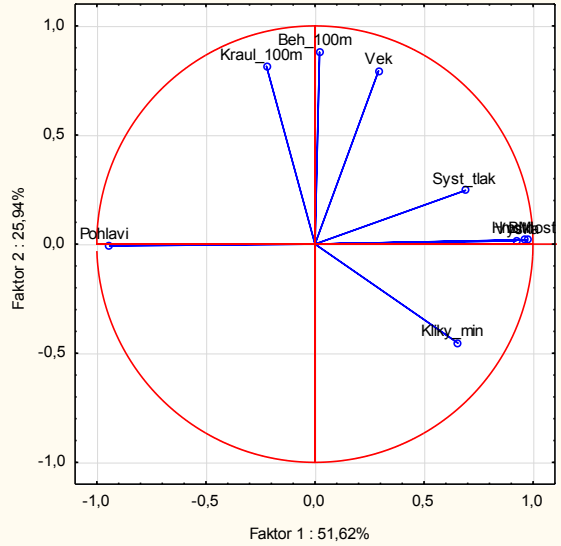
Analýza hlavních komponent

(Principal Component Analysis – PCA)



Brno, 2.5.2016
Klára Ambrožová

Motivační příklad

- Hasičský záchranný sbor Jihomoravského kraje prováděl nábor
- Přihlásilo se 32 uchazečů (15 žen a 17 mužů)
- Byly provedeny 3 testy (běh na 100 m, kraul na 100 m a počet kliků za minutu)
- Dále byly u všech uchazečů zjištěny některé osobní charakteristiky (pohlaví, věk, hmotnost, výška, BMI a krevní tlak)

Motivační příklad

- Jak vybrat uchazeče?

- Možnost stanovit limit – jenže u které proměnné a měl by být pro všechny stejný?

Data: osoby.sta (9s krát 32ř)

	1 Pohlavi	2 Vek	3 Vyska	4 Hmotnost	5 BMI	6 Syst_tlak	7 **Beh_100m**	8 Kraul_100m	9 Kliky_min
Jan	1	48	198	92	23,46699	140	14,2	99	28
Aleš	1	33	184	84	24,81096	150	13,8	85	32
Martin	1	37	183	83	24,78426	150	13,1	84	40
Helena	2	32	166	47	17,05618	90	14,5	95	20
Horymír	1	23	170	60	20,76125	110	12,9	90	33
Anna	2	24	172	64	21,63332	100	13,8	105	22
Jiří	1	35	182	80	24,15167	130	15,2	110	38
Petr	1	36	180	80	24,69136	160	14,4	112	34
Dana	2	24	169	51	17,85652	120	14,5	100	28
Zuzana	2	27	168	52	18,42404	110	12,8	85	38
Tomáš	1	37	183	81	24,18705	140	16,6	150	30

Pozn. Muži jsou v tabulce označeni číselm „1“, ženy číslem „2“ – důležité pro interpretaci!

Proměnné běh a kraul jsou v počtu sekund, tedy čím vyšší, tím horší výsledek.

Trocha teorie:

Analýza hlavních komponent (PCA)

- Metoda pro zhodnocení vícerozměrných dat (= u 1 jedince daného souboru jsou známy hodnoty více proměnných

- Pokud spolu proměnné navzájem korelují, je obtížné zjistit, jaké informace vlastně reálně získáváme

- Cílem je nahrazení velkého počtu proměnných menším počtem, které budou nekorelované (a ideálně je budeme schopni interpretovat ☺)

- **DOSTATEČNĚ SILNÁ KORELACE** MEZI PROMĚNNÝMI JE DŮLEŽITÝM **PŘEDPOKLADEM** POUŽITÍ PCA!

Trocha teorie:

Hlavní pojmy PCA

- **Hlavní komponenta – HK (faktor):**
  - Nová proměnná, kterou lze vypočítat jako lineární kombinaci původních proměnných

$$PC_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$

- **Vlastní vektor (např. 1 HK):**
  - Soubor čísel, které vstupují jako koeficienty do rovnice pro výpočet (první) hlavní komponenty a_1–a_x

- **Vlastní číslo/ hodnota (Eigenvalue):**
  - Charakteristické číslo pro danou HK
  - Udává, jakou část celkového rozptylu (v absolutních jednotkách!) vysvětluje daná hlavní komponenta

No.	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
1	3.487151	38.75	38.75	||||||||
2	2.130173	23.67	62.41	|||||
3	1.098958	12.21	74.63	|||
4	0.994483	11.05	85.68	|||
5	0.543218	6.04	91.71	||
6	0.383428	4.26	95.97	|
7	0.225754	2.51	98.48	|
8	0.136790	1.52	100.00	|
9	0.000046	0.00	100.00	|

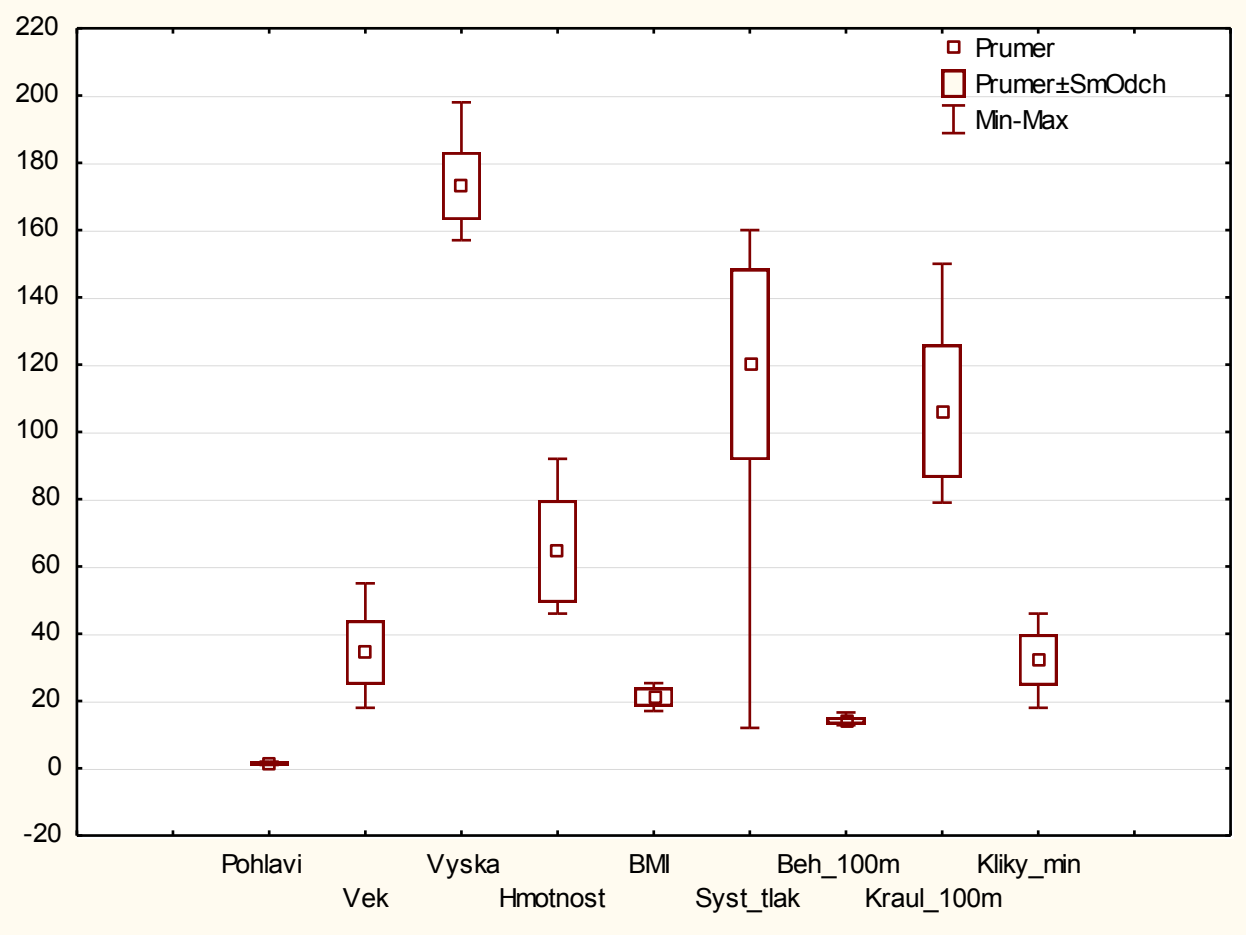
- **Zátěž (Loading):**
  - Číslo charakterizující vztah každé původní proměnné a každé hlavní komponenty
  - Udává, jak moc tato původní proměnná souvisí (koreluje) s touto hlavní komponentou

Variables	Factor1	Factor2	Factor3	Factor4
AGR	0.523791	0.053594	0.048674	0.028793
MIN	0.001323	0.617807	-0.201100	0.064085
MAN	-0.347495	0.355054	-0.150463	-0.346088
PS	-0.255716	0.261096	-0.561083	0.393309
CON	-0.325179	0.051288	0.153321	-0.668324

Trocha teorie:

Doporučený postup při PCA

1. Prohlédneme si data (*Grafy – 2D – Krabicový – Vícenásobný)*



Proč to děláme?

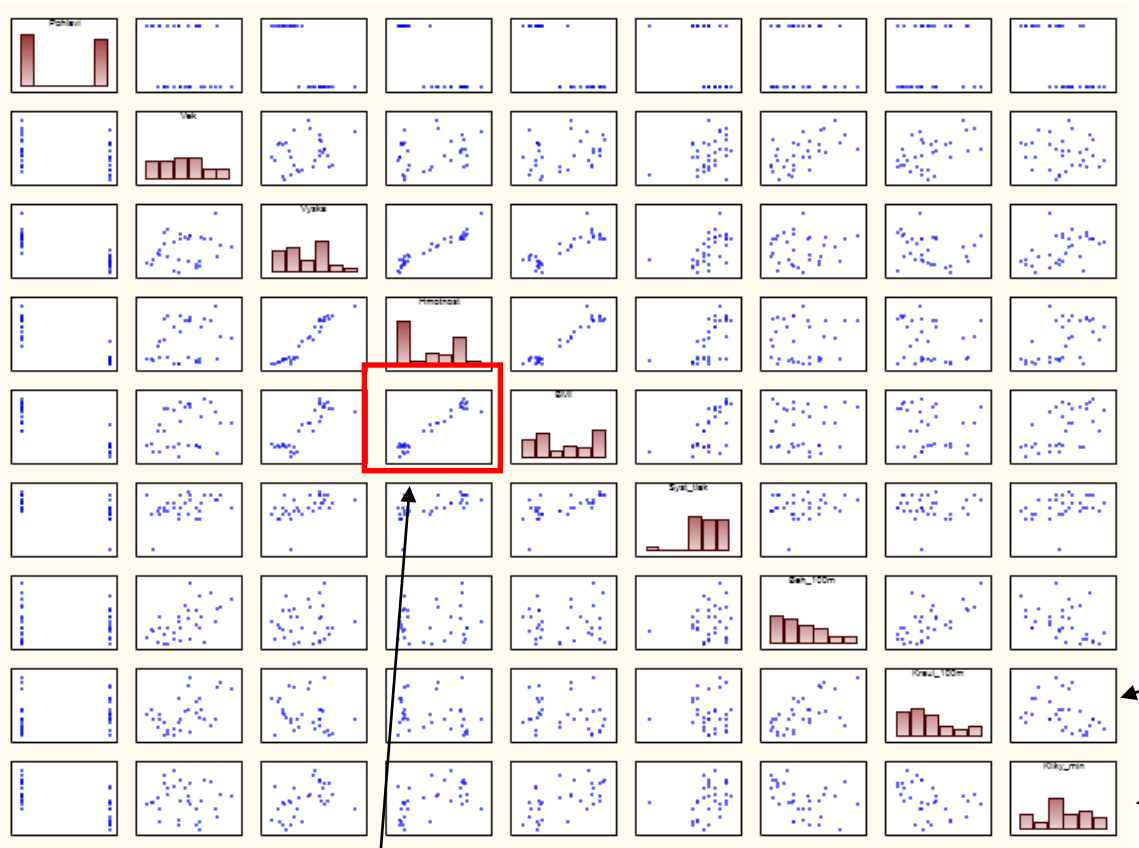
Potřebujeme zjistit, zda proměnné mají odlišnou variabilitu – pokud ano, budeme pracovat s **výběrovou korelační maticí R** (a nikoliv výběrovou kovariační maticí)

Trocha teorie:

Doporučený postup při PCA

2. Vytvoříme maticové grafy

(Grafy – Matice – Proměnné: Vše – Čtverec bodových)



„Hromadí“-li se někde v grafu body na přímce, máme korelaci a ze použít PCA

Proč to děláme?

Chceme zjistit, zda máme alespoň mezi některými proměnnými korelaci. Pokud ne, nemá smysl PCA provádět.

Bodový graf hodnot proměnné č. 8 a 9

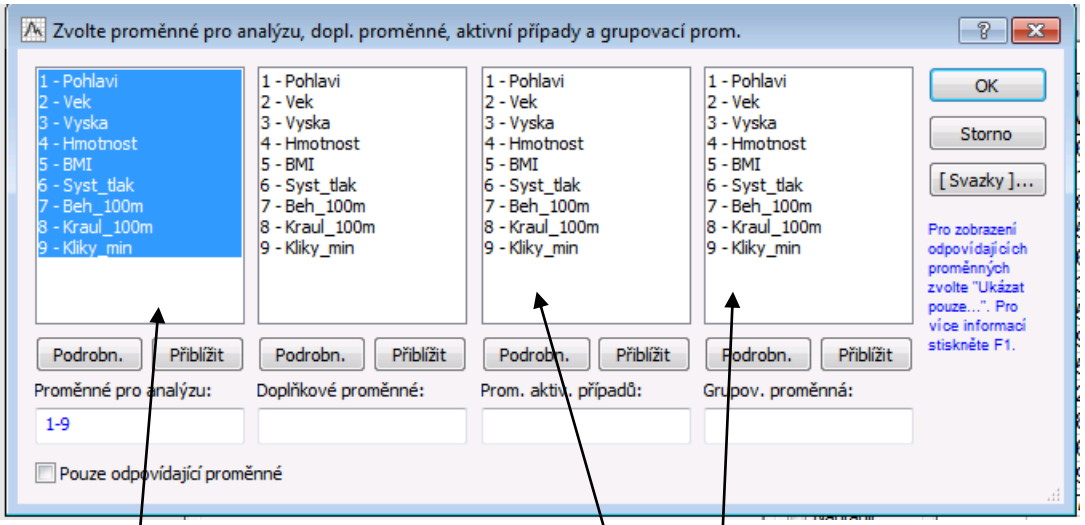
Histogram proměnné č. 9

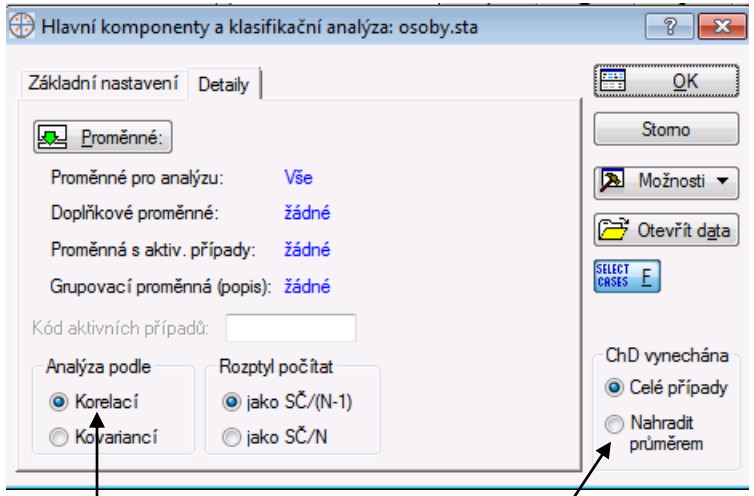
(Případně zájemci si mohou spočíst Gleason-Staelinovu míru redundance jako ověření vhodnosti použití PCA)

Trocha teorie:

Doporučený postup při PCA

- 3. Začneme s PCA v programu STATISTICA *(Statistiky – Vícerozměrné průzkumné techniky – Hlavní komponenty & klasifikační analýza – Proměnné pro analýzu: Všechny)*





Proměnné vstupující do PCA

Proměnné pro případ, že nechceme použít všechny případy, které máme uložené v souboru

Karta Detaily: Analýza podle Korelací (to je to zmiňované použití korelační matice)

Ve cvičení budou chybějící hodnoty = budeme chtít „ChD Nahradit průměrem“!

Trocha teorie:

Doporučený postup při PCA

4. Výběr vhodného počtu hlavních komponent *(Základní výsledky – Vlastní čísla)*

Pořadí vl.č.	Vlastní čísla korelační matice a související statistiky (o Pouze aktiv. proměnné **vl. číslo**	% celk. rozptylu	Kumulativ. vl. číslo	Kumulativ. %
1	4,645591	51,61768	4,645591	51,6177
2	2,334748	25,94165	6,980340	77,5593
3	0,581219	6,45799	7,561559	84,0173
4	0,551397	6,12664	8,112956	90,1440
5	0,383695	4,26328	8,496651	94,4072
6	0,286036	3,17818	8,782687	97,5854
7	0,121726	1,35252	8,904414	98,9379
8	0,094712	1,05235	8,999125	99,9903
9	0,000875	0,00972	9,000000	100,0000

1. Kaiserovo kritérium =

- zvolíme tolik HK, kolik jich má vlastní číslo větší než jedna
- v tomto případě by to byly jen 2 HK

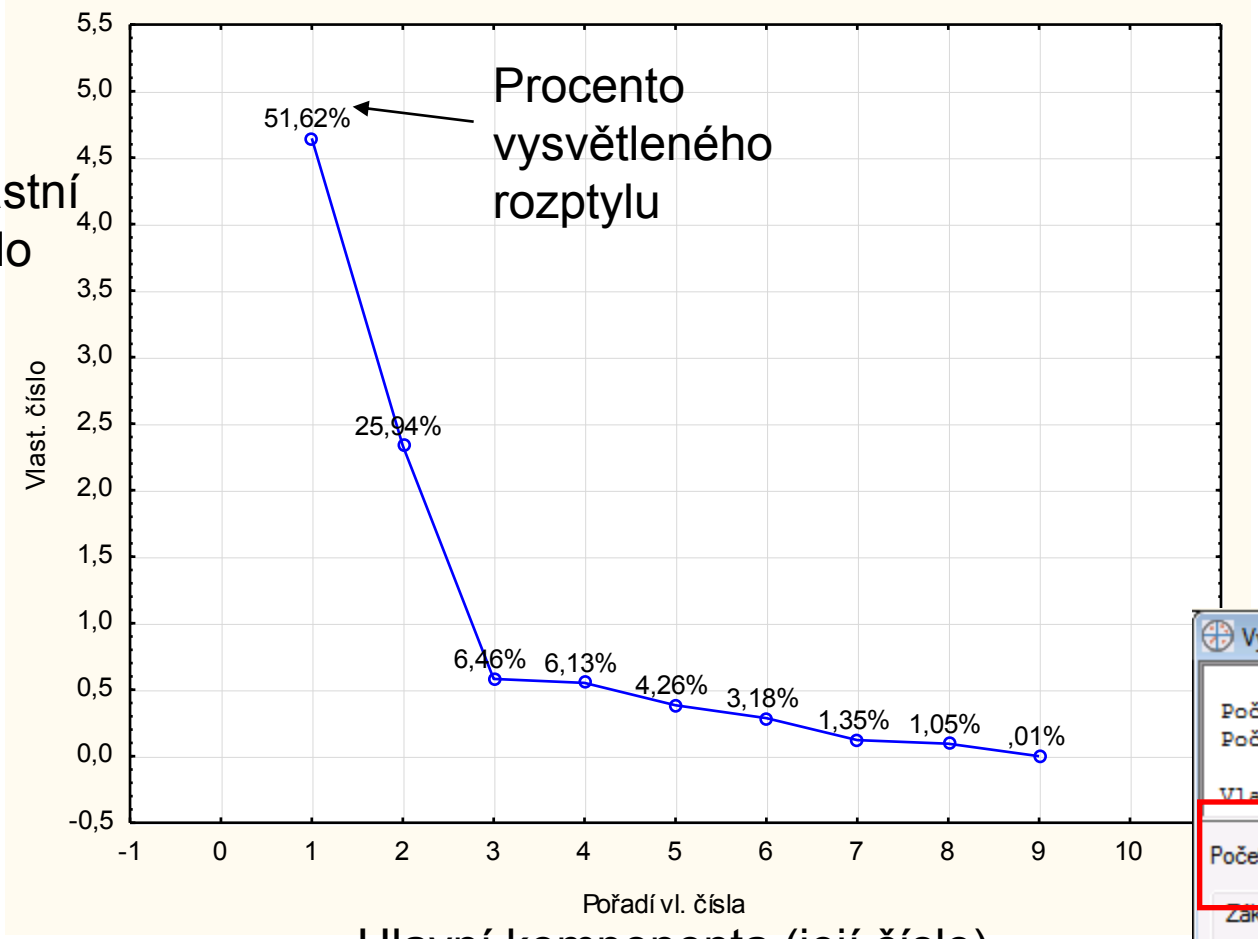
2. Kritérium založené na kumulativním procentu vysvětleného rozptylu =

- vezmeme tolik HK, aby procento vysvětleného rozptylu bylo alespoň 70 %
- opět to vypadá, že stačí 2 HK

Trocha teorie:

Doporučený postup při PCA

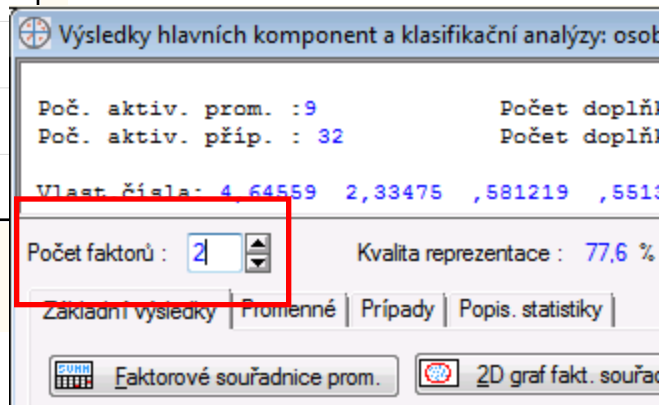
4. Výběr vhodného počtu hlavních komponent *(Základní výsledky – Sutinový graf)*



3. Sutinový graf

- Hledáme zlom v grafu
- Tady by to odpovídalo spíš 3 hlavním komponentám

Závěr: Na základě dvou kritérií zvolíme dvě hlavní komponenty



Trocha teorie:

Doporučený postup při PCA

5. Analýza výsledků ☺ : Zátěže (*Proměnné – Korelace faktorů a proměnných)*

Proměnná	Korelace faktorů a proměnn	
	Faktor 1	Faktor 2
Pohlavi	-0,945770	-0,007999
Vek	0,291324	0,793566
Vyska	0,921114	0,015365
Hmotnost	0,972724	0,018836
BMI	0,957395	0,019544
Syst_tlak	0,689992	0,246178
Beh_100m	0,021066	0,879782
Kraul_100m	-0,222529	0,814635
Kliky_min	0,654936	-0,453558

Co tím zjistíme?

Které původní proměné nejvíc ovlivňují danou hlavní komponentu.

2. hlavní komponenta je pozitivně korelovaná s věkem, rychlostí běhu i plavání. Platí tedy, že čím jsou lidé starší, tím pomaleji běží a plavou.

1. hlavní komponenta je negativně korelovaná s pohlavím a pozitivně korelovaná s výškou, hmotností a BMI. Pohlaví tedy značně determinuje zbylé tři proměnné.

Trocha teorie:

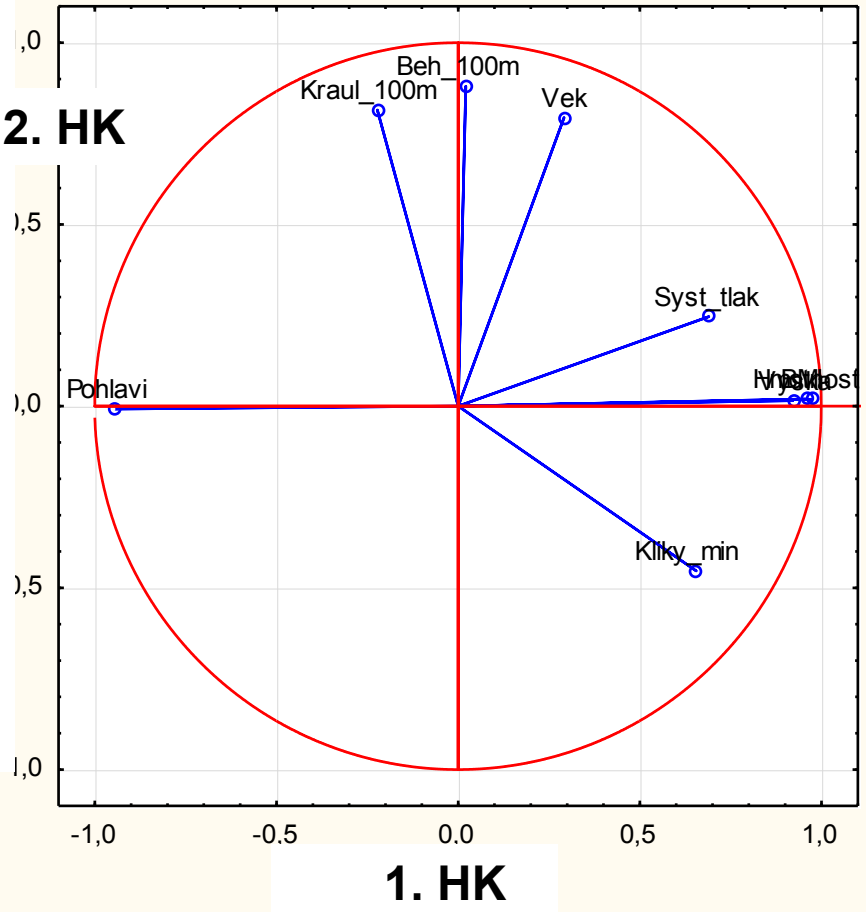
Doporučený postup při PCA

5. Analýza výsledků ☺ : Zátěže v grafu/ 2D graf faktorových souřadnic proměnných
(*Proměnné – 2D graf faktorových souřadnic proměnných)*

• Jsou-li proměnné v grafu v úhlu cca 90 , pak spolu téměř vůbec nekorelují

• Např. počet kliků za minutu a věk.

• Naopak, jsou-li blízko u sebe, korelují spolu pozitivně (výška, hmotnost, BMI) a jsou-li v úhlu kolem 180 , korelují spolu negativně (pohlaví a výška,..).



• Čím blíž je proměnná k obvodu jednotkové kružnice, tím je významnější.

• Např. hodnota systolického tlaku a počet kliků za minutu jsou poměrně nevýznamné).

Závěr: Můžeme zjistit a) které proměnné jsou důležité, b) které spolu nejvíce korelují.

Trocha teorie:

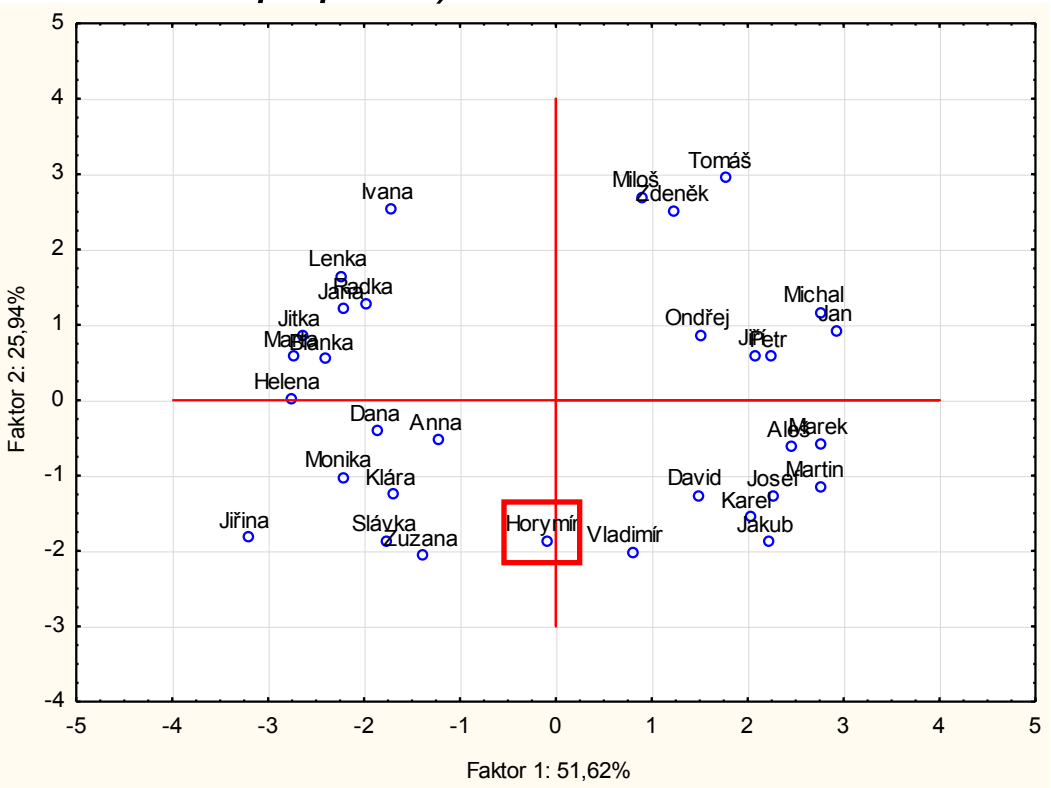
Doporučený postup při PCA

5. Analýza výsledků ☺ : 2D graf faktorových souřadnic případů
(Případy – Faktorové souřadnice případů)

• Hodnoty si lze zobrazit i v tabulce (*Faktorové souřadnice případů*)

• Vidíme, že 1. hlavní komponenta (= osa y) rozlišila muže a ženy.

• 2. hlavní komponenta (osa x) rozlišila účastníky podle věku a jejich fyzické výkonnosti (lze zkontrolovat v původním souboru dat).



Interpretace: Zdá se, že určitě má smysl vytvořit odlišnou hranici pro muže a ženy. Zároveň fyzická výkonnost zjevně souvisí s věkem, takže by mohlo mít smysl stanovit i věkovou hranici. Chceme-li rozlišit více a méně fyzicky zdatné účastníky, lépe nám poslouží běh a kraul než počet kliků (byl méně významný). Teoreticky by se dalo říci, že jedinci „pod osou x“ jsou fyzicky zdatnější (a tudíž vhodnější pro práci u hasičů…?).

Poznámka ke cvičení

- Cvičení bude obsahovat:
  - Krabicový graf proměnných a Maticové grafy
  - Tabulku vlastních čísel a Sutinový graf
  - 2D grafy souřadnic proměnných i případů

- Pozor!
  - Máte ve cvičení použít jednu proměnnou jako doplňkovou
  - Je možné, že vám vyjde, že budete muset použít 3 hlavní komponenty, tzn. je třeba vytvořit grafy pro
    - 1. vs. 2 HK,
    - 1. vs. 3. HK,
    - a 2. vs. 3 HK

  = (bude 3x víc grafů)

- Pro napsání závěru si přečíst požadavky profesora Dobrovolného v pdf zadání (body 5–8)

Zdroje:

- BUDÍKOVÁ, Marie. Snížení dimenze dat metodou hlavních komponent (přednáška). Brno: Masarykova univerzita, 2.5. 2016.

- DOBROVOLNÝ, Petr. Z2069 Statistické metody a zpracování dat II: Vícerozměrné metody (přednáška) Brno: Masarykova univerzita, 2.5. 2016.