Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Matemaatika ja statistika instituut

Kati Iher

Juhuslike jadade võrdlemine

Matemaatika

Bakalaureusetöö (9 EAP)

Juhendaja: prof. Jüri Lember

Tartu 2021

JUHUSLIKE JADADE VÕRDLEMINE

Bakalaureusetöö

Kati Iher

Lühikokkuvõte

Töös kirjeldatakse ja uuritakse mõnigaid juhuslike jadade võrdlemise meetodeid. Kõigepealt kirjeldatakse juhuslikke jadasid. Seejärel defineeritakse kaks jadade sarnasusmõõtu: pikima ühisjada pikkus ning H ehk ekstremaaljoonduste kaugus. Viimases osas uurime neid sarnasusmõõte Markovi ahelal põhineval mudelil. Simulatsioonide tulemusena leiame, et H kasv on logaritmiline sõltumatude ja teatud sõltumatute jadade korral. Samuti leiame, et osadel juhudel eristab pikima ühisjada pikkus sõltuvust paremini kui H, osadel vastupidi.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: Juhuslikud protsessid, Markovi ahelad, jadad.

RANDOM SEQUENCE COMPARISION

Bachelor thesis

Kati Iher

Abstract

In this thesis, we describe and investigate some methods of random sequence comparision. First we discribe the random sequences. Then we define two similarity measurement: the length of longest common subsequence (LCS) and H or the distance of extremal alignments. Finally we compare these measurements for a model based on Markov chains. By using simulations, we find that the growth of H is logarithmic for the unrelated case and for some

related case. In addition, we find that for some cases LCS is better than H for distinguishing relatedness and vice versa.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics.

Key Words: Random processes, Markov chains, sequences.

Sisukord

Sissejuhatus

Kahe jada omavahelise võrdlemisel on mitmeid praktilisi rakendusi. Üks peamisi valdkondi on bioinformaatikas DNA, RNA ja valkude võrdlemine, kus näiteks DNA jadade sarnasus võimaldab uurida eri liikide sugulust evolutsioonipuus.

Praegu on sellisel jadade võrdlemisel kasutusel peamiselt pikima ühisjada pikkus. Lember, Matzinger ja Vollmer, 2014 pakkusid alternatiivse variandina kasutada võrdlemiseks ülemise ja alumise joonduse vahelist kaugust. Bakalaureusetöö eesmärk on vaadelda selle alternatiivse mõõdiku käitumist ning uurida, kas see võiks teatud jadade puhul sõltuvust paremini eristada kui pikima ühisjada pikkus.

Bakalaureusetöö on jagatud kolmeks peatükiks.

Töö esimeses peatükis antakse teoreetiline ülevaade Markovi ahelatest ning eellasjadast jadadest. Tegemist on kahe võimalusega, kuidas on võimalik modelleerida juhuslikke jadasid.

Teises peatükis tutvustatakse jadade võrdlemise põhimõisted, seal hulgas defineeritakse pikima ühisjada pikkus ning ülemise ja alumise joonduse vaheline kaugus H. Viimase kohta antakse ka ülevaade seni leitud peamistest tulemustest.

Kolmandas peatükis uuritakse konkreetset Markovi ahelal põhinevat juhuslike jadade mudelit ning uuritakse simulatsioonidega eri sarnasusmõõdikute käitumist.

1 Juhuslikud jadad

Siin peatükis vaatame juhuslike jadadega seotud põhimõisteid ning anname teoreetilise ülevaate kahest erinevast juhusliku jada mudelist.

Definitsioon 1.1. *Nimetame lõplikku hulka* $\mathcal{X}$ tähestikuks *ning selle hulga elemente* tähtedeks.

Näide 1.1. *Tähestiku võivad moodustada neli nukleotiidi, millist koosnevad DNA ahelad:*

$$\mathcal{X} = \{A, C, G, T\}.$$

Näide 1.2. *Bitijadade elemendid on nullid ja ühed:*

$$\mathcal{X} = \{0, 1\}.$$

Näide 1.3. *Eesti tähestik:*

$$\mathcal{X} = \{A, B, D, E, F, \ldots\}.$$

Märkus 1.1. *Siin töös on naturaalarvude hulk* $\mathbb{N} = \{1, 2, 3 \ldots\}$*, st nulli me ei loe siin naturaalarvuks.*

Definitsioon 1.2. Juhuslikuks jadaks *nimetame jada* $\{X_n\}_{n=1}^{\infty}$*, mille puhul on iga liige* X_i*,* $i \in \mathbb{N}$ *juhuslik suurus.*

Siin töös eeldame edaspidi, et iga $i \in \mathbb{N}$ korral $X_i \in \mathcal{X}$ ehk X_i on diskreetne juhuslik suurus.

Lühemalt võime jada $\{X_n\}_{n=1}^{\infty}$ tähistada ka kui $\{X_n\}$.

Definitsioon 1.3. *Juhuslikku jada* $\{X_n\}$ *nimetatakse* statsionaarseks *(inglise keeles* stationary*), kui iga* $n \in \mathbb{N}$ *ja* $k \in \mathbb{N}$ *korral on vektorid*

$$(X_1, \ldots, X_n) \text{ ja } (X_{k+1}, \ldots, X_{k+n})$$

sama jaotusega.

See tähendab, et statsionaarse jada $\{X_n\}$ korral on juhuslikud suurused $X_1, X_2, \ldots$ sama jaotusega, samuti vektorid $(X_1, X_2), (X_2, X_3), \ldots$ on kõik sama jaotusega jne.

Definitsioon 1.4. *Juhusliku jada $\{X_n\}$ kõik liikmed on omavahel sõltumatud, kui iga $k \in \mathbb{N}$ ning $x_1, \ldots x_k \in \mathcal{X}$ korral*

$$P\{X_1 = x_1, \ldots X_k = x_k\} = P\{X_1 = x_1\} \cdot \ldots \cdot P\{X_k = x_k\}.$$

Definitsioon 1.5. *Juhuslikud jadad $\{X_n\}$ ja $\{Y_n\}$ on* sama jaotusega*, kui iga $k \in \mathbb{N}$ korral on juhuslikud vektorid*

$$(X_1, X_2, \ldots X_k) \text{ ja } (Y_1, Y_2, \ldots Y_k)$$

sama jaotusega.

Definitsioon 1.6. *Juhuslikud jadad $\{X_n\}$ ja $\{Y_n\}$ on omavahel* sõltumatud*, kui iga $n \in \mathbb{N}$ ja $x_1, \ldots x_n, y_1, \ldots y_n \in \mathcal{X}$ korral kehtib*

$$P\{X_1 = x_1, \ldots, X_k = x_k, Y_1 = y_1, \ldots, Y_n = y_n\} =$$
$$= P\{X_1 = x_1, \ldots, X_k = x_k\} \cdot P\{Y_1 = y_1, \ldots, Y_n = y_n\}.$$

1.1 Markovi ahelad

Käesolev alapeatükk tugineb järgmisel õpikul Pärna, 2013.

Olgu $S = \{E_1, E_2, \ldots\}$ mittetühi ülimalt loenduv hulk. Olgu $\{X_n\}$ juhuslik jada, mille kõik väärtused on hulgas S.

Definitsioon 1.7. *Hulga S elemente nimetatakse ka Markovi ahela* olekuteks *ehk* seisunditeks.

Edaspidi võime mugavuse mõttes seisundit E_i tähistada ka kui i.

Definitsioon 1.8. Markovi ahelaks *nimetatakse juhuslike suuruste jada* $\{X_n\}$, $n = 1, 2, \ldots$, *mille korral kehtib tinglike tõenäosuste võrdsus*

$$P\{X_{n+1} = j \mid \underbrace{X_1 = k_1, \ldots, X_{n-1} = k_{n-1}}_{minevik}, \underbrace{X_n = k_n}_{olevik}\} = \tag{1.1}$$

$$= P\{\underbrace{X_{n+1} = j}_{tulevik} \mid \underbrace{X_n = k_n}_{olevik}\}$$

iga $j, k_1, \ldots k_n \in S$ *ja* $n \in \mathbb{N}$ *korral.*

Tingimust 1.1 nimetatakse ka *Markovi omaduseks*. Seda võib mõista nii, et Markovi ahelate korral tulevik sõltub ainult olevikust, aga mitte minevikust. See tähendab ka, et seisundi X_{n+1} prognoosimiseks piisab ainult sellele eelneva ehk seisundi X_n teadmisest.

Definitsioon 1.9. *Markovi ahela seisundist* i *seisundisse* j ülemineku tõenäosuseks *sammul* n *nimetatakse tinglikku tõenäosust*

$$p_{ij}^{(n)} := P\{X_n = j \mid X_{n-1} = i\}$$

kus $n = 2, 3, \ldots$

Definitsioon 1.10. *Nimetame Markovi ahela* algtõenäosuste vektoriks *vektorit* π, *kus iga* $i \in S$, *korral kehtib*

$$\pi(i) = P\{X_1 = i\}.$$

Tegemist on tõenäosusjaotusega, seega $\sum_i \pi(i) = 1$.

Definitsioon 1.11. *Markovi ahelat nimetatakse* homogeenseks, *kui ülemineku tõenäosus* $p_{ij}^{(n)}$ *ei sõltu arvust* n.

$$p_{ij} := p_{ij}^{(n)}$$

Definitsioon 1.12. *Homogeense Markovi ahela korral nimetame* üleminekumaatriksiks *maatriksit*

$$P = (p_{ij}), \quad i, j = 1, 2, \ldots.$$

Maatriksi i. reaks on tinglike tõenäosuste jaotus üleminekul järgmisesse seisundisse j, kui olevikuseisund on i. Järelikult on iga rea summa 1 ehk $\sum_j p_{ij} = 1$.

Edaspidi eeldame, et vaadeldav Markovi ahel on homogeenne.

Olgu Markov ahel üleminekumaatriksiga $P = (p_{ij})$. Vaatame tõenäosust, et olles olekus i jõuab protsess täpselt k sammuga olekusse j (seejuures on lubatud ka, et mõni vahepealne olek j). Kuna eelduse kohaselt on tegemist homogeense ahelaga, siis see tõenäosus ei sõltu indeksist ehk saame tähistada

$$p_{ij}(k) := P\{X_{1+k} = j \mid X_1 = i\}.$$

Definitsioon 1.13. *Seisundit j nimetatakse* saavutatavaks *ehk* kättesaadavaks *seisundist i, kui seisundist i on mingi nullist suurema tõenäosusega võimalik positiivse arvu sammudega jõuda seisundisse j, see tähendab, et kehtib*

$$\exists k \in \mathbb{N} \quad p_{ij}(k) > 0. \tag{1.2}$$

Lühemalt võib seisundi j saavutatavust seisundist i tähistada $i \to j$ ehk $j \leftarrow i$. On ilmne, et tegemist, et saavutatavus on transitiivne seos ehk kui $i \to j$ ja $j \to k$, siis ka $i \to k$.

Juhul tingimus 1.2 ei kehti, siis see öeldakse, et seisund j *ei ole saavutatav* ehk *kättesaadav* seisundist i. Seda tähistame kui $i \not\to j$ ehk $j \not\leftarrow i$ ja sel juhul kehtib

$$\forall k \in \mathbb{N} \quad p_{ij}(k) = 0.$$

Definitsioon 1.14. *Seisundit i nimetatakse* ebaoluliseks*, kui leidub selline seisund j, niimoodi, et $i \to j$, aga $j \not\to i$.*

Niisiis ebaolulisest seisundist i on võimalik jõuda seisundisse, kust enam tagasi seisundisse i pole võimalik saada. Osutub, et varem või hiljem peab Markovi ahel

jõudma ebaolulisest seisundist sellisesse seisundisse, kust tagasi pole enam võimalik jõuda. Seetõttu saavad ebaolulised seisundid esineda ahelas ainult lõplik arv kordi.

Definitsioon 1.15. *Seisundit i, mis pole ebaoluline, nimetatakse* oluliseks. *See tähendab, et iga seisundi j korral, kui $i \to j$, siis kehtib ka $j \to i$.*

Definitsioon 1.16. *Seisundeid i ja j nimetatakse* kaasnevateks seisunditeks, *kui kehtib nii $i \to j$ kui ka $j \to i$*

Näide 1.4. *Olgu $S = \{E_1, E_2, E_3, E_4\}$ ja Markovi ahela $\{X_n\}$ üleminekumaatriks*

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{4} & 0 & \frac{1}{4} \end{pmatrix}$$

Siin seisundid E_3 ja E_4 on mõlemad ebaolulised, kuna neist on võimalik sattuda seisundisse E_1, kuid mitte jõuda tagasi.

Seevastu seisundid E_1 ja E_2 on mõlemad olulised ning omavahel kaasnevad seisundid, kuna mõlemast seisundist on võimalik jõuda teise.

Meenutame, et S on Markovi ahela seisundite hulk. Toome sisse järgmised tähised:

$$S^0 = \{i \in S \mid \text{seisund } i \text{ on ebaoluline}\},$$
$$S' = \{i \in S \mid \text{seisund } i \text{ on oluline}\}.$$

Vastavalt olulisuse definitsioonile on ilmne, seisundite hulk S on hulkade S^0 ja S' lõikumatu ühend.

Vaatame hulgal oluliste seisundite hulgal S' kaasnevuse seost. See tähendab, et elemendid i ja j on seoses parajasti siis kui need on kaasnevad seisundid. Näitame, et tegemist on ekvivalentsiseosega hulgal S':

- Reflektsiivsus. Olgu $i \in S'$ oluline seisund. Leidub seisund j nii, et $i \to j$. Selline seisund j eksisteerib, kuna $\sum_j p_{ij} = 1$, siis järelikult leidub j nii et $p_{ij} > 0$, mistõttu seisund j on saavutatav. Kuna seisund i on oluline, siis seoses $i \to j$ järeldub, et $j \to i$. Viimaks $i \to i$, sest kehtivad $i \to j$ ja $j \to i$.
- Transitiivsus järeldub seose $i \to j$ ehk saavutatavuse transitiivsusest.
- Sümmeetrilisus tuleneb vahetult kaasnevuse definitsioonist.

Niisiis kaasnevuse seose põhjal klassijaotuse oluliste seisundite hulgal S'

$$S' = S^1 \sqcup S^2 \sqcup \dots$$

ning lisades ebaoluliste seisundite hulga S^0 saame klassijaotuse ka hulgal S:

$$S' = S^0 \sqcup S^1 \sqcup S^2 \sqcup \dots.$$

Lause 1.1. *Kui Markovi ahel jõuab kaasnevate seisundite klassi S^i, siis ei välju see sealt kunagi.*

Tõestus. Olgu $i \in S^t$ seisund, kuhu Markovi ahel on jõudnud. Oletame vastuväiteliselt, et ahel jõuab seisundisse $j \notin S^t$ ehk $i \to j$. Peab kehtima ka $j \to i$, sest i on oluline seisund. Oleku j osas on aga kaks võimalust:

- j on oluline seisund. Et $i \to j$ ja $j \to i$, siis tegemist on kaasnevate seisunditega ja seega $j \in S^t$, mis on vastuolus eeldusega $j \notin S^t$.
- j on ebaoluline seisund ehk $j \in S^0$. Seisundi j ebaolulisuse tõttu leidub seisund k niimoodi, et $j \to k$, aga $k \not\to j$.

  Järelikult peab kehtima $i \to k$, sest $i \to j$ ja $j \to k$. Sellest seisundi i olulisuse tõttu ka $k \to i$. Nüüd aga kehtib $k \to i$ ja $i \to j$ tõttu $k \to j$.

  Niisiis saime vastuolu, et $k \not\to j$ ja $k \to j$.

Jõudsime mõlema variandi korral vastuoluni, seega sellist seisundit $j \notin S^t$ ei leidu ehk ahel jääb kogu aeg klassi S^t. □

Definitsioon 1.17. *Markovi ahelat nimetatakse* mittelahutuvaks *ehk* taandumatuks*, kui* $S = S^1$*. Kui Markovi ahel ei ole mittelahutuv, siis nimetatakse seda* lahutuvaks*.*

See tähendab, et mittelahutuvas Markovi ahelas on alati positiivse tõenäosusega jõuda igast seisundist suvalisse teise seisundisse.

Uurime millal on homogeenne Markovi ahel statsionaarne.

Lause 1.2. *Homogeenne Markovi ahel on statsionaarne parajasti siis, kui algtõenäoususte vektori* π *ja üleminekumaatriksi* P *korral kehtib*

$$\pi P = \pi.$$

Tõestus. Olgu π algtõenäosuste vektor ehk $\pi(i) = P\{X_1 = i\}$. Täistõenäosuse valemi kohaselt.

$$\begin{aligned} P\{X_2 = j\} &= \sum_{i=1}^{\infty} P\{X_1 = i\} P\{X_2 = j \mid X_1 = i\} \\ &= \sum_{i=1}^{\infty} \pi(i) p_{ij}. \end{aligned}$$

Saadud tulemus on j. komponent vektoris πP, niisiis X_2 vastab tõenäosusjaotus on πP. Analoogiliselt vastab X_3 jaotusele vektor $(\pi P)P = \pi P^2$ jne ehk üldistatult X_t jaotusele vastab vektor πP^{t-1}.

Piisavus. Kui Markovi ahel on statsionaarne, siis X_1 ja X_2 peavad olema sama jaotusega. Samas eelneva põhjal X_1 jaotusele vastab vektor π ja X_2 jaotusele πP. Selleks, et need oleksid sama jaotusega, peab kehtima

$$\pi P = \pi.$$

Tarvilikkus. Juhusliku suuruse X_t jaotus on eelneva põhjal πP^{t-1}. Niisiis kui kehtib $\pi P = \pi$, siis saame X_t jaotuseks iga $t \geq 2$ korral

$$\pi P^{t-1} = (\pi P)P^{t-2} = \pi P^{t-2} = \ldots = \pi.$$

Seega on kõik juhuslikud suurused X_i, $i \in \mathbb{N}$ sama jaotusega kui X_1 ehk jaotusest π.

Olgu $n \geq 2$ ja $k \in \mathbb{N}$. Näitame, et on vektorid $(X_1, \ldots X_n)$ ja $(X_{k+1}, \ldots X_{k+n})$. Olgu $s_1, s_2, \ldots s_n \in S$, siis kasutades tõenäosuste korrutamislauset ja Markovi omadust saame

$$\begin{aligned}
&P\{X_{s+1} = s_1, X_{s+2} = k_2, \ldots X_{s+n} = k_n\} \\
&P\{X_{s+1} = k_1\}P\{X_{s+2} = k_2 \mid X_{s+1} = k_1\}P\{X_{s+3} = k_3 \mid X_{s+1} = k_1, X_{s+2} = k_2\} \\
&\ldots P\{X_{s+n} = k_n \mid X_{s+1} = x_1, \ldots X_{s+n-1} = k_{n-1}\} \\
&= P\{X_{s+1} = k_1\}P\{X_{s+2} = k_2 \mid X_{s+1} = k_1\}P\{X_{s+3} = k_3 \mid X_{s+2} = k_2\} \\
&\ldots P\{X_{s+n} = k_n \mid X_{s+n-1} = k_{n-1}\}.
\end{aligned}$$

Kuna X_1 ja X_{s+1} on sama jaotusega (π) ning kasutades Markovi ahela homogeensusest saame edasi

$$\begin{aligned}
&P\{X_{s+1} = s_1, X_{s+2} = k_2, \ldots X_{s+n} = k_n\} \\
&= P\{X_1 = k_1\}P\{X_2 = k_2 \mid X_1 = k_1\}P\{X_3 = k_3 \mid X_2 = k_2\} \\
&\ldots P\{X_n = k_n \mid X_{n-1} = k_{n-1}\} \\
&= P\{X_1 = k_1\}P\{X_2 = k_2 \mid X_1 = k_1\}P\{X_3 = k_3 \mid X_1 = k_1, X_2 = k_2\} \\
&\ldots P\{X_n = k_n \mid X_1 = x_1, \ldots X_{n-1} = k_{n-1}\} \\
&= P\{X_1 = s_1, X_2 = k_2, \ldots X_n = k_n\}.
\end{aligned}$$

Sellest järeldubki, et ka vektorid $(X_1, \ldots X_n)$ ja $(X_{k+1}, \ldots X_{k+n})$, kus $n \geq 2$ on samast jaotusest, mistõttu on Markovi ahel statsionaarne. □

Algtõenäosuste vektorit π, mille korral Markovi ahel on statsionaarne ehk $\pi P = \pi$, nimetatakse *statsionaarseks algjaotuseks.*

Lause 1.3. *Kui Markovi ahel on mittelahutuv ja selle seisundite hulk on lõplik, siis leidub ühene statsionaarne algjaotus.*

Paneme tähele, et

$$\pi P = \pi \iff (\pi P)^T = \pi^T \iff P^T \pi^T = \pi^T$$

Niisiis on algtõenäosuste vektor statsionaarne algjaotus parajasti siis, kui see vastab maatriksi P^T omaväärtusele 1 vastavale omavektorile.

1.2 Eellasjadast jadad

Selles alapeatükis põhineb tööl (Sova, 2013).

Vaatame võimalust, kuidas ühest juhuslikust jadast saadakse teised, kasutades järgmisi teisendusi:

- mutatsioonid ehk mingid jada elemendid on asendunud teistega,
- kadumised ehk osa elementidest on jadast eemaldatud.

Meenutame, et $\mathcal{X}$ on lõplik tähestik.

Definitsioon 1.18. *Nimetame juhuslike suuruste jada* $\{X_n\}$, $X_i \in \mathcal{X}$ iid *jadaks (inglise keeles* independent and identically distributed*), kui:*

- *Kõik selle liikmed on ühe ja sama jaotusega ehk iga* $i, j \in \mathbb{N}$ *ja* $x \in \mathcal{X}$ *korral*

$$P\{X_i = x\} = P\{X_j = x\}$$

- *selle liikmed on omavahel sõltumatud ehk iga* $k \in \mathbb{N}$ *ja* $x_1, \dots x_k \in \mathcal{X}$ *korral*

$$P\{X_1 = x_1, \dots X_k = x_k\} = P\{X_1 = x_1\} \cdot \ldots \cdot P\{X_k = x_k\}$$

Definitsioon 1.19. Iid jada jaotuseks *nimetame selle suvalise elemendi jaotust.*

Definitsioon 1.20. *Olgu* $\xi \in \mathbb{R}$ *mingi juhuslik suurus ning*

$$f \colon \mathcal{X} \times \mathbb{R} \to \mathcal{X}$$

Mutatsiooniks *nimetatakse juhuslikku funktsiooni*

$$F \colon \mathcal{X} \to \mathcal{X},$$

mille puhul

$$F(i) = f(i, \xi), \quad j \in \mathcal{X}.$$

Mutatsiooni võime kirjeldada *üleminekumaatriksi* Q abil, kus

$$Q = (q_{ij}), \quad q_{ij} = P\{F(i) = j\}$$

ehk i. reas ja j. veerus on tõenäosus, et i vastav element muutub elemendiks j vastavaks elemendiks.

Näide 1.5. *Kui* Q *on ühikmaatriks, siis on tegemist mutatsiooniga, mis jätab kõik jada elemendid samaks.*

Olgu $\{\xi_n\}$, kus $\xi_i \in \mathbb{R}$ jada ning $\{Z_n\}$, kus $Z_i \in \mathcal{X}$ omavahel sõltumatud iid jadad. Olgu mutatsioonid $F_1, F_2, \dots$ sellised, et

$$F_k(i) := f(i, \xi_k), \quad k = 1, 2, \dots$$

Vaatame nüüd jada $\{X_n\}$, mis on saadud jadast $\{F_n(Z_n)\}$ mingite liikmete kus-

tutamisel. Selleks, vaatleme juhuslike suuruste jaotusega $Be(p)$ iid jada $\{D_n^x\}$: kui $D_i^x = 1$, siis $F(Z_i)$ jääb jadasse alles ja vastasel juhul kaob. Seejuures eeldame edaspidi, et $p > 0$, sest vastasel juhul $D_x^i = 0$ iga i korral ja kõik jada $\{F_n(Z_n)\}$ liikmed kustutataks.

Olgu X_i mingi jada $\{X_n\}$ element, seega $X_i = F_k(Z_k)$, kus $F_k(Z_k)$ on täpselt i. element, mis jadasse $\{F_n(Z_n)\}$ alles jääb. Kokkuvõttes tähendab see seda, et $X_i = F_k(Z_k)$ parajasti siis, kui

$$\sum_{j=1}^{k} D_j^x = i, \quad D_k^x = 1.$$

Kui $X_i = F_k(Z_k)$, siis juhuslikku suurust X_i nimetatakse juhusliku suuruse Z_k *järglaseks*. Teistpidiselt juhuslik suuruse X_i *eellaseks* nimetatakse suurust Z_k. Samuti vastab juhusliku suuruse X_i *eellase indeks* k, mis on samuti juhuslik suurus ja mida tähistame sümboliga K_i. Võime märgata, et ilmselt $K_i \geq i$.

Samuti nimetame jada $\{Z_n\}$ *eellasjadaks* ja jada $\{X_n\}$ *järglasjadaks*.

Osutub, et iid eellasjadast saadud järglasjada on samuti iid.

Lause 1.4. *(Sova, 2013, Lause 1.1) Jada $\{X_n\}$ on iid.*

Olgu $\{Y_n\}$ samuti jada $\{X_n\}$ järglasjada, mis on saadud analoogiliselt järglasjadaga $\{X_i\}$. Selle jaoks olgu

- $\{\eta_n\}$, $\eta_i \in \mathbb{R}$ on iid jada, mis on sama jaotusega nagu $\{\xi_n\}$ ning mis ei sõltu ülejäänutest juhuslikest suurustest;
- mutatsioonid G_n defineeritud kui

$$G_k(i) = f(i, \nu_k), \ k = 1, 2, \ldots;$$

- $\{D_n^y\}$ on jaotusega $Be(p)$ iid jada (st $\{D_n^y\}$ on sama jaotusega nagu iid jada $\{D_n^x\}$), mis ei sõltu ülejäänutest juhuslikest suurustest.

Jada $\{Y_n\}$ on saadakse jadast $\{G_n(Z_n)\}$ osade elementide kustutamisega. Seejuures kustutamine toimub juhusliku jada D_n^y põhjal, nagu saadi jada $\{X_n\}$.

Sarnaselt lausele 1.4 saame, et ka jada $\{Y_n\}$ on iid.

Kuna nii jada $\{X_n\}$ kui ka jada $\{Y_n\}$ sõltuvad üldiselt jadast $\{Z_n\}$, siis intuitiivselt on ilmne, et üldiselt jadad $\{X_n\}$ ja $\{Y_n\}$ ei ole omavahel sõltumatud (Sova, 2013, Omadus 1.3). Samuti kehtib järgmine tulemus.

Lause 1.5. *(Sova, 2013, Omadus 1.5, a) Kui $p < 1$, siis paarid (X_i, Y_i) ei ole üldiselt sama jaotusega.*

Sellest lausest järeldub vahetult, et kui $p < 1$, siis jada $\{(X_n, Y_n)\}$ ei ole üldiselt statsionaarne.

Näide 1.6. *Olgu $\mathcal{X} = \{A, C, G, T\}$ ning vaatame kuidas eellasjadast $\{Z_n\}$ saadakse järglasjadad $\{X_n\}$ ja $\{Y_n\}$:*

$\{Z_n\}$	A	C	C	A	T	G	A	A	$\dots$
$\{F_n(Z_n)\}$	A	T	C	A	T	C	A	C	$\dots$
$\{D_n^x\}$	1	1	0	1	0	1	1	1	$\dots$
$\{X_n\}$	A	T		A		C	A	C	$\dots$

$\{Z_n\}$	A	C	C	A	T	G	A	A	$\dots$
$\{G_n(Z_n)\}$	A	C	T	A	T	G	T	A	$\dots$
$\{D_n^y\}$	1	0	0	1	0	1	1	1	$\dots$
$\{Y_n\}$	A			A		G	T	A	$\dots$

Siin näiteks elementidel $X_3 = A$ ja $Y_2 = A$ on ühine eellane $Z_4 = A$ ning elementide $X_6 = C$ ja $Y_5 = A$ ühine eellane on $Z_8 = A$.

2 Juhuslike jadade võrdlemine

2.1 Tarvilikud eelteadmised

Meenutame, et $\mathcal{X}$ on lõplik hulk ehk tähestik.

Definitsioon 2.1. *Nimetame* lõplikuks jadaks *ehk* jadaks *tähestikus* $\mathcal{X}$ *pikkusega* n *hulga*

$$\mathcal{X}^n = \{(x_1, x_2, \ldots, x_n \mid x_i \in \mathcal{X}, 1 \leq i \leq n)$$

elemente.

Lühemalt võime lõplikku jada tähistada

$$x = x_1 x_2, \ldots x_n,$$

kus iga $i \in \{1, 2, \ldots n\}$ korral $x_i \in \mathcal{X}$ ning $n \in \mathbb{N}$ on *jada pikkus.*

Võtame lisaks tähestikule kasutusele ka sümboli nimetusega *indel*, mida tähistatakse $-$, seejuures eeldame, et $- \notin \mathcal{X}$.

Definitsioon 2.2. *Tähestiku* $\mathcal{X}$ laienduseks *ehk* laiendatud tähestikuks *nimetatakse tähestikku* $\mathcal{X}_+$*, millesse on lisatud indel*

$$\mathcal{X}_+ = \mathcal{X} \cup \{-\}.$$

Definitsioon 2.3. *Jada* $x \in \mathcal{X}^n$ *laiendatud jadaks nimetatakse jada* $x^* \in X_+^m$ *(*$n \leq m$*), mille korral leiduvad sellised indeksid*

$$1 \leq i_1 < i_2 < \ldots < l_n \leq m$$

niimoodi, et

$$x^*_{i_k} = x_k, \ k = 1, 2, \ldots n$$

ja kõik indeksitele $i_1, i_2, \ldots i_n$ mittevastavad elemendid jadas x^* on indelid ehk

$$l \notin \{i_1, i_2, \ldots i_n\} \Rightarrow x_l^* = -, \quad l = 1, 2, \ldots, m.$$

Teisisõnu on saadakse jadast laiendatud jada sellesse indlite lisamisel, seejuures lisatavate indlite arv võib ka olla null.

Näide 2.1. *Olgu selles ja järgnevates näidetes $\mathcal{X}$ eesti tähestik. Vaatame jada $x = HELEVANDU$, selle jada mõned laiendatud jadad on näiteks järgmised:*

- $HELEVANDU$,
- $HELE - VANDU$,
- $---HE - L - - - - EVAND - U - --$.

Definitsioon 2.4. *Jadade x ja y* joonduseks *pikkusega l nimetame paari (x^*, y^*), mille puhul kehtivad järgmised tingimused:*

1. *x^* ja y^* on vastavalt jadade x ja y laiendatud jadad,*
2. *jadade x^* ja y^* pikkus on mõlemal l,*
3. *laiendatud jadade samal indeksil ei asu korraga indlid ehk iga $1 \leq i \leq n$ korral $x_i^* \neq -$ või $y_i^* \neq -$.*

Jadades x^* ja y^* samadel indeksil asuvaid elemente nimetatame omavahel *vastandatud* elementideks.

Näide 2.2. *Vaatame jadasid $x = HELEVANDU$ ja $y = TUMEVANTSUD$. Nende jadade ükse võimalikke joondusi on näiteks*

$$(-HELEVAN - -DU-, TUM - EV - ANTSUD),$$

mida võib mugavamal kujul esitada ka järgmisel viisil

$-$	H	E	L	E	V	A	N	$-$	$-$	D	U	$-$
T	U	M	$-$	E	V	$-$	A	N	T	S	U	D

.

Samas näiteks $(HELEVAND-U, TUME---VANTSUD)$ *pole joondus, kuna laiendatud jadade pikkused pole võrdsed.*

Samuti ei sobi joonduseks $(---HELEVANDU, -TUMEVANTSUD)$*, sest laiendatud esimesel kohal vastandatakse omavahel indleid.*

Definitsioon 2.5. *Olgu* $y = y_1, y_2, \ldots y_n$ *jada tähestikus* $\mathcal{X}$*. Jada* $x = x_1, x_2, \ldots x_m$ *nimetatakse jada* y osajadaks *ehk* alamjadaks*, kui leiduvad indeksid*

$$1 \leq i_1 < i_2 < \ldots < i_m \leq n$$

nii, et

$$y_{i_k} = x_k, \quad k = 1, 2, \ldots m.$$

Näide 2.3. *Jada* $x = LEA$ *on jada* $y = ELEVANDU$ *osajada, kuna leiduvad indeksid* $i_1 = 2$*,* $i_2 = 3$ *ja* $i_3 = 5$*, nii et*

$$x_1 = y_{i_1} = y_2 = L,$$
$$x_2 = y_{i_2} = y_3 = E,$$
$$x_3 = y_{i_3} = y_5 = A.$$

Samas aga näiteks jada $x' = AVE$ *ei ole jada* $ELEVANDU$ *osajada.*

Niisiis intuitiivselt saame osajada esialgsest jadast tähti kustutades, kuid allesjäänute järjekorda muuta ei tohi.

Definitsioon 2.6. *Olgu* $x = x_1 x_2 \ldots x_m$ *ja* $y = y_1 y_2 \ldots y_n$ *jadad. Jadade* x *ja* y

ühisjadaks *nimetame jada* $z = z_1 z_2 \dots z_l$, *mille korral leiduvad sellised indeksid*

$$1 \leq i_1 < i_2 < \dots i_l \leq m$$

ja

$$1 \leq j_1 < j_2 < \dots j_l \leq n$$

niimoodi, et

$$x_{i_k} = y_{j_k} = z_k, k = 1, 2, \dots, l.$$

Teisisõnu on jadade x ja y ühisjada jada, mis on nii jada x kui ka jada y osajada.

Näide 2.4. *Jadade* $x = HELEVANDU$ *ja* $y = TUMEVANTSUD$ *ühisjadad on näiteks jadad* U, EVA, $EVAND$ *ja* $EVANU$. *Neist* $z = EVANU$ *on ühisjada, sest leiduvad indeksid*

$$1 \leq i_1 = 2 < i_2 = 5 < i_3 = 6 < i_4 = 7 < i_5 = 9 \leq 9$$

ja

$$1 \leq j_1 = 4 < j_2 = 5 < j_3 = 6 < j_4 = 7 < j_5 = 10 \leq 11$$

niimoodi, et

$$x_{i_1} = x_2 = y_{j_1} = y_4 = E$$
$$x_{i_2} = x_5 = y_{j_2} = y_5 = V$$
$$x_{i_3} = x_6 = y_{j_3} = y_6 = A$$
$$x_{i_4} = x_7 = y_{j_4} = y_7 = N$$
$$x_{i_5} = x_9 = y_{j_5} = y_{10} = U.$$

Definitsioon 2.7. *Kahe jada* pikimaks ühisjadaks *(inglise keeles* longest common subsequence*) nimetatakse ühisjada, mille pikkus on maksimaalne. See tähendab, et iga teise ühisjada pikkus pole suurem pikima ühisjada pikkusest.*

Näide 2.5. *Vaatame jadasid* $x = HELEVANDU$ *ja* $y = TUMEVANTSUD$, *nende pikimad ühisjadad on jadad* $EVAND$ *ja* $EVANU$.

Nagu näha ka näitest, siis pikim ühisjada ei pruugi olla üheselt määratud, ent on ilmne, et selle pikkus on alati ühene.

Pikima ühisjada pikkust on laialdaselt kasutusel jadade omavahelisel võrdlemisel. Järgneva osa eesmärk on defineerida sellele alternatiivne mõõdik H, mis iseloomustab seda, kui erinevad võivad olla kahe jada pikimad ühisjadad.

Paneme tähele, et igale ühisjadale saab vastavusse seada ühe või mitu joondust (ent igale joondusele ei vasta veel ühisjada). Ühisjadale vastavas joonduses ühisjadasse kuuluvad tähed on vastandatud sama tähega ning ülejäänutele vastandatakse indlid. Selles töös loeme joondused samadeks, kui vastandatakse omavahel samadel indeksitel olevaid tähti.

Eelnevas näites toodud jadad x ja y ühisjadale $EVANU$ kaks vastavat joondusust on näiteks

H	$-$	$-$	$-$	E	L	E	V	A	N	D	$-$	$-$	U	$-$
$-$	T	U	M	E	$-$	$-$	V	A	N	$-$	T	S	U	D

ja

$-$	$-$	H	$-$	E	L	E	V	A	N	$-$	D	$-$	U	$-$
T	U	$-$	M	E	$-$	$-$	V	A	N	T		S	U	D

.

Need joondused loeme siin töös samadeks, kuna mõlemas neis on vastandatud omavahel tähed x_2 ja y_4, x_5 ja y_5, x_6 ja y_6, x_7 ja y_7 ning x_9 ja y_{10}.

Kuna meid sellises ühisjada joonduses huvitab millised elemendid on omavahel vastandatud, siis sellist ühisjada joondust on võimalik kirjeldada indeksipaaride hulgana

$$\{(i_1, j_1), (i_2, j_2), \dots (i_l, j_l)\} \subset \mathbb{N}^2,$$

kus i_k, j_k, $k \in \mathbb{N}$ on indeksid ühisjada definitsioonist ja l on ühisjada pikkus.

Niisiis ülaltoodud ühisjada joondust saame tähistada hulgana

$$\{(2,4),(5,5),(6,6),(7,7),(9,10)\}.$$

Samas aga ühisjada $EVANU$ jaoks pole see ainuvõimalik joondus, sest võime omavahel vastandada tähtede x_2 ja y_4 asemel ka tähed x_4 ja y_4. Sel juhul saame joonduse

H	E	L	$-$	$-$	$-$	E	V	A	N	D	$-$	$-$	U	$-$
$-$	$-$	$-$	T	U	M	E	V	A	N	$-$	T	S	U	D

mida võime hulgana tähistada kui

$$\{(4,4),(5,5),(6,6),(7,7),(9,10)\}.$$

Ühisjada definitsioonist on ilmne, et jadade x ja y ühisjada on ka jadade y ja x ühisjada. Samas kui tähistada joondust $\mathbb{N}^2$ osahulgana, siis selline sümmeetria ei kehti.

Tulles tagasi eelmise näite juurde, leiame kõik võimalikud pikimate ühisjadade joondused:

- $\{(2,4),(5,5),(6,6),(7,7),(8,11)\}$ (vastab ühisjadale $EVAND$),
- $\{(4,4),(5,5),(6,6),(7,7),(8,11)\}$ (vastab ühisjadale $EVAND$),
- $\{(2,4),(5,5),(6,6),(7,7),(9,10)\}$ (vastab ühisjadale $EVANU$),
- $\{(4,4),(5,5),(6,6),(7,7),(9,10)\}$ (vastab ühisjadale $EVANU$).

Näeme, et nii nagu ka pikimaid ühisjadasid võib olla mitu, siis neile vastavaid joondusi võib olla veelgi rohkem. Pikimatele ühisjadade vastavaid joondusi nimetame ka *optimaalseteks joondusteks*.

Neis joondustes olevaid indeksitepaare võime vaadelda ka kui punkte järgmistel (koordinaat)tasanditel:

D								$*$	
U									
S									
T									
N							$*$		
A						$*$			
V					$*$				
E		$*$							
M									
U									
T									
	H	E	L	E	V	A	N	D	U

D								$*$	
U									
S									
T									
N							$*$		
A						$*$			
V					$*$				
E		$*$							
M									
U									
T									
	H	E	L	E	V	A	N	D	U

D									
U									$*$
S									
T									
N							$*$		
A						$*$			
V					$*$				
E		$*$							
M									
U									
T									
	H	E	L	E	V	A	N	D	U

D									
U									$*$
S									
T									
N							$*$		
A						$*$			
V					$*$				
E		$*$							
M									
U									
T									
	H	E	L	E	V	A	N	D	U

Juhul kui vaadeldavad jadad on väga pikad, siis ühendame punktid üheks pidevaks

graafikuks. Kanname kõik võimalikud optimaalsed joondused ühele joonisele

D								*	
U									*
S									
T									
N							*		
A						*			
V					*				
E		*		*					
M									
U									
T									
	H	E	L	E	V	A	N	D	U

.

Näeme, et üks joondustest on justkui kõige peal (sinised ja mustad punktid ehk ühisjadale $EVAND$ vastav joondus $\{(2,4),(5,5),(6,6),(7,7),(8,11)\}$) ning teine kõige all (punased ja mustad punktid ehk ühisjadale $EVANU$ vastav joondus $\{(4,4),(5,5),(6,6),(7,7),(9,10)\}$). Neid kahte joondust nimetamegi vastavalt *ülemiseks joonduseks* ja *alumiseks joonduseks*, mõlemaid kokku võib nimetada ka *ekstremaaljoondusteks*.

Anname alumisele ja ülemisele joondusele ka formaalsed definitsioonid. Selleks toome sisse järgmised tähistused.

Olgu x ja y jadad ning vaatleme nende jadade kõigi optimaalsete joonduste hulka A.

$$A = \{\{(i_1^\alpha, j_1^\alpha), \ldots (i_k^\alpha, j_k^\alpha)\} \mid \alpha \in A\},$$

kus $A = \{1, 2, \ldots |A|\}$ ja k on jadade x ja y pikima ühisjada pikkus.

Tähistame iga $\alpha \in A$ ja $t \in \{1, 2, \ldots k\}$ korral

$$j(i_t^\alpha) = j_t^\alpha \quad \text{ja} \quad i(j_t^\alpha) = i_t^\alpha.$$

Asume nüüd järgmise skeemi abil konstrueerima ülemist joondust

$$\left\{ \left(i_1^h, j_1^h\right), \ldots \left(i_k^h, j_k^h\right) \right\}.$$

Valime j_k^h kõigist võimalikest suurustest j maksimaalse. Joondustest, kus see maksimum realiseerub valime võimalikult väikese paarilise i_k^h ehk formaalselt

$$\begin{aligned} j_k^h &:= \max_{\alpha \in A} j_k^\alpha, \\ i_k^h &:= \min\{i_k^\alpha \mid j_k^\alpha = j_k^h, \alpha \in A\}. \end{aligned}$$

Olles kindlaks määranud (i_k^h, j_k^h), leiame järgmiseks paari (i_{k-1}^h, j_{k-1}^h). Valime j_{k-1}^h maksimaalse, nii sellele vastav i oleks väiksem kui i_l^k ning teise paari komponendi i_{k-1} võimalikest kõige väiksema:

$$\begin{aligned} j_{k-1}^h &:= \max\{j_t^\alpha \mid i(j_t^\alpha) < i_k^h, \alpha \in A, t = 1, 2, \ldots, k\}, \\ i_{k-1}^h &:= \min\{i_t^\alpha \mid j(i_t^\alpha) = j_{k-1}^h, \alpha \in A, t = 1, 2, \ldots, k\}. \end{aligned}$$

Jätkame samamoodi indeksipaaride koostamist ehk üldistatult tähendab see, et iga $s = k-1, k-2, \ldots 1$ korral valime paari $(i_s^h, j_s^h))$ järgmiselt

$$\begin{aligned} j_s^h &:= \max\{j_t^\alpha \mid i(j_t^\alpha) < i_{s+1}^h, \alpha \in A, t = 1, 2, \ldots, k\}, \\ i_s^h &:= \min\{i_t^\alpha \mid j(i_t^\alpha) = j_s^h, \alpha \in A, t = 1, 2, \ldots, k\}. \end{aligned}$$

Tulemuseks saadud joondust nimetatame *ülemiseks joonduseks.*

Võib tekkida küsimus, kas selline valik on üldse alati võimalik ja kas tulemuseks on optimaalne joondus. Seda kinnitab järgmine lause, mis pakub ka seejuures pisut

lihtsama mooduse, kuidas ülemine joondus kõigi optimaalsete joonduste seast leida.

Lause 2.1. *(Lember, Matzinger ja Vollmer, 2014, Proposition 2.1) Ülatoodud skeemi järgi leitud hulk*

$$\left\{\left(i_1^h, j_1^h\right), \ldots \left(i_k^h, j_k^h\right)\right\}$$

on jadade x ja y optimaalne joondus, kusjuures

$$j_t^h = \max\{j_t^\alpha \mid \alpha \in A\}, \quad i_t^h = \min\{i_t^\alpha \mid j(i_t^\alpha) = j_t^h \alpha \in A\}, \quad t = 1, 2 \ldots k.$$

Analoogiliselt eelnenud skeemiga saame leida ka alumise joonduse

$$\left\{\left(i_1^l, j_1^l\right), \ldots \left(i_k^l, j_k^l\right)\right\},$$

mille korral kehtib

$$j_t^l = \min\{i_t^\alpha \mid \alpha \in A\}, \quad i_t^l = \max\{j_t^\alpha \mid i(j_t^\alpha) = i_t^\alpha \alpha \in A\}, \quad t = 1, 2 \ldots k.$$

Üldiselt sarnaste jadade puhul on ülemine ja alumine joondus koordinaattasandil visuaalselt teineteisele lähemal kui sõltumatute jadade korral. Niisiis eeldatavasti võiks jadade ülemise ja alumise joonduse erinevus näidata midagi ka jadade endi sarnasuse kohta. Seetõttu soovime kuidagi mõõta kahe joonduse, ülemise ja alumise, omavahelist kaugust. Üks lihtne moodus oleks mõõta maksimaalset horisontaalset või vertikaalset kaugus, mis jääb kahe joonduse vahele.

Definitsioon 2.8. *Joonduste U ja V vaheliseks* maksimaalne horisontaalseks kauguseks *nimetame suurust*

$$\max\{0, |i_u - i_v| \colon (i_u, j_u) \in U, (i_v, j_v) \in V, j_u = j_v\},$$

Definitsioon 2.9. *Joonduste U ja V vaheliseks* maksimaalne vertikaalseks kaugu-

seks *nimetame suurust*

$$\max\{0, |j_u - j_v| \colon (i_u, j_u) \in U, (i_v, j_v) \in V, i_u = i_v\}.$$

Näide 2.6. *Jadade* $HELEVANDU$ *ja* $TUMEVANTSUD$ *ekstremaaljoonduste*

$$\{(2,4),(5,5),(6,6),(7,7),(8,11)\} \text{ ja } \{(4,4),(5,5),(6,6),(7,7),(9,10)\}$$

maksimaalne horisontaalne kaugus on 2 *ning vertikaalne* 0*.*

Samas aga maksimaalne horisontaalne või vertikaalne kaugus ei pruugi väga hästi kirjeldada, kuivõrd kaugel on kaks joondust.

Näide 2.7. *Vaatame kahte joondust, mis on tähistatud sümbolitega u ja v.*

									uv
				u				v	
			uv						
	uv								
uv									

Need joondused asetsevad teineteisele üsna lähedal, erinedes seejuures ainult ühe tähe juures, ent nende maksimaalne horisontaalne kaugus on suhteliselt suur (4).

Parema mõõdikuna võtame kasutusele Hausdorffi kauguse. Märgime, et Hausdorffi kaugus on universaalne mõõdik, millega saab mõõta kahe suvalise hulga omavahelist kaugust, kui on defineeritud selle elementide vaheline kaugus.

Definitsioon 2.10. *Joonduste* $U, V \subset \mathbb{N}^2$ *vaheliseks Hausdorffi kauguseks nimetame suurust*

$$h(U,V) := \max\{\sup_{u\in U}\inf_{v\in V} d(u,v), \sup_{v\in V}\inf_{u\in U} d(u,v)\},$$

kus d on mingi kaugus ruumis $\mathbb{R}^2$.

Siin töös kasutame kaugusena d maksimumkaugust ehk

$$d((x_1, y_1), (x_2, y_2))) = \max\{|x_1 - x_2|, |y_1 - y_2|\}.$$

Alternatiivina oleks võimalik on ka näiteks eukleidilist kauguse kasutamine.

Näites 2.7 toodud joonduste vaheline Hausdorffi kaugus on 1, eukleidilise kauguse kasutamise korral oleks tulemus $\sqrt{2}$.

Definitsioon 2.11. *Olgu jadade x ja y sarnasusmõõt H suurus*

$$H(x, y) = h(A_H, A_L),$$

kus A_H ja A_L on jadade x ja y vastavalt ülemine ja alumine joondus ning h on Hausdorffi kaugus (mille puhul d on maksimumkaugus).

Näide 2.8. *Jadade $x = HELEVANDU$ ja $y = TUMEVANTSUD$ korral on $H(x, y) = 2$.*

2.2 Teadaolevad tulemused

Järgmine alapeatükk on kokkuvõte artikli (Lember, Matzinger ja Vollmer, 2014) põhitulemustest.

Olgu $X = X_1, X_2, \ldots X_n \in \mathcal{X}^n$ ja $Y = Y_1, Y_2, \ldots Y_n \in \mathcal{X}^n$ juhuslike suuruste lõplikud jadad. Tähistame X ja Y pikima ühisjada pikkust tähisega L_n. Järeldusena Kingmani ergoodilisest subaditiivsuse teoreemist, kui X ja Y on sõltumatud ning iid jadad, siis kehtib teadaolevalt koondumine

$$\frac{L_n}{n} \to \gamma, \quad \text{p.k.} \tag{2.1}$$

Konstanti γ nimetatakse ka *Chvàtali-Sankoffi* konstandiks. γ sõltub jadade X ja Y jaotusest ning selle täpne väärtus senini teadmata.

Eeldame nüüd et lõplikud jadad X ja Y on saadud ühisest eellasjadast, nagu alapeatükis 1.2. Kui koondumises 2.1 oli eelduseks, et jadad X ja Y on sama pikad, siis artiklis näidatakse, et iid jadade, seega ja jada X ja Y puhul kehtib ka üldisem koondumine:

$$\frac{L(X_1, \dots X_n; Y_1, \dots Y_{\lfloor an \rfloor})}{n} \to \gamma_R(a) \quad \text{p. k.}$$

kus fikseeritud $a > 0$ korral on $\gamma_R(a)$ kontant (Lember, Matzinger ja Vollmer, 2014, Proposition 4.1). Edaspidi tähistame $\gamma_R := \gamma_R(1)$.

Üldjuhul X ja Y pole sõltumatud, kuid sõltumatuse korra $\gamma_R = \gamma$.

Toome sisse järgmised tähised:

$$\begin{aligned}
p_a &:= P\{X_i = a\}, \\
q &:= 1 - \min_{a \in \mathcal{X}} p_a, \\
p_0 &:= \sum_{a \in \mathcal{X}} p_a^2, \\
h(p) &:= -p \log_2 p - (1-p) \log_2 (1-p), \text{ kus } p \in [0,1], \\
\overline{p} &:= \max_{a \in \mathcal{X}} p_a, \\
\overline{q} &:= 1 - \min_{a,b \in \mathcal{X}} P\{X_1 = a \mid Y_1 = b, p = 1\}, \\
\rho &:= \frac{p_0 \overline{q}}{\overline{p} q}
\end{aligned}$$

(siin $P\{X_1 \mid Y_1 = b, p = 1\}$ on tinglik tõenäosus, et eeldusel, et pole toimunud kustutamist ehk juhuslikel suurustel X_1 ja Y_1 on ühine eellane.)

Märkus 2.1. *Sümbolid* $\vee$ *ja* $\wedge$ *märgivad siin ja edaspidi vastavalt funtksioone* max *ja* min.

Teoreem 2.1. *(Lember, Matzinger ja Vollmer, 2014, Theorem 1.3) Olgu X ja Y*

ühisest eellasjadast $Z = Z_1, Z_2, \ldots Z_n$ *on saadud lõplikud jadad. Kui kehtib*

$$\gamma_R \log_2 \overline{p} + (1 - \gamma_R \log_2(q\overline{q})) + ((1 - \gamma_R) \wedge \gamma_R) \log_2(\rho \vee 1) + 2h(\gamma_R) < 0, \quad (2.2)$$

siis leiduvad konstandid $C_r \in \mathbb{R}$ *ja* $D_r \in \mathbb{R}$ *niimoodi, et piisavalt suure* n *korral kehtib*

$$P\{H(X,Y) > C_r \ln n\} \leq D_r n^{-2}.$$

Osutub, et tingimuse 2.2 kehtimiseks, peavad jadad X ja Y olema omavahel sõltuvad.

Lemma 2.1. *Kui* X *ja* Y *on sõltumatud jadad, siis tingimus 2.2 ei kehti.*

Kokkuvõttes tähendab saadud tulemus, et teatud ühisest eellasjadadest saadud sõltuvate jadade X ja Y korral kehtib p. k. $H(X,Y) \in O(\ln n)$. Sarnasusmõõdu $H(X,Y)$ ning maksimaalse horisontaalse ja vertikaalse kauguse logaritmilist kasvu sõltuvate jadade korral kinnitavad empiiriliselt ka artiklis läbi viidud simulatsioonid. Nende põhjal pakutakse hüpoteesina ka, et sõltumatute jadade puhul on kasv lineaarne.

3 Simulatsioonid

Eelmises peatükis toodud tulemused sarnasusmõõdu H kohta põhinevad mudelil, kus jadad X ja Y on ühisest eellasjadadest saadud jadad. Samas lause 1.5 põhjal pole see mudel üldiselt statsionaarne, mis muudab selle uurimise matemaatiliselt keerukamaks. Järgnevalt uurime statsionaarset mudelit, mis põhineb Markovi ahelal, ja mille abil on võimalik samuti saada erineva sõltuvusastmega jadasid.

Seejärel uurime sellel mudelil simulatsioonidega ekstremaaljoonduste vahelisel kaugusel põhinevaid sarnasusmõõte (H, maksimaalne horisontaalne ja minimaalne kaugus) ning pikima ühisjada pikkust.

3.1 Mudel

Vaatame homogeenset Markovi ahelat $\{(X_n, Y_n)\}$ seisundite hulgal $\mathcal{X}\times\mathcal{X}$. Eeldusel, et $\mathcal{X}$ on lõplik hulk ja tegemist on mittelahutuva Markovi ahelaga, siis leidub sellel lause 1.3 põhjal ühene statsionaarne algjaotus.

Definitsioon 3.1. *Juhuslikke jadasid* $\{X_n\}$ *ja* $\{Y_n\}$ *nimetatakse Markovi ahela* $\{(X_n, Y_n)\}$ marginaaljadadeks.

Osutub, et ka statsionaarse algjaotuse korral võivad marginaaljadad $\{X_n\}$ ja $\{Y_n\}$ olla Markovi ahelad, kuid ei pruugi.

Näide 3.1. *Olgu* $\mathcal{X} = \{0, 1\}$ *ning vaatame Markovi ahelat* $\{(X_n, Y_n)\}$ *ülemineku-maatriks*

$$P = \begin{array}{c} (X,Y) \\ (0,0) \\ (0,1) \\ (1,0) \\ (1,2) \end{array} \begin{array}{c} \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\ \begin{pmatrix} \frac{5}{10} & \frac{5}{10} & 0 & 0 \\ \frac{3}{10} & \frac{2}{10} & \frac{5}{10} & 0 \\ 0 & 0 & \frac{1}{10} & \frac{9}{10} \\ \frac{2}{10} & \frac{3}{10} & \frac{4}{10} & \frac{1}{10} \end{pmatrix} \end{array}$$

kus vastavat väärtust üleminekumaatriksis tähistame

$$p_{(x_1,y_1)(x_2,y_2)} := P\{(X_n, Y_n) = (x_2, y_2) \mid (X_{n-1}, Y_{n-1}) = (x_1, y_1)\}$$

Sellisel juhul osutub, et statsionaarseks algjaotuseks on

$$\pi = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right),$$

kuna kehtib võrdus $\pi P = \pi$:

$$\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) \begin{pmatrix} \frac{5}{10} & \frac{5}{10} & 0 & 0 \\ \frac{3}{10} & \frac{2}{10} & \frac{5}{10} & 0 \\ 0 & 0 & \frac{1}{10} & \frac{9}{10} \\ \frac{2}{10} & \frac{3}{10} & \frac{4}{10} & \frac{1}{10} \end{pmatrix} = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right).$$

Vaatame nüüd tinglikku tõenäosust:

$$P\{X_3 = 0 \mid X_1 = 0, X_2 = 1\}$$

Paneme tähele, et kui $X_1 = 0$ ja $X_2 = 1$, siis peab kehtima $(X_1, Y_1) = (0, 1)$ ja $(X_2, Y_2) = (1, 0)$, sest kõik muud sobilikud üleminekutõenäosused on nullid:

$$p_{(0,0)(1,0)} = p_{(0,0)(1,1)} = p_{(0,1)(1,1)} = 0.$$

Samas aga olekust $(X_2, Y_2) = (1, 0)$ võimalik liikuda positiivse tõenäosusega ainult olekutesse, kus $X_3 \neq 0$. Seega, kui $X_1 = 0$ ja $X_2 = 1$, siis $X_3 \neq 0$, järelikult

$$P\{X_3 = 0 \mid X_1 = 0, X_2 = 1\} = 0.$$

Samas teisalt, kuna $P\{(X_2, Y_2) = (1, 1)\} = \pi(1, 1) > 0$ ja $p_{(1,1)(0,1)} > 0$, siis ilmselt

$$P\{X_3 = 0 \mid X_2 = 1\} > 0$$

Kokkuvõttes seega marginaaljada $\{X_n\}$ korral ei kehti Markovi omadus, kuna

$$P\{X_3 = 0 \mid X_1 = 0, X_2 = 1\} \neq P\{X_3 = 0 \mid X_2 = 1\}.$$

Järgnevas olgu $\mathcal{X} = \{0, 1\}$ ning vaatame oma mudelina Lember *et al.*, 2018 poolt kirjeldatud Markovi ahelat $\{(X_n, Y_n)\}$ seisundite hulgal $\mathcal{X} \times \mathcal{X}$, mille ülemineku-maatriks on

$$P = \begin{array}{c} (X,Y) \\ (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \end{array} \begin{array}{c} \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\ \begin{pmatrix} p\lambda_1 & p(1-\lambda_1) & (1-p)\theta_1 & (1-p)(1-\theta_1) \\ p\lambda_2 & p(1-\lambda_2) & (1-p)\theta_2 & (1-p)(1-\theta_2) \\ q\mu_1 & q(1-\mu_1) & (1-q)\rho_1 & (1-q)(1-\rho_1) \\ q\mu_2 & q(1-\mu_2) & (1-q)\rho_2 & (1-q)(1-\rho_2) \end{pmatrix} \end{array}$$

kus

$$(1-p)\theta_1 = p(1-\lambda_1) \iff \theta_1 = (1-\lambda_1)\frac{p}{1-p} \iff p = \frac{\theta_1}{1-\lambda_1+\theta_1} \tag{3.1}$$

$$(1-p)\theta_2 = q - p\lambda_2 \iff \theta_2 = \frac{q-p\lambda_2}{1-p} \iff q = (1-p)\theta_2 + p\lambda_2 \tag{3.2}$$

$$p - q\mu_1 = (1-q)\rho_1 \iff \rho_1 = \frac{p-q\mu_1}{1-q} \tag{3.3}$$

$$q(1-\mu_2) = (1-q)\rho_2 \iff \rho_2 = (1-\mu_2)\frac{q}{1-q} \tag{3.4}$$

Vastavat väärtust üleminekumaatriksis tähistame ka

$$p_{(x_1,y_1)(x_2,y_2)} = P\{(X_n, Y_n) = (x_2, y_2) \mid (X_{n-1}, Y_{n-1}) = (x_1, y_1)\}.$$

Lause 3.1. *Olgu $\{X_n, Y_n\}$ Markovi ahel üleminekumaatriksiga P seisundite hulgal $\mathcal{X} \times \mathcal{X}$. Sellisel juhul on marginaaljada $\{X_n\}$ Markovi ahel seisundite hulgal $\mathcal{X}$,*

kusjuures üleminekumaatriks on järgmine

$$\begin{pmatrix} p & 1-p \\ q & 1-q \end{pmatrix}.$$

Tõestus. Olgu $i, j, k \in \mathcal{X}$, siis kehtib ilmselt

$$\begin{aligned} P\{X_n = j \mid (X_{n-1}, Y_{n-1}) = (i,k)\} &= P\{(X_n, Y_n = (j,0) \mid (X_{n-1}, Y_{n-1}) = (i,k)\} \\ &\quad + P\{(X_n, Y_n = (j,1) \mid (X_{n-1}, Y_{n-1}) = (i,k)\} \\ &= p_{(i,k)(j,0)} + p_{(i,k),(j,1)}. \end{aligned}$$

Paneme tähele, et seejuures tinglik tõenäosus

$$P\{X_n = j \mid (X_{n-1}, Y_{n-1}) = (i,k)\}$$

ei sõltu selle üleminekumaatriksi korral k väärtusest. Tähistame need tinglikud tõenäosused:

$$\begin{aligned} p_{00} &:= P\{X_n = 0 \mid (X_{n-1}, Y_{n-1}) = (0,0)\} = P\{X_n = 0 \mid (X_{n-1}, Y_{n-1}) = (0,1)\} \\ &= p\lambda_1 + p(1-\lambda_1) = p\lambda_2 + p(1-\lambda_2) = p, \\ p_{01} &:= P\{X_n = 1 \mid (X_{n-1}, Y_{n-1}) = (0,0)\} = P\{X_n = 1 \mid (X_{n-1}, Y_{n-1}) = (0,1)\} \\ &= (1-p)\theta_1 + (1-p)(1-\theta_1) = (1-p)\theta_2 + (1-p)(1-\theta_2) = 1-p, \\ p_{10} &:= P\{X_n = 0 \mid (X_{n-1}, Y_{n-1}) = (1,0)\} = P\{X_n = 0 \mid (X_{n-1}, Y_{n-1}) = (1,1)\} \\ &= q\mu_1 + q(1-\mu_1) = q\mu_2 + q(1-\mu_1) = q, \\ p_{11} &:= P\{X_n = 1 \mid (X_{n-1}, Y_{n-1}) = (1,0)\} = P\{X_n = 1 \mid (X_{n-1}, Y_{n-1}) = (1,1)\} \\ &= (1-q)\rho_1 + (1-q)(1-\rho_1) = (1-q)\rho_1 + (1-q)(1-\rho_1) = 1-q. \end{aligned}$$

Näitame nüüd, et jada $\{X_n\}$ korral kehtib Markovi omadus. See tähendab, et iga

$k_1, \ldots, k_{n-2}, i, j \in \mathcal{X}$ ning $n \in \mathbb{N}$ korral

$$p_{ij} = P\{X_n = j \mid X_1 = k_1, \ldots, X_{n-2} = k_{n-2}, X_{n-1} = i\} = P\{X_n = j \mid X_{n-1} = i\}.$$

Tulenevalt täistõenäosuse valemist

$$\begin{aligned}
& P\{X_n = j \mid X_{n-1} = i\} \\
= & \sum_{k \in \mathcal{X}} P\{X_n = j \mid X_{n-1} = i, Y_{n-1} = k\} \cdot P\{Y_{n-1} = k \mid X_{n-1} = i\} \\
= & \sum_{k \in \mathcal{X}} P\{X_n = j \mid (X_{n-1}, Y_{n-1}) = (i, k)\} \cdot P\{Y_{n-1} = k \mid X_{n-1} = i\} \\
= & \sum_{k \in \mathcal{X}} p_{ij} \cdot P\{Y_{n-1} = k \mid X_{n-1} = i\} \\
= & \, p_{ij} \sum_{k \in \mathcal{X}} P\{Y_{n-1} = k \mid X_{n-1} = i\} \\
= & \, p_{ij} \cdot 1 = p_{ij}.
\end{aligned}$$

Seejuures võrdus $\sum_{k \in \mathcal{X}} P\{Y_{n-1} = k \mid X_{n-1} = i\} = 1$, kuna tegemist on tingliku tõenäosusjaotusega.

Paneme tähele, et võrdus jääb kehtima ka siis, kui sündmuse $\{X_{n-1} = i\}$ asemel on sündmus $\{X_1 = k_1, \ldots, X_{n-2} = k_{n-2}, X_{n-1} = i\}$. Seda seetõttu, et $\{X_n, Y_n\}$ on Markovi ahel ja

$$\begin{aligned}
& P\{X_n = j \mid X_1 = k_1, \ldots, X_{n-2} = k_{n-2}, (X_{n-1}, Y_{n-1}) = (i, k)\} \\
= & \, P\{X_n = j \mid (X_{n-1}, Y_{n-1}) = (i, k)\} = p_{ij}.
\end{aligned}$$

Seega saame teise vajaliku seose

$$P\{X_n = j \mid X_1 = k_1, \ldots, X_{n-2} = k_{n-2}, X_{n-1} = i\} = p_{ij},$$

mis kokkuvõttes põhjendab Markovi omaduse kehtivust jada $\{X_n\}$ puhul. □

Lause 3.2. *Olgu $\{(X_n, Y_n)\}$ Markovi ahel üleminekumaatriksiga P seisundite hulgal $\mathcal{X} \times \mathcal{X}$. Sellisel juhul on marginaaljada $\{Y_n\}$ Markovi ahel seisundite hulgal $\mathcal{X}$, kusjuures üleminekumaatriks on järgmine*

$$\begin{pmatrix} p & 1-p \\ q & 1-q \end{pmatrix}.$$

Tõestus. Vahetades seisundite järjekorda ning kasutades siis seoseid 3.1 kuni 3.4, saame üleminekumaatriksi viia kujule:

$$P = \begin{array}{c} (X,Y) \\ (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \end{array} \begin{array}{c} \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\ \begin{pmatrix} p\lambda_1 & p(1-\lambda_1) & (1-p)\theta_1 & (1-p)(1-\theta_1) \\ p\lambda_2 & p(1-\lambda_2) & (1-p)\theta_2 & (1-p)(1-\theta_2) \\ q\mu_1 & q(1-\mu_1) & (1-q)\rho_1 & (1-q)(1-\rho_1) \\ q\mu_2 & q(1-\mu_2) & (1-q)\rho_2 & (1-q)(1-\rho_2) \end{pmatrix} \end{array}$$

$$= \begin{array}{c} (X,Y) \\ (0,0) \\ (1,0) \\ (0,1) \\ (1,1) \end{array} \begin{array}{c} \begin{array}{cccc} (0,0) & (1,0) & (0,1) & (1,1) \end{array} \\ \begin{pmatrix} p\lambda_1 & (1-p)\theta_1 & p(1-\lambda_1) & (1-p)(1-\theta_1) \\ q\mu_1 & (1-q)\rho_1 & q(1-\mu_1) & (1-q)-\rho_1(1-q) \\ p\lambda_2 & (1-p)\theta_2 & p(1-\lambda_2) & (1-p)-\theta_2(1-p) \\ q\mu_2 & (1-q)\rho_2 & q(1-\mu_2) & (1-q)(1-\rho_2) \end{pmatrix} \end{array}$$

$$= \begin{array}{c} (X,Y) \\ (0,0) \\ (1,0) \\ (0,1) \\ (1,1) \end{array} \begin{array}{c} \begin{array}{cccc} (0,0) & (1,0) & (0,1) & (1,1) \end{array} \\ \begin{pmatrix} p\lambda_1 & p(1-\lambda_1) & (1-p)\theta_1 & (1-p)(1-\theta_1) \\ q\mu_1 & p-q\mu_1 & q(1-\mu_1) & (1-q)-(p-q\mu_1) \\ p\lambda_2 & q-p\lambda_2 & p(1-\lambda_2) & (1-p)-(q-p\lambda_2) \\ q\mu_2 & q(1-\mu_1) & (1-q)\rho_2 & (1-q)(1-\rho_2) \end{pmatrix} \end{array}$$

Nüüd saame samamoodi eelmise lause tõestusesega näidata, et ka $\{Y_n\}$ on Markovi ahel, kuna saame analoogselt tähistada üleminekumaatriksi elementide võrdsed

summad:

$$
\begin{aligned}
p_{00} &:= P\{Y_n = 0 \mid (X_{n-1}, Y_{n-1}) = (0,0)\} = P\{Y_n = 0 \mid (X_{n-1}, Y_{n-1}) = (1,0)\} \\
&= p\lambda_1 + p(1-\lambda_1) = q\mu_1 + p - q\mu_1 = p, \\
p_{01} &:= P\{Y_n = 1 \mid (X_{n-1}, Y_{n-1}) = (0,0)\} = P\{Y_n = 1 \mid (X_{n-1}, Y_{n-1}) = (1,0)\} \\
&= (1-p)\theta_1 + (1-p)(1-\theta_1) = q(1-\mu_1) + (1-q) - (p - q\mu_1) = 1 - p, \\
p_{10} &:= P\{Y_n = 0 \mid (X_{n-1}, Y_{n-1}) = (0,1)\} = P\{Y_n = 0 \mid (X_{n-1}, Y_{n-1}) = (1,1)\} \\
&= p\lambda_2 + q - p\lambda_2 = q\mu_2 + q(1-\mu_1) = q, \\
p_{11} &:= P\{Y_n = 1 \mid (X_{n-1}, Y_{n-1}) = (0,1)\} = P\{Y_n = 1 \mid (X_{n-1}, Y_{n-1}) = (1,1)\} \\
&= p(1-\lambda_2) + (1-p) - (q - p\lambda_2) = (1-q)\rho_2 + (1-q)(1-\rho_2) = 1 - q.
\end{aligned}
$$

□

Selleks, et jadade $\{X_n\}$ ja $\{Y_n\}$ üleminekumaatriks ei sisaldaks negatiivseid tõenäosusi, peab kehtima $p, q \in [0,1]$.

Olgu meil on $p, q \in [0,1]$ fikseeritud. Lisaks on meil üleminekumaatriksis kokku 8 parameetrit: $\lambda_i, \theta_i, \mu_i, \rho_i$, kus $i = 1,2$. Neist saame valida neli ja seoste 3.1 – 3.4 põhjal esituvad ülejäänud 4 parameetrit valitute funktsioonidena. Üldistust kitsendamata, olgu vabalt valitavateks parameetriteks λ_i, μ_1, $i = 1,2$. Osutub, et ka neid parameetreid ei saa valida täiesti vabalt, vaid need peavad rahuldama teatud tingimusi.

Lause 3.3. *Olgu $p, q \in [0,1]$. Markovi ahela (X_n, Y_n) korral üleminekumaatriksiga P parameetrite $\lambda_i, \mu_i, \theta_i$ ka ρ_i, $i = 1,2$ korral kehtima järgmised tingimused:*

1. $\lambda_1 \in \left[\dfrac{2p-1}{p} \vee 0, 1\right]$
2. $\lambda_2 \in \left[\dfrac{q+p-1}{p} \vee 0, \dfrac{q}{p} \wedge 1\right]$
3. $\mu_1 \in \left[\dfrac{p+q-1}{q} \vee 0, \dfrac{p}{q} \wedge 1\right]$

4. $\mu_2 \in \left[\dfrac{2q-1}{q} \vee 0, 1\right]$

Tõestus. Meenutame, et $\vee$ ja $\wedge$ on siin max ja min.

Esiteks peavad kõik tõenäosused maatriksis olema mittenegatiivsed, niisiis $p\lambda_1 \geq 0$ ja $p(1-\lambda_1) \geq 0$. Kuna eelduse kohaselt $p \geq 0$, siis $\lambda_1 \geq 0$ ja $1-\lambda_1 \geq 0$, millest $\lambda_1 \in [0,1]$.

Analoogselt maatriksi elementide mittenegatiivsuse põhjal ning kuna $p, q \in [0,1]$ on samas lõigus ka ülejäänud parameetrid:

$$\lambda_i, \theta_i, \mu_i, \rho_i \in [0,1], i = 1,2.$$

1. Lisaks seosele $\lambda_1 \in [0,1]$ piisab näidata, et $\lambda_1 \geq \dfrac{2p-1}{p}$.

   Eeldusest $(1-p)\theta_1 = p(1-\lambda_1)$ (3.1) saame

   $$(1-p)\theta_1 = p(1-\lambda_1) \iff \lambda_1 = 1 - \frac{(1-p)\theta_1}{p} = \frac{p-\theta_1(1-p)}{p}.$$

   Eelnevalt leitud tulemuse $\theta_1 \leq 1$ põhjal

   $$\lambda_1 = \frac{p-\theta_1(1-p)}{p} \geq \frac{p-1(1-p)}{p} = \frac{2p-1}{p}$$

   Niisiis kuna $\lambda_1 \leq 1$ ning $\lambda_1 \geq 0$ ja $\lambda_1 \geq \dfrac{2p-1}{p}$, millest $\lambda_1 \geq \dfrac{2p-1}{p} \vee 0$, siis kokkuvõttes

   $$\lambda_1 \in \left[\frac{2p-1}{p} \vee 0, 1\right].$$

2. Lisaks seosele $\lambda_2 \in [0,1]$ piisab näidata, et $\lambda_2 \geq \dfrac{q+p-1}{p}$ ja $\lambda_2 \leq \dfrac{q}{p}$.

   Võrdusest $(1-p)\theta_2 = q - p\lambda_2$ (3.2) saame

   $$(1-p)\theta_2 = q - p\lambda_2 \iff \lambda_2 = \frac{q-(1-p)\theta_2}{p}.$$

Seosest $\theta_2 \geq 0$ saame

$$\lambda_2 = \frac{q-(1-p)\theta_2}{p} \leq \frac{q-(1-p)0}{p} = \frac{q}{p}$$

ning seosest $\theta_2 \leq 1$ tuleneb

$$\lambda_2 = \frac{q-(1-p)\theta_2}{p} \geq \frac{q-(1-p)1}{p} = \frac{q+p-1}{p}.$$

3. Lisaks seosele $\mu_1 \in [0,1]$ piisab näidata, et $\mu_1 \geq \frac{p+q-1}{q}$ ja $\mu_1 \leq \frac{p}{q}$.

   Võrduse $p - q\mu_1 = (1-q)\rho_1$ (3.3) põhjal

   $$p - q\mu_1 = (1-q)\rho_1 \iff \mu_1 = \frac{p-(1-q)\rho_1}{q}$$

   Eelneva põhjal $\rho_1 \geq 0$, millest

   $$\mu_1 = \frac{p-(1-q)\rho_1}{q} \geq \frac{p-(1-q)0}{q} = \frac{p}{q}$$

   ja $\rho_1 \leq 1$, millest

   $$\mu_1 = \frac{p-(1-q)\rho_1}{q} \leq \frac{p-(1-q)1}{q} = \frac{p+q-1}{q}.$$

4. Lisaks seosele $\mu_1 \in [0,1]$ piisab näidata, et $\mu_2 \geq \frac{2q-1}{q}$.

   Võrdusest $q(1-\mu_2) = (1-q)\rho_2$ (3.4) saame

   $$q(1-\mu_2) = (1-q)\rho_2 \iff \mu_2 = \frac{q-(1-q)\rho_2}{q}$$

   ning edasi kuna $\rho_2 \leq 1$, siis

   $$\mu_2 = \frac{q-(1-q)\rho_2}{q} \geq \frac{q-(1-q)1}{q} = \frac{2q-1}{q}.$$

□

Lause 3.4. *Kui* $(p,q) \neq (1,0),(1,0)$*, siis jadade* $\{X_n\}$ *ja* $\{Y_n\}$ *ühene statsionaarne algjaotus on*

$$\pi = \left(\frac{q}{1-p+q}, \frac{1-p}{1-p+q}\right).$$

Tõestus. Kui $(p,q) \neq (1,0),(1,0)$ siis on tegemist mittelahutuva ahelaga, mistõttu leidub ühene statsionaarne algjaotus. Samas π on statsionaarse algjaotusega, sest üleminekumaatriksi P korral $\pi P = \pi$:

$$\left(\frac{q}{1-p+q}, \frac{1-p}{1-p+q}\right)\begin{pmatrix} p & 1-p \\ q & 1-q \end{pmatrix} = \left(\frac{q}{1-p+q}, \frac{1-p}{1-p+q}\right)$$

□

Lause 3.5. *Kui* $p = q$*, siis statsionaarse algjaotuse korral on marginaaljadad* $\{X_n\}$ *ja* $\{Y_n\}$ *iid jadad.*

Tõestus. Näitame, et jada $\{X_n\}$ on iid, tõestus jada $\{Y_n\}$ kohta on samasugune. Olgu $p = q$, siis üleminekumaatriks on kujul

$$\begin{pmatrix} p & 1-p \\ p & 1-p \end{pmatrix}.$$

Eelduse kohaselt on jadal $\{X_n\}$ statsionaarne algjaotus π. Eelmise lause põhjal

$$\pi = \left(\frac{q}{1-p+q}, \frac{1-p}{1-p+q}\right) = \left(\frac{p}{1-p+p}, \frac{1-p}{1-p+p}\right) = (p, 1-p)$$

Statsionaarsuse definitsioonist lähtuvalt on kõik jada elemendid X_i sama jaotusega, seega iga $i \in \mathbb{N}$ korral:

$$P\{X_i = 0\} = \pi(0) = p$$
$$P\{X_i = 1\} = \pi(1) = 1-p$$

Samuti üleminekutõenäousused langevad kokku algtõenäoustega ehk

$$p_{00} = P\{X_{n+1} = 0 \mid X_n = 0\} = p_{10} = P\{X_{n+1} = 0 \mid X_n = 0\} = p = \pi(0),$$
$$p_{01} = P\{X_{n+1} = 1 \mid X_n = 0\} = p_{11} = P\{X_{n+1} = 1 \mid X_n = 1\} = 1 - p = \pi(1).$$

Olgu $k \in \mathbb{N}$ ja $x_1, x_2, \ldots, x_k \in \mathcal{X}$. Näitame, et jada $\{X_n\}$ on sõltumatu, st peame näitama, et kehtib

$$P\{X_1 = x_1, \ldots, X_k = x_k\} = P\{X_1 = x_1\} \cdot \ldots \cdot P\{X_k = x_k\}.$$

Tulenevalt Markovi omadusest ning eelneva põhjal

$$\begin{aligned}
& P\{X_1 = x_1, \ldots, X_k = x_k\} \\
= {} & P\{X_1 = x_1\} \cdot P\{X_2 = x_2 \mid X_1 = x_1\} \cdot \ldots \\
& \ldots \cdot P\{X_k = x_k \mid X_1 = x_1, \ldots X_{k-1} = x_{k-1}\} \\
= {} & P\{X_1 = x_1\} \cdot P\{X_2 = x_2 \mid X_1 = x_1\} \cdot \ldots \\
& \ldots \cdot P\{X_k = x_k \mid X_{k-1} = x_{k-1}\} \\
= {} & \pi(x_1) \cdot p_{x_1 x_2} \cdot \ldots p_{x_{k-1} x_k} \\
= {} & \pi(x_1) \cdot \pi(x_2) \cdot \ldots \pi(x_k) \\
= {} & P\{X_1 = x_1\} \cdot \ldots \cdot P\{X_k = x_k\}.
\end{aligned}$$

Niisiis on jada X_i elemendid sama jaotusega ning sõltumatud ehk tegemist on iid jadaga. □

Lause 3.6. *Kui $\lambda_1 = \mu_1 = p$ ja $\lambda_2 = \mu_2 = q$, siis $\{X_n\}$ ja $\{Y_n\}$ sõltumatud Markovi ahelad.*

Tõestus. Leiame ülejäänud parameetrite väärtused

$$\theta_1 = (1 - \lambda_1)\frac{p}{1-p} = (1-p)\frac{p}{1-p} = p,$$

$$\theta_2 = \frac{q - p\lambda 2}{1-p} = \frac{q-pq}{1-p} = q,$$
$$\rho_1 = \frac{p - q\mu_1}{1-q} = \frac{p-qp}{1-q} = p,$$
$$\rho_2 = (1-\mu_2)\frac{q}{1-q} = (1-q)\frac{q}{1-q} = q.$$

Seega üleminekumaatriks P võtab kuju

$$P = \begin{array}{c} (X,Y) \\ (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \end{array} \begin{array}{c} \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\ \begin{pmatrix} p\lambda_1 & p(1-\lambda_1) & (1-p)\theta_1 & (1-p)(1-\theta_1) \\ p\lambda_2 & p(1-\lambda_2) & (1-p)\theta_2 & (1-p)(1-\theta_2) \\ q\mu_1 & q(1-\mu_1) & (1-q)\rho_1 & (1-q)(1-\rho_1) \\ q\mu_2 & q(1-\mu_2) & (1-q)\rho_2 & (1-q)(1-\rho_2) \end{pmatrix} \end{array}$$

$$= \begin{array}{c} (X,Y) \\ (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \end{array} \begin{array}{c} \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\ \begin{pmatrix} p^2 & p(1-p) & (1-p)p & (1-p)(1-p) \\ pq & p(1-q) & (1-p)q & (1-p)(1-q) \\ qp & q(1-p) & (1-q)p & (1-q)(1-p) \\ q^2 & q(1-q) & (1-q)q & (1-q)(1-q) \end{pmatrix} \end{array}$$

Sellise üleminekumaatriks statsionaarseks algjaotus π on

$$\left(\frac{q^2}{(1-p+q)^2}, \frac{(1-p)q}{(1-p+q)^2}, \frac{(1-p)q}{(1-p+q)^2}, \frac{(1-p)^2}{(1-p+q)^2}\right),$$

sest $\pi P = \pi$.

Kuna π on stasionaarne algjaotus, siis iga $i \in \mathbb{N}$ korral

$$P\{X_1 = 0\} = \pi(0,0) + \pi(0,1) = \frac{q^2}{(1-p+q)^2} + \frac{(1-p)q}{(1-p+q)^2} = \frac{q}{1-p+q},$$
$$P\{X_1 = 1\} = \pi(1,0) + \pi(1,1) = \frac{(1-p)q}{(1-p+q)^2} + \frac{(1-p)^2}{(1-p+q)^2} = \frac{1-p}{1-p+q},$$
$$P\{Y_1 = 0\} = \pi(0,0) + \pi(1,0) = \frac{q^2}{(1-p+q)^2} + \frac{(1-p)q}{(1-p+q)^2} = \frac{q}{1-p+q},$$

$$P\{Y_1 = 1\} = \pi(0,1) + \pi(1,1) = \frac{(1-p)q}{(1-p+q)^2} + \frac{(1-p)^2}{(1-p+q)^2} = \frac{1-p}{1-p+q}.$$

ning

$$\begin{aligned}
P\{X_1 = 0\}P\{Y_1 = 0\} &= \frac{q}{1-p+q} \cdot \frac{q}{1-p+q} = \frac{q^2}{(1-p+q)^2} = \pi(0,0),\\
P\{X_1 = 0\}P\{Y_1 = 1\} &= \frac{q}{1-p+q} \cdot \frac{1-p}{1-p+q} = \frac{(1-p)q}{(1-p+q)^2} = \pi(0,1),\\
P\{X_1 = 1\}P\{Y_1 = 0\} &= \frac{1-p}{1-p+q} \cdot \frac{q}{1-p+q} = \frac{(1-p)q}{(1-p+q)^2} = \pi(1,0),\\
P\{X_1 = 1\}P\{Y_1 = 1\} &= \frac{1-p}{1-p+q} \cdot \frac{1-p}{1-p+q} = \frac{(1-p)^2}{(1-p+q)^2} = \pi(1,1),
\end{aligned}$$

kusjuures viimasest neljast seosest järeldub, et iga $i, j \in X$ korral

$$P\{X_1 = i\} \cdot P\{Y_1 = j\} = \pi(i,j).$$

Meenutame, et lausete 3.1 ja 3.2 põhjal on marginaaljadad $\{X_n\}$ ja $\{Y_n\}$ mõlemad Markovi ahelad üleminekumaatriksiga

$$\begin{pmatrix} p & 1-p \\ q & 1-q \end{pmatrix}.$$

Tähistame selle maatriksi elemendid vastavalt p_{ij}. Näeme, et iga $i, j, k, l \in \mathcal{X}$ korral esitub üleminekumaatriksi P element järgmise korrutisena:

$$p_{(i,j),(k,l)} = p_{ik} \cdot p_{jk}.$$

Näitame nüüd, et Markovi ahelad $\{X_n\}$ ja $\{Y_n\}$ on sõltumatud. Olgu $n \in \mathbb{N}$ ja $x_1, \ldots x_n, y_1, \ldots y_n \in \mathcal{X}$. Sõltumatuse jaoks peab kehtima

$$\begin{aligned}
&P\{X_1 = x_1, \ldots, X_n = x_n, Y_1 = y_1, \ldots, Y_n = y_n\}\\
&= P\{X_1 = x_1, \ldots, X_n = x_n\} \cdot P\{Y_1 = y_1, \ldots, Y_n = y_n\}.
\end{aligned}$$

See võrdus kehtib eelpool leitud seoste põhjal

$$\begin{aligned}
&P\{X_1 = x_1, \ldots, X_n = x_n, Y_1 = y_1, \ldots, Y_n = y_n\} \\
&= P\{(X_1, Y_1) = (x_1, y_1) \ldots, (X_n, Y_n) = (x_n, y_n)\} \\
&= \pi(x_1, y_1) \cdot p_{(x_1,y_1),(x_2,y_2)} \cdot \ldots \cdot p_{(x_{n-1},y_{n-1}),(x_n,y_n)} \\
&= P\{X_1 = x_1\} P\{Y_1 = y_1\} p_{x_1x_2} p_{y_1y_2} \cdot \ldots \cdot p_{x_{n-1}x_n} p_{y_{n-1}y_n} \\
&= P\{X_1 = x_1\} p_{x_1x_2} \cdot \ldots \cdot p_{x_{n-1}x_n} \cdot P\{Y_1 = y_1\} p_{y_1y_2} \cdot \ldots \cdot p_{y_{n-1}y_n} \\
&= P\{X_1 = x_1, \ldots, X_n = x_n\} \cdot P\{Y_1 = y_1, \ldots, Y_n = y_n\}.
\end{aligned}$$

□

Lause 3.7. *Kui* $\lambda_1 = \mu_2 = 1$ *ja* $\pi(1,0) = \pi(0,1) = 0$ *siis* $X_t = Y_t$ *iga* $t \in \mathbb{N}$ *korral.*

Tõestus. Parameetrite λ_1 ja μ_2 väärtuste põhjal saame leida ka θ_1 ja ρ_2 väärtused:

$$\begin{aligned}
\theta_1 &= (1 - \lambda_1)\frac{p}{1-p} = (1-1)\frac{p}{1-p} = 0, \\
\rho_2 &= (1-1)\frac{q}{1-q} = (1-1)\frac{q}{1-q} = 0.
\end{aligned}$$

Seega üleminekumaatriks P võtab kuju:

$$P = \begin{array}{c} (X,Y) \\ (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \end{array} \begin{array}{c} \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\ \begin{pmatrix} p\lambda_1 & p(1-\lambda_1) & (1-p)\theta_1 & (1-p)(1-\theta_1) \\ p\lambda_2 & p(1-\lambda_2) & (1-p)\theta_2 & (1-p)(1-\theta_2) \\ q\mu_1 & q(1-\mu_1) & (1-q)\rho_1 & (1-q)(1-\rho_1) \\ q\mu_2 & q(1-\mu_2) & (1-q)\rho_2 & (1-q)(1-\rho_2) \end{pmatrix} \end{array}$$

$$
\begin{array}{c} \\ = \end{array}
\begin{array}{c}
(X,Y) \\ (0,0) \\ (0,1) \\ (1,0) \\ (1,1)
\end{array}
\begin{array}{c}
\begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\
\begin{pmatrix}
p1 & p(1-1) & (1-p)0 & (1-p)(1-0) \\
p\lambda_2 & p(1-\lambda_2) & (1-p)\theta_2 & (1-p)(1-\theta_2) \\
q\mu_1 & q(1-\mu_1) & (1-q)\rho_1 & (1-q)(1-\rho_1) \\
q1 & q(1-1) & (1-q)0 & (1-q)(1-0)
\end{pmatrix}
\end{array}
$$

$$
\begin{array}{c} \\ = \end{array}
\begin{array}{c}
(X,Y) \\ (0,0) \\ (0,1) \\ (1,0) \\ (1,1)
\end{array}
\begin{array}{c}
\begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\
\begin{pmatrix}
p & 0 & 0 & 1-p \\
p\lambda_2 & p(1-\lambda_2) & (1-p)\theta_2 & (1-p)(1-\theta_2) \\
q\mu_1 & q(1-\mu_1) & (1-q)\rho_1 & (1-q)(1-\rho_1) \\
q & 0 & 0 & 1-q
\end{pmatrix}
\end{array}
$$

Paneme tähele, et $S_1 := \{(0,0),(1,1)\}$ moodustavad kaasnevate seisundite klassi. Eelduse kohaselt $\pi(1,0) = \pi(0,1) = 0$, millest $(X_1, Y_1) \in S_1$. Järelikult lause 1.1 põhjal jääb ahel kogu aeg seisundite klassi S_1. Klassis S_1 aga kehtibki alati $X_i = Y_i$.

Kuna ahel ei välju klassist S_1, siis võime sellist ahelat vaadata ka kui Markovi ahelat seisundite hulgal S_1 üleminekumaatriksiga:

$$
\begin{array}{c}
(X,Y) \\ (0,0) \\ (1,1)
\end{array}
\begin{array}{c}
\begin{array}{cc} (0,0) & (1,1) \end{array} \\
\begin{pmatrix}
p & 1-p \\
q & 1-p
\end{pmatrix}
\end{array}
$$

Sellise üleminekumaatriksi korral on statsionaarseks algjaotuseks

$$\pi = \left(\frac{q}{q-p+1}, \frac{1-p}{q-p+1} \right),$$

sest

$$\pi \begin{pmatrix} p & 1-p \\ q & 1-q \end{pmatrix} = \pi.$$

Järelikult statsionaarse algjaotuse π korral peab kehtima

$$\pi(0,0) = \frac{q}{q-p+1},$$
$$\pi(1,1) = \frac{1-p}{q-p+1}.$$

□

Lause 3.8. *Kui* $q = 1-p$, $\lambda_2 = \mu_1 = 0$ *ja* $\pi(0,0) = \pi(1,1) = 0$ *siis* $X_t = 0$ *parajasti siis, kui* $Y_t = 1$ *ning vastupidi iga* $t \in \mathbb{N}$ *korral.*

Tõestus. Parameetrite λ_1 ja μ_2 väärtuste põhjal saame leida ka θ_1 ja ρ_1 väärtused:

$$\theta_2 = \frac{q-p\lambda_2}{1-p} = \frac{q-p0}{q} = 1,$$
$$\rho_1 = \frac{p-q\mu_1}{1-q} = \frac{p-q0}{1-(1-p)} = 1$$

Seega üleminekumaatriks P võtab kuju

$$P = \begin{array}{c} (X,Y) \\ (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \end{array} \begin{array}{c} \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\ \begin{pmatrix} p\lambda_1 & p(1-\lambda_1) & (1-p)\theta_1 & (1-p)(1-\theta_1) \\ p\lambda_2 & p(1-\lambda_2) & (1-p)\theta_2 & (1-p)(1-\theta_2) \\ q\mu_1 & q(1-\mu_1) & (1-q)\rho_1 & (1-q)(1-\rho_1) \\ q\mu_2 & q(1-\mu_2) & (1-q)\rho_2 & (1-q)(1-\rho_2) \end{pmatrix} \end{array}$$

$$= \begin{array}{c} (X,Y) \\ (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \end{array} \begin{array}{c} \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\ \begin{pmatrix} p\lambda_1 & p(1-\lambda_1) & (1-p)\theta_1 & (1-p)(1-\theta_1) \\ p0 & p(1-0) & (1-p)1 & (1-p)(1-1) \\ q0 & (1-p)(1-0) & (1-(1-p))1 & (1-q)(1-1) \\ q\mu_2 & q(1-\mu_2) & (1-q)\rho_2 & (1-q)(1-\rho_2) \end{pmatrix} \end{array}$$

$$
= \begin{array}{c} (X,Y) \\ (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \end{array}
\begin{array}{c}
\begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\
\begin{pmatrix}
p\lambda_1 & p(1-\lambda_1) & (1-p)\theta_1 & (1-p)(1-\theta_1) \\
0 & p & 1-p & 0 \\
0 & 1-p & p & 0) \\
q\mu_2 & q(1-\mu_2) & (1-q)\rho_2 & (1-q)(1-\rho_2)
\end{pmatrix}
\end{array}
$$

Analoogselt eelmise tõestusega paneme tähele, ahel jääb kaasnevate seisundite klassi $S_1 := \{(0,1),(0,1)\}$, kuna $\pi(0,0) = \pi(1,1) = 0$. Seega $(X_1, Y_1) \in S_1$, kus kehtibki vajalik tingimus, et $X_t = 0$ parajasti siis kui $Y_t = 1$ ja vastupidi.

Võime jällegi vaadata antud ahelat ka kui Markovi ahelat seisundite hulgal S_1 üleminekumaatriksiga,

$$
\begin{array}{c} (X,Y) \\ (0,1) \\ (1,0) \end{array}
\begin{array}{c}
\begin{array}{cc} (0,1) & (1,0) \end{array} \\
\begin{pmatrix}
p & 1-p \\
q & 1-p
\end{pmatrix}
\end{array}
$$

Sellise üleminekumaatriksi korral on statsionaarseks algjaotuseks $\left(\frac{1}{2}, \frac{1}{2}\right)$, sest

$$
\left(\frac{1}{2}, \frac{1}{2}\right) \begin{pmatrix} p & p-1 \\ 1-p & p \end{pmatrix} = \left(\frac{1}{2}, \frac{1}{2}\right).
$$

Järelikult statsionaarse algjaotuse π korral peab kehtima

$$
\begin{aligned}
\pi(0,1) &= \frac{1}{2}, \\
\pi(1,0) &= \frac{1}{2}.
\end{aligned}
$$

□

Niisiis lause 3.6 eeldustel on jadad $\{X_n\}$ ja $\{Y_n\}$ sõltumatud, samas lausete 3.7 ja 3.8 eeldustel on jadad $\{X_n\}$ ja $\{Y_n\}$ maksimaalselt sõltumatuvad, kuna üks

jada avaldub üheselt teise põhjal. Muude parameetrite väärtuste korral on seega sõltuvusaste vahepealne. Märgime lisaks, et mõisted sõltuvus ja sarnasus ei lange üldjuhul kokku (maksimaalselt sarnane ehk täpselt samad on jadad $\{X_n\}$ ja $\{Y_n\}$ ainult lause 3.7 tingimustel).

Seega selline mudel võimaldab meil fikseeritud p ja q korral genereerida Markovi ahelate paare $\{X_n\}$ ja $\{Y_n\}$ (ja kus mõlemad ahelad sama algjaotuse ning üleminekumaatrikiga), kuid muutes parameetreid μ_i, λ_i, $i = 1, 2$, saame muuta nende omavahelist sõltuvust.

3.2 Simulatsioonid

Järgnevas uurime simulatsioonide abil jadade optimaalsete joonduste kauguse ning pikima ühisjada pikkuse käitumist eelpool kirjeldatud Markovi ahelal põhineval mudelil.

Üks eesmärk on kontrollida, kas selle mudeli korral on teatud sõltuvate jadade korral H kasv logaritmiline, nii nagu kehtis ühisest eellasjadast saadud jadade korral (Lember, Matzinger ja Vollmer, 2014). Teine eesmärk on uurida, kas ekstremaalsete joonduste kaugustel põhinevad sarnasusmõõdikud suudavad mingeid sõltuvusastmeid paremini eristada kui pikima ühisjada pikkus.

Simulatsioonide läbi viimisel, iga fikseeritud p ja q korral:

- genereeritakse iga sõltuvusastme jaoks vastavate parameetritega 50 Markovi ahelat $\{X_n, Y_n\}$ pikkusega n, millest seejärel saadi jadad X ja Y, kus $n = 100, 250, 500, 750, 1000, 2000, 3000, 4000, 5000$;
- leiti nelja sarnasusmõõdiku väärtused;
- iga mõõdiku jaoks koostati graafikud, kus on näha sarnasusmõõdu vaatluste keskmise sõltuvus jadade pikkusest n eri sõltuvusastmete puhul. Lisaks

vaatluste keskmisele graafikul ära toodud punktidena ka üksikvaatlused ning +-sümblitega standardhälve;

- iga mõõdiku jaoks eraldi histogrammina tuuakse veel välja vaatlused, kus $n = 5000$.

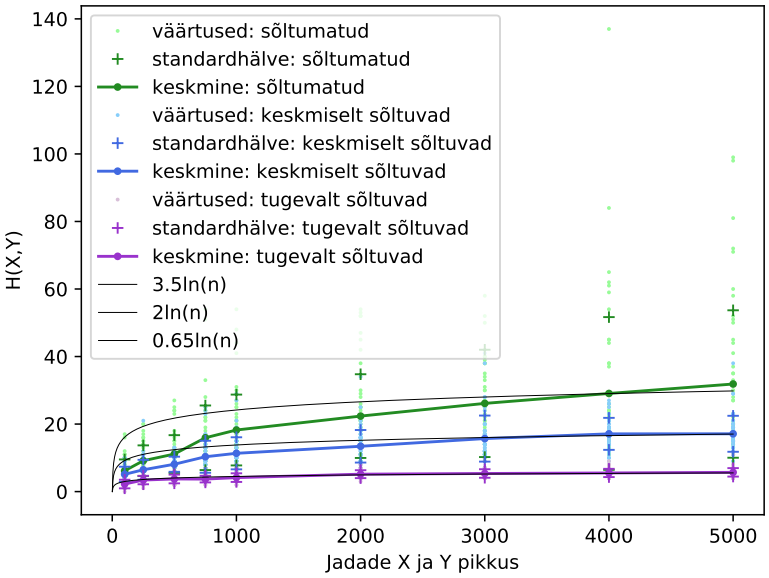
Vaatame kokku kolme erinevat p ja q väärtust.

- Fikseerime $p = q = 0{,}4$. Siis lause 3.5 kohaselt on $\{X_n\}$ ja $\{Y_n\}$ iid jadad ning lause 3.3 kohaselt peab kehtima $\lambda_i, \mu_i \in [0, 1], i = 1, 2$.

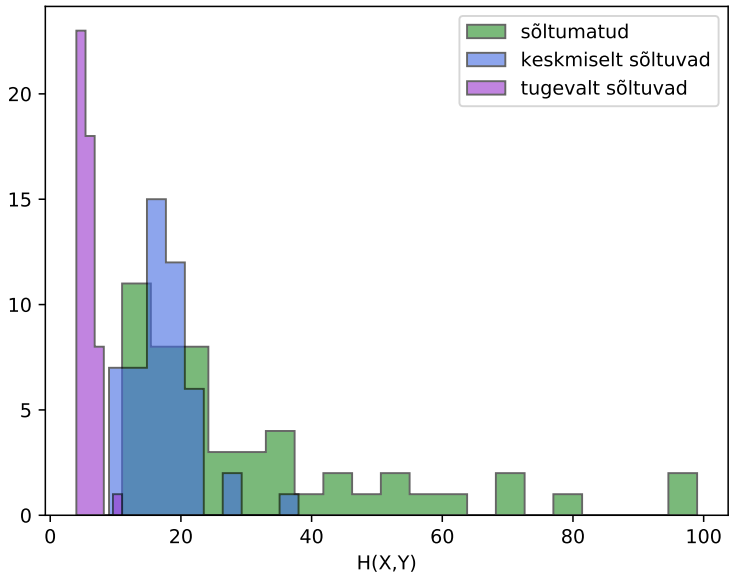
  Vaatame kolme eri sõltuvusastmega jadasid:

λ_1	λ_2	μ_1	μ_2	
0,4	0,4	0,4	0,4	sõltumatud
0,7	0,4	0,4	0,7	keskmine sõltuvusaste
0,95	0,4	0,4	0,95	tugev sõltuvusaste

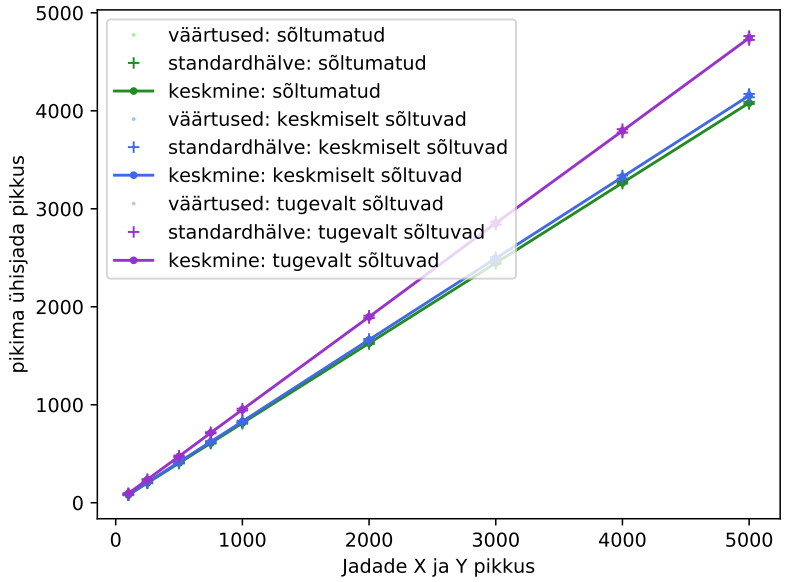
  1. $\lambda_1 = \mu_1 = p = 0{,}4$ ja $\lambda_2 = \mu_2 = q = 0{,}4$ ehk lause 3.6 kohaselt on X ja Y sõltumatud Markovi ahelad.
  2. $\lambda_1 = \mu_2 = 0{,}95$ ja $\lambda_2 = \mu_1 = 0{,}4$. Kui kehtiks $\lambda_1 = \mu_2 = 1$ ja $\pi(1, 0) = \pi(0, 1) = 0$, siis lause 3.7 põhjal on Markovi ahelad X ja Y langeksid kokku ehk sõltuvus on maksimaalne. Seega kuna λ_1 ja μ_2 on erinevad arvust 1 vähe, siis sõltuvusaste on siin suhteliselt suur.
  3. $\lambda_1 = \mu_2 = 0{,}7$ ja $\lambda_2 = \mu_1 = q = 0{,}4$. Siin jääb sõltuvusaste kahe eelmise vahele.



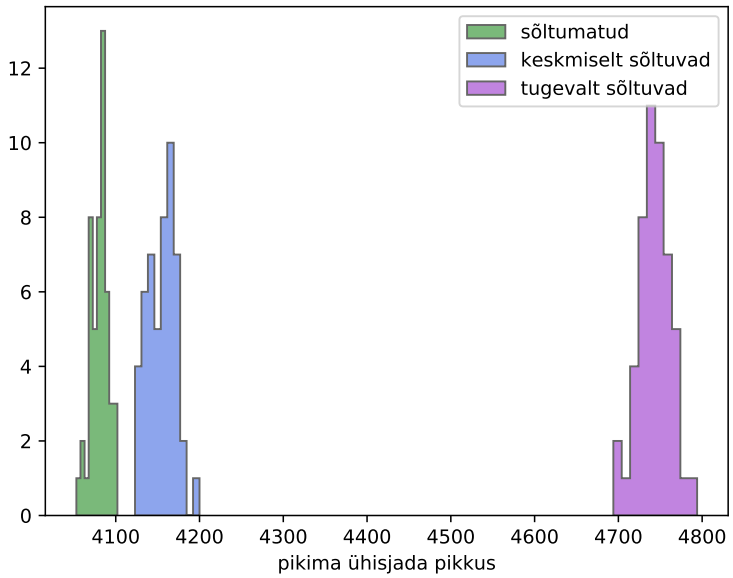
Joonis 1: $p = q = 0{,}4$. $H(X, Y)$.



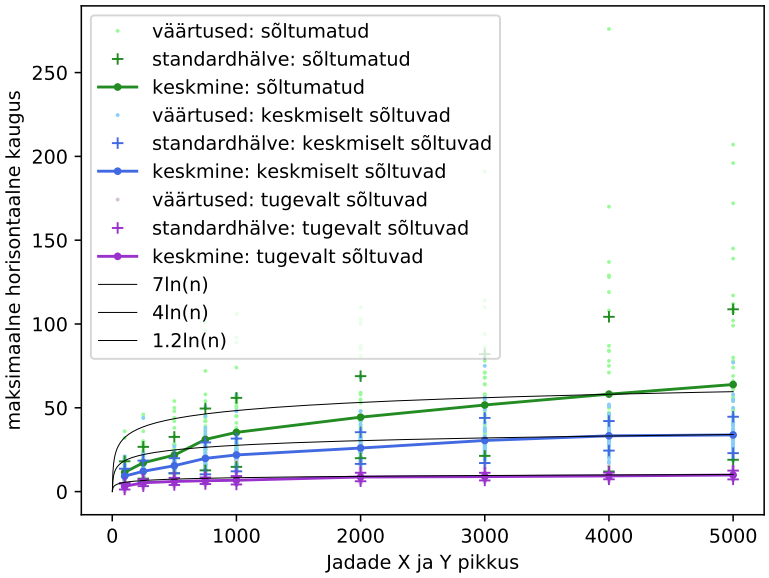
Joonis 2: $p = q = 0{,}4$. $H(X, Y)$.



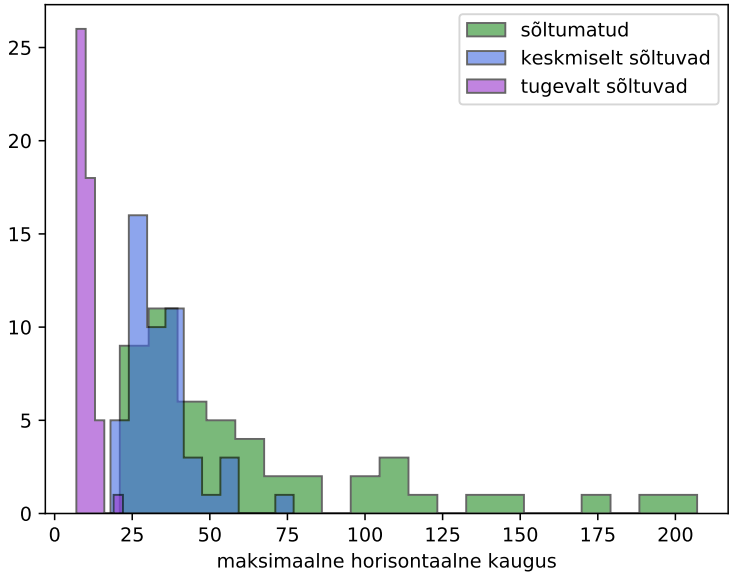
Joonis 3: $p = q = 0{,}4$. Pikima ühisjada pikkus.



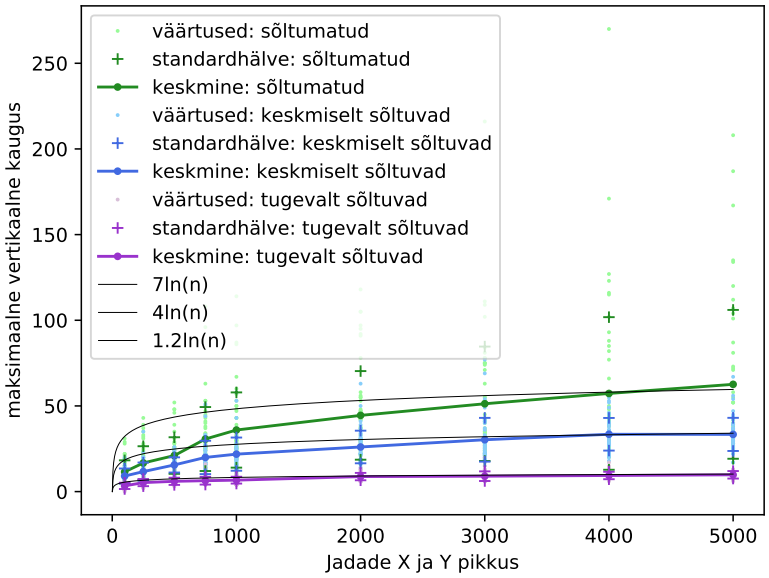
Joonis 4: $p = q = 0{,}4$. Pikim ühisjada pikkus. $n = 5000$



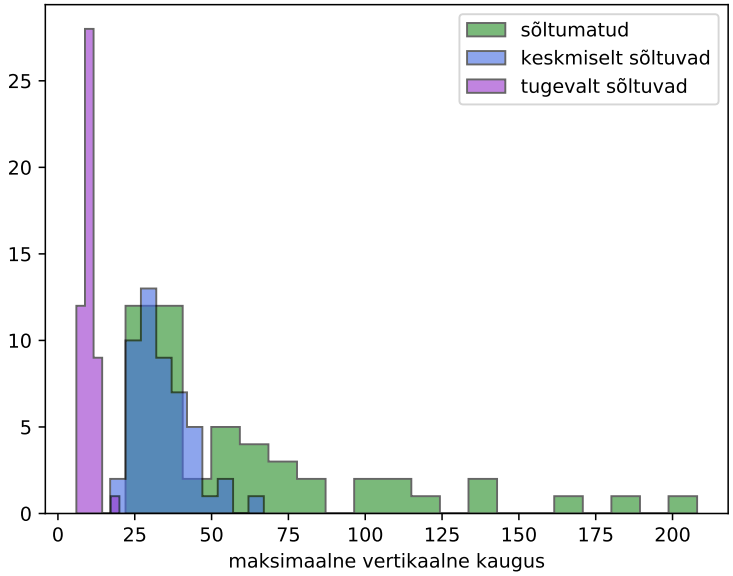
Joonis 5: $p = q = 0{,}4$. Maksimaalne horisontaalne kaugus.



Joonis 6: $p = q = 0{,}4$. Maksimaalne horisontaalne kaugus. $n = 5000$



Joonis 7: $p = q = 0{,}4$. Maksimaalne vertikaalne kaugus.



Joonis 8: $p = q = 0{,}4$. Maksimaalne vertikaalne kaugus. $n = 5000$

- Fikseerime $p = 0{,}4$ ja $q = 0{,}7$.

λ_1	λ_2	μ_1	μ_2	
0,4	0,7	0,4	0,7	sõltumatud
0,7	0,7	0,4	0,85	keskmine sõltuvusaste
0,95	0,7	0,4	0,95	tugev sõltuvusaste

Lause 3.3 kohaselt

$$\lambda_1 \in [0,1], \quad \lambda_2 \in [0{,}25,1], \quad \mu_1 \in \left[\frac{1}{7}, \frac{4}{7}\right], \quad \mu_2 \in \left[\frac{4}{7}, 1\right]$$

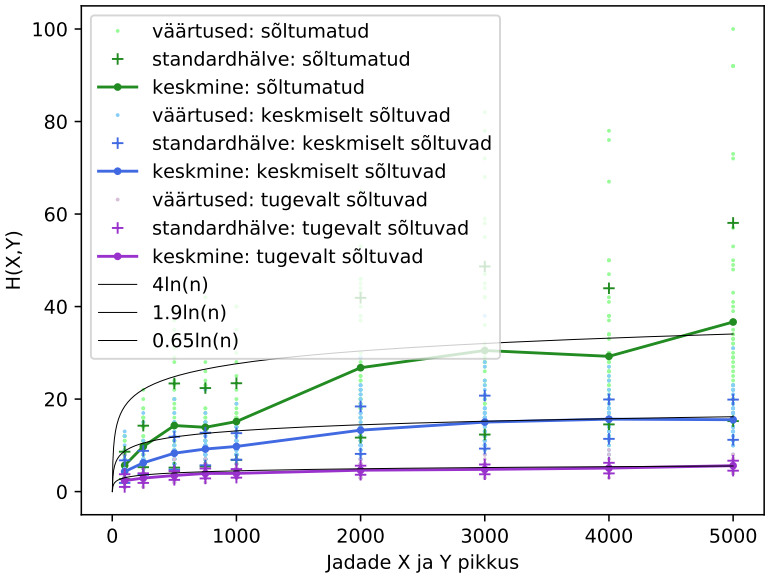
Simulatsioonides vaatame jällegi kolme juhtu

1. $\lambda_1 = \mu_1 = p = 0{,}4$ ja $\lambda_2 = \mu_2 = q = 0{,}7$ ehk lause 3.6 põhjal on X ja Y sõltumatud Markovi ahelad.

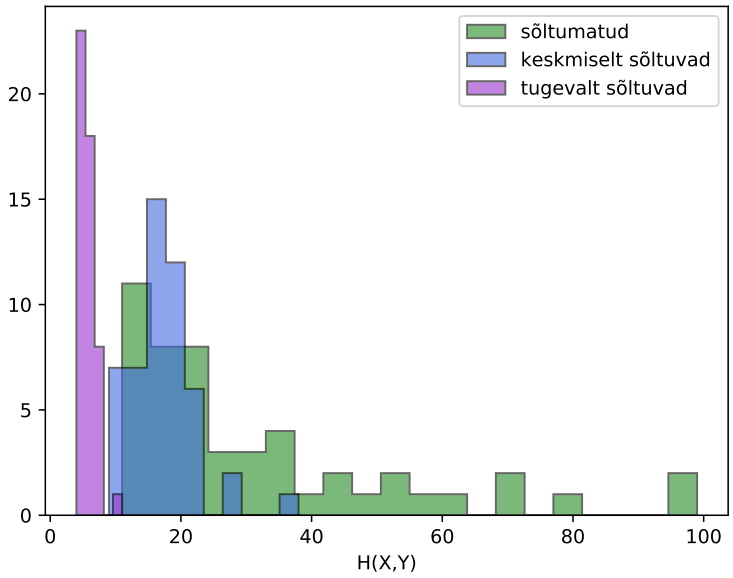
2. $\lambda_1 = \mu_2 = 0{,}95$ ja $\lambda_2 = 0{,}7$ ja $\mu_1 = 0{,}7$.

   Kui kehtiks $\lambda_2 = \mu_1 = 1$ ja $\pi(1,0) = \pi(0,1) = 0$, siis lause 3.7 põhjal on Markovi ahelad X ja Y langevad kokku, mis tähendab, et sõltuvus on maksimaalne. Seega kuna λ_2 ja μ_1 erinevad arvust 1 vähe, siis sõltuvusaste on siin suhteliselt suur.

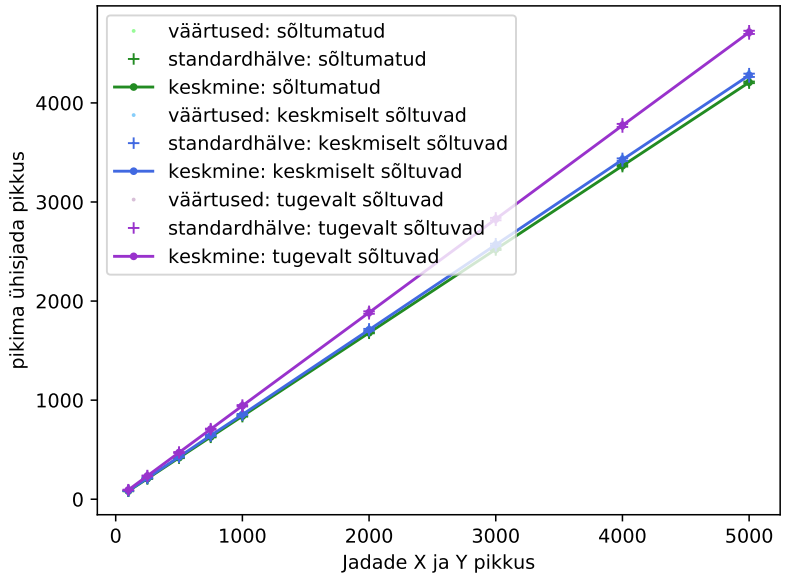
3. $\lambda_1 = 0{,}7$, $\mu_2 = 0{,}85$ ja $\lambda_2 = 0{,}7$ ja $\mu_1 = 0{,}4$. Siin jääb sõltuvusaste kahe eelmise vahele.



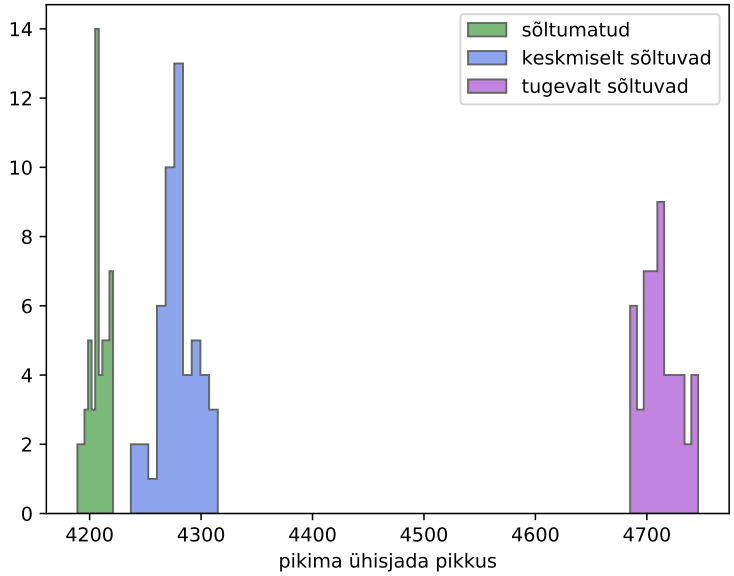
Joonis 9: $p = 0{,}4$, $q = 0{,}7$. $H(X, Y)$.



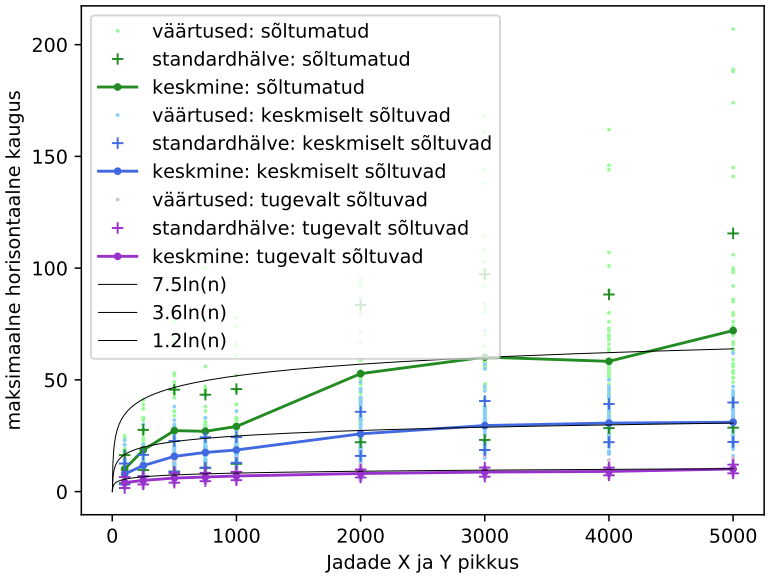
Joonis 10: $p = 0{,}4$, $q = 0{,}7$. $H(X, Y)$.



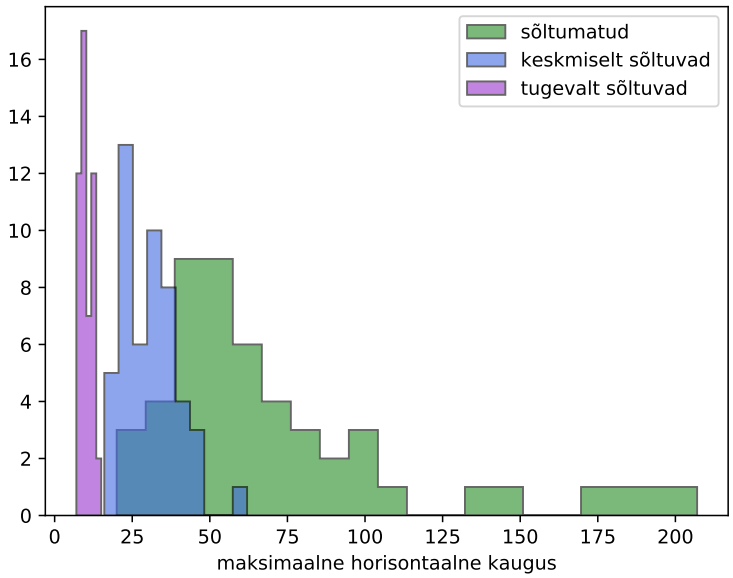
Joonis 11: $p = 0{,}4$, $q = 0{,}7$. Pikima ühisjada pikkus.



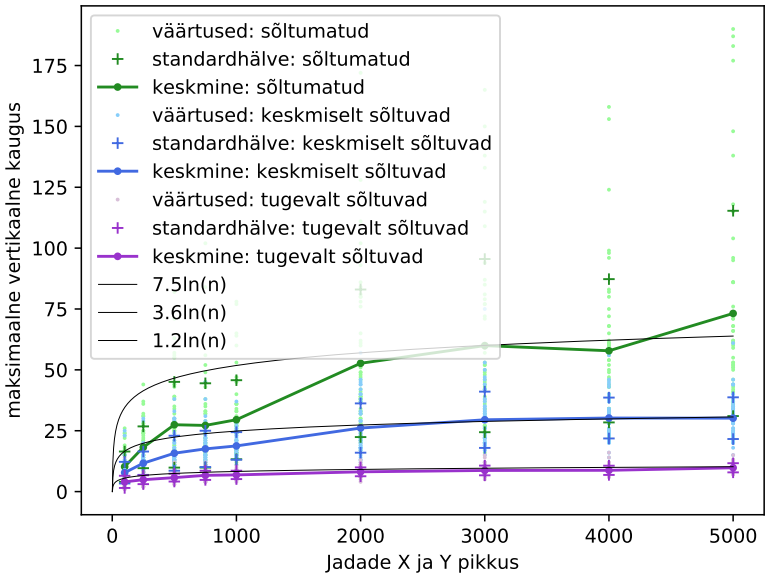
Joonis 12: $p = 0{,}4$, $q = 0{,}7$. Pikim ühisjada pikkus. $n = 5000$



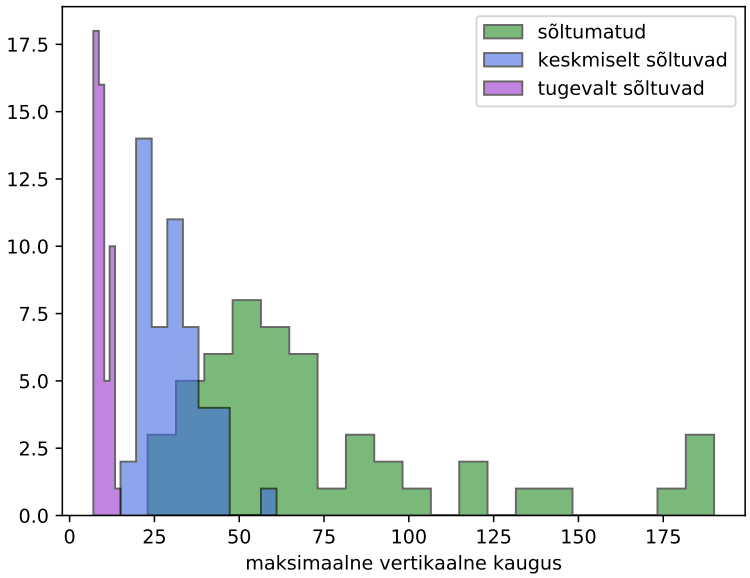
Joonis 13: $p = 0{,}4$, $q = 0{,}7$. Maksimaalne horisontaalne kaugus.



Joonis 14: $p = 0{,}4$, $q = 0{,}7$. Maksimaalne horisontaalne kaugus. $n = 5000$



Joonis 15: $p = 0{,}4$, $q = 0{,}7$. Maksimaalne vertikaalne kaugus.



Joonis 16: $p = q = 0{,}4$. Maksimaalne vertikaalne kaugus. $n = 5000$

- Fikseerime $p = 0{,}6$ ja $q = 0{,}4$.

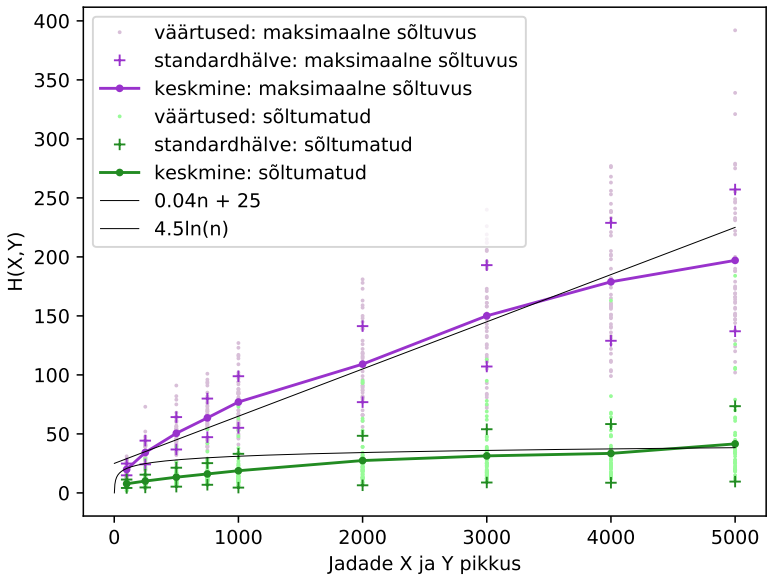
λ_1	λ_2	μ_1	μ_2	
0,6	0,4	0,6	0,4	sõltumatud
1/3	0	0	0	maksimaalne sõltuvus

Lause 3.3 kohaselt

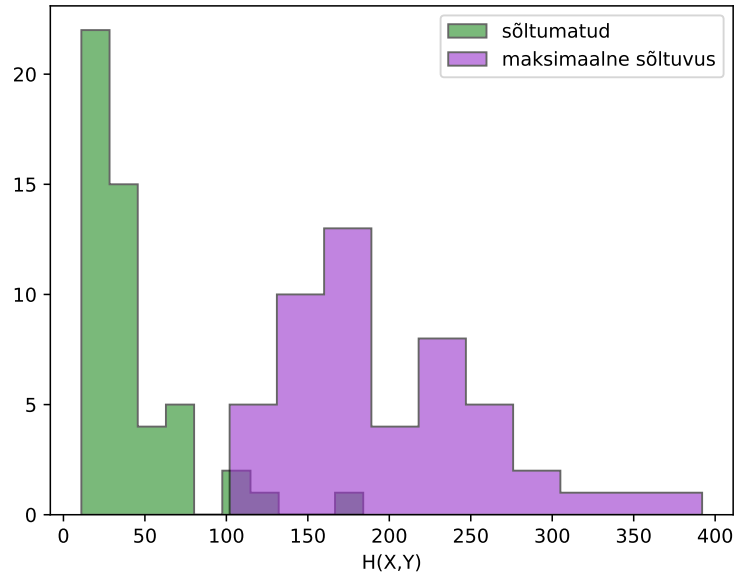
$$\lambda_1 \in \left[\frac{1}{3}, 1\right], \quad \lambda_2 \in \left[0, \frac{2}{3}\right], \quad \mu_1 \in [0,1], \quad \mu_2 \in [0,1]$$

1. $\lambda_1 = \mu_1 = p = 0{,}6$ ja $\lambda_2 = \mu_2 = q = 0{,}4$ ehk lause 3.6 põhjal on X ja Y sõltumatud Markovi ahelad.

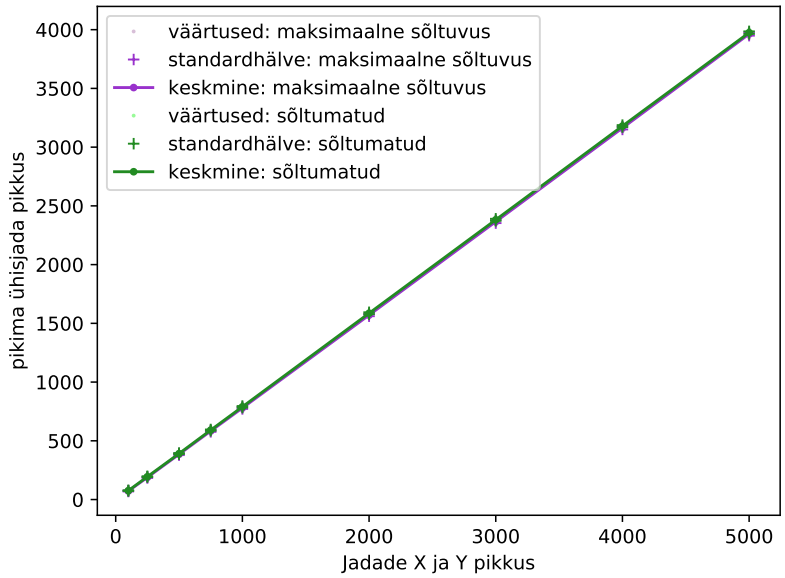
2. $\lambda_2 = \mu_1 = 0$. Kuna $p = 1 - q$ ja kui $\pi(0,0) = \pi(1,1) = 0$, siis lause 3.8 põhjal kehtib Markovi ahelate X ja Y puhul $X_t = 0$ parajasti siis kui $Y_t = 1$ ning vastupidi. See tähendab, et sõltuvus on maksimaalne, kuna üks ahel esitub teise funktsioonina (iga $t \in \mathbb{N}$ korral $Y_t = 1 - X_t$).



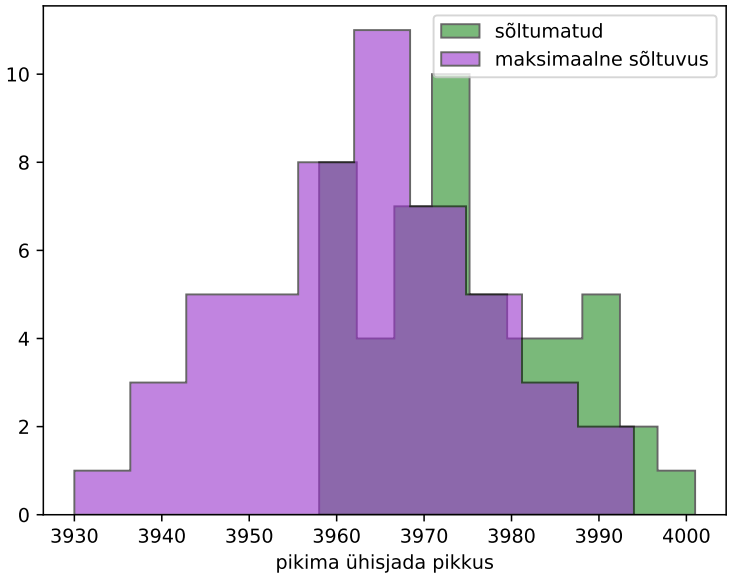
Joonis 17: $p = 0{,}6$, $q = 0{,}4$. $H(X,Y)$.



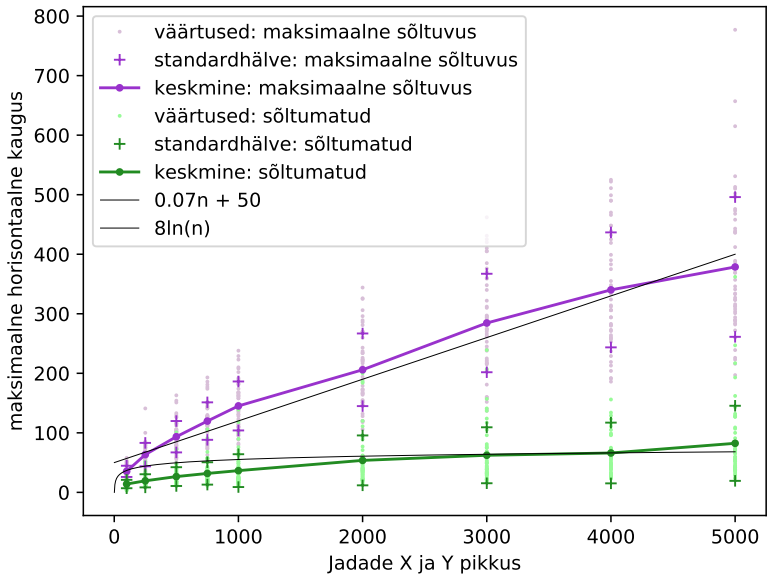
Joonis 18: $p = 0{,}6$, $q = 0{,}4$. $H(X, Y)$.



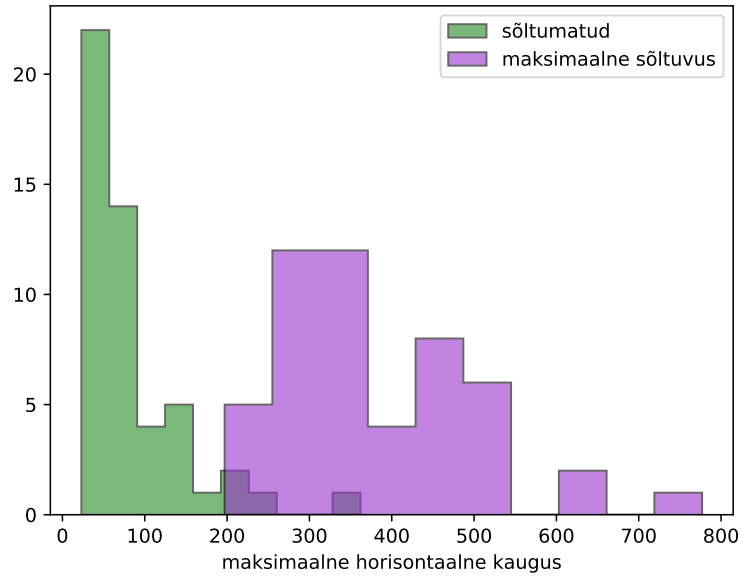
Joonis 19: $p = 0{,}6$, $q = 0{,}4$. Pikima ühisjada pikkus.



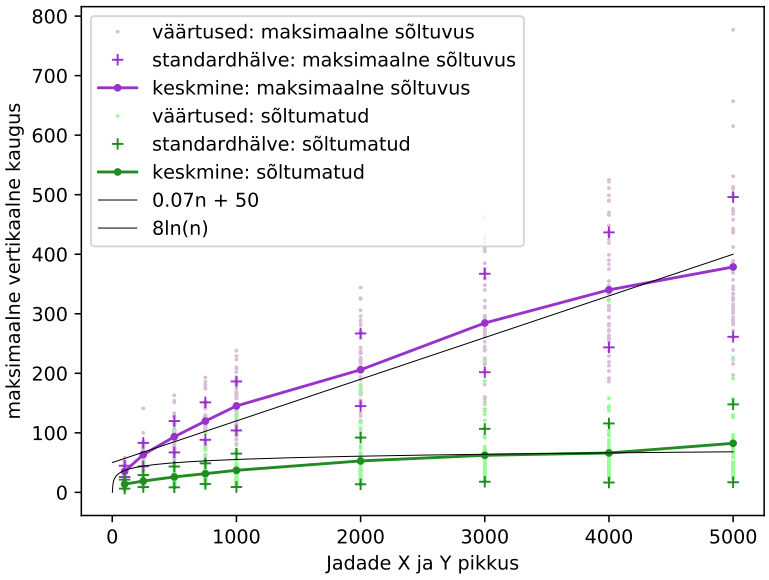
Joonis 20: $p = 0{,}6$, $q = 0{,}4$. Pikim ühisjada pikkus. $n = 5000$



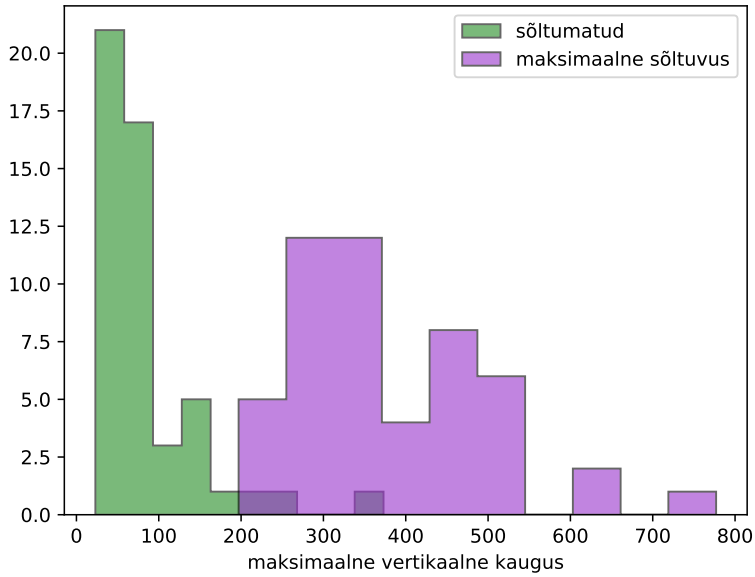
Joonis 21: $p = 0{,}6$, $q = 0{,}4$. Maksimaalne horisontaalne kaugus.



Joonis 22: $p = 0{,}6$, $q = 0{,}4$. Maksimaalne horisontaalne kaugus. $n = 5000$



Joonis 23: $p = 0{,}6$, $q = 0{,}4$. Maksimaalne vertikaalne kaugus.



Joonis 24: $p = 0{,}6$, $q = 0{,}4$. Maksimaalne vertikaalne kaugus. $n = 5000$

Lisa 1. Programmi kood sisaldab funktsioone, mida kasutati simulatsioonide läbiviimisel. Neist peamised on:

- `genereeriMarkoviAhelad(pi, P, seisundid, n)` genereerib Markovi ahelast $\{X_n, Y_n\}$ pikkusega n jadad $\{X_n\}$ ja $\{Y_n\}$ etteantud algjaotuse π ja üleminekumaatriksi P korral.
- `needlemanWunsch(jada1, jada2)` on modifitseeritud Needlemani-Wunsch algoritm (Needleman ja Wunsch, 1970), mis leiab alumise ja ülemise joonduse ning pikima ühisjada pikkuse.
- `maxHor(joondus1, joondus2)`, `maxVer(joondus1, joondus2)`, `h(x,y)` leiavad jadade eelpool defineeritud sarnasusmõõdud.

3.3 Arutelu

Esiteks kehtib hüpotees, et teatud sõltuvate jadade X ja Y puhul on sarnasusmõõdu H ning ka maksimaalse horisontaalse ja vertikaalse kauguse kasv logaritmiline jadade pikkuse suhtes. Vastavatele graafikutele on lisatud ka logaritmfunktsioonide lähendid (joonised 1, 5, 7, 9, 13, 15).

Samuti on kasv logaritmiline sõltumatute jadade puhul, mis lükkab ümber pakutud hüpoteesi, et sõltumatuse korral on kasv lineaarne.

Juhul, kus jadad X ja Y on maksimaalselt sõltuvad jadad, kus $X_t \neq Y_t$, paistab tegemist olevat lineaarse kasvuga (joonised 17, 21, 23).

Ilmse lineaarse kasvuga on kõigil juhtudel ka pikima ühisjada pikkus.

Osutub, et osade erineval määral sõltuvate ahelate sõltuvusastete eristamiseks sobib pikima ühisjada pikkus paremini kui ekstremaalsete joonuste vahe. Joonistel 4 ja 12 eristuvad kõik kolm sõltuvusklassid üheselt, ent joonistel 2, 6, 8 ning 10, 14, 16 saab selgelt eristada ainult kõige tugevamat sõltuvust.

Samas aga sõltumatute ning maksimaalselt sõltuvate jadade, kus $X_i \neq Y_i$, eristamisel jääb pikima ühisjada pikkus teistele mõõdikutele alla. Joonistelt 18, 22, 24 eristuvad kaks sõltuvusklassi oluliselt paremini, kui jooniselt 20.

Lisaks selgus, et kõik ekstremaalsete joonduste vahel põhinevad mõõdikud (H ning maksimaalse horisontaalse ja vertikaalse kauguse) käituvad üldiselt väga sarnaselt. Niisiis pisut keerulisemalt defineeritud Hausdorffi kauguse kasutamisel sarnasusmõõdu H puhul pole nähtavat eelist maksimaalse horisontaalse ja vertikaalse kauguse ees. Pigem siin isegi juhul $p = 0{,}4$, $q = 0{,}7$ eristab maksimaalse horisontaalse ja vertikaalse (joonised 14, 16) sõltuvusklasse pisut paremini kui H (joonis 18).

Kokkuvõte

Töös anti referatiivne ülevaade Markovi ahelatest ning eellasjadast saadud jadadest. Lisaks ka jadade võrdlemise põhimõistetest ning sarnasusmõõdust H.

Samuti kirjeldati Markovi ahelal põhinevat mudelit, mille abil on võimalik genereerida eri sõltuvusastmetega jadasid. Simulatsioonidega abil leiti sellel, et selle mudeli puhul:

- sarnasusmõõdu H ning maksimaalse horisontaalse ja vertikaalse kauguse kasv on logaritmiline teatud sõltuvate jadade ning ka sõltumatute jadade korral;
- pikima ühisjada pikkus eristab teatud erineval määral sõltuvate jadade sõltuvuastmeid oluliselt paremini kui teised vaadeldud mõõdikud.
- sarnasusmõõt H ning maksimaalse horisontaalse ja vertikaalse kauguse eristas paremini omavahel sõltumatuid jadasid ning selliseid maksimaalseid sõltuvaid jadasid, kus $X_t \neq Y_t$.

Lisa 1. Programmi kood

```
import numpy as np
import matplotlib.pyplot as plt
from numpy import linalg

# yleminekumaatriksi leidmine parameetritest
def yleminekumaatriks(p, q, lambda1, lambda2, mu1, mu2):
        P = np.zeros((4, 4))

        # Arvutame ülejäänud 4 parameetrit
        teeta1 = (1-lambda1)*p/(1-p)
        teeta2 = (q-p*lambda2)/(1-p)
        roo1 = (p-q*mu1)/(1-q)
        roo2 = (1-mu2)*q/(1-q)

        # täidame maatriksi:
        P[0][0] = p*lambda1
        P[0][1] = p*(1-lambda1)
        P[0][2] = (1-p)*teeta1
        P[0][3] = (1-p)*(1-teeta1)

        P[1][0] = p*lambda2
        P[1][1] = p*(1-lambda2)
        P[1][2] = (1-p)*teeta2
        P[1][3] = (1-p)*(1-teeta2)

        P[2][0] = q*mu1
        P[2][1] = q*(1-mu1)
```

```
        P[2][2] = (1-q)*roo1
        P[2][3] = (1-q)*(1-roo1)

        P[3][0] = q*mu2
        P[3][1] = q*(1-mu2)
        P[3][2] = (1-q)*roo2
        P[3][3] = (1-q)*(1-roo2)

        # Kui üleminekumaatriksis on negatiivseid
        # väärtusi, siis seame erindi.
        for i in range(4):
                for j in range(4):
                        if P[i][j] < 0:
                                raise RuntimeError('Negatiivne väärtus!')
        return P

# Leiame statsionaarse algjaotus pi, mille korral
# pi.P = pi ehk P^T omaväärtusele 1 vastav omavektor
def statsionaarneAlgjaotus(P):
        Pt = P.transpose()
        vaartused, vektorid = linalg.eig(Pt)
        for i in range(len(vaartused)):
                if (abs(vaartused[i] - 1) < 10**(-7)):
                        v = vektorid[:,i] #omaväärtusele ~1 vastav omavektor
                        k = np.sum(v)
                        # skaleerime omavektori selliseks,
                        # et selle elementide summa oleks 1
                        jaotus = ((1/k)*v).transpose()
                        return jaotus
```

```
def genereeriMarkoviAhelad(pi, P, seisundid, n):
        ahel = []

        #sõnastik, millest saame seisundile vastava indeksi
        seisundiIndeks = {}
        for i in range(len(seisundid)):
                seisundiIndeks[seisundid[i]] = i



        for i in range(n):
                if (i == 0):
                        # ahela esimese seisundi leiame
                        # juhuslikult algjaotusest
                        olevik = seisundid[np.random.choice(
                                                len(seisundid), p=pi)]
                else:
                        #ülejäänud seisundid tulenevad eelnevast
                        si = seisundiIndeks[olevik]
                        jaotus = P[si]
                        olevik = seisundid[np.random.choice(
                                                len(seisundid), p=jaotus)]
                ahel.append(olevik)
        return [x for x,y in ahel], [y for x,y in ahel]

#Needlemani-Wunschi algoritm pikima ühisjada leidmiseks
def needlemanWunsch(jada1, jada2):
        pikkus1 = len(jada1)
        pikkus2 = len(jada2)
```

```
        maatriks = np.zeros((pikkus1+1, pikkus2+1))
        for i in range(1, pikkus1+1):
                for j in range(1, pikkus2+1):
                        a = maatriks[i-1][j]
                        b = maatriks[i-1][j-1] + (jada1[i-1] == jada2[j-1])
                        c = maatriks[i][j-1]
                        maatriks[i][j] = max(a, b, c)
        ylemine = leiaJoondus(maatriks, pikkus1, pikkus2,
                        jada1, jada2, "ylemine")
        alumine = leiaJoondus(maatriks, pikkus1, pikkus2,
                        jada1, jada2, "alumine")
        return ylemine, alumine, int(maatriks[pikkus1][pikkus2])

# Maatriksist joonduse leidmine
def leiaJoondus(maatriks, i, j, jada1, jada2, tyyp):
        joondus = []
        while (i != 0 and j != 0):
                a = maatriks[i-1][j]
                b = maatriks[i-1][j-1] + (jada1[i-1] == jada2[j-1])
                c = maatriks[i][j-1]
                if (tyyp == "alumine"):
                        if (a >= b and a >= c):
                                i = i-1
                        elif (b >=a and b >= c):
                                joondus.append((i,j))
                                i = i-1
                                j = j-1
                        else:
                                j = j-1
```

```
                else:
                        if ( c >= a and c >= b):
                                j = j-1
                        elif (b >= a and b >= c):
                                joondus.append((i,j))
                                i = i-1
                                j = j-1
                        else:
                                i = i-1
        joondus.reverse()
        return joondus


# maksimaalne horisontaalne kaugus
def maxHor(joondus1, joondus2):
    maksimum = 0
    a, b, c, d = 0, 0, 0, 0
    i1 = 0
    i2 = 0
    while(i1 < len(joondus1) and i2 < len(joondus2)):
        x1, y1 = joondus1[i1]
        x2, y2 = joondus2[i2]
        if(y1 == y2):
            dis = abs(x1-x2)
            if maksimum < dis:
                maksimum = dis
                a, b, c, d = x1, y1, x2, y2
            i1+=1
            i2+=1
```

```
        elif(y1 < y2):
            i1+=1
        else:
            i2+=1
    return maksimum, a, b, c, d

# maksimaalne vertikaalne kaugus
def maxVer(joondus1, joondus2):
    maksimum = 0
    a, b, c, d = 0, 0, 0, 0
    i1 = 0
    i2 = 0
    while(i1 < len(joondus1) and i2 < len(joondus2)):
        x1, y1 = joondus1[i1]
        x2, y2 = joondus2[i2]
        if(x1 == x2):
            dis = abs(y1-y2)
            if maksimum < dis:
                maksimum = dis
                a, b, c, d = x1, y1, x2, y2
            i1+=1
            i2+=1
        elif(x1 < x2):
            i1+=1
        else:
            i2+=1
    return maksimum, a, b, c, d

# d on defineeritud kui maksimumkaugus
```

```
def d(a, b):
        (x1, y1) = a
        (x2, y2) = b
        return max(abs(x1-x2), abs(y1-y2))

# Hausdorffi kaugus - kuna tegemist on lõplike hulkadega,
# siis võime sup/infi asemel kasutada max/min
def h(x, y):
        supinfxy = max(map(lambda v : min(map(lambda u : d(u,v), y)), x))
        supinfyx = max(map(lambda u : min(map(lambda v : d(u,v), x)), y))
        return max(supinfxy, supinfyx)



P = yleminekumaatriks(p=0.5, q=0.5, lambda1=0.5, lambda2=0.5, mu1=0.5, mu2=0.5)
pi = statsionaarneAlgjaotus(P)
jada1, jada2 = genereeriMarkoviAhelad(pi, P, [(0,0), (0,1), (1,0), (1,1)], 200)
alumine, ylemine, lcs = needlemanWunsch(jada1, jada2)
maxh, x1h, y1h, x2h, y2h = maxHor(ylemine, alumine)
maxv, x1v, y1v, x2v, y2v = maxVer(ylemine, alumine)
hausdorff = h(ylemine, alumine)
print("maksimaalne horisontaalne kaugus: " + str(maxh))
print("maksimaalne vertikaalne kaugus: " + str(maxv))
print("Hausdorffi kaugus: " + str(hausdorff))
print("pikima ühisjada pikkus: " + str(lcs))
plt.plot([ i for i, j in ylemine ], [ j for i, j in ylemine],
        label = "ülemine joondus")
plt.plot([ i for i, j in alumine ], [ j for i, j in alumine],
        label = "alumine joondus")
```

```
plt.plot([x1h, x2h], [y1h, y2h],
        label = "maksimaalne horisontaalne kaugus")
plt.plot([x1v, x2v], [y1v, y2v],
        label = "maksimaalne vertikaalne kaugus")
plt.axis([0, len(jada1), 0, len(jada2)])
plt.legend(loc=0)
plt.show()
```

Kasutatud allikad

Lember, Jüri, Heinrich Matzinger, Joonas Sova ja Fabio Zucca (2018). "Lower bounds for moments of global scores of pairwise Markov chains". *Stochastic Processes and their Applications* 128, lk. 1678–1710.

Lember, Jüri, Heinrich Matzinger ja Anna Vollmer (2014). "Optimal alignments of longest common subsequences and their path properties". *Journal of the American Statistical Association* 52, lk. 1292–1343.

Needleman, Saul B. ja Christian D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". *Journal of Molecular Biology* 48, lk. 443–453.

Pärna, Kalev (2013). *Tõenäosusteooria algkursus*. Tartu Ülikooli kirjastus.

Sova, Joonas (2013). "Sõltuvate jadade mudel". Bakalaureusetöö. Tartu Ülikool.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kati Iher,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Juhuslike jadade võrdlemine“, mille juhendaja on Jüri Lember, 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.

Kati Iher
14.06.2021