12 Mnohorozměrné metody

Průvodce studiem

Na tuto rozsáhlou kapitolu počítejte nejméně s deseti hodinami usilovného studia s tím, že se k probírané látce budete ještě vracet po pochopení dalších souvislostí. Věnujte pozornost řešeným příkladům.

Dosud jsme se zabývali regresí, kdy jedna náhodná veličina je vysvětlována (nebo predikována) pomocí jiných veličin. Hledá se závislost podmíněné střední hodnoty náhodné veličiny na regresorech. Je to nejčastěji aplikovaná statistická metoda. Odhaduje se, že více jak 90% aplikací statistiky se opírá o regresi. Pochopení principů regrese je velmi užitečné pro pochopení ostatních metod analýzy mnohorozměrných dat. Regresní analýza bývá považována za zcela samostanou část stojící vedle mnohorozměrných metod (methods of multivariate analysis). Ve většině statistického software je regresní a korelační analýza uváděna jako samostatná položka stojící vedle mnohorozměrných metod. Také učebnice a monografie bývají věnovány samostatně regresi a samostatně zbývajícím metodám mnohorozměrné analýzy dat.

Mezi mnohorozměrné metody jsou zařazovány především:

- testy shody vektorů středních hodnot
  MANOVA (multivariate analysis of variance) - mnohorozměrná analogie analýzy rozptylu
- kanonické korelace, které můžeme považovat za jisté zobecnění lineární regrese, kdy vysvětlujeme ne jednu náhodnou veličinu, ale vektor náhodných veličin
- metody klasifikace, kdy předpokládáme, že data pocházejí z více populací a
  - hledáme pravidlo umožňující zařadit (klasifikovat) objekt charakterizovaný vektorem hodnot do jedné z populací (diskriminační analýza, logistická regrese, neuronové sítě atd.)
  - pokoušíme se najít v datech podmnožiny podobných objektů (shluková analýza – cluster analysis)
- metody redukce dimenze úlohy, kdy proměnlivost a závislosti v datech se pokoušíme vyjádřit pomocí méně veličin. Analýza hlavních komponent (principal components) vysvětluje rozptyl. Faktorová analýza vysvětluje kovarianční (korelační) strukturu.

12.1 Test shody vektoru středních hodnot

Pro test shody vektoru středních hodnot se užívá Hottelingův T^2 (čti té-kvadrát) test. Je to mnohorozměrná analogie t-testů. Ve stručnosti uvedeme základní myšlenky.

Jednovýběrový Hottelingův T^2 test:
Testuje se hypotéza, že p-rozměrný vektor středních hodnot $\boldsymbol{\mu}$ je roven nějakému danému konstantnímu vektoru. Předpokládá se, že výběr je z mnohorozměrného normálního rozdělení. Testovou statistikou je pak

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu})^T \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}). \tag{49}$$

Tato statistika má Hottelingovo rozdělení. Lze také užít statistiku

$$\frac{T^2}{n-1}\,\frac{n-p}{p} \sim F_{p,n-p} \tag{50}$$

Intervaly spolehlivosti pro p-rozměrný vektor středních hodnot odvodíme z (49) a (50).

$$P\left[\frac{T^2}{n-1}\,\frac{n-p}{p} < F_{1-\alpha}(p, n-p)\right] = 1-\alpha$$

Po úpravě dostaneme

$$(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu}) < \frac{n-1}{n}\frac{p}{n-p}F_{1-\alpha}(p, n-p),$$

kde $(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c$ znamená plochu elipsoidu se středem $\bar{\mathbf{x}}$, jehož tvar a velikost závisí na výběrové kovarianční matici $\mathbf{S}$. Volbou α určíme hodnotu c a můžeme určit intervaly spolehlivosti pro vektor středních hodnot $\boldsymbol{\mu}$.

Dvouvýběrový Hottelingův T^2 test:
Testujeme shodu dvou vektorů středních hodnot (mnohorozměrná analogie dvouvýběrového t-testu). Máme dva výběry z p-rozměrného normálního rozdělení o rozsazích $n_1,\ n_2$, $n_1 + n_2 = n$. Vektory výběrových průměrů jsou $\bar{\mathbf{x}}_1,\ \bar{\mathbf{x}}_2$. Za předpokladu shody kovariančních matic $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ můžeme z výběrových kovariančních matic $\mathbf{S}_1, \mathbf{S}_2$ odhadnout společnou výběrovou kovarianční matic

$$\mathbf{S} = \frac{(n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2}{n_1+n_2-2}$$

Označíme $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. Pak statistika

$$T^2 = \frac{n_1 n_2}{n}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\delta})^T \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\delta})$$

má Hottelingovo rozdělení a

$$\frac{n-p-1}{p}\frac{T^2}{n-2} \sim F(p, n-p-1),$$

kterou můžeme užít k testu hypotézy

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$

Pokud $\boldsymbol{\Sigma_1} \neq \boldsymbol{\Sigma_2}$, jedná se o mnohorozměrnou analogii dvouvýběrového t-testu s nestejnými rozptyly. Pak je test poněkud komplikovanější, viz např. Hebák a kol.

12.2 Diskriminační analýza

Diskriminační analýza je postup, který hledá vhodné pravidlo (rozhodovací funkci) umožňující na základě zadaných hodnot vektoru $\mathbf{x}$ zařadit objekt do některé, řekněme h-té skupiny. Např. velmi jednoduchou úlohou tohoto typu je rozhodnout podle změřené teploty osoby o tom, zda je zdravá či nemocná. V tomto případě $\mathbf{x}$ je skalár (teplota) a rozhodovací pravidlo velice jednoduché: je-li teplota vyšší než 37^0 C, pak zařaď do skupiny nemocných, jinak do skupiny zdravých. Tedy klasifikujeme osobu, u níž příslušnost do skupiny neznáme a užíváme rozhodovacího pravidla získaného z dat popisujících vztah teploty příslušnosti ke skupině.

Je jasné, že naším zájmem je najít takové pravidlo, které by klasifikovalo pokud možno správně. V reálném světě většinou není možné najít pravidlo klasifikující správně vždy. Proto dobré pravidlo bude takové, které minimalizuje pravděpodobnost chybných rozhodnutí. Jak uvidíme za chvilku, za jistých předpokladů je takovým pravidlem lineární diskriminační funkce. Odvození jejího tvaru si ukážeme pro klasifikaci do dvou skupin.

Zavedeme následující označení:

$h = 1, 2$ – index skupiny

A_h – jev „příslušnost k h-té skupině“

$P(A_h) = \pi_h$ – apriorní pravděpodobnost

$f_h(\mathbf{x})$ – sdružená hustota pro h-tou skupinu

$P(A_h|\mathbf{x})$ – aposteriorní pravděpodobnost, tj. pravděpodobnost příslušnosti k h-té skupině za podmínky daných hodnot $\mathbf{x}$

Hustotu můžeme zapsat $f_h(\mathbf{x}) = f(\mathbf{x}|A_h)$ pro $h = 1, 2$, tj. sdružená hustota pro h-tou skupinu je hustota za podmínky, že nastane jev A_h.

Podle Bayesova vzorce vyjádříme aposteriorní pravděpodobnost:

$$P(A_h|\mathbf{x}) = \frac{P(A_h)f_h(\mathbf{x}|A_h)}{P(A_1)f(\mathbf{x}|A_1) + P(A_2)f(\mathbf{x}|A_2)} = \frac{\pi_h f_h(\mathbf{x})}{\pi_1 f_1(\mathbf{x}) + \pi_2 f_2(\mathbf{x})}, \qquad (51)$$

$$h = 1, 2.$$

Klasifikovat budeme do skupiny s největší aposteriorní pravděpodobností.

Dále označme $\mathcal{S}$ – výběrový prostor (množinu všech možných výsledků $\mathbf{x}$). Naším cílem je rozdělit tento výběrový prostor na dvě části splňující podmínky:

$$\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2, \quad \mathcal{S}_1 \cap \mathcal{S}_2 = \oslash.$$

Pak když $\mathbf{x} \in \mathcal{S}_h$, zařadíme do h - té skupiny.

Pravděpodobnost chybného zařazení objektu z h-té skupiny do h'-té skupiny je

$$P(\mathbf{x} \in \mathcal{S}_{h'}|A_h) = \int_{\mathcal{S}_{h'}} f_h(\mathbf{x})\mathrm{d}\mathbf{x}, \quad h = 1, 2.$$

Podle věty o úplné pravděpodobnosti je celková pravděpodobnost chybné klasifikace

$$\omega = \pi_1 \int_{\mathcal{S}_2} f_1(\mathbf{x})\mathrm{d}\mathbf{x} + \pi_2 \int_{\mathcal{S}_1} f_2(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{52}$$

Pokud obě chyby klasifikace mají stejnou váhu, je optimální rozhodovací pravidlo, které minimalizuje ω dané vztahem (52). Chceme-li chybám klasifikace dát různou váhu, užijeme ztrátovou matici:

$$\mathbf{Z} = \begin{bmatrix} 0 & z(2|1) \\ z(1|2) & 0 \end{bmatrix}$$

Pak celková ztráta z chybné klasifikace je:

$$\tau = z(2|1)\pi_1 \int_{S_2} f_1(\mathbf{x})\mathrm{d}\mathbf{x} + z(1|2)\pi_2 \int_{S_1} f_2(\mathbf{x})\mathrm{d}\mathbf{x}$$

a optimální je postup, který minimalizuje τ.

Objekt řadíme do skupiny s vyšší aposteriorní pravděpodobností, např. z rov. (51) do skupiny 1 zařadíme objekt, když $\pi_1 f_1(x) > \pi_2 f_2(x)$ (jmenovatel je shodný pro obě skupiny). Klasifikační pravidlo pro zařazení do skupiny 1 je tedy

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2}{\pi_1} \tag{53}$$

Předpokládáme-li p-rozměrné normální rozdělení vektoru $\mathbf{x}$, tj. $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ v 1. skupině a $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ ve 2. skupině, pak hustota je:

$$f_h(\mathbf{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}_h|^{-\frac{1}{2}} \exp\left[-(\mathbf{x} - \boldsymbol{\mu}_h)^T \boldsymbol{\Sigma}_h^{-1} (\mathbf{x} - \boldsymbol{\mu}_h)/2\right]$$

Po dosazení do (53) a zlogaritmování dostaneme

$$\mathbf{x}^T \boldsymbol{\Gamma} \mathbf{x} + \boldsymbol{\eta}^T \mathbf{x} + \xi > 0,$$

kde

$$\boldsymbol{\Gamma} = 0,5(\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1}),$$

$$\boldsymbol{\eta}^T = \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2 \boldsymbol{\Sigma}_2^{-1}$$

$$\xi = \frac{1}{2} \ln \frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} - \ln \frac{\pi_2}{\pi_1} - \frac{1}{2} \left(\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2\right)$$

Jsou-li kovarianční matice v obou skupinách shodné, tj. $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$, pak odpadne kvadratický člen a rozhodovací pravidlo se podstatně zjednoduší:

$$\boldsymbol{\beta}^T \mathbf{x} + \gamma > 0,$$

kde

$$\boldsymbol{\beta}^T = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{\Sigma}^{-1}$$

a

$$\gamma = -\frac{1}{2}\boldsymbol{\beta}^T(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \frac{1}{2}\ln\frac{\pi_2}{\pi_1}$$

Funkce

$$L(\mathbf{x}) = \boldsymbol{\beta}^T \mathbf{x} \tag{54}$$

se nazývá *lineární diskriminační funkce*, zkratkou LDF.

Pokud $\mathbf{x}$ má p-rozměrné normální rozdělení, pak i $L(\mathbf{x})$ má normální rozdělení. Čtverec Mahalanobisovy vzdálenosti vektorů středních hodnot je

$$\triangle^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$

střední hodnoty LDF jsou pak pro skupinu 1 a skupinu 2 jsou

$$E_1[L(\mathbf{x})] = \frac{1}{2}\triangle^2 \qquad E_2[L(\mathbf{x})] = -\frac{1}{2}\triangle^2$$

Oba rozptyly jsou shodné

$$\mathrm{var}_1[L(\mathbf{x})] = \mathrm{var}_2[L(\mathbf{x})] = \triangle^2$$

Oba podprostory $\mathcal{S}_1$ a $\mathcal{S}_2$ v p-rozměrném podprostoru $\mathcal{S}$ odděluje nadrovina určená rovnicí

$$\boldsymbol{\beta}^T \mathbf{x} + \gamma = 0 \quad \text{čili} \quad L(\mathbf{x}) = -\gamma$$

LDF (54) lze vyjádřit také jako

$$L_h(\mathbf{x}) = \boldsymbol{\mu}_h^T \mathbf{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_h^T \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_h$$

a klasifikovat do té skupiny, pro kterou je $L_h(\mathbf{x})$ největší. Tak se postupuje, když se klasifikuje do více než dvou skupin.

LDF je optimální rozhodovací pravidlo pro klasifikaci do skupin, pokud náhodný vektor $\mathbf{x}$ má normální rozdělení a skupiny se liší jen vektorem středních hodnot, nikoliv kovarianční strukturou.

Procedura diskriminační analýzy z dat, u kterých je klasifikace známa, odhaduje hodnoty parametrů lineární diskriminační funkce $\boldsymbol{\beta}$. Pak LDF ve tvaru (54) s hod-

notami odhadů lze užít pro klasifikaci objektů, jejichž příslušnost do skupiny známa není.

Příklad 12.1 V souboru DISKRIM.XLS jsou na 30 objektech, které pocházejí ze dvou populací (veličina *skup*), změřeny hodnoty 10 spojitých veličin ($x1$ až $x10$). Naším úkolem je nalézt pravidlo pro klasifikaci objektů. Pravidlo má být co nejjednodušší (čím méně veličin, tím lépe). Použijeme jednak diskriminační analýzu (lineární diskriminační funkci), jednak logistickou regresi [14].

Abychom ověřili, zda vůbec můžeme lineární diskriminační funkci užít, je nutné, aby se populace lišily ve středních hodnotách. To lze zjistit pomocí dvouvýběrového Hotellinova testu:

```
Two-Sample Hotelling's T2 Report

Group     skup

Descriptive Statistics
                Means              Standard Deviations
Variable    0          1             0             1
x1       12.45333     17.25333     2.289375     2.207088
x2       14.996       13.638       2.427947     2.424848
x3       12.05333     17.23333     3.346612     2.916129
x4      183.5267     236.06       73.80299     62.96599
x5      180.28       232.2533     37.71711     47.72365
x6      187.74       233.14       43.81628     48.71579
x7      190.1267     240.2667     42.31836     47.44453
x8      188.4933     233.7267     37.21266     52.34204
x9      190.2333     238.74       31.271       54.68576
x10     188.08       231.4333     50.21964     61.33117
Count    15           15          15           15
```

Už letmý pohled na popisné statistiky výše naznačuje, že mezi průměry některých veličin ve skupinách jsou významné rozdíly. To potvrzuje jak Hotellingův test, tak dvouvýběrové t-testy pro jednotlivé veličiny.

```
Hotelling's T2 Test Section

Covariance                              Prob
Assumption     T2        DF1      DF2     Level
```

```
Equal       101.988     10      28     0.0002
Unequal     101.988     10      27     0.0002


Student's T-Test Section

                            Prob
Variable     |Student's T|  Level
All (T2)       101.988     0.0002
x1               5.846     0.0000
x2               1.533     0.1366
x3               4.520     0.0001
x4               2.097     0.0451
x5               3.309     0.0026
x6               2.684     0.0121
x7               3.055     0.0049
x8               2.728     0.0109
x9               2.982     0.0059
x10              2.118     0.0432
These individual t-test significance levels
should only be used when the overall T2 value is significant.
```

Stepwise procedura diskriminační analýzy poskytne následující výstup:

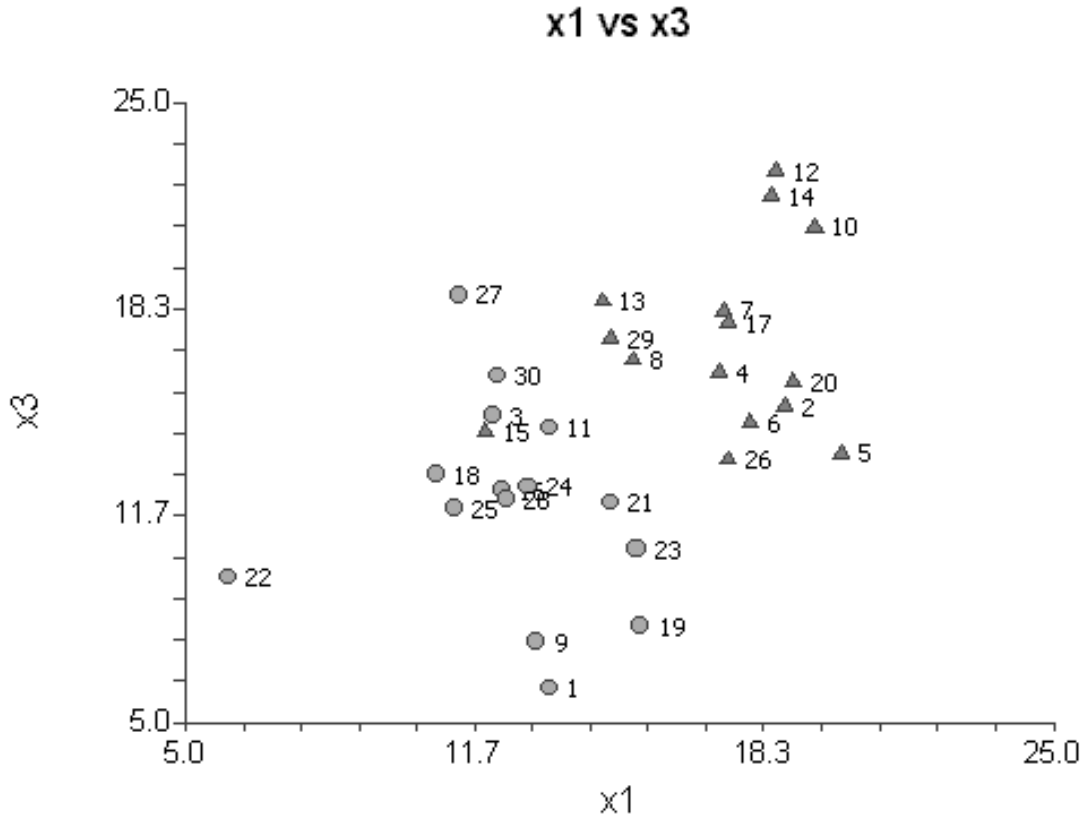
```
Discriminant Analysis Report

Dependent     skup

Variable-Selection Summary Section
           Action     Independent  Pct Chg In          Prob
Iteration  This Step  Variable     Lambda     F-Value  Level
0          None
1          Entered    x1          54.97       34.18   0.000003
2          Entered    x3          25.24        9.12   0.005481

Variable-Selection Detail Section - Step 2
       Independent  Pct Chg In           Prob       R-Squared
Status Variable     Lambda     F-Value   Level      Other X's
In      x1          41.77     19.37       0.000152   0.226687
In      x3          25.24     9.12        0.005481   0.226687
```

```
Out     x2          4.74     1.29        0.265589   0.118719
Out     x4          6.85     1.91        0.178413   0.749175
Out     x5          0.17     0.04        0.836049   0.398169
Out     x6          6.30     1.75        0.197650   0.505648
Out     x7          1.46     0.39        0.539842   0.485859
Out     x8          6.35     1.76        0.195840   0.505363
Out     x9          0.33     0.09        0.772237   0.485821
Out     x10         2.25     0.60        0.445960   0.320911
Overall Wilks' Lambda = 0.336670
Action this step:  None


Linear Discriminant Functions

                    skup
Variable        0            1
Constant    -22.93688    -44.95984
x1            2.481297     3.438486
x3            1.242258     1.775299

Classification Count Table for skup
          Predicted
Actual     0     1     Total
0         15     0     15
1          1    14     15
Total     16    14     30
```

Jako veličiny významně odlišující dvě skupiny byly do klasifikačního pravidla krokovou procedurou vybrány $x1$ a $x3$. Ze 30 objektů je pak 29 klasifikováno správně a jen jeden chybně. Toto empirické ověření spolehlivosti klasifikace je však nutno brát s opatrností, neboť spolehlivost klasifikačního pravidla je ověřována na datech, ze kterých byly koeficienty lineární diskriminační funkce spočítány, nikoliv na nezávislých pozorováních, kde musíme počítat s nižší úspěšností. Realističtější odhad očekávané úspěšnosti klasifikace lze získat buď tak, že data rozdělíme náhodně na skupinu učící a testovací, parametry lineární diskriminační funkce se spočítají jen z učící skupiny a úspěšnost klasifikace se odhadne ze zjištěné klasifikace testovací skupiny. Tento postup má ovšem nevýhodu, že parametry lineární diskriminační funkce se počítají z podstatně menšího počtu pozorování, tzn. nevyužije se informace v datech. V některých programech (v NCSS prozatím nikoliv) je proto *jackknife* procedura, která postupně spočítá parametry lineární diskriminační funkce z $n - 1$

objektů a úspěšnost klasifikace se vždy ověřuje na objektu, který byl vyjmut. Tak se získá odhad spolehlivosti klasifikace na nezávislých pozorováních.

Na obrázku, který následuje, vidíme nesprávně klasifikovaný objekt 15 (ze skupiny 1), který leží uvnitř shluku objektů skupiny 0 („kazí“ lineární separabilitu). Pro ostatní body v rovině lze skupiny od sebe oddělit lineární funkcí (přímkou).



Stejnou úlohu hledání klasifikačního pravidla pro klasifikaci do dvou skupin lze řešit i logistickou regresí. Příslušnost do skupiny je nutno označit { 0, 1 }. Klasifikace je pak založena na odhadu pravděpodobnosti, že pro dané hodnoty regresorů má veličina Y má hodnotu 1. Tvar klasifikační funkce lze snadno vyjádřit z modelu logistické regrese (35)

$$p = \frac{\exp(\mathbf{x}^T\boldsymbol{\beta})}{1 + \exp(\mathbf{x}^T\boldsymbol{\beta})}$$

Je-li p větší než zvolená hodnota (většinou 0,5), pak objekt klasifikujeme do skupiny 1, jinak do skupiny 0. Klasifikační pravidlo na rozdíl od lineární diskriminační funkce je složitější, především není lineární funkcí regresorů. To v některých případech je výhodou, neboť lze najít vhodné klasifikační pravidlo i pro skupiny, které nejsou lineárně separabilní nebo v situacích, kdy nejsou splněny poměrně přísné předpoklady pro aplikaci lineární diskriminační funkce (mnohorozměrné normální rozdělení, shoda kovariančních matic ve skupinách). Někdy je ovšem tato výhoda problematická, jak ukazuje následující výstup z postupné logistické regrese (metoda

forward, tj. postupné přidávání významných regresorů bez vylučování nadbytečných).

Příklad 12.2

```
Logistic Regression Report

Response    skup

Forward Variable-Selection
Action    Variable
Added      x1
Added      x3
Added      x5
Added      x2

Parameter Estimation Section
            Regression    Standard Chi-Square Prob
Variable    Coefficient    Error    Beta=0    Level
Intercept  -231.4874     68355.56     0.00    0.997
x1            4.703548    2773.934    0.00    0.998
x3            4.343548    1478.741    0.00    0.997
x5            2.484694     322.7473   0.00    0.993
x2           -27.89017     5084.675    0.00    0.995

Model Summary Section
Model        Model    Model        Model
R-Squared    D.F.    Chi-Square    Prob
0.624562      4       41.59        0.000000

Classification Table
                  Predicted
Actual             0     1   Total
0        Count    15     0    15
1        Count     0    15    15
Total    Count    15    15    30
Percent Correctly Classified=100
```

Logistickou regresí se podařilo nalézt pravidlo se čtyřmi regresory $x1$, $x3$, $x5$ a $x2$, které má stoprocentní úspěšnost klasifikace. Lineární diskrimininační funkce pro tato data stoprocentní úspěšnosti klasifikace nedosáhne ani při zařazení všech deseti veličin. Ale při podrobnějším pohledu na odhady parametrů logistického modelu a

statistiky s nimi spojené vidíme, že směrodatné odchylky odhadů parametrů jsou velmi vysoké v porovnání s hodnotami odhadů, takže vlastně žádný z odhadů parametrů nemůžeme považovat za významně odlišný od nuly. Klasifikační pravidlo je „ušito na míru“ datům. Pokud zpřísníme kritérium pro zařazování regresorů, dostaneme následující výstup:

```
Logistic Regression Report

Forward Variable-Selection
Action      Variable
Added        x1
Added        x3

Parameter Estimation Section
              Regression       Standard      Chi-Square  Prob
Variable      Coefficient      Error         Beta=0      Level
Intercept     -26.49609       10.43108        6.45      0.011082
x1              1.113222       0.5043868      4.87      0.027309
x3              0.7096213      0.3449301      4.23      0.039658

Model Summary Section
Model         Model      Model           Model
R-Squared      D.F.      Chi-Square      Prob
0.540694        2        31.78          0.000000

Classification Table
            Predicted
Actual              0      1      Total
0      Count       15      0      15
1      Count        1     14      15
Total  Count       16     14      30
Percent Correctly Classified=96.67
```

Toto klasifikační pravidlo obsahuje dva regesory (stejné jako LDF), všechny parametry tohoto logistického modelu můžeme považovat za nenulové a klasifikační pravidlo má stejnou úspěšnost jakou měla LDF.

12.3 Shluková analýza

Cílem shlukové analýzy je nalézt v datech podmnožiny podobných objektů. Mějme množinu m objektů, tuto množinu označme M.

Pro každé dva objekty $a, b \in M$ máme číslo $\sigma(a, b)$, kterému říkáme numerická podobnost.

$$\sigma : M \times M \to R$$

Požadavky na vlastnosti numerické podobnosti:

1. $0 \leq \sigma(a, b) \leq 1$
2. $\sigma(a, a) = 1$
3. $\sigma(a, b) = \sigma(b, a)$
4. $\min(\sigma(a, b), \sigma(b, c)) \leq \sigma(a, c)$ – slabší trojúhelníková nerovnost

Charakteristiku $(1 - \sigma(a, b))$ můžeme chápat jako normovanou vzdálenost dvou objektů $a,\ b$.

Úlohou shlukové analýzy je najít rozklad $\{M_i\}_{i=1}^{k}$, množiny M tj.

1. $\bigcup_{i=1}^{k} M_i = M$
2. $M_i \cap M_j = 0$ pro $i \neq j$
3. vágní kritérium: objekty uvnitř M_i jsou si podobnější mezi sebou než s objekty z množiny M_j, např. když $a, b \in M_i, c \in M_j$, pak $\sigma(a, b) \leq \sigma(a, c), \sigma(a, b) \leq \sigma(b, c)$

Je mnoho možností,

- jak definovat numerickou podobnost,
- jak formulovat postup zařazování objektů do podmnožin,

tedy existuje mnoho metod shlukování.

12.3.1 Hierarchické metody

Vychází se z matice podobnosti objektů (symetrická matice s jedničkami na diagonále), nejčastější je aglomerativní procedura, začne od m shluků (každý shluk je tvořen jedním objektem a spojuje ty shluky, které jsou si nejpodobnější, až skončí jedním shlukem, obsahujícím všech m objektů. Pro takovou posloupnost rozkladů $\{M_{ij}\}_{j=1}^{k}$, pro $i_1 < i_2$ platí $M_{i_1,j} \subseteq M_{i_2,j}$, tj. rozklady jsou do sebe zasunuty, objekty

jednou spojené do shluku zůstávají spolu. Posloupnost spojování můžeme je graficky znázornit *dendrogramem*. Podobnou úlohu řeší taxonomie v biologii.

Nejčastěji užívané strategie spojování shluků jsou:

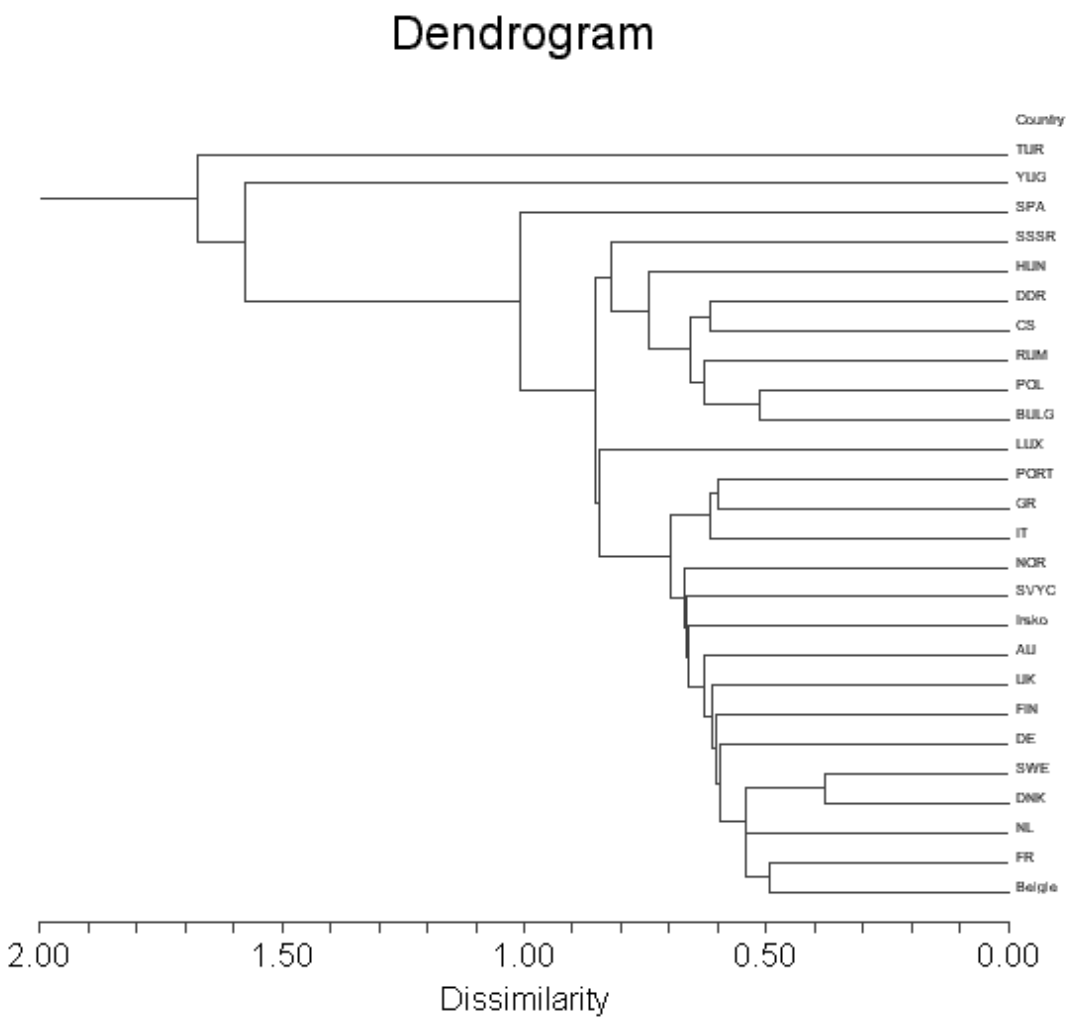
- *single linkage* (nejbližší soused, nearest neiborough) - shluk tvoří souvislý podgraf, tj. existuje aspoň jedna cesta mezi dvěma uzly podgrafu, nejméně přísná metoda na podobnost uvnitř shluků, shluky mají tvar „souhvězdí“
- *complete linkage* (nejvzdálenější soused, furthest neiborough) shluk tvoří úplný podgraf, tj. každé dva uzly podgrafu jsou spojeny hranou, nejpřísnější na podobnost uvnitř shluku
- *average linkage* - spojuje shluky podle jejich průměrné vzdálenosti
- *centroidní* - spojuje shluky podle vzdáleností jejich těžiště

Rozdíly mezi strategiemi shlukování ilustrují následující příklady.

Příklad 12.3 Data pro tento příklad jsou z knihy [18] a jsou uvedena i v souboru EMPLOY.XLS. Obsahují údaje o podílu zaměstnaných v devíti odvětvích ve 26 evropských zemí. Údaje jsou z konce 70. let 20. století, proto jsou v nich uvedeny i státy, které už nyní neexistují. Jednotlivé veličiny znamenají: AGR = agriculture (zemědělství), MIN = mining (těžba), MAN = manufacturing (těžký průmysl), PS = power supplies (energetika), CON = construction (stavebnictví), SER = service industries (lehký průmysl), FIN = finance, SPS = social and personal services (sociální služby), TC = transport and communications (doprava a spoje).

Ve všech ukázkách je zvolena Eukleidovská vzdálenost mezi objekty, všechny veličiny byly standardizovány, po standardizaci tedy mají jednotkový rozptyl. Výstupy se liší jen podle použité strategie (metody) shlukování.

```
Hierarchical Clustering Report
Variables            AGR to TC
Clustering Method    Single Linkage (Nearest Neighbor)
Distance Type        Euclidean
Scale Type           Standard Deviation
```



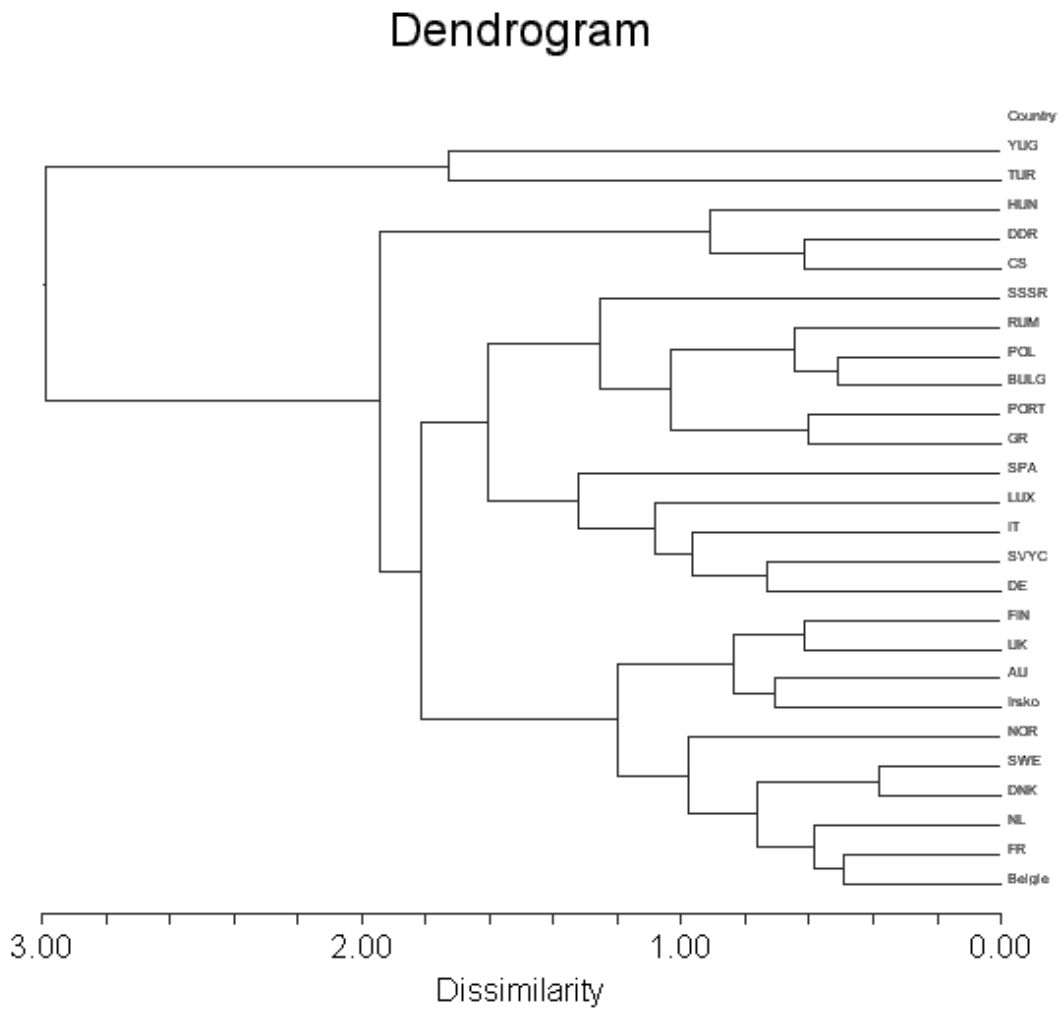
Na dendrogramu vidíme, jak postupně byly vytvářeny shluky. Jako první se spojily nejpodobnější objekty, tj. Dánsko a Švédsko, pak Belgie s Francií atd. Dendrogram ukazuje typické chování metody nejbližšího souseda, kdy k velkým shlukům jsou připojovány jednotlivé objekty nebo shluky s malým počtem objektů. Na úrovni nepodobnosti (vzdálenosti) zhruba 0,85 jsou země rozděleny do 6 shluků:

1. západoevropské země s výjimkou Španělska a Lucemburska
2. Lucembursko
3. země bývalého socialistického bloku
4. Španělsko
5. bývalá Jugoslávie
6. Turecko

Příklad 12.4

`Hierarchical Clustering Report`

```
Variables            AGR to TC
Clustering Method    Complete Linkage (Furthest Neighbor)
Distance Type        Euclidean
Scale Type           Standard Deviation
```

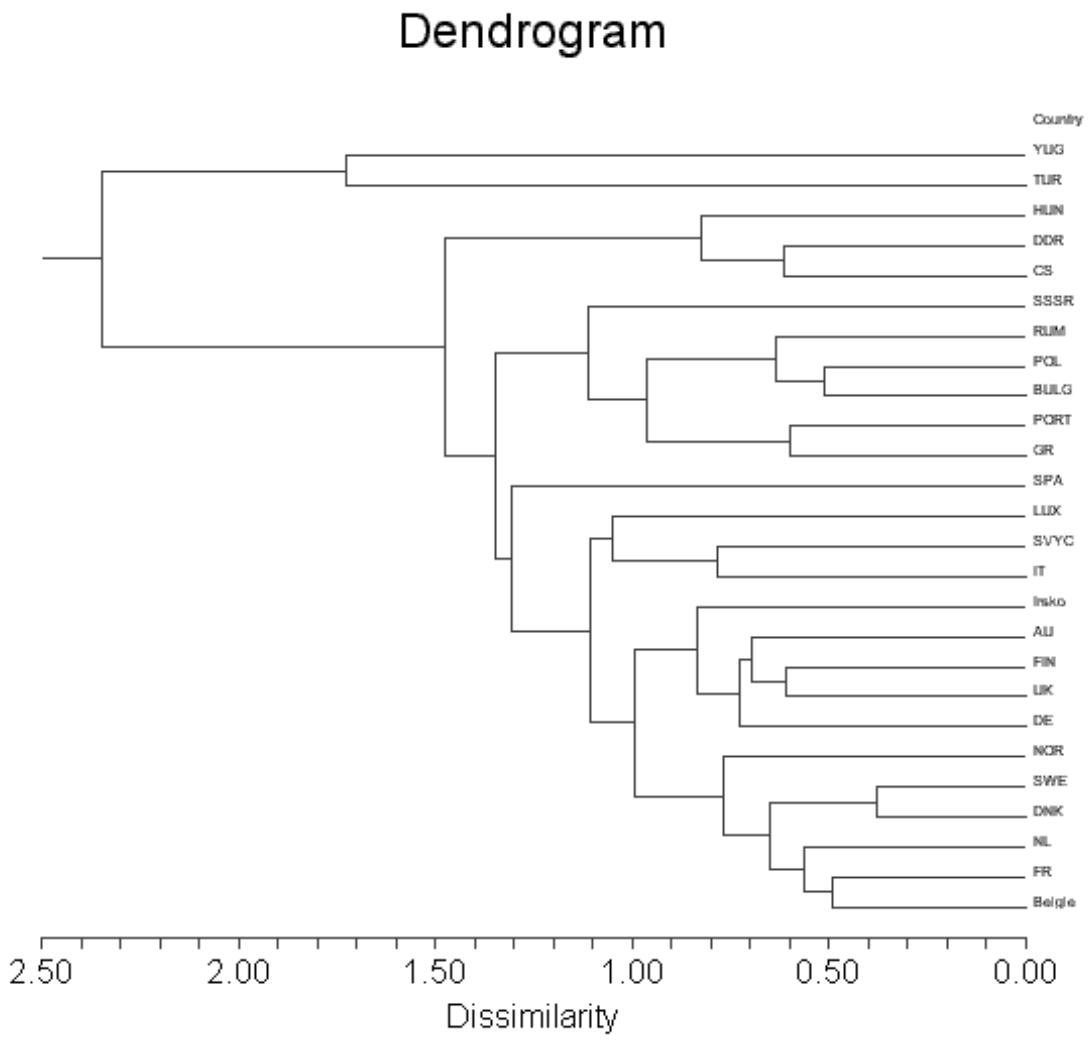


Jako první se opět spojily nejpodobnější objekty, tj. Dánsko a Švédsko, pak Belgie s Francií atd. Metoda nejvzdálenějšího souseda má tendenci vytvářet kompaktnější shluky, ve kterých je počet objektů rovnoměrnější. Na úrovni nepodobnosti (vzdálenosti) zhruba 0,95 jsou země rozděleny do 6 shluků:

1. západoevropské Belgie až Finsko, seznam viz dendrogram
2. SRN, Švýcarsko, Itálie, Lucembursko a Španělsko
3. Řecko, Portugalsko a přímořské země bývalého socialistického bloku s výjimkou NDR
4. Československo, NDR a Maďarsko
5. Turecko
6. bývalá Jugoslávie

Příklad 12.5

```
Hierarchical Clustering Report
Variables            AGR to TC
Clustering Method    Simple Average (Weighted Pair-Group)
Distance Type        Euclidean
Scale Type           Standard Deviation
```



Metoda průměrné vzdálenosti vedla k výsledkům, které jsou podobné metodě nejvzdálenějšího souseda, rozdíly jsou především uprostřed shlukovací procedury, např. Španělsko se ke shluku západoevropských zemím připojilo později než SSSR ke shluku Rumunska, Polska atd.

12.3.2 Nehierarchické metody

Mezi nejpopulárnější nehierarchické metody patří metoda k-means . Počet shluků k je předem znám, objekty se rozdělují do shluků tak, aby rozptyl uvnitř shluků (within sum of squares) byl co nejmenší. Jde tedy o to, abychom nalezli takové přiřazení objektů do shluků tak, aby stopa matice $\mathbf{W}$ byla minimální.

$$\mathbf{W} = \sum_{g=1}^{k} \mathbf{W}_g, \tag{55}$$

$\mathbf{W}_g$ je Wishartova matice pro shluk g, tj.

$$\mathbf{W}_g = \sum_{j=1}^{n_g} (\mathbf{x}_j^{(g)} - \bar{\mathbf{x}}^{(g)})(\mathbf{x}_j^{(g)} - \bar{\mathbf{x}}^{(g)})^T, \tag{56}$$

kde $\mathbf{x}_j^{(g)}$ je vektor hodnot veličin j-tého objektu v g-tém shluku, $\bar{\mathbf{x}}^{(g)} = \left(\sum_{j=1}^{n_g} \mathbf{x}_j^{(g)}\right)/n_g$ je vektor průměrů (centroid) g-tého shluku. Kritériem, jež má být minimalizovano, je pak

$$\mathrm{TRW} = \mathrm{tr}(\mathbf{W}). \tag{57}$$

Najít globální minimum je algoritmicky obtížný problém, který neumíme vyřešit v polynomiálním čase. Obvykle se užívá se Hartiganův algoritmus *k-means*, který umí najít přijatelné lokální minimum pro většinu jednodušších klasifikačních úloh nebo se v poslední době pro optimalizaci klasifikace využívají evoluční algoritmy.

Algoritmus *k-means* je velmi jednoduchý:

1. Nejdříve se k centroidů (těžišť shluků) zvolí náhodně, buď se vybere náhodně k objektů ze zadaných dat nebo se objekty náhodně klasifikují do k shluků a spočítají jejich těžiště (vektor průměrů).
2. Objekty se zařadí do shluku, jehož těžišti jsou nejbližší a spočítá se nové těžiště každého shluku.
3. Krok 2 se opakuje tak dlouho, dokud dochází ke změně klasifikace objektů.

Příklad 12.6 Využijeme opět data ze souboru EMPLOY.XLS. Stručné výsledky pro šest shluků následují.

```
K-Means Cluster Analysis Report

Iteration Section
Iteration No. of    Percent of    Bar Chart
No.       Clusters  Variation    of Percent
1          2        72.59        ||||||||||||||||||||
2          3        52.48        ||||||||||||||||
3          4        43.44        ||||||||||||||
4          5        36.60        ||||||||||||
5          6        33.70        ||||||||||||
```

Vidíme, jak s počtem shluků klesá podíl variability uvnitř shluků (within sum of squares) na celkové variabilitě. Nejvýraznější skok je mezi 2 a 3 shluky,pak už se rychlost snižování zmenšuje.

Průměry (tj. souřadnice těžiště shluků) jsou v následující tabulce. Podobnou tabulku volitelně obsahuje výstup z procedury k-means [14] i pro směrodatné odchylky uvnitř shluků.

```
Cluster Means
Variab Clust1    Clust2   Clust3   Clust4  Clust5  Clust6
AGR     31.7      12.9     20.13    9.76    6.78    57.75
MIN      0.95      0.97     2.45    1.20    0.4      1.1
MAN     25.17     26.75    31.68   31.9    24.04    12.35
PS       0.62      1.35     1.08    0.78    0.82     0.6
CON      9.17      7.7      8.33    8.98    8.44     3.85
SER     10.1      16.3      8.56   17.06   16.6      5.8
FIN      3.72      4.72     0.85    4.50    6.04     6.2
SPS     12.8      22.55    19.18   19.92   29.46     8.6
TC       5.72      6.77     7.71    5.88    7.46     3.6
Count    4         4        6       5        5       2
```

Země byly do shluků zařazeny takto:

1. Řecko, Portugalsko, Španělsko, Rumunsko
2. Irsko, Británie, Rakousko, Finsko
3. Bulharsko, Československo, NDR, Maďarsko, Polsko, SSSR
4. Francie, SNR, Itálie, Lucembursko, Švýcarsko

5. Belgie, Dánsko, Holandsko, Norsko, Švédsko
6. Turecko, Jugoslávie

Výsledky se s tím, co poskytly hierarchické procedury, shodují jen částečně, ovšem podstatné rysy možné klasifikace jsou společné. Právě porovnání výsledků více shlukovacích procedur a nalezení jejich společných rysů je užitečné pro úvahy o možné klasifikaci.

K této úloze se ještě vrátíme v kapitole o hlavních komponentách.

12.4 Analýza hlavních komponent

Analýza hlavních komponent (Principal components analysis, PCA) je jedna z metod redukce dimenze úlohy. Snaží se vysvětlit celkový rozptyl vektoru náhodných veličin, resp. jeho podstatnou část pomocí méně veličin.

$\mathbf{x} = (X_1, X_2, \ldots, X_p)^T$ náhodný vektor
$\mathbf{V} = [\mathrm{cov}(X_i, X_j)] = [\sigma_{ij}]$,
$i, j = 1, 2, \ldots, p$ kovarianční (varianční) matice

Kovarianční matice $\mathbf{V}$ typu $(p \times p)$ je symetrická (vlastní čísla jsou reálná) a pozitivně semidefinitní, tj. $\mathbf{y}^T\mathbf{V}\mathbf{y} \geq 0$ pro jakýkoliv vektor $\mathbf{y} \neq 0$. Tzn., že všechna vlastní čísla jsou nezáporná, – důkaz viz např. Anděl, str. 28 [2]

Vlastní čísla můžeme uspořádat:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$$

Matici $\mathbf{V}$ můžeme napsat jako

$$\mathbf{V} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T, \qquad \mathbf{U}\mathbf{U}^T = \mathbf{I}, \tag{58}$$

kde $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_p)$ a k – tý sloupec matice $\mathbf{U}$ je vlastní vektor $\mathbf{v}_k$ matice $\mathbf{V}$, který přísluší vlastnímu číslu λ_k a pro který platí $\mathbf{v}_k^T\mathbf{v}_k = 1$. Tedy $\mathbf{v}_k$ jsou ortonormální vlastní vektory kovarianční matice $\mathbf{V}$, platí

$$\mathbf{V}\mathbf{v}_k = \lambda_k\mathbf{v}_k$$
$$\mathbf{V} = \lambda_1\mathbf{v}_1\mathbf{v}_1^T + \lambda_2\mathbf{v}_2\mathbf{v}_2^T + \ldots + \lambda_p\mathbf{v}_p\mathbf{v}_p^T$$
$$\mathbf{I} = \mathbf{v}_1\mathbf{v}_1^T + \mathbf{v}_2\mathbf{v}_2^T + \ldots + \mathbf{v}_p\mathbf{v}_p^T$$

Vektory $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$ tvoří bázi prostoru $\mathcal{R}^p$, takže libovolný vektor $\mathbf{y} \in \mathcal{R}^p$ lze vyjádřit jako

$$\mathbf{y} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_p\mathbf{v}_p$$

Platí pro každý vektor $\mathbf{y} \in \mathcal{R}^p$ jednotkové délky ($\mathbf{y}^T\mathbf{y} = 1$), že kvadratická forma $\mathbf{y}^T\mathbf{V}\mathbf{y} \leq \lambda_1$, při čemž rovnost platí pro $\mathbf{c} = \mathbf{v}_1$, tj. pro první vlastní vektor – viz Anděl, str. 297 [2].

Celková variabilita náhodného vektoru $\mathbf{x} = (X_1, X_2, \ldots, X_p)^T$ je součet diagonálních prvků kovarianční matice (rozptylů jednotlivých náhodných veličin):

$$\sigma^2 = \mathrm{var}(X_1) + \mathrm{var}(X_2) + \ldots + \mathrm{var}(X_p)$$

Hledáme novou náhodnou veličinu, která vznikne lineární transformací vektoru $\mathbf{x} = (X_1, X_2, \ldots, X_p)^T$, tj. vlastně hledáme $\mathbf{c} \in \mathcal{R}^p$, $\mathbf{c}^T\mathbf{c} = 1$, aby náhodná veličina měla

co největší rozptyl. Tím tato nová veličina vyčerpá co největší část celkové variability, tzn. její rozptyl je roven λ_1. Tedy $\mathbf{c} = \mathbf{v}_1$. Náhodnou veličinu $Y_1 = \mathbf{v}_1^T\mathbf{x}$ nazýváme první hlavní komponentou.

Pak hledáme další náhodnou veličinu, tj. další $\mathbf{c} \in \mathcal{R}^p$, $\mathbf{c}^T\mathbf{c} = 1$, aby náhodná veličina $\mathbf{c}^T\mathbf{x}$ byla nekorelovaná s $Y_1 = \mathbf{v}_1^T\mathbf{x}$. Tomu vyhovuje $\mathbf{c} = \mathbf{v}_2$, takže druhá hlavní komponenta je $Y_2 = \mathbf{v}_2^T\mathbf{x}$ a její rozptyl je $\mathrm{var}(Y_2) = \mathrm{var}(\mathbf{v}_2^T\mathbf{x}) = \lambda_2$. Podobně i pro třetí a další hlavní komponenty.

Celková variabilita

$$\sigma^2 = \mathrm{var}(X_1) + \mathrm{var}(X_2) + \ldots + \mathrm{var}(X_p) =$$

$$= \mathrm{var}(Y_1) + \mathrm{var}(Y_2) + \ldots + \mathrm{var}(Y_p) = \sum_{i=1}^{p} \lambda_i$$

Relativní podíl celkové variability vysvětlovaný i-tou hlavní komponentou je λ_i/σ^2. Prvních k hlavních komponent vysvětluje $\sum_{i=1}^{k} \lambda_i/\sigma^2$ z celkové variability.

V praktických aplikacích se vychází z výběrové kovarianční matice nebo častěji z výběrové korelační matice (abychom se vyhnuli vlivu volby jednotek, ve kterých měříme veličiny, na hodnoty výběrových rozptylů). Korelační matice je vlastně kovarianční matice standardizovaných veličin. Pak celková variabilita je rovna počtu veličin,

$$\sigma^2 = \sum_{i=1}^{p} \lambda_i = p$$

Příklad 12.7 Základní možnosti analýzy hlavních komponent ukážeme na příkladu o struktuře zaměstnanosti ve 26 evropských zemích, data jsou v souboru EMPLOY.XLS. Vycházíme z výběrové korelační matice.

```
Principal Components Report

Database   C:\avdat\employ.S0

Eigenvalues
                  Individual Cumulative
No.   Eigenvalue   Percent   Percent   Scree Plot
1      3.487151    38.75      38.75    ||||||||
2      2.130173    23.67      62.41    |||||
3      1.098958    12.21      74.63    |||
4      0.994483    11.05      85.68    |||
5      0.543218     6.04      91.71    ||
6      0.383428     4.26      95.97    |
7      0.225754     2.51      98.48    |
8      0.136790     1.52     100.00    |
9      0.000046     0.00     100.00    |
```

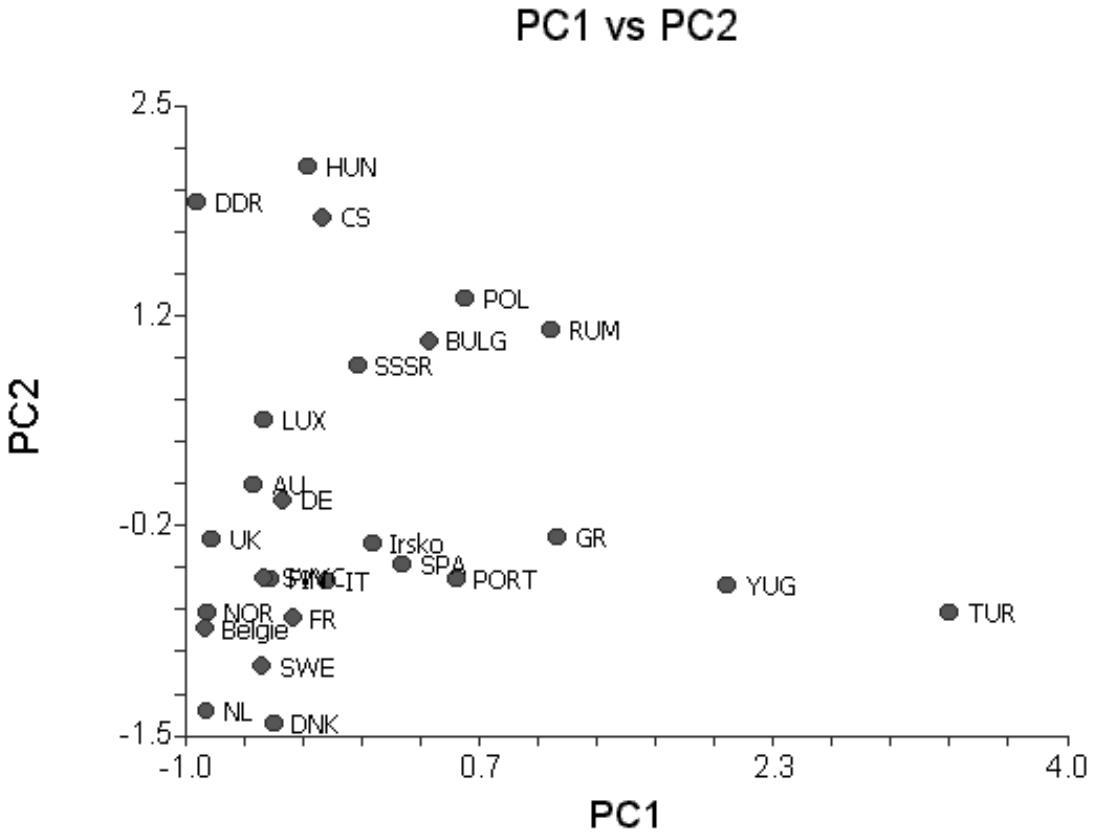
Vidíme, že tři vlastní čísla jsou větší než 1, čtvrté téměř rovno jedné. První dvě hlavní komponenty vysvětlují 62 % z celkové variability, první čtyři už 85 % celkové variability.

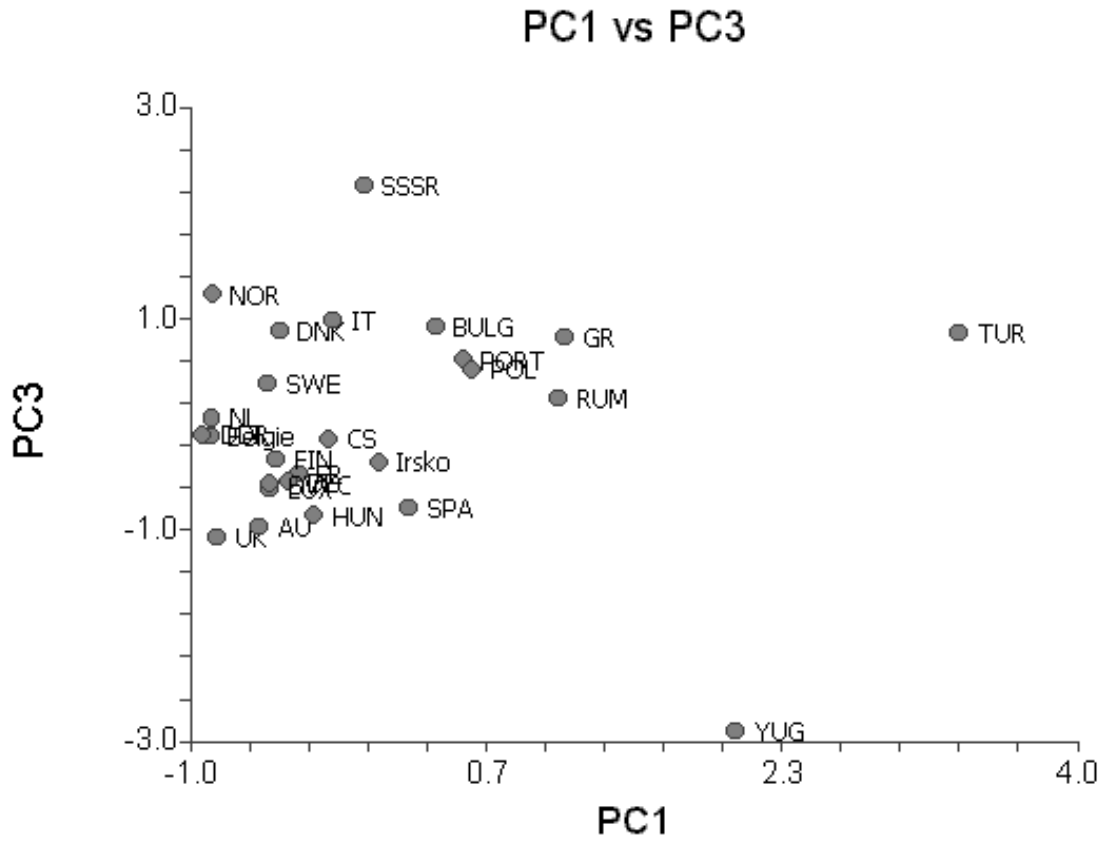
První čtyři vlastní vektory jsou v následující tabulce:

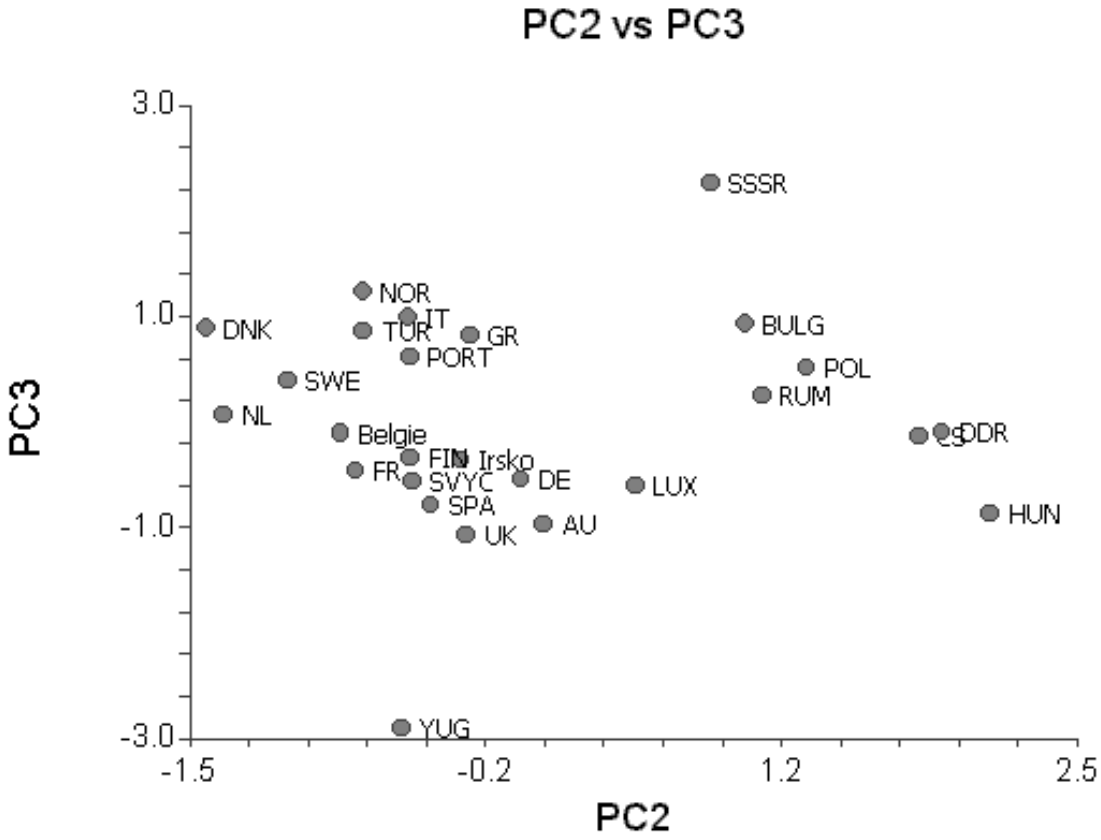
```
Eigenvectors
Variables   Factor1     Factor2     Factor3    Factor4
AGR          0.523791    0.053594    0.048674   0.028793
MIN          0.001323    0.617807   -0.201100   0.064085
MAN         -0.347495    0.355054   -0.150463  -0.346088
PS          -0.255716    0.261096   -0.561083   0.393309
CON         -0.325179    0.051288    0.153321  -0.668324
SER         -0.378920   -0.350172   -0.115096  -0.050157
FIN         -0.074374   -0.453698   -0.587361  -0.051567
SPS         -0.387409   -0.221521    0.311904   0.412230
TC          -0.366823    0.202592    0.375106   0.314372
```

Na dvourozměrných grafech v rovinách dvojic prvních tří hlavních komponent vidíme, že při takto podstatném snížení dimenze, ve kterých objekty zobrazujeme, lze sledovat podobnosti a odlišnosti zobrazených objektů. Zejména na obrázku prvních

dvou hlavních komponent jsou viditelné shluky objektů (zemí) v souladu s klasifikací nalezenou shlukovou analýzou.






PC2 vs PC3
PC3
PC2
3.0
1.0
-1.0
-3.0
-1.5
-0.2
1.2
2.5
SSSR
NOR
IT
DNK
TUR
GR
BULG
PORT
POL
SWE
RUM
NL
Belgie
DDR
FIN
Irsko
FR
SVYC
DE
LUX
SPA
HUN
AU
UK
YUG

12.5 Faktorová analýza

Faktorová analýza je jedna z metod redukce dimenze úlohy. Snaží se vysvětlit kovarianční strukturu (korelační matici) vektoru náhodných veličin, pomocí méně tzv. faktorů, tj. jakýchsi skrytých veličin, které nemůžeme nebo neumíme přímo měřit.

Uvažujme, že naměřená data jsou matice $\mathbf{X}$ typu $n \times p$ (n objektů, p veličin). Standardizovaná matice dat je matice $\mathbf{Z}$ opět typu $n \times p$, kde

$$z_{ij} = \frac{x_{ij} - \bar{x_j}}{s_j}$$

$\bar{x_j}$, s_j jsou výběrový průměr a směrodatná odchylka j-té veličiny.

Model faktorové analýzy můžeme pak zapsat

$$z_{ij} = a_{j1}f_{i1} + a_{j2}f_{i2} + \ldots + a_{jm}f_{im} + e_{ij}, \quad m < p$$

tj. naměřenou hodnotu j-té veličiny na i-tém objektu vysvětlujeme jako vážený součet m faktorů a nějaké složky, která těmito faktory vysvětlit nelze.

Zavedeme následující matice:

- $\mathbf{A}$ je typu $(p \times m)$, má prvky a_{jk}, označují se jako faktorové zátěže (sycení, saturace, loadings)
- $\mathbf{F}$ je typu $(n \times m)$, má prvky f_{ik}, říká se jim faktorové skóry
- $\mathbf{E}$ je typu $(n \times p)$, má prvky e_{ij}, jsou to rezidua, tj. to z hodnot z_{ij}, co nemůžeme vysvětlit pomocí faktorů

Maticově pak můžeme model faktorové analýzy zapsat takto:

$$\mathbf{Z} = \mathbf{F}\mathbf{A}^T + \mathbf{E}$$

Předpokládejme, že faktory jsou ortonormální (nekorelované, jednotkové délky), tzn. $\mathbf{F}^T\mathbf{F} = \mathbf{I}$. Označme $\mathbf{U} = \mathbf{E}^T\mathbf{E}$. $\mathbf{U}$ je diagonální matice typu $p \times p$, tedy $\mathbf{U} = diag(u_1, u_2, \ldots, u_p)$, kde $u_j = \sum_{i=1}^{n} e_{ij}^2$, tj. variabilita j-té veličiny, kterou nelze vysvětlit faktory, tzv. specificita.

Pak

$$h_j = 1 - u_j = \sum_{k=1}^{m} a_{jk}^2$$

je tzv. komunalita j-té veličiny, tj. variabilita vysvětlitelná faktory.

Korelační matici můžeme vyjádřit jako

$$\mathbf{R} = \mathbf{A}\mathbf{A}^T + \mathbf{U}$$

Matice $\mathbf{A}\mathbf{A}^T$ vysvětluje korelační matici až na diagonální prvky, kde místo jedniček jsou komunality.

Faktorová analýza hledá model, aby

- příspěvek specifických faktorů (u_j) byl co nejmenší
- faktorové zátěže byly v absolutní hodnotě co nejbližší jedné nebo nule
- počet faktorů byl co nejmenší (podstatně menší než počet veličin)

Tyto požadavky jsou ve vzájemném rozporu a „umění“ faktorové analýzy spočívá v nalezení vhodného a přijatelného kompromisu:

extrakce faktorů - stanovit jejich počet, např. z vlastních čísel korelační matice

problém komunalit $R_j^2 \leq h_j \leq 1$, kde R_j^2 je koeficient mnohonásobné korelace (j-tá veličina na ostatních)

rotace faktorů tj. nalezení „ jednoduché struktury“, aby faktorové zátěže v absolutní hodnotě byly co nejbližší jedné nebo nule

Ortogonální rotace znamená najít takovou transformaci matice $\mathbf{A}$ na $\mathbf{B} = \mathbf{AT}$, aby $\mathbf{B}$ splňovalo požadavek jednoduché struktury, $\mathbf{T}$ je ortogonální matice, tj. $\mathbf{T}^T\mathbf{T} = \mathbf{I}$, takže $\mathbf{A}\mathbf{A}^T = \mathbf{B}\mathbf{B}^T$.

Je mnoho možných metod takové rotace faktorů, nejběžnější je tzv. VARIMAX, založená na maximalizaci výrazu

$$\frac{1}{p}\sum_{j=1}^{m}\sum_{i=1}^{p}(a_{ij}^2 - a_{\cdot j}^2)^2,$$

kde $a_{\cdot j}^2 = \frac{1}{p}\sum_{i=1}^{p} a_{ij}^2$, tj. průměr čtverců zátěží pro j-tý faktor.

Po rotaci můžeme nahlédnout, která veličina „patří“ kterému faktoru (faktorové zátěže v absolutní hodnotě blízké jedné), případně faktorové zátěže jednotlivých veličin vynést do grafů pro dvojice faktorů. Lze pak spočítat i faktorové skóry a na grafech faktorových skórů hledat, zda zobrazené objekty nevytvářejí nějaké shluky signalizující možný rozklad pozorovaných objektů do dvou či více podskupin.

Příklad 12.8 Data pro tento příklad jsou převzata z knihy [6]. Původní data byla výsledky dosažené ve výběru 220 chlapců v šesti předmětech - galštině, angličtině, dějepisu, aritmetice, algebře a geometrii. K dispozici pro analýzu máme výběrovou korelační matici (dolní trojúhelník, předměty jsou ve výše uvedeném pořadí):

```
1.00
0.44   1.00
```

```
0.41   0.35   1.00
0.29   0.35   0.16   1.00
0.33   0.32   0.19   0.59   1.00
0.25   0.33   0.18   0.47   0.46   1.00
```

Abychom si uvědomili rozdíly mezi analýzou hlavních komponent a faktorovou analýzou, uvádíme nejdříve stručné výsledky analýzy hlavních komponent, která vysvětluje celkový rozptyl, tedy hlavní diagonálu korelační matice.

```
Principal Components Report

Database     C:\avdat\subject.S0

Eigenvalues
                      Individual    Cumulative
No.     Eigenvalue    Percent       Percent    Scree Plot
1        2.728683      45.48         45.48     ||||||||||
2        1.128792      18.81         64.29     ||||
3        0.615291      10.25         74.55     |||
4        0.602809      10.05         84.59     |||
5        0.522514       8.71         93.30     ||
6        0.401910       6.70        100.00     ||


Factor Loadings
Variables     Factor1       Factor2
Gaelic       -0.660803     -0.444475
English      -0.688465     -0.289771
History      -0.516356     -0.639552
Arithmetic   -0.735620      0.417018
Algebra      -0.741868      0.372759
Geometry     -0.678168      0.354100
```

Dvě vlastní čísla jsou větší než jedna, pro faktorovou analýzu tedy zvolíme počet faktorů 2, rotaci *varimax* a dostaneme následující výsledky:

```
Factor Analysis Report

Database         C:\avdat\subject.S0

Eigenvalues after Varimax Rotation
                    Individual   Cumulative
No.    Eigenvalue   Percent      Percent     Scree Plot
1      1.596863      56.94        56.94      ||||||||||||
2      1.207981      43.08       100.02      |||||||||
3      0.050820       1.81       101.83      |
4      0.011910       0.42       102.26      |
5     -0.008657      -0.31       101.95      |
6     -0.054642      -1.95       100.00      |
```

Vlastní čísla se týkají matice $\mathbf{A}\mathbf{A}^T = \mathbf{R} - \mathbf{U}$ po rotaci, cílem faktorové analýzy je především vysvětlit korelační strukturu, tj. mimodiagonální prvky korelační matice. Celkový vysvětlený rozptyl je roven jen součtu komunalit.

Po rotaci jsou faktorové zátěže následující:

```
Factor Loadings after Varimax Rotation
Variables      Factor1       Factor2
Gaelic        -0.233132     -0.659253
English       -0.322810     -0.552071
History       -0.084713     -0.589192
Arithmetic    -0.765986     -0.170657
Algebra       -0.718105     -0.214689
Geometry      -0.573340     -0.214994
```

Absolutní hodnoty faktorových zátěží jsou znázorněny graficky:

```
Bar Chart of Absolute Factor Loadings
after Varimax Rotation

Variables      Factor1             Factor2
Gaelic         |||||               |||||||||||||
English        |||||||             ||||||||||||
History        ||                  ||||||||||||
Arithmetic     ||||||||||||||||    ||||
Algebra        ||||||||||||||||    |||||
```

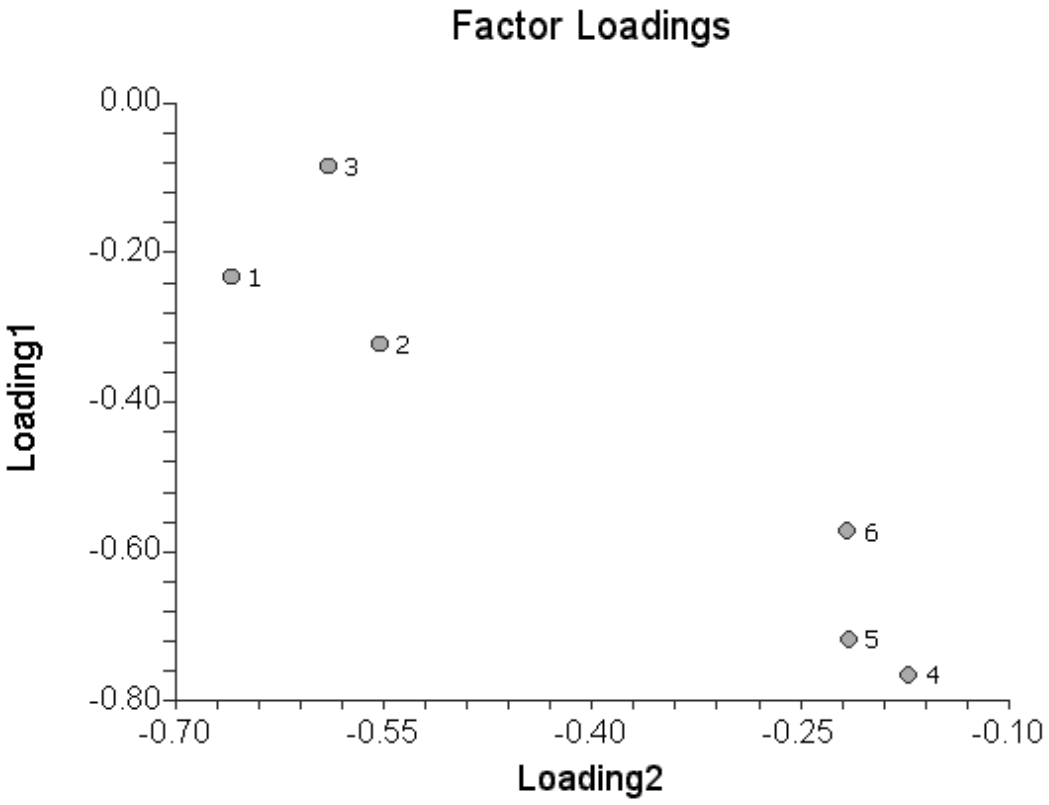
```
Geometry        ||||||||||||          |||||
```

Faktor 1 je sycen veličinami aritmetika, algebra a geometrie, faktor 2 veličinami galština, angličtina a dějepis.

Podíly faktorů na komunalitách ukazuje následující tabulka:

```
Communalities after Varimax Rotation
     Factors
Variables      Factor1       Factor2       Communality
Gaelic         0.054350      0.434614      0.488965
English        0.104206      0.304783      0.408989
History        0.007176      0.347147      0.354324
Arithmetic     0.586735      0.029124      0.615859
Algebra        0.515675      0.046091      0.561766
Geometry       0.328719      0.046222      0.374942
```

Vztah veličin k faktorům je graficky znázorněn v grafu faktorových zátěží, ve kterém vidíme shluk veličin 1, 2 a 3 s nízkými absolutními hodnotami zátěží prvního faktoru, které sytí především druhý faktor, a shluk veličin 4, 5 a 6 s nízkými absolutními hodnotami zátěží druhého faktoru, sytících především první faktor.



Korelační strukturu pozorovaných dat (studijních výsledků v šesti předmětech) lze tedy uspokojivě vysvětlit dvě faktory. První faktor vyjadřuje matematickou dispozici žáka, druhý dispozici jazykově-humanitní.

Shrnutí

- *úlohy řešené mnohorozměrnými metodami*
- *test shody vektorů středních hodnot*
- *metody klasifikace objektů, diskriminační analýza, logistická regrese*
- *metody shlukování, numerická podobnost, hierarchické a nehierarchické metody shlukování*
- *redukce dimenze, analýza hlavních komponent, faktorová analýza*

Kontrolní otázky

1. *V čem jsou shodné a v čem odlišné cíle diskriminační analýzy a shlukové analýzy?*
2. *Proč se při hledání lineární diskriminační funkce musí lišit střední hodnoty skupin?*
3. *Jaké jsou výhody a nevýhody lineární diskriminační funkce oproti jiným klasifikačním pravidlům (např. logistická regrese, neuronové sítě ap.)?*
4. *Jaké jsou rozdíly mezi hierarchickými a nehierarchickými metodami?*
5. *V čem se liší faktorová analýza od analýzy hlavních komponent?*
6. *Jak zjistíte souřadnice jednotlivých objektů (řádků datové matice) v rovině prvních dvou hlavních komponent?*
7. *Jak určit počet faktorů?*
8. *Co je to komunalita?*
9. *Co jsou faktorové zátěže? Co lze vyčíst z grafu faktorových zátěží?*
10. *Jakou část celkové variability vysvětluje první hlavní komponenta?*

Korespondenční úloha

Korespondenční úlohy budou zadávány ke každému kursu samostatně.