Kompleksnost skrivenih Markovljevih modela

Horvatek, Tea

Master's thesis / Diplomski rad

2018

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:619343

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-04-25**

Repository / Repozitorij:

Repository of the Faculty of Science - University of Zagreb

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Tea Horvatek

KOMPLEKSNOST SKRIVENIH MARKOVLJEVIH MODELA

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, siječanj 2018.

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 roditeljima, bratu i suprugu na strpljenju i podršci tokom studiranja.
Veliko hvala mentoru doc. dr. sc. Pavlu Goldsteinu na posvećenom vremenu i savjetima tokom izrade ovog rada.

Sadržaj

Sadržaj **iv**

Uvod **1**

1 Osnovni pojmovi **2**
1.1 Vjerojatnost 2
1.2 Statistika 4
1.3 Shannonova entropija 7

2 Skriveni Markovljev model **10**
2.1 Markovljevi lanci 10
2.2 Skriveni Markovljev model (HMM) 11
2.3 Primjer skrivenog Markovljevog modela 12

3 Algoritmi za analizu HMM **14**
3.1 Viterbijev algoritam 15
3.2 Viterbijevo treniranje 16
3.3 Deterministično kaljenje 16

4 Rezultati **18**
4.1 Simulacija i optimizacija 18
4.2 AIC i BIC 21
4.3 Logaritamsko-polinomijalni kriterij kompleksnosti 23

Bibliografija **31**

Uvod

Cilj ovog diplomskog rada jest dati kratki pregled teorije skrivenih Markovljevih modela te predložiti metodu za određivanje njihove kompleksnosti. Skriveni Markovljevi modeli se danas primjenjuju u raznim područjima, poput prepoznavanja govora i rukopisa, analizi vremenskih nizova i dr. Specijalno, model povremeno nepoštene kockarnice, opisan u ovom radu, često se primjenjuje za modeliranje genoma u bioinformatici.

U prvom poglavlju dan je kratki pregled osnovnih pojmova iz vjerojatnosti i statistike te definicija Shannonove entropije. U drugom poglavlju definirani su Markovljevi lanci te je dana formalna definicija skrivenih Markovljevih modela, uz jednostavan primjer. Treće poglavlje sadrži opis nekih algoritama korištenih pri simulaciji i procjeni parametara modela. U četvrtom poglavlju kratko su definirani neki informacijski kriteriji te pokazana njihova primjena na našim rezultatima, te je na kraju predložena bolja metoda procjene kompleksnosti skrivenih Markovljevih modela.

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 otvorenih intervala $(a, b) = \{x \in \mathbb{R}, a < x < b\}, a, b \in \mathbb{R}$, 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.*

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****.*

Definicija 1.2.7. *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 1.2.8. *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 1.2.9. *Ako jednadžba vjerodostojnosti ima jedinstvenu stacionarnu točku* $\hat{\theta}_n \xrightarrow{\mathbb{P}_{\theta_0}} \theta_0$*, tada Teorem* 1.2.8 *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 1.2.10. *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 Shannonova entropija

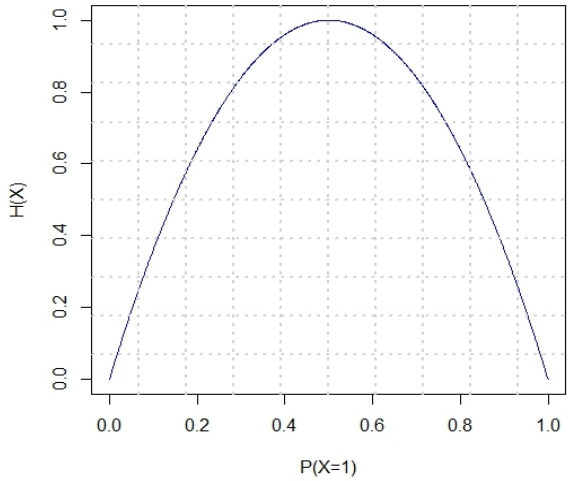
Definicija 1.3.1. *Za danu slučajnu varijablu X sa vjerojatnostima* $\mathbb{P}(x_i)$ *i za diskretan skup događaja* $x_1, ..., x_K$ *definiramo* ***Shannonovu entropiju*** *s*

$$H(X) = -\sum_{i=1}^{K} \mathbb{P}(x_i) \log(\mathbb{P}(x_i)) \tag{1.3}$$

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 (1.1). 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 1.1: 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 1.3.2. *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 1.3.3. *Uniformno distribuirani parametri imaju maksimalnu entropiju.*

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

Teorem 1.3.4. *(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 1.3.5. *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 1.3.6. *(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 (1.3.3)): 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 *Rolleovom 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 $p_M = (p_1, ..., p_k)$

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

. □

Poglavlje 2

Skriveni Markovljev model

2.1 Markovljevi lanci

Definicija 2.1.1. *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 2.1.2. *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*

$$\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = i_0) = \mathbb{P}(X_{n+1} = j | X_n = i) \tag{2.1}$$

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 (2.1) naziva se *Markovljevim svojstvom*.

Definicija 2.1.3. *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**.

2.2 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 2.2.1. ***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.
$$P(Q_t|Q_{t-1}, O_{t-1}, ..., Q_1, O_1) = \mathbb{P}(Q_t|Q_{t-1}) \tag{2.2}$$

2.
$$\mathbb{P}(O_t|Q_T, O_T, ..., Q_{t+1}, O_{t+1}, Q_t, Q_{t-1}, O_{t-1}, ..., Q_1, O_1) = P(O_t|Q_t) \tag{2.3}$$

Uočili smo da, osim niza stanja kroz koja proces prolazi (pripadne slučajne varijable označili smo sa Q_i), promatramo i niz opažanja (simbola, pripadne slučajne varijable označili smo sa O_i).

Da pojasnimo, relacija (2.2) 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 (2.3) povlači da realizacija nekog opažanja u sadašnjem stanju ovisi samo o tom stanju. Vjerojatnosti iz relacije (2.3) nazivamo **emisijska vjerojatnost**.

Skriveni Markovljev model zadan je sljedećim parametrima:

- N - broj stanja u kojima se proces može nalaziti

$$S = \{1, ..., N\} \tag{2.4}$$

  S - skup svih stanja procesa

- M - broj mogućih opažanja

$$B = \{b_1, ..., b_M\} \tag{2.5}$$

  B - skup svih opaženih vrijednosti

- L - duljina opaženog niza

$$X = (x_1, ..., x_L) \tag{2.6}$$

  X - opaženi niz

- A - matrica tranzicijskih vjerojatnosti

$$A = \{a_{ij}\}, a_{ij} = \mathbb{P}(Q_{t+1} = j | Q_t = i), 1 \leq i, j \leq N \tag{2.7}$$

- E - matrica emisijskih vjerojatnosti

$$E = \{e_j(k)\}, e_j(k) = \mathbb{P}(O_t = b_k | Q_t = j), 1 \leq j \leq N, 1 \leq k \leq M \tag{2.8}$$

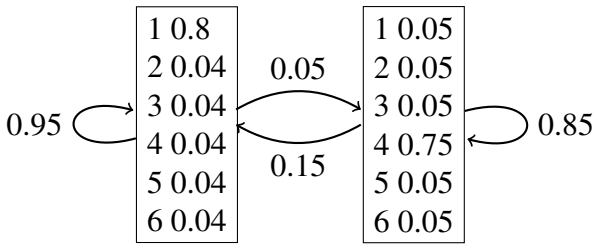
Primijetimo da nam je kod Markovljevog modela nultog reda niz stanja koji emitira neki niz simbola poznat.

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.

2.3 Primjer skrivenog Markovljevog modela

Imamo dvije nepoštene igraće kocke. Jedna kocka, koju označavamo K_1, ima vjerojatnost da dobijemo jedinicu $\frac{4}{5}$, a vjerojatnost preostalih ishoda je $\frac{1}{25}$, dok druga kocka, u oznaci K_4 ima vjerojatnost da padne četvorka $\frac{3}{4}$, a vjerojatnost preostalih ishoda je $\frac{1}{20}$.

Pretpostavimo da počinjemo sa K_1. Vjerojatnost da ćemo ponovo koristiti K_1 je 95%, dok je vjerojatnost da ćemo je zamijeniti sa K_4 5%. Kad smo jednom K_1 zamijenili sa K_4, u 85% slučajeva ćemo je i nastaviti koristiti. Vjerojatnost da je zamijenimo sa K_1 je 15%.



Slika 2.1

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

- N=2
$$S = \{K_1, K_4\}$$
- M=6
$$B = \{1, 2, 3, 4, 5, 6\}$$
- Matrica tranzicijskih vrijednosti je dana s:
$$A = \begin{pmatrix} 0.95 & 0.05 \\ 0.15 & 0.85 \end{pmatrix}$$
gdje je $a_{11} = \mathbb{P}(K_1|K_1)$ - vjerojatnost da je nakon K_1 ponovo bačena K_1, $a_{12} = \mathbb{P}(K_4|K_1)$ - vjerojatnost bacanja K_4, ako je prethodno bačena K_1, $a_{21} = \mathbb{P}(K_1|K_4)$ - vjerojatnost da je nakon bacanja K_4 bačena K_1 i $a_{22} = \mathbb{P}(K_4|K_4)$ - vjerojatnost da je nakon K_4 opet bačena K_4.
- Matrica emisijskih vjerojatnosti je:
$$E = \begin{pmatrix} 0.8 & 0.04 & 0.04 & 0.04 & 0.04 & 0.04 \\ 0.05 & 0.05 & 0.05 & 0.75 & 0.05 & 0.05 \end{pmatrix}$$
Prvi redak čine emisijske vjerojatnosti elemenata iz B u stanju K_1, a drugi redak emisijske vjerojatnosti elemenata iz B u stanju K_4

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*.

Poglavlje 3

Algoritmi za analizu HMM

Kao što se vidi iz prethodnog primjera, ako imamo niz simbola, odnosno opaženih vrijednosti, npr. $X = (1, 2, 1, 2, 5, 6, 4, 4, 3)$, nepoznat nam je *niz stanja* - on je skriven. Ipak, pomoću niza simbola moguće je:

- odrediti *najvjerojatniji niz stanja* za dani niz simbola koristeći **Viterbijev algoritam**.
- procijeniti *parametre uvjetne maksimalne vjerodostojnosti* koristeći **Viterbijevo treniranje**.
- bolje procijeniti *parametre uvjetne maksimalne vjerodostojnosti* koristeći **Viterbijevo treniranje modificirano deterministčkim kaljenjem**.

Uvedimo neke oznake. Niz opaženih simbola ćemo označiti s $X = (x_1, ..., x_n)$. Pripadajući niz skrivenih stanja nazivamo stazom $\pi = (\pi_1, ..., \pi_n)$. Za tranzicijske i emisijske vjerojatnosti koristit ćemo već spomenute oznake a_{kl} i $e_k(b)$.

Staza slijedi Markovljev lanac tako da vjerojatnost stanja u trenutku i ovisi o prethodnom stanju u trenutku $i - 1$. Lanac je karakteriziran tranzicijskim vjerojatnostima

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

pri čemu tranzicijsku vjerojatnost a_{0k} možemo smatrati vjerojatnošću da počnemo u stanju k. Emisijska vjerojatnost, odnosno vjerojatnost da je simbol b vidljiv u stanju k, definirana je s

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

3.1 Viterbijev algoritam

Cilj Viterbijevog algoritma je danom nizu opaženih simbola pridružiti najvjerojatniju stazu π^*, tj. onu stazu za koju vrijedi

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

pri čemu 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 $a_{0\pi_1} = a_{\pi_n \pi_{n+1}} = 1$. Takvu stazu π^* dobivenu Viterbijevim algoritmom nazivamo **Viterbijev put** ili **Viterbijev prolaz**.

Pretpostavimo da je za sva stanja k poznata vjerojatnost najvjerojatnije staze koja u stanju k završava sa 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 rekurzivno:

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

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):

$$\pi^*_{i-1} = ptr_i(\pi^*_i)$$

Napomenimo još da se Viterbijev algoritam u praksi uvijek izvodi u log-prostoru, odnosno, računamo $log(v_l(i))$. Time se množenje mnogo malih vjerojatnosti pretvara u zbrajanje te brojevi ostaju 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. Neka je zadan 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 iz X pridružimo stanje. Sada možemo odrediti emisijske i tranzicijske frekvencije. Prema lemi 1.2.10, relativne frekvencije su procjenitelji maksimalne vjerodostojnosti, one su novi parametri modela i proces se iterira. Opisana procedura pronalazi vrijednost θ koja maksimizira vjerodostojnost najvjerojatnijeg niza skrivenih stanja, ako se ne zaustavi u nekom lokalnom maksimumu.

3.3 Deterministično kaljenje

Kaljenje provodimo na konveksnoj kombinaciji parametara maksimalne entropije i deterministički izračunatih parametara. U početku je doprinos deterministički određenih parametara jako mali, tj. zadajemo visoku vrijednost parametra kaljenja γ na nekom intervalu, koju kroz iteracije smanjujemo te time mijenjamo omjer u korist deterministički određenih parametara. Dodavanje parametara maksimalne entropije služi izbjegavanju lokalnih optimuma.

U prvoj iteraciji zadajemo inicijalne parametre modela. Ulazni parametri za Viterbijevo treniranje su konveksne kombinacije parametara maksimalne entropije i inicijalnih parametara. Kada je izračunata najvjerojatnija staza kroz model, računamo tranzicijske i emisijske relativne frekvencije. U svakom sljedećem koraku ulazni parametri su konveksne kombinacije parametara maksimalne entropije i relativnih frekvencija izračunatih u prethodnom koraku.

Konkretnije, deterministično kaljenje provodi se 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

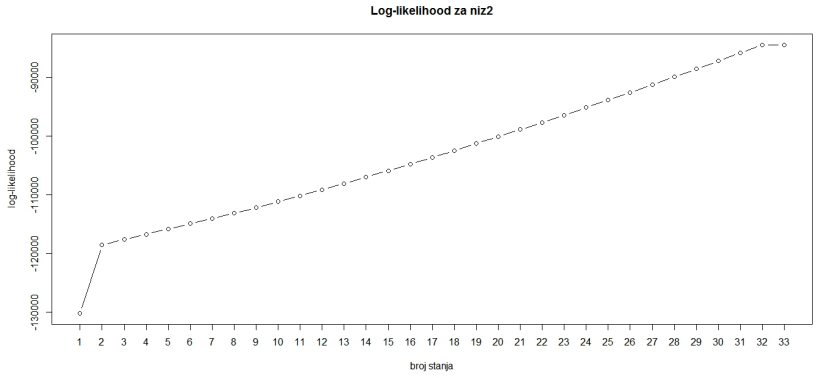
Simulirani su nizovi duljine 60000 koristeći, redom, $k = 2, 5, 15, 25$ "kocki" s 32 strane. Dakle, za niz simuliran s, primjerice, $k = 5$ "kocki", skriveni Markovljev model dan je peteročlanim skupom stanja $S = \{K_1, K_2, K_3, K_4, K_5\}$, skupom simbola $B = \{1, 2, ..., 32\}$, matricom tranzicijskih vjerojatnosti $A5$ te matricom emisijskih vjerojatnosti $E5$.

Primijetimo da je entropiju lako dovesti u vezu sa log-vjerodostojnosti. Prema definiciji, za skup simbola B, entropija je negativna suma umnožaka vjerojatnosti pojavljivanja simbola $b_i \in B$ i logaritamske vrijednosti te vjerojatnosti, dok je log-vjerodostojnost suma relativnih frekvencija emitiranja simbola, uzimajući u obzir stanje, sa logaritamskom vrijednosti vjerojatnosti emitiranja tog simbola, također uz uvjet stanja u kojem se proces nalazi, i logaritamske vrijednosti tranzicijskih vjerojatnosti među stanjima kroz koja proces prolazi. Vidimo da će za parametre za koje je entropija maksimalna, log-vjerodostojnost biti minimalna, i obrnuto. Budući da procesom optimizacije želimo maksimizirati log-vjerodostojnost, tranzicijski i emisijski parametri za "kocke" birani su tako da to budu parametri male entropije, tj. tako da svaka "kocka" ima jedan dominantan simbol te da je vjerojatnost ostanka u istoj "kocki", tj. istom stanju, značajno veća od vjerojatnosti promjene "kocke". Primjerice, niz simuliran s $k = 5$ "kocki", dobiven je iz parametara

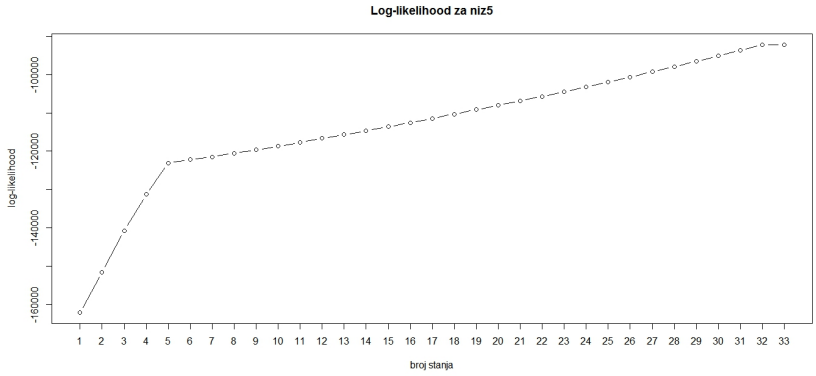
$$A5 = \begin{pmatrix} 0.9 & 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.9 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.9 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.9 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 & 0.9 \end{pmatrix}$$

$$E5 = \begin{pmatrix} 0.69 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & ... & 0.01 \\ 0.01 & 0.69 & 0.01 & 0.01 & 0.01 & 0.01 & ... & 0.01 \\ 0.01 & 0.01 & 0.69 & 0.01 & 0.01 & 0.01 & ... & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.69 & 0.01 & 0.01 & ... & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.69 & 0.01 & ... & 0.01 \end{pmatrix}$$

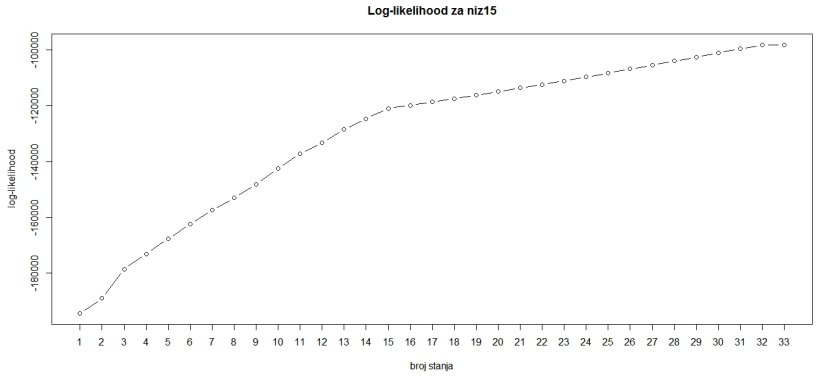
U procesu optimizacije parametre maksimalne vjerodostojnosti i samu maksimalnu vjerodostojnost modela odredili smo pomoću Viterbijevog treniranja modificiranog determinističkim kaljenjem, pri čemu je broj iteracija postavljen na 1000, no do prekida izvođenja programa moglo je doći i ranije, ako bi se log-vjerodostojnost u nekoj iteraciji smanjila. Ova se greška ne bi trebala događati jer je metoda Viterbijevog treniranja modificiranog determinističkim kaljenjem optimizacijski proces maksimizacije uvjetne vjerodostojnosti te bi vjerodostojnost trebala rasti sa svakom iteracijom. Do greške dolazi zbog postavljanja inicijalne vrijednosti svih tranzicijskih i emisijskih parametara na vrijednost različitu od nule, ali blizu nuli, npr. 0.01. Time smo pretpostavili i osigurali nemogućnost postojanja nul-parametara. Rezultati optimizacijskog procesa za $k = 2, 5, 15, 25$ prikazani su sljedećim grafovima:



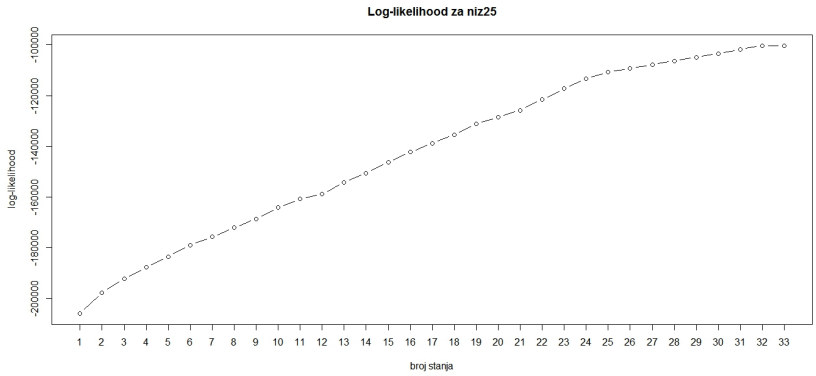
Slika 4.1: Log-vjerodostojnost za niz2



Slika 4.2: Log-vjerodostojnost za niz5



Slika 4.3: Log-vjerodostojnost za niz15



Slika 4.4: Log-vjerodostojnost za niz25

Primijetimo lakat za *niz2* i *niz5* za stanja 2 i 5. Za te slučajeve možemo iz ovih grafova doći do dobrih zaključaka o kompleksnosti, tj. broju stanja, danih modela, dok već za niz simuliran s 15 kocki broj stanja nije tako očit.

Za kraj, komentirajmo nepromijenjenu vrijednost log-vjerodostojnosti za 32 i 33 stanja. Dokle god je broj stanja manji od duljine simuliranog niza, očekujemo povećanje vjerodostojnosti s povećanjem broja stanja. Međutim, optimizacijski proces teže pronalazi takve parametre za broj stanja veći od veličine skupa simbola. Optimizacijski proces je "dodijelio" svakom stanju jedan dominantan simbol, međutim, kada imamo jedno stanje više, parametri maksimalne vjerodostojnosti više nisu tako jednostavni, optimizacijski proces ih teže nalazi, te ne dobivamo očekivano povećanje vjerodostojnosti. U našem slučaju veličina skupa simbola B je 32, te ovo svojstvo optimizacijskog procesa vidimo kada povećamo broj stanja s 32 na 33.

4.2 AIC i BIC

Dva često korištena kriterija za odabir najboljeg statističkog modela su AIC (Akaike Information Criterion) i BIC (Bayesian Information Criterion). Oba se temelje na funkciji vjerodostojnosti, a mjere koliko "dobro" model opisuje podatke. Definirani su sljedećim jednadžbama:

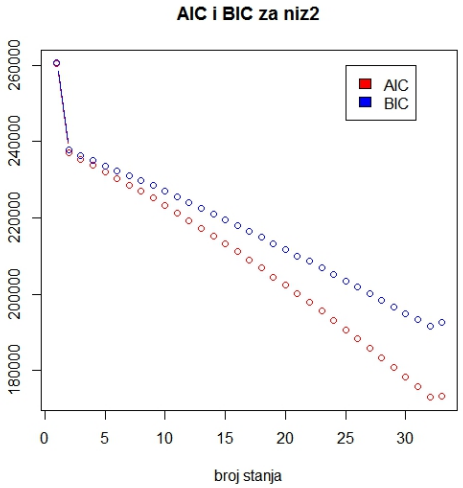
$$AIC = -2\log(L) + 2j$$

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

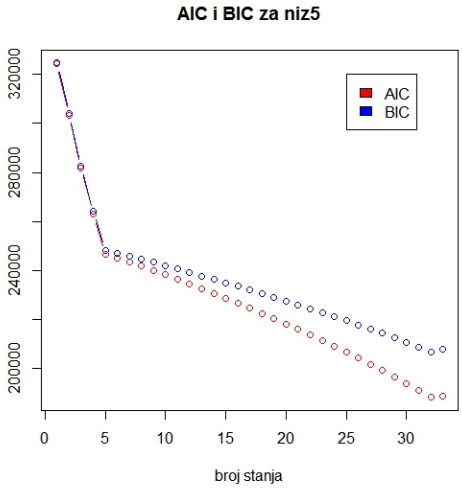
gdje je L maksimalna vjerodostojnost modela, m duljina niza, a j broj slobodnih parametara.

Cilj je imati najbolji mogući model, što se postiže dodavanjem parametara, ali želimo i da taj model bude što jednostavniji, tj. da ima što manje parametara. Zato informacijski kriteriji sadrže i penalizaciju na kompleksnost modela (tj. količinu parametara). Time ujedno mogu spriječiti 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 biti će onaj s najmanjim AIC-om (BIC-om).

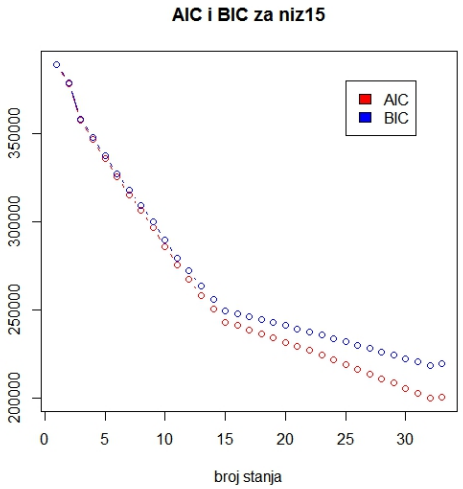
U našem slučaju duljina svakog niza je $m = 60000$, a broj slobodnih parametara $j = i(i-1) + 32i$, pri čemu i označava i-to stanje, $i = 1, 2, \ldots, 33$. AIC i BIC za nizove simulirane s $2, 5, 15, 25$ "kocki" prikazani su sljedećim grafovima:



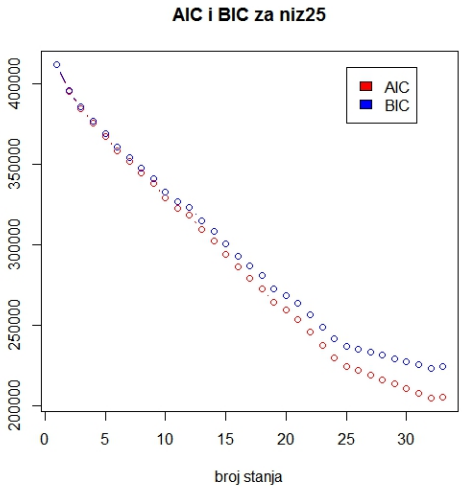
(a) AIC i BIC za niz2



(b) AIC i BIC za niz5



(c) AIC i BIC za niz15



(d) AIC i BIC za niz25

Slika 4.5

Očito da AIC i BIC u ovom slučaju nisu dobri kriteriji jer se njihova vrijednost smanjuje sa svakim dodatnim stanjem dovodeći do zaključka da je najbolja procjena broja stanja za svaki simulirani niz upravo 32.

4.3 Logaritamsko-polinomijalni kriterij kompleksnosti

Kao što smo vidjeli, grafička metoda lakta te AIC i BIC ne daju zadovoljavajuće zaključke o kompleksnosti modela. Ideja logaritamsko-polinomijalnog kriterija je naći krivulju koja najbolje opisuje očekivanu vrijednost log-vjerodostojnosti za neki opaženi niz i pretpostavljeni broj stanja $k = 1, \ldots, 33$. Tada će broj stanja čija log-vjerodostojnost dobivena optimizacijom ima najveće pozitivno odstupanje od krivulje biti upravo kompleksnost zadanog modela, tj. moći ćemo zaključiti da je upravo taj broj stanja traženi skriveni broj stanja.

Log-vjerodostojnost je suma logaritamskih vrijednosti emisijskih i tranzicijskih vjerojatnosti množenih nekim koeficijentima, koji za emisijske vjerojatnosti ovise o relativnoj frekvenciji nekog simbola u nekom stanju, a za tranzicijske ovise o položaju simbola u nizu. Upravo je zbog te ovisnosti o položaju simbola u nizu (tj. ovisnosti trenutnog stanja koje emitira opaženi simbol o prethodnom stanju), teško odrediti očekivanu log-vjerodostojnost za neki niz.

Neka funkcija $g(k)$, $k = 1, 2, \ldots, 33$ predstavlja vrijednost log-vjerodostojnosti dobivenu optimizacijskim procesom za neki niz. Definiramo funkciju

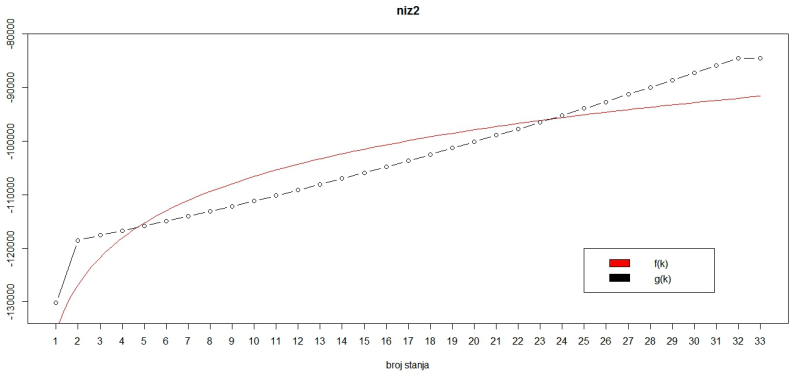
$$f(k) = \alpha \ln \frac{k}{32} + \beta,$$

gdje je k broj stanja, $k = 1, \ldots, 33$. Parametre α i β procijenjujemo metodom najmanjih kvadrata, tj. tražimo α i β takve da je vrijednost

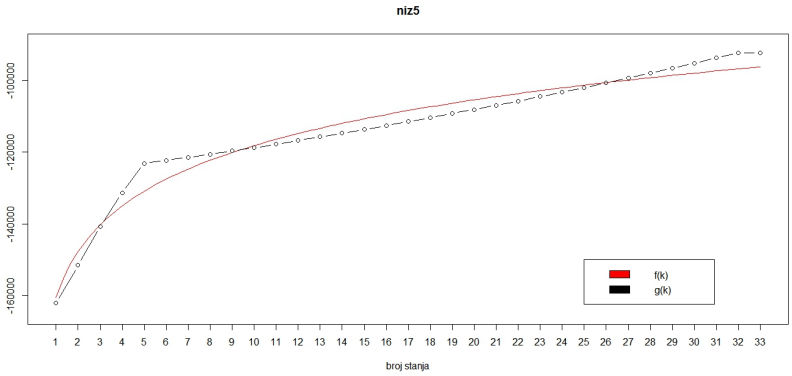
$$\sum_{k=1}^{33} (\alpha \ln \frac{k}{32} + \beta - g(k))^2$$

minimalna.

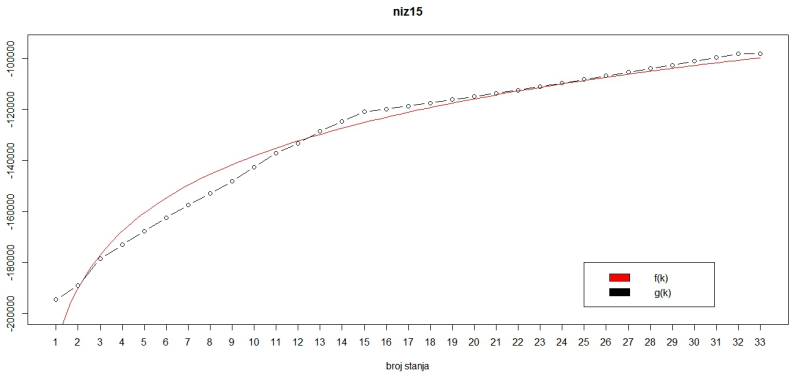
Grafovi funkcija $g(k)$ i $f(k)$, $k = 1, \ldots, 33$, za nizove simulirane sa redom $2, 5, 15, 25$ "kocki" izgledaju ovako:



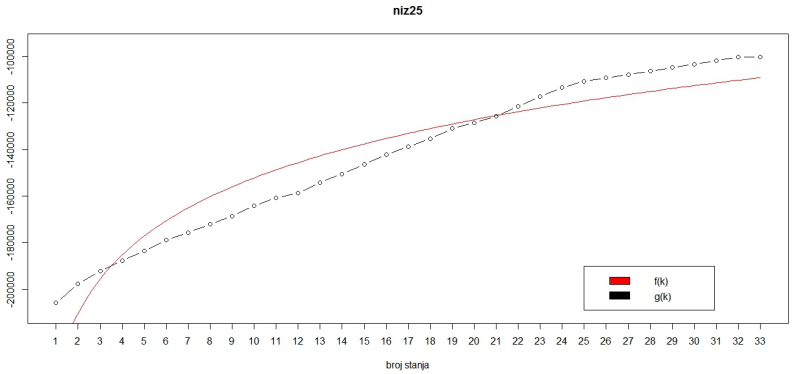
Slika 4.6: $f(k) = \alpha \ln \frac{k}{32} + \beta$ i $g(k)$ za niz2



Slika 4.7: $f(k) = \alpha \ln \frac{k}{32} + \beta$ i $g(k)$ za niz5



Slika 4.8: $f(k) = \alpha \ln \frac{k}{32} + \beta$ i $g(k)$ za niz15



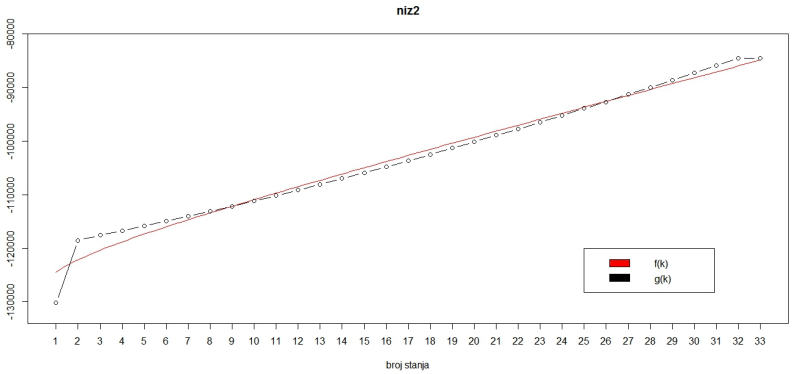
Slika 4.9: $f(k) = \alpha \ln \frac{k}{32} + \beta$ i $g(k)$ za niz25

Vidimo da krivulja ne opisuje sasvim dobro dane podatke o log-vjerodostojnosti, iako već sada možemo doći do približno dobrih zaključaka o kompleksnosti modela, zanemarimo li odstupanja za $k = 1$.

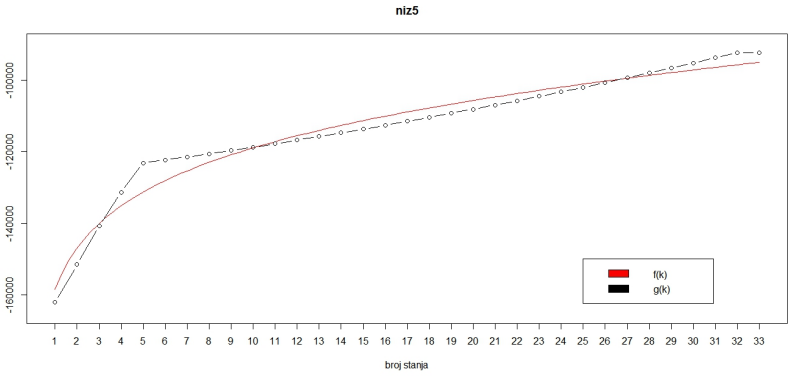
U svrhu poboljšanja krivulje, definirajmo $f(k)$ ovako:

$$f(k) = \alpha \ln \frac{k}{32} + \beta + \gamma k, k = 1, \ldots, 33.$$

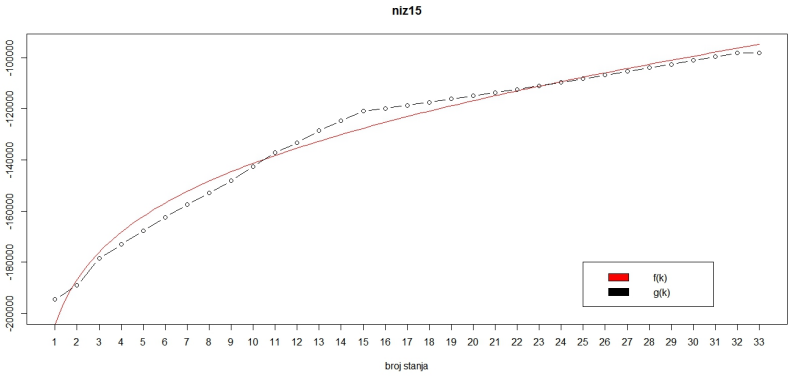
Želja nam je krivulju malo "izravnati", zato dodajemo komponentu linearne zavisnosti funkcije $f(k)$ o broju stanja k. Nakon provođenja metode najmanjih kvadrata, grafovi za ovako definiranu $f(k)$ su sljedeći:



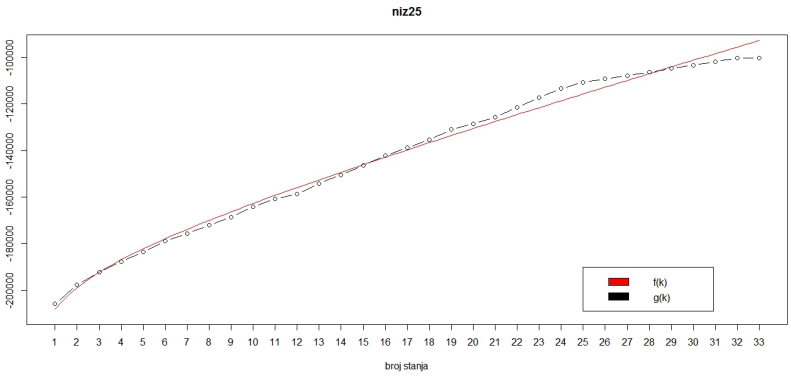
Slika 4.10: $f(k) = \alpha \ln \frac{k}{32} + \beta + \gamma k$ i $g(k)$ za niz2



Slika 4.11: $f(k) = \alpha \ln \frac{k}{32} + \beta + \gamma k$ i $g(k)$ za niz5



Slika 4.12: $f(k) = \alpha \ln \frac{k}{32} + \beta + \gamma k$ i $g(k)$ za niz15

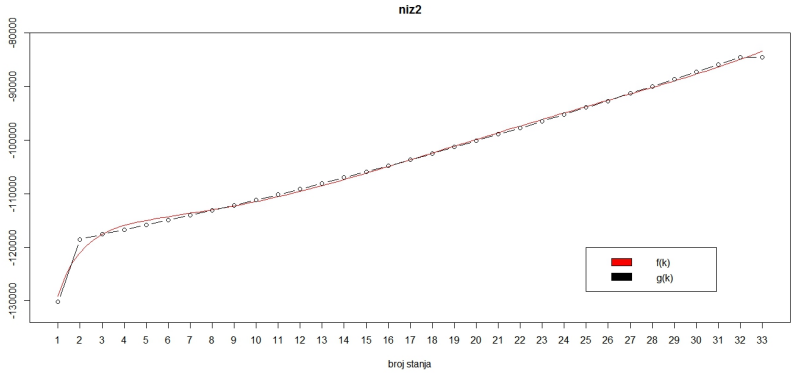


Slika 4.13: $f(k) = \alpha \ln \frac{k}{32} + \beta + \gamma k$ i $g(k)$ za niz25

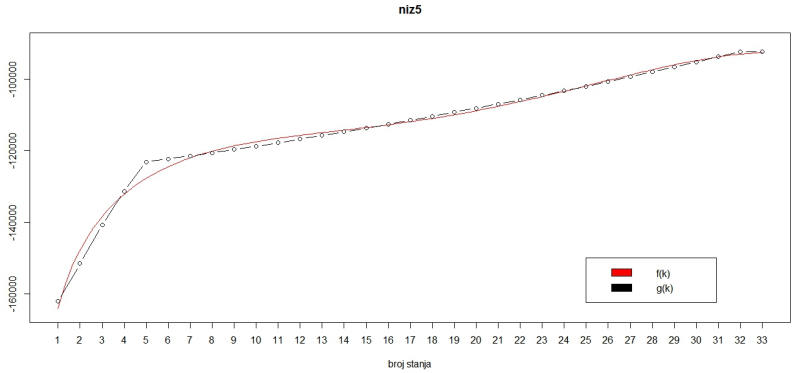
Iako je poboljšanje očigledno, još nismo sasvim zadovoljni procjenom, pogotovo u rubnim brojevima stanja (poput $k = 1$ za *niz15*). Definirajmo stoga krivulju $f(k)$ ovako:

$$f(k) = \alpha \ln \frac{k}{32} + \beta + \gamma k + \delta k^2 + \epsilon k^3 + \zeta k^4, k = 1, \ldots, 33.$$

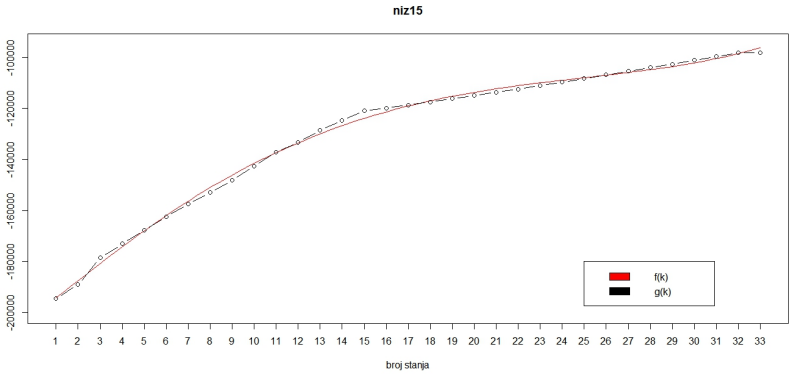
Sada grafovi izgledaju ovako:



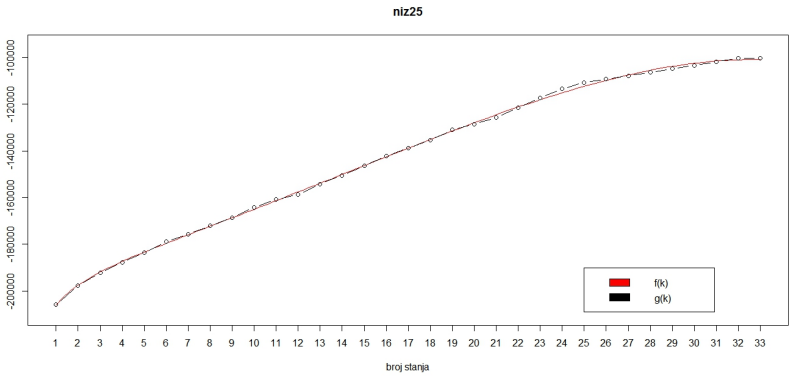
Slika 4.14: $f(k) = \alpha \ln \frac{k}{32} + \beta + \gamma k + \delta k^2 + \epsilon k^3 + \zeta k^4$ i $g(k)$ za niz2



Slika 4.15: $f(k) = \alpha \ln \frac{k}{32} + \beta + \gamma k + \delta k^2 + \epsilon k^3 + \zeta k^4$ i $g(k)$ za niz5



Slika 4.16: $f(k) = \alpha \ln \frac{k}{32} + \beta + \gamma k + \delta k^2 + \epsilon k^3 + \zeta k^4$ i $g(k)$ za niz15



Slika 4.17: $f(k) = \alpha \ln \frac{k}{32} + \beta + \gamma k + \delta k^2 + \epsilon k^3 + \zeta k^4$ i $g(k)$ za niz25

Očito je procjena log-vjerodostojnosti funkcijom $f(k)$ vrlo dobra. Međutim, za nizove *niz15* i *niz25* nije iz samih grafova na prvu vidljivo koji je broj stanja najbolji. Pogledajmo stoga tablice konkretnih vrijednosti funkcija $g(k)$, $f(k)$ te reziduala $g(k) - f(k)$, za vrijednosti k koje su nam od interesa, tj. za koje je iz grafova očito $g(k) > f(k)$:

k	$g(k)$	$f(k)$	$g(k) - f(k)$
1	-130182.91	-129190.93	-991.97373
2	-118516.61	-120934.86	**2418.24575**
3	-117595.48	-117617.78	22.29801

Tablica 4.1: niz2

k	$g(k)$	$f(k)$	$g(k) - f(k)$
4	-131315.52	-132187.64	872.12207
5	-123079.40	-127734.27	**4654.86837**
6	-122255.29	-124444.10	2188.80975
7	-121463.31	-121967.16	503.85347

Tablica 4.2: niz5

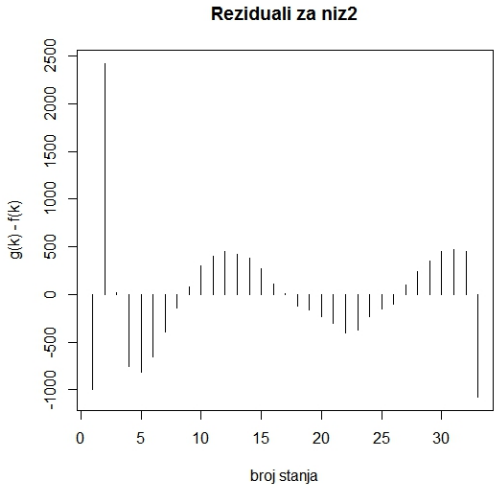
k	$g(k)$	$f(k)$	$g(k) - f(k)$
2	-188984.19	-187637.09	-1347.09623
3	-178575.76	-180877.38	2301.61991
4	-173065.02	-174296.18	1231.15864
13	-128533.99	-129938.10	1404.11502
14	-124631.96	-126754.19	2122.22960
15	-120900.65	-123889.88	**2989.22349**
16	-119846.30	-121328.50	1482.20098

Tablica 4.3: niz15

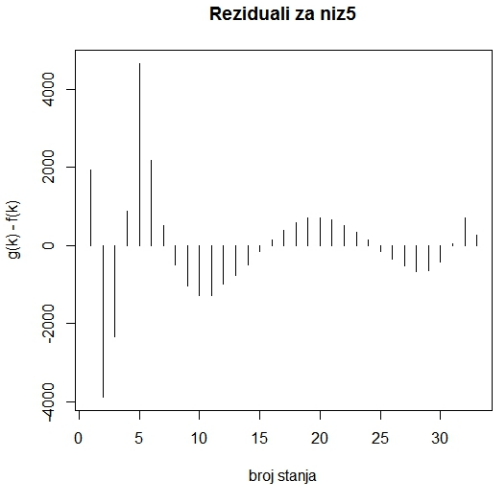
k	$g(k)$	$f(k)$	$g(k) - f(k)$
10	-164155.5	-164968.4	812.842988
11	-160786.1	-161282.3	496.153892
19	-131042.0	-131498.9	456.842403
23	-117261.8	-118056.4	794.578810
24	-113456.0	-115078.9	1622.936801
25	-110669.0	-112305.3	**1636.301808**
26	-109283.8	-109762.7	478.900203

Tablica 4.4: niz25

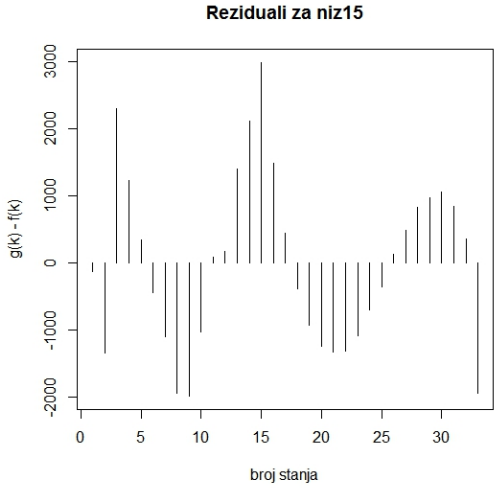
Iz tablica vidimo da je najveće odstupanje postignuto upravo za vrijednosti $k = 2, 5, 15, 25$ i to su traženi brojevi stanja koji određuju kompleksnosti opaženih modela. Isto je vidljivo i ako pogledamo grafove rezidualnih vrijednosti za opažene nizove:



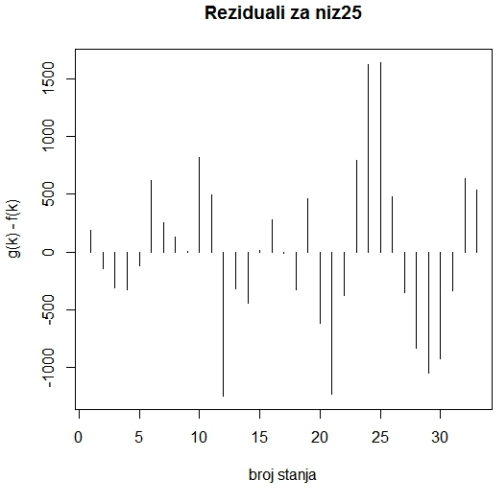
(a) Reziduali za niz2



(b) Reziduali za niz5



(c) Reziduali za niz15



(d) Reziduali za niz25

Slika 4.18

Bibliografija

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

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

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

[4] A. Mišura, *Kompleksnost skrivenih Markovljevih modela*, diplomski rad, PMF-MO, Zagreb, 2016.

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

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

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

Sažetak

U ovom diplomskom radu bavili smo se analizom skrivenih Markovljevih modela. Dana je njihova formalna definicija te opisani neki algoritmi za njihovu analizu. Algoritme smo implementirali u programskom jeziku Python te primijenili na modelu nepoštene kockarnice s 32-stranim "kockama". Prikazano je nekoliko poznatih statističkih metoda za odabir najboljeg modela, međutim, nijedna nije dala zadovoljavajući rezultat pri primjeni na procjenu kompleksnosti skrivenih Markovljevih modela. Zato u radu predlažemo novi kriterij, koji daje bolje rezultate.

Summary

This thesis is concerned with analysis of hidden Markov models. We give a formal definition of an HMM and describe several algorithms used in their analysis. We implement these algorithms in Python and show their application on an occasionally dishonest casino model with 32-sided “cubes”. We describe several known statistical methods for choosing the best model. However, none give satisfying results when used to assess the complexity of given hidden Markov model. Therefore, we propose a new criterion which gives better results.

Životopis

Rođena sam 31. kolovoza 1992. u Zagrebu. Od 1999. do 2007. pohađam Osnovnu školu Ivana Cankara u Zagrebu, a od 2007. do 2011. Gimnaziju Lucijana Vranjanina u Zagrebu. 2011. upisujem preddiplomski studij Matematika na PMF-MO u Zagrebu kojeg završavam 2015. godine, te iste godine upisujem diplomski studij Financijska i poslovna matematika na PMF-MO u Zagrebu. Udana sam i u iščekivanju kćerkice.