Bi8600: Vícerozměrné metody
4. cvičení

Analýza hlavních komponent (PCA)

Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, L. Brožová

Analýza hlavních komponent – jaký je cíl?

Analýza hlavních komponent – jaký je cíl?

- V převážné většině případů existují mezi dimenzemi **korelační vztahy**, tedy dimenze se **navzájem vysvětlují** a pro popis kompletní informace v datech **není třeba všech dimenzí vstupního souboru.**

1. Popis a vizualizace vztahů mezi proměnnými
2. Výběr neredundantních proměnných pro další analýzy
3. Vytvoření zástupných faktorových os
4. Identifikace shluků v datech spjatých s variabilitou dat
5. Identifikace vícerozměrně odlehlých objektů

Analýza hlavních komponent – vstup?

Analýza hlavních komponent – vstup?

- Pracuje s asociační maticí korelací/kovariancí.
- Kdy použijeme kterou matici?
- Jaká bude dimenze matic?

Jaký je vztah mezi kovariancí a korelací?

- **Kovariance** popisuje vztah dvou proměnných; její rozsah závisí na variabilitě dat

$$C(x_1, x_2) = \frac{\sum_{i=1}^{n}(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)}{n-1}; C \in (-\infty; \infty)$$

- **Korelace** = kovariance standardizovaná na rozptyl proměnných.

$$r(x_1, x_2) = \frac{C(x_1, x_2)}{\sqrt{D(x_1)}\sqrt{D(x_2)}}; r \in \langle -1; 1 \rangle$$

- Jaké hodnoty se nachází na diagonále korelační matice?
- Má smysl použít metody redukce dimenzionality dat v situaci, kdy jsou hodnoty kovariance/korelace blízké nule?
- Čemu odpovídá kovariance na standardizovaných datech?

→ Pokud $D(x_1)=D(x_2)=1$ → kovariance = korelace

Analýza hlavních komponent – předpoklady?

Analýza hlavních komponent – předpoklady?

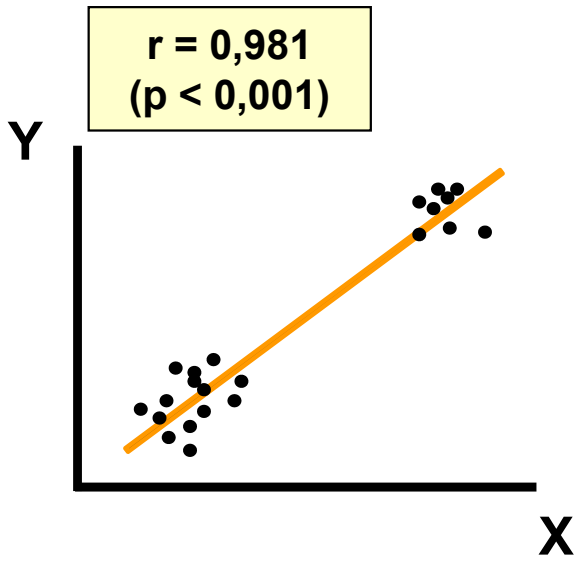
- Více objektů než proměnných (obvykle se uvádí 10x větší počet objektů než proměnných)
- Vícerozměrná technika – 100% vyplněnost dat (jedna chybějící hodnota vede k odstranění celého objektu z analýzy)
- Souvisí s výpočtem asociační matice – korelace/kovariance vyžadují zhruba normální rozdělení proměnných.

ALE! Jaké mohou být výjimky?

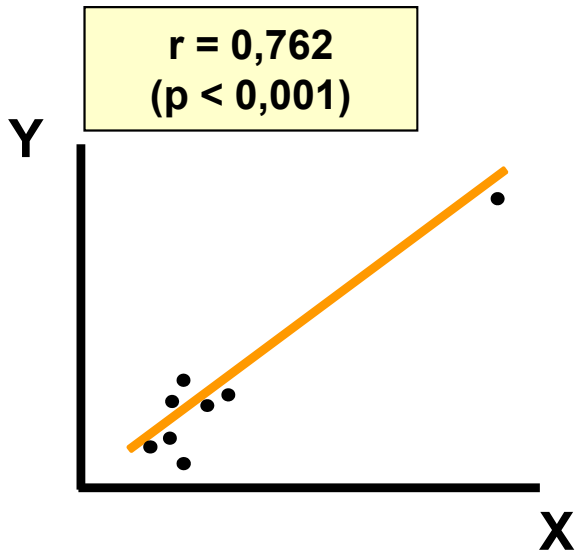
Problémy s výpočtem korelačního koeficientu

- Výjimkou jsou situace, kdy provádíme analýzu za účelem identifikace shluků / odlehlých hodnot.

Identifikace shluků

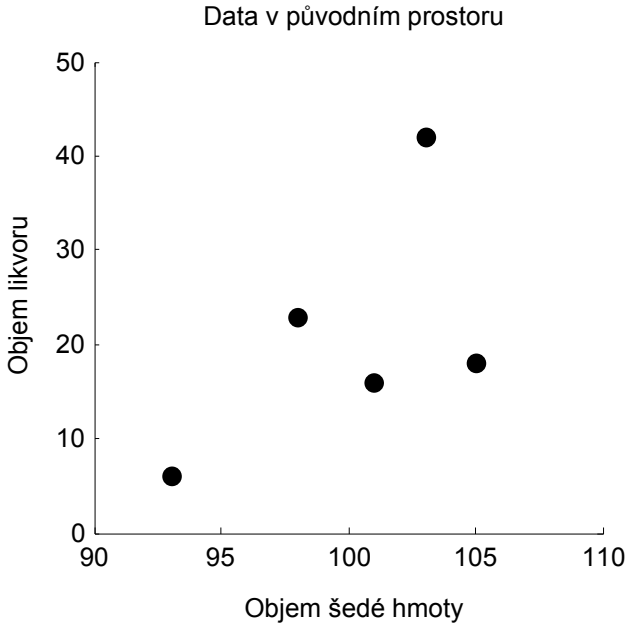


Identifikace odlehlých hodnot



Popis výstupů - příklad

- Bylo provedeno měření objemu šedé hmoty $\mathbf{x}_1$ (v cm³) a objemu likvoru $\mathbf{x}_2$ (v cm³) u pěti dětí. Naměřené hodnoty byly zaznamenány do matice $\mathbf{X}$:



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

- Jelikož jsou vstupní data měřena ve stejných jednotkách, analýza bude provedena na kovarianční matici, vstupní data jsou centrována průměrem →

$$\mathbf{U} = \mathbf{X} - \overline{\boldsymbol{x}} = \begin{bmatrix} 1 & -5 \\ 5 & -3 \\ 3 & 21 \\ -2 & 2 \\ -7 & -15 \end{bmatrix}.$$

Popis výstupů - příklad

- Jelikož jsou proměnné hodnoceny ve stejných jednotkách, analýza je provedena na kovarianční matici C:

$$\mathbf{C} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} 22 & 36 \\ 36 & 176 \end{bmatrix}$$

$$(\mathbf{C} - \lambda \mathrm{I})\ \mathrm{u} = 0 \Rightarrow \begin{bmatrix} 22 - \lambda & 36 \\ 36 & 176 - \lambda \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- Spočítáme-li determinant matice $(\mathbf{C} - \lambda \mathrm{I})$, dostáváme **vlastní čísla** $\boldsymbol{\lambda_1}$ a $\boldsymbol{\lambda_2}$:

$\lambda_1 = 184$ → % rozptylu, které popisuje osa: 184/(184+14) *100 = 92.9 %

$\lambda_2 = 14$ → % rozptylu, které popisuje osa: 14/(184+14) *100 = 7.1 %

184+14=22+176 → PCA přerozděluje rozptyl původních dat

- Po dosazení vlastních čísel spočítáme vlastní vektory:

$$\mathbf{U} = [\mathbf{u}_1^T \quad \mathbf{u}_2^T] = \begin{bmatrix} 0{,}2169 & 0{,}9762 \\ 0{,}9762 & -0{,}2169 \end{bmatrix}$$

→ vlastní vektor asociovaný s λ_1

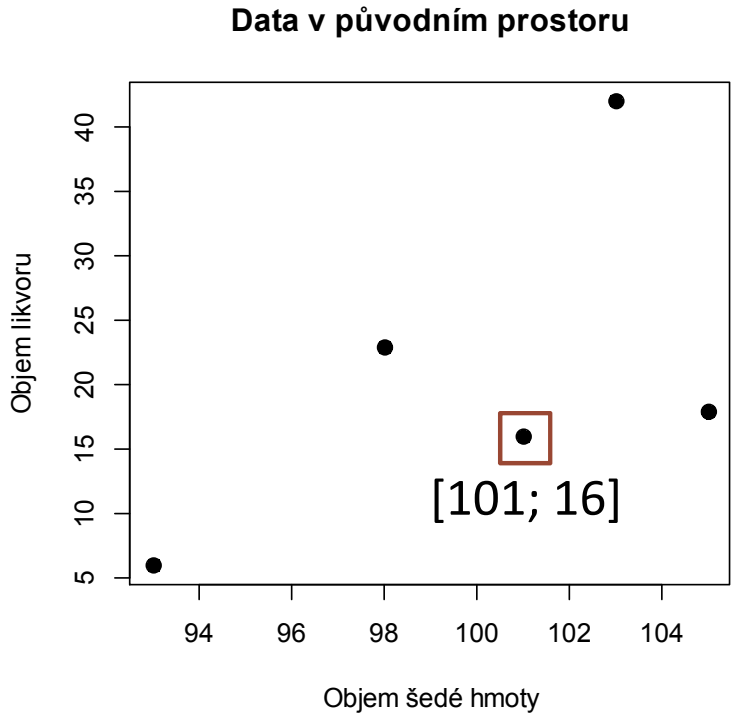
Popis výstupů - příklad

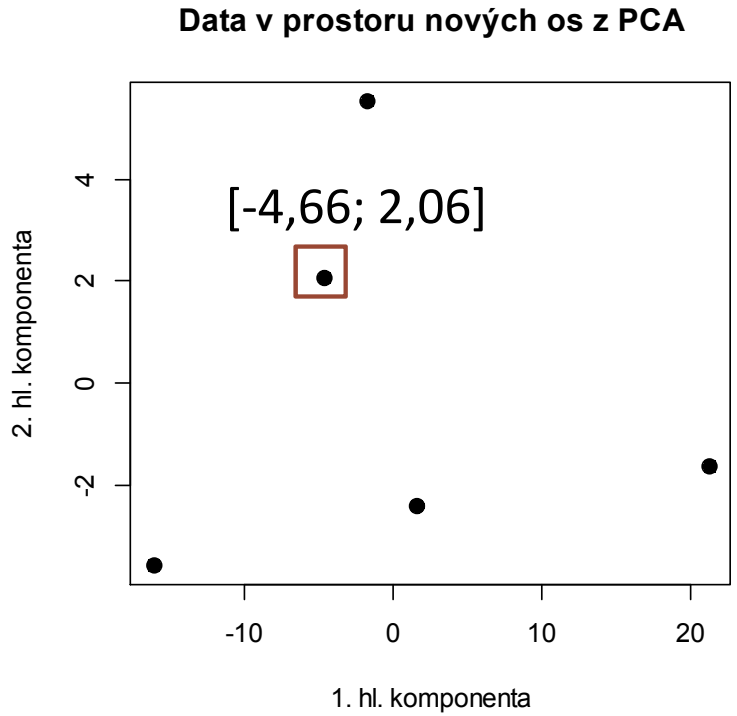
$$\mathbf{U} = [\mathbf{u}_1^T \quad \mathbf{u}_2^T] = \begin{bmatrix} 0{,}2169 & 0{,}9762 \\ 0{,}9762 & -0{,}2169 \end{bmatrix}$$

- Nové osy ($\mathbf{y_1}$, $\mathbf{y_2}$) jsou lineární kombinací původních proměnných:

$$\mathbf{y}_1 = 0{,}2169 \cdot (\mathrm{x}_1 - \overline{x_1}) + 0{,}9762 \cdot (x_2 - \overline{x_2}) = 0{,}2169 \cdot 1 + 0{,}9762 \cdot (-5) = -4{,}66$$

$$\mathbf{y}_2 = 0{,}9762 \cdot (\mathrm{x}_1 - \overline{x_1}) + (-0{,}2169) \cdot (\mathrm{x}_2 - \overline{x_2}) = 0{,}9762 \cdot 1 + (-0{,}2169) \cdot (-5) = 2{,}06$$

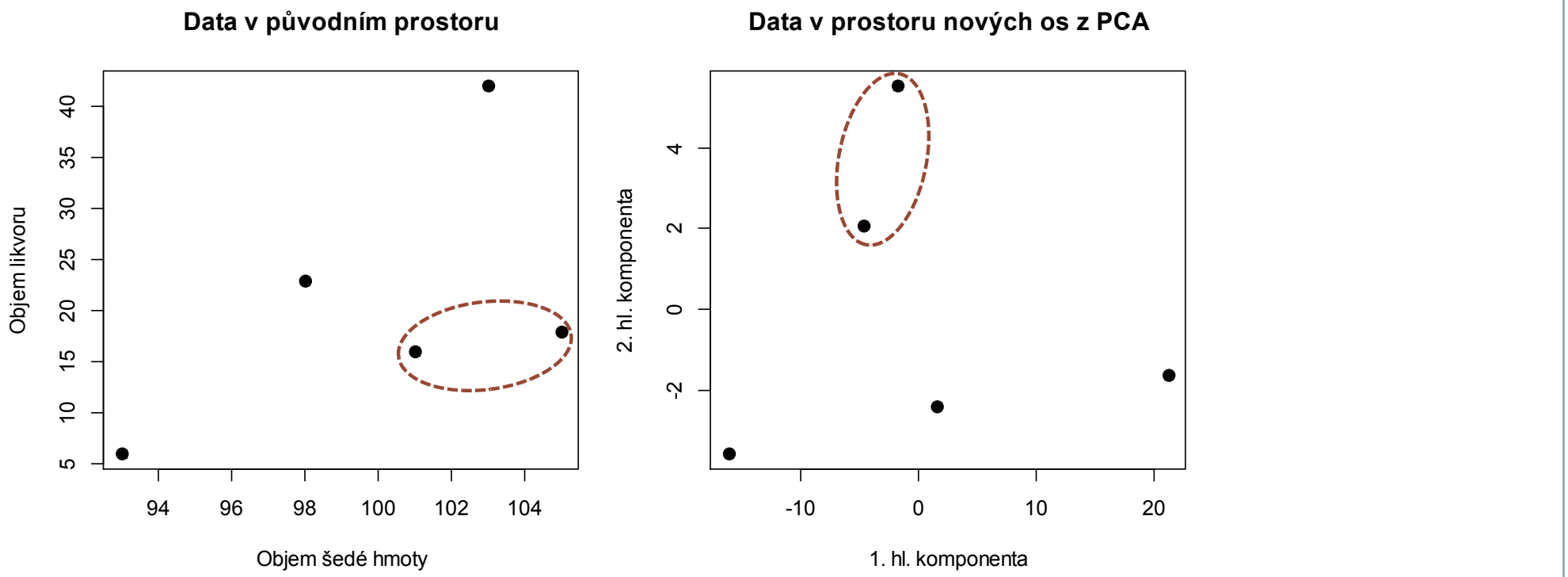




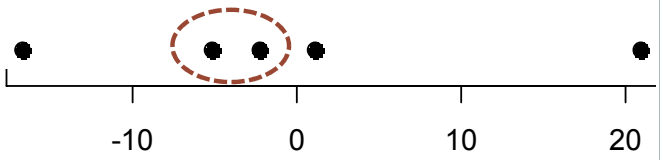
- PCA natočí datový prostor a vytvoří nové osy tak, aby popisovaly maximum variability původních dat.

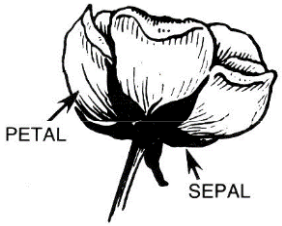
Popis výstupů - příklad

- Každá další osa popisuje rozptyl, který nebyl popsán osami předchozími – každá další osa je nezávislá = kolmá na osy předchozí.



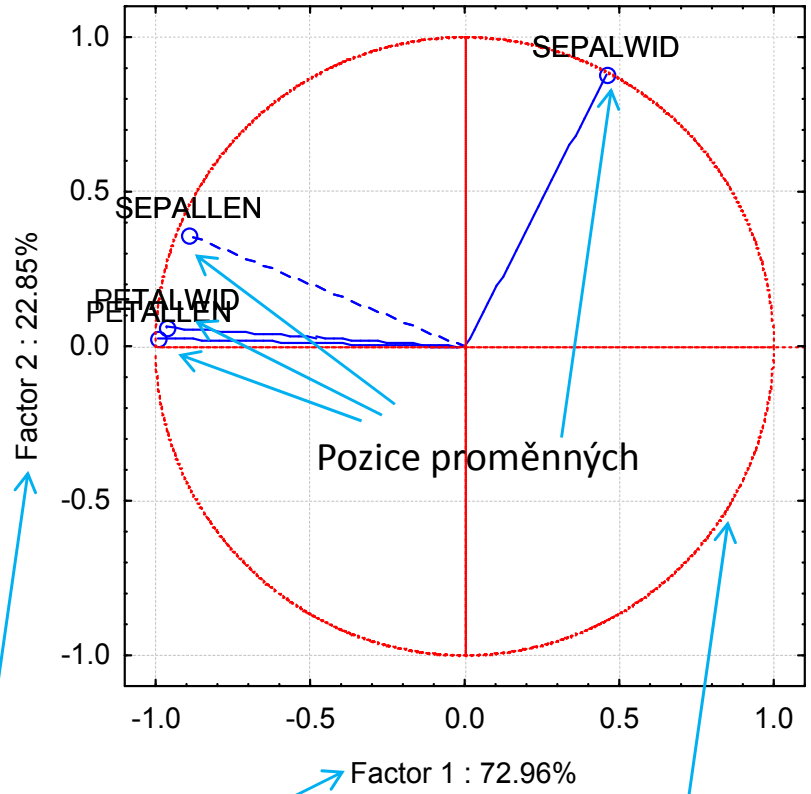
- Výběrem faktorových os přicházíme o určité % variability původních dat





Grafické výstupy

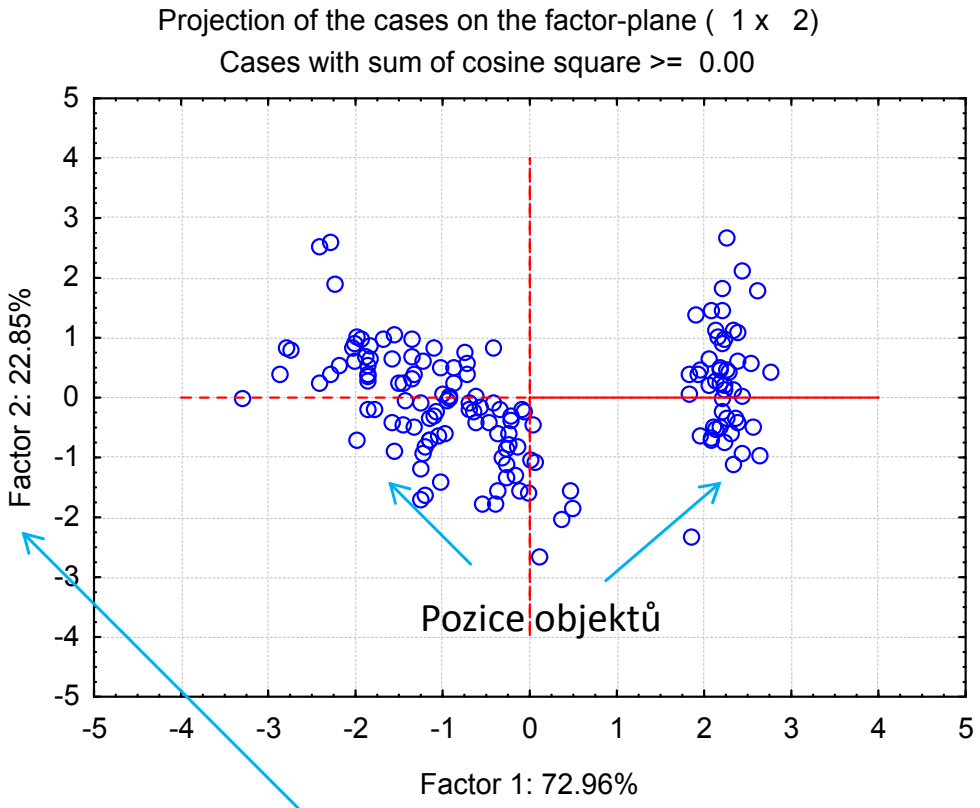
Biplot korelací



Variabilita vyčerpaná faktorovými osami

Jednotková kružnice - Hranice příspěvku k definici faktorové osy

Biplot vzdáleností



Variabilita vyčerpaná faktorovými osami

Jaký počet os popisuje dostatečně datový soubor?

Jaký počet os popisuje dostatečně datový soubor?

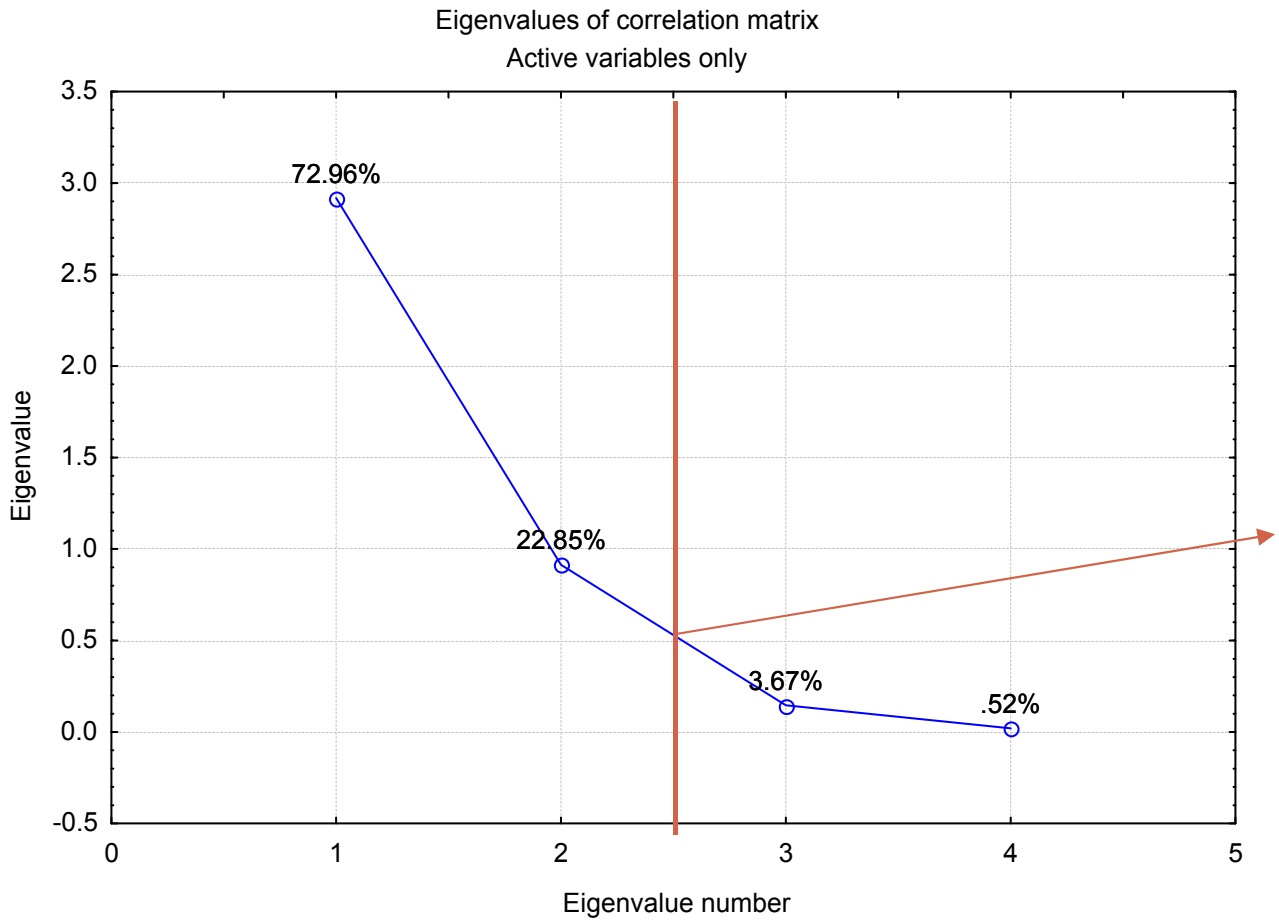
- Ideálně 2-3 osy, je však potřeba brát ohled na % rozptylu původních dat, který vybranými osami popíšeme.

- **Kaiser-Gutmanovo kritérium**

  - ✓ Pro další analýzu jsou vybrány osy s vlastním číslem >1 (korelace) nebo větším než je průměrné eigenvalue (kovariance)
  - ✓ Logika je vybírat osy, které přispívají k vysvětlení variability dat více než připadá rovnoměrným rozdělením variability

Jaký počet os popisuje dostatečně datový soubor?

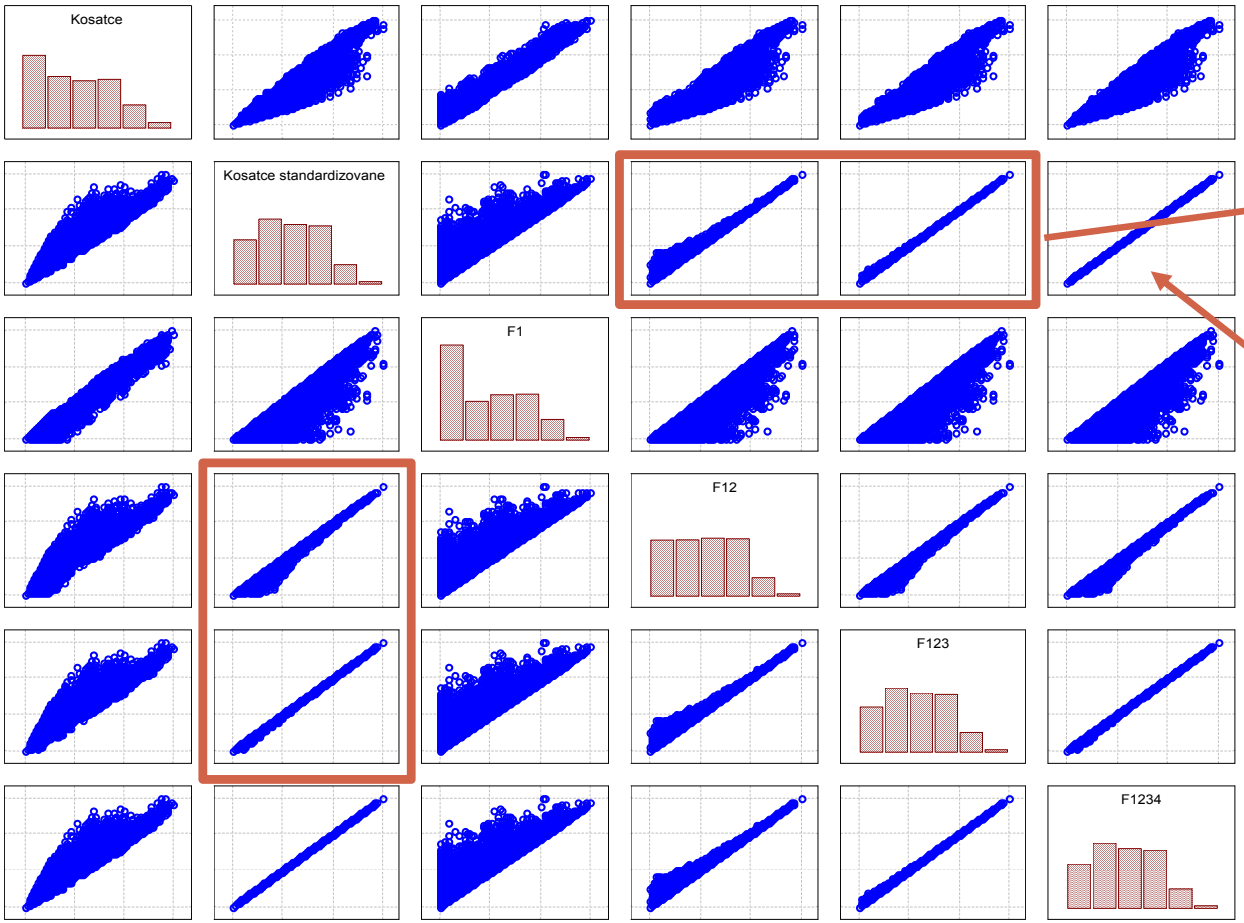
- **Scree plot**
  - ✓ Grafický nástroj hledající zlom ve vztahu počtu os a vyčerpané variability



- Zlom ve vztahu mezi počtem nových os a popsanou variabilitou – pro další analýzu budou použity první dvě faktorové osy.
- Tyto osy popisují téměř 96 % rozptylu původních dat.

Jaký počet os popisuje dostatečně datový soubor?

- **Sheppardův diagram**
  - ✓ Vykresluje vzdálenosti v prostoru původních proměnných proti vzdálenostem na nových osách



Za optimální z hlediska zachování vzdáleností objektů lze považovat dvě nebo tři dimenze.

Při použití všech dimenzí jsou vzdálenosti perfektně zachovány.