Tartu Ülikool

Matemaatika-informaatikateaduskond

Matemaatilise Statistika Instituut

Joonas Sova

Homoloogsete jadade sõltuvusmõõdud

Magistritöö

Juhendaja: Jüri Lember

Tartu 2015

Homoloogsete jadade sõltuvusmõõdud

LÜHIKOKKUVÕTE

Käesolevas töös uuritakse erinevaid meetodeid juhuslike jadade sõltuvuse mõõtmiseks. Kahe jada sõltuvus võib tähendada väga erinevaid asju. Antud töös uuritakse *homoloogseid* jadasid – see tähendab, et jadad on tekkinud mingist ühisest eellasjadast.

Töö alguses defineerime formaalselt meie sõltuvate jadade mudeli. Selle mudeli kohaselt tekivad kaks üldjuhul sõltuvat jada ühisest eellasjadast, elementide muutmise ja eemaldamise teel.

Me tutvustame kolme tüüpi jadade sõltuvusmõõte: sagedustabelipõhised statistikud, pikima ühisjada pikkus (LLCS) ning ekstremaalsete joonduste vaheline Hausdorffi kaugus (HD).

Me genereerime erinevatest sõltuvusklassidest sõltuvate jadade paare ning võrdleme karpdiagrammide abil ülalmainitud sõltuvusmõõte. Lisaks uurime simulatsioonide abil ka LLCS-i piirjaotust.

Tuletame McDiarmidi ja Efron-Stein'i võrratuste abil ülemise tõkke sõltuvate jadade LLCS-i dispersioonile. Nagu sõltumatute jadade korralgi, on see tõke lineaarset järku. Leiame ülemise tõkke ka teistele LLCS-i tsentreeritud absoluutsetele momentidele. Momendile järguga k saame ülemise tõkke järguga $\frac{k}{2}$ $(k > 0)$.

Uurime ka HD tsentreerimata momente. Varasemast artiklist teadaolevat tulemust kasutades näitame, et teatud tingimustel on HD k-ndat järku momendil EH_m ülemine tõke järguga $[\ln(m)]^k$ $(k > 0)$. Siin m on eellasjada pikkus. Seega (teatud tingimustel) on HD keskväärtusel ja standardhälbel logaritmilist järku ülemine tõke. Uurime HD keskväärtuse ja standardhälbe kasvu ka simulatsionide abil.

Märksõnad: pikima ühisjada pikkus, jadade joondus, ekstremaalsed joondused, optimaalsed joondused, homoloogsed jadad.

Measures of relatedness for homologous random sequences

ABSTRACT

The purpose of this thesis is to study different methods to measure the relatedness of two random sequences. Two sequences can be related in very different ways. In this thesis we study *homologuous* sequences – this means that the sequences are related through some common ancestor.

We start by formally defining our model of relatedness. In this model, two generally related sequences are obtained by changing and removing some of the elements of a common ancestor sequence. Next, we introduce three kinds of methods to measure the relatedness of the sequences: commonly known contingency table based statistics, length of longest common subsequence (LLCS) and Hausdorff distance between extremal alingments (HD).

We simulate related sequences from different dependance classes and use box plots to compare the aforementioned methods. The limit distribution of LLCS is also studied with simulations.

The variance of the LLCS of the related sequences is studied theoretically. We prove that McDiarmid and Efron-Stein inequalities can be used to obtain an upper bound to the variance. As with unrelated sequences, this upper bound is of linear order. We generalize this approach, and also find an upper bound to all centered absolute moments of LLCS. For the k-th moment, we obtain an upper bound of order $\frac{k}{2}$ $(k > 0)$.

We also study the uncentered moments of HD. Using a theorem known from a previous article, we show that under certain conditions the k-th moment EH_m of HD has an upper bound of order $[\ln(m)]^k$ $(k > 0)$. Here m is the length of the ancestor sequence. Therefore (under certain conditions) the expected value and standard deviation of HD has an upper bound with logarithmic order. We further study the growth of the expected value and standard deviation using simulations.

Keywords: longest common subsequence, sequence alignments, extremal alignments, optimal alignments, homologous sequences.

Sisukord

Sissejuhatus

Käesolevas töös uuritakse juhuslike jadade sõltuvusmõõte. Üldiselt võib kahe jada sõltuvus tähendada väga erinevaid asju. Antud töös keskendume jadade *homoloogsusest* tingitud sõltuvusele. Jadade homoloogsus tähendab seda, et jadad on tekkinud mingist ühisest eellasjadast.

Kohe töö alguses defineeritakse formaalselt kahe jada sõltuvust kirjeldav mudel. Mudeli kohaselt tekivad kaks üldjuhul sõltuvat *iid* jada ühisest *iid* eellasjadast osade liikmete muutmise ja eemaldamise teel. Selline sõltuvusstruktuur on oluline praktika seisukohast. Taolise sõltuvusstruktuuriga jadade sõltuvuse mõõtmisega tegeldakse bioloogias (DNA- ja RNA-jadad), kõnetuvastuses, lingvistikas ja ka mujal.

Töö on jagatud kolmeks peatükiks. Esimeses peatükis tutvustame erinevaid meetodeid homoloogsete jadade sõltuvuse mõõtmiseks. Esmalt vaatame üle sellised tuntud sagedustabelipõhised statistikud nagu χ^2-statistik ja empiiriline vastastikune informatsioon. Seejärel tutvume pikima ühisjada pikkusega. Tutvustame ka ühte uuemat ja vähemtuntud statistikut nimega "ekstremaalsete joonduste vaheline kaugus".

Kui esimeses peatükis on nii teooriat kui ka simulatsioone, siis teine peatükk on puhtteoreetiline. Siin uurime, kuidas leida pikima ühisjada pikkuse dispersioonile ülemist tõket. Kolmandas peatükis tegeleme ekstremaalsete joonduse vahelise kauguse dispersiooni ning tsentreerimata absoluutsete momentide hindamisega.

Seega esimene peatükk tegeleb üldise sõltuvuse mõõtmise probleemiga ning teine ja kolmas peatükk tegelevad juba spetsiifilisemate momentide hindamisega seotud küsimustega.

Töös on teoreemide, lausete, lemmade ja järelduste kujul esitatud mitmeid tulemusi. Osades tulemustes on toodud viide vastavale allikale – see tähendab, et selle tulemuse tõestus (kui see on ära toodud) on ka võetud samast allikast. Juhul, kui tulemuses viidet ei ole, on vastav tõestus autori enda välja mõeldud. Siiski paljud sellised tulemused on tegelikult juba varem teada. Esimeses peatükis ei ole ühtegi tulemust, mis ei oleks varem teada – samuti mitte teise peatüki sõltumatuid jadasid puudutavas osas. Teises peatükis tuletatud dispersioonitõkked sõltuvate jadade mudeli kontekstis on aga uued ja ehk käesoleva töö oluliseim panus teadusele. Kolmanda peatüki tulemused kujutavad endas aga ühe varasemalt teadaoleva teoreemi loomulikku edasiarendust.

Peatükk 1

Kuidas mõõta jadade sõltuvust?

1.1 Sõltuvate jadade mudel

Olgu $\mathcal{A}$ lõplik tähestik ja $X_1, X_2, ..., Y_1, Y_2, ...$ juhuslikud suurused, mis võtavad väärtuseid sellest tähestikust. Defineerime

$$X^\infty := X_1, X_2, ..., \quad Y^\infty := Y_1, Y_2, ...$$

Meie eesmärk on defineerida mudel, mille kohaselt tekivad jadad X^∞, Y^∞ ühisest eellasjadast osade liikmete muutmise ja eemaldamise teel.

Olgu $Z_1, Z_2, ...$ *iid* jada (eellasjada), mille liikmed võtavad väärtusi tähestikust $\mathcal{A}$. Lisaks olgu $\xi_1, \xi_2, ..., \eta_1, \eta_2, ...$ reaalarvulised *iid* juhuslikud suurused (näiteks normaaljaotusest), mis on sõltumatud jadast $Z_1, Z_2, ...$ Olgu $h : \mathcal{A} \times \mathbb{R} \to \mathcal{A}$ mingi funktsioon. Iga $i \in \mathbb{N}$ korral saame tähest Z_i kaks muteerunud tähte $M_i^x := h(Z_i, \xi_i)$ ja $M_i^y := h(Z_i, \eta_i)$. Jadad $M_1^x, M_2^x, ...$ ja $M_1^y, M_2^y, ...$ on mõlemad *iid*, kusjuures juhuslikud suurused M_1^x, M_1^y on sama jaotusega. Jadad $M_1^x, M_2^x, ...$ ja $M_1^y, M_2^y, ...$ ei ole üldiselt sõltumatud, aga paarid $(M_1^x, M_1^y), (M_2^x, M_2^y), ...$ on. Tähega Q tähistame mutatsioone kirjeldavat üleminekumaatriksit:

$$Q(a, b) := \mathbf{P}(h(a, \xi_1) = b) = \mathbf{P}(h(a, \eta_1) = b) \quad \forall a, b \in \mathcal{A}.$$

Siin $Q(a, b)$ tähistab maatriksi Q tähele a vastava rea ja tähele b vastava veeru elementi.

Jada X^∞ saamiseks eemaldatakse mutatsioonide jadast $M_1^x, M_2^x, ...$ osa elemente. Eemaldamine toimub *iid* Bernoulli jaotusega juhuslike suuruste $D_1^x, D_2^x, ...$ abil, $D_1^x \sim Be(p)$. Kui $D_i^x = 1$, siis täht M_i^x jääb alles, vastasel juhul kaob; parameeter $p > 0$ on tähe allesjäämise tõenäosus. Analoogiliselt saadakse jada Y^∞ jadast $M_1^y, M_2^y, ...$ osade tähtede eemaldamise teel. Eemaldamine toimub jälle *iid* Bernoulli jaotusega juhuslike suuruste $D_1^y, D_2^y, ...$ abil, $D_1^y \sim Be(p)$. Eeldame, et juhuslikud suurused

$$Z_1, Z_2, ..., \xi_1, \xi_2, ..., \eta_1, \eta_2, ..., D_1^x, D_2^x, ..., D_1^y, D_2^y, ...$$

on kõik omavahel sõltumatud.

Formaalselt saame juhuslikud suurused X_i defineerida järgmiselt: $X_i := M_j^x$, kus j rahuldab tingimust

$$\sum_{k=1}^{j} D_k^x = i \text{ ja } D_j^x = 1. \tag{1.1}$$

Kui kehtib (1.1), siis ütleme, et element i on elemendi j järglane ja et element j on elemendi i eellane, ning kirjutame $a^x(i) = j$. Seega $X_i = M^x_{a^x(i)}$. Juhuslikud suurused Y_i saame defineerida analoogiliselt: $Y_i := M^y_j$, kus j rahuldab tingimust

$$\sum_{k=1}^{j} D^y_k = i \text{ ja } D^y_j = 1. \tag{1.2}$$

Kui kehtib (1.2), siis ütleme, et element i on elemendi j järglane ja et element j on elemendi i eellane, ning kirjutame $a^y(i) = j$. Seega $Y_i = M^y_{a^y(i)}$. Kui X- ja Y-jadade vastavatel elementidel i', j' on sama eellane ehk kui $a^x(i') = a^y(j')$, ütleme, et nad on sugulased ja et paar (i', j') on sugulaspaar.

Olgu näiteks tähestikuks $\mathcal{A} = \{\text{A, C, T, G}\}$ (DNA nukleotiidid) ja

i	1	2	3	4	5
M^x_i	A	C	A	G	T
D^x_i	1	0	1	1	0
M^y_i	A	T	A	G	A
D^y_i	1	1	1	0	1

.

Siis $X_1X_2X_3 = \text{AAG}$ ja $Y_1Y_2Y_3Y_4 = \text{ATAA}$. X-jada elementide $1, 2, 3$ eellased on vastavalt $1, 3, 4$ ja Y-jada elementide $1, 2, 3, 4$ eellased on vastavalt $1, 2, 3, 5$. X-jada elemendid $1, 2$ on sugulased vastavalt Y-jada elementidega $1, 3$.

Esitame mõned tulemused toodud mudeli kohta. Tulemuste tõestused leiate autori bakalaureusetööst [1].

1. Jadad X^∞ ja Y^∞ on mõlemad *iid*, kusjuures juhuslikud suurused X_1, Y_1 on mõlemad sama jaotusega kui M^x_1 (või M^y_1).

2. Kui $p < 1$, siis üldjuhul juhuslikud suurused X_i, Y_j, kus $i, j \in \mathbb{N}$, on sõltuvad, kuid kui maatriksi Q read on võrdsed, siis on nad sõltumatud. Seega meie sõltuvate jadade mudel sisaldab endas ka sõltumatuse juhtu.

3. Kui $p < 1$, siis

$$\mathbf{P}[a^x(k) = a^y(k)] \xrightarrow{k} 0.$$

   Seega tõenäosus, et juhuslikud suurused X_k, Y_k on sugulased kahaneb nulli protsessis $k \to \infty$, kui $p < 1$. See annab intuitiivse põhjenduse järgmisele punktile.

4. Kui $p < 1$, siis kõikide $a, b \in \mathcal{A}$ korral

$$\mathbf{P}(X_k = a, Y_k = b) \xrightarrow{k} \mathbf{P}(X_1 = a)\mathbf{P}(Y_1 = b).$$

   Sisuliselt tähendab see, et kui k on piisavalt suur, siis juhuslikud suurused X_k, Y_k on praktiliselt sõltumatud (kui $p < 1$).

5. Kui $p < 1$ ja jadad X^∞, Y^∞ on sõltuvad, siis kahedimensionaalne protsess $\{(X_i, Y_i)\}_{i=1}^\infty$ pole statsionaarne.

Vastavalt punktile 1 ülaltoodud loetelus juhuslikud suurused $X_1, X_2, ..., Y_1, Y_2, ...$ on kõik sama marginaaljaotusega; vastavat jaotust tähistame tähega P. Teisisõnu $X_i \sim P$ ja $Y_i \sim P$ iga $i \in \mathbb{N}$ korral.

Defineerime

$$X^k := X_1, ..., X_k, \quad Y^k := Y_1, ..., Y_k, \quad k = 1, 2, ...$$

Peatüki keskne küsimus on järgmine: kuidas mõõta jadade X^n ja Y^m, $n, m \in \mathbb{N}$, sõltuvust?

Märgime veel, et kasutame kogu töö vältel miinimumi ja maksimumi tähistamiseks vastavalt sümboleid $\wedge$ ja $\vee$.

1.2 Sissejuhatus simulatsioonidesse

Töös teostame ka mitmeid simulatsioone, kus genereeritakse eelnevalt kirjeldatud mudelile vastavaid jadapaare. Simulatsioonides alati $|\mathcal{A}| = 4$, st tähestik koosneb neljast tähest. Genereeritavad jadapaarid võib jagada järgmisse kolme klassi.

1. Sõltumatud jadad pikkusega n. Genereeritakse sõltumatud jadad X^n, Y^n, mille elementide marginaaljaotus P ühtlane. Erijuhul võib P olla ka ebaühtlane, kuid sellisel juhul on see eraldi välja toodud.

2. Sõltuvad jadad pikkusega n. Genereeritakse üldjuhul sõltuvad jadad X^n, Y^n, mille elementide marginaaljaotus P on ühtlane. Tähe allesjäämise tõenäosuseks võtame $p = 0.95$. Üleminekumaatriksiks Q võtame

$$\begin{pmatrix} 1-3\epsilon & \epsilon & \epsilon & \epsilon \\ \epsilon & 1-3\epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & 1-3\epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & 1-3\epsilon \end{pmatrix}. \tag{1.3}$$

   Siin $0 \leq \epsilon \leq \frac{1}{4}$ on mutatsioone kirjeldav parameeter, mis täpsustatakse konkreetsete simulatsioonide juures. Märgime, et kui Q on ülaltoodud kujul ja juhuslike suuruste $Z_1, Z_2, ...$ marginaaljaotus on ühtlane, siis on ühtlased ka juhuslike suuruste $M_1^x, M_2^x, ..., M_1^y, M_2^y, ...$ marginaaljaotus ja jaotus P (vt punkt 1 jaotise 1.1 loetelus). Veel paneme tähele, et kui võtta $\epsilon = \frac{1}{4}$, siis jadad X^n, Y^n on tegelikult sõltumatud ehk vastavad punktile 1 (vt punkt 2 jaotise 1.1 loetelus). Kui aga $0 \leq \epsilon < \frac{1}{4}$, siis jadad X^n, Y^n on sõltuvad. Intuitiivselt võib öelda, et mida väiksem on parameeter $0 \leq \epsilon < \frac{1}{4}$, seda tugevam on sõltuvus jadade X^n, Y^n vahel.

3. Sõltuvad jadad juhuslike pikkustega eellasjada pikkusega k. Genereeritakse eellastele $Z_1, ..., Z_k$ vastavad järglased. Täpsemalt: genereeritakse jadad $X^{\sum_{i=1}^k D_i^x}, Y^{\sum_{i=1}^k D_i^y}$. Nende jadade pikkused on binoomjaotusega $Bin(p, k)$ juhuslikud suurused. Tähe allesjäämise tõenäosus on jälle $p = 0.95$. Jadad X^∞, Y^∞ on üldjuhul sõltuvad, P on ühtlane. Üleminekumaatriks Q avaldub parameetrist $0 \leq \epsilon \leq \frac{1}{4}$ sõltuvalt kujul (1.3).

Tänu ülaltoodud loetelule saame simulatsioone edaspidi lühidalt kirjeldada. Näiteks lause "Genereerime 1000 juhuslike pikkustega sõltuvate jadade paari, võttes eellasjada pikkuseks 1000 ja $\epsilon = 0.05$." on nüüd üheselt mõistetav.

Simulatsioonide teostamiseks kasutati statistikatarkvara R [2] ja selle pakette ggplot2 [3], Biostrings [4] ja Rcpp [5, 6]. Simulatsioonides kasutatavatest programmidest on täpsemalt kirjutatud töö lisas.

1.3 Sagedustabelipõhised statistikud jadade sõltuvuse mõõtmiseks ja sõltumatuse testimiseks

Üks võimalus jadade X^n, Y^m sõltuvuse mõõtmiseks on kasutada klassikalisi sagedustabelipõhiseid meetodeid. Saamaks jadadest X^n, Y^m sagedustabelit, peame moodustame nende elementidest paarid. Vast kõige lihtsam lähenemine oleks moodustada paarid

$$(X_1, Y_1), ..., (X_{n \wedge m}, Y_{n \wedge m}).$$

Näiteks, kui jadadeks X^n, Y^m on vastavalt

$$\begin{matrix} 1\ 1\ 1\ 1\ 1 \\ 1\ 0\ 0\ 0, \end{matrix}$$

saame sellise lähenemisega järgmise sagedustabeli:

X \ Y	0	1
0	0	0
1	3	1

Jadade X^n, Y^m sõltumatuse testimiseks võime sagedustabeli pealt arvutada näiteks χ^2-statistiku või Fisheri teststatistiku. Teisalt, neid statistikuid võib kasutada ka heuristiliste sõltuvusmõõtudena.

Sagedustabelipõhist lähenemist saab kasutada ka seoses vastastikuse informatsiooniga. Vastastikune informatsioon on informatsiooniteoorias tuntud suurus, mis mõõdab kahe diskreetse juhusliku suuruse vahelist sõltuvust. Defineerime $0 \log_2 0 := 0$. Juhuslike suuruste $X, Y \in \mathcal{A}$ vastastikuseks informatsiooniks nimetatakse suurust

$$I(X;Y) := \sum_{x,y \in \mathcal{A}} \mathbf{P}(X = x, Y = y) \log_2 \left(\frac{\mathbf{P}(X = x, Y = y)}{\mathbf{P}(X = x)\mathbf{P}(Y = y)} \right).$$

Vastastikune informatsioon on sümmeetriline ja mittenegatiivne. Lisaks $I(X;Y) = 0$ parajasti siis, kui X ja Y on sõltumatud. Sisuliselt võib vastastikust informatsiooni interpreteerida kui informatsioonihulga suurust, mida ühe juhusliku suuruse teadmine annab teise kohta. Vastastikuse informatsiooni ühik sõltub logaritmi baasist definitsioonis; meil on logaritmi baasiks 2, seega ühikuks on bitt. Märgime veel, et vastastikune informatsioon $I(X;Y)$ on juhuslike suuruste X, Y sõltumatusele vastava ühisjaotuse – st marginaaljaotiste korrutise – Kullback-Leibleri kaugus nende tegelikust ühisjaotusest.

Kuidas aga kasutada vastastikust informatsiooni jadade sõltuvuse mõõtmiseks? Kui meie sõltuvate jadade mudelis kadumisi poleks ehk tähe allesjäämise tõenäosus p oleks

alati võrdne ühega, oleks asi lihtne. Sellisel juhul paaride jada $(X_1,Y_1),(X_2,Y_2),...$ oleks *iid* ning jadade X^∞, Y^∞ sõltuvusmäära saaks defineerida kui $I(X_1;Y_1)$. Kõige lihtsamaks hinnanguks suurusele $I(X_1;Y_1)$ oleks nõndanimetatud empiiriline vastastikune informatsioon

$$\hat{I}_{n\wedge m} := \sum_{x,y\in\mathcal{A}} \hat{p}_{X,Y}(x,y)\log_2\left(\frac{\hat{p}_{X,Y}(x,y)}{\hat{p}_X(x)\hat{p}_Y(y)}\right),$$

kus $\hat{p}_{X,Y}(x,y)$ on paari (x,y) esinemise sagedus jadas $(X_1,Y_1),...,(X_{n\wedge m},Y_{n\wedge m})$, $\hat{p}_X(x)$ on tähe x esinemise sagedus jadas $X^{n\wedge m}$ ja $\hat{p}_Y(y)$ on tähe y esinemise sagedus jadas $Y^{n\wedge m}$. Suurte arvude seadusest saame, et suuruse $n\wedge m$ kasvades toimub peaaeug kindlasti koondumine $\hat{I}_{n\wedge m}\to I(X_1,Y_1)$.

Kui aga $p<1$, siis paarid $(X_1,Y_1),(X_2,Y_2),...$ pole üldjuhul sõltumatud ega sama jaotusega. Sellest hoolimata võib statistik $\hat{I}_{n\wedge m}$ olla kasutatav heuristilise sõltuvusmõõduna.

Veel üks võimalus jadade X^n, Y^m sõltuvuse mõõtmiseks oleks astakkorrelatsioonikordajate kasutamine. Selleks me kõigepealt järjestame tähestiku $\mathcal{A}$ tähed, seades igale tähele $a\in\mathcal{A}$ vastavusse erineva arvu $g(a)$, kus $g:\mathcal{A}\to\{1,...,|\mathcal{A}|\}$ on bijektiivne funktsioon. Seejärel moodustame paarid $[g(X_1),g(Y_1)],[g(X_2),g(Y_2)],...$ ja arvutame nende pealt astakkorrelatsioonikordaja, näiteks Spearmanni ρ või Kendalli τ. Et negatiivne korrelatsioon siin tähendust ei oma, siis jadade X^n, Y^m sõltuvusmõõduks võtame astakkorrelatsiooni absoluutväärtuse. Meetodi puuduseks on see, et saadud sõltuvusmõõt sõltub tähtede järjestusest ehk funktsioonist g. Olgu näiteks $\mathcal{A}=\{\text{A,B,C}\}$ ja olgu X- ja Y- jadadeks vastavalt ABC ja AAB. Võttes $g(\text{A})=1, g(\text{B})=2, g(\text{C})=3$, saame saadud arvupaaride Spearmani ρ-ks 0.866 ja Kendalli τ-ks 0.816. Teisalt, võttes $g(\text{A})=1, g(\text{B})=3, g(\text{C})=2$, saame nii Spearmani ρ-ks kui ka Kendalli τ-ks nulli.

1.4 Pikima ühisjada pikkus

1.4.1 Tutvustus

Oleme nüüd vaadanud mitmeid sagedustabelipõhiseid meetodeid sõltuvuse mõõtmiseks ja sõltumatuse testimiseks. Kõigil neil meetoditel on aga ühine puudus: nad saavad informatsiooni jadade sõltuvuse kohta vaid paaridest $(X_1,Y_1),(X_2,Y_2),...$, mitte aga teistest paaridest. Toome ühe näite, mis seda probleemi illustreerib. Olgu $p=1$ ja Q ühikmaatriks. Sellisel juhul $X_i\equiv Y_i$ iga $i\in\mathbb{N}$ korral. Oletame, et soovime testida, kas jadad $X_2,...,X_{n+1}$ ja $Y_1,...,Y_n$ on sõltumatud (võib mõelda, et element X_1 on mõõtmise käigus ära kadunud). On selge, et need kaks jada on väga tugevalt sõltuvad (eeldades, et n pole väga väike). Sõltumatuse testimiseks moodustame paarid $(X_2,Y_1),...,(X_{n+1},Y_n)$ ja rakendame neile χ^2-testi. Paraku X_{i+1},Y_i on sõltumatud iga i korral, seega ei suuda me suure tõenäosusega tuvastada jadade $X_2,...,X_{n+1}$ ja $Y_1,...,Y_n$ vahelist sõltuvust.

Seega klassikalised sagedustabelipõhised meetodid jadade sõltuvuse mõõtmiseks ei pruugi olla kõige paremini toimivad ning mõistlik on otsida alternatiive. Üheks selliseks alternatiiviks on jadade pikima ühisjada pikkus. Kahe jada ühisjadaks nimetatakse jada, mis on nende mõlema osajadaks. Pikimaks ühisjada pikkuseks nimetatakse maksimaalse pikkusega ühisjada pikkust. Esitame ka formaalse definitsiooni.

Definitsioon 1.1 (pikima ühisjada pikkus)**.** *Jadade $x_1,...,x_n$ ja $y_1,...,y_m$ pikima ühisjada pikkuseks nimetatakse maksimumi hulgast*

$$\{k \,|\, x_{i_1},...,x_{i_k} = y_{j_1},...,y_{j_k};\, 1 \le i_1 < ... < i_k \le n;\, 1 \le j_1 < ... < j_k \le m\} \cup \{0\}.$$

Näiteks järgmise kahe jada pikimaks ühisjadaks on AAAGA ja pikima ühisjada pikkuseks on 5:

G **A** C **A A G** T **A** G T
A A A C **G A** C C

Märgime, et kahe jada pikim ühisjada ei pruugi olla üheselt määratud, aga pikima ühisjada pikkus on.

Kuidas aga leida jadade $x_1,...,x_n$ ja $y_1,...,y_m$ pikima ühisjada pikkust? Naiivne meetod oleks iga $k \in \{1,...,n \wedge m\}$ korral ja kõikide $1 \le i_1 < ... < i_k \le n$ ja $1 \le j_1 < ... < j_k \le m$ korral kontrollida, kas $x_{i_1},...,x_{i_k} = y_{j_1},...,y_{j_k}$. Selliselt saame konstrueerida pikima ühisjada pikkuse definitsioonis toodud hulga, millest maksimum ongi otsitud suurus. Kui võtta lihtsuse huvides $n = m$, siis võrreldavate tähepaaride arv on kirjeldatud algoritmi korral

$$\sum_{k=1}^{n} k \binom{n}{k}^2.$$

Kui $n \ge 2$, siis saame ülaltoodud suurusele järgmise jämeda alumise tõkke:

$$\sum_{k=1}^{n} k \binom{n}{k}^2 \ge \sum_{k=0}^{n} \binom{n}{k}^2 \ge \sum_{k=0}^{n} \binom{n}{k} = 2^n = \left(2^{\log_2 10}\right)^{\frac{n}{\log_2 10}} = 10^{\frac{n}{\log_2 10}}.$$

Selle tõkke abil saame näiteks, et kui $n = m = 300$, siis kirjeldatud algoritm peab võrdlema rohkem kui 10^{90} tähepaari. Võrdluseks märgime, et universumis olevate aatomite arv arvatakse olevat ligikaudu suurusjärku 10^{80}. On selge, et vähegi suuremate jadapikkuste korral ei ole naiivse algoritmi kasutamine võimalik. Õnneks on aga olemas dünaamilist planeerimist kasutav algoritm, mis suudab pikkustega n ja m jadade pikima ühisjada pikkuse leida ajaga $O(n \cdot m)$ – see algoritm avaldati 1970. aastal Saul B. Needlemani and Christian D. Wunschi poolt ja seda nimetatakse Needleman-Wunschi algoritmiks.

Defineerime ka mõisted "joondus" ja "optimaalne joondus".

Definitsioon 1.2 (joondus ja optimaalne joondus)**.** *Jadade $x_1,...,x_n$ ja $y_1,...,y_m$ joonduseks pikkusega $k \ge 1$ nimetame hulka*

$$\{(i_1,j_1),...,(i_k,j_k) \,|\, x_{i_1},...,x_{i_k} = y_{j_1},...,y_{j_k};\, 1 \le i_1 < ... < i_k \le n;\, 1 \le j_1 < ... < j_k \le m\}.$$

Kui k on jadade $x_1,...,x_n$ ja $y_1,...,y_m$ pikima ühisjada pikkus, nimetame toodud hulka optimaalseks joonduseks.

Seega kahe jada joondus on sisuliselt nende jadade mingile ühisjadale vastav indeksite komplekt. Optimaalne joondus on aga nende jadade pikimale ühisjadale vastav indeksite komplekt. Vaatame uuesti pikima ühisjada definitsiooni juures toodud näidet. Lisame jadadele indeksite rea:

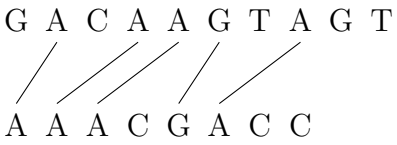
1 2 3 4 5 6 7 8 9 10

G **A** C **A A G** T **A** G T
A A A C **G A** C C

Pikimale ühisjadale AAAGA vastav optimaalne joondus on

$$\{(2,1),(4,2),(5,3),(6,5),(8,6)\}.$$

Kahe jada joonduse moodustamist võib vaadelda ka kui võrdsete tähtede ühendamist teatud viisil. Näiteks ülaltoodud näites vastavad optimaalsele joondusele järgmised ühendused:



Et ühendused vastaksid mingile joondusele, ei tohi ühendavad jooned lõikuda ja iga täht tohib olla ühendatud maksimaalselt ühe tähega.

Kahe jada pikima ühisjada pikkust tagastavat funktsiooni tähistame tähega L. Hulgale $A \subset \mathbb{N}^2$ elemendi $(i,j) \in \mathbb{N}^2$ liitmise defineerime järgmiselt:

$$A + (i,j) = \{(i+x, j+y) \,|\, (x,y) \in A\}.$$

Tõestame funktsiooni L kohta ühe abitulemuse.

Lemma 1.1. *Kõikide $x_1,...,x_n,y_1,...,y_n \in \mathcal{A}$ ja mis tahes $k \in \{1,...,n-1\}$ korral*

$$L(x_1,...,x_n;\, y_1,...,y_n) \geq L(x_1,...,x_k;\, y_1,...,y_k) + L(x_{k+1},...,x_n;\, y_{k+1},...,y_n).$$

Tõestus. Olgu $k \in \{1,...,n-1\}$. Veel olgu

$$\begin{aligned}
&\widetilde{L}_{1,n} := L(x_1,...,x_n\,;\, y_1,...,y_n),\\
&\widetilde{L}_{1,k} := L(x_1,...,x_k\,;\, y_1,...,y_k),\\
&\widetilde{L}_{k+1,n} := L(x_{k+1},...,x_n\,;\, y_{k+1},...,y_n),\\
&x'_1,...,x'_{n-k} := x_{k+1},...,x_n,\\
&y'_1,...,y'_{n-k} := y_{k+1},...,y_n.
\end{aligned}$$

ja

$$C_1 := \begin{cases} \text{jadade } x_1,...,x_k \text{ ja } y_1,...,y_k \text{ optimaalne joondus, kui } \widetilde{L}_{1,k} \geq 1 \\ \emptyset, \text{ vastasel korral} \end{cases},$$

$$C_2 := \begin{cases} \text{jadade } x'_1,...x'_{n-k} \text{ ja } y'_1,...,y'_{n-k} \text{ optimaalne joondus, kui } \widetilde{L}_{k+1,n} \geq 1 \\ \emptyset, \text{ vastasel korral} \end{cases}.$$

Kui kehtivad võrdused

$$C_1 = C_2 = \emptyset, \tag{1.4}$$

siis

$$\widetilde{L}_{1,n} \geq \widetilde{L}_{1,k} + \widetilde{L}_{k+1,n} = 0.$$

Kui võrdused (1.4) ei kehti, siis hulk

$$C := C_1 \cup [C_2 + (k,k)]$$

on jadade $x_1,...,x_n$ ja $y_1,...,y_n$ joondus. Seega

$$\widetilde{L}_{1,n} \geq |C| = |C_1| + |C_2| = \widetilde{L}_{1,k} + \widetilde{L}_{k+1,n}.$$

□

Toodud lemmat saab rakendada järgmise teoreemi abil.

Teoreem 1.1 (Kingmani subaditiivne ergoodiline teoreem)**.** [7, Teoreem 6.10] *Olgu* $W_1, W_2, ...$ iid *jada. Olgu funktsioonid* f_k, $k = 1,2,...,$ *sellised, et*

$$E|f_k(W_1, W_2, ...)| < \infty \quad \forall k \in \mathbb{N}$$

ja

$$f_{k+l}(W_1, W_2, ...) \leq f_k(W_1, W_2, ...) + f_l(W_{k+1}, W_{k+2}, ...) \quad \forall k, l \in \mathbb{N}.$$

Siis leidub konstant c *nii, et*

$$\frac{1}{k} f_k(W_1, W_2, ...) \xrightarrow{k} c \quad \textit{p.k. ja ruumis } L_1.$$

Tegelikult kehtib Kingmani subaditiivne ergoodiline teoreem palju üldisematel eeldusel, aga see langeb käesoleva töö raamest välja. Defineerime

$$L_k := L(X^k; Y^k), \quad k = 1,2,...$$

Kasutades lemmat 1.1 ja ülaltoodud teoreemi, saame järgmise tulemuse.

Järeldus 1.1. *Kui jadad* X^∞, Y^∞ *on sõltumatud, siis leidub konstant* $\gamma \in [0,1]$ *nii, et*

$$\frac{L_k}{k} \xrightarrow{k} \gamma \quad \textit{p.k. ja ruumis } L_1. \tag{1.5}$$

Lisaks

$$\gamma = \lim_k \left(\frac{EL_k}{k} \right) = \sup_k \left(\frac{EL_k}{k} \right). \tag{1.6}$$

Tõestus. Teame, et jadad X^∞, Y^∞ on mõlemad *iid*; eeldame veel, et nad on sõltumatud. Siis paaride jada $(X_1,Y_1),(X_2,Y_2),...$ on *iid*. Defineerime funktsioonid f_k:

$$f_k[(x_1,y_1),(x_2,y_2),...] := -L(x_1,...,x_k; y_1,...,y_k), \quad k = 1,2,....$$

Kuna iga $k \in \mathbb{N}$ korral $-k \leq f_k \leq 0$, siis

$$E|f_k[(X_1,Y_1),(X_2,Y_2)...]| < \infty \quad \forall k \in \mathbb{N}.$$

Tänu lemmale 1.1 mis tahes $k,l \in \mathbb{N}$ korral

$$L(X_1,...,X_{k+l};\, Y_1,...,Y_{k+l}) \geq L(X_1,...,X_k;\, Y_1,...,Y_k) + L(X_{k+1},...,X_{k+l};\, Y_{k+1},...,Y_{k+l}). \tag{1.7}$$

Korrutades ülaltoodud võrratuses mõlemad pooled arvuga -1, saame, et mis tahes $k,l \in \mathbb{N}$ korral

$$f_{k+l}[(X_1,Y_1),(X_2,Y_2),...] \leq f_k[(X_1,Y_1),(X_2,Y_2),...] + f_l[(X_{k+1},Y_{k+1}),(X_{k+2},Y_{k+2}),...].$$

Näeme, et Kingmani subaditiivse ergoodilise teoreemi (teoreem 1.1) eeldused on täidetud. Seega leidub konstant $\gamma \in [0,1]$ nii, et

$$\frac{1}{k} f_k[(X_1,Y_1),(X_2,Y_2),...] \xrightarrow{k} -\gamma \quad \text{p.k. ja ruumis } L_1.$$

Teisisõnu

$$-\frac{L_k}{k} \xrightarrow{k} -\gamma \quad \text{p.k. ja ruumis } L_1.$$

Siit järeldub koondumine (1.5).

Koondumisest ruumis L_1 järeldub keskväärtuste koondumine, st

$$\gamma = \lim_k \left(\frac{EL_k}{k} \right). \tag{1.8}$$

Võttes võrratuse (1.7) mõlemast poolest keskväärtuse, saame, et mis tahes $k,l \in \mathbb{N}$ korral

$$EL_{k+l} \geq EL_k + EL_l.$$

Teisisõnu jada (EL_k) on superaditiivne. Rakendades nüüd Fekete lemmat, saame

$$\lim_k \left(\frac{EL_k}{k} \right) = \sup_k \left(\frac{EL_k}{k} \right). \tag{1.9}$$

Seostest (1.8) ja (1.9) saamegi seose (1.6). □

Konstant γ ülaltoodud järelduses sõltub juhuslike suuruste $X_1, X_2, ..., Y_1, Y_2, ...$ marginaaljaotusest P. Seda konstanti nimetatakse Chvatal-Sankoffi konstandiks, samanimeliste autorite 1975. aasta teedrajava artikli järgi [8]. Kuigi Chvatal-Sankoffi konstanti on uuritud juba mitukümmend aastat pole selle täpne väärtus teada ühegi jaotuse P korral. Simuleerimise teel on aga võimalik teada saada γ ligikaudne väärtus. Näiteks kui $P = Be(0.5)$, on γ ligikaudu võrdne arvuga 0.81.

Tegelikult kehtib koondumine (1.5) ka ilma sõltumatuse eelduseta, st alati leidub konstant $\gamma^* \in [0,1]$ nii, et

$$\frac{L_k}{k} \xrightarrow{k} \gamma^* \quad \text{p.k. ja ruumis } L_1.$$

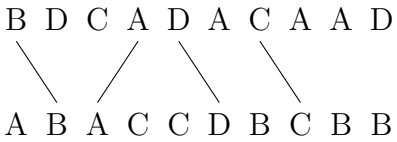
Toodud väite tõestuse leiab lugeja kirjandusest [9, Proposition 4.1]. Kuna koondumisest ruumis L_1 järeldub keskväärtuste koondumine, siis

$$\gamma^* = \lim_k \left(\frac{EL_k}{k} \right).$$

Konstant γ^* sõltub juhuslike suuruste $Z_1, Z_2, ...$ marginaaljaotusest, tähe allesjäämise tõenäosusest p ning üleminekumaatriksist Q.

Uurime nüüd simulatsioonide abil konstandi γ^* käitumist. Esiteks vaatame, kuidas sõltub γ^* jaotusest P. Genereerime iga $\lambda \in \{0.25, 0.26, ..., 1\}$ korral jaotusest $(\lambda, \frac{1-\lambda}{3}, \frac{1-\lambda}{3}, \frac{1-\lambda}{3})$ tähtedega sõltumatute jadade paari pikkusega 1000. Mida suurem on parameeter $\lambda \in [0.25, 1]$ seda ebaühtlasem on jaotus P; kui $\lambda = 1$, siis jaotus P koosneb ainult ühest tähest, st vastab konstandile. Konstanti $\gamma^* = \gamma$ saame hinnata suuruse $L_{1000}/1000$ abil. Simulatsioonide tulemused on toodud joonisel 1.1a. Näeme, et mida ebaühtlasem on jaotus P, seda suurem on ka konstant γ^* . Seda võib põhjendada järgmiselt. Kui jaotus P on ebaühtlane, siis X- ja Y-jadas on mingeid tähti rohkem kui teisi. Näiteks, kui $\mathcal{A} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$, siis võib ebaühtlane P tähendada seda, et X- ja Y-jada koosnevad peamiselt tähest A. See aga tähendab omakorda, et me saame pikima ühisjada moodustamisel ühendada mitmeid A-tähti. Selle asjaolu illustreerimiseks toome alljärgnevalt kaks ebaühtlase jaotusega jada pikkusega 10 ja ühele nende jadade optimaalsele joondusele vastavad ühendused.

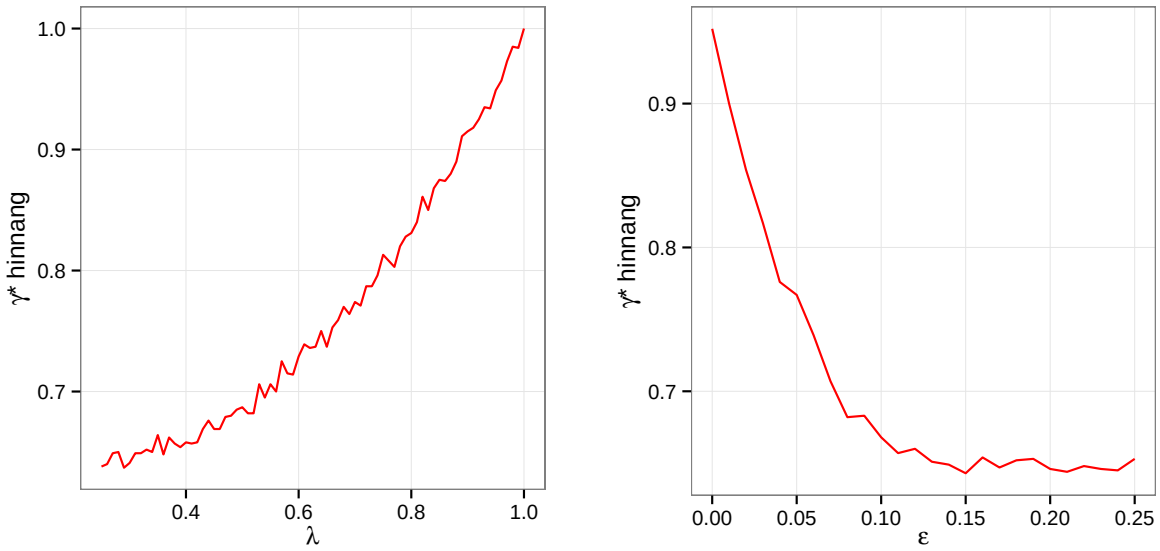
Näeme, et nende jadade optimaalsele joondusele vastav ühenduste arv ehk pikima ühisjada pikkus on seitse. Lisaks ühendustele vastav pikim ühisjada – ABAAAAA – koosneb peamiselt A-tähest. Võrdluseks toome ka kaks sama pikka ühtlase jaotusega jada koos ühele optimaalsele joondusele vastavate ühendustega.



Näeme, et nende jadade pikima ühisjada pikkus neli on palju väiksem kui eelneva jadapaari oma.

Meenutame nüüd, et sõltuvate jadade simuleerimisel avaldub üleminekumaatriks Q parameetri ϵ kaudu kujul (1.3). Uurime, kuidas sõltub konstant γ^* parameetrist ϵ. Genereerime iga $\epsilon \in \{0, 0.01, ..., 0.25\}$ korral sõltuvate jadade paari fikseeritud pikkusega 1000. Mida suurem on parameeter $\epsilon \in [0, 0.25]$, seda väiksem on sõltuvus genereeritavate X-ja Y-jadade vahel; kui $\epsilon = 0.25$, siis X- ja Y-jadad on sõltumatud. Konstanti γ^* hindame jällegi suuruse $L_{1000}/1000$ abil. Simulatsioonide tulemused on toodud joonisel 1.1b. Näeme, et mida väiksem on sõltuvus X- ja Y-jadade vahel, seda väiksem on ka konstant γ^*. Seda asjaolu võib põhjendada järgmiselt. Mida väiksem on parameeter $\epsilon \in [0, 0.25]$, seda suurem on tõenäosus, et mingi sugulaspaari elemendid on võrdsete tähtedega (kui $\epsilon = 0$, siis see tõenäosus on 1). Seega, mida väiksem on $\epsilon \in [0, 0.25]$, seda rohkem kipub olema sugulaspaari, mida saab pikima ühisjada moodustamisel ühendada. Rohkem ühendusi tähendabki aga pikemat pikima ühisjada pikkust.

Seega fikseeritud jaotuse P korral võib ehk jadade X^n, Y^m sõltuvusmõõduna kasutada



(a) (b)

Joonis 1.1: konstandi γ^* sõltuvus parameetritest λ (a) ja ϵ (b)

statistikut

$$\frac{L_{n\wedge m}}{n\wedge m}.$$

Seda statistikut võib ka vaadelda kui valimihinnangut suurusele γ^*.

1.4.2 Asümptootilisest jaotusest

Kui me soovime pikima ühisjada pikkuse põhjal konstrueerida mingit statilist testi, oleks hea teada statistiku L_n täpset jaotust. Paraku vähegi suure n korral on täpse jaotuse leidmine osutunud väga keeruliseks ülesandeks. Seega on ehk mõistlikum uurida selle statistiku asümptootilist jaotust. Paraku pole ka asümptootilise jaotuse kohta palju teada. Üks hüpotees on, et statistik

$$\frac{L_n - EL_n}{\sqrt{\mathrm{Var}L_n}}$$

on asümptootiliselt standardse normaaljaotusega. Hiljuti tõestati see hüpotees erijuhul, kus jadad X^∞, Y^∞ on sõltumatud ja P on väga ebaühtlane Bernoulli jaotus [10].

Me püstitame veel hüpoteesi, et leidub meie sõltuvate jadade mudelist sõltuv positiivne konstant $\sigma < \infty$ nii, et

$$\frac{\mathrm{Var}L_n}{n} \xrightarrow{n} \sigma^2. \tag{1.10}$$

Nimetatud koondumist pole tõestatud ühegi erijuhu korral. Siiski on mõningaid teoreetilisi tulemusi, mis selle kehtivusele vihjavad. Näiteks jaotises 2.2.2 tõestab autor, et

$$\mathrm{Var}L_n = O(n)$$

Lisaks on veel tõestatud [11], et kui jadad X^∞, Y^∞ on sõltumatud ja P on väga ebaühtlane Bernoulli jaotus, siis

$$\mathrm{Var}L_n = \Omega(n) \quad \overset{\text{def}}{\Leftrightarrow} \quad n = O(\mathrm{Var}L_n),$$

st dispersioon $\mathrm{Var}L_n$ on asümptootiliselt tõkestatud mingi lineaarse alumise tõkkega.

Kui mõlemad ülaltoodud hüpoteesid kehtivad, st kui statistik (1.10) on asümptootiliselt standardse normaaljaotusega ja kui piirväärtus σ^2 leidub, siis tänu Slutsky lemmale

$$\frac{L_n - EL_n}{\sqrt{\mathrm{Var}L_n}} \cdot \sqrt{\frac{\mathrm{Var}L_n}{n}} \Rightarrow X_{\mathcal{N}(0,1)} \cdot \sigma,$$

kus $X_{\mathcal{N}(0,1)}$ on standardse normaaljaotusega juhuslik suurus. Teisisõnu, kui mõlemad ülaltoodud hüpoteesid kehtivad, siis

$$\frac{L_n - EL_n}{\sqrt{n}} \Rightarrow \mathcal{N}(0, \sigma^2). \tag{1.11}$$

Me oleme nüüd esitanud mitmeid hüpoteese. Ülevaatlikkuse huvides moodustame nendest järgmise loendi.

HÜP 1. Kehtib koondumine

$$\frac{L_n - EL_n}{\sqrt{\mathrm{Var}L_n}} \Rightarrow \mathcal{N}(0,1)$$

(tõestatud juhul, kus jadad X^∞, Y^∞ on sõltumatud ja P on väga ebaühtlane Bernoulli jaotus).

HÜP 2. Leidub meie sõltuvate jadade mudelist sõltuv positiivne konstant $\sigma < \infty$ nii, et

$$\frac{\mathrm{Var}L_n}{n} \xrightarrow{n} \sigma^2.$$

HÜP 3. Leidub meie sõltuvate jadade mudelist sõltuv positiivne konstant $\sigma < \infty$ nii, et

$$\frac{L_n - EL_n}{\sqrt{n}} \Rightarrow \mathcal{N}(0, \sigma^2).$$

Veendusime, et kehtib implikatsioon

$$\text{HÜP 1 ja HÜP 2} \Rightarrow \text{HÜP 3}.$$

Jaotises 1.6.1 kontrollime simulatsioonide abil toodud hüpoteeside kehtivust.

1.5 Ekstremaalsed joondused

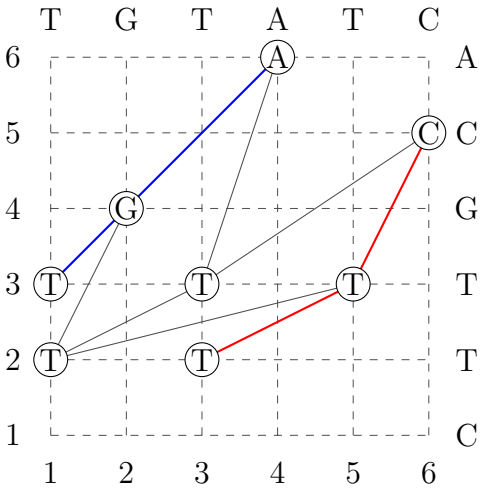
1.5.1 Idee

Ekstremaalsete joonduste idee jadade sõltuvuse mõõtmiseks käidi välja 2014. aasta artiklis [9]. Kahel jadal on üldjuhul mitmeid optimaalseid joondusi – kõik need saame esitada kahedimensionaalsel graafikul. Näiteks joonisel 1.2 on toodud kõik jadade TGTATC ja CTTGCA optimaalsed joondused; iga kasvav joon graafikul kujutab endast ühte optimaalset joondust. Näeme, et nimetatud jadadel on kuus erinevat optimaalset joondust ja et nende pikima ühisjada pikkuseks on 3.

Toodud näite varal üritame selgitada ka ekstremaalsete joonduste mõistet. Paneme tähele, et üks joondus asub joonisel teiste peal (sinine joon) ja üks teiste all (punane joon) – neid nimetame vastavalt ülemiseks ja alumiseks joonduseks. Ülemist ja alumist joondust nimetamegi kokkuvõtvalt ekstremaalseteks joondusteks. Antud näites on ülemiseks ja alumiseks joonduseks seega vastavalt

$$\{(1,3),(2,4),(4,6)\}, \quad \{(3,2),(5,3),(6,5)\}$$

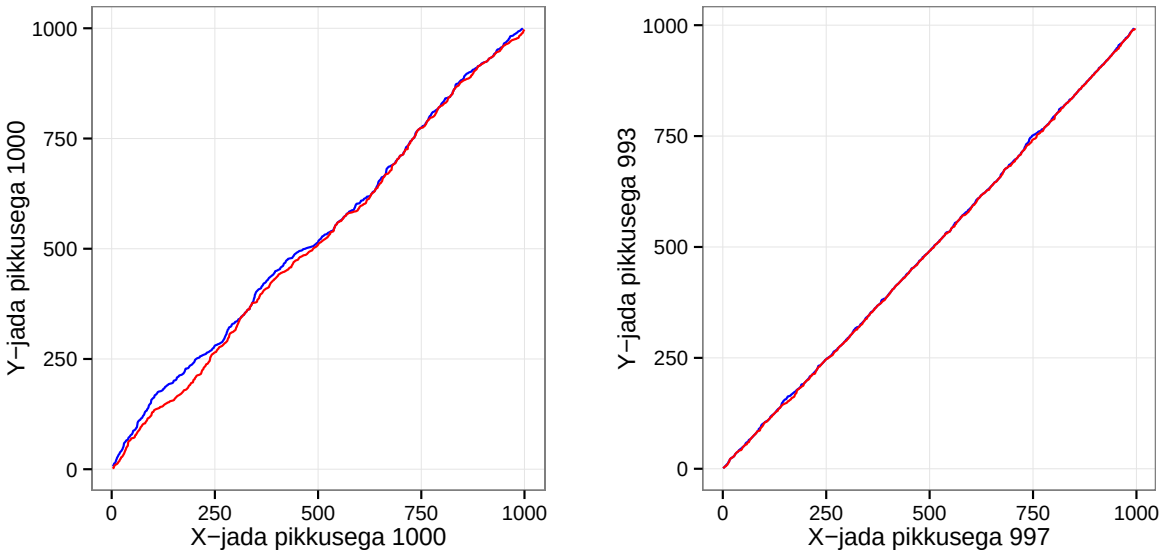
ning ülemisele ja alumisele joondusele vastavad ühisjadad on TGA ja TTC. Loomulik küsimus on, kas mis tahes jadapaaril ekstremaalsed joondused alati leiduvad. Järgmises jaotises esitame ekstremaalsete joonduste formaalse definitsiooni ning tõestame, et kui jadapaari pikima ühisjada pikkus on vähemalt üks, siis sellel leiduvad ka ekstremaalsed joondused.



Joonis 1.2: jadade TGTATC ja CTTGCA optimaalsed joondused

Järgnevalt genereerime ühe sõltumatute jadade paari ja ühe sõltuvate jadade paari ning võrdleme joonise abil nende ekstremaalseid joonduseid. Sõltumatud jadad genereerime pikkusega 1000. Sõltuvad jadad genereerime juhuslike pikkustega (fikseeritud pikkustega sõltuvad jadad toovad ekstremaalsete joonduste teoreetilises uurimises juurde teatavat tehnilist lisakeerukust, mida antud töös üritame vältida), võttes eellasjada pikkuseks 1053 ning $\epsilon = 0.05$ (meenutame, et üleminekumaatriks Q avaldus parameetri ϵ kujul (1.3)). Genereeritavate sõltuvate jadade keskmine pikkus on ligikaudu $1053 \cdot 0.95 \approx$

1000. Ekstremaalsed joondused leiame Needleman-Wunschi algoritmi abil. Ühendades ülemisele ja alumisele joondusele vastavad punktid, saame ühe jadapaari ekstremaalsed joondused esitada kahe joonena. Tulemus on toodud joonistel 1.3a (sõltumatud jadad) ja 1.3b (sõltuvad jadad). Võrreldes neid jooniseid, näeme, et sõltuvate jadade korral asuvad



(a) (b)

Joonis 1.3: ekstremaalsed joondused sõltumatute (a) ja sõltuvate (b) jadade korral

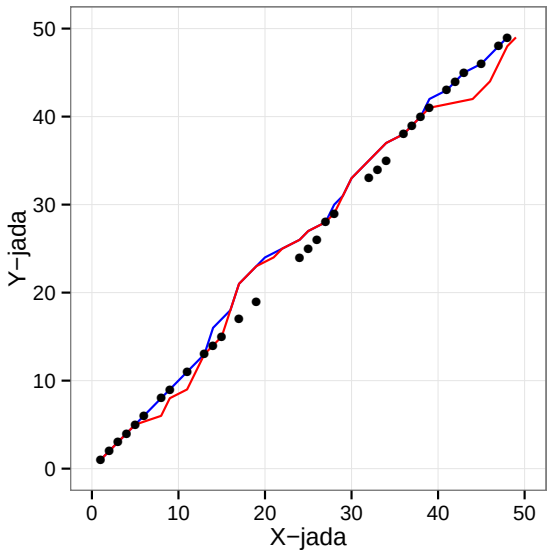
ekstremaalsed joondused üksteisest rohkem lahus. Hiljem esitame karpdiagramme, mis kinnitavad, et mida tugevam on jadade sõltuvus, seda väiksem kipub olema lahknevus ekstremaalsete joonduste vahel. Mis võiks olla aga sellise seaduspära põhjus?

Meenutame, et paar (i, j) on meie definitsiooni kohaselt sugulaspaar parajasti siis, kui X-jada elemendil i ja Y-jada elemendil j on sama eellane. Ütleme, et paar (i, j) on võrdsete sugulaste paar – lühidalt VSP –, kui ta on sugulaspaar ja $X_i = Y_j$. Intuitiivselt pole raske mõista, et mida suurem on kahe jada vaheline sõltuvus, seda suurem kipub fikseeritud jadapikkuste korral olema VSP-de arv. Joonisel 1.4 on toodud joonise 1.3 sõltuvate jadade ekstremaalsete joonduse algusosa. Lisaks on joonisel mustade täppidega märgitud VSP-d. Näeme, et paljud punktid asuvad ekstremaalsete joonduste peal. Simulatsioonid on näidanud, et see ongi ekstremaalsetele joondustele omane: nad läbivad tihti VSP-sid. Kui jadade sõltuvus on suur, siis neil on ka palju VSP-sid, mistõttu on ka ekstremaalsed joondused üksteisele lähedal – samas kui jadade sõltuvus on väike, on lahknevus tavaliselt suurem.

Kuidas aga mõõta ekstremaalsete joonduste vahelist lahknevust? Võimalusi on mitmeid, aga lihtsuse huvides piirdume käesolevas töös vaid ühega, nimelt Hausdorffi kaugusega. Anname selle kauguse definitsiooni.

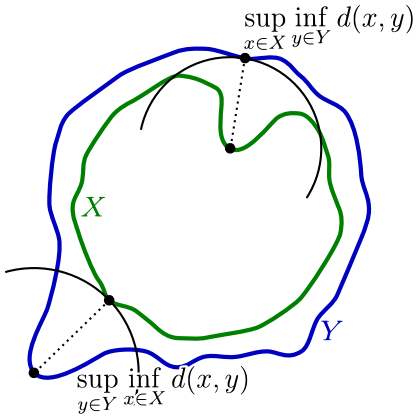
Definitsioon 1.3 (Hausdorffi kaugus)**.** *Olgu X, Y kaks mittetühja hulka meetrilises ruumis kaugusega d. Hausdorffi kauguseks nende hulkade vahel nimetatakse suurust*

$$d_H(X,Y) := \max\{\sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(x,y)\}$$



Joonis 1.4: osa sõltuvate jadade ekstremaalses joondusest koos VSP-dega

Meil on definitsioonis toodud kauguseks d alati eukleidiline kaugus. Joonisel 1.5 on toodud näide hulkade $X, Y \in \mathbb{R}^2$ vahelise Hausdorffi kauguse kahest arvutuslikust komponendist.



Joonis 1.5: näide Hausdorffi kauguse arvutuslikest komponentidest[1]

Märgime, et kui pikima ühisjada pikkusele põhineva statistiku $L_{n\wedge m}/(n \wedge m)$ suur väärtus näitas suurt sõltuvust ja väike väärtus väikest sõltuvust, siis siin on olukord vastupidine: suur Hausdorffi kaugus ekstremaalsete joonduste vahel näitab väikest sõltuvust jadade vahel ja väike kaugus suurt sõltuvust.

1.5.2 Formaalne definitsioon

Me defineerime nüüd formaalselt jadade $x \in \mathcal{A}^n$ ja $y \in \mathcal{A}^m$ ekstremaalsed joondused. Olgu nende jadade pikima ühisjada pikkuseks $t \geq 1$ ja olgu neil k erinevat optimaalset joondust.

[1] Autor: Rocchini; CC-BY-3.0 (http://creativecommons.org/licenses/by/3.0), Wikimedia Commons.

Olgu

$$\{\{(i_1^1,j_1^1),...,(i_t^1,j_t^1)\},...,\{(i_1^k,j_i^k),...,(i_t^k,j_t^k)\}\}$$

jadade x,y kõigi optimaalsete joonduste hulk. Defineerime $A:=\{1,...,k\}$. Eeldame, et meie optimaalsete joonduste elemendid on järjestatud, st $i_1^\alpha<...<i_t^\alpha$ iga $\alpha\in A$ korral (millest järeldub, et $j_1^\alpha<...<j_t^\alpha$ iga $\alpha\in A$ korral). Olgu

$$j_t^\uparrow:=\max_{\alpha\in A} j_t^\alpha=\max\{j_s^\alpha\,|\,\alpha\in A,s\in\{1,...,t\}\},\tag{1.12}$$

$$i_t^\uparrow:=\min\{i_t^\alpha\,|\,j_t^\alpha=j_t^\uparrow,\alpha\in A\}.\tag{1.13}$$

Edasi võtame iga $l\in\{1,...,t-1\}$ korral.

$$j_l^\uparrow:=\max\{j_s^\alpha\,|\,i_s^\alpha<i_{l+1}^\uparrow,\alpha\in A,s\in\{1,...,t\}\},$$
$$i_l^\uparrow:=\min\{i_s^\alpha\,|\,j_s^\alpha=j_l^\uparrow,\alpha\in A,s\in\{1,...,t\}\}.$$

Hulka $\{(i_1^\uparrow,j_1^\uparrow),...,(i_t^\uparrow,j_t^\uparrow)\}$ nimetamegi jadade x,y ülemiseks joonduseks. Me peame nüüd tõestame, et toodud definitsioon on korrektne, st paarid $(i_s^\uparrow,j_s^\uparrow)$ leiduvad. Seda ütlebki meile muuhulgas järgmine lause.

Lause 1.1. [9, Proposition 2.1] *Hulk* $\{(i_1^\uparrow,j_1^\uparrow),...,(i_t^\uparrow,j_t^\uparrow)\}$ *on jadade* x,y *optimaalseks joonduseks, kusjuures iga* $s\in\{1,...,t\}$ *korral*

$$j_s^\uparrow=\max_{\alpha\in A} j_s^\alpha,\quad i_s^\uparrow=\min\{i_s^\alpha\,|\,j_s^\alpha=j_s^\uparrow,\alpha\in A\}.\tag{1.14}$$

Tõestus. Piisab vaid tõestada, et (1.14) kehtib iga $s\in\{1,...,t\}$ korral. Selleks kasutame induktsiooni. Definitsioonide (1.12), (1.13) põhjal (1.14) kehtib $s=t$ korral. Olgu $u\in\{1,...,t-1\}$. Eeldame, et (1.14) kehtib $s=u+1$ korral, seega leidub selline $\alpha\in A$, et

$$j_{u+1}^\uparrow=j_{u+1}^\alpha,\quad i_{u+1}^\uparrow=i_{u+1}^\alpha.\tag{1.15}$$

Näitame, et (1.14) kehtib $s=u$ korral. Tänu eeldusele (1.15) paar $(i_u^\uparrow,j_u^\uparrow)$ leidub, seega leidub selline paar $(\beta,v)\in A\times\{1,...,t\}$, et $(i_u^\uparrow,j_u^\uparrow)=(i_v^\beta,j_v^\beta)$. Näitame kõigepealt, et $v=u$. Oletame vastuväiteliselt, et $v>u$. Kuna $i_v^\beta<i_{u+1}^\alpha$ ja $j_v^\beta<j_{u+1}^\alpha$, siis hulk

$$\{(i_1^\beta,j_1^\beta),...,(i_v^\beta,j_v^\beta),(i_{u+1}^\alpha,j_{u+1}^\alpha),...,(i_t^\alpha,j_t^\alpha)\}$$

moodustab joonduse, mille pikkus on vähemalt $t+1$ – vastuolu. Oletame nüüd vastuväiteliselt, et $v<u$. Paneme tähele, et $i_u^\alpha<i_{u+1}^\alpha\le i_u^\beta$. Teisalt $j_u^\alpha\le j_v^\beta<j_u^\beta$. Seega hulk

$$\{(i_1^\alpha,j_1^\alpha),...,(i_u^\alpha,j_u^\alpha),(i_u^\beta,j_u^\beta),...,(i_t^\beta,j_t^\beta)\}$$

moodustab joonduse pikkusega $t+1$ – vastuolu. Oleme näidanud, et $v=u$, millest järeldub $j_u^\uparrow=\max\{j_u^{\alpha'}\,|\,i_u^{\alpha'}<i_{u+1}^\uparrow,\alpha'\in A\}$. Kui nüüd vastuväiteliselt oletada, et (1.14) ei kehti $s=u$ korral, saame $j_u^\uparrow<\max_{\alpha'\in A} j_u^{\alpha'}$. Eelnevast järeldub, et leidub β' nii, et $j_u^{\beta'}>j_u^\uparrow=j_u^\beta$ ja $i_u^{\beta'}\ge i_{u+1}^\uparrow>i_u^\uparrow=i_u^\beta$. Seega hulk

$$\{(i_1^\beta,j_1^\beta),...,(i_u^\beta,j_u^\beta),(i_u^{\beta'},j_u^{\beta'}),...,(i_t^{\beta'},j_t^{\beta'})\}$$

moodustab joonduse pikkusega $t+1$ – vastuolu. □

Analoogiliselt saame defineerida ka jadade x, y vasakpoolseima joonduse kui hulga

$$\{(i_1^{\leftarrow}, j_1^{\leftarrow}), ..., (i_t^{\leftarrow}, j_t^{\leftarrow})\},$$

kus

$$\begin{aligned} i_1^{\leftarrow} &:= \min_{\alpha \in A} i_1^{\alpha} = \min\{i_s^{\alpha} \,|\, \alpha \in A, s \in \{1, ..., t\}\}, \\ j_1^{\leftarrow} &:= \max\{j_1^{\alpha} \,|\, i_1^{\alpha} = i_1^{\leftarrow}, \alpha \in A\} \end{aligned}$$

ning iga $l \in \{2, ..., t\}$ korral

$$\begin{aligned} i_l^{\leftarrow} &:= \min\{i_s^{\alpha} \,|\, j_s^{\alpha} > j_{l-1}^{\leftarrow}, \alpha \in A, s \in \{1, ..., t\}\}, \\ j_l^{\leftarrow} &:= \max\{j_s^{\alpha} \,|\, i_s^{\alpha} = i_l^{\leftarrow}, \alpha \in A, s \in \{1, ..., t\}\}. \end{aligned}$$

Analoogiliselt lause 1.1 tõestusega saab näidata, et

$$i_s^{\leftarrow} = \min_{\alpha \in A} i_s^{\alpha}, \quad j_s^{\leftarrow} = \max\{j_s^{\alpha} \,|\, i_s^{\alpha} = i_s^{\leftarrow}, \alpha \in A\}, \quad s = 1, ..., t. \tag{1.16}$$

Osutub, et vasakpoolseim joondus on sama mis ülemine joondus. Formuleerime vastava tulemuse lausena.

Lause 1.2. [9, lk 1035-1036] *Iga* $s \in \{1, ..., t\}$ *korral* $(i_s^{\leftarrow}, j_s^{\leftarrow}) = (i_s^{\uparrow}, j_s^{\uparrow})$.

Tõestus. Tõestame lause seoste (1.14) ja (1.16) abil. Olgu $s \in \{1, ..., t\}$ suvaline. Kehtivad samaväärsused

$$\begin{aligned} i_s^{\leftarrow} = i_s^{\uparrow} \quad &\Leftrightarrow \quad \min_{\alpha' \in A} i_s^{\alpha'} = \min\{i_s^{\alpha'} \,|\, j_s^{\alpha'} = j_s^{\uparrow}, \alpha' \in A\} \\ &\Leftrightarrow \quad \text{leidub } \alpha \in A \text{ nii, et } i_s^{\alpha} = \min_{\alpha' \in A} i_s^{\alpha'} \text{ ja } j_s^{\alpha} = j_s^{\uparrow} \\ &\Leftrightarrow \quad \text{leidub } \alpha \in A \text{ nii, et } j_s^{\alpha} = \max_{\alpha' \in A} j_s^{\alpha'} \text{ ja } i_s^{\alpha} = i_s^{\leftarrow} \\ &\Leftrightarrow \quad \max\{j_s^{\alpha'} \,|\, i_s^{\alpha'} = i_s^{\leftarrow}, \alpha' \in A\} = \max_{\alpha' \in A} j_s^{\alpha'} \\ &\Leftrightarrow \quad j_s^{\leftarrow} = j_s^{\uparrow}. \end{aligned}$$

Samas seoste (1.14) ja (1.16) põhjal $i_s^{\leftarrow} \le i_s^{\uparrow}$ ja $j_s^{\leftarrow} \le j_s^{\uparrow}$. Kui nüüd vastuväiteliselt oletada, et $(i_s^{\leftarrow}, j_s^{\leftarrow}) \neq (i_s^{\uparrow}, j_s^{\uparrow})$, saame eelneva põhjal, et $i_s^{\leftarrow} < i_s^{\uparrow}$ ja $j_s^{\uparrow} < j_s^{\leftarrow}$. See aga annaks meile optimaalse joonduse pikkusega $t+1$ – vastuolu. □

Sarnaselt ülaltooduga defineerime jadade x, y alumise ehk parempoolseima joonduse kui hulga $\{(i_1^{\downarrow}, j_1^{\downarrow}), ..., (i_t^{\downarrow}, j_t^{\downarrow})\}$, kus

$$j_s^{\downarrow} := \min_{\alpha \in A} j_s^{\alpha}, \quad i_s^{\downarrow} := \max\{i_s^{\alpha} \,|\, j_s^{\alpha} = j_s^{\downarrow}, \alpha \in A\}, \quad s = 1, ..., t.$$

Paneme tähele, et jadade x, y vasakpoolseim joondus on sama mis jadade y, x alumine joondus. Seega jadade x, y ülemine joondus on sama mis jadade y, x alumine joondus.

1.6 Simulatsioonid (1)

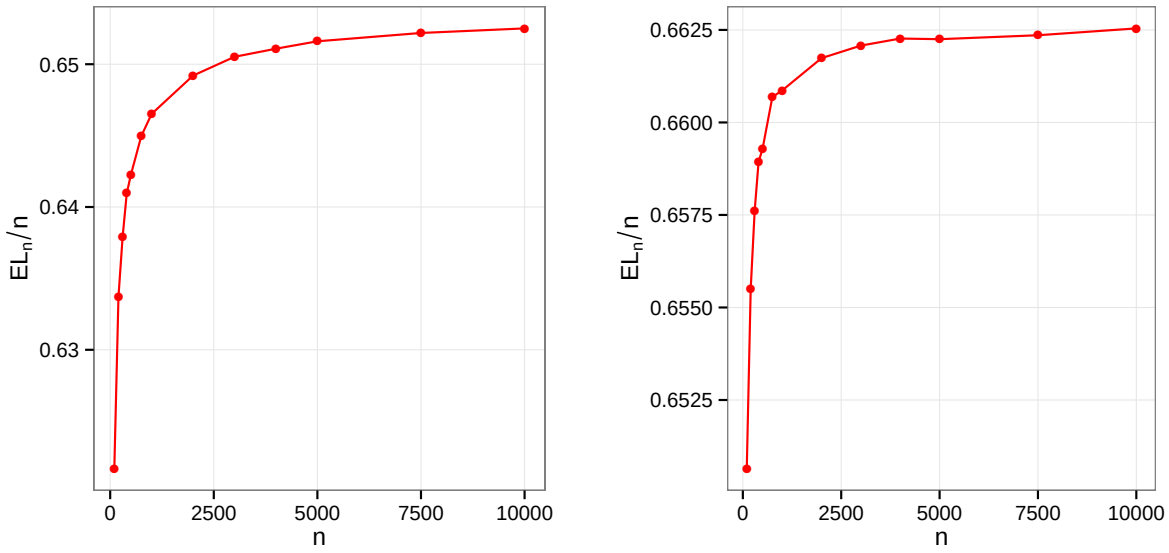
1.6.1 Pikima ühisjada pikkus

Uurime nüüd pikima ühisjada pikkuse asümptootilisi omadusi simulatsioonide abil. Genereerime

$$n = 100, 200, 300, 400, 500, 750, 1000, 2000, 3000, 4000, 5000, 7500, 10\,000$$

korral 1000 sõltumatute ja sõltuvate jadade paari fikseeritud pikkusega n. Sõltuvate jadade korral võtame $\epsilon = 0.05$. Tuhande jadapaari pealt saame hinnata pikima ühisjada pikkuse keskväärtust ja dispersiooni (kasutades dispersiooni nihketa hinnangut).

Joonisel 1.6 näeme suuruse EL_n/n hinnangulist käitumist sõltumatute (a) ja sõltuvate (b) jadade korral. Joonisele 1.6a on veel lisatud vastav Chvatal-Sankoffi konstant (γ). Joonis kinnitab meie varasemat teooriat, et jada (EL_n/n) koondub mingiks positiivseks konstandiks (mis sõltumatuse juhul on konstant γ).



(a) (b)

Joonis 1.6: suuruse EL_n/n hinnanguline käitumine sõltumatute (a) ja sõltuvate (b) jadade korral

Meenutame nüüd, et jaotises 1.4.2 esitasime kolm hüpoteesi. Kontrollime simulatsioonide abil nende hüpoteeside kehtivust. Esimene hüpotees (HÜP 1) oli, et statistik

$$\frac{L_n - EL_n}{\sqrt{\mathrm{Var} L_n}}$$

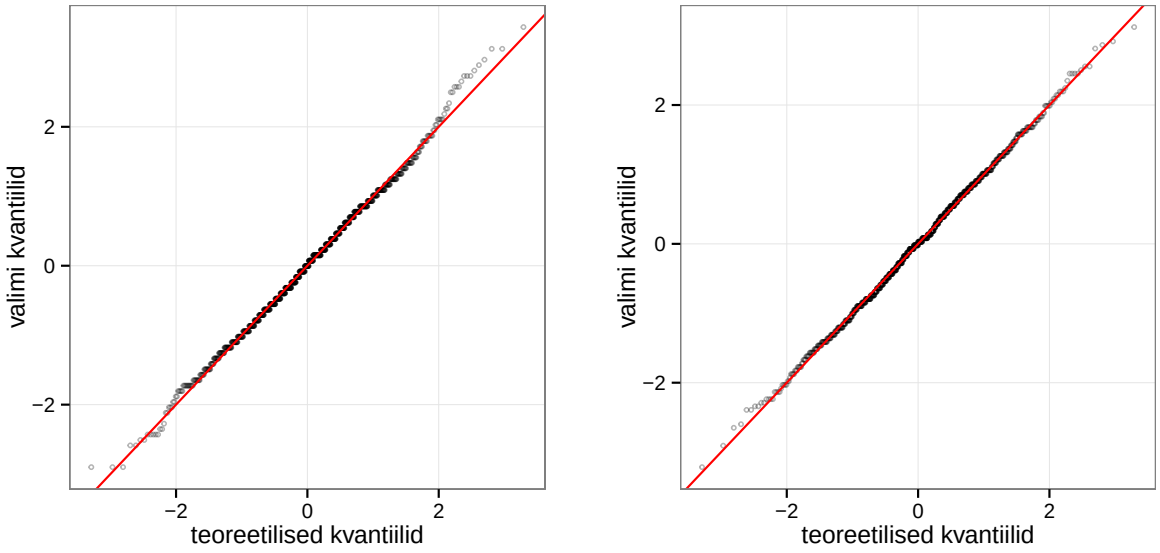
on asümptootiliselt standardse normaaljaotusega. Joonisel 1.7 on $n = 10\,000$ korral võrreldud ülaltoodud statistiku valimikvantiile standardse normaaljaotuse kvantiilidega. Näeme, et nii sõltumatuse (a) kui ka sõltuvuse (b) korral on valimikvantiilid üpriski lähedased standardse normaaljaotuse kvantiilidele. Joonis seega kinnitab hüpoteesi HÜP 1.

Hüpotees HÜP 2 ütles, et jada $(\mathrm{Var}L_n/n)$ koondub mingiks positiivseks konstandiks. Joonisel 1.8 on näha suuruse $\mathrm{Var}L_n/n$ hinnanguline käitumine. Joonis kinnitab hüpoteesi HÜP 2.

Hüpotees HÜP 3 ütles, et statistik

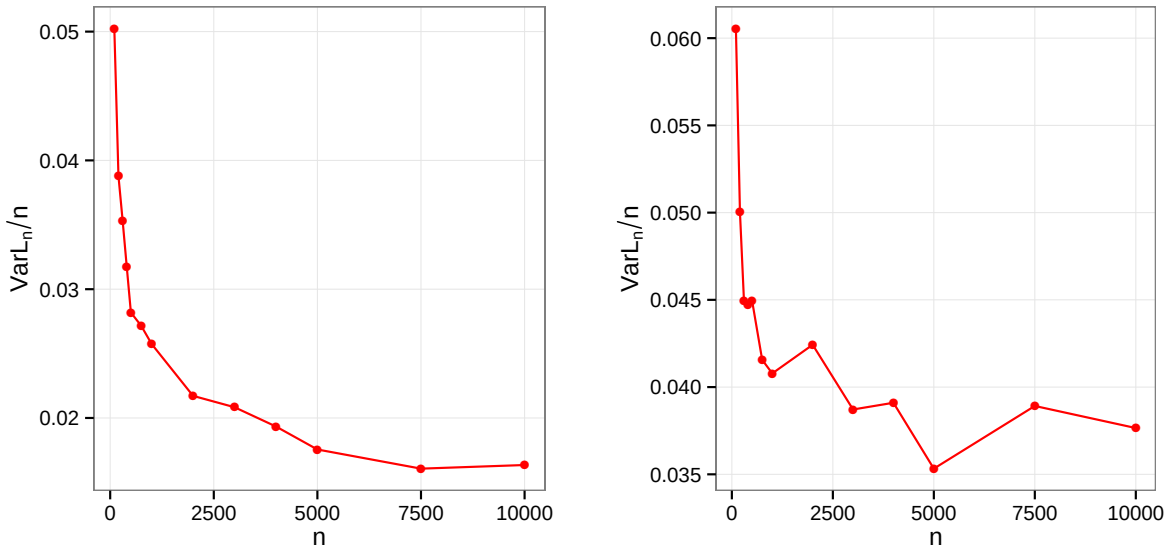
$$\frac{L_n - EL_n}{\sqrt{n}}$$

on asümptootiliselt normaaljaotusega. Selle kontrollimiseks joonistame jälle $n = 10\,000$ korral selle statistiku kvantiilgraafikud. Vastavad graafikud on toodud joonisel 1.9; graafik (a) vastab sõltumatuse juhule ja graafik (b) sõltuvuse juhule. Joonis kinnitab hüpoteesi ülaltoodud statistiku asümptootilise normaaljaotuse kohta.



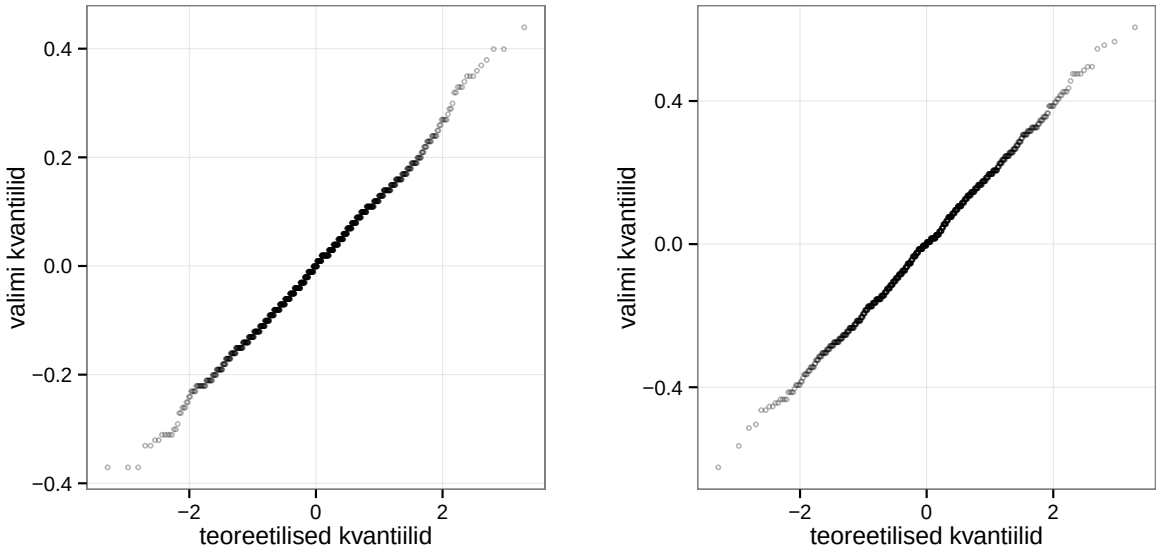
(a) (b)

Joonis 1.7: kvantiilgraafikud hüpoteesi HÜP 1 kontrollimiseks sõltumatute (a) ja sõltuvate (b) jadade korral $(n = 10\,000)$



(a) (b)

Joonis 1.8: suuruse $\mathrm{Var}L_n/n$ hinnanguline käitumine sõltumatute (a) ja sõltuvate (b) jadade korral



(a) (b)

Joonis 1.9: kvantiilgraafikud hüpoteesi HÜP 3 kontrollimiseks sõltumatute (a) ja sõltuvate (b) jadade korral ($n = 10\,000$)

1.6.2 Sõltuvusmõõtude võrdlus karpdiagrammide abil

Järgnevalt uurime karpdiagrammide abil, kui hästi suudavad erinevad sõltuvusmõõdud sõltuvusklasse eristada. Vaatleme nelja sõltuvusmõõtu:

1. empiiriline vastastikune informatsioon;
2. astakkorrelatsioonikordaja "Kendalli τ" absoluutväärtus (korrelatsioonikordaja arvutamiseks nummerdatakse tähestiku $\mathcal{A}$ tähted);
3. pikima ühisjada pikkusel põhinev statistik $L_{n\wedge m}/(n \wedge m)$;
4. ekstremaalsete joonduste Hausdorffi kaugus (eukleidilises mõttes).

Meenutame, et esimest kahte sõltuvusmõõtu on täpsemalt kirjeldatud jaotises 1.3. Edaspidi nimetame pikima ühisjada pikkust ka lühidalt LLCS-iks (vastava ingliskeelse termini *length of longest common subsequence* järgi). Loendis kolmandat sõltuvusmõõtu nimetame seetõttu ka LLCS-statistikuks. Loendis neljanda sõltuvusmõõdu kohta ütleme lühidalt HD (ingliskeelse termini *Hausdorff distance* järgi). X- ja Y-jadad genereerime järgmisest neljast klassist:

A sõltumatud jadad pikkusega 100;

B $\epsilon = 0.05$, juhuslike pikkustega jadad, eellasjada pikkus 105 (keskmine järglasjada pikkus $105 \cdot 0.95 \approx 100$);

C $\epsilon = 0.02$, juhuslike pikkustega jadad, eellasjada pikkus 105;

D sõltumatud jadad pikkusega 100, $P = (0.05, 0.05, 0.05, 0.85)$.

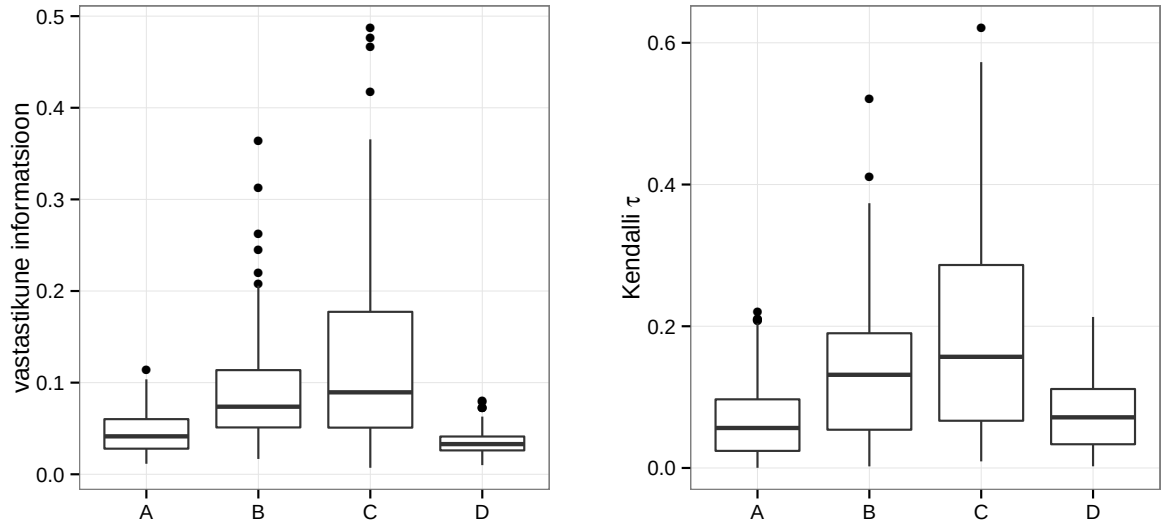
Klassi A ja D kuuluvad sõltumatute jadade paarid, klassi B "keskmiselt" sõltuvate jadade paarid ning klassi C "tugevalt" sõltuvate jadade paarid. Klasside A,B,C korral on P ühtlane, klassi D korral ebaühtlane. Kõigist neljast klassist genereerime 100 jadapaari. Tulemused on toodud joonisel 1.10.

Interpreteerime kõigepealt ühtlasele jaotusele – st klassidele A,B,C – vastavaid karpdiagramme. Näeme, et kõige paremini suudab sõltuvusklasse eristada LLCS-statistik – see on ka ainus sõltuvusmõõt, mille korral klasside A ja C karpdiagrammid üldse ei kattu. Paremuselt järgmiselt suuda klasse eristada HD. Vastastikune informatsioon ja Kendalli τ eristavad klasse enam-vähem sama hästi.

Vaatleme nüüd klassi D. Et see klass vastab sõltumatusele, siis ideaalis peaks see enam-vähem kattuma teise sõltumatuse klassiga A. Näeme, et vastastikuse informatsiooni ja Kendalli τ puhul see nii ongi. LLCS-statistik on aga jaotuse muutuse suhtes väga tundlik – see ei suuda eristada klassi D tugeva sõltuvuse klassist C. HD käitub jaotuse muutuse suhtes natuke paremini kui LLCS-statistik – selle puhul kattub klassi D karpdiagramm klassi B omaga.

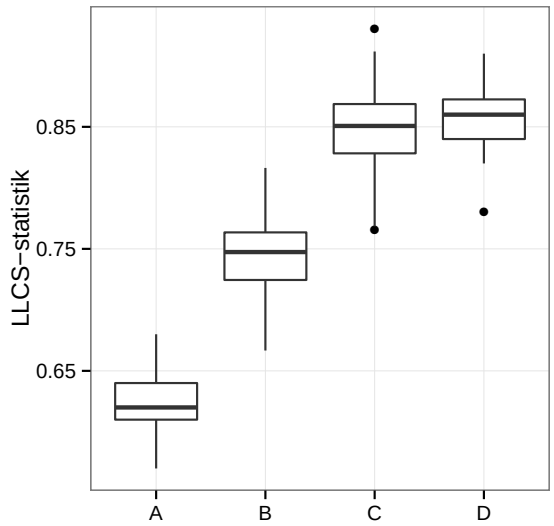
Genereerime veelkord igast klassist 100 jadapaari, võttes seekord keskmiseks jadapikkuseks 1000. Täpsemalt: klasside A ja D korral on jadapikkus 1000, ülejäänud klasside korral on järglasjadade pikkus juhuslik ja eellasjada pikkus 1053. Vastavad karpdiagrammid on toodud joonisel 1.11. Märgime, et joonisel 1.11d on üheksa vaatluspunkti klassist A jäänud joonise ülemisest piirist väljaspoole.

Paneme tähele, et kuigi keskmine jadapikkus on võrreldes eelmise joonisega 10 korda suurem, on vastastikuse informatsiooni ja Kendalli τ suutlikkus sõltuvusklasse eristada

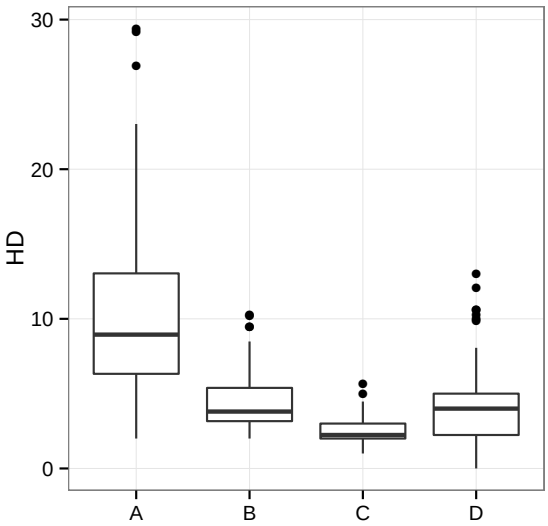


(a)

(b)

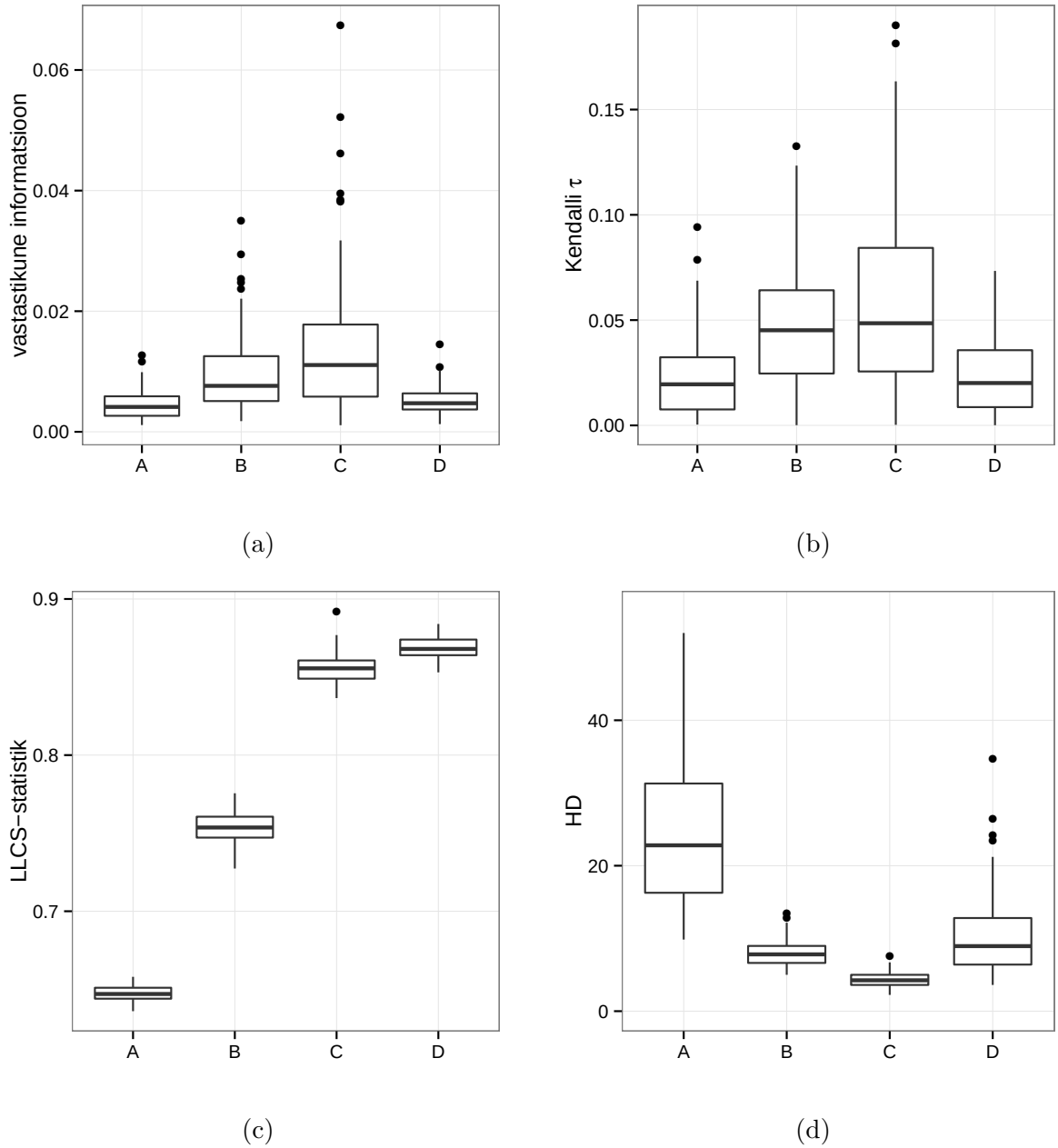


(c)



(d)

Joonis 1.10: erinevate sõltuvusmõõtude karpdiagrammid – keskmine jadapikkus 100



(a) (b)

(c) (d)

Joonis 1.11: erinevate sõltuvusmõõtude karpdiagrammid – keskmine jadapikkus 1000

enam-vähem sama. Samas LLCS-statistiku ja HD puhul on näha sõltuvusklasside eristusvõime olulist paranemist. Seda võib põhjendada järgmiselt. Meenutame (vt jaotis 1.1), et kui $p < 1$, siis

$$\mathbf{P}(X_k \text{ ja } Y_k \text{ on sugulased}) \xrightarrow{k} 0.$$

Samas just sugulased sisaldavadki informatsiooni jadade X^n, Y^m sõltuvuse suuruse kohta. Seega piisavalt suure k korral paarid (X_k, Y_k) enam informatsiooni jadade X^n, Y^m sõltuvuse kohta eriti ei sisalda (kui $p < 1$). Vastastikune informatsioon ja Kendalli τ erinevalt LLCS-statistikust ja HD-st aga kasutavadki vaid paarides (X_k, Y_k) olevat informatsiooni.

Kokkuvõttes võib öelda, et sagedustabelipõhistel statistikutel ei näi olevat väga head asümptootilised omadused. Lisaks, LLCS-statistikul näis ühtlase P korral olevat teistest statistikutest parem eristusvõime. Ebaühtlase P korral võib aga HD paremini toimida kui LLCS-statistik.

Peatükk 2

Pikima ühisjada pikkuse dispersiooni hindamine

2.1 Sõltumatud jadad

2.1.1 Dispersiooni alumisest tõkkest

Üks hüpotees jaotisest 1.4.2 ütles, et jada $(\mathrm{Var}L_n/n)$ koondub positiivseks konstandiks. Tarvilik tingimus selliseks koondumiseks on, et

$$\mathrm{Var}L_n = O(n) \quad \text{ja} \quad \mathrm{Var}L_n = \Omega(n),$$

st dispersioon on ülalt ja alt asümptootiliselt tõkestatud lineaarse tõkkega (meenutame, et mingi funktsiooni f korral tähendab $f(n) = \Omega(n)$, et $n = O[f(n)]$).

Olgu $T_1, T_2, ..., S_1, S_2, ...$ sõltumatud, kuid mitte tingimata sama jaotusega juhuslikud suurused, mis võtavad väärtusi tähestikust $\mathcal{A}$. Defineerime

$$L^*_{n,m} := L(T_1, ..., T_n; S_1, ..., S_m),$$
$$L^*_n := L(T_1, ..., T_n; S_1, ..., S_n).$$

Meie eesmärgiks on uurida juhusliku suuruse $L^*_{n,m}$ dispersiooni. Nimetatud dispersiooni alumine tõke on osutunud raskemini uuritavaks kui ülemine tõke. Tõke

$$\mathrm{Var}L^*_n = \Omega(n)$$

on tõestatud järgmisel kolmel juhul:

- $T_1, ..., T_n$ on *iid* jada, mille tähed on Bernoulli jaotusega parameetriga $\frac{1}{2}$, ning $S_1, ..., S_n$ on mittejuhuslik perioodiline binaarne jada [12];
- $T_1, ..., T_n$ on *iid* jada, mille tähed on Bernoulli jaotusega parameetriga $\frac{1}{2}$, ning $S_1, ..., S_n$ on kolmetähelisest jaotusest tähtedega *iid* jada [13];
- $T_1, ..., T_n, S_1, ..., S_n$ on *iid* jada, mille tähed on väga ebaühtlasest Bernoulli jaotusest [11].

Ülaltoodud loetelus viimane punkt on artiklis [14] üldistatud juhule, kus tähed on küll väga ebaühtlasest jaotusest, kuid tähestiku suurus võib olla suurem kui 2.

Järgnevalt konstrueerime juhusliku suuruse $L^*_{n,m}$ dispersioonile ülemisi tõkkeid.

2.1.2 Funktsiooni L tõkestatud muutude omadus

Me alustame järgmise definitsiooniga.

Definitsioon 2.1 (funktsiooni tõkestatud muutude omadus)**.** *Ütleme, et funktsioonil $f : \mathcal{X}^n \to \mathbb{R}$ on tõkestatud muutude omadus (ingl. k.* bounded differences property*) konstantidega $c_1, ..., c_n$, kui*

$$\sup_{x_1,...,x_n,x' \in \mathcal{X}} |f(x_1, ..., x_n) - f(x_1, ..., x_{i-1}, x', x_{i+1}, ..., x_n)| \leq c_i, \quad i = 1, ..., n.$$

Kui $c_1 = ... = c_n = c$, siis ütleme, et funktsioonil on tõkestatud muutude omadus konstandiga c.

Teisisõnu, funktsioonil on tõkestatud muutude omadus, kui funktsiooni ühe argumendi muutmisel muutub funktsiooni väärtus maksimaalselt mingi fikseeritud konstandi võrra.

Olgu $x_1, ..., x_n, y_1, ..., y_m \in \mathcal{A}$. Edaspidi kasutame tähistusi

$$x^n := x_1, ..., x_n,$$
$$y^m := y_1, ..., y_m.$$

Alljärgnevalt näitame, et pikima ühisjada pikkusel on tõkestatud muutude omadus konstandiga 1.

Lause 2.1 (funktsiooni L tõkestatud muutude omadus)**.** *Iga $z \in \mathcal{A}$ korral*

$$|L(x^n; y^m) - L(x_1, ..., x_{i-1}, z, x_{i+1}, ..., x_n; y^m)| \leq 1, \quad i = 1, ..., n, \tag{2.1}$$

ja

$$|L(x^n; y^m) - L(x^n; y_1, ..., y_{i-1}, z, y_{i+1}, ..., y_m)| \leq 1, \quad i = 1, ..., m. \tag{2.2}$$

Tõestus. Olgu $i \in \{1, ..., n\}$ fikseeritud. Tähistame

$$x' := x'_1, ..., x'_n := x_1, ..., x_{i-1}, z, x_{i+1}, ..., x_n,$$
$$t := L(x^n; y^m).$$

Seega jada x' võib jadast x^n erineda ainult i-nda elemendi poolest. Kui $t = 0$, siis võrratused (2.1) ja (2.2) on triviaalsed. Olgu nüüd $t \geq 1$ ja

$$C := \{(i_1, j_1), ..., (i_t, j_t)\}$$

jadade x^n, y^m (mingi) optimaalne joondus. Juhul, kui $i \notin \{i_1, ..., i_t\}$, on C jadade x', y^m joondus ja sellisel juhul

$$L(x'; y^m) \geq L(x^n; y^m).$$

Kui aga $i \in \{i_1, ..., i_t\}$ ehk leidub $s \in \{1, ..., t\}$ nii, et $i = i_s$, siis $C \setminus \{(i_s, j_s)\}$ on jadade x', y^m joondus ja

$$L(x'; y^m) \geq L(x^n; y^m) - 1.$$

Seega alati

$$L(x^n; y^m) - L(x'; y^m) \leq 1.$$

Ülaltooduga täpselt samasugust argumentatsiooni kasutades saab näidata, et ka

$$L(x';y^m)-L(x^n;y^m)\le 1.$$

Kokkuvõttes

$$|L(x^n;y^m)-L(x';y^m)|\le 1.$$

Võrratused (2.1) on tõestatud. Võrratused (2.2) kehtivad tänu funktsiooni L sümmeetriale. □

2.1.3 Ülemine tõke dispersioonile McDiarmidi võrratuse abil

Tuletame meelde McDiarmidi võrratuse.

Teoreem 2.1 (McDiarmidi võrratus)**.** [15, lk 4] *Olgu $W_1,...,W_n$ sõltumatud juhuslikud suurused, mis võtavad väärtusi hulgast $\mathcal{X}$. Kui funktsioonil $f:\mathcal{X}^n\to\mathbb{R}$ on tõkestatud muutude omadus konstantidega $c_1,...,c_n$, siis iga $t>0$ korral*

$$\mathbf{P}\left(|f(W_1,...,W_n)-Ef(W_1,...,W_n)|\ge t\right)\le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Tänu funktsiooni L tõkestatud muutude omadusele seega

$$\mathbf{P}\left(|L^*_{n,m}-EL^*_{n,m}|\ge t\right)\le 2\exp\left(-\frac{2t^2}{n+m}\right)\quad\forall t>0.$$

Meenutame, et kui W on mittenegatiivne juhuslik suurus, siis

$$EW=\int_0^\infty \mathbf{P}(W>t)\,dt.$$

Seega

$$\begin{aligned}\mathrm{Var}L^*_{n,m}&=\int_0^\infty \mathbf{P}\left(|L^*_{n,m}-EL^*_{n,m}|>\sqrt{t}\right)dt\\&\le\int_0^\infty 2\exp\left(-\frac{2t}{n+m}\right)dt\\&=-\frac{2(n+m)}{2}\exp\left(-\frac{2t}{n+m}\right)\Big|_{t=0}^{\infty}\\&=n+m.\end{aligned}\tag{2.3}$$

Saadud tõket saab parandada, arvestades, et leidub $x>0$ nii, et iga $t\in[0,x)$ korral funktsioon $2\exp\left(-\frac{2t}{n+m}\right)$ on ühest suurem. Leiame punkti x:

$$2\exp\left(-\frac{2x}{n+m}\right)=1\quad\Leftrightarrow\quad-\frac{2x}{n+m}=\ln\frac{1}{2}\quad\Leftrightarrow\quad x=\frac{1}{2}(n+m)\ln 2.$$

Ülemiseks tõkkeks saame

$$\begin{aligned}\mathrm{Var}L^*_{n,m} &\le \int_0^x 1\,dt + \int_x^\infty 2\exp\left(-\frac{2t}{n+m}\right)dt\\ &= x - \frac{2(n+m)}{2}\exp\left(-\frac{2t}{n+m}\right)\Big|_{t=x}^{\infty}\\ &= \frac{1}{2}(n+m)\ln 2 + \frac{(n+m)}{2}\\ &= \frac{1+\ln 2}{2}(n+m) < 0.847(n+m). \end{aligned} \tag{2.4}$$

2.1.4 Ülemine tõke kõikidele momentidele McDiarmidi võrratuse abil

Ülaltoodut saame üldistada ka teistele momentidele. Juhusliku suuruse W k-ndat tsentreeritud absoluutset momenti tähistame järgmiselt:

$$m_k(W) := E|W - EW|^k.$$

Leiame ülemise tõkke:

$$m_k(L^*_{n,m}) = \int_0^\infty \mathbf{P}\left(|L_{n,m} - EL_{n,m}| > t^{\frac{1}{k}}\right)dt \le 2\int_0^\infty \exp\left(-\frac{2t^{\frac{2}{k}}}{n+m}\right)dt \quad \forall k > 0.$$

Muutuja vahetuse

$$\begin{aligned} u &= \frac{2t^{\frac{2}{k}}}{n+m}, \quad t = \left(\frac{u(n+m)}{2}\right)^{\frac{k}{2}}, \quad du = \frac{4t^{\frac{2}{k}-1}}{k(n+m)}dt,\\ dt &= \frac{k(n+m)}{4}\left(\frac{u(n+m)}{2}\right)^{\frac{k}{2}-1}du = \frac{k}{2\cdot 2^{\frac{k}{2}}}(n+m)^{\frac{k}{2}}u^{\frac{k}{2}-1}du \end{aligned}$$

abil saame

$$m_k(L^*_{n,m}) \le \frac{k}{2^{\frac{k}{2}}}(n+m)^{\frac{k}{2}}\int_0^\infty e^{-u}u^{\frac{k}{2}-1}du = \frac{k}{2^{\frac{k}{2}}}\Gamma\left(\frac{k}{2}\right)(n+m)^{\frac{k}{2}} \quad \forall k > 0.$$

Juhul $k = 2$ saame siit tõkke (2.3) ehk $n+m$.

Jällegi, leidub $x_k > 0$ nii, et iga $t \in [0, x_k)$ korral funktsioon $2\exp\left(-\frac{2t^{\frac{2}{k}}}{n+m}\right)$ on ühest suurem. Leiame punkti x_k:

$$\begin{aligned} 2\exp\left(-\frac{2x_k^{\frac{2}{k}}}{n+m}\right) = 1 \quad &\Leftrightarrow \quad -\frac{2x_k^{\frac{2}{k}}}{n+m} = \ln\frac{1}{2}\\ &\Leftrightarrow \quad x_k^{\frac{2}{k}} = -\frac{1}{2}(n+m)\ln\frac{1}{2}\\ &\Leftrightarrow \quad x_k = \left(\frac{1}{2}(n+m)\ln 2\right)^{\frac{k}{2}}. \end{aligned}$$

Kasutades sama muutuja vahetust nagu eelmise integraali juures, saame parandatud ülemiseks tõkkeks

$$\begin{aligned} m_k(L^*_{n,m}) &\le x_k + 2\int_{x_k}^{\infty} \exp\left(-\frac{2t^{\frac{2}{k}}}{n+m}\right) dt \\ &= \left(\frac{1}{2}(n+m)\ln 2\right)^{\frac{k}{2}} + \frac{k}{2^{\frac{k}{2}}}(n+m)^{\frac{k}{2}} \int_{\ln 2}^{\infty} e^{-u} u^{\frac{k}{2}-1}\, du \\ &= \frac{1}{2^{\frac{k}{2}}}\left((\ln 2)^{\frac{k}{2}} + k\int_{\ln 2}^{\infty} e^{-u} u^{\frac{k}{2}-1}\, du\right)(n+m)^{\frac{k}{2}} \quad \forall k > 0. \end{aligned} \tag{2.5}$$

Lugeja võib veenduda, et juhul $k = 2$ saame siit tõkke (2.4).

2.1.5 Ülemine tõke dispersioonile Efron-Stein'i võrratuse abil

Järgmine tulemus võimaldab saada dispersioonile parema ülemise tõkke kui McDiarmidi võrratus.

Teoreem 2.2 (Efron-Stein'i võrratus)**.** [15, Theorem 3.1] *Olgu $W_1, ..., W_n$ sõltumatud juhuslikud suurused, mis võtavad väärtusi hulgast $\mathcal{X}$. Olgu $f : \mathcal{X}^n \to \mathbb{R}$ selline funktsioon, et juhuslikul suurusel $F := f(W_1, ..., W_n)$ leidub lõplik teine moment. Olgu $W'_1, ..., W'_n$ juhuslike suuruste $W_1, ..., W_n$ sõltumatud koopiad. Defineerime*

$$F'_i := f(W_1, ..., W_{i-1}, W'_i, W_{i+1}, ..., W_n), \quad i = 1, ..., n.$$

Siis

$$\mathrm{Var}F \le \frac{1}{2}\sum_{i=1}^{n} E(F - F'_i)^2 \tag{2.6}$$

ning

$$\mathrm{Var}F \le \inf_{F_i} \sum_{i=1}^{n} E(F - F_i)^2, \tag{2.7}$$

kus infimum *on võetud üle kõigi $\sigma(W_1, ..., W_{i-1}, W_{i+1}, ..., W_n)$-mõõtuvate juhuslike suuruste F_i, millel leidub lõplik teine moment.*

Järgmine järeldus aitab meil ülaltoodud teoreemi rakendada.

Järeldus 2.1. [15, Corollary 3.2] *Olgu $W_1, ..., W_n$ sõltumatud juhuslikud suurused, mis võtavad väärtusi hulgast $\mathcal{X}$. Olgu funktsioonil $f : \mathcal{X}^n \to \mathbb{R}$ tõkestatud muutude omadus konstantidega $c_1, ..., c_n$ ning $F := f(W_1, ..., W_n)$. Siis*

$$\mathrm{Var}F \le \frac{1}{4}\sum_{i=1}^{n} c_i^2.$$

Tõestus. Tõkestatud muutude omadusest järelduvalt

$$\sup_{x,y \in \mathcal{X}^n} |f(x) - f(y)| \le \sum_{i=1}^{n} c_i. \tag{2.8}$$

Seega juhuslik suurus F on tõkestatud ja tal leidub lõplik teine moment. Defineerime iga $i \in \{1,...,n\}$ korral

$$F_i := \frac{1}{2}\left(\sup_{w' \in \mathcal{X}} f(W_1,...,W_{i-1},w',W_{i+1},...,W_n) + \inf_{w' \in \mathcal{X}} f(W_1,...,W_{i-1},w',W_{i+1},...,W_n)\right).$$

Seose (2.8) põhjal on ka juhuslikud suurused $F_1,...,F_n$ tõkestatud ja neil leidub lõplik teine moment. Lisaks F_i on $\sigma(W_1,...,W_{i-1},W_{i+1},...,W_n)$-mõõtuv iga $i \in \{1,...,n\}$ korral. Nüüd saamegi Efron-Stein'i võrratusest kujul (2.7), et

$$\mathrm{Var}F \leq \sum_{i=1}^{n} E(F-F_i)^2 \leq \sum_{i=1}^{n}\left(\frac{c_i}{2}\right)^2 = \frac{1}{4}\sum_{i=1}^{n} c_i^2.$$

□

Rakendades ülaltoodud järeldust funktsioonile L, saame

$$\mathrm{Var}L_{n,m}^* \leq \frac{n+m}{4}.$$

Kui $n = m$, lihtsustub ülaltoodud tõke kujule

$$\mathrm{Var}L_n^* \leq \frac{n}{2}. \tag{2.9}$$

2.1.6 Ülemine tõke dispersioonile *iid* juhul

Käesolevas jaotises eeldame, et jadad $T_1,...,T_n$ ja $S_1,...,S_m$ on mõlemad *iid* jadad vastavate marginaaljaotustega P_T, P_S. Olgu $T_1',...,T_n',S_1',...,S_m'$ juhuslike suuruste $T_1,...,T_n,S_1,...,S_m$ sõltumatud koopiad. Defineerime

$$\begin{aligned} L_i^T &:= L(T_1,...,T_{i-1},T_i',T_{i+1},...,T_n;\, S_1,...,S_m), \quad i = 1,...,n \\ L_i^S &:= L(T_1,...,T_n;\, S_1,...,S_{i-1},S_i',S_{i+1},...,S_m), \quad i = 1,...,m. \end{aligned}$$

Rakendades Efron-Stein'i võrratust kujul (2.6), saame

$$\begin{aligned} \mathrm{Var}L_{n,m}^* &\leq \frac{1}{2}\left(\sum_{i=1}^{n} E(L_{n,m}^* - L_i^T)^2 + \sum_{i=1}^{m} E(L_{n,m}^* - L_i^S)^2\right) \\ &= \frac{1}{2}\left(\sum_{i=1}^{n} \mathbf{P}(|L_{n,m}^* - L_i^T| = 1) + \sum_{i=1}^{m} \mathbf{P}(|L_{n,m}^* - L_i^S| = 1)\right) \\ &\leq \frac{1}{2}\left(\sum_{i=1}^{n}[1-\mathbf{P}(T_i = T_i')] + \sum_{i=1}^{m}[1-\mathbf{P}(S_i = S_i')]\right) \\ &= \frac{1}{2}\left(n+m-\sum_{a\in\mathcal{A}}\left(nP_T(a)^2 + mP_S(a)^2\right)\right). \end{aligned}$$

Kui $n = m$ ja $P_T = P_S =: P^*$, saame siit tõkke

$$\mathrm{Var}L_n^* \leq n\left(1 - \sum_{a\in\mathcal{A}} P^*(a)^2\right). \tag{2.10}$$

Selle tõkke tõestas Steele juba aastal 1986 [16]. Juhul, kui jaotus P^* on ühtlane, saame ülaltoodust

$$\mathrm{Var}L_n^* \leq n\left(1-\frac{1}{|\mathcal{A}|}\right),$$

mis ei ole parem tõke kui (2.9). Juhul, kui jaotus P^* on ebaühtlane, võib tõke (2.10) anda ka parema hinnangu kui (2.9).

2.1.7 Funktsiooni L ennasttõkestavus ja ülemine tõke tema dispersioonile läbi keskväärtuse

Toome sisse ennasttõkestava funktsiooni mõiste.

Definitsioon 2.2 (funktsiooni ennasttõkestavus)**.** *Ütleme, et mittenegatiivne funktsioon $f:\mathcal{X}^n\to[0,\infty)$ on ennasttõkestav (ingl. k.* self-bounding*), kui leiduvad sellised funktsioonid $f_i:\mathcal{X}^{n-1}\to\mathbb{R}$ ja positiivne konstant B, et kõikide $x_1,...,x_n\in\mathcal{X}$ ja iga $i\in\{1,...,n\}$ korral*

$$0\leq f(x_1,...,x_n)-f_i(x_1,...,x_{i-1},x_{i+1},...,x_n)\leq 1$$

ja

$$\sum_{i=1}^{n}[f(x_1,...,x_n)-f_i(x_1,...,x_{i-1},x_{i+1},...,x_n)]\leq Bf(x_1,...,x_n).$$

Järgmine järeldus tuleneb Efron-Steini võrratusest.

Järeldus 2.2. *Olgu $W_1,...,W_n$ sõltumatud juhuslikud suurused, mis võtavad väärtusi hulgast $\mathcal{X}$. Lisaks olgu funktsioon $f:\mathcal{X}^n\to[0,\infty)$ ennasttõkestav funktsioonidega $f_1,...,f_n$ ja konstandiga B. Eeldame, et juhuslikul suurusel $F:=f(W_1,...,W_n)$ ning juhuslikel suurustel*

$$F_i:=f_i(W_1,...,W_{i-1},W_{i+1},...,W_n),\quad i=1,...,n,$$

leidub lõplik teine moment. Siis

$$\mathrm{Var}F\leq B\cdot EF.$$

Tõestus. Ennasttõkestava funktsiooni definitsioonist järelduvalt

$$\sum_{i=1}^{n}(F-F_i)^2\leq BF.$$

Võttes nüüd mõlemast poolest keskväärtuse ja rakendades Efron-Stein'i võrratust kujul (2.7), saamegi otsitud seose. □

Toome sisse mõned uued tähistused. Defineerime

$$x^{(i)}:=x_1^{(i)},...,x_{n-1}^{(i)}:=x_1,...,x_{i-1},x_{i+1},...,x_n,\quad i=1,...,n$$
$$y^{(i)}:=y_1^{(i)},...,y_{n-1}^{(i)}:=y_1,...,y_{i-1},y_{i+1},...,y_m,\quad i=1,...,m.$$

Mis tahes joonduse C korral defineerime funktsioonid

$$x(C) := \{i \,|\, (i,j) \in C\},$$
$$y(C) := \{j \,|\, (i,j) \in C\}.$$

Seega $i \in x(C)$ parajasti siis, kui leidub $j \in \mathbb{N}$ nii, et $(i,j) \in C$. Analoogiliselt $j \in y(C)$ parajasti siis, kui leidub $i \in \mathbb{N}$ nii, et $(i,j) \in C$.

Suvalise hulga A korral on indikaatorfunktsioon I_A defineeritud järgmiselt:

$$I_A(x) := \begin{cases} 1, & \text{kui } x \in A, \\ 0, & \text{vastasel korral.} \end{cases}$$

Meenutame veel, et hulgale $A \subset \mathbb{N}^2$ elemendi $(i,j) \in \mathbb{N}^2$ liitmise defineerisime järgmiselt:

$$A + (i,j) = \{(i+x, j+y) \,|\, (x,y) \in A\}.$$

Järgmine lemma ütleb, et funktsiooni L väärtus ei suurene, kui eemaldada üks argumentidest.

Lemma 2.1. *Kehtivad võrratused*

$$L(x^n; y^m) \geq L(x^{(i)}; y^m), \quad i = 1, ...n, \tag{2.11}$$
$$L(x^n; y^m) \geq L(x^n; y^{(i)}), \quad i = 1, ...m. \tag{2.12}$$

Tõestus. Olgu $i \in \{1, ..., n\}$ fikseeritud ja

$$t := L(x^{(i)}; y^m).$$

Juhul $t = 0$ on väide triviaalne. Vaatleme juhtu, kus $t \geq 1$. Olgu

$$C := \{(i_1, j_1), ..., (i_t, j_t)\}$$

jadade $x^{(i)}, y^m$ optimaalne joondus. Defineerime

$$C_1 := \{(x,y) \in C \,|\, x < i\}$$
$$C_2 := \{(x,y) \in C \,|\, x \geq i\}$$

Seega $C_1 \cup C_2 = C$. Hulk

$$C_1 \cup [C_2 + (1,0)]$$

on jadade x^n, y^m joondus. Võrratused (2.11) on tõestatud. Võrratused (2.12) järelduvad funktsiooni L sümmeetriast. □

Järgmine lemma ütleb muuhulgas, et funktsiooni L väärtus väheneb maksimaalselt ühe võrra, kui eemaldada üks argumentidest.

Lemma 2.2. *Olgu C jadade x^n, y^m suvaline optimaalne joondus. Kehtivad võrratused*

$$0 \leq L(x^n; y^m) - L(x^{(i)}; y^m) \leq I_{x(C)}(i) \leq 1, \quad i = 1, ...n, \tag{2.13}$$
$$0 \leq L(x^n; y^m) - L(x^n; y^{(i)}) \leq I_{y(C)}(i) \leq 1, \quad i = 1, ...m. \tag{2.14}$$

Tõestus. Fikseerime $i \in \{1,...,n\}$. Olgu

$$t := L(x^n, y^m).$$

Kui $t = 0$, siis lemma 2.1 kohaselt

$$L(x^{(i)}; y^m) = L(x^n; y^{(i)}) = 0.$$

Olgu nüüd $t \geq 1$ ja

$$C := \{(i_1, j_1), ..., (i_t, j_t)\}$$

jadade x^n, y^m (mingi) optimaalne joondus. Lisaks olgu

$$C_1 := \{(x, y) \in C \,|\, x < i\},$$
$$C_2 := \{(x, y) \in C \,|\, x \geq i\}.$$

Seega $C_1 \cup C_2 = C$.

Oletame, et $i \notin \{i_1, ..., i_t\}$. Siis hulk

$$C_1 \cup [C_2 - (1, 0)]$$

on jadade $x^{(i)}, y^m$ joondus. Seega

$$L(x^{(i)}; y^m) \geq L(x^n, y^m).$$

Ülaltoodust ja lemmast (2.1) saame

$$L(x^n, y^m) - L(x^{(i)}; y^m) = 0.$$

Oletame nüüd, et $i \in \{i_1, ..., i_t\}$, st leidub $s \in \{1, ..., t\}$ nii, et $i = i_s$. Hulk

$$C_1 \cup [C_2 \setminus \{(i_s, j_s)\} - (1, 0)]$$

on jadade $x^{(i)}, y^m$ joondus. Seega

$$L(x^{(i)}; y^m) \geq L(x^n, y^m) - 1.$$

Ülaltoodust ja lemmast 2.1 saame

$$0 \leq L(x^n, y^m) - L(x^{(i)}; y^m) \leq 1.$$

Võrratused (2.13) on tõestatud. Võrratuste (2.14) tõestus on analoogiline. □

Ülaltoodud lemmaga varustatult on nüüd lihtne tõestada, et funktsioon L on ennasttõkestav.

Lause 2.2. *Funktsioon L on ennasttõkestav konstandiga 2.*

Tõestus. Olgu C jadade x^n, y^m optimaalne joondus. Tänu lemmale 2.2

$$0 \le L(x^n, y^m) - L(x^{(i)}; y^m) \le 1, \quad i = 1, ..., n,$$
$$0 \le L(x^n, y^m) - L(x^n; y^{(i)}) \le 1, \quad i = 1, ..., m$$

ja

$$\sum_{i=1}^{n}[L(x^n, y^m) - L(x^{(i)}; y^m)] + \sum_{i=1}^{m}[L(x^n, y^m) - L(x^n; y^{(i)})] \le \sum_{i=1}^{n} I_{x(C)}(i) + \sum_{i=1}^{m} I_{y(C)}(i)$$
$$= 2L(x^n, y^m).$$

□

Rakendades järeldust (2.2), saame nüüd

$$\mathrm{Var}L^*_{n,m} \le 2EL^*_{n,m}.$$

Juhul, kui jadad X^∞, Y^∞ on sõltumatud ja Chvatal-Sankoffi konstandiga γ, saame tänu triviaalsele seosele $L_{n,m} \le L_{n\vee m}$ ja järelduse 1.1 seosele (1.6)

$$\mathrm{Var}L_{n,m} \le 2EL_{n,m} \le 2EL_{n\vee m} \le 2\gamma(n \vee m).$$

Juhul $n = m$ lihtsustub ülaltoodud tõke kujule

$$\mathrm{Var}L_n \le 2\gamma n.$$

2.2 Sõltuvad jadad

2.2.1 Funktsioon L' ja juhuslik suurus L'_n

Tuleme nüüd tagasi (üldjuhul) sõltuvate jadade X^∞, Y^∞ juurde. Soovime leida ülemise tõkke juhusliku suuruse $L_n = L(X^n; Y^n)$ dispersioonile (lihtsuse huvides vaatame võrdse pikkusega jadasid). Jadade X^n, Y^n sõltuvus toob sisse mõningast tehnilist lisakeerukust, kuid tuleb välja, et ka sõltuvate jadade korral saab dispersiooni $\mathrm{Var}L_n$ hindamiseks kasutada McDiarmidi ja Efron-Stein'i võrratust.

Meenutame, et p on tähe allesjäämise tõenäosus meie mudelis. Fikseerime konstandi δ, $0 < \delta < p$. Defineerime

$$m(n) := \left\lceil \frac{n}{p-\delta} \right\rceil$$
$$n_x(n) := \sum_{i=1}^{m(n)} D_i^x, \quad n_y(n) := \sum_{i=1}^{m(n)} D_i^y,$$
$$X := X^{n_x(n)\wedge n}, \quad Y := Y^{n_y(n)\wedge n},$$
$$U_i := (Z_i, \xi_i, \eta_i, D_i^x, D_i^y), \quad i = 1, 2, ...$$

Et jada $U_1, ..., U_{m(n)}$ sisaldab kogu informatsiooni jadade X, Y kohta, siis leidub funktsioon L' nii, et

$$L'(U_1, ..., U_{m(n)}) \equiv L(X; Y) =: L'_n.$$

Vektorit U_i vaatleme funktsiooni L' ühe argumendina – seega funktsioonil L' on $m(n)$ argumenti. Tõestame, et funktsioonil L' on tõkestatud muutude omadus.

Lemma 2.3. *Funktsioonil L' on tõkestatud muutude omadus konstandiga 2.*

Tõestus. Tõestuseks arutleme natuke selle üle, mis juhtub jadadega X, Y, kui muuta funktsiooni L' ühte argumenti.

Kui eellasjada $Z_1, Z_2, \ldots$ elemendil i ei ole järglast jadas X, siis funktsiooni L' i-nda argumendi muutmine jada X ei mõjuta. Olgu nüüd eellasjada elemendil i jadas X järglane olemas ja olgu selleks i_x. Funktsiooni L' i-nda argumendi muutmine võib endaga tuua kaasa ühe järgmistest muudatustest jadas X:

1. jada elemendi i_x väärtus muutub;
2. jada element i_x kaob, mille tagajärjel lisatakse jada lõppu võib-olla üks element juurde.

Esimene sündmus muudab jadade X, Y pikima ühisjada pikkust maksimaalselt ühe võrra (vt lause 2.1). Oletame nüüd, et toimub teine sündmus. Siis elemendi kadumine kas jätab LLCS-i (pikima ühisjada pikkuse) samaks või vähendab seda ühe võrra (vt lemma 2.2). Samas, elemendi lisamine jada lõppu võib kas LLCS-i muutmata jätta või seda suurendada ühe võrra (järeldub samuti lemmast 2.2). Seega teine sündmus võib samuti LLCS-i muuta maksimaalselt ühe võrra.

Muudatused, mida funktsiooni L' i-nda argumendi muutmine võib endaga tuua kaasa jadas Y on analoogilised ülaltooduga. Seega väide kehtib. □

Tänu funktsiooni L' tõkestatud muutude omadusele, saame dispersiooni $\mathrm{Var}L'_n$ hindamiseks kasutada Efron-Stein'i võrratust. Kasutades järeldust 2.1, saame

$$\mathrm{Var}L'_n \leq m(n) \leq \frac{n}{p-\delta} + 1. \tag{2.15}$$

Tõestame veel järgmise abitulemuse.

Lemma 2.4. *Kõikide $a, b, c \in \mathbb{R}^+$ korral*

$$(a+b)^c \leq 2^{(c-1)\vee 0}(a^c + b^c).$$

Tõestus. Olgu $f(x) := x^c$. Vaatleme kõigepealt juhtu, kus $0 < c \leq 1$. Pole raske veenduda, et sellisel juhul f on nõgus intervallis $(0, \infty)$. Olgu nüüd $x > 0$, $y > 0$ ja $\alpha \in [0,1]$. Nõgususest saame

$$\alpha f(x) + (1-\alpha) f(y) \leq f(\alpha x + (1-\alpha) y).$$

Võtame mõlemast poolest piirväärtuse:

$$\alpha f(x) + (1-\alpha) \lim_{y \to 0+} f(y) \leq \lim_{y \to 0+} f(\alpha x + (1-\alpha) y).$$

Ülaltoodust saame

$$\alpha f(x) \leq f(\alpha x).$$

Kui $a>0$ ja $b>0$, siis eelneva põhjal

$$\begin{aligned} f(a+b) &= \frac{a}{a+b}f(a+b)+\frac{b}{a+b}f(a+b) \\ &\le f\left(\frac{a}{a+b}(a+b)\right)+f\left(\frac{b}{a+b}(a+b)\right) \\ &= f(a)+f(b). \end{aligned}$$

Vaatleme nüüd juhtu, kus $c>1$. Pole raske veenduda, et sellisel juhul f on kumer intervallis $(0,\infty)$. Eeldades jälle, et $a>0$ ja $b>0$, saame

$$2^{-c}f(a+b) = f\left(\frac{1}{2}a+\frac{1}{2}b\right) \le \frac{1}{2}[f(a)+f(b)].$$

Siit saame

$$f(a+b) \le 2^{c-1}[f(a)+f(b)].$$

Lõpuks paneme tähele, et kui $0<c\le 1$, siis $2^{(c-1)\vee 0}=1$, ja kui $c>1$, siis $2^{(c-1)\vee 0}=2^{c-1}$. □

Et funktsioon L' on tõkestatud muutude omadusega, siis saame McDiarmidi võrratuse (teoreem 2.1) abil leida ülemise tõkke juhusliku suuruse L'_n tsentreeritud absoluutsetele momentidele $m_k(L'_n)$. Saame kasutada juba varem leitud valemit (2.5) – peame vaid asendama suuruse $n+m$ suurusega $4m(n)$. Saame

$$m_k(L'_n) \le c_k^*[4m(n)]^{\frac{k}{2}},$$

kus

$$c_k^* := \frac{1}{2^{\frac{k}{2}}}\left((\ln 2)^{\frac{k}{2}}+k\int_{\ln 2}^{\infty} e^{-u}u^{\frac{k}{2}-1}\,du\right) \quad \forall k>0.$$

Kasutades lemmat 2.4, saame seega

$$\begin{aligned} m_k(L'_n) &\le c_k^*\left(4\left\lceil\frac{n}{p-\delta}\right\rceil\right)^{\frac{k}{2}} \\ &\le c_k^*\left[4\left(\frac{n}{p-\delta}+1\right)\right]^{\frac{k}{2}} \\ &\le c_k^* 2^{(\frac{k}{2}-1)\vee 0}\left(\frac{4}{p-\delta}\right)^{\frac{k}{2}} n^{\frac{k}{2}}+o(n^{\frac{k}{2}}) \quad \forall k>0. \end{aligned} \tag{2.16}$$

Defineerides

$$c_k^{**} := c_k^* 2^{(\frac{k}{2}-1)\vee 0}\left(\frac{4}{p-\delta}\right)^{\frac{k}{2}} \quad \forall k>0,$$

saame

$$m_k(L'_n) \le c_k^{**} n^{\frac{k}{2}} + o(n^{\frac{k}{2}}) \quad \forall k > 0. \tag{2.17}$$

Suurus c_2^* avaldub järgmiselt:

$$\begin{aligned} c_2^* &= \frac{1}{2}\left(\ln 2 + 2\int_{\ln 2}^{\infty} e^{-u}\,du\right) \\ &= \frac{1}{2}\left[\ln 2 + 2\left(-e^{-u}\Big|_{u=\ln 2}^{\infty}\right)\right] \\ &= \frac{1}{2}\left(\ln 2 + 2e^{-\ln 2}\right) \\ &= \frac{1+\ln 2}{2}. \end{aligned}$$

Tänu ülaltoodud seosele ja seosele (2.17), saame dispersioonile $\mathrm{Var}L'_n$ järgmise McDiarmidi võrratust kasutava tõkke:

$$\begin{aligned} \mathrm{Var}L'_n &\le c_2^{**} n + o(n) \\ &= c_2^* \cdot \frac{4}{p-\delta} \cdot n + o(n) \\ &= \frac{2(1+\ln 2)}{p-\delta} \cdot n + o(n). \end{aligned} \tag{2.18}$$

Tõestame nüüd, et funktsioon $\frac{L'}{2}$ on ennasttõkestav.

Lemma 2.5. *Funktsioon $\frac{L'}{2}$ on ennasttõkestav konstandiga 2.*

Tõestus. Defineerime iga $i \in \{1,...,m(n)\}$ korral funktsiooni $L^{(i)}$ järgmiselt:

$$\begin{aligned} &L^{(i)}(U_1,...,U_{i-1},U_{i+1},...,U_{m(n)}) \\ &:= \begin{cases} L(X;Y), \\ \text{kui eellasel } i \text{ ei leidu ei jadas } X \text{ ega ka jadas } Y \text{ järglast;} \\ L(X_1,...,X_{i_x(i)-1},X_{i_x(i)+1},...,X_{n\wedge n_x(n)};Y), \\ \text{kui eellasel } i \text{ leidub jadas } X \text{ järglane } i_x(i), \text{ aga ei leidu järglast jadas } Y; \\ L(X;Y_1,...,Y_{i_y(i)-1},Y_{i_y(i)+1},...,Y_{n\wedge n_y(n)}), \\ \text{kui eellasel } i \text{ leidub jadas } Y \text{ järglane } i_y(i), \text{ aga ei leidu järglast jadas } X; \\ L(X_1,...,X_{i_x(i)-1},X_{i_x(i)+1},...,X_{n\wedge n_x(n)};Y_1,...,Y_{i_y(i)-1},Y_{i_y(i)+1},...,Y_{n\wedge n_y(n)}), \\ \text{kui eellasel } i \text{ leiduvad jadas } X,Y \text{ vastavalt järglased } i_x(i), i_y(i). \end{cases} \end{aligned}$$

Lugeja võib võtta hetke aega veendumaks, et selline funktsioon leidub. Olgu C jadade X,Y suvaline optimaalne joondus. Edasi pole raske veenduda lemma 2.2 abil, et

$$0 \le \frac{L'(U_1,...,U_{m(n)})}{2} - \frac{L^{(i)}(U_1,...,U_{i-1},U_{i+1},...,U_{m(n)})}{2} \le 1, \quad i = 1,...,m(n),$$

ja

$$\sum_{i=1}^{m(n)}\left(\frac{L'(U_1,...,U_{m(n)})}{2}-\frac{L^{(i)}(U_1,...,U_{i-1},U_{i+1},...,U_{m(n)})}{2}\right)$$
$$\leq\frac{1}{2}\left(\sum_{i=1}^{n\wedge n_x(n)}I_{x(C)}(i)+\sum_{i=1}^{n\wedge n_y(n)}I_{y(C)}(i)\right)$$
$$=2\cdot\frac{L'(U_1,...,U_{m(n)})}{2}.$$

□

Järeldusest 2.2 saame seega

$$\mathrm{Var}\left(\frac{L'_n}{2}\right)\leq 2E\left(\frac{L'_n}{2}\right).$$

Seega

$$\mathrm{Var}L'_n\leq 4EL'_n. \tag{2.19}$$

2.2.2 Ülemine tõke juhusliku suuruse L_n dispersioonile ja teistele absoluutsetele tsentreeritud momentidele

Hoeffdingi võrratuse abil saame

$$\begin{aligned}\mathbf{P}(L'_n\neq L_n)&=\mathbf{P}(n_x(n)<n)+\mathbf{P}(n_y(n)<n)-\mathbf{P}(n_x(n)\geq n,n_y(n)\geq n)\\&\leq\mathbf{P}(n_x(n)\leq n)+\mathbf{P}(n_y(n)\leq n)\\&\leq\mathbf{P}\left(n_x(n)\leq(p-\delta)\left\lceil\frac{n}{p-\delta}\right\rceil\right)+\mathbf{P}\left(n_y(n)\leq(p-\delta)\left\lceil\frac{n}{p-\delta}\right\rceil\right)\\&=\mathbf{P}\left(\sum_{i=1}^{m(n)}D_i^x\leq(p-\delta)m(n)\right)+\mathbf{P}\left(\sum_{i=1}^{m(n)}D_i^y\leq(p-\delta)m(n)\right)\\&\leq 2\exp[-2\delta^2m(n)]\\&\leq 2\exp\left(-2\delta^2\frac{n}{p-\delta}\right)\\&=2\exp(-c_1n),\end{aligned} \tag{2.20}$$

kus $c_1:=\dfrac{2\delta^2}{p-\delta}$.

Paneme tähele, et $L'_n\leq L_n\leq n$. Tänu seosele (2.20) saame

$$\begin{aligned}EL_n-EL'_n&=\mathbf{P}(L_n-L'_n=1)\cdot 1+...+\mathbf{P}(L_n-L'_n=n)\cdot n\cdot\\&\leq n\cdot[\mathbf{P}(L_n-L'_n=1)+...+\mathbf{P}(L_n-L'_n=n)]\\&=n\cdot\mathbf{P}(L'_n\neq L_n)\\&\leq 2n\exp(-c_1n).\end{aligned} \tag{2.21}$$

Olgu $t_0 > 0$. Tänu seosele (2.21) leidub $n_0 \in \mathbb{N}$ nii, et

$$|EL_n - EL'_n| = EL_n - EL'_n \leq t_0 \quad \forall n \geq n_0. \tag{2.22}$$

Tänu seostele (2.20), (2.22) saame iga $t > 0$ ja $n \geq n_0$ korral

$$\begin{aligned}
\mathbf{P}(|L_n - EL_n| > t) &= \mathbf{P}(|L_n - EL_n| > t,\, L_n = L'_n) + \mathbf{P}(|L_n - EL_n| > t,\, L_n \neq L'_n) \\
&\leq \mathbf{P}(|L'_n - EL_n| > t) + \mathbf{P}(L_n \neq L'_n) \\
&\leq \mathbf{P}(|L'_n - EL_n| - |EL'_n - EL_n| > t - t_0) + 2\exp(-c_1 n) \\
&\leq \mathbf{P}(||L'_n - EL_n| - |EL'_n - EL_n|| > t - t_0) + 2\exp(-c_1 n) \\
&\leq \mathbf{P}(|L'_n - EL_n - (EL'_n - EL_n)| > t - t_0) + 2\exp(-c_1 n) \\
&= \mathbf{P}(|L'_n - EL'_n| + t_0 > t) + 2\exp(-c_1 n).
\end{aligned} \tag{2.23}$$

Seega iga $t > 0$ ja $n \geq n_0$ korral

$$\mathbf{P}(|L_n - EL_n| > \sqrt{t}) \leq \mathbf{P}(|L'_n - EL'_n| + t_0 > \sqrt{t}) + 2\exp(-c_1 n).$$

Võttes mõlemalt poolt integraali rajades $(0, n^2)$, saame

$$\int_0^{n^2} \mathbf{P}(|L_n - EL_n| > \sqrt{t})\, dt \leq \int_0^{n^2} \mathbf{P}(|L'_n - EL'_n| + t_0 > \sqrt{t})\, dt + \int_0^{n^2} 2\exp(-c_1 n)\, dt.$$

Siit saame seose (2.17) abil

$$\begin{aligned}
\mathrm{Var} L_n &\leq E(|L'_n - EL'_n| + t_0)^2 + 2n^2 \exp(-c_1 n) \\
&= E|L'_n - EL'_n|^2 + 2t_0 \cdot E|L'_n - EL'_n| + t_0^2 + o(n) \\
&= \mathrm{Var} L'_n + 2t_0 \cdot m_1(L'_n) + o(n) \\
&= \mathrm{Var} L'_n + o(n).
\end{aligned} \tag{2.24}$$

Seoste (2.15) ja (2.24) abil saame juhusliku suuruse L_n dispersioonile järgmise Efron-Stein'i võrratust kasutava tõkke:

$$\mathrm{Var} L_n \leq \frac{1}{p - \delta} \cdot n + o(n), \tag{2.25}$$

kus $0 < \delta < p$.

Seoste (2.18) ja (2.24) abil saame juhusliku suuruse L_n dispersioonile ainult McDiarmidi võrratust kasutava tõkke:

$$\mathrm{Var} L_n \leq \frac{2(1 + \ln 2)}{p - \delta} \cdot n + o(n).$$

Paneme tähele, et tegemist on halvema tõkkega kui (2.25).

Kolmas võimalus on kasutada funktsiooni $\frac{L'}{2}$ ennasttõkestavuse omadust. Seostest (2.19) ja (2.24) saame

$$\mathrm{Var} L_n \leq 4EL_n + o(n).$$

Analoogiliselt eelnevaga saame leida ülemise tõkke juhusliku suuruse L_n mis tahes tsentreeritud absoluutsele momendile $m_k(L_n)$, $k > 0$. Seosest (2.23) saame, et iga $t > 0$ ja $n \geq n_0$ korral

$$\mathbf{P}\left(|L_n - EL_n| > t^{\frac{1}{k}}\right) \leq \mathbf{P}\left(|L'_n - EL'_n| + t_0 > t^{\frac{1}{k}}\right) + 2\exp(-c_1 n).$$

Võttes mõlemalt poolt integraali rajades $(0, n^k)$, saame

$$\int_0^{n^k} \mathbf{P}\left(|L_n - EL_n| > t^{\frac{1}{k}}\right) dt \leq \int_0^{n^k} \mathbf{P}\left(|L'_n - EL'_n| + t_0 > t^{\frac{1}{k}}\right) dt + \int_0^{n^k} 2\exp(-c_1 n)\, dt.$$

Siit saame tänu lemmale 2.4 ja seosele (2.17)

$$\begin{aligned} m_k(L_n) &\leq E(|L'_n - EL'_n| + t_0)^k + 2n^k \exp(-c_1 n) \\ &\leq E[2^{(k-1)\vee 0}(|L'_n - EL'_n|^k + t_0^k)] + o(n^{\frac{k}{2}}) \\ &= 2^{(k-1)\vee 0} m_k(L'_n) + o(n^{\frac{k}{2}}) \\ &\leq 2^{(k-1)\vee 0} c_k^{**} n^{\frac{k}{2}} + o(n^{\frac{k}{2}}) \quad \forall k > 0. \end{aligned}$$

Peatükk 3

Ekstremaalsete joonduste vahelise kauguse momentide asümptootika

3.1 Dispersiooni hindamise probleem sõltumatute jadade korral

Meenutame, et $T_1, ..., T_n, S_1, ..., S_m \in \mathcal{A}$ on sõltumatud juhuslikud suurused. Olgu H on funktsioon, mis tagastab kahe jada ekstremaalsete joonduste vahelise Hausdorffi kauguse eukleidilises mõttes. Oleme huvitatud dispersiooni

$$\mathrm{Var} H(T_1, ..., T_n; S_1, ..., S_m)$$

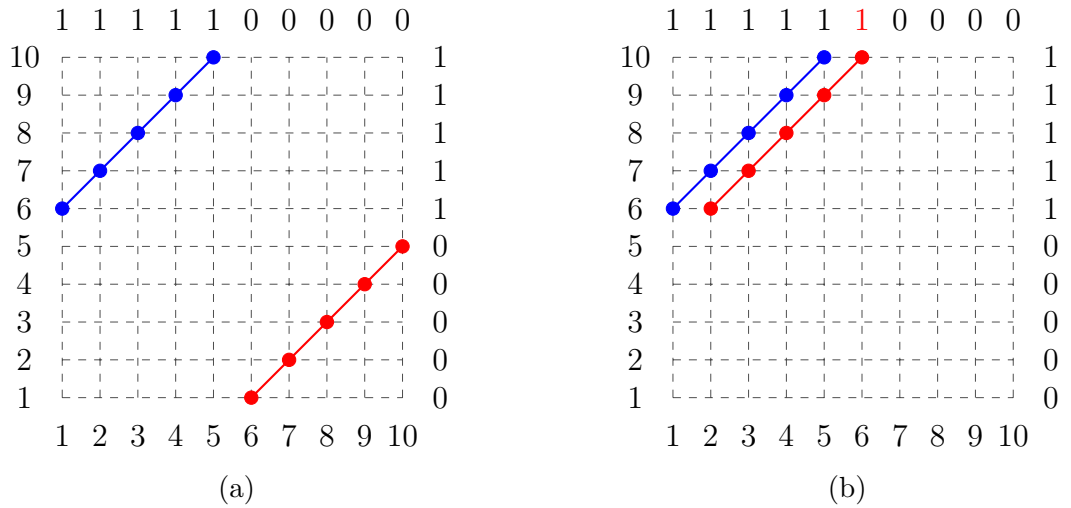
hindamisest. Võib-olla saab selleks kasutada meile juba tuttavat tõkestatud muutude omadust?

Paraku ei leidu n-ist ja m-ist sõltumatut konstanti, millega funktsioonil H on tõkestatud muutude omadus. Selleks näitamiseks konstrueerime kontranäite. Olgu n paarisarv ja $n = m$. Lisaks olgu

$$\begin{aligned} k &= \frac{n}{2} = \frac{m}{2}, \\ x &= \underbrace{1, ..., 1}_{k}, \underbrace{0, ..., 0}_{k}, \\ x' &= \underbrace{1, ..., 1}_{k+1}, \underbrace{0, ..., 0}_{k-1}, \\ y &= \underbrace{0, ..., 0}_{k}, \underbrace{1, ..., 1}_{k}. \end{aligned}$$

Näeme, et jada x' erineb jadast x vaid ühe elemendi poolest.

Üritame kõigepealt kontranäite ideed selgitada joonise abil. Jadade x, y optimaalse joonduse moodustamiseks on kaks võimalust: ühendada kõik ühed või kõik nullid. Seega neil jadadel on ainult kaks optimaalset joondust. Joonisel 3.1a on kujutatud jadade x, y ekstremaalsed joondused $n = 10$ korral. Näeme, et ekstremaalsed joondused asuvad üksteisest küllaltki kaugel. Jadade x', y optimaalsetes joondustes nulle aga enam ei ühendata. Joonisel 3.1b on kujutatud jadade x', y ekstremaalsed joondused $n = 10$ korral.



Joonis 3.1: jadade x, y (a) ja jadade x', y (b) ekstremaalsed joondused kahedimensionaalse graafikuna $n = 10$ korral

Näeme, et kõigest ühe elemendi muutmine võib ekstremaalsete joonduste vahelist kaugust oluliselt muuta.

Esitame nüüd kontranäite analüütilise kirjelduse. Nagu juba nägime, on jadadel x, y kaks optimaalset joondust:

$$\{(1,k+1),...,(k,k+k)\} =: \{a_1^1,...,a_k^1\},$$
$$\{(k+1,1),...,(k+k,k)\} =: \{a_1^2,...,a_k^2\}.$$

Leiame elementide a_i^1, a_j^2 vahelise kauguse:

$$\begin{aligned}\|a_i^1 - a_j^2\|_2 &= \|(i,k+i)-(k+j,j)\|_2 \\ &= \sqrt{(i-j-k)^2+(k+i-j)^2} \\ &= \sqrt{2(i-j)^2+2k^2}.\end{aligned}$$

Seega

$$\min_{i=1,...,k} \|a_i^1 - a_j^2\|_2 = \min_{j=1,...,k} \|a_i^1 - a_j^2\|_2 = k\sqrt{2}$$

ja

$$H(x;y) = k\sqrt{2} = \frac{n}{\sqrt{2}}. \tag{3.1}$$

Jadadel x', y on $k+1$ erinevat joondust:

$$\{(1,k+1),...,(s-1,k+s-1),(s+1,k+s),...,(k+1,k+k)\} =: \{b_1^s,...,b_k^s\},$$
$$s = 1,...,k+1.$$

Olgu $u, v \in \{1,...,k+1\}, u \neq v$. Kui $i \notin [u \wedge v, u \vee v - 1]$, siis

$$\min_{j=1,...,k} \|b_i^u - b_j^v\|_2 = 0.$$

Kui $i \in [u \wedge v, u \vee v - 1]$, siis

$$\min_{j=1,\dots,k} \|b_i^u - b_j^v\|_2 = 1.$$

Seega $H(x';y) = 1$ ja

$$H(x;y) - H(x';y) = \frac{n}{\sqrt{2}} - 1.$$

Seega tõkestatud muutude omadust dispersiooni $\mathrm{Var}H(T_1,...,T_n;\, S_1,...,S_m)$ hindamiseks otseseselt kasutada ei saa.

Nimetatud dispersiooni hindamine jääb lahtiseks küsimuseks.

3.2 Sõltuvad jadad

3.2.1 Momentide asümptootika

Käesolevas jaotises tegeleme ainult juhuslike pikkustega järglasjadadega, kuna neid on tehniliselt lihtsam käsitleda kui fikseeritud pikkusega jadasid. Defineerime

$$H_m := H(X^{\sum_{i=1}^m D_i^x}; Y^{\sum_{i=1}^m D_i^y}).$$

Seega juhuslik suurus H_m on Hausdorffi kaugus selliste juhuslike pikkustega järglasjadade ekstremaalsete joonduste vahel, mille eellasjada pikkus on m.

Ütleme, et meie sõltuvate jadade mudel rahuldab tingimust (*), kui täidetud on tingimus [9, (1.8)]. Kuna nimetatud tingimus on tehnilist laadi, siis lihtsuse huvides ei hakka me seda siinkohal täpsustama. Kas tingimus (*) on täidetud või mitte, sõltub meie mudeli parameetritest p ja Q ning eellasjada $Z_1, Z_2, ...$ marginaaljaotusest. Esitame käesoleva jaotise võtmeteoreemi.

Teoreem 3.1. *Rahuldagu meie sõltuvate jadade mudel tingimust (*). Siis leiduvad sellised konstandid $k_0 < \infty$, $K < \infty$, $b > 0$, et*

$$\mathbf{P}(H_m > k) \le Kme^{-bk} \quad \forall k \ge k_0, \quad \forall m \in \mathbb{N}.$$

Teoreemi tõestus on analoogiline toereemi [9, Theorem 3.1] tõestusega. Teoreemi konstandid k_0, K, b sõltuvad meie mudeli parameetritest Q ja p ning jada $Z_1, Z_2, ...$ marginaaljaotusest.

Kehtigu nüüd tingimus (*). Toodud teoreemi kohaselt leiduvad konstandid k_0, K, b nii, et

$$\mathbf{P}(H_m > k) \le Kme^{-bk} \quad \forall k \ge k_0, \quad \forall m \in \mathbb{N}. \tag{3.2}$$

Olgu $s > 0$ ja $A_s := \frac{s+1}{b}$. Siis piisavalt suure m korral

$$\begin{aligned} \mathbf{P}(H_m > A_s \ln m) &\le Kme^{-bA_s \ln m} \\ &= Km \exp(\ln m^{-bA_s}) \\ &= Km \cdot m^{-bA_s} \\ &= Km^{-s}. \end{aligned} \tag{3.3}$$

Kui $s > 1$, siis Boreli-Cantelli lemmast saame

$$\mathbf{P}(H_m \le A_s \ln m, \quad \text{ev.}) = 1,$$

st peaaegu kindlasti leidub $M = M(X^\infty, Y^\infty) \in \mathbb{N}$ nii, et $H_m \le A_s \ln m$ iga $m \ge M$ korral. Sõnastame saadud tulemuse teoreemina.

Teoreem 3.2. *Rahuldagu meie sõltuvate jadade mudel tingimust (*). Siis leidub konstant $C > 0$ nii, et peaaegu kindlasti*

$$H_m \le C \ln m, \quad \textit{ev.}$$

Tänu seosele (3.2) saame hinnata juhusliku suuruse H_m keskväärtust. Leidub punkt $x = x(m) > 0$ nii, et $Kme^{-bt} > 1$ iga $t \in (0, x)$ korral. Avaldame suuruse x:

$$\begin{aligned} Kme^{-bx} = 1 \quad &\Leftrightarrow \quad \ln(Kme^{-bx}) = 0 \\ &\Leftrightarrow \quad -bx = -\ln(Km) \\ &\Leftrightarrow \quad x = \frac{\ln(Km)}{b}. \end{aligned}$$

Kui m on nii suur, et $x \ge k_0$, siis

$$\begin{aligned} EH_m &= \int_0^\infty \mathbf{P}(H_m > t)\, dt \\ &\le x + \int_x^\infty \mathbf{P}(H_m > t)\, dt \\ &\le x + Km \int_x^\infty e^{-bt}\, dt \\ &= x + Km \left(\frac{e^{-bt}}{-b} \Big|_x^\infty \right) \\ &= x + \frac{Kme^{-bx}}{b} \\ &= \frac{\ln(Km)}{b} + \frac{Km(Km)^{-1}}{b} \\ &= \frac{1}{b} \ln m + \frac{1}{b} \ln K - \ln b + b^{-1}. \end{aligned}$$

Seega oleme tõestanud järgmise teoreemi.

Teoreem 3.3. *Rahuldagu meie sõltuvate jadade mudel tingimust (*). Siis*

$$EH_m = O(\ln m).$$

Järgnevalt püüame hinnata suvalist tsentreerimata momenti EH_m^s, $s > 0$. Leidub $x_s = x_s(m) > 0$ nii, et $Km \exp\left(-bt^{\frac{1}{s}}\right) > 1$ iga $t \in (0, x_s)$. Leiame punkti x_s:

$$\begin{aligned} Km \exp\left(-bx_s^{\frac{1}{s}}\right) = 1 \quad &\Leftrightarrow \quad \ln\left[Km \exp\left(-bx_s^{\frac{1}{s}}\right)\right] = 0 \\ &\Leftrightarrow \quad -bx_s^{\frac{1}{s}} = -\ln(Km) \\ &\Leftrightarrow \quad x_s = \left(\frac{\ln(Km)}{b}\right)^s. \end{aligned}$$

Kui m on nii suur, et $x_s \geq k_0^s$, siis

$$\begin{aligned} EH_m^s &= \int_0^\infty \mathbf{P}\left(H_m > t^{\frac{1}{s}}\right) dt \\ &\leq x_s + \int_{x_s}^\infty \mathbf{P}\left(H_m > t^{\frac{1}{s}}\right) dt \\ &\leq x_s + Km \int_{x_s}^\infty \exp\left(-bt^{\frac{1}{s}}\right) dt. \end{aligned}$$

Teostades muutuja vahetuse

$$u = bt^{\frac{1}{s}}, \quad t = \left(\frac{u}{b}\right)^s, \quad du = \frac{b}{s} t^{\frac{1}{s}-1}\, dt, \quad dt = \frac{s}{b}\left(\frac{u}{b}\right)^{s-1} du,$$

saame

$$EH_m^s \leq x_s + \frac{Kms}{b^s} \int_{\ln(Km)}^\infty e^{-u} u^{s-1}\, du = x_s + \frac{Kms}{b^s} \Gamma[s, \ln(Km)], \tag{3.4}$$

kus

$$\Gamma : \left(\mathbb{R}^+\right)^2 \to \mathbb{R}, \quad \Gamma(a,z) := \int_z^\infty e^{-t} t^{a-1}\, dt$$

on ülemine mittetäielik gammafunktsioon (ingl. k. *upper incomplete gamma function*). Nimetatud funktsiooni hindamiseks kasutame järgmist teoreemi.

Teoreem 3.4. [17, Theorem 2.1] *(i) Kui $z-(a-1)>0$, siis iga $a \geq 1$ korral*

$$\Gamma(a,z) \leq \frac{z^a e^{-z}}{z-(a-1)}.$$

(ii) Iga $0 < a \leq 1$ korral

$$\Gamma(a,z) \leq z^{a-1} e^{-z}.$$

Tõestus. Muutuja vahetuse

$$u = \frac{t-z}{z}, \quad t = z(1+u), \quad du = \frac{1}{z}\, dt, \quad dt = z\, du$$

abil saame

$$\Gamma(a,z) = \int_0^\infty e^{-z(1+u)} z^{a-1} (1+u)^{a-1} z\, du = z^a e^{-z} \int_0^\infty e^{-zu} (1+u)^{a-1}\, du.$$

On teada, et $1+u \leq e^u$. Seega, kui $a \geq 1$, siis

$$\begin{aligned} \Gamma(a,z) &\leq z^a e^{-z} \int_0^\infty e^{-zu} \left(e^u\right)^{a-1} du \\ &= z^a e^{-z} \int_0^\infty e^{u(a-1-z)}\, du \\ &= z^a e^{-z} \left(\frac{e^{u(a-1-z)}}{a-1-z} \Big|_{u=0}^\infty \right) \\ &= \frac{z^a e^{-z}}{z-(a-1)}. \end{aligned}$$

Kui $0 < a \leq 1$, siis $(1+u)^{a-1} \leq 1$ iga $u \geq 0$ korral ning

$$\Gamma(a,z) \leq z^a e^{-z} \int_0^\infty e^{-zu} 1\, du = z^a e^{-z} \left(\frac{e^{-zu}}{-z} \Big|_{u=0}^{\infty} \right) = z^{a-1} e^{-z}.$$

□

Vaatleme juhtu $s \geq 1$. Kui m on nii suur, et $\ln(Km) - (s-1) > 0$, saame tänu ülaltoodud teoreemile

$$\begin{aligned}
\Gamma[s, \ln(Km)] &\leq \frac{[\ln(Km)]^s e^{-\ln(Km)}}{\ln(Km) - (s-1)} \\
&= (Km)^{-1} \cdot (\ln K + \ln m)^{s-1} \frac{\ln(Km)}{\ln(Km) - (s-1)} \\
&= (Km)^{-1} \cdot o[(\ln m)^s].
\end{aligned}$$

Juhul $0 < s \leq 1$ saame teoreemi abil sama tulemuse:

$$\Gamma[s, \ln(Km)] \leq [\ln(Km)]^{s-1} e^{-\ln(Km)} = (Km)^{-1} \cdot o[(\ln m)^s].$$

Seega seosest (3.4) saame piisavalt suure m korral

$$\begin{aligned}
EH_m^s &\leq \left(\frac{\ln(Km)}{b} \right)^s + \frac{Kms}{b^s} \cdot (Km)^{-1} \cdot o[ln^s(m)] \\
&= \left(\frac{\ln(Km)}{b} \right)^s + o[(\ln m)^s] \\
&= \frac{1}{b^s} (\ln m)^s + o[(\ln m)^s]. \qquad (3.5)
\end{aligned}$$

Seega oleme tõestanud järgmise teoreemi.

Teoreem 3.5. *Rahuldagu meie sõltuvate jadade mudel tingimust (*). Siis iga $s > 0$ korral*

$$EH_m^s = O[(\ln m)^s].$$

Erijuhul $s = 1$ saame ülaltoodud teoreemist teoreemi 3.3. Ülaltoodud teoreemist järeldavalt

$$\mathrm{Var} H_m = O[(\ln m)^2].$$

Teoreemini 3.5 oleksime saanud jõuda ka lihtsamal viisil. Paneme tähele, et alati $H_m \leq \sqrt{2m}$. Meenutame, et $A_s = \frac{s+1}{b}$. Lisaks ütles seos (3.3), et piisavalt suure m korral

$\mathbf{P}(H_m^s > A_s \ln m^s) \le Km^{-s}$. Seega suure m korral

$$\begin{aligned}
EH_m^s &= \int_0^\infty \mathbf{P}(H_m^s > t)\,dt \\
&= \int_0^{(A_s \ln m)^s} \mathbf{P}(H_m^s > t)\,dt + \int_{(A_s \ln m)^s}^{\left(\sqrt{2}m\right)^s} \mathbf{P}(H_m^s > t)\,dt + \int_{\left(\sqrt{2}m\right)^s}^{\infty} \mathbf{P}(H_m^s > t)\,dt \\
&\le \int_0^{(A_s \ln m)^s} 1\,dt + \int_{(A_s \ln m)^s}^{\left(\sqrt{2}m\right)^s} \mathbf{P}(H_m^s > (A_s \ln m)^s)\,dt + 0 \\
&\le (A_s \ln m)^s + Km^{-s} \int_{(A_s \ln m)^s}^{\left(\sqrt{2}m\right)^s} 1\,dt \\
&= (A_s \ln m)^s (1 - Km^{-s}) + \left(\sqrt{2}m\right)^s Km^{-s} \\
&\le (A_s \ln m)^s + K\left(\sqrt{2}\right)^s \\
&= \left(\frac{s+1}{b}\right)^s (\ln m)^s + o[(\ln m)^s]
\end{aligned}$$

Saadud tõke on siiski halvema konstandiga kui tõke (3.5).

3.2.2 Sõltuvuse test

Tänu teoreemile 3.1 võime seega öelda, et leiduvad sellised konstandid K', b', k', et

$$\mathbf{P}(H_m > k) \le K' m e^{-b'k} \quad \forall k \ge k_0', \quad \forall m \in \mathbb{N}. \tag{3.6}$$

Vältides liigset formalismi, ütleme, et jadad $X^* := X^{\sum_{i=1}^m D_i^x}$ ja $Y^* := Y^{\sum_{i=1}^m D_i^y}$ on "tugevalt" sõltuvad, kui ülaltoodud seos kehtib. Tahame testida hüpoteeside paari

$$\begin{aligned}
H_0 &: \text{jadade } X^*, Y^* \text{ sõltuvus on tugev,} \\
H_1 &: \text{jadade } X^*, Y^* \text{ sõltuvus ei ole tugev.}
\end{aligned}$$

Võtame olulisuse nivooks α. Defineerime funktsiooni

$$\kappa(x) := \frac{\ln K' + \ln x - \ln \alpha}{b'} \vee k_0'.$$

Kui $\kappa(m) = \dfrac{\ln K' + \ln m - \ln \alpha}{b'}$, siis

$$K' m e^{-b'\kappa(m)} = K' m e^{-\ln K' - \ln m + \ln \alpha} = \alpha.$$

Kui $\kappa(m) = k_0'$, siis $K' m e^{-b'k'} \le \alpha$. Seega seosest (3.6) saame, et nullhüpoteesi kehtides

$$\mathbf{P}[H_m > \kappa(m)] \le \alpha,$$

Siit saamegi hüpoteeside testimiseks otsustusreegli

$$H_m > \kappa(m) \quad \to \quad H_1.$$

Saadud test on siiski puhtteoreetiline, sest pole teada konstantide K', b', k_0', m väärtused.

3.3 Simulatsioonid (2)

Saamaks ülevaadet ekstremaalsete joonduste keskväärtuse ja standardhälbe asümptootilisest käitumisest, teostame mõned simulatsioonid. Meenutame, et üleminekumaatriks Q avaldub parameetri ϵ kaudu kujul (1.3). Genereerime jadasid järgmisest kolmest klassist:

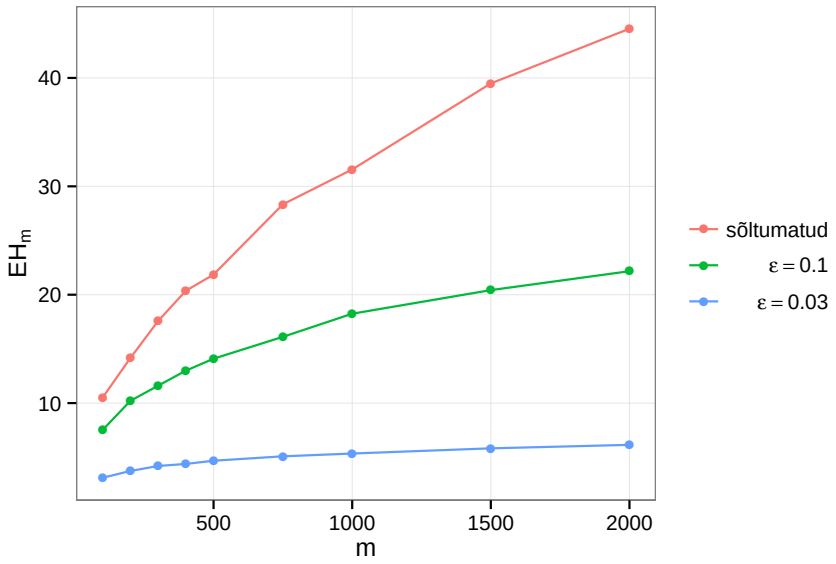
1. sõltumatud jadad pikkusega m;
2. juhuslike pikkustega jadad, $\epsilon = 0.1$, eellasjada pikkus m ("nõrgalt" sõltuvad jadad);
3. juhuslike pikkustega jadad, $\epsilon = 0.03$, eellasjada pikkus m ("tugevalt" sõltuvad jadad).

Saab näidata, et loendis viimase klassi puhul tingimus (*) kehtib. Iga

$$m = 100, 200, 300, 400, 500, 750, 1000, 1500, 2000$$

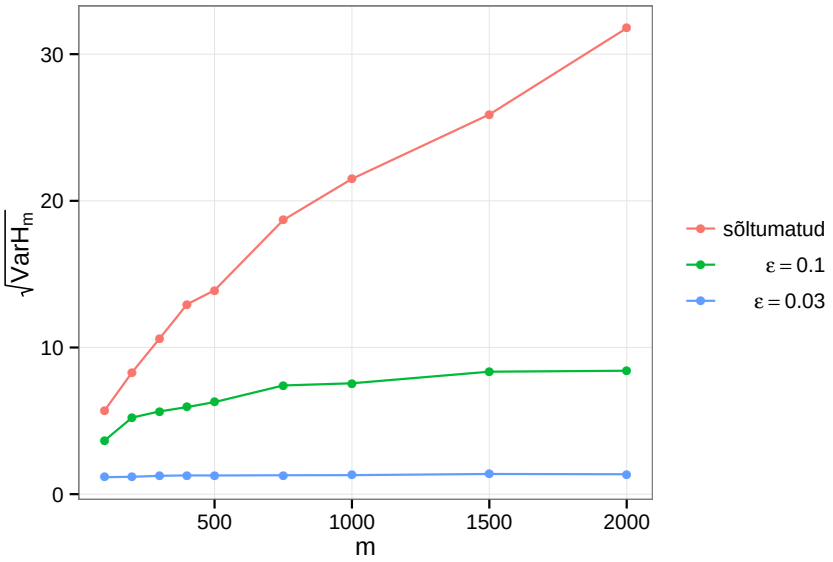
korral genereerime igast klassist 1000 tuhat jadapaari. Seega kokku genereerime $3 \cdot 9 \cdot 1000 = 27\,000$ jadapaari.

Joonisel 3.2 on toodud valimikeskmise abil saadadud juhusliku suuruse H_m keskväärtuste hinnangud. Autori meelest näib kõigi kolme sõltuvustaseme korral keskväärtuse kasv olevat logaritmiline. Keskväärtuse logaritmiline kasv tugeva sõltuvustaseme korral on kooskõlas teoreemiga 3.3.



Joonis 3.2: Juhusliku suuruse H_m keskväärtuse hinnanguline kasv

Joonisel 3.3 on toodud dispersiooni nihketa hinnangu abil arvutatud juhusliku suuruse H_m standardhälbe hinnangud. Jällegi, kõigi kolme sõltuvustaseme korral näib standardhälbe kasv olevat logaritmiline. Standardhälbe logaritmiline kasv tugeva sõltuvustaseme korral on kooskõlas teoreemiga 3.5.



Joonis 3.3: Juhusliku suuruse H_m standardhälbe hinnanguline kasv

Kokkuvõte

Tutvusime järgmise kolme võimalusega homoloogsete jadade sõltuvuse mõõtmiseks:

- sagedustabelipõhised statistikud,
- pikima ühisjada pikkus (LLCS),
- ekstremaalsete joonduste vaheline kaugus (HD).

Simuleerisime jadasid erinevatest sõltuvusklassidest jadasid ja võrdlesime karpdiagrammide abil ülaltoodud meetodite suutlikkust sõltuvusklasse eristada. Nägime, et tänu meie sõltuvate jadade mudeli sõltuvustruktuurile ei ole sagedustabelipõhistel statistikutel eriti head asümptootilised omadused. Võrreldes HD-ga näis LLCS-il olevat vähemalt ühtlase jadade marginaaljaotuse korral parem sõltuvusklasside eristusvõime. Samas nägime ka, et LLCS on marginaaljaotuse suhtes väga tundlik ning seetõttu võib HD LLCS-ist mõnikord ka paremini käituda.

Uurisime simulatsioonide abil ka LLCS-i piirjaotust. Simulatsioonid kinnitavad hüpoteesi, et LLCS on meie sõltuvate jadade mudeli kontekstis asümptootiliselt normaaljaotusega.

Teoreetiliselt uurisime LLCS-i dispersiooni. Sõltumatute jadade korral saab LLCS-i dispersiooni hindamiseks kasutada McDiarmidi ja Efron-Stein'i võrratust. Nägime, et need võrratused annavad dispersioonile jadade pikkuse suhtes lineaarse tõkke. Sõltuvate jadade korral on olukord keerulisem, kuid selgus, et McDiarmidi ja Efron-Stein'i võrratusi on siiski võimalik rakendada. Sõltuvate jadade korral saime dispersioonile samuti lineaarset järku ülemise tõkke. Analoogiliselt dispersiooni hindamisega leidsime ülemise tõkke ka kõigile absoluutsetele tsentreeritud momentidele. Nägime, et nii sõltuvate kui ka sõltumatute jadade LLCS-i k-ndat järku momendil on $\frac{k}{2}$-järku ülemine tõke ($k > 0$).

Uurisime ka HD tsentreerimata momente. Veendusime, et teatud tingimustel on sõltuvate jadade HD k-ndat järku (tsentreerimata) momendil EH_m ülemine tõke, mis on eellasjada pikkuse m suhtes järku $[\ln(m)]^k$ ($k > 0$). Seega (teatud tingimustel) on sõltuvate jadade HD keskväärtus ja standardhälve logaritmilist järku ülemise tõkkega. Simulatsioonid kinnitasid, et HD keskväärtus ja standardhälve kipuvad kasvama eellasjada pikkuse suhtes logaritmiliselt.

Lisa A

Simulatsioonides kasutatud programmid

A.1 Sõltuvate jadade genereerimine

Käesolevas peatükis eeldame, et tähestikuks on $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{T}, \mathrm{G}\}$. Simulatsioonide teostamiseks kasutame statistikatarkvara R [2]. Sõltumatute jadade genereerimiseks saab R-is kasutada *sample*-käsku. Sõltuvate jadade genereerimiseks kasutame faili, kus on funktsioon tähtede muteerimiseks ja funktsioon kadumiste teostamiseks. Vastav fail "funktsioonid.R" on toodud koodis A.1.

Kood A.1: fail "funktsioonid.R" funktsioonidega mutatsioonide ja kadumiste teostamiseks

```
#Tähtede järjekord maatriksis Q: ACTG
mutatsioon=function(Z, Q){

  Z1=Z
  Z1[Z=="A"]=sample(x=c("A","C","T","G"), size=sum(Z=="A"), replace=T,
    prob=Q[1,])
  Z1[Z=="C"]=sample(x=c("A","C","T","G"), size=sum(Z=="C"), replace=T,
    prob=Q[2,])
  Z1[Z=="T"]=sample(x=c("A","C","T","G"), size=sum(Z=="T"), replace=T,
    prob=Q[3,])
  Z1[Z=="G"]=sample(x=c("A","C","T","G"), size=sum(Z=="G"), replace=T,
    prob=Q[4,])

  return(Z1)
}

kadumised=function(FZ, p){      #p - allesjäämise toenaosus

  #Moodustatakse tõevektor kadumiste jaoks
  allesjäämised=rbinom(length(FZ), 1, p)
  tõevektor=allesjäämised==1

  return(FZ[tõevektor])
}
```

Juhuslike pikkustega sõltuvate jadade genereerimiseks on nüüd vaja vaid genereerida eellasjada ja rakendada sellele mutatsioone ja kadumisi. Vastav näidisprogramm on toodud koodis A.2. Näidisprogrammis on üleminekumaatriksi Q peadiagonaali elementideks 0.85 ja ülejäänud elementideks 0.05. Eellasjada on ühtlase marginaaljaotusega ja pikkusega 100. Kadumise tõenäosuseks on nagu ikka $p = 0.05$.

Kood A.2: näidisprogrammm juhuslike pikkustega sõltuvate jadade genereerimiseks

```
setwd("...")
source("Funktsioonid.R")

Q=matrix(0.05, nrow=4, ncol=4)
diag(Q)=1-0.05

p=0.05

eellane=sample(x=c("A","C","T","G"), size=100, replace=T)
X=kadumised(mutatsioon(eellane,Q),p)
Y=kadumised(mutatsioon(eellane,Q),p)
```

Fikseeritud pikkustega sõltuvate jadade genereerimine on analoogiline. Vaja on ainult genereerida piisavalt pikk eellasjada ning eemaldada järglasjadade lõpust üleliigsed elemendid. On võimalik, et järglasjadade pikkus tuleb liiga väike – selle vältimiseks lisame koodi ka *while*-tsükli. Näidisprogrammis A.3 genereerime kaks järglasjada pikkusega 100.

Kood A.3: näidisprogrammm fikseeritud pikkustega sõltuvate jadade genereerimiseks

```
setwd("...")
source("Funktsioonid.R")

Q=matrix(0.05, nrow=4, ncol=4)
diag(Q)=1-0.05

p=0.05

while(T){
  eellane=sample(x=c("A","C","T","G"), size=120, replace=T)
  X1=kadumised(mutatsioon(eellane,Q),p)
  Y1=kadumised(mutatsioon(eellane,Q),p)

  if(length(X1) >=100 & length(Y1) >=100){
    break
  }
}

X=X1[1:100]
Y=Y1[1:100]
```

A.2 Pikima ühisjada pikkuse ja ekstremaalsete joonduste vahelise kauguse leidmine

Nagu juba varem mainitud, saab kahe jada pikimat ühisjada leida Needlemann-Wunschi algoritmi abil. See algoritm on juba olemas R-i paketis Biostrings [4]. Näidisprogram-

mis A.4 genereerime kaks sõltumatut jada pikkusega 100 ja leiame nende jadade pikima ühisjada pikkuse.

Kood A.4: näidisprogrammm kahe jada pikima ühisjada pikkuse leidmiseks paketi Biostrings abil

```
library(Biostrings)

X1=sample(x=c("A","C","T","G"), size=100, replace=T)
X1=sample(x=c("A","C","T","G"), size=100, replace=T)

#teeme jadad üheks stringiks
X=paste(X1, collapse="")
Y=paste(X1, collapse="")

#sarnasusmaatriks NW-algoritmi jaoks
sarnasus=matrix(rep(-1,16), nrow=4)
diag(sarnasus)=rep(1,4)
rownames(sarnasus) = c("A","C","T", "G")
colnames(sarnasus)=c("A","C","T", "G")

#leiame pikima ühisjada pikkuse
LCS=pairwiseAlignment(c2s(X), c2s(Y), substitutionMatrix=sarnasus,
  gapOpening=0, gapExtension=0, scoreOnly=TRUE)
```

Ekstremaalsete joonduste leidmiseks pidi autor vajaliku koodi ise kirjutama. Selleks oli vaja Needlemann-Wunschi algoritmi veidi modifitseerima. Kuna R-i tsüklid on väga aeglased, siis efektiivsuse huvides kirjutas autor vajaliku koodi C++ keeles ja kasutas R-i ja C++ ühildamiseks paketti Rcpp [5, 6]. Kirjutatud koodi paigutas autor faili "NW.cpp". Faili sisu on toodud koodis A.5. Et koodist täielikult aru saada, on vaja mõista Needlemann-Wunschi algoritmi. Huvitatud lugeja leiab selle algoritmi selgituse veebist hõlpsasti üles.

Kood A.5: faili "NW.cpp" sisu

```
#include <Rcpp.h>
#include <vector>

using namespace Rcpp;

// [[Rcpp::export]]
std::vector<int> Ekstrem(std::wstring jada1, std::wstring jada2){
  int n = jada1.length();
  int m = jada2.length();

  //Järgmised 3 tõeväärtusmaatriksit moodustavad sisuliselt Backtrace
  //maatriksi
  std::vector< std::vector<bool> > vasakule(n+1, std::vector<bool>
    (m+1));
  std::vector< std::vector<bool> > yles(n+1, std::vector<bool>(m+1));
  std::vector< std::vector<bool> > diagonaalis(n+1, std::vector<bool>
    (m+1));


  for(int i = 1; i < n+1; ++i){
    yles[i][0]=1;
```

```
}

for(int i = 1; i < m+1; ++i){
  vasakule[0][i]=1;
}

//Vektor skoorimaatriksi täitmiseks
IntegerVector S(3);

//Mälu kokkuhoidmiseks kasutatakse skoorimaatriksi asemel kahte
//vektorit
std::vector<int> vektor1(m+1);
std::vector<int> vektor2(m+1);
//igas tsüklis skoorimaatriksisse sisestatav element
//(maksimum elementidest s1,s2,s4)):
int maksimum=0;

String str1="";
String str2="";

for(int i = 1; i < n+1; ++i){
  for(int j = 1; j < m+1; ++j){

    S(2)=vektor1[j];
    S(0)=vektor2[j-1];

    str1=jada1.substr(i-1,1);
    str2=jada2.substr(j-1,1);

    if(str1==str2){
      S(1)=vektor1[j-1]+1;
    } else {S(1)=-1;}

    maksimum=max(S);

    //Backtrace maatriksite täitmine
    if(S(0)==maksimum){vasakule[i][j]=1;}
    if(S(1)==maksimum){diagonaalis[i][j]=1;}
    if(S(2)==maksimum){yles[i][j]=1;}

    vektor2[j]=maksimum;

  }
  //Vektor2-e elemendid vektor1-te
  for(int k=1; k<m+1; ++k){vektor1[k]=vektor2[k];}
}

//Saime pikima ühisjada pikkuse.
int Ln=vektor2[m];


//Loome vektori ekstremaalsete joonduste jaoks.
//elementide järjestus:
//ülemise joonduse x-koordinaadid (0 : Ln-1),
//ülemise joonduse y-koordinaadid (Ln : 2*Ln-1),
```

```
  //    alumise joonduse x-koordinaadid (2*Ln : 3*Ln-1),
  //    alumise joonduse y-koordinaadid (3*Ln1 : 4*Ln-1).
  std::vector<int> koord(4*Ln);
  int i=n;
  int j=m;
  int n1=Ln-1;

  //Ülemine joondus
  while(n1>=0){
    if(yles[i][j]){--i;}
    else if(diagonaalis[i][j]){
      koord[n1]=i;
      koord[Ln+n1]=j;
      --j;--i;--n1;
    }
    else{--j;}
  }

  i=n;
  j=m;
  n1=Ln-1;
  //Alumine joondus
  while(n1>=0){
    if(vasakule[i][j]){--j;}
    else if(diagonaalis[i][j]){
      koord[2*Ln+n1]=i;
      koord[3*Ln+n1]=j;
      --j;--i; --n1;
    }
    else{--i;}
        }

  return koord;

}
```

Failis “NW.cpp” sisalduv funktsioon tagastab kahe jada ekstremaalsete joonduste koordinaadid ühe vektorina. Vaja on veel programmi, mis tagastaks ekstremaalsete joonduste Hausdorffi kauguse. Vastava programmi kirjutas autor jällegi C++ keeles ja paigutas selle faili “Hausdorff.cpp”. Selle faili sisu on toodud koodis A.6.

Kood A.6: faili “Hausdorff.cpp” sisu

```
#include <Rcpp.h>
#include <vector>
#include <math.h>

using namespace Rcpp;

// [[Rcpp::export]]
double euklNorm(int a, int b) {
  return sqrt(pow(a,2)+pow(b,2));
}

// [[Rcpp::export]]
std::vector<double> Hausdorff(std::vector<int> Xx, std::vector<int> Xy,
```

```
std::vector<int> Yx, std::vector<int> Yy) {
//Xx - hulga X x-koordinaadid, Xy - hulga X y-koordinaadid,
//Yx - hulga Y x-koordinaadid, Yy - hulga Y y-koorindaadid.
int nX=Xx.size();
int nY=Yx.size();

double miinimum=-1;
double maksimum=0;
double kaugus=0;
int j;
int maksimum_x; //maksimumile vastav x-koordinaat
int maksimum_y; //maksimumile vastav y-koordinaat

//abimuutuja Hausdorffi kaugusele vastavate koordinaatide
//salvestamiseks
int abi;

//sup_{x in X} inf_{y in Y} d(x,y)
for(int i = 0; i < nX; ++i) {
  j=i;


  while(Xx[i]<=Yx[j] && j>0){
    kaugus=euklNorm(Xx[i]-Yx[j], Xy[i]-Yy[j]);
    if(kaugus<miinimum || miinimum==-1){
      miinimum=kaugus;
      abi=j;
    }
    --j;
  }

  kaugus=euklNorm(Xx[i]-Yx[j], Xy[i]-Yy[j]);
  if(kaugus<miinimum || miinimum==-1){
    miinimum=kaugus;
    abi=j;
  }


  j=i+1;

  while(Xy[i]>=Yy[j] && j<nY-1){
    kaugus=euklNorm(Xx[i]-Yx[j], Xy[i]-Yy[j]);
    if(kaugus<miinimum){
      miinimum=kaugus;
      abi=j;
    }
    ++j;
  }

  kaugus=euklNorm(Xx[i]-Yx[j], Xy[i]-Yy[j]);
  if(kaugus<miinimum){
    miinimum=kaugus;
    abi=j;
  }
```

```
    if(maksimum<miinimum){
      maksimum=miinimum;
      maksimum_x=i;
      maksimum_y=abi;
    }

    miinimum=-1;
  }


  //sup_{y in Y} inf_{x in X} d(x,y)
  for(int i = 0; i < nY; ++i) {
    j=i;

    while(Yy[i]<=Xy[j] && j>0){
      kaugus=euklNorm(Xx[j]-Yx[i], Xy[j]-Yy[i]);
      if(kaugus<miinimum || miinimum==-1){
        miinimum=kaugus;
        abi=j;
      }
      --j;
    }

    kaugus=euklNorm(Xx[j]-Yx[i], Xy[j]-Yy[i]);
    if(kaugus<miinimum || miinimum==-1){
      miinimum=kaugus;
      abi=j;
    }


    j=i+1;

    while(Yx[i]>=Xx[j] && j<nX-1){
      kaugus=euklNorm(Xx[j]-Yx[i], Xy[j]-Yy[i]);
      if(kaugus<miinimum){
        miinimum=kaugus;
        abi=j;
      }
      ++j;
    }

    kaugus=euklNorm(Xx[j]-Yx[i], Xy[j]-Yy[i]);
    if(kaugus<miinimum){
      miinimum=kaugus;
      abi=j;
    }

    if(maksimum<miinimum){
      maksimum=miinimum;
      maksimum_x=i;
      maksimum_y=abi;
    }

    miinimum=-1;
  }
```

```
  std::vector<double> v(3);

  v[0]=maksimum;
  v[1]=maksimum_x+1;
  v[2]=maksimum_y+1;

  //tagastatakse kaugus ja vastavad indeksid ülemises ja alumises
  //joonduses
  return v;

}
```

Nüüd on meil olemas kõik vajalikud funktsioonid kahe jada ekstremaalsete joonduste vahelise kauguse leidmiseks. Näidisprogrammis A.7 genereeritakse kaks sõltumatut jada pikkusega 100, leitakse nende jadade ekstremaalsed joondused ja nende joonduste vaheline Hausdorffi kaugus.

Kood A.7: näidisprogramm kahe jada ekstremaalsete joonduste vahelise Hausdorffi kauguse leidmiseks

```
library(Rcpp)
setwd("...")
sourceCpp("Hausdorff.cpp")
sourceCpp("NW.cpp")

X1=sample(x=c("A","C","T","G"), size=100, replace=T)
X1=sample(x=c("A","C","T","G"), size=100, replace=T)

#teeme jadad üheks stringiks
X=paste(X1, collapse="")
Y=paste(X1, collapse="")

Vektor=Ekstrem(X,Y)

Ln=length(Vektor)/4  #pikima ühisjada pikkus

Xx=Vektor[1:Ln]             #Ülemise joonduse x-koordinaadid
Xy=Vektor[(Ln+1):(2*Ln)]    #Ülemise joonduse y-koordinaadid
Yx=Vektor[(2*Ln+1):(3*Ln)] #Alumise joonduse x-koordinaadid
Yy=Vektor[(3*Ln+1):(4*Ln)] #Alumise joonduse y-koordinaadid

Haus=Hausdorff(Xx,Xy,Yx,Yy)[1]  #Hausdorffi kaugus
```

Märgime, et kahe jada pikima ühisjada pikkust saab leida ka ekstremaalsete joonduste kaudu. Siiski pakett Biostrings leiab pikima ühisjada pikkuse kiiremini kui autori ekstremaalsete joonduste leidmise programm.

Viited

[1] J. Sova. Sõltuvate jadade mudel. Bakalaureusetöö, Tartu Ülikool, 2013. http://dspace.utlib.ee/dspace/handle/10062/31723.

[2] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Viin, Austria, 2014.

[3] H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

[4] H. Pages, P. Aboyoun, R. Gentleman ja S. DebRoy. *Biostrings: String objects representing biological sequences, and matching algorithms*. R package version 2.34.1.

[5] D. Eddelbuettel ja R. Francois. Rcpp: Seamless r and c++ integration. *J. Stat. Softw.*, 40(8):1–18, 2011.

[6] D. Eddelbuettel. *Rcpp: Seamless R and C++ Integration*. Springer New York, 2013].

[7] J. Lember. Informatsiooniteooria 2011. http://www.ms.ut.ee/sites/default/files/ms/informatsiooniteooria_2011.pdf.

[8] V. Chvatal ja D. Sankoff. Longest common subsequences of two random sequences. *J. Appl. Probab.*, 12:306–315, 1975.

[9] J. Lember, H. Matzinger ja A. Vollmer. Optimal alignments of longest common subsequences and their path properties. *Bernoulli*, 20:1292–1343, 2014.

[10] J. Lember, H. Matzinger ja F. Torres. General approach to the fluctuations problem in random sequence comparison. (avaldamata), 2015.

[11] J. Lember ja H. Matzinger. Standard deviation of the longest common subsequence. *Ann. Probab.*, 37:1192–1235, 2009.

[12] C. Durringer, J. Lember ja H. Matzinger. Deviation from the mean in sequence comparison with a periodic sequence. *ALEA*, 3:1–29, 2007.

[13] F. Bonetto ja H. Matzinger. Fluctuations of the longest common subsequence in the case of 2- and 3- letter aphabets. *ALEA*, 2:195–216, 2006.

[14] C. Houdre ja J. Ma. On the order of the central moments of the length of the longest common subsequence. 2015. [arXiv:1212.3265v2].

[15] S. Boucheron, G. Lugosi ja P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[16] J. M. Steele. An efron-stein inequality for non-symmetric statistics. *Ann. Stat.*, 14:753–758, 1986.

[17] Chan O. Borwein, J. M. Uniform bounds for the complementary incomplete gamma function. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.250&rep=rep1&type=pdf.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Joonas Sova (sünnikuupäev: 01.06.1987),

1. Annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose "Homoloogsete jadade sõltuvusmõõdud', mille juhendaja on Jüri Lember

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace'i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, **12.05.2015**.