

# **Osnova přednášky „Analýza hlavních komponent“**

1. Motivace
2. Cíl PCA
3. Geometrický význam PCA
4. Označení a ilustrační příklad
5. Základní pojmy v PCA
6. Doporučený postup při PCA
7. Ilustrační příklad
8. Nejdůležitější problémy v PCA
9. Příklad

**1. Motivace:** V řadě situací se setkáváme s tím, že počet sledovaných proměnných je příliš vysoký a pro interpretaci nepřehledný. Pro snazší hodnocení výsledků je vhodné zkoumat, zda by se velký počet původních proměnných nedal nahradit menším počtem nově vytvořených proměnných bez podstatné ztráty informace. K tomuto účelu můžeme použít metodu hlavních komponent (Principal Component Analysis – PCA), kterou popsal v r. 1901 Karl Pearson a ve 30. letech 20. století ji dále rozvinul Harold Hotelling, (1895 – 1973), americký matematik a statistik.



PCA je považována za nejužitečnější metodu pro posouzení a prověření kvality vícerozměrných dat.

## 2. Cíl PCA:

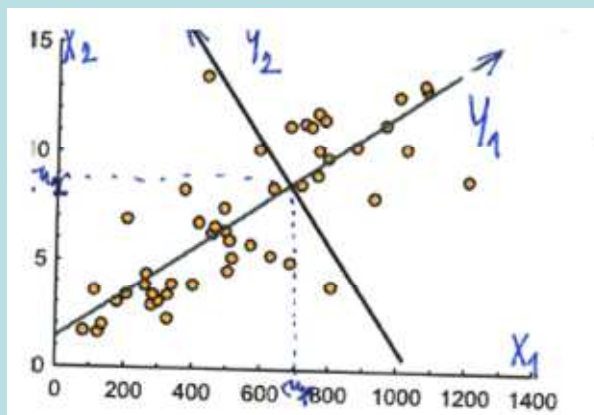
vyjádřit informace o variabilitě obsažené v datovém souboru pomocí několika málo nových znaků získaných jako lineární kombinace znaků původních.

1. Nové znaky (**hlavní komponenty**) jsou uspořádané podle svého klesajícího rozptylu.
2. Hlavní komponenty jsou nekorelované.
3. První hlavní komponenta je nejdůležitější, vysvětlí co nejvíce z celkové variability.
4. Každá další hlavní komponenta vysvětlí co nejvíce ze zbývajících variability, takže poslední hlavní komponenta je nejméně důležitá.
5. Je-li  $p$  počet původních znaků a rozhodneme-li se použít právě  $m$  ( $m \leq p$ ) hlavních komponent, pak požadujeme, aby těchto  $m$  hlavních komponent vysvětlovalo dostatečnou část celkové variability. (O kritériích pro stanovení vhodného  $m$  se zmíníme později. Zkušenosti s používáním PCA ukazují, že případ, kdy  $m = 1$  až 4 je poměrně častý.)
6. Hlavní komponenty lze interpretovat jako hlavní osy  $p$ -rozměrného elipsoidu  $\mathbf{x}'\mathbf{S}^{-1}\mathbf{x} = \text{konst.}$ , kde  $\mathbf{x}$  je vektor původních znaků a  $\mathbf{S}$  je jeho varianční matice.

**Důležitý předpoklad použití PCA:** V datovém souboru musí existovat mezi znaky **dostatečně silná korelace**, aby bylo možno tuto redukci provést.

### 3. Geometrický význam PCA

Analýza hlavních komponent může být chápána jako transformace z původního do nového souřadnicového systému, jehož osy jsou tvořeny hlavními komponentami. Osy procházejí směry maximálního rozptylu, protože podmínka nezávislosti komponent vede ke kolmosti os.



*Pozorované body můžeme vyjádřit v původních souřadnicích  $X_1, X_2$ , nebo v nových souřadnicích  $Y_1, Y_2$ . Je vidět, že směr největší variability v datech je totožný se směrem osy  $Y_1$ . Na ni kolmá osa  $Y_2$  je ve směru nejmenší možné zbylé variability. (Body jsou zobrazeny v dimenzi  $p = 2$ , tedy více směrů nového souřadného systému nemůže být.) Pokud bychom chtěli snížit dimenzi prostoru, pak bychom všechny body vyjádřili pouze prostřednictvím souřadnice  $Y_1$  i když bychom tím část informace o variabilitě souboru ztratili.*

## 4. Označení

Máme p-rozměrný datový soubor ve formě matice n x p:

$$\begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \cdots & \cdots & \cdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix}.$$

$$\mathbf{x}_i = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{ip} \end{pmatrix} - \text{vektor pozorování } i\text{-tého objektu, } i = 1, 2, \dots, n$$

$$m_j = \frac{1}{n} \sum_{i=1}^n X_{ij} - \text{průměr } j\text{-tého znaku, } j = 1, 2, \dots, p$$

$$s_j^2 = \frac{1}{n-1} \sum_{i=1}^n (X_{ij} - m_j)^2 - \text{rozptyl } j\text{-tého znaku, } j = 1, 2, \dots, p$$

$$z_{ij} = \frac{X_{ij} - m_j}{s_j} - (i,j)\text{-tá standardizovaná hodnota, } i = 1, 2, \dots, n, j = 1, 2, \dots, p$$

$$\mathbf{z}_i = \begin{pmatrix} z_{i1} \\ \vdots \\ z_{ip} \end{pmatrix} - \text{vektor standardizovaných pozorování } i\text{-tého objektu, } i = 1, 2, \dots, n$$

$$\mathbf{m} = \begin{pmatrix} m_1 \\ \vdots \\ m_p \end{pmatrix} - \text{vektor průměrů}$$

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n \begin{pmatrix} x_{i1} - m_1 \\ \vdots \\ x_{ip} - m_p \end{pmatrix} (x_{i1} - m_1, \dots, x_{ip} - m_p) - \text{výběrová varianční matice}$$

$$\mathbf{R} = \frac{1}{n-1} \sum_{i=1}^n \begin{pmatrix} z_{i1} \\ \vdots \\ z_{ip} \end{pmatrix} (z_{i1}, \dots, z_{ip}) - \text{výběrová korelační matice}$$

( $\mathbf{S}$  a  $\mathbf{R}$  jsou čtvercové symetrické matice řádu  $p$ .)

**Příklad:** Na pěti objektech byly zjišťovány hodnoty dvou znaků. Datový soubor je tvaru

$$\begin{pmatrix} 3 & 7 \\ 5 & 6 \\ 6 & 8 \\ 7 & 10 \\ 9 & 9 \end{pmatrix}.$$

Vypočítejte výběrové průměry, výběrové rozptyly, vektor průměrů, výběrovou varianční matici a výběrovou korelační matici.

**Řešení:**

**Vektor průměrů:**

Popisné statistiky (Dva_znaky.sta)	
Proměnná	Průměr
X1	6
X2	8

**Varianční matice:**

Kovariance (Dva_znaky.sta)		
Proměnná	X1	X2
X1	5,0	2,5
X2	2,5	2,5

**Korelační matice:**

Korelace (Dva_znaky.sta)		
Proměnná	X1	X2
X1	1,000000	0,707107
X2	0,707107	1,000000

## 5. Základní pojmy v metodě hlavních komponent

$\mathbf{A}$  - čtvercová matice řádu  $p$ .

**Vlastní číslo matice  $\mathbf{A}$**  – takové číslo  $\lambda$ , které pro libovolný nenulový vektor  $\mathbf{v}$  typu  $p \times 1$  splňuje rovnici  $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ .

**Vlastní vektor matice  $\mathbf{A}$**  – vektor  $\mathbf{v}$ .

**Charakteristický polynom matice  $\mathbf{A}$**  - determinant  $|\mathbf{A} - \lambda\mathbf{I}|$ .

**Stopa matice  $\mathbf{A}$**  - součet jejích diagonálních prvků (značí se  $\text{Tr}(\mathbf{A})$ ).

### Výpočet vlastních čísel matice $\mathbf{A}$

Rovnici  $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$  upravíme na tvar  $(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}$ .

Tato soustava  $p$  rovnic má netriviální řešení, právě když charakteristický polynom matice  $\mathbf{A}$  je roven 0.



## Získání hlavních komponent

Nechť výběrová varianční matice  $\mathbf{S}$  má vlastní čísla  $l_1, \dots, l_p$  a vlastní vektory  $\mathbf{v}_1, \dots, \mathbf{v}_p$ , přičemž

$$v_{j1}^2 + v_{j2}^2 + \dots + v_{jp}^2 = 1, v_{j1}v_{k1} + v_{j2}v_{k2} + \dots + v_{jp}v_{kp} = 0 \text{ pro } j \neq k.$$

(Znamená to, že vektory  $\mathbf{v}_1, \dots, \mathbf{v}_p$  jsou ortonormální.)

Bez újmy na obecnosti předpokládáme, že  $l_1 > l_2 > \dots > l_p$ .

**1. hlavní komponenta**  $Y_1$  vznikne jako lineární kombinace znaků  $X_1, \dots, X_p$ , kde koeficienty této lineární kombinace jsou souřadnice vlastního vektoru  $\mathbf{v}_1$ , tedy

$$Y_1 = v_{11}X_1 + \dots + v_{1p}X_p.$$

Rozptyl 1. hlavní komponenty je  $l_1$ .

Dosadíme-li za  $X_1, \dots, X_p$  vektory pozorování  $\mathbf{x}_i$ ,  $i = 1, \dots, n$ , dostaneme **vektor souřadnic**  $\mathbf{y}_1 = (y_{11}, \dots, y_{1n})^T$ , kde  $y_{1i} = v_{11}x_{i1} + v_{12}x_{i2} + \dots + v_{1p}x_{ip}$ ,  $i = 1, \dots, n$ .

**2. hlavní komponenta** vznikne jako lineární kombinace znaků  $X_1, \dots, X_p$ , kde koeficienty této lineární kombinace jsou souřadnice vlastního vektoru  $\mathbf{v}_2$ , tedy

$$Y_2 = v_{21}X_1 + \dots + v_{2p}X_p.$$

Přitom  $v_{11}v_{21} + v_{12}v_{22} + \dots + v_{1p}v_{2p} = 0$ , tj. 1. a 2. hlavní komponenta jsou lineárně nezávislé.

Rozptyl 2. hlavní komponenty je  $l_2$ .

Dosadíme-li za  $X_1, \dots, X_p$  vektory pozorování  $\mathbf{x}_i$ ,  $i = 1, \dots, n$ , dostaneme **vektor souřadnic**  $\mathbf{y}_2 = (y_{21}, \dots, y_{2n})^T$ , kde  $y_{2i} = v_{21}x_{i1} + v_{22}x_{i2} + \dots + v_{2p}x_{ip}$ ,  $i = 1, \dots, n$ .

.....

**j-tá hlavní komponenta** vznikne jako lineární kombinace znaků  $X_1, \dots, X_p$ , kde koeficienty této lineární kombinace jsou souřadnice vlastního vektoru  $\mathbf{v}_j$ , tedy

$$Y_j = v_{j1}X_1 + \dots + v_{jp}X_p.$$

Přitom  $v_{j1}v_{k1} + v_{j2}v_{k2} + \dots + v_{jp}v_{kp} = 0$ ,  $j = 1, \dots, k-1$ , tj. j-tá hlavní komponenta je lineárně nezávislá se všemi ostatními hlavními komponentami. Její rozptyl je  $l_j$ .

Dosadíme-li za  $X_1, \dots, X_p$  vektory pozorování  $\mathbf{x}_i$ ,  $i = 1, \dots, n$ , dostaneme **vektor souřadnic**  $\mathbf{y}_j = (y_{j1}, \dots, y_{jn})^T$ , kde  $y_{ji} = v_{j1}x_{i1} + v_{j2}x_{i2} + \dots + v_{jp}x_{ip}$ ,  $i = 1, \dots, n$ .

Vektory souřadnic všech  $p$  hlavních komponent uspořádáme do matice

$$\mathbf{T} = \begin{pmatrix} y_{11} & \cdots & y_{p1} \\ \cdots & \cdots & \cdots \\ y_{1n} & \cdots & y_{pn} \end{pmatrix}.$$

Prvek  $y_{ij}$  této matice se nazývá komponentní skóre  $j$ -té hlavní komponenty příslušné  $i$ -tému objektu.

Lze dokázat, že celková variabilita obsažená v datech je rovna stopě matice  $\mathbf{S}$ , tj. součtu vlastních

čísel  $l_1 + \dots + l_p$ . 1. hlavní komponenta tedy vyčerpává  $\frac{l_1}{l_1 + \dots + l_p} 100\%$  celkové variability. Pokud

je číslo  $\frac{l_1}{l_1 + \dots + l_p}$  dostatečně blízké 1, znamená to, že 1. hlavní komponenta dobře nahrazuje ce-

lý datový soubor. Je-li toto číslo podstatně menší než 1, musíme vzít tolik hlavních komponent, aby jejich součet dělený stopou matice  $\mathbf{S}$  byl dostatečně blízký 1. V mnoha aplikacích se stává, že i při velkém počtu znaků stačí poměrně malý počet hlavních komponent.

**Upozornění:** Před provedením metody hlavních komponent je třeba se rozhodnout, zda budeme pracovat s původními hodnotami znaků nebo standardizovanými hodnotami. Použití standardizovaných hodnot vede na analýzu výběrové korelační matice místo výběrové varianční matice. Hodí se zvláště v těch případech, kdy znaky jsou uváděny v nestejných měřicích jednotkách nebo znaky mají velmi odlišné rozptyly.

**Koeficient korelace**  $i$ -tého znaku  $X_i$  s  $k$ -tou hlavní komponentou  $Y_k$  lze vyjádřit jako

$R(X_i, Y_k) = \frac{v_{ki} \sqrt{\lambda_k}}{s_i}$ . Popisuje vztah mezi  $i$ -tým znakem a  $k$ -tou hlavní komponentou.

**Reprodukce výchozí varianční matice:**

V teorii matic se dokazuje vzorec  $S = \sum_{i=1}^p \lambda_i \mathbf{v}_i \mathbf{v}_i^T$  (tzv. **spektrální rozklad matice S**).

Rozhodneme-li se uvažovat právě  $m$  hlavních komponent ( $m \leq p$ ), pak pomocí tohoto vztahu můžeme posoudit, jak těchto  $m$  hlavních komponent reprodukuje rozptyly a kovariance původních proměnných. Lze posoudit i reziduální matici, tj. matici, kterou získáme jako rozdíl výchozí varianční matice a reprodukované varianční matice.

## 6. Doporučený postup při analýze hlavních komponent

a) Provedeme tabulkové a grafické zpracování datového souboru, abychom se blíže seznámili s daty.

b) Sestavíme korelační matici a prověříme, zda jsou korelace natolik silné, aby mělo smysl provádět analýzu hlavních komponent. K tomu slouží např. **Bartlettův test sféricity**, kde nulová hypotéza tvrdí, že výběrová korelační matice je matice jednotková.

Testová statistika  $T_0 = -n \left( 1 - \frac{2p+11}{6n} \right) \ln |\mathbf{R}|$  se za platnosti  $H_0$  asymptoticky řídí rozložením  $\chi^2 \left( \frac{p(p-1)}{2} \right)$ .

Kritický obor:  $W = \left( \chi^2_{1-\alpha} \left( \frac{p(p-1)}{2} \right), \infty \right)$

Jestliže  $t_0 \in W$ ,  $H_0$  zamítáme na hladině významnosti  $\alpha$ .

Nezamítneme-li nulovou hypotézu, neměli bychom analýzu hlavních komponent vůbec provádět.

c) Rozhodneme, kolika hlavními komponentami lze popsat datový soubor bez podstatné ztráty informace. Označme tento vhodný počet jako  $m$ . Při stanovení  $m$  můžeme použít tato pomocná kritéria:

- **Kaiserovo kritérium** - za  $m$  volíme počet těch vlastních čísel matice  $\mathbf{R}$ , která jsou větší než 1.
- **Sutinový test** (scree test) – grafická metoda, která spočívá v subjektivním posouzení vzhledu sutinového grafu (scree plot), tj. grafu znázorňujícího velikosti sestupně uspořádaných vlastních čísel matice  $\mathbf{R}$ . Objeví-li se v grafu určité zploštění, pak za  $m$  vezmeme to pořadové číslo, kde se zploštění projevilo.
- **Kritérium založené na kumulativním procentu vysvětleného rozptylu**. Požadujeme, aby vybrané hlavní komponenty vysvětlily aspoň 70 % celkového rozptylu.
- **Kritérium založené na reziduální korelační či kovarianční matici**. Požadujeme, aby prvky reziduální matice byly co možná nejmenší.

d) Pokusíme se o interpretaci prvních  $m$  hlavních komponent. Zkoumáme přitom, jak jsou jednotlivé vybrané hlavní komponenty utvořeny z původních znaků a jak s nimi korelují.

e) Vypočítáme vektory souřadnic a následně sestrojíme dvourozměrné tečkové diagramy.

**7. Příklad:** Na 24 objektech byly pozorovány znaky  $X_1$ ,  $X_2$  a  $X_3$ .

Z datového souboru byla vypočtena výběrová varianční matice  $\mathbf{S} = \begin{pmatrix} 451,39 & 271,17 & 168,70 \\ 271,17 & 171,73 & 103,29 \\ 168,70 & 103,29 & 66,69 \end{pmatrix}$ .

Vlastní čísla získaná řešením rovnice  $|\mathbf{S} - \lambda \mathbf{I}| = 0$  a jim odpovídající vlastní vektory jsou:

$$l_1 = 680,411,$$

$$l_2 = 6,5016,$$

$$l_3 = 2,8573,$$

$$\mathbf{v}_1 = (0,8126; 0,4955; 0,3068)^T,$$

$$\mathbf{v}_2 = (0,5454; -0,8321; -0,1009)^T,$$

$$\mathbf{v}_3 = (0,2053; 0,2493; -0,9464)^T.$$

Vyjádřete hlavní komponenty a určete, kolik procent variability obsažené v matici  $\mathbf{S}$  každá z nich vyčerpává. Najděte koeficienty korelace mezi původními znaky a hlavními komponentami. Pomocí první hlavní komponenty vypočtete reprodukovanou varianční matici.

### Řešení:

Stopa matice  $\mathbf{S}$ :  $\text{st}(\mathbf{S}) = l_1 + l_2 + l_3 = 680,411 + 6,5016 + 2,8573 = 689,77$

1. vlastní vektor:  $\mathbf{v}_1 = (0,8126; 0,4955; 0,3068)^T$

1. HK:  $Y_1 = v_{11}X_1 + \dots + v_{1p}X_p = 0,8126X_1 + 0,4955X_2 + 0,3068X_3$ , vyčerpává

$\frac{l_1}{\text{st}(\mathbf{S})} 100\% = \frac{680,411}{689,77} 100\% = 98,65\%$  variability obsažené v datovém souboru.

Výpočet koeficientů korelace:

$$R(X_1, Y_1) = \frac{v_{11}\sqrt{l_1}}{s_1} = \frac{0,8126\sqrt{680,411}}{\sqrt{451,39}} = 0,9977$$

$$R(X_2, Y_1) = \frac{v_{12}\sqrt{l_1}}{s_2} = \frac{0,4955\sqrt{680,411}}{\sqrt{171,73}} = 0,9863$$

$$R(X_3, Y_1) = \frac{v_{13}\sqrt{l_1}}{s_3} = \frac{0,3068\sqrt{680,411}}{\sqrt{66,69}} = 0,9799$$

Vidíme, že první hlavní komponenta je vysoce korelována se všemi třemi proměnnými.



2. vlastní vektor:  $\mathbf{v}_2 = (0,5454; -0,8321; -0,1009)^T$

2. HK:  $Y_2 = v_{21}X_1 + \dots + v_{2p}X_p = 0,5454X_1 - 0,8321X_2 - 0,1009X_3$ , vyčerpává

$\frac{l_2}{st(S)} 100\% = \frac{6,5016}{689,77} 100\% = 0,94\%$  variability obsažené v datovém souboru.

Výpočet koeficientů korelace:

$$R(X_1, Y_2) = \frac{v_{21}\sqrt{l_2}}{s_1} = \frac{0,5454\sqrt{6,5016}}{\sqrt{451,39}} = 0,0655$$

$$R(X_2, Y_2) = \frac{v_{22}\sqrt{l_2}}{s_2} = \frac{-0,8321\sqrt{6,5016}}{\sqrt{171,73}} = -0,1619$$

$$R(X_3, Y_2) = \frac{v_{23}\sqrt{l_2}}{s_3} = \frac{-0,1009\sqrt{6,5016}}{\sqrt{66,69}} = -0,0315$$

Druhá hlavní komponenta je pouze slabě záporně korelována s druhou proměnnou.

3. vlastní vektor:  $\mathbf{v}_3 = (0,2053; 0,2493; -0,9464)^T$

3. HK:  $Y_3 = v_{31}X_1 + \dots + v_{3p}X_p = 0,2053 X_1 + 0,2493 X_2 - 0,9464 X_3$ , vyčerpává

$$\frac{I_3}{\text{st}(S)} 100\% = \frac{2,8573}{689,77} 100\% = 0,41\% \text{ variability obsažené v datovém souboru.}$$

Výpočet koeficientů korelace:

$$R(X_1, Y_3) = \frac{v_{31} \sqrt{I_3}}{s_1} = \frac{0,2053 \sqrt{2,8573}}{\sqrt{451,39}} = 0,0163$$

$$R(X_2, Y_3) = \frac{v_{32} \sqrt{I_3}}{s_2} = \frac{0,2493 \sqrt{2,8573}}{\sqrt{171,73}} = 0,0322$$

$$R(X_3, Y_3) = \frac{v_{33} \sqrt{I_3}}{s_3} = \frac{-0,9464 \sqrt{2,8573}}{\sqrt{66,69}} = -0,1959$$

Třetí hlavní komponenta je pouze slabě záporně korelována s třetí proměnnou.

### Tabulka korelací původních proměnných a hlavních komponent

proměnná	komponenta		
	Y <sub>1</sub>	Y <sub>2</sub>	Y <sub>3</sub>
X <sub>1</sub>	0,9977	0,0655	0,0163
X <sub>2</sub>	0,9863	-0,1619	0,0322
X <sub>3</sub>	0,9799	-0,0315	-0,1959

## Výpočet reprodukované varianční matice založené na 1. HK:

$$l_1 \mathbf{v}_1 \mathbf{v}_1^T =$$

$$680,411 \begin{pmatrix} 0,8126 \\ 0,4955 \\ 0,3068 \end{pmatrix} (0,8126 \quad 0,4955 \quad 0,3068) = \begin{pmatrix} 449,2881 & 273,9629 & 169,6303 \\ 273,9629 & 167,0547 & 103,4357 \\ 169,6303 & 103,4357 & 64,0445 \end{pmatrix}$$

$$\text{Původní varianční matice: } \mathbf{S} = \begin{pmatrix} 451,39 & 271,17 & 168,70 \\ 271,17 & 171,73 & 103,29 \\ 169,70 & 103,29 & 66,69 \end{pmatrix}.$$

$$\text{Reziduální matice: } \mathbf{S} - l_1 \mathbf{v}_1 \mathbf{v}_1^T = \begin{pmatrix} 2,1019 & -2,7929 & -0,9303 \\ -2,7929 & 4,6753 & -0,1457 \\ -0,9303 & -0,1457 & 2,6055 \end{pmatrix}$$

Vidíme, že 1. hlavní komponenta velmi dobře reprodukuje rozptyly a kovariance původních tří proměnných.

## 8. Nejdůležitější problémy v metodě hlavních komponent

1. Data neobsahují předpokládanou informaci: nemá smysl provádět PCA.
2. Bylo vybráno příliš málo hlavních komponent: „podceněný“ model způsobí povrchní popis datové struktury.
3. Bylo vybráno příliš mnoho hlavních komponent: „přeceněný“ model způsobí, že šum je nesprávně zahrnut do modelu.
4. Neoprávněné ponechání vybočujících pozorování: do modelu jsou zahrnuty hrubé chyby.
5. Nesprávné odstranění vybočujících pozorování: ztratila se důležitá informace, model je zkreslený.
6. Graf faktorových souřadnic proměnných byl vytvořen se špatným počtem hlavních komponent: může dojít k neoprávněnému odstranění důležitých proměnných.
7. Objekty jsou roztrženy do několika dobře oddělených skupin: to se projeví v rozmístění objektů na ploše prvních dvou hlavních komponent. V takovém případě se soubor rozdělí na skupiny a ty jsou analyzovány PCA odděleně.

## 9. Příklad: Máme datový soubor Lide.sta, který obsahuje údaje o 32 lidech:

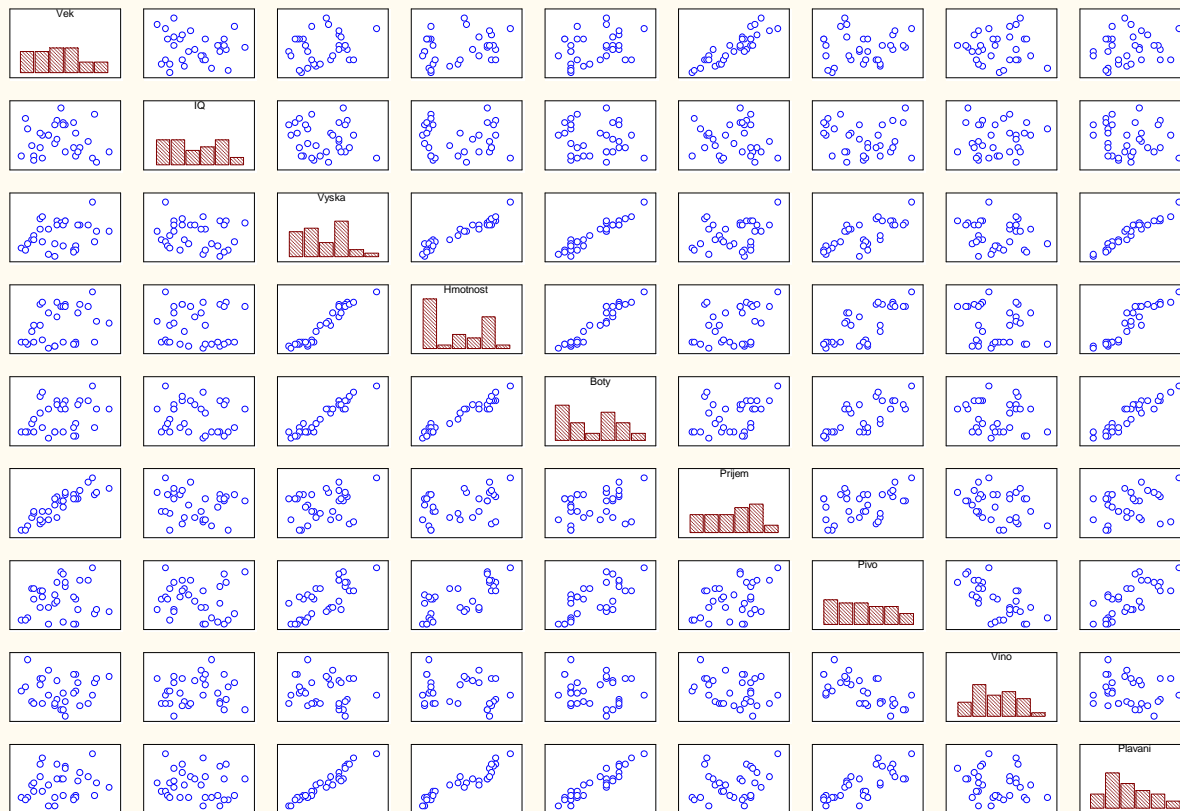
	1 Sex	2 Vlasy	3 Vek	4 IQ	5 Vyska	6 Hmotnost	7 Boty	8 Prijem	9 Pivo	10 Vino	11 Plavani	12 Puvod
	muz	kratke	48	100	198	92	48	45000	420	115	98	Skandinavie
	muz	kratke	33	130	184	84	44	33000	350	102	92	Skandinavie
	muz	kratke	37	127	183	83	44	34000	320	98	91	Skandinavie
	zena	kratke	32	112	166	47	36	28000	270	78	75	Skandinavie
	zena	dlouhe	23	110	170	60	38	20000	312	99	81	Skandinavie
	zena	dlouhe	24	102	172	64	39	22000	308	91	82	Skandinavie
	muz	kratke	35	140	182	80	42	30000	398	65	85	Skandinavie
	muz	kratke	36	129	180	80	43	30000	388	63	84	Skandinavie
	zena	dlouhe	24	98	169	51	36	23000	250	89	78	Skandinavie
	zena	dlouhe	27	100	168	52	37	23500	260	86	78	Skandinavie
	muz	kratke	37	105	183	81	42	35000	345	45	90	Skandinavie
	zena	dlouhe	32	127	157	47	36	32000	235	92	70	Skandinavie
	zena	dlouhe	41	101	164	50	38	34000	255	134	76	Skandinavie
	zena	dlouhe	40	108	162	49	37	34000	265	124	75	Skandinavie
	muz	kratke	43	109	180	82	44	37000	355	82	88	Skandinavie
	muz	kratke	46	113	180	81	44	42000	362	90	86	Skandinavie
	muz	kratke	26	109	185	82	45	16000	295	180	92	Stredomori
	muz	kratke	27	119	187	84	46	16500	299	178	95	Stredomori
	zena	dlouhe	49	135	168	50	37	34000	170	162	76	Stredomori
	zena	dlouhe	21	123	166	49	36	14000	150	245	75	Stredomori
	zena	dlouhe	30	119	158	46	34	18000	120	120	70	Stredomori
	muz	kratke	26	120	177	65	41	18000	209	160	86	Stredomori
	muz	kratke	33	115	180	72	43	19000	236	175	85	Stredomori
	muz	kratke	42	105	181	75	43	31000	198	161	83	Stredomori
	zena	dlouhe	18	102	163	50	36	11000	143	136	75	Stredomori
	zena	dlouhe	20	132	162	50	36	11500	133	146	74	Stredomori
	muz	kratke	50	96	176	68	42	36000	195	177	82	Stredomori
	muz	dlouhe	55	105	175	67	42	38000	185	187	80	Stredomori
	zena	dlouhe	36	126	165	51	36	26000	121	129	76	Stredomori
	zena	dlouhe	41	120	161	48	35	31500	116	196	75	Stredomori
	muz	kratke	30	118	178	75	42	24000	203	208	81	Stredomori
	zena	dlouhe	40	129	160	48	35	31000	118	198	74	Stredomori

Z 12 sledovaných proměnných jsou 3 alternativní (Sex, Vlasy, Původ), 9 je poměrového typu. Proměnná Příjem udává roční příjem v eurech, Pivo a Víno roční spotřebu v litrech a proměnná Plavání obsahuje naměřený čas na uplavání 50 m. Analyzujte tato data metodou hlavních komponent.

### **Řešení:**

Nejprve sestrojíme dvourozměrné tečkové diagramy pro všechny dvojice proměnných poměrového typu.

Maticový graf  
Lide.sta 12v\*32c



Je patrné, že silná přímá lineární závislost existuje mezi libovolnými dvojicemi z proměnných Výška, Hmotnost, Boty, Plavání. Rovněž vidíme dosti silnou přímou závislost mezi proměnnými Věk a Příjem. Středně silnou nepřímou lineární závislost pak mají proměnné (Pivo, Vínó).

Dále vypočteme výběrovou korelační matici všech 12 proměnných:

Proměnná	Korelace (Lide.sta)											
	Sex	Vlasy	Vek	IQ	Vyska	Hmotnost	Boty	Prijem	Pivo	Vino	Plavani	Puvod
Sex	1,000	0,875	-0,354	0,010	-0,878	-0,918	-0,921	-0,324	-0,537	0,025	-0,816	-0,000
Vlasy	0,875	1,000	-0,200	-0,026	-0,821	-0,834	-0,823	-0,252	-0,596	0,165	-0,772	0,125
Vek	-0,354	-0,200	1,000	-0,078	0,241	0,254	0,323	0,885	0,128	0,027	0,158	-0,047
IQ	0,010	-0,026	-0,078	1,000	-0,122	-0,034	-0,120	-0,107	-0,107	0,068	-0,116	0,162
Vyska	-0,878	-0,821	0,241	-0,122	1,000	0,960	0,961	0,301	0,715	-0,138	0,962	-0,177
Hmotnost	-0,918	-0,834	0,254	-0,034	0,960	1,000	0,969	0,335	0,738	-0,197	0,937	-0,215
Boty	-0,921	-0,823	0,323	-0,120	0,961	0,969	1,000	0,354	0,697	-0,089	0,933	-0,155
Prijem	-0,324	-0,252	0,885	-0,107	0,301	0,335	0,354	1,000	0,417	-0,297	0,252	-0,452
Pivo	-0,537	-0,596	0,128	-0,107	0,715	0,738	0,697	0,417	1,000	-0,654	0,725	-0,772
Vino	0,025	0,165	0,027	0,068	-0,138	-0,197	-0,089	-0,297	-0,654	1,000	-0,166	0,837
Plavani	-0,816	-0,772	0,158	-0,116	0,962	0,937	0,933	0,252	0,725	-0,166	1,000	-0,217
Puvod	-0,000	0,125	-0,047	0,162	-0,177	-0,215	-0,155	-0,452	-0,772	0,837	-0,217	1,000

Některé korelační koeficienty jsou v absolutní hodnotě dostatečně velké a zřejmě tedy bude mít smysl provést analýzu hlavních komponent. Ověříme to Bartlettovým testem sféricity.



$$\text{Testová statistika } t_0 = -n \left( 1 - \frac{2p+11}{6n} \right) \ln |\mathbf{R}| = -32 \left( 1 - \frac{2 \cdot 12 + 11}{6 \cdot 32} \right) (-21,472392) = 561,8609$$

$$\text{Kritický obor: } W = \left\langle \chi^2_{1-\alpha} \left( \frac{p(p-1)}{2} \right), \infty \right\rangle = \left\langle \chi^2_{0,95} (66), \infty \right\rangle = \langle 85,965, \infty \rangle$$

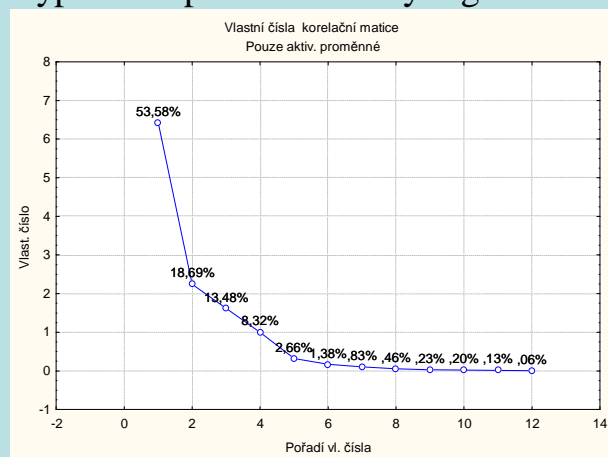
Testová statistika se realizuje v kritickém oboru, tedy hypotézu o úplné nezávislosti sledovaných 12 proměnných zamítáme na asymptotické hladině významnosti 0,05.

Má smysl provádět analýzu hlavních komponent.

## Vlastní čísla výběrové korelační matice a procento vysvětleného rozptylu:

Vlastní čísla korelační matice a související statistiky (Lide.sta) Pouze aktiv. proměnné				
Pořadí vl.č.	vl. číslo	% celk. rozptylu	Kumulativ. vl. číslo	Kumulativ. %
1	6,429692	53,58077	6,42969	53,5808
2	2,242551	18,68792	8,67224	72,2687
3	1,617699	13,48083	10,28994	85,7495
4	0,997988	8,31657	11,28793	94,0661
5	0,318660	2,65550	11,60659	96,7216
6	0,165229	1,37691	11,77182	98,0985
7	0,099393	0,82828	11,87121	98,9268
8	0,054994	0,45828	11,92621	99,3850
9	0,027449	0,22874	11,95365	99,6138
10	0,024139	0,20116	11,97779	99,8149
11	0,015199	0,12666	11,99299	99,9416
12	0,007007	0,05839	12,00000	100,0000

Výpočet doplníme sutinovým grafem:



První zlom je pozorovatelný u indexu 2, zvolíme tedy první dvě hlavní komponenty, které vysvětlují 72,3 % variability obsažené v datovém souboru.

Počet faktorů snížíme na 2. Dále vypočítáme vlastní vektory:

Proměnná	Vlastní vektory korelační matice (Lide.sta) Pouze aktiv. proměnné	
	Faktor 1	Faktor 2
Sex	0,351783	0,231671
Vlasy	0,337773	0,150163
Vek	-0,142945	0,061463
IQ	0,044067	-0,122604
Vyska	-0,375286	-0,135459
Hmotnost	-0,381136	-0,111447
Boty	-0,377697	-0,150806
Prijem	-0,190466	0,286893
Pivo	-0,324666	0,308285
Vino	0,124149	-0,554200
Plavani	-0,364904	-0,112425
Puvod	0,144121	-0,595259

1. hlavní komponenta:

$$Y_1 = 0,35\text{Sex} + 0,33\text{Vlasy} - 0,14 \text{Vek} + 0,04 \text{IQ} - 0,38\text{Vyska} - 0,38\text{Hmotnost} - 0,38\text{Boty} - 0,19\text{Prijem} - 0,32\text{Pivo} + 0,12\text{Vino} - 0,36\text{Plavani} + 0,14\text{Puvod} ,$$

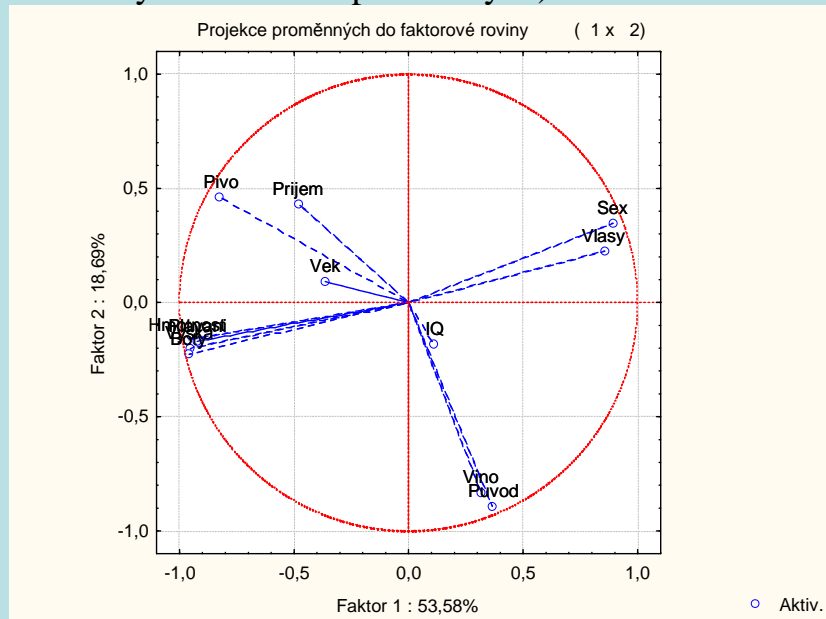
2. hlavní komponenta:

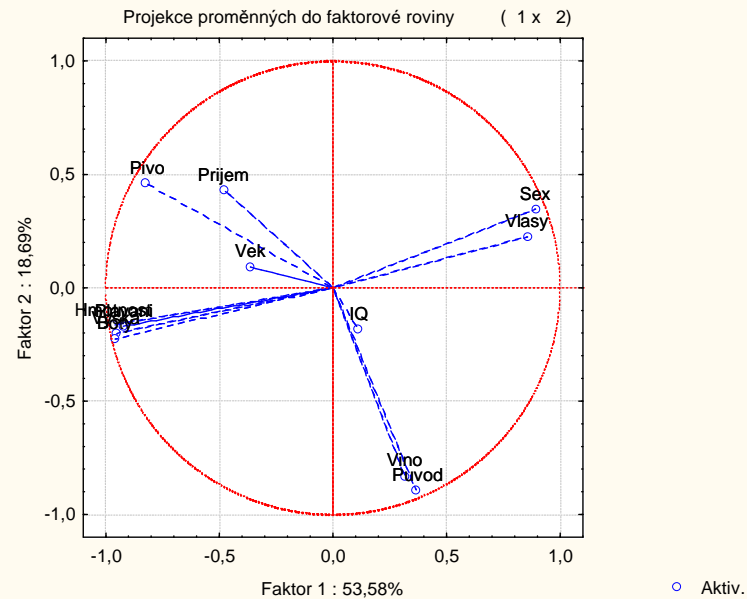
$$Y_2 = 0,23\text{Sex} + 0,15\text{Vlasy} + 0,06 \text{Vek} - 0,12 \text{IQ} - 0,13\text{Vyska} - 0,11\text{Hmotnost} - 0,15\text{Boty} + 0,29\text{Prijem} + 0,31\text{Pivo} - 0,55\text{Vino} - 0,11\text{Plavani} - 0,6\text{Puvod}$$

Vypočítáme koeficienty korelace 1. a 2. hlavní komponenty a původních 12 proměnných:

Proměnná	Faktor 1	Faktor 2
Sex	0,892009	0,346931
Vlasy	0,856487	0,224872
Vek	-0,362464	0,092041
IQ	0,111741	-0,183602
Vyska	-0,951606	-0,202851
Hmotnost	-0,966440	-0,166894
Boty	-0,957720	-0,225834
Prijem	-0,482963	0,429627
Pivo	-0,823250	0,461662
Vino	0,314802	-0,829923
Plavani	-0,925280	-0,168358
Puvod	0,365446	-0,891409

Znázorníme proměnné na ploše prvních dvou hlavních komponent (v systému STATISTICA se tento graf nazývá 2D graf faktorových souřadnic proměnných)





Každý bod v grafu odpovídá jedné proměnné. V grafu se porovnávají vzdálenosti mezi proměnnými. Malá vzdálenost mezi proměnnými znamená silnou korelaci

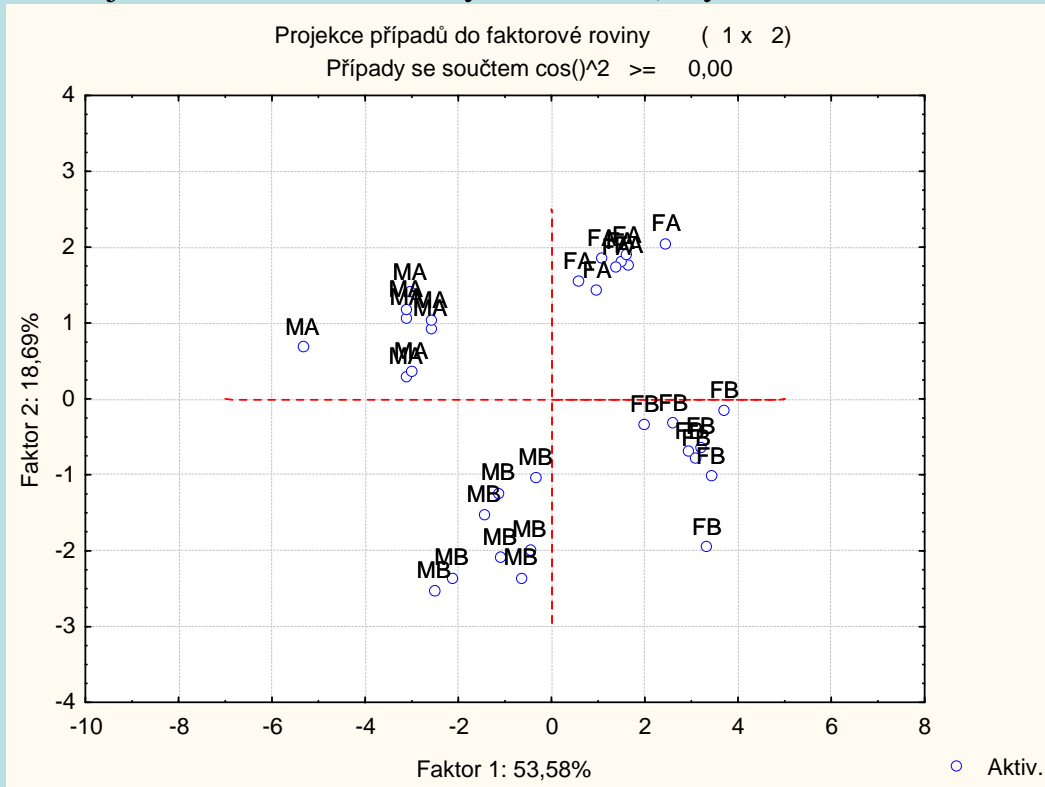
Pomocí grafu faktorových souřadnic proměnných lze posoudit tyto skutečnosti:

**Důležitost původních proměnných** – důležité proměnné leží daleko od počátku, málo důležité proměnné naopak leží blízko počátku.

**Korelace a kovariance** – proměnné s malým úhlem mezi svými průvodiči a na stejné straně vůči počátku mají vysokou kladnou korelaci či kovarianci. Naopak proměnné s velkým úhlem mezi průvodiči jsou záporně korelovány.

V našem případě jsou důležité proměnné Výška, Hmotnost, Boty, Plavání, Pivo, Vínó, Původ, Sex, méně důležité jsou Příjem, Vlasy a nedůležité pak Věk a IQ.

Podívejme se rovněž na vektory souřadnic (v systému STATISTICA se jim říká faktorové souřadnice případů):



Reprodukované korelace (Lide.sta)												
Extrakce: Hlavní komponenty												
Proměnná	Sex	Vlasy	Věk	IQ	Vyska	Hmotnost	Boty	Příjem	Pivo	Vino	Plavani	Puvod
Sex	0,92	0,84	-0,29	0,04	-0,92	-0,92	-0,93	-0,28	-0,57	-0,01	-0,88	0,02
Vlasy	0,84	0,78	-0,29	0,05	-0,86	-0,87	-0,87	-0,32	-0,60	0,08	-0,83	0,11
Věk	-0,29	-0,29	0,14	-0,06	0,33	0,33	0,33	0,21	0,34	-0,19	0,32	-0,21
IQ	0,04	0,05	-0,06	0,05	-0,07	-0,08	-0,07	-0,13	-0,18	0,19	-0,07	0,20
Vyska	-0,92	-0,86	0,33	-0,07	0,95	0,95	0,96	0,37	0,69	-0,13	0,91	-0,17
Hmotnost	-0,92	-0,87	0,33	-0,08	0,95	0,96	0,96	0,40	0,72	-0,17	0,92	-0,20
Boty	-0,93	-0,87	0,33	-0,07	0,96	0,96	0,97	0,37	0,68	-0,11	0,92	-0,15
Příjem	-0,28	-0,32	0,21	-0,13	0,37	0,40	0,37	0,42	0,60	-0,51	0,37	-0,56
Pivo	-0,57	-0,60	0,34	-0,18	0,69	0,72	0,68	0,60	0,89	-0,64	0,68	-0,71
Vino	-0,01	0,08	-0,19	0,19	-0,13	-0,17	-0,11	-0,51	-0,64	0,79	-0,15	0,85
Plavani	-0,88	-0,83	0,32	-0,07	0,91	0,92	0,92	0,37	0,68	-0,15	0,88	-0,19
Puvod	0,02	0,11	-0,21	0,20	-0,17	-0,20	-0,15	-0,56	-0,71	0,85	-0,19	0,93

Reziduální korelace (Lide.sta)												
Extrakce: Hlavní komponenty												
(Označená rezidua jsou > ,100000)												
Proměnná	Sex	Vlasy	Věk	IQ	Vyska	Hmotnost	Boty	Příjem	Pivo	Vino	Plavani	Puvod
Sex	0,08	0,03	-0,06	-0,03	0,04	0,00	0,01	-0,04	0,04	0,03	0,07	-0,02
Vlasy	0,03	0,22	0,09	-0,08	0,04	0,03	0,05	0,06	0,00	0,08	0,06	0,01
Věk	-0,06	0,09	0,86	-0,02	-0,09	-0,08	-0,00	<b>0,67</b>	<b>-0,21</b>	<b>0,22</b>	<b>-0,16</b>	<b>0,17</b>
IQ	-0,03	-0,08	-0,02	0,95	-0,05	0,04	-0,05	0,03	0,07	<b>-0,12</b>	-0,04	-0,04
Vyska	0,04	0,04	-0,09	-0,05	0,05	0,01	0,00	-0,07	0,03	-0,01	0,05	-0,01
Hmotnost	0,00	0,03	-0,08	0,04	0,01	0,04	0,01	-0,06	0,02	-0,03	0,01	-0,01
Boty	0,01	0,05	-0,00	-0,05	0,00	0,01	0,03	-0,01	0,01	0,03	0,01	-0,01
Příjem	-0,04	0,06	<b>0,67</b>	0,03	-0,07	-0,06	-0,01	0,58	<b>-0,18</b>	<b>0,21</b>	<b>-0,12</b>	<b>0,11</b>
Pivo	0,04	0,00	<b>-0,21</b>	0,07	0,03	0,02	0,01	<b>-0,18</b>	0,11	-0,01	0,04	-0,06
Vino	0,03	0,08	<b>0,22</b>	<b>-0,12</b>	-0,01	-0,03	0,03	<b>0,21</b>	-0,01	0,21	-0,01	-0,02
Plavani	0,07	0,06	<b>-0,16</b>	-0,04	0,05	0,01	0,01	<b>-0,12</b>	0,04	-0,01	0,12	-0,03
Puvod	-0,02	0,01	<b>0,17</b>	-0,04	-0,01	-0,01	-0,01	<b>0,11</b>	-0,06	-0,02	-0,03	0,07

Vysoké hodnoty reziduální korelace vidíme především u proměnných Věk a Příjem.





Vypočteme korelační matici:

```
R<-cor(lide)
format(R,digits=1)
```

	Sex	Vlasy	vek	IQ	Vyska	Hmotnost	Boty	Prijem	Pivo	Vino	Plavani	Puvod
Sex	" 1.00"	" 0.88"	"-0.35"	" 0.01"	"-0.88"	"-0.92"	"-0.92"	"-0.32"	"-0.54"	" 0.03"	"-0.82"	" 0.00"
Vlasy	" 0.88"	" 1.00"	"-0.20"	"-0.03"	"-0.82"	"-0.83"	"-0.82"	"-0.25"	"-0.60"	" 0.16"	"-0.77"	" 0.12"
vek	"-0.35"	"-0.20"	" 1.00"	"-0.08"	" 0.24"	" 0.25"	" 0.32"	" 0.88"	" 0.13"	" 0.03"	" 0.16"	"-0.05"
IQ	" 0.01"	"-0.03"	"-0.08"	" 1.00"	"-0.12"	"-0.03"	"-0.12"	"-0.11"	"-0.11"	" 0.07"	"-0.12"	" 0.16"
Vyska	"-0.88"	"-0.82"	" 0.24"	"-0.12"	" 1.00"	" 0.96"	" 0.96"	" 0.30"	" 0.72"	"-0.14"	" 0.96"	"-0.18"
Hmotnost	"-0.92"	"-0.83"	" 0.25"	"-0.03"	" 0.96"	" 1.00"	" 0.97"	" 0.34"	" 0.74"	"-0.20"	" 0.94"	"-0.22"
Boty	"-0.92"	"-0.82"	" 0.32"	"-0.12"	" 0.96"	" 0.97"	" 1.00"	" 0.35"	" 0.70"	"-0.09"	" 0.93"	"-0.15"
Prijem	"-0.32"	"-0.25"	" 0.88"	"-0.11"	" 0.30"	" 0.34"	" 0.35"	" 1.00"	" 0.42"	"-0.30"	" 0.25"	"-0.45"
Pivo	"-0.54"	"-0.60"	" 0.13"	"-0.11"	" 0.72"	" 0.74"	" 0.70"	" 0.42"	" 1.00"	"-0.65"	" 0.73"	"-0.77"
Vino	" 0.03"	" 0.16"	" 0.03"	" 0.07"	"-0.14"	"-0.20"	"-0.09"	"-0.30"	"-0.65"	" 1.00"	"-0.17"	" 0.84"
Plavani	"-0.82"	"-0.77"	" 0.16"	"-0.12"	" 0.96"	" 0.94"	" 0.93"	" 0.25"	" 0.73"	"-0.17"	" 1.00"	"-0.22"
Puvod	" 0.00"	" 0.12"	"-0.05"	" 0.16"	"-0.18"	"-0.22"	"-0.15"	"-0.45"	"-0.77"	" 0.84"	"-0.22"	" 1.00"

Ověření, zda má smysl provést analýzu hlavních komponent, provedeme pomocí Bartlettova testu sféricity. Načteme balíček psych:

```
library(psych)
```

Zavoláme funkci cortest.bartlett:

```
cortest.bartlett(R,n=32,diag=T)
```

```
$chisq
[1] 561.8609
$p.value
[1] 9.566259e-80
$df
[1] 66
```

Pomocí funkce `prcomp` provedeme analýzu hlavních komponent:

```
l1ide.PCA <- prcomp(l1ide, center=T, scale.=T)
summary(l1ide.PCA)
```

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
Standard deviation	2.5357	1.4975	1.2719	0.99899	0.56450	0.40648	0.31527	0.23451	0.16568	0.15537	0.12329	0.08371
Proportion of Variance	0.5358	0.1869	0.1348	0.08317	0.02655	0.01377	0.00828	0.00458	0.00229	0.00201	0.00127	0.00058
Cumulative Proportion	0.5358	0.7227	0.8575	0.94066	0.96722	0.98098	0.98927	0.99385	0.99614	0.99815	0.99942	1.00000

Na 1. řádku jsou uvedeny odmocniny z vlastních čísel výběrové korelační matice  $R$ . Lze to ověřit tak, že pomocí funkce `eigen` vypočteme vlastní čísla matice  $R$  a porovnáme je s 2. mocninami čísel na 1. řádku výstupu funkce `summary`:

```
(v1.cisla <- eigen(R)$values)
```

```
[1] 6.429691844 2.242550718 1.617699043 0.997987976 0.318659978 0.165228771 0.099393064 0.054993966 0.027448879
0.024139246 0.015199249
[12] 0.007007266
```

```
(v1.cisla.pca <- l1ide.PCA$sdev^2)
```

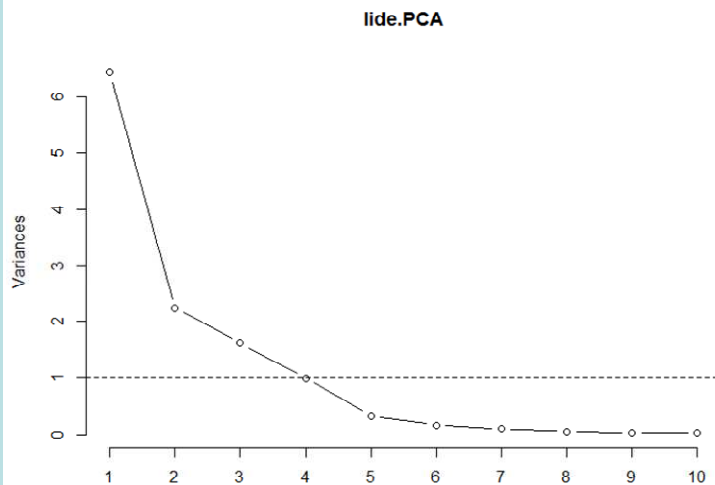
```
[1] 6.429691844 2.242550718 1.617699043 0.997987976 0.318659978 0.165228771 0.099393064 0.054993966 0.027448879
0.024139246 0.015199249
[12] 0.007007266
```

Na 2. řádku je uveden podíl vysvětlené variability pro jednotlivé hlavní komponenty. Vidíme, že 1. HK vysvětluje 53,58 % celkové variability, 2. HK 18,69 % atd. až 12. HK 0,7 %.

Na 3. řádku je kumulativní podíl vysvětlené variability.

Pro rozhodnutí, kolik HK máme uvažovat, použijeme sutinový graf, v němž ve výšce 1 zakreslíme vodorovnou čáru:

```
plot(l1ide.PCA, type='l')
abline(h=1, lty=2)
```



V sutinovém grafu se objeví zploštění u indexu 2. První dvě HK vysvětlují 72,27 % celkové variability. Omezíme se na první dvě HK.

Pro vyjádření prvních dvou HK potřebujeme první dva vlastní vektory:

`lide.PCA$rotation[,1:2]`

	PC1	PC2
Sex	0.35178259	0.23167089
Vlasy	0.33777346	0.15016340
vek	-0.14294536	0.06146276
IQ	0.04406728	-0.12260414
Vyska	-0.37528576	-0.13545854
Hmotnost	-0.38113572	-0.11144718
Boty	-0.37769704	-0.15080610
Prijem	-0.19046640	0.28689320
Pivo	-0.32466579	0.30828523
Vino	0.12414873	-0.55420028
Plavani	-0.36490373	-0.11242522
Puvod	0.14412125	-0.59525899

1. hlavní komponenta:

$$Y_1 = 0,35\text{Sex} + 0,33\text{Vlasy} - 0,14\text{Vek} + 0,04\text{IQ} - 0,38\text{Vyska} - 0,38\text{Hmotnost} - 0,38\text{Boty} - 0,19\text{Prijem} - 0,32\text{Pivo} + 0,12\text{Vino} - 0,36\text{Plavani} + 0,14\text{Puvod}$$

2. hlavní komponenta:

$$Y_2 = 0,23\text{Sex} + 0,15\text{Vlasy} + 0,06\text{Vek} - 0,12\text{IQ} - 0,13\text{Vyska} - 0,11\text{Hmotnost} - 0,15\text{Boty} + 0,29\text{Prijem} + 0,31\text{Pivo} - 0,55\text{Vino} - 0,11\text{Plavani} - 0,6\text{Puvod}$$

Abychom mohli vypočítat koeficienty korelace hlavních komponent a původních 12 proměnných, musíme uložit komponentní skóre (tj. vektory souřadnic HK) a poté spočítat korelace:

```
l1de.in.PCA<-l1de.PCA$x
cor(l1de,l1de.in.PCA[,1:2])
```

	PC1	PC2
Sex	0.8920094	0.34693060
Vlasy	0.8564867	0.22487192
Vek	-0.3624642	0.09204139
IQ	0.1117407	-0.18360152
Vyska	-0.9516060	-0.20285118
Hmotnost	-0.9664396	-0.16689380
Boty	-0.9577202	-0.22583437
Prijem	-0.4829625	0.42962683
Pivo	-0.8232498	0.46166171
Vino	0.3148019	-0.82992315
Plavani	-0.9252804	-0.16835844
Puvod	0.3654459	-0.89140917

Rovněž vykreslíme rozmístění případů na ploše prvních dvou HK společně se znázorněním proměnných:

```
biplot(l1de.PCA, scale=0, xlab=c('MA', 'MA', 'MA', 'FA', 'FA', 'FA', 'MA', 'MA',
'FA', 'FA', 'MA', 'FA', 'FA', 'FA', 'MA', 'MA', 'MB', 'MB', 'FB', 'FB',
'FB', 'MB', 'MB', 'MB', 'FB', 'FB', 'MB', 'MB', 'FB', 'FB', 'MB', 'FB'),
+      xlab='1. hlavni komponenta', ylab='2. hlavni komponenta' )
> abline(h=1, lty=2)
> abline(h=0,v=0, lty=2)
```



Podíváme se ještě na reprodukovanou korelační matici a reziduální korelační matici:

```
R.reproduced <- pc[,1:2] %*% diag(vl.cisla.pca[1:2]) %*% t(pc[,1:2])
format(R.reproduced,digits=1)
```

	Sex	Vlasy	Vek	IQ	Vyska	Hmotnost	Boty	Prijem	Pivo	Vino	Plavani	Puvod
Sex	" 0.916"	" 0.842"	"-0.291"	" 0.036"	"-0.919"	"-0.920"	"-0.933"	"-0.282"	"-0.574"	"-0.007"	"-0.884"	" 0.017"
Vlasy	" 0.842"	" 0.784"	"-0.290"	" 0.054"	"-0.861"	"-0.865"	"-0.871"	"-0.317"	"-0.601"	" 0.083"	"-0.830"	" 0.113"
Vek	"-0.291"	"-0.290"	" 0.140"	"-0.057"	" 0.326"	" 0.335"	" 0.326"	" 0.215"	" 0.341"	"-0.190"	" 0.320"	"-0.215"
IQ	" 0.036"	" 0.054"	"-0.057"	" 0.046"	"-0.069"	"-0.077"	"-0.066"	"-0.133"	"-0.177"	" 0.188"	"-0.072"	" 0.204"
Vyska	"-0.919"	"-0.861"	" 0.326"	"-0.069"	" 0.947"	" 0.954"	" 0.957"	" 0.372"	" 0.690"	"-0.131"	" 0.915"	"-0.167"
Hmotnost	"-0.920"	"-0.865"	" 0.335"	"-0.077"	" 0.954"	" 0.962"	" 0.963"	" 0.395"	" 0.719"	"-0.166"	" 0.922"	"-0.204"
Boty	"-0.933"	"-0.871"	" 0.326"	"-0.066"	" 0.957"	" 0.963"	" 0.968"	" 0.366"	" 0.684"	"-0.114"	" 0.924"	"-0.149"
Prijem	"-0.282"	"-0.317"	" 0.215"	"-0.133"	" 0.372"	" 0.395"	" 0.366"	" 0.418"	" 0.596"	"-0.509"	" 0.375"	"-0.559"
Pivo	"-0.574"	"-0.601"	" 0.341"	"-0.177"	" 0.690"	" 0.719"	" 0.684"	" 0.596"	" 0.891"	"-0.642"	" 0.684"	"-0.712"
Vino	"-0.007"	" 0.083"	"-0.190"	" 0.188"	"-0.131"	"-0.166"	"-0.114"	"-0.509"	"-0.642"	" 0.788"	"-0.152"	" 0.855"
Plavani	"-0.884"	"-0.830"	" 0.320"	"-0.072"	" 0.915"	" 0.922"	" 0.924"	" 0.375"	" 0.684"	"-0.152"	" 0.884"	"-0.188"
Puvod	" 0.017"	" 0.113"	"-0.215"	" 0.204"	"-0.167"	"-0.204"	"-0.149"	"-0.559"	"-0.712"	" 0.855"	"-0.188"	" 0.928"

```
R.residual <- R - R.reproduced
format(R.residual,digits=1)
```

	Sex	Vlasy	Vek	IQ	Vyska	Hmotnost	Boty	Prijem	Pivo	Vino	Plavani	Puvod
Sex	" 0.084"	" 0.033"	"-0.062"	"-0.026"	" 0.042"	" 0.002"	" 0.012"	"-0.042"	" 0.037"	" 0.032"	" 0.068"	"-0.017"
Vlasy	" 0.033"	" 0.216"	" 0.090"	"-0.081"	" 0.040"	" 0.031"	" 0.048"	" 0.065"	" 0.005"	" 0.082"	" 0.058"	" 0.012"
Vek	"-0.062"	" 0.090"	" 0.860"	"-0.020"	"-0.085"	"-0.081"	"-0.003"	" 0.670"	"-0.213"	" 0.217"	"-0.161"	" 0.168"
IQ	"-0.026"	"-0.081"	"-0.020"	" 0.954"	"-0.053"	" 0.044"	"-0.055"	" 0.026"	" 0.069"	"-0.120"	"-0.044"	"-0.043"
Vyska	" 0.042"	" 0.040"	"-0.085"	"-0.053"	" 0.053"	" 0.006"	" 0.004"	"-0.071"	" 0.025"	"-0.007"	" 0.047"	"-0.010"
Hmotnost	" 0.002"	" 0.031"	"-0.081"	" 0.044"	" 0.006"	" 0.038"	" 0.006"	"-0.060"	" 0.020"	"-0.031"	" 0.015"	"-0.011"
Boty	" 0.012"	" 0.048"	"-0.003"	"-0.055"	" 0.004"	" 0.006"	" 0.032"	"-0.011"	" 0.013"	" 0.025"	" 0.009"	"-0.006"
Prijem	"-0.042"	" 0.065"	" 0.670"	" 0.026"	"-0.071"	"-0.060"	"-0.011"	" 0.582"	"-0.179"	" 0.212"	"-0.123"	" 0.108"
Pivo	" 0.037"	" 0.005"	"-0.213"	" 0.069"	" 0.025"	" 0.020"	" 0.013"	"-0.179"	" 0.109"	"-0.012"	" 0.041"	"-0.059"
Vino	" 0.032"	" 0.082"	" 0.217"	"-0.120"	"-0.007"	"-0.031"	" 0.025"	" 0.212"	"-0.012"	" 0.212"	"-0.014"	"-0.018"
Plavani	" 0.068"	" 0.058"	"-0.161"	"-0.044"	" 0.047"	" 0.015"	" 0.009"	"-0.123"	" 0.041"	"-0.014"	" 0.116"	"-0.029"
Puvod	"-0.017"	" 0.012"	" 0.168"	"-0.043"	"-0.010"	"-0.011"	"-0.006"	" 0.108"	"-0.059"	"-0.018"	"-0.029"	" 0.072"