Bi8600: Vícerozměrné metody
3. cvičení

Analýza hlavních komponent (PCA)

Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, S. Littnerová, 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/odlehlých objektů

Analýza hlavních komponent – vstup?

Analýza hlavních komponent – vstup?

- Pracuje s asociační maticí korelací/kovariancí.
- Jaký je vztah mezi kovariancí a korelací?
- 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

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

- **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?

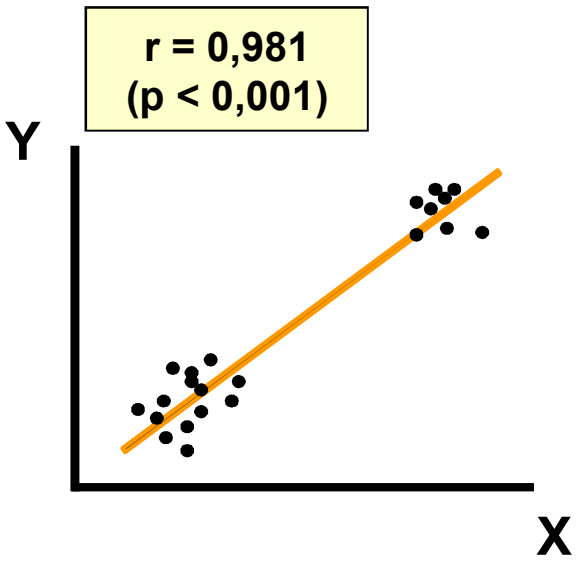
- Více objektů než proměnných (obvykle se uvádí 10x větší počet objektů než proměnných)
- Vícerozměrná technika – 100% vyplněnost dat (jedna chybějící hodnota vede k odstranění celého objektu z analýzy)
- Souvisí s výpočtem asociační matice – korelace/kovariance vyžadují zhruba normální rozdělení proměnných.

ALE! Jaké mohou být výjimky?

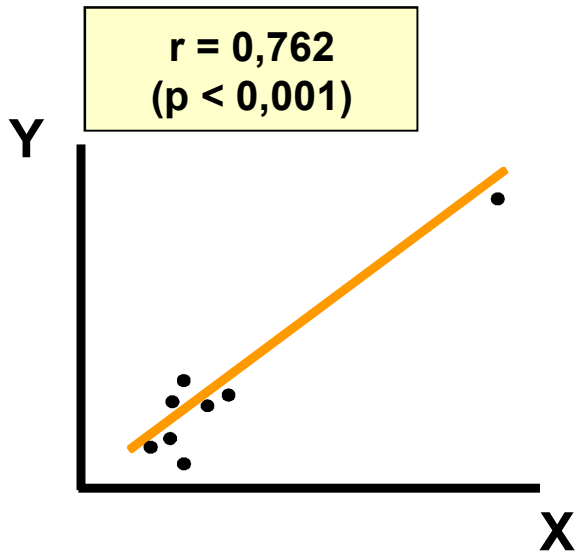
Problémy s výpočtem korelačního koeficientu

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

Identifikace shluků



Identifikace odlehlých hodnot



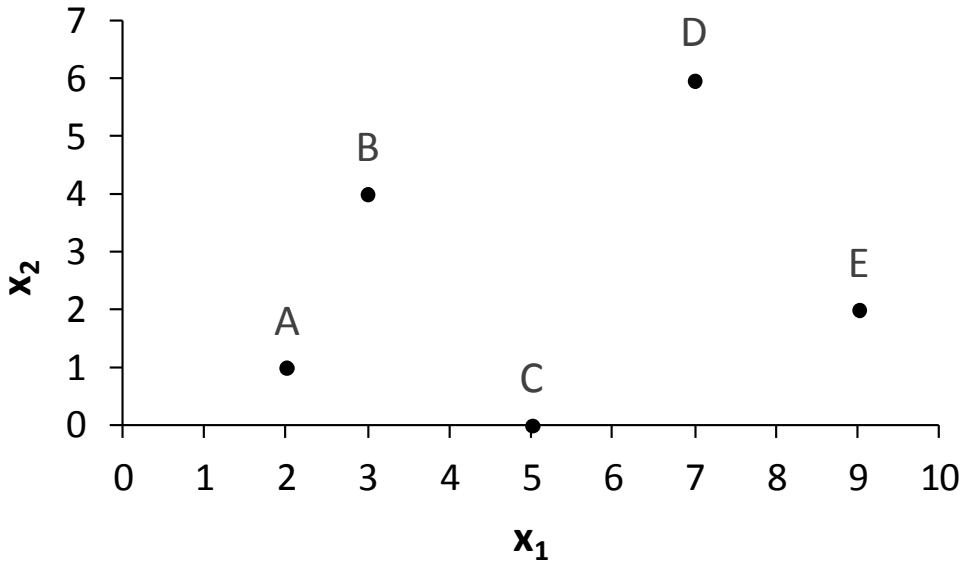
Postup výpočtu PCA – primární data

- Bylo provedeno měření dvou parametrů (,) u pěti objektů (A-D). Naměřené hodnoty byly zaznamenány do matice :

Datový soubor

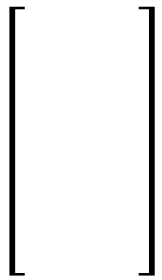
ID	x_1	x_2
A	2	1
B	3	4
C	5	0
D	7	6
E	9	2

Pozice objektů v původním prostoru



Postup výpočtu PCA - standardizace

- Proměnné jsou hodnoceny ve stejných jednotkách – proměnné jsou centrovány.



Postup výpočtu PCA – kovarianční matice.

- Proměnné jsou hodnoceny ve stejných jednotkách, analýza je provedena na kovarianční matici **S**:

$$\begin{bmatrix} \quad \end{bmatrix} \begin{bmatrix} \quad \end{bmatrix}$$

- Spočítáme-li determinant kovarianční matice, dostáváme **vlastní čísla** a a následně **vlastní vektory** $\mathbf{v}_1$ a $\mathbf{v}_2$ příslušné vlastním číslům.

Jaký význam mají vlastní čísla a vlastní vektory?

Postup výpočtu PCA – vlastní čísla, vlastní vektory

- Proměnné jsou hodnoceny ve stejných jednotkách, analýza je provedena na kovarianční matici **S**:

$$\begin{bmatrix} & \end{bmatrix} \begin{bmatrix} & \end{bmatrix}$$

- Z kovarianční matice spočítáme **vlastní čísla** a a následně **vlastní vektory** $\mathbf{v_1}$ a $\mathbf{v_2}$ příslušné vlastním číslům:

5 → % rozptylu, které popisuje první osa: 9/(9 + 5) *100 = 64,3 %
→ % rozptylu, které popisuje druhá osa: 5/(9 + 5) *100 = 35,7 %

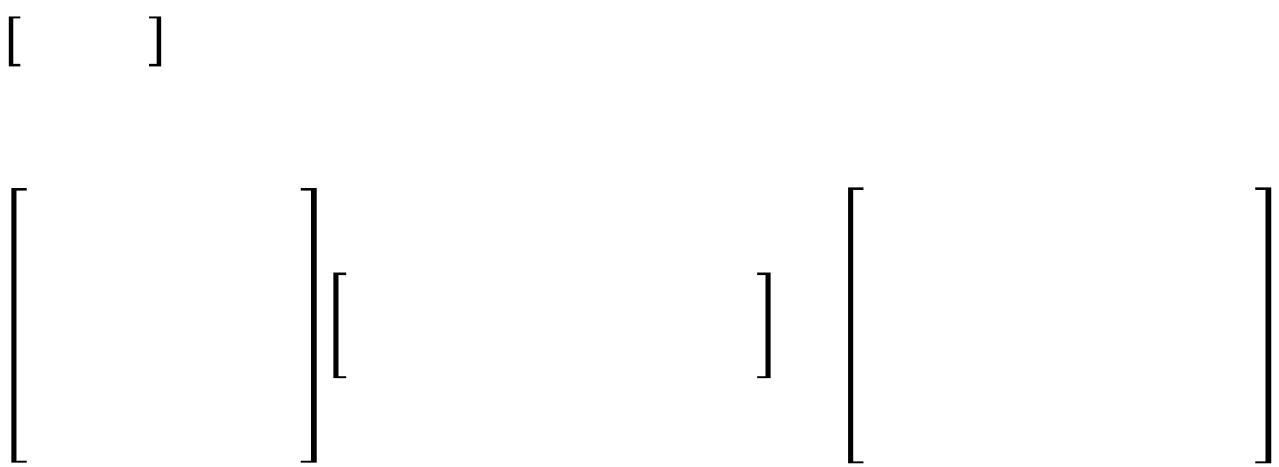
PCA pouze přerozděluje rozptyl původních dat do nových os

- Součet vlastních čísel je roven součtů rozptylů jednotlivých proměnných (pokud je vstupem korelační matice, odpovídá počtu proměnných, protože rozptyl každé proměnné je roven jedné).
- Vlastní vektory popisují směr nových faktorových os:

$$\begin{bmatrix} & \end{bmatrix} \begin{bmatrix} & \end{bmatrix}$$

Postup výpočtu PCA – pozice na nových osách - výpočet

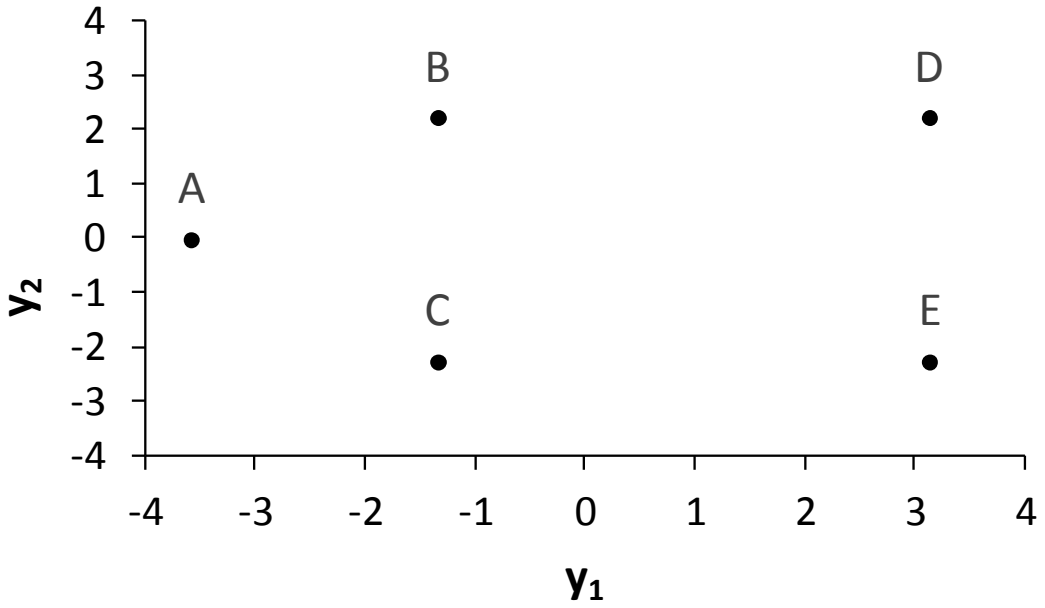
- Nové osy ($\mathbf{y_1}$, $\mathbf{y_2}$) jsou lineární kombinací původních proměnných:



Postup výpočtu PCA – pozice na nových osách - vizualizace

- PCA natočí datový prostor a vytvoří nové osy tak, aby popisovali maximum variability původních dat.
- Každá další osa popisuje rozptyl, který nebyl popsán osami předchozími – každá další osa je nezávislá = kolmá na osy předchozí.

Pozice objektů v novém prostoru



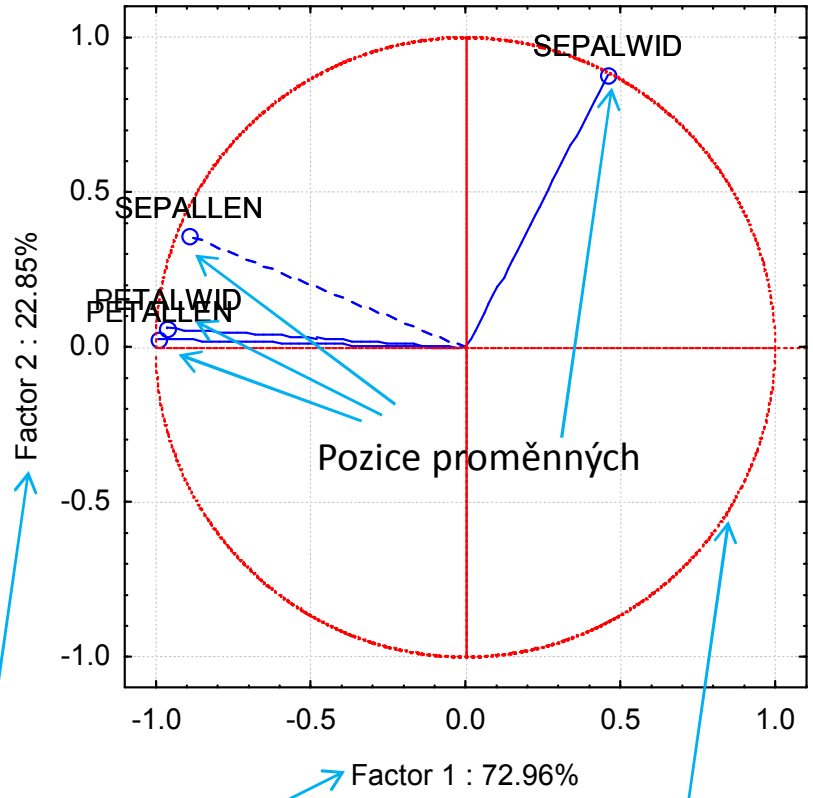
Výběrem faktorových os přicházíme o určité % variability původních dat

Datový soubor

ID	x_1	x_2	y_1
A	2	1	-3.578
B	3	4	-1.342
C	5	0	-1.342
D	7	6	3.130
E	9	2	3.130

Grafické výstupy

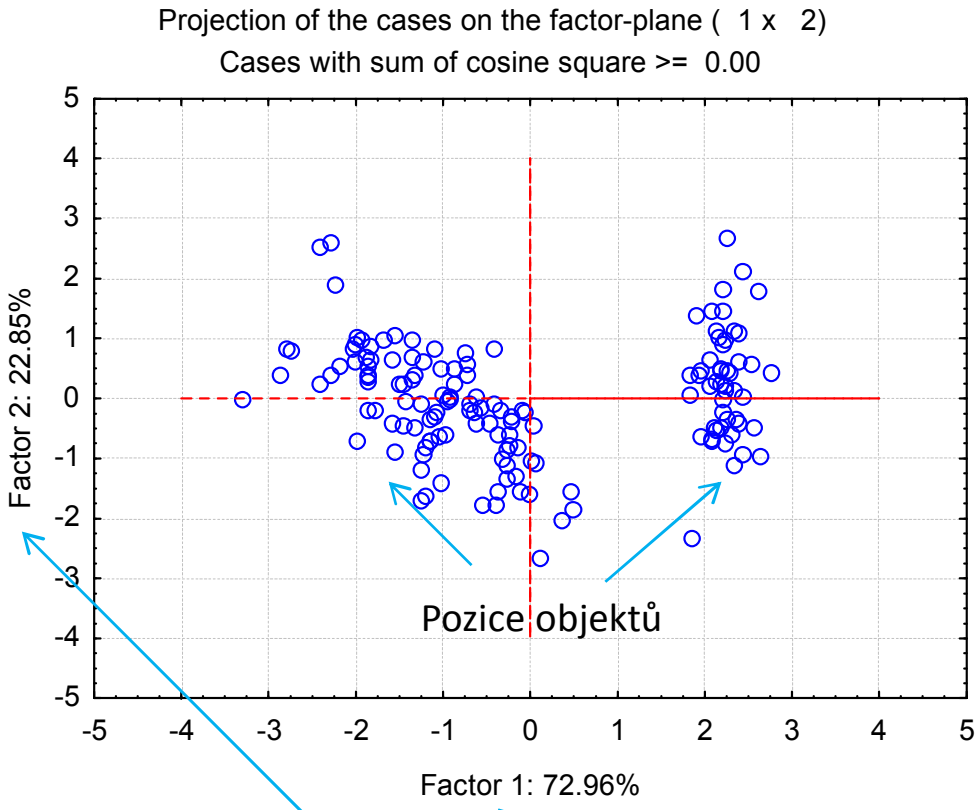
Biplot korelací



Variabilita vyčerpaná faktorovými osami

Jednotková kružnice - Hranice příspěvku k definici faktorové osy

Biplot vzdáleností



Variabilita vyčerpaná faktorovými osami

Otázka: jaký počet os vybrat?

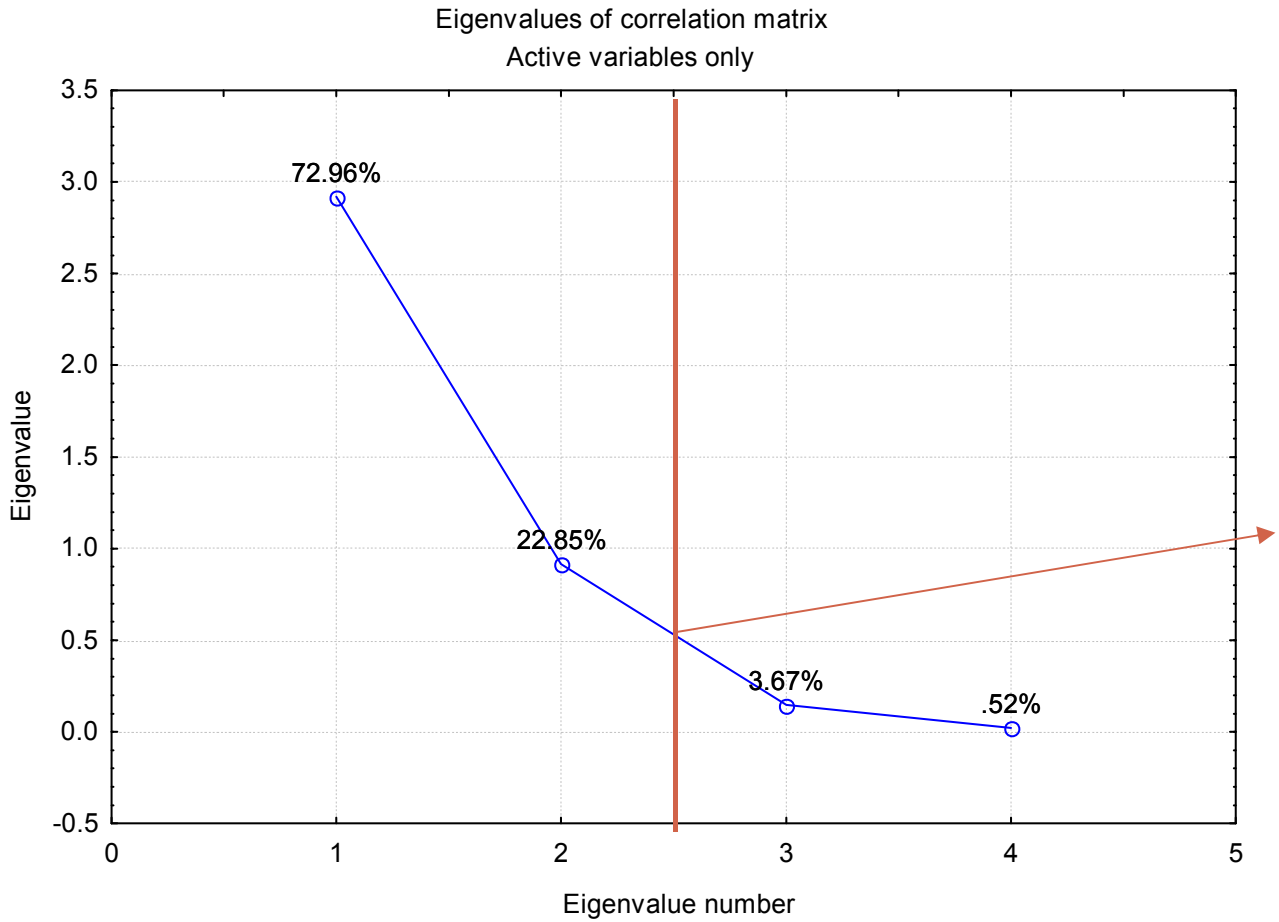
- Pokud je cílem vizualizace dat: ideálně 2-3 osy.
- Pokud chceme data zredukovat do menšího počtu nových proměnných, které budou vstupovat do další analýzy, definujeme počet os hlavně na základě % rozptylu původních dat, který vybranými osami popíšeme.

- **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?

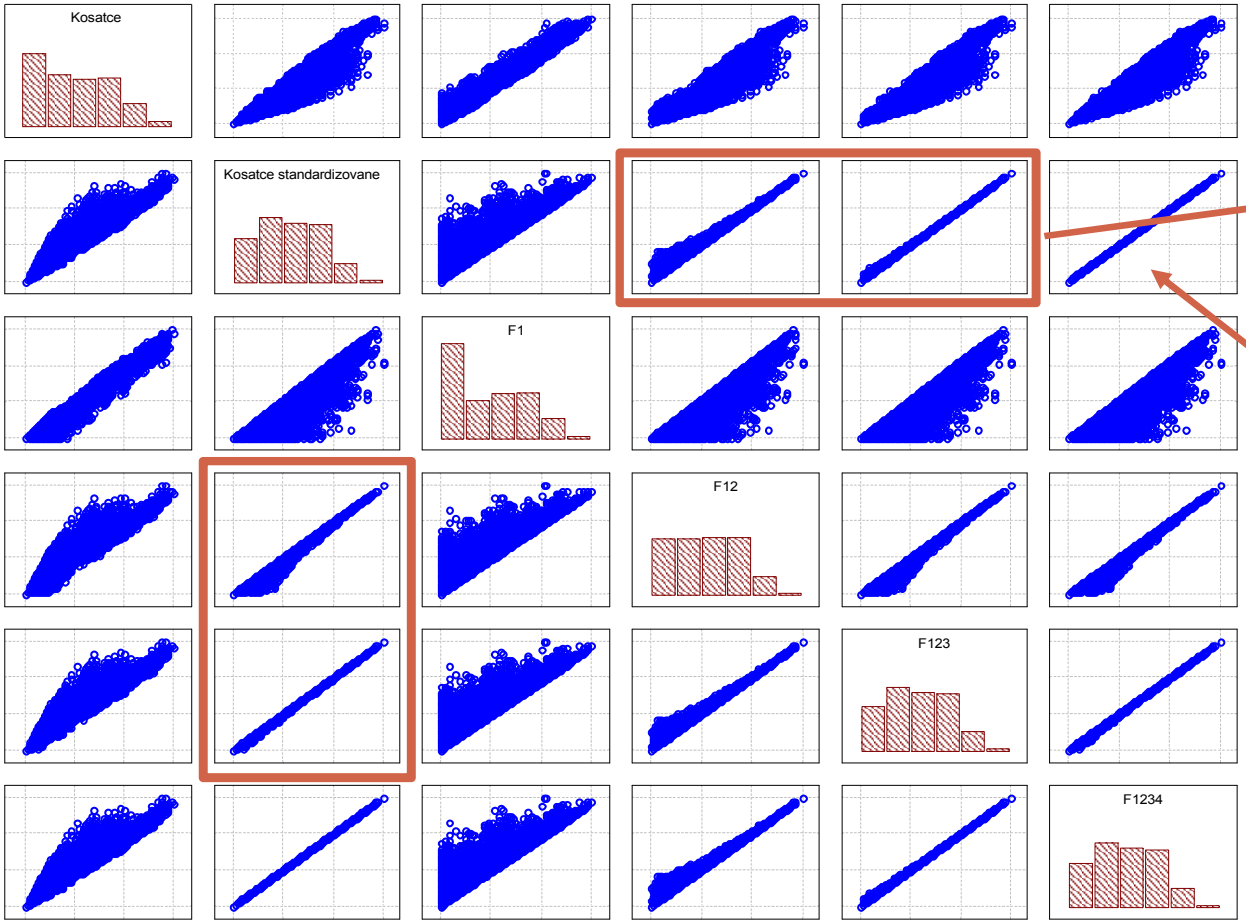
- **Scree plot**
  - ✓ Grafický nástroj hledající zlom ve vztahu počtu os a vyčerpané variability



- 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.