# Lõputööd segujaotustest, parameetrite hindamisest ja segmeteerimisest

Tööde eesmärk on tutvuda segujaotuse ja tema parameetrite hindamise meetooodidega (EM-algoritm, Viterbi treening ja parandatud Viterbi treening), samuti segmeteerimise ja Bayesi leähenemisviisiga. Teemad on valdavalt teoreetilist laadi kuid sisaldavad ka arvutisimulatsioone. Muuhulgas tutvub üliõpilane suurima tõepära meetodiga ja mudeli ühesuse küsimustega.

Alljärgnev tekst on samaaegselt nii õppematerjal kui ka juhend lõputööde kirjutamiseks.

## 1 Segujaotus

**(Diskreetne) segujaotus** (*mixture distribution*) on ruumil $\mathbb{R}^d$ antud tõenäosjaotus $P$, mis esitub kujul

$$P = \sum_{i=1}^{m} p_i P_i, \tag{1}$$

kus $m \in \mathbb{N}$, $(p_1, \ldots, p_m)$ on tõenäosuste vektor (mittenegatiivesd arvud, mille summa on 1) ja iga $i$ korral on $P_i$ samuti tõenäosusjaotus ruumil $\mathbb{R}^d$. Summa (1) tähendab seda, et iga (mõõtuva hulga) $B \subset \mathbb{R}^d$ korral

$$P(B) = \sum_{i=1}^{m} p_i P_i(B).$$

Sellest järeldub, et kui $F$ on jaotuse $P$ jaotusfunktsioon ning $F_i$ on jaotuse $P_i$ jaotusfunktsioon, siis iga $x \in \mathbb{R}^d$ korral

$$F(x) = \sum_{i=1}^{m} p_i F_i(x).$$

Samuti järeldub definitsioonist (1) (ja seda üliõpilane näitab formaalselt), et jaotused $P_i$ on absoluutselt pidevad (st neil on tihedus $f_i$) parajasti siis kui $P$ on absoluutslet pidev, kusjuures $P$ tihedusfunktsioon $f$ avaldub järgmiselt

$$f(x) = \sum_{i=1}^{m} p_i f_i(x).$$

Jaotusi $P_i$, $i = 1, \dots, m$ nimetatakse **emissioonijaotusteks** ja vastavaid tõenäosusi $(p_1, \dots, p_m)$ **kaaludeks**. Praktikas on oluline erijuht, kui $P_i$ on normaaljaotusega, näited normaaljaotuste segu kohta vaata [1],

**Ühesus.** On kerge näha (ja siinkohal näide), et lahutus (1) ei pruugi olla ühene: võib olla, et leiduvad jaotused $P'_i$ $(i = 1, \dots, l)$ ja tõenäosused $p'_1, \dots p'_l$ nii, et

$$P = \sum_{i=1}^{m} p_i P_i = \sum_{j=1}^{l} p'_j P'_j.$$

Enamasti vaadeldakse segujaotusi, kus komponendid kuuluvad ühte klassi $\mathcal{P}$ (näiteks normaaljaotused, Poissoni jaotused). Kas segu on ühene või mitte sõltub klassist $\mathcal{P}$. Öeldakse, et emissioonijaotuste klass $\mathcal{P}$ on **ühene** (*identifiable*) kui iga $m, l \in \mathbb{N}$, kõikide kaaluvektorite $(\alpha_1, \dots, \alpha_m)$, $(\beta_1, \dots, \beta_l)$ ja kõikide jaotuste $P_1, \dots, P_k, P'_1, \dots, P'_l \in \mathcal{P}$ korral kehtib

$$\sum_{i=1}^{m} \alpha_i P_i = \sum_{i=1}^{l} \beta_i P'_i \quad \Leftrightarrow \quad m = l, \quad \alpha_i = \beta_i, \quad P_i = P'_i, \quad \forall i. \tag{2}$$

Loomulikult ülaltooduv võrdus kehtib permutatsiooni täpsusega (st võrdus kehtib mingi komponentide järjestuse/permutatsiooni korral).

Seega ühesus on jaotuste hulga $\mathcal{P}$ omadus. Olgu $\mathcal{P}$ mingi diskreetsete jaotuste lõplik hulk, kusjuures kõikidel neil jaotustel on samad aatomid $(a_1, \dots, a_k)$. Siis iga jaotus $P \in \mathcal{P}$ on samastatav tõenäosuste vektoriga $(p_1, \dots, p_k)$ ja klass $\mathcal{P}$ on samastatav maatriksiga. Näita, et sellisel juhul on segujaotuse ühesus sama, mis vektorite hulga $\mathcal{P}$ lineaarne sõltumatus.

Olgu nüüd $\mathcal{P}$ suvalise võimsusega (mitte ilmtingimata diskreetsete) jaotuste klass. Iga $k$ korral defineeritakse mõõtude $P_1, \dots, P_k$ lineaarne sõltumatus analoogiliselt (kuidas?) ja klassi $\mathcal{P}$ lineaarne sõltumatus defineeritakse järgmiselt: iga $k$ korral on suvalised $k$ erinevat jaotust $P_1, \dots, P_k$ lineaarselt sõltumatud. On lihtne näha (ja see on Theorem 1 artiklis [4]), et klass $\mathcal{P}$ on ühene parajasti siis, kui ta on lineaarselt sõltumatu. Üliõpilane tõestab selle väite.

Artiklis [4] on näidatud, et väga paljud praktikas kasutatavad jaotuste klassid on ühesed (lineaarelt sõltumatud), neid tulemusi tuleb refereerida.

**Segujaotus kui varjatud tunnusega mudel.** Olgu $Y$ diskreetne juhuslik suurus, mis võtab $m$ erinevat väärtust $(1,\ldots,m)$ tõenäosutega $(p_1,\ldots,p_m)$. Olgu $X$ juhuslik suurus, mille saame järgmiselt: kui juhuslik suurus $Y$ võtab väärtuse $i$, genereeritake $X$ väärtus jaotusest $P_i$. Seega $Y$ otsustab $X$ jaotuse. Siis on $X$ segujaotusega (1), üliõpilane näitab seda. Juhusliku suuruse $Y$ väärtusi nimetatakse **varjatud tunnusteks**, vt ka [1].

# 2 Segmenteerimine

Olgu nüüd $X_1,\ldots,X_n$ sõltumatud juhuslikud suurused segujaotusest (1) ja $x_1,\ldots,x_n$ nende realiastsioonid ehk **valim.** Valimi saame kas otse segujaotusest vaatlusi genereerides või varjatud tunnuste kaudu järgmiselt: olgu $y_1,\ldots,y_n$ iid juhulike suuruste (varjatud tunnuste) $Y_1,\ldots,Y_n$ realisatsioon. Vaatlus $x_t$, kus $t=1,\ldots,n$ genereeritakse jaotusest $P_{y_t}$, sõltumate teistest vaatlustest.

**Segmenteerimine** on varjatud tunnuste jada $(y_1,\ldots,y_n)$ hindamine/prognoosimine vaatluste $x_1,\ldots,x_n$ põhjal. Oletame nüüd, et emissioonijaotustel $P_i$, $i=1,\ldots,m$ on tihedus/tõnäosusfunktsioon $f_i(x)$. Vaatleme hinnaangut $(\hat{y}_1,\ldots,\hat{y}_n)$, kus

$$\hat{y}_t = \arg\max_{i=1,\ldots,m} p_i f_i(x_t).$$

Selline hinnang on hea mitmes mõttes:

- Maksimiseerib tinglikku tõenäosust:

  $$(\hat{y}_1,\ldots,\hat{y}_n) = \arg\max_{(a_1,\ldots,a_n)\in\{1,\ldots,m\}^n} P(Y_1=a_1,\ldots,Y_n=a_n|X_1=x_1,\ldots,X_n=x_n)$$

- Minimiseerib keskmist segmenteerimisvigade arvu:

  $$(\hat{y}_1,\ldots,\hat{y}_n) = \arg\min_{(a_1,\ldots,a_n)\in\{1,\ldots,m\}^n} E\big[L(Y_1,\ldots,Y_n;a_1,\ldots,a_n)|X_1=x_1,\ldots,X_n=x_n],$$

  kus $L(y_1,\ldots,y_n;a_1,\ldots,a_n)=\sum_{t=1}^n I_{y_t\neq a_t}$.

Üliõpilane tõestab need lihtsad väited.

Sõltumatu valimi korral on segmenteerimine tihedalt seotud tehisõppest tuntud klassifitseerimisprobleemiga (*pattern recognition*), vt [2], ptk 1.

# 3 Parameetrite hindamine, STP

**Segujaotuse parameetrid.** Kõigi $m$-dim kaaluvektorite hulk on

$$\mathcal{P}^m := \{q = (p_1, \ldots, p_m) : p_i \geq 0, \sum_{i=1}^m p_i = 1\}.$$

Olgu

$$\mathcal{P} = \{P_\lambda\}_{\lambda \in \Lambda}$$

emissioonijaotuste klass. Tähistame

$$\Theta = \cup_{m \geq 1} \mathcal{P}^m \times \Lambda^m$$

ja selle hulga elemente $\theta = (q, \lambda)$ nimetame jaotuse **parameetriteks**. Igale parameetrile $\theta = (q, \lambda)$ vastab segujaotus $P_\theta = \sum_{i=1}^m p_i P_{\lambda_i}$. Kui $\mathcal{P}$ on ühene, on kõikvõimalike segujaotuste hulga, olgu see hulk $\mathcal{Q}$, ja parameetrite hulga $\Theta$ vahel üksühene seos (miks?), vastasel juhul võib olla aga nii, et kaks erinevat parameetrit $\theta$ ja $\theta'$ annavad ühe ja sama segujaotuse.

**Segujaotuse/parameetrite hindamine.** Olgu $x_1, \ldots, x_n$ valim mingist segujaotusest $P$. **Parameetrite hindamise probleem** on jaotuse $P$ leidmine/ära arvamine/hindamine valimi $x_1, \ldots, x_n$ põhjal. Kui hulkade $\mathcal{Q}$ ja $\Theta$ vahel on üksühene vastavus, siis on jaotuse hindamine ekvivalentne parameetri $\theta$ hindamisega.

Oletame nüüd, et igal emissioonijaotusel $P_\lambda$ on tihedus/tõenäosusfunktsioon $f_\lambda$. Teame, et sellisel juhul on ka igal segujaotusel hulgast $\mathcal{Q}$ tihedus. Seega on $\mathcal{Q}$ (segujaotuse) tiheduste hulk. **Suurima tõepära printsiip (STP)** valib hulgast $\mathcal{Q}$ sellise jaotuse/tiheduse – **suurima tõepära (STP) hinnangu** – mis maksimiseerib valimi tihedust (miks korrutis?):

$$\hat{f} = \arg\max_{f \in \mathcal{Q}} \prod_{t=1}^n f(x_t). \tag{3}$$

Et igale parameetrile $\theta \in \Theta$ vastab tihedus $f_\theta$, võime ülaltoodud maksimiseerimisülesande parameetrite abil kirjutada

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \prod_{t=1}^n f_\theta(x_t). \tag{4}$$

Kui parameetrite hulga $\Theta$ ja segujaotuste hulga $\mathcal{Q}$ vahel on üksühene vastavus (st emissioonijaotuste klass $\mathcal{P}$ on ühene ehk lineaaarselt sõltumatu), siis on ülaltoodud kaks optimeerimisülesannet (3) ja (4) ekvivalentsed, vastasel juhul ei pruugi (4) lahend olla ühene.

Tegelikult ei pruugi suurima tõepära hinnang $\hat{f}$ (ekvivalentselt $\hat{\theta}$) olla ühene ka siis, kui emissioonijaotuste klass on ühene. Too näide segujaotustest, mille korral suurima tõepära hinnang $\hat{f}$ pole ühene.

Veel enam – suurima tõepära hinnngut ei pruugi üldse leiduda, st ülesannetel (3) ja (4) ei pruugi leiduda lahend. Too näide.

**EM-algoritm.** Et

$$f_\theta(x) = \sum_{i=1}^{m} p_i f_{\lambda_i}(x),$$

saame ülesandele (4) kuju

$$(\hat{m}, \hat{q}, \hat{\lambda}) = \arg \max_{m \geq 1, q \in \mathcal{P}^m, \lambda \in \Lambda^m} \prod_{t=1}^{n} \big( \sum_{i=1}^{m} p_i f_{\lambda_i}(x_t) \big).$$

Seda ülesannet ei saa üldiselt analüütiliselt lahendada. Tihti on komponentide arv $m$ teada, kuid see ei tee ülesannet kergemaks.

Selleks, et STP hinnangut $\hat{\theta}$ leida, kasutatakse nn iteratiivseid meetode. Iteratiivsus tähendab, et alustatakse mingist prameetite algähendist $\theta^{(0)}$, algoritm väljastab järgmise lähendi $\theta^{(1)}$, see sisestatakse algoritmi, mis seejärel väljastab $\theta^{(2)}$. Nii saadakse jada $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots$, mis loodetavasti koondub piirväärtuseks $\hat{\theta}$. See piirväärtus (mis üldiselt sõltub alglähendist) ongi algoritmi väljund – hinnang.

Tuntuim iteratiivne ülesande (4) lahendusalgoritm on **EM-algoritm**. Üliõpilane kirjeldab EM-algoritmi segujaotuste parameetrite hindamiseks (vt [3]) ja tõestab, et

$$\prod_{t=1}^{n} f_{\theta^{(i)}}(x_t) \leq \prod_{t=1}^{n} f_{\theta^{(i+1)}}(x_t).$$

Ülaltoodud võrratus tähendab, et EM-algoritmi iga iteratsioon suurendab tõepära (omptimiseeritav kriteerium). Sellest tehakse tihti ekslik järeldus, et jada $\theta^{(i)}$ koondub mingiks tõepärafunktsiooni lokaalseks maksimumiks. Väide iseenesest pole alati vale.

# 4 Teema 1: Viterbi treening ja parandatud Viterbi treening

**Viterbi treening (VT).**

**Sisend:** valim $x_1, \ldots, x_n$, komponentide arv $m$, emissioonijaotused $\{f_\lambda\}_{\lambda \in \Lambda}$;

**Start:** algaväärtused $\theta^{(0)}$ ehk $q^{(0)} = (p_1^{(0)}, \ldots, p_m^{(0)})$ ja $\lambda^{(0)} = (\lambda_1^{(0)}, \ldots, \lambda_m^{(0)})$

**Iteratsioon:** antud $q^{(i)}$ ja $\lambda^{(i)}$ korral:

- Leia varjatud tunnuste hinnangud $\hat{y}^{(i)} = \hat{y}_1, \ldots, \hat{y}_n$ ning jaga valim nende abil osavalimiteks $T_j = \{t \in (1, \ldots, n) : \hat{y}_t = j\}$.
- Leia iga osavalimi korral emissiooniparameetrite STP hinnangud
$$\lambda_j^{(i+1)} = \arg\max_{\lambda \in \Lambda} \prod_{t \in T_j} f_\lambda(x_t), \quad j = 1, \ldots, m.$$
- Leia kaalude STP hinnangud:
$$q^{(i+1)} = \arg \max_{(p_1,\ldots,p_m) \in \mathcal{P}^m} \prod_{t=1}^{n} p_{\hat{y}_t} = \arg \max_{(p_1,\ldots,p_m) \in \mathcal{P}^m} \prod_{j=1}^{m} (p_j)^{|T_j|}$$
ehk
$$p^{(i+1)} = \frac{|T_j|}{n}.$$

**Lõpp:** Lõpeta kui $\theta^{(i+1)} = \theta^{(i)}$;

**Väljund:** hinnang $\hat{\theta} = \theta^{(i+1)}$.

Defineerime iga jada $y = (y_1, \ldots, y_n) \in \{1, \ldots, m\}^n$ ja valimi $x = (x_1, \ldots, x_n)$ korral

$$p(x|y, \lambda) = \prod_{t=1}^{n} f_{\lambda_{y_t}}(x_t) \quad q(y) = \prod_{t=1}^{n} p_{y_t}; \quad p(x, y|\theta) = p(x|y, \lambda)q(y), \quad \lambda \in \Lambda, q \in \mathcal{P}_m.$$

Veendu, et

$$p(x, \hat{y}^{(i)}|\theta^{(i)}) \leq p(x, \hat{y}^{(i)}|\theta^{(i+1)}) \leq p(x, \hat{y}^{(i+1)}|\theta^{(i+1)}) \leq \cdots .$$

Näita, et VT sisend võib olla ka jada $y^{(0)}$. Kuidas algoritm näeb välja sellisel juhul?

**Simulatsioonid:** Veendu, et VT töötab (samade alglähendite korral) kiiremini ja efektiivsmealt kui EM, kuid annab üldjuhul palju halvema hinnangu.

## 4.1 Asümptootilise püsipunkti omadus ja AVT

Selles peatükis eeldame, et emissioonijaotuste/tiheduste hulk $\{f_\lambda\}_{\lambda\in\Lambda}$ on ühene, st segujaotuste ja parameetrite vahel on üksühene vastavus.

Veel üks eeldus hulga $\{f_\lambda\}_{\lambda\in\Lambda}$ kohta:

Olgu $X_1, X_2, \ldots$ iid vaatlused jaotusest $Q$, kusjuures leidub ainult üks $\lambda'$ (sõltub jaotusest $Q$) nii, et

$$\lambda' = \arg\max_{\lambda\in\Lambda} \int \ln f_\lambda(x) Q(dx).$$

Siis kehtib koondumine

$$\arg\max_{\lambda\in\Lambda} \frac{1}{n}\sum_{i=1}^{n} \ln f_\lambda(X_i) \to \lambda', \quad \text{p.k.} \tag{5}$$

Too näiteid emissioonitihedustest $\{f_\lambda\}_{\lambda\in\Lambda}$, mille korral see eeldus on täidetud.

Olgu $P$ tegelik jaotus (millest pärinevad iid juhuslikud suurused $X_1, X_2 \ldots$) ja olgu $\theta$ vastav (tegelik) parameeter. Rakendame Viterbi treeningut võttes alglähendiks parameetri $\theta$, st $\theta^{(0)} = \theta$ ja uurime VT esimest iteratsiooni $\theta_n^{(1)}$ (siin $n$ viitab valimile). Et $\theta_n^{(1)} = (q_n^{(1)}, \lambda_n^{(1)})$ on juhuslike suuruste $X_1, \ldots, X_n$ funktsioon, on ta ise ka juhuslik vektor. Leidub vektor $q^{(1)} \in \mathcal{P}^m$ nii, et

$$q_n^{(1)} \to q^{(1)}, \quad \text{p.k..} \tag{6}$$

Üliõpilane tõestab selle koondumise. Vektori $q^{(1)}$ analüütiline kuju normaaljaotuste segu korral ning juhul kui $m = 2$ on toodud artiklis [1]. Leia see vektor kui $m > 2$ ja/või ka juhul kui komponendid pole normaaljaotusega.

Kui kehtib eeldus (5), siis leidub vektor $\lambda^{(1)} \in \Lambda^m$ nii, et

$$\lambda_n^{(1)} \to \lambda^{(1)} \quad \text{p.k..}$$

Üliõpilane tõestab ka selle koondumise. Seega, tehtud eeldustel kehtib koondumine

$$\theta_n^{(1)} \to \theta^{(1)}, \quad \text{p.k.,}$$

kus $\theta^{(1)} = (q^{(1)}, \lambda^{(1)}$. Siinkohal paneme tähele, et $\theta_n^{(1)}$ ja ka piirväärtus $\theta^{(1)}$ sõltuvad parameetrist $\theta$. Tähistame

$$\Delta(\theta) = \theta - \theta^{(1)}(\theta).$$

Üliõpilane arvutab funktsiooni $\theta \mapsto \Delta(\theta)$ lihtsate näidete korral ja veendub, et see pole 0. Seega Viterbi treeningul on järgmine omadus: alustates tegelikust parameetrist $\theta$, väljastab algoritm midagi muud ja seda isegi siis, kui $n = \infty$. Seega VT algoritmil puudub asümptootiline püsipunkti omadus.

Üliõpilane tõestab, et samadel eeldustel on EM-algoritmil asümptootiline püsipunkti omadus.

Asümptootilise püsipunkti omadust saab formaliseerida järgmiselt. Olgu $T_\theta : \Theta \to \Theta$ operaator, mis vastab algoritmi (EM või VT) ühele iteratsioonile juhul kui $n = \infty$. See tähendab $T_\theta(\theta^{(i)}) = \theta^{(i+1)}$. Algoritm sõltub tegelikust jaotusest ehk parameetrist $\theta$, seda näitab alaindeks. Asümptootiline püsipunkti omadus on see, et $T_\theta(\theta) = \theta$.

**Parandatud Viterbi treening (AVT).** AVT tekitab asümptootilise püsipunkti omaduse sel moel, et igal iteratssioonil lisatakse VT väljundile $\theta_{VT}^{(i)}$ korrektsioon $\Delta(\theta^{(i)})$ (vt [6]). AVT treeningul on asümptootilise püsipunkti omadus.

## 4.2 Simulatsioonid

Simulatsioonide eesmärk on EM, VT ja AVT treeningalgoritmide võrdlus. See tähendab, et tudeng leiab ja programeerib kujutised $T_\theta^{VT}, T_\theta^{EM}, T_\theta^{AVT}$ mingi suhteliselt lihtsal juhul ($m = 2$ või $m = 3$, emissioonijaotued on näiteks normaaljaotused). Kui $\theta^{(0)} \neq \theta$, siis väljastavad kõik algoritmid jada $\theta^{(0)}$,

$\theta^{(1)}, \theta^{(2)}, \ldots$. Kas (ja miks) see jada koondub? Kui koondub, siis piirväärtus on hinnang $\hat{\theta}$, mis sõltub $\theta^{(0)}$-st. Ja sellisel juhul on meil kujutis

$$S_\theta : \Theta \to \Theta, \quad S_\theta(\theta^{(0)}) = \hat{\theta}.$$

Millise algoritmi korral on koondumine kõige kiirem? Millise algoritmi korral on lõpp-hinnang $\hat{\theta}$ kõige täpsem ehk viga $S_\theta(\theta^{(0)}) - \theta$ kõige väiksem? Kui tundlik on viga alglähendi $\theta^{(0)}$ suhtes? Kui $\theta \approx \theta^{(0)}$, kas siis $S_\theta(\theta^{(0)}) \approx \theta$?

Algoritme võib ka võrrelda simuleeritud andmetel nag on tehtud artiklis [6].

# 5 Teema 2: Bayesi lähenemisviis ja segmenteerimine

Tundmatute parameetrite hindamine enne statistiliste järelduste tegemist on statistikas klassikaline lähenemisviis (*frequentist inference*). Bayesi lähenemisviisil vaadeldakse parameetreid teatava jaotusega, nn eeljaotusega juhuslike suurustena. Oletame (esialgu), et emissiooniparameetrid $\lambda$ on teada. Seega kaalud $q = (p_1, \ldots, p_m)$ on juhuslikud, nende täpne väärtus pole teada, küll aga nende jaotus (eeljaotus). Bayese statistikas modelleeritakse kaale enamasti Dirichlet jaotusega $\mathrm{Dir}(\alpha_1, \ldots, \alpha_m)$. Dirichlet jaotusel on palju häid omadusi, sellest jaotusest loe täpsemalt konspekist [7]. Kui $m = 2$, siis Dirichlet jaotust nimetatakse Beta jaotuseks. Seega meie mudel on järgmine

$$\begin{aligned} q &\sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_k) \qquad (7) \\ Y_1, \ldots, Y_n | q &\overset{i.i.d.}{\sim} \sum_{i=1}^{k} p_i \delta_i \\ X_i | Y_i, &\overset{ind}{\sim} P_{\lambda_{Y_i}}, \quad i = 1, \ldots, n. \end{aligned}$$

Fikseeritud $q$ korral on $Y_1, \ldots, Y_n$ iid jaotusega $\sum_i \delta_{\lambda_i} p_i$ juhuslikud suurused, kuid meie mudelis nad enam pole sõltumatud [7]. Juhuslikku protsessi $Y_1, Y_2, \ldots$ nimetatakse **Dirichlet-kategooriliseks protsessiks**, juhul kui $m = 2$, siis nimetatakse seda ka **Beta-Beroulli protsessiks**. Nende protsesside omadused on väga erinevad iid juhuslike suuruste omadustest, vaata [7, 1].

Kuigi parameeterid ei pea enam hindama, teeb Bayesi lähenemine segmenteerimise märksa keerulisemaks. Näiteks tingliku tõenäosuse maksimiseerimine ja keskmise vigade arvu mineerimine pole enam üks ja seesama – üliõpilane näitab seda, samuti pole kumbagi kriteeriumi enam lihtne optimeerida. Bakatöö keskendub suurima tõepäraga hinnangu $\hat{y}_1, \ldots, \hat{y}_n$ omadustele. Nagu näidatud artiklis [1], saab selle ülesande esitada nn karistusliikmega kaofunktsiooni optimeerimisele (üliõpilane üldistab seost (6)). Juhul kui komponendid on normaaljaotusega, on ülesanne intutitiivselt hästi arusaadav. Artiklis [1] toodud argumendid ja simulatsioonid näitavad, et Bayesi lähenemisviisil saadud hinnang $\hat{y}_1, \ldots, \hat{y}_n$ on märksa konservatiivsem kui iid mudeli korral saadud hinnang, selle põhjus on Dirichlet-kategoorilise protsessi omadused. Konservatiivsus tähendab, et jada $\hat{y}_1, \ldots, \hat{y}_n$ muutub (hüppab) vähe ning tihti on konstantne: $\hat{y}_1 = \cdots = \hat{y}_n$. Artiklis [1] toodud simulatsioonid näitavad, et kui $m = 2$, mõlemad komponendid on sama dispersiooniga $\sigma^2$ ja normaaljaotusega (keskväärtused 0 ja 1), siis muutub hinnang konstantseks, kui $\sigma \geq \sigma_o$, kusjuures $\sigma_o \approx 0.8$ (sõltumata tegeliku parameetri $q$ väärtustest).

Probleem: näita või lükka kontranäitega ümber, et selline $\sigma_o$ alati leidub (ei sõltu tegelikust parameetrist $q$). Kas on võimalik leida $\sigma_o$ analüütiline kuju? Kas selline fenomen on ka teiste mudelite ($m > 2$, dipersioonid erinevad, teised jaotused jne) korral?

Bayesi lähenemise korral ei saa jada $\hat{y}_1, \ldots, \hat{y}_n$ ei saa enam analüütiliselt leida. Üliõpilane uurib meetode selle probleemi lahendamiseks, näiteks iteratiivseid algoritme artiklist [8].

# References


[1] Lember, J. Bayesi talumatu kergus *EMS Aastaraamat 2020-21*, 2025

[2] Lember, J. Tehisõpe (loengkonspekt), 2025.

[3] Bilmes, J. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models *CTIT technical reports series*, 1998

[4] Yakowitz, S. and Spragins, J. On the identifiability of finite mixtures, *The Annals of Mathematical Statistics*, pp. 209–214, 1968

[5] Teicher, H. Identifiability of mixtures of product measures *The Annals of Mathematical Statistics* Vol 38 (4)4, pp. 1300-1302, 1967

[6] Lember, J., Kolodenko, A. Adjusted Viterbi training: proof of concept *Probability in the Engineering and Informational Sciences* vol 21(3), pp- 451–475, 2007

[7] Lember, J. Introduction to Bayesian modelling, loengukonspekt, 2023.

[8] Lember,J., Gasbarra, D., Koloydenko, A., Kuljus, K. Estimation of Viterbi path in Bayesian hidden Markov models *Metron*, pp. 137–169, 2019.