Příklad

Bylo provedeno měření výšky $\mathbf{x}_1$ (v cm) a váhy $\mathbf{x}_2$ (v kg) u pěti dětí. Naměřené hodnoty byly zaznamenány do matice $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} 101 & 16 \\ 105 & 18 \\ 103 & 42 \\ 98 & 23 \\ 93 & 6 \end{bmatrix}.$$

U tohoto datového souboru proveďte analýzu hlavních komponent.

Řešení:

U analýzy hlavních komponent potřebujeme nejprve spočítat kovarianční matici $\mathbf{S} = \begin{bmatrix} \mathrm{s}_{11} & \mathrm{s}_{12} \\ \mathrm{s}_{21} & \mathrm{s}_{22} \end{bmatrix}$. Pro výpočet kovarianční matice potřebujeme znát průměrnou výšku a váhu u $n = 5$ dětí:

$$\bar{\mathbf{x}} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_{i1} & \frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_{i2} \end{bmatrix} = \begin{bmatrix} \frac{1}{5}(101 + 105 + 103 + 98 + 93) & \frac{1}{5}(16 + 18 + 42 + 23 + 6) \end{bmatrix} = \begin{bmatrix} 100 & 21 \end{bmatrix}$$

Jednotlivé prvky kovarianční matice poté spočítáme následujícím způsobem:

Rozptyl výšky: $\mathrm{s}_{11} = \frac{1}{n-1}((\mathrm{x}_{11} - \bar{\mathrm{x}}_1)^2 + (\mathrm{x}_{21} - \bar{\mathrm{x}}_1)^2 + (\mathrm{x}_{31} - \bar{\mathrm{x}}_1)^2 + (\mathrm{x}_{41} - \bar{\mathrm{x}}_1)^2 + (\mathrm{x}_{51} - \bar{\mathrm{x}}_1)^2) = \frac{1}{5-1}((101 - 100)^2 + (105 - 100)^2 + (103 - 100)^2 + (98 - 100)^2 + (93 - 100)^2) = \frac{1}{4}(1 + 25 + 9 + 4 + 49) = \frac{1}{4} \cdot 88 = 22$

Rozptyl váhy: $\mathrm{s}_{22} = \frac{1}{n-1}((\mathrm{x}_{12} - \bar{\mathrm{x}}_2)^2 + (\mathrm{x}_{22} - \bar{\mathrm{x}}_2)^2 + (\mathrm{x}_{32} - \bar{\mathrm{x}}_2)^2 + (\mathrm{x}_{42} - \bar{\mathrm{x}}_2)^2 + (\mathrm{x}_{52} - \bar{\mathrm{x}}_2)^2) = \frac{1}{5-1}((16 - 21)^2 + (18 - 21)^2 + (42 - 21)^2 + (23 - 21)^2 + (6 - 21)^2) = \frac{1}{4}(25 + 9 + 441 + 4 + 225) = \frac{1}{4} \cdot 704 = 176$

Kovariance výšky a váhy: $\mathrm{s}_{12} = \mathrm{s}_{21} = \frac{1}{n-1}\big((\mathrm{x}_{11} - \bar{\mathrm{x}}_1)(\mathrm{x}_{12} - \bar{\mathrm{x}}_2) + (\mathrm{x}_{21} - \bar{\mathrm{x}}_1)(\mathrm{x}_{22} - \bar{\mathrm{x}}_2) + (\mathrm{x}_{31} - \bar{\mathrm{x}}_1)(\mathrm{x}_{32} - \bar{\mathrm{x}}_2) + (\mathrm{x}_{41} - \bar{\mathrm{x}}_1)(\mathrm{x}_{42} - \bar{\mathrm{x}}_2) + (\mathrm{x}_{51} - \bar{\mathrm{x}}_1)(\mathrm{x}_{52} - \bar{\mathrm{x}}_2)\big) = \frac{1}{5-1}\big((101 - 100)(16 - 21) + (105 - 100)(18 - 21) + (103 - 100)(42 - 21) + (98 - 100)(23 - 21) + (93 - 100)(6 - 21)\big) = \frac{1}{4}(-5 - 15 + 63 - 4 + 105) = \frac{1}{4} \cdot 144 = 36$

Kovarianční matice je tedy: $\mathbf{S} = \begin{bmatrix} \mathrm{s}_{11} & \mathrm{s}_{12} \\ \mathrm{s}_{21} & \mathrm{s}_{22} \end{bmatrix} = \begin{bmatrix} 22 & 36 \\ 36 & 176 \end{bmatrix}$.

Nyní spočítáme vlastní čísla a vlastní vektory kovarianční matice – tzn., spočítáme následující determinant: $\begin{vmatrix} 22 - \lambda & 36 \\ 36 & 176 - \lambda \end{vmatrix}$

Vypočteme charakteristický polynom: $(22-\lambda)\cdot(176-\lambda)-36^2=3872-176\lambda-22\lambda+\lambda^2-1296=\lambda^2-198\lambda+2576$

A jeho kořeny, které odpovídají vlastním číslům:

$\lambda_1=\frac{198+\sqrt{(-198)^2-4\cdot1\cdot2576}}{2\cdot1}=\frac{198+170}{2}=184$

$\lambda_2=\frac{198-\sqrt{(-198)^2-4\cdot1\cdot2576}}{2\cdot1}=\frac{198-170}{2}=14$

Následně spočítáme vlastní vektor $\mathbf{v}_1$ odpovídající prvnímu vlastnímu číslu $\lambda_1=184$:

$$\begin{bmatrix}22-184 & 36\\ 36 & 176-184\end{bmatrix}\sim\begin{bmatrix}-162 & 36\\ 36 & -8\end{bmatrix}\sim\begin{bmatrix}-162 & 36\\ -162 & 36\end{bmatrix}\sim\begin{bmatrix}-4{,}5 & 1\\ 0 & 0\end{bmatrix}$$

$\mathrm{v}_{12}=a$; $(-4{,}5)\cdot\mathrm{v}_{11}+\mathrm{v}_{12}=0\rightarrow\mathrm{v}_{11}=\frac{a}{4{,}5}$; např. pro $a=4{,}5$ pak dostáváme: $\mathbf{v}_1=[1\quad 4{,}5]$, který je po normalizaci roven $\mathbf{v}_1=\left[\frac{1}{\sqrt{1^2+4{,}5^2}}\quad\frac{4{,}5}{\sqrt{1^2+4{,}5^2}}\right]=[0{,}2169\quad 0{,}9762]$. Kontrola, že vektor má jednotkovou délku: $(0{,}2169)^2+(0{,}9762)^2=1$.

Spočítáme vlastní vektor $\mathbf{v}_2$ odpovídající druhému vlastnímu číslu $\lambda_2=14$:

$$\begin{bmatrix}22-14 & 36\\ 36 & 176-14\end{bmatrix}\sim\begin{bmatrix}8 & 36\\ 36 & 162\end{bmatrix}\sim\begin{bmatrix}36 & 162\\ 36 & 162\end{bmatrix}\sim\begin{bmatrix}1 & 4{,}5\\ 0 & 0\end{bmatrix}$$

$\mathrm{v}_{22}=b$; $\mathrm{v}_{21}+4{,}5\cdot\mathrm{v}_{22}=0\rightarrow\mathrm{v}_{21}=-4{,}5b$; např. pro $b=1$ pak dostáváme: $\mathbf{v}_2=[-4{,}5\quad 1]$, který je po normalizaci roven $\mathbf{v}_2=\left[\frac{-4{,}5}{\sqrt{(-4{,}5)^2+1^2}}\quad\frac{1}{\sqrt{(-4{,}5)^2+1^2}}\right]=[-0{,}9762\quad 0{,}2169]$. Kontrola, že vektor má jednotkovou délku: $(-0{,}9762)^2+(0{,}2169)^2=1$.

Vlastní vektory můžeme uspořádat do matice $\mathbf{V}=[\mathbf{v}_1^T\quad\mathbf{v}_2^T]=\begin{bmatrix}0{,}2169 & -0{,}9762\\ 0{,}9762 & 0{,}2169\end{bmatrix}$, přičemž pořadí vlastních vektorů odpovídá pořadí vlastních čísel seřazených od největšího k nejmenšímu.

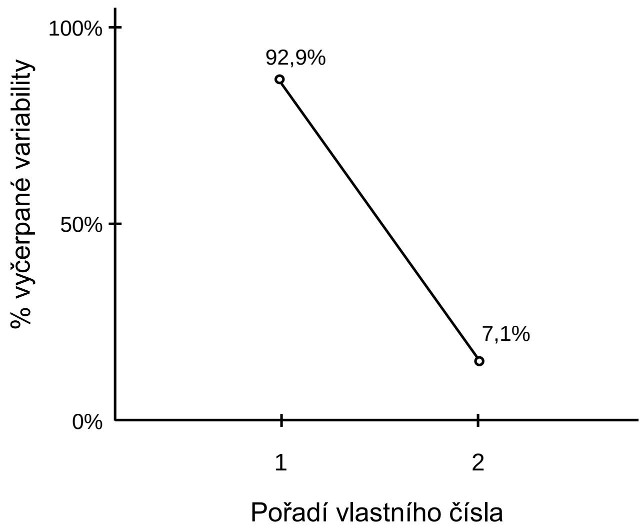
Nyní vyjádříme hlavní komponenty odpovídající vlastním číslům seřazeným od největšího k nejmenšímu – hlavní komponenty jsou lineární kombinace původních proměnných, přičemž koeficienty jsou souřadnice příslušného vlastního vektoru:

1. hlavní komponenta: $\mathbf{y}_1=0{,}2169\cdot\mathbf{x}_1+0{,}9762\cdot\mathbf{x}_2$ (pro $\lambda_1=184$)
2. hlavní komponenta: $\mathbf{y}_2=-0{,}9762\cdot\mathbf{x}_1+0{,}2169\cdot\mathbf{x}_2$ (pro $\lambda_2=14$)

Výpočet procent vyčerpané variability:

1. hlavní komponenta vyčerpává: $\frac{\lambda_1}{\lambda_1+\lambda_2}=\frac{184}{14+184}=0{,}9293$ (tzn., 92,93% variability v datech)
2. hlavní komponenta vyčerpává: $\frac{\lambda_2}{\lambda_1+\lambda_2}=\frac{14}{14+184}=0{,}0707$ (tzn., 7,07% variability v datech)

Vyčerpanou variabilitu můžeme znázornit i pomocí sutinového grafu:



Dále spočítáme korelace hlavních komponent s původními proměnnými:

$$R(\mathbf{x}_1, \mathbf{y}_1) = \frac{\mathrm{v}_{11} \cdot \sqrt{\lambda_1}}{\sqrt{s_{11}}} = \frac{0{,}2169 \cdot \sqrt{184}}{\sqrt{22}} = 0{,}6274$$

$$R(\mathbf{x}_2, \mathbf{y}_1) = \frac{\mathrm{v}_{12} \cdot \sqrt{\lambda_1}}{\sqrt{s_{22}}} = \frac{0{,}9762 \cdot \sqrt{184}}{\sqrt{176}} = 0{,}9981$$

$$R(\mathbf{x}_1, \mathbf{y}_2) = \frac{\mathrm{v}_{21} \cdot \sqrt{\lambda_2}}{\sqrt{s_{11}}} = \frac{-0{,}9762 \cdot \sqrt{14}}{\sqrt{22}} = -0{,}7787$$

$$R(\mathbf{x}_2, \mathbf{y}_2) = \frac{\mathrm{v}_{22} \cdot \sqrt{\lambda_2}}{\sqrt{s_{22}}} = \frac{0{,}2169 \cdot \sqrt{14}}{\sqrt{176}} = 0{,}0612$$

První hlavní je vysoce korelována s váhou a středně korelována s výškou. Druhá hlavní komponenta je středně záporně korelována s výškou.

Na závěr vypočítáme nové souřadnice původních bodů po transformaci pomocí obou hlavních komponent spočítaných pomocí PCA: $\mathbf{Y} = \mathbf{X} \cdot \mathbf{V} = \begin{bmatrix} 101 & 16 \\ 105 & 18 \\ 103 & 42 \\ 98 & 23 \\ 93 & 6 \end{bmatrix} \begin{bmatrix} 0{,}2169 & -0{,}9762 \\ 0{,}9762 & 0{,}2169 \end{bmatrix} =$

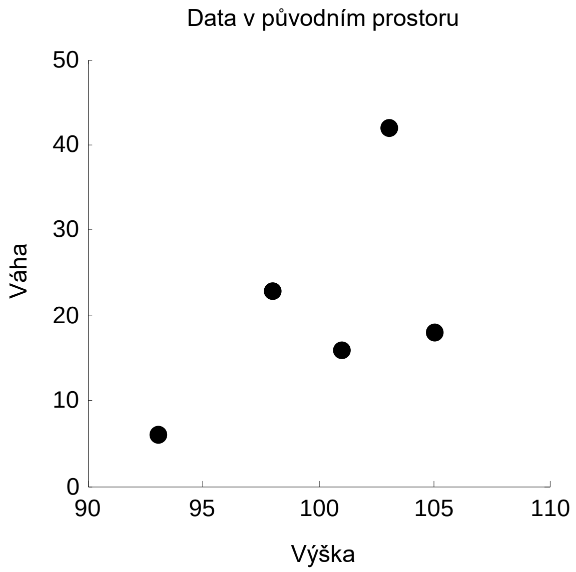
$$\begin{bmatrix} 101 \cdot 0{,}2169 + 16 \cdot 0{,}9762 & 101 \cdot (-0{,}9762) + 16 \cdot 0{,}2169 \\ 105 \cdot 0{,}2169 + 18 \cdot 0{,}9762 & 105 \cdot (-0{,}9762) + 18 \cdot 0{,}2169 \\ 103 \cdot 0{,}2169 + 42 \cdot 0{,}9762 & 103 \cdot (-0{,}9762) + 42 \cdot 0{,}2169 \\ 98 \cdot 0{,}2169 + 23 \cdot 0{,}9762 & 98 \cdot (-0{,}9762) + 23 \cdot 0{,}2169 \\ 93 \cdot 0{,}2169 + 6 \cdot 0{,}9762 & 93 \cdot (-0{,}9762) + 6 \cdot 0{,}2169 \end{bmatrix} = \begin{bmatrix} 37{,}5 & -95{,}1 \\ 40{,}3 & -98{,}6 \\ 63{,}3 & -91{,}4 \\ 43{,}7 & -90{,}8 \\ 26{,}0 & -89{,}5 \end{bmatrix}$$

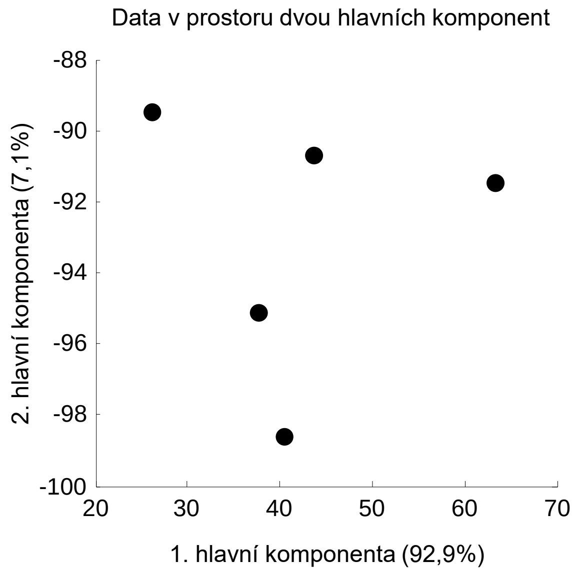
Souřadnice subjektů můžeme přímo získat i z hlavních komponent – např. pro první subjekt:

$$y_{11} = 0{,}2169 \cdot x_{11} + 0{,}9762 \cdot x_{12} = 0{,}2169 \cdot 101 + 0{,}9762 \cdot 16 = 37{,}5$$

$$y_{12} = -0{,}9762 \cdot x_{11} + 0{,}2169 \cdot x_{12} = -0{,}9762 \cdot 101 + 0{,}2169 \cdot 16 = -95{,}1$$

Původní data i data po transformaci pomocí PCA si znázorníme:





Pokud bychom k transformaci použili pouze první vlastní vektor, získáváme data v prostoru první hlavní komponenty: