Kompleksnost skrivenih Markovljevih modela

Mišura, Ankica

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Ankica Mišura

KOMPLEKSNOST SKRIVENIH MARKOVLJEVIH MODELA

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, rujan 2016.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Hvala mentoru doc.dr.sc. Pavlu Goldsteinu na strpljenju, savjetima i pomoći pri izradi diplomskog rada.
Hvala mojoj obitelji i prijateljima na podršci i razumijevanju tokom svih ovih godina studiranja.

Sadržaj

Uvod

Tema ovog diplomskog rada je bioinformatičke prirode. Bioinformatika je znanost koja se bavi analizom bioloških podataka o nizovima, sadržaju i organizaciji genoma te predviđa strukture i funkcije makromolekula uz pomoć tehnika iz primijenjene matematike, statistike i računarstva. U ovom diplomskom radu bavimo se skrivenim Markovljevim modelima, statističkim alatom koji danas ima primjenu u različitim područjima: prepoznavanju govora, rukopisa i gesta, računalnom prevođenju, analizi vremenskih nizova te bioinformatici.

Ovim radom želimo pružiti kratki pregled teorije skrivenih Markovljevih modela, dati primjer njihove implementacije i pojasniti postupke za procjenu parametara modela. Ujedno, promatramo neke metode za procjenu kompleksnosti skrivenih Markovljevih modela.

U prvom poglavlju je dan pregled osnovnih pojmova iz teorije vjerojatnosti, statistike i Markovljevih lanaca, te smo definirali Shannnovu entropiju i informacijske kriterije. U drugom poglavlju formalno definiramo skrivene Markovljeve modele i dajemo primjer, dok u trećem opisujemo algoritme koje smo koristili za procjenu parametara modela i maksimizaciju vjerodostojnosti. U četvrtom poglavlju prezentiramo rezultate i zaključke rada.

Poglavlje 1

Osnovni pojmovi

1.1 Vjerojatnost

Definicija 1.1.1. *Pod **slučajnim pokusom** podrazumijevamo takav pokus čiji **ishodi**, odnosno **rezultati** nisu jednoznačno određeni uvjetima u kojima izvodimo pokus. Rezultate slučajnog pokusa nazivamo **događajima**.*

Definicija 1.1.2. *Neka je A događaj vezan uz neki slučajni pokus. Pretpostavimo da smo taj pokus ponovili n puta i da se u tih n ponavljanja događaj A pojavio točno n_A puta. Tada broj n_A zovemo **frekvencija** događaja A, a broj $\frac{n_A}{n}$ **relativna frekvencija** događaja A.*

Definicija 1.1.3. *Osnovni objekt u teoriji vjerojatnosti jest neprazan skup Ω koji zovemo **prostor elementarnih događaja** i koji reprezentira skup svih ishoda slučajnih pokusa. Ako je Ω konačan ili prebrojiv, govorimo o **diskretnom** prostoru elementarnih događaja. Prostor elementarnih događaja je **kontinuiran** ako je Ω neprebrojiv skup.*
*Točke ω iz skupa Ω zvat ćemo **elementarni događaji***

Označimo sa $\mathcal{P}(\Omega)$ partitivni skup od Ω.

Definicija 1.1.4. *Familija $\mathcal{F}$ podskupova od Ω ($\mathcal{F} \subset \mathcal{P}(\Omega)$) jest σ – **algebra skupova** (na Ω) ako je:*

F1. $\emptyset \in \mathcal{F}$

F2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$

F3. $A_i \in \mathcal{F},\ i \in \mathbb{N} \Rightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{F}$

Definicija 1.1.5. *Neka je $\mathcal{F}$ σ-algebra na skupu Ω. Uređen par $(\Omega, \mathcal{F})$ se zove **izmjeriv prostor***

Sad možemo definirati vjerojatnost.

Definicija 1.1.6. *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor. Funkcija $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ jest* ***vjerojatnost*** *ako vrijedi:*

P1. $\mathbb{P}(\Omega) = 1$ *(normiranost vjerojatnosti)*

P2. $\mathbb{P}(A) \geq 0,\ A \in \mathcal{F}$ *(nenegativnost vjerojatnosti)*

P3. $A_i \in \mathcal{F},\ i \in \mathbb{N}\ \ i\ \ A_i \cap A_j = \emptyset\ \ za\ \ i \neq j \Rightarrow \mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ *(prebrojiva ili σ - aditivnost vjerojatnosti)*

Definicija 1.1.7. *Uređena trojka $(\Omega, \mathcal{F}, \mathbb{P})$ gdje je $\mathcal{F}$ $\sigma-$algebra na Ω i $\mathbb{P}$ vjerojatnost na $\mathcal{F}$, zove se* ***vjerojatnosni prostor****.*

Definicija 1.1.8. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Elemente $\sigma-$algebre zovemo* ***događaji****, a broj $\mathbb{P}(A),\ A \in \mathcal{F}$ se zove* ***vjerojatnost događaja*** *A.*

Budući da radimo sa slučajnim varijablama, potrebno je definirati otvoreni skup.

Definicija 1.1.9. *Neka je $x \in \mathbb{R}^n$ i $r > 0$. Skup*

$$\begin{aligned} K(x, r) &= \{y \in \mathbb{R}^n : d(x, y) < r\} \\ &= \left\{ y \in \mathbb{R}^n : \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} < r \right\} \end{aligned}$$

nazivamo ***otvorena kugla oko x radijusa r****. Skup $A \subset \mathbb{R}^n$ je* ***otvoren*** *ako vrijedi*

$$\forall x \in A, \exists r > 0, K(x, r) \subset A.$$

Otvorena okolina *točke $x \in \mathbb{R}^n$ je svaki otvoreni skup koji sadrži točku x.*

Definicija 1.1.10. *Označimo sa $\mathcal{B}$ $\sigma-$algebru generiranu familijom svih otvorenih skupova na skupu realnih brojeva $\mathbb{R}$. $\mathcal{B}$ zovemo $\sigma -$* ***algebra skupova na*** *$\mathbb{R}$, a elemente $\sigma-$algebre $\mathcal{B}$ zovemo* ***Borelovi skupovi****.*

Budući da je svaki otvoreni skup na $\mathbb{R}$ prebrojiva unija intervala, lako je dokazati da vrijedi

$$\mathcal{B} = \sigma\{(a, b);\ a, b \in \mathbb{R},\ a < b\}$$

Definicija 1.1.11. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Funkcija $X : \Omega \to \mathbb{R}$ jest* ***slučajna varijabla*** *(na Ω) ako je $X^{-1}(B) \in \mathcal{F}$ za proizvoljno $B \in \mathcal{B}$, odnosno $X^{-1}(\mathcal{B}) \subset \mathcal{F}$.*

Definicija 1.1.12. *Neka je* $(\Omega, \mathcal{F}, \mathbb{P})$ *proizvoljan vjerojatnosni prostor i* $A \in \mathcal{F}$ *takav da je* $\mathbb{P}(A) > 0$. *Definiramo funkciju* $P_A : \mathcal{F} \to [0,1]$ *ovako:*

$$P_A(B) = P(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)},\ B \in \mathcal{F}. \tag{1.1}$$

Lako je provjeriti da je P_A *vjerojatnost na* $\mathcal{F}$ *i nju zovemo* ***vjerojatnost od B uz uvjet A****.*

Definicija 1.1.13. *Neka je* $(\Omega, \mathcal{F}, \mathbb{P})$ *vjerojatnosni prostor i* $A_i \in \mathcal{F},\ i \in I$ *proizvoljna familija događaja. Kažemo da je to* ***familija nezavisnih događaja*** *ako za svaki konačan podskup različitih indeksa* $i_1, i_2, ..., i_k$ *vrijedi*

$$\mathbb{P}(\cap_{i=1}^{k} A_{i_j}) = \prod_{j=1}^{k} \mathbb{P}(A_{i_j}). \tag{1.2}$$

Neka je X slučajna varijabla na diskretnom vjerojatnosnom prostoru $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ i neka je

$$X = \begin{pmatrix} a_1 & a_2 & \dots \\ p_1 & p_2 & \dots \end{pmatrix}$$

njena distribucija, odnosno vrijedi $\mathbb{P}(a_i) = p_i$.

Definicija 1.1.14. ***Funkcija gustoće vjerojatnosti*** *od X ili, kraće,* ***gustoća*** *od X jest funkcija* $f_X = f : \mathbb{R} \to \mathbb{R}_+$ *definirana sa*

$$f(x) = \mathbb{P}\{X = x\} = \begin{cases} 0, & x \neq a_i \\ p_i, & x = a_i \end{cases}, x \in \mathbb{R}$$

Definicija 1.1.15. ***Funkcija distribucije slučajne varijable*** *X jest funkcija* $F_X = F : \mathbb{R} \to [0,1]$ *definirana sa*

$$F(x) = \mathbb{P}\{X \leq x\} = \mathbb{P}\{\omega; X(\omega) \leq x\},\ x \in \mathbb{R}.$$

1.2 Statistika

Definicija 1.2.1. *Za model* $\mathcal{T} = \{f(\cdot;\theta) : \theta \in \Theta\}, f(\cdot;\theta) : \mathbb{R} \to [0, +\infty >, \Theta \subset \mathbb{R}$ *kažemo da je regularan ako su zadovoljeni sljedeći uvjeti:*

i) $\sup f(\cdot;\theta) = \{x \in \mathbb{R} : f(x;\theta) > 0\}$ *ne ovisi o* $\theta \in \Theta$

ii) Θ *je otvoreni interval u* $\mathbb{R}$

iii) $\forall x \in \mathbb{R}$, $\theta \to f(x;\theta)$ *je diferencijabilna na* Θ

iv) Za slučajnu varijablu X kojoj je f funkcija gustoće vrijedi:

$$0 < I(\theta) := \mathbb{E}_\theta[(\frac{\partial}{\partial\theta}\log f(x;\theta))^2] < \infty$$

Broj $I(\theta)$ *se zove* ***Fisherova informacija****.*

v) $\forall\theta \in \Theta$, $\frac{d}{d\theta}\int_{\mathbb{R}} f(x;\theta)dx = \int_{\mathbb{R}} \frac{\partial}{\partial\theta} f(x;\theta)dx = 0$*, ako se radi o neprekidnoj slučajnoj varijabli, odnosno*
$\forall\theta \in \Theta$, $\frac{d}{d\theta}\sum_x f(x;\theta) = \sum_x \frac{\partial}{\partial\theta} f(x;\theta) = 0$*, ako je riječ o diskretnoj slučajnoj varijabli*[1].

Definicija 1.2.2. *Neka je* $(\Omega, \mathcal{F})$ *izmjeriv prostor i* $\mathcal{P}$ *familija vjerojatnosnih mjera na* $(\Omega, \mathcal{F})$*. Uređena trojka* $(\Omega, \mathcal{F}, \mathcal{P})$ *se zove* ***statistička struktura****.*

Definicija 1.2.3. *n-dimenzionalni* ***slučajni uzorak*** *na statističkoj strukturi* $(\Omega, \mathcal{F}, \mathcal{P})$ *je niz* $(X_1, ..., X_n)$ *slučajnih varijabli na izmjerivom prostoru* $(\Omega, \mathcal{F})$ *takav da su slučajne varijable* $X_1, \ldots, X_n$ *nezavisne i jednako distribuirane* $\forall\mathbb{P} \in \mathcal{P}$*.*

Definicija 1.2.4. *Neka je* $X = (X_1, ..., X_n)$ *slučajan uzorak iz modela* $\mathcal{P}$*,* $\mathcal{P} = \{f(\cdot;\theta) : \theta \in \Theta\}, \Theta \subset \mathbb{R}^m$*. Ako je* $X = (X_1, ..., X_n)$ *jedna realizacija od* $\mathbb{X}$*, tada je* ***vjerodostojnost*** *funkcija* $L : \Theta \to \mathbb{R}$

$$L(\theta) = L(\theta|\mathbb{X}) := \prod_{i=1}^{n} f(X_i;\theta)$$

Statistika $\hat{\theta} = \hat{\theta}(\mathbb{X})$ *je procjenjitelj maksimalne vjerodostojnosti (****MLE****) ako vrijedi*

$$L(\hat{\theta}|\mathbb{X}) = \max_{\theta\in\Theta} L(\theta|\mathbb{X})$$

Definicija 1.2.5. *Za opaženu vrijednost x od* $\mathbb{X}_n$*,* $l : \Theta \to \mathbb{R}$*,*

$$l(\theta) = l(\theta|\mathbb{X}) = \log L(\theta|\mathbb{X}) = \sum_{i=1}^{n} \log f(x_i : \theta)$$

je ***log-vjerodostojnost****.*

Definicija 1.2.6. *Procjenitelj* $T = t(X)$ *za* $\tau(\theta) \in \mathbb{R}$ *je* ***nepristran*** *ako vrijedi*

$$\forall\theta \in \Theta,\ \mathbb{E}_\theta(T) = \tau(\theta)$$

. Procjenitelj koji nije nepristran je ***pristran****.*

[1] Prisjetimo se: slučajna varijabla je diskretna ako je definirana na diskretnom vjerojatnosnom prostoru, a neprekidna ukoliko joj je funkcija gustoće nenegativna realna funkcija

Definicija 1.2.7. *T je **efikasan** procjenitelj za $\tau(\theta)$ ako je nepristran i vrijedi*

$$Var_\theta = \frac{[\tau'(\theta)]^2}{nI(\theta)}, \forall\theta \in \Theta$$

Definicija 1.2.8. *Niz procjenitelja $(T_n : n \in \mathbb{N})$ je **konzistentan** procjenitelj za θ ako za proizvoljni $\epsilon > 0$ vrijedi*

$$\lim_{n\to\infty} \mathbb{P}_\theta\{|T_n - \theta| \geq \epsilon\} = 0$$

Teorem 1.2.9. *Neka je $\mathbb{X}_n = (X_1, ..., X_n)$ slučajan uzorak iz regularnog modela $\mathcal{P}$, uz dodatnu pretpostavku da je $\theta \to f(x;\theta)$ neprekidno diferencijabilna. Tada jednadžba vjerodostojnosti*

$$\frac{\partial}{\partial\theta} l(\theta|\mathbb{X}_n) = 0$$

na događaju čija vjerojatnost teži ka 1 za $n \to \infty$ ima korjen $\hat{\theta}_n = \hat{\theta}_n(X_n)$ takav da je $\hat{\theta}_n \xrightarrow{P_\theta} \theta$, za $n \to \infty$.

Napomena 1.2.10. *Ako jednadžba vjerodostojnosti ima jedinstvenu stacionarnu točku $\hat{\theta}_n \xrightarrow{\mathbb{P}_{\theta_0}} \theta_0$, tada Teorem* 1.2.9 *tvrdi da ona mora biti konzistentan procjenitelj za θ_0. Ako je MLE jedinstvena stacionarna točka kao točka lokalnog maksimuma, onda je MLE konzistentan procjenitelj za θ.*

Lema 1.2.11. *Neka je $X \sim B(n, \theta)$ gdje je θ vjerojatnost uspjeha. Tada je procjenitelj maksimalne vjerodostojnosti za θ relativna frekvencija uspjeha.*

Dokaz. Označimo sa n broj pokušaja, a sa k broj uspjeha. Tada je vjerojatnost da smo imali točno k uspjeha dana s

$$f(\theta) = P(X = k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}, k = 0, 1, 2, ..., n$$

Nađimo stacionarne točke koje su kandidati za lokalni maksimum:

$$\begin{aligned} f'(\theta) &= \binom{n}{k}[k\theta^{k-1}(1-\theta)^{n-k} - \theta^k(n-k)(1-\theta)^{n-k-1}] \\ &= \binom{n}{k}[\theta^{k-1}(1-\theta)^{n-k-1}(k(1-\theta) - (n-k)\theta)] \\ &= 0 \end{aligned}$$

$\Rightarrow k - k\theta - n\theta + k\theta = 0$
$\Rightarrow n\theta = k$
$\Rightarrow \theta = \frac{k}{n}$ □

1.3 Markovljevi lanci

Definicija 1.3.1. *Neka je S skup.* ***Slučajan proces*** *s diskretnim vremenom i prostorom stanja S je familija $X = (X_n : n \geq 0)$ slučajnih varijabli definiranih na nekom vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ s vrijednostima u S.*
Dakle, za svaki $n \geq 0$ je $X_n : \Omega \to S$ slučajna varijabla.

Definicija 1.3.2. *Neka je S prebrojiv skup. Slučajni proces $X = (X_n : n \geq 0)$ definiran na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ s vrijednostima u skupu S je*
Markovljev lanac prvog reda *ako vrijedi*

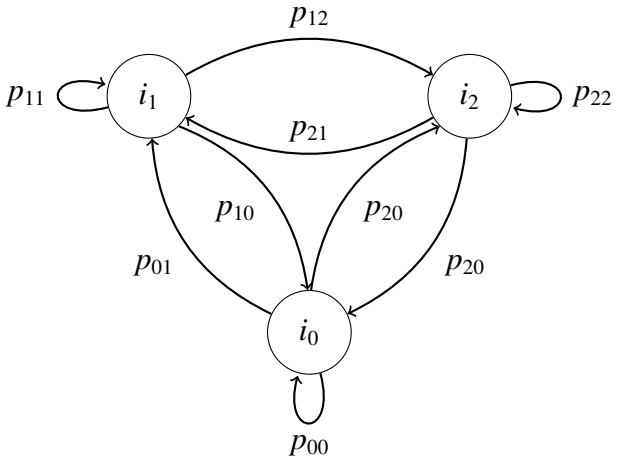
$$\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = i_0) = \mathbb{P}(X_{n+1} = j | X_n = i) \tag{1.3}$$

za svaki $n \geq 0$ i za sve $i_0, ..., i_{n-1}, i, j \in S$ za koje su obje uvjetne vjerojatnosti dobro definirane.

Svojstvo u relaciji (1.3) naziva se *Markovljevim svojstvom*.

Definicija 1.3.3. *Označimo sa $p_{ij} = \mathbb{P}(X_{t+1} = j \mid X_t = i)$ vjerojatnost da slučajna varijabla X prijeđe u stanje j u trenutku $t + 1$, ako je u trenutku t bila u stanju i. Vrijednost p_{ij} nazivamo* ***prijelazna (tranzicijska) vjerojatnost****.*

Markovljev lanac zajedno sa zadanim prijelaznim vjerojatnostima nazivamo **Markovljevim modelom**.



Slika 1.1: Markovljev lanac

1.4 Shannonova entropija

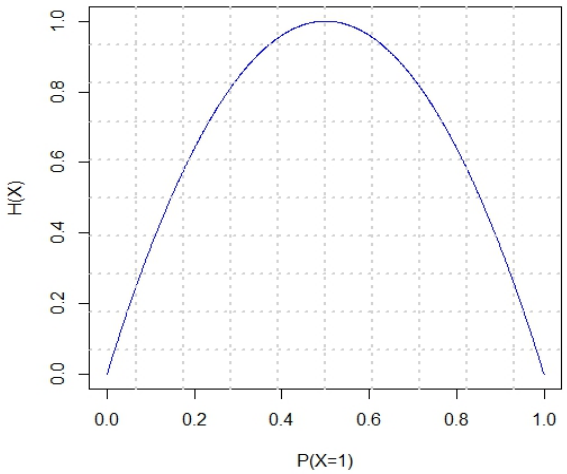
Definicija 1.4.1. *Za danu slučajnu varijablu X sa vjerojatnostima* $\mathbb{P}(x_i)$ *i za diskretan skup događaja* $x_1, ..., x_K$ *definiramo* ***Shannonovu entropiju*** *s*

$$H(X) = -\sum_{i=1}^{K} \mathbb{P}(x_i) \log(\mathbb{P}(x_i)) \tag{1.4}$$

Da bismo intuitivno shvatili o čemu je riječ razmotrimo primjer bacanja novčića:
U ovom slučaju, imamo dva moguća simbola (K = 2), i oba se pojavljuju s vjerojatnošću $p(x_i) = \frac{1}{2}$.

Jednostavnim uvrštavanjem u formulu entropije dobivamo $H(X) = 1$ bit/simbol. Dakle, vrijednost entropije u ovisnosti o vjerojatnosti pojave pisma/glave kod bacanja novčića je 1 bit/simbol.

Za slučaj "nepoštenog" novčića koji uvijek daje pismo, imamo $p(x_1) = 1,\ p(x_2) = 0$, dobivamo očekivano H(X) = 0 bit/simbol ($0 \log 0 = 0$, jer vrijedi $x \log x \to 0$ kada $x \to 0$). Uvrštavanjem svih mogućih vjerojatnosti pojave pisma u formulu entropije, dobivamo graf ovisnosti vrijednosti entropije o toj vjerojatnosti (1.2). Maksimum (1 bit/simbol) je postignut kada je vjerojatnost pisma jednaka vjerojatnosti glave ($p = \frac{1}{2}$). Primijetimo simetriju ovog grafa. Svejedno je pojavljuje li se s većom vjerojatnošću pismo ili glava.



Slika 1.2: Vrijednost entropije u ovisnosti o vjerojatnosti pojave pisma kod bacanja novčića

Pretpostavimo da su zadane dvije funkcije više varijabli $f, \varphi : \mathcal{D} \to \mathbb{R}$ definirane na skupu $\mathcal{D} \subseteq \mathbb{R}^k$. Funkciji φ pridružimo implicitnu jednadžbu $\varphi(y_1, ..., y_k) = 0$ i pripadajući skup $S \subseteq \mathcal{D}$ definiran tom jednadžbom $S = \{(y_1, ..., y_k) \in \mathcal{D} \mid \varphi(y_1, ..., y_k) = 0\}$.

Definicija 1.4.2. *Ako za točku $T_0 = (x_{10}, ..., x_{k0}) \in S$ postoji okolina $K(T_0, \delta) \subseteq \mathcal{D}$ tako da je*

$$f(x_1, ..., x_k) < f(x_{10}, ..., x_{k0}),\ \forall (x_1, ..., x_k) \in S \cap K(T_0, \delta) \setminus \{T_0\}$$

onda kažemo da funkcija f u točki T_0 ima uvjetni lokalni maksimum uz uvjet $\varphi(x_1, ..., x_k) = 0$.

Problem uvjetnog lokalnog maksimuma

$$\begin{cases} z = f(x_1, ..., x_k) \to \max \\ \varphi(x_1, ..., x_k) = 0 \end{cases}$$

često rješavamo uvođenjem Lagrangeove funkcije $L(x_1, ..., x_k, \lambda)$:

$$L(x_1, .., x_k, \lambda) = f(x_1, ..., x_k) + \lambda\varphi(x_1, ..., x_k),\ (x_1, ..., x_k) \in \mathcal{D},\ \lambda \in \mathbb{R}.$$

Parametar λ zove se **Lagrangeov multiplikator**.

Lema 1.4.3. *Uniformno distribuirani parametri imaju maksimalnu entropiju.*

Prije samog dokaza prisjetimo se Bolzano-Weierstrassova i Rolleova teorema:

Teorem 1.4.4. *(Bolzano-Weierstrass): Neka je funkcija $f : [a, b] \to \mathbb{R}$ neprekidna na segmentu $[a, b] \subset \mathbb{R}$. Tada je $f([a, b]) = [m, M]$ također segment.*

Napomena 1.4.5. *Tvrdnja teorema može se razdvojiti na tri dijela:*

1. *f je ograničena na $[a, b]$, odnosno postoje $m = \inf_{[a,b]} f$ i $M = \sup_{[a,b]} f$.*
2. *funkcija f postiže svoj minimum i maksimum na $[a, b]$, odnosno postoje $x_m, x_M \in [a, b]$ takvi da vrijedi $f(x_m) = m$ i $f(x_M) = M$.*
3. *za svaki $C \in (m, M)$, postoji $c \in [a, b]$ takav da je $f(c) = C$.*

Teorem 1.4.6. *(Rolle): Neka je $f : I \to \mathbb{R}$, diferencijabilna na otvorenom intervalu $I \subset \mathbb{R}$ i neka za $a, b \in I$, $a < b$, vrijedi $f(a) = f(b) = 0$. Tada postoji $c \in (a, b)$ takav da je $f'(c) = 0$*

Dokaz. (Lema (1.4.3)): Definiramo funkcije $f : [0,1]^k \to \mathbb{R}$ i $\varphi : [0,1]^k \to \mathbb{R}$ s

$$f(p_1, ..., p_k) = -\sum_{i=1}^{k} p_i \log p_i$$

$$\varphi(p_1, ..., p_k) = \sum_{i=1}^{k} p_i - 1.$$

Neka je λ Lagrangeov multiplikator. Definiramo funkciju $g : \mathbb{R}^k \to \mathbb{R}$ sa

$$g(p_1, ..., p_k) = f(p_1, ..., p_k) + \lambda\varphi(p_1, ..., p_k)$$

Funkcija g je klase C^∞ na zatvorenom skupu $[0,1]^k$, znači da je ujedno i neprekidna pa prema *Bolzano-Weierstrassovom teoremu* poprima minimum m i maksimum M na tom skupu. Budući da funkcija g nije konstantna funkcija na $[0,1]^k$ barem jedna od te dvije vrijednosti se nalazi unutar otvorenog skupa $(0,1)^k$.
Funkcija g je strogo pozitivna na $(0,1)^k$, u rubovima je jednaka 0, stoga će prema *Rollovom teoremu* stacionarna točka biti maksimum.

Tražimo stacionarne točke te funkcije.

$$\frac{dg}{dp_i} = -\log p_i - 1 + \lambda = 0$$

$$\log p_i = \lambda - 1$$

$$p_i = \exp(\lambda - 1)$$

$$\sum_{i=1}^{k} p_i = 1 \;\Rightarrow\; k\exp(\lambda - 1) = 1$$

Slijedi da funkcija g postiže maksimum u točki $p_M = (p_1, ..., p_k)$

$$p_i = \frac{1}{k},\ i = 1, ..., k$$

. □

1.5 AIC i BIC

Osim promatranja funkcije vjerodostojnosti ili omjera vjerodostojnosti, za odabir najboljeg statističkog modela mogu se koristiti jos neki kriteriji. Dva često koristena kriterija su AIC

(Akaike Information Criterion) i BIC (Bayesian Information Criterion). AIC i BIC mjere koliko "dobro" model opisuje podatke. Dani su sljedećim jednadžbama:

$$AIC = -2\log(L) + 2k$$

$$BIC = -2\log(L) + k\log(m)$$

gdje je L maksimalna vjerodostojnost modela, m duljina niza, a k broj slobodnih parametara. Informacijski se kriteriji temelje na funkciji vjerodostojnosti. Osim vjerodostojnosti, sadržavaju i penalizaciju za kompleksnost modela. Znamo, općenito, da pri odredivanju modela isti možemo poboljšati dodajući mu parametre. No s druge strane, cilj nam je da model bude što jednostavniji tj. da ima što manje parametara. Primjećujemo da i AIC i BIC penaliziraju količinu parametara (BIC više nego AIC), a time ujedno mogu riješiti i problem overfittinga do kojeg može doći dodamo li previše parametara u model. Budući da AIC i BIC zapravo procjenjuju koliko je informacija izgubljeno modeliranjem podataka, najbolji model bit će onaj s najmanjim AIC-om (BIC-om).

Poglavlje 2

Skriveni Markovljev model

2.1 Skriveni Markovljev model (HMM)

Skriveni Markovljevi modeli su statistički modeli koji imaju široku primjenu u molekularnoj biologiji, prepoznavanju govora i računalnom prevođenju.

Kod običnog Markovljevog modela niz stanja koji emitira neki niz opažanja nam je uvijek poznat. Kod skrivenog Markovljevog modela, imamo niz stanja i niz simbola. Svaki simbol ovisi jedino o trenutnom stanju u kojem se proces nalazi. Zato generiranje simbola iz stanja modeliramo **Markovljevim lancem nultog reda** što je upravo *niz nezavisnih događaja*. Niz stanja skrivenog Markovljevog modela modeliran je Markovljevim lancem prvog reda, tj. vjerojatnost da se nalazimo u nekom stanju ovisi samo o prethodnom stanju.
Formalno rečeno:

Definicija 2.1.1. ***Skriveni Markovljev model prvog reda*** *(eng. Hidden Markov model, HMM) je skup slučajnih varijabli koji se sastoji od dva podskupa, Q i O:*

- $Q = Q_1, ..., Q_N$ *- skup slučajnih varijabli koje poprimaju diskretne vrijednosti*
- $O = O_1, ..., O_N$ *- skup slučajnih varijabli koje poprimaju diskretne ili kontinuirane vrijednosti.*

Te varijable zadovoljavaju sljedeće uvjete:

1.
$$P(Q_t|Q_{t-1}, O_{t-1}, ..., Q_1, O_1) = \mathbb{P}(Q_t|Q_{t-1}) \tag{2.1}$$

2.
$$\mathbb{P}(O_t|Q_T, O_T, ..., Q_{t+1}, O_{t+1}, Q_t, Q_{t-1}, O_{t-1}, ..., Q_1, O_1) = P(O_t|Q_t) \tag{2.2}$$

Uočili smo da, osim niza stanja kroz koja proces prolazi (označili smo ih sa Q_i), promatramo i niz opažanja (simbola, označili smo ih sa O_i).
Da pojasnimo, relacija (2.1) predstavlja vjerojatnost da smo, za neko $t \in \{1, 2, ..., N\}$, u stanju Q_t uz uvjet da su se dogodila sva prethodna stanja $Q_1, ..., Q_{t-1}$ i emitirali simboli $O_1, ..., O_{t-1}$ jednaka **tranzicijskoj vjerojatnosti** iz stanja Q_{t-1} u stanje Q_t.
Relacija (2.2) povlači da realizacija nekog opažanja u sadašnjem stanju ovisi samo o tom stanju. Vjerojatnosti iz relacije (2.2) nazivamo **emisijska vjerojatnost** i kažemo da neko stanje Q_t **emitira** simbol O_t.

Skriveni Markovljev model zadan je sljedećim parametrima:

- N - broj stanja u kojima se proces može nalaziti

$$S = \{1, ..., N\} \tag{2.3}$$

  S - skup svih stanja procesa

- M - broj mogućih opažanja

$$B = \{b_1, ..., b_M\} \tag{2.4}$$

  B - skup svih opaženih vrijednosti

- L - duljina opaženog niza

$$X = (x_1, ..., x_L) \tag{2.5}$$

  X - opaženi niz

- A - matrica tranzicijskih vjerojatnosti

$$A = \{a_{ij}\}, a_{ij} = \mathbb{P}(Q_{t+1} = j | Q_t = i), 1 \leq i, j \leq N \tag{2.6}$$

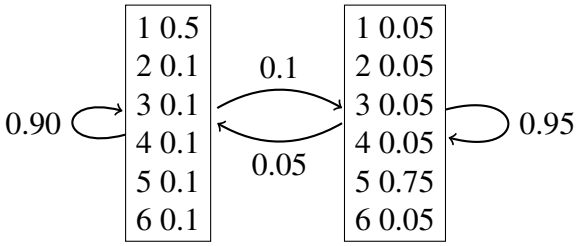
- E - matrica emisijskih vjerojatnosti

$$E = \{e_j(k)\}, e_j(k) = \mathbb{P}(O_t = b_k | Q_t = j), 1 \leq j \leq N, 1 \leq k \leq M \tag{2.7}$$

Primijetimo da nam je kod Markovljevog modela nultog reda niz stanja koji emitira neki niz simbola poznat.
Kod skrivenog Markovljeva modela stanja su "skrivena", odnosno ne znamo ih pri opažanju nekog niza vrijednosti ili simbola. Međutim, taj niz vrijednosti nam je poznat i pomoću njega možemo donijeti neke zaključke o nizu stanja koji nam je nepoznat.

2.2 Primjer skrivenog Markovljevog modela

Imamo dvije nepoštene igraće kocke. Jedna kocka, koju označavamo K_1, ima vjerojatnost da dobijemo jedinicu $\frac{1}{2}$, a vjerojatnost preostalih ishoda je $\frac{1}{10}$, dok druga kocka, u oznaci K_5 ima vjerojatnost da padne petica $\frac{3}{4}$, a vjerojatnost preostalih ishoda je $\frac{1}{20}$.
Pretpostavimo da počinjemo sa K_1. Vjerojatnost da ćemo ponovo koristiti K_1 je 90%, dok je vjerojatnost da ćemo je zamijeniti sa K_5 10%. Kad smo jednom K_1 zamijenili sa K_5, u 95% slučajeva ćemo je i nastaviti koristiti. Vjerojatnost da je zamijenimo sa K_1 je 5%.



Slika 2.1

Koristimo li notaciju za HMM, naš model zapisujemo na sljedeći način:

- N=2
$$\mathcal{S} = \{K_1, K_5\}$$
- M=6
$$B = \{1, 2, 3, 4, 5, 6\}$$
- Matrica tranzicijskih vrijednosti je dana s:
$$A = \begin{pmatrix} 0.90 & 0.1 \\ 0.05 & 0.95 \end{pmatrix}$$
gdje je $a_{11} = \mathbb{P}(K_1|K_1)$ - vjerojatnost da je nakon K_1 ponovo bačena K_1, $a_{12} = \mathbb{P}(K_5|K_1)$ - vjerojatnost bacanja K_5, ako je prethodno bačena K_1, $a_{21} = \mathbb{P}(K_1|K_5)$ - vjerojatnost da je nakon bacanja K_5 bačena K_1 i $a_{22} = \mathbb{P}(K_5|K_5)$ - vjerojatnost da je nakon K_5 opet bačena K_5.
- Matrica emisijskih vjerojatnosti je:
$$E = \begin{pmatrix} 0.5 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.05 & 0.05 & 0.05 & 0.05 & 0.75 & 0.05 \end{pmatrix}$$
Prvi redak čine emisijske vjerojatnosti elemenata iz B u stanju K_1, a drugi redak emisijske vjerojatnosti elemenata iz B u stanju K_5

Proces koji modelira izbor kocki je Markovljev proces prvog reda sa stanjima u $\mathcal{S}$. Kocke su stanja i prijelaz iz jedne kocke u drugu se može opisati Markovljevim lancem. Emisijske vjerojatnosti simbola iz B su u svakom od stanja različite i ne ovise o prijašnjim stanjima. Možemo reći da smo dali primjer *skrivenog Markovljevog modela prvog reda*.
Ako imamo niz simbola, odnosno opaženih vrijednosti, primjerice $X = (1, 3, 1, 2, 5, 6, 6, 4, 3)$ ne znamo koja kocka stoji iza pojedine opažene vrijednosti. Dakle, *niz stanja* je skriven.

Poglavlje 3

Analiza skrivenih Markovljevih modela

U skrivenom Markovljevom modelu je niz stanja nepoznat, ali pomoću niza opaženih vrijednosti možemo nešto zaključiti o nizu stanja. Pomoću niza simbola moguće je:

- odrediti *najvjerojatniji niz stanja* za dani niz simbola koristeći **Viterbijev algoritam**.
- procijeniti *parametre uvjetne maksimalne vjerodostojnosti* koristeći **Viterbijevo treniranje**.
- procijeniti parametre maksimalne vjerodostojnosti modela koristeći **Baum-Welchov algoritam**.

Niz opaženih simbola ćemo označiti s $x = (x_1, ..., x_n)$. Pripadajući niz skrivenih stanja nazivamo stazom $\pi = (\pi_1, ..., \pi_n)$. Sama staza slijedi Markovljev lanac, tako da vjerojatnost stanja ovisi o prethodnom stanju. Lanac je karakteriziran parametrima

$$a_{kl} = \mathbb{P}(\pi_i = l | \pi_{i-1} = k).$$

Tranzicijska vjerojatnost a_{0k} se može smatrati vjerojatnošću da počnemo u stanju k. Budući da smo razdvojili simbole b od stanja k, moramo uvesti novi skup parametara za model, $e_k(b)$ koje definiramo sa

$$e_k(b) = \mathbb{P}(x_i = b | \pi_i = k),$$

tj. kao vjerojatnost da je simbol b vidljiv u stanju k. Ove vjerojatnosti su poznate i kao emisijske vjerojatnosti.

3.1 Viterbijev algoritam

Viterbijev algoritam je algoritam dinamičkog programiranja za pronalaženje najvjerojatnije staze π^* u skrivenom Markovljevom modelu koji emitira zadani niz simbola x.

Takvu stazu π^* dobivenu Viterbijevim algoritmom nazivamo **Viterbijev put** ili **Viterbijev prolaz**. Dakle, tražimo onu stazu s najvećom vjerojatnošću,

$$\pi^* = arg\max_{\pi} \mathbb{P}(x, \pi) = arg\max_{\pi} \mathbb{P}(\pi|x).$$

Pri tome je vjerojatnost $\mathbb{P}(x, \pi)$ definirana kao:

$$\mathbb{P}(x, \pi) = a_{0\pi_1} \prod_{i=1}^{n} e_{\pi_i}(x_i) a_{\pi_i \pi_{i+1}} \tag{3.1}$$

gdje $a_{0\pi_1}$ i $a_{\pi_n \pi_{n+1}}$ koristimo za modeliranje početka i kraja te stavljamo $a_{0\pi_1} = a_{\pi_n \pi_{n+1}} = 1$ Pretpostavimo da je za sva stanja k poznata vjerojatnost najvjerojatnije staze koja u stanju k završava s simbolom x_i,

$$v_k(i) = \mathbb{P}(x_1, \ldots, x_i|\pi_t = k).$$

Tada se vjerojatnost optimalne staze kroz model gdje su emitirani $x_1, \ldots, x_{i+1}$ i koji završava u stanju l može izraziti rekruzivno:

$$v_l(i+1) = e_l(x_{i+1}) \max_k (v_k(i) a_{kl}).$$

Svi nizovi moraju početi u stanju 0, tako da je početni uvjet $v_0(0) = 1$.
Viterbijev algoritam se sastoji od četiri koraka:

1. **Inicijalizacija** (i=0):
$$v_0(0) = 1,\ v_k(0) = 0,\ k > 0$$

2. **Rekurzija** (i=1,...,n):
$$v_l(i) = e_l(x_i) \max_k (v_k(i-1) a_{kl})$$
$$ptr_i(l) = arg\max_k (v_k(i-1) a_{kl})$$

3. **Kraj**:
$$P(x, \pi^*) = \max_k v_k(n) a_{k0}$$
$$\pi_n^* = arg\max_k (v_k(n) a_{k0})$$

4. **Povratak unazad** (i=n,...,1):
$$\pi_{i-1}^* = ptr_i(\pi_i^*)$$

Napomena 3.1.1. *U praksi je problem što množenje mnogo malih vjerojatnosti uvijek vodi malim brojevima, a to daje "underflow" greške na računalu. Zbog toga Viterbijev algoritam uvijek treba izvoditi u log-prostoru. Odnosno, trebamo računati* $\log(v_l(i))$. *Tako će se produkti pretvoriti u sume i brojevi će ostati razumni.*

3.2 Viterbijevo treniranje

Funkcija cilja u procjeni maksimalne vjerodostojnosti je maksimizacija relacije (3.1) preko svih staza π za dani niz simbola X. Viterbijevo treniranje je iterativan proces koji garantira monotoni rast vjerodostojnosti kroz skup ponovo procjenjenih parametara. Neka je zadan fiksan model M, te neki inicijalni parametri θ. Viterbijevim algoritmom pronađemo najbolju stazu $\pi^* = arg\max_\pi \mathbb{P}(x, \pi)$ kroz model M. Na taj način svakom simbolu od X pridružimo stanje. Sad možemo odrediti emisijske i tranzicijske frekvencije. Prema lemi 1.2.11, relativne frekvencije su procjenitelji maksimalne vjerodostojnosti, one su novi parametri modela i proces se iterira. Budući da parametri ovise o stazi, a broj staza je konačan, ta metoda konvergira. Opisana procedura pronalazi vrijednost θ koja maksimizira vjerodostojnost najvjerojatnijeg niza skrivenih stanja. Viterbijevo treniranje je metoda sa širokom primjenom i može se koristiti kad je primarna svrha HMMova proizvesti dekodiranje preko Viterbijevog poravnanja.

3.3 Baum-Welchov algoritam

Kao i Viterbijevo treniranje, Baum-Welchov algoritam je iterativni postupak za određivanje parametara modela na temelju niza opažanja. Za razliku od Viterbijevog treniranja, Baum-Welchov algoritam parametre procjenjuje tako da maksimizira očekivanje promatranog niza uzimajući u obzir sve staze, a ne samo najbolju. Budući da nam je staza π nepoznata, ne možemo jednostavno prebrojati tranzicije i emisije nego moramo izračunati njihove očekivane vrijednosti. U svakoj iteraciji dobivamo novi skup vrijednosti parametara iz očekivanog broja emisija i tranzicija, uzimajući u obzir sve moguće staze. Algoritam obično zaustavljamo nakon određenog broja iteracija ili kad promjena logaritma funkcije maksimalne vjerodostojnosti postane dovoljno mala. Može se pokazati da Baum-Welchov algoritam konvergira u lokalni optimum. Uobičajeno je da sustav ima mnogo lokalnih ekstrema pa konvergencija u jedan od njih uvelike ovisi o zadanim početnim parametrima.

3.4 Deterministično kaljenje

Značajno poboljšanje naspram standardnim nadziranim i nenadziranim metodama učenja je primjena metode determinističkog kaljenja. Ova metoda ima dva važna svojstva:

1. mogućnost izbjegavanja lokalnih optimuma
2. primjenjivost na mnogo različitih struktura

Kod determinističkog kaljenja, proces je deterministički što znači da ne želimo "slučajno" lutati po prostoru parametara dok postepeno napredujemo u maksimizaciji vjerodostoj-

nosti. S druge strane, to je još uvijek metoda kaljenja koja teži globalnom optimumu, umjesto da pohlepno zapne u obližnjem lokalnom optimumu.
U deterministčkom kaljenju pristup se formalno temelji na principima teorije informacija i teorije vjerojatnosti.

U početku zadajemo visoku vrijednost parametra kaljenja γ na nekom intervalu i kaljenje se provodi na konveksnoj kombinaciji parametara maksimalne entropije i deterministički izračunatih parametara (u simuliranom kaljenju parametri su odabrani na slučajan način). U početku je doprinos deterministički određenih parametara jako mali, a što se više približavamo cilju, omjer se mijenja u njihovu korist. Dodavanje parametara maksimalne entropije služi izbjegavanju lokalnih optimuma.
Konkretno, u prvoj iteraciji zadamo inicijalne parametre modela. Ulazni parametri za Viterbijevo treniranje su konveksne kombinacije parametara maksimalne entropije i inicijalnih parametara. Kad je izračunata najvjerojatnija staza kroz model, računamo tranzicijske i emisijske relativne frekvencije. U svakom sljedećem koraku su ulazni parametri konveksne kombinacije parametara maksimalne entropije i relativnih frekvencija izračunatih u prethodnom koraku.

Deterministčko kaljenje se provodi na sljedeći način:
Inicijalizacija:

$$T_u = \text{zadani tranzicijski parametri}$$
$$E_u = \text{zadani emisijski parametri}$$

flT i flE su tranzicijski odnosno emisijski parametri maksimalne entropije

Petlja:
Dok je brojač manji od ukupnog broja iteracija radi:

1. $$\begin{cases} T = \gamma flT + (1-\gamma)T_u \\ E = \gamma flE + (1-\gamma)E_u \end{cases}$$

   γ *- parametar kaljenja*

2. Viterbijevim algoritmom računamo najvjerojatniju stazu kroz model i maksimalnu vjerodostojnost

3. Računamo relativne tranzicijske i emisijske frekvencije

4.
$$\begin{cases} T_u = \textit{relativne tranzicijske frekvencije izračunate u koraku 3.} \\ E_u = \textit{relativne emisijske frekvencije izračunate u koraku 3.} \end{cases}$$

5. Provjera uvjeta petlje

Kraj:
Imamo matrice relativnih tranzicijskih i emisijskih parametara i maksimalnu vjerodostojnost koju smo dobili u koraku 2.

Poglavlje 4

Rezultati

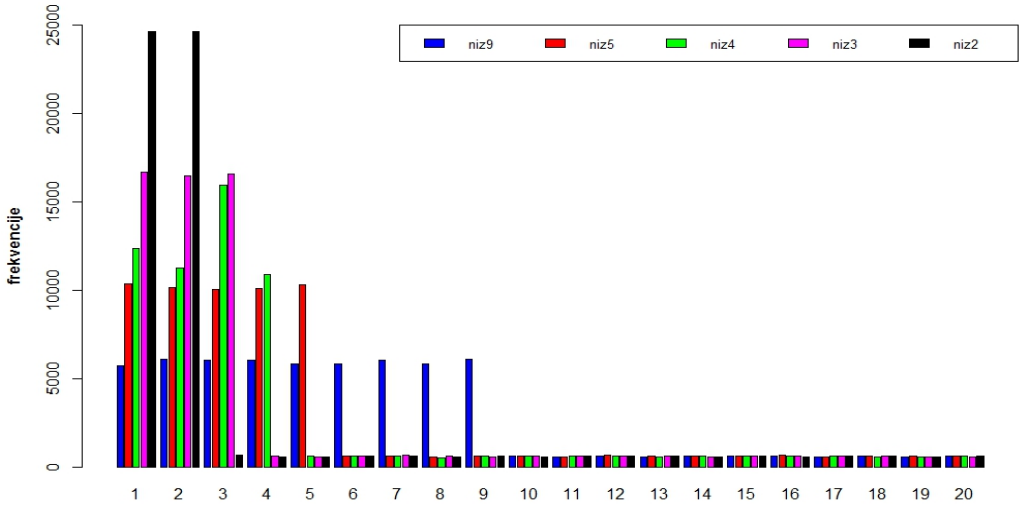
U radu su korišteni programski jezici Python i R.

4.1 Simulacija i optimizacija

Neka je skup opaženih brojeva $B = \{1, 2, \ldots, 20\}$. U ovom radu želimo saznati možemo li iz niza dobivenih brojeva saznati broj stanja.

Za procjenu parametara maksimalne vjerodostojnosti može se koristiti nekoliko algoritama. Iako Baum-Welchov algoritam u općenitom slučaju daje bolje rezultate nego Viterbijevo treniranje, u ovom smo se radu odlučili na korištenje Viterbijevog treniranja modificirano determinističkim kaljenjem zbog njegove manje složenosti i bitno jednostavnije implementacije.

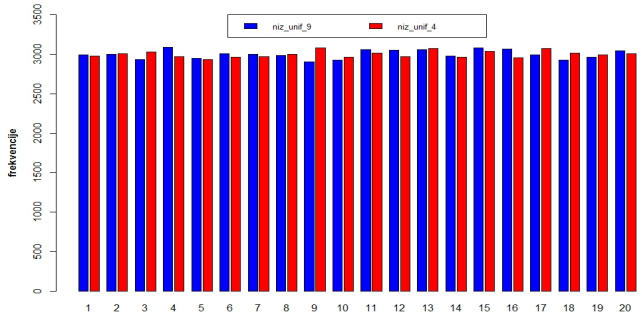
Zadali smo nekoliko skrivenih Markovljevih modela te iz njih smo simulirali smo nizove duljine 60000. Parametri za simulaciju nizova nalaze se u Dodatku (4.2). Izabrali smo parametre male entropije jer proces optimizacije povećava vjerodostojnost, a smanjuje entropiju. Parametri male entropije zapravo govore da svako stanje ima jedan dominantan simbol. Sa matricom emisijskih vrijednosti $E2$ i matricom tranzicijskih vrijednosti $T2$ zadali smo skriveni Markovljevi model sa 2 stanja i skupom opažanjem $B = \{1, 2, \ldots, 20\}$. Iz njega smo simulirali *niz2*. Sa matricom emisijskih vrijednosti $E3$ i matricom tranzicijskih vrijednosti $T3$ zadali smo skriveni Markovljevi model sa 3 stanja. Iz njega smo simulirali *niz3*. Slično, *niz4* sa $E4$ i $T4$, niz5 sa $E5$ i $T5$ te *niz9* sa $E9$ i $T9$. Prikažimo frekvencije opaženih nizova histogramom:



Slika 4.1

Primijetimo da bi iz histograma za svaki niz mogli otkriti broj stanja.

Simulirali smo i dva niza duljine 60000 sa uniformno distribuiranim parametrima, tj. parametrima maksimalne entropije. Sa matricom emisijskih vrijednosti $EU4$ i matricom tranzicijskih vrijednosti $TU4$ zadali smo skriveni Markovljevi model sa 4 stanja. Iz njega smo simulirali *niz_unif_4*. Slično, *niz_unif_9* iz Markovljeva modela sa 9 stanja, $EU9$ i $TU9$.



Slika 4.2

Iz histograma možemo vidjeti da nema velike razlike između ta dva niza. χ^2-testom možemo pokazati da su oba niza uniformno distibuirana. Možemo zaključiti da su nizovi slični i nema razlike iz kojeg su Markovljevog modela sa 4 ili 9 stanja simulirani. Stoga ovaj slučaj sa parametrima maksimalne entropije nećemo promatrati.

Za sve simulirane nizove *niz2*, *niz3*, *niz4*, *niz5* i *niz9* smo procijenili parametre i izračunali maksimalnu vjerodostojnost modela pomoću Viterbijevog treniranja modificiranog determinističkim kaljenjem. Za broj iteracija smo uzeli 1000.

Kako bismo odredili najbolji model za svaki niz, promatrat ćemo vjerodostojnost, log-omjer vjerodostojnosti i informacijske kriterije.

Vjerodostojnost

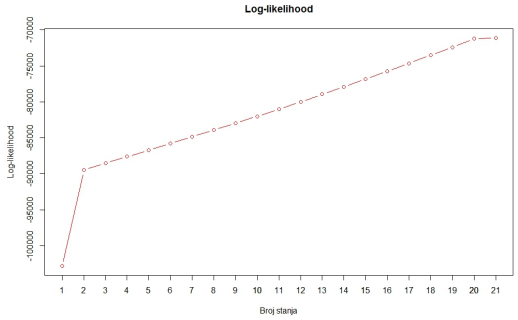
Na temelju svakog niza pokušali smo odrediti broj stanja tako što smo modelirali početni problem s $i = 1, 2, \ldots, 21$ stanja. Primijetili smo da funkcija vjerodostojnosti počinje padati nakon određenog broja iteracija, što se u teoriji ne bi smjelo događati. Razlog tome je što smo pri korištenju modificiranog Viterbijeva treniranja postavili pseudozbroj relativnih frekvencija na inicijalnu vrijednost 0.01 (umjesto 0). U programu smo maksimizirali vjerodostojnost koliko god je moguće, a kad je počela padati, prekinuli smo proces.

Rezultati se nalaze u tablici:

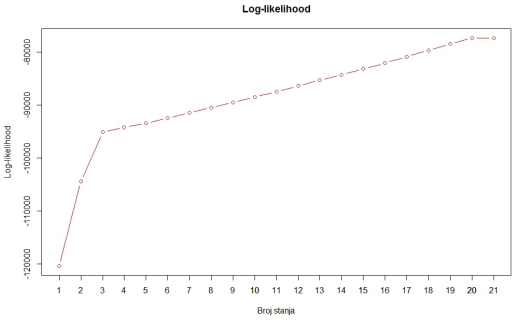
i	niz2	niz3	niz4	niz5	niz9
1	-102819.514971	-120411.092176	-131058.177589	-140704.172606	-161801.242698
2	-89480.366538	-104405.601045	-116091.348708	-125471.993042	-149437.736222
3	-88527.415070	-95106.199531	-105815.995647	-115863.634912	-141743.965541
4	-87613.955896	-94232.739465	-96748.173081	-106801.136634	-134151.873971
5	-86725.946936	-93391.202573	-95890.384635	-97485.566893	-124918.756752
6	-85797.594161	-92448.364082	-94965.444018	-98396.980929	-117532.202302
7	-84853.756762	-91481.585620	-94044.089908	-96493.377265	-111605.657688
8	-83927.128716	-90451.572977	-93160.652744	-95558.274532	-104516.257547
9	-82938.386595	-89509.537902	-92155.594307	-94561.722045	-98456.535886
10	-82002.791074	-88473.811665	-91121.586280	-93531.627792	-97406.149718
11	-81006.172752	-87462.629817	-90087.629335	-92529.654970	-96414.858283
12	-80003.721887	-86376.642738	-88968.774283	-91371.611443	-95335.515384
13	-78951.226121	-85268.658412	-87913.525282	-90244.855298	-94309.717700
14	-77894.180075	-84261.246973	-86839.220277	-89118.818377	-93176.632622
15	-76823.474428	-83155.390928	-85699.807436	-87964.422484	-91994.738145
16	-75760.282537	-82041.434574	-84588.957438	-86782.721689	-90853.761403
17	-74636.829165	-80859.873137	-83427.883289	-85675.595616	-89743.035898
18	-73515.037935	-79649.113762	-82273.636308	-84465.822478	-88592.970199
19	-72391.833529	-78441.120701	-81171.812040	-83268.280231	-87432.493839
20	-71227.012652	-77301.951699	-79952.836138	-82076.049662	-86210.334532
21	-71111.528357	-77302.073430	-79858.125900	-82076.147347	-86210.417468

Tablica 4.1

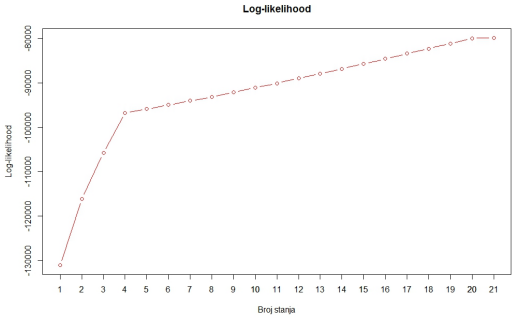
Prikažimo rezultate grafički.



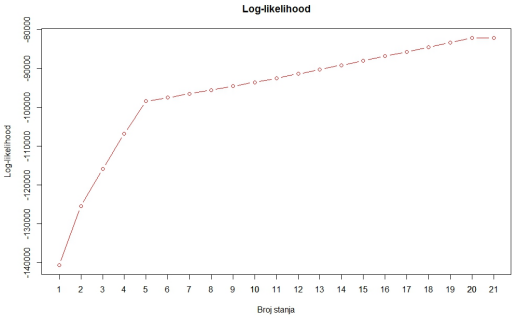
(a) Log-likelihood za niz2



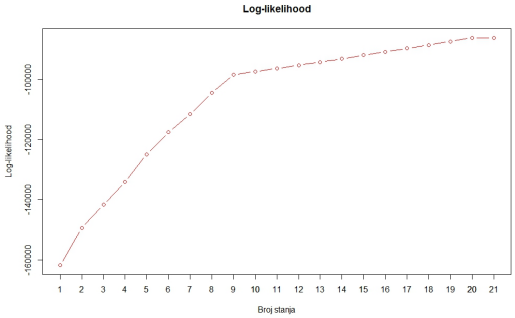
(b) Log-likelihood za niz3



(c) Log-likelihood za niz4



(d) Log-likelihood za niz5



(e) Log-likelihood za niz9

Slika 4.3

Iz Slike 4.3*a* vidimo da funkcija vjerodostojnosti ima najveći skok sa stanja 1 na stanje 2, a nakon stanja 2 funkcija vjerodostojnosti jednako linearno raste do stanja 20, a onda se sa stanja 20 na stanje 21 smanjuje. Primijetimo da najudaljenija točka od pravca koji prolazi prvom i zadnjom točkom na grafu je zapravo stanje 2. Budući da smo *niz2* simulirali iz skrivenog Markovljeva modela sa 2 stanja, možemo zaključiti da je ovo dobar kriterij.

Iz Slike 4.3*b* vidimo da funkcija vjerodostojnosti raste skoro linearno do stanja 3,a onda se mijenja koefcijent linearnosti do stanja 20, pa se funkcija vjerodostojnosti smanjuje sa stanja 20 na stanje 21. I tu možemo primijetiti da najudaljenija točka od pravca koji prolazi prvom i zadnjom točkom na grafu zapravo opisuje broj stanja sa koliko je niz simuliran(u ovom slučaju 3 stanja).

Slično i za ostala tri grafa. Možemo zaključiti da najudaljenija točka od pravca koji prolazi prvom i zadnjom točkom na grafu opisuje broj stanja sa koliko je niz simuliran.

Log-omjer vjerodostojnosti

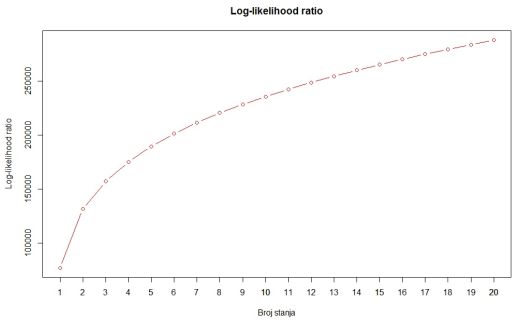
Kako bi odredili najbolji model za svaki niz, promatrat ćemo log-omjer vjerodostojnosti (log-likelihood ratio, oznaka: LLR) koji je definiran sa:

$$LLR = \log \frac{\mathbb{P}(X|M)}{\mathbb{P}(X|\mathcal{R})}$$

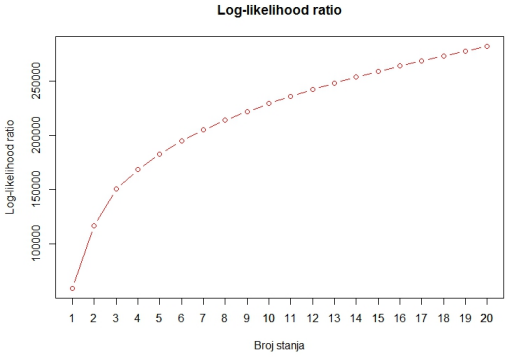
gdje je

- $\mathbb{P}(X|M)$ vjerodostojnost za svaki od modela(s $1, 2, \ldots, 20$ stanja) izračunata Viterbijevim treniranjem modificiranim deterministčkim kaljenjem,
- $\mathbb{P}(X|\mathcal{R}) = \left(\frac{1}{20}\right)^m \left(\frac{1}{i}\right)^{m-1}$, pri čemu je duljina niza $m = 60000$, a i broj stanja u modelu, $i = 1, 2, \ldots, 20$.

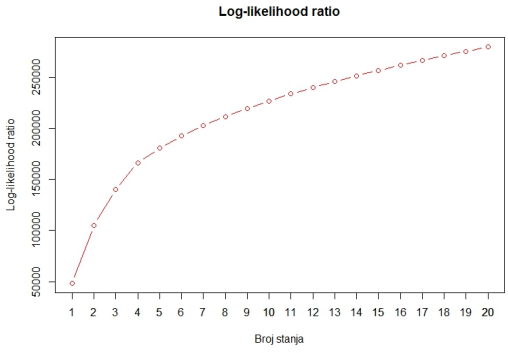
Rezultate prikazujemo grafički:



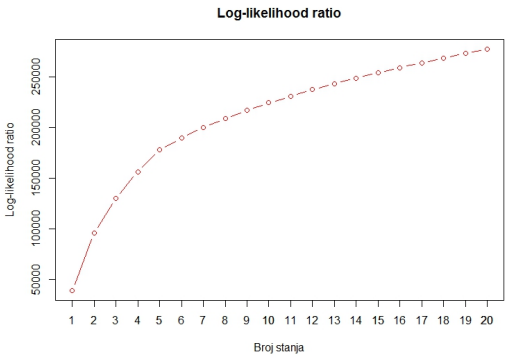
(a) Log-likelihood ratio za niz2



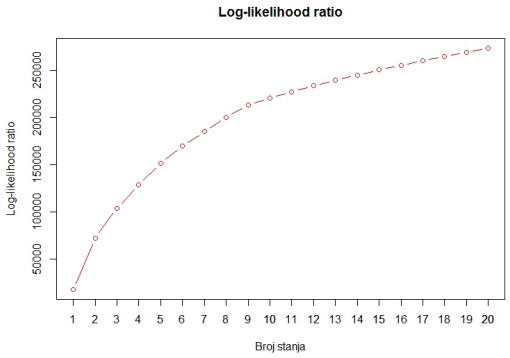
(b) Log-likelihood ratio za niz3



(c) Log-likelihood ratio za niz4



(d) Log-likelihood ratio za niz5



(e) Log-likelihood ratio za niz9

Slika 4.4

Primjetimo maleni lakat redom u stanju $2, 3, 4, 5, 9$ za dane nizove.

Informacijski kriteriji

Za svaki niz izračunali smo AIC i BIC kako bismo za taj niz odredili najbolji model. Podsjetimo se, u prvom poglavlju smo definirali AIC i BIC sljedećim jednadžbama:

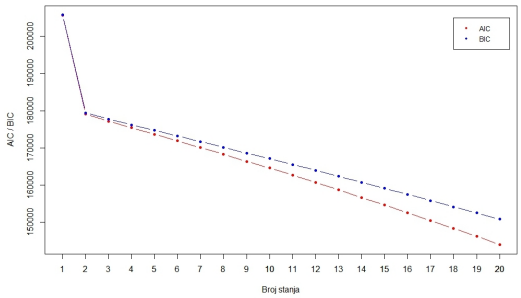
$$AIC = -2\log(L) + 2k$$

$$BIC = -2\log(L) + k\log(m)$$

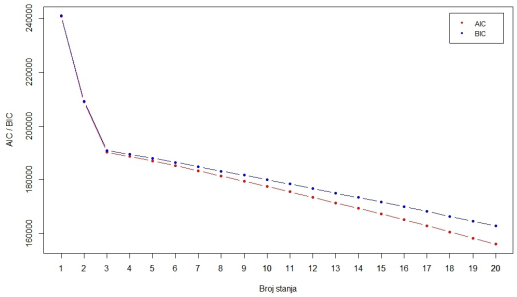
gdje je u našem slučaju:

- L maksimalna vjerodostojnost predloženog modela,
- duljina svakog niza je $m = 60000$,
- broj slobodnih parametara $k = i(i-1) + 19i$, pri čemu i označava i-to stanje, $i = 1, 2, \ldots, 20$.

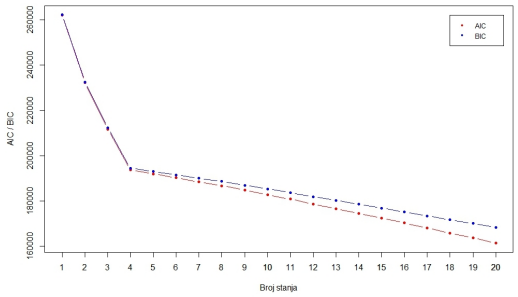
Dobivene rezultate prikazat ćemo grafički.



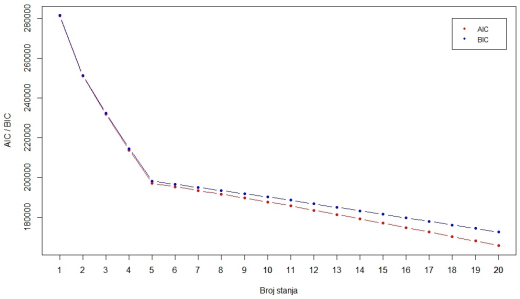
(a) AIC i BIC za niz2



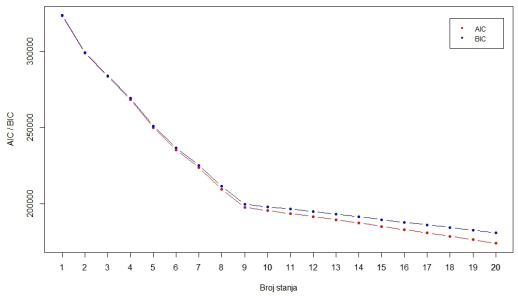
(b) AIC i BIC za niz3



(c) AIC i BIC za niz4



(d) AIC i BIC za niz5



(e) AIC i BIC za niz9

Slika 4.5

Za sve nizove vidimo da su AIC i BIC najmanji za model sa 20 stanja. S obzirom da su podaci simulirani sa redom 2,3,4,5,9 stanja, možemo zaključiti da AIC i BIC nisu dobri kriteriji.

4.2 Dodatak:popis parametara za simulaciju

$$T2 = \begin{pmatrix} 0.80 & 0.20 \\ 0.20 & 0.80 \end{pmatrix}$$

$$E2 = \begin{pmatrix} 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \end{pmatrix}$$

$$T3 = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}$$

$$E3 = \begin{pmatrix} 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \end{pmatrix}$$

$$T4 = \begin{pmatrix} 0.79 & 0.07 & 0.07 & 0.07 \\ 0.07 & 0.79 & 0.07 & 0.07 \\ 0.07 & 0.07 & 0.79 & 0.07 \\ 0.07 & 0.07 & 0.07 & 0.79 \end{pmatrix}$$

$$E4 = \begin{pmatrix} 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \end{pmatrix}$$

$$T5 = \begin{pmatrix} 0.80 & 0.05 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.80 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.80 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.80 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.05 & 0.80 \end{pmatrix}$$

$$E5 = \begin{pmatrix} 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \end{pmatrix}$$

$$T9 = \begin{pmatrix} 0.800 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.800 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.800 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.800 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 & 0.800 & 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.800 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.800 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.800 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.800 \end{pmatrix}$$

$$E9 = \begin{pmatrix} 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.81 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \end{pmatrix}$$

$$TU4 = \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix}$$

$$EU4 = \begin{pmatrix} 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \end{pmatrix}$$

$$TU9 = \begin{pmatrix} 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 \end{pmatrix}$$

$$EU9 = \begin{pmatrix}
0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \\
0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \\
0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \\
0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \\
0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \\
0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \\
0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \\
0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \\
0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05
\end{pmatrix}$$

Bibliografija

[1] A. Allahverdyan, A. Galstyan,*Comparative Analysis of Viterbi Training and Maximum Likelihood Estimation for HMMs*, USC Information Sciences Institute, USA

[2] R. Durbin, S. Eddy, A. Krogh, G. Mitchinson, *Biological sequence analysis*, Cambridge University Press, 1998.

[3] J. Ernst, M. Kellis, *Discovery and characterization of chromatin states for systematic annotation of the human genome, http://www.nature.com/nbt/journal/v28/n8/full/nbt.1662.html* (lipanj 2015.)

[4] B. Guljaš, *Matematička analiza I & II*, PMF-MO predavanja, 2014.

[5] M. Huzak, *Matematička statistika*, PMF-MO predavanja, 2012.

[6] M. Tepić, *Kompleksnost skrivenih Markovljevih modela*, diplomski rad, PMF-MO, Zagreb, 2015.

[7] K. Rose, *Deterministic Annealing for Clustering, Compresion, Classification, Regression and Related Optimization Problems*, IEEE, (1998.), 2210-2239

[8] I. Valčić, *Analiza kompleksnosti skrivenih Markovljevih modela*, diplomski rad, PMF-MO, Zagreb, 2015.

[9] N. Sarapa, *Teorija vjerojatnosti*, Školska knjiga, Zagreb, 2002.

[10] Z. Vondraček, *Markovljevi lanci*, PMF-MO skripta, 2008.

[11] https://element.hr/artikli/file/1357

Sažetak

U ovom diplomskom radu smo se bavili skrivenim Markovljevim modelima, statističkim alatom koji je namijenjen modeliranju nizova koje generira neki skriveni proces. Dali smo formalnu definiciju skrivenog Markovljevog modela, opisali neke algoritme za rad sa skrivenim Markovljevim modelima i implementirali ih u programskom jeziku Python. Konstruirali smo primjer skrivenog Markovljeva modela. Proveli smo nekoliko simulacija i pokušali smo naći najbolji model koji bi opisao te podatke. Koristili smo nekoliko statističkih metoda za odabir najboljeg modela: maksimizacija vjerodostojnosti, log-omjer vjerodostojnosti i informacijske kriterije.

Summary

This thesis is concerned with a statistical model called hidden Markov model (HMM), statistical tool designed for modelling sequences generated by hidden processes. We give a formal definition of the hidden Markov model, describe several algorithms used in their analysis and present their in Python. We also construct example of hidden Markov model. We simulate data and attempt to find the best model for it. We use several statistical methods: likelihood maximization, log-likelihood ratio and information criteria.

Životopis

- Rođena sam 24.kolovoza 1992. u Prozoru.
- Od 1999. do 2007. pohađam Osnovnu školu fra Jeronima Vladića u Ripcima.
- Od 2007. do 2011. pohađam Gimnaziju Lucijana Vranjanina u Zagrebu.
- Od 2011. do 2014. pohađam preddiplomski studij Matematika na PMF-MO u Zagrebu.
- Od 2014. do 2016. pohađam diplomski studij Matematička statistika na PMF-MO u Zagrebu.