analýza.*
(3) Hledání struktury *v objektech* v obou škálách: *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.

3. Charakteristiky vícerozměrných náhodných veličin

3.1 Intenzita vztahu mezi proměnnými

K charakterizaci polohy j-té proměnné ξ_j tj. j-tého sloupce matice X se používá **střední hodnota** $E(\xi_j) = \mu_j$ a pro charakterizaci rozptýlení **rozptyl** $D(\xi_j) = \sigma_j^2$. Dále je třeba definovat *míru intenzity* vztahu mezi proměnnými ξ_i a ξ_j, $j = i$. Vhodnou charakteristikou je *druhý smíšený centrální moment*, nazývaný **kovariance** $cov(\xi_i, \xi_j)$, definovaný vztahem

$$cov(\xi_i, \xi_j) = E(\xi_i \, \xi_j) - E(\xi_i) \, E(\xi_j)$$

Kovariance má vlastnosti:

a) Její znaménko ukazuje na typ stochastické vazby mezi j-tým a i-tým sloupcem matice.

b) Je v absolutní hodnotě shora ohraničená součinem $\sigma_i \sigma_j$, tj. $|\, cov(\xi_i, \xi_j)| \leq \sigma_i \sigma_j$.

c) Je symetrickou funkcí svých argumentů.

d) Nemění se posunem počátku, ale změna měřítka se projeví úměrně jeho velikosti. Pro čísla a_1, a_2, b_1, b_2 pak platí, že

$$cov(a_1 \xi_i + b_1, a_2 \xi_j + b_2) = a_1 a_2 cov(\xi_i, \xi_j) .$$

e) Pro nekorelované náhodné veličiny je $cov(\xi_i, \xi_j) = 0$ a mohou nastat dva případy:

1. $E(\xi_i \xi_j) = 0$ a zároveň $E(\xi_i) = E(\xi_j) = 0$, což je případ *centrovaných ortogonálních* náhodných veličin, ne nutně nezávislých.
2. $E(\xi_i \xi_j) = E(\xi_i) E(\xi_j)$, což je případ *nezávislých* náhodných veličin.

f) Je *mírou intenzity lineární závislosti.*

Nevýhodou kovariance je fakt, že její hodnoty závisí na měřítku, ve kterém jsou vyjádřeny proměnné ξ_i a ξ_j. Její velikost lze hodnotit vzhledem k součinu $\sigma_i \sigma_j$. Je proto přirozené provést standardizaci podělením tímto součinem. Vzniklá veličina $\rho_{ij} = \rho(\xi_i, \xi_j)$ se nazývá *párový korelační koeficient*

$$\rho(\xi_i, \xi_j) = \rho_{ij} = \frac{cov(\xi_i, \xi_j)}{\sigma_i \sigma_j}$$

Je zřejmé, že párový korelační koeficient leží v rozmezí $-1 \leq \rho_{ij} \leq 1$. Pokud je $\rho_{ij} > 0$, jde o *pozitivně korelované* náhodné veličiny, a pokud je $\rho_{ij} < 0$, jde o *negativně korelované* náhodné veličiny.

Párový korelační koeficient má vlastnosti:

a) Rovnost $|\rho_{ij}| = 1$ ukazuje, že mezi ξ_i a ξ_j existuje přesně lineární vztah.

b) Pokud jsou náhodné veličiny ξ_i a ξ_j vzájemně nekorelované, je $\rho_{ij} = 0$.

c) V případě, že ξ_i a ξ_j pocházejí z vícerozměrného normálního rozdělení a $\rho_{ij} = 0$, znamená to, že jsou *vzájemně nezávislé.*

d) Platí, že i pro nelineárně závislé náhodné veličiny může být $\rho_{ij} = 0$.

e) Korelační koeficient ρ_{ii} náhodné veličiny ξ_i samotné se sebou je roven jedné.

f) Korelační koeficient je invariantní vůči lineární transformaci náhodných proměnných ξ_i, ξ_j. Pro čísla a_1, a_2, b_1, b_2 platí vztah

$$\rho(a_1 \xi_i + b_1, a_2 \xi_j + b_2) = sign(a_1 a_2)\, \rho(\xi_i, \xi_j)$$

kde $sign(x)$ je znaménková funkce, pro kterou platí

$$sign(x) = \begin{cases} -1 & \text{pro } x < 0 \\ 0 & \text{pro } x = 0 \\ 1 & \text{pro } x > 0 \end{cases} .$$

3.2 Odhady parametrů polohy a rozptýlení

Z vícerozměrného výběru objektů o velikosti n, definovaného n-ticí m-rozměrných objektů $\boldsymbol{x}^T_i = (x_{i,1}, ..., x_{i,m})^T$, $i = 1, ..., n$, je možno stanovit *výběrový vektor středních hodnot* $\hat{\boldsymbol{\mu}}$ určený vztahem

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i^T .$$

Podobně pro *odhad kovarianční matice* $\boldsymbol{S}^0$ platí rovnice

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

Pro vektor výběrových středních hodnot platí

$$E(\hat{\boldsymbol{\mu}}) = \boldsymbol{\mu} \quad \text{a} \quad D(\hat{\boldsymbol{\mu}}) = \frac{1}{n} \boldsymbol{C} .$$

Odhad $\hat{\boldsymbol{\mu}}$ je tedy nevychýlený. Pro odhad kovarianční matice platí, že

$$E(\boldsymbol{S}^0) = \frac{n-1}{n} \boldsymbol{C}$$

a jde o vychýlený odhad. Proto se používá *výběrová korigovaná kovarianční matice*

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

která je již nevychýleným odhadem kovarianční matice $\boldsymbol{C}$. Matice $\boldsymbol{S}^0$ je *výběrová kovarianční matice.* Odhady $\hat{\boldsymbol{\mu}}$ a $\boldsymbol{S}^0$ jsou maximálně věrohodné pro případ, že náhodný výběr, charakterizovaný maticí $\boldsymbol{X}$ pochází z normálního rozdělení $N(\boldsymbol{\mu}, \boldsymbol{C})$. Za stejných podmínek má $\hat{\boldsymbol{\mu}}$ rozdělení $N(\boldsymbol{\mu}, \boldsymbol{C} / n)$.

3.3 Standardizace proměnných

Standardizace čili *normování* znamená škálování proměnné, spočívající v jejím převedení na náhodnou veličinu s jednotkovým rozptylem a nulovou střední hodnotou,

$$\xi_1^* = \frac{\xi_1 - E(\xi_1)}{\sqrt{D(\xi_1)}} \quad \text{a} \quad \xi_2^* = \frac{\xi_2 - E(\xi_2)}{\sqrt{D(\xi_2)}} .$$

Míra polohy náhodného vektoru se charakterizuje pomocí *vektoru středních hodnot* $\boldsymbol{\mu}^T = [E(\xi_1), ..., E(\xi_m)]$, a míra rozptýlení pomocí *kovarianční matice* řádu $m \times m$

$$C = \begin{bmatrix} D(\xi_1) & cov(\xi_1, \xi_2) & \cdots & cov(\xi_1, \xi_i) & \cdots & cov(\xi_1, \xi_m) \\ cov(\xi_1, \xi_2) & D(\xi_2) & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & D(\xi_i) & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ cov(\xi_1, \xi_m) & cov(\xi_2, \xi_m) & \cdots & cov(\xi_i, \xi_m) & \cdots & D(\xi_m) \end{bmatrix} .$$

Místo kovarianční matice můžeme použít také její normovanou verzi, tj. *korelační matici*

$$\boldsymbol{R} = \begin{bmatrix} 1 & \varrho_{12} & \cdots & \varrho_{1i} & \cdots & \varrho_{1m} \\ \varrho_{12} & 1 & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & 1 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \varrho_{1m} & \varrho_{2m} & \cdots & \varrho_{im} & \cdots & 1 \end{bmatrix} .$$

Korelační matice má na diagonále samé jedničky a mimodiagonální prvky jsou jednotlivé *párové korelační koeficienty*. Kovarianční matice $\boldsymbol{C}$ i korelační matice $\boldsymbol{R}$ jsou symetrické. Pokud máme dva vektory, ξ_1 a ξ_2, které jsou nezávislé a stejně rozdělené se střední hodnotou $\boldsymbol{\mu}$ a kovarianční maticí $\boldsymbol{C}$, je *vícerozměrná šikmost* dána vztahem

$$g_{1,m} = E[(\xi_1 - \boldsymbol{\mu})^T \boldsymbol{C}^{-1} (\xi_2 - \boldsymbol{\mu})]^3$$

a pro *vícerozměrnou špičatost* platí

$$g_{2,m} = E[(\xi_1 - \boldsymbol{\mu})^T \boldsymbol{C}^{-1} (\xi_1 - \boldsymbol{\mu})]^2$$

K vyjádření funkcí $g_{1,m}$ a $g_{2,m}$ lze využít i vícerozměrných centrálních momentů. Speciálně pro případ vícerozměrného normálního rozdělení pak platí, že

$$g_{1,m} = 0 \text{ , a } g_{2,m} = m\,(m + 2) .$$

Vzorová úloha

Na úloze *Účinky neuroleptik při tlumení rozličných psychóz* (**B4.02** v ref. [30]) si ukážeme řadu pomůcek vícerozměrné analýzy dat. Neuroleptika se liší nejenom ve svých účincích ale i ve vedlejších účincích. Nalezneme strukturu a vazby v proměnných a objektech. Účelem je provést klasifikaci neuroleptik do shluků podobných účinků jako je např. potlačení nervozity B402X2, potlačení stereotypního chování B402X3, potlačení záchvatu a třesu B402X4, a konečně i velikost dávky smrtícího účinku neuroleptika B402X5. K analýze užijeme také škálovaná data.

Data: převrácená hodnota mediánové účinné dávky 1/ED50 [kg/mg] pro potlačení

nervozity B402X2, pro potlačení stereotypního chování B402X3, pro potlačení záchvatu a třesu B402X4, a smrtící dávka B402X5.

B402X1	B402X2	B402X3	B402X4	B402X5
1 Chlorphromazine	3.846	3.333	1.111	1.923
2 Promazine	0.323	0.213	0.108	1.429
3 Trifluperazine	27.027	17.857	0.562	0.140
4 Fluphenazine	17.857	15.385	1.695	1.075
5 Perphenazine	27.027	27.027	1.961	2.083
6 Thioridazine	0.244	0.185	0.093	1.333
7 Pifluthixol	142.857	142.857	20.408	163.934
8 Thiothixene	4.348	4.348	0.047	0.345
9 Chorprothixene	5.882	2.941	4.545	4.167
10 Spiperone	62.500	47.619	11.765	0.847
11 Haloperidol	52.632	62.500	1.282	0.568
12 Azaperone	2.941	1.282	2.222	3.030
13 Pipamperone	0.327	0.187	1.724	0.397
14 Pimozide	20.408	20.408	0.107	0.025
15 Metitepine	15.385	10.204	10.204	27.027
16 Clozapine	0.161	0.093	0.327	0.323
17 Perlapine	0.323	0.323	0.370	0.067
18 Sulpiride	0.047	0.047	0.003	0.001
19 Butaclamol	10.204	9.091	1.471	0.025
20 Molindone	7.692	7.692	0.140	0.006

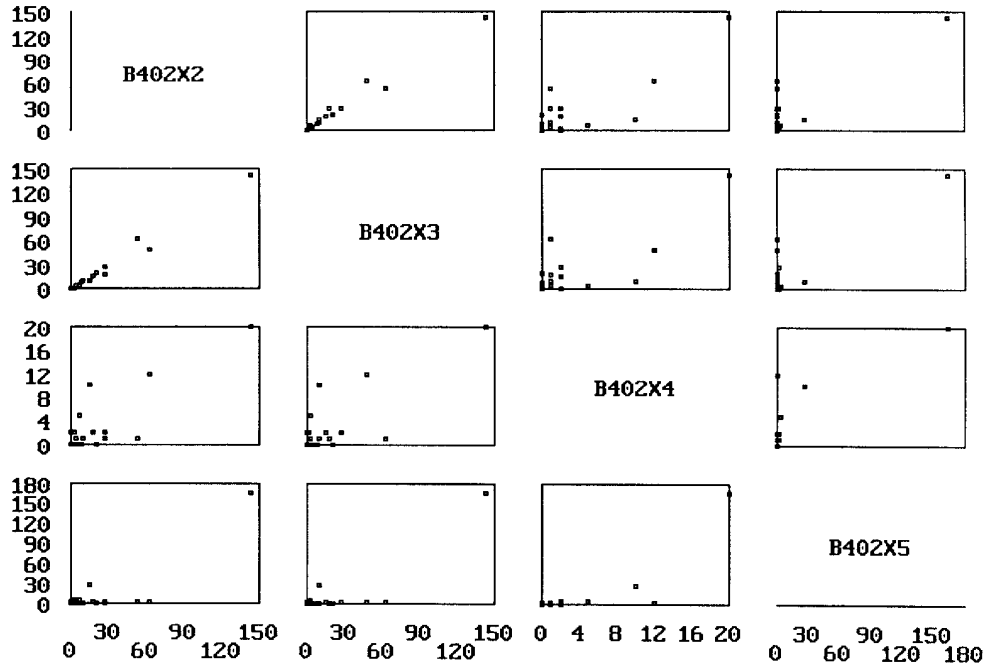
Korelační matice R

	B402X2	B402X3	B402X4	B402X5
B402X2	1.0000	0.9905	0.8359	0.8445
B402X3	0.9905	1.0000	0.7864	0.8518
B402X4	0.8359	0.7864	1.0000	0.8238
B402X5	0.8445	0.8518	0.8238	1.0000

Výklad: korelace mezi dvěma proměnnými vystihuje míru lineární závislosti mezi dvěma proměnnými.

5. Exploratorní analýza podobnosti objektů (EDA)

Průzkumová analýza vícerozměrných dat je stejně jako u jednorozměrných dat založena na grafických diagnostikách. K tomuto účelu se využívá různých technik zobrazování vícerozměrných dat. Pro případ, kdy jsou jednotlivé sloupce matice X málo korelované postačují *rozptylové diagramy* pro jednotlivé kombinace složek vektoru x a pro nekorelované pak sloupce matice X.



Obr. 1 Rozptylový diagram pro 20 objektů a 4 proměnné B402X2, B402X3, B402X4, B402X5 nestandardizovaných dat úlohy B402, STATGRAPHICS.
Výklad: *Z diagramu je patrná vzájemná podobnost objektů a vysoká korelovanost zejména prvních dvou proměnných. Jsou patrné i odlehlé objekty, představované body vzdálenými od ostatních.*

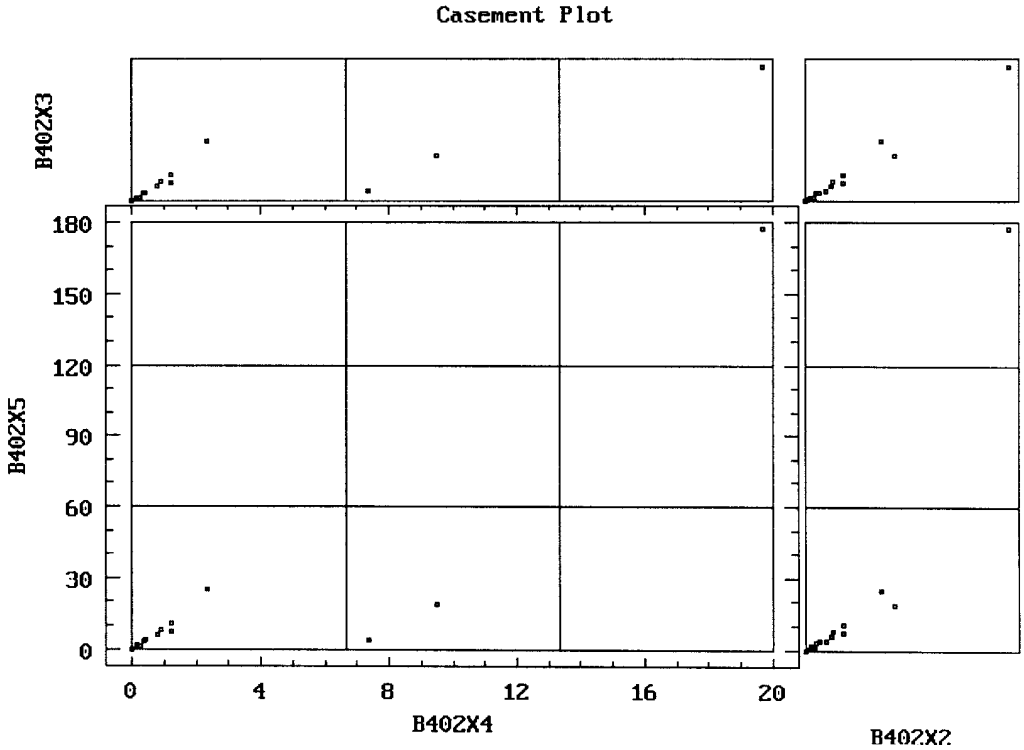
Rychlé posouzení podobnosti mezi jednotlivými objekty čili řádky datové matice usnadňují především *symbolové grafy.* Jednotlivé proměnné jsou v nich "kódovány" s ohledem na jejich konkrétní hodnoty do určitých geometrických tvarů, *symbolů.* Každému objektu x_i (např. autu) tak odpovídá jistý obrazec zvaný symbol. Vlastnosti dat se posuzují s ohledem na vizuální rozdíly mezi symboly. Tím lze v jednom grafu rozlišit 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*.

1. **Profily** představují dvourozměrné zobrazení m-rozměrných objektů. Každý objekt x_i je charakterizován m proměnnými, zobrazenými zde vertikálními úsečkami. Jejich velikost je úměrná hodnotě odpovídající proměnné x_{ij}, $j = 1, ..., m$. Profil pak vzniká spojením koncových bodů těchto úseček. Je vhodné použít standardizované proměnné,

$$x_{ij}^* = x_{ij}/(\max_i |x_{ij}|),$$

kde max $| x_{ij} |$ je maximální hodnota absolutní velikosti proměnné x_j vektoru x_i^T přes všechny body, $i = 1, ..., n$. Profily jsou jednoduché a umožňují snadné určení rozdílů

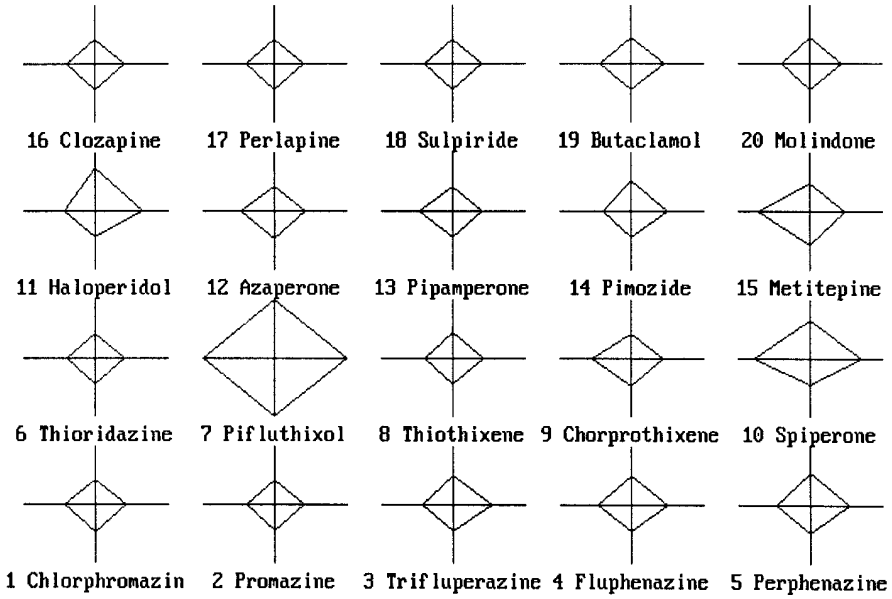
mezi jednotlivými objekty x_i a x_j. Snadno lze takto identifikovat vybočující objekt.



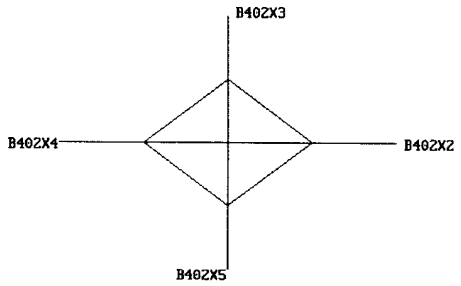
Obr. 2 Korelační diagram pro 20 objektů a 4 proměnné B402X2, B402X3, B402X4, B402X5 nestandardizovaných dat.
Výklad: *Z diagramu je patrná vysoká korelovanost čtyř sledovaných proměnných. V pravém horním rohu jsou patrné odlehlé objekty.*

2. **Polygony** jsou vlastně profily v polárních souřadnicích, kdy každá proměnná objektu x_i^T odpovídá délce paprsku vycházejícího ze společného středu. Paprsky dělí kružnici ekvidistantně, proměnné jsou standardizovány do intervalu [0, 1]. Mezi polygony patří *graf slunečních paprsků* a *hvězdicový graf.*

(a) **Graf slunečních paprsků** má tvar "hvězdice", která sestává z paprsků, začínajících ve společném bodě a spojujících úseček mezi paprsky, které tak tvoří polygon. Zde každá proměnná x_{ij} objektu x_i^T odpovídá délce paprsku vycházejícího ze středu hvězdice.



Obr. 3a Polygony: graf slunečních paprsků pro 20 objektů a 4 proměnné B402X2, B402X3, B402X4, B402X5 standardizovaných dat.
Výklad: *řada objektů je velice podobných vlastností, protože jsou si jejich sluníčka podobná. Z množiny 20 objektů zřetelně vybočuje objekt 7.*
Objekty 10, 11 a 15 se také odlišují od ostatních, ale méně než objekt 7.



Obr. 3b Polygony: klíč ke grafu slunečních paprsků pro proměnné B402X2, B402X3, B402X4, B402X5 standardizovaných dat.

Paprsky jsou rozmístěny ekvidistantně, ve stejných vzdálenostech na kružnici, a proto se provádí lineární transformace do intervalu $[a, 1]$, kde a je zvolená spodní mez, obyčejně $a = 0$. Pro tuto transformaci platí, že

$$x_{ij}^{*} = \frac{(1 - a)(x_{ij} - \min_i x_{ij})}{\max_i x_{ij} - \min_i x_{ij}} + a$$

kde $\min x_{ij}$ je minimální a $\max x_{ij}$ maximální hodnota j-té proměnné objektu x_i^T přes

všechny objekty x_i^T, $i = 1, ..., n$. K určení směrů jednotlivých paprsků se definuje jejich úhel α_j, pro který platí

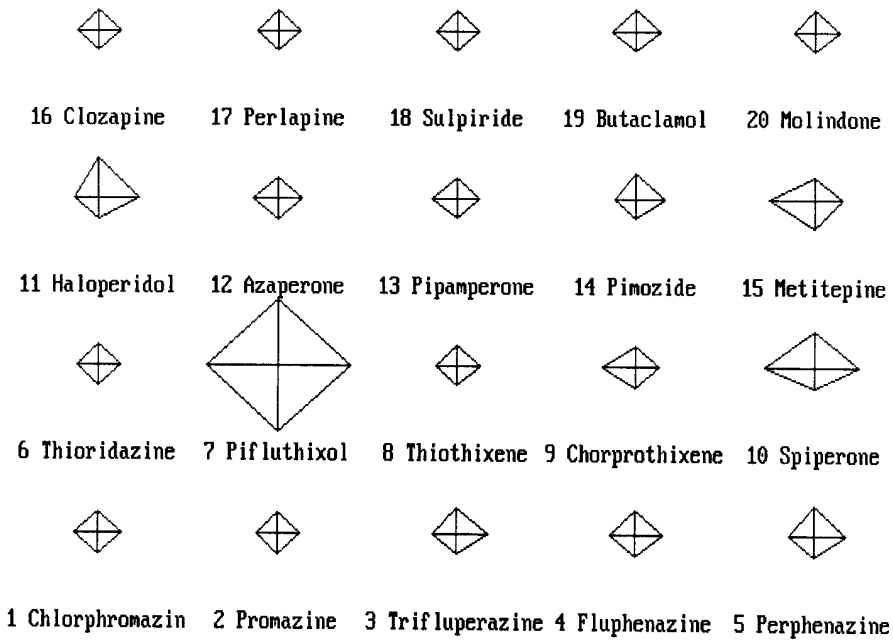
$$\alpha_j = \frac{2\pi(j-1)}{m}, \qquad j = 1, ..., m$$

Za společný střed paprsků se obyčejně volí počátek souřadnic. Pokud má být maximální délka paprsků rovna R, je polygon pro objekt x_i^T spojnicí m bodů p_{ij} o souřadnicích

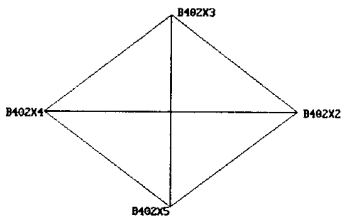
$$p_{ij} = (x_{ij}^* R \cos\alpha_j, x_{ij}^* R \sin\alpha_j).$$

Aby vznikl uzavřený obrazec, spojují se ještě první a poslední bod p_{i1} a p_{im}. Vzájemné porovnání polygonů slouží k vizuálnímu posouzení podobnosti objektů. V případě velkého počtu proměnných, např. $m > 6$, bývá však výsledný obrázek polygonů nepřehledný.

(b) **Hvězdicový graf** vypadá na první pohled jako předchozí graf. Sestává z paprsků, reprezentujících relativní hodnoty proměnných u jednotlivých objektů, které se pro každý objekt spojují v jednom centrálním bodě.



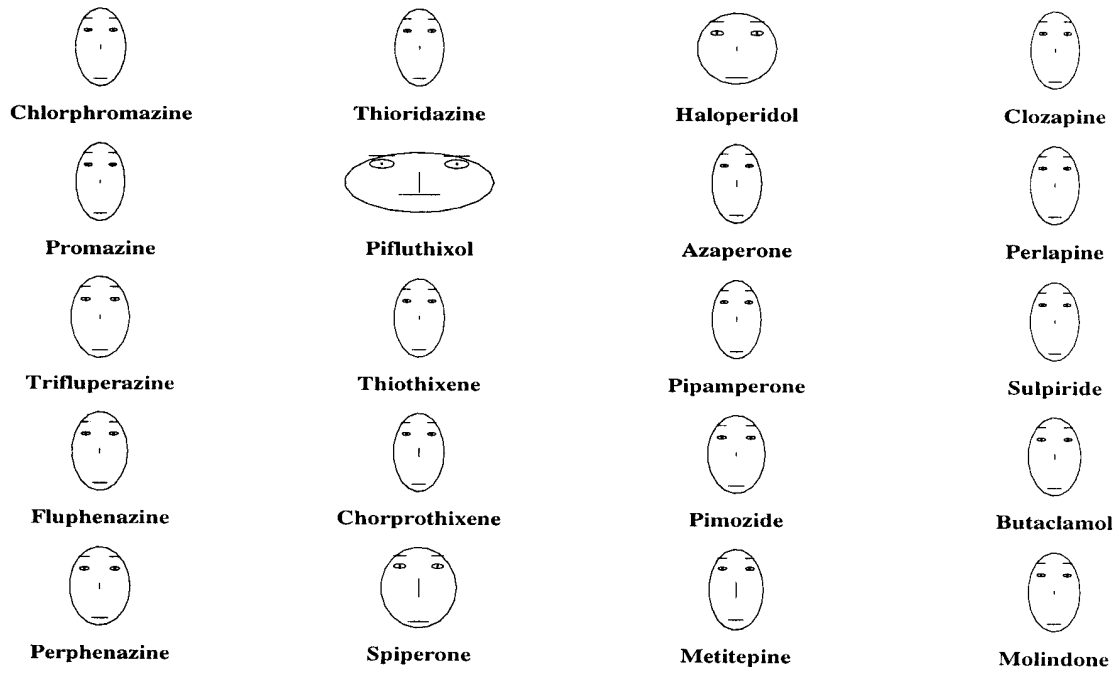
Obr. 4a Polygony: hvězdičkový graf pro 20 objektů a 4 proměnné B402X2, B402X3, B402X4, B402X5 standardizovaných dat.



Obr. 4b Polygony: klíč ke hvězdičkovému grafu pro 4 proměnné B402X2, B402X3, B402X4, B402X5 standardizovaných dat.

Stejně směřující paprsky u různých objektů se liší svojí délkou. *Nejkratší paprsek* indikuje, že u objektu nabývá příslušná proměnná nejmenší hodnoty z celého výběru. Podobně *nejdelší paprsek* informuje o nejvyšší hodnotě příslušné proměnné. Délky ostatních paprsků se pohybují podle relativní velikosti hodnot proměnné u příslušného objektu mezi těmito dvěma krajními mezemi.

3. **Tváře** charakterizují každou proměnnou x_{ij} objektu $\boldsymbol{x}_i^T$ nějakým znakem. Mezi znaky patří tvar tváře, délka nosu, velikost očí, tvar úst, atp. Tvar tváře závisí na použitém pořadí proměnných, které ovlivňuje snadnost interpretace dat.



Obr. 5 Polygony: tváře nestandardizovaných dat pro 20 objektů a 4 proměnné B402X2, B402X3, B402X4, B402X5, S-Plus.
Výklad: *lze nalézt řadu vzájemně podobných tváří, ukazujících na podobnost objektů. Tvář Pifluthixolu se jeví silně odlišná od ostatních.*

4. **Křivky** využívají transformace každého objektu $\boldsymbol{x}_i^T$ do spojité křivky, která je lineární kombinací všech jeho proměnných. Andrews[12] volí pro vyjádření křivky f_i odpovídajícího objektu $\boldsymbol{x}_i^T$ konečnou Fourierovu řadu

$$f_{x_i}(t) = f_i = \frac{x_{i1}}{\sqrt{2}} + x_{i2}\sin(t) + x_{i3}\cos(t) + x_{i4}\sin(2t) + x_{i5}\cos(2t) + \ldots$$

Křivky f_i, $i = 1, \ldots, n$, se vynášejí jako funkce proměnné t v intervalu $-\pi \leq t \leq \pi$. Funkce f_i, mají řadu výhodných vlastností:

a) Funkce f_i zachovávají *průměr*. To znamená, že pokud je $\bar{\boldsymbol{x}}$ průměrem z celkového počtu n vícerozměrných dat $\boldsymbol{x}_i$, je funkce rovna

$$f_{\bar{x}}(t) = \frac{1}{n} \sum_{i=1}^{n} f_{x_i}(t),$$

kde funkce $f_{\bar{x}}(t)$ je "průměrná" křivka.

b) Funkce f_i zachovávají *vzdálenosti*. To znamená, že celková vzdálenost mezi křivkami f_i a f_j definovaná jako integrální kvadratická odchylka, odpovídá vzdálenosti mezi objekty $\boldsymbol{x}_i^T$ a $\boldsymbol{x}_j^T$. Blízké křivky ukazují na nepříliš vzdálené objekty.

c) Pro zvolenou hodnotu t_0 je funkce $f_{x_i}(t_0)$ projekcí objektu $\boldsymbol{x}_i$ na vektor $\boldsymbol{p}_0$ o složkách

$$\boldsymbol{p}_0 = (\frac{1}{\sqrt{2}}, \sin(t_0), \cos(t_0), \sin(2t_0), \cos(2t_0), ...).$$

Tato projekce do jednoho bodu umožňuje odhalení vybočujících objektů či skupin objektů, které mohou být ve více dimenzích špatně identifikovatelné. Křivka $f_{x_i}(t)$ je složena ze všech projekcí na daném intervalu hodnot t;

d) Funkce f_i zachovávají *rozptyl*. To znamená, že pokud jsou proměnné x_j objektu $\boldsymbol{x}_i^T$ nekorelované náhodné veličiny se stejným rozptylem σ^2, je

$$D(f_i) = \sigma^2 (0.5 + \sin^2(t) + \cos^2(t) + \sin^2(2t) + \cos^2(2t) + ...)$$

Pro liché m je $D(f_i) = 0.5\ \sigma^2\ m$ a pro sudé m je $0.5\ \sigma^2\ (m - 1) < D(f_i) < 0.5\ \sigma^2\ (m + 1)$. Rozptyl funkce f_i je téměř konstantní v celém rozmezí veličiny t.

V praktických úlohách je běžné, že složky objektu $\boldsymbol{x}$, tj. jednotlivé proměnné, jsou silně korelované a mají nestejné rozptyly. Pak je výhodné převést objekty původních dat $\boldsymbol{x}_i$ na objekty $\boldsymbol{y}_i$, kde y_{ij} odpovídá transformaci do j-té hlavní komponenty. Veličiny y_{ij} jsou již nekorelované. Snadno lze provést i jejich standardizaci tak, aby měly konstantní rozptyly. Nevýhodou křivek je to, že jejich tvar závisí na pořadí složek. Na druhé straně lze pomocí křivek snadno indikovat vybočující objekty nebo skupiny objektů a konstruovat i konfidenční křivky. Pro větší počty objektů ($n > 10$) dochází ke splývání křivek, což ztěžuje jejich interpretaci. Pak je možné vynášet pouze zvolené podskupiny objektů.

5. **Stromy** jsou vhodné pro případy, kdy je počet proměnných m objektu $\boldsymbol{x}_i^T$ veliký. Jednotlivé složky $\boldsymbol{x}_j$ představují délku větví schematického stromu. Jeho struktura čili rozmístění větví se provádí na základě předběžného hierarchického shlukování proměnných (viz. *shluková analýza*). Předběžná shluková analýza se dá použít také při výběru pořadí složek objektu $\boldsymbol{x}$ při konstrukci ostatních symbolových grafů.

6. Určení struktury a vazeb v proměnných

Určením struktury a vzájemných vazeb mezi proměnnými se zabývají techniky redukce proměnných na latentní proměnné, metoda *analýzy hlavních komponent (PCA)* a *metoda faktorové analýzy (FA)*.

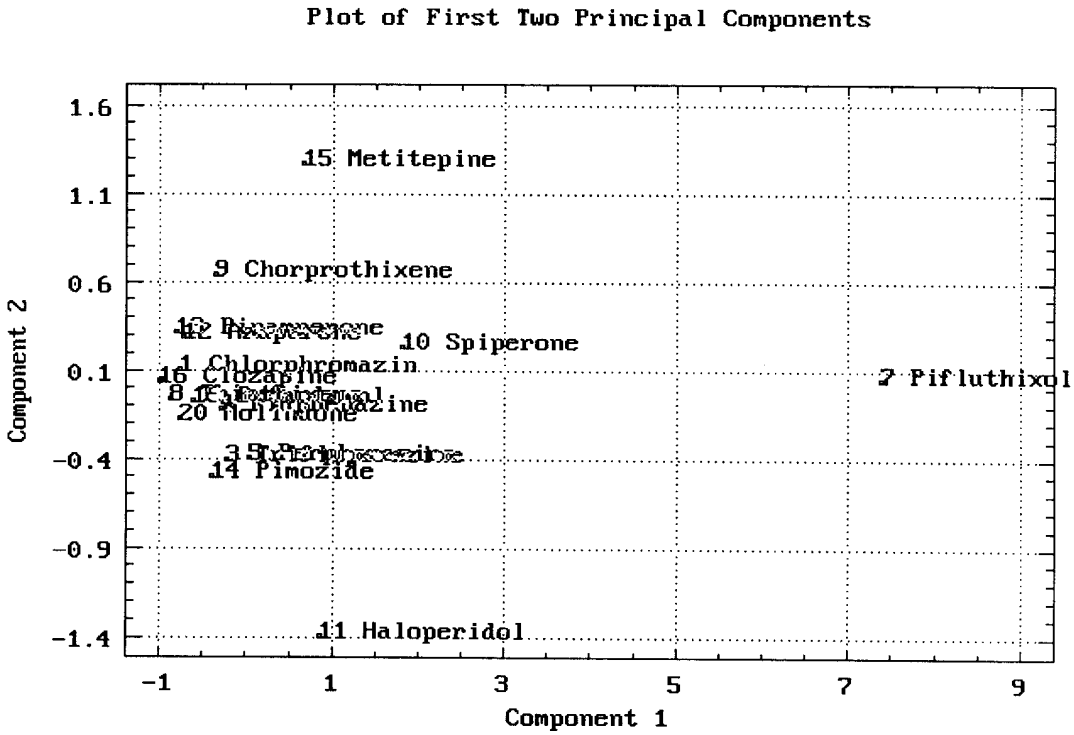
6.1 Metoda analýzy hlavních komponent (PCA)

Principem metody je náhrada původních proměnných x_i tzv. *latentními proměnnými* y_i, které mají vhodnější vlastnosti, totiž je jich výrazně menší počet, ačkoliv vystihují téměř celou *proměnlivost* původních proměnných a jsou vzájemně nekorelované (korelační koeficient mezi latentními proměnnými $y_1, ..., y_m$ je 0). Latentní proměnné se nazývají hlavní komponenty a jsou lineárními kombinacemi původních proměnných. *První hlavní komponenta*, tj. y_1 popisuje největší část proměnlivosti čili rozptylu původních dat, *druhá hlavní komponenta*, tj. y_2 zase největší část rozptylu neobsaženého v y_1, atd.

Matematicky řečeno, *první hlavní komponenta* je takovou lineární kombinací vstupních proměnných, která obsahuje největší rozptyl mezi všemi lineárními kombinacemi. Má tvar

$$y_1 = \sum_{j=1}^{m} v_{1j}\, x_j = \boldsymbol{v}_1^T\, \boldsymbol{x},$$

kde objekt $\boldsymbol{x}$ obsahuje proměnné $x_1, ..., x_m$. Pro vektor koeficientů $\boldsymbol{v}_1^T = (v_{11}, ..., v_{1m})^T$ platí, že rozptyl $D(y_1) = \boldsymbol{v}_1^T\, \boldsymbol{S}\, \boldsymbol{v}_1$ je maximální, přičemž $\boldsymbol{S}$ značí kovarianční matici původních dat $\boldsymbol{X}$.



Obr. 6 Metoda hlavních komponent: rozptylový diagram pro 20 objektů a 4 proměnné B402X2, B402X3, B402X4, B402X5 standardizovaných dat.

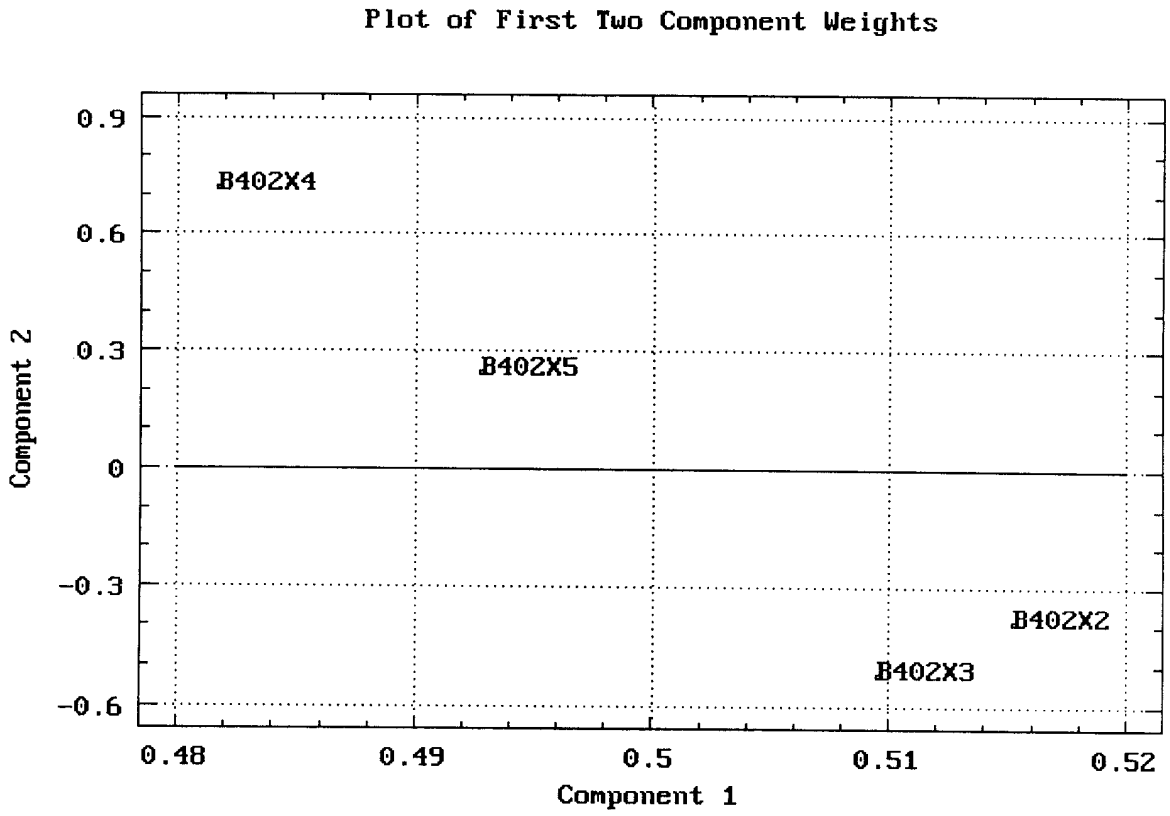
Výklad: *Kromě tří objektů: 7, 11 a 15 leží zbývajících 17 objektů v jediném shluku. Objekty 7, 11 a 15 jsou odlehlé body. Nejvíce odlišný objekt od ostatních je 7, protože je odlehlý na hlavní komponentě 1, popisující většinu rozptylu. První hlavní komponenta 1 vlastně popisuje rozdíl mezi Pifluthixolem a ostatními objekty. Na druhé straně objekty 11 a 15 jsou extrémy na druhé hlavní komponentě a udávají její směr. Ostatní objekty tvoří v rovině prvních dvou hlavních komponent homogenní shluk.*

Zcela analogicky jsou konstruovány další hlavní komponenty, jejichž celkový počet roven menšímu z čísel n (počet objektů) a m (počet proměnných). Protože platí, že součet rozptylů všech hlavních komponent je roven součtu rozptylů vstupujících proměnných, můžeme z podílu rozptylů jednotlivých hlavních komponent usuzovat na část proměnlivosti, vysvětlenou dotyčnou hlavní komponentou. Jestliže součet prvních (nejvyšších) k podílů proměnlivosti je dostatečně blízký jedné (obvykle však stačí 0.9 - 0.95), postačí brát v úvahu právě těchto prvních k hlavních komponent pro "dostatečné" vysvětlení původních proměnných. I při velkém počtu původních proměnných (m) může být k velmi malé, často 2 až 5.

Maximalizací při zavedení normalizační podmínky $v_1^T \mathbf{v} = 1$ vyjde, že

$$(S - \lambda_1 E)\, v_1 = \mathbf{0} ,$$

kde **0** označuje nulový vektor, λ_1 je největší *vlastní číslo* a v_1 je odpovídající *vlastní vektor* kovarianční matice S. Po dosazení vyjde $D(y_1) = v_1^T S\, v_1\, \lambda_1$. Analogicky lze odvodit, že vektor koeficientů v_2 ve vztahu $y_2 = \sum_{j=1}^{m} v_{2j}\, x_j$, maximalizující $D(y_2)$ za podmínky, že $cov(y_1, y_2) = 0$, odpovídá vlastnímu vektoru, příslušejícímu druhému největšímu vlastnímu číslu λ_2.

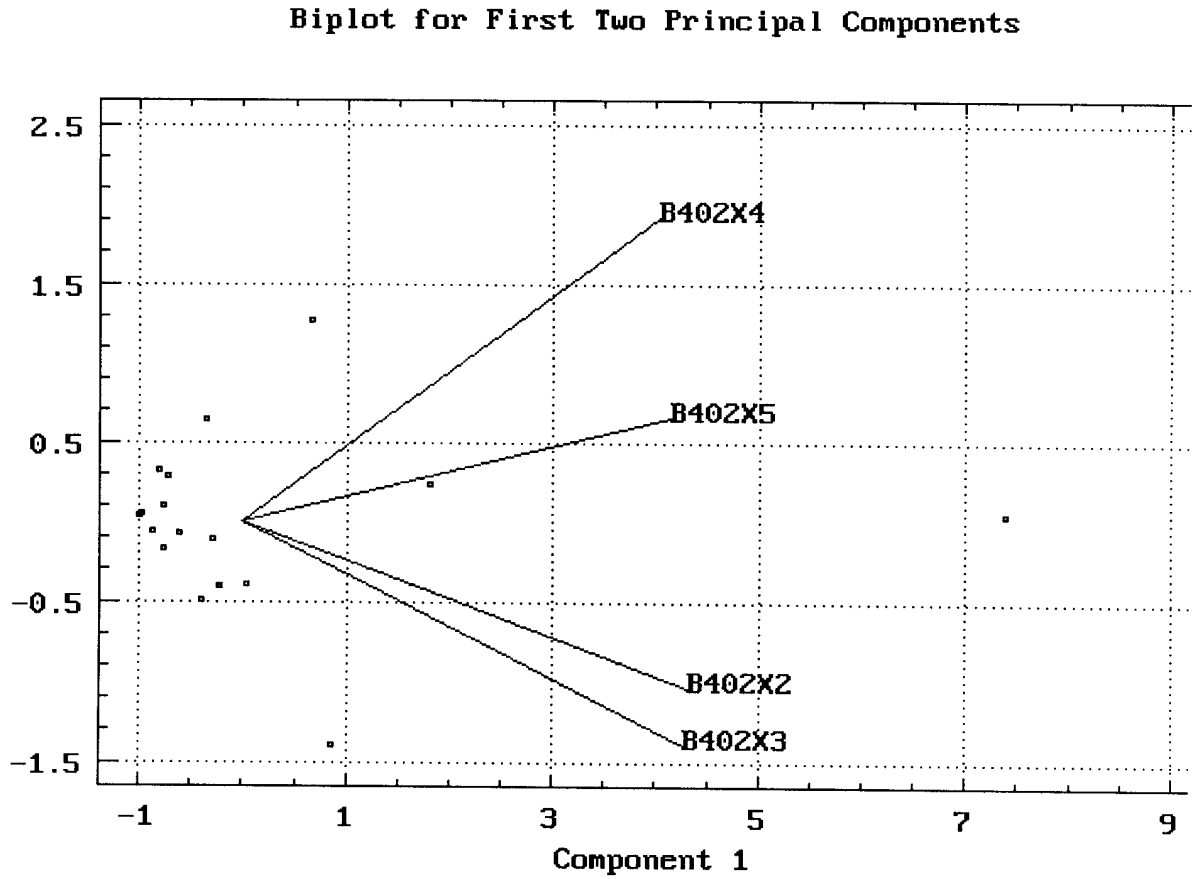


Obr. 7 Metoda hlavních komponent: graf komponentních vah pro 20 objektů a 4 proměnné B402X2, B402X3, B402X4, B402X5 ze standarizovaných dat.
Výklad: *Proměnné B402X2 a B402X3 leží v diagramu blízko sebe, a proto spolu silně korelují. Proměnné B402X4 a B402X5 jsou dál od sebe, proto daleko méně korelují. Méně korelují rovněž se zbývajícími dvěma proměnnými B402X2 a B402X3, jsou totiž daleko od nich.*

Provedeme-li rozklad kovarianční matice S na vlastní čísla $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_m$, jsou odpovídající vlastní vektory $v_1, v_2, \ldots, v_m$ přímo koeficienty hlavních komponent $y_1, \ldots, y_m$. Hlavní komponenty mají řadu zajímavých vlastností. Lze je interpretovat jako hlavní osy m-rozměrného elipsoidu $x^T S^{-1} x$ = konst . K odstranění závislosti na jednotkách původních proměnných se lépe užívá standardizovaných proměnných x^* s prvky $x_j^* = (x_j - \bar{x}_j)/\sigma_j$. Pro j-tou hlavní komponentu pak platí

$$y_j^* = \sum_{k=1}^{m} v_{jk}^* x_k^*,$$

kde v_j^* je vlastní vektor *korelační* matice R odpovídající j-tému největšímu vlastnímu číslu λ_j^*. Hlavní komponenty y_j^*, určené z korelační matice, jsou však hůře interpretovatelné. Platí, že $v_j^{*T} v_j^* = \lambda_j$, nikoliv rovno jedné. Navíc je jejich statistická analýza komplikovanější. Pro účely zobrazení vícerozměrných dat různého měřítka jsou však vhodnější standardizované y_j^* než původní y_j.



Obr. 8 Metoda hlavních komponent: dvojný graf pro 20 objektů a 4 proměnné B402X2, B402X3, B402X4, B402X5 standardizovaných dat.

Výklad: *Úhel mezi dvěma průvodiči dvou proměnných je nepřímo úměrný velikosti korelace mezi těmito proměnnými. Mezi průvodiči B402X2 a B402X3 je malý úhel, což svědčí o silné korelaci. Úhly mezi těmito dvěma průvodiči a průvodiči B402X4 a B402X5 jsou pak větší, což ukazuje na jejich slabší korelaci.*

Graf hlavních komponent ukazuje na ose y hodnoty i-té hlavní komponenty y_{i1} (resp. y^*_{i1}) a na ose x hodnoty j-té hlavní komponenty y_{i2} (resp. y^*_{i2}). Obecně se i-tá souřadnice y_{i1} určí dosazením i-tého bodu $\boldsymbol{x}_i$ místo $\boldsymbol{x}$. Stejně se vypočtou i souřadnice y_{i2}, resp. y^*_{i1}, y^*_{i2}. Z grafu hlavních komponent lze snadno určit jak vybočující objekty, tak i shluky objektů. Graficky lze výsledek analýzy hlavních komponent zobrazit trojím způsobem:

(a) **Rozptylový diagram** (Scatterplot) zobrazuje *komponentní skóre*, tj. hodnoty dvou hlavních komponent u jednotlivých objektů navzájem. Dokonalé rozptýlení objektů v rovině obou hlavních komponent vede k rozlišení objektů při jejich popisu pomoci y_1 a y_2. Snadno lze v rovině nalézt shluk vzájemně podobných objektů.

(b) **Graf komponentních vah** (Plot Components Weights) zobrazí komponentní váhy pro první dvě hlavní komponenty. V tomto grafu se porovnávají vzdálenosti mezi proměnnými. Krátká vzdálenost mezi dvěma proměnnými znamená silnou korelaci. Lze nalézt i shluk podobných proměnných, jež spolu korelují.

(c) **Dvojný graf** (Biplot) kombinuje předchozí dva grafy. Úhel mezi průvodiči dvou proměnných je nepřímo úměrný velikosti korelace mezi těmito proměnnými. Čím menší úhel, tím větší korelace. Každý průvodič má své souřadnice na první a na druhé hlavní komponentě. Délka této souřadnice je úměrná příspěvku původní proměnné x_j do hlavní komponety čili je úměrná komponentní váze.

Eigenvalues

No.	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
1	3394.339	92.62	92.62	¦¦¦¦¦¦¦¦¦¦¦¦¦¦¦¦¦¦¦
2	252.286	6.88	99.50	¦¦
3	15.8825	0.43	99.94	¦
4	2.295	0.06	100.00	¦

Vlastní čísla slouží k určení počtu "využitelných" hlavních komponent, jež si zvolíme v analýze k dalšímu užívání. Procento a kumulativní procento popisuje proměnlivost v původních proměnných, popsanou dotyčnou hlavní komponentou. Bereme obvykle k dalšímu popisu proměnlivosti tolik hlavních komponent, aby bylo jimi popsáno 90 až 99% celkové proměnlivosti. V tomto případě stačí užít první dvě. Scree Plot je vlastně sloupcový diagram vlastních čísel. Zobrazuje relativní velikost jednotlivých vlastních čísel. Řada autorů ho s oblibou využívá k určení počtu "užitečných" hlavních komponent. Cattel vysvětluje scree jako zlomové místo mezi kolmou stěnou a vodorovným dnem. Vybrané "užitečné" hlavní komponenty (nebo také faktory) pak tvoří kolmou stěnu a "neužitečné" hlavní komponenty (nebo faktory) představují vodorovné dno. Užitečné komponenty jsou tak odděleny zlomovým místem.

Vlastní vektory jsou váhy, jež umožňují kombinovat komponentní proměnné, které byly předem normovány vzorcem $(x_i - \bar{x})/\sigma_i$. Např. první hlavní komponenta

Component1 je vážený průměr normovaných proměnných, kdy váha každé proměnné je dána odpovídajícím prvkem prvního vlastního vektoru

$$Component1 = v_{11}x_1 + v_{12}x_2 + \ldots + v_{1m}x_m .$$

Koeficienty v této rovnici vystihují relativní důležitost každé proměnné při tvorbě hlavní komponenty. Vlastní vektory bývají často normovány, takže rozptyl komponentního skóre je roven jedné.

7. Určení struktury a vzájemných vazeb v objektech

Hledáním struktury a vzájemných vazeb v objektech se zabývají především klasifikační metody vícerozměrné statistické analýzy. *Klasifikační metody* jsou postupy, pomocí kterých se jeden objekt zařadí do jedné existující třídy (*diskriminační analýza*), nebo pomocí nichž lze neuspořádanou skupinu objektů uspořádat do několika vnitřně sourodých tříd (*analýza shluků*). Postup klasifikace je založen na určitých předpokladech o vlastnostech klasifikovaných objektů, např. normální rozdělení náhodného vektoru charakterizujícího objekty, a pak hovoříme o *parametrických klasifikačních metodách.* Není-li klasifikace založena na znalostech rozdělení náhodného vektoru mluvíme o *neparametrických klasifikačních metodách.*

7.1 Analýza shluků

Analýza shluků patří mezi metody, zabývající se vyšetřováním podobnosti *vícerozměrných objektů* (tj. objektů, u nichž je změřeno větší množství proměnných) a jejich roztříděním do skupin čili *shluků.* Hodí se zejména tam, kde objekty projevují přirozenou tendenci se seskupovat.

Hierarchické postupy jsou založeny na postupném spojování objektů a jejich shluků do dalších, větších shluků. Nejprve se vypočte základní matice vzdáleností mezi objekty. Dva objekty, jejichž vzdálenost je nejmenší, se 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 celek. Celý postup se opakuje tak dlouho, dokud všechny objekty netvoří jeden velký shluk nebo dokud nezůstane určitý předem zadaný počet shluků. Přitom vznikají dva základní problémy:

(a) *způsob měření vzdáleností mezi objekty:* i když existuje celá řada měr vzdáleností (vícerozměrných metrik), nejčastěji se užívá *euklidovská metrika*, která je přirozeným zobecnením běžného pojmu vzdálenosti;

(b) *volba vhodné shlukovací procedury*, dle zvoleného způsobu metriky. Metody shlukování jsou

Metoda průměrová (Average): vzdálenost dvou shluků se počítá jako průměr z možných mezishlukových vzdáleností dvou objektů, kdy mezishlukovou vzdáleností objektů se rozumí vzdálenost dvou objektů, z nichž každý patří do jiného shluku.

Metoda centroidní (Centroid): vzdálenost shluků se počítá jako euklidovská vzdálenost jejich centroidů, tj. průměrů proměnných v jednotlivých shlucích.

Metoda nejbližšího souseda (Nearest): kritériem pro spojování shluků je minimum z možných mezishlukových vzdáleností objektů.

Metoda nejvzdálenějšího souseda (Furthest): počítá vzdálenost dvou shluků jako maximum z možných mezishlukových vzdáleností objektů.

Metoda mediánová (Median): jde o jisté vylepšení centroidní metody, neboť se snaží odstranit rozdílné "váhy", které centroidní metoda dává různě velkým shlukům.

Nehierarchické shlukové metody: *metoda typických bodů* (Seeded), kdy uživatel na základě svých věcných znalostí určí, které objekty mají být "typickými" představiteli nově vytvořených shluků a systém rozdělí objekty do shluků podle jejich euklidovské vzdálenosti od těchto typických objektů.

Diagram shluků se objeví pouze v případě, že jsme zadali hodnoty původních proměnných a nikoli matici vzdáleností. Výsledkem je zobrazení hodnot ve dvojrozměrném prostoru, kde osy tvoří zadané proměnné. Objeví se také "obkroužení" objektů v jednotlivých shlucích.

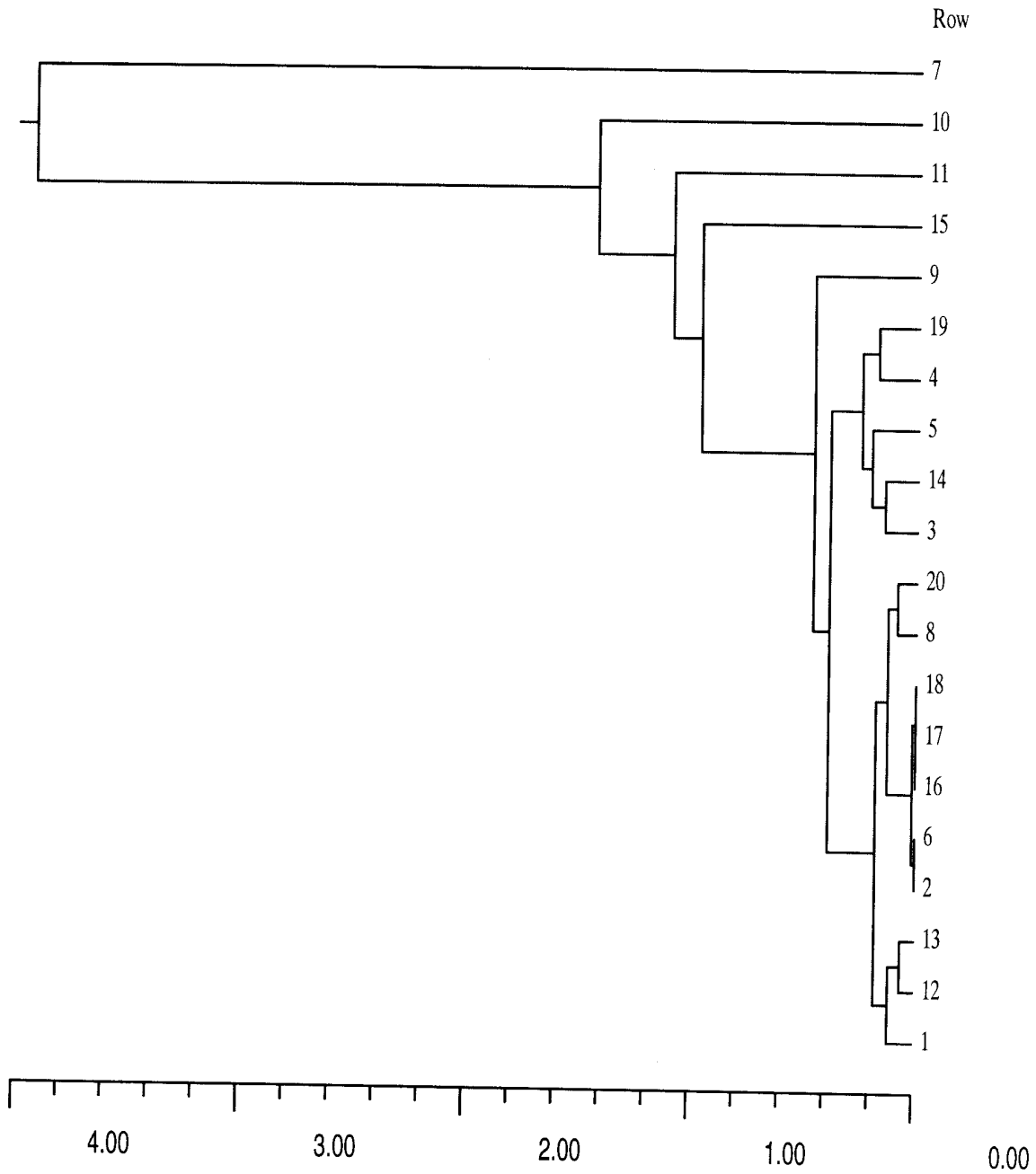
Dendrogram je standardní výstup hierarchických shlukovacích metod, ze kterého je patrná struktura objektů ve shlucích.

Number Link	Clusters	Distance Value	Distance Bar	Rows Linked
19	1	3.919240	¦IIIIIIIIIIIIIIIIIIIIIIIIIIII	1,12,13,2,6,16,17,18,8,20,3,14,5,4,19,9,15, 11,10,7
18	2	1.425636	¦IIIIIIIIII	1,12,13,2,6,16,17,18,8,20,3,14,5,4,19,9,15,11,10
17	3	1.088240	¦IIIIIIII	1,12,13,2,6,16,17,18,8,20,3,14,5,4,19,9,15,11
16	4	0.961214	¦IIIIIII	1,12,13,2,6,16,17,18,8,20,3,14,5,4,19,9,15
15	5	0.459505	¦IIII	1,12,13,2,6,16,17,18,8,20,3,14,5,4,19,9
14	6	0.387292	¦III	1,12,13,2,6,16,17,18,8,20,3,14,5,4,19
13	7	0.250467	¦II	3,14,5,4,19
12	8	0.204587	¦II	3,14,5
11	9	0.177166	¦I	1,12,13,2,6,16,17,18,8,20
10	10	0.176610	¦I	4,19
9	11	0.144072	¦I	3,14
8	12	0.125591	¦I	2,6,16,17,18,8,20
7	13	0.113314	¦I	1,12,13
6	14	0.083506	¦I	8,20
5	15	0.062059	¦	12,13
4	16	0.013643	¦	2,6,16,17,18
3	17	0.000000	¦	16,17,18
2	18	0.000000	¦	16,17
1	19	0.000000	¦	2,6

Cophenetic Correlation 0.989977

Delta(0.5) 0.119003

Delta(1.0) 0.110925



Obr. 9 Dendrogram, metoda průměru a Eukleidovské vzdálenosti pro 20 objektů a 4 proměnné B402X2, B402X3, B402X4, B402X5 ze standardizovaných dat úlohy B402, ***NCSS60***

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.
[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.