TARTU ÜLIKOOL

MATEMAATIKA-INFORMAATIKATEADUSKOND

Matemaatilise statistika instituut

Joonas Sova

Sõltuvate jadade mudel

Bakalaureusetöö

Juhendaja: Jüri Lember

Tartu 2013

SISUKORD

SISSEJUHATUS

Kui evolutsiooni käigus tekib ühest liigist kaks uut, on uute liikide genoomid omavahel sarnased. Selline uute liikide tekkimine tähendab, et eellasliigi DNA-jadaga on toimunud teisendusi, täpsemalt:

- mõned jada elemendid on asendunud teistega (mutatsioonid),
- jada varasemate elementide vahele on sisestatud uusi elemente (sisestused),
- jadast on elemente kaduma läinud (kadumised).

Seega, mida lähemal on liigid evolutsioonipuus üksteisele, seda sarnasemad on nende genoomid.

Töö on jagatud kaheks peatükiks.

Esimesene peatükk on puhtteoreetiline. Siin konstrueeritakse mudel jada kahe järglasjada moodustumiseks koos mutatsioonide ja kadumistega. Lisaks esitatakse mõned teoreetilised tulemused mudeli kohta.

Teine peatükk keskendub simulatsioonidele. Siin tutvustatakse üht lihtsamat pikima ühisjada pikkusel põhinevat jadade sarnasusmõõtu ning uuritakse, kuidas esimeses peatükis konstrueeritud järglasjadade sarnasus sõltub nende vahelisest sõltuvusmäärast.

Töös esitatud tõestuskäigud on autor kas ise leidnud või leidnud etteantud skeemi või idee põhjal.

1. MUDEL

1.1 Mutatsioonid

Olgu $\mathcal{A}$ lõplik tähestik.

Definitsioon 1.1. *Olgu*

$$f : \mathcal{A} \times \mathbb{R} \to \mathcal{A}.$$

ja olgu ξ juhuslik suurus mingist fikseeritud jaotusest. Mutatsiooniks nimetame sellist juhuslikku funktsiooni

$$F : \mathcal{A} \to \mathcal{A},$$

et iga $a \in \mathcal{A}$ korral

$$F(a) := f(a, \xi).$$

Sellise juhusliku funktsiooni F jaotuse määrab üleminekumaatriks

$$Q : \quad Q(a, b) = P(F(a) = b), \quad a, b \in \mathcal{A},$$

st $Q(a, b)$ on tõenäosus, et täht a muutub täheks b ($Q(a, b)$ on maatriksi Q element tähele a vastavast reast ning tähele b vastavast veerust). Kui Q on ühikmaatriks, siis iga täht jääb iseendaks.

Definitsioon 1.2. *Juhuslikku jada (st juhuslike suuruste jada) nimetatakse iid (ingl k* independent and identically distributed*) jadaks, kui kõik selle liikmed on sama jaotusega ja sõltumatud.*

Olgu $\xi_1, \xi_2, ...$ sõltumatud ja sama jaotusega (iid) juhuslikud suurused. Defineerime mutatsioonid $F_1, F_2, ...$ järgmiselt: $F_i(a) := f(a, \xi_i)$ iga $a \in \mathcal{A}$ korral. Kõik mutatsioonid $F_1, F_2, ...$ on siis sama üleminekumaatriksiga.

Olgu $a_1, a_2, ...$ etteantud tähed. Paarid $(a_1, \xi_1), (a_2, \xi_2), ...$ on sõltumatud, kuid pole üldjuhul sama jaotusega. Seega juhuslikud suurused

$$F_1(a_1), F_2(a_2), ... = f(a_1, \xi_1), f(a_2, \xi_2), ...$$

on sõltumatud, kuid ei pruugi olla samast jaotusest.

Olgu $Z_1, Z_2, ...$ iid juhuslikud suurused tähestikul $\mathcal{A}$, mis on sõltumatud juhuslikest suurustest $\xi_1, \xi_2, ...$ Rakendades selle jada juhuslikele suurustele mutatsioone $F_1, F_2, ...$, saame muteerunud juhuslikud suurused $F_1(Z_1), F_2(Z_2), ...$

Omadus 1.1. *$F_1(Z_1), F_2(Z_2), ...$ on iid.*

Tõestus. Kuna $Z_1, Z_2, ...$ on samast jaotusest, $\xi_1, \xi_2, ...$ on samast jaotusest ning $Z_1, Z_2, ..., \xi_1, \xi_2, ...$ on sõltumatud, siis paarid $(Z_1, \xi_1), (Z_2, \xi_2), ...$ on sõltumatud ja sama jaotusega. Seega juhuslikud suurused

$$F_1(Z_1), F_2(Z_2), ... = f(Z_1, \xi_1), f(Z_2, \xi_2), ...$$

on sõltumatud ja samast jaotusest. ■

Milline peab olema üleminekumaatriks Q, et $F_i(Z_i)$ oleks samast jaotusest kui Z_i? Ühe võimalusena võib Q olla ühikmaatriks. Vaatleme veel ühte võimalust. Olgu tähestikuks $\mathcal{A} = \{z_1, ..., z_n\}$. Kasutame lühendatud kirjaviisi:

$$\{Z_i = z_j\} = \{z_j\} \quad \forall\, j \in \{1, ..., n\}.$$

Olgu

$$Q = \begin{array}{c} \\ z_1 \\ z_2 \\ \vdots \\ z_n \end{array} \begin{array}{c} \begin{array}{cccc} z_1 & z_2 & ... & z_n \end{array} \\ \begin{pmatrix} P(z_1) & P(z_2) & ... & P(z_n) \\ P(z_1) & P(z_2) & ... & P(z_n) \\ \vdots & \vdots & \ddots & \vdots \\ P(z_1) & P(z_2) & ... & P(z_n) \end{pmatrix} \end{array}.$$

Siis

$$P(F_i(Z_i) = z_j) = P(z_1)P(z_j) + ... + P(z_n)P(z_j) = P(z_j) \cdot 1$$

ehk $F_i(Z_i)$ on samast jaotusest kui Z_i.

1.2 Kadumised

Jadast $F_1(Z_1), F_2(Z_2), ...$ saadakse peale osade tähtede kustutamist jada $X_1, X_2, ...$ Kustutamine toimub jaotusega $Be(p)$ iid juhuslike suuruste $D_1^x, D_2^x, ...$ abil – kui $D_i^x = 1$, siis $F_i(Z_i)$ jääb alles, vastasel juhul kaob. Kogu järgneva töö jooksul eeldame, et $p > 0$, vastasel korral ei jää kadumiste tagajärjel alles ühtegi tähte.

Olgu alles jäänud juhuslikust suurusest $F_k(Z_k)$ jadas eespool pool kaduma läinud s ($0 \leq s \leq k-1$) juhuslikku suurust, st $X_{k-s} = F_k(Z_k)$. Samaväärselt võime kirjutada, et

$$\sum_{j=1}^{k} D_j^x = k - s \quad \text{ja} \quad D_k^x = 1.$$

Võtame $i = k - s$. Seega

$$X_i = F_k(Z_k) \quad \text{parajasti siis, kui} \quad D_k^x = 1 \quad \text{ja} \quad \sum_{j=1}^{k} D_j^x = i.$$

Kui $X_i = F_k(Z_k)$, siis nimetame juhuslikku suurust Z_k juhusliku suuruse X_i *eellaseks* ning juhuslikku suurust X_i juhusliku suuruse Z_k *järglaseks*. Juhusliku suuruse X_i eellase indeks on juhuslik suurus, mida tähistame sümboliga K_i. Pole raske näha, et $K_i \geq i$.

Näide 1.1. Olgu $D_1^x, ..., D_8^x$ antud järgmise tabeliga:

i	1	**2**	3	4	5	**6**	7	**8**
D_i^x	0	**1**	0	0	0	**1**	0	**1**

Siis $K_1 = 2, K_2 = 6, K_3 = 8$ ning $X_1 = F_2(Z_2), X_2 = F_6(Z_6), X_3 = F_8(Z_8)$.

1.3 Järglased

Olgu $\eta_1, \eta_2, ...$ iid juhuslikud suurused, mis on sama jaotusega kui ξ_i ning sõltumatud juhuslikest suurustest $Z_1, Z_2, ..., \xi_1, \xi_2, ...$ Lisaks olgu $G_i(a) := f(a, \eta_i)$. Siis $G_1, G_2, ...$ on sõltumatud mutatsioonid ja $G_1(Z_1), G_2(Z_2), ...$ iid jada.

Kuna juhuslik suurus Z_i pole iseendast sõltumatu (eeldame, et ta ei ole konstant), siis paarid $(Z_i, \xi_i), (Z_i, \eta_i)$ pole sõltumatud (kuid on sama jaotusega). Seetõttu juhuslikud suurused $F_i(Z_i) = f(Z_i, \xi_i)$ ja $G_i(Z_i) = f(Z_i, \eta_i)$ on sama jaotusega, kuid pole üldjuhul sõltumatud.

Defineerime funktsiooni g järgmiselt:

$$g(Z, \xi, \eta) = (f(Z, \xi), f(Z, \eta)).$$

Kuna kolmikute jada $(Z_1, \xi_1, \eta_1), (Z_2, \xi_2, \eta_2), ...$ on iid, siis ka paaride jada

$$(F_1(Z_1), G_1(Z_1)),\ (F_2(Z_2), G_2(Z_2)),\ ... = g(Z_1, \xi_1, \eta_1), g(Z_2, \xi_2, \eta_2), ...$$

on iid.

Jadast $G_1(Z_1), G_2(Z_2), ...$ saame jada $Y_1, Y_2, ...$ peale osade tähtede kustutamist. See toimub iid Bernoulli jaotusega juhuslike suuruste $D_1^y, D_2^y, ...$ abil, allesjäämise tõenäosus on endiselt p.

Kui $Y_i = G_k(Z_k)$, siis nimetame juhuslikku suurust Z_k juhusliku suuruse Y_i *eellaseks* ning juhuslikku suurust Y_i juhusliku suuruse Z_k *järglaseks*. Juhusliku suuruse Y_i eellase indeks on juhuslik suurus, mida tähistame sümboliga L_i.

Jadad $X_1, X_2, ...$ ja $Y_1, Y_2, ...$ sõltuvad juhuslikest suurustest

$$Z_1, Z_2, ..., \xi_1, \xi_2, ..., \eta_1, \eta_2, ..., D_1^x, D_2^x, ..., D_1^y, D_2^y, ..., \tag{1.1}$$

kõik need juhuslikud suurused on omavahel sõltumatud. Jadasid $X_1, X_2, ...$ ja $Y_1, Y_2, ...$ nimetatakse jada $Z_1, Z_2, ...$ *järglasteks*; jada $Z_1, Z_2, ...$ nimetatakse jadade $X_1, X_2, ...$ ja $Y_1, Y_2, ...$ *eellaseks*.

Näide 1.2. Olgu meil lõplik jada $DDACBA$. Alljärgnevalt on näidatud järglaste moodustumine sellest jadast. Kahe otsaga noolega on näidatud, milleks jada element muutus mutatsiooni tagajärjel (ilma nooleta jada liikmed jäid samaks). Ühe otsaga noolega on näidatud kadumised. Järglasjadades on paksus kirjas märgitud ühise eellasega juhuslikud suurused (antud juhul on selliseid paare ainult üks).

$$\begin{array}{ccccc} & B \quad D & & & \\ & \updownarrow \quad \updownarrow & & \uparrow \quad \uparrow & \\ DDACBA & \longrightarrow & DBACBD & \longrightarrow & DBCD \\ & C & & & \\ & \updownarrow & & \uparrow\uparrow \quad \uparrow & \\ DDACBA & \longrightarrow & DDCCBA & \longrightarrow & CCB \end{array}$$

Lause 1.1. $P(X_i = x) = P(F_1(Z_1) = x);\ P(Y_i = y) = P(F_1(Z_1) = y)$.

Tõestus. Täistõenäosuse valemi järgi

$$P(X_i = x) = \sum_{k=1}^{\infty} P(X_i = x \mid K_i = k) \cdot P(K_i = k) = \sum_{k=1}^{\infty} P(F_k(Z_k) = x) \cdot P(K_i = k).$$

Omaduse 1.1 põhjal aga teame, et $F_1(Z_1), F_2(Z_2), ...$ on sama jaotusega, seega

$$P(X_i = x) = \sum_{k=1}^{\infty} P(F_1(Z_1) = x) \cdot P(K_i = k) = P(F_1(Z_1) = x) \cdot 1.$$

Analoogiliselt $P(Y_i = y) = P(G_1(Z_1) = y)$. Kuna $F_1(Z_1)$ ja $G_1(Z_1)$ on sama jaotusega, siis

$$P(Y_i = y) = P(F_1(Z_1) = y).$$

■

Märkus 1.1. Juhusliku suuruse (või juhusliku vektori) võimalike väärtuste hulka märgime sama sümboliga kui juhuslikku suurust (juhuslikku vektorit) ennast.

Omadus 1.2. *Jada $X_1, X_2, ...$ on iid ja jada $Y_1, Y_2, ...$ on iid.*

Tõestus. Lause 1.1 põhjal $X_1, X_2, ...$ on sama jaotusega.

Olgu $j \in \{2, 3, ...\}$, $1 \leq i_1 < ... < i_j$ ning $x_1, ..., x_j \in \mathcal{A}$. Olgu $K := (K_{i_1}, ..., K_{i_j})$. Täistõenäosuse valemi järgi

$$P(X_{i_1} = x_1,\ ...,\ X_{i_j} = x_j) = \sum_{(k_1,...,k_j)\in K} P^*(k_1, ..., k_j) \cdot P(K = (k_1, ..., k_j)), \tag{1.2}$$

kus

$$\begin{aligned} P^*(k_1, ..., k_j) &:= P(X_{i_1} = x_1,\ ...,\ X_{i_j} = x_j \mid K = (k_1, ..., k_j)) \\ &= P(F_{k_1}(Z_{k_1}) = x_1,\ ...,\ F_{k_j}(Z_{k_j}) = x_j). \end{aligned}$$

Omaduse 1.1 põhjal teame, et $F_1(Z_1), F_2(Z_2), ...$ on sõltumatud ja sama jaotusega, seega

$$P^*(k_1, ..., k_j) = P(F_1(Z_1) = x_1) \cdot ... \cdot P(F_1(Z_1) = x_j) =: P^*.$$

Samasuses (1.2) saame avaldise $P^* = P^*(k_1, ..., k_j)$ summa ette tuua:

$$P(X_{i_1} = x_1,\ ...,\ X_{i_j} = x_j) = P^* \cdot 1 \overset{\text{lause 1.1}}{=} P(X_{i_1} = x_1) \cdot ... \cdot P(X_{i_j} = x_j).$$

Seega $X_1, X_2, ...$ on iid. Analoogiliselt saab näidata, et $Y_1, Y_2, ...$ on iid. ■

Definitsioon 1.3. *Punktide $a, b \in \mathbb{R}$ kumeraks kombinatsiooniks nimetatakse punkti*

$$\lambda a + (1 - \lambda)b, \quad 0 \leq \lambda \leq 1.$$

Kui $0 < \lambda < 1$, siis nimetatakse seda punkti rangeks kumeraks kombinatsiooniks.

Lause 1.2. *(i) Range kumer kombinatsioon kujul*

$$\lambda a + (1-\lambda)b, \quad 0 < \lambda < 1$$

on võrdne punktiga b parajasti siis, kui a = b.

(ii) Kumerad kombinatsioonid kujul

$$\lambda_1 a + (1-\lambda_1)b, \quad \lambda_2 a + (1-\lambda_2)b \quad (\lambda_1 \neq \lambda_2), \tag{1.3}$$

on võrdsed parajasti siis, kui a = b.

Tõestus. (i) Piisavus.

$$a = b \;\; \Rightarrow \;\; \lambda a + (1-\lambda)b = \lambda b + (1-\lambda)b = b.$$

Tarvilikkus. Olgu $a \neq b$. Oletame vastuväiteliselt, et $\lambda a + (1-\lambda)b = b$. Siis

$$\lambda a + (1-\lambda)b = \lambda b + (1-\lambda)b \;\; \Rightarrow \;\; a = b,$$

mis on vastuolu.

(ii) Piisavus. Vt osa (i) piisavuse tõestus.

Tarvilikkus. Olgu $a \neq b$ ja $\lambda_2 = \lambda_1 + \epsilon$, kus $\epsilon \neq 0$. Oletame vastuväiteliselt, et kumerad kombinatsioonid kujul (1.3) on võrdsed. Siis

$$(\lambda_1 + \epsilon)a + (1 - \lambda_1 - \epsilon)b = \lambda_1 a + (1-\lambda_1)b \;\; \Rightarrow \;\; \epsilon a - \epsilon b = 0 \;\; \Rightarrow \;\; a = b,$$

mis on vastuolu. ■

Märkus 1.2. Järgnevas kasutame lühendatud kirjaviisi:

$$\{X_i = x_i\} = \{x_i\},$$
$$\{Y_i = y_i\} = \{y_i\},$$
$$\{X = x\} = \{x\},$$
$$\{Y = y\} = \{y\}.$$

Omadus 1.3. *Kui $p < 1$ (meenutame, et $p = P(D_1^x = 1) = P(D_1^y = 1)$), siis X_i ja Y_j pole üldiselt sõltumatud.*

Tõestus. Olgu $p < 1$. Olgu $K := K_i$, $X := X_i$, $L := L_j$ ning $Y := Y_j$. Tähistame

$$P_1 := P(x, y,\ K = L),$$
$$P_2 := P(x, y,\ K \neq L).$$

Paneme tähele, et

$$P(x, y) = P_1 + P_2.$$

Avaldame P_1.

$$P_1 = \sum_{k=1}^{\infty} P_1^*(k) \cdot P(K = L = k), \tag{1.4}$$

kus

$$P_1^*(k) := P(x, y \mid K = L = k).$$

Paneme tähele, et juhuslik suurus K sõltub vaid juhuslikest suurustest $D_1^x, D_2^x, ...$ ning juhuslik suurus L sõltub vaid juhuslikest suurustest $D_1^y, D_2^y, ...$ Juhuslike suuruste (1.1) sõltumatuse tõttu

$$P_1^*(k) = \sum_{z \in \mathcal{A}} P(Z_k = z) \cdot P(f(z, \xi_k) = x) \cdot P(f(z, \eta_k) = y).$$

Kuna $\xi_1, \xi_2, ...$ on sama jaotusega, $\eta_1, \eta_2, ...$ on sama jaotusega ning $Z_1, Z_2, ...$ on sama jaotusega, siis

$$P_1^*(k) = P_1^* := \sum_{z \in \mathcal{A}} P(Z_1 = z) \cdot P(f(z, \xi_1) = x) \cdot P(f(z, \eta_1) = y). \tag{1.5}$$

Summast (1.4) saab $P_1^* = P_1^*(k)$ sulgude ette tuua, saame samasuse

$$P_1 = P_1^* \cdot P(K = L).$$

Avaldame P_2.

$$P_2 = \sum_{k \neq l} P_2^*(k, l) \cdot P(K = k, L = l), \tag{1.6}$$

kus

$$P_2^*(k, l) := P(x, y \mid K = k,\ L = l).$$

Kui $k \neq l$, siis $F_k(Z_k)$ ja $G_l(Z_l)$ on sõltumatud:

$$P_2^*(k, l) = P(F_k(Z_k) = x) \cdot P(G_l(Z_l) = y) = P(F_1(Z_1) = x) \cdot P(F_1(Z_1) = y) =: P_2^*.$$

Summast (1.6) saab $P_2^* = P_2^*(k,l)$ sulgude ette tuua. Rakendades lauset 1.1, saame samasuse

$$P_2 = P(x)P(y) \cdot P(K \neq L).$$

Kokkuvõttes

$$P(x,y) = P_1 + P_2 = P_1^* \cdot P(K = L) + P(x)P(y) \cdot P(K \neq L). \tag{1.7}$$

Juhusliku suuruse K väärtus on üheselt määratud ainult siis, kui $p = 1$; vastasel korral on ta võimalike väärtuste hulgaks $\{i, i+1, ...\}$. Analoogiliselt on juhusliku suuruse L väärtus üheselt määratud samuti ainult siis, kui $p = 1$; vastasel korral on ta võimalike väärtuste hulgaks $\{j, j+1, ...\}$. Seega, tõenäosus $P(K = L)$ saab võrduda arvuga 1 või arvuga 0 ainult siis, kui $p = 1$. Antud tõestuses eeldame, et $p < 1$. Seega avaldise (1.7) näol on tegemist range kumera kombinatsiooniga, mistõttu saame rakendada lause 1.2 osa (i). Saame samaväärsuse

$$P = P(x)P(y) \quad \Leftrightarrow \quad P_1^* = P(x)P(y). \tag{1.8}$$

Lause 1.1 põhjal

$$\begin{aligned} &P(x)P(y) = P(F_1(Z_1) = x) \cdot P(F_1(Z_1) = y) \\ &= \left(\sum_{z \in \mathcal{A}} P(Z_1 = z) \cdot P(f(z, \xi_1) = x) \right) \left(\sum_{z \in \mathcal{A}} P(Z_1 = z) \cdot P(f(z, \eta_1) = y) \right). \end{aligned} \tag{1.9}$$

Tähistame $p(z) := P(Z_1 = z),\ q(z,x) := P(f(z,\xi_1) = x)$. Kuna ξ_1, η_1 on sama jaotusega, siis $P(f(z,\eta_1) = y) = q(z,y)$. Kasutades suuruse $P(x)P(y)$ esitust kujul (1.9), suuruse P_1^* esitust kujul (1.5), ning arvestades, et kehtib samaväärsus (1.8), piisab näidata, et võrdus

$$\left(\sum_{z \in \mathcal{A}} p(z)q(z,x) \right) \left(\sum_{z \in \mathcal{A}} p(z)q(z,y) \right) = \sum_{z \in \mathcal{A}} p(z)q(z,x)q(z,y) \tag{1.10}$$

ei kehti. Tõepoolest, võttes tähestiku $\mathcal{A}$ pikkuseks näiteks 2 ning

$$\begin{aligned} &x = z_1, && y = z_2, \\ &p(z_1) = 0.4, && p(z_2) = 0.6, \\ &q(z_1, z_1) = 0.1, && q(z_2, z_1) = 0.2, \\ &q(z_1, z_2) = 0.9, && q(z_2, z_2) = 0.8, \end{aligned}$$

saame võrduse (1.10) vasakpoolse avaldise väärtuseks

$$(0.4 \cdot 0.1 + 0.6 \cdot 0.2)(0.4 \cdot 0.9 + 0.6 \cdot 0.8) \approx 0.134$$

ja parempoolse avaldise väärtuseks

$$0.4 \cdot 0.1 \cdot 0.9 + 0.6 \cdot 0.2 \cdot 0.8 = 0.132.$$

■

Järeldus 1.1. *Kui maatriksi Q read on võrdsed, siis jadad $X_1, X_2, ...$ ja $Y_1, Y_2, ...$ on sõltumatud.*

Tõestus. Olgu maatriksi Q read võrdsed. See tähendab, et iga $x \in \mathcal{A}$ korral

$$q(z_i, x) = q(z_j, x) \quad \forall z_i, z_j \in \mathcal{A}.$$

Jagame võrduse (1.10) suurusega $q(x) := q(z, x)$ läbi. Kuna $\sum_{z \in \mathcal{A}} p(z) = 1$, siis võrdus (1.10) kehtib iga $x, y \in \mathcal{A}$ korral. Seega X_i, Y_j on sõltumatud iga i ja j korral; sellest järeldub, et jadad $X_1, X_2, ...$ ja $Y_1, Y_2, ...$ on sõltumatud. ■

Märkus 1.3. Juhuslikke suurusi X_i, Y_j nimetatakse *sugulasteks*, kui neil on ühine eellane (st $K_i = L_j$). Juhul $p = 1$ on X_i ja Y_j sugulased parajasti siis, kui $i = j$. Seose (1.7) põhjal on avaldis (1.5) $X = X_i$ ja $Y = Y_j$ ühisjaotus tingimusel, et nad on sugulased, ning avaldis (1.9) X_i ja Y_j ühisjaotus tingimusel, et nad ei ole sugulased.

Omadus 1.4. *Kui $p < 1$, siis paarid $(X_1, Y_1), (X_2, Y_2), ...$ pole üldiselt sõltumatud.*

Tõestus. Omaduse 1.3 põhjal paarid $(X_i, Y_i), (X_j, Y_j)$ pole üldiselt sõltumatud, kui $p < 1$; sellest järeldub tõestatav väide. ■

Lause 1.3. *Kui $p = 1$, siis siis kõik paarid $(X_1, Y_1), (X_2, Y_2), ...$ on sõltumatud, aga iga i korral X_i, Y_i üldiselt pole.*

Tõestus. Kui $p = 1$, siis iga i korral $X_i = F(Z_i)$ ja $Y_i = G(Z_i)$. Juhuslikud suurused $F(Z_i)$ ja $G(Z_i)$ üldiselt pole sõltumatud, aga paarid

$$(F_1(Z_1), G_1(Z_1)), (F_2(Z_2), G_2(Z_2)), ...$$

on sõltumatud. ■

Lause 1.4. *Kui $p < 1$, siis iga $m \in \{0, 1, ...\}$ korral $P(K_n = L_{n+m}) \to 0$ protsessis $n \to \infty$.*

Tõestus. a) Tõestame lause $m = 0$ korral. Olgu $K_0 := 0$ ja $L_0 := 0$. Defineerime:

$$T_i^x = K_i - K_{i-1},$$
$$T_i^y = L_i - L_{i-1},$$

$i = 1, 2, ...$ Paneme tähele, et

$$K_n = \sum_{i=1}^{n} T_i^x, \quad L_n = \sum_{i=1}^{n} T_i^y.$$

Väite tõestamiseks $m = 0$ korral näitame, et $P\left(\sum_{i=1}^{n}(T_i^x - T_i^y) = 0\right) \to 0$ protsessis $n \to \infty$.

Jaotises "Juhuslikud suurused $T_1, T_2, ...$" näidatakse, et juhuslikud suurused

$$T_1, T_2, ... = T_1^x, T_2^x, ...$$

on sama geomeetrilise jaotusega ja sõltumatud. Analoogiliselt saab näidata, et juhuslikud suurused $T_1^y, T_2^y, ...$ on sama geomeetrilise jaotusega ja sõltumatud, kusjuures T_i^x, T_i^y on sama jaotusega. Kuna juhuslikud suurused $T_1^x, T_2^x, ...$ sõltuvad vaid vektorist $D^x := (D_1^x, D_2^x, ...)$ ja juhuslikud suurused $T_1^y, T_2^y, ...$ sõltuvad vaid vektorist $D^y := (D_1^y, D_2^y, ...)$ ning vektorid D^x, D^y on sõltumatud, siis jadad $T_1^x, T_2^x, ...$ ja $T_1^y, T_2^y, ...$ on sõltumatud. Eelneva põhjal on paarid $(T_1^x, T_1^y), (T_2^x, T_2^y), ...$ sõltumatud. Olgu

$$\zeta_i := (T_i^x - T_i^y), \quad i = 1, 2, ...$$

Jada $\zeta_1, \zeta_2, ...$ on iid, kusjuures $E\zeta_i = ET_i^x - ET_i^y = 0$. Kuna ζ_i on kahe geomeetrilise jaotusega juhusliku suuruse vahe, siis tal leidub standardhälve – olgu selleks σ. Tsentraalse piirteoreemi kohaselt

$$\frac{\sum_{i=1}^{n} \zeta_i}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

protsessis $n \to \infty$. Seega

$$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^{n} \zeta_i}{\sigma\sqrt{n}} = 0\right) = 0$$

ehk

$$\lim_{n\to\infty} P\left(\sum_{i=1}^{n} \zeta_i = 0\right) = 0.$$

b) Tõestame lause $m \geq 1$ korral. Olgu

$$\omega_n := \sum_{i=n+1}^{n+m} T_i^y.$$

Väite tõestamiseks $m \geq 1$ korral näitame, et $\lim_{n\to\infty} P\left(\sum_{i=1}^{n} \zeta_i = \omega_n\right) = 0$. Avaldame:

$$\lim_{n\to\infty} P\left(\sum_{i=1}^{n} \zeta_i = \omega_n\right) = \lim_{n\to\infty} \sum_{k=m}^{\infty} P\left(\sum_{i=1}^{n} \zeta_i = k \mid \omega_n = k\right) \cdot P(\omega_n = k).$$

Paneme tähele, et juhuslikud suurused $\zeta_1, ..., \zeta_n$ ei sõltu juhuslikust suurusest ω_n. Lisaks on juhuslikud suurused $\omega_0, \omega_1, ...$ sama jaotusega. Seega

$$\lim_{n\to\infty} P\left(\sum_{i=1}^{n} \zeta_i = \omega_n\right) = \lim_{n\to\infty} \sum_{k=m}^{\infty} P\left(\sum_{i=1}^{n} \zeta_i = k\right) \cdot P(\omega_0 = k).$$

Paneme tähele, et

$$P\left(\sum_{i=1}^{n} \zeta_i = k\right) \cdot P(\omega_0 = k) \leq P(\omega_0 = k) \quad \forall n, k$$

ja

$$\sum_{k=m}^{\infty} P(\omega_0 = k) = 1 < \infty.$$

Lisaks

$$\lim_{n\to\infty} P\left(\sum_{i=1}^{n} \zeta_i = k\right) \cdot P(\omega_0 = k) = 0 \quad \forall k,$$

sest

$$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^{n} \zeta_i}{\sigma\sqrt{n}} = \frac{k}{\sigma\sqrt{n}}\right) = 0 \quad \forall k.$$

Seega domineeritud koondumise teoreemi kohaselt

$$\lim_{n\to\infty} P\left(\sum_{i=1}^{n} \zeta_i = \omega_n\right) = \sum_{k=m}^{\infty} 0 = 0.$$

■

Järeldus 1.2. *Kui $p < 1$, siis iga $m \in \{0, 1, ...\}$ korral $P(L_n = K_{n+m}) \to 0$ protsessis $n \to \infty$.*

Tõestus. Tõestus on sümmeetriline lause 1.4 tõestusega. ■

Järeldus 1.3. *Kui $p < 1$, siis iga $m \in \{1, 2, ...\}$ korral*

$$P(K_n \in \{L_{n-m}, L_{n-m+1}, ..., L_{n+m-1}, L_{n+m}\}) \to 0.$$

protsessis $n \to \infty$.

Tõestus. Lause 1.4 ja järelduse 1.2 põhjal

$$\lim_{n\to\infty} P(K_n \in \{L_{n-m}, L_{n-m+1}, ..., L_{n+m-1}, L_{n+m}\}) = \sum_{i=n-m}^{n+m} \lim_{n\to\infty} P(K_n = L_i) = 0.$$

■

Märkus 1.4. Lause 1.4 ja järeldus 1.2 tähendavad kokku võttes sisuliselt, et iga $m \in \mathbb{Z}$ korral tõenäosus, et juhuslik suurus X_i on juhusliku suuruse Y_{i+m} sugulane, koondub nulli protsessis $i \to \infty$. Järeldus 1.3 tähendab sisuliselt, et iga $m \in \{1, 2, ...\}$ korral tõenäosus, et juhuslik suurus X_i on sugulane mõne juhusliku suurusega hulgast $\{Y_{i-m}, Y_{i-m+1}, ..., Y_{i+m-1}, Y_{i+m}\}$, koondub nulli protsessis $i \to \infty$.

Omadus 1.5. *Kui $p < 1$, siis*

(i) paarid $(X_1, Y_1), (X_2, Y_2), ...$ pole üldiselt sama jaotusega,

(ii) kahedimensionaalne protsess $\{(X_i, Y_i)\}_i$ pole üldiselt statsionaarne.

Tõestus. (i) Samasuse (1.7) põhjal

$$P(x_i, y_i) = P_1 \cdot P(K_i = L_i) + P_2 \cdot (1 - P(K_i = L_i)), \tag{1.11}$$

kus

$$P_1 := \sum_{z \in \mathcal{A}} P(Z_1 = z) \cdot P(f(z, \xi_1) = x_i) \cdot P(f(z, \mu_1) = y_i).$$

ja

$$\begin{aligned} P_2 &:= P(x_i)P(y_i) \\ &\overset{(1.9)}{=} \left(\sum_{z \in \mathcal{A}} P(Z_1 = z) \cdot P(f(z, \xi_1) = x_i) \right) \left(\sum_{z \in \mathcal{A}} P(Z_1 = z) \cdot P(f(z, \mu_1) = y_i) \right). \end{aligned}$$

Avaldise (1.11) näol on tegemist kumera kombinatsiooniga. Lause 1.4 põhjal leiduvad n, m nii, et $P(K_n = L_n) \neq P(K_m = L_m)$. Lisaks on P_1 ja P_2 iga i korral samad. Seega saame rakendada lause 1.2 osa (ii). Et $P_1 \neq P_2$, siis $P(x_n, y_n) \neq P(x_m, y_m)$.

(ii) Piisab näidata, et leiduvad (X_n, Y_n) ja (X_m, Y_m) nii, et $P(x_n, y_n) \neq P(x_m, y_m)$. See järeldub osast (i). ■

Omadus 1.6. *Kui $p < 1$, siis iga $m \in \mathbb{Z}$ korral $|P(x_i, y_{i+m}) - P(x_i)P(y_{i+m})| \to 0$ protsessis $i \to \infty$.*

Tõestus. Samasuse (1.7) põhjal

$$P(x_i, y_{i+m}) = P(x_i, y_{i+m} \mid K_i = L_{i+m}) \cdot P(K_i = L_{i+m}) + P(x_i)P(y_{i+m}) \cdot (1 - P(K_i = L_{i+m})).$$

Lause 1.4 ja järelduse 1.2 põhjal $P(K_i = L_{i+m}) \to 0$ protsessi $i \to \infty$. ■

Märkus 1.5. Olgu $p < 1$. Omadus 1.6 ütleb, et juhuslikud suurused X_i, Y_{i+m} lähenevad iga $m \in \mathbb{Z}$ korral sõltumatusele protsessis $i \to \infty$. Olgu lisaks

$$m_1 \leq m_2, n_1 \leq n_2, \quad m_1, m_2, n_1, n_2 \in \mathbb{Z},$$
$$D_i^x(m_1, m_2) := (X_{i+m_1}, X_{i+m_1+1}, ..., X_{i+m_2-1}, X_{i+m_2}),$$
$$D_i^y(n_1, n_2) := (Y_{i+n_1}, Y_{i+n_1+1}, ..., Y_{i+n_2-1}, Y_{i+n_2}).$$

Omadust 1.6 kasutades saab näidata, et

$$|P(D_i^x(m_1, m_2) = d_1,\ P(D_i^y(n_1, n_2) = d_2) - P(D_i^x(m_1, m_2) = d_1) \cdot P(D_i^y(n_1, n_2) = d_2)| \to 0$$

protsessis $i \to \infty$. Teisisõnu vektorid $D_i^x(m_1, m_2), D_i^y(n_1, n_2)$ lähenevad sõltumatusele protsessis $i \to \infty$.

Omadus 1.7. $|P(X_i = x,\ Y_n = y) - P(X_i = x) \cdot P(Y_n = y)| \to 0$ *protsessis* $n \to \infty$.

Tõestus. Kui $p = 1$, siis $X_i = F(Z_i)$ ja $Y_n = G(Z_n)$ iga n korral ning väide kehtib triviaalselt. Olgu $p < 1$. Samasuse (1.7) põhjal

$$P(X_i = x,\ Y_n = y)$$
$$= P(X_i = x,\ Y_n = y \mid K_i = L_n) \cdot P(K_i = L_n) + P(X_i = x) \cdot P(Y_n = y) \cdot (1 - P(K_i = L_n)).$$

Piisab näidata, et $P(K_i = L_n) \to 0$ protsessis $n \to \infty$. Et iga n korral $L_n \geq n$ ning juhuslikud suurused L_n, K_i on sõltumatud, siis

$$\lim_{n\to\infty} P(K_i = L_n) \leq \lim_{n\to\infty} P(K_i \geq n,\ L_n \geq n) = \lim_{n\to\infty} P(K_i \geq n) \cdot \lim_{n\to\infty} P(L_n \geq n) = 0 \cdot 1 = 0.$$

■

Märkus 1.6. Omadus 1.7 ütleb, et juhuslikud suurused X_i, Y_n lähenevad sõltumatusele protsessis $n \to \infty$. Olgu

$$m_1, m_2, n_1, n_2, D_i^x(m_1, m_2), D_n^x(n_1, n_2)$$

sellised nagu nad on märkuses 1.5 defineeritud. Omadust 1.7 kasutades saab näidata, et

$$|P(D_i^x(m_1, m_2) = d_1,\ P(D_n^y(n_1, n_2) = d_2) - P(D_i^x(m_1, m_2) = d_1) \cdot P(D_n^y(n_1, n_2) = d_2)| \to 0$$

protsessis $n \to \infty$. Teisisõnu vektorid $D_i^x(m_1, m_2), D_n^y(n_1, n_2)$ lähenevad sõltumatusele protsessis $n \to \infty$.

2. SIMULATSIOONID

2.1 Põhimõisted

Definitsioon 2.1. *Jada $y_1, ..., y_k$ nimetatakse jada $x_1, ..., x_m$ osajadaks, kui leiduvad indeksid $n_1 < ... < n_k \leq m$ nii, et $y_1, ..., y_k = x_{n_1}, ..., x_{n_k}$ $(k \leq m)$.*

Teisisõnu jadast $x_1, ..., x_m$ saadakse osajada 0 kuni $m - 1$ tähe eemaldamise teel.

Definitsioon 2.2. *Kahe lõpliku pikkusega jada ühisjadaks nimetatakse jada, mis on nende mõlema osajadaks. Pikimaks ühisjadaks nimetatakse maksimaalse võimaliku pikkusega ühisjada.*

Näide 2.1. Kahe järgneva jada ühisjada on $ABCD$:

$$GAHABJTHCRDMW$$
$$QKLAODBCKBLCMDOBB$$

Vaatleme jadasid $x_1, ..., x_{k_x}$ ja $y_1, ..., y_{k_y}$ ning nende ühisjada $z_1, ..., z_k$ $(k \leq \min\{k_x, k_y\})$. Ühisjada definitsiooni kohaselt leiduvad indeksid $m_1 < ... < m_k$ ja $n_1 < ... < n_k$ nii, et

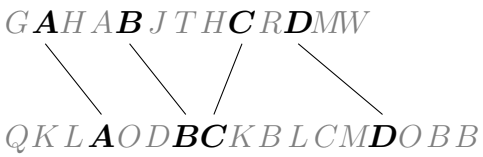
$$x_{m_1}, ..., x_{m_k} = y_{n_1}, ..., y_{n_k}.$$

Moodustame võrdsete elementidega paarid

$$(x_{m_1}, y_{n_1}), ..., (x_{m_k}, y_{n_k}).$$

Selliste paaride moodustamist nimetame vastavate elementide *ühendamiseks*, selliseid paare nimetame *ühendusteks*.

Asetades kaks jada üksteise kohale, võime nende vahelised ühendused kujutada joontena ühendatud elementide vahel. Siis ükski joon ei lõiku. Joonisel 2.1 on kujutatud näites 2.1 toodud jadad ning nende ühisjadale $ABCD$ vastavad ühendused.



Joonis 2.1: ühisjadale $ABCD$ vastavad ühendused

Kahe jada ühisjada pikkust võib vaadelda nende sarnasusskoorina. Edaspidi kasutamegi mõisteid "ühisjada pikkus" ja "sarnasusskoor" samas tähenduses.

Iid jada jaotuse all mõistame ta mis tahes liikme jaotust.

Olgu funktsioon L selline, et $L(X;Y)$ on jadade X, Y pikima ühisjada pikkus mis tahes lõplike juhuslike jadade X, Y korral. Olgu lõplikud iid jadad X, Y sõltumatud ning mõlemad jaotusega G. Siis järeldusena Kingmani subaditiivsest ergoodilisest teoreemist (vt [1]) leidub konstant γ^* nii, et

$$\frac{L(X;Y)}{n} \xrightarrow{\text{p.k.}} \gamma^*$$

protsessis $n \to \infty$.

Suurust γ^* nimetatakse Chvátal-Sankoffi konstandiks. Chvátal-Sankoffi konstandi täpne väärtus pole teada ühegi G väärtuse korral. Simulatsioonide teel on aga kindlaks tehtud, et näiteks $G = Be(0.5)$ korral on Chvátal-Sankoffi konstant ligikaudu väärtusega 0.81.

Olgu juhuslikud suurused

$$X_1, X_2, ..., Y_1, Y_2, ..., Z_1, Z_2, ...$$

sellised nagu nad on 1. peatükis defineeritud. Olgu iga n korral

$$X^n := X_1, ..., X_n,$$
$$Y^n := Y_1, ..., Y_n,$$
$$Z^n := Z_1, ..., Z_n$$

ja

$$L_n := L(X^n; Y^n).$$

Kuigi protsess $\{X_i, Y_i\}_i$ pole statsionaarne ning Kingmani subaditiivne ergoodiline teoreem antud juhul ei rakendu, leidub siiski konstant γ nii, et

$$\frac{L_n}{n} \xrightarrow{\text{p.k.}} \gamma$$

protsessis $n \to \infty$ [2]. Suurus γ sõltub juhuslike suuruste $Z_1, Z_2, ...$ jaotusest, üleminekumaatriksist Q ning tähe säilimistõenäosusest p.

2.2 Sissejuhatus simulatsioonidesse

Simulatsioonide eesmärk on kõigepealt kontrollida suuruse L_n/n koondumist konstandiks γ protsessis $n \to \infty$ ning seejärel uurida, kuidas funktsioon $\gamma = \gamma(p, Q)$ sõltub parameetritest p ja Q. Simulatsioonides võtame jada $Z_1, Z_2, ...$ jaotuseks $Be(0.5)$.

Tähestikule $\{0, 1\}$ vastava üleminekumaatriksi Q kirjeldamiseks kasutame suurusi ϵ_1, ϵ_2, kusjuures maatriks Q avaldub ϵ_1 ja ϵ_2 kaudu järgmiselt:

$$Q = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 1-\epsilon_1 & \epsilon_1 \\ \epsilon_2 & 1-\epsilon_2 \end{pmatrix} \end{array}. \tag{2.1}$$

Kui $\epsilon_1 = \epsilon_2$, siis kasutame nende mõlema märkimiseks tähist ϵ.

Paneme tähele, et kui kujul (2.1) toodud maatriks on sümmeetriline ning juhuslike suuruste $Z_1, Z_2, ...$ jaotuseks on $Be(0.5)$, siis

$$P(X_i = 0) = P(Y_i = 0) = \frac{1}{2}(1 - \epsilon) + \frac{1}{2}\epsilon = \frac{1}{2}$$

ja juhuslikud suurused $X_1, X_2, ..., Y_1, Y_2, ...$ on jaotusega $Be(0.5)$.

Simulatsioonides kasutatakse programmeerimiskeelt R [3]. Pikima ühisjada leidmiseks kasutatakse Needleman-Wunschi algoritmi, mis realiseeritakse paketi Biostrings [4] funktsiooni pairwiseAlignment kaudu.

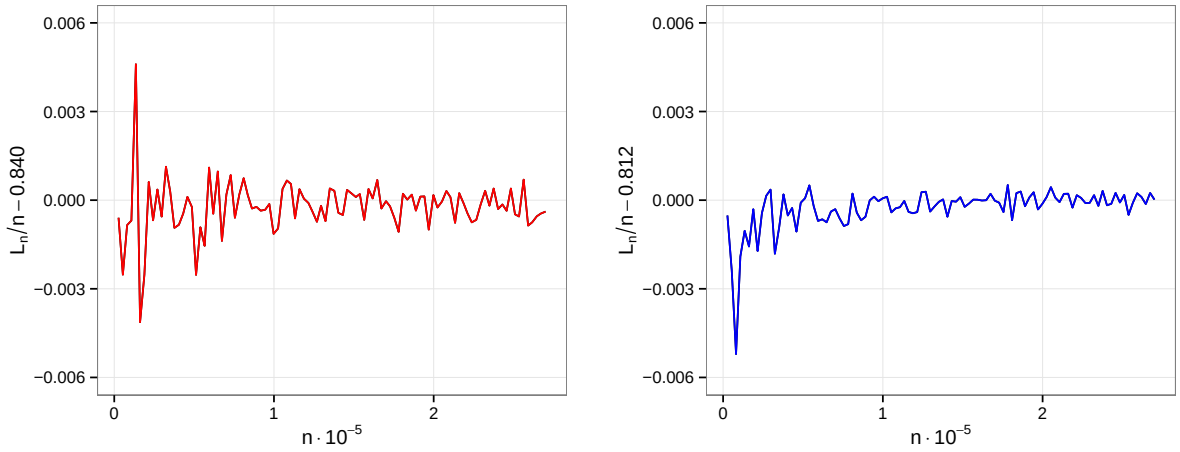
2.3 L_n/n koondumine

Simulatsioonid koondumise kohta on kujutatud joonisel 2.2. Simuleerimiseks genereeriti jadad Z^m (jaotusega $Be(0.5)$) – joonise 2.2a puhul

$$m = 3000, 6000, ..., 300\,000$$

ning joonise 2.2b puhul

$$m = 2700, 5400, ..., 270\,000.$$



(a) $\epsilon = 0.9,\ p = 0.9$ *(b)* $\epsilon = 0.5,\ p = 1$

Joonis 2.2: suuruse L_n/n koondumine

Jadast Z^m saadakse mutatsioonide ja kadumiste tagajärjel jadad X^{N_X}, Y^{N_Y}. Saamaks võrdse pikkusega jadad X^n ja Y^n, eemaldame vajadusel jada X^{N_X} või jada Y^{N_Y} lõpust elemente. Teisisõnu

$$n = \min\{N_X, N_Y\}.$$

Kuna joonise 2.2b puhul kadumisi pole, siis seal $n = m$.

Vaatleme joonist 2.2b. Kuna siin $\epsilon = 0.5$, siis jadad X^n ja Y^n on sõltumatud (järelduse 1.1 kohaselt on jadad $X_1, X_2, ...$ ja $Y_1, Y_2, ...$ sõltumatud, kui maatriksi Q read on võrdsed) ja jaotusega $Be(0.5)$ (sest Q on sümmeetriline). Seega joonisel 2.1b on kujutatud suuruse L_n/n koondumine sõltumatute jadade korral. Näeme, et koondumine tundub toimuvat oodatavalt ligikaudselt väärtuseks 0.81.

Vaatleme joonist 2.2a. Jällegi on jadad X^n ja Y^n jaotusega $Be(0.5)$, kuid seekord nad ei ole sõltumatud. Näeme, et joonis kinnitab koondumist, kusjuures koondumine tundub toimuvat ligikaudu väärtuseks 0.84. Selgitame, miks selline tulemus on võib-olla mõnevõrra üllatav. Kujutleme jadasid X^n, Y^n asetsemas üksteise kohal. Võiks arvata, et kui juhuslik suurus X_i asub oma sugulasest piisavalt kaugel, siis neid ei ühendata. Kui see on nii, siis järelduse 1.3 kohaselt tõenäosus, et X_i ühendatakse tema sugulasega, koondub nulli protsessis $i \to \infty$. Siis peaks aga L_n/n koonduma ligikaudu suuruseks 0.81. Miks koondus L_n/n suuruseks 0.84? Võib-olla sugulasi ühendatakse ka siis, kui nendevaheline vahemaa on suur? Järgmises jaotises näitame täpsemalt, et see on tõepoolest nii.

2.4 Alumine tõke suurusele γ

Käesolevas jaotises konstrueerime alumise tõkke suurusele $\gamma = \lim_{n\to\infty} L_n/n$ ning näitame seda tõket kasutades teoreetiliselt, et kui jadad X^n ja Y^n on jaotusega $Be(0.5)$, siis see ei tähenda, et γ on ligikaudu võrdne suurusega 0.81. Tehnilisem osa alumise tõkke tõestusest on toodud käesoleva jaotise lõpus.

Peale alumise tõkke konstrueerimise tutvustame sarnasusskoori D_n ja esitame simulatsioone tõkke kohta.

Esmalt defineerime aga alumise tõkke konstrueerimisel kasutatava funktsiooni M.

2.4.1 Ühisjada tagastav funktsioon M

Rakendades jadale Z^{K_n} mutatsioone $F_1, ..., F_{K_n}$ ja kadumisi $D_1^x, ...D_{K_n}^x$, saame jada X^n. Olgu

$$n_y := \sum_{j=1}^{K_n} D_j^y.$$

Rakendades jadale Z^{K_n} mutatsioone $G_1, ..., G_{K_n}$ ja kadumisi $D_1^y, ...D_{K_n}^y$, saame jada Y^{n_y}.

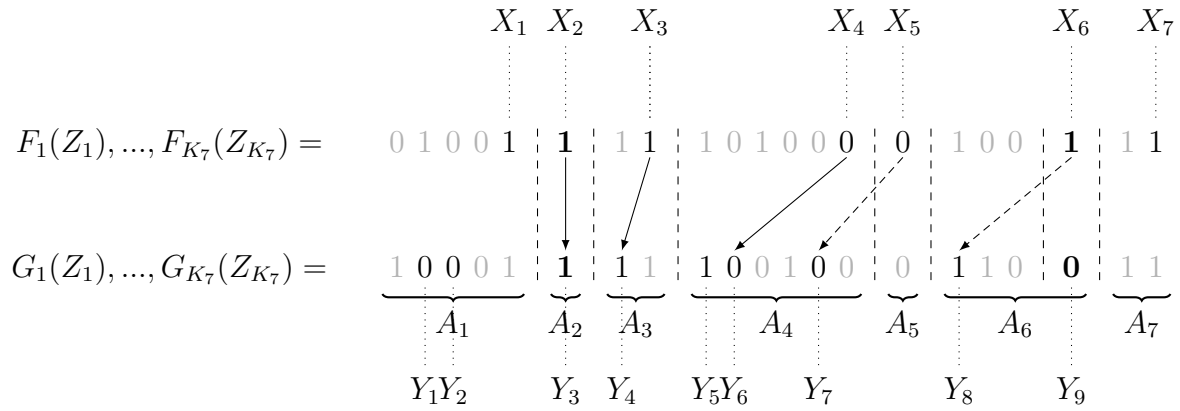
Defineerime $K_0 := 0$. Olgu hulgad A_i $(i = 1, 2, ...)$ defineeritud järgmiselt:

$$A_i := \{G_{K_{i-1}+1}(Z_{K_{i-1}+1}), ..., G_{K_i}(Z_{K_i})\}.$$

Jadade X^n ja Y^{n_y} ühisjada leidmist funktsiooni M abil illustreerib joonisel 2.3 toodud näide. Seal $n = 7$ ja $n_y = 9$. Halli värviga on märgitud elemendid, mis kaovad ära, ülejäänud jäävad alles. Paksus kirjas on märgitud elemendid, mille sugulane jääb alles. Pidevate nooltega on märgitud ühendused elementide vahel. Katkendlikud püstjooned on barjäärid ühenduste jaoks, st ükski ühendusjoon ei tohi nendega lõikuda. Paneme tähele, et funktsioon M ei pruugi tagastada pikimat ühisjada: katkendlike nooltega on tähistatud võimalikud ühendused, mis ei läheks konflikti teiste ühendustega, kuid mida siiski funktsiooni M eeskirja kohaselt ei lubata. Näites saadakse ühisjadaks 110.

Funktsioon M leiab jadade X^n ja Y^{n_y} ühisjada, konstrueerides ühendused järgmiselt:

1. kui juhuslikul suurusel X_i on temaga võrdne sugulane, siis nad ühendatakse $(i = 1, ..., n)$;



Joonis 2.3: jadade X^n ja Y^{n_y} ühisjada leidmine funktsiooniga M ($n = 7$)

2. kui juhuslikul suurusel X_i pole sugulast ning hulgas A_i leidub temaga võrdne mittekaduv element, siis X_i ühendatakse selle elemendiga ($i = 1, ..., n$).

Paneme tähele, et kui juhuslikul suurusel X_i on sugulane, siis on see $G_{K_i}(Z_{K_i})$. Seega juhuslikku suurust X_i saab funktsiooni M eeskirja järgi ühendada vaid hulka A_i kuuluva elemendiga.

2.4.2 Tõkke konstrueerimine

Toome sisse juhuslikud suurused $V_1, V_2, ...$ ja $W_1, W_2, ...$ Formaalselt defineeritakse need juhuslikud suurused vastavalt jaotistes "Juhuslikud suurused $V_1, V_2, ...$" ja "Juhuslikud suurused $W_1, W_2, ...$". Siin esitame vaid nende sisulise tähenduse.

Juhusliku suuruse V_i väärtus on 1, kui juhuslikul suurusel X_i leidub võrdse väärtusega sugulane; vastasel korral on V_i väärtus 0 ($i = 1, 2, ...$). Jaotistes "Juhuslikud suurused $V_1, V_2, ...$" leitakse juhuslike suuruste $V_1, V_2, ..$ jaotus (näidatakse, et nad on sama jaotusega) ning näidatakse, et nad on sõltumatud. Etteruttavalt:

$$P(V_1 = 1) = p \cdot p_z,$$

kus $p_z = P(F_i(Z_i) = G_i(Z_i))$.

Juhusliku suuruse W_i väärtus on 1, kui hulgas A_i leidub alles jääv X_i-ga võrdse väärtusega element ning X_i sugulane ei jää alles; vastasel korral on W_i väärtus 0 ($i = 1, 2, ...$). Jaotises "Juhuslikud suurused $W_1, W_2, ...$" leitakse juhuslike suuruste $W_1, W_2, ...$ jaotus (näidatakse, et nad on sama jaotusega) ning näidatakse, et nad on sõltumatud. Jällegi

etteruttavalt:

$$P(W_1 = 1) = 1 - p - \frac{1-p}{1 + q_z - pq_z},$$

kus $q_z = P(F_i(Z_i) = G_j(Z_j)),\ i \neq j$.

Seega funktsiooniga M leitud jadade X^n ja Y^{n_y} ühisjada pikkus (teisisõnu funktsiooniga M teostatud ühenduste arv) avaldub kujul

$$B_n^x := \sum_{i=1}^{n} (V_i + W_i).$$

Suurte arvude seaduse põhjal

$$\frac{B_n^x}{n} \xrightarrow{\text{p.k.}} EV_1 + EW_1 = p \cdot p_z + 1 - p - \frac{1-p}{1 + q_z - pq_z} =: \alpha$$

protsessis $n \to \infty$.

Analoogiliselt ülaltooduga saab näidata, et

$$\frac{B_n^y}{n} \xrightarrow{\text{p.k.}} \alpha \quad \text{protsessis } n \to \infty,$$

kus B_n^y on jadade Y^n ja X^{n_x} ühisjada pikkus. Siin

$$n_x := \sum_{j=1}^{L_n} D_j^x.$$

Paneme tähele, et kui mingite jadade x, y osajadadel leidub ühisjada z, siis z on ühisjadaks ka jadadele x, y. Vaatleme nüüd jadasid X^n ja Y^n. Kui $n_y \leq n$, siis eelneva põhjal leidub neil ühisjada pikkusega B_n^x.

Kui $n_y > n$, siis $L_n < K_n$ ning

$$n = \sum_{j=1}^{K_n} D_j^x \geq \sum_{j=1}^{L_n} D_j^x = n_x$$

(tegelikult kehtib ka range võrratus, kuid see ei oma antud juhul tähtsust). Seega juhul $n_y > n$ leidub jadade X^n ja Y^n ühisjada pikkusega B_n^y.

Eelneva põhjal leidub jadade X^n, Y^n ühisjada pikkusega $B_n := \min\{B_n^x, B_n^y\}$. Kuna

$$\frac{B_n^x}{n} \xrightarrow{\text{p.k.}} \alpha \quad \text{ja} \quad \frac{B_n^y}{n} \xrightarrow{\text{p.k.}} \alpha \quad \text{protsessis } n \to \infty,$$

siis

$$\frac{B_n}{n} \xrightarrow{\text{p.k.}} \alpha \quad \text{protsessis } n \to \infty.$$

Kokku võttes, suuruse γ alumiseks tõkkeks on

$$\alpha = p \cdot p_z + 1 - p - \frac{1-p}{1+q_z-pq_z},$$

kus $p_z = P(F_i(Z_i) = G_i(Z_i))$ ning $q_z = P(F_i(Z_i) = G_j(Z_j))$ $(i \neq j)$.

Suurused p_z, q_z sõltuvad vaid juhuslike suuruste $Z_1, Z_2, ...$ jaotusest ning maatriksist Q. Kui $Z_1, Z_2, ...$ on jaotusega $Be(0.5)$, siis kasutades maatriksi Q esitust kujul (2.1) saame samasused

$$p_z = \frac{1}{2}((1-\epsilon_1)^2 + \epsilon_1^2) + \frac{1}{2}(\epsilon_2^2 + (1-\epsilon_2)^2),$$
$$q_z = \frac{1}{4}((1-\epsilon_1)^2 + \epsilon_1^2) + \frac{1}{4}(\epsilon_2^2 + (1-\epsilon_2)^2) + \frac{1}{2}((1-\epsilon_1)\epsilon_2 + \epsilon_1(1-\epsilon_2)).$$

Olgu Q sümmeetriline ja juhuslikud suurused $Z_1, Z_2, ...$ jaotusega $Be(0.5)$. Siis $\epsilon_1 = \epsilon_2 = \epsilon$ ning jadad X^n, Y^n on jaotusega $Be(0.5)$. Lisaks $p_z \to 1$ ja $q_z \to 0.5$ protsessis $\epsilon \to 0$. Seega $\alpha \to 1$ protsessis $(p, \epsilon) \to (1, 0)$. Seega oleme tõestanud, et kui jadad X^n, Y^n on jaotusega $Be(0.5)$ ja $p < 1$, siis γ ei ole alati ligikaudselt võrdne väärtusega 0.81.

2.4.3 Jadade X^n, Y^n sarnasusskoor D_n

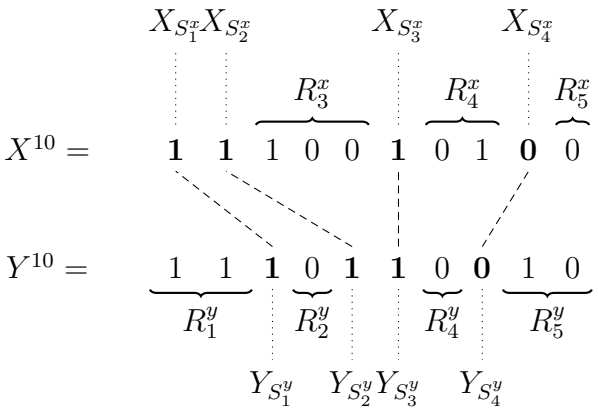
Meenutame, et mõiste "sarnasusskoor" all mõistame ühisjada pikkust. Siiamaani oleme käsitlenud jadade X^n ja Y^n sarnasusskoore L_n ja B_n. Vaatleme nüüd veel ühte jadade X^n, Y^n sarnasusskoori, mida me märgime tähisega D_n.

Nagu sarnasusskoor B_n, on ka D_n selline sarnasusskoor, mille puhul ühendatakse kõik võrdsed sugulased. Erinevalt sarnasusskoorist B_n ühendatakse D_n puhul aga võrdsete sugulaste vahel nii palju elemente kui võimalik. Kuna piirväärtuse $\lim_{n\to\infty} D_n/n$ teoreetiline avaldamine käib töö autoril üle jõu, piirdume vaid suuruse D_n/n empiirilise arvutamisega simulatsioonides.

Defineerime sarnasusskoori D_n formaalselt.

Olgu $S_1^x < ... < S_M^x$ nende jada X^n elementide indeksid, millel on jadas Y^n võrdne sugulane, ning analoogiliselt $S_1^y < ... < S_M^y$ nende jada Y^n elementide indeksid, millel on jadas X^n võrdne sugulane. Olgu

$$S_0^x := 0, \quad S_0^y := 0, \quad S_{M+1}^x := n+1, \quad S_{M+1}^y := n+1.$$



Joonis 2.4: näide sarnasusskoori D_n leidmise kohta

Defineerime järgmised jada X^n alamlõigud:

$$R_i^x := \begin{cases} X_{S_{i-1}^x+1}, ..., X_{S_i^x-1}, & \text{kui } S_i^x - S_{i-1}^x \geq 2; \\ \emptyset, & \text{mujal.} \end{cases}, \quad i = 1, ..., M+1.$$

Analoogiliselt defineerime järgmised jada Y^n alamlõigud:

$$R_i^y := \begin{cases} Y_{S_{i-1}^y+1}, ..., Y_{S_i^y-1}, & \text{kui } S_i^y - S_{i-1}^y \geq 2; \\ \emptyset, & \text{mujal.} \end{cases}, \quad i = 1, ..., M+1.$$

Meenutame, et L on pikima ühisjada tagastav funktsioon. Olgu $L(\emptyset; A) := 0, L(A; \emptyset) := 0$ ja $L(\emptyset; \emptyset) := 0$ mis tahes jada A korral. Defineerime sarnasusskoori D_n järgmiselt:

$$D_n := \sum_{i=1}^{M+1} L(R_i^x; R_i^y) + M, \tag{2.2}$$

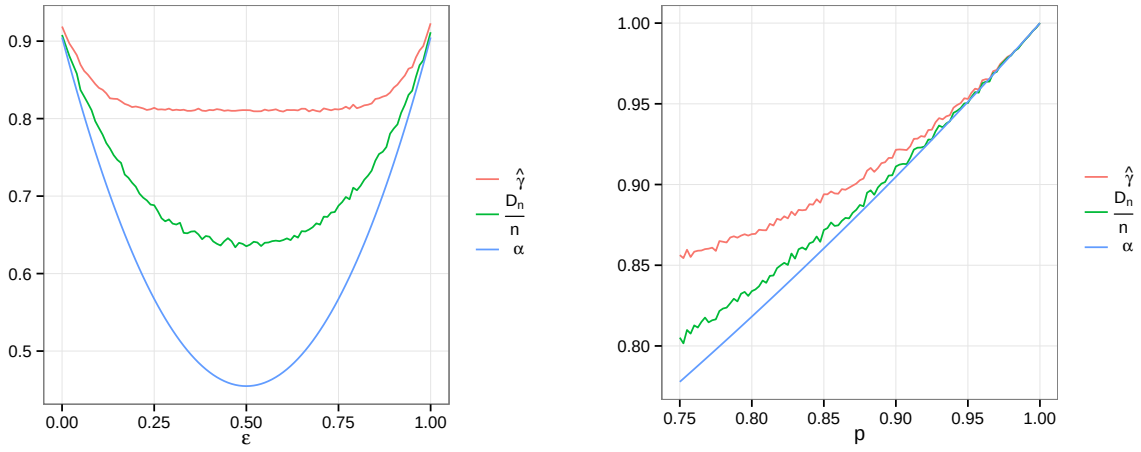
Joonisel 2.4 toodud näites on valemis (2.2) esinevad jadad ära näidatud. Seal $n = 10$, $M = 4$ ning $R_1^x = R_2^x = R_3^y = \emptyset$. Võrdsed sugulased on ühendatud katkendlike joontega.

Pole raske näha, et

$$\frac{B_n}{n} \leq \frac{D_n}{n} \leq \frac{L_n}{n}.$$

Suure n korral

$$\alpha \approx \frac{B_n}{n}.$$



(a) $p = 0.9$ *(b)* $\epsilon = 0$

Joonis 2.5: suuruste $\hat{\gamma}, D_n/n, \alpha$ sõltuvus parameetritest ϵ ja p

2.4.4 Simulatsioonid tõkke kohta

Simulatsioonide tulemused on kujutatud joonisel 2.5. Suuruse γ hinnangu ja D_n/n leidmiseks genereeritakse jada Z^m (jaotusega $Be(0.5)$) – joonise 2.5a puhul $m = 15\,000$ ning joonise 2.5b puhul $m = 25\,000$. Pärast mutatsioonide ja kadumiste teostamist saame jadad X^{N_X} ja Y^{N_Y}. Taas

$$n = \min\{N_X, N_Y\}.$$

Suuruse γ hinnang leitakse järgmiselt:

$$\hat{\gamma} = \frac{L(X^n; Y^n)}{n} = \frac{L_n}{n}.$$

Märgime, et $\hat{\gamma}$ ja D_n/n on leitud alati samade jadade pealt.

Vaatleme joonist 2.5a.

Siin $\epsilon_1 = \epsilon_2 = \epsilon$, st Q on alati sümmeetriline. Mida lähemal on ϵ arvule 0.5, seda väiksem on tõenäosus, et sugulased on võrdsed – see selgitab, miks γ on $\epsilon = 0.5$ korral minimaalne.

Suurus D_n/n kaugeneb γ-st parameetri ϵ lähenemisel arvule 0.5. Seega on tõenäosus, et pikima ühisjada konstrueerimisel võrdsed sugulased ühendatakse, seda väiksem, mida vähem võrdseid sugulasi jadades X^n, Y^n on.

Tõke α kaugeneb suurusest D_n/n parameetri ϵ lähenedes arvule 0.5. Selgitame, miks see nii on. Võrdsete sugulaste arvukus jadades X^n, Y^n väheneb, kui ϵ läheneb arvule 0.5.

Sarnasusskoori D_n puhul ühendatakse võrdsete sugulaste vahel nii palju elemente, kui võimalik – sarnasusskoori B_n puhul aga ühendatakse võrdsete sugulaste vahel üldjuhul vähem elemente, kui võimalik (meenutame, et $\lim_{n\to\infty} B_n/n = \alpha$).

Vaatleme joonist 2.5b.

Siin $\epsilon = 0$, st mutatsioone pole.

Interpreteerime γ vähenemist suuruse p kahanedes. See on tingitud sellest, et mida väiksem on p, seda suurem on tõenäosus, et jada X^n (või jada Y^n) elemendi sugulane on ära kadunud.

2.4.5 Juhuslikud suurused $T_1, T_2, ...$

Defineerime $K_0 := 0$ ning

$$T_i := K_i - K_{i-1}, \quad i = 1, 2, ...$$

$T_1, T_2, ...$ on sama jaotusega:

$$\begin{aligned} P(T_i = t) &= \sum_{k=i-1}^{\infty} P(T_i = t \mid K_{i-1} = k) \cdot P(K_{i-1} = k) \\ &= \sum_{k=i-1}^{\infty} P(D^x_{k+1} = ... = D^x_{k+t-1} = 0,\ D^x_{k+t} = 1) \cdot P(K_{i-1} = k) \\ &= p(1-p)^{t-1}, \quad t = 1, 2, ... \end{aligned}$$

Olgu $i \geq 2$ ja $(t_1, ..., t_{i-1})$ vektori $(T_1, ..., T_{i-1})$ suvaline väärtus. Olgu $s := t_1 + ... + t_{i-1}$. Avaldame:

$$\begin{aligned} P(T_i = t \mid (T_1, ..., T_{i-1}) = (t_1, ..., t_{i-1})) &= P(D^x_{s+1} = ... = D^x_{s+t-1} = 0,\ D^x_{s+t} = 1) \\ &= p(1-p)^{t-1} = P(T_i = t). \end{aligned}$$

Seega juhuslik suurus T_i ei sõltu iga $i \geq 2$ korral juhuslikest suurustest $T_1, ..., T_{i-1}$. Tõenäosusteooria järgi tähendab see, et juhuslikud suurused $T_1, T_2, ...$ on sõltumatud.

2.4.6 *Juhuslikud suurused $V_1, V_2, ...$*

Olgu

$$V_i := \begin{cases} 1, & \text{kui } D^y_{K_i} = 1 \text{ ja } F_{K_i}(Z_{K_i}) = G_{K_i}(Z_{K_i}) \\ 0, & \text{mujal} \end{cases}, \quad i = 1, 2, ...$$

Ehk $V_i = 1$ parajasti siis, kui juhuslikul suurusel X_i leidub temaga võrdne alles jäänud sugulane.

Paneme tähele, et

$$K_i, D^y_1, D^y_2, ..., (F_1(Z_1), G_1(Z_1)), (F_2(Z_2), G_2(Z_2)), ... \text{ on sõltumatud iga } i \text{ korral.} \quad (2.3)$$

Olgu $p_z := P(F_i(Z_i) = G_i(Z_i))$. $V_1, V_2, ...$ on sama jaotusega:

$$P(V_i = 1) = \sum_{k=i}^{\infty} P(D^y_{K_i} = 1,\ F_{K_i}(Z_{K_i}) = G_{K_i}(Z_{K_i}) \mid K_i = k) \cdot P(K_i = k) \overset{(2.3)}{=} p \cdot p_z.$$

Meenutame, et juhusliku suuruse (või juhusliku vektori) võimalike väärtuste hulka tähistame sama sümboliga kui juhuslikku suurust (juhuslikku vektorit) ennast.

Lause 2.1. *Olgu antud diskreetsed juhuslikud vektorid X, Y, Z. Kui X, Y on mis tahes fikseeritud Z väärtuse korral tinglikult sõltumatud ning vektorid Y ja Z on sõltumatud, siis vektorid X, Y on sõltumatud.*

Tõestus. Avaldame:

$$\begin{aligned} P(X = x,\ Y = y) &= \sum_{z \in Z} P(X = x,\ Y = y \mid Z = z) \cdot P(Z = z) \\ &= \sum_{z \in Z} P(X = x \mid Z = z) \cdot P(Y = y \mid Z = z) \cdot P(Z = z) \\ &= P(Y = y) \sum_{z \in Z} P(X = x \mid Z = z) \cdot P(Z = z) \\ &= P(X = x) \cdot P(Y = y). \end{aligned}$$

■

Lause 2.2. *Olgu X, Y diskreetsed juhuslikud suurused, kusjuures X on binaarne ehk X-il on kaks võimalikku väärtust. Olgu $x_1 \in X$. Eeldame, et iga $y \in Y$ korral kehtib samasus*

$$P(X = x_1 \mid Y = y) = P(X = x_1).$$

Siis X, Y on sõltumatud.

Tõestus. Kehtigu väites esitatud eeldused. Olgu $x_2 \in X \backslash \{x_1\}$. Siis

$$P(X = x_2 \mid Y = y) = 1 - P(X = x_1 \mid Y = y) = 1 - P(X = x_1) = P(X = x_2).$$

Seega X, Y on sõltumatud. ■

Olgu $i \geq 2$. Fikseeritud $K_i = k$ korral juhuslikud suurused $(V_1, ..., V_{i-1})$ sõltuvad ainult juhuslikust vektorist

$$D_1 := (D_1^x, ..., D_{k-1}^x, D_1^y, ..., D_{k-1}^y, F_1(Z_1), ..., F_{k-1}(Z_{k-1}), G_1(Z_1), ..., G_{k-1}(Z_{k-1}))$$

ning juhuslik suurus V_i sõltub vaid juhuslikust vektorist

$$D_2 := (D_k^y, F_k(Z_k), G_k(Z_k)).$$

Kuna vektorid D_1, D_2 on sõltumatud, siis fikseeritud K_i korral on $(V_1, ..., V_{i-1}), V_i$ tinglikult sõltumatud. Lisaks

$$P(V_i = 1 \mid K_i = k) = P(D_{K_i}^y = 1,\ F_{K_i}(Z_{K_i}) = G_{K_i}(Z_{K_i}) \mid K_i = k) \overset{(2.3)}{=} p \cdot p_z = P(V_i = 1).$$

Seega lause 2.2 põhjal juhuslikud suurused V_i, K_i on sõltumatud. Lause 2.1 põhjal $(V_1, ..., V_{i-1}), V_i$ on sõltumatud. Seega juhuslikud suurused $V_1, V_2, ...$ on sõltumatud.

2.4.7 Juhuslikud suurused $U_1, U_2, ...$

Olgu nüüd

$$U_i := \begin{cases} 1, & \text{kui leidub } j \in \{K_{i-1} + 1, ..., K_i - 1\} \text{ nii, et } F_{K_i}(Z_{K_i}) = G_j(Z_j) \text{ ja } D_j^y = 1 \\ 0, & \text{mujal} \end{cases},$$

$i = 1, 2, ...$ Ehk $U_i = 1$ parajasti siis, kui juhuslikul suurusel X_i leidub hulgas A_i temaga võrdne alles jääv mittesugulasest element.

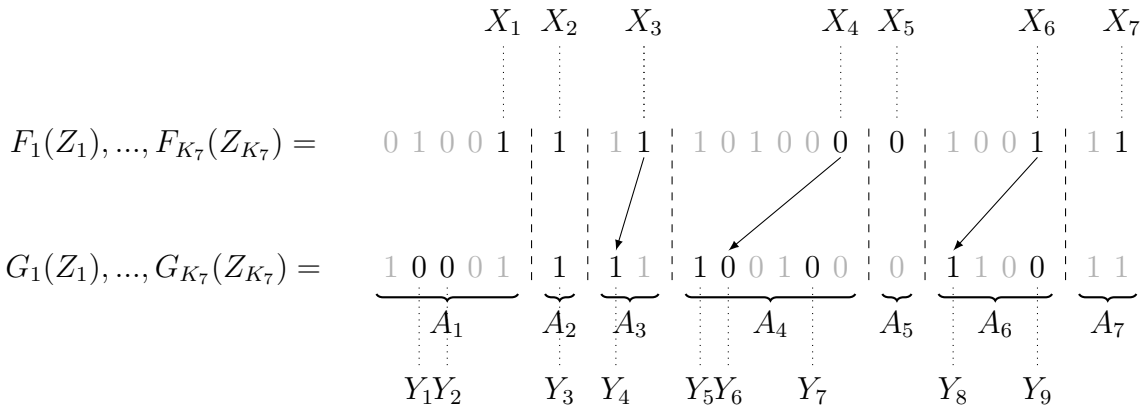
Ülaltoodut illustreerib joonises 2.6 toodud näide. Halli värviga on taas märgitud kaduvad elemendid. Nooltega märgitud ühendused näitavad, et $U_3 = U_4 = U_6 = 1$. Samas $U_1 = U_2 = U_5 = U_7 = 0$.

Leiame juhusliku suuruse U_i jaotuse. Meenutame, et $T_j = K_j - K_{j-1}$ $(j = 1, 2, ...)$. Kuna

$$P(T_j = t \mid K_{j-1} = k) = P(D_{k+1}^x = ... = D_{k+t-1}^x = 0,\ D_{k+t}^x = 1) = (1-p)^{t-1}p = P(T_j = t),$$

siis

$$T_j, K_{j-1} \text{ on sõltumatud iga } j \geq 2 \text{ korral.} \tag{2.4}$$



Joonis 2.6: juhuslike suuruste $U_1, ..., U_7$ väärtustamine

Avaldame:

$$P(U_i = 0) = \sum_{t=1}^{\infty} P_1(t) \cdot P(T_i = t),$$

kus

$$P_1(t) = P(U_i = 0 \mid T_i = t) = \begin{cases} \sum_{k=i-1}^{\infty} P_2(t,k) \cdot P(K_{i-1} = k \mid T_i = t), & \text{kui } i > 1 \\ P_2(t,0), & \text{kui } i = 1 \end{cases}$$
$$\overset{(2.4)}{=} \begin{cases} \sum_{k=i-1}^{\infty} P_2(t,k) \cdot P(K_{i-1} = k), & \text{kui } i > 1 \\ P_2(t,0), & \text{kui } i = 1 \end{cases}.$$

Siin

$$P_2(t,k) = P(U_i = 0 \mid K_{i-1} = k,\ T_i = t).$$

Olgu $q_z := P(F_i(Z_i) = G_j(Z_j))$ $(i \neq j)$. Kuna juhuslikud suurused

$$G_1(Z_1), ..., G_{n-1}(Z_{n-1}), F_n(Z_n), D_1^y, D_2^y, ...$$

on sõltumatud iga n korral ning ka sündmused

$$\{G_1(Z_1) \neq F_n(Z_n)\}, ..., \{G_{n-1}(Z_{n-1}) \neq F_n(Z_n)\}$$

on sõltumatud iga n korral, siis

$$\begin{aligned} P_2(t,k) &= P(\{\{D_{k+1}^y = 1,\ G_{k+1}(Z_{k+1}) \neq F_{k+t}(Z_{k+t})\} \cup \{D_{k+1}^y = 0\}\}, ... \\ &\quad \{\{D_{k+t-1}^y = 1,\ G_{k+t-1}(Z_{k+t-1}) \neq F_{k+t}(Z_{k+t})\} \cup \{D_{k+t-1}^y = 0\}\}) \\ &= (p(1-q_z) + 1 - p)^{t-1} = (1 - pq_z)^{t-1} =: P_2(t). \end{aligned} \tag{2.5}$$

Seega $P_1(t) = P_2(t)$ ning

$$P(U_i = 0) = \sum_{t=1}^{\infty} P_2(t) \cdot P(T_i = t) \tag{2.6}$$
$$= p \sum_{t=1}^{\infty} ((1 - pq_z)(1 - p))^{t-1}.$$

Geomeetrilise rea summa valemi järgi

$$P(U_i = 0) = \frac{p}{1 - (1 - p - pq_z + p^2 q_z)} = \frac{1}{1 + q_z - pq_z}.$$

Seega $U_1, U_2, ...$ on sama jaotusega.

Olgu $i \geq 2$. Fikseeritud $K_{i-1} = k$ korral juhuslikud suurused $U_1, ...U_{i-1}$ sõltuvad ainult juhuslikust vektorist

$$D_1 := (D_1^x, ..., D_{k-1}^x, D_1^y, ..., D_{k-1}^y, F_1(Z_1), ..., F_k(Z_k), G_1(Z_1), ..., G_{k-1}(Z_{k-1}))$$

ning U_i sõltub vaid juhuslikust vektorist

$$D_2 := (D_{k+1}^x, D_{k+2}^x, ..., D_{k+1}^y, D_{k+2}^y, ..., F_{k+1}(Z_{k+1}), F_{k+2}(Z_{k+2}), ...,$$
$$G_{k+1}(Z_{k+1}), G_{k+2}(Z_{k+2}), ...).$$

Kuna vektorid D_1, D_2 on sõltumatud, siis fikseeritud K_{i-1} korral on $(U_1, ..., U_{i-1}), U_i$ tinglikult sõltumatud. Lisaks

$$P(U_i = 0 \mid K_{i-1} = k) = \sum_{t=1}^{\infty} P_2(t, k) \cdot P(T_i = t \mid K_{i-1} = k) \overset{(2.5),(2.4)}{=} \sum_{t=1}^{\infty} P_2(t) \cdot P(T_i = t)$$
$$\overset{(2.6)}{=} P(U_i = 0).$$

Seega lause 2.2 põhjal

$$U_j, K_{j-1} \text{ on sõltumatud iga } j \geq 2 \text{ korral.} \tag{2.7}$$

Lause 2.1 põhjal $(U_1, ..., U_{i-1}), U_i$ on sõltumatud. Seega juhuslikud suurused $U_1, U_2, ...$ on sõltumatud.

2.4.8 Juhuslikud suurused $W_1, W_2, ...$

Olgu iga i korral

$$W_i := U_i(1 - D_{K_i}^y).$$

Seega $W_i = 1$ parajasti siis, kui $U_i = 1$ ja $D^y_{K_i} = 0$.

$$P(W_i = 1) = \sum_{k=i}^{\infty} P(U_i = 1,\ D^y_{K_i} = 0 \mid K_i = k) \cdot P(K_i = k).$$

Fikseeritud $K_i = k$ korral U_i, D^y_k on tinglikult sõltumatud. Lisaks D^y_k, K_i on iga k korral sõltumatud. Seega

$$\begin{aligned} P(W_i = 1) &= \sum_{k=i}^{\infty} P(U_i = 1 \mid K_i = k) \cdot P(D^y_k = 0 \mid K_i = k) \cdot P(K_i = k) \\ &= (1-p) \sum_{k=i}^{\infty} P(U_i = 1 \mid K_i = k) \cdot P(K_i = k) = (1-p) \cdot P(U_1 = 1) \\ &= (1-p)\left(1 - \frac{1}{1 + q_z - pq_z}\right) = 1 - p - \frac{1-p}{1+q_z-pq_z}. \end{aligned}$$

Seega $W_1, W_2, ...$ on sama jaotusega.

Olgu $i \geq 2$. Fikseeritud $K_{i-1} = k$ korral juhuslikud suurused $W_1, ..., W_{i-1}$ sõltuvad ainult juhuslikust vektorist

$$D_1 := (D^x_1, ..., D^x_{k-1}, D^y_1, ..., D^y_k, F_1(Z_1), ..., F_k(Z_k), G_1(Z_1), ..., G_{k-1}(Z_{k-1}))$$

ning W_i sõltub ainult juhuslikust vektorist

$$D_2 := (D^x_{k+1}, D^x_{k+2}, ..., D^y_{k+1}, D^y_{k+2}, ..., F_{k+1}(Z_{k+1}), F_{k+2}(Z_{k+2})..., G_{k+1}(Z_{k+1}), G_{k+2}(Z_{k+2}), ...).$$

Kuna D_1, D_2 on sõltumatud, siis $(W_1, ..., W_{i-1}), W_i$ on fikseeritud K_{i-1} korral tinglikult sõltumatud. Paneme tähele, et

$$\begin{aligned} &P(D^y_{K_i} = 1 \mid K_{i-1} = k) \\ &= \sum_{k^*=k+1}^{\infty} P(D^y_{k^*} = 1 \mid K_i = k^*,\ K_{i-1} = k) \cdot P(K_i = k^* \mid K_{i-1} = k) = p. \end{aligned}$$

Samas

$$P(D^y_{K_i} = 1) = \sum_{k^*=i}^{\infty} P(D^y_{k^*} = 1 \mid K_i = k^*) \cdot P(K_i = k^*) = p.$$

Seega

$$P(D^y_{K_i} = 1 \mid K_{i-1} = k) = P(D^y_{K_i} = 1)$$

ning lause 2.2 põhjal $D^y_{K_i}, K_{i-1}$. on sõltumatud. Lisaks (2.7) kohaselt U_i, K_{i-1} on sõltumatud. Olgu

$$g(U, D) := U(1 - D).$$

Kuna $W_i = g(U_i, D^y_{K_i})$, siis eelneva põhjal W_i, K_{i-1} on sõltumatud. Lause 2.1 põhjal $(W_1, ..., W_{i-1}), W_i$ on sõltumatud. Seega juhuslikud suurused $W_1, W_2, ...$ on sõltumatud.

2.5 Suuruse γ sõltuvus maatriksist Q

2.5.1 Sissejuhatus

Meenutame, et $\hat{\gamma} = L(X^n, Y^n)/n = L_n/n$, kus L on pikima ühisjada tagastav funktsioon. Seega $\hat{\gamma}$ on jadade X^n, Y^n normeeritud sarnasusmõõt: mida sarnasemad on jadad X^n, Y^n, seda suurem on $\hat{\gamma}$. Kui X^n, Y^n on identsed, siis $\hat{\gamma} = 1$. Praktikas ei huvita meid niivõrd jadade sarnasus, vaid nende sõltuvus – jadade sarnasuse mõõtmine on tihtipeale vaid vahend nende sõltuvuse hindamiseks.

Eelnevalt nägime, et vähemasti siis, kui jada $Z_1, Z_2, ...$ jaotuseks on $Be(0.5)$ ja Q on sümmeetriline, funktsioon $\gamma = \gamma(p, Q)$ tõepoolest sõltub parameetritest p ja Q ning ei ole konstant. Mudeli seisukohast on selline sõltuvus tegelikult vajalik, sest kui γ ei sõltuks jadade $X_1, X_2, ...$ ja $Y_1, Y_2, ...$ fikseeritud jaotuse korral parameetritest p, Q, siis ei sõltuks γ ka jadade X^n, Y^n vahelise sõltuvuse suurusest ning mudelil ei oleks praktikas erilist väärtust.

Tõenäosus, et juhusliku suuruse sugulane jääb alles on p. Seega p vähenedes väheneb ka jadade $X_1, X_2, ...$ ja $Y_1, Y_2, ...$ vaheline sõltuvus. Seega võiks arvata, et p vähenedes ka γ väheneb – seda kinnitavad ka joonisel 2.5b toodud simulatsioonid.

Jadade $X_1, X_2, ...$ ja $Y_1, Y_2, ...$ vahelise sõltuvuse suurus sõltub ka maatriksist Q. Käesolevas jaotises tutvustame informatsiooniteooriast pärinevat mõistet "vastastikune informatsioon". Vastastikune informatsioon võimaldab meil leida maatriksi Q funktsiooni, mis mõõdab jadade $X_1, X_2, ...$ ja $Y_1, Y_2, ...$ vahelist sõltuvust fikseeritud p korral. Seejärel võrdleme simulatsioonide abil selle funktsiooni käitumist suuruse γ käitumisega.

Kasutame lühendatud kirjaviisi:

$$\{X = x\} = \{x\},$$
$$\{Y = y\} = \{y\},$$
$$\{Z = z\} = \{z\}.$$

Olgu $\log := \log_2$. Järgnev teave vastastikuse informatsiooni kohta on saadud kirjandusest [1].

Definitsioon 2.3. *Diskreetsete juhuslike suuruste X, Y vastastikuseks informatsiooniks*

nimetatakse suurust

$$I(X;Y) := \sum_{x\in X,\ y\in Y} P(x,y)\cdot \log \frac{P(x,y)}{P(x)P(y)}.$$

Vastastikune informatsioon on mittenegatiivne. Lisaks on ta sümmeetriline, st $I(X;Y) = I(Y;X)$. Vastastikune informatsioon mõõdab kahe juhusliku suuruse sõltuvust, kusjuures $I(X;Y) = 0$ parajasti siis, kui X, Y on sõltumatud.

Vastastikuse informatsiooni saab avaldada informatsiooniteooria baassuuruste "entroopia" ja "tinglik entroopia" kaudu. Entroopia $H(X)$ mõõdab juhusliku suuruse X juhuslikkust. Entroopia on mittenegatiivne, kusjuures $H(X) = 0$ parajasti siis, kui X on konstant. Tinglik entroopia $H(X|Y)$ mõõdab juhusliku suuruse X keskmist juhuslikkust tingimusel, et Y väärtus on teada. Kehtib seos

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X).$$

Seega vastastikune informatsioon näitab, kui palju väheneb keskmiselt ühe juhusliku suuruse väärtuse teadasaamisel teise juhusliku suuruse juhuslikkus – ehk kui palju annab ühe juhusliku suuruse väärtuse teadasaamine keskmiselt teise juhusliku suuruse kohta informatsiooni. Juhul, kui $X = Y$, siis $H(X|Y) = H(Y|X) = 0$ ning $I(X;Y) = H(X) = H(Y)$.

Olgu juhusliku suuruse $Z := Z_i$ järglasteks $X := X_{C_X}$ ja $Y := Y_{C_Y}$, st C_X, C_Y on sellised, et $K_{C_X} = L_{C_Y} = i$. Lisaks lepime kokku, et $0 \log 0 = 0$. Siis

$$I(X;Z) = \sum_{x,z\in\mathcal{A}} P(z)Q(z,x)\cdot \log \frac{Q(z,x)}{\sum_{z^*\in\mathcal{A}} P(Z=z^*)Q(z^*,x)}$$

ning

$$I(X;Y) = \sum_{x,y\in\mathcal{A}} \left(\sum_{z\in\mathcal{A}} P(z)Q(z,x)Q(z,y) \right) \log \frac{\sum_{z\in\mathcal{A}} P(z)Q(z,x)Q(z,y)}{\left(\sum_{z\in\mathcal{A}} P(z)Q(z,x)\right)\left(\sum_{z\in\mathcal{A}} P(z)Q(z,y)\right)}.$$

Fikseeritud p korral $I(X;Z)$ mõõdab jadade $X_1, X_2, \ldots$ ja $Z_1, Z_2, \ldots$ vahelist sõltuvust ning $I(X;Y)$ mõõdab jadade $X_1, X_2, \ldots$ ja $Y_1, Y_2, \ldots$ vahelist sõltuvust.

Tõestame mõne tulemuse.

Lause 2.3. *(i) Juhuslikud suurused X, Z on sõltumatud parajasti siis, kui maatriksi Q read on võrdsed.*

(ii) Juhuslikud suurused Y, Z on sõltumatud parajasti siis, kui maatriksi Q read on võrdsed.

Tõestus. (i) Piisavus. Paneme tähele, et

$$P(z,x) = P(z)Q(z,x)$$

ja

$$P(x) = \sum_{z^* \in Z} P(z^*)Q(z^*,x).$$

Seega Z, X on sõltumatud parajasti siis, kui iga $z \in Z$ ja $x \in X$ korral

$$Q(z,x) = \sum_{z^* \in Z} P(z^*)Q(z^*,x). \tag{2.8}$$

Kui $Q(z,x)$ ei sõltu z väärtusest (ehk kui maatriksi Q read on võrdsed), saame selle ülaltoodud summas ette tuua ja samasus (2.8) tõepoolest kehtib.

Tarvilikkus. Kuna võrduse (2.8) parem pool on fikseeritud x korral alati sama, siis ka $Q(z,x)$ ei tohi võrduse kehtimiseks z väärtusest sõltuda.

(ii) Tõestus on analoogiline osaga (i). ■

Lause 2.4. *Jada X, Z, Y on Markovi ahel.*

Tõestus. Avaldame:

$$\begin{aligned} P(y \mid z,x) &= \frac{P(f(z,\eta_i) = y,\ Z = z,\ f(z,\xi_i) = x)}{P(Z = z,\ f(z,\xi_i) = x)} = \frac{P(f(z,\eta_i) = y,\ Z = z)}{P(Z = z)} \\ &= P(y \mid z). \end{aligned}$$

■

Järeldus 2.1. *Juhuslikud suurused X, Y on sõltumatud parajasti siis, kui maatriksi Q read on võrdsed.*

Tõestus. Kuna lause 2.4 kohaselt jada X, Z, Y on Markovi ahel, siis X, Y on sõltumatud parajasti siis, kui Y, Z on sõltumatud. Lause 2.3 kohaselt Y, Z on sõltumatud parajasti siis, kui maatriksi Q read on võrdsed. ■

2.5.2 Simulatsioonid

Simuatsioonides võtame lihtsuse huvides $p = 1$ (st kadumisi pole). Meenutame, et X, Y on sugulased eellasega Z, et $\hat{\gamma} = L_n/n$ ja et jada Z^n elemendid genereeritakse jaotusega $Be(0.5)$. Suuruseks n võeti 10 000.

Joonisel 2.7 on toodud simulatsioonid sümmeetriliste Q-de korral. Sümmeetrilise Q korral

$$P(X=0) = P(Y=0) = \frac{1}{2}(1-\epsilon) + \frac{1}{2}\epsilon = \frac{1}{2}.$$

Jooniselt 2.7b näeme, kuidas nii $I(X;Z)$ kui ka $I(X;Y)$ on maksimaalsed, kui maatriksi Q mõlemas veerus on üks 0 ja üks 1, ning kuidas $I(X;Z)$ ja $I(X;Y)$ saavutavad miinimumi, kui maatriksi Q read on võrdsed ehk kui juhuslikud suurused X, Y, Z on sõltumatud.

Informatsiooniteooriast on teada järgmine tulemus.

Lause 2.5. (Andmetöötlusvõrratus) [1] *Kui jada X_1, X_2, X_3 on Markovi ahel, siis*

$$I(X_1; X_2) \geq I(X_1; X_3),$$

kusjuures võrdus kehtib parajasti siis, kui jada X_1, X_3, X_2 on Markovi ahel.

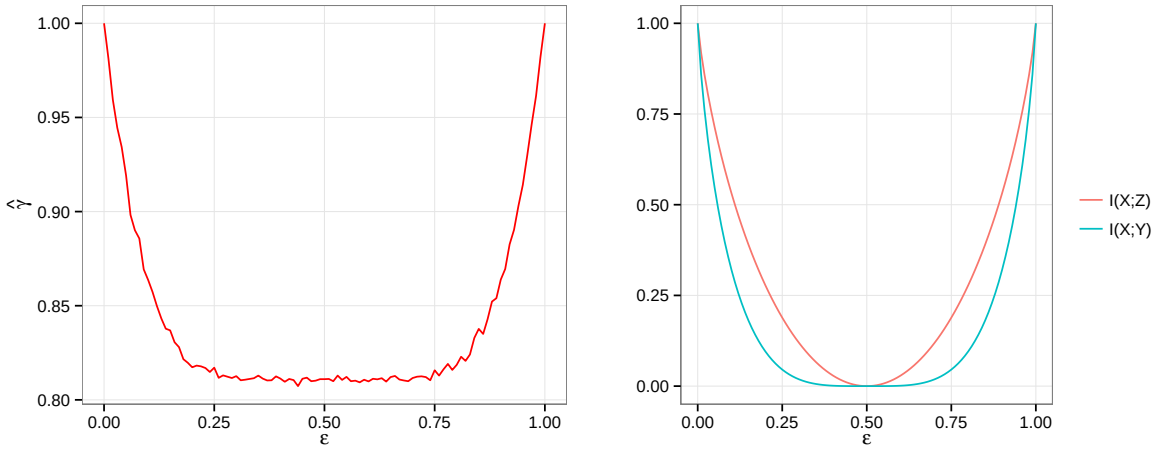
Lause 2.4 kohaselt jada X, Z, Y on Markovi ahel. Seega $I(X;Z) \geq I(X;Y)$ – seda kinnitab ka joonis. Võrdus $I(X;Z) = I(X;Y)$ kehtib andmetöötlusvõrratuse kohaselt parajasti siis, kui X, Y, Z on Markovi ahel – joonise järgi tundub see võrdus kehtivat kolmel juhul: kui $\epsilon = 0.5$, $\epsilon = 0$ ja $\epsilon = 1$. Esimesel juhul on X, Z, Y sõltumatud, seega jada X, Y, Z on tõepoolest Markovi ahel. Kui $\epsilon = 0$ või $\epsilon = 1$, on fikseeritud Y väärtuse korral suurused X, Z konstandid, seega ka nendel juhtudel jada X, Y, Z on Markovi ahel.

Jooniselt 2.7 on näha, kuidas suurus γ käitub samamoodi kui $I(X;Z)$ ja $I(X;Y)$: kui ϵ läheneb arvule 0.5, siis γ väheneb; kui ϵ eemaldub arvust 0.5, siis γ kasvab. Lisaks käituvad suurused γ, $I(X;Z)$, $I(X;Y)$ punkti $\epsilon = 0.5$ suhtes sümmeetriliselt. Seega nii γ ja $I(X;Z)$ kui ka γ ja $I(X;Y)$ vahel on üksühene seos. See üksühene seos eksisteerib tänu sellele, et jadade $X_1, X_2, ...$ ja $Y_1, Y_2, ...$ jaotus on fikseeritud (sümmeetrilise Q korral juhuslikud suurused $X_1, X_2, ..., Y_1, Y_2, ...$ on mis tahes ϵ korral jaotusega $Be(0.5)$).

Joonistel 2.8 ja 2.9 on näidatud γ ja $I(X;Y)$ käitumine (üldiselt) ebasümmeetriliste Q-de korral. Sellisel juhul jadade $X_1, X_2, ...$ ja $Y_1, Y_2, ...$ jaotus pole fikseeritud. Järgnevalt interpreteerime jooniseid 2.8 ja 2.9 erinevatel juhtudel.

a) Kõik maatriksi Q elemendid on lähedal arvule 0.5

ehk $\epsilon_1 \approx 0.5$ ja $\epsilon_2 \approx 0.5$. Suurused $I(X;Y)$ ja γ on mõlemad väikesed, sest juhuslikud suurused X, Y on lähedal sõltumatusele (sest maatriksi Q read on lähedal juhule, kus nad võrdsed) ja jadades $X_1, X_2, ...$ ja $Y_1, Y_2, ...$ on nullide ja ühtede proportsioon ligikaudu võrdne arvuga 0.5.



(a) $p = 1$ *(b)*

Joonis 2.7: suuruse γ ning $I(X;Z), I(X;Y)$ käitumine $\epsilon_1 = \epsilon_2$ korral

b) Maatriks Q on lähedal juhule, kus mõlemas veerus on üks 1 ja üks 0

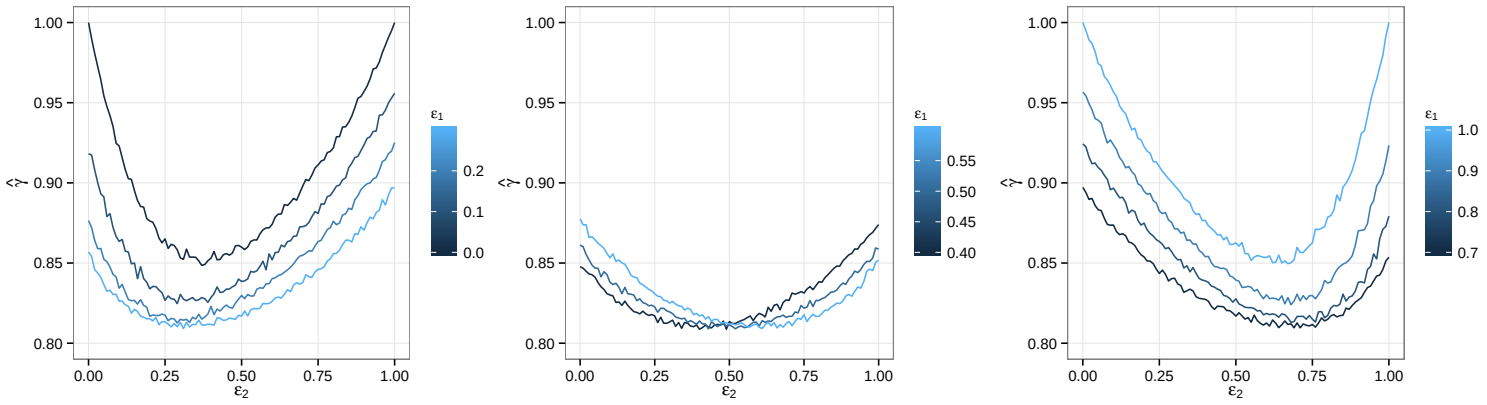
ehk $(\epsilon_1, \epsilon_2) \approx (0,0)$ või $(\epsilon_1, \epsilon_2) \approx (1,1)$. Suurused $I(X;Y)$ ja γ on juhuslike suuruste X, Y vahelise tugeva sõltuvuse tõttu mõlemad suured.

c) Maatriks Q on lähedal juhule, kus ühes veerus on nullid ja teises ühed

ehk $(\epsilon_1, \epsilon_2) \approx (0,1)$ või $(\epsilon_1, \epsilon_2) \approx (1,0)$. Suurus $I(X;Y)$ on väike, sest juhuslikud suurused X, Y on lähedal juhule, kus nad on konstandid. Suurus γ on suur, sest nullide või ühtede osakaal jadades $X_1, X_2, ...$ ja $Y_1, Y_2, ...$ on suur.

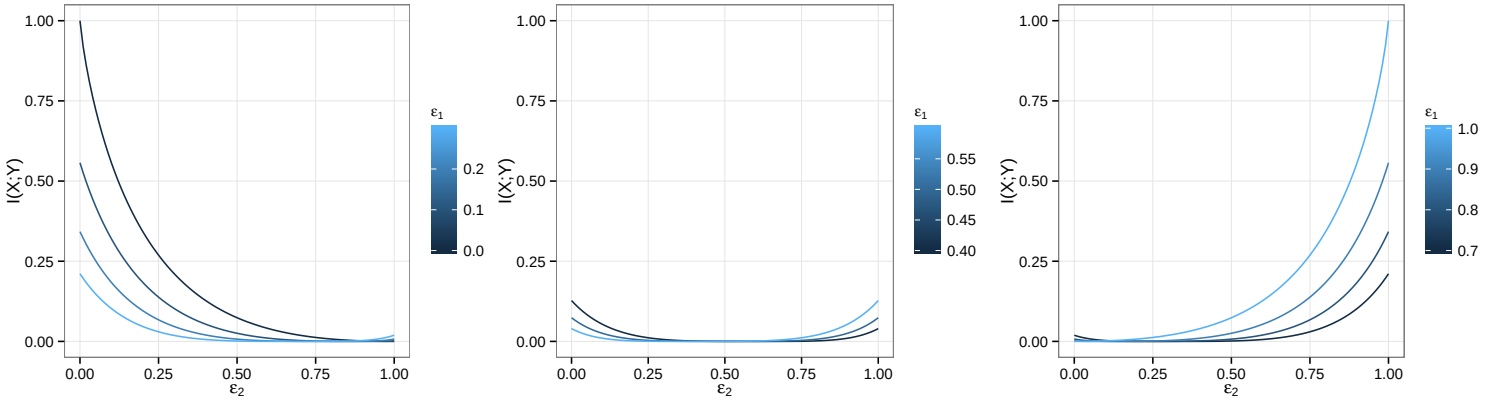
Nagu nägime punktis c), võib ebasümmeetriliste maatriksite korral olla γ suur, kuid $I(X;Y)$ väike.

Kokku võttes võib käesolevas jaotises toodud simulatsioonide põhjal öelda, et kui p ning jadade $X_1, X_2, ...$ ja $Y_1, Y_2, ...$ jaotus on fikseeritud, siis mida suurem on jadade $X_1, X_2, ...$ ja $Y_1, Y_2, ...$ vaheline sõltuvus, seda suurem on γ. Võttes arvesse joonisel 2.5b esitatud simulatsioone p ja γ vahelise seose kohta, võime saadud tulemuse üldistada juhule, kus p ei ole fikseeritud; teisisõnu, kui jadade $X_1, X_2, ...$ ja $Y_1, Y_2, ...$ jaotus on fikseeritud, siis mida suurem on jadade $X_1, X_2, ...$ ja $Y_1, Y_2, ...$ vaheline sõltuvus, seda suurem on γ.



(a) $\epsilon_1 = 0, 0.1, 0.2, 0.3$ *(b)* $\epsilon_1 = 0.4, 0.5, 0.6$ *(c)* $\epsilon_1 = 0.7, 0.8, 0.9, 1$

Joonis 2.8: suuruse γ käitumine üldjuhul ($p = 1$)



(a) $\epsilon_1 = 0, 0.1, 0.2, 0.3$ *(b)* $\epsilon_1 = 0.4, 0.5, 0.6$ *(c)* $\epsilon_1 = 0.7, 0.8, 0.9, 1$

Joonis 2.9: $I(X;Y)$ käitumine üldjuhul

A MODEL FOR RELATED SEQUENCES

Bachelor thesis

Joonas Sova

Summary

At the beginning of the thesis a model for generating two random sequences X, Y was constructed. Both X and Y are obtained by mutating (i.e altering) and deleting some elements in an independent and identically distributed (iid) sequence Z. The probability that any given element in Z is mutated is determined by transition matrix Q. The probability that any given element in Z is not deleted is p. It was shown that both X and Y are iid sequences, and that X and Y are generally related. Some other theoretical results were shown.

Further on, the notion of longest common subsequence was introduced. A simple similarity measure for X and Y was constructed using that notion. As the length of X and Y approaches to infinity, this similarity measure converges to a constant $\gamma = \gamma(Q, p)$. A lower bound for γ was constructed. Using this lower bound, it was shown that function $\gamma(Q, p)$ is not constant when the distribution of X and Y is fixed.

One of the purposes of the thesis is to see how γ depends on the relatedness of X and Y. The relatedness of X and Y depends on both transition matrix Q and deletion parameter p. The relation between p and γ was studied with simulations. Using the notion of mutual information from the information theory, a function of matrix Q was constructed, that measures the relatedness of X and Y when p is fixed. That function was compared to estimate of γ in simulations.

The final conclusion of the thesis is that, when the distribution of X and Y is fixed, the greater the relatedness of X and Y, the greater the γ.

KIRJANDUS

[1] J. Lember. Konspekt aines "Informatsiooniteooria". http://www.math.ut.ee/orb.aw/class=file/action=preview/id=1028872/Informatsiooniteooria.pdf, 2011.

[2] J. Lember, H. Matzinger, A. Vollmer. Path properties of LCS-optimal alignments. http://www.math.uni-bielefeld.de/sfb701/files/preprints/sfb07077.pdf, 2007.

[3] R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

[4] H. Pages, P. Aboyoun, R. Gentleman and S. DebRoy. Biostrings: String objects representing biological sequences, and matching algorithms. R package version 2.26.2.

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 "Sõltuvate jadade mudel", 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, **06.05.2013**.

```
#Funktsioon mutatsioonide teostamiseks Bernoulli jaotusega jadas.
#Maatriks Q on kujul (2.1).
mutatsioon=function(Z, Q){
    #Moodustatakse tõevektor ühtede muteerimise jaoks
    ühtedeMuteerimine=rep(0, length(Z))

    ühtedeMuteerimine[Z==1]=rbinom(length(Z[Z==1]), 1, Q[2,1])

    ühtedeMuteerimine=ühtedeMuteerimine==1

    #Moodustatakse tõevektor nullide muteerimise jaoks
    nullideMuteerimine=rep(0, length(Z))

    nullideMuteerimine[Z==0]=rbinom(length(Z[Z==0]), 1, Q[1,2])

    nullideMuteerimine=nullideMuteerimine==1

    #Ühtede muutmine nullideks
    Z[ühtedeMuteerimine]=0

    #Nullide muutmine ühtedeks
    Z[nullideMuteerimine]=1

    return(Z)
}

#Funktsioon kadumiste teostamiseks Bernoulli jaotusega jadas.
kadumised=function(FZ, p){          #p - allesjäämise tõenäosus
    #Moodustatakse tõevektor kadumiste jaoks
    allesjäämised=rbinom(length(FZ), 1, p)
    tõevektor=allesjäämised==1

    return(FZ[tõevektor])
}

#Funktsioon vastastikuse informatsioonid I(X;Z) arvutamiseks.
infXZ=function(e1,e2){
    infXZ=rep(NA,length(e1))
    #e1=epsilon1
    #e2=epsilon2
    for (i in 1:length(e1)){
        summa0=0.5*(1-e1[i])+0.5*e2[i]
```

```
        summa1=0.5*e1[i]+0.5*(1-e2[i])
        summa=0

        #z=0, x=0
        if (1-e1[i]!=0){      #0*log0=0
            summa=0.5*(1-e1[i])*log2((1-e1[i])/summa0)
        }

        #0,1
        if (e1[i]!=0){        #0*log0=0
            summa=summa+0.5*e1[i]*log2(e1[i]/summa1)
        }

        #1,0
        if (e2[i]!=0){        #0*log0=0
            summa=summa+0.5*e2[i]*log2(e2[i]/summa0)
        }

        #1,1
        if (1-e2[i]!=0){      #0*log0=0
            summa=summa+0.5*(1-e2[i])*log2((1-e2[i])/summa1)
        }

        infXZ[i]=summa
    }

    return(infXZ)
}

#Funktsioon vastastikuse informatsioonid I(X;Y) arvutamiseks.
infXY=function(e1,e2){
    #e1 - epsilon1
    #e2 - epsilon2

    M=matrix(c(0,0,1,1,0,1,0,1),ncol=2)
    inf=rep(NA,length(e1))
    for (i in 1:length(e2)){
        Q=matrix(c(1-e1[i],e2[i],e1[i],1-e2[i]),ncol=2)
        summa=0
        for (j in 1:4){
            a1=0.5*Q[1,M[j,1]+1]*Q[1,M[j,2]+1] + 0.5*Q[2,M[j,1]+1]*Q[2,M[j,2]+1]
            a2=0.5*Q[1,M[j,1]+1]+0.5*Q[2,M[j,1]+1]
```

```
            a3=0.5*Q[1,M[j,2]+1] + 0.5*Q[2,M[j,2]+1]

            if (a1!=0){       #0*log0=0
                summa=summa+a1*log2(a1/(a2*a3))
            }
        }

        inf[i]=summa
    }

    return(inf)
}

#Funktsioon genereerib järglasjadad X_1,...,X_n ja Y_1,...,Y_n
#ning tagastab D_n/n ja L_n/n.
#s - eellasjada pikkus
F=function(s,p,Q){
    #Sarnasusmaatriksi loomine
    sarnasus=matrix(c(1,0, 0, 1), nrow=2)

    rownames(sarnasus)=c("0", "1")
    colnames(sarnasus)=c("0", "1")

    Z=rbinom(s, 1, 0.5)                    #eellase genereerimine

    muteerunudZ1=mutatsioon(Z, Q)          #jada F_1(Z_1),...,F_s(Z_s)
    muteerunudZ2=mutatsioon(Z,Q)           #jada G_1(Z_1),...,G_s(Z_s)

    allesjäämised1=rbinom(s,1,p)

    #allesjäämiste tõeväärtusvektor järglase 1 jaoks
    tõevektor1=allesjäämised1==1

    allesjäämised2=rbinom(s,1,p)

    #allesjäämiste tõeväärtusvektor järglase 2 jaoks
    tõevektor2=allesjäämised2==1


    #Soovime saada ühepikkused järglasjadad. Selleks vaatame
    #kummal tõeväärtusvektoril on rohkem tõeseid väärtusi.
    #Vastaval vektoril muudame lõpust alustades vajalikul
```

```
#hulgal tõeseid väärtusi vääradeks. Tulemuseks saame
#tõeväärtusvektorid, millel on ühepalju tõeseid väärtusi.

vahe=sum(tõevektor1)-sum(tõevektor2)
if(vahe>0){
    b=length(which(tõevektor1))
    a=b-vahe+1
    lõpuIndeksid=which(tõevektor1)[a:b]
    tõevektor1[lõpuIndeksid]=FALSE
}

if(vahe<0){
    vahe=vahe*-1
    b=length(which(tõevektor2))
    a=b-vahe+1
    lõpuIndeksid=which(tõevektor2)[a:b]
    tõevektor2[lõpuIndeksid]=FALSE
}


#leiame sugulaste indeksid
sugulasteIndeksid=intersect(which(tõevektor1),which(tõevektor2))

sugulasedX=muteerunudZ1[sugulasteIndeksid]
sugulasedY=muteerunudZ2[sugulasteIndeksid]

#võrdsete sugulaste indeksid (VSIndeksid)
VSIndeksid=sugulasteIndeksid[sugulasedX==sugulasedY]

#Suuruse Dn leidmiseks peame summeerima jada VSIndeksid pikkuse (M)
#ja suurused L(R_1^x;R_1^y),...,L(R_{M+1}^x;R_{M+1}^y)

Dn=length(VSIndeksid)


#leitakse L(R_1^x;R_1^y)
if(VSIndeksid[1]>1){
    lõik1=muteerunudZ1[1:(VSIndeksid[1]-1)]
    lõik2=muteerunudZ2[1:(VSIndeksid[1]-1)]

    tõevektoriLõik1=tõevektor1[1:(VSIndeksid[1]-1)]
    tõevektoriLõik2=tõevektor2[1:(VSIndeksid[1]-1)]
```

```
    #kontrollitakse, ega R_1^x või R_1^y pole tühihulgad
    if(TRUE %in% tõevektoriLõik1 & TRUE %in% tõevektoriLõik2){

        #alles jäävate elementide osajada lõigust 1 (R^x_1)
        lõik1=lõik1[tõevektoriLõik1]

        #alles jäävate elementide osajada lõigust 2 (R^y_1)
        lõik2=lõik2[tõevektoriLõik2]

        lõik1=c2s(lõik1)
        lõik2=c2s(lõik2)

        Dn=Dn+pairwiseAlignment(lõik1, lõik2, substitutionMatrix=sarnasus,
                                gapOpening=0, gapExtension=0, scoreOnly=TRUE)
    }
}


#leitakse L(R_2^x;R_2^y),...,L(R_M^x;R_M^y)
for(i in 2:length(VSIndeksid)){

    if(VSIndeksid[i]-VSIndeksid[i-1]>1){

        lõik1=muteerunudZ1[(VSIndeksid[i-1]+1):(VSIndeksid[i]-1)]
        lõik2=muteerunudZ2[(VSIndeksid[i-1]+1):(VSIndeksid[i]-1)]

        tõevektoriLõik1=tõevektor1[(VSIndeksid[i-1]+1):(VSIndeksid[i]-1)]
        tõevektoriLõik2=tõevektor2[(VSIndeksid[i-1]+1):(VSIndeksid[i]-1)]


        if(TRUE %in% tõevektoriLõik1 & TRUE %in% tõevektoriLõik2){

            lõik1=lõik1[tõevektoriLõik1]
            lõik2=lõik2[tõevektoriLõik2]

            lõik1=c2s(lõik1)
            lõik2=c2s(lõik2)

            Dn=Dn+pairwiseAlignment(lõik1, lõik2, substitutionMatrix=sarnasus,
                                    gapOpening=0, gapExtension=0, scoreOnly=TRUE)
        }
```

```
        }
    }


    #leitakse L(R_{M+1}^x;R_{M+1}^y)
    i=length(VSIndeksid)
    if(s>VSIndeksid[i]){

        lõik1=muteerunudZ1[(VSIndeksid[i]+1):s]
        lõik2=muteerunudZ2[(VSIndeksid[i]+1):s]

        tõevektoriLõik1=tõevektor1[(VSIndeksid[i]+1):s]
        tõevektoriLõik2=tõevektor2[(VSIndeksid[i]+1):s]

        if(TRUE %in% tõevektoriLõik1 & TRUE %in% tõevektoriLõik2){

            lõik1=lõik1[tõevektoriLõik1]
            lõik2=lõik2[tõevektoriLõik2]

            lõik1=c2s(lõik1)
            lõik2=c2s(lõik2)

            Dn=Dn+pairwiseAlignment(lõik1, lõik2, substitutionMatrix=sarnasus,
                                    gapOpening=0, gapExtension=0, scoreOnly=TRUE)
        }
    }


    #Arvutatakse L_n
    Ln=pairwiseAlignment(c2s(muteerunudZ1[tõevektor1]), c2s(muteerunudZ2[tõevektor2]),
            substitutionMatrix=sarnasus, gapOpening=0, gapExtension=0, scoreOnly=TRUE)

    n=length(which(tõevektor1))

    return(c(Dn/n,Ln/n))
}
```

```
library(Biostrings)
library(seqinr)

#Sarnasusmaatriksi loomine
sarnasus=matrix(c(1, 0, 0, 1), nrow=2)

rownames(sarnasus)=c("0", "1")
colnames(sarnasus)=c("0", "1")

#Üleminekumaatriks
Q=matrix(c(0.9, 0.1, 0.1, 0.9), nrow=2)

p=0.9
m=10000

Z=rbinom(m, 1, 0.5)

FZ1 = mutatsioon(Z, Q)
järglane1=kadumised(FZ1, p)

FZ2 = mutatsioon(Z, Q)
järglane2=kadumised(FZ2, p)

n=min(length(järglane1), length(järglane2))
järglane1=järglane1[1:n]
järglane2=järglane2[1:n]

järglane1=c2s(chars=järglane1)
järglane2=c2s(chars=järglane2)

Ln=pairwiseAlignment(järglane1, järglane2, substitutionMatrix=sarnasus,
                     gapOpening=0, gapExtension=0, scoreOnly=TRUE)
gamma=Ln/n
```