Přednáška č. 2.: 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. Tyto nové znaky, kterým se říká hlavní komponenty, jsou nekorelované a jsou uspořádané podle svého klesajícího rozptylu. Většina informace o variabilitě původních dat je tedy soustředěna v první hlavní komponentě a nejméně informace je obsaženo v poslední komponentě. Ukazuje se, že pouze několik prvních hlavních komponent má dostatečně velký rozptyl. Ostatní pak můžeme zanedbat, čímž docílíme snížení dimenze dat. 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.

Podstata metody hlavních komponent

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

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

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

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

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

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

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

Pro výpočet výběrové korelační matice potřebujeme vektory standardizovaných hodnot:

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

Pak

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

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

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

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

Výpočet pomocí systému SPSS:

Potřebujeme datový soubor o dvou proměnných X1, X2 a 5 případech
Získání vektoru průměrů: Analyze – Descriptive Statistics – Descriptives – Variables X1, X2 – Options – ponecháme jen Mean – Continue - OK

Descriptive Statistics

	N	Mean
X1	5	6,00
X2	5	8,00
Valid N (listwise)	5	

Získání varianční a korelační matice: Analyze – Correlate – Bivariate - Variables X1, X2 – Options – zaškrtneme Cross-product and covariances – Continue – OK

Correlations

		X1	X2
X1	Pearson Correlation	1	,707
	Sig. (2-tailed)		,182
	Sum of Squares and Cross-products	20,000	10,000
	Covariance	5,000	2,500
	N	5	5
X2	Pearson Correlation	,707	1
	Sig. (2-tailed)	,182	
	Sum of Squares and Cross-products	10,000	10,000
	Covariance	2,500	2,500
	N	5	5

Základní pojmy v metodě hlavních komponent

Nechť **A** je čtvercová matice řádu p. Číslo λ se nazývá vlastní číslo matice **A**, jestliže pro libovolný nenulový vektor **v** typu p x 1 splňuje rovnici $\mathbf{Av} = \lambda\mathbf{v}$. Vektor **v** se nazývá vlastní vektor matice **A**. Determinant $|\mathbf{A} - \lambda\mathbf{I}|$ se nazývá charakteristický polynom matice **A**. Stopou matice **A** rozumíme součet jejích diagonálních prvků.
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. Do-

staneme 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 $S = \sum_{i=1}^{p} l_i v_i v_i^T$ (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 $\chi^2(p(p-1)/2)$. Nulo-

vou 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).

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

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. HK: $Y_1 = v_{11}X_1 + ... + v_{1p}X_p = 0{,}8126X_1 + 0{,}4955X_2 + 0{,}3068X_3$, 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. 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. HK: $Y_3 = v_{31}X_1 + ... + v_{3p}X_p = 0{,}2053\ X_1 + 0{,}2493\ X_2 - 0{,}9464\ X_3$, 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:

$$l_1\mathbf{v}_1\,\mathbf{v}_1^T = 680{,}411\begin{pmatrix}0{,}8126\\0{,}4955\\0{,}3068\end{pmatrix}(0{,}8126 \quad 0{,}4955 \quad 0{,}3068) = \begin{pmatrix}449{,}2881 & 273{,}9629 & 169{,}6303\\273{,}9629 & 167{,}0547 & 103{,}4357\\169{,}6303 & 103{,}4357 & 64{,}0445\end{pmatrix}$$

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 k dispozici datový soubor z roku 1979 o 26 evropských zemích, který obsahuje údaje o procentuálním zastoupení ekonomicky činného obyvatelstva v různých odvětvích národního hospodářství:

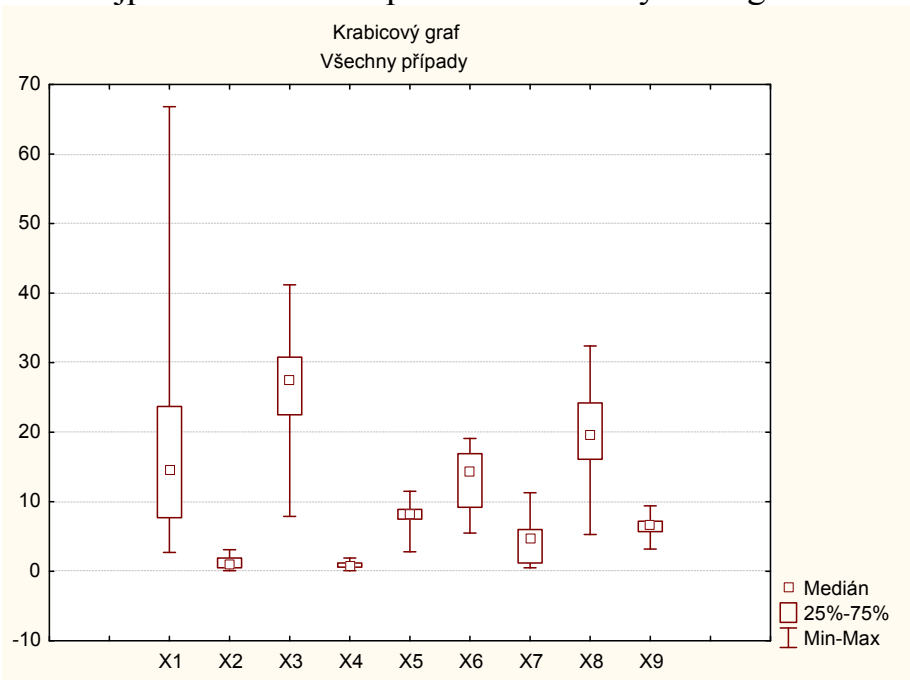
X_1 … zemědělství
X_2 … těžba
X_3 … průmyslová výroba
X_4 … energetika
X_5 … stavebnictví
X_6 … místní hospodářství
X_7 … finanční sektor
X_8 … služby
X_9 … doprava a komunikace.

	1 Stát	2 X1	3 X2	4 X3	5 X4	6 X5	7 X6	8 X7	9 X8	10 X9
1	Belgie	3,3	0,9	27,6	0,9	8,2	19,1	6,2	26,6	7,2
2	Dánsko	9,2	0,1	21,8	0,6	8,3	14,2	6,5	32,2	7,1
3	Francie	10,8	0,8	27,5	0,9	8,9	16,8	6	22,6	5,7
4	Záp. Německo	6,7	1,3	35,8	0,9	7,3	14,4	5	22,5	6,1
5	Irsko	23,2	1	20,7	1,3	7,5	16,8	2,8	20,6	6,1
6	Itálie	15,9	0,6	27,6	0,5	10	18,1	1,5	20,1	5,7
7	Lucembursko	7,7	3,1	30,8	0,8	9,2	18,5	4,5	19,2	6,2
8	Nizozemsko	6,3	0,1	22,5	1	9,9	18	6,9	28,5	6,8
9	Velká Británie	2,7	1,4	30,2	1,4	6,9	16,9	5,8	28,3	6,4
10	Rakousko	12,7	1,1	31,4	1,4	8	16,8	4,9	16,7	7
11	Finsko	13	0,4	25,9	1,3	7,4	14,7	5,5	24,2	7,6
12	Řecko	41,4	0,6	17,6	0,6	8,1	11,5	2,4	11,1	6,7
13	Norsko	9	0,5	22,4	0,8	8,6	16,9	4,7	27,7	9,4
14	Portugalsko	27,8	0,3	24,5	0,6	8,4	13,3	2,7	16,7	5,7
15	Španělsko	22,9	0,8	28,5	0,7	11,5	9,7	8,5	11,9	5,5
16	Švédsko	6,1	0,4	25,9	0,8	7,2	14,4	6	32,4	6,8
17	Švýcarsko	7,7	0,2	37,8	0,8	9,5	17,5	5,3	15,5	5,7
18	Turecko	66,8	0,7	7,9	0,1	2,8	5,5	1,1	11,9	3,2
19	Bulharsko	23,6	1,9	32,3	0,6	7,9	8	0,7	18,2	6,8
20	Československo	16,5	2,9	35,5	1,2	8,7	9,2	0,9	17,9	7,2
21	Vých. Německo	4,2	2,9	41,2	1,3	7,6	11,2	1,2	22,1	8,3
22	Maďarsko	21,7	3,1	29,6	1,9	8,2	9,4	0,9	17,2	8
23	Polsko	31,1	2,5	25,7	0,9	8,4	7,5	0,9	16,1	6,9
24	Rumunsko	34,7	2,1	30,1	0,6	8,7	5,9	1,3	11,6	5
25	Sovětský svaz	23,7	1,4	25,8	0,6	9,2	6,1	0,5	23,4	9,3
26	Jugoslávie	48,7	1,5	16,8	1,1	4,9	6,4	11,3	5,3	4

Tento datový soubor analyzujte metodou hlavních komponent.

Řešení v systému STATISTICA:

Data nejprve znázorníme pomocí krabicových diagramů:



Proměnné vykazují značně rozdílnou variabilitu. Analýzu tedy založíme na výběrové korelační matici **R**:

Statistiky – Vícerozměrné průzkumné techniky – Hlavní komponenty&klasifikační analýza – Proměnné X1 až X19, OK – OK – Popisné statistiky – Korelační matice.

	Korelace (staty1979.sta)								
Proměnná	X1	X2	X3	X4	X5	X6	X7	X8	X9
X1	1,00	0,04	-0,67	-0,40	-0,53	-0,73	-0,22	-0,75	-0,56
X2	0,04	1,00	0,44	0,41	-0,02	-0,40	-0,44	-0,28	0,16
X3	-0,67	0,44	1,00	0,39	0,48	0,21	-0,15	0,15	0,36
X4	-0,40	0,41	0,39	1,00	0,03	0,20	0,11	0,13	0,37
X5	-0,53	-0,02	0,48	0,03	1,00	0,33	0,01	0,17	0,38
X6	-0,73	-0,40	0,21	0,20	0,33	1,00	0,36	0,57	0,17
X7	-0,22	-0,44	-0,15	0,11	0,01	0,36	1,00	0,11	-0,25
X8	-0,75	-0,28	0,15	0,13	0,17	0,57	0,11	1,00	0,56
X9	-0,56	0,16	0,36	0,37	0,38	0,17	-0,25	0,56	1,00

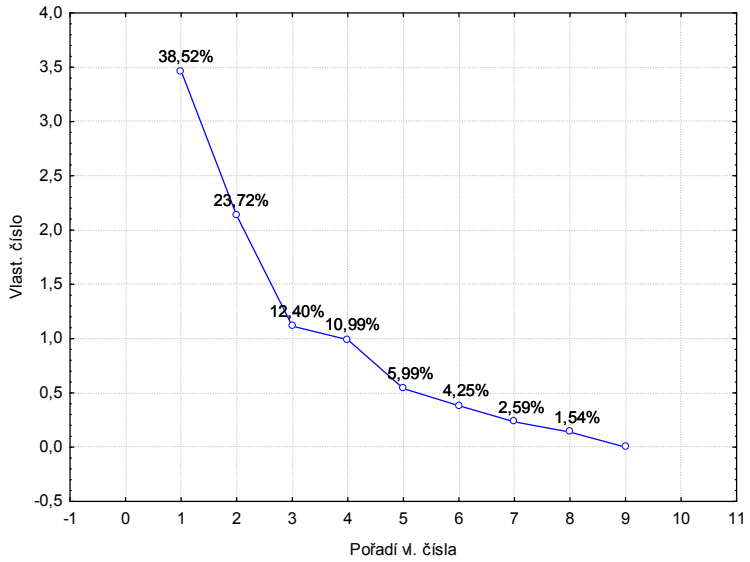
Tato korelační matice má bohužel determinant blízký 0 (říkáme, že je špatně podmíněná), nelze tedy provést Bartlettův test. Je však vidět, že některé korelační koeficienty jsou v absolutní hodnotě dostatečně velké a zřejmě tedy bude mít smysl provést analýzu hlavních komponent.

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

Pořadí vl.č.	vl. číslo	% celk. rozptylu	Kumulativ. vl. číslo	Kumulativ. %
1	3,466490	38,51655	3,466490	38,5166
2	2,135004	23,72227	5,601494	62,2388
3	1,115581	12,39534	6,717075	74,6342
4	0,989394	10,99326	7,706468	85,6274
5	0,539211	5,99123	8,245679	91,6187
6	0,382111	4,24568	8,627790	95,8643
7	0,233226	2,59140	8,861015	98,4557
8	0,138985	1,54428	9,000000	100,0000

První hlavní komponenta tedy vysvětluje 38,52% variability obsažené v devíti sledovaných proměnných, druhá 23,72%, třetí 12,40% atd. Celkové procento variability vysvětlené prvními třemi hlavními komponentami je 74,63%.

Sestrojíme sutinový graf (scree plot): na záložce Základní výsledky vybereme Sutinový graf.



Počet m hlavních komponent zvolíme tři na základě sutinového grafu, na základě vysvětleného rozptylu a na základě Kaiserova kritéria (první tři vlastní čísla jsou větší než 1). V nabídce Výsledky hlavních komponent snížíme počet faktorů na 3.

Vypočteme korelační koeficienty prvních tří hlavních komponent a původních devíti proměnných: na záložce Proměnné vybereme Korelace faktorů & proměnných.

	Korelace faktorů a proměnných (faktor. zátěže) podle korelací (staty1979.sta)		
Proměnná	Faktor 1	Faktor 2	Faktor 3
X1	0,978776	0,081725	-0,049455
X2	-0,000898	0,901105	0,216344
X3	-0,652174	0,513343	0,112868
X4	-0,474888	0,378598	0,649962
X5	-0,595263	0,073032	-0,304047
X6	-0,698213	-0,513734	0,119592
X7	-0,136193	-0,663299	0,589451
X8	-0,727506	-0,327637	-0,251642
X9	-0,684094	0,304809	-0,337074

Podívejme se rovněž na vektory souřadnic (v systému STATISTICA se jim říká faktorové souřadnice případů): na záložce Případy vybereme Faktorové souřadnice případů.

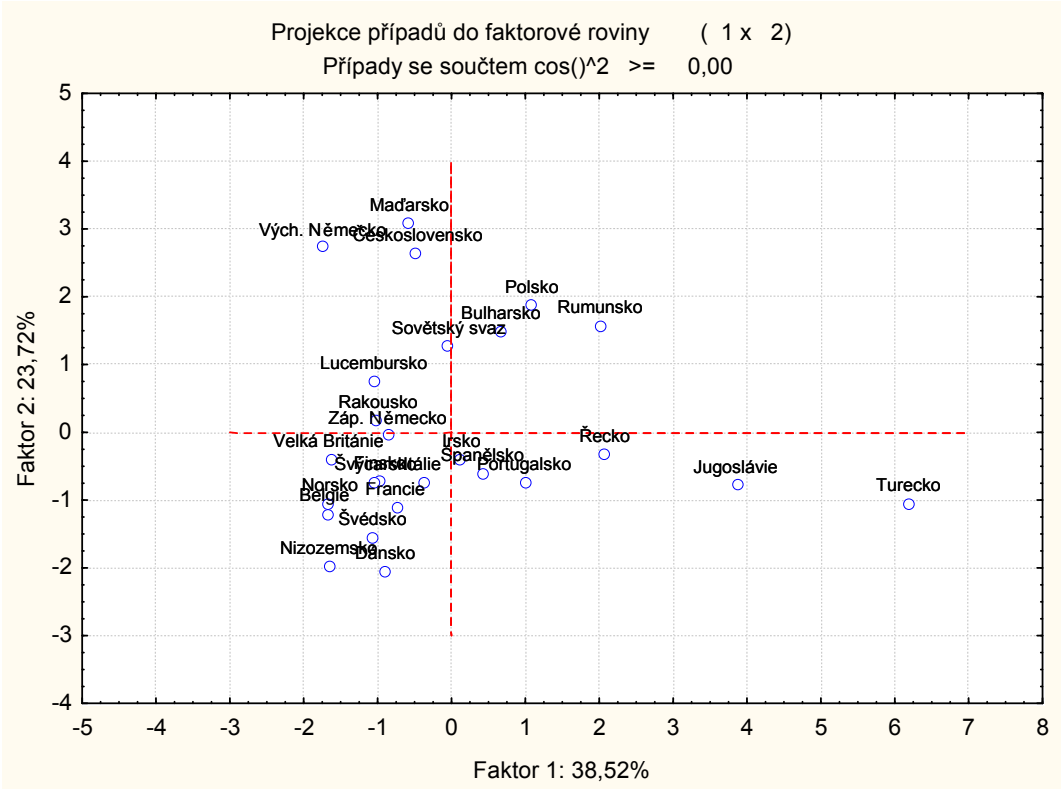
Případ	Faktorové souřadnice případů podle korelací (staty1979.sta		
	Faktor 1	Faktor 2	Faktor 3
Belgie	-1,68273	-1,20656	0,16668
Dánsko	-0,90831	-2,05598	-0,85147
Francie	-0,74050	-1,11048	0,38553
Záp. Německo	-0,85647	-0,03165	0,56466
Irsko	0,11153	-0,40400	0,53134
Itálie	-0,36366	-0,74902	-1,29050
Lucembursko	-1,04022	0,74294	0,46327
Nizozemsko	-1,65732	-1,98866	-0,08729
Velká Británie	-1,61201	-0,39776	1,35031
Rakousko	-1,01103	0,16508	1,16804
Finsko	-0,97223	-0,73166	0,54475
Řecko	2,07154	-0,33521	-0,92274
Norsko	-1,66538	-1,05092	-1,14341
Portugalsko	0,99709	-0,74259	-0,75474
Španělsko	0,43244	-0,60818	0,31825
Švédsko	-1,07387	-1,55390	-0,22815
Švýcarsko	-1,04031	-0,74707	0,28216
Turecko	6,19519	-1,04930	-0,64265
Bulharsko	0,67558	1,48159	-1,03101
Československo	-0,48005	2,63421	0,07902
Vých. Německo	-1,73669	2,73412	0,26970
Maďarsko	-0,57526	3,07981	1,09460
Polsko	1,08637	1,87264	-0,54684
Rumunsko	2,01536	1,57550	-0,48595
Sovětský svaz	-0,04779	1,26246	-2,30671
Jugoslávie	3,87872	-0,78542	3,07316

1. HK vysoce kladně koreluje s proměnnou X_1 (zemědělství) a záporně se všemi ostatními proměnnými. Tato hlavní komponenta tedy rozlišuje země na zemědělské a průmyslové. Povšimněte si, že souřadnice této hlavní komponenty jsou nejvyšší u Turecka (6,2) a Jugoslávie (3,9).

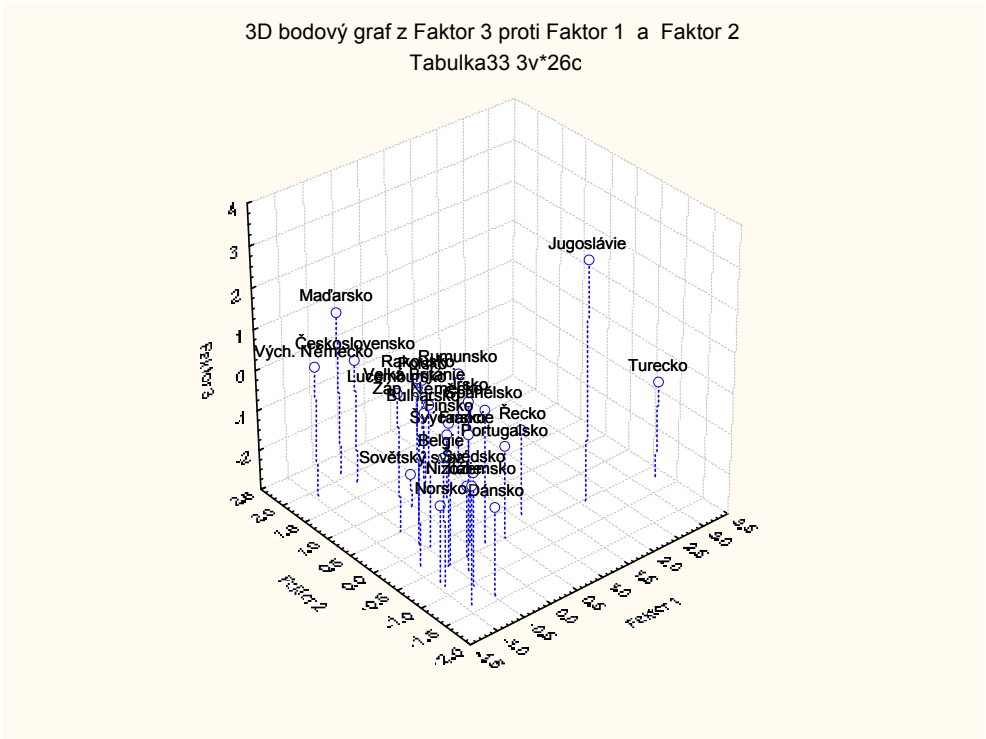
2. HK vysoce kladně koreluje s proměnnou X_2 (těžba) a podstatně slaběji s proměnnou X_3 (průmyslová výroba). Vysoké hodnoty souřadnic této hlavní komponenty najdeme u Maďarska, Východního Německa a Československa.

3. HK středně silně koreluje s proměnnou X_4 (energetika) a X_7 (finanční sektor). Nejvyšší hodnotu najdeme u Jugoslávie.

Nyní znázorníme rozmístění zemí na ploše prvních dvou hlavních komponent: Na záložce Případy vybereme 2D graf fakt. Souřadnic příp.

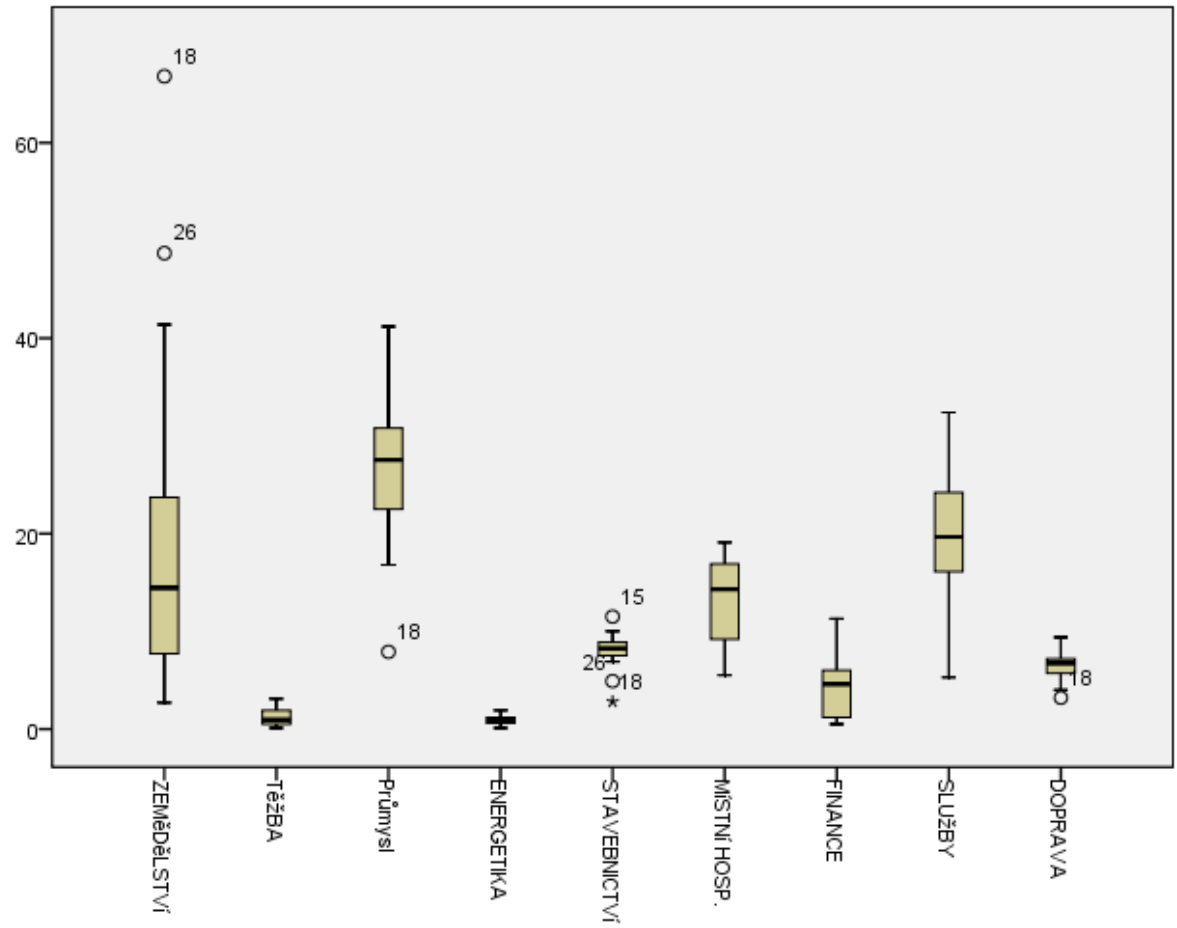


Můžeme se ještě pokusit o znázornění zemí v prostoru prvních tří hlavních komponent: přepneme se v pracovním sešitě na tabulku Faktorové souřadnice případů dle korelací. Označíme myší 3 hlavní komponenty. Klikneme pravým tlačítkem, vybereme Grafy bloku dat – Vlastní graf bloku podle sloupce – 3D XYZ grafy – Bodové grafy – Běžný – OK, 2x klikneme na pozadí grafu – Popisy bodů – zaškrtneme Zobrazovat popisy bodů.



Řešení v systému SPSS:

Hodnoty všech 9 proměnných znázorníme pomocí krabicových diagramů.
Graphs – Legacy Dialogs – Box plot – zaškrtneme Summaries of separate variables – Define – Boxes represent x1 až x9 - OK



Výpočet výběrové korelační matice: Analyze – Correlate – Bivariate – Variables x1 až x9 – OK
Postup při provedení Bartlettova testu: Analyze – Data Reduction – Factor - Variables x1 až x9 – Descriptives – zaškrtneme KMO and Bartlett's test of sphericity. V našem případě výsledek neobdržíme, protože korelační matice má determinant blízký 0.
Nyní získáme vlastní čísla výběrové korelační matice a procento vysvětleného rozptylu:
Analyze – Data Reduction – Factor - Variables x1 až x9 – OK

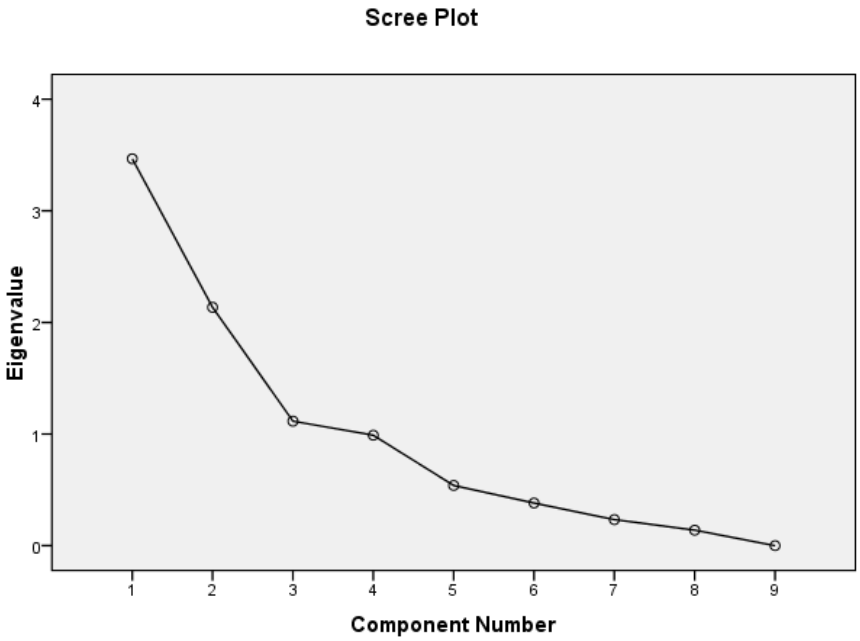
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,466	38,517	38,517	3,466	38,517	38,517
2	2,135	23,722	62,239	2,135	23,722	62,239
3	1,116	12,395	74,634	1,116	12,395	74,634
4	,989	10,993	85,627			
5	,539	5,991	91,619			
6	,382	4,246	95,864			
7	,233	2,591	98,456			
8	,139	1,544	100,000			
9	8,077E-17	8,975E-16	100,000			

Extraction Method: Principal Component Analysis.

(Systém SPSS je implicitně nastaven tak, že uvažuje tolik hlavních komponent, kolik je vlastních čísel větších než 1.)

Chceme-li znázornit sutinový graf, ve volbě Extraction zaškrtneme Scree plot.



Ve výstupu se objeví rovněž korelační koeficienty prvních tří hlavních komponent a původních devíti proměnných:

Component Matrix[a]

	Component		
	1	2	3
ZEMěDěLSTVí	-,979	,082	-,049
TěžBA	,001	,901	,216
Průmysl	,652	,513	,113
ENERGETIKA	,475	,379	,650
STAVEBNICTVí	,595	,073	-,304
MíSTNí HOSP.	,698	-,514	,120
FINANCE	,136	-,663	,589
SLUžBY	,728	-,328	-,252
DOPRAVA	,684	,305	-,337

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

(Na rozdíl od systému STATISTICA dostáváme korelační koeficienty s opačnými znaménky.)

Můžeme získat rovněž reprodukovanou korelační matici a reziduální matici: ve volbě Descriptives zaškrtneme Correlation Matrix Reprodused.

Reproduced Correlations

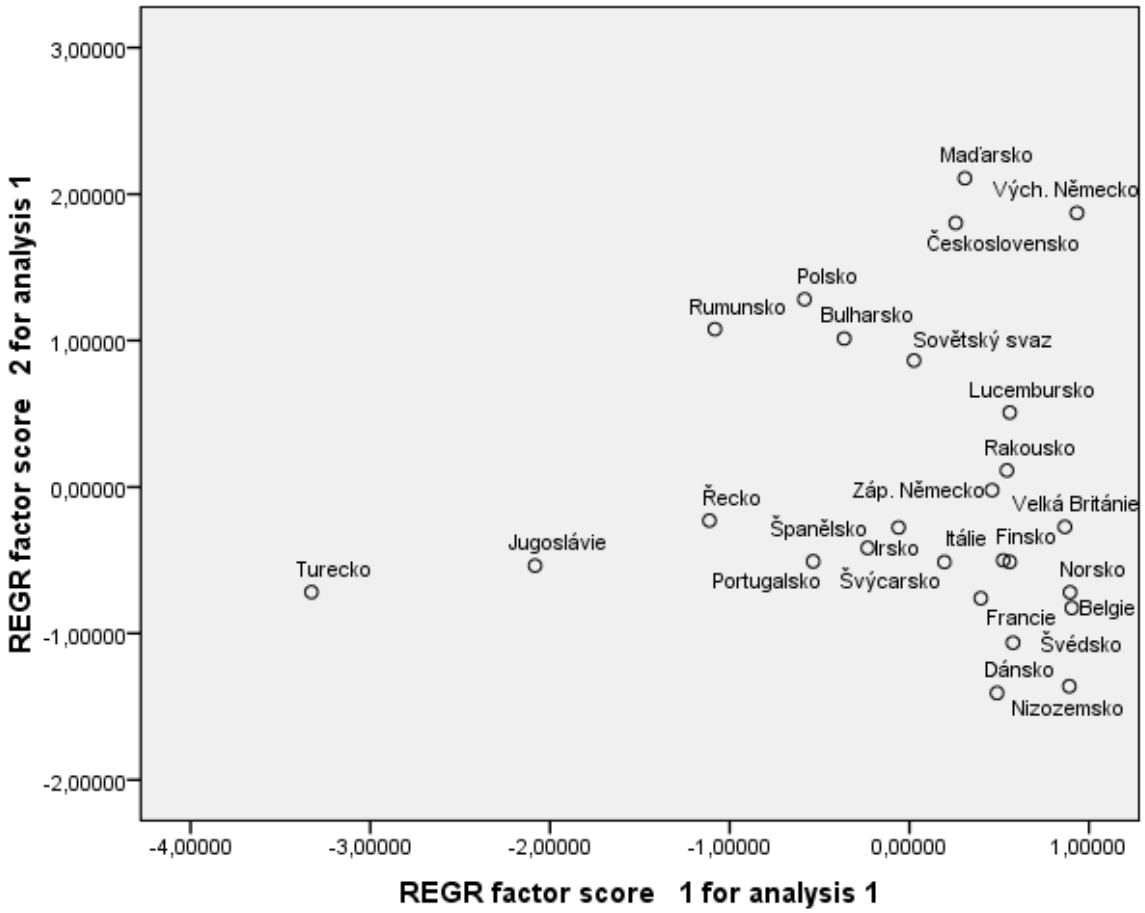
		ZEMěDěLSTVí	TěžBA	Průmysl	ENERGETIKA	STAVEBNICTVí	MíSTNí HOSP.	FINANCE	SLUžBY	DOPRAVA
Reproduced Correlation	ZEMěDěLSTVí	,967[a]	,062	-,602	-,466	-,562	-,731	-,217	-,726	-,628
	TěžBA	,062	,859[a]	,488	,482	,001	-,436	-,470	-,349	,202
	Průmysl	-,602	,488	,702[a]	,577	,391	,205	-,185	,278	,565
	ENERGETIKA	-,466	,482	,577	,791[a]	,113	,215	,197	,058	,221
	STAVEBNICTVí	-,562	,001	,391	,113	,452[a]	,342	-,147	,486	,532
	MíSTNí HOSP.	-,731	-,436	,205	,215	,342	,766[a]	,506	,646	,281
	FINANCE	-,217	-,470	-,185	,197	-,147	,506	,806[a]	,168	-,308
	SLUžBY	-,726	-,349	,278	,058	,486	,646	,168	,700[a]	,483
	DOPRAVA	-,628	,202	,565	,221	,532	,281	-,308	,483	,675[a]
Residual[b]	ZEMěDěLSTVí		-,026	-,070	,066	,031	,001	-,004	-,023	,065
	TěžBA	-,026		-,045	-,077	-,022	,040	,026	,066	-,039
	Průmysl	-,070	-,045		-,185	,092	,000	,031	-,125	-,209
	ENERGETIKA	,066	-,077	-,185		-,084	-,015	-,083	,072	,154
	STAVEBNICTVí	,031	-,022	,092	-,084		-,011	,153	-,314	-,147
	MíSTNí HOSP.	,001	,040	,000	-,015	-,011		-,146	-,078	-,106
	FINANCE	-,004	,026	,031	-,083	,153	-,146		-,054	,057
	SLUžBY	-,023	,066	-,125	,072	-,314	-,078	-,054		,081
	DOPRAVA	,065	-,039	-,209	,154	-,147	-,106	,057	,081	

Extraction Method: Principal Component Analysis.

a. Reproduced communalities

b. Residuals are computed between observed and reproduced correlations. There are 22 (61,0%) nonredundant residuals with absolute values greater than 0.05.

Znázorníme rozmístění zemí na ploše prvních dvou hlavních komponent: ve volbě Scores zaškrtneme Save as variables. K datovému souboru se přidají tři nové proměnné - vektory souřadnic, nazvané FAC1_1, FAC_2_1, FAC_3_1. Graphs – Legacy Dialogs – Scatter/Dot – ponecháme Simple Scatter – Define – Y Axis factor score 2, X Axis factor score 1, Label Cases by stat, Options - zaškrtneme Display chart with case labels.


3,00000
2,00000
1,00000
0,00000
-1,00000
-2,00000
REGR factor score 2 for analysis 1
-4,00000
-3,00000
-2,00000
-1,00000
0,00000
1,00000
REGR factor score 1 for analysis 1
Maďarsko
Vých. Německo
Československo
Polsko
Rumunsko
Bulharsko
Sovětský svaz
Lucembursko
Rakousko
Řecko
Záp. Německo
Velká Británie
Španělsko
Irsko
Itálie
Finsko
Jugoslávie
Turecko
Portugalsko
Švýcarsko
Norsko
Belgie
Francie
Švédsko
Dánsko
Nizozemsko