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

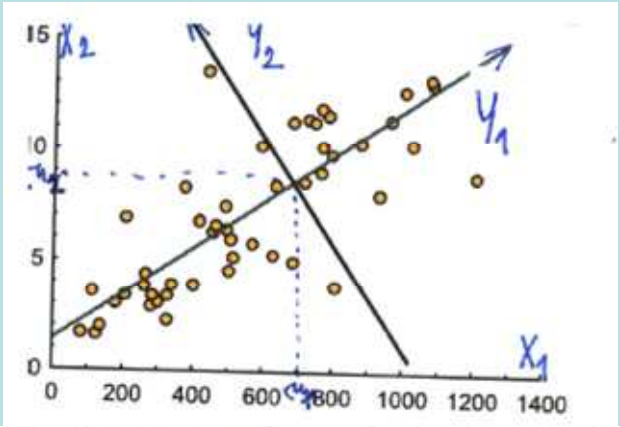
Cíl PCA:

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.)
6. Hlavní komponenty lze interpretovat jako hlavní osy p-rozměrného elipsoidu $\mathbf{x'S^{-1}x}$ = konst., kde **x** je vektor původních znaků a **S** je jeho varianční matice.

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

$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - m_j\right)^2$ - 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

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

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

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

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

Pro výpočet výběrové varianční matice potřebujeme vektory centrovaných hodnot:

$$\begin{pmatrix} 3-6 \\ 7-8 \end{pmatrix} = \begin{pmatrix} -3 \\ -1 \end{pmatrix}, \begin{pmatrix} 5-6 \\ 6-8 \end{pmatrix} = \begin{pmatrix} -1 \\ -2 \end{pmatrix}, \begin{pmatrix} 6-6 \\ 8-8 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 7-6 \\ 10-8 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 9-6 \\ 9-8 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$$

Pak

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

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

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

V našem případě:

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

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

$$= \frac{1}{4}[(3-6)\cdot(7-8) + (5-6)\cdot(6-8) + (6-6)\cdot(8-8) + (7-6)\cdot(10-8) + (9-6)\cdot(9-8)] =$$

$$= \frac{10}{4} = 2{,}5$$

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

Pro výpočet výběrové korelační matice potřebujeme vektory standardizovaných hodnot:

$$\begin{pmatrix} \frac{3-6}{\sqrt{5}} \\ \frac{7-8}{\sqrt{2,5}} \end{pmatrix} = \begin{pmatrix} \frac{-3}{\sqrt{5}} \\ \frac{-1}{\sqrt{2,5}} \end{pmatrix}, \begin{pmatrix} \frac{5-6}{\sqrt{5}} \\ \frac{6-8}{\sqrt{2,5}} \end{pmatrix} = \begin{pmatrix} \frac{-1}{\sqrt{5}} \\ \frac{-2}{\sqrt{2,5}} \end{pmatrix}, \begin{pmatrix} \frac{6-6}{\sqrt{5}} \\ \frac{8-8}{\sqrt{2,5}} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{7-6}{\sqrt{5}} \\ \frac{10-8}{\sqrt{2,5}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{2,5}} \end{pmatrix}, \begin{pmatrix} \frac{9-6}{\sqrt{5}} \\ \frac{9-8}{\sqrt{2,5}} \end{pmatrix} = \begin{pmatrix} \frac{3}{\sqrt{5}} \\ \frac{1}{\sqrt{2,5}} \end{pmatrix}$$

Pak

$$\frac{1}{4}\left[\begin{pmatrix} \frac{-3}{\sqrt{5}} \\ \frac{-1}{\sqrt{2,5}} \end{pmatrix} \cdot \left(\frac{-3}{\sqrt{5}}, \frac{-1}{\sqrt{2,5}}\right) + \begin{pmatrix} \frac{-1}{\sqrt{5}} \\ \frac{-2}{\sqrt{2,5}} \end{pmatrix} \cdot \left(\frac{-1}{\sqrt{5}}, \frac{-2}{\sqrt{2,5}}\right) + \begin{pmatrix} \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{2,5}} \end{pmatrix} \cdot \left(\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{2,5}}\right) + \begin{pmatrix} \frac{3}{\sqrt{5}} \\ \frac{1}{\sqrt{2,5}} \end{pmatrix} \cdot \left(\frac{3}{\sqrt{5}}, \frac{1}{\sqrt{2,5}}\right)\right] =$$

$$\mathbf{R} = = \frac{1}{4}\left[\begin{pmatrix} \frac{9}{5} & \frac{3}{\sqrt{12,5}} \\ \frac{3}{\sqrt{12,5}} & \frac{1}{2,5} \end{pmatrix} + \begin{pmatrix} \frac{1}{5} & \frac{2}{\sqrt{12,5}} \\ \frac{2}{\sqrt{12,5}} & \frac{4}{2,5} \end{pmatrix} + \begin{pmatrix} \frac{1}{5} & \frac{2}{\sqrt{12,5}} \\ \frac{2}{\sqrt{12,5}} & \frac{4}{2,5} \end{pmatrix} + \begin{pmatrix} \frac{9}{5} & \frac{3}{\sqrt{12,5}} \\ \frac{3}{\sqrt{12,5}} & \frac{1}{2,5} \end{pmatrix}\right] =$$

$$\frac{1}{4}\begin{pmatrix} \frac{20}{5} & \frac{10}{\sqrt{12,5}} \\ \frac{10}{\sqrt{12,5}} & \frac{10}{2,5} \end{pmatrix} = \begin{pmatrix} 1 & 0,707 \\ 0,707 & 1 \end{pmatrix}$$

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:

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

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

	Popisné statistiky (Dva_znaky.sta)
Proměnná	Průměr
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

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

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

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

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 $|\mathbf{A} - \lambda\mathbf{I}|$.

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 $\lambda_1, ..., \lambda_p$. 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
$v_{j1}^2 + v_{j2}^2 + \dots v_{jp}^2 = 1$, $v_{j1}v_{k1} + v_{j2}v_{k2} + \dots + v_{jp}v_{kp} = 0$ 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 $y_{1i} = v_{11}x_{i1} + v_{12}x_{i2} + \dots + v_{1p}x_{ip}$, i = 1, …, 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 $v_{11}v_{21} + v_{12}v_{22} + \dots + v_{1p}v_{2p} = 0$, tj. 1. a 2. hlavní komponenta jsou lineárně nezávislé.
Rozptyl 2. hlavní komponenty je l_2.

Dosadíme-li za $X_1, ..., X_p$ vektory pozorování $\mathbf{x}_i$, i = 1, ..., n, dostaneme vektor souřadnic $\mathbf{y}_2 = (y_{21}, ..., y_{2n})^T$, kde $y_{2i} = v_{21}x_{i1} + v_{22}x_{i2} + \dots + v_{2p}x_{ip}$, i = 1, …, 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 $v_{j1}v_{k1} + v_{j2}v_{k2} + \dots + v_{jp}v_{kp} = 0$, 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 $y_{ji} = v_{j1}x_{i1} + v_{j2}x_{i2} + \dots + v_{jp}x_{ip}$, i = 1, …, 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 $\frac{l_1}{l_1 + \ldots + l_p} 100\%$ celkové variability. Pokud je číslo $\frac{l_1}{l_1 + \ldots + l_p}$ 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 $R(X_i, Y_k) = \frac{v_{ki}\sqrt{l_k}}{s_i}$.

Reprodukce výchozí kovarianční matice:

V teorii matic se dokazuje vzorec $S = \sum_{i=1}^{p} l_i v_i \mathbf{v}_i^T$ (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 $\chi^2 = \frac{11+2p-6n}{6}\ln|\mathbf{R}|$. 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 $\chi^2 \geq \chi^2_{1-\alpha}(p(p-1)/2)$. Nezamítneme-li nulovou hypotézu, neměli bychom analýzu hlavních komponent vůbec provádět (Bartlettův test je implementován např. v systému SPSS). Test je použitelný pro n > 150.

Lze spočítat též Gleasonovu – Staelinovu míru redundance $\Phi = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^{\,2} - p}{p(p-1)}}$. Nabývá hodnot mezi 0 a 1, 0 značí, že mezi proměnnými není žádná korelace, 1 znamená perfektní korelaci.

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.

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 roztříděny 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.

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 $|\mathbf{S} - l\mathbf{I}| = 0$ 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

$\frac{l_1}{st(S)}100\% = \frac{680{,}411}{689{,}77}100\% = 98{,}65\%$ 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: $Y_2 = v_{21}X_1 + ... + v_{2p}X_p = 0{,}5454X_1 - 0{,}8321X_2 - 0{,}1009X_3$, 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: $Y_3 = v_{31}X_1 + \ldots + v_{3p}X_p = 0{,}2053\ X_1 + 0{,}2493\ X_2 - 0{,}9464\ X_3$, vyčerpává

$\frac{l_3}{st(S)}100\% = \frac{2{,}8573}{689{,}77}100\% = 0{,}41\%$ variability obsažené v datovém souboru.

Výpočet koeficientů korelace:

$$R(X_1, Y_3) = \frac{v_{31}\sqrt{l_3}}{s_1} = \frac{0{,}2053\sqrt{2{,}8573}}{\sqrt{451{,}39}} = 0{,}0163$$

$$R(X_2, Y_3) = \frac{v_{32}\sqrt{l_3}}{s_2} = \frac{0{,}2493\sqrt{2{,}8573}}{\sqrt{171{,}73}} = 0{,}0322$$

$$R(X_3, Y_3) = \frac{v_{33}\sqrt{l_3}}{s_3} = \frac{-0{,}9464\sqrt{2{,}8573}}{\sqrt{66{,}69}} = -0{,}1959$$

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

Tabulka korelací původních proměnných a hlavních komponent

proměnná	komponenta		
	Y_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:

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

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

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

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.

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

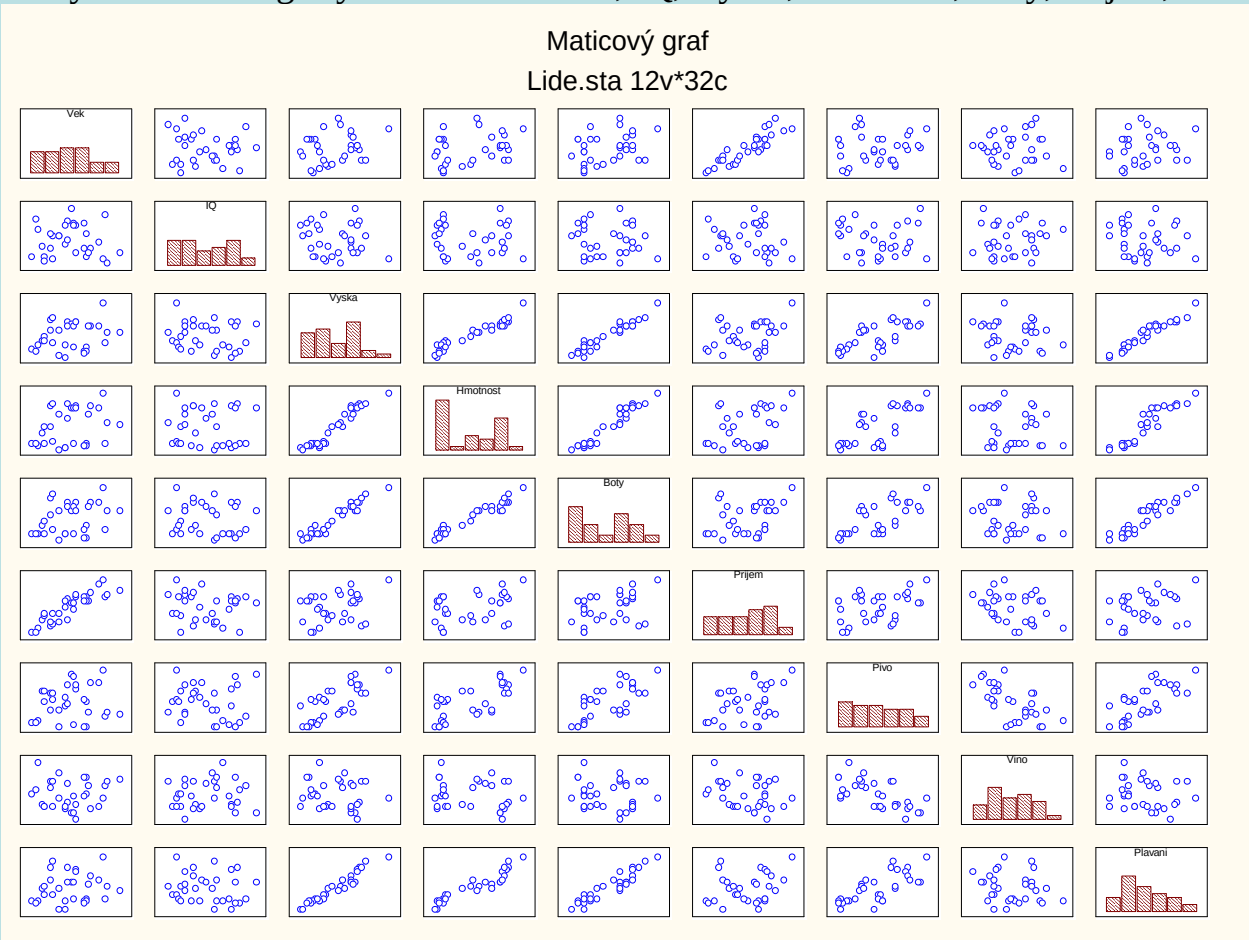
	1 Sex	2 Vlasy	3 Vek	4 IQ	5 Vyska	6 Hmotnost	7 Boty	8 Prijem	9 Pivo	10 Vino	11 Plavani	12 Puvod
	muz	kratke	48	100	198	92	48	45000	420	115	98	Skandinavie
	muz	kratke	33	130	184	84	44	33000	350	102	92	Skandinavie
	muz	kratke	37	127	183	83	44	34000	320	98	91	Skandinavie
	zena	kratke	32	112	166	47	36	28000	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 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.

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 výpočtem Gleasonovy – Staelinovy míry redundance $\Phi = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^{2} - p}{p(p-1)}}$.

K výstupní tabulce, v níž je uložena korelační matice, přidáme novou proměnnou, která bude obsahovat součty kvadrátů korelačních koeficientů. Do jejího Dlouhého jména napíšeme:

=v1^2+v2^2+v3^2+v4^2+v5^2+v6^2+v7^2+v8^2+v9^2+v10^2+v11^2+v12^2

Pomocí Statistiky – Blok sloupců – Součty získáme součet této proměnné. Přidáme další proměnnou a do jejího Dlouhého jména napíšeme: =sqrt((v1-12)/132)

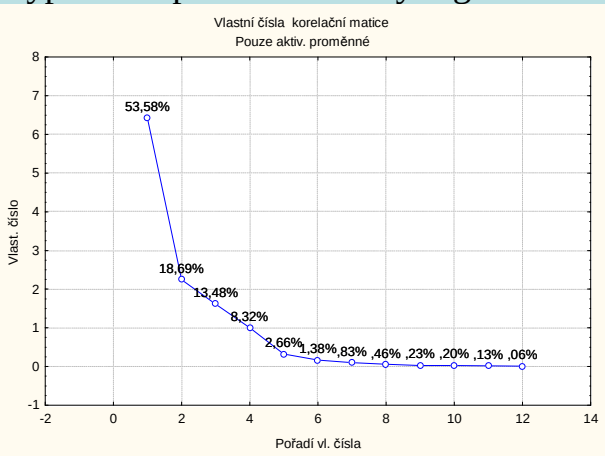
	Korelace (Lide.sta)	
Proměnná	1 NProm	2 NProm
SOUČET případy 1-12	50,1262654	0,53743404

Vidíme, že koeficient Φ = 0,5374 nabývá dostatečně velké hodnoty pro prokázání korelace v datech.

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

Pořadí vl.č.	Vlastní čísla korelační matice a související statistiky (Lide.sta) Pouze aktiv. proměnné			
	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.

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.

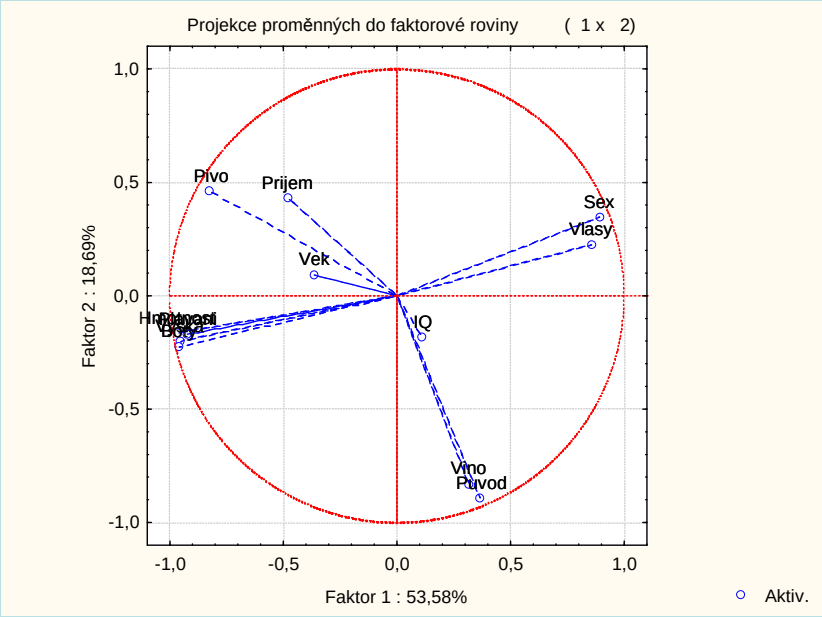
	Vlastní vektory korelační matice (Lide.sta) Pouze aktiv. proměnné	
Proměnná	Faktor 1	Faktor 2
Sex	0,351783	0,231671
Vlasy	0,337773	0,150163
Vek	-0,142945	0,061463
IQ	0,044067	-0,122604
Vyska	-0,375286	-0,135459
Hmotnost	-0,381136	-0,111447
Boty	-0,377697	-0,150806
Prijem	-0,190466	0,286893
Pivo	-0,324666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

1. hlavní komponenta:
Y_1 = 0,35Sex + 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	0,892009	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

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

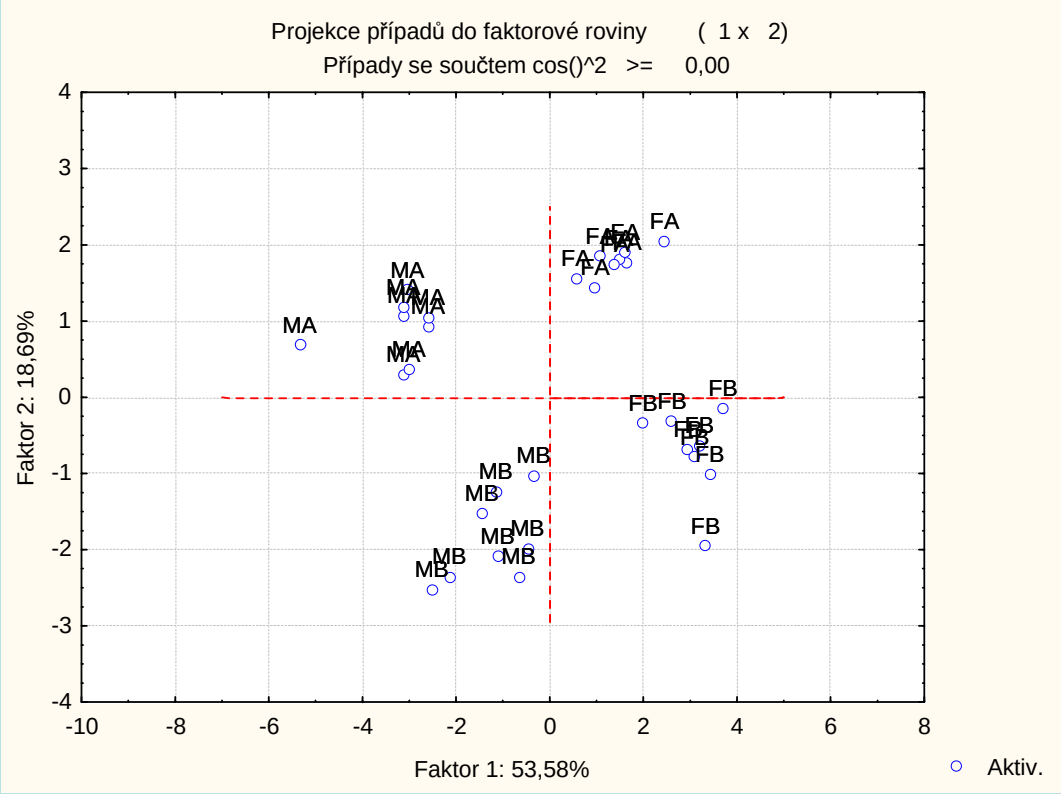
Pomocí grafu faktorových souřadnic proměnných lze posoudit tyto skutečnosti:

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.

Proměnná	Reprodukované korelace (Lide.sta) Extrakce: Hlavní komponenty											
	Sex	Vlasy	Vek	IQ	Vyska	Hmotnost	Boty	Prijem	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
Vek	-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
Prijem	-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

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

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