Multivarijatne metode

Kratak pregled i primjeri

Multivarijatne metode - metode za analizu setova podataka koji se sastoje od više (zavisnih i/ili nezavisnih) varijabli

Linearni modeli sa više prediktora (obrađeni na ranijim predavanjima) se uglavnom ubrajaju u „multivarijatne" metode (spominjali smo i multikolinearnost)

Biološki, ovakvi setovi podataka imaju puno smisla (analizirana svojstva su često povezana i utječu jedno na drugo) no često zahtijevaju prilagođene metode za vizualizaciju (ne znamo prikazati više od 3 dimenzije odjednom) i analizu

Tehnički, prikupljanje podataka o više varijabli odjednom je efikasno, a uz modernu tehnologiju i lako (HPLC, genetika, sustavi za fenotipizaciju...)

Problemi multivarijatnih analiza

Problem velikog broja varijabli – gotovo sve metode koje spominjemo zahtijevaju 5-10 puta više opažanja (jedinki) nego varijabli!!

Problem potpunih observacija – većina metoda ne radi ako neki podatci nedostaju

Problem različitih skala – varijable u istom setu podataka su često u različitim skalama (primjer s Irisima) – varijable se mogu skalirati na z vrijednost (udaljenost od prosjeka izražena u standardnim devijacijama) – R funkcija scale()

```
> head(data)
  SepalLength SepalWidth PetalLength PetalWidth Species
1         5.1        3.5         1.4        0.2  setosa
2         4.9        3.0         1.4        0.2  setosa
3         4.7        3.2         1.3        0.2  setosa
4         4.6        3.1         1.5        0.2  setosa
5         5.0        3.6         1.4        0.2  setosa
6         5.4        3.9         1.7        0.4  setosa
```

```
> head (scale(data[1:4]))
     SepalLength  SepalWidth PetalLength PetalWidth
[1,]  -0.8976739  1.01560199   -1.335752  -1.311052
[2,]  -1.1392005 -0.13153881   -1.335752  -1.311052
[3,]  -1.3807271  0.32731751   -1.392399  -1.311052
[4,]  -1.5014904  0.09788935   -1.279104  -1.311052
[5,]  -1.0184372  1.24503015   -1.335752  -1.311052
[6,]  -0.5353840  1.93331463   -1.165809  -1.048667
```

Multivarijatne metode služe za (*ovo nije „službena”, potpuna, niti jedina podjela + kategorije se često preklapaju*):

1) Redukciju (smanjenje) broja varijabli (dimension reduction) – uglavnom tvore sintetske varijable koje nastoje opisati svu varijabilnost izvornih varijabli – za vizualizaciju ili daljnje analize manjeg broja varijabli - analiza glavnih sastavnica (PCA), multidimenzionalno skaliranje (MDS), kanonička korelacija (CCA)…

2) Klasificiranje – metode koje razvrstavaju jedinke u poznat ili nepoznat broj grupa (kategorija) – po sličnosti ili nekom drugom algoritmu – metode klasteriranja (temeljene na modelu ili ne; hijerarhijske ili ne...) – diskriminantna analiza (LDA, QDA…), k means clustering, hijerarhijski klasteri, classification trees, razne metode strojnog učenja…

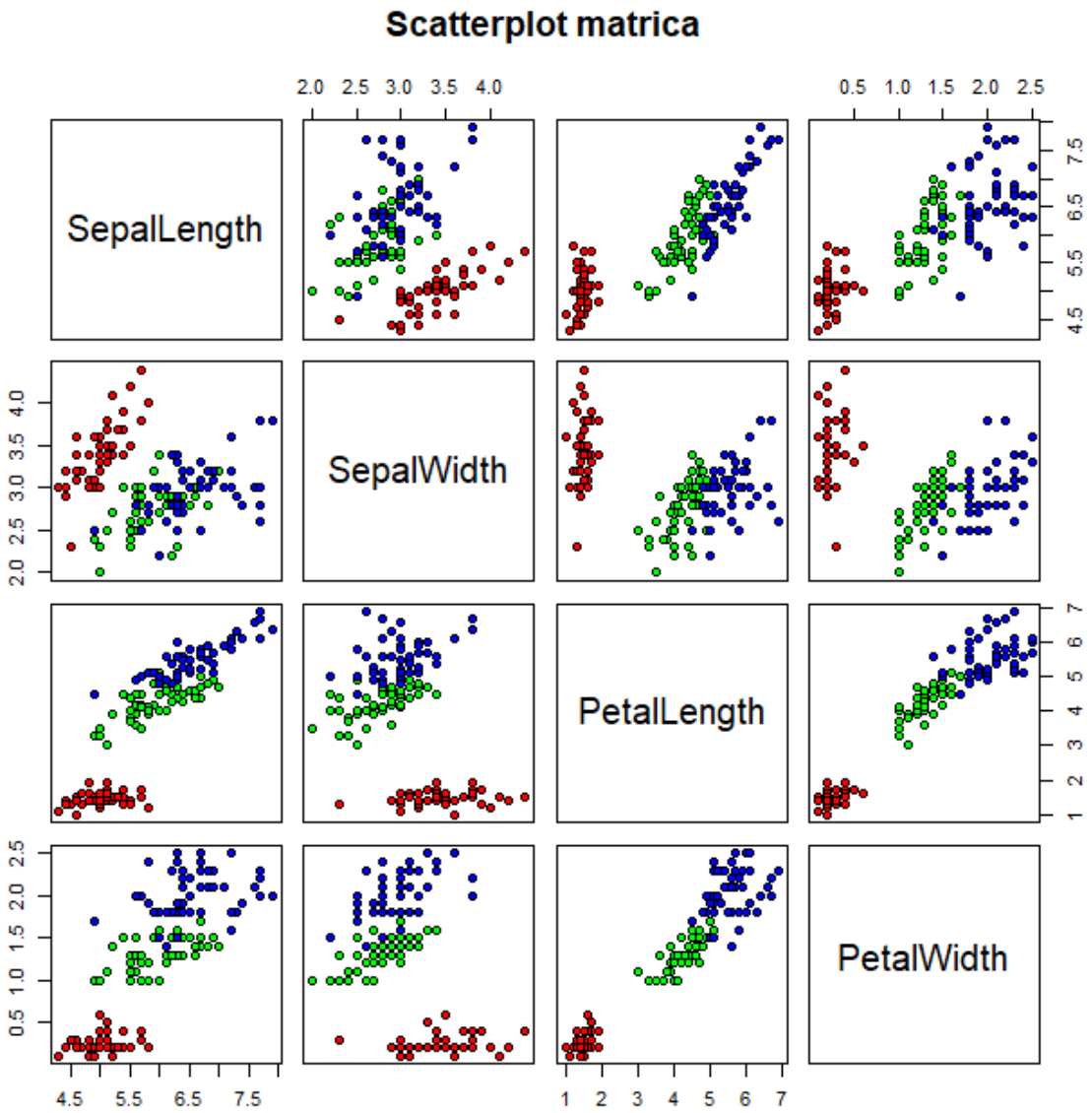
3) Testiranje hipoteza – modeli – Multivarijatna ANOVA (MANOVA); dekompozicija matrice udaljenosti (ANOSIM; AMOVA)…

Korelacija – mjera „povezanosti” 2 varijable – temelj metoda za smanjenje broja varijabli

Korelacijski koeficijent – r – (-1<=r<=1) – smje (predznak) i jačina (apsolutna vrijednost) povezanosti

r se testira t-testom (generalno, jaka korelacij će gotovo uvijek biti signifikantna)

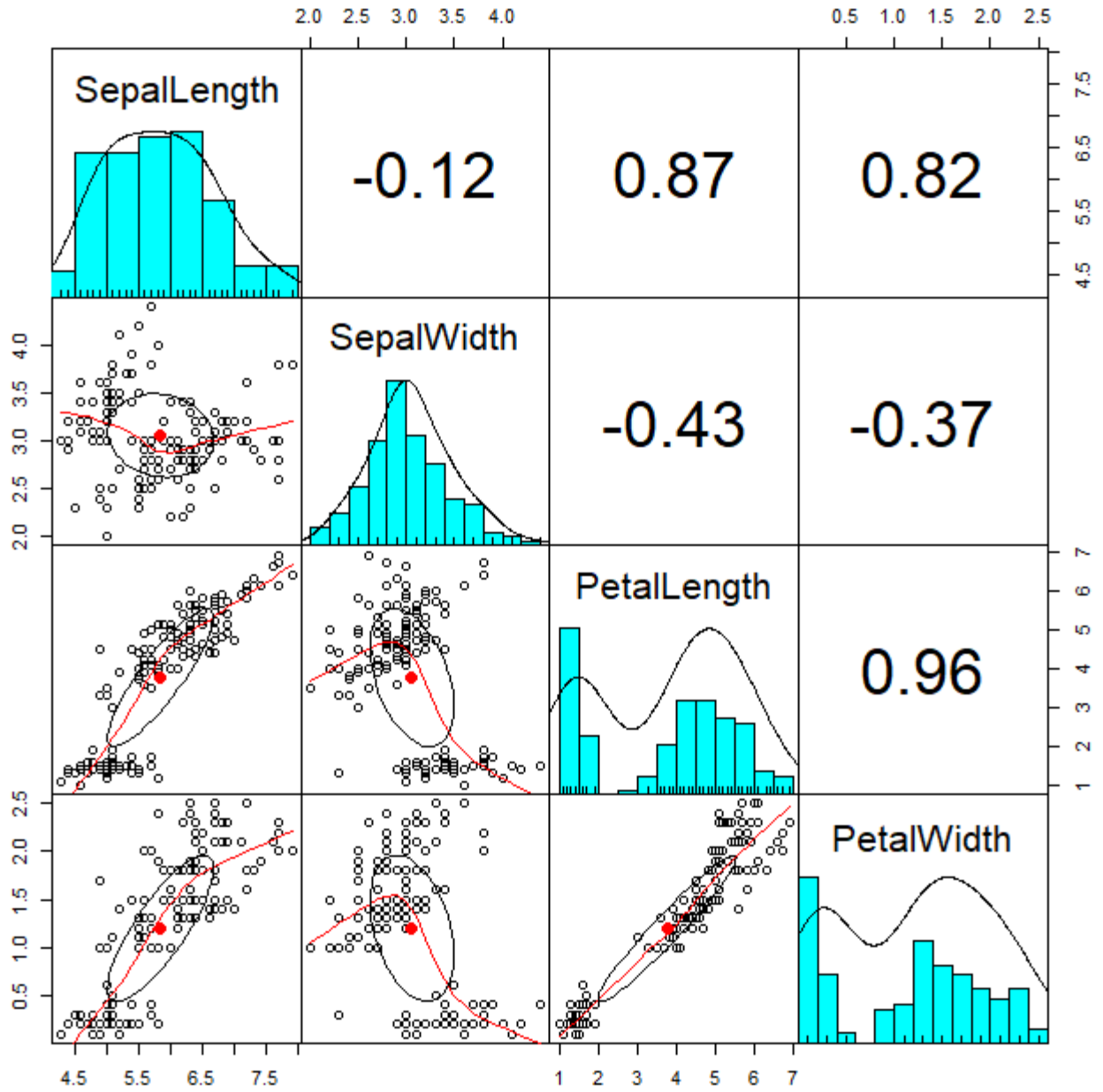
Scatter plot matrix – točkasti grafovi svih parova varijabli – ako opišemo elipsu oko točaka, njena „širina” okvirno procjenjuje korelaciju varijabli



Korelacija – mjera „povezanosti” 2 varijable

```
p_values<-rcorr(as.matrix(data[1:4]))
p_values
```

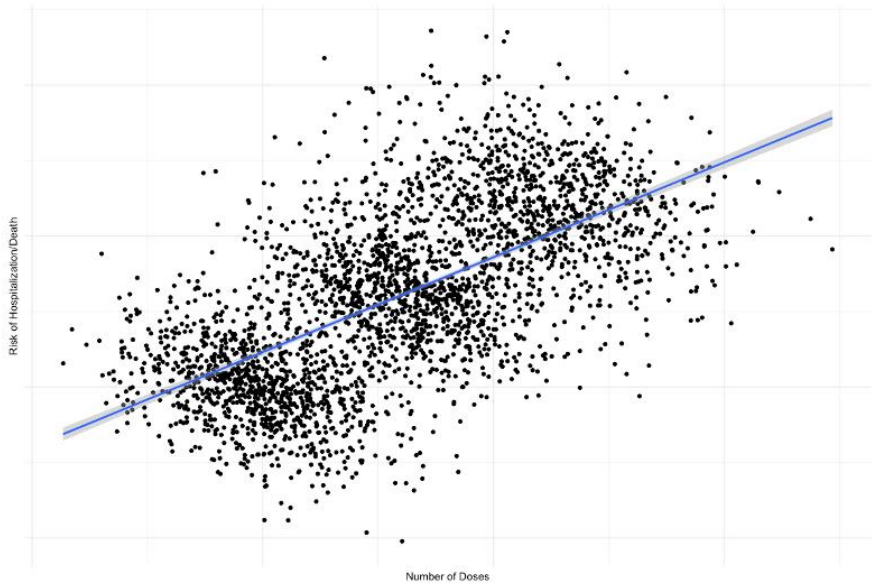
```
  SepalLength SepalWidth PetalLength PetalWidth
SepalLength        1.00       -0.12        0.87       0.82
SepalWidth        -0.12        1.00       -0.43      -0.37
PetalLength        0.87       -0.43        1.00       0.96
PetalWidth         0.82       -0.37        0.96       1.00

n= 150


P
            SepalLength SepalWidth PetalLength PetalWidth
SepalLength             0.1519     0.0000      0.0000
SepalWidth  0.1519                 0.0000      0.0000
PetalLength 0.0000      0.0000                 0.0000
PetalWidth  0.0000      0.0000     0.0000
```

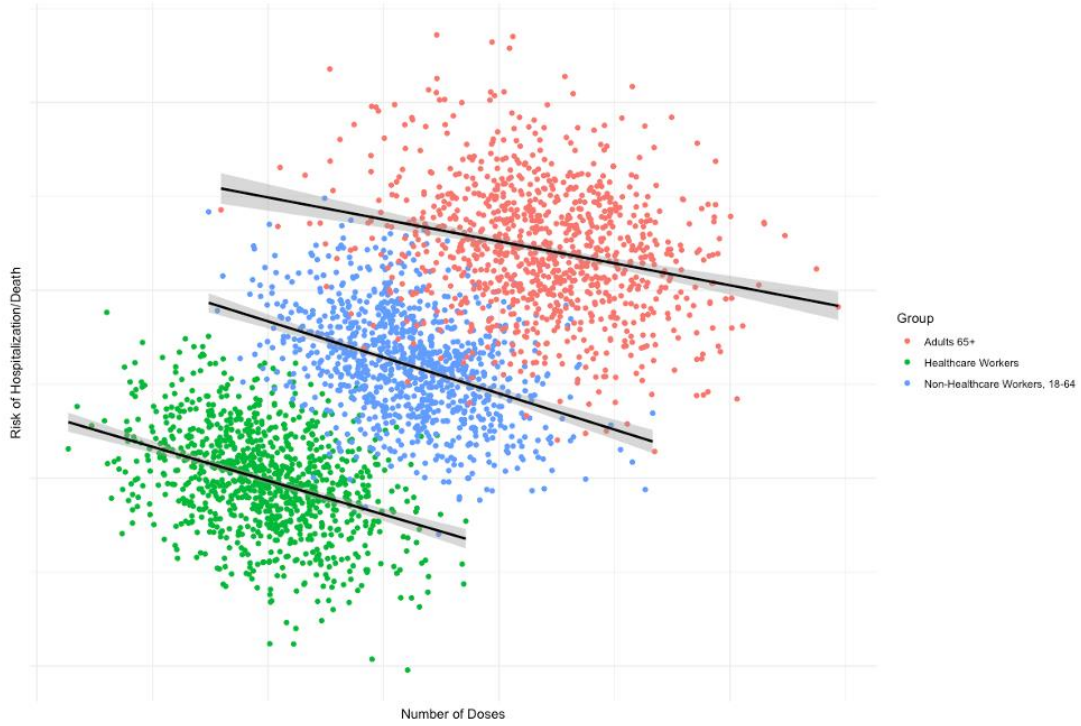


Problem strukturiranih podataka (nisu svi iz iste populacije)

Simpsonov paradoks - statistički fenomen u kojem povezanost između dvije varijable u populaciji nestaje ili postaje obrnuta kada se populacija podijeli u subpopulacije

Posljedično – potencijalni problem za neke od metoda (npr PCA)

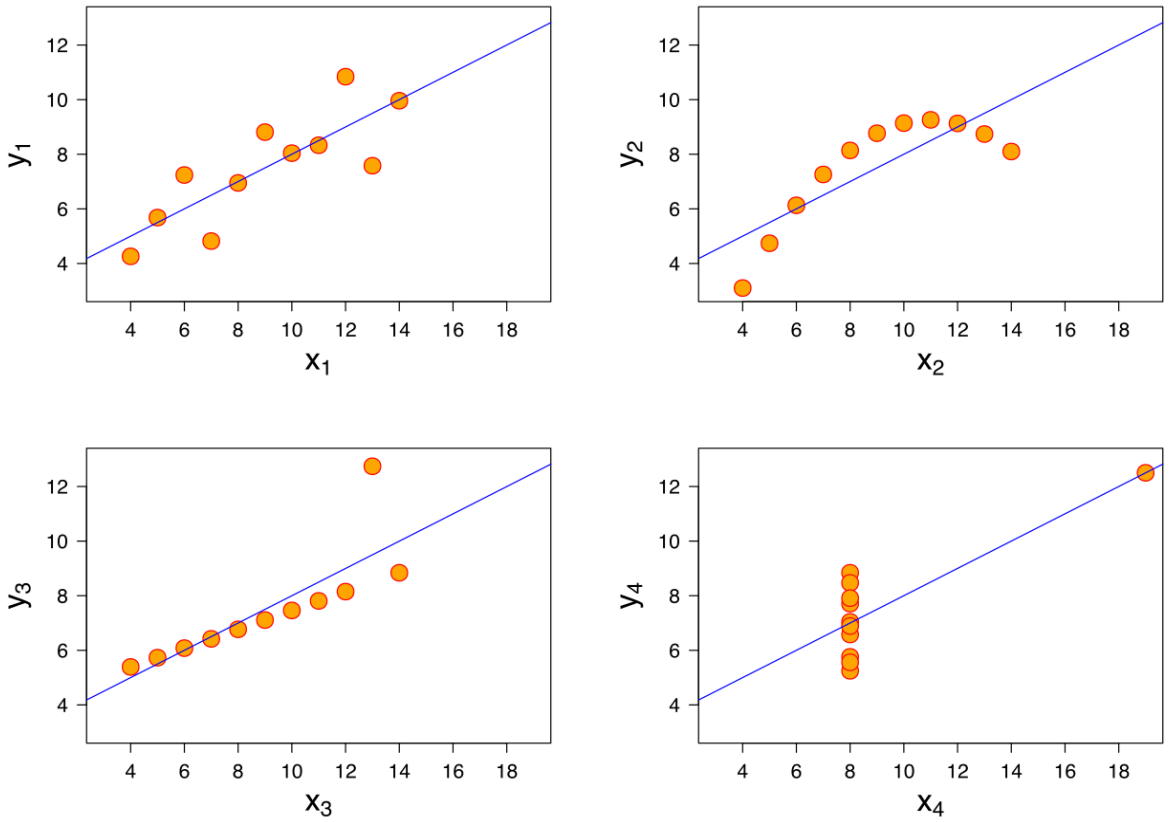




Problem strukturiranih podataka

Anscombeov kvartet - 4 seta podataka od po 2 varijable sa približno jednakim mjerilima sredine i disperzije, ali sasvim različitih distribucija (i grafova)

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x: s^2 x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y: s^2 y	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression: R^2	0.67	to 2 decimal places



Analiza glavnih sastavnica (komponenata) - PCA

Iz k originalnih varijabli tvori k novih, **sintetskih**, koje su **nekorelirane** i obuhvaćaju najveći dio neobjašnjene varijabilnosti (dakle svaka nakon prve objašnjava sve manji dio)

Sintetske varijable (komponente) „izdvajaju" zajedničku dio varijabilnosti opisan koreliranim varijablama – cilj je „dobiti" što manje komponenata koje objašnjavaju što veći dio izvorne varijabilnosti podataka (nikada neće biti potpun ako se ne uzmu sve komponente)

Nije statistička metoda (ne testira nikakvu hipotezu, ne omogućava generalizaciju) već samo transformira podatke (ovo se često zaboravlja!)

Analiza glavnih sastavnica (komponenata) - PCA

R ima više funkcija za PCA – prcomp(), princomp(), paket Factoextra… - neke koriste matricu korelacija a druge matricu kovarijanci

Bitna napomena – PCA koristi samo numeričke varijable! (dakle ne i kategoričke – problem strukturiranih podataka!)

```
> pca_res <- prcomp(data [1:4], scale. = TRUE)
> print(pca_res)
Standard deviations (1, .., p=4):
[1] 1.7083611 0.9560494 0.3830886 0.1439265

Rotation (n x k) = (4 x 4):
                    PC1         PC2        PC3        PC4
SepalLength  0.5210659 -0.37741762  0.7195664  0.2612863
SepalWidth  -0.2693474 -0.92329566 -0.2443818 -0.1235096
PetalLength  0.5804131 -0.02449161 -0.1421264 -0.8014492
PetalWidth   0.5648565 -0.06694199 -0.6342727  0.5235971
```

Kvadrati st dev su eigenvalues (svojstvene vrijednosti)!
To je bitno kod odluke koliko osi nam je „bitno” (ev>1)

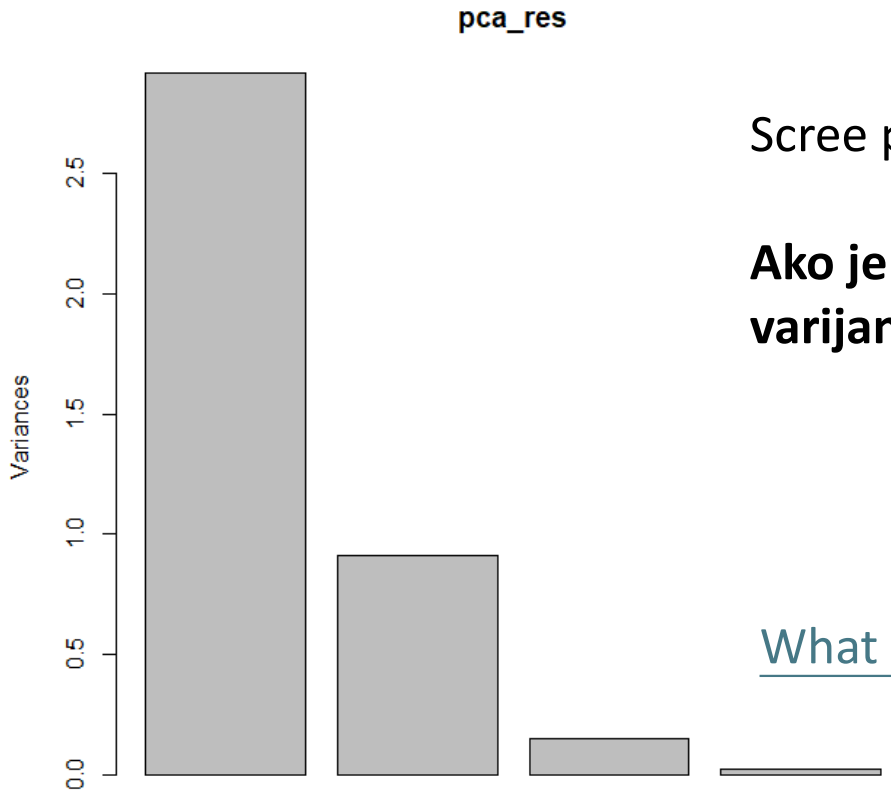
Ovo su korelacije izvornih varijabli sa svakom od PC osi (sintetskih varijabli)

```
> summary(pca_res)
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.7084 0.9560 0.38309 0.14393
Proportion of Variance 0.7296 0.2285 0.03669 0.00518
Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
```

Udio objašnjene varijance za svaku PC

Kumulativni dio objašnjene varijance

Prve dvije PC osi zajedno objašnjavaju 95.8% ukupne varijabilnosti

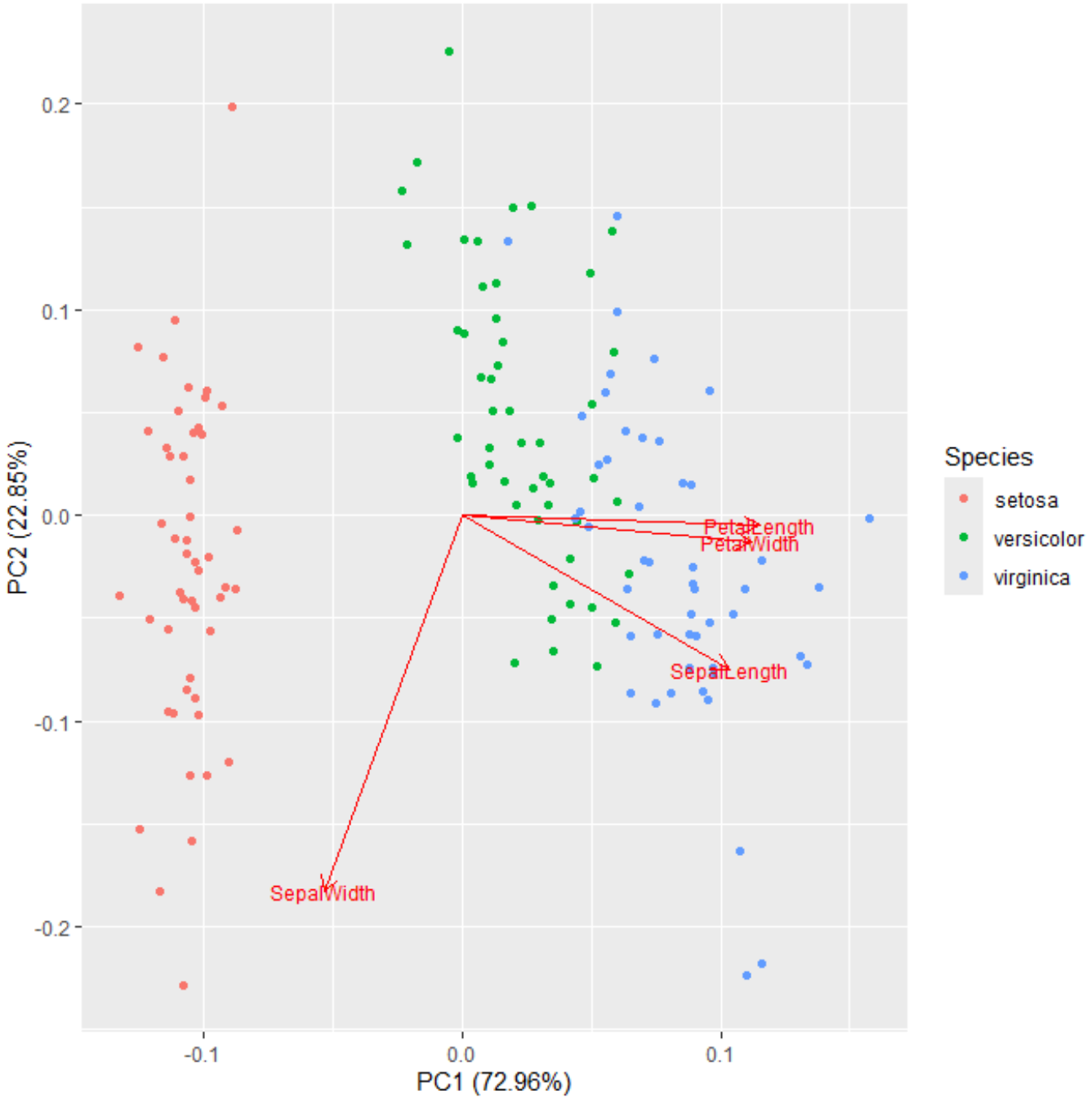


Scree plot

Ako je varijanca neke osi veća od 1 – ta os objašnjava više varijance nego jedna izvorna varijabla

prve dvije PC osi objašnjavaju gotovo svu izvornu varijabilnost

What is the Use of Eigenvalues & Eigenvectors in PCA? (Example)



Prikaz prve 2 PC osi

Boje su dodane samo na graf – vrste nisu dio analize!!
Nema rasprave o tome kako se vrste „razdvajaju”!!

Vektori pokazuju smjer i jačinu korelacije izvornih varijabli s PC osima (i međusobnu korelaciju izvornih varijabli)

Setosa ima kraće i uže latice od ostalih vrsta, ali šire lapove.

Možemo raditi ANOVA-u na PC1 i PC2

(Linearna) diskrimantna analiza - LDA

Definira **diskriminantne funkcije koje najbolje razdvajaju zadane** (definirane) kategorije – maksimiziraju razliku između kategorija

Diskriminantnih funkcija ima i-1 (i je broj zadanih grupa)

Kategorijska varijabla je dio modela (to je **zavisna** varijabla u modelu)

Definiranu diskriminantnu funkciju može se primijeniti na jedinke za koje se ne zna kojoj kategoriji pripadaju, i to pokušati procijeniti – procjenjuje vjerojatnost da nepoznata jedinka pripada svakoj od definiranih kategorija

Pomaže procijeniti koliko svaka od izvornih varijabli pridonosi razlikovanju definiranih kategorija

```
> lda_res <- lda(Species ~ SepalLength + SepalWidth + PetalLength + PetalWidth, data = data)
> lda_res
Call:
lda(Species ~ SepalLength + SepalWidth + PetalLength + PetalWidth,
    data = data)

Prior probabilities of groups:
    setosa versicolor  virginica
 0.3333333  0.3333333  0.3333333

Group means:
           SepalLength SepalWidth PetalLength PetalWidth
setosa           5.006      3.428       1.462      0.246
versicolor       5.936      2.770       4.260      1.326
virginica        6.588      2.974       5.552      2.026

Coefficients of linear discriminants:
                    LD1         LD2
SepalLength  0.8293776 -0.02410215
SepalWidth   1.5344731 -2.16452123
PetalLength -2.2012117  0.93192121
PetalWidth  -2.8104603 -2.83918785

Proportion of trace:
   LD1    LD2
0.9912 0.0088
```

Unaprijed definirane vjerojatnosti pripadanja svakoj od grupa – default=1/broj grupa

LD1 i LD2 su diskriminantne funkcije (2 su jer su 3 grupe)

Ovo su koeficijenti linearnih funkcija za svaki prediktor (izvornu varijablu)

```
> lda_pred <- predict(lda_res)
> head(lda_pred$posterior)
  setosa   versicolor    virginica
1      1 3.896358e-22 2.611168e-42
2      1 7.217970e-18 5.042143e-37
3      1 1.463849e-19 4.675932e-39
4      1 1.268536e-16 3.566610e-35
5      1 1.637387e-22 1.082605e-42
6      1 3.883282e-21 4.566540e-40
```

Nakon što definiramo LDA, primijenimo ju na jedinke koje želimo klasificirati

Rezultat – **vjerojatnost pripadanja svakoj od 3 vrste** za prvih 6 jedinki (samo prvih 6 je prikazano)

```
> mean(lda_pred$class==data$Species)
[1] 0.98
```

Točnost predikcije – usporedba unaprijed definiranih vrsta i rezultata klasifikacije
98% jedinki je točno razvrstano

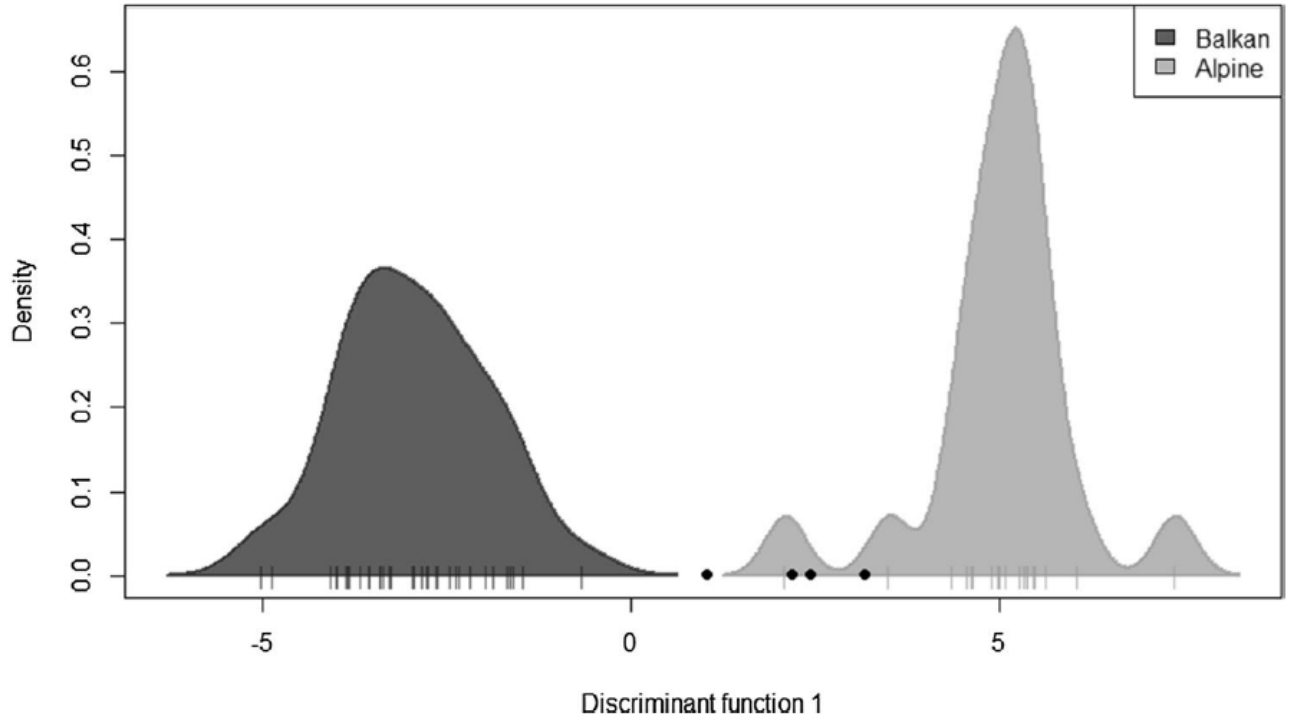
```
> tab <- table(pred = lda_pred$class, true = data$Species)
> tab
            true
pred         setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         48         1
  virginica       0          2        49
```

Tablični prikaz točnosti predikcije za svaku vrstu

Microsatellite based assignment reveals history of extirpated mountain ungulate

Toni Safner[1,2] · Elena Buzan[3,4] · Laura Iacolina[5] · Sandra Potušek[3] · Andrea Rezić[5] · Magda Sindičić[6] · Krešimir Kavčić[5] · Nikica Šprem[5]

4 muzejska uzorka DNA divokoze sa Velebita starija od 1940 (istrijebljena početkom 2 stoljeća) genotipizirana mikrosatelitnim biljezima
Sadašnje populacije Balkanske i Alpske divokoze genotipizirane istim biljezima
Diskriminantna funkcija kreirana na „sadašnjim" uzorcima, i primijenjena na muzejske



K means clustering

Grupira podatke u K (zadano) grupa prema sličnosti.

Algoritam je jednostavan – najprije se nasumično generira k centroida (jezgri, središta) clustera, te se svaka jedinka pridaje clusteru čijem centroidu je najsličnija

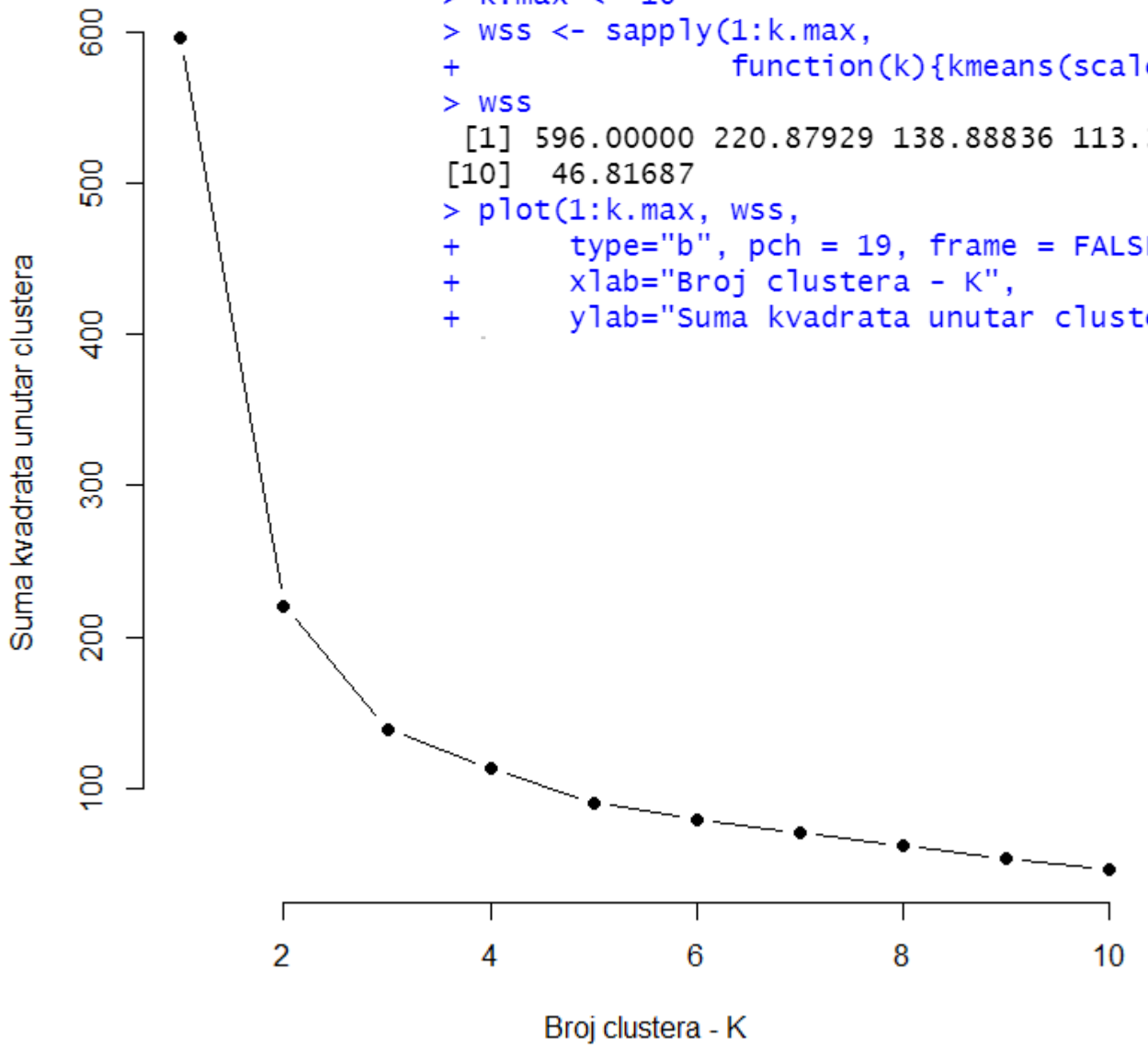
Zbog nasumičnosti, dobro je proceduru ponoviti više puta i odabrati „najbolji” rezultat (R to radi automatski)

Kriterij odabira – najmanja suma kvadrata unutar klastera (wss)

Problem – kako znamo koji k zadati?

Ako nemamo predefiniran broj na umu, najpovoljniji K se odredi ponavljanjem postupka za različite vrijednosti K, i odabira najbolje opcije usporedbom suma kvadrata unutar klastera

```
> set.seed(123)
> k.max <- 10
> wss <- sapply(1:k.max,
+               function(k){kmeans(scaled_data, k, nstart=50,iter.max = 15 )$tot.withinss})
> wss
 [1] 596.00000 220.87929 138.88836 113.33162  90.20190  79.46554  70.18758  62.05221  53.89983
[10]  46.81687
> plot(1:k.max, wss,
+      type="b", pch = 19, frame = FALSE,
+      xlab="Broj clustera - K",
+      ylab="Suma kvadrata unutar clustera")
```



Tražimo „koljeno” na grafu

K=3

Ponovimo analizu sa K=3

```
K-means clustering with 3 clusters of sizes 50, 47, 53

Cluster means:
  SepalLength  SepalWidth PetalLength PetalWidth
1 -1.01119138  0.85041372  -1.3006301 -1.2507035
2  1.13217737  0.08812645   0.9928284  1.0141287
3 -0.05005221 -0.88042696   0.3465767  0.2805873

Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [48] 1 1 1 2 2 2 3 3 3 2 3 3 3 3 3 3 3 2 3 3 3 3 2 3 3 3 3 2 2 2 3 3 3 3 3 3 2 2 3 3 3 3
 [95] 3 3 3 3 3 3 2 3 2 2 2 2 3 2 2 2 2 2 2 3 3 2 2 2 2 3 2 3 2 3 2 2 3 2 2 2 2 2 2 3 3 2
[142] 2 3 2 2 2 3 2 2 3

Within cluster sum of squares by cluster:
[1] 47.35062 47.45019 44.08754
 (between_SS / total_SS =  76.7 %)
```

Tablica s rezultatom (isti postupak kao kod LD)

```
> tab <- table(clust = dd$cluster, true = dd$Species)
> tab
     true
clust setosa versicolor virginica
    1     50          0         0
    2      0         11        36
    3      0         39        14
```