12 Shluková analýza

Příklad 1. V souboru Howell.csv máme k dispozici kraniometrické údaje z různých populací. Nás zajímají muži (kategorie M proměnné Sex) z populací (proměnná Population) - AUSTRALI, BUSHMAN, ZALAVAR, BERG, ZULU, NORSE, EGYPT, EASTER I, PERU, BURIAT. Konkrétně chceme na základě průměrných hodnot následujících proměnných v každé populaci identifikovat pomocí metod shlukové analýzy, které populace jsou si v těchto rozměrech podobné:

- XFB - maximální transversální šířka čela,
- ZYB - bizygomatická šířka,
- BPL - délka obličejové části lebky,
- NPH - výška horní části obličejového skeletu,
- NLH - výška nosu,
- OBH - výška očnice,
- OBB - šířka očnice,
- NLB - šířka nosu,
- ZMB - zygomatikomaxilární šířka.

Načteme datový soubor. Protože v databázi jsou chybějící pozorování kódovány jako 0, je potřeba při načítání zadat, aby se 0 braly jako NA. Vybereme pozorování a proměnné, které nás zajímají, a zbavíme se nyní prázdných kategorií proměnné Population.

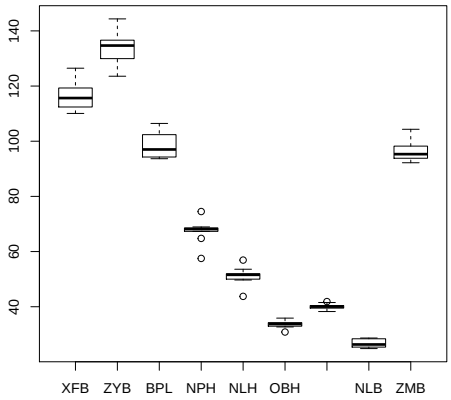
```
cranio <- read.csv('DATA/Howell.csv',header=T, na.strings='0')
our.pop <- c('AUSTRALI', 'BUSHMAN', 'ZALAVAR', 'BERG', 'ZULU', 'NORSE', 'EGYPT',
             'EASTER I', 'PERU', 'BURIAT')
data <- cranio[cranio$Sex == 'M' & cranio$Population %in% our.pop,
               c('Population', 'XFB', 'ZYB', 'BPL', 'NPH', 'NLH', 'OBH', 'OBB', 'NLB', 'ZMB')]
data$Population <- factor(data$Population)
```

Nyní je potřeba pro každou populaci vypočítat průměrné hodnoty sledovaných proměnných.

```
m.all <- c()
for (pop in levels(data$Population)){
  m.pop <- colMeans(data[data$Population==pop,-1])
  m.all <- rbind(m.all, m.pop)
}

howells <- data.frame(Population=levels(data$Population), m.all, row.names=NULL)
```

Vykreslíme si krabicový diagram všech proměnných.

```
boxplot(howells[,-1])
```

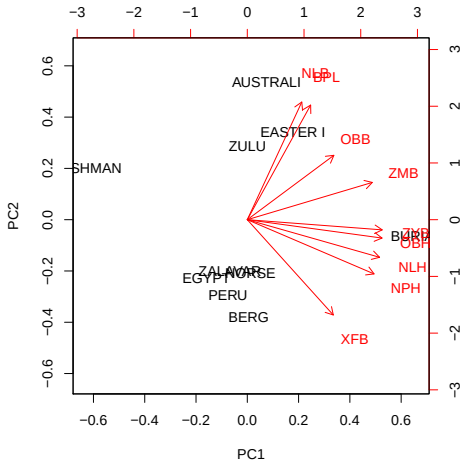


Kvůli rozdílné variabilitě budeme pracovat se standardizovanými daty, ke standardizaci lze použít funkci scale().

```
howells.scaled <- scale(howells[,-1])
```

Vykreslíme si data v rovině prvních dvou hlavních komponent.

```
cluster.pca <- prcomp(howells.scaled)
biplot(cluster.pca, xlabs=howells$Population)
```



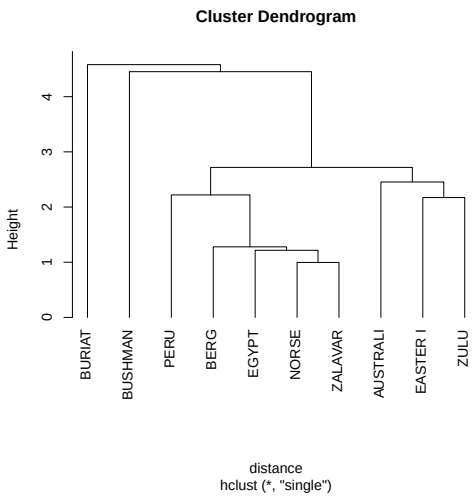
Vypadá to, že se populace shlukují do čtyř skupin - Australi, Easter Island a Zulu; Zalavar, Norse, Egypt, Peru a Berg; populace Buriat a Bushman jsou obě daleko od ostatních a budou tedy zřejmě každá tvořit samostatnou skupinu.

Nejprve vyzkoušíme metody hierarchického slučování, k tomu slouží funkce hclust(). Nejprve je potřeba vypočítat matici vzdáleností. Protože pracujeme se spojitými proměnnými, použijeme euklidovské vzdálenosti.

```
distance <- dist(howells.scaled, method='euclidean')
```

Použijeme metodu nejbližšího souseda (anglicky single linkage) a vykreslíme si tzv. dendogram - diagram, který vykresluje, jak probíhalo slučování.

```
method1 <- hclust(distance, method='single')
plot(method1, hang=-1, labels=howells$Population)
```



Vidíme, že nejprve byly sloučeny populace Zalavar a Norse, poté k nim přibyla populace Egypt a Berg, poté byly společně sloučeny populace Zulu a Easter Island, v dalším kroku byla k prvnímu shluku přidána populace z Peru, poté k druhému shluku přibyla populace Australi. Další krok tyto dva shluky sloučil a teprve poté k nim přibyly populace Bushman a Buriat.

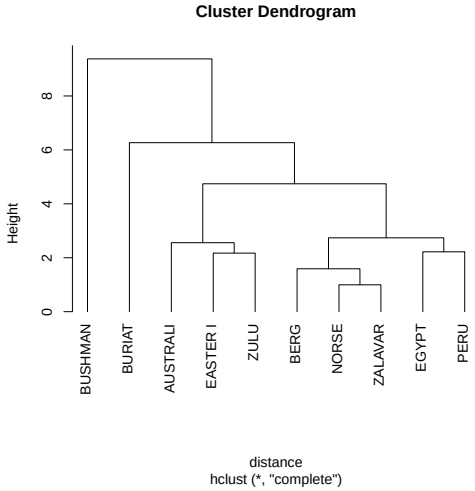
Pokud chceme populace rozdělit do 4 shluků, znamená to, že ve vhodném místě dendogram "uřízneme"- můžeme použít funkci cutree(), když si její výstup vypíšeme společně se jmény populací, uvidíme, jak byly zařazeny. Ve větším, nepřehledném dendogramu bychom to vizuálně nemuseli zvládnout.

```
clusters1 <- cutree(method1, 4)
data.frame(howells$Population, clusters1)

##    howells.Population clusters1
## 1            AUSTRALI         1
## 2                BERG         2
## 3              BURIAT         3
## 4             BUSHMAN         4
## 5            EASTER I         1
## 6               EGYPT         2
## 7               NORSE         2
## 8                PERU         2
## 9             ZALAVAR         2
## 10               ZULU         1
```

První shluk obsahuje populace Australi, Zulu a Easter Island, druhý shluk tvoří pouze populace Bushman, třetí shluk obsahuje populace Zalavar, Berg, Norse, Egypt a Peru, poslední shluk je samostatně populace Buriat.

Vyzkoušíme nyní metodu nevzdálenějšího souseda (complete linkage)

```
method2 <- hclust(distance, method='complete')
plot(method2, hang=-1, labels=howells$Population)
```

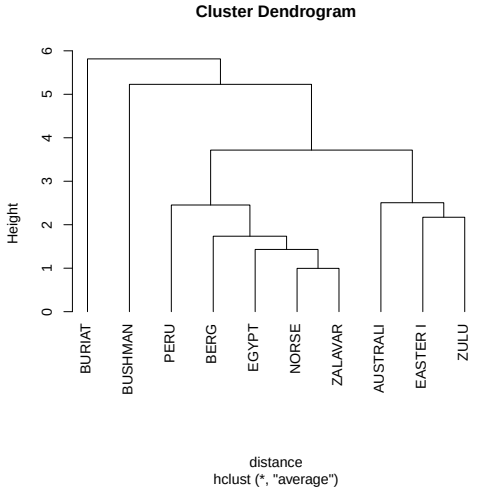


Shlukování probíhalo v jiném pořadí, ale pokud si populace rozdělíme do čtyř shluků, jsou stejné jako u předchozí metody. To ale není obecné pravidlo, jen v našem konrkétním případě to tak vychází.

```
clusters2 <- cutree(method2, 4)
data.frame(howells$Population, clusters2)

##    howells.Population clusters2
## 1            AUSTRALI         1
## 2                BERG         2
## 3              BURIAT         3
## 4             BUSHMAN         4
## 5            EASTER I         1
## 6               EGYPT         2
## 7               NORSE         2
## 8                PERU         2
## 9             ZALAVAR         2
## 10               ZULU         1
```

Další metodou, kterou lze při shlukování použít, je metoda průměrné vazby (average linkage).

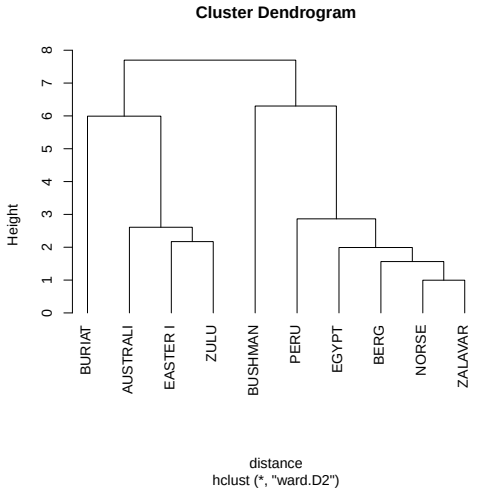
```
method3<- hclust(distance, method='average')
plot(method3, hang=-1, labels=howells$Population)
```




```
clusters3 <- cutree(method3, 4)
data.frame(howells$Population, clusters3)

##    howells.Population clusters3
## 1            AUSTRALI         1
## 2                BERG         2
## 3              BURIAT         3
## 4             BUSHMAN         4
## 5            EASTER I         1
## 6               EGYPT         2
## 7               NORSE         2
## 8                PERU         2
## 9             ZALAVAR         2
## 10               ZULU         1
```

I zde v našem případě vychází stejné 4 shluky. Poslední hierarchickou metodou, kterou vyzkoušíme, je Wardova metoda.

```
method4 <- hclust(distance, method='ward.D2')
plot(method4, hang=-1, labels=howells$Population)
```



Zde vidímě, že při shlukování postupovala docela odlišně od jiných metod. Rozdělení do 4 shluků vypadá stejně.

```
clusters4 <- cutree(method4, 4)
data.frame(howells$Population, clusters4)

##    howells.Population clusters4
## 1            AUSTRALI         1
## 2                BERG         2
## 3              BURIAT         3
## 4             BUSHMAN         4
## 5            EASTER I         1
## 6               EGYPT         2
## 7               NORSE         2
## 8                PERU         2
## 9             ZALAVAR         2
## 10               ZULU         1
```

Pokud bychom ale chtěli jen 2 shluky, metoda nejbližšího souseda a metoda průměrné vazby by nechaly samostatně populaci Buriat a ostatní sloučila; metoda nejvzdálenějšího souseda by nechala samostatně populaci Bushman a ostatní sloučila; Wardova metoda by v jednom shluku nechala Easter Island, Zulu, Australi a Buriat a v druhém Norse, Zalavar, Berg, Egypt, Peru a Bushman.

Porovnejme nyní hierarchické metody pomocí kofenetického koeficientu korelace. Nejprve pro každou metodu vypočítáme kofenetické vzdálenosti pomocí funkce cophenetic(), a poté korelace těchto vzdáleností s původními.

```
coph1 <- cophenetic(method1)
cor(distance, coph1)

## [1] 0.8763084

coph2 <- cophenetic(method2)
cor(distance, coph2)

## [1] 0.8359245

coph3 <- cophenetic(method3)
cor(distance, coph3)

## [1] 0.8944812

coph4 <- cophenetic(method4)
cor(distance, coph4)

## [1] 0.7075275
```

Hodnota kofenetického koeficientu korelace pro metodu nejbližšího souseda: .....
Hodnota kofenetického koeficientu korelace pro metodu nejvzdálenějšího souseda: .....
Hodnota kofenetického koeficientu korelace pro metodu průměrné vazby: .....
Hodnota kofenetického koeficientu korelace pro Wardovu metodu: .....
Závěr: .....

Vyzkoušejme nyní nehierarchické shlukování, konkrétně metodu k-průměrů. Ta potřebuje dopředu specifikovat, kolik chceme shluků.

```
method5 <- kmeans(howells.scaled, 4)
data.frame(howells$Population, method5$cluster)

##    howells.Population method5.cluster
## 1            AUSTRALI               4
## 2                BERG               3
## 3              BURIAT               1
## 4             BUSHMAN               2
## 5            EASTER I               4
## 6               EGYPT               3
## 7               NORSE               3
## 8                PERU               3
## 9             ZALAVAR               3
## 10               ZULU               4
```

Metoda k-průmerů zařadila do prvního shluku populaci Bushman, do druhého shluku Australi, Zulu a Easter Island, třetí shluk tvoří populace Buriat a čtvrtý shluk Zalavar, Berg, Norse, Egypt a Peru. Pokud si přiřazení do clusterů přidáme jako nový sloupec k datové matici, můžeme se pomocí analýzy rozptylu podívat, které proměnné mají vliv na zařazení do shluků. ANOVA potřebuje zařazení do shluků jako kategoriální proměnnou neboli faktor, při vytváření sloupce tedy použijeme tedy as.factor().

```
howells$kmeans.cluster <- as.factor(method5$cluster)
anova( aov(XFB ~ kmeans.cluster, data=howells) )

## Analysis of Variance Table
##
## Response: XFB
##                Df  Sum Sq Mean Sq F value  Pr(>F)
## kmeans.cluster  3 206.282  68.761  5.5171 0.03684 *
## Residuals       6  74.779  12.463
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova( aov(ZYB ~ kmeans.cluster, data=howells) )

## Analysis of Variance Table
##
## Response: ZYB
##                Df  Sum Sq Mean Sq F value Pr(>F)
## kmeans.cluster  3 218.325  72.775  7.3347 0.0197 *
## Residuals       6  59.532   9.922
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova( aov(BPL ~ kmeans.cluster, data=howells) )

## Analysis of Variance Table
##
## Response: BPL
##                Df  Sum Sq Mean Sq F value   Pr(>F)
## kmeans.cluster  3 180.624  60.208  19.014 0.001817 **
## Residuals       6  18.999   3.167
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova( aov(NPH ~ kmeans.cluster, data=howells) )
```

```
## Analysis of Variance Table
##
## Response: NPH
##                Df  Sum Sq Mean Sq F value    Pr(>F)
## kmeans.cluster  3 153.133  51.044  36.829 0.0002931 ***
## Residuals       6   8.316   1.386
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova( aov(NLH ~ kmeans.cluster, data=howells) )
```

```
## Analysis of Variance Table
##
## Response: NLH
##                Df Sum Sq Mean Sq F value   Pr(>F)
## kmeans.cluster  3 88.253 29.4178  15.989 0.002878 **
## Residuals       6 11.039  1.8398
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova( aov(OBH ~ kmeans.cluster, data=howells) )
```

```
## Analysis of Variance Table
##
## Response: OBH
##                Df Sum Sq Mean Sq F value   Pr(>F)
## kmeans.cluster  3 13.036  4.3452  13.095 0.004828 **
## Residuals       6  1.991  0.3318
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova( aov(OBB ~ kmeans.cluster, data=howells) )
```

```
## Analysis of Variance Table
##
## Response: OBB
##                Df Sum Sq Mean Sq F value Pr(>F)
## kmeans.cluster  3 4.8970 1.63234  1.9656 0.2206
## Residuals       6 4.9826 0.83044
```

```r
anova( aov(NLB ~ kmeans.cluster, data=howells) )
```

```
## Analysis of Variance Table
##
## Response: NLB
##                Df  Sum Sq Mean Sq F value    Pr(>F)
## kmeans.cluster  3 21.3662  7.1221  73.443 4.035e-05 ***
## Residuals       6  0.5818  0.0970
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

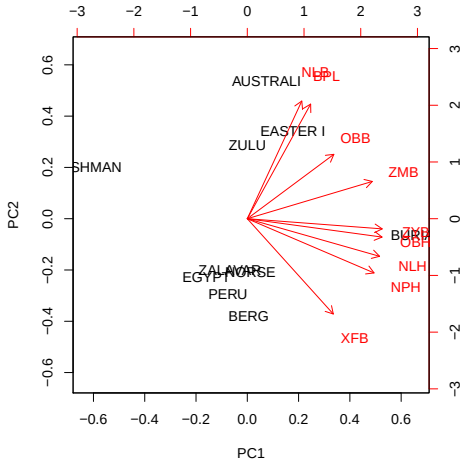
```r
anova( aov(ZMB ~ kmeans.cluster, data=howells) )
```

```
## Analysis of Variance Table
##
## Response: ZMB
##                Df  Sum Sq Mean Sq F value   Pr(>F)
## kmeans.cluster  3 100.473  33.491  16.889 0.002491 **
## Residuals       6  11.898   1.983
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Na hladině významnosti 0.05 není významná pouze proměnná OBB. Největší vliv budou mít proměnné NPH a NLB.

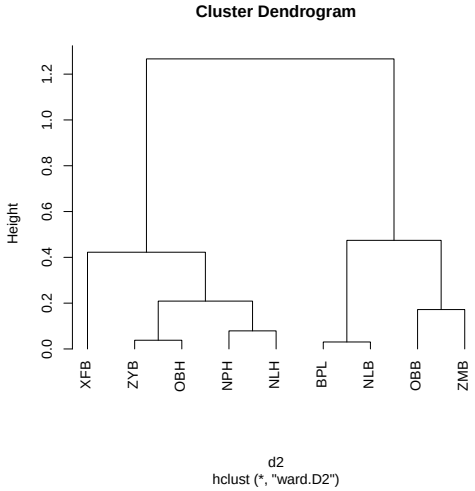
Provedeme nyní shlukovou analýzu pro proměnné. K tomu využijeme první dvě sestavené hlavní komponenty.

```
biplot(cluster.pca, xlabs=howells$Population)
```



V našem případě jsou proměnné blízko, přesto zkusíme proměnné rozdělit do dvou shluků. Nejprve vypočítáme vzdálenosti proměnných v souřadnicích prvních dvou komponent, poté provedeme shlukovou analýzu. Použijeme Wardovu metodu.

```
d2 <- dist(cluster.pca$rotation[,1:2])

m.variables <- hclust(d2, method='ward.D2')
plot(m.variables, hang=-1)
```

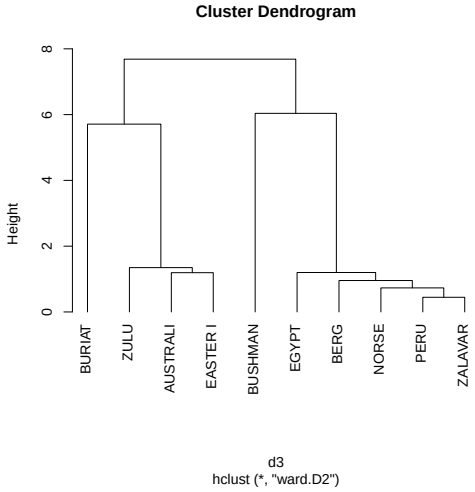



```
cutree(m.variables, 2)

## XFB ZYB BPL NPH NLH OBH OBB NLB ZMB
##   1   1   2   1   1   1   2   2   2
```

Metoda do prvního shluku zahrnula XFB, ZYB, NPH, NLH, OBH, a do druhého BPL, OBB, NLB, ZMB.

Shlukovou metodu lze provést i pomocí hlavních komponent. Vypočítáme vzdálenosti populací v prvních několika hlavních komponentách, a provedeme shlukování Wardovou metodou. V tomto případě zvolíme dvě hlavní komponenty - vysvětlují více než 80 % variability a také podle Kaiserova kritéria jsou dvě komponenty dostatečné. Opět při dělení na 4 shluky získáváme stejné dělení.

```
d3 <- dist(cluster.pca$x[,1:2])

m.pca <- hclust(d3, method='ward.D2')
plot(m.pca, hang=-1, labels=howells$Population)
```




```
clusters.pca <- cutree(method4, 4)
data.frame(howells$Population, clusters.pca)
```

```
##    howells.Population clusters.pca
## 1            AUSTRALI            1
## 2                BERG            2
## 3              BURIAT            3
## 4             BUSHMAN            4
## 5            EASTER I            1
## 6               EGYPT            2
## 7               NORSE            2
## 8                PERU            2
## 9             ZALAVAR            2
## 10               ZULU            1
```