# 2. laboratorijska vježba

Multivarijatna analiza podataka

Ivan Klabucar

ak. god. 2021/2022

## 1. Uvod i upute za predaju

Cilj ove laboratorijske vježbe je primijeniti osnovne koncepte multivarijatne analize podataka, istražiti podatke te ispitati hipoteze. Preduvjet za rješavanje vježbe je osnovno znanje programskog jezika *R* i rad s *R Markdown* dokumentima. Sama vježba je koncipirana kao projekt u kojem istražujete i eksperimentirate koristeći dane podatke - ne postoji nužno samo jedan točan način rješavanja svakog podzadatka.

Rješavanje vježbe svodi se na čitanje uputa u tekstu ovog dokumenta, nadopunjavanje blokova kôda (možete dodavati i dodatne blokove kôda ukoliko je potrebno) i ispisivanje rezultata (u vidu ispisa iz funkcija, tablica i grafova). Vježbu radite samostalno, a svoje rješenje branite na terminima koji su vam dodijeljeni u kalendaru. Pritom morate razumjeti teorijske osnove u okviru onoga što je obrađeno na predavanjima i morate pokazati da razumijete sav kôd koji ste napisali.

Vaše rješenje potrebno je predati u sustav *Moodle* u obliku dvije datoteke:

1. Ovaj .Rmd dokument s Vašim rješenjem (naziva IME_PREZIME_JMBAG.rmd),

2. PDF ili HTML dokument kao izvještaj generiran iz vašeg .Rmd rješenja (također naziva IME_PREZIME_JMBAG).

Rok za predaju je **15. svibnja 2022. u 23:59h**. **Jedan od uvjeta za prolaz predmeta je minimalno ostvarenih 50% bodova na svim laboratorijskim vježbama. Nadoknade laboratorijskih vježbi neće biti organizirane.** Za sva dodatna pitanja svakako se javite na email adresu predmeta: $map@fer.hr$.

## 2. Podatkovni skup

U laboratorijskoj vježbi razmatra se dinamika cijena vrijednosnica na financijskim tržištima. Dane su povijesne tjedne cijene ETF-ova (eng. exchange traded fund) koji prate određene dioničke, obvezničke ili druge indekse. Konkretno, radi se o sljedećim fondovima:

- `AGG` (iShares Core U.S. Aggregate Bond ETF) - obveznice s američkog tržišta,
- `IEF` (iShares 7-10 Year Treasury Bond ETF) - srednjeročne državne obveznice,
- `LQD` (iShares iBoxx $ Investment Grade Corporate Bond ETF) - korporativne obveznice,
- `SHY` (iShares 1-3 Year Treasury Bond ETF) - kratkoročne državne obveznice,
- `TIP` (iShares TIPS Bond ETF) - državne obveznice zaštićene od inflacije,
- `TLT` (iShares 20+ Year Treasury Bond ETF) - dugoročne državne obveznice,
- `DBC` (Invesco DB Commodity Index Tracking Fund) - sirovine i roba,
- `GLD` (SPDR Gold Trust) - zlato,
- `USO` (United States Oil Fund) - nafta,
- `IJH` (iShares Core S&P Mid-Cap ETF) - dionice tvrtki s američkog tržišta,
- `IWM` (iShares Russell 2000 ETF) - dionice američkih tvrtki s malim kapitalom,
- `SPY` (SPDR S&P 500 ETF Trust) - dionice tvrtki s američkog tržišta,
- `VTV` (Vanguard Value ETF) - dionice tvrtki s američkog tržišta,

- `XLB` (Materials Select Sector SPDR Fund) - dionice tvrtki za materijale,
- `XLE` (Energy Select Sector SPDR Fund) - dionice tvrtki energetskog sektora,
- `XLF` (Financial Select Sector SPDR Fund) - dionice tvrtki financijskog sektora,
- `XLI` (Industrial Select Sector SPDR Fund) - dionice tvrtki industrijskog sektora,
- `XLK` (Technology Select Sector SPDR Fund) - dionice tvrtki iz tehnološkog sektora,
- `XLP` (Consumer Staples Select Sector SPDR Fund) - dionice tvrki za necikličku potrošačku robu,
- `XLU` (Utilities Select Sector SPDR Fund) - dionice tvrtki komunalnih djelatnosti,
- `XLV` (Health Care Select Sector SPDR Fund) - dionice tvrtki iz zdravstvenog sektora,
- `XLY` (Consumer Discretionary Select Sector SPDR Fund) - dionice tvrtki za cikličku potrošačku robu,
- `IYR` (iShares U.S. Real Estate ETF) - dionice tvrtki iz područja nekretnina,
- `VNQ` (Vanguard Real Estate Index Fund) - dionice tvrtki iz područja nekretnina.

Pri modeliranju zajedničkog kretanja i rizika vrijednosnica, najčešće se koriste povrati: $R(t) = \frac{S(t)-S(t-1)}{S(t-1)}$, gdje je $S(t)$ cijena vrijednosnice u tjednu $t$.

### 2.1. Učitavanje podataka i korelacijska analiza

Podaci se nalaze u datoteci "ETFprices.csv". Učitajte ih, provjerite ispravnost, izračunajte tjedne povrate te vizualizirajte matricu korelacije povrata - razmislite o grupama i korelacijskim strukturama koje u njoj vidite. U ostatku laboratorijske vježbe također koristite povrate, a ne cijene.

```
ETF.prices = read.csv(file = 'ETFprices.csv')
ETF.prices$Time = as.Date(ETF.prices$Time, "%d-%b-%Y")

summary(ETF.prices)
```

```
##       Time                 AGG              IEF              LQD
##  Min.   :2006-04-09   Min.   : 64.93   Min.   : 56.54   Min.   : 54.42
##  1st Qu.:2009-06-17   1st Qu.: 78.45   1st Qu.: 74.56   1st Qu.: 69.36
##  Median :2012-08-26   Median : 94.34   Median : 92.17   Median : 94.16
##  Mean   :2012-08-26   Mean   : 90.16   Mean   : 86.75   Mean   : 90.01
##  3rd Qu.:2015-11-04   3rd Qu.:101.10   3rd Qu.:100.00   3rd Qu.:105.13
##  Max.   :2019-01-13   Max.   :106.69   Max.   :108.30   Max.   :117.26
##       SHY              TIP              TLT              DBC
##  Min.   :65.21   Min.   : 68.62   Min.   : 54.14   Min.   :11.92
##  1st Qu.:77.13   1st Qu.: 83.69   1st Qu.: 71.00   1st Qu.:17.04
##  Median :80.57   Median :103.62   Median : 96.30   Median :23.43
##  Mean   :78.57   Mean   : 96.78   Mean   : 92.61   Mean   :22.68
##  3rd Qu.:81.80   3rd Qu.:108.71   3rd Qu.:113.83   3rd Qu.:26.30
##  Max.   :83.62   Max.   :112.45   Max.   :134.72   Max.   :45.11
##       GLD              USO              IJH              IWM
##  Min.   : 56.99   Min.   :  8.33   Min.   : 35.47   Min.   : 30.54
##  1st Qu.: 93.73   1st Qu.: 14.71   1st Qu.: 69.84   1st Qu.: 62.51
##  Median :117.84   Median : 34.58   Median : 89.81   Median : 74.99
##  Mean   :114.75   Mean   : 34.99   Mean   :107.04   Mean   : 89.68
##  3rd Qu.:127.50   3rd Qu.: 39.56   3rd Qu.:139.99   3rd Qu.:112.25
##  Max.   :183.24   Max.   :117.39   Max.   :203.47   Max.   :171.99
##       SPY              VTV              XLB              XLE
##  Min.   : 56.2   Min.   : 22.35   Min.   :14.74   Min.   :30.90
##  1st Qu.:105.1   1st Qu.: 43.96   1st Qu.:28.78   1st Qu.:50.52
##  Median :123.8   Median : 52.52   Median :33.75   Median :61.05
##  Mean   :151.0   Mean   : 61.32   Mean   :36.80   Mean   :59.53
##  3rd Qu.:194.2   3rd Qu.: 77.25   3rd Qu.:45.13   3rd Qu.:67.59
##  Max.   :290.3   Max.   :111.68   Max.   :62.84   Max.   :88.75
##       XLF              XLI              XLK              XLP
```

```
##  Min.   : 3.200   Min.   :12.55   Min.   :11.56   Min.   :14.99
##  1st Qu.: 8.322   1st Qu.:27.42   1st Qu.:19.73   1st Qu.:20.74
##  Median :12.913   Median :32.44   Median :26.26   Median :29.86
##  Mean   :13.782   Mean   :40.52   Mean   :32.07   Mean   :33.05
##  3rd Qu.:16.881   3rd Qu.:51.81   3rd Qu.:40.36   3rd Qu.:45.49
##  Max.   :29.614   Max.   :79.51   Max.   :75.02   Max.   :57.01
##       XLU              XLV              XLY              IYR
##  Min.   :15.88   Min.   :18.74   Min.   : 14.06   Min.   :14.88
##  1st Qu.:23.73   1st Qu.:27.15   1st Qu.: 29.06   1st Qu.:41.81
##  Median :28.92   Median :35.05   Median : 41.46   Median :51.47
##  Mean   :32.57   Mean   :46.27   Mean   : 52.11   Mean   :53.08
##  3rd Qu.:39.97   3rd Qu.:67.28   3rd Qu.: 74.06   3rd Qu.:66.79
##  Max.   :56.34   Max.   :95.44   Max.   :116.76   Max.   :81.89
##       VNQ
##  Min.   :13.86
##  1st Qu.:39.18
##  Median :49.90
##  Mean   :52.62
##  3rd Qu.:69.29
##  Max.   :82.43
```

```
head(ETF.prices)
```

```
##         Time      AGG      IEF      LQD      SHY      TIP      TLT      DBC
## 1 2006-04-09 65.09827 56.76505 60.70945 65.21442 68.62030 55.09219 23.18837
## 2 2006-04-16 65.11148 57.07544 60.98924 65.30439 69.29211 55.32708 24.19736
## 3 2006-04-23 65.32363 57.01193 61.00089 65.36169 69.18821 54.90950 23.71601
## 4 2006-04-30 65.17391 56.81649 60.83233 65.35593 69.04548 54.75883 23.79006
## 5 2006-05-07 64.93439 56.54039 60.71520 65.35593 69.11513 54.13662 24.85459
## 6 2006-05-14 65.36685 57.07851 61.15432 65.46262 69.48401 55.30253 23.53087
##     GLD   USO      IJH      IWM       SPY      VTV      XLB      XLE      XLF
## 1 59.50 68.82 65.63398 62.55068  99.18915 42.67178 24.63047 42.91956 15.42653
## 2 63.20 72.81 67.69395 64.32922 101.06951 43.65933 25.71436 45.94183 15.73049
## 3 65.09 69.62 67.35899 63.92653 101.31612 43.97420 25.04160 43.94236 16.12946
## 4 67.99 68.00 68.38899 65.03390 102.12528 44.59678 25.89375 45.26510 16.22445
## 5 71.12 69.11 66.19505 61.66143  99.59763 43.43749 25.25837 43.63475 15.76849
## 6 65.58 65.57 64.09319 60.31911  97.94840 42.66463 24.10720 41.54298 15.48352
##        XLI      XLK      XLP      XLU      XLV      XLY      IYR      VNQ
## 1 25.99025 18.21236 16.62211 18.97177 24.89937 27.91007 40.64932 36.12251
## 2 26.78085 18.17940 16.73646 19.60541 25.01174 27.96821 42.06361 37.39742
## 3 26.50452 18.00634 17.02945 19.64933 24.87528 28.31710 41.89862 37.17645
## 4 27.38724 18.06403 17.19381 20.18887 24.70672 28.74904 42.19325 37.41442
## 5 26.81155 17.28938 16.91511 19.61169 24.39367 28.31710 40.82611 36.26982
## 6 26.12841 16.97623 16.87938 19.50503 24.30538 27.97652 40.23091 35.74854
```

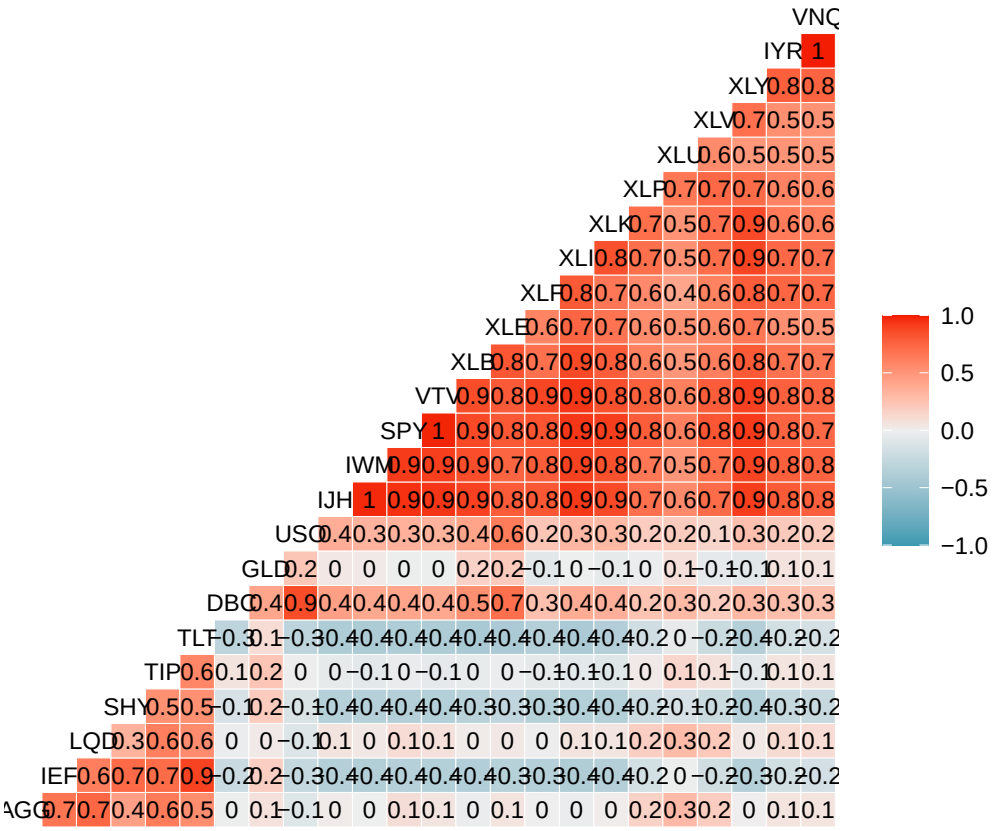
```
n = nrow(ETF.prices)
p = ncol(ETF.prices)

ETF_returns = ((data.matrix(ETF.prices[2:n,2:p]) - data.matrix(ETF.prices[1:(n-1),2:p]))/data.matrix(ETF

ETF_returns = cbind(ETF.prices$Time[2:n],as.data.frame(ETF_returns))
names(ETF_returns)[1] <- "Time"
head(ETF_returns)
```

```
##         Time           AGG          IEF           LQD           SHY
```

```
## 2 2006-04-16  0.0002029854  0.005468119  0.0046086562  1.379541e-03
## 3 2006-04-23  0.0032582579 -0.001112755  0.0001910173  8.773836e-04
## 4 2006-04-30 -0.0022919577 -0.003428160 -0.0027633038 -8.812502e-05
## 5 2006-05-07 -0.0036751668 -0.004859505 -0.0019255222  0.000000e+00
## 6 2006-05-14  0.0066600305  0.009517515  0.0072324729  1.632430e-03
## 7 2006-05-21  0.0018327944  0.001736643 -0.0002868154  1.629770e-03
##            TIP          TLT         DBC          GLD         USO          IJH
## 2  0.009790208  0.004263435  0.04351289  0.062184891  0.05797730  0.031385785
## 3 -0.001499406 -0.007547480 -0.01989275  0.029904984 -0.04381260 -0.004948152
## 4 -0.002062938 -0.002743879  0.00312249  0.044553728 -0.02326922  0.015291069
## 5  0.001008842 -0.011362752  0.04474696  0.046036257  0.01632354 -0.032080195
## 6  0.005337094  0.021536346 -0.05325876 -0.077896524 -0.05122270 -0.031752524
## 7  0.002705270 -0.002605885  0.02478361 -0.007319365  0.03324691  0.001959927
##            IWM          SPY          VTV          XLB         XLE          XLF
## 2  0.028433553  0.018957346  0.023142976  0.044006110  0.07041725  0.019704242
## 3 -0.006259722  0.002439954  0.007212043 -0.026162816 -0.04352184  0.025362649
## 4  0.017322447  0.007986479  0.014157800  0.034029699  0.03010175  0.005889347
## 5 -0.051857111 -0.024750474 -0.025994947 -0.024538233 -0.03601792 -0.028103259
## 6 -0.021769201 -0.016558939 -0.017792487 -0.045575589 -0.04793810 -0.018072367
## 7  0.010292443  0.010071201  0.010566879  0.009612687  0.01999286  0.010429283
##           XLI          XLK          XLP          XLU          XLV          XLY
## 2  0.03041930 -0.001809924  0.006879332  0.033399474  0.004513127  0.0020832268
## 3 -0.01031827 -0.009519512  0.017506275  0.002239841 -0.005455758  0.0124745559
## 4  0.03330440  0.003203872  0.009651340  0.027458601 -0.006776284  0.0152536095
## 5 -0.02102023 -0.042883509 -0.016209326 -0.028589021 -0.012670681 -0.0150244326
## 6 -0.02547943 -0.018112389 -0.002112136 -0.005438440 -0.003619504 -0.0120273611
## 7  0.00293776  0.007767273  0.019898066  0.017369210  0.008916627 -0.0002970705
##            IYR          VNQ
## 2  0.034792391  0.035294060
## 3 -0.003922369 -0.005908803
## 4  0.007032022  0.006401258
## 5 -0.032401913 -0.030592375
## 6 -0.014578809 -0.014372443
## 7  0.016552173  0.018544674
```

```
R = cor(ETF_returns[,2:p])
ggcorr(ETF_returns,label = TRUE, label_size=3, cex=3)
```

```
## Warning in ggcorr(ETF_returns, label = TRUE, label_size = 3, cex = 3): data in
## column(s) 'Time' are not numeric and were ignored
```




## 3. Analiza glavnih komponenti Cilj ovog zadatka je analizirati kretanje danih ETF-ova i izračunati glavne komponente koje objašnjavaju njihovu dinamiku.

### 3.1. Glavne komponente

Izračunajte glavne komponente matrice korelacije i izračunajte koliki udio varijance objašnjavaju. Odredite broj glavnih komponenti koje ćete zadržati u analizi. Grafički prikažite i usporedite koeficijente prvih nekoliko komponenti.

```
ev_R = eigen(R)

lambda_R = ev_R$values
e_R = ev_R$vectors

lambda_R
```

```
##  [1] 12.267970767  4.080795585  2.110540958  1.031704443  0.732784909
##  [6]  0.659925665  0.500415286  0.398476759  0.359793030  0.296415864
## [11]  0.291520255  0.243574145  0.231955205  0.186011616  0.121919175
## [16]  0.115569222  0.099948510  0.095157238  0.075167525  0.051743893
## [21]  0.021456721  0.018277559  0.005052089  0.003823582
```

```
sum(lambda_R)
```

```
## [1] 24
```

```
tr(R)
```

```
## [1] 24
```

```r
sum(lambda_R[1:3])/sum(lambda_R)
```

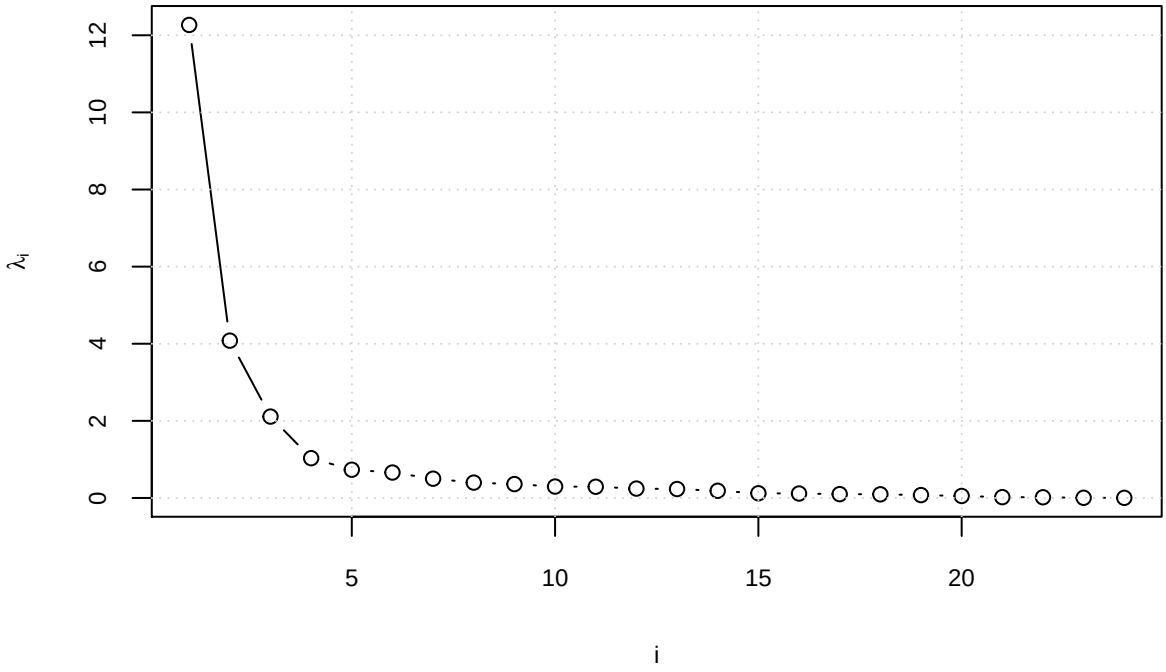
```
## [1] 0.7691378
```

```r
sum(lambda_R[1:4])/sum(lambda_R)
```

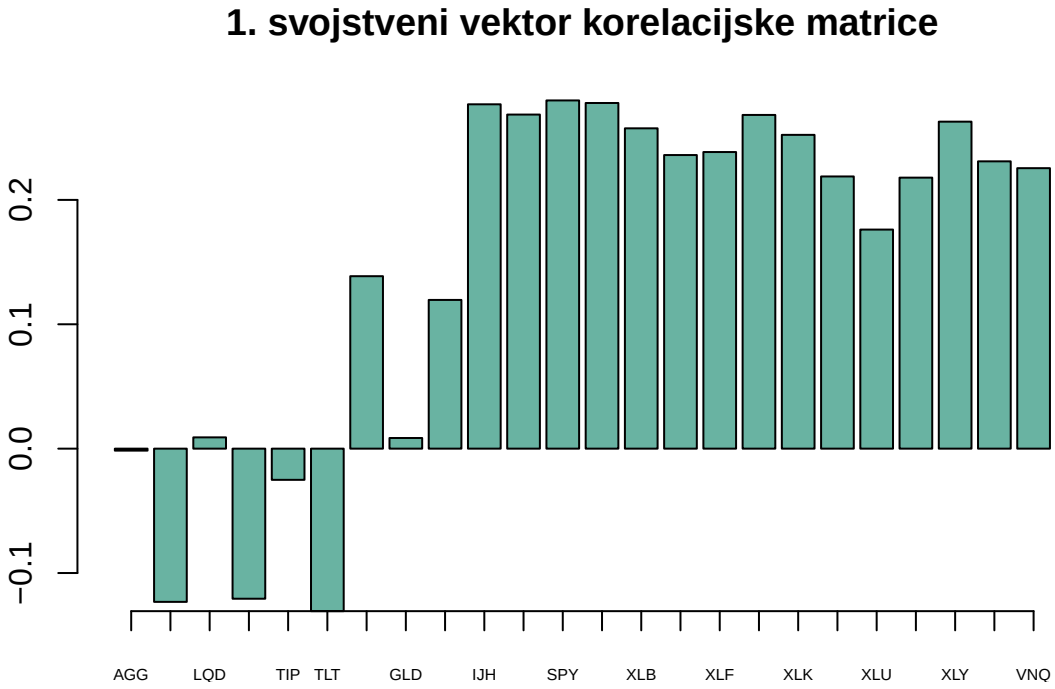
```
## [1] 0.8121255
```

```r
#scree plot za glavne komponente kovarijance
plot(lambda_R, type = "b", cex.lab=0.75, cex.main=0.75, cex.axis=0.75, xlab="i", ylab=expression(lambda
grid()
```

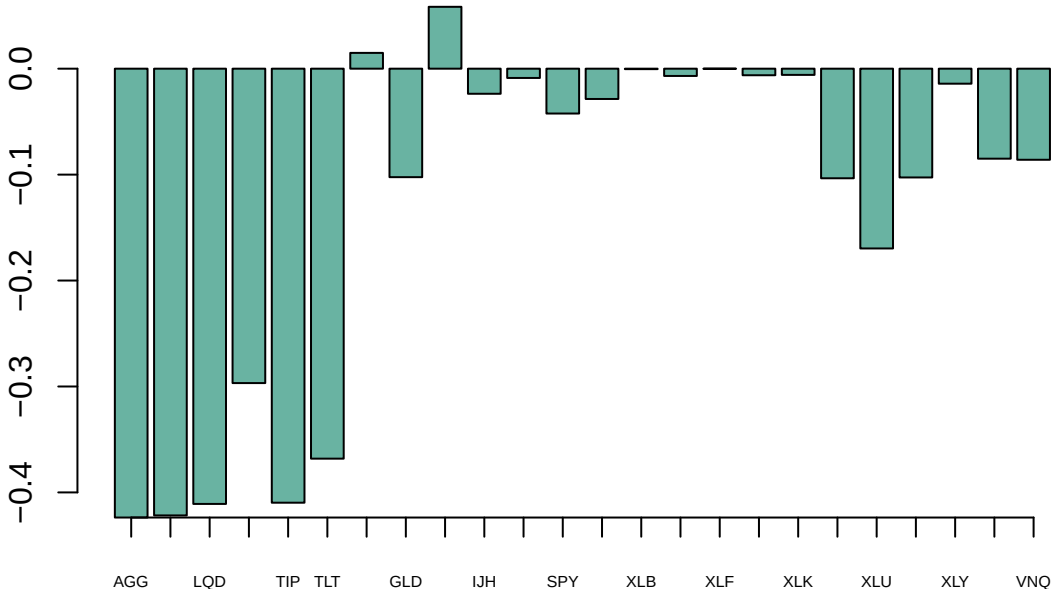
**Scree plot svojstvenih vrijednosti korelacijske matrice**



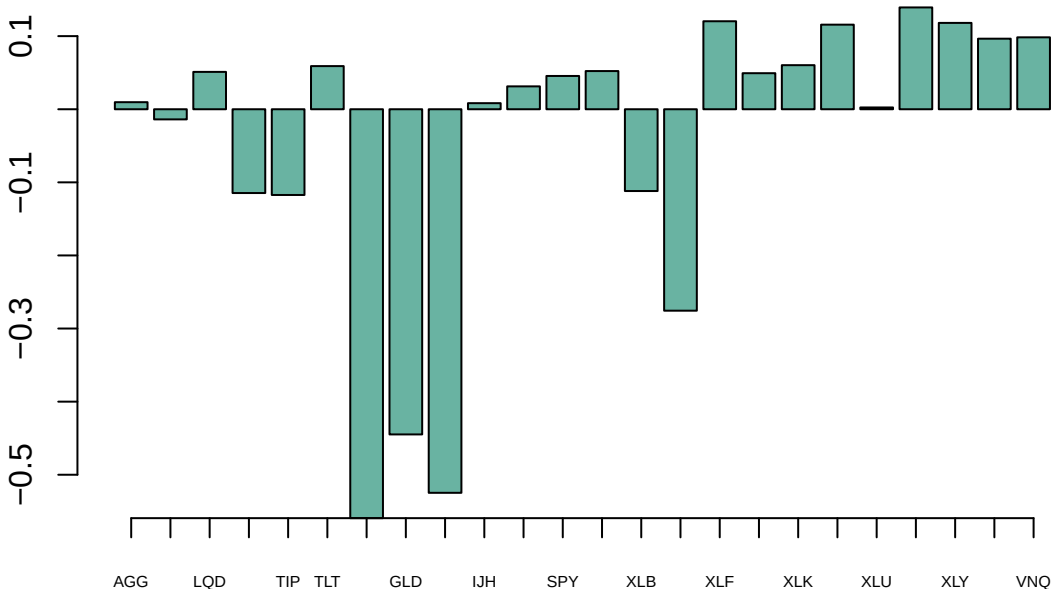

```r
midpts <- barplot(e_R[,1], col="#69b3a2", main="1. svojstveni vektor korelacijske matrice") # assign re
axis(1, at = midpts, labels=colnames(ETF_returns[,2:p]), cex.axis=0.47) # shrinks axis labels
```



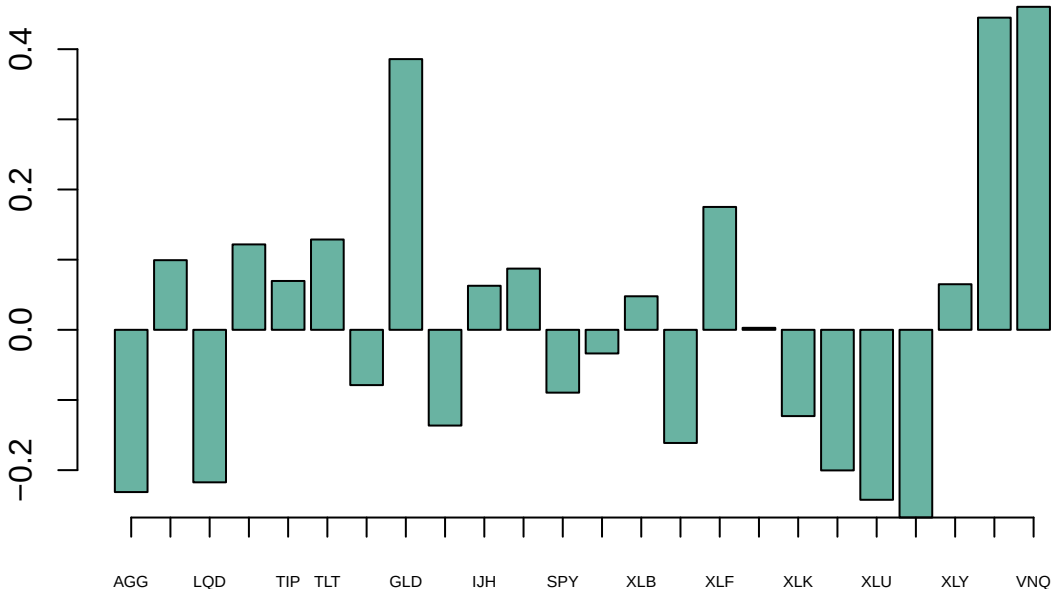

```
midpts <- barplot(e_R[,2], col="#69b3a2", main="2. svojstveni vektor korelacijske matrice") # assign re
axis(1, at = midpts, labels=colnames(ETF_returns[,2:p]), cex.axis=0.47) # shrinks axis labels
```




```
midpts <- barplot(e_R[,3], col="#69b3a2", main="3. svojstveni vektor korelacijske matrice") # assign re
axis(1, at = midpts, labels=colnames(ETF_returns[,2:p]), cex.axis=0.47) # shrinks axis labels
```




```
midpts <- barplot(e_R[,4], col="#69b3a2", main="4. svojstveni vektor korelacijske matrice") # assign re
axis(1, at = midpts, labels=colnames(ETF_returns[,2:p]), cex.axis=0.47) # shrinks axis labels
```




Prikažite graf raspršenja prve dvije glavne komponente i proučite možete li primijetiti neke grupe fondova.

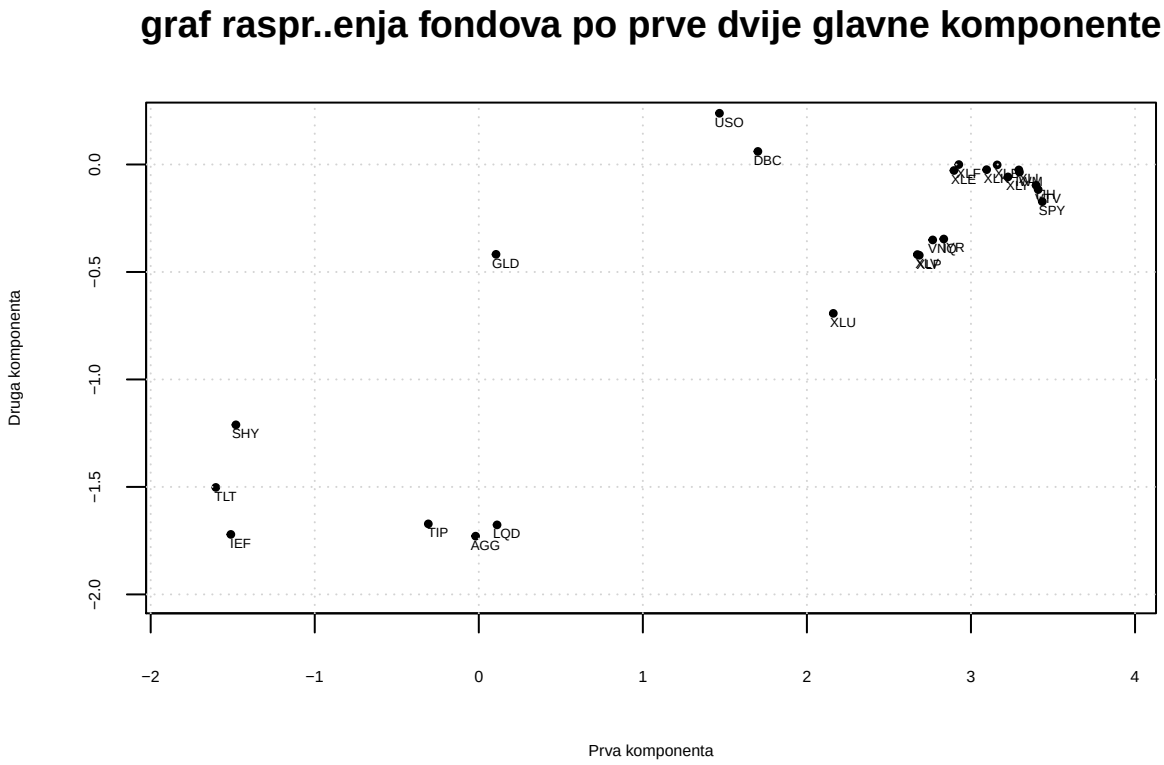
```
Y = R%*%e_R
plot(Y[,1],Y[,2], pch = 20, main="graf raspršenja fondova po prve dvije glavne komponente", cex=0.7, ce
```

```
## Warning in title(...): conversion failure on 'graf raspršenja fondova po prve
## dvije glavne komponente' in 'mbcsToSbcs': dot substituted for <c5>
```

```
## Warning in title(...): conversion failure on 'graf raspršenja fondova po prve
```

```
## dvije glavne komponente' in 'mbcsToSbcs': dot substituted for <a1>
```

```
grid()
#moguce je dodati i tekst - pritom treba pripaziti na citljivost
text(Y[,1]+0.06,Y[,2]-0.04, colnames(ETF_returns[,2:p]),cex=0.4)
```




### 3.2. Svojstveni portfelji

U primjeni PCA i svojstvenoj dekompoziciji kovarijance u financijama, svojstveni vektori se često zovu i tzv. svojstveni portfelji. Općenito, portfelj je vektor $w = [w_1, ..., w_N]$ u kojem svaki element predstavlja težinu ili udio kapitala u određenoj vrijednosnici. Često je dobro pomnožiti njihove težine s predznakom njihove sume - na taj način zapravo samo "okrećemo" predznak svojstvenog vektora tako da mu je suma pozitivna (konačni PCA rastav je i dalje isti ako svojstveni vektor pomnožimo s -1). Također, dobro je i skalirati svojstvene portfelje sa sumom njihovih apsolutnih vrijednosti: $\tilde{w}_i = \frac{w_i}{\sum_j^N |w_j|}$. Na taj način se osigurava da visoke magnitude pojedinih elemenata ne uzrokuju velike razlike u volatilnostima svojstvenih portfelja. Ukoliko znamo povrate $R \in \mathbb{R}^{T \times N}$ (gdje je $R_i \in \mathbb{R}^T$ vektor povrata za vrijednosnicu $i$) za $N$ vrijednosnica u nekom vremenskom periodu od $T$ dana, povrate portfelja $w$ u tom istom periodu možemo izračunati kao: $R_p = \sum R_i w_i = R \cdot w$. Izračunajte skalirane svojstvene portfelje $\tilde{w}$ koji proizlaze iz prve dvije glavne komponente. Za ta dva svojstvena portfelja izračunajte povijesne povrate kroz razmatrani period. Grafički prikažite vremensko kretanje njihovih vrijednosti tako da njihove povrate "vratite" natrag u cijene, s tim da početna cijena bude jednaka za oba portfelja, npr. $V_0 = 100$. Vrijednost portfelja u trenutku $t$ možemo izračunati po formuli: $V_t = V_{t-1} \cdot (1 + R_t)$.

```
#  Vaš kôd ovdje
```

## 4. Faktorska analiza

### 4.1. Metode procjena koeficijenata modela

Na danim podacima odredite broj faktora te procijenite faktorski model pomoću metode glavnih komponenti i metode najveće izglednosti. Usporedite procjene ove dvije metode. Koja Vam se čini bolja? Što možete

zaključiti iz vrijednosti faktora? Pronađite procjenu vrijednosti faktora koja daje najbolju interpretabilnost.

```
#  Vaš kôd ovdje
```

### 4.2. Specifične varijance faktora

Izračunajte specifične varijance faktora za model s dva faktora i model s tri faktora. Pomoću stupčastog dijagrama prikažite i usporedite dobivene vrijednosti.

```
#  Vaš kôd ovdje
```

## 5. Diskriminantna analiza

Financijska tržišta su od listopada 2007. do srpnja 2009. godine bila u krizi. U datoteci "crisis.csv" za svaki tjedan iz prethodno učitanih povijesnih tjednih cijena možete pronaći je li tržište tada bilo u krizi ili ne - 1 predstavlja krizu, 0 predstavlja period bez krize. Učitajte nove podatke te ih spojite s tablicom povrata.

```
#  Vaš kôd ovdje
```

### 5.1. Diskriminantna analiza pomoću povrata

Provedite diskriminantnu analizu koja tjedne odvaja na krizne i one bez krize pomoću povrata fondova. Pomoću stupčastog dijagrama prikažite vektore srednjih vrijednosti u krizi i izvan nje. Također, na isti način prikažite korelaciju fonda AGG (Aggregate Bond ETF-a) s ostalim fondovima u krizi i izvan krize. Usporedite rezultate linearne diskriminantne analize (funkcija u R-u: `lda`) i kvadratne diskriminantne analize (funkcija u R-u: `qda`) pomoću tablica konfuzije i mjere APER (eng. apparent error rate). Razmislite o tome koji je razlog razlike u rezultatima ove dvije metode.

```
#  Vaš kôd ovdje
```

### 5.2. Diskriminantna analiza pomoću glavnih komponenti

Provedite diskriminantnu analizu kao u prošlom podzadatku, no ovaj put koristeći glavne komponente izračunate u 3. zadatku kao varijable. Provjerite i usporedite uspješnost klasifikacije koristeći tablice konfuzije i APER za različit broj komponenti.

```
#  Vaš kôd ovdje
```