

# Ülevaade EM-algoritmist

Jelena Zaitseva  
jellen@ut.ee

Andmekaevandamise uurimisseminar MTAT.03.169.  
Arvutiteaduse instituut, Tartu Ülikool  
Detsember 2003, lk. 148–160

## **Kokkuvõte**

EM-algoritm on üldine meetod leidmaks tõenäosusjaotuse parameetrite suurima tõepära hinnaguid etteantud ebatäieliku või puuduvate väärtustega andmehulga põhjal.

Artikli eesmärgiks on anda ülevaade EM-algoritmi tööpõhimõtetest. Ära on toodud olulisemad definitsioonid ja meetodi kirjeldus neist lähtuvalt. Lisaks on algoritmi tööd kirjeldatud lihtsate näidete abil.

## 1. Sissejuhatus

Oletame, et meil on olemas andmed mingitest parameetritest sõltuva tõenäosusjaotusega kirjeldatavast üldkogumist. Tundmatute jaotusparameetrite hindamine reaalse andmete alusel võib olla vägagi keerukas ja töömahukas protsess. Üks võimalikke variante on leida parameetrite suurima tõepära hinnangud EM-algoritmi abil <sup>1</sup>.

EM meetodit võib vaadelda kui üldist iteratiivset optimeerimise algoritmi tõepärafunktsiooni maksimumi leidmiseks etteantud puuduvate andmetega tõenäosusliku mudeli põhjal.

Algoritmi sissejuhatava sammuna antakse ette parameetrite algühendid. Nendele vastavalt leitakse esimese, nõ E-sammuna ( $E = \textit{expectation}$ ), oodatav lisainformatsioon. Viimase alusel arvutatakse hinnatavate parameetrite suurima tõepära hinnangud (nn M-samm,  $M = \textit{maximization}$ ). Saadud hinnangutega pöördutakse tagasi E-sammu juurde ja nii edasi.

Käesolev ülevaade on üles ehitatud järgmiselt: osa 2 sisaldab suurima tõepära meetodi definitsiooni; algoritmi olemus on kirjeldatud 3. osas; 4. osa seletab EM-algoritmi mündi viskamise näitel; osa 5. tutvustab EM-algoritmi kasutamist segumudelite parameetrite hindamisel; algoritmi nõrgad küljed on esitatud osas 6; osa 7 on ülevaate kokkuvõtteks.

## 2. Suurima tõepära meetod

Suurima tõepära meetod (*maximum-likelihood*, ML) on laialt kasutatav meetod parameetrite hindamiseks. Olgu antud

- mudel  $X = \{X_1, \dots, X_n\}$ , kus iga  $X_i$  on juhuslik muutuja (üksik väärtus või väärtuste vektor),

---

<sup>1</sup>Termin EM esitatakse esmakordselt artiklis (Dempster, Laird, & Rubin 1977), kus on ära toodud tõestused algoritmi käitumise põhiliste tulemuste kohta (muu hulgas ka tulemus, et logaritmilise tõepärafunktsiooni väärtus iteratsiooniprotsessi käigus ei kahane) ning suur arv algoritmi rakendusi.

Tegelikult tekkis EM-algoritmi idee palju varem, nimetatud artikkel kujutab enesest lihtsalt varasemate tööde tulemuste üldistamist ja täiendamist. Selliste varasemate tööde autorite hulgas on Baum jt (Baum *et al.* 1970), kes kirjutasid väga kompaktselt ning Dempsteriga samadel põhimõtetel baseeruva artikli, kuid nende artikkel ei saanud nii tuntuks, nagu Dempsteri oma seitse aastat hiljem. Suurt tähelepanu väärrib ka Sundbergi töö (Sundberg 1972). Sundberg formuleeris EM-algoritmi alused baseeruvana eksponentsiaalsete jaotuste perest pärinevaid andmetel ja illustreeris oma tulemusi mitmete näidetega. Tema töös aga puudus selgesõnaline tulemus tõepärafunktsioonidel põhineva lähenemise monotoonsuse kohta, mis oli tõestatud näiteks Baumi artiklis (Meng & van Dyk 1997).

- parameetrite vektor  $\theta$ , mille abil on võimalik defineerida andmete *tõepärafunktsioon*  $P(X | \theta)$ . On võimalik ka defineerida *logaritmiline tõepära*

$$\mathcal{L}(X | \theta) = \ln P(X | \theta)$$

Kuna tavaliselt  $X_i$ -d on sõlumatud ja sama jaotusega, võime kirjutada

$$\mathcal{L}(X | \theta) = \sum_{i=1}^n \ln P(X_i | \theta)$$

Kui  $\Omega$  on parameeterruum, siis parameetrite  $\theta$  suurima tõepära hinnang  $\theta_{ML}$  on defineeritud seosega:

$$\theta_{ML} = \max_{\theta \in \Omega} \mathcal{L}(X | \theta)$$

**Näide** Oletame, et viskame münti 6 korda, ning  $X_i = 1$ , kui saame kulli, ja  $X_i = 0$ , kui tulemuseks on kiri. Olgu meie katse tulemus  $x = \{1, 0, 0, 0, 1, 0\}$ . Ja oletame, et tõenäosus saada juhuslikul mündiviskel tulemuseks kull on  $p$  ning tõenäosus saada kiri on  $1 - p$ ; seega,  $\theta = \{p\}$ . Siis

$$\begin{aligned} \mathcal{L}(X = x | \theta) &= \sum_{i=1}^n \ln P(X_i = x_i | p) \\ &= 2 \ln p + 4 \ln(1 - p) \end{aligned}$$

Logaritmilist tõepärafunktsiooni  $\mathcal{L}(\theta)$  maksimiseeriva parameetri  $\theta$  väärtuse leidmiseks võrdsustame  $\mathcal{L}(\theta)$  parameetri  $\theta$  järgi leitud tuletise 0-ga:

$$\frac{\mathcal{L}(X = x | \theta)}{dp} = \frac{2}{p} - \frac{4}{1 - p} = 0$$

Lahendades viimase võrduse  $p$  suhtes, saame  $p = \frac{1}{3}$ , mis on intuiitivseks hinnanguks  $p$  jaoks ehk kullide proportsiooniks meie näites.

### 3. EM-algoritm

Selle peatüki sisu põhineb M.Beali (Beal 2003) tööil.

## Probleemi püstitus

EM-algoritm on üsna tähelepanuväärne algoritm puuduvate andmetega ülesannete lahendamiseks tõepära kontekstis.

Vaatleme mudelit varjatud (vaatlusandmed puuduvad) muutujatega  $X$  ning teadaolevate muutujatega  $Y$ . Parameetrite vektor, mis kirjeldab muutujate vahelisi (potentsiaalseid) stohhastilisi sõltuvusi, on esitatud  $\theta$  abil.

Olgu antud mudel, mille alusel on produtseeritud  $n$  sõltumatust ja sama jaotustega elemendist koosnev andmehulk  $Y = \{Y_1, \dots, Y_n\}$  ja nn varjatud muutujate hulk  $X = \{X_1, \dots, X_n\}$ . Kirjeldagu meie mudeli tõepärafunktsioon:

$$P(Y | \theta) = \prod_{i=1}^n P(Y_i | \theta) = \prod_{i=1}^n \int P(X_i, Y_i | \theta) dX_i$$

Integreerimine üle varjatud muutujate  $X_i$  on vajalik moodustamiseks üksnes vaadeldud andmetest  $Y_i$  sõltuvat tõepärafunktsiooni. Seejuures eeldame üldisust kitsendamata, et varjatud muutujad on pidevad.

Suurima tõepära meetodi rakendamisel püüatakse leida selliseid parameetreid  $\theta_{ML}$ , mille korral tõepärafunktsiooni väärtus on maksimaalne. Piisab, kui vaatleme logaritmilist tõepärafunktsiooni, sest see saavutab maksimumi samal parameetri väärtusel.

$$\mathcal{L}(\theta) = \ln P(Y | \theta) = \ln \sum_{i=1}^n P(Y_i | \theta) = \sum_{i=1}^n \ln \int P(X_i, Y_i | \theta) dX_i$$

Lihtsama üleskirjutuse huvides märgime logaritmilist tõepära  $\mathcal{L}$  kui ainult parameetritest  $\theta$ -st sõltuva funktsiooni (sõltuvus  $Y$ -st on siin kaasa arvatud).

Meie ülesanne on kahekordselt raske – nii parameetrid  $\theta$  kui ka varjatud andmed  $X$  on teadmata.  $\mathcal{L}(\theta)$  maksimiseerimise probleem  $\theta$  järgi lihtsustub tuues sisse varjatud muutujate tõenäosusjaotuse. Iga juhuslike suuruste  $X_i$  tõenäosusjaotuse fikseerimine toob kaasa  $\mathcal{L}(\theta)$  võimalike väärtuste alumise piiri tõusu. Seades igale elemendile  $Y_i$  vastavusse tõenäosusjaotuse  $q_{X_i}(X_i)$ , saame logaritmilise tõepärafunktsiooni esitada kujul:

$$\begin{aligned}
\mathcal{L}(\theta) &= \sum_i \ln \int P(X_i, Y_i | \theta) dX_i \\
&= \sum_i \ln \int q_{X_i}(X_i) \frac{P(X_i, Y_i | \theta)}{q_{X_i}(X_i)} dX_i \\
&\geq \sum_i \int q_{X_i}(X_i) \ln \frac{P(X_i, Y_i | \theta)}{q_{X_i}(X_i)} dX_i \tag{1} \\
&= \sum_i \left( \int q_{X_i}(X_i) \ln P(X_i, Y_i | \theta) dX_i - \int q_{X_i}(X_i) \ln q_{X_i}(X_i) dX_i \right) \\
&= \mathcal{F}(q_{X_1}(X_1), \dots, q_{X_n}(X_n), \theta) = \mathcal{F}(q_{X_i}(X_i), \theta)
\end{aligned}$$

Võrratus toodud võrduste ahelas on tingitud logaritmfunksiooni kume-  
rusest (Jenseni võrratus<sup>2</sup>). Märgistus  $q_X(X)$  tähistab hulka  $\{q_{X_i}(X_i)\}_{i=1}^n$ . Et-  
teantud tõenäosusjaotuste  $q_{X_i}(X_i)$  ja parameetri  $\theta$  funktsioon  $\mathcal{F}(q_{X_i}(X_i), \theta)$   
on  $\mathcal{L}(\theta)$  alumiseks piiriks.

## EM-algoritmi olemus

EM algorim koosneb kahest sammust. E-sammul leitakse varjatud muutuja-  
te aposterioorne tõenäosusjaotus fikseeritud parameetrite  $\theta$  (ja andmete  $Y$ )  
korral maksimeerides  $\mathcal{F}(q_X(X), \theta)$  iga  $q_{X_i}(X_i)$  suhtes. M-sammul maksimee-  
ritakse  $\mathcal{F}(q_X(X), \theta)$  väärtuse  $\theta$  suhtes kasutades eelneval E-sammul hinnatud  
tõenäosusjaotusi  $q_{X_i}(X_i)$ . Kasutades ülaindeksit ( $t$ ) iteratsiooni numbriga tähis-  
tamiseks (ning alustades suvaliste algparameetritega  $\theta^{(0)}$ ), saab ülaltoodud  
mõttekäigu esitada kujul:

$$\mathbf{E \text{ samm:}} \quad q_{X_i}^{(t+1)} \leftarrow \max_{q_{X_i}} \mathcal{F}(q_X(X), \theta^{(t)}), \forall i \in \{1, \dots, n\}$$

$$\mathbf{M \text{ samm:}} \quad \theta^{(t+1)} \leftarrow \max_{\theta} \mathcal{F}(q_X^{(t+1)}(X), \theta)$$

---

<sup>2</sup>(Jensen's Inequality, 1906): olgu  $f$  pidev reaalkäitumistega funktsioon intervallil  $I$ . Kui  $x_1, x_2, \dots, x_n \in I$ , siis

- $\frac{\sum f(x_i)}{n} \leq f\left(\frac{\sum x_i}{n}\right)$  kui  $f$  on nõgus,
- $\frac{\sum f(x_i)}{n} \geq f\left(\frac{\sum x_i}{n}\right)$  kui  $f$  on kumer,
- $\frac{\sum f(x_i)}{n} = f\left(\frac{\sum x_i}{n}\right)$  siis ja ainult siis, kui  $x_1 = x_2 = \dots = x_n$ .

Osutub, et E-sammul leitava piiri (1) maksimum üle  $q_{X_i}(X_i)$  avaldub võrdusena:

$$q_{X_i}^{(t+1)} = P(X_i | Y_i, \theta^{(t)}), \forall i, \quad (2)$$

mille puhul võrratus (1) asendub võrdusega. Seda saab tõestada valemi (2) otsese asendamisega võrdusse (1):

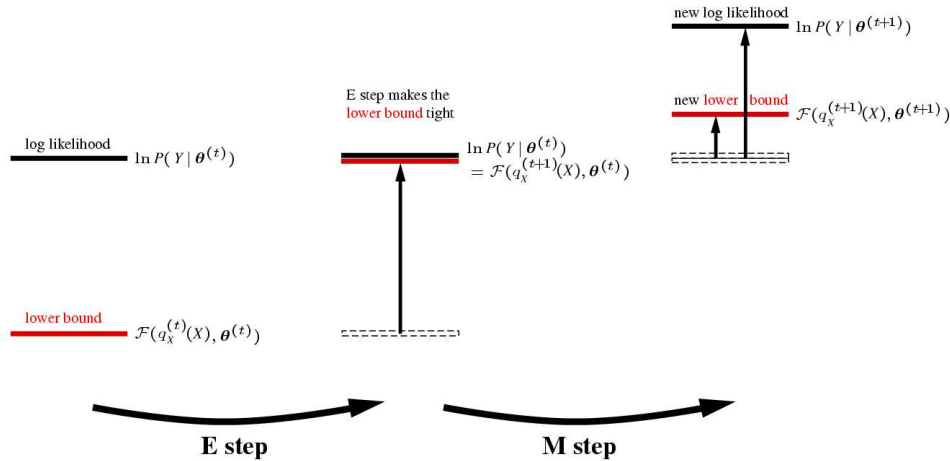
$$\begin{aligned} \mathcal{F}(q_X^{(t+1)}(X), \theta^{(t)}) &= \sum_i \int q_{X_i}^{(t+1)}(X_i) \ln \frac{P(X_i, Y_i | \theta^{(t)})}{q_{X_i}^{(t+1)}(X_i)} dX_i \\ &= \sum_i \int P(X_i | Y_i, \theta^{(t)}) \ln \frac{P(X_i, Y_i | \theta^{(t)})}{P(X_i | Y_i, \theta^{(t)})} dX_i \\ &= \sum_i \int P(X_i | Y_i, \theta^{(t)}) \ln \frac{P(Y_i | \theta^{(t)}) P(X_i | Y_i, \theta^{(t)})}{P(X_i | Y_i, \theta^{(t)})} dX_i \\ &= \sum_i \int P(X_i | Y_i, \theta^{(t)}) \ln P(Y_i | \theta^{(t)}) dX_i \\ &= \sum_i \ln P(Y_i | \theta^{(t)}) \int P(X_i | Y_i, \theta^{(t)}) dX_i \\ &= \sum_i \ln P(Y_i | \theta^{(t)}) = \mathcal{L}(\theta^{(t)}) \end{aligned}$$

kus viimase rea saame tänu sellele, et  $\ln P(Y_i | \theta)$  ei ole  $X$ -i funktsioon. Seega ei saa tõepärafunktsiooni väärtus EM-algoritmi rakendamise käigus väheneda.

M sammul leitakse maksimum avaldise (1) osatuletise (parameetrite  $\theta$  järgi) nulliga võrdsustamise teel, kuna  $q_X$  ei sõltu  $\theta$ -st.

$$\mathbf{M \ samm:} \theta^{(t+1)} \leftarrow \max_{\theta} \sum_i \int P(X_i | Y_i, \theta^{(t)}) \ln P(X_i, Y_i | \theta) dX_i$$

EM-algortm on skemaatiliselt esitatud joonisel 1.



Joonis 1: EM-algoritmi skemaatiline esitlus

## 4. EM-algoritmi töö demonstreerimine mündi viskamise näitel

Näide on võtnud M.Collinsi (Collins 1997) tööst.

### Algandmed

Vaatleme mündi viskamise katseseeriat. Olgu inimesel taskus kaks münti, olgu kulli saamise tõenäosus esimese mündi korral  $p_1$  ja teise mündi korral  $p_2$ . Mingil hetkel valitakse esimene münt tõenäosusega  $\lambda$  või teine münt tõenäosusega  $1 - \lambda$  ning valitud münti visatakse kolm korda. Saadakse mündiviske tulemuste kolmikute jada  $Y = (\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle)$  ( $H$  – kull,  $T$  – kiri). Juhul, kui me võiksime vaadata täielikke andmeid  $X$ , siis me näeksime ka münti, mis oli valitud igal sammul:  $X = (\langle HHH, 1 \rangle, \langle TTT, 2 \rangle, \langle HHH, 1 \rangle, \langle TTT, 2 \rangle)$ .

### Analüütiline osa

Oletame, et  $X$  on varjatud. Siis EM sammud on järgmised:

- E samm. Defineerime  $\tilde{p}_i = P(X_i = \langle Y_i, 1 \rangle | \theta)$  – tõenäosus, et vaadeldavate andmete ja antud parameetrite väärtuste korral valiti esimene münt. Kui  $P_c(Y_i | p)$  on katse tulemuse  $Y_i$  tõenäosus ilmutamise tõenäosusega  $p$ , siis

$$\begin{aligned}
P(Y_i | \theta) &= \lambda P_c(Y_i | p_1) + (1 - \lambda) P_c(Y_i | p_2), \\
q_{X_i}(X_i) = \tilde{p}_i &= \frac{\lambda P_c(Y_i | p_1)}{P(Y_i | \theta)} \\
&= \frac{\lambda P_c(Y_i | p_1)}{\lambda P_c(Y_i | p_1) + (1 - \lambda) P_c(Y_i | p_2)}
\end{aligned}$$

Tähistagu  $\tilde{p}_i$  aposterioorset tõenäosust, et  $i$ -s katsetulemus on saadud esimese mündi viskamise tulemusel. Defineerime  $H_i$  kui kullide arvu juhuslikus suuruses  $Y_i$ , siis

$$P_c(Y_i | p) = \binom{3}{H_i} p^{H_i} (1 - p)^{3 - H_i}$$

Kuna binomiaalkoefitsient ei mõjuta tõepära maksimiseerivaid parameetreid  $\theta' = \{\lambda', p'_1, p'_2\}$ , kasutame edaspidi avaldist  $p^{H_i} (1 - p)^{3 - H_i}$ . Katsetulemused on sõltumatud ja sama jaotusega, seega võime kirjutada:

$$\begin{aligned}
\mathcal{L}(\theta) &= \sum_i (\tilde{p}_i \ln \lambda' P_c(Y_i | p'_1) + (1 - \tilde{p}_i) \ln(1 - \lambda') P_c(Y_i | p'_2)) \\
&= \sum_i (\tilde{p}_i \ln \lambda' p_1'^{H_i} (1 - p'_1)^{3 - H_i} + (1 - \tilde{p}_i) \ln(1 - \lambda') p_2'^{H_i} (1 - p'_2)^{3 - H_i}) \\
&= \sum_i (\tilde{p}_i \ln \lambda' + (1 - \tilde{p}_i) \ln(1 - \lambda') + \tilde{p}_i \ln p_1'^{H_i} (1 - p'_1)^{3 - H_i} + \\
&\quad (1 - \tilde{p}_i) \ln p_2'^{H_i} (1 - p'_2)^{3 - H_i}) \\
&= \sum_i (\tilde{p}_i \ln \lambda' + (1 - \tilde{p}_i) \ln(1 - \lambda') + \tilde{p}_i (H_i \ln p'_1 + \\
&\quad (3 - H_i) \ln(1 - p'_1)) + (1 - \tilde{p}_i) (H_i \ln p'_2 + (3 - H_i) \ln(1 - p'_2))) \\
&= \sum_i (\tilde{p}_i (\ln \lambda' + H_i \ln p'_1 + (3 - H_i) \ln(1 - p'_1)) + \\
&\quad (1 - \tilde{p}_i) (\ln(1 - \lambda') + H_i \ln p'_2 + (3 - H_i) \ln(1 - p'_2)))
\end{aligned}$$

- M samm. Kuna meil on väga lihtne juht, võime analüütiliselt leida tõepärafunktsiooni väärtust maksimeerivad parameetrite hinnangud. Funktsiooni maksimiseerimine (võrdsustades vastavalt  $\lambda'$ ,  $p'_1$ ,  $p'_2$  järgi leitud osatuletised nulliga) annab järgmiseid valemid:

$$\lambda' = \frac{\sum \tilde{p}_i}{n}$$

$$p'_1 = \frac{\sum \frac{H_i}{3} \tilde{p}_i}{\sum \tilde{p}_i}$$

$$p'_2 = \frac{\sum \frac{H_i}{3} (1 - \tilde{p}_i)}{\sum (1 - \tilde{p}_i)}$$

Saadud valemitel on ilus intuitiivne interpretatsioon:

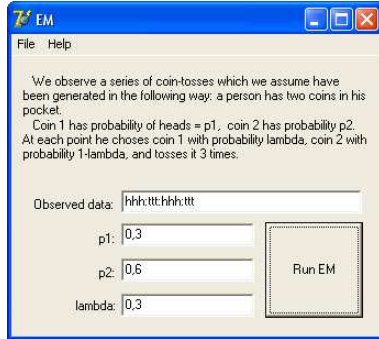
- $\lambda$  on keskmine aposterioorne tõenäosus selleks, et  $i$ -s katsetulemus on saadud esimese mündi viskamisel,
- $p_1$  on tavaliste suurima tõepära hinnangute  $\frac{H_i}{3}$  kaalutud keskmine üle katsetulemuste  $Y_i$ , kus kaal on vastavuses  $\tilde{p}_i$ -ga,
- $p_2$  on kaalutud keskmine üle katsetulemuste, kus kaal on vastavuses  $1 - \tilde{p}_i$ -ga.

## Praktiline osa

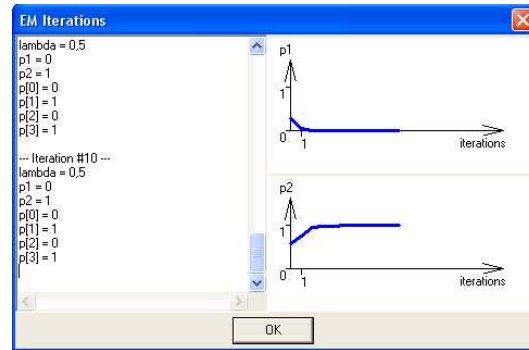
Kasutades ülaltoodud arutelu tulemusi on lihtne kirjutada programm, mis demonstreeriks EM-i tööd antud ülesande jaoks. Ülevaate autori teostus oli järgmine:

- kasutaja käest küsitatakse (joonis 2) andmeid mündiviske katsete tulemuste kohta (*observed data*), kummagi mündi kulli saamise lähtetõenäosusi ( $p_1$ ,  $p_2$ ) ning esimese mündi valiku algtõenäosust (*lambda i.e.*  $\lambda$ );
- peale algoritmi käivitamist, arvutatakse igal iteratsiooni sammul parameetrite hinnangud ning kujutatakse tõenäosuste  $p_1$  ja  $p_2$  hinnangute muutusi graafiliselt (joonis 3). Programmis kasutatakse ainult 10 iteratsiooni sammu. Tabel 1 illustreerib parameetrite väärtusi iga iteratsioonil.

Joonistel 2, 3 ja tabelis 1 toodud andmed lähtuvad algtingimustest  $p_1 = 0, 3$ ,  $p_2 = 0, 6$  ja  $\lambda = 0, 3$ . Nendest nähtub algoritmi koonduvuskiirus parameetrite  $p_1$ ,  $p_2$  ja  $\lambda$  hinnangute leidmisel antud ülesande tingimusel.



Joonis 2: Programmi ekraanivorm parameetrite ekraanivorm parameetrite algväärtuste sisestamiseks



Joonis 3: EM simulatsiooni resultaadid ette antud mündiviske tulemuste  $Y = (\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle)$  korral. Vasakul on parameetrite väärtused, paremal - parameetrite  $p_1$  ja  $p_2$  koonduvuste graafikud

Iteration	$\lambda$	$p_1$	$p_2$	$\tilde{p}_0$	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$
1	0,3	0,3	0,6	0,0508	0,6966	0,0508	0,6966
2	0,3737	0,0680	0,7578	0,0004	0,9714	0,0004	0,9714
3	0,4859	0,0004	0,9722	8,99E-11	0,9999	8,99E-11	0,9999
4	0,4999	8,99E-11	0,9999	7,28E-31	0,9999	7,28E-31	0,9999
5	0,4999	7,28E-31	0,9999	3,86E-91	1	3,86E-91	1
6	0,5	3,86E-91	1	5,7E-272	1	5,7E-272	1
7	0,5	5,7E-272	1	0	1	0	1
8	0,5	0	1	0	1	0	1
9	0,5	0	1	0	1	0	1
10	0,5	0	1	0	1	0	1

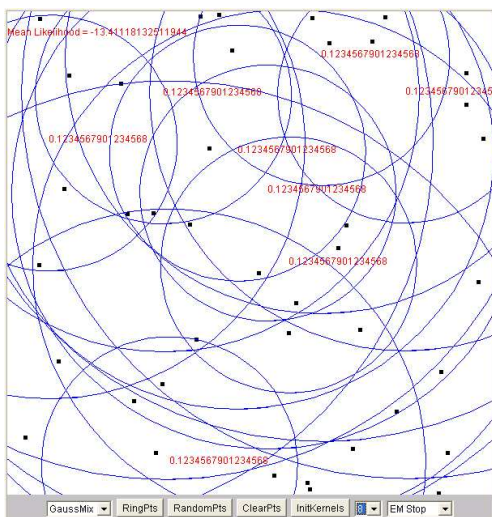
Tabel 1: Parameetrite hinnangud igal EM-algoritmi sammul mündiviske tulemuste  $Y = (\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle)$  korral

## 5. EM segujaotuste parameetrite hindamisel

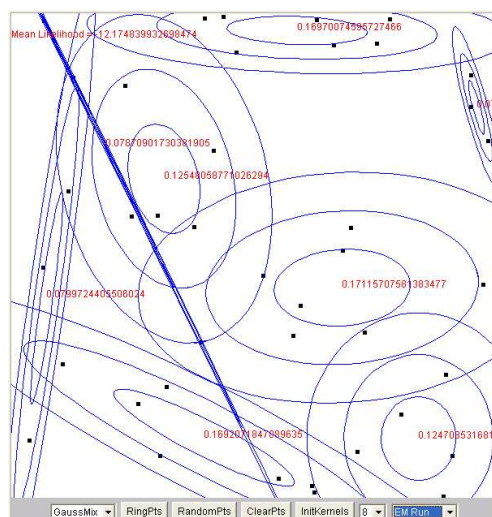
Üheks tuntumaks EM-algoritmi rakenduseks on selle kasutamine jaotuste segude analüüsil (vt. näiteks rakendusi ühemõõtmeliste normaal- ja Poissoni jaotusega mudelite tarvis - (Hand, Mannila, & Smyth 2001)).

Segujaotuste puhul ei tea me ei üksikute segusse kuuluvate jaotuste parameetrid ega ka andmepunktide päritolu (millisesse gruppi, ehk siis millisesse jaotusesse, andmepunkt kuulub). Sellise kaheastmelise probleemi lahendamine on oma olemuselt vastavuses EM-algoritmiga – me saame anda igale andmepunktile algtõenäosuse kuuluda mingisse gruppi (olla pärit mingist jaotusest), hinnata seejärel oma andmete ja etteantud tõenäosuste alusel segusse oletatavalt kuuluvate jaotuste parameetrid ning püüda viimastele tuginedes määrata uuesti vaatluspunktide kuuluvus. Nii jätkates (hinnates kordamööda vaatluspunktide mingisse jaotuse kuulumise tõenäosusi ning nende jaotuste parameetreid) saame viimaks EM hinnangud segujaotuste parameetritele.

Joonistel 4 ja 5 on illustreeritud EM-algoritmi rakendust Gaussi segumudeli korral Akaho ja Micheli loodud apletti abil (Akaho & Michel):



Joonis 4: Algsandmed



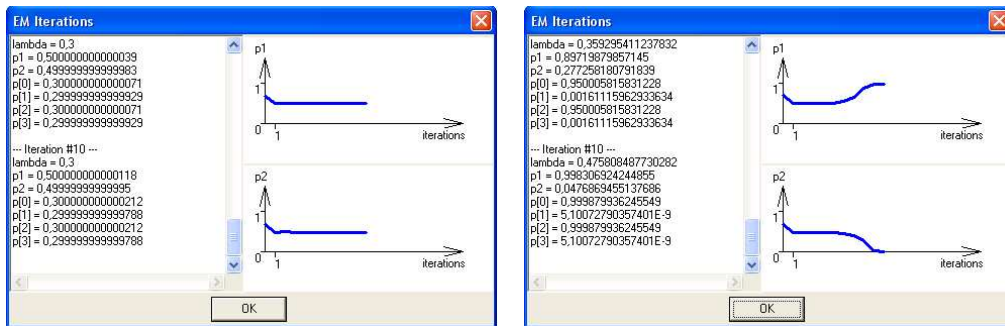
Joonis 5: EM simulatsiooni tulemused

Geneetikas kasutatakse kirjeldatud algoritmi näiteks komplekssses segregatsioonanalüüsis, kus püütakse uuritava tunnuse fenotüübijaotust lähendada erinevatele genotüüpidele vastavate jaotuste seguga (Kaart 2002).

## 6. EM-algoritmi nõrgad küljed

Realse algoritmi rakendamiseks on vaja valida algväärtused (näiteks, valida juhuslikult kas  $q$  või  $\theta$ ) ja meetodi koondumiskriteerium (näiteks, fikseerida ühe väärtuse,  $q$ ,  $\theta$  või  $\mathcal{L}(\theta)$ , jaoks maksimaalne lubatav erinevus kahel järjestikusel iteratsioonisammul leitud hinnangute vahel).

EM meetod on tundlik algtingimuste suhtes (vaata joonist 6), seepärast võivad algtingimuste erinevad valikud viia erinevate lokaalsete maksimumide juurde. Seetõttu tuleks praktikas käivitada EM-algoritm erinevate algväärtuste komplektidega (ning valida pärast suurimale tõepärale vastavad parameetrite väärtused) (Hand, Mannila, & Smyth 2001).



Joonis 6: Sõltuvus algtingimuste väärtustest.

Mündiviske tulemused on

$$Y = (\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle).$$

Vasakul parameetrite kombinatsioon

$$\lambda = 0, 3, p1 = 0, 7, p2 = 0, 7; \text{ paremal}$$

$$\lambda = 0, 3, p1 = 0, 7001, p2 = 0, 7.$$

EM-algoritmi arvutuslik keerukus sõltub lähendamiseks vajalike iteratsioonide arvust ning E- ja M-sammude keerukusest (mis omakorda sõltuvad andmete mudelist).

EM-algoritmi kiirendamiseks on välja pakutud palju meetodeid (Jamshidian & Jennrich 1997), mis on tihti edukad koonduvuse mõttes, aga samas palju keerulisemad kui EM-algoritm ning raskemini lahendatavad. Seetõttu ei ole nad saanud praktikas nii populaarseks.

## 7. Kokkuvõtte

EM-algoritm on lihtsalt realiseeritav vahend suurima tõepära hinnangute leidmiseks ülesannetes, kus andmed on ebatäielikud.

Meetod alustab oma tööd ette antud algparameetritega ning muudab neid iteratiivselt vaadeldavate andmete tõepära suurendamiseks.

Ülevaade tutvustas lugejat algoritmi loomusega, tõestas selle stabiilse koondumise igal iteratsiooni sammul ning tõi lihtsad näited selle töö demonstreerimiseks.

## Viited

Akaho, S., and Michel, O. Gaussian Mixture Model EM algorithm. <http://diwww.epfl.ch/mantra/tutorial/english/gaussian/html/>.

Baum, L. E.; Petrie, T.; Soules, G.; and Weiss, N. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.* 41:164–171.

Beal, M. 2003. *Variational Algorithms for Approximate Bayesian Inference*. Ph.D. Dissertation, Gatsby Computational Neuroscience Unit, University College London.

Collins, M. 1997. The EM Algorithm. <http://www.ai.mit.edu/people/mcollins/papers/>.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39(1):1–38.

Hand, D.; Mannila, H.; and Smyth, P. 2001. *Principles of Data Mining*. Adaptive computation and machine learning. The MIT Press.

Jamshidian, M., and Jennrich, R. I. 1997. Acceleration of the EM Algorithm by using Quasi-Newton Methods. *Journal of the Royal Statistical Society. Series B (Methodological)* 59(3):569–587.

Kaart, T. 2002. Geenitehnoloogia statistilised mudelid. [http://www.eau.ee/~ktanel/kool\\_ja\\_too/geenitehn\\_stat\\_mudelid/](http://www.eau.ee/~ktanel/kool_ja_too/geenitehn_stat_mudelid/).

Meng, X.-L., and van Dyk, D. 1997. The EM Algorithm – An Old Folk-Song Sung to a Fast New Tune. *Journal of the Royal Statistical Society. Series B (Methodological)* 59(3):511–567.

Sundberg, R. 1972. *Maximum likelihood theory and applications for distributions generated when observing a function of an exponential variable*. Ph.D. Dissertation, Institute of Mathematics and Statistics, Stockholm University, Stockholm.