

# Bi8600: Pokročilé statistické metody

## 4. cvičení



### Analýza hlavních komponent (PCA)

# 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/odlehklých objektů

# 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 kovariancemi a korelací?



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

$$C(x_1, x_2) = \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{n-1}; C \in (-\infty; \infty)$$

- **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  $D(x_1)=D(x_2)=1 \rightarrow$  kovariance = korelace

# 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.

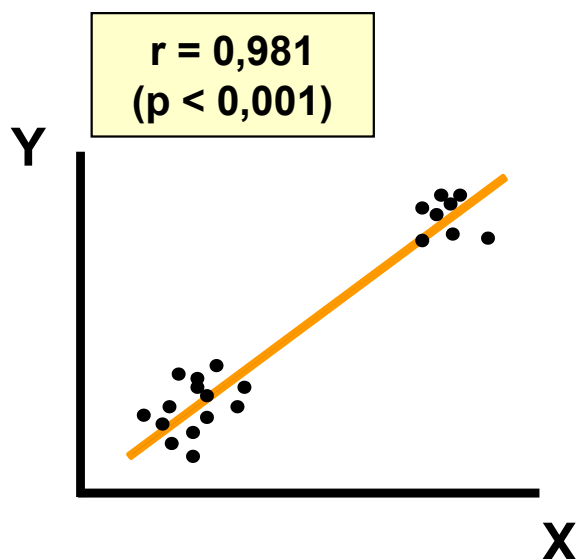
**ALE! Jaké mohou být výjimky?**

# Problémy s výpočtem korelačního koeficientu

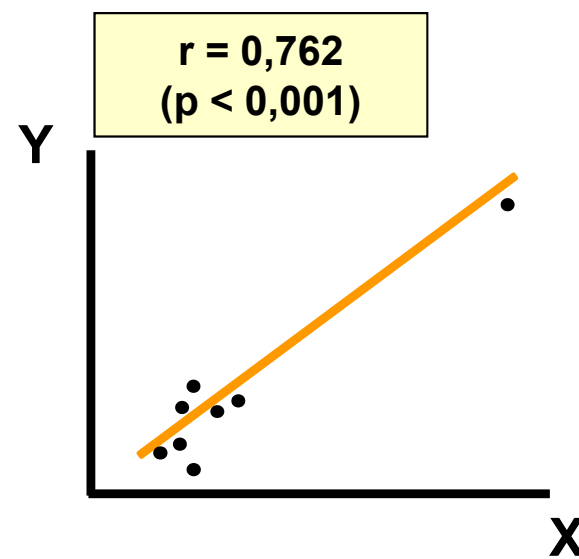


- Výjimkou jsou situace, kdy provádíme analýzu za účelem identifikace shluků / odlehlých hodnot.

## Identifikace shluků



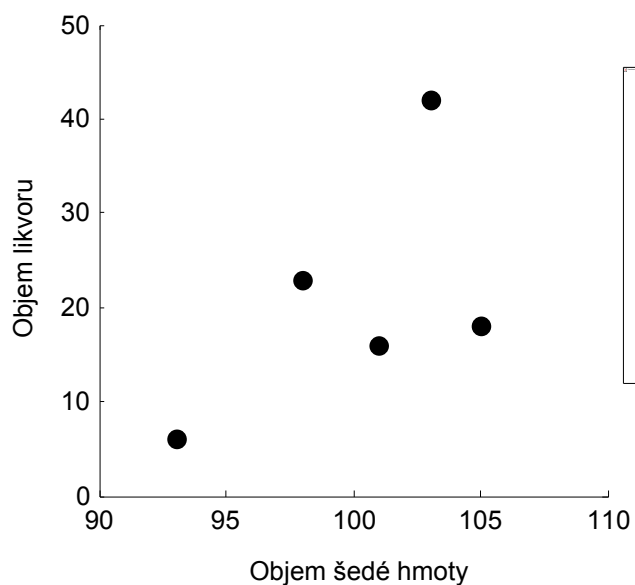
## Identifikace odlehlých hodnot



# Popis výstupů - příklad

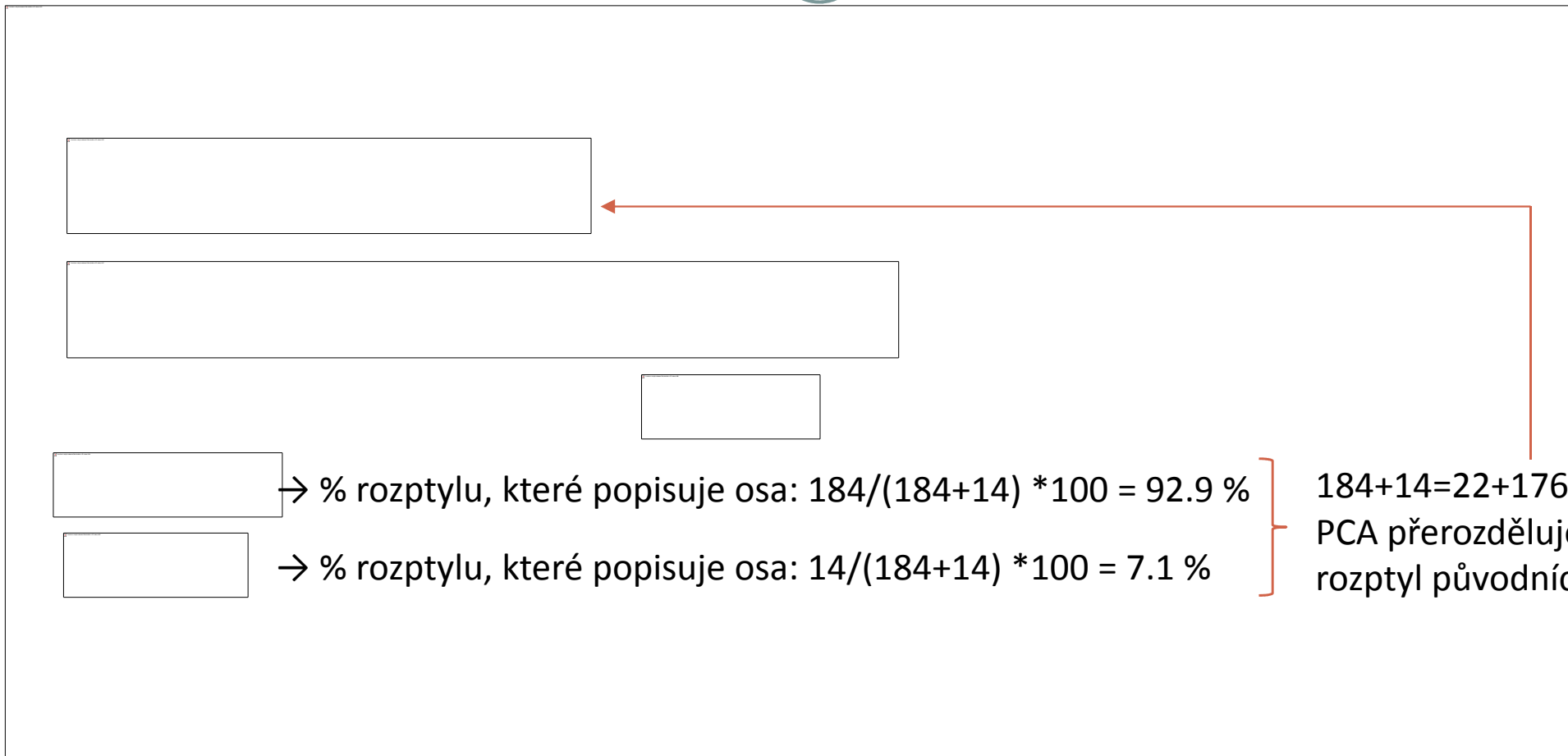


Data v původním prostoru



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

# Popis výstupů - příklad



[Empty box]

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

[Empty box]

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

}  $184+14=22+176 \rightarrow$   
PCA přerozděluje  
rozptyl původních dat

[Empty box] [Dashed red box]

[Empty box]

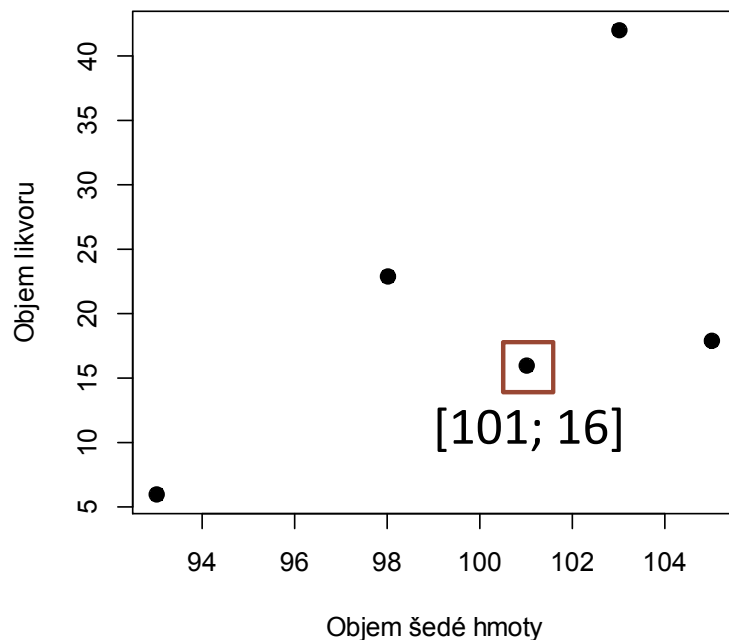


# Popis výstupů - příklad

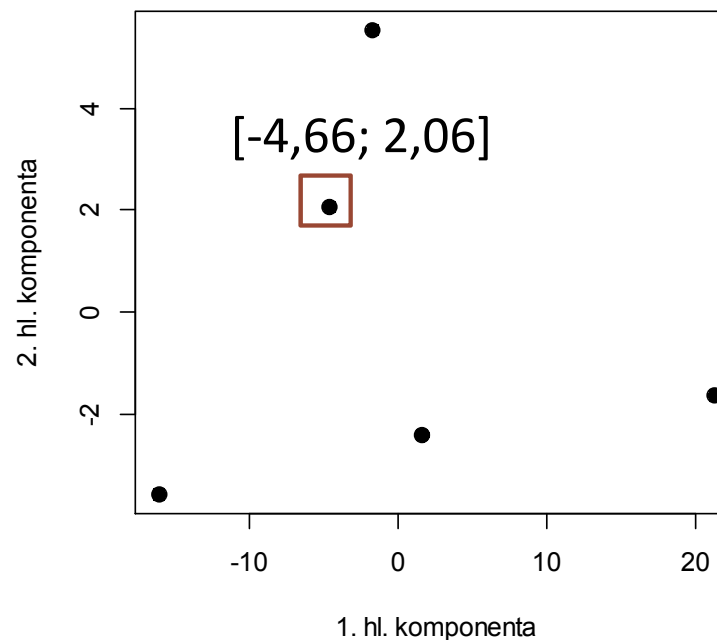


- Nové osy ( $y_1, y_2$ ) jsou lineární kombinací původních proměnných:

Data v původním prostoru



Data v prostoru nových os z PCA



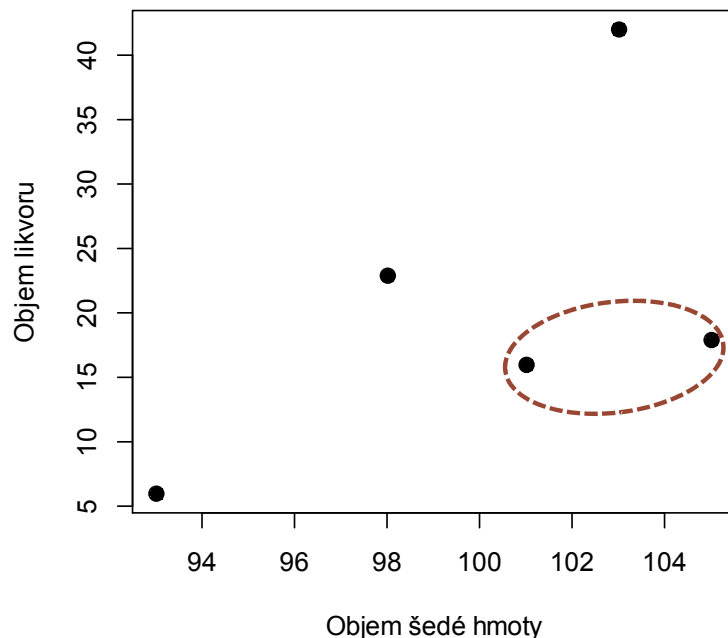
- PCA natočí datový prostor a vytvoří nové osy tak, aby popisovaly maximum variability původních dat.

# Popis výstupů - příklad

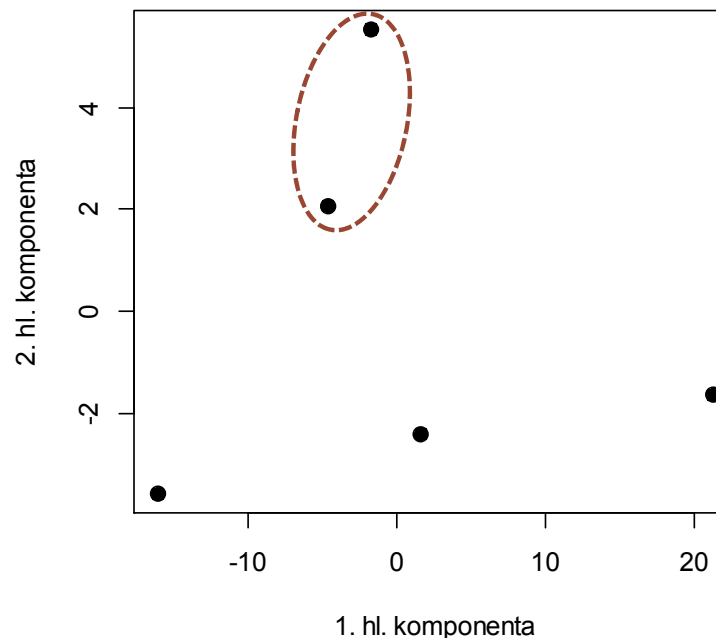


- Každá další osa popisuje rozptyl, který nebyl popsán osami předchozími – každá další osa je nezávislá = kolmá na osy předchozí.

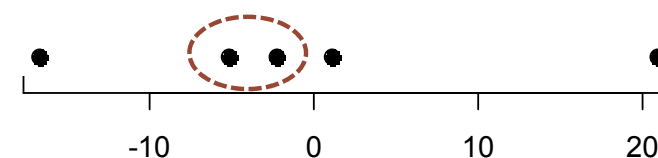
Data v původním prostoru

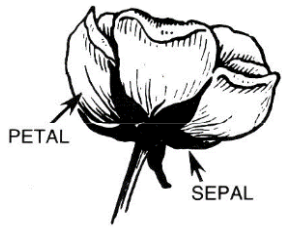


Data v prostoru nových os z PCA



- Výběrem faktorových os přicházíme o určité % variability původních dat

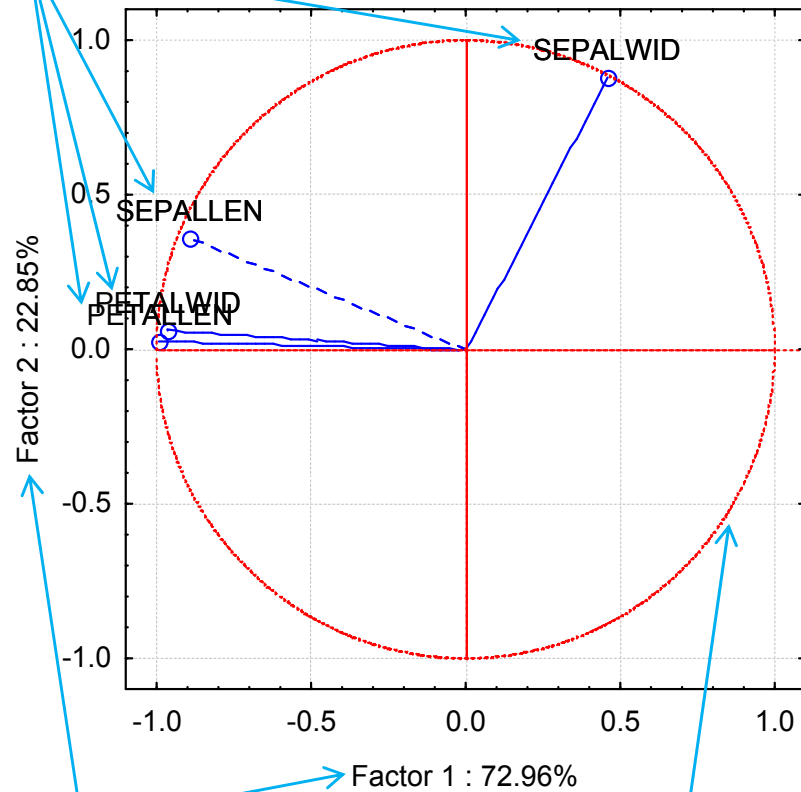




# Grafické výstupy

Pozice proměnných

## Biplot korelací

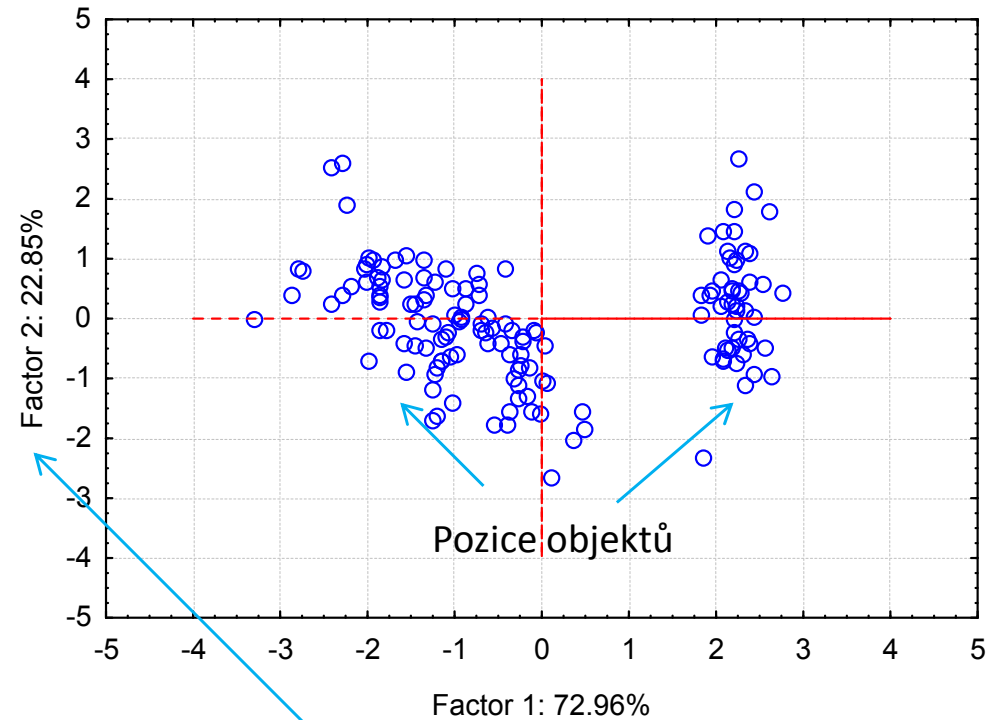


Variabilita vyčerpaná faktorovými osami

Jednotková kružnice -  
Hranice příspěvku k  
definici faktorové osy

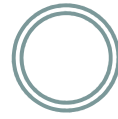
## Biplot vzdáleností

Projection of the cases on the factor-plane ( 1 x 2)  
Cases with sum of cosine square  $\geq 0.00$



Variabilita vyčerpaná faktorovými osami

# Jaký počet os popisuje dostatečně datový soubor?



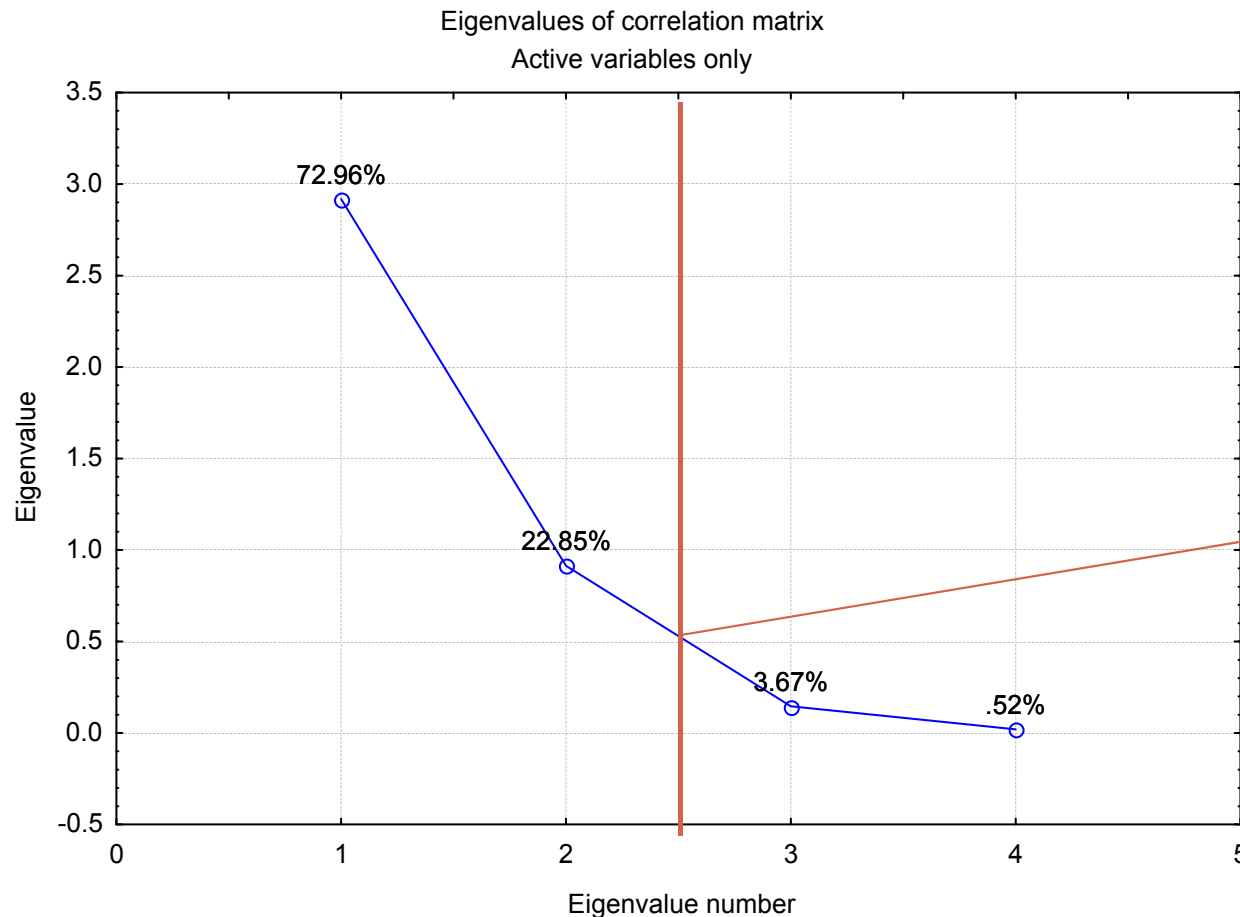
- Ideálně 2-3 osy, je však potřeba brát ohled na % rozptylu původních dat, který vybranými osami popíšeme.
- **Kaiser-Gutmanovo kritérium**
  - ✓ Pro další analýzu jsou vybrány osy s vlastním číslem  $>1$  (korelace) nebo větším než je průměrné eigenvalue (kovariance)
  - ✓ Logika je vybírat osy, které 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



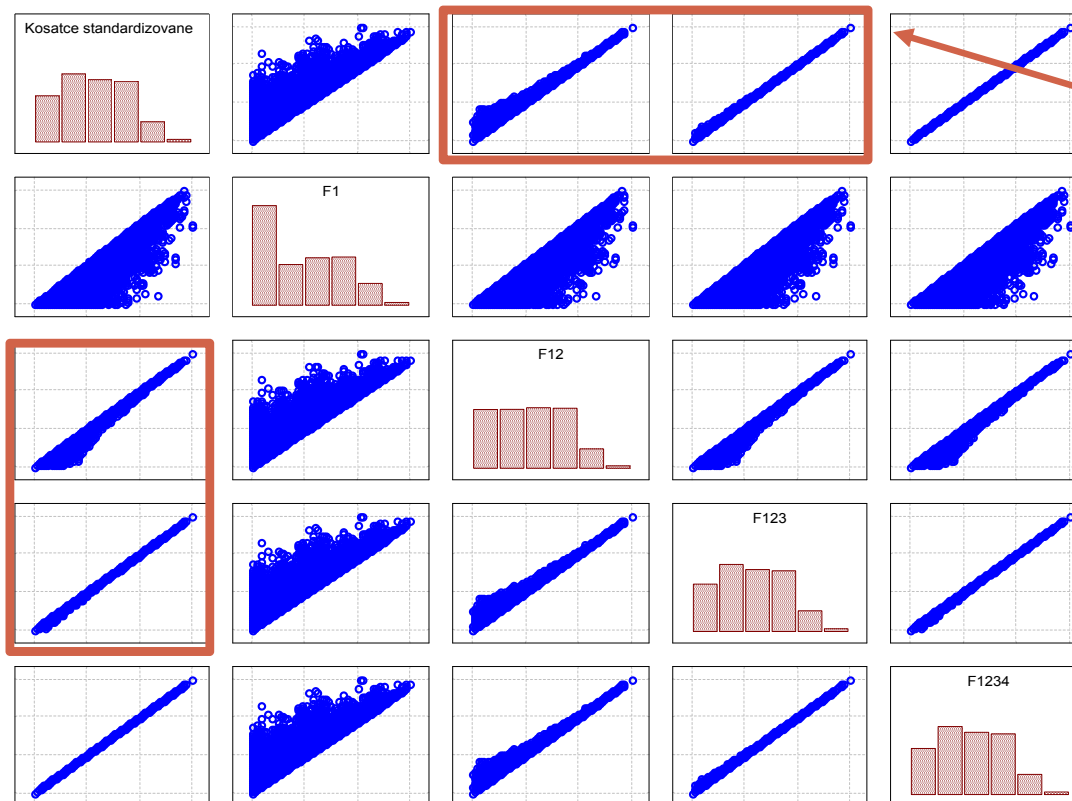
- Zlom ve vztahu mezi počtem nových os a popsanou variabilitou – pro další analýzu budou použity první dvě faktorové osy.
- Tyto osy popisují téměř 96 % rozptylu původních dat.

# Jaký počet os popisuje dostatečně datový soubor?



- **Shepardův 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.