Analiza multivarijatnih podataka - primjene u financijama

Stjepan Begušić, Zvonko Kostanjčar

Sveučilište u Zagrebu
Fakultet elektrotehnike i računarstva

Ak. god. 2023./2024.

 v1.0

Danas. . .

1. Uvod u financijska tržišta – podatci, povrat i rizik investicija
2. Neprekidne slučajne varijable – pojmovi iz vjerojatnosti i statistike
3. Portfelji i diverzifikacija rizika
4. Financijski vremenski nizovi i njihova statistička svojstva
5. Analiza glavnih komponenti

Financijska tržišta – uvod

- **Tržište** – mjesto gdje se susreću kupci i prodavatelji kako bi razmjenili dobra ili usluge
- **Financijska tržišta** – razmjena kapitala (novac, dionice, obveznice, financijske vrijednosnice i izvedenice)
- **Burza** – organizirano (najčešće elektroničko) tržište gdje se trgovanje odvija po posebnim pravilima

Zašto investitori sudjeluju u financijskim tržištima:

- rast kapitala
- upravljanje rizikom

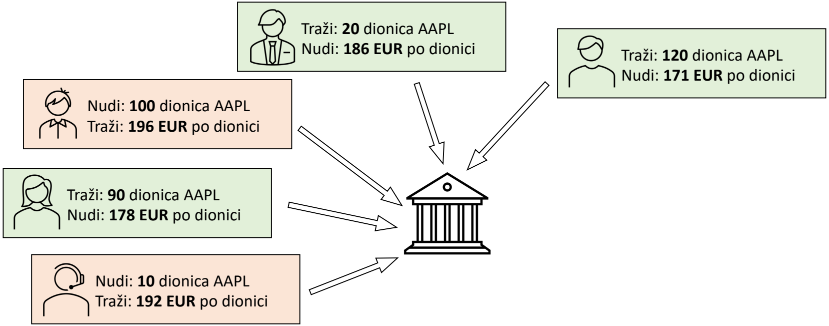
Čemu financijska tržišta doprinose u modernoj ekonomiji:

- likvidnost za uspješnu razmjenu kapitala
- definiranje cijena (price discovery)

Financijska tržišta – burze nekad i danas

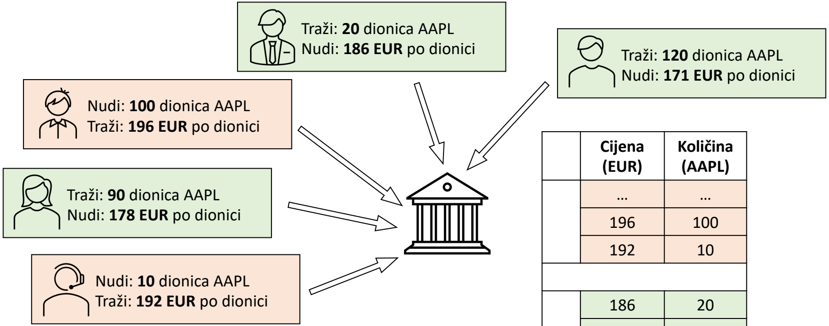
Financijska tržišta – formiranje cijena u elektroničkim burzama

Ograničeni nalozi (limit order):



Financijska tržišta – formiranje cijena u elektroničkim burzama

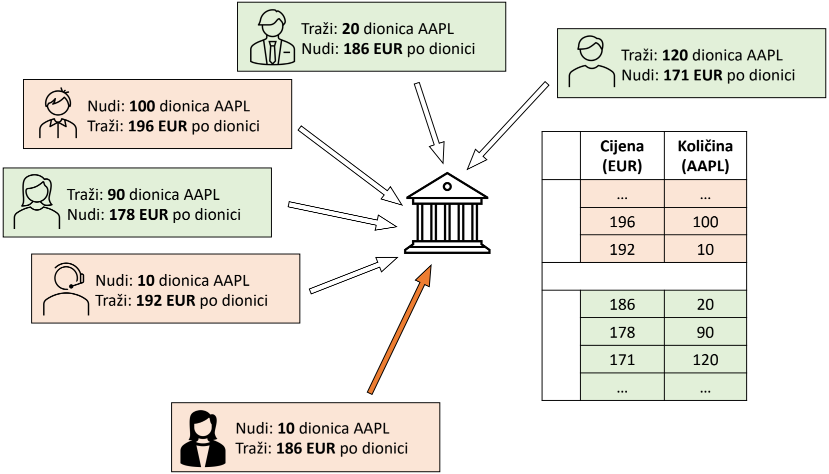
Knjiga ograničenih naloga (limit order book):



	Cijena (EUR)	**Količina (AAPL)**
	…	…
	196	100
	192	10
	186	20
	178	90
	171	120
	…	…

Financijska tržišta – formiranje cijena u elektroničkim burzama

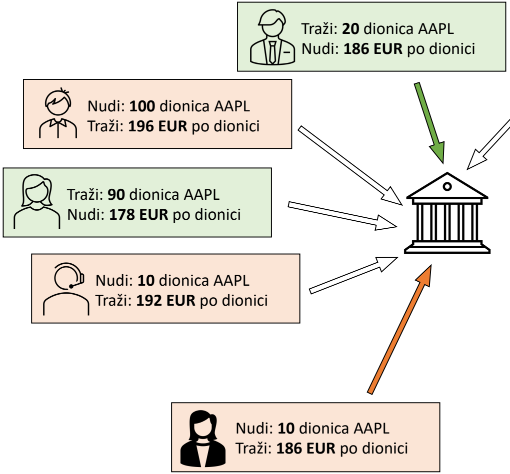
Burza djeluje kao posrednik koji spaja kupca i prodavatelja:



	Cijena (EUR)	**Količina (AAPL)**
	…	…
	196	100
	192	10
	186	20
	178	90
	171	120
	…	…

Financijska tržišta – formiranje cijena u elektroničkim burzama

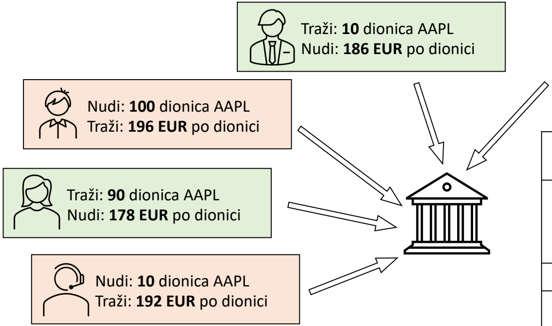
Burza djeluje kao posrednik koji spaja kupca i prodavatelja:



	Cijena (EUR)	**Količina (AAPL)**
	…	…
	196	100
	192	10
	186	**20**
	178	90
	171	120
	…	…

Financijska tržišta – formiranje cijena u elektroničkim burzama

Trgovanje:



	Cijena (EUR)	**Količina (AAPL)**
	...	...
	196	100
	192	10
	186	10
	178	90
	171	120
	...	...

Trgovina: **10** dionica AAPL prodano/kupljeno po cijeni **186 EUR** po dionici.

Financijska tržišta – formiranje cijena u elektroničkim burzama

Slika: Preuzeto s `http://finance.yahoo.com`.

Financijska tržišta – cijene i povrati

Uzorkovanje cijena u fiksnim vremenskim intervalima (minute, sati, dani, mjeseci itd.):

- Open, High, Low, Close (OHLC)
- U kvantitativnoj analizi često se koriste zadnje cijene u intervalima

Povrati

Aritmetički (linearni) povrat:

$$R(t) = \frac{P(t) - P(t-1)}{P(t-1)}$$

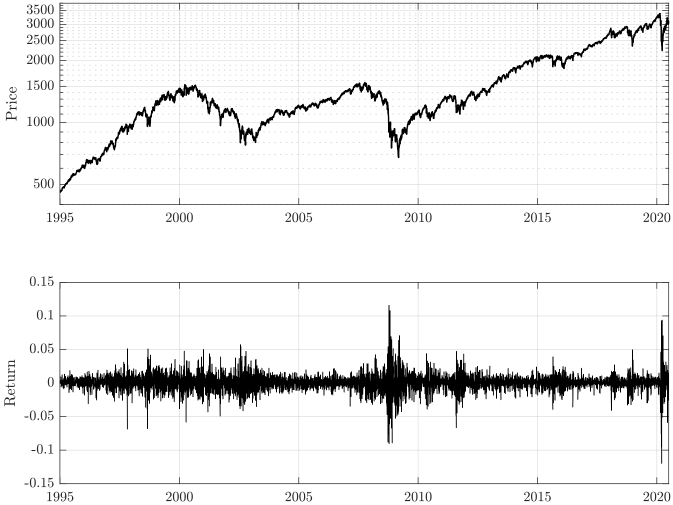
Logaritamski (kontinuirani) povrat:

$$r(t) = \log\left(\frac{P(t)}{P(t-1)}\right)$$

Zašto u analizi koristimo povrate, a ne cijene?

- Razina cijene vrijednosnice (npr. dionice) ne govori ništa o veličini ili vrijednosti kompanije
- Cijene s vremenom rastu ili padaju (nisu stacionarne)

Financijska tržišta – cijene i povrati



Slika: Cijene (iznad) i povrati (ispod) indeksa S&P 500

Svojstva aritmetičkih i logaritamskih povrata

Izračun cijene (iz prethodne cijene i povrata)

- Aritmetički:

$$P(t) = P(t-1)(1 + R(t))$$

- Logaritamski:

$$P(t) = P(t-1)\exp(r(t))$$

Ukupni (kumulativni) povrat kroz T vremenskih perioda

- Aritmetički:

$$R_{uk} = (1 + R(1)) \cdot (1 + R(2)) \cdot \ldots \cdot (1 + R(T)) - 1 = \prod_{t=1}^{T} (1 + R(t)) - 1$$

- Logaritamski:

$$r_{uk} = r(1) + r(2) + \ldots + r(T) = \sum_{t=1}^{T} r(t)$$

Investiranje

Investicija: Odricanje (danas) od sredstava (novac) na neko vrijeme kako bi se ostvarili budući povrati koji će kompenzirati investitora za

- **Vrijeme** na koji su sredstva uložena
  - 100 EUR danas "vrijedi" više nego 100 EUR za godinu dana
- Očekivanu stopu **inflacije**
  - Želimo da naša investicija pokrije rast cijena i troškova koje tim sredstvima plaćamo
- **Rizik**: nesigurnost budućih isplata
  - Za investiciju s nesigurnijim budućim isplatama (većim rizikom) tražimo veći povrat
  - Od dvije investicije jednakog (očekivanog) povrata preferiramo onu s manjom nesigurnošću (rizikom)

Povrat i rizik

- Početna vrijednost investicije: 100.000 EUR, horizont: 1 godina.
- Investicija A

Scenarij	povrat	vjerojatnost povrata	krajnja vrijednost
Optimistični	100%	0.2	200.000
Neutralni	0%	0.6	100.000
Pesimistični	-50%	0.2	50.000

- Investicija B

Scenarij	povrat	vjerojatnost povrata	krajnja vrijednost
Optimistični	20%	0.3	120.000
Neutralni	0%	0.6	100.000
Pesimistični	-10%	0.1	90.000

Povrat i rizik

- U ovim primjerima povrat je diskretna slučajna varijabla

$$R \sim \begin{pmatrix} r_1 & r_2 & \dots & r_N \\ p_1 & p_2 & \dots & p_N \end{pmatrix}$$

- Očekivani povrat investicije:

$$\mathrm{E}[R] = \sum_{i=1}^{N} r_i p_i$$

- Rizik investicije – standardna devijacija (volatilnost):

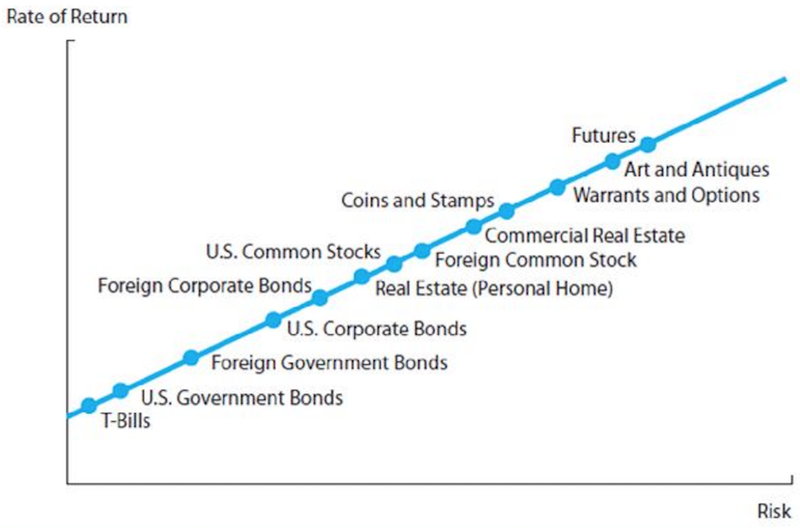
$$\sqrt{\mathrm{Var}[R]} = \sqrt{\sum_{i=1}^{N} p_i (r_i - \mathrm{E}[r_i])^2}$$

- Koliki su očekivani povrati i rizici investicije A i B? Koju investiciju odabrati?
- Općenito se povrat modelira s neprekidnom slučajnom varijablom s odgovarajućom gustoćom

Osnovne investicijske klase

- Instrumenti fiksnog prinosa (obveznice)
  - suštinski slično kreditu
  - ugovorena dinamika isplata investitoru
- Dionice
  - vlasnički udjel u trgovačkom društvu
  - prihod od dividende i/ili kapitalne dobiti
- Nekretnine
  - vlasnički udjel u stambenom/poslovnom objektu ili zemlji
  - prihod od najma i/ili kapitalne dobiti
- Izvedenice
  - vrijednost im ovisi o nekom drugom financijskom instrumentu
  - opcije, unaprijedni ugovori, swapovi, itd
- Alternativna ulaganja
  - venture investicije, umjetnine, collectibles
  - niska likvidnost

Investicijske klase



Danas. . .

Funkcija distribucije

Neprekidna slučajna varijabla

Neka je $(\Omega, \mathfrak{F}, \mathrm{P})$ vjerojatnosni prostor. Preslikavanje $X : \Omega \to \mathbb{R}$ nazivamo slučajna varijabla ako je za svaki $x \in \mathbb{R}$ skup $\{\omega \in \Omega : X(\omega) < x\}$ događaj, dakle element od $\mathfrak{F}$.

Funkcija distribucije slučajne varijable

Neka je $(\Omega, \mathfrak{F}, \mathrm{P})$ vjerojatnosni prostor te neka je $X : \Omega \to \mathbb{R}$ slučajna varijabla. Funkcija $F = F_X : \mathbb{R} \in [0, 1]$ definirana formulom

$$F(x) = F_X(x) = \mathrm{P}(X < x), x \in \mathbb{R},$$

zove se funkcija distribucije slučajne varijable X.

Funkcija gustoće

Funkcija gustoće slučajne varijable

Slučajna varijabla $X : \Omega \to \mathbb{R}$ je neprekidna slučajna varijabla ako postoji funkcija $f = f_X : \mathbb{R} \to [0, \infty]$ takva da vrijedi

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt, \forall x \in \mathbb{R}$$

i f_X se zove vjerojatnosna funkcija gustoće.

Egzistencija i zadavanje neprekidne slučajne varijable

Svaka funkcija gustoće neprekidne slučajne varijable ima svojstva

$$f(x) \geq 0, \forall x \in \mathbb{R} \texttt{ te } \int_{-\infty}^{\infty} f(x)dx = 1.$$

Vrijedi i obrat.

Očekivanje i varijanca

Očekivanje

Neka je $X : \Omega \to \mathbb{R}$ neprekidna slučajna varijabla s funkcijom gustoće f_X. Ako integral $\int_{-\infty}^{\infty} |x| f_X(x) dx$ postoji, tada kažemo da slučajna varijabla ima matematičko očekivanje i definiramo

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx.$$

Varijanca

Neka je $X : \Omega \to \mathbb{R}$ neprekidna slučajna varijabla s funkcijom gustoće f_X. Ako integral $\int_{-\infty}^{\infty} x^2 f_X(x) dx$ postoji, tada kažemo da slučajna varijabla ima varijancu i definiramo

$$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right].$$

Kovarijanca

Kovarijanca

Neka su $X, Y : \Omega \to \mathbb{R}$ slučajne varijable koje imaju varijancu, tada se njihova kovarijanca definirana s

$$\mathrm{Cov}(X, Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right].$$

Svojstva očekivanja i varijance

Neka su $X, Y : \Omega \to \mathbb{R}$ neprekidne slučajne varijable koje imaju očekivanje, odnosno varijancu. Tada vrijedi

$$\mathrm{E}\left[aX + bY\right] = a\mathrm{E}[X] + b\mathrm{E}[Y].$$

$$\mathrm{Var}\left[aX + bY\right] = a^2\mathrm{Var}[X] + b^2\mathrm{Var}[Y] + 2ab\mathrm{Cov}(X, Y).$$

Korelacija

Pearsonov koeficijent korelacije

Neka su $X, Y : \Omega \to \mathbb{R}$ slučajne varijable koje imaju varijance σ_X^2, σ_Y^2, tada je njihov Pearsonov koeficijent korelacije dan s

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Svojstva koeficijenta korelacije

Neka je ρ koeficijent korelacije slučajnih varijabli X i Y. Tada je $\rho \in [-1, 1]$ te vrijedi:

- $\rho = 0$ - slučajne varijable X i Y nisu linearno povezane (kažemo da su nekorelirane)
- $\rho = 1$ - slučajne varijable su pozitivno korelirane i postoji linearna veza između njih
- $\rho = -1$ - slučajne varijable su negativno korelirane i postoji linearna veza između njih

Očekivanje, (ko)varijanca i korelacija slučajnog vektora

Za slučajni vektor $X = [X_1, X_2, ...X_N]'$ definiramo:

- Vektor očekivanja

  $$\mu_X = [\mu_1, \mu_2, ..., \mu_N]'$$

  gdje je $\mu_i = \mathrm{E}[X_i]$.

- Matrica kovarijance

  $$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,N} \\ \sigma_{1,2} & \sigma_2^2 & \cdots & \sigma_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,N} & \sigma_{2,N} & \cdots & \sigma_N^2 \end{bmatrix}$$

  gdje je $\sigma_i^2 = \mathrm{Var}[X_i], \sigma_{i,j}^2 = \mathrm{Cov}(X_i, X_j)$.

- Korelacijska matrica

  $$\mathbf{C} = \begin{bmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,N} \\ \rho_{1,2} & 1 & \cdots & \rho_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,N} & \rho_{2,N} & \cdots & 1 \end{bmatrix}$$

Danas. . .

Diverzifikacija rizika – portfelji

- Nesigurnost budućih isplata ili vrijednosti pojedinih investicija može se smanjiti kombiniranjem više investicija
- Npr. postoji vjerojatnost da će propasti kompanija u čiju dionicu investiramo – no ako odaberemo 100 različitih dionica/kompanija, vjerojatnost da će sve kompanije propasti je puno manja!

Primjer - razmotrimo sljedeću investiciju (portfelj):

- U trenutku t_0: 1 dionica A vrijednosti 100 EUR i 1 dionica B vrijednosti 300 EUR
- U trenutku t_1 dionica A padne 10%, dionica B naraste 4%
- Koja je vrijednost portfelja u t_1?
- Koliki je povrat portfelja?

Portfelj

Portfelj

Portfelj je linearna kombinacija investicija s udjelima $w = [w_1, w_2, ..., w_N]'$, čiji povrat je dan s

$$R_p = \sum_{i=1}^{N} w_i R_i, \quad \sum_{i=1}^{N} w_i = 1,$$

gdje je R_i povrat i-te investicije, a w_i udjel i-te investicije.

Očekivani povrat portfelja:

$$\mathrm{E}[R_p] = \sum_{i=1}^{N} w_i \mathrm{E}[R_i] = w' \mu_R$$

Rizik portfelja (standardna devijacija, volatilnost):

$$\sqrt{\mathrm{Var}[R_p]} = \sqrt{\sum_{i=1}^{N} w_i^2 \mathrm{Var}(R_i) + \sum_{i,j=1, i \neq j}^{N} w_i w_j \mathrm{Cov}(R_i, R_j)} = \sqrt{w' \mathbf{\Sigma} w}$$

Napomena: ovi izrazi vrijede samo za aritmetičke povrate!

Diverzifikacija rizika u portfelju

Primjer: razmotrimo portfelj jednakih težina $w = [1/N, 1/N, ..., 1/N]$ od ukupno N nekoreliranih vrijednosnica koje imaju jednaku volatilnost σ. Kolika je volatilnost portfelja?

Svojstva varijance aritmetičke sredine slučajnih varijabli

Neka su X_i slučajne varijable koje imaju varijancu te neka je $Y = \frac{1}{n}\sum_{i=1}^{i=n} X_i$. Tada vrijedi

1. ako su varijable X_i nekorelirane

$$Var(Y) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} X_i = \frac{Var(X)}{n}$$

2. ako su varijable korelirane s prosječnom korelacijom ρ

$$Var(Y) = \frac{Var(X)}{n} + \frac{n-1}{n}\rho Var(X)$$

Moderna teorija portfelja

Optimizacijski problem:

$$\min_w w'\mathbf{\Sigma}w - \gamma w'\mu_R \quad \text{s.t. } 1'w = 1,$$

uz parametar $\gamma > 0$.

- Za svaku vrijednost γ postoji jedan optimalan portfelj
- Svi optimalni portfelji tvore **efikasnu granicu** (*efficient frontier*)

Druge formulacije problema:

- Minimizacija rizika za danu razinu očekivanog povrata:

$$\min_w w'\mathbf{\Sigma}w \quad \text{s.t. } w'\mu_R = \mu_P, 1'w = 1$$

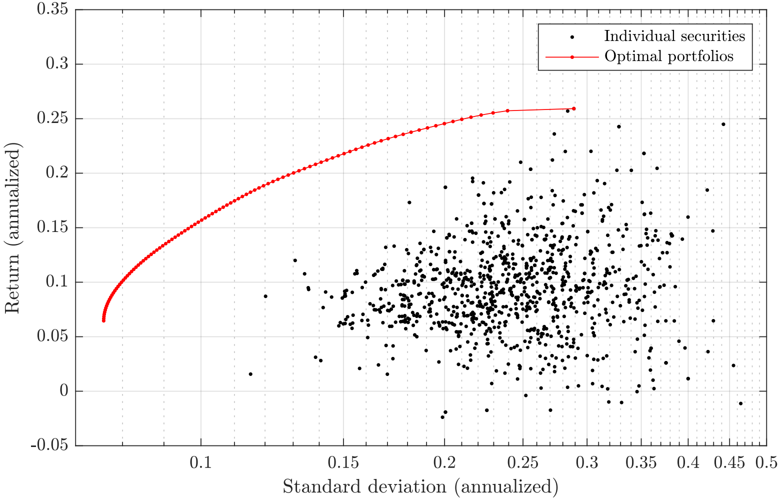
,

- Maksimizacija očekivanog povrata uz danu razinu rizika

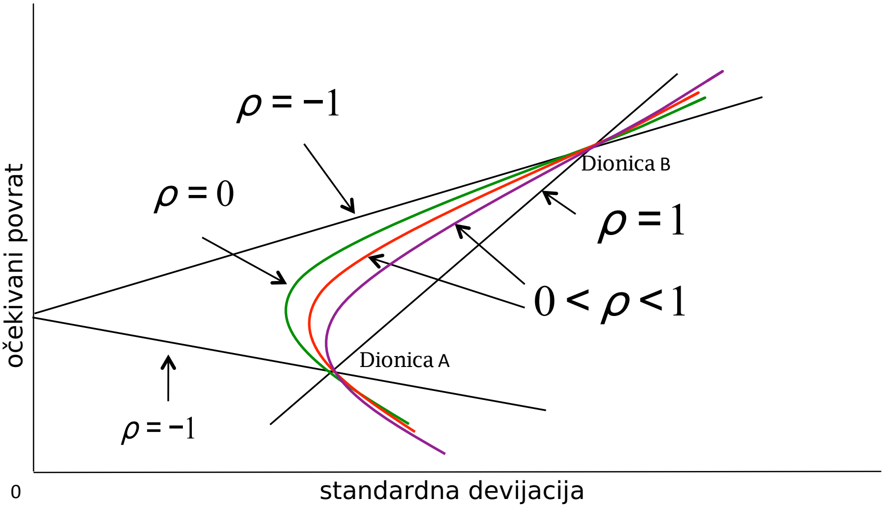
$$\max_w w'\mu_R \quad \text{s.t. } w'\mathbf{\Sigma}w = \sigma_P^2, 1'w = 1$$

.

Moderna teorija portfelja – efikasna granica



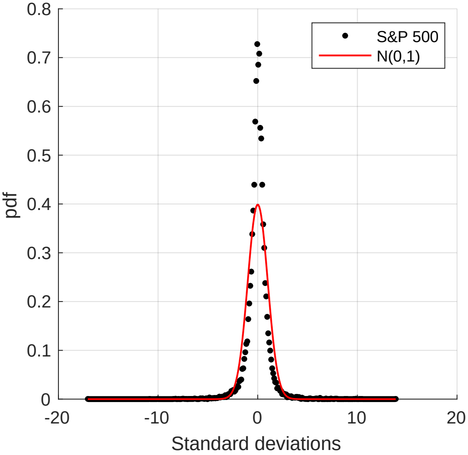
Primjer – efikasna granica za dvije dionice

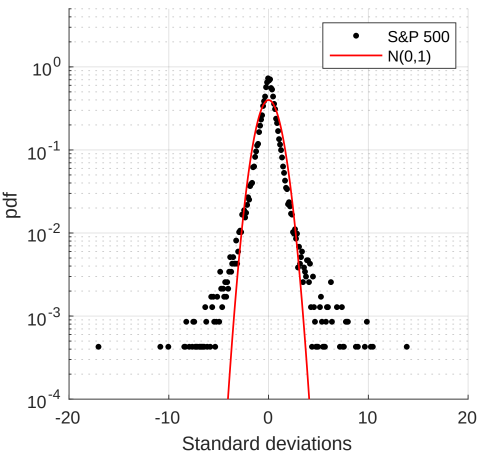


Danas. . .

1. Uvod u financijska tržišta – podatci, povrat i rizik investicija
2. Neprekidne slučajne varijable – pojmovi iz vjerojatnosti i statistike
3. Portfelji i diverzifikacija rizika
4. Financijski vremenski nizovi i njihova statistička svojstva
5. Analiza glavnih komponenti

Teški repovi u razdiobama povrata





Vremenska zavisnost povrata

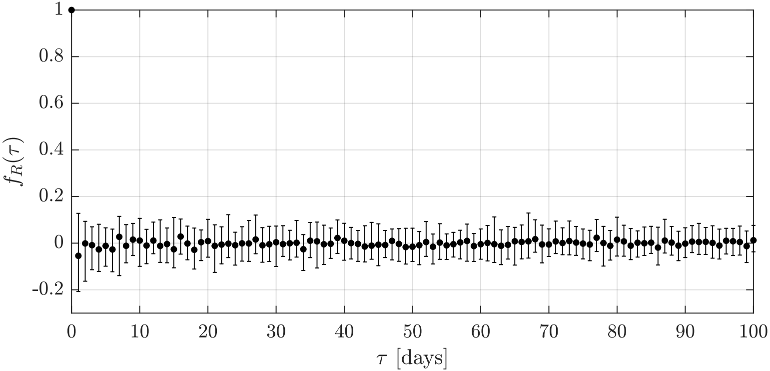
Normalizirana autokorelacijska funkcija

Neka je $\{X_i, i = 1, \ldots, T\}$ niz jednako distribuiranih slučajnih varijabli koje imaju očekivanje μ i varijancu σ^2. Normalizirana autokorelacijska funkcija definirana je izrazom

$$f_X(\tau) = \frac{Cov(X_t, X_{t+\tau})}{\sigma^2}$$

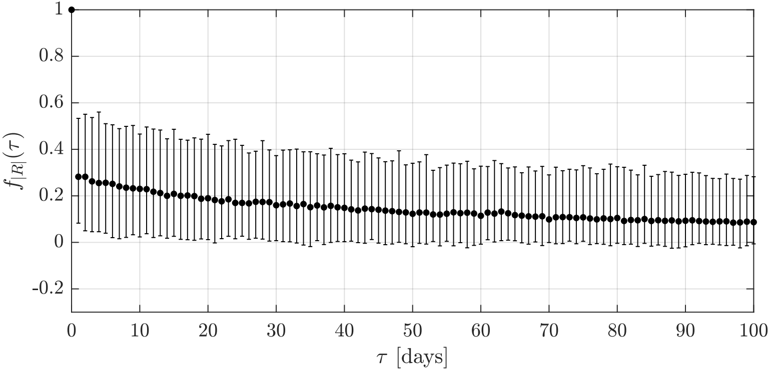
Vremenska zavisnost povrata

Autokorelacija povrata:



Vremenska zavisnost povrata

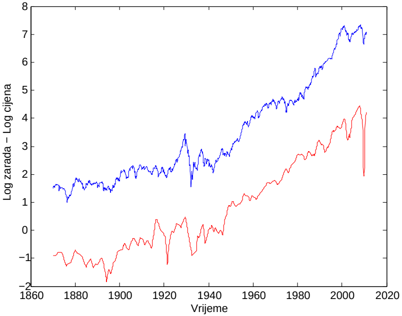
Autokorelacija apsolutnih povrata:



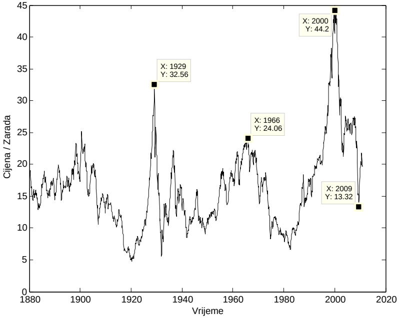
Financijski signali ili financijski vremenski nizovi

- Cijene financijskih instrumenata
  - cijene dionica
  - cijene obveznica
  - cijene opcija
  - itd
- Makroekonomske varijable
  - kamatne stope
  - tečajne liste (valute)
  - bruto domaći proizvod (BDP)
  - itd
- Podatci o kompanijama
  - zarada
  - promet
  - dug
  - itd

Odnos cijene i zarade

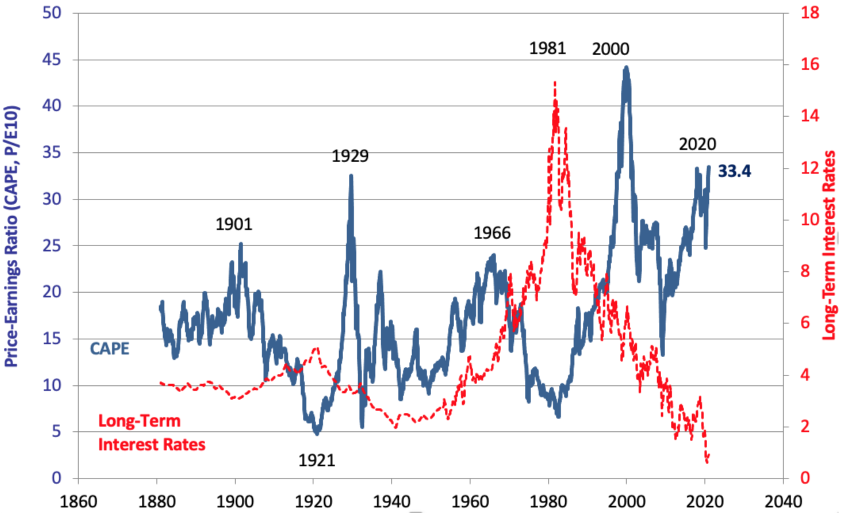


(a) Logaritam S&P 500 indeksa i logaritam odgovarajućih zarada



(b) P/E za S&P 500

Odnos cijene i zarade



Danas...

1. Uvod u financijska tržišta – podatci, povrat i rizik investicija
2. Neprekidne slučajne varijable – pojmovi iz vjerojatnosti i statistike
3. Portfelji i diverzifikacija rizika
4. Financijski vremenski nizovi i njihova statistička svojstva
5. Analiza glavnih komponenti

Analiza glavnih komponenti - uvod

Analiza glavnih komponenti bavi se objašnjavanjem kovarijancijske strukture skupa varijabli s nekoliko linearnih kombinacija tih varijabli.

Glavni ciljevi analize glavnih komponenti su:

- redukcija dimenzionalnosti
- interpretacija (idenfitikacija zajedničkih izvora varijabilnosti)

Algebarski, glavne komponente su određene linearne kombinacije slučajnih varijabli $X_1, \ldots, X_p$.
Geometrijski, te linearne kombinacije predstavljaju nove koordinatne osi dobivene rotacijom originalnog sustava s $X_1, \ldots, X_p$ kao originalnim osima.

Analiza glavnih komponenti - uvod

Linearne kombinacije slučajnih varijabli

- Slučajni vektor $\mathbf{X} = [X_1, X_2, ..., X_p]'$, s kovarijancom $\mathbf{\Sigma}$
- Razmatramo linearne kombinacije:

$$Y_1 = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p = \mathbf{a_1}'\mathbf{X}$$

$$Y_2 = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p = \mathbf{a_2}'\mathbf{X}$$

$$\vdots$$

$$Y_p = a_{p1}X_1 + a_{p2}X_2 + ... + a_{pp}X_p = \mathbf{a_p}'\mathbf{X}$$

- Varijanca i kovarijanca:

$$\mathrm{Var}(Y_i) = \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_i$$

$$\mathrm{Cov}(Y_i, Y_j) = \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_j$$

Glavne komponente

- Prva glavna komponenta:
  - linearna kombinacija $\mathbf{a_1}'\mathbf{X}$ koja maksimizira $\mathrm{Var}(Y_1) = \mathbf{a_1}'\mathbf{\Sigma}\mathbf{a_1}$,
  - uz uvjet $\mathbf{a_1}'\mathbf{a_1} = 1$.
- Druga glavna komponenta:
  - linearna kombinacija $\mathbf{a_2}'\mathbf{X}$ koja maksimizira $\mathrm{Var}(Y_2) = \mathbf{a_2}'\mathbf{\Sigma}\mathbf{a_2}$,
  - uz uvjete $\mathbf{a_2}'\mathbf{a_2} = 1$ i $\mathrm{Cov}(Y_1, Y_2) = \mathbf{a_1}'\mathbf{\Sigma}\mathbf{a_2} = 0$.
- i-ta glavna komponenta:
  - linearna kombinacija $\mathbf{a}_i'\mathbf{X}$ koja maksimizira $\mathrm{Var}(Y_i) = \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_i$,
  - uz uvjete $\mathbf{a}_i'\mathbf{a}_i = 1$ i $\mathrm{Cov}(Y_i, Y_j) = \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_j = 0, \quad \forall j < i$.

Glavne komponente

Glavne komponente - svojstvena dekompozicija kovarijance

Neka je $\mathbf{\Sigma}$ kovarijacijska matrica slučajnog vektora $\mathbf{X} = [X_1, X_2, ..., X_p]'$, s parovima svojstvenih vrijednosti i svojstvenih vektora: $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), ..., (\lambda_p, \mathbf{e}_p)$, tako da vrijedi $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$. Vrijedi svojstvena dekompozicija kovarijance:

$$\mathbf{\Sigma} = \sum_{i=1}^{p} \lambda_i \mathbf{e}_i \mathbf{e}_i' = \mathbf{E}\mathbf{\Lambda}\mathbf{E}', \quad \mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_p], \mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_p)$$

Onda je i-ta glavna komponenta:

$$Y_i = e_{i1}X_1 + e_{i2}X_2 + ... + e_{ip}X_p = \mathbf{e}_i'\mathbf{X}.$$

Varijanca i kovarijanca komponenti:

$$\mathrm{Var}(Y_i) = \mathbf{e}_i'\mathbf{\Sigma}\mathbf{e}_i = \lambda_i, \quad i = 1, 2, ..., p.$$

$$\mathrm{Cov}(Y_i, Y_j) = \mathbf{e}_i'\mathbf{\Sigma}\mathbf{e}_j = 0, \quad \forall i \neq j.$$

Glavne komponente

Glavne komponente - dekompozicija varijance

Zbroj varijanci glavnih komponenti jednak je zbroju varijanci originalnih varijabli:

$$\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_p^2 = \sum_{i=1}^{p} \mathrm{Var}(X_i) = \lambda_1 + \lambda_2 + \ldots + \lambda_p = \sum_{i=p}^{p} \mathrm{Var}(Y_i).$$

- Postotak varijance koju objašnjava i-ta glavna komponenta:

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}.$$

- Korelacijski koeficijent komponente Y_i i varijable X_j:

$$\rho_{Y_i, X_j} = \frac{e_{ij}\sqrt{\lambda_i}}{\sqrt{\sigma_{jj}}}$$

Svojstvene vrijednosti i vektori - izračun

Za svojstveni vektor $\mathbf{e}$ i pripadajuću svojstvenu vrijednost λ_i neke matrice $\mathbf{\Sigma}$ vrijedi:

$$\mathbf{\Sigma}\mathbf{e} = \lambda\mathbf{e}$$

odnosno: $(\mathbf{\Sigma} - \lambda I)\mathbf{e} = 0$. Sustav ima netrivijalno rješenje kad vrijedi $|\mathbf{\Sigma} - \lambda I| = 0$.

- Odredimo determinantu s nepoznatom λ, rješavanjem jednadžbe odredimo svojstvene vrijednosti.
- Svaku λ_i uvrstimo i rješavanjem sustava uz $\mathbf{e}'\mathbf{e} = 1$ izračunamo elemente pripadajućeg svojstvenog vektora.

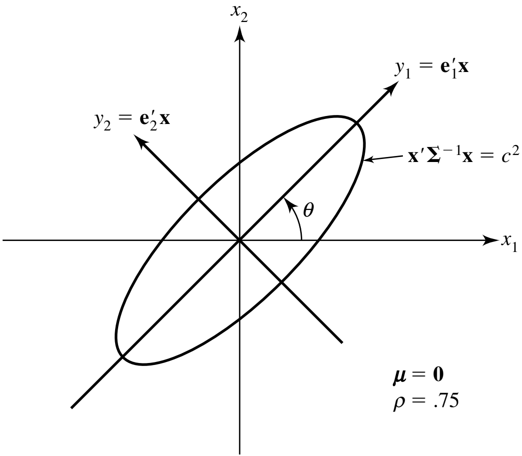
Primjer

- Kovarijanca dvije varijable X_1 i X_2:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 3 \\ 3 & 9 \end{bmatrix}$$

- Odredite glavne komponente.
- Koliki postotak ukupne varijance objašnjava pojedina komponenta?
- Koja je korelacija pojedine komponente i varijable?

Glavne komponente multivarijatne normalne varijable



Glavne komponente standardiziranih varijabli

- Razmotrimo standardizirane varijable:

$$Z_1 = \frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}}$$

$$Z_2 = \frac{X_2 - \mu_2}{\sqrt{\sigma_{22}}}$$

$$\vdots$$

$$Z_p = \frac{X_p - \mu_p}{\sqrt{\sigma_{pp}}}$$

- U matričnoj notaciji: $\mathbf{Z} = \mathbf{V}^{-1/2}(\mathbf{X} - \mu)$, gdje je $\mathbf{V}$ dijagonalna matrica s varijancama σ_i^2 na dijagonali.
- Matrica kovarijance: $\mathrm{Cov}(\mathbf{Z}) = \mathbf{V}^{-1/2}(\mathbf{\Sigma})\mathbf{V}^{-1/2} = \mathbf{C}$, gdje je $\mathbf{C}$ korelacijska matrica slučajnog vektora $\mathbf{X}$.

Glavne komponente standardiziranih varijabli

Glavne komponente standardiziranih varijabli

Neka je $\mathbf{R}$ korelacijska matrica slučajnog vektora $\mathbf{X}$, s parovima svojstvenih vrijednosti i svojstvenih vektora: $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), ..., (\lambda_p, \mathbf{e}_p)$, tako da vrijedi $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$.
Onda je i-ta glavna komponenta standardiziranih varijabli $\mathbf{Z} = [Z_1, Z_2, ..., Z_p]'$:

$$Y_i = \mathbf{e}_i'\mathbf{Z} = \mathbf{e}_i'\mathbf{V}^{-1/2}(\mathbf{X} - \mu).$$

- Zbroj varijanci komponenti:

$$\sum_{i=1}^{p} \mathrm{Var}(Y_i) = \sum_{i=1}^{p} \mathrm{Var}(Z_i) = p$$

- Korelacijski koeficijent komponente Y_i i varijable Z_j:

$$\rho_{Y_i,Z_j} = e_{ij}\sqrt{\lambda_i}$$

Primjer

- Kovarijanca dvije varijable X_1 i X_2:

$$\mathbf{\Sigma} = \begin{bmatrix} 4 & 3 \\ 3 & 9 \end{bmatrix}$$

- Odredite korelacijsku matricu i glavne komponente standardiziranih varijabli (iz korelacijske matrice)? Kako se mijenjaju u odnosu na glavne komponente iz kovarijance?
- Mijenja li se postotak opisane varijance?
- Mijenjaju li se korelacije komponenti i pojedinih varijabli?

Glavne komponente uzorka

Populacijsku matricu kovarijance ili korelacije ne poznajemo – ali ih možemo procijeniti iz podataka:

- Podatci: n opservacija p-dimenzionalnog slučajnog vektora $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}]$
- Uzoračka matrica kovarijance:

$$\hat{\mathbf{\Sigma}} = \frac{1}{n-1} \sum_{i=1}^{n} \left[(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' \right], \quad \overline{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

- Uzoračka matrica korelacije:

$$\hat{\mathbf{C}} = \mathbf{D}^{-1} \hat{\mathbf{\Sigma}} \mathbf{D}^{-1}, \quad \mathbf{D} = \mathrm{diag}(\hat{\sigma}_1, \hat{\sigma}_2, ..., \hat{\sigma}_p)$$

- Koristimo svojstvenu dekompoziciju uzoračkih procjena matrica u analizi.

Broj glavnih komponenti

- Komponente koje objašnjavaju jako malo varijance nisu bitne u analizi.
  - Što ako je $n < p$, koliko ne-nul svojstvenih vrijednosti postoji?
- Pri odabiru broja glavnih komponenti dobro je imati na umu:
  - Količina i udio objašnjene varijance
  - Relativne vrijednosti objašnjene varijance
  - Interpretacija komponenti
- Nekoliko metoda za odabir broja komponenti:
  - Zadržavamo sve komponente koje objašnjavaju bar neki udio varijance (npr. barem 5%).
  - Zadržavamo broj komponenti koji ukupno objašnjava neki udio ukupne varijance (npr. 90%).
  - *Scree plot* - tražimo nagli pad u postotku objašnjene varijance.

Svojstva glavnih komponenti

- Rotacije komponenti:
  - Moguće je promijeniti predznak svojstvenim vektorima (pomnožiti s -1).
  - Radi interpretacije, vektori se mogu rotirati tako da je većina elemenata pozitivno, ili da je zbroj elemenata pozitivan.
- Rekonstrukcija podataka:
  - Svaka observacija $\mathbf{x}_j$ se može rekonstruirati kao linearna kombinacija prvih q glavnih komponenti:

$$\hat{\mathbf{x}}_j = \hat{y}_{j1}\hat{\mathbf{e}}_1 + \hat{y}_{j2}\hat{\mathbf{e}}_2 + \ldots + \hat{y}_{jq}\hat{\mathbf{e}}_q = (\mathbf{x}_j'\hat{\mathbf{e}}_1)\hat{\mathbf{e}}_1 + (\mathbf{x}_j'\hat{\mathbf{e}}_2)\hat{\mathbf{e}}_2 + \ldots + (\mathbf{x}_j'\hat{\mathbf{e}}_q)\hat{\mathbf{e}}_q$$

  - Rekonstrukcija će biti savršena ako koristimo svih $q = p$ komponenti, u protivnom, za $q < p$ će postojati rekonstrukcijska greška koja iznosi $\hat{y}_{jq+1}\hat{\mathbf{e}}_{q+1} + \ldots + \hat{y}_{jp}\hat{\mathbf{e}}_p$.
  - Rekonstrukcijska greška korisna je pri analizi potencijalnih stršećih vrijednosti.

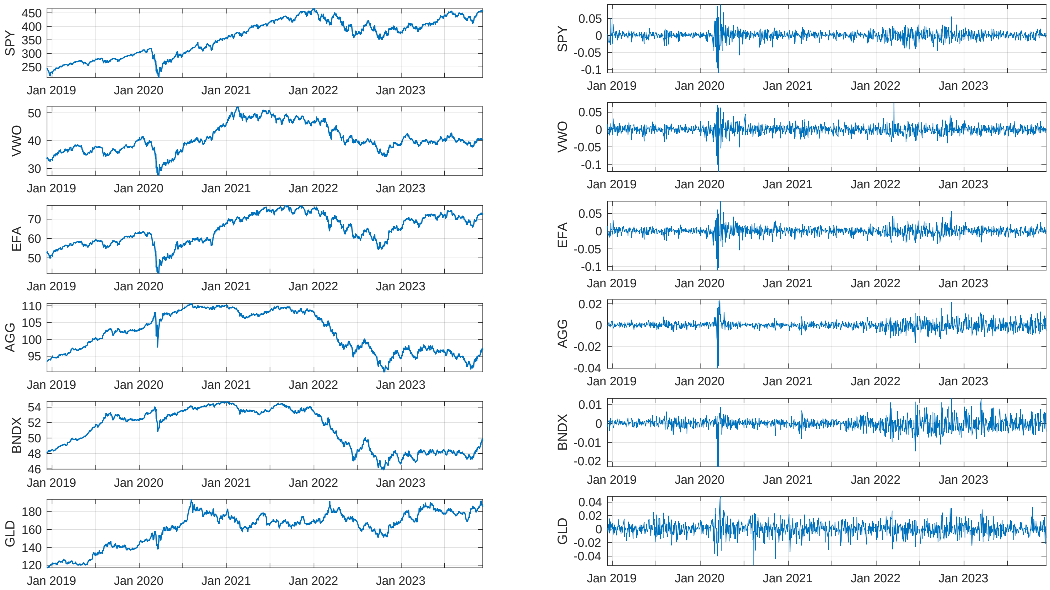
PCA u financijama

Razmatramo skup imovina (fondova):

Oznaka	Naziv	Vrsta imovine
SPY	*SPDR S&P 500 ETF Trust*	Dionice - USA
AGG	*iShares Core U.S. Aggregate Bond ETF*	Obveznice - USA
VWO	*Vanguard FTSE Emerging Markets ETF*	Dionice - tržišta u razvoju
EFA	*iShares MSCI EAFE ETF*	Dionice - razvijena tržišta izvan USA
BNDX	*Vanguard Total International Bond ETF*	Obveznice - razvijena tržišta izvan USA
GLD	*SPDR Gold Trust*	Zlato

- Analiziramo dnevne cijene u zadnjih 5 godina (ukupno 1258 dnevnih cijena za svaki fond)
- Koristimo (aritmetičke) dnevne povrate koje računamo iz cijena.

PCA u financijama - cijene i povrati



PCA u financijama - kovarijanca i korelacija povrata

Uzoračka kovarijanca povrata:

$$\begin{bmatrix} 0.1785 & 0.1378 & 0.1465 & 0.0084 & 0.0067 & 0.0126 \\ & 0.1809 & 0.1401 & 0.0080 & 0.0057 & 0.0232 \\ & & 0.1552 & 0.0097 & 0.0062 & 0.0217 \\ & & & 0.0168 & 0.0096 & 0.0150 \\ & & & & 0.0093 & 0.0101 \\ & & & & & 0.0882 \end{bmatrix} \cdot 10^{-3}$$

- Teško interpretirati, u praksi se često preferiraju korelacije ili anualizirane volatilnosti

PCA u financijama - kovarijanca i korelacija povrata

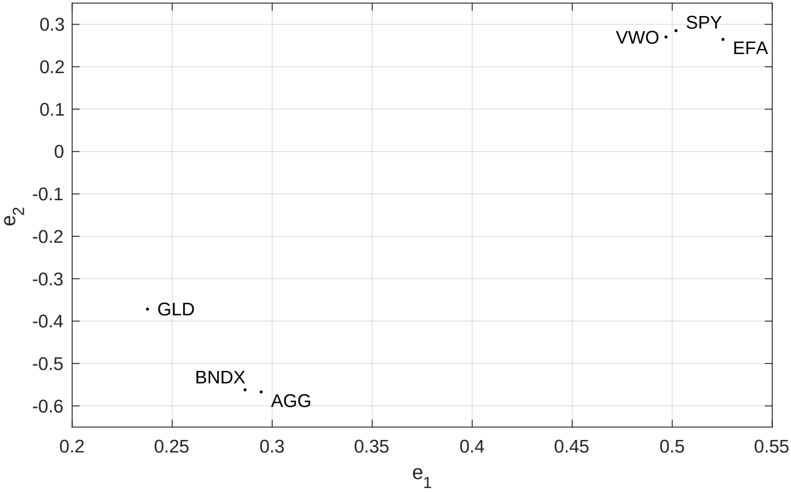
Uzoračka korelacija:

PCA u financijama - svojstvena dekompozicija korelacijske matrice

- Koeficijenti glavnih komponenti:

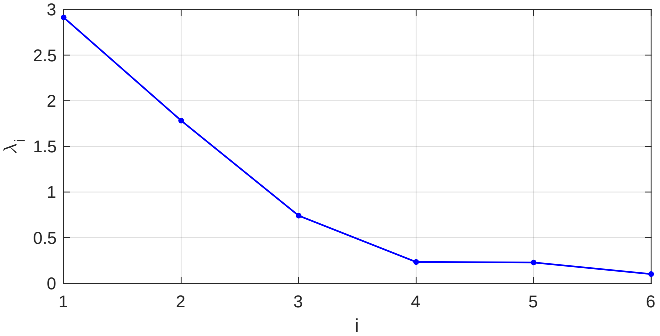
	$\mathbf{e}_1$	$\mathbf{e}_2$	$\mathbf{e}_3$	$\mathbf{e}_4$	$\mathbf{e}_5$	$\mathbf{e}_6$
SPY	0.50	0.29	-0.11	-0.49	-0.33	0.55
VWO	0.50	0.27	0.05	0.57	0.53	0.25
EFA	0.53	0.26	0.00	-0.07	-0.19	-0.78
AGG	0.29	-0.57	-0.28	0.49	-0.52	0.09
$BNDX$	0.29	-0.56	-0.34	-0.42	0.55	-0.09
GLD	0.24	-0.37	0.89	-0.11	-0.03	0.05
Varijanca	2.91	1.78	0.74	0.24	0.23	0.10
Varijanca [%]	48.5%	29.7%	12.4%	3.9%	3.8%	1.7%
Varijanca - kum. [%]	48.5%	78.2%	90.6%	94.5%	98.3%	100.0%

PCA u financijama - vizualizacija podataka u novom prostoru



PCA u financijama - broj komponenti u analizi

Scree plot:



PCA u financijama - komponente kao portfelji

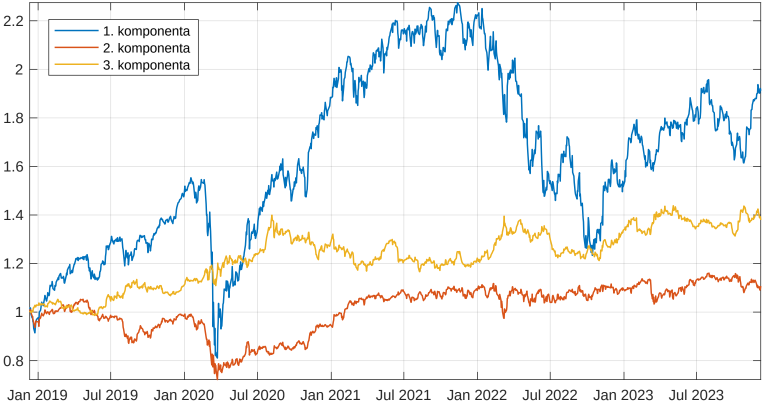
Glavne komponente su portfelji (tzv. svojstveni portfelji):

- Negativne težine predstavljaju kratke pozicije.
- Za potrebe analize koeficijente glavnih komponenti možemo rotirati:
  - tako da suma težina portfelja bude pozitivna
  - tako da srednji povrat portfelja bude pozitivan
- Ponekad želimo dodatno skalirati težine (npr. tako da vrijedi $\sum w_i = 1$ ili $\sum |w_i| = 1$).

Primjer – težine rotirane tako da srednji povrat svakog portfelja bude pozitivan, bez skaliranja:

	$\mathbf{w}_1$	$\mathbf{w}_2$	$\mathbf{w}_3$
SPY	0.50	0.29	−0.11
VWO	0.50	0.27	0.05
EFA	0.53	0.26	0.00
AGG	0.29	−0.57	−0.28
$BNDX$	0.29	−0.56	−0.34
GLD	0.24	−0.37	0.89

PCA u financijama - komponente kao portfelji



Literatura

- Johnson, R. A., and D. W. Wichern, Applied Multivariate Statistical Analysis - poglavlja 8.1 - 8.4