# Domača naloga 1

Neza Krzan, Tom Rupnik

## Kazalo



## Slike

# Tabele

# Cilji naloge

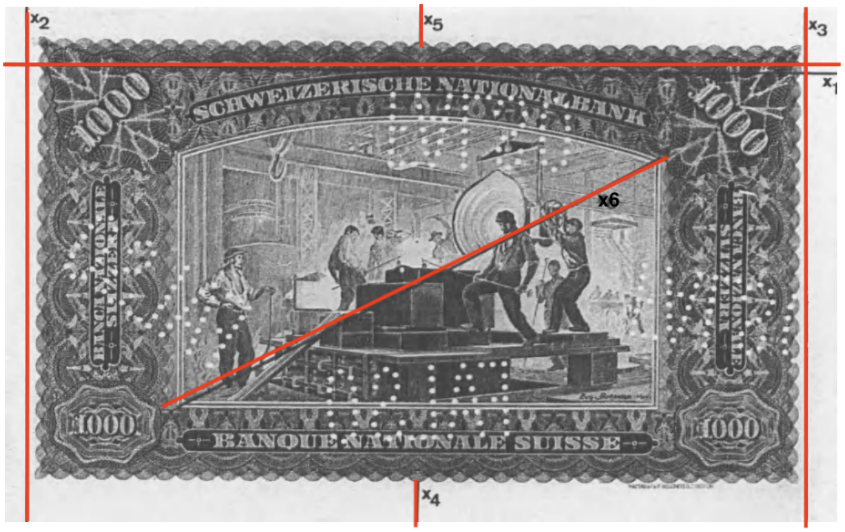
V nalogi vova poskušala razvrstiti enote v skupine tako, da si bodo enote znotraj skupin čim bolj podobne in enote v različnih skupinah čim bolj različne glede na več spremenljivk.

# Podatki

Uporabila bova podatke *Swiss banknotes data*, ki vsebujejo šest meritev, opravljenih na 100 pravih in 100 ponarejenih starih švicarskih bankovcih za 1000 frankov.

Podatki vsebujejo 7 spremenljivk - 6 številskih in eno opisno. Vsebujejo različne izmerjene dolžine in širine bankovca v milimetrih:

- `length`: dolžina bankovca(na sliki $x_1$),
- `left`: dolžina levega roba(na sliki $x_2$),
- `right`: dolžina desnega roba(na sliki $x_3$),
- `bottom`: dolžina spodnjega roba(na sliki $x_4$) in
- `top`: dolžina zornjega roba(na sliki $x_5$) ter
- `diag`: dolžina diagonale bankovca(na sliki $x_6$).




Slika 1: Označene mere na bankovcu.

Opisna spremenlivka `status` pa določa ali je bankovec pravi(`genuine`) ali ponarejen(`counterfeit`). V tabeli imamo torej meritve za 200 različnih bankovcev.

## Urejanje podatkov

Imena spremenljivk in vrednosti kategorične spremenljivke sva preimenovala v slovenska imena ter, kot sva že napisala zgoraj, sva podatke skalirala.

Preimenovane spremenljivke:

- `length`: `dolžina`,
- `left`: `levi.rob`,
- `right`: `desni.rob`,
- `bottom`: `spodnji.rob`,
- `top`: `zgornji.rob`,

Tabela 1: Opisne statistike za številske spremenljivke v podatkovnem okviru `Swiss banknotes data`.

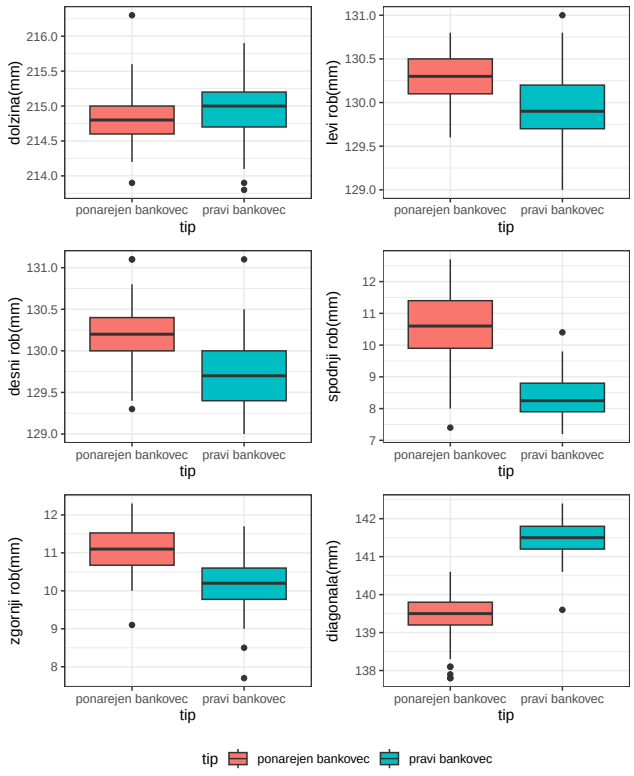
| Variable | N | Mean | Std. Dev. | Min | Pctl. 25 | Pctl. 50 | Pctl. 75 | Max |
|---|---|---|---|---|---|---|---|---|
| dolzina | 200 | 215 | 0.4 | 214 | 215 | 215 | 215 | 216 |
| levi.rob | 200 | 130 | 0.4 | 129 | 130 | 130 | 130 | 131 |
| desni.rob | 200 | 130 | 0.4 | 129 | 130 | 130 | 130 | 131 |
| spodnji.rob | 200 | 9 | 1 | 7 | 8 | 9 | 11 | 13 |
| zgornji.rob | 200 | 11 | 0.8 | 8 | 10 | 11 | 11 | 12 |
| diagonala | 200 | 140 | 1 | 138 | 140 | 140 | 142 | 142 |

- `diag`: `diagonala` in
- `status` : `tip`, kjer je potem `counterfeit`:`ponarejen bankovec` in `genuine`:`pravi bankovec`.

Za lažjo predstavo si poglejmo opisne statistike številskih spremenljivk, da bomo vedeli s kakšnimi podatki imamo opravka.

Spremenljivke imajo različen razpon vrednosti, zato jih bova, skalirala. Tako bodo imele spremenljivke povprečje 0 in standardni odklon 1. S tem doseževa enakovreden vpliv spremenljivk na razvrstitev. Vidimo pa tudi, da nimamo mankajočih vrednosti v podatkih.

Poglejmo si še porazdelitve spremenljivk.




Slika 2: Porazdelitve spremenljivk v podatkovnem okviru `Swiss banknotes data`.

Opazna je razlika med pravimi bankovci in ponarejenimi pri vseh spremenljivkah.

### Analiza povezanosti med spremenljivkami

### Konstrukcija in analiza Likertovih spremenljivk

Za razvrščanje bova uporabljala samo številske spremenljivke, in sicer `dolzina`, `levi.rob`, `desni.rob`, `spodnji.rob`, `zgornji.rob`; za analizo pa spremenljivki `tip` in `diagonala`. Ker je `diagonala` edina številska spremenljivka pri analizi, le ta ne bo skalirana.

## Hierarhično razvrščanje

Pri hierarhičnem razvrščanju začnemo s tem, da je vsaka enota v svoji skupini. Potem pa se na vsakem koraku, glede na izračunane matrike različnosti, v kateri so razdalje med pari skupin, združujejo skupine, ki so si najbližje. Nato se izračunajo različnosti novih združenih skupin od ostalih, kar se nadaljuje dokler niso vse enote v eni skupini. Dobra lastnost hierarhičnega razvrščanja je, da uporabniku ni potrebno vnaprej določiti števila skupin.

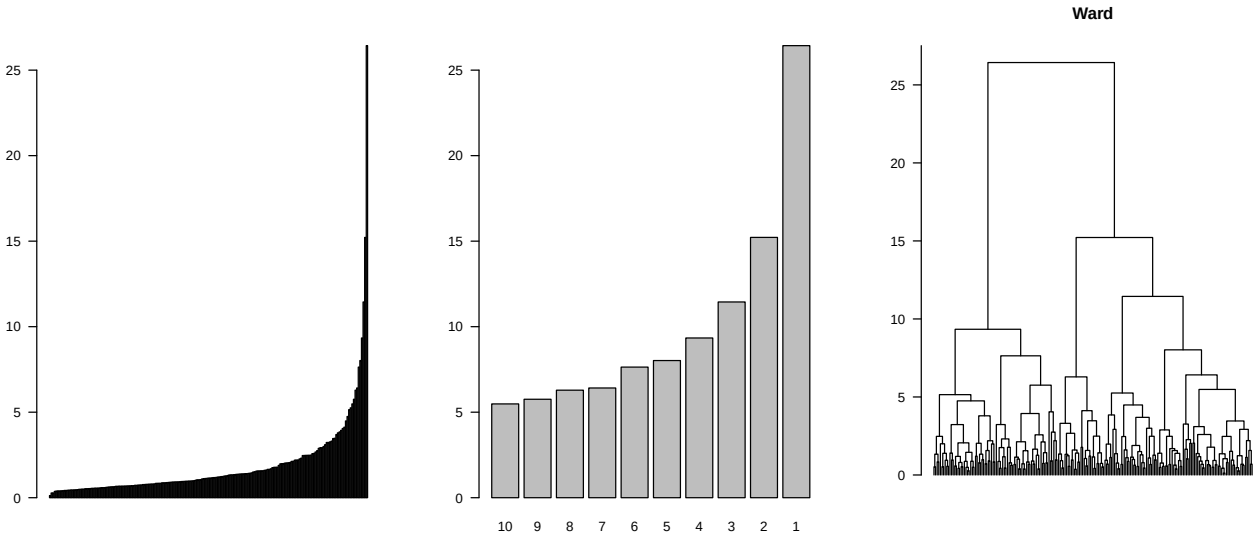
Kot mero različnosti bova uporabila evklidsko razdaljo.

Torej za razvrščanje uporabljava spremenljivke `dolzina`, `levi.rob`, `desni.rob`, `spodnji.rob` in `zgornji.rob` ter primerjala bova tri različne metode in sicer, Wardovo metodo, minimalno metoda (single linkage) in maksimalno metoda (complete linkage).

Število skupin lahko določimo na podlagi dendograma, ki grafično prikazuje potek združevanja v skupine. Število skupin pa določimo tako na podlagi vidnejšega zmanjšanja razdalj skupinami.

### Wardowa metoda

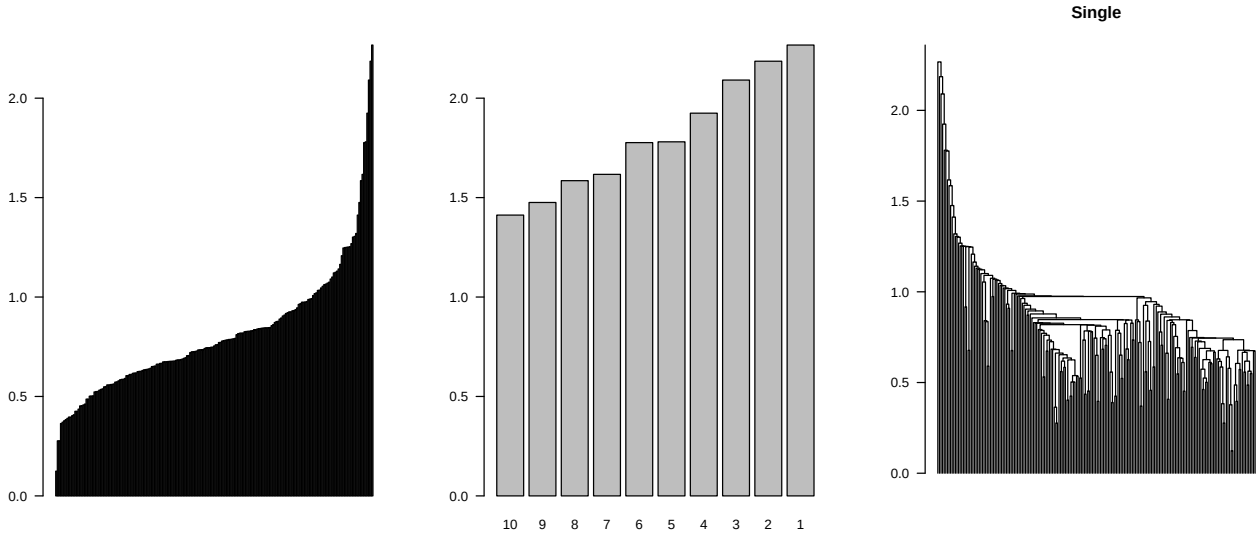
Wardova metoda je primera za eliptične skupine.




Slika 3: Dendogrami Wardowe metode razvrščanja v skupine.

### Minimalna metoda

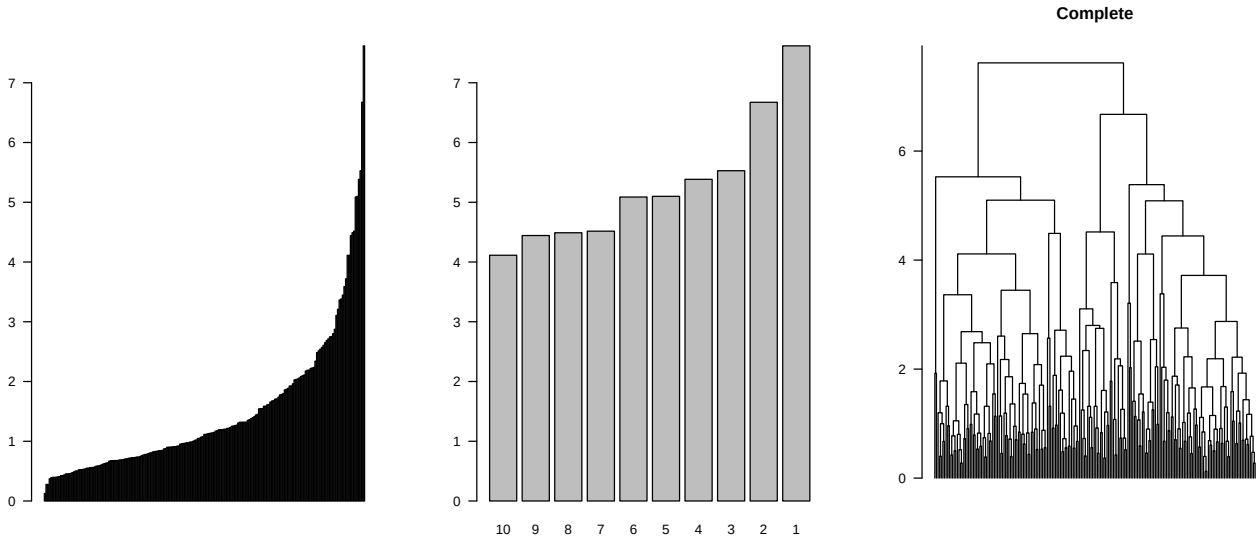
Minimalna metoda (enojna povezanost - single linkage) je primerna za dolge in neeliptične skupine, ki so jasno ločene med seboj. Kadar skupine med seboj niso jasno ločene pri minimalni metodi pride do problema veriženja. Na takem dendogramu ne moremo določiti števila skupin in zato rečemo, da je skupina zgolj ena.




Slika 4: Dendogrami minimalne metode razvrščanja v skupine.

## Maksimalna metoda

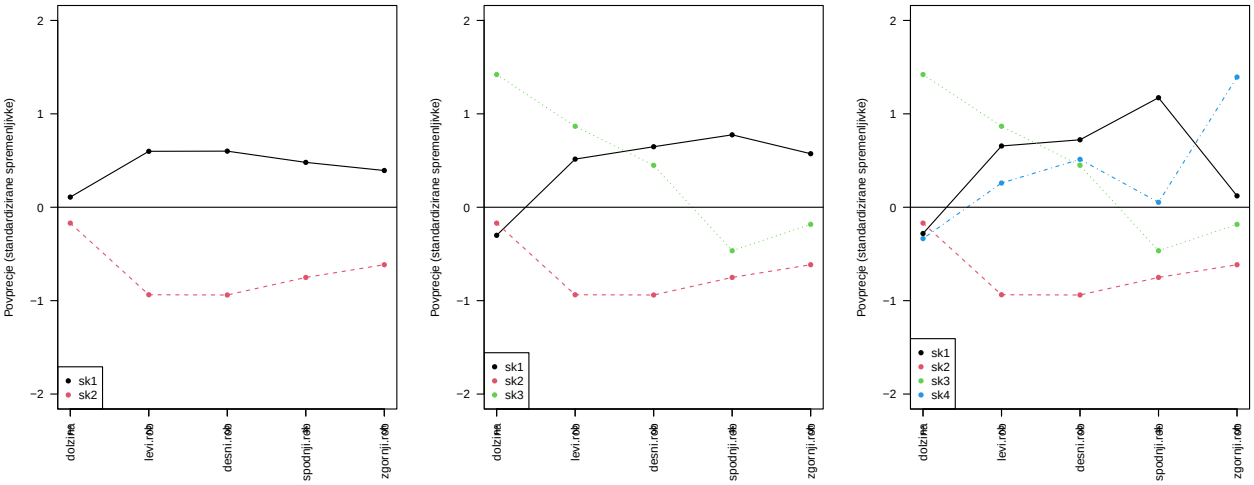
Maksimalna metoda (polna povezanost - complete linkage) pa je primerna za okrogle skupine.




Slika 5: Dendogrami maksimalne metode razvrščanja v skupine.

## Analiza

Glede na izgled grafov (razvrstitve) sva se odločila, da je najbolj primerna razvrstitev po Wardowi metodi. Pri ostalih dveh metodah so različnosti dokaj majhne (ni tako izrazitih skokov v višini). Grafe bomo narisali za *2*, *3* in *4* skupine, saj so tu razlike bolj izrazite.




Slika 6: Povprečja po skupinah za Wardowo metodo.

Če si pogledamo skupino 2 na vseh treh grafih, vidimo, da zavzema podpovprečne vrednosti. Ravno obratno vidimo pri skupini 1, ki na prvem grafu zavzema nadpovprečne vrednosti, na drugih dveh pa zavzema podpovprečne vrednosti samo pri dolžini bankovca. Skupina 3 pa je v nekaterih primerih nadpovprečna v nekaterih pa podpovprečna(`spodnji.rob`, `zgornji.rob`). Pri zadnjem garfu se skupina 4 pri spremenljivki `dolzina` pribliza povprecju zelo dobro, pri vseh ostalih spremenljivkah je nadpovprecna in pri zadnji mocno podpovprecna.

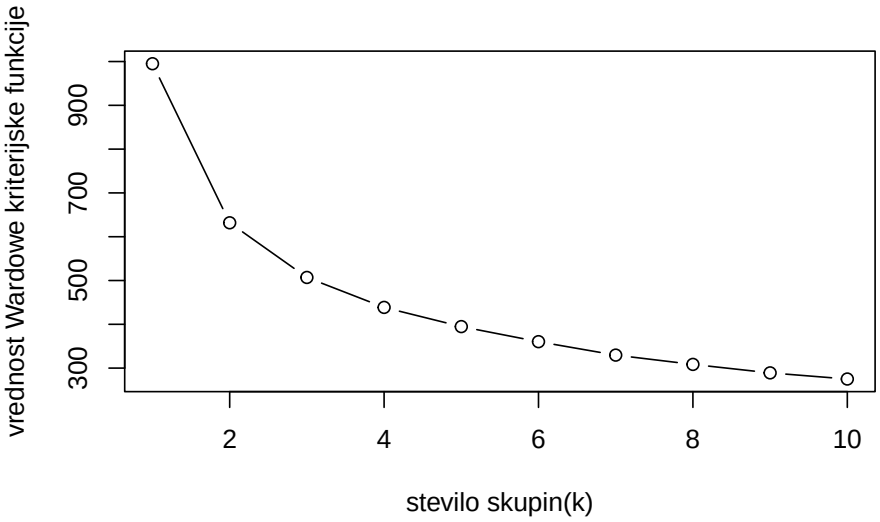
# Nehierarhično razvrščanje

## Razvrščanje K-means

K-means je metoda voditeljev oz. nehierarhičnega razvrščanja. Voditelji so "predstavniki skupin", vsaka enota pa pripada skupini, kateremu voditelju je najbližje(razdalja je evklidska) oz. mu je najbolj podobna; voditelj predstavlja povprečje skupine. Spremenljivke pri metodi *k-means* morajo biti vsaj intervalne.

Tukaj pri tej metodi mora biti število skupin podanoi vnaprej, kar je morda slaba lastnost in se glede tega razlikuje od npr. Wardove metode. Na začetku določimo voditelje, potem pa na vsakem koraku vsako enoto priredimo voditelju oz. skupini, kateremu je najbljižja glede na evklidsko razdaljo. Na vsakem koraku se izračunajo novi voditelji kot povprečja skupin. Postopek se zaključi, ko so novi voditelji enaki starim.

Izberemo tisto razvrstitev, ki ima najmanjšo vrednost Wardove kriterijske funkcije, za katero vemo, da pada z naraščanjem števila skupin. Torej za optimalno število skupin ponavadi vzamemo tisto vrednost, kjer se zgori t.i. "koleno" funkcije. Če to "koleno" ni jasno razvidno, lahko sklepamo, da skupine niso jasno ločene. Postopek običajno večkrat ponovimo, saj za različne začetne voditelje lahko dobimo različne rešitve, torej razvrstitve v skupine.

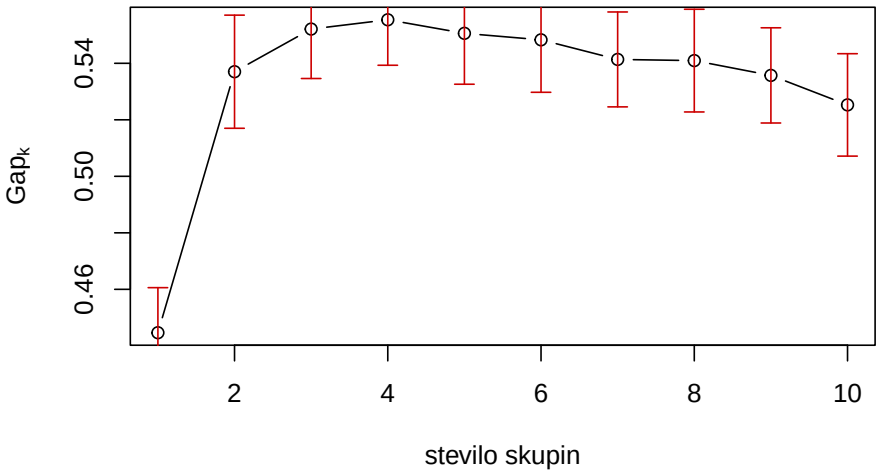



Slika 7: Vrednost Wardove kriterijske funkcije.

Sprememba naklona funkcije izgleda največja pri **2** ali **4** skupinah oziroma je tam "koleno" najbolj razvidno.

## GAP statistika

Pri določevanju števila skupin si lahko pomagamo tudi z GAP statistiko, kjer iščemo skupine, ki so podatki bolj homogeni, kot kjer ni skupin. Gre za primerjavo razdalj znotraj skupin z razdaljami na podatkih brez skupin. Izberemo pa tisto najmanjše število skupin k , kjer je vrednost $GAP(k)$ statistike vsaj tolikšna kot $GAP(k+1) - SE(GAP(k+1))$; $SE$ je standardna napaka GAP statistike.

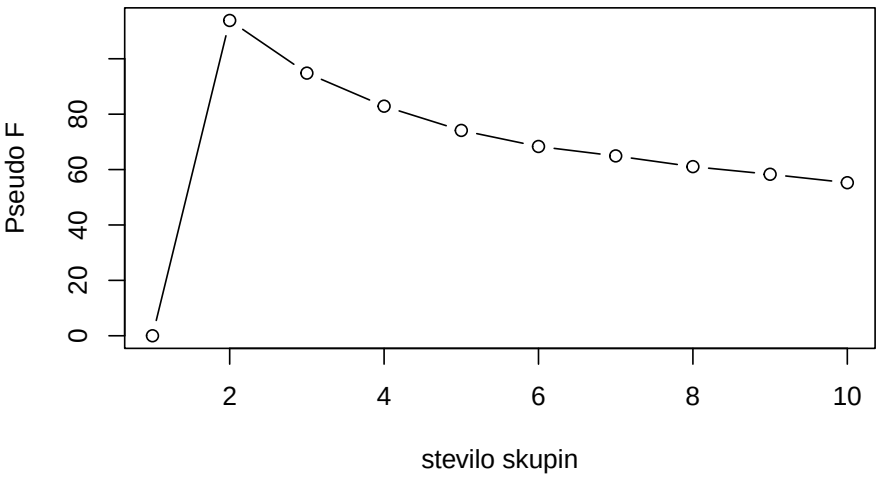



Slika 8: Vrednost GAP statistike.

Na podlagi grafičnega prikaza vrednosti GAP statistike pri različnem številu skupin se odločimo za **4** skupine, saj tam doseže najvišjo točko.

## Pseudo F (Calinski - Harabasz indeks)

Uporabimo pa lahko tudi indeks Calinski-Harabasz, ki ocenjuje razmerje med razpršenostjo znotraj skupin in razpršenostjo med skupinami. Uporabljamo ga za oceno primernosti števila skupin v metodi gručenja (angl. *clustering*). Višje vrednosti indeksa Calinski-Harabasz označujejo boljše gručenje, pri čemer optimalno število skupin običajno doseže maksimum tega indeksa.




Slika 9: Vrednost Pseudo F oz. Calinski - Harabasz indeksa.

Tukaj je maksimum dosežen pri **2** skupinah.

Torej, če povzameva celotno analizo, bi, glede na posamezen graf, izbrala

- WSS: sprememba naklona izgleda največja pri **4** skupinah,
- Pseudo F: maksimum doseze pri **2** skupinah,
- gap statistika: najvišjo točko preden začne padati doseže pri **4** skupinah.

Na podlagi zgornjih analiz in ugotovitev pri hierarhičnem razvrščanju, kjer smo e odločali med 2, 3 ali 4 skupinami, bi se tu določili raje za **4** skupine, kot za 2, saj težimo k večjemu številu skupin kot je 2.

Tabela 2: Primerjava vrednosti kriterijske funkcije za Wardowo metodo in K-means.

| | k=1 | k=2 | k=3 | k=4 |
|---|---|---|---|---|
| Ward | 995 | 645.6782 | 529.8618 | 464.3683 |
| Kmeans | 995 | 631.7882 | 506.9825 | 438.5950 |

## Primerjava vrednosti kriterijske funkcije za Wardowo metodo in K-means

Primerjala sva tudi vrednosti kriterijskih funkcij za Wardowo metodo in metodo K-means, ker sta podobno oziroma delujeta na isti princip. Je pa metoda K-means boljša, ker išče lokalne minimume, za razliko do Wardove, ki deluje hierarhično in vedno poda enak rezultat. Ocenjujeva sva pa po principu, da ima boljša razvrtitev manjšo vrednost karaketristične funkcije. Pomembno pa je tudi to, da so podatki standardizirani, saj drugače med seboj ne bi bilo primerljivo.

Vidimo da ima v vseh primerih (z izjemo prvega kjer sta enaka) K-means manjšo vrednost, kar si tudi želimo. Primerjavo razvrtitev bomo naredili na številu skupin $\mathbf{k = 4}$.

```
##
##      1  2  3  4
##   1 27  0 17  4
##   2  1 69  0  0
##   3  3  3 43  0
##   4  2  6  0 25
```

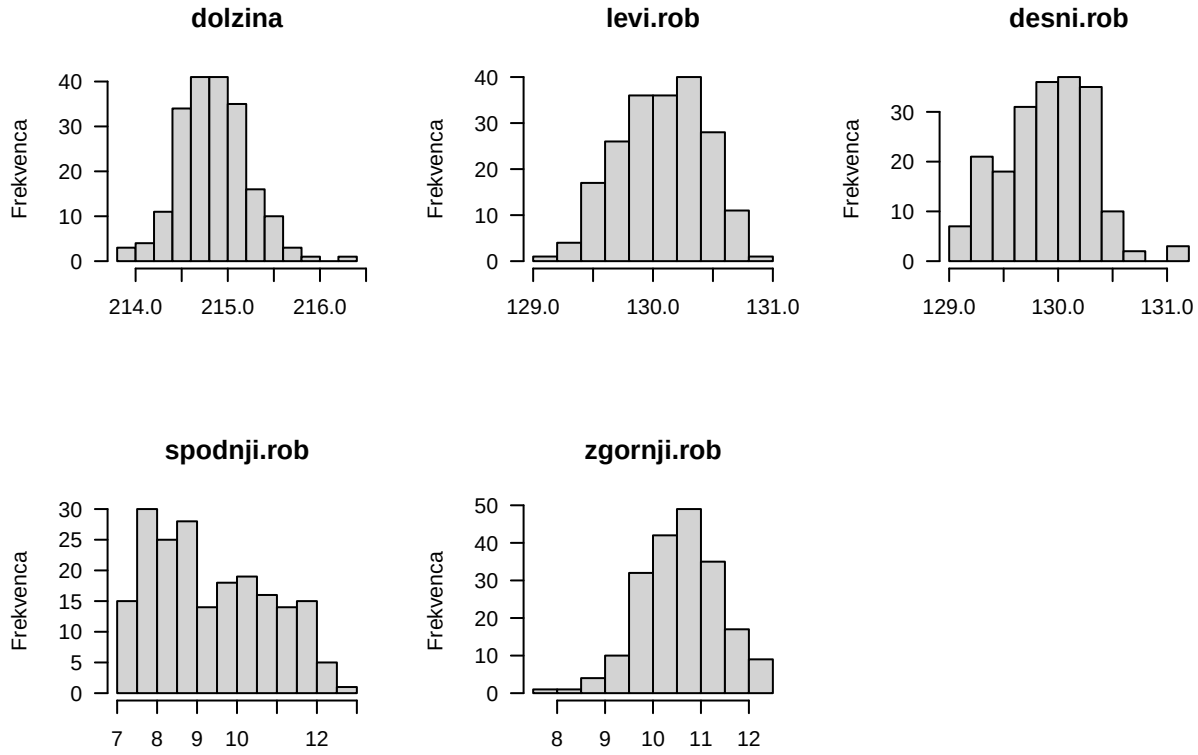
```
## [1] 0.6442817
```

Največje elemente imamo na diagonali kontingenčne tabele, tudi te vrednosti niso ekstremno velike(npr. 400). Za izvendiagonalne elemente si želimo, da bi bili čim manjši oziroma zelo blizu 0, kar pa po večini so, ali pa so celo kar enaki 0(iztopa le ena vrednost - 17).

Poglejmo si še Randow indeks, ki predstavlja delež parov enot, ki so si v obeh razbitjih usklajeni - v obeh razbitjih v isti skupini ali pa v obeh razbitjih v različnih skupinah. Pogledala sva si popravljen Randow indeks, zaradi boljše primerljivosti. Enak je 0.644, kar je blizu 0,5, torej gre skoraj za neko srednjo podobnost, sicer pa večji kot je, boljše je - vrednost 1 pomeni identični razbitji, vrednost 0 pa, da sta si razbitji tako podobni po slučaju.

# Razvrščanje na podlagi modelov

Tukaj predpostavimo, da so podatki generirani iz multivariatnih normalnih porazdelitev z različnimi parametri oziroma komponentami; vsaka skupina ima svojo multivariatno normalno porazdelitev. Skupina je večja po volumnu, če ima večjo variabilnost, omejimo pa se z domnevami oziroma predhodnim znanjem, kakšne naj bi ti skupine bile. Zato si poglejmo porazdelitve spremenljivk ne glede na tip bankovca.



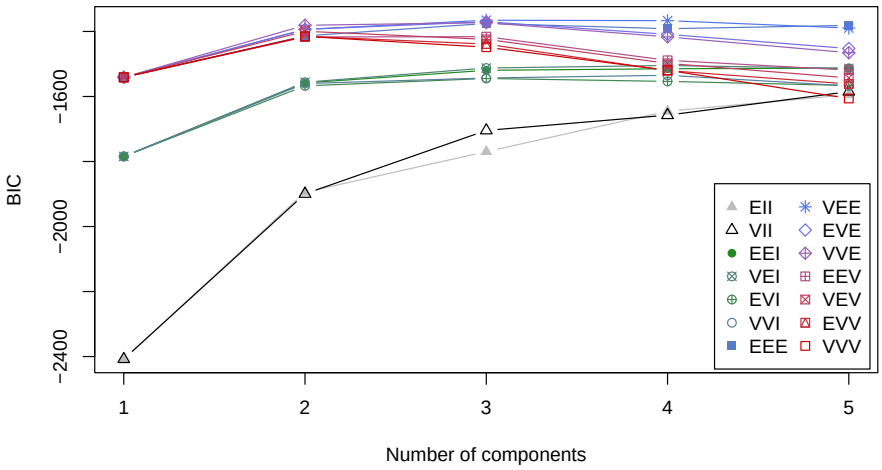

Slika 10: Porazdelitve spremenljivk.

Spremenljivka `spodnji.rob` in tudi `desni.rob` nista porazdeljeni po normalni porazdelitvi, zato ne moremo trditi, da je zadoščen ta pogoj. Ostale so porazdeljene po normalni, nekatere asimetrične v denso(npr. spremenljivka `dolzina`) in nekatere v levo(npr. spremenljivka `levi.rob`).

Tukaj ocenimo število skupin in parametre za vsako skupino ter kateri skupini posamezna enota pripada. V najinem primeru, kjer je predpostavka o multivariatni normalni porazdelitvi kršena, se simulacija ne izkaže za optimalno. Razvrstitev se dela na originalnih podatkih oz. nestandardiziranih podatkih, ker s tem omogočimo različno velikost skupin.

## BIC(Bayes Information Criterion) kriterij

Naredimo torej razvrstitev na originalnih, nestandardiziranih podatkih, kjer funkcija sama izbere naprimernejši model.

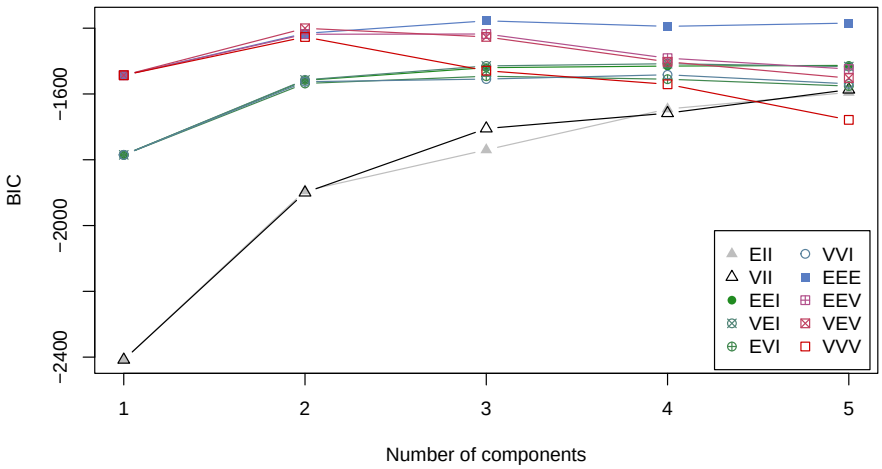



Slika 11: BIC kriterij za originalne podatke.

Na podlagi BIC kriterija (Bayesian Information Criterion), ki zavzame vrednost -592.64 izberemo model VEE s tremi skupinami, kar pomeni, da gre za elipsoidne(angl. *ellipsoidal*) skupine, ki so različno velike, različnih oblik in enako usmerjene.

Kriterij temelji na "Bayesovski" statistiki, zato lahko določimo tudi apriorne verjetnosti(torej neko naše predhodno znanje oziroma prepričanja).

```
# priorControl
mcP <- Mclust(data=df, G=1:5, prior = priorControl())
summary(mcP)
plot(mcP, what = "BIC")
```

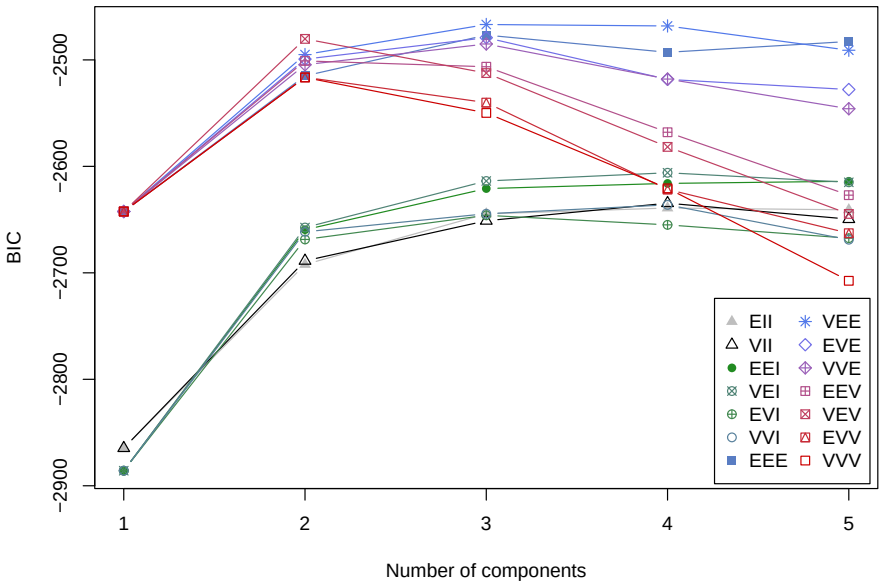



Slika 12: BIC kriterij (priorControl) za originalne podatke.

Na podlagi BIC kriterija z uporabljenim argumentom o apriornih verjetnostih se odločimo za model EEE s tremi skupinami, kar pomeni, da gre za različno velike skupine, različnih oblik in enake usmerjenosti.

## BIC kriterij na standariziranih podatkih

Poglejmo si še, iz radovednosti, kako je z oceno modela na standariziranih podatkih, ampak vrednosti BIC kriterija niso primerljive med standariziranimi in nestandariziranimi podatki.

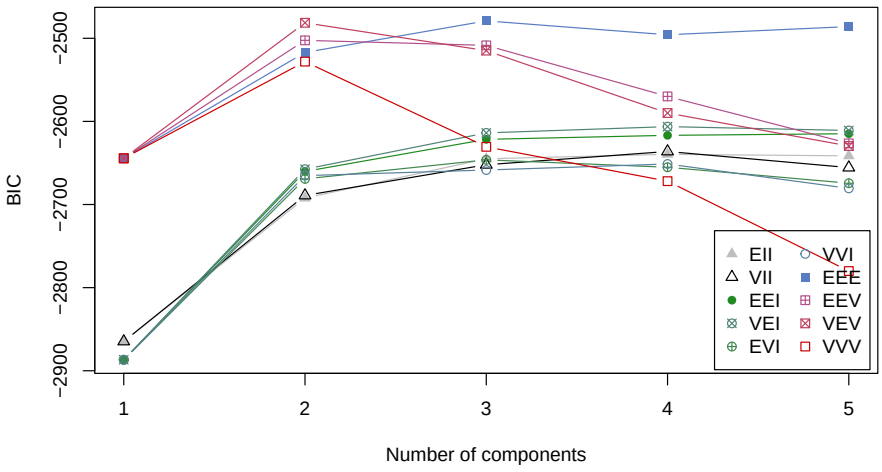



Slika 13: BIC kriterij za standardizirane podatke.

Na podlagi BIC kriterija (Bayesian Information Criterion), ki zavzame vrednost -1143.31 ponovno izberemo model VVE s tremi skupinami.

Tudi tukaj lahko primerjamo z vključitvijo apriornih verjetnosti.

```
# priorControl
mcP <- Mclust(data=dfz, G=1:5, prior = priorControl())
summary(mcP)
plot(mcP, what = "BIC")
```




Slika 14: BIC kriterij (priorControl) za standardizirane podatke.

Tudi tukaj se na podlagi BIC kriterija z uporabljenim argumentom o apriornih verjetnostih se odločimo za model EEE s tremi skupinami, kar pomeni, da gre za različno velike skupine, različnih oblik in enake usmerjenosti.

Na podlagi vseh štirih kriterijev se zaradi enostavnosti odločimo za model VEE s tremi skupinami - torej različno velike skupine, različnih oblik in enakih umserjenosti.

## Najboljša razvrstitev in predstavitev skupin

Tukaj naju pa zanima kako podobne so si naše razvrstitve, ki sva jih v prejšnjih poglavjih izbrala na podlagi različnih modelov.