1 Suboptimaalsed joondused ja nende skoor

Olgu $\mathcal{X}$ lõplik tähestik (edaspidi värvid) ning olgu P_x ning P_y kaks jaotust tähestikul $\mathcal{X}$. Olgu $X_1, X_2, \ldots$ ja $Y_1, Y_2, \ldots$ sõltumatud iid jadad, $X_i \sim P_x$, $Y_i \sim P_y$. Mida see täpselt tähendab, vaata (Vimm, 22). Töö eesmärk on uurida jadade $X_1, \ldots, X_n$ ja $Y_1, \ldots, Y_n$ sarnasuskooride asümptootilist käitumist. Sarnasuskoor $B(X_1, \ldots, X_n; Y_1, \ldots, Y_n)$ on mingil meetodil saadud jadade sarnasust isaloomustav arv, mida suurem skoor, seda sarnasemad jadad. Klassikaline skoor on pikima ühisjada pikkus, mida dünaamilise planeerimise algoritmidega (Needleman-Wunsh) saab küll arvutada, kuid millele puudub analüütiline kuju. Sarnasuskoor defineeritakse läbi joonduse, mistõttu räägime joonduse skoorist. Et pikima ühisjada pikkust on teoreetiliselt raske uurida ning üsna arvutusmahukas leida (kompleksus $O(n^2)$) pakuvad praktikas ning ka teoorias huvi nn suboptimaalsed joondused, mille skoor on küll väiksem pikima ühisjada pikkusest (see on teatavas mõttes parim/optimaalne skoor), aga mida on kergem leida ning kergem analüütiliselt uurida. Bakalaureusetöö käsitleb nn naiivseid joondusi ja nende skoore. Täpsemalt loe bakalaureusetöödest (Vimm, 22), ptk 1 või (Toots, 08).

Olgu nüüd $B(X_1, \ldots, X_n; Y_1, \ldots, Y_n)$ mingi joonduse skoor. Et jadad on juhuslikud, on ka skoor juhuslik suurus.

Eesmärk: tõestada, keskmine/suhteline skoor koondub piirkonstandiks γ:

$$\frac{B(X_1, \ldots, X_n; Y_1, \ldots, Y_n)}{n} \to \gamma, \quad \text{p.k.}$$

(edaspidi "koondumine"). Piirväärtus $\gamma(P_x, P_y)$ on konstant, mis sõltub jaotustest P_x ja P_y ning soovime leida selle funktsiooni (konstandi) analüütilist kuju. Ülaltoodud koondumine tähendab, et suure n korral on antud joonduse skoor ligikaudu γn. Teades erinevatele joondustele vastavaid konstante γ võime leida neist suurima ning praktikas kasutada suurimale konstandile vastavat joondust. Niis saame joonduste headust võrrelda.

Pane tähele (näita), et ülaltoodud koondumisest järeldub ka keskmiste (suhteliste) skooride koondumine:

$$\frac{EB(X_1, \ldots, X_n; Y_1, \ldots, Y_n)}{n} \to \gamma.$$

2 Ülevaade senitehtust

2.1 Naiivsed joondused, kahevärvilised jadad: $\mathcal{X} = \{0, 1\}$

Bitikaupa Hamming. Jadade elementide paarikaupa võrdlemine. Bakatöös (Vimm, 22) leiti piirväärtus γ_H juhul kui

$$B(X_1, \ldots, X_n; Y_1, \ldots, Y_n) = \sum_{t=1}^{n} \ell(X_t, Y_t), \quad \ell(x, y) = \begin{cases} 1, & \text{kui } x = y; \\ -1, & \text{kui } x \neq y \ . \end{cases}$$

Blokikaupa Hamming. Kahe sama värvi bloki joondamise skoor on lühima bloki pikkus. Joonduse formaalne definitsioon, piirkonstandi γ_{BH} analüütiline kuju ning koondumise formaalne tõestus bakatöös (Vimm, 22) (teoreem 7). Sama konstant on leitud ja tõestatud (mitte küll nii formaalselt) M. Tootsi magistritöös (Toots, 12) (ptk 3.1.3, valem (3.10)).

Järjestikune joondus. Järjestikune joondus X-jada järgi: X_1 joondatakse esimese Y-jada sama värvi elemendiga, X_2 joondatakse esimese Y-jada sama värvi elemendiga nii, et olemasolevat joondust ei muudeta jne. Järjestikune joondus Y-jada järgi analoogiline. Joonduse formaalne definitsioon, piirkonstandi γ_{CA} analüütiline kuju ning koondumise tõestus bakatöös (Vimm, 22) (teoreem 8).

Eelistatud tähtede meetod (cellwise alignment). Kahe värvi korral kaks võimalust: kõigepealt joondatakse kõik ühed ja siis nullid (rikkumata ühtede joondust) või vastupidi: alguses kõik nullid ja siis ühed. Artikli (Klement, Lember, 17) tulemustest järeldub, et kui $P(X_i = 1) = p_x$, $P(Y_i = 1) = p_y$, siis järjestades esimesena ühed, saame

$$\gamma_{\text{prior1}} = \frac{p_x \wedge p_y}{p_x + p_y - p_x p_y}.$$

2.2 Naiivsed joondused, mitmevärvilised jadad: $\mathcal{X} = \{1, 2, \ldots, K\}$

Bitikaupa Hamming. Jadade elementide paarikaupa võrdlemine. Tõestus sama, mis kahe värvi korral, piirväärtus γ_H on toodud (Toots, 12), valem (3.1). NB! Toots defineerib

$$\ell(x, y) = \begin{cases} 1, & \text{kui } x = y; \\ 0, & \text{kui } x \neq y\ . \end{cases}$$

Blokikaupa Hamming. Sama, mis bitikaupa Hamming, kuid blokkidega. NB! erinevus kahest värvist: esimesi värve ei panda kokku sobituma. St kui kahevärvilisel juhul $X_1 \neq Y_1$, siis blokikaupa Hammingu skoor on 0! Piirkonstandi γ_{BH} analüütiline kuju, koondumise tõestus (koos suurte hälvete võrratusega) on magistritöös (Toots, 12) (teoreem 3.1).

Järjestikune joondus. Definitsioon sama, mis kahe värvi korral, piirkonstandi analüütiline kuju, koondumise tõestus (koos suurte hälvete võrratusega) on magistritöös (Toots, 12) (teoreem 3.2).

Blokikaupa järjestikune joondus. Blokikaupa järjestikune joondus X-jada järgi: X-jada esimene blokk joondatakse esimese Y jada sama värvi blokiga, X-jada järgmine blokk joondatakse jälle esimese Y jada sama värvi blokiga, olemasolevat joondust rikkumata jne. Kahe värvi korral sama, mis blokikaupa Hamming. Piirkonstandi analüütiline kuju, koondumise tõestus (koos suurte hälvete võrratusega) on magistritöös (Toots, 12) (teoreem 3.3).

Eelistatud tähtede meetod (cellwise alignment) Kõigepealt joondatakse ühed, siis (olemasolevat ühtede joondust rikkumata) joondatakse kõik kahed jne (priority letter alignment). Seda joondust on uuritud artiklis (Klement, Lember, 17), kus on tõestatud koondumine (koos suurte hälvete võrratusega) ning toodud rekursiivne valem piirkonstandi γ arvutamiseks.

2.3 Eeltöötlus: kahevärvilised jadad, ebasümmeetria

Artiklis (Barder, et al, 10) käsitletakse ebasümmeetrilisi jaotusi

$$\epsilon := P(X_i = 1) = P(Y_i = 0), \quad \epsilon < 0.5$$

st ühes jadas on ühtedel väike tõenäosus ning teises jadas on nullidel sama väike tõnäosus. Olgu

$$N_x := \sum_{t=1}^{n} X_t, \quad N_y := \sum_{t=1}^{n} (1 - Y_t), \quad M_n := N_x + N_y$$

Seega N_x on ühtede arv X-jadas (neid on keskmiselt vähem), N_y on nullide arv Y-jadas ning M_n on nende summa (vähemuses olevate elementide summa). On selge, et M_n on ülemine tõke pikima ühisjada pikkusele. Kui jadad algavad sama värviga ning i-s 1-blokk X-jadas on väiksem i 1-blokist Y jadas ning iga 0-blokk Y jadas on väiksem vastavast 0 blokist X jadas, siis blokikaupa Hammingu skoor $B_{BH}(X_1, \ldots, X_n, Y_1, \ldots, Y_n) = M_n$ ja sellisel juhul annab blokikaupa Hamming meile pikima ühisjada. Sellest tulenevalt defineerime kaotatud bittide arvu

$$N_n := M_n - B_{BH}(X_1, \ldots, X_n, Y_1, \ldots, Y_n).$$

Kui N_n on väike, siis blokikaupa Hamming töötab suhteliselt hästi. Ülaltoodust järeldub, et

$$\frac{N_n}{n} \to 2\epsilon - \gamma_{BH}(\epsilon) =: \nu(\epsilon), \quad \text{.a.s.}$$

Avalda $\gamma_{BH}(\epsilon)$ (Vimm, 10 trm. 7) antud spetsiifilisel juhul ja veendu, et ülaldefineeritud $\nu(\epsilon)$ ühtib valemiga (2.4) artiklist (Barder et al, 10).

K-eeltöötluse idee on elimineerida liiga väikesed suured blokid. Olgu N_n^K kaotatud bittide arv peale K-eeltöötlust. Artiklis (Barder et al, 10) on näidatud (teoreem 2.1):

$$\frac{N_n^K}{n} \to \nu_K(\epsilon), \quad \text{p.k.}, \quad \nu_K(\epsilon) := 2\epsilon^2 \frac{\left(1 - (1-\epsilon)^{K+1}\right)^{K+1}}{\epsilon + (1-\epsilon)^{K+2}}. \tag{2.1}$$

Artiklis on uuritud ka seda, milline on parim/optimaalne $K(\epsilon)$. Artikkel lõpeb nn "online"-eeltöötlusega, kus blokkide kustutamine ühes jadas sõltub teisest jadast (K-eeltöötlust saab läbi viia mõlemas jadas eraldi), vastav $\nu(\epsilon)$ on tuletatud (rea summana).

3 Bakalaureusetööde uurimissuunad/teemad

3.1 Teema 1: eeltöötlus üldisemal juhul

Olgu endiselt $\mathcal{X} = \{0,1\}$, kuid jaotused P_x ja P_y on üldisemad ja mitte nii rangelt ebasümmeetrilised. Bakatöös (Vimm, 22) läbiviidud analüüs (ptk 2) näitas, et blokikaupa Hamming on parim naiivne joondus paljudel muudel juhtudelgi (üllataval kombel ka siis kui $p_x = p_y$). Seega vaatleme olukorda kui teatav ebasümmeetria on olemas $p_x < 0.5$ ja $p_y > 0.5$ (tuleta meelde, et p_x ja p_y on ühtede tõenäosused), kuid mitte ilmtingimata $p_x = 1 - p_y$ nagu enne. Eesmärk on tõestada koondumine (2.1) ja üldistada valemit $\nu_K(\epsilon)$ üldisemale juhule (nüüd siis $\nu_K(p_x, p_y)$). Juhul kui $K = 0$, saame (näita!)

$$\nu(p_x, p_y) = p_x + (1 - p_y) - \gamma_{BH}(p_x, p_y).$$

Kui $\nu_K(p_x, p_y)$ leitud, siis leia/hinda optimaalne $K(p_x, p_y)$. Leia $\nu_o(p_x, p_y)$ "online" eeltöötluse korral.

Kui konstandid teoreetiliselt leitud ja koondumised tõestatud, siis vii läbi simulatsioonid:

- Kontrolli simulatsioonidega, kas tõestatud koondumised pravad paika. Kas simulatsioonid kinnitavad teoreetilisi tulemusi – $\nu_K(p_x, p_y)$ kuju.
- Vii läbi kõikide meetoodide võrdlus: etteantud (p_x, p_y) leia kõikidel naiivsetel ja eeltöötlusega leitud meetodidte asümptootilised skoorid (kaasa arvatud erinevad eeltöötluse K-d, erinevad eelistatud tähtede järjestused) ning leia: 1) parim meetod (antud jaotuste korral); 2) kaugus Chvatal-Sankovi konstandist $\gamma^*(p_x, p_y)$ (see on see arv, mida üritame lähendada ja seega tahaks teada, kui lähedale oleme jõudnud). Vaata (Vimm, 22, ptk 2). Kas eeltöötlusega saab rohelist ala joonisel 2 suurendada (sel joonisel pole eelistatud tähtede meetodi kaasatud).

3.2 Teema 2: Sõltumatud Markovi ahelad

Olgu endiselt $\mathcal{X} = \{0,1\}$, kuid loobume eeldusest, et $X_1, X_2, \ldots$ ja $Y_1, Y_2, \ldots$ on iid jadad. Olgu nad mõlemad statsionaarsed ja homogeensed Markovi ahelad üleminekumaatrikistega

$$P_x = \begin{pmatrix} p^x_{00} & 1 - p^x_{00} \\ 1 - p^x_{11} & p^x_{11} \end{pmatrix}, \quad P_y = \begin{pmatrix} p^y_{00} & 1 - p^y_{00} \\ 1 - p^y_{11} & p^y_{11} \end{pmatrix}.$$

Markovi ahelad võimaldavad blokkide struktuuri palju paindlikumalt modelleerida. Veendu selles. Erijuhul saame muidugi iid jadad (kuidas?). Olgu jadad endiselt sõltumatud. Bakatöö käsitleb naiivseid joondusi ja eesmärk on tõestada naiivsete skooride koondumine piirkonstandiks $\gamma(p^x_{00}, p^x_{11}, p^y_{00}, p^y_{11})$. Täpsemalt: leia $\gamma(p^x_{00}, p^x_{11}, p^y_{00}, p^y_{11})$ ja tõesta koondumine bitikaupa Hammingu, blokikaupa Hammingu, järjestikuse joonduse ja eelistatud tähtede joonduse korral. Teisisõnu, bakatöös (Vimm, 22) läbiviidud tõestused tuleb üldistada Markovi ahelatele. Sõltumatuse erijuhul pead saama olemasolevad valemid.

Kui konstandid teoreetiliselt leitud ja koondumised tõestatud, siis vii läbi simulatsioonid:

- Kontrolli simulatsioonidega, kas tõestatud koondumised pravad paika. Kas simulatsioonid kinnitavad teoreetilisi tulemusi – $\gamma(p_{00}^x, p_{11}^x, p_{00}^y, p_{11}^y)$ kuju.
- Vii läbi meetodidte võrdlus nagu (Vimm, 22, ptk 2).

Kirjandus

Vimm, K (2022). Suboptimaalsed meetodid jadade sarnasuse võrdlemiseks kahetähelise tähestiku puhul. Bakalureusetöö. TÜ

Toots, M. (2008). Chvatal-Sankovi konstandi hindamine simulatsioonide abil. Bakalaureusetöö. TÜ.

Toots, M. (2012). Suboptimal Alignments and Similarity of Random Sequences. Magistritöö. TÜ.

Barder, S., J. Lember, H. Matzinger ja M. Toots (2012). On Suboptimal LCS-Alignments for Independent Bernoulli Sequences with Asymmetric Distributions. *Methodol Comput Appl Probab* **14**, lk. 357 – 382.

Klement, R. ja J. Lember (2017). On expected score of cellwise alignments. *Acta et Comm. Universitetis Tartuensis* **21(1)**, lk. 141 – 165