

Pravděpodobnostní metody a matematická statistika

Prof. RNDr. Gejza Dohnal, CSc.

XV. Závěrem



<https://sms.nipax.cz/pas>

Bootstrap

Metoda „syntetického“ výběru

- chceme odhadnout parametry rozdělení nějaké statistiky, které neznáme
- v případech, kdy sice rozdělení statistiky známe, ale analytické odvození hledaného rozdělení uvažované statistiky je příliš náročné nebo zdlouhavé
- v některých případech tato metoda umožňuje dosažení lepších výsledků než klasické analytické metody tím, že se vyhne běžně používaným aproximacím (například při konstrukci intervalových odhadů)

Bootstrap

Metoda „syntetického“ výběru

- chceme odhadnout parametry rozdělení nějaké statistiky, které neznáme
- v případech, kdy sice rozdělení statistiky známe, ale analytické odvození hledaného rozdělení uvažované statistiky je příliš náročné nebo zdlouhavé
- v některých případech tato metoda umožňuje dosažení lepších výsledků než klasické analytické metody tím, že se vyhne běžně používaným aproximacím (například při konstrukci intervalových odhadů)



Bootstrap

Metoda „syntetického“ výběru

- Máme náhodný výběr: $X_1, X_2, \dots, X_k$

Bootstrap

Metoda „syntetického“ výběru

- Máme náhodný výběr: $X_1, X_2, \dots, X_k \rightarrow F_n(x)$

Bootstrap

Metoda „syntetického“ výběru

- Máme náhodný výběr: $X_1, X_2, \dots, X_k \rightarrow F_n(x)$
- Zajímá nás rozdělení pravděpodobnosti statistiky $T(X_1, X_2, \dots, X_k)$

Bootstrap

Metoda „syntetického“ výběru

- Máme náhodný výběr: $X_1, X_2, \dots, X_k \rightarrow F_n(x)$
- Zajímá nás rozdělení pravděpodobnosti statistiky $T(X_1, X_2, \dots, X_k)$
- Provedeme náhodný výběr (s vracením) z původního výběru $X_1^1, X_2^1, \dots, X_k^1$ a spočteme $T_1(X_1^1, X_2^1, \dots, X_k^1)$

Bootstrap

Metoda „syntetického“ výběru

- Máme náhodný výběr: $X_1, X_2, \dots, X_k \rightarrow F_n(x)$
- Zajímá nás rozdělení pravděpodobnosti statistiky $T(X_1, X_2, \dots, X_k)$
- Provedeme náhodný výběr (s vracením) z původního výběru
 $X_1^1, X_2^1, \dots, X_k^1$ a spočteme $T_1(X_1^1, X_2^1, \dots, X_k^1)$
-

Bootstrap

Metoda „syntetického“ výběru

- Máme náhodný výběr: $X_1, X_2, \dots, X_k \rightarrow F_n(x)$
- Zajímá nás rozdělení pravděpodobnosti statistiky $T(X_1, X_2, \dots, X_k)$
- Provedeme náhodný výběr (s vracením) z původního výběru
 $X_1^1, X_2^1, \dots, X_k^1$ a spočteme $T_1(X_1^1, X_2^1, \dots, X_k^1)$
-
- Provedeme náhodný výběr (s vracením) z původního výběru
 $X_1^n, X_2^n, \dots, X_k^n$ a spočteme $T_n(X_1^n, X_2^n, \dots, X_k^n)$

Bootstrap

Metoda „syntetického“ výběru

- Máme náhodný výběr: $X_1, X_2, \dots, X_k \rightarrow F_n(x)$
- Zajímá nás rozdělení pravděpodobnosti statistiky $T(X_1, X_2, \dots, X_k)$
- Provedeme náhodný výběr (s vracením) z původního výběru
 $X_1^1, X_2^1, \dots, X_k^1$ a spočteme $T_1(X_1^1, X_2^1, \dots, X_k^1)$
-
- Provedeme náhodný výběr (s vracením) z původního výběru
 $X_1^n, X_2^n, \dots, X_k^n$ a spočteme $T_n(X_1^n, X_2^n, \dots, X_k^n)$



Bootstrap

```
> b.samples <- function(chi, B){  
+   samples <- array(dim = c(B,length(chi)))  
+   for(i in 1:B){  
+     samples[i,]<-sample(chi,size=length(chi),replace=TRUE,prob=NULL)  
+   }  
+   return <- samples  
+ }  
> (n <- 1:10)
```

```
[1] 1 2 3 4 5 6 7 8 9 10
```

```
> (b.samples(n, 5))
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
[1,]	4	2	7	8	5	1	1	4	5	7
[2,]	6	5	7	7	4	8	1	9	3	2
[3,]	5	2	2	7	6	3	5	4	3	1
[4,]	1	7	3	4	8	10	1	7	10	2
[5,]	7	9	6	10	8	1	8	5	5	2

Bootstrap

```
> b.samples <- function(chi, B){  
+   samples <- array(dim = c(B,length(chi)))  
+   for(i in 1:B){  
+     samples[i,]<-sample(chi,size=length(chi),replace=TRUE,prob=NULL)  
+   }  
+   return <- samples  
+ }  
> (n <- 1:10)
```

```
[1] 1 2 3 4 5 6 7 8 9 10
```

```
> (b.samples(n, 5))
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
[1,]	4	2	7	8	5	1	1	4	5	7
[2,]	6	5	7	7	4	8	1	9	3	2
[3,]	5	2	2	7	6	3	5	4	3	1
[4,]	1	7	3	4	8	10	1	7	10	2
[5,]	7	9	6	10	8	1	8	5	5	2



Bootstrap

Příklad: Odhad parametru σ^2 Uvažujme výběr z $N(0,1)$.
Nestranným odhadem σ^2 je výběrový rozptyl s^2 .

Ten spočteme na základě výběru $X_1, X_2, \dots, X_k$ jako

$$s^2(x_1, \dots, x_k) = \frac{1}{k-1} \sum_{i=1}^k (x_i - \bar{x})^2$$

1) vygenerujeme náhodný výběr z $N(0,1)$:

```
> vyber <- rnorm(n = 30, mean = 0, sd = 1)
```

2) provedeme z něj 100 bootstrapových výběrů:

```
> M <- b.samples(vyber, 100)
```

3) pro každý bootstrapový výběr spočteme výběrový rozptyl:

```
> R.b <- apply(M, 1, var)
```

Bootstrap

Příklad: Odhad parametru σ^2 Uvažujme výběr z $N(0,1)$.
Nestranným odhadem σ^2 je výběrový rozptyl s^2 .

Ten spočteme na základě výběru $X_1, X_2, \dots, X_k$ jako

$$s^2(x_1, \dots, x_k) = \frac{1}{k-1} \sum_{i=1}^k (x_i - \bar{x})^2$$

1) vygenerujeme náhodný výběr z $N(0,1)$:

```
> vyber <- rnorm(n = 30, mean = 0, sd = 1)
```

2) provedeme z něj 100 bootstrapových výběrů:

```
> M <- b.samples(vyber, 100)
```

3) pro každý bootstrapový výběr spočteme výběrový rozptyl:

```
> R.b <- apply(M, 1, var)
```



Bootstrap

Příklad: Odhad parametru σ^2

Bootstrap

Příklad: Odhad parametru σ^2

1) vygenerujeme náhodný výběr z $N(0,1)$:

```
> vyber <- rnorm(n = 30, mean = 0, sd = 1)
```


Bootstrap

Příklad: Odhad parametru σ^2

1) vygenerujeme náhodný výběr z $N(0,1)$:

```
> vyber <- rnorm(n = 30, mean = 0, sd = 1)
```

2) provedeme z něj 100 bootstrapových výběrů:

```
> M <- b.samples(vyber, 100)
```

Bootstrap

Příklad: Odhad parametru σ^2

1) vygenerujeme náhodný výběr z $N(0,1)$:

```
> vyber <- rnorm(n = 30, mean = 0, sd = 1)
```

2) provedeme z něj 100 bootstrapových výběrů:

```
> M <- b.samples(vyber, 100)
```

3) pro každý bootstrapový výběr spočteme výběrový rozptyl:

```
> R.b <- apply(M, 1, var)
```

Bootstrap

Příklad: Odhad parametru σ^2

1) vygenerujeme náhodný výběr z $N(0,1)$:

```
> vyber <- rnorm(n = 30, mean = 0, sd = 1)
```

2) provedeme z něj 100 bootstrapových výběrů:

```
> M <- b.samples(vyber, 100)
```

3) pro každý bootstrapový výběr spočteme výběrový rozptyl:

```
> R.b <- apply(M, 1, var)
```

4) spočteme průměr takto spočtených rozptylů:

```
> R.estimate <- mean(R.b)
```

Bootstrap

Příklad: Odhad parametru σ^2

1) vygenerujeme náhodný výběr z $N(0,1)$:

```
> vyber <- rnorm(n = 30, mean = 0, sd = 1)
```

2) provedeme z něj 100 bootstrapových výběrů:

```
> M <- b.samples(vyber, 100)
```

3) pro každý bootstrapový výběr spočteme výběrový rozptyl:

```
> R.b <- apply(M, 1, var)
```

4) spočteme průměr takto spočtených rozptylů:

```
> R.estimate <- mean(R.b)
```

```
> R.estimate
```

```
[1] 1.152983
```

Bootstrap

Příklad: Odhad parametru σ^2

1) vygenerujeme náhodný výběr z $N(0,1)$:

```
> vyber <- rnorm(n = 30, mean = 0, sd = 1)
```

2) provedeme z něj 100 bootstrapových výběrů:

```
> M <- b.samples(vyber, 100)
```

3) pro každý bootstrapový výběr spočteme výběrový rozptyl:

```
> R.b <- apply(M, 1, var)
```

4) spočteme průměr takto spočtených rozptylů:

```
> R.estimate <- mean(R.b)
```

```
> R.estimate
```

```
[1] 1.152983
```

```
> var(vyber)
```

```
[1] 1.170146
```

Bootstrap

Příklad: Odhad parametru σ^2

1) vygenerujeme náhodný výběr z $N(0,1)$:

```
> vyber <- rnorm(n = 30, mean = 0, sd = 1)
```

2) provedeme z něj 100 bootstrapových výběrů:

```
> M <- b.samples(vyber, 100)
```

3) pro každý bootstrapový výběr spočteme výběrový rozptyl:

```
> R.b <- apply(M, 1, var)
```

4) spočteme průměr takto spočtených rozptylů:

```
> R.estimate <- mean(R.b)
```

```
> R.estimate
```

```
[1] 1.152983
```

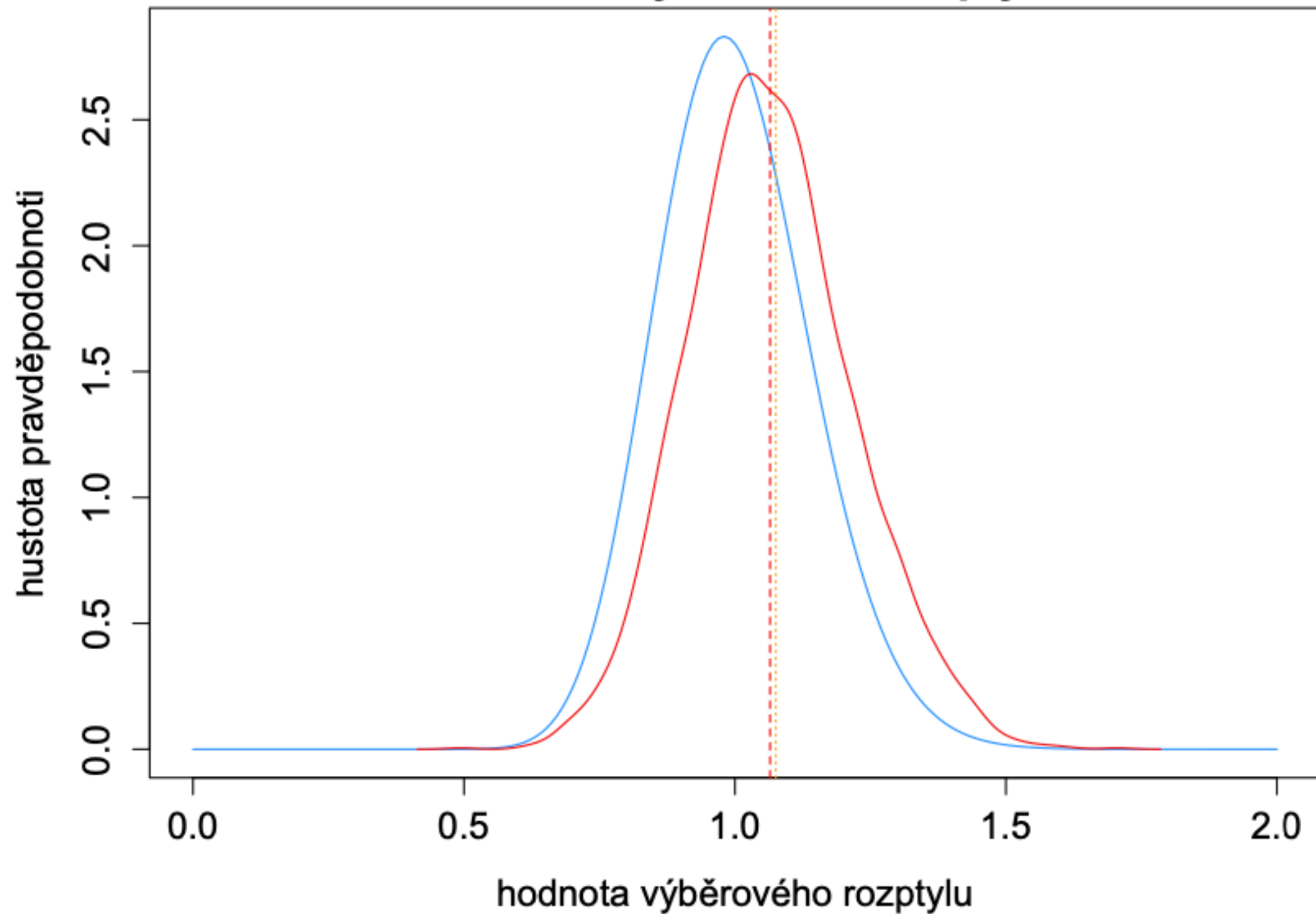
```
> var(vyber)
```

```
[1] 1.170146
```



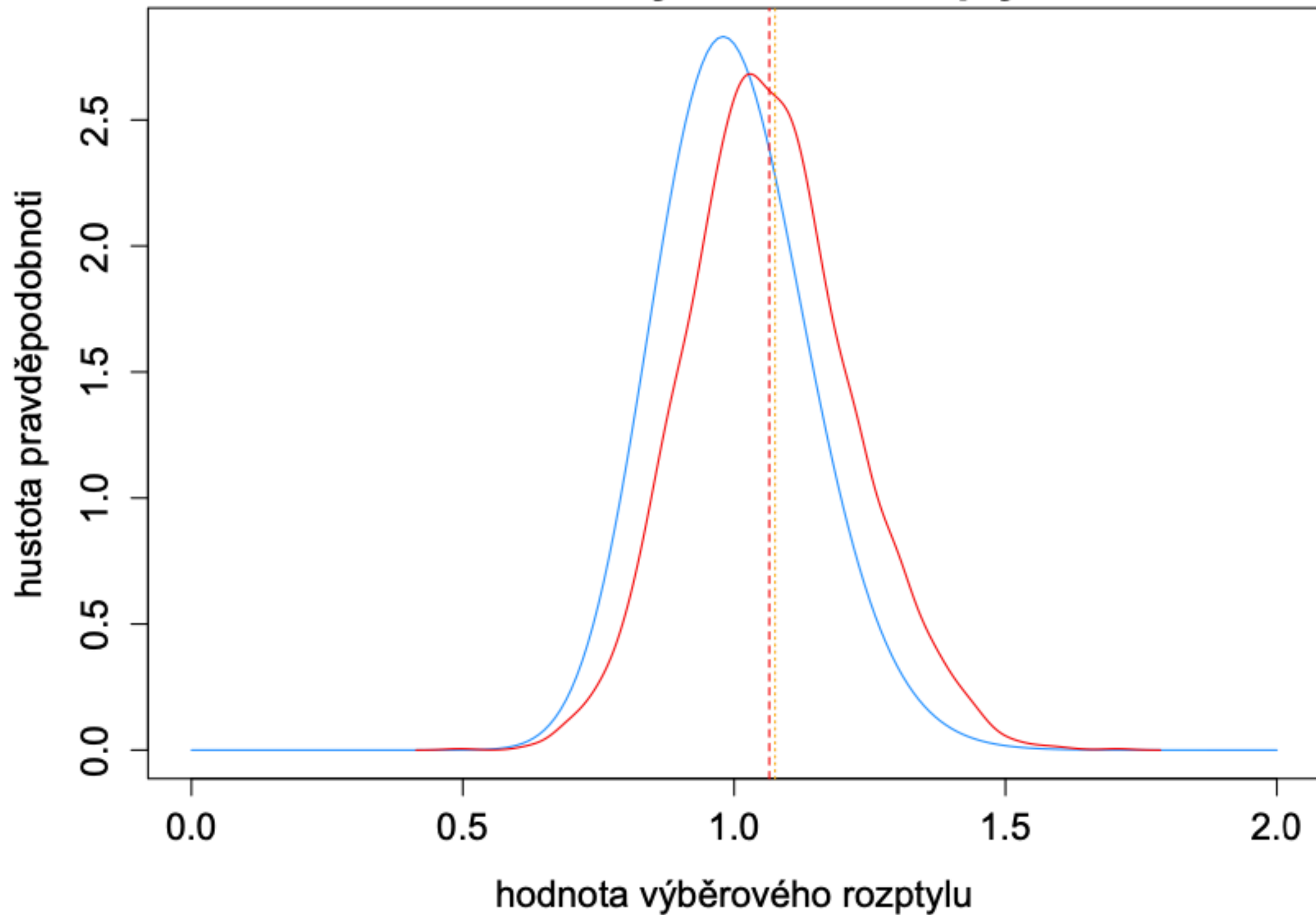
Bootstrap

Rozdělení výběrového rozptylu



Bootstrap

Rozdělení výběrového rozptylu



Vícerozměrná analýza dat (MDA)

- Standardizace, odhady parametrů polohy, rozptýlenosti, korelační analýza
- Exporatorní analýza dat (EDA), vybočující pozorování, podobnost, grafická analýza
- Snižování dimenze, hledání latentních proměnných, hledání struktury a vzájemných vazeb mezi proměnnými
- Klasifikace, diskriminační analýza, shluková analýza, zařazování objektů do tříd, hledání hierarchického uspořádání, ...

Vícerozměrná analýza dat (MDA)

- Standardizace, odhady parametrů polohy, rozptýlenosti, korelační analýza
- Exporatorní analýza dat (EDA), vybočující pozorování, podobnost, grafická analýza
- Snižování dimenze, hledání latentních proměnných, hledání struktury a vzájemných vazeb mezi proměnnými
- Klasifikace, diskriminační analýza, shluková analýza, zařazování objektů do tříd, hledání hierarchického uspořádání, ...



Snižování dimenze

- Selektce proměnných - vybíráme pouze ty, které jsou relevantní pro náš model
- Extrakce příznaků - nahrazuje pozorovaná data jejich lineárními či nelineárními kombinacemi
 - Metoda hlavních komponent (PCA)
 - Korespondenční analýza (pro kategoriální data)
 - Mnohorozměrné škálování
 - Faktorová analýza - předpokládá, že pozorovaná data lze vysvětlit pomocí malého počtu faktorů
 - Lineární diskriminační analýza (LDA) - klasifikační metoda

Snižování dimenze

- Selektce proměnných - vybíráme pouze ty, které jsou relevantní pro náš model
- Extrakce příznaků - nahrazuje pozorovaná data jejich lineárními či nelineárními kombinacemi
 - Metoda hlavních komponent (PCA)
 - Korespondenční analýza (pro kategoriální data)
 - Mnohorozměrné škálování
 - Faktorová analýza - předpokládá, že pozorovaná data lze vysvětlit pomocí malého počtu faktorů
 - Lineární diskriminační analýza (LDA) - klasifikační metoda



Metoda hlavních komponent (PCA)

- Analýza hlavních komponent je často využívána u vícerozměrných metod jako první krok při velkém počtu měření (případů) nebo proměnných typicky s úkolem provést jejich redukci.

Metoda hlavních komponent (PCA)

- Analýza hlavních komponent je často využívána u vícerozměrných metod jako první krok při velkém počtu měření (případů) nebo proměnných typicky s úkolem provést jejich redukci.
- Vyjadřuje informace o variabilitě obsažené v datovém souboru pomocí několika málo nových znaků získaných jako lineární kombinace znaků původních (latentní proměnné)

Metoda hlavních komponent (PCA)

- Analýza hlavních komponent je často využívána u vícerozměrných metod jako první krok při velkém počtu měření (případů) nebo proměnných typicky s úkolem provést jejich redukci.
- Vyjadřuje informace o variabilitě obsažené v datovém souboru pomocí několika málo nových znaků získaných jako lineární kombinace znaků původních (latentní proměnné)
- V podstatě se jedná o transformaci z původního do nového souřadnicového systému, jehož osy jsou tvořeny tzv. „hlavními komponentami“

Metoda hlavních komponent (PCA)

- Analýza hlavních komponent je často využívána u vícerozměrných metod jako první krok při velkém počtu měření (případů) nebo proměnných typicky s úkolem provést jejich redukci.
- Vyjadřuje informace o variabilitě obsažené v datovém souboru pomocí několika málo nových znaků získaných jako lineární kombinace znaků původních (latentní proměnné)
- V podstatě se jedná o transformaci z původního do nového souřadnicového systému, jehož osy jsou tvořeny tzv. „hlavními komponentami“
- Komponenty jsou zařazovány v pořadí takovém, že první vysvětluje největší procento celkové variability, a jsou řazeny v pořadí podle vysvětlení původní variability.

Metoda hlavních komponent (PCA)

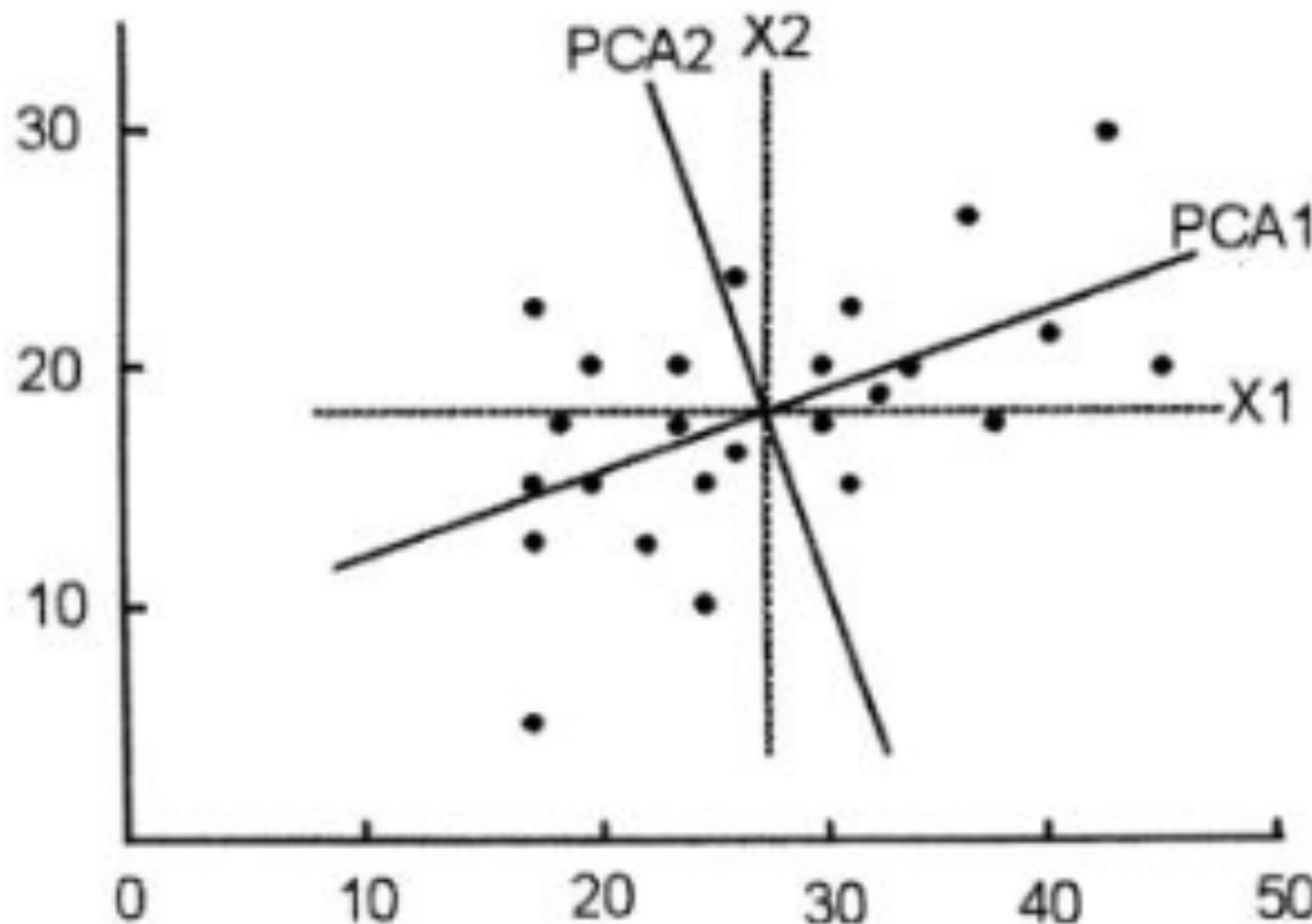
- Analýza hlavních komponent je často využívána u vícerozměrných metod jako první krok při velkém počtu měření (případů) nebo proměnných typicky s úkolem provést jejich redukci.
- Vyjadřuje informace o variabilitě obsažené v datovém souboru pomocí několika málo nových znaků získaných jako lineární kombinace znaků původních (latentní proměnné)
- V podstatě se jedná o transformaci z původního do nového souřadnicového systému, jehož osy jsou tvořeny tzv. „hlavními komponentami“
- Komponenty jsou zařazovány v pořadí takovém, že první vysvětluje největší procento celkové variability, a jsou řazeny v pořadí podle vysvětlení původní variability.
- Každá další komponenta vysvětluje co nejvíce ze zbývajících celkové variability.

Metoda hlavních komponent (PCA)

- Analýza hlavních komponent je často využívána u vícerozměrných metod jako první krok při velkém počtu měření (případů) nebo proměnných typicky s úkolem provést jejich redukci.
- Vyjadřuje informace o variabilitě obsažené v datovém souboru pomocí několika málo nových znaků získaných jako lineární kombinace znaků původních (latentní proměnné)
- V podstatě se jedná o transformaci z původního do nového souřadnicového systému, jehož osy jsou tvořeny tzv. „hlavními komponentami“
- Komponenty jsou zařazovány v pořadí takovém, že první vysvětluje největší procento celkové variability, a jsou řazeny v pořadí podle vysvětlení původní variability.
- Každá další komponenta vysvětluje co nejvíce ze zbývajících celkové variability.
- Komponenty již nejsou vzájemně korelované.

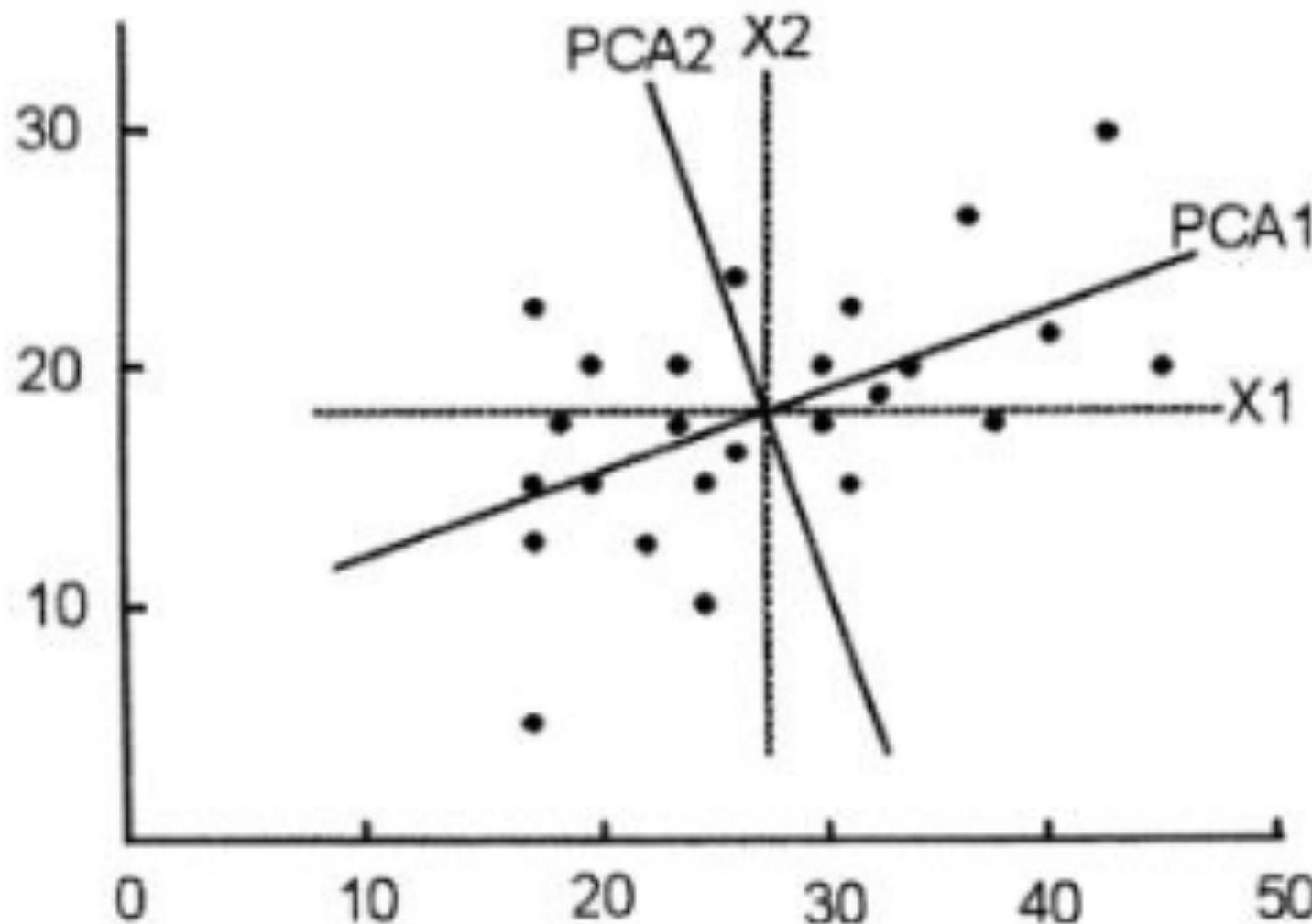
Metoda hlavních komponent (PCA)

- Analýza hlavních komponent je často využívána u vícerozměrných metod jako
 - Vyjadřuje i pomocí ně
 - V podstatě souřadnic
 - Komponent největší pr
 - Každá dal
 - Komponenty již nejsou vzájemně korelované.
- nebo
uboru
rní
ho
vními
vysvětluje
dí podle
cí celkové



Metoda hlavních komponent (PCA)

- Analýza hlavních komponent je často využívána u vícerozměrných metod jako
- Vyjadřuje i pomocí ně kombinace nebo
- V podstatě souřadnic komponent uboru
urní
- Komponenty největší pr vysvětlení ho
vními
- Každá další variability vysvětluje
dí podle
- Komponenty již nejsou vzájemně korelované. cí celkové



Metoda hlavních komponent (PCA)

Příklad: U třiceti čtyř vybraných domácností byly zjištěny údaje o počtu členů domácnosti (X_1), počtu dětí (X_2), průměrném věku (X_3), měsíčním příjmu domácnosti (X_4) a měsíčních výdajích za potraviny a nápoje (X_5). Cílem je zredukovat počet vstupních proměnných pro další analýzu.

Domácnosti	Počet členů	Počet dětí	Průměr věku	Příjem	Vydání						
1	2	0	60,5	14290	5177	17	4	2	33,5	13998	5155
2	2	0	60,0	13459	5840	18	4	2	32,5	19728	7130
3	1	0	55,0	5586	1732	19	3	1	36,0	12136	4855
4	3	1	38,5	19593	6908	20	2	0	28,0	20484	2357
5	4	2	35,0	15193	5389	21	2	0	59,5	19187	4339
6	4	2	34,5	14741	5683	22	6	4	33,5	20462	3786
7	2	0	62,5	12055	5073	23	1	0	21,0	4434	991
8	2	0	57,0	19265	4101	24	1	0	21,0	4089	1936
9	2	0	60,5	7908	4584	25	1	0	49,0	8707	1614
10	2	0	54,5	11451	4666	26	4	2	38,0	24229	5467
11	3	1	45,5	14379	5164	27	2	0	39,0	13737	2451
12	4	2	34,5	16236	5178	28	1	0	22,0	5365	1517
13	5	2	47,0	20032	8475	29	1	0	23,0	4810	832
14	4	2	30,5	15875	5284	30	4	2	37,0	18995	3194
15	2	0	57,5	12724	3548	31	4	2	33,5	12460	4292
16	3	1	28,5	16829	2393	32	4	2	36,0	20146	3597
17	4	2	33,5	13998	5155	33	2	0	28,0	10163	3124
						34	1	0	28,0	13998	1552

Metoda hlavních komponent (PCA)

Příklad: U třiceti čtyř vybraných domácností byly zjištěny údaje o počtu členů domácnosti (X_1), počtu dětí (X_2), průměrném věku (X_3), měsíčním příjmu domácnosti (X_4) a měsíčních výdajích za potraviny a nápoje (X_5). Cílem je zredukovat počet vstupních proměnných pro další analýzu.

Domácnost	Počet členů	Počet dětí	Průměr věku	Příjem	Vydání						
1	2	0	60,5	14290	5177	17	4	2	33,5	13998	5155
2	2	0	60,0	13459	5840	18	4	2	32,5	19728	7130
3	1	0	55,0	5586	1732	19	3	1	36,0	12136	4855
4	3	1	38,5	19593	6908	20	2	0	28,0	20484	2357
5	4	2	35,0	15193	5389	21	2	0	59,5	19187	4339
6	4	2	34,5	14741	5683	22	6	4	33,5	20462	3786
7	2	0	62,5	12055	5073	23	1	0	21,0	4434	991
8	2	0	57,0	19265	4101	24	1	0	21,0	4089	1936
9	2	0	60,5	7908	4584	25	1	0	49,0	8707	1614
10	2	0	54,5	11451	4666	26	4	2	38,0	24229	5467
11	3	1	45,5	14379	5164	27	2	0	39,0	13737	2451
12	4	2	34,5	16236	5178	28	1	0	22,0	5365	1517
13	5	2	47,0	20032	8475	29	1	0	23,0	4810	832
14	4	2	30,5	15875	5284	30	4	2	37,0	18995	3194
15	2	0	57,5	12724	3548	31	4	2	33,5	12460	4292
16	3	1	28,5	16829	2393	32	4	2	36,0	20146	3597
17	4	2	33,5	13998	5155	33	2	0	28,0	10163	3124
						34	1	0	28,0	13998	1552

Metoda hlavních komponent (PCA)

Příklad: U třiceti čtyř vybraných domácností byly zjištěny údaje o počtu členů domácnosti (X_1), počtu dětí (X_2), průměrném věku (X_3), měsíčním příjmu domácnosti (X_4) a měsíčních výdajích za potraviny a nápoje (X_5). Cílem je zredukovat počet vstupních proměnných pro další analýzu.

$$\mathbf{X} = \begin{pmatrix} X_1^1 & X_2^1 & X_3^1 & X_4^1 & X_5^1 \\ X_1^2 & X_2^2 & X_3^2 & X_4^2 & X_5^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_1^{34} & X_2^{34} & X_3^{34} & X_4^{34} & X_5^{34} \end{pmatrix}$$

Metoda je založena na tzv. „spektrálním rozkladu empirické kovarianční matice

$$\mathbf{S} = \frac{1}{33} \sum_{i=1}^{34} (\mathbf{X}^i - \bar{\mathbf{X}})(\mathbf{X}^i - \bar{\mathbf{X}})^T$$

pomocí něhož dostaneme vlastní čísla $\lambda_1, \dots, \lambda_5$ a vlastní vektory $\mathbf{q}_1, \dots, \mathbf{q}_5$.

Platí, že součet vlastních čísel je roven součtu rozptylů jednotlivých proměnných.

Metoda hlavních komponent (PCA)

Příklad: U třiceti čtyř vybraných domácností byly zjištěny údaje o počtu členů domácnosti (X_1), počtu dětí (X_2), průměrném věku (X_3), měsíčním příjmu domácnosti (X_4) a měsíčních výdajích za potraviny a nápoje (X_5). Cílem je zredukovat počet vstupních proměnných pro další analýzu.

$$\mathbf{X} = \begin{pmatrix} X_1^1 & X_2^1 & X_3^1 & X_4^1 & X_5^1 \\ X_1^2 & X_2^2 & X_3^2 & X_4^2 & X_5^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_1^{34} & X_2^{34} & X_3^{34} & X_4^{34} & X_5^{34} \end{pmatrix}$$

Metoda je založena na tzv. „spektrálním rozkladu empirické kovarianční matice

$$\mathbf{S} = \frac{1}{33} \sum_{i=1}^{34} (\mathbf{X}^i - \bar{\mathbf{X}})(\mathbf{X}^i - \bar{\mathbf{X}})^T$$

pomocí něhož dostaneme vlastní čísla $\lambda_1, \dots, \lambda_5$ a vlastní vektory $\mathbf{q}_1, \dots, \mathbf{q}_5$.

Platí, že součet vlastních čísel je roven součtu rozptylů jednotlivých proměnných.



Metoda hlavních komponent (PCA)

Příklad: U třiceti čtyř vybraných domácností byly zjištěny údaje o počtu členů domácnosti (X_1), počtu dětí (X_2), průměrném věku (X_3), měsíčním příjmu domácnosti (X_4) a měsíčních výdajích za potraviny a nápoje (X_5). Cílem je zredukovat počet vstupních proměnných pro další analýzu.

Barevná matice absolut. hodnoty r (domácnosti) N=34 (Celé případy vynechány u ChD) abs(r)>=		0	0,10	0,20	0,30	0,40	0,5	0,60	0,70	0,80	0,90	1
Proměnná	Počet členů X1	Počet dětí X2	Průměr. věk X3	Příjem X4	Vydání X5							
Počet členů X1	Vlastní čísla korelační matice a související statistiky (domácnosti) Pouze aktiv. proměnné											
Počet dětí X2												
Průměr. věk X3	Pořadí vl. č.	vl. číslo	% celk. rozptylu	Kumulativ. vl. číslo	Kumulativ. %							
Příjem X4	1	2,980	59,591	2,980	59,591							
Vydání X5	2	1,330	26,594	4,309	86,185							
	3	0,408	8,162	4,717	94,348							
	4	0,265	5,296	4,982	99,643							
	5	0,018	0,357	5,000	100,000							

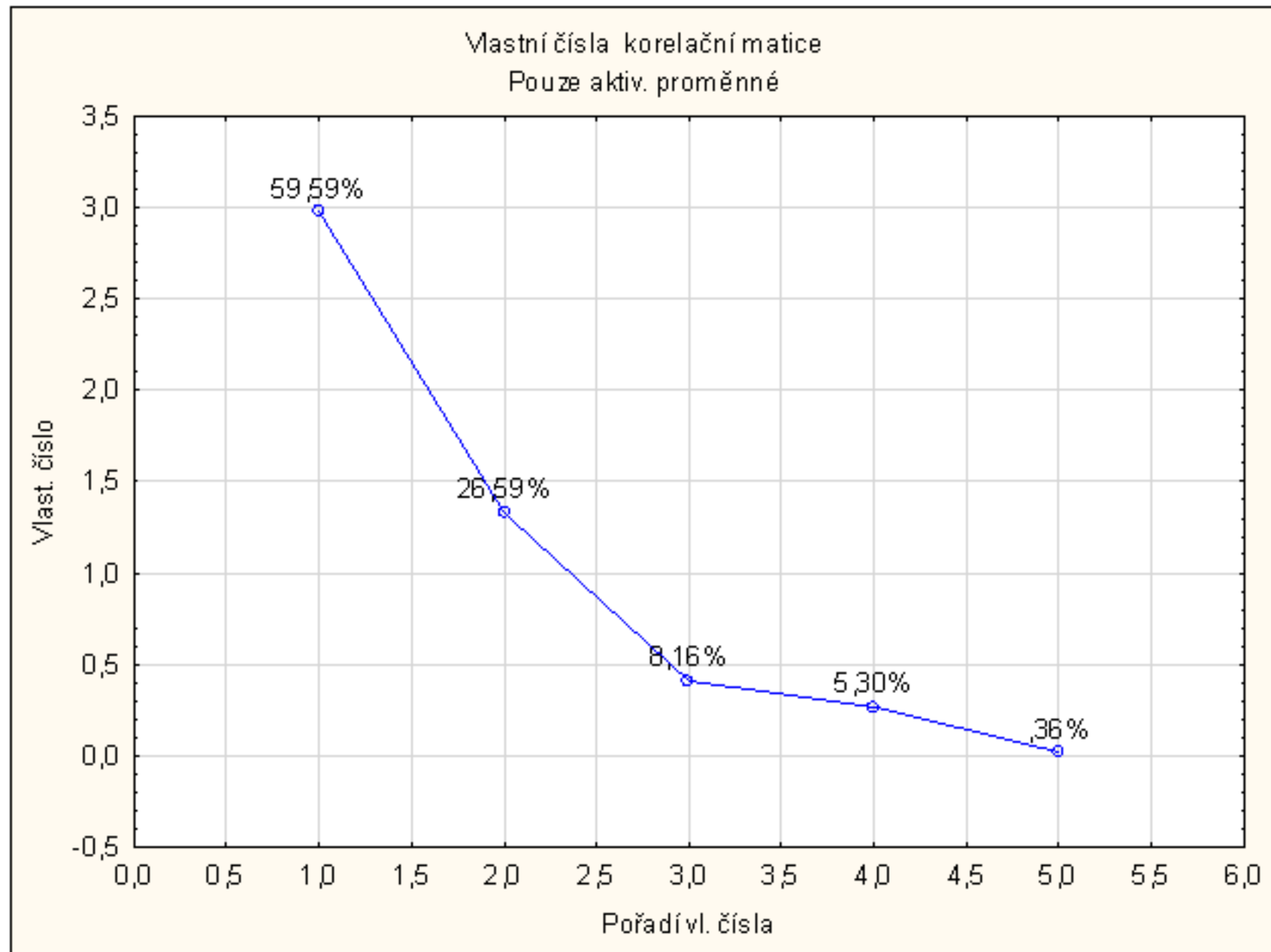
Metoda hlavních komponent (PCA)

Příklad: U třiceti čtyř vybraných domácností byly zjištěny údaje o počtu členů domácnosti (X_1), počtu dětí (X_2), průměrném věku (X_3), měsíčním příjmu domácnosti (X_4) a měsíčních výdajích za potraviny a nápoje (X_5). Cílem je zredukovat počet vstupních proměnných pro další analýzu.

Barevná matice absolut. hodnoty r (domácnosti) N=34 (Celé případy vynechány u ChD) abs(r)>=		0	0,10	0,20	0,30	0,40	0,5	0,60	0,70	0,80	0,90	1
Proměnná	Počet členů X1	Počet dětí X2	Průměr. věk X3	Příjem X4	Vydání X5							
Počet členů X1	Vlastní čísla korelační matice a související statistiky (domácnosti) Pouze aktiv. proměnné											
Počet dětí X2												
Průměr. věk X3	Pořadí vl. č.	vl. číslo	% celk. rozptylu	Kumulativ. vl. číslo	Kumulativ. %							
Příjem X4	1	2,980	59,591	2,980	59,591							
Vydání X5	2	1,330	26,594	4,309	86,185							
	3	0,408	8,162	4,717	94,348							
	4	0,265	5,296	4,982	99,643							
	5	0,018	0,357	5,000	100,000							

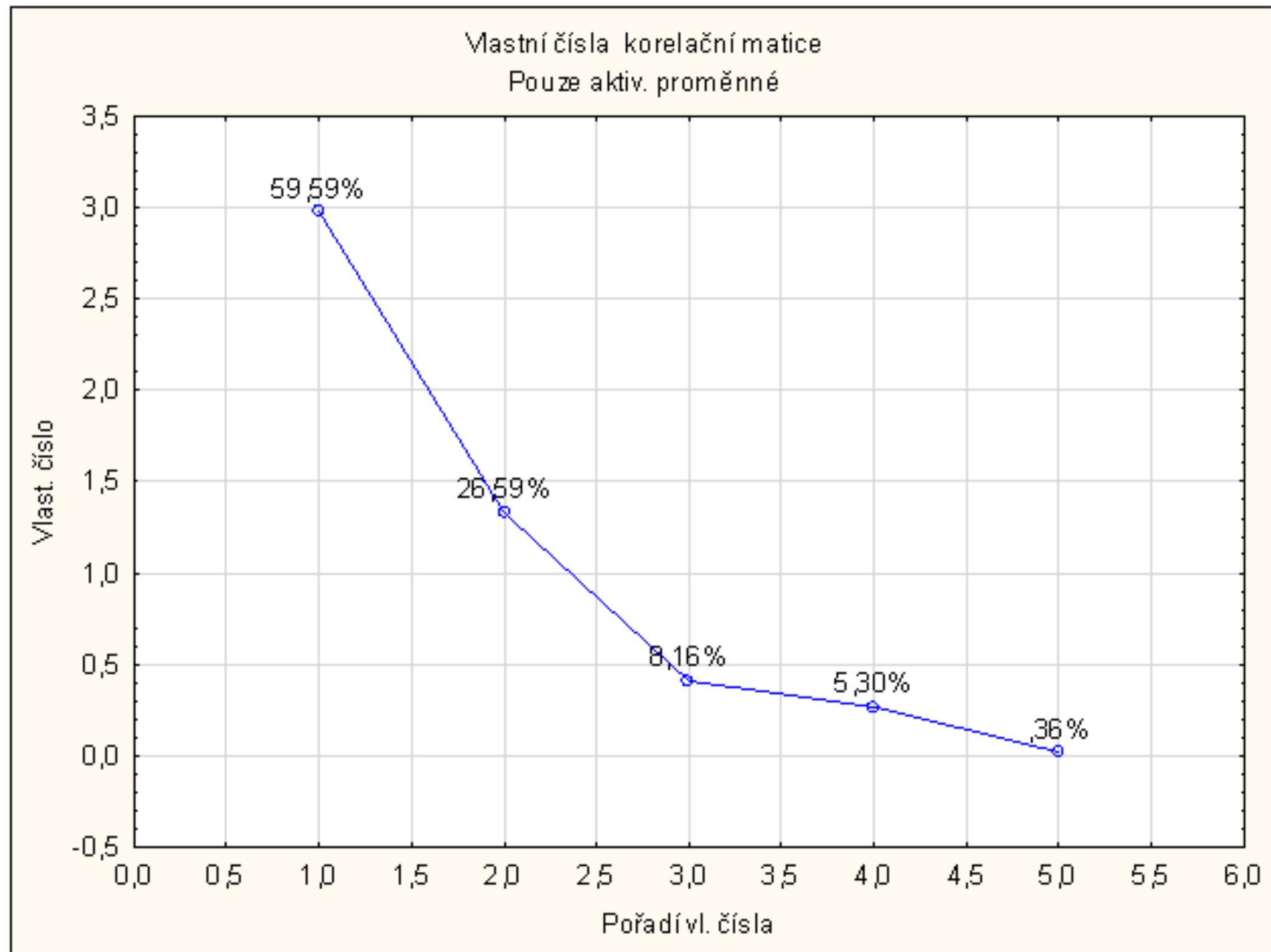


Metoda hlavních komponent (PCA)



Scree diagram

Metoda hlavních komponent (PCA)



Scree diagram



Metoda hlavních komponent (PCA)

Příklad: U třiceti čtyř vybraných domácností byly zjištěny údaje o počtu členů domácnosti (X_1), počtu dětí (X_2), průměrném věku (X_3), měsíčním příjmu domácnosti (X_4) a měsíčních výdajích za potraviny a nápoje (X_5). Cílem je zredukovat počet vstupních proměnných pro další analýzu.

Faktor „sociální úrovně“

Faktor. zátěže (Varimax pr.) (domácnosti) Extrakce: Hlavní komponenty (Označené zátěže jsou >,690000)				
Proměnná	Faktor 1	Faktor 2		
Počet členů X1	0,979918	-0,033381		
Počet dětí X2	0,926672	-0,255325		
Průměr. věk X3	-0,124075	0,950098		
Příjem X4	0,791642	0,266056		
Vydání X5	0,699283	0,565199		
Výkl.roz	2,950048	1,359226		
Prp.celk	0,500010	0,271845		

Faktor „věkové struktury domácnosti“

$$F_1 = 0,979918 X_1 + 0,926672 X_2 - 0,124075 X_3 + 0,791642 X_4 + 0,699283 X_5$$

$$F_2 = -0,033381 X_1 - 0,255325 X_2 + 0,950098 X_3 + 0,266056 X_4 + 0,565199 X_5$$

Metoda hlavních komponent (PCA)

Příklad: U třiceti čtyř vybraných domácností byly zjištěny údaje o počtu členů domácnosti (X_1), počtu dětí (X_2), průměrném věku (X_3), měsíčním příjmu domácnosti (X_4) a měsíčních výdajích za potraviny a nápoje (X_5). Cílem je zredukovat počet vstupních proměnných pro další analýzu.

Faktor „sociální úrovně“

Faktor. zátěže (Varimax pr.) (domácnosti) Extrakce: Hlavní komponenty (Označené zátěže jsou >,690000)				
Proměnná	Faktor 1	Faktor 2		
Počet členů X1	0,979918	-0,033381		
Počet dětí X2	0,926672	-0,255325		
Průměr. věk X3	-0,124075	0,950098		
Příjem X4	0,791642	0,266056		
Vydání X5	0,699283	0,565199		
Výkl.roz	2,950048	1,359226		
Prp.celk	0,500010	0,271845		

Faktor „věkové struktury domácnosti“

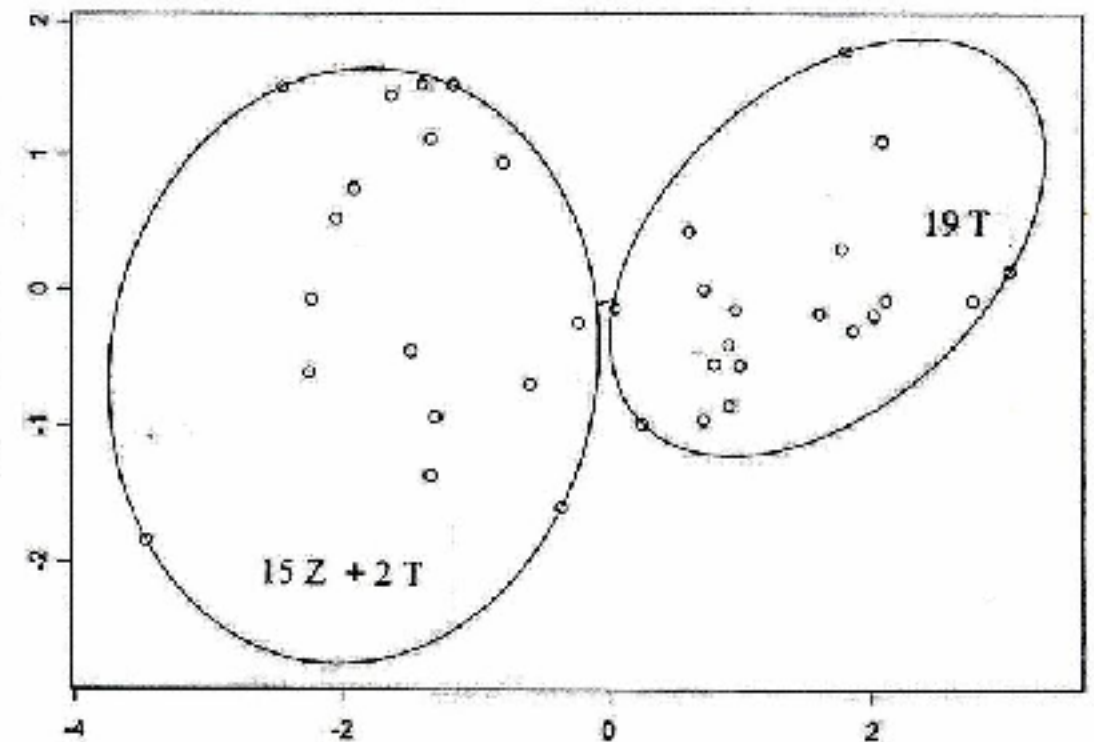
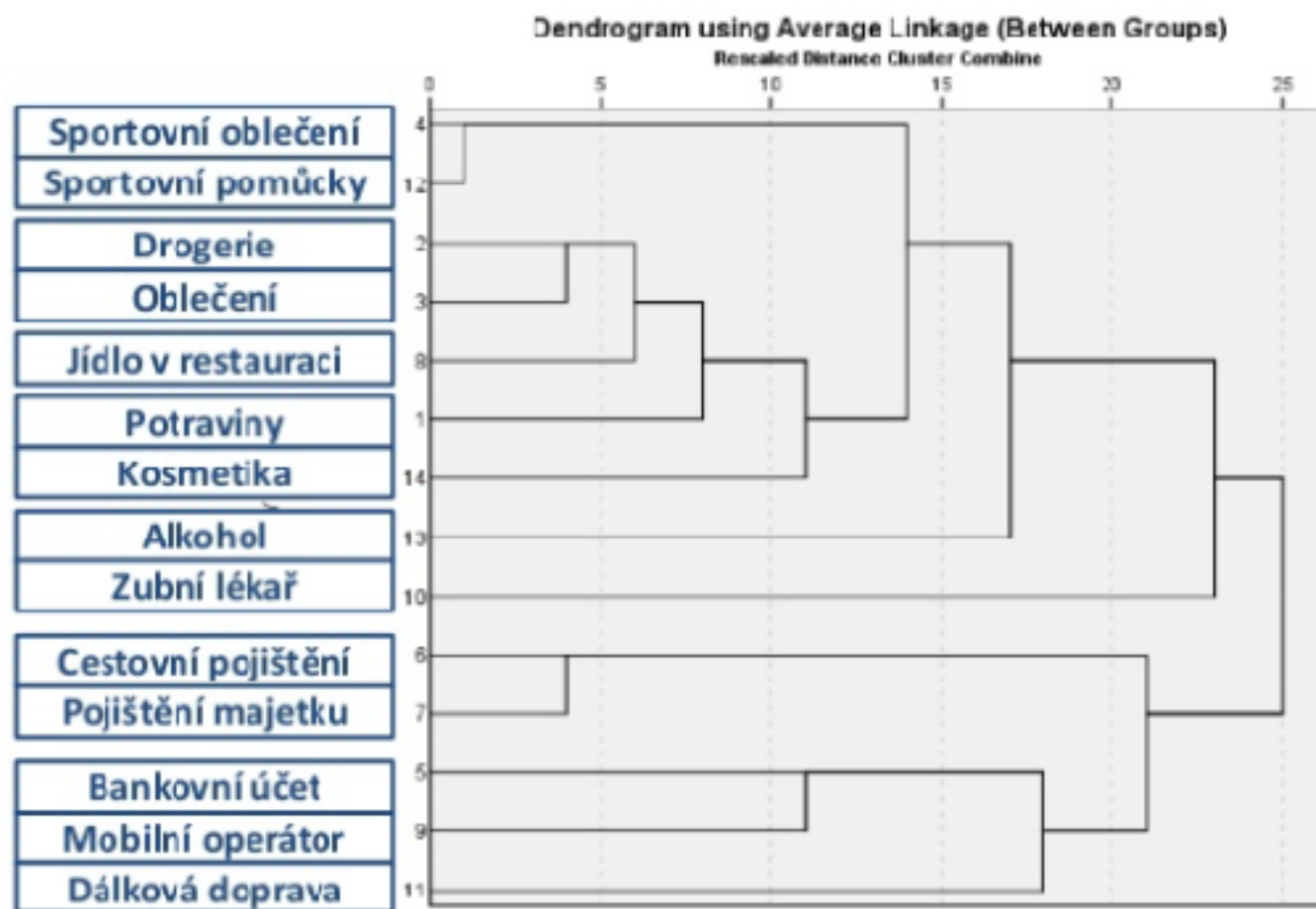
$$F_1 = 0,979918 X_1 + 0,926672 X_2 - 0,124075 X_3 + 0,791642 X_4 + 0,699283 X_5$$

$$F_2 = -0,033381 X_1 - 0,255325 X_2 + 0,950098 X_3 + 0,266056 X_4 + 0,565199 X_5$$



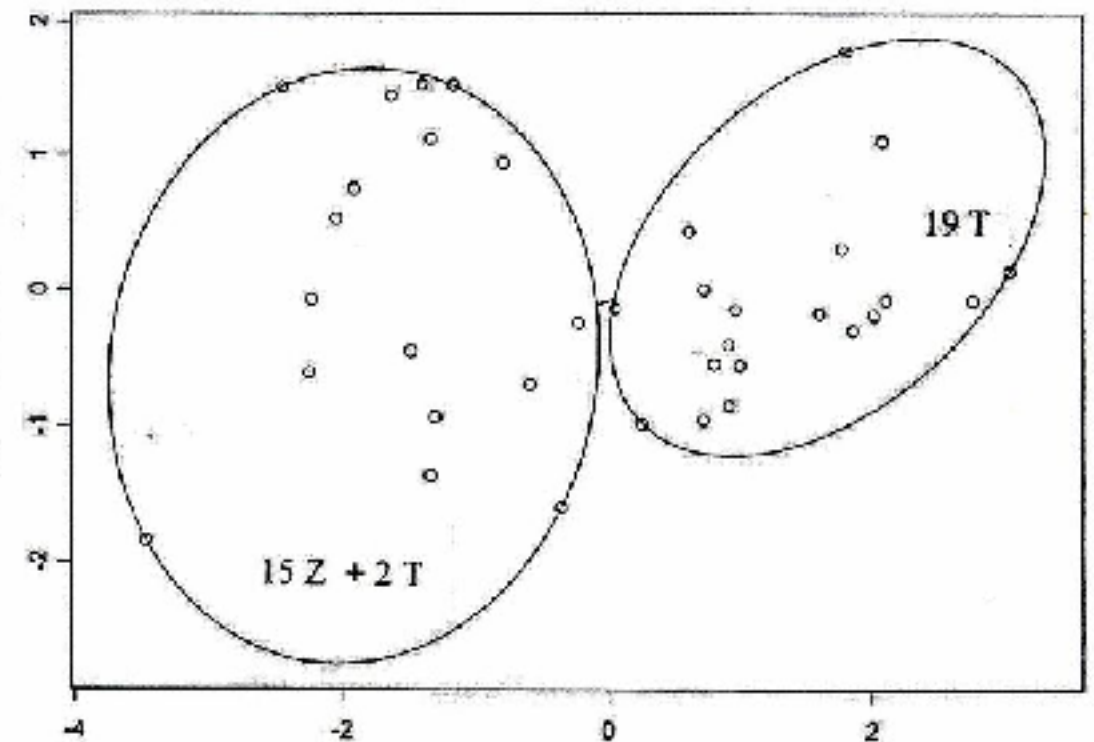
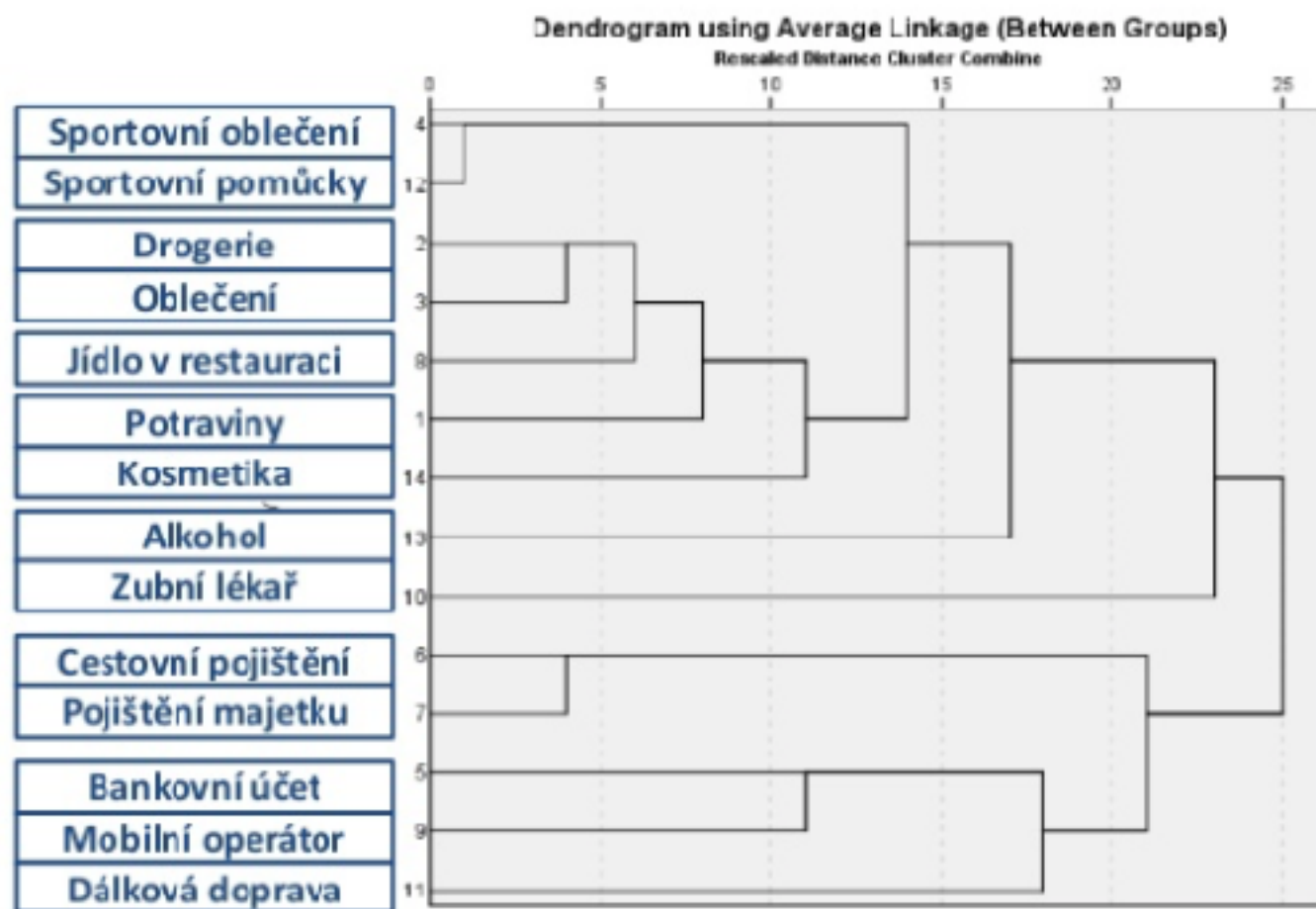
Shluková (clusterová) analýza

- Zabývá se vyšetřováním podobnosti vícerozměrných objektů a jejich klasifikací do tříd (shluků).
- Dělení do tříd může být hierarchické (dendrogram) nebo nehierarchické (disjunktní shluky).



Shluková (clusterová) analýza

- Zabývá se vyšetřováním podobnosti vícerozměrných objektů a jejich klasifikací do tříd (shluků).
- Dělení do tříd může být hierarchické (dendrogram) nebo nehierarchické (disjunktní shluky).



Shluková (clusterová) analýza

Hlavním kriteriem tvorby shluků je míra podobnosti objektů (kvantitativní vyjádření podobnosti).

- korelační míry (Pearsonův či Spearmanův korelační koeficient, ...)
- míry vzdálenosti (Eukleidovská metrika, Manhattanská vzdálenost, Mahalanobisova metrika, ...)
- míry asociace (pro binární či kategoriální proměnné)

Shluk je potom skupina objektů, jejichž podobnost (korelace, vzdálenost, asociace) je větší než podobnost s objekty do shluku nepatřícími.

Shluková (clusterová) analýza

Hlavním kritériem tvorby shluků je míra podobnosti objektů (kvantitativní vyjádření podobnosti).

- korelační míry (Pearsonův či Spearmanův korelační koeficient, ...)
- míry vzdálenosti (Eukleidovská metrika, Manhattanská vzdálenost, Mahalanobisova metrika, ...)
- míry asociace (pro binární či kategoriální proměnné)

Shluk je potom skupina objektů, jejichž podobnost (korelace, vzdálenost, asociace) je větší než podobnost s objekty do shluku nepatřícími.



Závěrem

Jsou tři druhy lží:

- a) lež
- b) zavrženíhodná lež
- c) statistika

Závěrem

Jsou tři druhy lží:

a) lež

b) zavrženíhodná lež

~~c) statistika~~

Závěrem

Jsou tři druhy lží:

a) lež

b) zavrženíhodná lež

~~c) statistika~~

Jsou pouze různí lidé s různými zájmy, kteří mají různé pohledy na svět kolem nás a často se snaží k tomu zneužít statistiku.

Závěrem

Jsou tři druhy lží:

a) lež

b) zavrženíhodná lež

~~c) statistika~~

Jsou pouze různí lidé s různými zájmy, kteří mají různé pohledy na svět kolem nás a často se snaží k tomu zneužít statistiku.



Děkuji za pozornost;)