Cvičení 09 - PCA

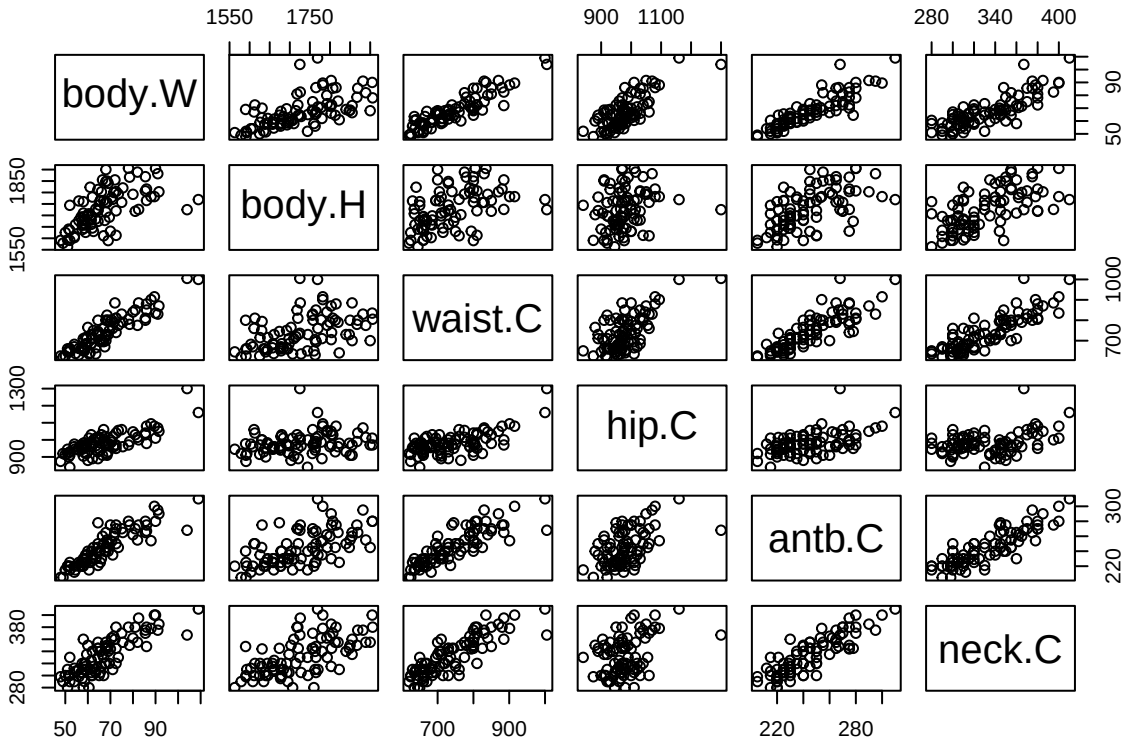
Zdeňka Geršlová

2023-04-11

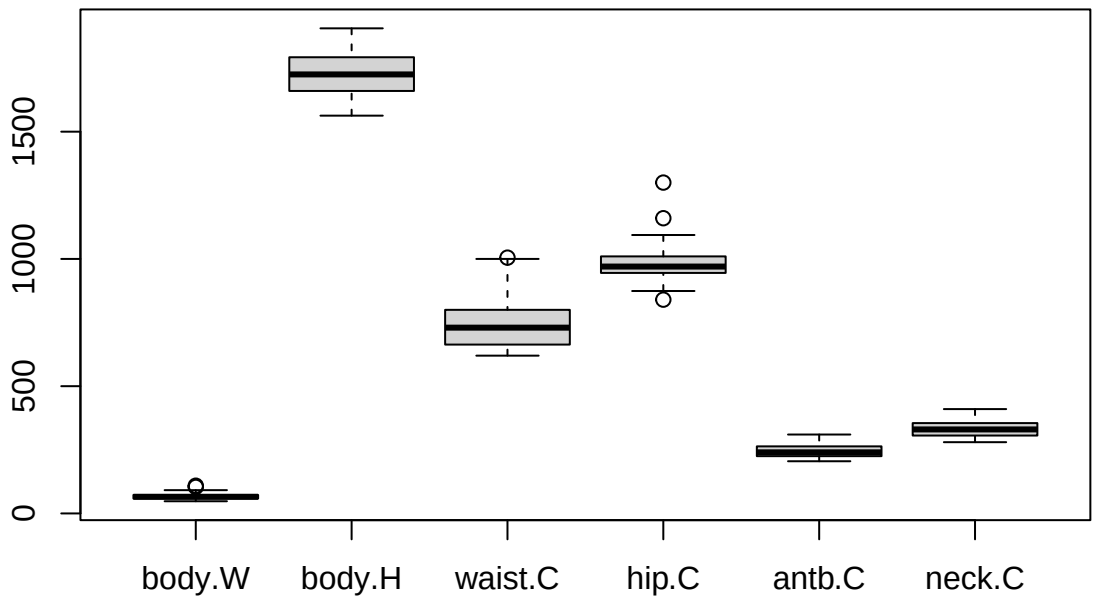
Analýza hlavních komponent (PCA)

Datový soubor cneck.txt obsahuje antropometrické údaje mladých dospělých lidí (převážně studentů VŠ z Brna a Ostravy). U jedinců známe následující hodnoty: identifikační číslo pozorování (id), pohlaví (sex), tělesná hmotnost (body.W, v kg), tělesná výška (body.H, v mm), obvod pasu (waist.C, v mm), obvod boků (hip.C, v mm), obvod předloktí (antb.C, v mm) a obvod krku (neck.C, v mm). Analyzujte spojité proměnné pomocí metody hlavních komponent:

1. Prozkoumejte závislost mezi spojitými veličinami pomocí dvourozměrného bodového diagramu. Pomocí funkce plot()



2. Vykreslete krabicové diagramy pro tyto proměnné. Pouze ilustrační - pro lepší orientaci bychom museli nejprve standardizovat data nebo vykreslit pro každou proměnnou zvlášť.



3. Vypočítejte korelační matici. Otestujte hypotézu o úplné nezávislosti proměnných.

```
##            body.W    body.H   waist.C     hip.C    antb.C    neck.C
## body.W  1.0000000 0.6086383 0.9047087 0.7604090 0.8810742 0.8235417
## body.H  0.6086383 1.0000000 0.4591687 0.2303759 0.5851208 0.6223121
## waist.C 0.9047087 0.4591687 1.0000000 0.6539080 0.8520787 0.8494347
## hip.C   0.7604090 0.2303759 0.6539080 1.0000000 0.5251877 0.3963821
## antb.C  0.8810742 0.5851208 0.8520787 0.5251877 1.0000000 0.8597562
## neck.C  0.8235417 0.6223121 0.8494347 0.3963821 0.8597562 1.0000000
```

```
## $chisq
## [1] 597.0546
##
## $p.value
## [1] 1.499778e-117
##
## $df
## [1] 15
```

4. Proveďte analýzu hlavních komponent na základě korelační matice. Na 1. řádku jsou uvedeny odmocniny z vlastních čísel výběrové korelační matice R (odmocniny z rozptylu).

```
## Standard deviations (1, .., p=6):
## [1] 2.1027616 0.9252278 0.6800690 0.3736955 0.2869147 0.1946411
##
## Rotation (n x k) = (6 x 6):
```

```
##                PC1         PC2        PC3         PC4           PC5        PC6
## body.W  -0.4637988  0.13093254  0.1075173 -0.04223989 -0.0208601720 -0.8683133
## body.H  -0.3177807 -0.63299118  0.6691236  0.07729945 -0.1416139744  0.1567850
## waist.C -0.4443724  0.15285395 -0.3101131  0.38122322 -0.6993939281  0.2202628
## hip.C   -0.3269653  0.69349461  0.4767675  0.03906470  0.2730634504  0.3297907
## antb.C  -0.4423606 -0.09399514 -0.2581029 -0.82212632 -0.0006748173  0.2301581
## neck.C  -0.4287774 -0.26276348 -0.3880969  0.41169371  0.6448211887  0.1058305
```

5. Zjistete podíl variability a kumulativní podíl variability pro jednotlivé komponenty.

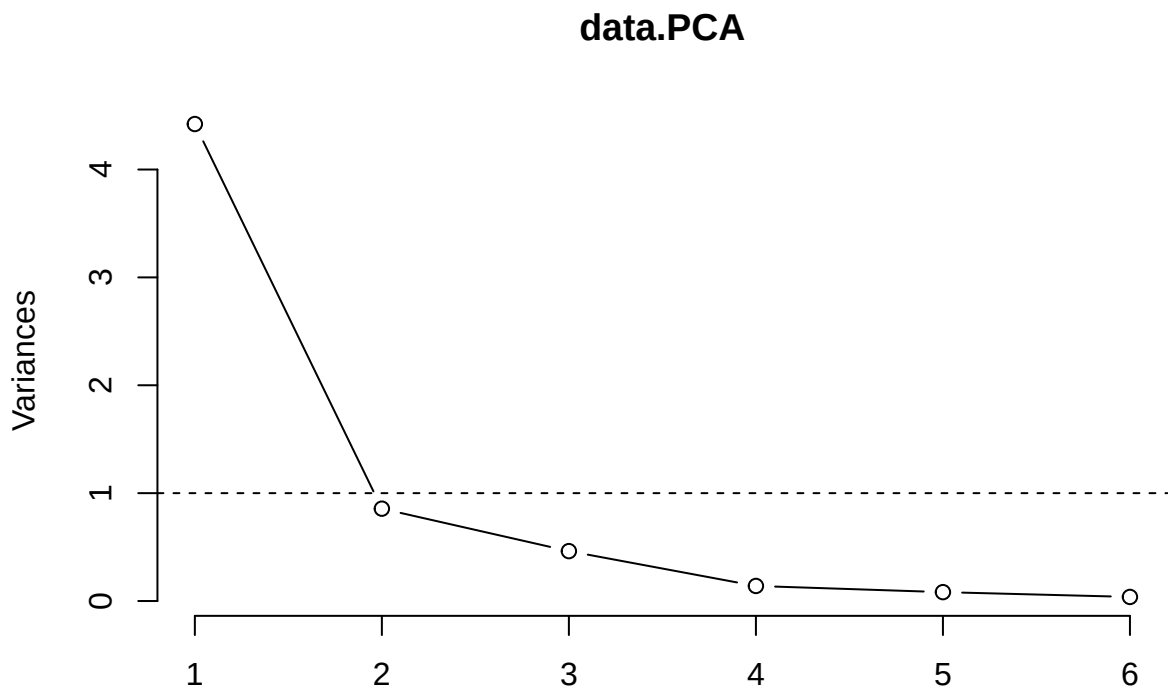
```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6
## Standard deviation     2.1028 0.9252 0.68007 0.37370 0.28691 0.19464
## Proportion of Variance 0.7369 0.1427 0.07708 0.02327 0.01372 0.00631
## Cumulative Proportion  0.7369 0.8796 0.95669 0.97997 0.99369 1.00000
```

6. Počty komponent Kolik komponent by vybralo Kaiserovo kritérium? Kaiserovo kritérium - vybíráme komponenty, jejichž příslušná vl. čísla jsou větší než 1.

```
data.PCA$sdev ^ 2
```

```
## [1] 4.42160614 0.85604641 0.46249388 0.13964833 0.08232006 0.03788517
```

Kolik by jich bylo vybráno podle zploštění sutinového grafu?



Kolik by jich bylo vybráno, pokud bychom požadovali vysvětlení alespoň 80 % variability? Při další práci se omezte na tento počet.

```r
data.in.pc <- data.PCA$x
# komponentni skore = pozorovani v souradnicich hlavnich komponent

# Dale spocteme korelaci techto pozorovani pro prvni dve komponenty s datovym
# souborem. Opet vyuzijeme funkci cor().
cor(data, data.in.pc[,1:2])
```

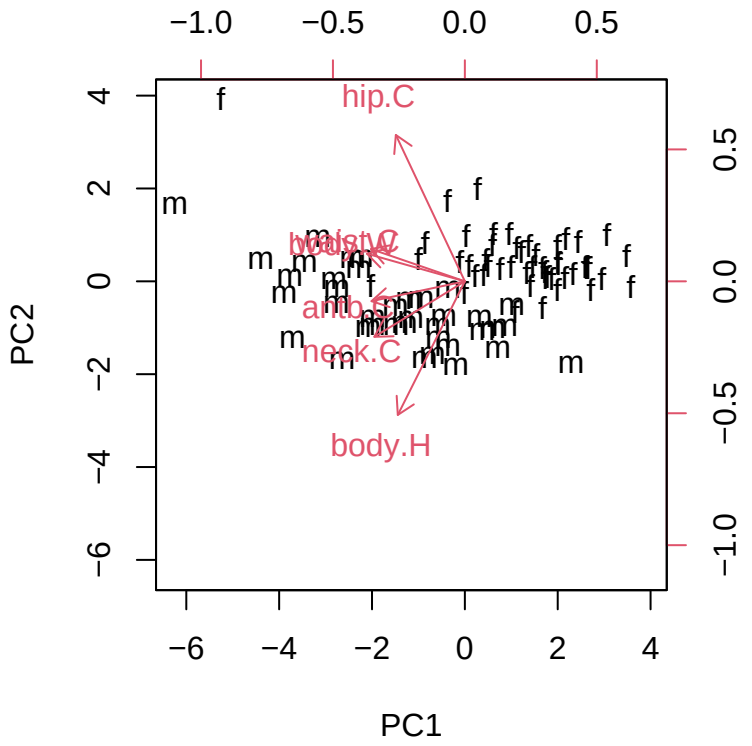
7. Podívejte se na korelace původních proměnných s těmito komponentami.

```
##                PC1         PC2
## body.W  -0.9752584  0.12114242
## body.H  -0.6682169 -0.58566101
## waist.C -0.9344093  0.14142472
## hip.C   -0.6875301  0.64164047
## antb.C  -0.9301788 -0.08696692
## neck.C  -0.9016165 -0.24311606
```

První komponenta je výrazně negativně korelovaná se všemi proměnnými, nejvíce s tělesnou hmotností a obvody pasu, předloktí a krku. Druhá komponenta je výrazně pozitivně korelovaná s obvodem boků a výrazně negativně korelovaná s tělesnou výškou.

8. Vykreslete pozorování a proměnné v rovině prvních dvou komponent. Místo čísel označte pozorování pohlavím jedince (stačí značka f nebo m z proměnné sex).

```r
# Pro lepsi nahled na celou situaci si vykrelime i pozorovani a promenne v rovine
# prvnich dvou hlavnich komponent, tzv. biplot. Argumentem stejnojmene funkce je
# list obsahujici hlavni komponenty a parametr scale, kterym graf skalujeme.
# Nastavime jej na nulu, pouzivame-li korelacni matici pro vypocet komponent.
# Rovnez nastavime popisky pozorovani pomoci xlabs. Chceme aby se pozorovani
# rozlisila podle pohlavi, vyuzijeme tedy faktoru sex v puvodni datove tabulce
# cneck.
# cim "vice jsou sipky rovnobeznejsi", tim vetsi ma dana promenna korelaci s PC
biplot(data.PCA, scale=0, xlabs=cneck$sex)
```



9. Vypočítejte reprodukovanou korelační matici a reziduální korelační matici. Pozorujete vysoké nebo nízké reziduální hodnoty? Nezapomeňte nastavit správný počet použitých hlavních komponent.

```
R.reproduced
```

```
##            body.W     body.H   waist.C      hip.C    antb.C    neck.C
## body.W  0.9658044 0.58073578 0.9284230 0.74824936 0.8966293 0.8498574
## body.H  0.5807358 0.78951270 0.5415612 0.08363545 0.6724944 0.7448590
## waist.C 0.9284230 0.54156116 0.8931216 0.73317831 0.8568685 0.8080962
## hip.C   0.7482494 0.08363545 0.7331783 0.88440012 0.5837245 0.4638954
## antb.C  0.8966293 0.67249440 0.8568685 0.58372445 0.8727959 0.8598077
## neck.C  0.8498574 0.74485904 0.8080962 0.46389539 0.8598077 0.8720178
```

```
R.residual
```

```
##              body.W      body.H     waist.C       hip.C        antb.C
## body.W   0.03419561  0.02790250 -0.02371425  0.01215960 -1.555512e-02
## body.H   0.02790250  0.21048730 -0.08239248  0.14674044 -8.737363e-02
## waist.C -0.02371425 -0.08239248  0.10687838 -0.07927027 -4.789730e-03
## hip.C    0.01215960  0.14674044 -0.07927027  0.11559988 -5.853672e-02
## antb.C  -0.01555512 -0.08737363 -0.00478973 -0.05853672  1.272041e-01
## neck.C  -0.02631573 -0.12254698  0.04133845 -0.06751329 -5.147386e-05
##                neck.C
## body.W  -2.631573e-02
## body.H  -1.225470e-01
## waist.C  4.133845e-02
```

```
## hip.C   -6.751329e-02
## antb.C  -5.147386e-05
## neck.C   1.279822e-01
```

Antropologický závěr: Analýza hlavních komponent sady tělesných rozměrů u dospělých osob smíchaného pohlaví poskytla dvě první hlavní komponenty, z nichž první je sycena především hmotností a obvodovými rozměry, druhá zejména výškou postavy. Na grafu prvních dvou hlavních komponent s odlišením značek pro každé pohlaví vidíme, že se shluky mužů a žen téměř úplně oddělily od sebe, přičemž hranice je diagonální – vlevo dole jsou muži (což souvisí s větší výškou postavy a obvodů krku a předloktí), zatímco ženy jsou vpravo nahoře (což souvisí s nižší výškou postavy a vyšším obvodem boků). Vidíme tedy, že dimorfismus v těchto tělesných rozměrech je u člověka natolik výrazný, že už kombinace několika z nich vede k odlišení pohlaví ordinační metodou, aniž bychom údaj o pohlaví vůbec použili v analýze.