Snížení dimenze dat metodou hlavních komponent

Motivace: Metodu hlavních komponent (Principal Component Analysis – PCA) popsal v r. 1901 Karl Pearson a ve 30. letech 20. století ji dále rozvinul Harold Hotelling.

Harold Hotelling (1895 – 1973), americký matematik a statistik

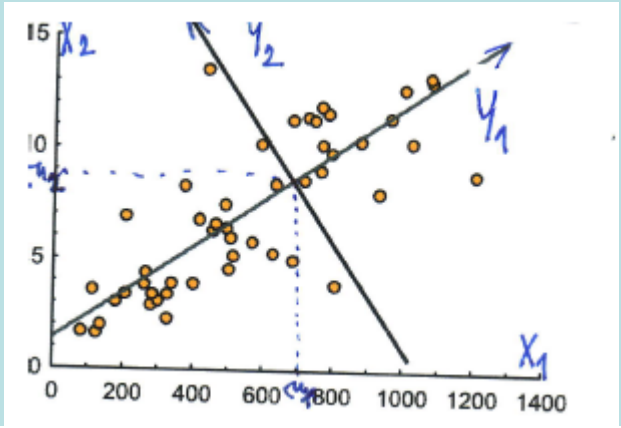
Cíl PCA:

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 zbývající 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 požíváním PCA ukazují, že případ, kdy $m = 1$ až 4 je poměrně častý.)

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

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.

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

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

Označení

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

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

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

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

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

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

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

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

(**S** a **R** jsou čtvercové symetrické matice řádu p.)

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

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

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

Řešení:

Nejprve vypočteme průměry 1. a 2. znaku:

$m_1 = \frac{1}{5}(3 + 5 + 6 + 7 + 9) = 6$, $m_2 = \frac{1}{5}(7 + 6 + 8 + 10 + 9) = 8$, tedy

vektor průměrů má tvar $\mathbf{m} = \begin{pmatrix} 6 \\ 8 \end{pmatrix}$.

Dále spočteme výběrové rozptyly 1. a 2. znaku:

$s_1^2 = \frac{1}{4}[(3-6)^2 + (5-6)^2 + (6-6)^2 + (7-6)^2 + (9-6)^2] = \frac{1}{4}(9 + 1 + 0 + 1 + 9) = 5$

$s_2^2 = \frac{1}{4}[(7-8)^2 + (6-8)^2 + (8-8)^2 + (10-8)^2 + (9-8)^2] = \frac{1}{4}(1 + 4 + 0 + 4 + 1) = 2{,}5$

Pro výpočet výběrové varianční matice potřebujeme vektory centrovaných hodnot:

$$\begin{pmatrix} 3-6 \\ 7-8 \end{pmatrix}, \begin{pmatrix} 5-6 \\ 6-8 \end{pmatrix}, \begin{pmatrix} 6-6 \\ 8-8 \end{pmatrix}, \begin{pmatrix} 7-6 \\ 10-8 \end{pmatrix}, \begin{pmatrix} 9-6 \\ 9-8 \end{pmatrix}$$

Pak

$$\mathbf{S} = \frac{1}{4}\left[\begin{pmatrix} -3 \\ -1 \end{pmatrix}(-3\ -1) + \begin{pmatrix} -1 \\ -2 \end{pmatrix}(-1\ -2) + \begin{pmatrix} 0 \\ 0 \end{pmatrix}(0\ 0) + \begin{pmatrix} 1 \\ 2 \end{pmatrix}(1\ 2) + \begin{pmatrix} 3 \\ 1 \end{pmatrix}(3\ 1)\right]$$

$$= \frac{1}{4}\begin{pmatrix} 9 & 3 \\ 3 & 1 \end{pmatrix} + \frac{1}{4}\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} + \frac{1}{4}\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{4}\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} + \frac{1}{4}\begin{pmatrix} 9 & 3 \\ 3 & 1 \end{pmatrix} = \frac{1}{4}\begin{pmatrix} 20 & 10 \\ 10 & 10 \end{pmatrix} = \begin{pmatrix} 5 & 2{,}5 \\ 2{,}5 & 2{,}5 \end{pmatrix}$$

Upozornění: K výpočtu výběrové varianční matice můžeme přistoupit i jinak. Na hlavní diagonále této matice jsou rozptyly, mimo hlavní diagonálu kovariance.

V našem případě:

$$\begin{pmatrix} 3 & 7 \\ 5 & 6 \\ 6 & 8 \\ 7 & 10 \\ 9 & 9 \end{pmatrix}, \; m_1 = 6, \; m_2 = 8, \; s_1^2 = 5, \; s_2^2 = 2{,}5$$

$$s_{12} = \frac{1}{n} \sum_{i=1}^{n} (x_{i1} - m_1)(x_{i2} - m_2) =$$

$$= \frac{1}{5} \left[(3-6)(7-8) + (5-6)(6-8) + (6-6)(8-8) + (7-6)(10-8) + (9-6)(9-8) \right] =$$

$$= \frac{12}{5} = 2{,}5$$

$$\mathbf{S} = \begin{pmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{pmatrix} = \begin{pmatrix} 5 & 2{,}5 \\ 2{,}5 & 2{,}5 \end{pmatrix}$$

Pro výpočet výběrové korelační matice potřebujeme vektory standardizovaných hodnot:

$$\left(\frac{3,6}{5}\right)\left(-\frac{3}{5}\right)\frac{5,6}{5}\left(-\frac{1}{5}\right)\frac{6,6}{5}\;0\;\frac{7,6}{5}\left(\frac{1}{5}\right)\frac{9,6}{5}\left(\frac{3}{5}\right)$$

$$\frac{7,8}{25}\left(-\frac{1}{25}\right)\frac{6,8}{25}\left(-\frac{2}{25}\right)\frac{8,8}{25}\;0\;\frac{10,8}{25}\left(\frac{2}{25}\right)\frac{9,8}{25}\left(\frac{1}{25}\right)$$

Pak

$$\frac{1}{4}\left[\left(-\frac{3}{5}\right)\left(-\frac{3}{5\sqrt{25}}\right)+\left(-\frac{1}{5}\right)\left(-\frac{2}{5\sqrt{25}}\right)+\left(\frac{1}{5}\right)\left(\frac{2}{5\sqrt{25}}\right)+\left(\frac{3}{5}\right)\left(\frac{1}{5\sqrt{25}}\right)\right]=$$

$$\mathbf{R}=\frac{1}{4}\left[\begin{pmatrix}\frac{9}{5} & \frac{3}{\sqrt{125}}\\ \frac{3}{125} & \frac{1}{25}\end{pmatrix}+\begin{pmatrix}\frac{1}{5} & \frac{2}{\sqrt{125}}\\ \frac{2}{125} & \frac{4}{25}\end{pmatrix}+\begin{pmatrix}\frac{1}{5} & \frac{2}{\sqrt{125}}\\ \frac{2}{125} & \frac{4}{25}\end{pmatrix}+\begin{pmatrix}\frac{9}{5} & \frac{3}{\sqrt{125}}\\ \frac{3}{125} & \frac{1}{25}\end{pmatrix}\right]=$$

$$\frac{1}{4}\begin{pmatrix}\frac{20}{5} & \frac{10}{\sqrt{125}}\\ \frac{10}{125} & \frac{10}{25}\end{pmatrix}=\begin{pmatrix}1 & 0{,}707\\ 0{,}707 & 1\end{pmatrix}$$

Upozornění: K výpočtu výběrové korelační matice můžeme přistoupit i jinak. Na hlavní diagonále této matice jsou jedničky, mimo hlavní diagonálu koeficienty korelace.
V našem případě:

$r_{12} = \frac{s_{12}}{s_1 s_2} = \frac{2,5}{\sqrt{5}\sqrt{25}} = 0,707$, $\mathbf{R} = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0,707 \\ 0,707 & 1 \end{pmatrix}$

Výpočet pomocí systému STATISTICA:
Potřebujeme datový soubor o dvou proměnných X1, X2 a 5 případech
Získání vektoru průměrů: Statistiky – Základní statistiky/tabulky – Popisné statistiky – Proměnné X1, X2 – ponecháme zaškrtnutý jen průměr – OK

Proměnná	Popisné statistiky (Dv… Prům…
X1	6
X2	8

Získání varianční matice: Statistiky – Vícerozměrná regrese – Proměnné - Závislá proměnná X2, Seznam nezáv. proměnných X1 – OK – OK Residua/předpoklady/předpovědi – Popisné statistiky – Další statistiky - Kovariance

Proměnná	Kovariance (Dva… X1	X2
X1	5,	2,
X2	2,	2,

Získání korelační matice: Statistiky – Vícerozměrná regrese – Proměnné - Závislá proměnná X2, Seznam nezáv. proměnných X1 – OK – OK Residua/předpoklady/předpovědi – Popisné statistiky – Korelace

Proměnná	Korelace (Dva… X1	X2
X1	1,000	0,707
X2	0,707	1,000

Základní pojmy v metodě hlavních komponent

A - čtvercová matice řádu p.

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

Vlastní vektor matice **A** – vektor **v.**

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

Stopa matice **A** - součet jejích diagonálních prvků (značí se Tr(**A**)).

Výpočet vlastních čísel matice A

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

Tato soustava p rovnic má netriviální řešení, právě když charakteristický polynom matice **A** je roven 0.

Dostaneme rovnici p-tého stupně. Jejím řešením jsou vlastní čísla $\lambda_1, \ldots, \lambda_p$. Jejich součet je roven stopě matice **A**.

Získání hlavních komponent

Nechť výběrová varianční matice **S** má vlastní čísla $l_1, ..., l_p$ a vlastní vektory $\mathbf{v}_1, ..., \mathbf{v}_p$, přičemž
$v_{j1}^2 + v_{j2}^2 + \ldots v_{jp}^2 = 1$, $v_{j1}v_{k1} + v_{j2}v_{k2} + \ldots + v_{jp}v_{kp} = 0$ pro $j \neq k$.
(Znamená to, že vektory $\mathbf{v}_1, ..., \mathbf{v}_p$ jsou ortonormální.)
Bez újmy na obecnosti předpokládáme, že $l_1 > l_2 > ... > l_p$.

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

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

Rozptyl 1. hlavní komponenty je l_1.

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

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

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

Přitom $v_{11}v_{21} + v_{12}v_{22} + \ldots + v_{1p}v_{2p} = 0$, tj. 1. a 2. hlavní komponenta jsou lineárně nezávislé.
Rozptyl 2. hlavní komponenty je l_2.

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

.....

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

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

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

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

Vektory souřadnic všech p hlavních komponent uspořádáme do matice

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

Lze dokázat, že celková variabilita obsažená v datech je rovna stopě matice **S**, tj. součtu vlastních čísel $l_1 + ... + l_p$. 1. hlavní komponenta tedy vyčerpává $\frac{l_1}{l_1 + \cdots +} 100\%$ celkové variability. Pokud je číslo $\frac{l_1}{l_1 + \cdots +}$ dostatečně blízké 1, znamená to, že 1. hlavní komponenta dobře nahrazuje celý datový soubor. Je-li toto číslo podstatně menší než 1, musíme vzít tolik hlavních komponent, aby jejich součet dělený stopou matice **S** byl dostatečně blízký 1. V mnoha aplikacích se stává, že i při velkém počtu znaků stačí poměrně malý počet hlavních komponent.
(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_j s k-tou hlavní komponentou Y_k lze vyjádřit jako $R(X_i, Y_k) = \frac{v_{ik}\sqrt{l_k}}{s_i}$.

Reprodukce výchozí kovarianční matice:

V teorii matic se dokazuje vzorec $\mathbf{S} = \sum_{i=1}^{p} l_i \mathbf{v}_i \mathbf{v}_i^T$ (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 kovarianční matice a reprodukované kovarianční matice.

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, kde nulová hypotéza tvrdí, že výběrová korelační matice je matice jednotková. Testová statistika je dána vzorcem $\chi^2 = -\left(n - 1 - \frac{2p+5}{6}\right)\ln|\mathbf{R}|$. Platí-li nulová hypotéza, testová statistika se asymptoticky řídí rozložením $\chi^2(p(p-1)/2)$. Nulovou hypotézu tedy zamítáme na asymptotické hladině významnosti α, když $\chi^2 > \chi^2_{1-\alpha}(p(p-1)/2)$. Nezamítneme-li nulovou hypotézu, neměli bychom analýzu hlavních komponent vůbec provádět (Bartlettův test je implementován např. v systému SPSS).

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 **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 **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.

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

Z datového souboru byla vypočtena výběrová varianční matice $\mathbf{S} = \begin{pmatrix} 45,39 & 27,17 & 16,80 \\ 27,17 & 17,73 & 10,29 \\ 16,80 & 10,29 & 6,69 \end{pmatrix}$.

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 **S** každá z nich vyčerpává. Najděte koeficienty korelace mezi původními znaky a hlavními komponentami. Pomocí první hlavní komponenty vypočtěte reprodukovanou kovarianční matici.

Řešení:

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

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

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

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

Výpočet koeficientů korelace:

$$r(X_1, Y_1) = \frac{v_{11}\sqrt{l_1}}{s_1} = \frac{0{,}8126\sqrt{680{,}411}}{\sqrt{45{,}39}} = 0{,}997$$

$$r(X_2, Y_1) = \frac{v_{21}\sqrt{l_1}}{s_2} = \frac{0{,}4955\sqrt{680{,}411}}{\sqrt{17{,}73}} = 0{,}986$$

$$r(X_3, Y_1) = \frac{v_{31}\sqrt{l_1}}{s_3} = \frac{0{,}3068\sqrt{680{,}411}}{\sqrt{6{,}669}} = 0{,}979$$

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

$\frac{l_2}{\text{stopa}} \cdot 100\% = \frac{\text{65016}}{\text{97}} \cdot 100\% =$ variability obsažené v datovém souboru.

Výpočet koeficientů korelace:

$r(X_1, Y_2) = \frac{v_{21}\sqrt{l_2}}{s_1} = \frac{0{,}5454\sqrt{65016}}{\sqrt{45{,}39}} =$

$r(X_2, Y_2) = \frac{v_{22}\sqrt{l_2}}{s_2} = \frac{-0{,}8321\sqrt{65016}}{\sqrt{17{,}73}} =$

$r(X_3, Y_2) = \frac{v_{23}\sqrt{l_2}}{s_3} = \frac{-0{,}1009\sqrt{65016}}{\sqrt{6{,}669}} =$

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

$\frac{l_3}{s_t^2} 100\% = \frac{28573}{8,97} 100\% = 41\%$ variability obsažené v datovém souboru.

Výpočet koeficientů korelace:

$$r(X_1, Y_3) = \frac{v_{31}\sqrt{l_3}}{s_1} = \frac{0{,}2053\sqrt{2,28573}}{\sqrt{45,39}} = 0{,}016$$

$$r(X_2, Y_3) = \frac{v_{32}\sqrt{l_3}}{s_2} = \frac{0{,}2493\sqrt{2,28573}}{\sqrt{17,73}} = 0{,}032$$

$$r(X_3, Y_3) = \frac{v_{33}\sqrt{l_3}}{s_3} = \frac{-0{,}9464\sqrt{2,28573}}{\sqrt{6,69}} = -0{,}695$$

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

Tabulka korelací původních proměnných a hlavních komponent

proměnná	komponenta		
	Y_1	Y_2	Y_3
X_1	0,9977	0,0655	0,0163
X_2	0,9863	-0,1619	0,0322
X_3	0,9799	-0,0315	-0,1959

Výpočet reprodukované kovarianční matice založené na 1. HK:

$$l_1 \mathbf{v}_1 \mathbf{v}_1^T = 68{,}4 \begin{pmatrix} 0{,}8126 \\ -0{,}4955 \\ 0{,}3068 \end{pmatrix} \begin{pmatrix} 0{,}8126 & -0{,}4955 & 0{,}3068 \end{pmatrix} = \begin{pmatrix} 44{,}288 & -27{,}000 & 16{,}720 \\ -27{,}000 & 16{,}463 & -10{,}192 \\ 16{,}720 & -10{,}192 & 6{,}311 \end{pmatrix}$$

Původní varianční matice: $\mathbf{S} = \begin{pmatrix} 45{,}39 & 27{,}17 & 16{,}80 \\ 27{,}17 & 17{,}73 & 10{,}29 \\ 16{,}80 & 10{,}29 & 6{,}69 \end{pmatrix}$

Reziduální matice: $\mathbf{S} - l_1 \mathbf{v}_1 \mathbf{v}_1^T = \begin{pmatrix} 1{,}102 & -0{,}170 & 0{,}080 \\ -0{,}170 & 1{,}267 & -0{,}097 \\ 0{,}080 & -0{,}097 & 0{,}379 \end{pmatrix}$

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

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

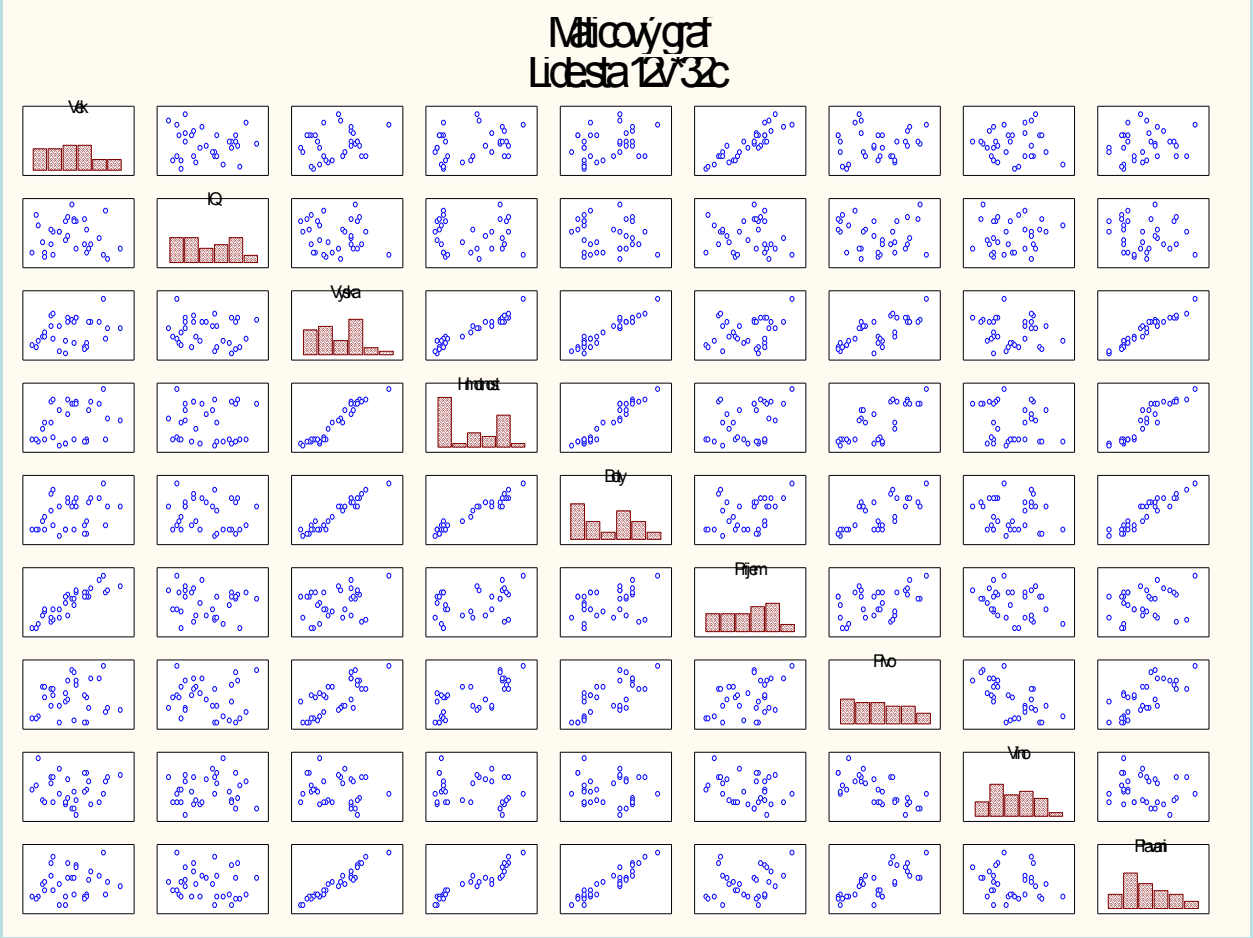
1 Sex	2 Vlasy	3 Vek	4 IQ	5 Výska	6 Hmotn	7 Boty	8 Prije	9 Pivo	10 Vino	11 Plava	12 Puvod
muz	kratke	48	10	19	92	48	4500	42	11	9	Skandir
muz	kratke	33	13	18	84	44	3300	35	10	9	Skandir
muz	kratke	3	12	18	83	44	3400	32	98	9	Skandir
zena	kratke	32	11	16	4	36	2800	27	78	7	Skandir
zena	dlouhe	23	11	17	60	38	2000	31	95	8	Skandir
zena	dlouhe	24	10	17	64	39	2200	30	97	8	Skandir
muz	kratke	35	14	18	80	42	3000	39	65	8	Skandir
muz	kratke	36	12	18	80	43	3000	38	63	8	Skandir
zena	dlouhe	24	98	16	57	36	2300	25	85	7	Skandir
zena	dlouhe	2	10	16	52	3	2350	26	80	7	Skandir
muz	kratke	3	10	18	87	42	3500	34	45	9	Skandir
zena	dlouhe	32	12	15	4	36	3200	23	92	7	Skandir
zena	dlouhe	47	10	16	50	38	3400	25	13	7	Skandir
zena	dlouhe	40	10	16	45	3	3400	26	12	7	Skandir
muz	kratke	43	10	18	82	44	3700	35	82	8	Skandir
muz	kratke	46	11	18	87	44	4200	36	90	8	Skandir
muz	kratke	26	10	18	82	45	1600	29	18	9	Stredon
muz	kratke	2	11	18	84	46	1650	29	17	9	Stredon
zena	dlouhe	49	13	16	50	3	3400	17	16	7	Stredon
zena	dlouhe	27	12	16	45	36	1400	15	24	7	Stredon
zena	dlouhe	30	11	15	40	34	1800	12	12	7	Stredon
muz	kratke	26	12	17	65	41	1800	20	16	8	Stredon
muz	kratke	33	11	18	72	43	1900	23	17	8	Stredon
muz	kratke	42	10	18	75	43	3100	19	16	8	Stredon
zena	dlouhe	18	10	16	50	36	1100	14	13	7	Stredon
zena	dlouhe	20	13	16	50	36	1150	13	14	7	Stredon
muz	kratke	50	96	17	68	42	3600	19	17	8	Stredon
muz	dlouhe	55	10	17	6	42	3800	18	18	8	Stredon
zena	dlouhe	36	12	16	57	36	2600	12	12	7	Stredon
zena	dlouhe	41	12	16	48	35	3150	11	19	7	Stredon
muz	kratke	30	11	17	75	42	2400	20	20	8	Stredon
zena	dlouhe	40	12	16	48	35	3100	11	19	7	Stredon

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

Výpočet pomocí systému STATISTICA

Nejprve sestrojíme dvourozměrné tečkové diagramy pro všechny dvojice proměnných poměrového typu:
Grafy – Maticové grafy – Proměnné Věk, IQ, Výška, Hmotnost, Boty, Příjem, Pivo, Víno, Plavání – OK – OK.

Grafy – Maticové grafy – Proměnné Věk, IQ, Výška, Hmotnost, Boty, Příjem, Pivo, Víno, Plavání – OK – OK.



Je patrné, že silná přímá lineární závislost existuje mezi libovolnými dojicemi 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, Víno).

Dále vypočteme výběrovou korelační matici všech 12 proměnných:
Statistiky – Vícerozměrné průzkumné techniky – Hlavní komponenty&klasifikační analýza – Proměnné 1 - 12, OK – OK – Popisné statistiky – Korelační matice.

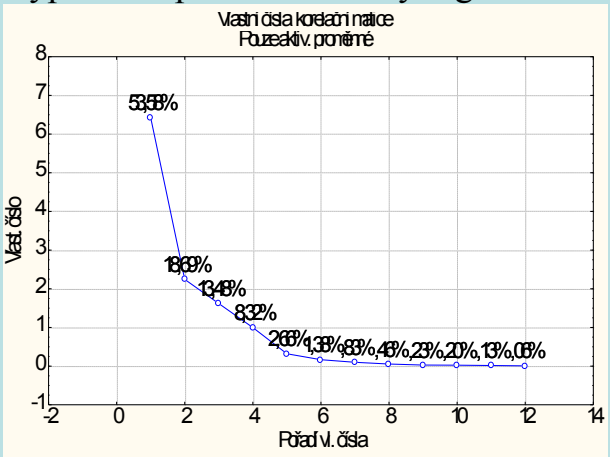
Proměnné	Korelace (Lide.sta)											
	Sex	Vlas	Vek	IQ	Vysk	Hmotn	Boty	Prije	Pivo	Vino	Plava	Puvo
Sex	1,0	0,8	-0,3	0,0	-0,8	-0,9	-0,9	-0,3	-0,5	0,0	-0,8	-0,0
Vlasy	0,8	1,0	-0,2	-0,0	-0,8	-0,8	-0,8	-0,2	-0,5	0,1	-0,7	0,1
Vek	-0,3	-0,2	1,0	-0,0	0,2	0,2	0,3	0,8	0,1	0,0	0,1	-0,0
IQ	0,0	-0,0	-0,0	1,0	-0,1	-0,0	-0,1	-0,1	-0,1	0,0	-0,1	0,1
Vyska	-0,8	-0,8	0,2	-0,1	1,0	0,9	0,9	0,3	0,7	-0,1	0,9	-0,1
Hmotn	-0,9	-0,8	0,2	-0,0	0,9	1,0	0,9	0,3	0,7	-0,1	0,9	-0,2
Boty	-0,9	-0,8	0,3	-0,1	0,9	0,9	1,0	0,3	0,6	-0,0	0,9	-0,1
Prijem	-0,3	-0,2	0,8	-0,1	0,3	0,3	0,3	1,0	0,4	-0,2	0,2	-0,4
Pivo	-0,5	-0,5	0,1	-0,1	0,7	0,7	0,6	0,4	1,0	-0,6	0,7	-0,7
Vino	0,0	0,1	0,0	0,0	-0,1	-0,1	-0,0	-0,2	-0,6	1,0	-0,1	0,8
Plavan	-0,8	-0,7	0,1	-0,1	0,9	0,9	0,9	0,2	0,7	-0,1	1,0	-0,2
Puvod	-0,0	0,1	-0,0	0,1	-0,1	-0,2	-0,1	-0,4	-0,7	0,8	-0,2	1,0

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

Nyní získáme vlastní čísla výběrové korelační matice a procento vysvětleného rozptylu: na záložce Základní výsledky vybereme Vlastní čísla.

Pořadí	Vlastní čísla korelační matice a souvis Pouze aktiv. proměnné			
	vl. čís	% cel rozpt	Kumula vl. čísl	Kumula %
1	6,429	53,58	6,42	53,5
2	2,242	18,68	8,67	72,2
3	1,617	13,48	10,28	85,7
4	0,997	8,31	11,28	94,0
5	0,318	2,65	11,60	96,7
6	0,165	1,37	11,77	98,0
7	0,099	0,82	11,87	98,9
8	0,054	0,45	11,92	99,3
9	0,027	0,22	11,95	99,6
10	0,024	0,20	11,97	99,8
11	0,015	0,12	11,99	99,9
12	0,007	0,05	12,00	100,0

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.

V nabídce Výsledky hlavních komponent snížíme počet faktorů na 2.
Dále vypočítáme vlastní vektory: na záložce Proměnné vybereme Vlastní vektory a v získané tabulce odstraníme proměnné 3 – 12.

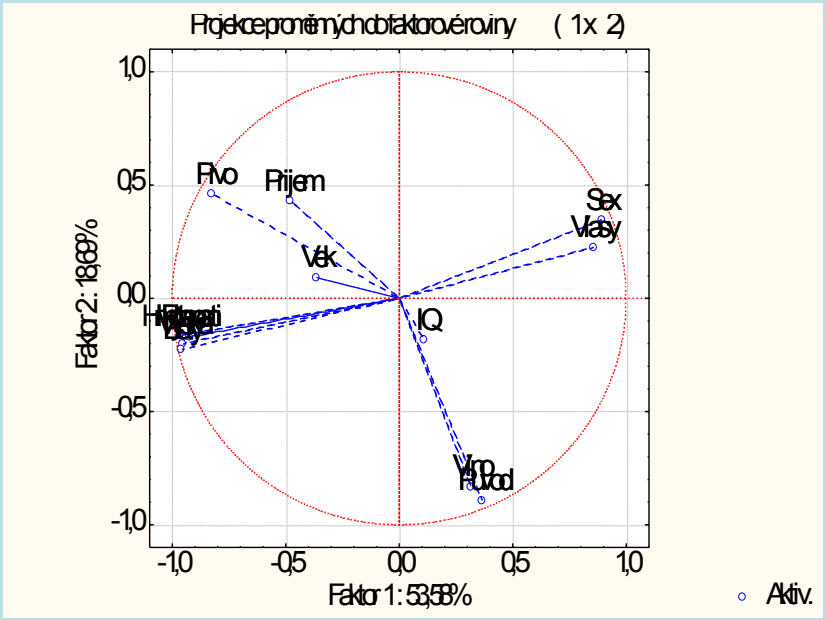
	Vlastní vektory korelační r	
	Pouze aktiv. proměnné	
Proměnná	Faktor 1	Faktor 2
Sex	0,351	0,231
Vlasy	0,337	0,150
Vek	-0,142	0,061
IQ	0,044	-0,122
Vyska	-0,375	-0,135
Hmotn	-0,381	-0,111
Boty	-0,377	-0,150
Prijem	-0,190	0,286
Pivo	-0,324	0,308
Vino	0,124	-0,554
Plavan	-0,364	-0,112
Puvod	0,144	-0,595

1. hlavní komponenta:
Y_1 = 0,35Sex + 0,33Vlasy - 0,14 Vek + 0,04 IQ - 0,38Vyska – 0,38Hmotnost – 0,38Boty - 0,19Prijem – 0,32Pivo + 0,12Vino – 0,36Plavani + 0,14Puvod ,
2. hlavní komponenta:
Y_2 = 0,23Sex + 0,15Vlasy + 0,06 Vek – 0,12 IQ - 0,13Vyska – 0,11Hmotnost – 0,15Boty + 0,29Prijem + 0,31Pivo - 0,55Vino – 0,11Plavani – 0,6Puvod

Výpočet koeficientů korelace 1. a 2. hlavní komponenty a původních čtyř proměnných: na záložce Proměnné vybereme Korelace faktorů & proměnných

Proměnná	Faktor 1	Faktor 2
Sex	0,892	0,346
Vlasy	0,856	0,224
Vek	-0,362	0,092
IQ	0,111	-0,183
Vyska	-0,951	-0,202
Hmotn	-0,966	-0,166
Boty	-0,957	-0,225
Prijem	-0,482	0,429
Pivo	-0,823	0,461
Vino	0,314	-0,829
Plavan	-0,925	-0,168
Puvod	0,365	-0,891

Znázornění proměnných 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

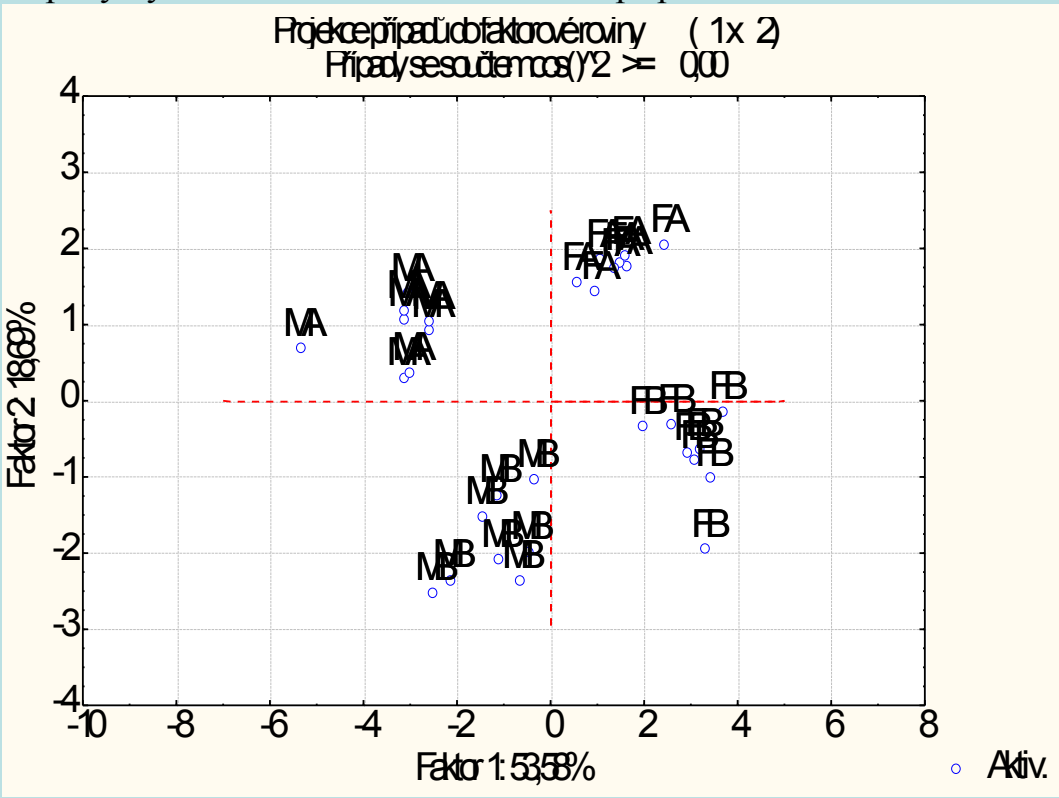
Pomocí grafu faktorových souřadnic proměnných lze posoudit tyto skutečnosti:

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, Víno, 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): na záložce Případy vybereme Faktorové souřadnice případů.



Vidíme, že 1. hlavní komponenta rozlišila pohlaví (muži jsou nalevo, ženy napravo) a 2. hlavní komponenta rozlišila původ (osoby ze Středomoří jsou dole, ze Skandinávie nahoře).

Nakonec posoudíme reprodukovanou a reziduální korelační matici:
Statistiky – Vícerozměrné průzkumné techniky – Faktorová analýza – Proměnné 1 – 12, OK – Max. počet faktorů 2 – OK – Výklad rozptylu – Reproduk./ rezid. korelace.

Reprodukované korelace (Lide.sta)
Extrakce: Hlavní komponenty

Proměn	Sex	Vlas	Vek	IQ	Vysk	Hmotn	Boty	Prije	Piv	Vin	Plava	Puvo
Sex	0,9	0,8	-0,2	0,0	-0,9	-0,9	-0,9	-0,2	-0,5	-0,0	-0,8	0,0
Vlasy	0,8	0,7	-0,2	0,0	-0,8	-0,8	-0,8	-0,3	-0,6	0,0	-0,8	0,1
Vek	-0,2	-0,2	0,1	-0,0	0,3	0,3	0,3	0,2	0,3	-0,1	0,3	-0,2
IQ	0,0	0,0	-0,0	0,0	-0,0	-0,0	-0,0	-0,1	-0,1	0,1	-0,0	0,2
Vyska	-0,9	-0,8	0,3	-0,0	0,9	0,9	0,9	0,3	0,6	-0,1	0,9	-0,1
Hmotn	-0,9	-0,8	0,3	-0,0	0,9	0,9	0,9	0,2	0,7	-0,1	0,9	-0,2
Boty	-0,9	-0,8	0,3	-0,0	0,9	0,9	0,9	0,3	0,6	-0,1	0,9	-0,1
Prijem	-0,2	-0,3	0,2	-0,1	0,3	0,4	0,3	0,4	0,6	-0,5	0,3	-0,5
Pivo	-0,5	-0,6	0,3	-0,1	0,6	0,7	0,6	0,6	0,8	-0,6	0,6	-0,1
Vino	-0,0	0,0	-0,1	0,1	-0,1	-0,1	-0,1	-0,5	-0,6	0,7	-0,1	0,8
Plavan	-0,8	-0,8	0,3	-0,0	0,9	0,9	0,9	0,3	0,6	-0,1	0,8	-0,1
Puvod	0,0	0,1	-0,2	0,2	-0,1	-0,2	-0,1	-0,5	-0,1	0,8	-0,1	0,9

Reziduální korelace (Lide.sta)
Extrakce: Hlavní komponenty
(Označená rezidua jsou > ,100000)

Proměn	Sex	Vlas	Vek	IQ	Vysk	Hmotn	Boty	Prije	Piv	Vin	Plava	Puvo
Sex	0,0	0,0	-0,0	-0,0	0,0	0,0	0,0	-0,0	0,0	0,0	0,0	-0,0
Vlasy	0,0	0,2	0,0	-0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Vek	-0,0	0,0	0,8	-0,0	-0,0	-0,0	-0,0	0,6	-0,2	0,2	-0,1	0,1
IQ	-0,0	-0,0	-0,0	0,9	-0,0	0,0	-0,0	0,0	0,0	-0,1	-0,0	-0,0
Vyska	0,0	0,0	-0,0	-0,0	0,0	0,0	0,0	-0,0	0,0	-0,0	0,0	-0,0
Hmotn	0,0	0,0	-0,0	0,0	0,0	0,0	0,0	-0,0	0,0	-0,0	0,0	-0,0
Boty	0,0	0,0	-0,0	-0,0	0,0	0,0	0,0	-0,0	0,0	0,0	0,0	-0,0
Prijem	-0,0	0,0	0,6	0,0	-0,0	-0,0	-0,0	0,5	-0,1	0,2	-0,1	0,1
Pivo	0,0	0,0	-0,2	0,0	0,0	0,0	0,0	-0,1	0,1	-0,0	0,0	-0,0
Vino	0,0	0,0	0,2	-0,1	-0,0	-0,0	0,0	0,2	-0,0	0,2	-0,0	-0,0
Plavan	0,0	0,0	-0,1	-0,0	0,0	0,0	0,0	-0,1	0,0	-0,0	0,1	-0,0
Puvod	-0,0	0,0	0,1	-0,0	-0,0	-0,0	-0,0	0,1	-0,0	-0,0	-0,0	0,0

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