Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Hans Rahi

Juhuslike jadade võrdlemine kolmekaupa Markovi mudeli korral

matemaatiline statistika

Bakalaureusetöö (9 EAP)

Juhendaja: PhD Joonas Sova

Tartu 2026

JUHUSLIKE JADADE VÕRDLEMINE KOLMEKAUPA MARKOVI MUDELI KORRAL

Bakalaureusetöö

Hans Rahi

Lühikokkuvõte

Töös kirjeldatakse ja uuritakse kahe juhusliku jada võrdlemist kolmekaupa Markovi mudeli korral. Alguses tutvustatakse kolmekaupa Markovi mudelit ja pikima ühisjada pikkuse sarnasusmõõtu. Seejärel kirjeldatakse varjatud Markovi mudeli (VMM) erijuhtu kolmekaupa Markovi mudelist. Viimases peatükis uuritakse varjatud Markovi mudeleid, kus varjatud protsess on juhuslik ekslemine täisarvudel. Töö käigus läbi viidud simulatsioonid näitasid, et positiivselt korreleeritud emissioonide saartega mudeli ning positiivselt ja negatiivselt korreleeritud emissioonidega seisundites eksleva mudeli korral on pikima ühisjada pikkus suurem kui *iid*-jadade korral.

CERCS teaduseriala: P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Märksõnad: Juhuslikud protsessid, Markovi ahelad, jadad

COMPARISON OF RANDOM SEQUENCES IN THE CASE OF A TRIPLET MARKOV MODEL

Bachelor thesis

Hans Rahi

Abstract

In this thesis, we describe and explore comparison of two random sequences in the case of a triplet Markov model. At first, we describe the triplet Markov

model and the length longest common subsequence (LCS) similarity measurement. Then we explore the hidden Markov model (HMM) as a special case of the triplet Markov model. Finally, we explore HMM-s where the hidden process is a random walk on the integer line. The simulations conducted during this study demonstrated that for both the model with islands of positively correlated emissions and the model randomly walking between states of positively and negatively correlated emissions, the length of the longest common subsequence is greater than for i.i.d sequences.

CERCS research specialisation: P160 Statistics, operation research, programming, actuarial mathematics

Key Words: Stochastic processes, Markov chains, sequences

Sisukord

Sissejuhatus

Kahe jada sarnasuse mõõtmine on uurimisprobleem, mis leiab rakendust mitmes teadusharus. Üks olulisimaid valdkondi on bioinformaatikas DNA-ahelate võrdlemine, kus genoomide sarnasuse kaudu saab uurida kahe isendi või liigi sugulust.

Teoreetilises käsitluses on enamasti uuritud olukordi, kus võrreldavad jadad on teineteisest sõltumatud või moodustavad ühise Markovi ahela, ehk paarikaupa Markovi mudeli. Paarikaupa Markovi mudelit saame üldistada kolmekaupa Markovi mudeliks, kus võrreldavatele jadadele lisandub varjatud protsess, mis võimaldab modelleerida keerulisemaid sõltuvusstruktuure.

Bakalaureusetöö koosneb kolmest peatükist.

Esimeses peatükis antakse ülevaade kolmekaupa Markovi mudelist ja pikima ühisjada pikkuse sarnasusmõõdikust.

Teises peatükis tutvustatakse varjatud Markovi mudeli erijuhtu ning uuritakse simulatsioonide abil pikima ühisjada pikkuse käitumist ühe konkreetse varjatud Markovi mudeli korral.

Kolmandas peatükis uuritakse varjatud Markovi mudeleid, kus varjatud protsess on juhuslik ekslemine täisarvude hulgal ning uuritakse simulatsioonide abil pikima ühisjada pikkuse käitumist ühe konkreetse juhusliku ekslemise tüüpi mudeli korral.

1 Juhuslike jadade võrdlemise mudel

Eeldame, et käesoleva töö lugeja on tuttav Markovi ahelate teooria põhiteadmistega, millega võib tutvuda näiteks õpiku „Tõenäosusteooria algkursus“ (Pärna, 2013) 5. peatükis.

1.1 Markovi ahelate omadusi

Käesolev peatükk põhineb konspekti (Aldridge, 2021) 9. ja 10. peatükil ning õpiku (Pärna, 2013) 5. peatükil.

Olgu $X = X_1, X_2, ...$ diskreetse ajaga homogeenne Markovi ahel, mis võtab väärtusi seisundite ruumis $\mathcal{X}$. Tähistame tõenäosuse, et olles seisundis j, on Markovi ahel k sammu pärast seisundis i

$$p_{ij}(k) := P(X_{n+k} = j | X_n = i),$$

ja tõenäosuse, et seisundist j lähtuv Markovi ahel naaseb sellesse seisundisse esimest korda k sammu pärast

$$f_j(k) := P(X_{n+k} = j, X_{n+k-1} \neq j, ..., X_{n+1} \neq j | X_n = j).$$

Tuletame meelde korduva seisundi definitsiooni

Definitsioon 1.1. Markovi ahela seisund j on *korduv*, kui

$$F_j := \sum_{k=1}^{\infty} f_j(k) = 1.$$

Kui $F_j < 1$, siis on seisund j *mööduv*.

Toome välja ka alternatiivse tingimuse korduvuse kontrollimiseks

Lause 1.1. *Markovi ahela seisund j on korduv parajasti siis, kui*

$$\sum_{k=1}^{\infty} p_j(k) = \infty$$

Tõestus. Õpikus (Pärna, 2013, Teoreem 5.2). □

Tähistame esimese aja, mil Markovi ahel X jõuab esimest korda pärast n. sammu seisundisse j

$$T_n(j) = \min\{k \geqslant 1 | X_{n+k} = j\}.$$

Paneme tähele, et $T_n(j)$ on juhuslik suurus, kusjuures statsionaarse Markovi ahela korral ei sõltu $T_n(j)$ jaotus arvu n valikust. Kui X on korduv, siis arvestades, et $\sum_{k=1}^{\infty} f_j(k) = 1$, saame esitada keskmise naasmisaja seisundisse j kujul

$$\begin{aligned}\mu_j := E(T_n(j)|X_n = j) &= \sum_{k=n}^{\infty} kP(X_{k+n} = j, X_k \neq j, ..., X_2 \neq j | X_n = j) \\ &= \sum_{k=n}^{\infty} k f_j(n),\end{aligned}$$

kusjuures Markovi ahela X homogeensuse tõttu ei sõltu juhusliku suuruse $T_n(j)$ tinglik jaotus tingimusel $X_n = j$ arvu n valikust. Keskmise naasmisaja järgi saame jagada korduvaid seisundeid kaheks.

Definitsioon 1.2. Olgu j Markovi ahela korduv seisund. Kui $\mu_j < \infty$, siis j on *positiivselt korduv (ingl. k. positive recurrent)*. Kui $\mu_j = \infty$, siis j on *null-korduv (ingl. k. null recurrent)*.

Toome välja kaks piisavat tingimust, et kaasnevate seisundite klass oleks positiivselt korduv.

Lause 1.2. *Kehtivad järgmised omadused:*

(i) (Solidaarsusteoreem) Korduvas kaasnevate seisundite klassis $\mathcal{S} \subseteq \mathcal{X}$ on kõik seisundid positiivselt korduvad või kõik seisundid null-korduvad.

(ii) Kui $\mathcal{S} \subseteq \mathcal{X}$ on lõplik kaasnevate ja oluliste seisundite klass, siis klass $\mathcal{S}$ on positiivselt korduv.

(Aldridge, 2021, Alapeatükis 9.3)

Järgneva tulemuse põhjal saab positiivselt korduvuse kaudu teha kindlaks statsionaarse jaotuse olemasolu.

Lause 1.3. *Mittelahutuval Markovi ahelal X leidub statsionaarne jaotus $\pi_X = (\pi_X(1), \pi_X(2), ...)$ parajasti siis, kui X on positiivselt korduv. Seejuures statsionaarsed tõenäosused on üheselt määratud võrdustega $\pi_X(j) = \frac{1}{\mu_j}$.*

Tõestus. Konspektis (Aldridge, 2021, Teoreem 10.1). □

1.2 Kolmekaupa Markovi mudel

Üks lihtsamaid mudeleid, mida juhuslike jadade võrdlemisel kasutatakse on *iid*-mudel (inglise keeles *independent and identically distributed*)

Definitsioon 1.3. Käesolevas töös nimetame *iid-mudeliks* juhuslike jadade $X = X_1, X_2, ...$ ja $Y = Y_1, Y_2, ...$ paari, kus jadade X ja Y elemendid on iga $k \in \mathbb{N}$ korral sõltumatud juhuslikud suurused jaotusega $X_k \sim Be(0{,}5),\ Y_k \sim Be(0{,}5)$.

Kolmekaupa Markovi mudeliks nimetame homogeenset Markovi ahelat $W = W_1, W_2, ... = (X_1, Y_1, Z_1), (X_2, Y_2, Z_2), ...,$ kus marginaalprotsessid X ja Y

võtavad väärtusi vastavatest lõplikest seisundite ruumidest $\mathcal{X}$ ja $\mathcal{Y}$ ning marginaalprotsess Z võtab väärtusi ülimalt loenduvast seisundite ruumist $\mathcal{Z}$. Markovi ahel W võtab väärtusi ruumis $\mathcal{W} \subseteq \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Kui $\mathcal{Z} = \emptyset$, saame erijuhu, kus $W = (X, Y)$ on *paarikaupa Markovi mudel*. Paarikaupa Markovi mudelitest on pikemalt kirjutatud bakalaureusetöös (Iher, 2021). Kasutame edaspidi vektorite esitamisel tähistust

$$a_{1:n} := (a_1, ..., a_n).$$

Käesolevas töös käsitleme olukordi, kus täheldatakse ainult lõplikke jadasid $X_{1:n} \in \mathcal{X}^n$ ja $Y_{1:n} \in \mathcal{Y}^n$. Protsess Z on varjatud ehk latentne protsess, millega modelleerime erinevaid sõltuvusstruktuure.

Jadade võrdlemisel kasutame ühisjada mõistet.

Definitsioon 1.4. Jada $x_{1:n}$ osajadaks nimetatakse jada $z_{1:k}$, mille korral leiduvad indeksid

$$1 \leqslant i_1 < i_2 < ... < i_k \leqslant n$$

selliselt, et

$$x_{i_l} = z_l, \quad l = 1, 2, ...k.$$

Definitsioon 1.5. Jadade $x_{1:n}$ ja $y_{1:m}$ *ühisjadaks* nimetatakse ühist osajada, ehk jada $z_{1:k}$, mille korral leiduvad indeksid

$$1 \leqslant i_1 < i_2 < ... < i_k \leqslant n$$

ja

$$1 \leqslant j_1 < j_2 < ... < j_k \leqslant m$$

selliselt, et

$$x_{i_l} = y_{j_l} = z_l, \quad l = 1, 2, ...k.$$

Ühisjadade kaudu on defineeritud laialdaselt kasutatav jadade võrdlemise sarnasusmõõdik.

Definitsioon 1.6. Jadade $x_{1:n}$ ja $y_{1:m}$ *pikimaks ühisjadaks* nimetatakse nende jadade ühisjada, mille pikkus on maksimaalne ehk millest pikemat ühisajada ei leidu.

Näide 1.1. Jadade $x_{1:10} = HEINAMAADE$ ja $y_{1:12} = HEERINGAVAAL$ pikim ühisjada on $z_{1:7} = HEINAAA$.

Edaspidi uurime jadade $X_{1:n}$ ja $Y_{1:n}$ pikima ühisjada pikkust, mida tähistame $L_n = L(X_{1:n}, Y_{1:n})$. Esitame olulise tulemuse pikima ühisjada pikkuse kohta.

Teoreem 1.1. *Olgu W statsionaarne, mittelahutuv ja positiivselt korduv ruumis $\mathcal{W}$. Siis leidub konstant γ nii, et*

$$\lim_{n\to\infty} \frac{L_n}{n} = \gamma \quad p.k, \tag{1}$$

ja

$$\lim_{n\to\infty} \frac{E(L_n)}{n} = \sup_n \frac{E(L_n)}{n} = \gamma. \tag{2}$$

Tõestus. Kuna pikima ühisjada pikkus on superaditiivne funktsioon, saab antud eeldustel saab rakendada Kingmani subaditiivset ergoodilist teoreemi ja Fekete lemmat. (Vt detaile magistritööst (Sova, 2015, Järeldus 1.1)) □

Eelnevas teoreemis kirjeldatud konstanti γ nimetatakse *Chvatal-Sankoffi* konstandiks. On teada, et *iid*-mudeli korral on γ väärtus ligikaudu 0,812 (Bukh ja Cox, 2022). Kui W on mittelahutuv ja positiivselt korduv, siis lause (1.3) järgi

leidub sellel statsionaarne jaotus π_W. Seega, võttes W algjaotuseks statsionaarse jaotuse, on protsess W statsionaarne ja täidab teoreemi (1.1) eeldusi. Samuti on teoreemi eeldused täidetud, kui $\mathcal{W}$ on lõplik ja W on mittelahutuv, sest lause (1.2) järgi on W positiivselt korduv.

Mõlemad eelmises lõigus kirjeldatud eritingimused säilitasid eelduse, et W on mittelahutuv. Statsionaarse, kuid lahutuva Markovi ahela korral ei pruugi L_n/n koonduda konstandiks.

Näide 1.2. Olgu W paarikaupa Markovi mudel seisundite ruumil $\mathcal{W} = \{(0,0),(1,0)\}$. Võtame W algjaotuse vektoriks

$$\pi_W = (P(W_1 = (0,0)), P(W_1 = (1,0))) = (0{,}5,\ 0{,}5)$$

ja üleminekumaatriksiks

$$P = \begin{array}{c} \\ (0,0) \\ (0,1) \end{array} \begin{array}{c} \begin{array}{cc} (0,0) & (1,0) \end{array} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array}.$$

Seega W on lahutuv Markovi ahel, mis jääb kordama alguses võetud seisundit. Näeme, et W on statsionaarne, sest

$$\pi_W P = (0{,}5 \cdot 1 + 0,\ 0 + 0{,}5 \cdot 1) = (0{,}5;\ 0{,}5) = \pi_W.$$

Seejuures iga $n \in \mathbb{N}$ korral

$$\begin{aligned} P(L_n = 0) &= P(X_1 \neq Y_1, \ldots X_n \neq Y_n) = P(W_1 = \ldots = W_n = (1,0)) \\ &= P(W_1 = (1,0)) = 0{,}5, \end{aligned}$$

ning analoogiliselt

$$P(L_n = 1) = P(W_1 = (0,0)) = 0{,}5.$$

Järelikult

$$\lim_{n\to\infty} \frac{L_n}{n} = \frac{L_n}{n} = Be(0{,}5).$$

Teoreemi (1.1) abil võime tõestada tulemusi ka mittejuhuslike jadade kohta.

Järeldus 1.1.1. *Olgu $x_1, x_2, ...$ ja $y_1, y_2, ...$ naturaalarvude jadad, mõlemad perioodiga k. Siis jada $(L(x_{1:n}, y_{1:n})/n)$ koondub.*

Tõestus. Olgu (x_n) ja (y_n) naturaalarvude jadad perioodiga k. Siis leiduvad naturaalarvud $a_1, a_2, ..., a_k$ ja $b_1, b_2, ..., b_k$ nii, et $x_{mk+i} = a_i$ ja $y_{mk+i} = b_i$ iga $m \in \mathbb{N} \cup \{0\},\ i \in \{1, 2, ..., k\}$ korral. Vaatame jadasid paarikaupa Markovi mudelina $W = (X, Y)$, mis võtab väärtusi lõplikul hulgal $\{(a_i, b_i) | 1 \leqslant i \leqslant k\}$. Kasutame arvupaaride jaoks tähistust $c_i = (a_i, b_i)$, $i \in \{1, 2, ..., k\}$. Võtame üleminekumaatriksiks

$$P = \begin{array}{c} W \\ c_1 \\ c_2 \\ ... \\ c_{k-1} \\ c_k \end{array} \begin{array}{c} \begin{array}{ccccccc} c_1 & c_2 & c_3 & c_4 & ... & c_{k-1} & c_k \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & ... & 0 & 0 \\ 0 & 0 & 1 & 0 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & 0 & ... & 0 & 1 \\ 1 & 0 & 0 & 0 & ... & 0 & 0 \end{pmatrix} \end{array}. \tag{3}$$

Markovi ahel (X, Y) on mittelahutuv ja lause 1.2 *(ii)* järgi positiivselt korduv. Kuna $\mu_{c_i} = k$ iga $i \in \{1, 2, ..., k\}$ korral, omab (X, Y) lause 1.3 järgi statsionaarset algjaotust $\pi = \left(\frac{1}{k}, \frac{1}{k}, ..., \frac{1}{k}\right)$. Kui võtta algjaotuseks π, on see

Markovi ahel statsionaarne. Seega on teoreemi (1.1) eeldused täidetud ja jada $(L(X_{1:n}, Y_{1:n})/n)$ koondub peaaegu kindlasti.

Veendume, et jada $(L(x_{1:n}, y_{1:n})/n)$ koondub. Oletame vastuväiteliselt, et see jada ei koondu. Sellisel juhul ei koondu ka Markovi ahel W algseisundi $W_1 = (x_1, y_1) = c_1$, korral. Kuna

$$\begin{aligned} P(W_{1:n} = (x_{1:n}, y_{1:n})) &= P(W_n = (x_n, y_n)|W_{n-1} = (x_{n-1}, y_{n-1}))... \\ &...P(W_2 = (x_2, y_2)|W_1 = (x_1, y_1)P(W_1 = (x_1, y_1)) \\ &= P(W_1 = (x_1, y_1)) = \pi(c_1) = \frac{1}{k}, \end{aligned}$$

on jada $(L(X_{1:n}, Y_{1:n})/n)$ koondumise tõenäosus ülalt tõkestatud

$$P\left(\lim_{n\to\infty} \frac{L(X_{1:n}, Y_{1:n})}{n} = \gamma\right) \leqslant 1 - \frac{1}{k} < 1, \quad \text{kus } \gamma \in [0, 1].$$

Saadud tulemus on vastuolus asjaoluga, et $(L(X_{1:n}, Y_{1:n})/n)$ koondub peaaegu kindlasti. Seega jada $(L(x_{1:n}, y_{1:n})/n)$ koondub. □

2 Varjatud Markovi mudel

Kasutame tähistusi

$$\mathcal{U} = \{(x, y) : (x, y, z) \in \mathcal{W}\}$$

ja

$$U = U_1, U_2, ... = (X_1, Y_1),(X_2, Y_2), ...$$

Vaatame niinimetatud *varjatud Markovi mudelit*, kus Z_n sõltub ainult Z_{n-1}-st ja U_n sõltub vaid Z_n-st. Esitame ka formaalse definitsiooni.

Definitsioon 2.1. Markovi ahelat W nimetame varjatud Markovi mudeliks (VMM), kui tema üleminekutõenäosused faktoriseeruvad kujul

$$q_W(u, z|u', z') = q_Z(z|z')f(u|z),$$

kus

$$\sum_{u \in \mathcal{U}} f(u|z) = 1$$

ja

$$q_W(u, z|u', z') = P(W_k = (u, z)|W_{k-1} = (u', z')), \quad k = 2, 3, ...$$

on Markovi ahela W üleminekutõenäosused.

2.1 Varjatud Markovi Mudeli omadused

Esitame olulised tulemused varjatud Markovi mudeli üleminekutõenäosuste kohta.

Lause 2.1. *Olgu W varjatud Markovi mudel. Siis kehtivad järgmised omadused:*

(i) Iga $k \geqslant 2$ korral

$$P(W_k = (u, z)|Z_{k-1} = z') = q_Z(z|z')f(u|z). \tag{4}$$

(ii) Mis tahes $k \geqslant 2$ korral kehtivad võrdused

$$q_Z(z|z') = P(Z_k = z|Z_{k-1} = z'), \tag{5}$$

$$f(u|z) = P(U_k = u|Z_k = z). \tag{6}$$

(iii) Juhuslik protsess Z on homogeenne Markovi ahel üleminekutõenäosustega $q_Z(z|z')$.

Tõestus. (i) Olgu $k \geqslant 2, (u, z) \in \mathcal{W}$, ja $u' \in \mathcal{U}$ suvalised. Võrdus (4) kehtib, sest

$$\begin{aligned}
&P(W_k = (u, z)|Z_{k-1} = z') = \sum_{u' \in \mathcal{U}} P(W_k = (u, z), U_{k-1} = u'|Z_{k-1} = z') \\
&= \sum_{u' \in \mathcal{U}} P(W_k = (u, z)|W_{k-1} = (u', z'))P(U_{k-1} = u'|Z_{k-1} = z') \\
&= \sum_{u' \in \mathcal{U}} q_W(u, z|u', z')P(U_{k-1} = u'|Z_{k-1} = z') \\
&= q_Z(z|z')f(u|z) \sum_{u' \in \mathcal{U}} P(U_{k-1} = u'|Z_{k-1} = z') = q_Z(z|z')f(u|z).
\end{aligned}$$

(ii) Olgu $k \geqslant 2$. Näeme, et iga $z, z' \in \mathcal{Z}$ korral

$$\begin{aligned}
&P(Z_k = z|Z_{k-1} = z') = \sum_{u \in \mathcal{U}} P(W_k = (u, z)|Z_{k-1} = z') \\
&= \sum_{u \in \mathcal{U}} \sum_{u' \in \mathcal{U}} P(W_k = (u, z), U_{k-1} = u'|Z_{k-1} = z')
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{u\in\mathcal{U}}\sum_{u'\in\mathcal{U}} P(W_k=(u,z)|W_{k-1}=(u',z'))P(U_{k-1}=u'|Z_{k-1}=z') \\
&= \sum_{u\in\mathcal{U}}\sum_{u'\in\mathcal{U}} q_W(u,z|u',z')P(U_{k-1}=u'|Z_{k-1}=z') \\
&= \sum_{u\in\mathcal{U}}\sum_{u'\in\mathcal{U}} q_Z(z|z')f(u|z)P(U_{k-1}=u'|Z_{k-1}=z') \\
&= q_Z(z|z')\sum_{u\in\mathcal{U}} f(u|z)\sum_{u'\in\mathcal{U}} P(U_{k-1}=u'|Z_{k-1}=z') \\
&= q_Z(z|z')\sum_{u\in\mathcal{U}} f(u|z) = q_Z(z|z').
\end{aligned}
$$

Seega võrdus (5) kehtib. Tõestame võrduse (6). Paneme tähele, et iga $(u,z)\in\mathcal{W}$, $z'\in\mathcal{Z}$ korral

$$
\begin{aligned}
P(W_k=(u,z)) &= \sum_{z'\in\mathcal{Z}}\sum_{u'\in\mathcal{U}} P(W_k=(u,z)|W_{k-1}=(u',z'))P(W_{k-1}=(u',z')) \\
&= \sum_{z'\in\mathcal{Z}}\sum_{u'\in\mathcal{U}} q_Z(z|z')f(u|z)P(W_{k-1}=(u',z')) \\
&= f(u|z)\sum_{z'\in\mathcal{Z}} q_Z(z|z')\sum_{u'\in\mathcal{U}} P(W_{k-1}=(u',z')) \\
&= f(u|z)\sum_{z'\in\mathcal{Z}} q_Z(z|z')P(Z_{k-1}=z') \\
&= f(u|z)\sum_{z'\in\mathcal{Z}} P(Z_k=z|Z_{k-1}=z')P(Z_{k-1}=z') \\
&= f(u|z)P(Z_k=z).
\end{aligned}
$$

Järelikult

$$
P(U_k=u|Z_k=z) = \frac{P(W_k=(u,z))}{P(Z_k=z)} = \frac{f(u|z)P(Z_k=z)}{P(Z_k=z)} = f(u|z).
$$

(iii) Olgu $n \in \mathbb{N}$ suvaline. Näeme, et iga $z_{1:n} \in \mathcal{Z}^n$ korral

$$
\begin{aligned}
&P(Z_n = z_n, ..., Z_1 = z_1) = \sum_{u_{1:n} \in \mathcal{U}^n} P(W_n = (u_n, z_n), ..., W_1 = (u_1, z_1)) \\
&= \sum_{u_{1:n} \in \mathcal{U}^n} P(W_n = (u_n, z_n) | W_{n-1} = (u_{n-1}, z_{n-1}))... \\
&\qquad ...P(W_2 = (u_2, z_2) | W_1 = (u_1, z_1)) P(W_1 = (u_1, z_1)) \\
&= \sum_{u_{1:n} \in \mathcal{U}^n} q_W(u_n, z_n | u_{n-1}, z_{n-1}) ... q_W(u_2, z_2 | u_1, z_1) P(W_1 = (u_1, z_1)) \\
&= \sum_{u_{1:n} \in \mathcal{U}^n} q_Z(z_n | z_{n-1}) f(u_n | z_n) ... q_Z(z_2 | z_1) f(u_2 | z_2) f(u_1 | z_1) P(Z_1 = z_1) \\
&= q_Z(z_n | z_{n-1}) ... q_Z(z_2 | z_1) P(Z_1 = z_1) \sum_{u_{1:n} \in \mathcal{U}^n} f(u_n | z_n) ... f(u_1 | z_1) \\
&= q_Z(z_n | z_{n-1}) ... q_Z(z_2 | z_1) P(Z_1 = z_1),
\end{aligned}
$$

Kuna võrduse (5) põhjal on $z \mapsto q_Z(z|z')$ tõenäosusjaotus iga $z' \in \mathcal{Z}$ korral, on Z homogeenne Markovi ahel. □

Markovi ahelat Z nimetatakse ka varjatud Markovi mudeli *režiimiks*, jada U elemente nimetatakse varjatud Markovi mudeli *emissioonideks*. Varjatud Markovi mudeli korral võime vaikimisi eeldada, et W on määratud ruumis

$$\mathcal{W} = \{(u, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} : f(u|z) > 0\}.$$

Kasutame edaspidi tähistusi

$$
\begin{aligned}
p^Z_{z'z}(n) &:= P(Z_{k+n} = z | Z_k = z'), \\
p^W_{(u',z')(u,z)}(n) &:= P(W_{k+n} = (u, z) | W_k = (u', z')),
\end{aligned}
$$

kus k on suvaline naturaalarv. Tõestame abitulemuse tõenäosuste $p^Z_{z'z}(n)$ ja

$p^W_{(u',z')(u,z)}(n)$ kohta.

Lemma 2.1. *Iga* $(u,z),(u',z') \in \mathcal{W}, n \in \mathbb{N}$ *korral*

$$p^W_{(u',z')(u,z)}(n) = f(u|z)p^Z_{z'z}(n) \tag{7}$$

Tõestus. Olgu $(u,z),(u',z') \in \mathcal{W}$ ja $n \in \mathbb{N}$ suvalised. Täistõenäosuse valemi järgi üle kõigi võimalike jada $W_k, W_{k+1}, ..., W_{k+n},\ (k \in \mathbb{N})$ realisatsioonide summeerides saame

$$\begin{aligned}
&p^W_{(u',z')(u,z)}(n) = P(W_{k+n} = (u,z)|W_k = (u',z')) \\
&= \sum_{((u_1,z_1),...,(u_{n-1},z_{n-1}))\in\mathcal{W}^{n-1}} P(W_{k+n} = (u,z), W_{k+n-1} = (u_{n-1},z_{n-1}), ... \\
&\qquad\qquad ..., W_{k+1} = (u_1,z_1)|W_k = (u',z')) \\
&= \sum_{((u_1,z_1),...,(u_{n-1},z_{n-1}))\in\mathcal{W}^{n-1}} P(W_{k+n} = (u,z)|W_{k+n-1} = (u_{n-1},z_{n-1}))... \\
&\qquad\qquad ...P(W_{k+1} = (u_1,z_1)|W_k = (u',z')) \\
&= \sum_{((u_1,z_1),...,(u_{n-1},z_{n-1}))\in\mathcal{W}^{n-1}} q_W(u,z|u_{n-1},z_{n-1})...q_W(u_1,z_1|u',z') \\
&= f(u|z) \sum_{z_{1:(n-1)}\in\mathcal{Z}^{n-1}} q_Z(z|z_{n-1})...q_Z(z_1|z') \sum_{u_{1:(n-1)}\in\mathcal{U}^{n-1}} f(u_{n-1}|z_{n-1})...f(u_1|z_1) \\
&= f(u|z) \sum_{z_{1:(n-1)}\in\mathcal{Z}^{n-1}} q_Z(z|z_{n-1})...q_Z(z_1|z') = f(u|z)p^Z_{z'z}(n) > 0.
\end{aligned}$$

Arvestades, et

$$\begin{aligned}
&\sum_{u_{1:(n-1)}\in\mathcal{U}^{n-1}} f(u_{n-1}|z_{n-1})...f(u_1|z_1) = \\
&= \sum_{u_1\in\mathcal{U}_{z_1}} f(u_1|z_1) \sum_{u_2\in\mathcal{U}_{z_2}} f(u_2|z_2) \ ... \sum_{u_{n-1}\in\mathcal{U}_{z_{n-1}}} f(u_{n-1}|z_{n-1})
\end{aligned}$$

$$= \sum_{u_1 \in \mathcal{U}} f(u_1|z_1) \sum_{u_2 \in \mathcal{U}} f(u_2|z_2) \ ... \sum_{u_{n-1} \in \mathcal{U}} f(u_{n-1}|z_{n-1}) = 1,$$

saame

$$f(u|z) \sum_{z_{1:(n-1)} \in \mathcal{Z}^{n-1}} q_Z(z|z_{n-1})...q_Z(z_1|z') \sum_{u_{1:(n-1)} \in \mathcal{U}^{n-1}} f(u_{n-1}|z_{n-1})...f(u_1|z_1)$$
$$= f(u|z) \sum_{z_{1:(n-1)} \in \mathcal{Z}^{n-1}} q_Z(z|z_{n-1})...q_Z(z_1|z') = f(u|z)p^Z_{z'z}(n).$$

Seega võrdus (7) kehtib. □

Kui eeldada, et Z on mittelahutuv, saame tõestada W kohta täiendavaid omadusi.

Lause 2.2. *Olgu W varjatud Markovi mudel nii, et Z on mittelahutuv ruumis $\mathcal{Z}$. Siis Kehitvad järgmised omadused*

(i) W on mittelahutuv;

(ii) W on korduv parajasti siis, kui Z on korduv;

(iii) W on positiivselt korduv parajasti siis, kui Z on positiivselt korduv;

(iv) W on statsionaarse jaotusega

$$\pi_W(u, z) = f(u|z)\pi_Z(z), \quad (u, z) \in \mathcal{W}, \tag{8}$$

kus $\pi_Z(z)$ on Markovi ahela Z statsionaarne jaotus.

Tõestus. (i) Olgu Z mittelahutuv. Siis iga $k \in \mathbb{N}$, $(u, z), (u', z') \in \mathcal{W}$ korral leiduvad $n, m \in \mathbb{N}$ nii, et $p^Z_{z'z}(n) > 0$ ja $p^Z_{zz'}(m) > 0$. Lemma (2.1) järgi

$$p^W_{(u',z')(u,z)}(n) = f(u|z)p^Z_{z'z}(n) > 0,$$

$$p^W_{(u,z)(u',z')}(m) = p^Z_{zz'}(m) f(u'|z') > 0,$$

seega W on mittelahutuv.

(ii) Kasutame tõestuses lause (1.1) poolt antud tingimust.

Tarvilikkus. Eeldame, et W on korduv. Olgu $(u, z) \in \mathcal{W}$. Siis iga $n \in \mathbb{N}$ korral

$$p^W_{(u,z)(u,z)}(n) \leqslant \sum_{u' \in \mathcal{U}} p^W_{(u,z)(u',z)}(n) = \sum_{u' \in \mathcal{U}} p^Z_{zz}(n) f(u'|z) = p^Z_{zz}(n).$$

Markovi ahela W on korduvusest järeldub, et

$$\sum_{n=1}^{\infty} p^Z_z(n) \geqslant \sum_{n=1}^{\infty} p^W_{(u,z)}(n) = \infty$$

Seega Z on korduv.

Piisavus. Olgu Z korduv. Siis

$$\sum_{n=1}^{\infty} p^W_{(u,z)(u,z)}(n) = \sum_{n=1}^{\infty} f(u|z) p^Z_{zz}(n) = f(u|z) \sum_{n=1}^{\infty} p^Z_z(n) = \infty.$$

Järelikult W on korduv.

(iii) Kasutame lausest (1.3) teadmist, et Markovi ahel on positiivselt korduv parajasti siis, kui sellel leidub statsionaarne jaotus.

Tarvilikkus. Olgu W positiivselt korduv. Siis leidub protsessil W statsionaarne jaotus, mis on määratud tõenäosustega $\pi_W(u, z)$, kus $(u, z) \in \mathcal{W}$. Näitame, et võrdustega

$$\pi_Z(z) = \sum_{u \in \mathcal{U}} \pi_W(u, z)$$

määratud tõenäosused moodustavad protsessi Z statsionaarse jaotuse, ehk

täidavad tingimust

$$\pi_Z(z) = \sum_{z' \in \mathcal{Z}} q_Z(z|z')\pi_Z(z').$$

Näeme, et iga $z \in \mathcal{Z}$ korral

$$\begin{aligned}
\pi_Z(z) &= \sum_{u \in \mathcal{U}} \pi_W(u, z) \\
&= \sum_{u \in \mathcal{U}} \sum_{(u', z') \in \mathcal{W}} \pi_W(u', z') q_W(u, z|u', z') \\
&= \sum_{u \in \mathcal{U}} \sum_{u' \in \mathcal{U}} \sum_{z' \in \mathcal{Z}} \pi_W(u', z') q_Z(z|z') f(u|z) \\
&= \sum_{z' \in \mathcal{Z}} \sum_{u' \in \mathcal{U}} \sum_{u \in \mathcal{U}} \pi_W(u', z') q_Z(z|z') f(u|z) \\
&= \sum_{z' \in \mathcal{Z}} q_Z(z|z') \sum_{u' \in \mathcal{U}} \pi_W(u', z') \sum_{u \in \mathcal{U}} f(u|z) \\
&= \sum_{z' \in \mathcal{Z}} q_Z(z|z') \sum_{u' \in \mathcal{U}} \pi_W(u', z') \\
&= \sum_{z' \in \mathcal{Z}} q_Z(z|z') \pi_Z(z').
\end{aligned}$$

Seega protsess Z omab statsionaarset jaotust ning on järelikult positiivselt korduv.

Piisavus. Olgu protsess Z positiivselt korduv. Siis leidub sellel statsionaarne jaotus π_Z. Piisab näidata, et tõenäosused

$$\pi_W(u, z) = f(u|z)\pi_Z(z)$$

moodustavad protsessi W statsionaarse jaotuse. Näeme, et

$$\begin{aligned}
\pi_W(u, z) &= f(u|z)\pi_Z(z) \\
&= f(u|z) \sum_{z' \in \mathcal{Z}} q_Z(z|z')\pi_Z(z')
\end{aligned}$$

$$
\begin{aligned}
&= f(u|z) \sum_{z' \in \mathcal{Z}} q_Z(z|z') \pi_Z(z') \sum_{u' \in \mathcal{U}} f(u'|z') \\
&= \sum_{(u',z') \in \mathcal{W}} f(u|z) q_Z(z|z') \pi_Z(z') f(u'|z') \\
&= \sum_{(u',z') \in \mathcal{W}} q_W(u, z|u', z') \pi_W(u', z').
\end{aligned}
$$

Seega protsess W omab statsionaarset jaotust ning on järelikult positiivselt korduv.

(iv) Alajaotuses (iii) nägime, et Markovi ahelal W leidub statsionaarne jaotus parajasti siis, kui Markovi ahelal Z leidub statsionaarne jaotus, kusjuures tõenäosused $\pi_W(u, z)$ on määratud võrdusega (8). □

Eelnevalt nägime, et varjatud Markovi mudeli W režiim Z on Markovi ahel. Emissioonide jada U kohta see üldjuhul ei kehti.

Näide 2.1. Olgu varjatud Markovi mudel W määratud ruumil

$$\mathcal{W} = \{(0,0,0), (1,1,0), (0,0,1), (1,1,1)\},$$

kusjuures Markovi ahela Z üleminekumaatriks on

$$P = \begin{pmatrix} 0{,}9 & 0{,}1 \\ 0{,}3 & 0{,}7 \end{pmatrix},$$

režiimi algjaotus on statsionaarne jaotus

$$(\pi_Z(0), \pi_Z(1)) = (0{,}75, 0{,}25)$$

ja emissioonide tinglikud esinemistõenäosused on

$$f((0,0)|0) = f((1,1)|1) = 0{,}9; \qquad f((1,1)|0) = f((0,0)|1) = 0{,}1.$$

Sellisel juhul

$$\begin{aligned}
&P(U_3 = (0,0)|U_2 = (0,0), U_1 = (0,0)) \\
&= \frac{P(U_3 = (0,0), U_2 = (0,0), U_1 = (0,0))}{P(U_2 = (0,0), U_1 = (0,0))} \\
&= \frac{\sum_{z_{1:3} \in \mathcal{Z}^3} \pi_Z(z_1) q_Z(z_2|z_1) q_Z(z_3|z_2) f((0,0)|z_1) f((0,0)|z_2) f((0,0)|z_3)}{\sum_{z_{1:2} \in \mathcal{Z}^2} \pi_Z(z_1) q_Z(z_2|z_1) f((0,0)|z_1) f((0,0)|z_2)} \\
&= \frac{0{,}45676}{0{,}562} \approx 0.81274,
\end{aligned}$$

samas

$$\begin{aligned}
&P(U_3 = (0,0)|U_2 = (0,0)) = \frac{P(U_3 = (0,0), U_2 = (0,0))}{P(U_2 = (0,0))} \\
&= \frac{\sum_{z_2, z_3 \in \mathcal{Z}} \pi_Z(z_2) q_Z(z_3|z_2) f((0,0)|z_2) f((0,0)|z_3)}{\sum_{z_2 \in \mathcal{Z}} \pi_Z(z_2) f((0,0)|z_2)} \\
&= \frac{0{,}562}{0{,}7} \approx 0{,}80286.
\end{aligned}$$

Seega jada U ei täida Markovi tingimust, sest

$$P(U_3 = (0,0)|U_2 = (0,0), U_1 = (0,0)) \neq P(U_3 = (0,0)|U_2 = (0,0)).$$

2.2 Sõltuvuse saartega mudel

Vaatame käesolevas ning järgmises alapeatükis varjatud Markovi mudelit W, kus $\mathcal{W} = \{0,1\} \times \{0,1\} \times \{0,1\}$ ja Markovi ahela Z üleminekumaatriks on

$$P = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & \quad 1 \end{array} \\ \begin{pmatrix} p & 1-p \\ 1-q & q \end{pmatrix} \end{array}, \tag{9}$$

kus $p, q \in (0,1)$. Lause (1.2) põhjal on Z positiivselt korduv ning lause (1.3) järgi saame anda sellele statsionaarse algjaotuse $\pi_Z = (\pi_Z(0), \pi_Z(1))$, mis on leitav valemitega

$$\pi_Z(0) = \frac{1-q}{2-q-p}, \qquad \pi_Z(1) = \frac{1-p}{2-q-p}$$

Emissioonijaotuse $f(u|z)$ moodustame selliselt, et X_k ja Y_k on Z_k-st sõltumatud $Be(0{,}5)$ jaotusega juhuslikud suurused, mis $Z_K = 0$ korral on teineteisest sõltumatud ning $Z_k = 1$ korral positiivselt korreleeritud.

$$f(u|0) = \frac{1}{4} \quad \text{iga } u \in \mathcal{U} \text{ korral,}$$

$$f(u|1) = \begin{cases} \frac{1}{2} - \alpha, & \text{kui } z \in \{(0,0),(1,1)\}, \\ \alpha, & \text{kui } z \in \{(0,1),(1,0)\}, \end{cases} \qquad \alpha \in \left[0, \frac{1}{4}\right).$$

Emissioonide X_k ja Y_k kovariatsioon on statsionaarse jaotuse korral

$$\begin{aligned} \operatorname{cov}(X_k, Y_k) &= E(X_k Y_k) - E(X_k)E(Y_k) \\ &= E[E(X_k Y_k | Z_k)] - E(X_k)E(Y_k) \\ &= \pi_Z(1)\left(\frac{1}{2} - \alpha\right) + \frac{1}{4}\pi_Z(0) - \frac{1}{4} = \pi_Z(1)\left(\frac{1}{2} - \alpha\right) - \frac{1-\pi_Z(0)}{4}. \end{aligned}$$

Arvestades, et $DX_k = DY_k = 0{,}25$, on X_k ja Y_k vaheline Pearsoni korrelatsioonikordaja

$$\rho_{X_kY_k} = \frac{\mathrm{cov(X_k Y_k)}}{\sqrt{DX_k DY_k}} = 4\mathrm{cov(X_k Y_k)}$$

Üheks huviks sellisest mudelist saadud vaatluste $X_{1:n}$ ja $Y_{1:n}$ tõlgendamisel on varjatud jada $Z_{1:n}$ järjestikuste 0-ide ja 1-de saarte pikkuste tuvastamine.

Definitsioon 2.2. Olgu Z jada, mis võtab väärtusi ülimalt loenduval hulgal $\mathcal{Z}$, olgu $z \in \mathcal{Z}$. Jada Z *z-ide saareks* nimetame osajada $(Z_k, Z_{k+1}, ..., Z_{k+l})$, kus $k, l \in \mathbb{N}$ ja

$$Z_i = z \text{ iga } k \leqslant i \leqslant l \text{ korral,}$$

kusjuures

$$Z_{k-1} \neq z \text{ või } k = 1$$

ning

$$Z_{k+l+1} \neq z.$$

Tuletame järjestikuste 0-ide ja 1-de saarte pikkuste jaotuse.

Väide 2.1. *Olgu Z Markovi ahel üleminekumaatriksiga (9). Siis*

(i) 0-ide saarte pikkused on juhuslikud suurused jaotusega $Geom(1-p)$.

(ii) 1-de saarte pikkused on juhuslikud suurused jaotusega $Geom(1-q)$.

Tõestus. Tõestame ainult väite (i), teise väite tõestus on analoogiline. Olgu T_1^0 esimese nullide saare esimene indeks, ehk

$$T_1^0 := \min\{t \in \mathbb{N} | Z_t = 0\}$$

Olgu $k \geqslant 2$, siis k-nda nullide saare esimese indeksi T_k^0 saame defineerida induktiivselt

$$T_k^0 := \min\{t > T_{k-1}^0 | Z_{t-1} = 1, Z_t = 0\}.$$

Iga $k \in \mathbb{N}$ korral on k-nda nullide saare pikkus defineeritud järgnevalt:

$$N_k^0 = t \Leftrightarrow Z_{T_k^0} = 0, Z_{T_k^0+1} = 0, ..., Z_{T_k^0+t-1} = 0, Z_{T_k^0+t} = 1.$$

Veendume, et mis tahes $t \in \mathbb{N}$ korral $P(N_k^0 = t) = p^{t-1}(1-p)$. Tõepoolest,

$$\begin{aligned}
&P(N_k^0 = t) = P(Z_{T_k^0} = Z_{T_k^0+1} = ... = Z_{T_k^0+t-1} = 0, Z_{T_k^0+t} = 1) \\
&= \sum_{m\in\mathbb{N}} P(Z_{T_k^0} = Z_{T_k^0+1} = ... = Z_{T_k^0+t-1} = 0, Z_{T_k^0+t} = 1 | T_k^0 = m) P(T_k^0 = m) \\
&= \sum_{m\in\mathbb{N}} P(Z_m = Z_{m+1} = ... = Z_{m+t-1} = 0, Z_{m+t} = 1 | T_k^0 = m) P(T_k^0 = m) \\
&= \sum_{m\in\mathbb{N}} P(Z_m = Z_{m+1} = ... = Z_{m+t-1} = 0, Z_{m+t} = 1 | T_k^0 = m, Z_m = 0)... \\
&\qquad\qquad ...P(Z_m = 0 | T_k^0 = m) P(T_k^0 = m),
\end{aligned}$$

kus teine võrdus tuleneb täistõenäosuse valemist ja neljas võrdus asjaolust, et sündmus $\{Z_m = 0\}$ järeldub sündmusest $\{T_k^0 = m\}$. Arvestades, et sündmuse $\{T_k^0 = m\}$ toimunine on üheselt määratud juhuslike suuruste $Z_1, Z_2, ..., Z_m$ poolt, saame Markovi omaduse ja tõenäosuste korrutamise lause põhjal

$$\begin{aligned}
&\sum_{m\in\mathbb{N}} P(Z_m = Z_{m+1} = ... = Z_{m+t-1} = 0, Z_{m+t} = 1 | T_k^0 = m, Z_m = 0)... \\
&\qquad\qquad ...P(Z_m = 0 | T_k^0 = m) P(T_k^0 = m) \\
&= \sum_{m\in\mathbb{N}} P(Z_{m+1} = Z_{m+2} = ... = Z_{m+t-1} = 0, Z_{m+t} = 1 | Z_m = 0) P(T_k^0 = m)
\end{aligned}$$

$$= \sum_{m\in\mathbb{N}} P(Z_{m+1}=0|Z_m=0)P(Z_{m+2}=0|Z_{m+1}=0)...$$

$$...P(Z_{m+t-1}=0|Z_{m+t-2}=0)P(Z_{m+t}=1|Z_{m+t-2}=0)P(T_k^0=m)$$

$$= p^{t-1}(1-p)\sum_{m\in\mathbb{N}} P(T_k^0=m) = p^{t-1}(1-p).$$

Samamoodi saab veenduda, et $P(Z_n = 1, Z_{n+1} = 1, Z_{n+2} = 1, ...) = 0$. Järelikult $P(T_k^0 < \infty) = 1$ ja seega $P(N_k^0 = t) = p^{t-1}(1-p)$. □

Simuleerime üht konkreetset näidet sõltuvuse saartega mudelist.

Näide 2.2. Vaatame VMM-i, kus $p = 0{,}9$, $q = 0{,}8$ ja $\alpha = 0{,}025$. Siis režiimi üleminekumaatriks on

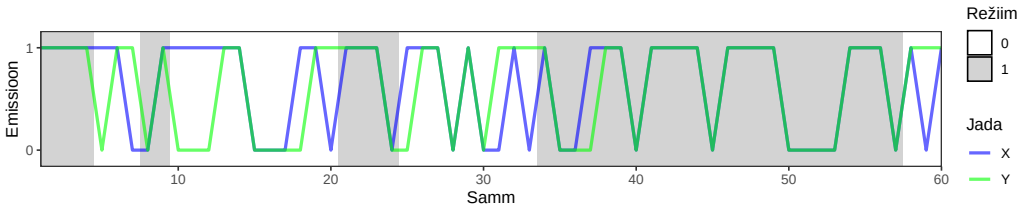
$$P = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 0{,}9 & 0{,}1 \\ 0{,}2 & 0{,}8 \end{pmatrix} \end{array}$$

ja emissioonide tõenäosused on

$$f(u|0) = \frac{1}{4} \quad \text{iga } z \in \mathcal{U} \text{ korral,}$$

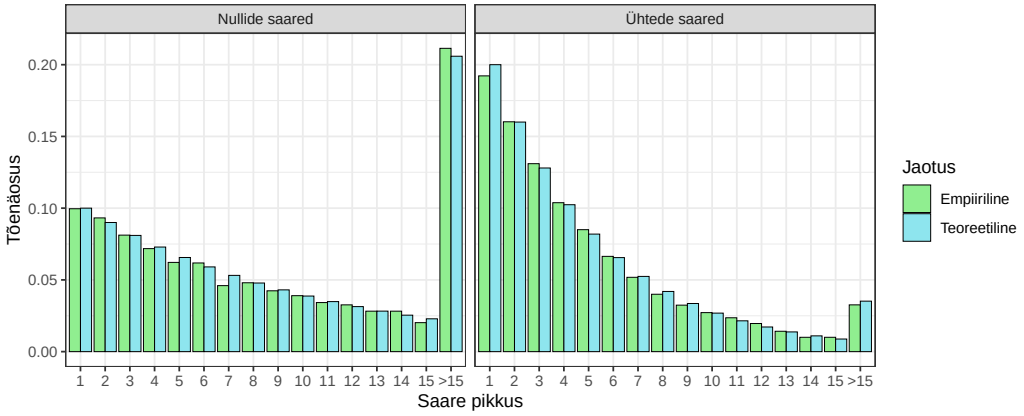
$$f(u|1) = \begin{cases} 0{,}475, & \text{kui } u \in \{(0,0),(1,1)\}, \\ 0{,}025, & \text{kui } u \in \{(0,1),(1,0)\}. \end{cases}$$

Ülal kirjeldatud Markovi mudelit simuleeriti, kuni režiimi olek vahetus 10 000. korda. Joonisel (1) on kujutatud jadade $X_{1:60}$ ja $Y_{1:60}$ graafikuid, kus valge taust tähistab režiimi 0 ja hall taust režiimi 1. Joonisel näeme, et $Z_k = 0$ korral võtavad X_k ja Y_k sama väärtuse ligikaudu pooltel juhtudel ning $Z_k = 1$ korral on enamik X_k ja Y_k väärtustest võrdsed. Näite koostamiseks kasutati programmi VMM_naide.R (Lisa 1).



Joonis 1: Sõltuvuse ja sõltumatuse saartega VMM-i esimesed 60 olekut.

Simulatsiooni jätkamisel, kuni režiim oli muutunud 10 000 korda, ehk järjestikuste 0-ide ja järjestikuste 1-de saari oli mõlemaid tekkinud 5 000 tükki. Saadudu saarte pikkuste empiirilise jaotuse võrdlemisel $Geom(0{,}1)$ ja $Geom(0{,}2)$ jaotustega näeme, et simulatsiooni tulemused on väitega (2.1) kooskõlas. (Joonis 2)



Joonis 2: Järjestikuste 1-de ja 0-ide saarte empiirilised ja teoreetilised jaotused.

2.3 Simulatsioonid

Eelnevas alapeatükis kirjeldatud VMM W on lause (2.2) järgi mittelahutuv ja positiivselt korduv ning sellele saab anda statsionaarse algjaotuse π_W. Seega täidab mudel W teoreemi (1.1) eeldusi ning L_n/n koondub peaaegu kindlasti

mingiks konstandiks γ.

Hindame simulatsioonide abil, kuidas sõltub γ parameetrite p ja q väärtustest, kui $\alpha = 0{,}025$. Käsitleme nelja mudelit, mille parameetrid p ja q, režiimi statsionaarne jaotus π_Z, emissioonide kovariatsioon $\mathrm{cov}(X, Y)$ ja korrelatsioon ρ_{XY} ning 0-ide ja 1-de saarte pikkuste N_k^0 ja N_k^1 keskväärtused on toodud tabelis (1).

Tabel 1: Ülevaade simuleeritavatest sõltuvuse saartega mudelitest.

Mudel	p	q	π_Z	$\mathrm{cov}(X,Y)$	ρ_{XY}	$E(N_k^0)$	$E(N_k^1)$
1	0,8	0,8	$\left(\frac{1}{2}, \frac{1}{2}\right)$	0,1125	0,45	5	5
2	0,9	0,9	$\left(\frac{1}{2}, \frac{1}{2}\right)$	0,1125	0,45	10	10
3	0,9	0,8	$\left(\frac{2}{3}, \frac{1}{3}\right)$	0,075	0,3	10	5
4	0,95	0,9	$\left(\frac{2}{3}, \frac{1}{3}\right)$	0,075	0,3	20	10

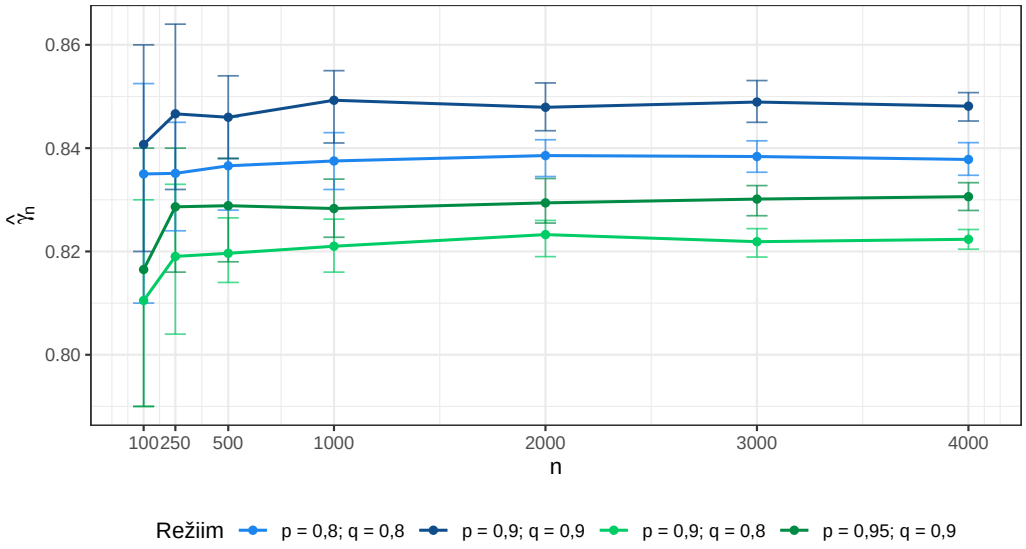
Mudelid 1 ja 2 viibivad režiimis 1 kauem, kui mudelid 3 ja 4, seega peaks ka γ väärtus olema esimese kahe mudeli korral suurem. Lisaks saab mudeleid 1 ja 2 ning mudeleid 3 ja 4 omavahel võrreldes näha, kas sama α ja π_Z korral ning järelikult ka sama π_W korral mõjutavad erinevad režiimi 0-ide ja 1-de saarte pikkused γ väärtust. Simulatsioonide läbiviimisel kasutati programme VMM_simulatsioonid.R ja NW2.cpp (Lisa).

Simulatsioonid viidi läbi järgnevalt:

- Iga mudeli ja $n \in \{100, 250, 500, 1000, 2000, 3000, 4000\}$ korral genereeriti valim 100-st jadade $X_{1:n}, Y_{1:n}$ paarist.
- Needleman-Wunschi algoritmiga leiti iga genereeritud jadade paari korral L_n väärtus.

- Iga mudeli ja $n \in \{100, 250, 500, 1000, 2000, 3000, 4000\}$ korral salvestati andmestikku valimikeskmine $\hat{\gamma}_n$, standardhälve $\hat{\sigma}$, alumine kvartiil $\hat{q}_{0,25}$ ja ülemine kvartiil $\hat{q}_{0,75}$.

Simulatsioonide tulemusel saadud valimikeskmised ja kvartiilide vahemikud on toodud graafikul (3). Näeme, et kõigi mudelite korral L_n/n koondub ning $n = 1000$-st edasi ei esine märgatavaid valimikeskmiste kõikumisi. Näeme, et mudelitel 1 ja 2 on $\hat{\gamma}_n$ väärtused suuremad, kui mudelitel 3 ja 4. Lisaks on sama statsionaarse jaotusega mudelite korral $\hat{\gamma}_n$ väärtus suurem, kui 0-ide ja 1-de saared on pikemad.



Joonis 3: $\hat{\gamma}_n$ väärtused ja valimi kvartiilid.

Tabelis (2) on toodud $\hat{\gamma}_n$ ja $\hat{\sigma}_n$ väärtused $n = 4000$ korral. Tabelisse on lisatud ka ligikaudne γ väärtus *iid*-mudeli korral. Näeme, et sõltuvuse saarte lisamine suurendab $\hat{\gamma}_n$ väärtust.

Tabel 2: $\hat{\gamma}_n$ väärtused ja valimi standardvead $n = 4000$ korral

Mudel	1	2	3	4	*iid*-mudel
$\hat{\gamma}_n$	0,838	0,848	0,822	0,831	0,812
$\hat{\sigma}_n$	$4{,}32 \cdot 10^{-3}$	$4{,}90 \cdot 10^{-3}$	$3{,}29 \cdot 10^{-3}$	$4{,}60 \cdot 10^{-3}$	

3 Juhusliku ekslemise tüüpi mudelid

Selles peatükis vaatame varjatud Markovi mudeleid, mille režiim Z on juhuslik ekslemine täisarvude hulgal.

3.1 Lihtne juhusliku ekslemise tüüpi mudel

Käsitleme käesolevas alapeatükis lihtsat juhusliku ekslemise tüüpi mudelit, kus režiim on määratud juhuslike suurustega

$$Z_1 = S_1, \quad Z_k = Z_{k-1} + S_k, \quad k = 2, 3, ..., \tag{10}$$

kus $S_1, S_2, ...$ on sõltumatud juhuslikud suurused jaotusega

$$P(S_k = 1) = p, \quad P(S_k = -1) = 1 - p, \quad p \in (0, 1), \tag{11}$$

ja emissioonid on määratud võrdusega

$$U_k = \begin{cases} (1,1) & \text{kui } Z_k > 0, \\ (2,3) & \text{vastasel korral,} \end{cases} \qquad k = 1, 2, ... \tag{12}$$

Tähistame

$$M_n := |\{1 \leqslant k \leqslant n : Z_k > 0\}|, \tag{13}$$

ehk M_n on juhusliku ekslemise Z positiivsetes seisundites oldud sammude arv esimese n sammu jooksul. Veendume, et antud mudeli korral $M_n/n = L_n/n$

Lemma 3.1. *Olgu Markovi ahela W üleminekud määratud seostega (10),*

(11) ja (12). Siis L_n/n on proportsionaalne aeg, mil $Z_k > 0$, ehk

$$\frac{L_n}{n} = \frac{M_n}{n}, \tag{14}$$

Tõestus. Võrduse (12) järgi on M_n jadades $X_{1:n}$ ja $Y_{1:n}$ esinevate 1-de arv. Kui $m = 0$, siis $X_i = 2$ ja $Y_i = 3$ iga $i = 1, 2, ..., n$ korral, seega $L_n = M_n = 0$. Kui $1 \leqslant M_n \leqslant n$, siis leiduvad indeksid $1 \leqslant i_1 < i_2 < ... < i_{M_n} \leqslant n$, mille korral $X_{i_k} = Y_{i_k} = 1$ $(k = 1, 2, ..., M_n)$. Seega jadadel $X_{1:n}$ ja $Y_{1:n}$ leidub ühisjada pikkusega M_n. Arvestades, et $\mathcal{X} = \{1, 2\}$ ja $\mathcal{Y} = \{1, 3\}$, saavad jadade $X_{1:n}$ ja $Y_{1:n}$ ühisjadad koosneda vaid 1-dest, mistõttu $L_n \leqslant M_n$. Järelikult $L_n = M_n$ ja võrdus (14) kehtib. □

On teada, et Z on sümmeetrilise juhusliku ekslemise $(p = \frac{1}{2})$ korral null-korduv ja mittesümmeetrilise juhusliku ekslemise $(p \neq \frac{1}{2})$ korral mööduv (Aldridge, 2021, Peatükk 8). Uurime esmalt piirväärtust $\lim_{n\to\infty}(L_n/n)$ mittesümmeetrilise juhusliku ekslemise korral.

Väide 3.1. *Olgu Markovi ahela W üleminekud määratud seostega (10), (11) ja (12). Kehtivad järgmised väited:*

(a) kui $p > \frac{1}{2}$, siis $(L_n/n) \to 1$ p.k;

(b) kui $p < \frac{1}{2}$, siis $(L_n/n) \to 0$ p.k.

Tõestus. Juhusliku ekslemise Z sammude S_i $(i \in \mathbb{N})$ keskväärtus on

$$E(S_i) = 1 \cdot p - 1 \cdot (1 - p) = 2p - 1.$$

Tugeva suurte arvude seaduse järgi leiab aset koondumine

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} S_i = 2p - 1 \quad p.k.$$

Arvestades, et

$$Z_n = \sum_{i=1}^{n} S_i,$$

saame peaaegu kindlasti koondumise definitsiooni järgi

$$P\left(\lim_{n\to\infty} \frac{1}{n} Z_n = 2p - 1\right) = 1$$

Olgu M_n defineeritud võrdusega (13) ning olgu M'_n juhusliku ekslemise Z mittepositiivsetes seisundites oldud sammude arvu esimese n sammu jooksul, ehk

$$M_n := |\{1 \leqslant k \leqslant n | Z_k > 0\}|, \quad M'_n := |\{1 \leqslant k \leqslant n | Z_k \leqslant 0\}|$$

On selge, et $M_n = n - M'_n$. Seega lemma (3.1) järgi kehtivad võrdused

$$\frac{L_n}{n} = \frac{M_n}{n} = \frac{n - M'_n}{n}. \tag{15}$$

(a) Oletame, et $p > \frac{1}{2}$. Olgu T varaseim aeg nii, et $Z_k > 0$ iga $k \geqslant T$ korral, ehk

$$T = \begin{cases} \min\{N \in \mathbb{N} : n \geqslant N \Rightarrow Z_n > 0\}, & \text{kui } \{N \in \mathbb{N} : n \geqslant N \Rightarrow Z_n > 0\} \neq \emptyset \\ \infty & \text{muidu.} \end{cases}$$

Veendume, et $T < \infty$ (ehk $T \in \mathbb{N}$) tõenäosusega 1. Tõepoolest,

$$\begin{aligned} 1 &\geqslant P(T < \infty) = P(\exists n \in \mathbb{N} : k \geqslant n \Rightarrow Z_k > 0) \\ &= P\left(\exists n \in \mathbb{N} : k \geqslant n \Rightarrow \frac{Z_k}{k} > 0\right) \\ &\geqslant P\left(\exists n \in \mathbb{N} : k \geqslant n \Rightarrow 2(2p-1) > \frac{Z_k}{k} > 0\right) \\ &= P\left(\exists n \in \mathbb{N} : k \geqslant n \Rightarrow \left|\frac{Z_k}{k} - (2p-1)\right| < 2p - 1\right) \end{aligned} \tag{16}$$

$$\geqslant P\left(\forall \varepsilon > 0\ \exists n \in \mathbb{N} : k \geqslant n \Rightarrow \left|\frac{Z_k}{k} - (2p-1)\right| < \varepsilon\right) \tag{17}$$
$$= P\left(\lim_{n\to\infty} \frac{1}{n} Z_n = 2p - 1\right) = 1.$$

Seejuures võrratused (16) ja (17) kehtivad tõenäosusmõõdu monotoonsuse tõttu. Kui $T < \infty$, siis $0 \leqslant M'_n \leqslant T-1$, seega piirväärtuse keskmise muutuja omaduse ning seose (15) teise võrduse järgi

$$\lim_{n\to\infty} \frac{L_n}{n} = \lim_{n\to\infty} \frac{n - M'_n}{n} = \lim_{n\to\infty} \frac{n}{n} = \lim_{n\to\infty} \frac{n - T + 1}{n} = 1 \quad p.k.$$

(b) Oletame, et $p < \frac{1}{2}$. Olgu T varaseim aeg nii, et $Z_k \leqslant 0$ iga $k \geqslant T$ korral, ehk

$$T = \begin{cases} \min\{N \in \mathbb{N} : n \geqslant N \Rightarrow Z_n \leqslant 0\}, & \text{kui } \{N \in \mathbb{N} : n \geqslant N \Rightarrow Z_n \leqslant 0\} \neq \emptyset \\ \infty & \text{muidu.} \end{cases}$$

Veendume, et $T < \infty$ (ehk $T \in \mathbb{N}$) tõenäosusega 1. Tõepoolest,

$$1 \geqslant P(T < \infty) = P(\exists n \in \mathbb{N} : k \geqslant n \Rightarrow Z_k \leqslant 0)$$
$$= P\left(\lim_{n\to\infty} \frac{1}{n} Z_n = 2p - 1\right) = 1.$$

Kui $T < \infty$, siis $0 \leqslant M_n \leqslant T - 1$, seega piirväärtuse keskmise muutuja omaduse ning seose (15) esimese võrduse järgi

$$\lim_{n\to\infty} \frac{L_n}{n} = \lim_{n\to\infty} \frac{M_n}{n} = \lim_{n\to\infty} \frac{0}{n} = \lim_{n\to\infty} \frac{T-1}{n} = 0 \quad p.k.$$

□

Eelnevast tulemusest järeldub, et konstant γ võib leiduda ka mudelil, mis

teoreemi (1.1) kõiki eelduseid ei täida.

Naiivselt võiks arvata, et konstant γ leidub ka sümmeetrilise juhusliku ekslemise ehk $p = \frac{1}{2}$ korral. Kuna sümmeetriline juhuslik ekslemine on null-korduv, ei ole teoreemi (1.1) eeldused täidetud ning selgub, et Z koondub L_n/n jaotuse järgi arkussiinusjaotusega juhuslikuks suuruseks, mitte konstandiks.

Definitsioon 3.1. Juhuslik suurus X on *arkussiinusjaotusega*, ehk

$X \sim Arcsin(0,1)$, kui tema jaotusfunktsioon avaldub seosega

$$F(x) = \begin{cases} 0, & \text{kui } x < 0, \\ \frac{2}{\pi} \arcsin \sqrt{x}, & \text{kui } x \in [0,1], \\ 1, & \text{kui } x > 1, \end{cases}$$

ja tihedusfunktsioon avaldub seosega

$$f(x) = \begin{cases} \frac{1}{\pi\sqrt{x(1-x)}}, & \text{kui } x \in (0,1), \\ 0, & \text{mujal.} \end{cases}$$

Väide 3.2. *Kui $p = \frac{1}{2}$, leiab aset koondumine*

$$\frac{L_n}{n} \xrightarrow{d} Arcsin(0,1). \tag{18}$$

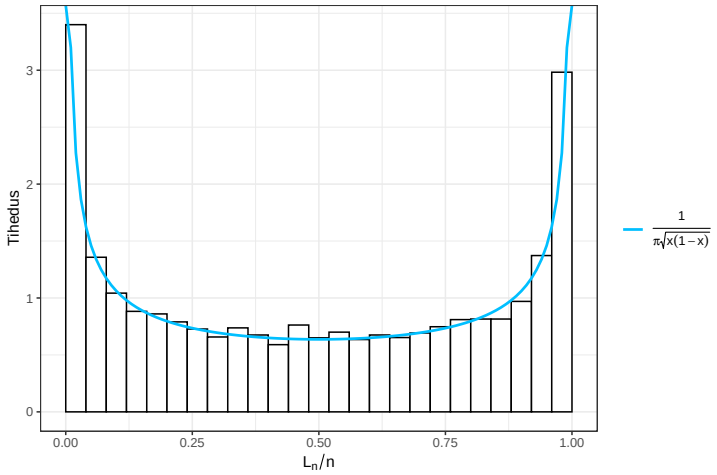
Tõestus. Tekstis (Ackelsberg, 2018) tõestatakse, et

$$\frac{M_n}{n} \xrightarrow{d} Arcsin(0,1),$$

kus suurus M_n Lemma (3.1) järgi on see samaväärne väitega (18). □

Kontrollime seda tulemust ka simulatsioonide abil. Programmiga Arcsin_juhuslik_ekslemine.R

(Lisa 1) Genereeriti 10 000 realisatsiooni juhusliku ekslemise Z esimesest $n = 5000$ olekust. Iga jada korral võeti vastavalt lemmale (3.1) L_n/n väärtuseks positiivsetes seisundite osakaal M_n/n. Saadud L_n/n väärtuste empiiriline jaotus sarnaneb arkussiinusjaotusega. (Joonis 4)



Joonis 4: L_n/n väärtuste histogramm koos arkussiinusjaotuse tihedusfunktsiooniga.

3.2 Positiivselt korduv juhusliku ekslemise tüüpi mudel

Eelmises alapeatükis kirjeldatud mudel ei täitnud teoreemi (1.1) eeldust, et W on positiivselt korduv, sest ühegi $p \in (0, 1)$ korral ei olnud Z positiivselt korduv.

Käesolevas alapeatükis koostame positiivselt korduva juhusliku ekslemise Z, mis ei triiviks jäädavalt nullist eemale. Jätame esialgu seisundi Z_1 jaotuse ning emissioonide U_k tinglikud jaotused määramata. Üleminekud määrame

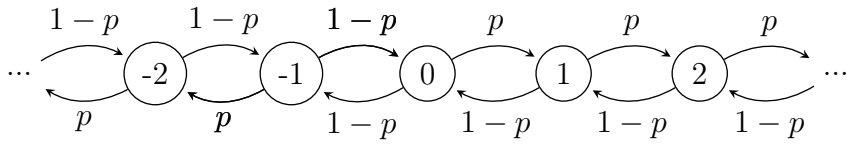
võrdustega

$$Z_k = \begin{cases} Z_{k-1} + S_k, & \text{kui } Z_{k-1} \geq 0 \\ Z_{k-1} - S_k, & \text{muidu,} \end{cases} \qquad k = 2, 3, ..., \tag{19}$$

kus $S_2, S_3, ...$ on sõltumatud juhuslikud suurused jaotusega

$$P(S_k = 1) = p, \quad P(S_k = -1) = 1 - p, \quad p \in (0, 1), \quad k = 2, 3, ... \tag{20}$$

Antud juhusliku ekslemise üleminekud on kujutatud joonisel (5).



Joonis 5: Juhusliku ekslemise (19) üleminekute diagramm

Järgnevalt tõestame, et $p \in \left(0, \frac{1}{2}\right)$ korral on juhuslik ekslemine Z positiivselt korduv. Kuna Z on mittelahutuv, piisab tõestada, et seisund 0 on positiivselt korduv. Kuna seisundi positiivselt korduvus on defineeritud keskmise naasmisaja kaudu, huvitab meid erinevate juhusliku ekslemise Z võimalike teekondade arv, mis algavad seisundist 0 ning naasevad $2n$ sammu pärast esimest korda seisundisse 0. Sellised teekonnad rahuldavad üht kahest järgnevast tingimusest

1. $z_2 = z_{2n} = 1$, ja $z_k \geq 1$ iga $k \in \{2, 3, ..., 2n\}$ korral;

2. $z_2 = z_{2n} = -1$, ja $z_k \leq -1$ iga $k \in \{2, 3, ..., 2n\}$ korral.

Iga esimest tingimust rahuldava teekonna tõenäosus on $p^n(1-p)^n$ ning iga

teist tingimust rahuldava teekonna tõenäosus on $p^{n-1}(1-p)^{n+1}$. Tähistame

$$N_n^{\neq 0}(a) = \left| \left\{ z_{1:(n+1)} \in \mathbb{Z}^{n+1} : \begin{array}{r} P(Z_{1:(n+1)} = z_{1:(n+1)} | Z_1 = z_1) > 0 \\ Z_1 = a,\ Z_{n+1} = a \\ Z_i \neq 0\ \forall i \in \{2, 3, ..., n\} \end{array} \right\} \right|$$

Ehk $N_n^{\neq 0}(a)$ on erinevate võimalike n sammu pikkuste juhusliku ekslemise Z seisundis a algavate ja lõppevate ning vahepeal seisundit 0 mitte külastavate teekondade arv. Seisundist 0 algavate ning esimest korda $2n$ sammu pärast seisundisse 0 naasvate teekondade arvu annab järgnev abitulemus.

Lemma 3.2. *Kehtivad järgmised seosed*

$$N_{2n}^{\neq 0}(0) = \frac{1}{2n-1}\binom{2n}{n} \tag{21}$$

$$N_{2n}^{\neq 0}(0) = 2 \cdot N_{2(n-1)}^{\neq 0}(1) \tag{22}$$

$$N_{2(n-1)}^{\neq 0}(-1) = N_{2(n-1)}^{\neq 0}(1) = \frac{1}{n}\binom{2(n-1)}{n-1} \tag{23}$$

Tõestus. Võrdused (21), (22) ja seose (23) teine võrdus on tõestatud tekstis (Alm, 2006, Alapeatükk 4.1). Seose (23) esimene võrdus tuleneb sümmeetriast. □

Kasutame edaspidi teadmist, et arvjada $\{N_{2(n-1)}^{\neq 0}(1)\}_{n=1}^{\infty}$ puhul on tegemist Catalani arvudega.

Definitsioon 3.2. (Guichard, 2025, Alapeatükk 3.5) *Catalani arvudeks* nimetatakse arve

$$C_n = \frac{1}{n+1}\binom{2n}{n}, \quad n \in \mathbb{N} \cup \{0\}$$

On selge, et $N_{2(n-1)}^{\neq 0}(1) = N_{2(n-1)}^{\neq 0}(-1) = C_{n-1}$ ning $N_{2n}^{\neq 0}(0) = 2C_{n-1}$. Seega saame esimest korda esimest korda $2n$ sammu järel seisundisse 0 naasmise

tõenäosuseks

$$P(T = 2n|Z_1 = 0) = C_{n-1}(p^n(1-p)^n + p^{n-1}(1-p)^{n+1}) = C_{n-1}p^{n-1}(1-p)^n,$$

kus T on aeg, mil Z naaseb esimest korda seisundisse 0, ehk

$$T = \min\{k \geq 1 : Z_{k+1} = 0\}.$$

Järgnevas tõestuses kasutame genereeriva funktsiooni mõistet

Definitsioon 3.3. Arvjada $\{a_n\}_{n=0}^{\infty}$ *genereerivaks funktsiooniks* nimetatakse astmerida

$$a(x) = \sum_{n=0}^{\infty} a_n x^n.$$

Võtame teadmiseks, et funktsioon

$$\frac{1 - \sqrt{1-4x}}{2x} = \sum_{n=0}^{\infty} C_n x^n$$

on Catalani arvude jada $\{C_n\}_{n=0}^{\infty}$ genereeriv funktsioon ning funktsioon

$$\frac{1}{\sqrt{1-4x}} = \sum_{n=0}^{\infty} \binom{2n}{n} x^n$$

on arvjada $\left\{\binom{2n}{n}\right\}_{n=0}^{\infty}$ genereeriv funktsioon (Wilf, 1994, Alapeatükk 2.5).

Kui $p = \frac{1}{2}$, on tegemist sümmeetrilise juhusliku ekslemisega, mida käsitlesime eelmises alapeatükis. Järgmises tõestuses vaatame vaid olukordi, kus $p \neq \frac{1}{2}$.

Väide 3.3. *Olgu juhusliku ekslemise Z üleminekud määratud seostega (19) ja (20). Kehtivad järgmised väited:*

1. *Kui $p > \frac{1}{2}$, siis Z on mööduv.*

2. Kui $p < \frac{1}{2}$, siis Z on positiivselt korduv.

Tõestus. Solidaarsusteoreemi põhjal piisab veenduda, et seisund 0 on mööduv, kui $p > \frac{1}{2}$ ja positiivselt korduv, kui $p < \frac{1}{2}$. Kuna Z on homogeenne, eeldame üldisust kitsendamata, et $Z_1 = 0$. Näitame esmalt, et Z on mööduv, kui $p > \frac{1}{2}$, ja korduv, kui $p < \frac{1}{2}$. Seisundi korduvuse definitsiooni järgi peame näitama, et

$$\sum_{n=1}^{\infty} P(T = n|Z_1 = 0) < 1, \text{ kui } p > \frac{1}{2},$$
$$\sum_{n=1}^{\infty} P(T = n|Z_1 = 0) = 1, \text{ kui } p < \frac{1}{2}.$$

Arvestades, et juhuslik ekslemine ei saa paaritu arvu sammudega samasse seisundisse naasta, saame

$$\sum_{n=1}^{\infty} P(T = n|Z_1 = 0) = \sum_{n=1}^{\infty} P(T = 2n|Z_1 = 0) = \sum_{n=1}^{\infty} C_{n-1}p^{n-1}(1-p)^n$$
$$= (1-p)\sum_{n=0}^{\infty} C_n(p(1-p))^n.$$

Rakendame Catalani arvude jada genereerivat funktsiooni, kus $x = p(1-p)$, siis

$$\sum_{n=1}^{\infty} P(T = n|Z_1 = 0) = (1-p)\sum_{n=0}^{\infty} C_n(p(1-p))^n =$$
$$= \frac{(1-p)(1-\sqrt{1-4p(1-p)})}{2p(1-p)} = \frac{1-\sqrt{(2p-1)^2}}{2p} =$$
$$= \frac{1-|2p-1|}{2p} = \begin{cases} \frac{1-p}{p} < 1, & \text{kui } p > \frac{1}{2}; \\ \frac{p}{p} = 1, & \text{kui } p < \frac{1}{2}. \end{cases}$$

Esimene väide on tõestatud. Olgu $p < \frac{1}{2}$. Näitame, et Z on positiivselt korduv, ehk

$$E(T|Z_1 = 0) < \infty.$$

Näeme, et

$$\begin{aligned} E(T|Z_1 = 0) &= \sum_{n=1}^{\infty} 2nP(T = 2n|Z_1 = 0) = 2(1-p)\sum_{n=0}^{\infty}(n+1)C_n(p(1-p))^n \\ &= 2(1-p)\sum_{n=0}^{\infty}\binom{2n}{n}(p(1-p))^n \end{aligned}$$

Rakendame jada $\left\{\binom{2n}{n}\right\}_{n=0}^{\infty}$ genereerivat funktsiooni. Võttes $x = p(1-p)$, saame

$$\begin{aligned} E(T|Z_1 = 0) &= 2(1-p)\sum_{n=0}^{\infty}\binom{2n}{n}(p(1-p))^n \\ &= \frac{2(1-p)}{\sqrt{1-4p(1-p)}} = \frac{2(1-p)}{|2p-1|} = \frac{2-2p}{1-2p} = 1 + \frac{1}{1-2p} < \infty. \end{aligned}$$

Seega juhul $p < \frac{1}{2}$ on Z positiivselt korduv. □

Oletame edaspidi, et Z on varjatud Markovi mudeli W režiim ja $p < \frac{1}{2}$. Lause (1.3) järgi leidub juhuslikul ekslemisel Z statsionaarne jaotus, mille võtame algseisundi Z_1 jaotuseks. Sellisel juhul on režiimiga Z VMM W statsionaarne, mittelahutuv ja positiivselt korduv, mistõttu teoreemi (1.1) järgi koondub sellise mudeli korral L_n/n peaaegu kindlasti mingiks konstandiks γ.

Mudelil W võivad olla mis tahes emissioonijaotused $f(u|z)$, kuid konkreetsuse mõttes vaatame olukorda, kus $\mathcal{X} = \mathcal{Y} = \{0, 1\}$, $\mathcal{W} = \mathcal{X} \times \mathcal{Y} \times \mathbb{Z}$ ning

emissioonide tinglik jaotus on määratud järgnevalt:

$$f(u|z) = \begin{cases} \frac{1}{4} + \frac{\arctan\frac{2z+1}{4}}{2\pi}, & \text{kui } u \in \{(0,0),(1,1)\}, \\ \frac{1}{4} - \frac{\arctan\frac{2z+1}{4}}{2\pi}, & \text{kui } u \in \{(0,1),(1,0)\}. \end{cases} \tag{24}$$

Kuna $\arctan(x)$ on rangelt kasvav funktsioon piirväärtustega

$$\lim_{x\to\infty} \arctan(x) = \frac{\pi}{2}, \quad \text{ja} \quad \lim_{x\to-\infty} \arctan(x) = -\frac{\pi}{2},$$

saame emissioonitõenäosuste piirväärtusteks

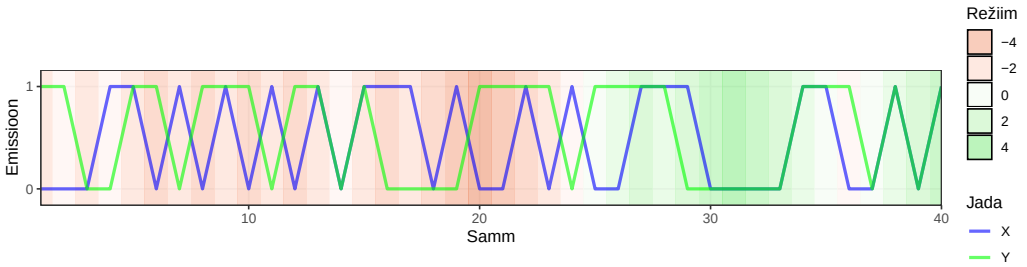
$$\begin{aligned} &\lim_{z\to\infty} f(u|z) = \frac{1}{2}, && \lim_{z\to-\infty} f(u|z) = 0, && \text{kui } u \in \{(0,0),(1,1)\}, \\ &\lim_{z\to\infty} f(u|z) = 0, && \lim_{z\to-\infty} f(u|z) = \frac{1}{2}, && \text{kui } u \in \{(0,1),(1,0)\}. \end{aligned}$$

Seega emissioonitõenäosusi (24) kasutades saame mudeli W, kus X_k ja Y_k on negatiivse režiimi korral negatiivselt korreleeritud ning mittenegatiivse režiimi korral positiivselt korreleeritud, seejuures korrelatsiooni tugevus sõltub Z_k absoluutväärtusest. Kuigi paaride (X_k, Y_k) ühisjaotused on Z_k-st sõltuvad, on X_k ja Y_k Z_k-st sõltumatud juhuslikud suurused jaotusega $Be(0{,}5)$. Tabelis (3) on esitatud emissioonide tõenäosused, X_k ja Y_k tinglikud kovariatsioonid ning Pearsoni korrelatsioonikordaja väärtused $-3 \leqslant z \leqslant 3$ korral.

Tabel 3: $f(u|z)$, $\mathrm{cov}(X, Y|Z = z)$ ja $\rho_{X,Y|Z=z}$ väärtused $-3 \leqslant z \leqslant 3$ korral.

z	-3	-2	-1	0	1	2	3
$f((0,0)\|z)$ $f((1,1)\|z)$	0,107	0,148	0,211	0,289	0,352	0,393	0,417
$f((0,1)\|z)$ $f((1,0)\|z)$	0,393	0,352	0,289	0,211	0,148	0,107	0,083
$\mathrm{cov}(X, Y\|Z = z)$	$-0{,}143$	$-0{,}102$	$-0{,}039$	0,039	0,102	0,143	0,167
$\rho_{X,Y\|Z=z}$	$-0{,}572$	$-0{,}408$	$-0{,}156$	0,156	0,408	0,572	0,668

Näide 3.1. Vaatame ühte simuleeritud näidet mudelist W. Kuigi mudelil W leidub statsionaarne jaotus, oleks selle leidmine keeruline. Selleks, et jada simuleerimisel statsionaarsuse eeldust rahuldada, võime alustada simulatsiooni fikseeritud režiimi alguspunktist $Z_1 = 0$ ning eemaldada saadud jada algusest mingi arvu väärtuseid (nn *burn-in* perioodi). Joonisel (6) on toodud näide sellise mudeli realisatsioonist. Emissioonid X_k ja Y_k võtavad enamjaolt positiivse režiimi korral samu väärtuseid ning negatiivse režiimi korral erinevaid väärtuseid. Näite koostamiseks kasutati programmi PK_juhuslik_ekslemine_naide.R (Lisa 1).



Joonis 6: Positiivselt korduva juhusliku ekslemise režiimiga VMM-i 40 olekut.

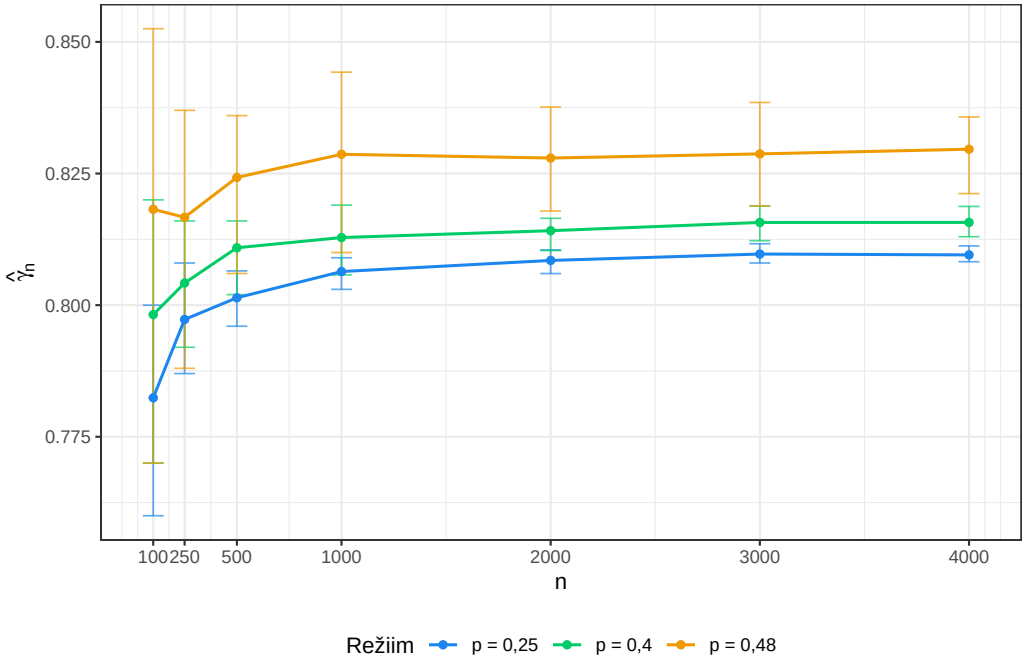
3.3 Simulatsioonid (2)

Hindame simulatsioonide abil, kuidas sõltub eelmises alapeatükis kirjeldatud mudeli korral γ väärtus parameetri p väärtusest. Võrreldi kolme erineva p väärtusega mudelit: $p = 0{,}25$, $p = 0{,}4$, $p = 0{,}48$. Väikse p korral viibib mudeli režiim seisundites 0 ja -1 või nende lähiümbruses, seega emissioonid X_k ja Z_k on enamasti nõrgalt korreleeritud. Suure p korral triivib režiim tihedamini 0-st eemale ning seega on tõenäoliselt ka suurem osa emissioonidest tugevalt korreleeritud. Simulatsioonide läbiviimisel kasutati programme PK_Juhuslik_ekslemine_simulatsioonid.R ja NW2.cpp (Lisa 1).

Simulatsioonid viidi läbi samamoodi, nagu alapeatükis (2.3):

- Iga mudeli ja $n \in \{100, 250, 500, 1000, 2000, 3000, 4000\}$ korral genereeriti valim 100-st jadade $X_{1:n}, Y_{1:n}$ paarist.
- Needleman-Wunschi algoritmiga leiti iga genereeritud jadade paari korral L_n väärtus.
- Iga mudeli ja $n \in \{100, 250, 500, 1000, 2000, 3000, 4000\}$ korral salvestati andmestikku valimikeskmine $\hat{\gamma}_n$, standardhälve $\hat{\sigma}$, alumine kvartiil $\hat{q}_{0,25}$ ja ülemine kvartiil $\hat{q}_{0,75}$.

Simulatsioonide tulemusel saadud valimikeskmised ja kvartiilide vahemikud on toodud graafikul (7). Näeme, et kõigi mudelite korral $\hat{\gamma}_n$ koondub ning $n = 2000$-st edasi ei esine märgatavaid valimikeskmiste kõikumisi. L_n/n koondumisele vihjavad ka kahanevad kvartiilide vahemikud. Näeme, et suurema p korral on ka $\hat{\gamma}_n$ väärtus suurem.



Joonis 7: $\hat{\gamma}_n$ väärtused ja valimi kvartiilid.

Tabelis (4) on toodud hinnangute $\hat{\gamma}_n$ ja valimidispersioonide $\hat{\sigma}$ väärtused $n = 4000$ korral. Tabelisse on lisatud ka ligikaudne γ väärtus *iid*-mudeli korral. Näeme, et mudeli W saavutatav $\hat{\gamma}_n$ väärtus on $p = 0{,}25$ korral ligikaudu sama, mis *iid*-mudelil, ning suuremate p väärtuste korral ületab $\hat{\gamma}_n$ väärtus *iid*-mudeli oma.

Tabel 4: $\hat{\gamma}_n$ väärtused ja valimi standardvead $n = 4000$ korral

Mudel	$p = 0{,}25$	$p = 0{,}4$	$p = 0{,}48$	*iid*-mudel
$\hat{\gamma}_n$	0,809	0,817	0,830	0,812
$\hat{\sigma}$	$2{,}14 \cdot 10^{-3}$	$3{,}90 \cdot 10^{-3}$	$1{,}08 \cdot 10^{-2}$	

Kokkuvõte

Töös anti ülevaade pikima ühisjada pikkuse sarnasusmõõdikust L_n, kolmekaupa Markovi mudelist ning juhuslike jadade võrdlemisel kasulikest varjatud Markovi mudeli omadustest.

Lähemalt käsitleti varjatud Markovi mudelit, mille emissioonid (X_k, Y_k) on sõltumatud režiimi seisundi $Z_k = 0$ korral ning positiivselt korreleeritud režiimi seisundi $Z_k = 1$ korral. Simulatsioonidega leiti, sellisel mudelil on L_n/n piirväärtus suurem, kui *iid*-mudelil, kusjuures pikima ühisjada pikkus on suurem siis, kui režiimi seisundi 1 statsionaarne tõenäosus on suurem ja režiimi saared on pikemad.

Samuti käsitleti varjatud Markovi mudeleid, mille režiim Z on juhuslik ekslemine. Kirjeldati erijuhtu, mille korral L_n/n koondub jaotuse järgi $Arcsin(0, 1)$ jaotusega juhuslikuks suuruseks. Koostati positiivselt korduva juhusliku ekslemise režiimiga mudel, mille emissioonid (X_k, Y_k) on $Z_k < 0$ korral negatiivselt korreleeritud ning $Z_k \geqslant 0$ korral positiivselt korreleeritud. Simulatsioonidega näidati, et ka sellise mudeli korral võib L_n/n koonduda suuremaks väärtuseks, kui *iid*-mudeli korral.

Kasutatud allikad

Ackelsberg, Ethan (2018). *What is the Arcsine Law?* URL: https://math.osu.edu/sites/math.osu.edu/files/What_is_2018_Arcsine_Law.pdf.

Aldridge, Matthew (2021). *Introduction to Markov Processes*. URL: https://mpaldridge.github.io/math2750/math2750.pdf.

Alm, Sven Erick (2006). *Simple random walk*. URL: https://www2.math.uu.se/~sveralm/kurser/stokprocmn1/slumpvandring_eng.pdf.

Bukh, B. ja C. Cox (2022). "Periodic words, common subsequences and frogs". *The Annals of Applied Probability* 32.2, lk. 1295–1332. DOI: 10.1214/21-AAP1709. URL: https://doi.org/10.1214/21-AAP1709.

Eddelbuettel, Dirk ja James Joseph Balamuta (2018). "Extending R with C++: A Brief Introduction to Rcpp". *The American Statistician* 72.1, lk. 28–36. DOI: 10.1080/00031305.2017.1375990.

Guichard, David (2025). *An Introduction to Combinatorics and Graph Theory*. David Guichard. URL: https://www.whitman.edu/mathematics/cgt_online/cgt.pdf.

Iher, Kati (2021). *Juhuslike jadade võrdlemine*. Bakalaureusetöö. Tartu Ülikool.

Pärna, Kalev (2013). *Tõenäosusteooria Algkursus*. Tartu Ülikooli Kirjastus.

R Core Team (2026). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. DOI: 10.32614/R.manuals. URL: https://www.R-project.org/.

Sova, Joonas (2015). "Homoloogsete jadade sõltuvusmõõdud". Magistritöö. Tartu Ülikool.

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller ja Davis Vaughan (2026). *dplyr: A Grammar of Data Manipulation*. R package version 1.2.1. DOI: 10.32614/CRAN.package.dplyr. URL: https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Davis Vaughan ja Maximilian Girlich (2025). *tidyr: Tidy Messy Data*. R package version 1.3.2. DOI: 10.32614/CRAN.package.tidyr. URL: https://CRAN.R-project.org/package=tidyr.

Wilf, Herbert S. (1994). *generatingfunctionology*. 2. väljaanne. Internet Edition. URL: https://www2.math.upenn.edu/~wilf/gfologyLinked2.pdf.

Lisa 1. Kasutatud programmid

Simulatsioonide läbiviimiseks ja nende tulemustest jooniste koostamiseks kirjutati programmid keeles R (R Core Team, 2026). C++ keeles kirjutatud Needleman-Wunschi algoritmi ühildamiseks R-ga kasutati paketti Rcpp (Eddelbuettel ja Balamuta, 2018). Simulatsioonidest saadud andmete töötlemiseks kasutati pakette tidyr (Wickham, Vaughan ja Girlich, 2025) ja dplyr (Wickham *et al.*, 2026). Jooniste koostamiseks kasutati paketti ggplot2 (Wickham, 2016). Programmide loetelu ja lühikirjeldused on järgnevad:

- VMM_naide.R - Näites (2.2) kirjeldatud sõltuvuse saartega varjatud Markovi mudeli simuleerimine ja jooniste koostamine.
- VMM_simulatsioonid.R - Alapeatükis (2.3) sõltuvuse saartega mudeli simuleerimine ja graafiku koostamine.
- Arcsin_Juhuslik_ekslemine.R - Sümmeetrilise juhusliku ekslemise režiimiga mudeli simuleerimine ja joonise (4) koostamine.
- PK_Juhuslik_ekslemine_naide.R - Näites (3.1) kirjeldatud positiivselt korduva juhusliku ekslemise režiimiga varjatud Markovi mudeli simuleerimine ja jooniste koostamine.
- PK_Juhuslik_ekslemine_naide.R - Alapeatükis (3.3) positiivselt korduva juhusliku ekslemise režiimiga mudeli simuleerimine ja graafiku koostamine.
- NW2.cpp - C++ keeles kirjutatud Needleman-Wunschi algoritmiga pikima ühisjada pikkuse leidmise programm magistritööst (Sova, 2015).

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Hans Rahi,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "Juhuslike jadade võrdlemine kolmekaupa Markovi mudeli korral", mille juhendaja on Joonas Sova, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Hans Rahi
12.05.2026