

Juhuslikud tükeldused ehk liikide arvukuse mudelid

Olgu $[n] := \{1, \dots, n\}$. Uurime hulga $[n]$ tükeldusi: $\{A_1, \dots, A_k\}$, st $A_i \cap A_j = \emptyset$ ($i \neq j$) ja $\cup_j A_j = [n]$. Pane tähele: kirjaipilt $\{A_1, \dots, A_k\}$ eeldab, et tükid on järjestatud. Iga lõplikku hulka saab alati järjestada, kuid edaspidi peame silmas sellist tükelduste järjestust, kus A_1 on see tükk, mis sisaldab arvu 1, A_2 on see tükk, mis sisaldab $[n]$ elementi, mis pole hulgas A_1 jne. Olgu $n_i = |A_i|$ ühe tüki võimsus. Seega igale tükeldusele $\{A_1, \dots, A_k\}$ vastab võimsuste vektor $(n_1, \dots, n_k)$ kuid mitte vastupidi, st vektorile $(n_1, \dots, n_k)$ vastab (üldiselt) mitu samade võimsustega ja samas järjestuses tükeldust. Vektorid $(n_1, \dots, n_k)$ rahuldavad tingimust $\sum_{i=1}^k n_i = n$ ja selliseid vektoreid nimetatakse arvu n *kompositsioonideks*.

Juhuslik tükeldus – igal tükeldusel on oma tõenäosus. Meid huvitavad nn *vahetatavad* (*infinitely exchangeable*) juhuslikud tükeldused – kõigil sama kompositsiooniga tükeldustel on sama tõenäosus $p(n_1, \dots, n_k)$, mis ei sõltu vektori $(n_1, \dots, n_k)$ komponentide järjestusest (funktsioon p on sümmeetriline). On lihtne veenduda, et juhuslik tükeldus on vahetatav, kui tema jaotus (tükelduste tõenäosused) on invariantne hulga $[n]$ permutatsioonide suhtes. Seega vahetatava juhusliku tükelduse korral on iga tükelduse tõenäosus määratud ainult tükelduste võimsustega ning $p(n_1, \dots, n_k)$ on iga sellise tükelduse tõenäosus, kus mingi (mitte ilmtingimata esimese) tüki suurus on n_1 , mingi teise tüki suurus n_2 jne.

Olgu $0 < m < n$ ning $\{A_1, \dots, A_k\}$ mingi $[n]$ tükeldus. Eemaldades tükkidest A_j elemendid $\{m+1, \dots, n\}$ ning visates välja tühjad hulgad, saame hulga $[m]$ tükelduse $\{B_1, \dots, B_l\}$, kus $B_1 = A_1 \cap [m]$, kui see ühisosa pole tühi ($B_1 = A_2 \cap [m]$ kui eelmine ühisosa on tühi, $B_1 = A_3 \cap [m]$, kui mõlemad eelmised ühisosad on tühjad jne). Kui hulga $[n]$ juhuslik tükeldus – olgu see $\mathcal{P}_n$ – on vahetatav, siis on vahetatav ka temast saadud juhuslik tükeldus $\mathcal{P}_m$ hulgal $[m]$. Lõpmatu vahetatav juhuslik tükeldus on vahetatavate juhuslike tükelduste jada $\{\mathcal{P}_n\}$, kus hulga $[n-1]$ juhuslik tükeldus $\mathcal{P}_{n-1}$ on saadud tükeldusest $\mathcal{P}_n$ arvu n välja jätmisel – tükeldused on kooskõlas. Lõputöö tegelebki lõpmatute vahelduvate tükeldustega.

Juhuslik protsess $X_1, X_2, \dots$ on vahetatav (*infinitely exchangeable*), kui iga $n > 1$ korral on $(X_1, \dots, X_n)$ jaotus invariantne komponentide järjekorra vahetamise suhtes. See tähendab, et mistahes permutatsiooni (bijektsiooni) $\sigma : [n] \rightarrow [n]$ korral $(X_1, \dots, X_n) \sim (X_{\sigma(1)}, \dots, X_{\sigma(n)})$, kus $\sim$ tähendab sama jaotust. Muuhulgas tähendab see, et kõik juhuslikud suurused, kõik paarid (X_i, X_j) , kolmikud jne on sama jaotusega jne. On selge, et sõltumatud sama jaotusega juhuslikud suurused (*i.i.d.* protsess) on vahetatav, samuti on vahetatav selliste protsesside segu. Kuulsast deFinetti teoreemist selgub, et iga vahetatav protsess on tegelikult *i.i.d.* protsesside segu. See tähendab, et iga vahetatava protsessi

saab esitada/genereerida järgmiselt:

$$\begin{aligned} P &\sim \pi \\ X_1, X_2, \dots &\stackrel{\text{i.i.d.}}{\sim} P \end{aligned} \tag{1}$$

Kirjelduses (1) on π mingi tõenäosusmõõt kõikide tõenäosusmõõtude hulgal (täpsemalt vaata [2], ptk 1). Sellest jaotusest genereeritakse tõenäosusmõõt P ja juhuslik protsess on siis $X_1, X_2, \dots$ on i.i.d. protsess, $X_i \sim P$. Seega iga vahetatav protsess on sisuliselt määratud moodsuga π .

Olgu $X_1, X_2, \dots$ vahetatav juhuslik protsess. Defineerime hulgal $\mathbb{N}$ ekvivalentsiseose:

$$i \sim j \quad \leftrightarrow \quad X_i = X_j.$$

Iga ekvivalentsiseos tekitab tükelduse (ekvivalentsiklassid) ja nii oleme defineerinud lõpmatu juhusliku tükelduse ülaltoodud definitsiooni mõttes (iga n korral moodustavad $[n]$ tükelduse juhuslikud ekvivalentsiklassid $X_1, \dots, X_n$ põhjal, tükelduste kooskõla on ilmne). Et jada $X_1, X_2, \dots$ on vahetatav, on tükeldused ka vahetatavad.

Seega iga vahetatav juhuslik protsess $X_1, X_2, \dots$ tekitab lõpmatu vahetatava tükelduse. Kingmani esitusteoreem ([1], Theorem 14.7) väidab, et iga lõpmatu vahetatav tükeldus on defineeritud mingi vahetatava protsessi poolt ning juhusliku moodsu P esituses (1) võib defineerida järgmiselt:

$$P = \sum_{i=1}^{\infty} W_i \delta_{\theta_i} + \left(1 - \sum_{i=1}^{\infty} W_i\right) G, \tag{2}$$

kus G on mingi aatomiteta tõenäosusmõõt, $\theta_1, \theta_2, \dots$ on i.i.d. juhuslikud suurused jaotusest G ning $W_1, W_2, \dots$ on juhuslik mittenegatiivsete liikmetega jada, mille summa on maksimaalslt 1 (kui summa = 1, siis (W_i) on tõenäosusjaotus). Moodsul P on aatomid W_i ja (kui summa < 1), siis ka mingi pidev osa. Et G on aatomiteta, siis on $\theta_1, \theta_2, \dots$ p.k. erinevad, mistõttu $X_k = X_l$ parajasti siis, kui $X_k = X_l = \theta_j$ ning θ_j korral. Vahetatav juhuslik protsess (1), kus juhuslik mõõt P on defineeritud kui (2) on *species sampling model (SSM)*.

Seega iga lõpmatu vahetatav juhuslik tükeldus $\{\mathcal{P}_n\}$ on saadud mingi SSM mudeli poolt. Fikseritud n korral on $\mathcal{P}_n$ jaotus defineeritud tõenäosuste $p(n_1, \dots, n_k)$ (n kompositsioonid) kaudu. Liikide ja populatsiooni terminoloogias on n populatsiooni suurus ning n_j on erinevate liikide suurused populatsioonis. On selge, et kui $\mathcal{P}_n = \{A_1, \dots, A_k\}$, kus $|A_j| = n_k$, siis $\mathcal{P}_{n+1}$ kas lisab ühe elemendi mõnele olemasolevale tükile (see juhtub siis, kui $X_{n+1} \in \{X_1, \dots, X_n\}$) või olemasolevad tükid jäävad samaks ning lisandub uus üheelemendiline tükk (see juhtub siis, kui $X_{n+1} \notin \{X_1, \dots, X_n\}$). Sündmuse " i -s tükk saab ühe elemendi juurde" tinglik tõenäosus on

$$p_i(n_1, \dots, n_k) = \frac{p(n_1, \dots, n_{i-1}, n_i + 1, n_{i+1}, \dots, n_k)}{p(n_1, \dots, n_k)}$$

ja uue tüki tekkimise tinglik tõenäosus on

$$p_{k+1}(n_1, \dots, n_k) = \frac{p(n_1, \dots, n_k, 1)}{p(n_1, \dots, n_k)}.$$

Liikide terminoloogias on $p_i(n_1, \dots, n_k)$ tõenäosus, et antud struktuuriga (liikide arvud) populatsiooni järgmine liige on sellest liigist, mille arvukus on n_i ning $p_{k+1}(n_1, \dots, n_k)$ on tõenäosus, et sünnib uus liik (mutatsioon). Nende tinglike tõenäosuste kaudu saab kirjeldada vahetatavate lõpmatute tükelduste jaotusi ning modellerida liikide tekkimise protsesse.

Hiina restorani protsess on selline lõpmatu vahetatav tükeldus, mille korral tinglikud tõenäosused avalduvad

$$p_i(n_1, \dots, n_k) = \begin{cases} \frac{n_i}{n+M}, & i = 1, \dots, k; \\ \frac{M}{M+n}, & i = k+1. \end{cases}$$

Siin $M > 0$ on parameeter. Seega tõenäosus, et sündiv isend on mingist liigist sõltub vaid selle liigi isendite arvust populatsioonis ja ei millestki muust, mutatsiooni tõenäosus ei sõltu üldse populatsiooni struktuurist. Sellise mudeli korral on SSP protsessi või juhuslike tükelduste genereerimine (sämplimine) ülilihtne, vastavat algoritmi nimetatakse *üldistatud Polya urniks*. Hiina restorani protsessile vastav SSP on selline, kus juhuslik mõõt P esituses (2) on *Dirichlet protsess*. Täpsemalt, juhuslikud kaalud $W_1, W_2, \dots$ saadakse nn tokimurdmise (*stick breaking*) protseduuri käigus. Dirichlet protsess on keskne eelmõõt (*prior*) mitteparameetrilises Bayesi statistikas; Dirichlet protsessist, Hiina restorani protsessist ning Polya urnidest loe ([1], ptk 4: [2], [3], ptk 7).

Olgu K_n tükelduste arv juhuslikus tükelduses $\mathcal{P}_n$ (ekvivalentselt erinevate väärtuste arv tükeldust tekitavas jadas $X_1, \dots, X_n$ või erinevate liikide arv populatsioonis). Seega K_n on juhuslik suurus, mis võtab väärtusi hulgas $[n]$. Hiina restorani protsessi korral

$$EK_n = \sum_{i=1}^n \frac{M}{N+i-1} \sim M \ln(n),$$

ning $K_n/\ln n \rightarrow M$, p.k. ([1], Prop 4.9). Seega kasvab erinevate liikide arv logaritmiliselt. Mõnikord on logaritmiline kasv liiga kiire või liiga aeglane ja siis Hiina restorani/Dirichlet protsess ei sobi.

Pitman-Yori protsess on selline lõpmatu vahetatav tükeldus, kus

$$p_i(n_1, \dots, n_k) = \begin{cases} \frac{(n_i-\sigma)}{n+M}, & i = 1, \dots, k; \\ \frac{M+k\sigma}{M+n}, & i = k+1. \end{cases}$$

Siin $\sigma < 1$ ja M on parameetrid, kusjuures kui $\sigma < 0$, siis $M \in \{-2\sigma, -3\sigma, \dots\}$ ja kui $\sigma \in [0, 1)$, siis $M > -\sigma$.

Kui $\sigma < 0$ ja $M = -m\sigma$, siis juhuslik tükeldus on genereeritud Dirichlet-kategoorilise protsessi poolt, st SSM on järgmine

$$(p_1, \dots, p_m) \sim \text{Dir}(m, -\sigma, \dots, -\sigma)$$

$$X_1, X_2, \dots \stackrel{\text{i.i.d.}}{\sim} (p_1, \dots, p_m)$$

(vt [3], ptk 2). Sellisel juhul $K_n \leq m$, üliõpilane leiab EK_n .

Kui $\sigma = 0$, siis Pitman-Yori protsess on Hiina restorani (Dirichlet) protsess, kui $\sigma > 0$, siis

$$EK_n \sim n^\sigma$$

ning $K_n/n^\sigma \rightarrow Y$, p.k., kus Y on juhuslik suurus.

Pitman-Yori protsess ja selle üldistused on lõputöö peamine teema. Töö on referatiivne. Peamine allikas ([1], ptk 14).

References

- [1] S. Ghosal, A. Van der Vaart *Fundamentals of Nonparametric Bayesian Inference*. Cambridge, 2017
- [2] M. Raihhelgauz *Dirichlet protsess* (Bakalureusetöö) Tartu, 2020
- [3] J. Lember *Introduction to Bayesian modelling* (Lecture notes), Tartu 2023.