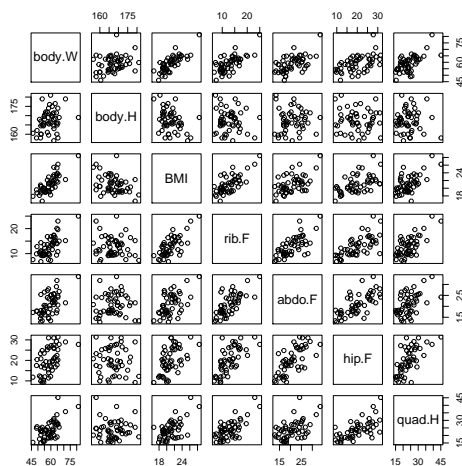


10 Analýza hlavních komponent (PCA)

Příklad 1. V souboru `mlrm-fat.txt` máme k dispozici antropometrická data mladých zdravých dospělých žen (převážně studentek vysokých škol z Brna): tělesnou hmotnost (proměnná `body.W`, v *kg*), tělesnou výšku (proměnná `body.H`, v *cm*), BMI (proměnná `BMI`, v kg/m^2), tloušťku kožní řasy ve výši 10. žebra (proměnná `rib.F`, v *mm*), tloušťku kožní řasy na břicho (proměnná `abdo.F`, v *mm*), tloušťku kožní řasy na boku (proměnná `hip.F`, v *mm*) a tloušťku kožní řasy nad čtyřhlavým svalem stehenním (proměnná `quad.H`, v *mm*).

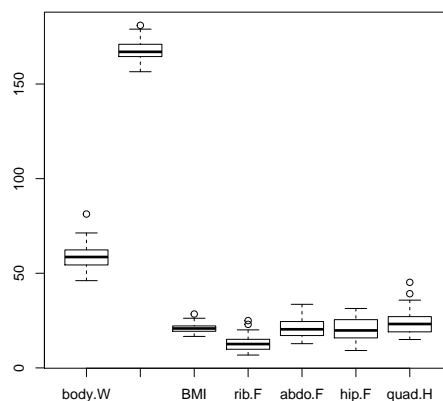
Načteme datový soubor a posoudíme vazby mezi proměnnými pomocí bodových diagramů.

```
fat <- read.table('DATA/mlrm-fat.txt', header=T)
plot(fat)
```



Data znázorníme pomocí krabicových diagramů.

```
boxplot(fat)
```



Podíváme se na korelační matici.

```
R <- cor(fat)
R
##          body.W      body.H      BMI      rib.F      abdo.F
## body.W 1.0000000 0.284939369 0.7899114 0.69037471 0.58028551
## body.H 0.2849394 1.000000000 -0.3588399 -0.08726129 0.03318397
## BMI    0.7899114 -0.358839858 1.0000000 0.72745544 0.53489013
## rib.F  0.6903747 -0.087261293 0.7274554 1.00000000 0.70532553
## abdo.F 0.5802855 0.033183970 0.5348901 0.70532553 1.00000000
## hip.F  0.5327246 -0.053201429 0.5416473 0.68277901 0.69316107
## quad.H 0.6207775 0.007551012 0.6024704 0.68879439 0.50761994
##          hip.F      quad.H
## body.W 0.53272456 0.620777543
## body.H -0.05320143 0.007551012
## BMI    0.54164729 0.602470351
## rib.F  0.68277901 0.688794385
## abdo.F 0.69316107 0.507619937
## hip.F  1.00000000 0.656692703
## quad.H 0.65669270 1.000000000
```

Vidíme, že korelační koeficienty mezi některými proměnnými jsou v absolutní hodnotě dostatečně velké, má tedy smysl přistoupit k analýze hlavních komponent. Provedeme Bartlettův test o úplné nezávislosti proměnných.

```
n <- nrow(fat)
k <- 7 #pocet promennych

( test.stat <- -n*log(det(R))*(1- (2*k+11)/(6*n)) )

## [1] 407.9905

( kvantil <- qchisq(0.95, df=k*(k-1)/2) )

## [1] 32.67057
```

Hodnota testovací statistiky

Kritický obor

Závěr

K provedení analýzy hlavních komponent použijeme funkci `prcomp()`, protože máme proměnné v různých jednotkách, nastavíme argumenty `center=T`, `scale.=T`, abychom pracovali s korelační maticí:

```
fat.PCA <- prcomp(fat, center=T, scale.=T)
fat.PCA

## Standard deviations (1, ..., p=7):
## [1] 2.04761717 1.10336637 0.80357610 0.68875541 0.51374229 0.45148562
## [7] 0.04424491
##
## Rotation (n x k) = (7 x 7):
##          PC1      PC2      PC3      PC4      PC5
## body.W 0.40951111 -0.285871559 -0.508568602 -0.17057416 0.242304902
```

```
## body.H -0.02344427 -0.903889345 -0.038432845 0.01241701 0.003333954
## BMI 0.41202400 0.298452358 -0.486269446 -0.16308369 0.211789814
## rib.F 0.44004864 0.049959964 0.034830865 -0.06621216 -0.569614937
## abdo.F 0.39004710 -0.087292192 0.466719822 -0.55059136 -0.234828665
## hip.F 0.39892312 0.007004983 0.533241456 0.18444630 0.664115894
## quad.H 0.39631811 -0.044974425 -0.005990611 0.77627227 -0.275257507
##
## PC6 PC7
## body.W -0.03769926 0.634592494
## body.H 0.09074296 -0.415400713
## BMI -0.09423121 -0.651439719
## rib.F 0.68829736 0.007037453
## abdo.F -0.51389332 -0.005852338
## hip.F 0.28495337 -0.014001572
## quad.H -0.40302226 0.009004806
```

Ve výpisu vidíme směrodatné odchylky komponent, které souvisí s vlastními čísly korelační matice, konkrétně se jedná o odmocninu vlastních čísel. Můžeme si to ověřit tak, že vypočítáme vlastní čísla korelační matice pomocí funkce `eigen()` a srovnáme je s druhou mocninou směrodatných odchylek, které nám poskytuje funkce `prcomp()`:

```
(vl.cisla <- eigen(R)$values)

## [1] 4.192736088 1.217417336 0.645734546 0.474384014 0.263931143 0.203839262
## [7] 0.001957612

(vl.cisla.pca <- fat.PCA$sdev^2)

## [1] 4.192736088 1.217417336 0.645734546 0.474384014 0.263931143 0.203839262
## [7] 0.001957612
```

Ve výpisu dále vidíme vlastní vektory neboli hlavní komponenty, pokud bychom s nimi dále chtěli pracovat, můžeme si je uložit samostatně:

```
pc <- fat.PCA$rotation
```

Funkce `summary()` vypíše pro jednotlivé komponenty jejich směrodatnou odchylku, podíl vysvětleného rozptylu a kumulativní podíl vysvětleného rozptylu.

```
summary(fat.PCA)

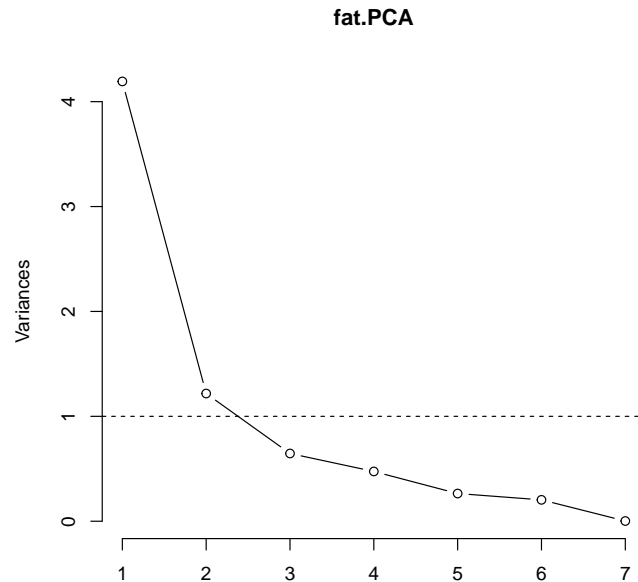
## Importance of components:
##
## Standard deviation      PC1    PC2    PC3    PC4    PC5    PC6    PC7
## Proportion of Variance 0.599 0.1739 0.09225 0.06777 0.0377 0.02912 0.00028
## Cumulative Proportion 0.599 0.7729 0.86513 0.93290 0.9706 0.99972 1.00000
```

Vidíme, že první komponenta vysvětluje% variability, druhá komponenta% variability a třetí vysvětluje% variability. Dohromady vysvětlují % variability.

Počet m hlavních komponent můžeme volit na základě několika pravidel. Pokud bychom požadovali, aby m hlavních komponent vysvětlovalo alespoň 70 % variability, vybrali bychom hlavní komponenty.

Další kritéria si můžeme zobrazit graficky v tzv. sutinovém grafu. Pokud do něj přidáme vodovou čáru ve výšce 1, můžeme tak zhodnotit zároveň i Kaiserovo kritérium:

```
plot(fat.PCA, type='l')
abline(h=1, lty=2)
```



Počet hlavních komponent na základě Kaiserova kritéria:

Počet hlavních komponent na základě zploštění v sutinovém grafu:

Omezíme se na dvě hlavní komponenty.

Abychom mohli první dvě komponenty interpretovat, vypočítáme si korelace pozorování v původních proměnných a v souřadnicích vybraného počtu hlavních komponent.

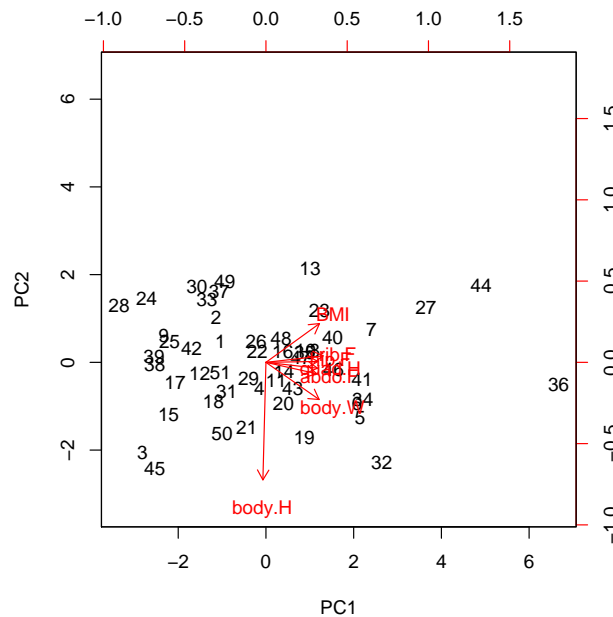
```
fat.in.pc <- fat.PCA$x #pozorovani v souradnicich hlavnich komponentach
cor(fat,fat.in.pc[,1:2])
```

```
##          PC1          PC2
## body.W  0.83852198 -0.315421063
## body.H -0.04800489 -0.997321101
## BMI     0.84366741  0.329302293
## rib.F   0.90105116  0.055124143
## abdo.F  0.79866714 -0.096315268
## hip.F   0.81684182  0.007729063
## quad.H  0.81150777 -0.049623267
```

První komponenta má vysoké korelace s tělesnou vahou, BMI, a tloušťkou kožní řasy na žebrou, bříše, boku i stehně, které byly i v původním souboru mezi sebou vysoce korelované. Druhá komponenta má vysokou korelaci s tělesnou výškou, bude tedy odlišovat sledované jedince na základě výšky.

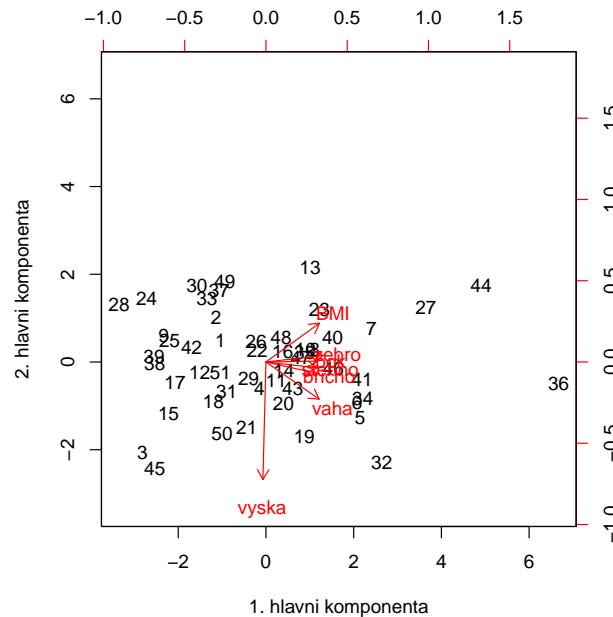
Pozorování si můžeme vykreslit v rovině prvních dvou komponent pomocí funkce `biplot()`, která nám zároveň vykreslí i původní proměnné v rovině prvních dvou komponent.

```
biplot(fat.PCA, scale=0)
```



I z tohoto grafu vidíme, že proměnná tělesná výška není příliš korelovaná s ostatními proměnnými. Pozn.: Pokud by naše pozorování měla nějaká jména/značky, můžeme je v `biplot()` nastavit pomocí argumentu `xlabs=`. Stejně tak lze změnit jména proměnných pomocí argumentu `ylabs=`. Popisky os lze změnit klasickým způsobem.

```
biplot(fat.PCA, scale=0, ylabs=c('vaha', 'vyska', 'BMI', 'zebro', 'bricho', 'bok', 'stehno'),
       xlab='1. hlavni komponenta', ylab='2. hlavni komponenta' )
```



Dále je možné si vybrat dvojici komponent, v kterých chceme vykreslovat, pomocí argumentu `choices=`. Pokud by nás zajímala například 1. a 3. komponenta, nastavili bychom `choices=c(1,3)`.

Podíváme se ještě na reprodukovanou korelační matici a reziduální korelační matici.

```
(R.reproduced <- pc[,1:2] %*% diag(vl.cisla.pca[1:2]) %*% t(pc[,1:2]))

##          body.W      body.H      BMI      rib.F      abdo.F      hip.F
## body.W 0.8026096 0.27432292 0.6035648 0.73816389 0.70007982 0.68250192
## body.H 0.2743229 0.99695385 -0.3689203 -0.09823134 0.05771732 -0.04692076
## BMI    0.6035648 -0.36892029 0.8202147 0.77834000 0.64209260 0.69168802
## rib.F  0.7381639 -0.09823134 0.7783400 0.81493186 0.71433065 0.73644233
## abdo.F 0.7000798 0.05771732 0.6420926 0.71433065 0.64714583 0.65164029
## hip.F  0.6825019 -0.04692076 0.6916880 0.73644233 0.65164029 0.66729030
## quad.H 0.6961193 0.01053399 0.6683016 0.72847457 0.65290406 0.66248994
##          quad.H
## body.W 0.69611933
## body.H 0.01053399
## BMI    0.66830160
## rib.F  0.72847457
## abdo.F 0.65290406
## hip.F  0.66248994
## quad.H 0.66100733

(R.residual <- R - R.reproduced)

##          body.W      body.H      BMI      rib.F      abdo.F
## body.W 0.19739044 0.010616447 0.18634658 -0.047789179 -0.119794303
## body.H 0.01061645 0.003046151 0.01008043 0.010970044 -0.024533348
## BMI    0.18634658 0.010080433 0.17978530 -0.050884566 -0.107202467
## rib.F  -0.04778918 0.010970044 -0.05088457 0.185068141 -0.009005123
## abdo.F -0.11979430 -0.024533348 -0.10720247 -0.009005123 0.352854172
## hip.F  -0.14977736 -0.006280666 -0.15004073 -0.053663318 0.041520776
## quad.H -0.07534178 -0.002982975 -0.06583125 -0.039680188 -0.145284128
##          hip.F      quad.H
## body.W -0.149777355 -0.075341783
## body.H -0.006280666 -0.002982975
## BMI    -0.150040730 -0.065831249
## rib.F  -0.053663318 -0.039680188
## abdo.F 0.041520776 -0.145284128
## hip.F  0.332709699 -0.005797239
## quad.H -0.005797239 0.338992674
```

V reziduální korelační matici vidíme hodnoty v absolutní hodnotě menší než 0.2.