The Expectation-Maximization (EM) Algorithm

Algoritmul EM, fundamentare teoretică: pasul E [şi pasul M]

CMU, 2008 fall, Eric Xing, HW4, pr. 1.1-3

Algoritmul EM (Expectation-Maximization) permite crearea unor modele probabiliste care pe de o parte depind de un set de parametri θ iar pe de altă parte includ pe lângă variabilele obişnuite ("observabile" sau "vizibile") x şi variabile necunoscute ("neobservabile", "ascunse" sau "latente") z.

În general, în astfel de situaţii/modele, nu se poate face în mod direct o estimare a parametrilor modelului (θ), în aşa fel încât să se garanteze atingerea maximului verosimilităţii datelor observabile x.

În schimb, algoritmul EM procedează în manieră iterativă, constituind astfel o modalitate foarte convenabilă de estimare a parametrilor θ.

Definim log-verosimilitatea datelor observabile (x) ca fiind $\log P(x \mid \theta)$, iar log-verosimilitatea datelor *complete* (observabile, x, şi neobservabile, z) ca fiind $\log P(x, z \mid \theta)$.

***Observaţie*: Pe tot parcursul acestui exerciţiu se va considera funcţia $\log$ ca având baza supraunitară, fixată.**

a. Log-verosimilitatea datelor *observabile* (x) se poate exprima în funcţie de datele neobservabile (z), astfel:[a]

$$\ell(\theta) \stackrel{not.}{=} \log P(x \mid \theta) = \log \left(\sum_z P(x, z \mid \theta) \right)$$

În continuare vom nota cu q o funcţie / distribuţie de probabilitate definită peste variabilele ascunse/neobservabile z.

Folosiţi *inegalitatea lui Jensen* pentru a demonstra că are loc următoarea inegalitate:

$$\log P(x \mid \theta) \geq \sum_z q(z) \log \left(\frac{P(x, z \mid \theta)}{q(z)} \right) \tag{1}$$

pentru orice x (fixat), pentru orice valoare a parametrului θ şi pentru orice distribuţie probabilistă q definită peste variabilele neobservabile z.

[a] Reţineţi că x, vectorul de date observabile, este fixat (dat), în vreme ce z, vectorul de date neobservabile, este liber (variabil).

Observaţie (1)

<u>Semnificaţia</u> inegalităţii (1) este următoarea:

Funcţia (de fapt, orice funcţie de forma)

$$F(q, \theta) \stackrel{\textit{def.}}{=} \sum_z q(z) \log \left(\frac{P(x, z \mid \theta)}{q(z)} \right)$$

constituie o <u>margine inferioară</u> pentru funcţia de log-verosimilitate a datelor incomplete / observabile, $\ell(\theta) \stackrel{\textit{not.}}{=} \log P(x \mid \theta)$**.**

Remarcaţi faptul că F **este o funcţie de două variabile, iar prima variabilă nu este de tip numeric (cum este** θ**), ci este de tip funcţional.**

Mai mult, se observă că expresia funcţiei F **este de fapt o <u>medie</u>,**

$$E_{q(z)} \left[\log \frac{P(x, z \mid \theta)}{q(z)} \right],$$

atunci când x**,** q **şi parametrul** θ **se consideră fixaţi, iar** z **este lăsat să varieze.**

Soluţie

Inegalitatea lui Jensen, în contextul teoriei probabilităţilor:

considerând X este o variabilă aleatoare (unară),
dacă φ este o funcţie (reală) convexă, atunci $\varphi(E[X]) \leq E[\varphi(X)]$;

dacă φ este funcţie concavă, atunci $\varphi(E[X]) \geq E[\varphi(X)]$.

Aici vom folosi funcţia $\log$ cu bază supraunitară (funcţie concavă), deci aplicând inegalitatea lui Jensen vom obţine: $\log(E[X]) \geq E[\log(X)]$.

Log-verosimilitatea datelor observabile este:

$$
\begin{aligned}
\ell(\theta) \stackrel{not.}{=} \log P(x \mid \theta) &= \log\left(\sum_z P(x,z \mid \theta)\right) = \log\left(\sum_z q(z)\frac{P(x,z \mid \theta)}{q(z)}\right) \\
&\stackrel{def.}{=} \log\left(E_{q(z)}\left[\frac{P(x,z \mid \theta)}{q(z)}\right]\right)
\end{aligned}
$$

Conform inegalităţii lui Jensen (înlocuind X de mai sus cu $\dfrac{P(x,z \mid \theta)}{q(z)}$), rezultă:

$$
\ell(\theta) \stackrel{not.}{=} \log P(x \mid \theta) \geq E_{q(z)}\left[\log\frac{P(x,z \mid \theta)}{q(z)}\right] \stackrel{def.}{=} \sum_z q(z)\log\frac{P(x,z \mid \theta)}{q(z)},
$$

Notaţie

În continuare, pentru a vă aduce mereu aminte că distribuţia q se referă la datele neobservabile z, vom folosi notaţia $q(z)$ în loc de q.

În consecinţă, în cele ce urmează, în funcţie de context, $q(z)$ va desemna fie distribuţia q, fie valoarea acestei distribuţii pentru o valoare oarecare [a variabilei neobservabile] z.

(Este adevărat că această lejeră ambiguitate poate induce în eroare cititorul neexperimentat.)

b. Vă reamintim definiţia entropiei relative (numită şi divergenţa Kullback-Leibler):

$$KL(q(z) \,||\, P(z \mid x, \theta)) = -\sum_z q(z) \log \left(\frac{P(z \mid x, \theta)}{q(z)} \right)$$

Arătaţi că

$$\log P(x \mid \theta) = F(q(z), \theta) + KL(q(z) \,||\, P(z \mid x, \theta)).$$

Observaţie (2):

Semnificaţia egalităţii care trebuie demonstrată la acest punct este foarte interesantă: diferenţa dintre funcţia obiectiv $\ell(\theta) \stackrel{not.}{=} \log P(x \mid \theta)$ şi marginea sa inferioară $F(q(z), \theta)$ — a se vedea punctul a — este $KL(q(z) \,||\, P(z \mid x, \theta))$. Tocmai pe această chestiune se va "construi" punctul final, şi cel mai important, al problemei noastre.

Observaţie (3)

Ideile de bază ale algoritmului EM sunt două:

1. În loc să calculeze maximul funcţiei de log-verosimilitate $\log P(x \mid \theta)$ în raport cu θ, algoritmul EM va maximiza marginea sa inferioară, $F(q(z), \theta)$, în raport cu ambele argumente, $q(z)$ şi θ.

2. Pentru a căuta maximul (de fapt, un maxim local al) marginii inferioare $F(q(z), \theta)$, algoritmul EM aplică metoda *creşterii pe coordonate* (engl., coordinate ascent): după ce iniţial se fixează $\theta^{(0)}$ eventual aleatoriu, se maximizează *iterativ* funcţia $F(q(z), \theta)$, în mod *alternativ*: mai întâi în raport cu distribuţia $q(z)$ şi apoi în raport cu parametrul θ.

Pasul E: $\quad q^{(t)}(z) = \underset{q(z)}{\operatorname{argmax}} F(q(z), \theta^{(t)})$

Pasul M: $\quad \theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}} F(q^{(t)}(z), \theta)$

Soluţie

$$
\begin{aligned}
F(q(z),\theta) &\stackrel{def.}{=} \sum_z q(z)\log\left(\frac{P(x,z\mid\theta)}{q(z)}\right)\\
&= \sum_z q(z)\log\left(\frac{P(z\mid x,\theta)\cdot P(x\mid\theta)}{q(z)}\right)\\
&= \sum_z q(z)\left[\log\frac{P(z\mid x,\theta)}{q(z)} + \log P(x\mid\theta)\right]\\
&= \sum_z q(z)\log\left(\frac{P(z\mid x,\theta)}{q(z)}\right) + \sum_z q(z)\log P(x\mid\theta)\\
&= -KL(q(z)\mid\mid P(z\mid x,\theta)) + \log P(x\mid\theta)\cdot\underbrace{\sum_z q(z)}_{=1}
\end{aligned}
$$

$$
\Rightarrow \qquad \log P(x\mid\theta) = F(q(z),\theta) + KL(q(z)\mid\mid P(z\mid x,\theta)).
$$

Observaţie (4)

Conform proprietăţii $KL(p \parallel q) \geq 0$ **pentru** $\forall p, q$**, rezultă** $KL(q(z) \parallel P(z \mid x, \theta)) \geq 0$**.**

Aşadar, din egalitatea care tocmai a fost demonstrată la punctul b **obţinem (din nou!, după rezultatul de la punctul** a**) că** $F(q(z), \theta)$ **este o margine inferioară pentru log-verosimilitatea datelor observabile,** $\ell(\theta) \stackrel{not.}{=} \log P(x \mid \theta)$**.**

c. Fie $\theta^{(t)}$ valoarea obţinută pentru parametrul / parametrii θ la iteraţia t a algoritmului EM. Considerând această valoare fixată, arătaţi că maximul lui F în raport cu argumentul / distribuţia $q(z)$ este atins pentru distribuţia $P(z \mid x, \theta^{(t)})$, iar valoarea maximului este:

$$\max_{q(z)} F(q(z), \theta^{(t)}) = E_{P(z|x,\theta^{(t)})}[\log P(x, z \mid \theta^{(t)})] + H(P(z \mid x, \theta^{(t)}))$$

Soluţie

Trebuie să maximizăm $F(q(z), \theta^{(t)})$ — marginea inferioară a log-verosimilităţii datelor observabile x — în raport cu distribuţia $q(z)$.

Pe de o parte, rezultatul de la punctul a ne spune că $F(q(z), \theta) \leq \log P(x \mid \theta)$, pentru orice valoare a lui θ; în particular, pentru $\theta^{(t)}$ avem

$$\log P(x \mid \theta^{(t)}) \geq F(q(z), \theta^{(t)})$$

Pe de altă parte, dacă în egalitatea demonstrată la punctul b se înlocuieşte θ cu $\theta^{(t)}$, rezultă:

$$\log P(x \mid \theta^{(t)}) = F(q(z), \theta^{(t)}) + KL(q(z) || P(z \mid x, \theta^{(t)}))$$

În fine, dacă alegem $q(z) = P(z \mid x, \theta^{(t)})$, atunci termenul $KL(q(z) || P(z \mid x, \theta^{(t)}))$ din dreapta egalităţii de mai sus devine zero (vezi exerciţiul [PS-31] de la capitolul de Probabilităţi şi statistică).

Aşadar, valoarea $\max_{q(z)} F(q(z), \theta^{(t)})$ se obţine pentru distribuţia $q(z) = P(z \mid x, \theta^{(t)})$.

Acum vom calcula această valoare maximă:

$$
\begin{aligned}
\max_{q(z)} F(q(z), \theta^{(t)}) &= \log P(x \mid \theta^{(t)}) \stackrel{def.\ F}{=} \sum_z P(z \mid x, \theta^{(t)}) \log \left(\frac{P(x, z \mid \theta^{(t)})}{P(z \mid x, \theta^{(t)})} \right) \\
&= E_{P(z|x,\theta^{(t)})} \left[\log \frac{P(x, z \mid \theta^{(t)})}{P(z \mid x, \theta^{(t)})} \right] \\
&= E_{P(z|x,\theta^{(t)})} \left[\log P(x, z|\theta^{(t)}) - \log P(z|x, \theta^{(t)}) \right] \\
&= E_{P(z|x,\theta^{(t)})} \left[\log P(x, z|\theta^{(t)}) \right] - E_{P(z|x,\theta^{(t)})} \left[\log P(z|x, \theta^{(t)}) \right] \\
&= E_{P(z|x,\theta^{(t)})} \left[\log P(x, z|\theta^{(t)}) \right] + H[P(z|x, \theta^{(t)})] \\
&= Q(\theta^{(t)} \mid \theta^{(t)}) + H[P(z \mid x, \theta^{(t)})]
\end{aligned}
$$

unde $Q(\theta \mid \theta^{(t)}) \stackrel{not.}{=} E_{P(z|x,\theta^{(t)})}[\log P(x, z \mid \theta)]$**.**

Observaţie (5)

Notând $G_t(\theta) \stackrel{def.}{=} F(P(z \mid x, \theta^{(t)}), \theta)$**, din calculul de mai sus rezultă că** $G_t(\theta^{(t)}) = \log P(x \mid \theta^{(t)}) = Q(\theta^{(t)} \mid \theta^{(t)}) + H(P(z \mid x, \theta^{(t)}))$**. Se poate demonstra uşor — procedând similar cu calculul de mai sus — egalitatea**

$$G_t(\theta) = Q(\theta \mid \theta^{(t)}) + H[P(z \mid x, \theta^{(t)})]$$

Observând că termenul $H[P(z \mid x, \theta^{(t)})]$ **din această ultimă egalitate nu depinde de** θ**, rezultă imediat că**

$$\operatorname*{argmax}_{\theta} G_t(\theta) = \operatorname*{argmax}_{\theta} Q(\theta \mid \theta^{(t)})$$

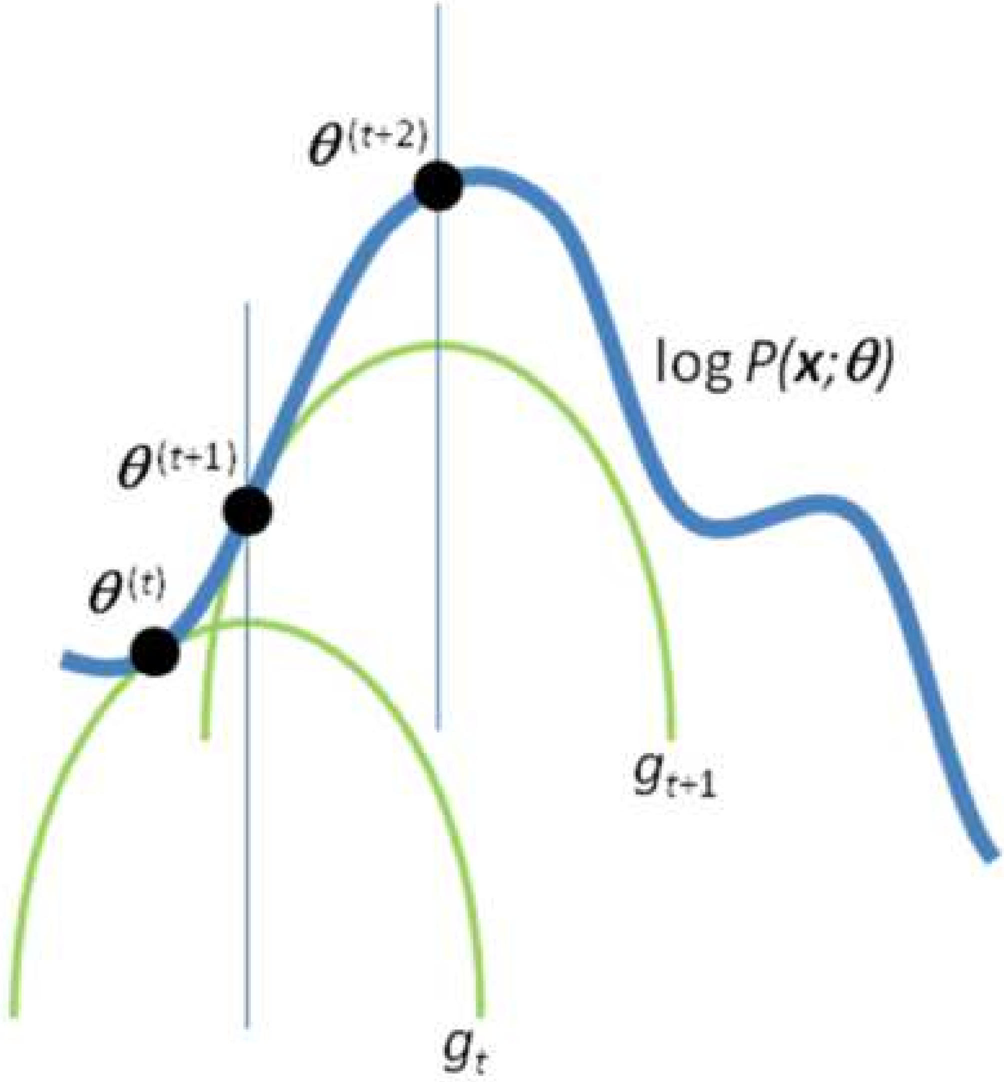
În consecinţă,

$$\theta^{(t+1)} \stackrel{def.}{=} \operatorname*{argmax}_{\theta} F(P(z \mid x, \theta^{(t)}), \theta) = \operatorname*{argmax}_{\theta} G_t(\theta) = \operatorname*{argmax}_{\theta} Q(\theta \mid \theta^{(t)})$$

From:

What is the expectation maximization algorithm?

Chuong B. Do, Serafim Batzoglou,
Nature Biotechnology,
vol. 26, no. 8, 2008, pp. 897-899



Egalitatea precedentă este responsabilă pentru următoarea reformulare (cea uzuală!) a algoritmului EM:

Pasul E′: **calculează** $Q(\theta \mid \theta^{(t)}) = E_{P(z \mid x, \theta^{(t)})}[\log P(x, z \mid \theta)]$

Pasul M′: **calculează** $\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta \mid \theta^{(t)})$

Algoritmul EM:
corectitudine / convergenţă

prelucrare de Liviu Ciortuz, după

en.wikipedia.org/wiki/Expectation-maximization

Pentru a maximiza funcţia de log-verosimilitate a datelor observabile, i.e. $\ell(\theta) \stackrel{\textit{def.}}{=} \log P(x \mid \theta)$, unde baza logaritmului (nespecificată) este considerată supraunitară, algoritmul EM procedează în mod iterativ, optimizând la pasul M al fiecărei iteraţii (t) o funcţie "auxiliară"

$$Q(\theta \mid \theta^{(t)}) \stackrel{\textit{def.}}{=} E_{P(z|x,\theta^{(t)})}[\log P(x, z \mid \theta)],$$

reprezentând media log-verosimilităţii datelor complete (observabile şi neobservabile) în raport cu distribuţia condiţională $P(z \mid x, \theta^{(t)})$.

Vom considera iteraţiile $t = 0, 1, \ldots$ şi $\theta^{(t+1)} = \operatorname{argmax}_\theta Q(\theta \mid \theta^{(t)})$, cu $\theta^{(0)}$ ales în mod arbitrar.

Demonstraţi că pentru orice t fixat (arbitrar) şi pentru orice θ astfel încât $Q(\theta \mid \theta^{(t)}) \geq Q(\theta^{(t)} \mid \theta^{(t)})$ are loc inegalitatea:

$$\log P(x \mid \theta) - \log P(x \mid \theta^{(t)}) \geq Q(\theta \mid \theta^{(t)}) - Q(\theta^{(t)} \mid \theta^{(t)}) \tag{2}$$

Observaţii

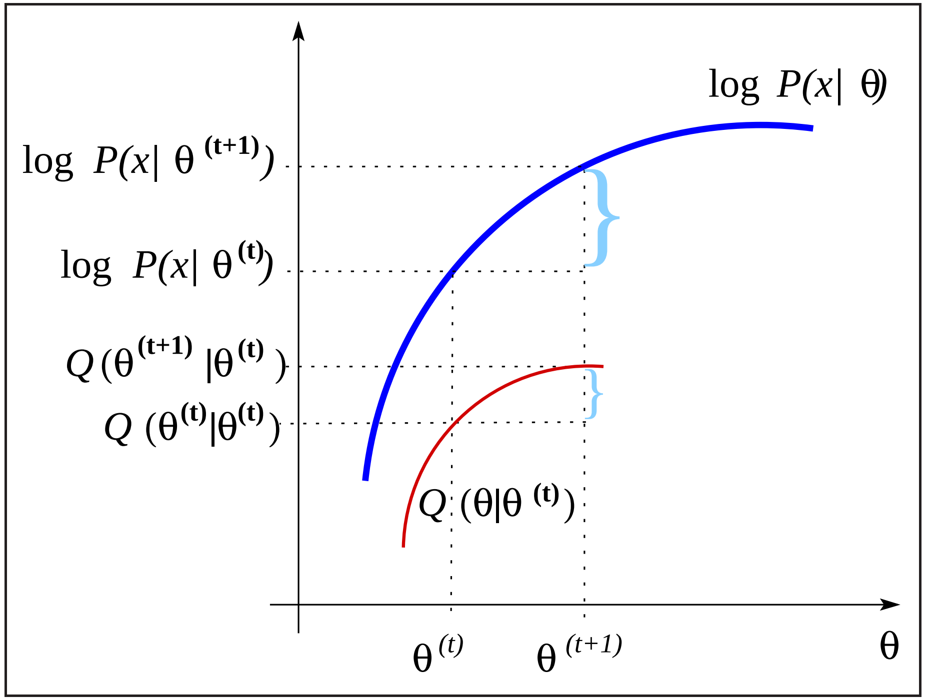
1. Semnificaţia imediată a relaţiei (2):
Orice îmbunătăţire a valorii funcţiei auxiliare $Q(\theta \mid \theta^{(t)})$ conduce la o îmbunătăţire cel puţin la fel de mare a valorii funcţiei obiectiv, $\ell(\theta)$.

2. Dacă în inegalitatea (2) se înlocuieşte θ cu $\theta^{(t+1)} = \text{argmax}_\theta\, Q(\theta \mid \theta^{(t)})$, va rezulta

$$\log P(x \mid \theta^{(t+1)}) \geq \log P(x \mid \theta^{(t)}).$$

În final, vom avea
$\ell(\theta^{(0)}) \leq \ell(\theta^{(t)}) \leq \ell(\theta^{(t+1)}) \leq \ldots$.

Şirul acesta (monoton) este mărginit superior de 0 (vezi definiţia lui ℓ), deci converge la o anumită valoare ℓ^*. În anumite cazuri / condiţii, această valoare este un maxim (în general, local) al funcţiei de log-verosimilitate.



Observaţii (cont.)

3. Conform aceleiaşi inegalităţi (2), la pasul M de la iteraţia t a algoritmului EM, este suficient ca în loc să se ia $\theta^{(t+1)} = \mathrm{argmax}_\theta\, Q(\theta \mid \theta^{(t)})$, să se aleagă $\theta^{(t+1)}$ astfel încât $Q(\theta^{(t+1)} \mid \theta^{(t)}) > Q(\theta^{(t)} \mid \theta^{(t)})$. Aceasta constituie *versiunea "generalizată"* a algoritmului EM.

Demonstraţia relaţiei (2)

$$P(x, z \mid \theta) = P(z \mid x, \theta) \cdot P(x \mid \theta) \Rightarrow \log P(x \mid \theta) = \log P(x, z \mid \theta) - \log P(z \mid x, \theta) \Rightarrow$$

$$\underbrace{\sum_z P(z \mid x, \theta^{(t)})}_{1} \cdot \log P(x \mid \theta) =$$

$$\sum_z P(z \mid x, \theta^{(t)}) \cdot \log P(x, z \mid \theta) - \sum_z P(z \mid x, \theta^{(t)}) \cdot \log P(z \mid x, \theta) \Rightarrow$$

$$\log P(x \mid \theta) = Q(\theta \mid \theta^{(t)}) - \sum_z P(z \mid x, \theta^{(t)}) \cdot \log P(z \mid x, \theta)$$

Ultimul termen din egalitatea aceasta reprezintă o cross-entropie, pe care o vom nota cu $CH(\theta \mid \theta^{(t)})$. Aşadar,

$$\log P(x \mid \theta) = Q(\theta \mid \theta^{(t)}) + CH(\theta \mid \theta^{(t)})$$

Această egalitate este valabilă pentru toate valorile posibile ale parametrului θ. În particular pentru $\theta = \theta^{(t)}$, vom avea:

$$\log P(x \mid \theta^{(t)}) = Q(\theta^{(t)} \mid \theta^{(t)}) + CH(\theta^{(t)} \mid \theta^{(t)})$$

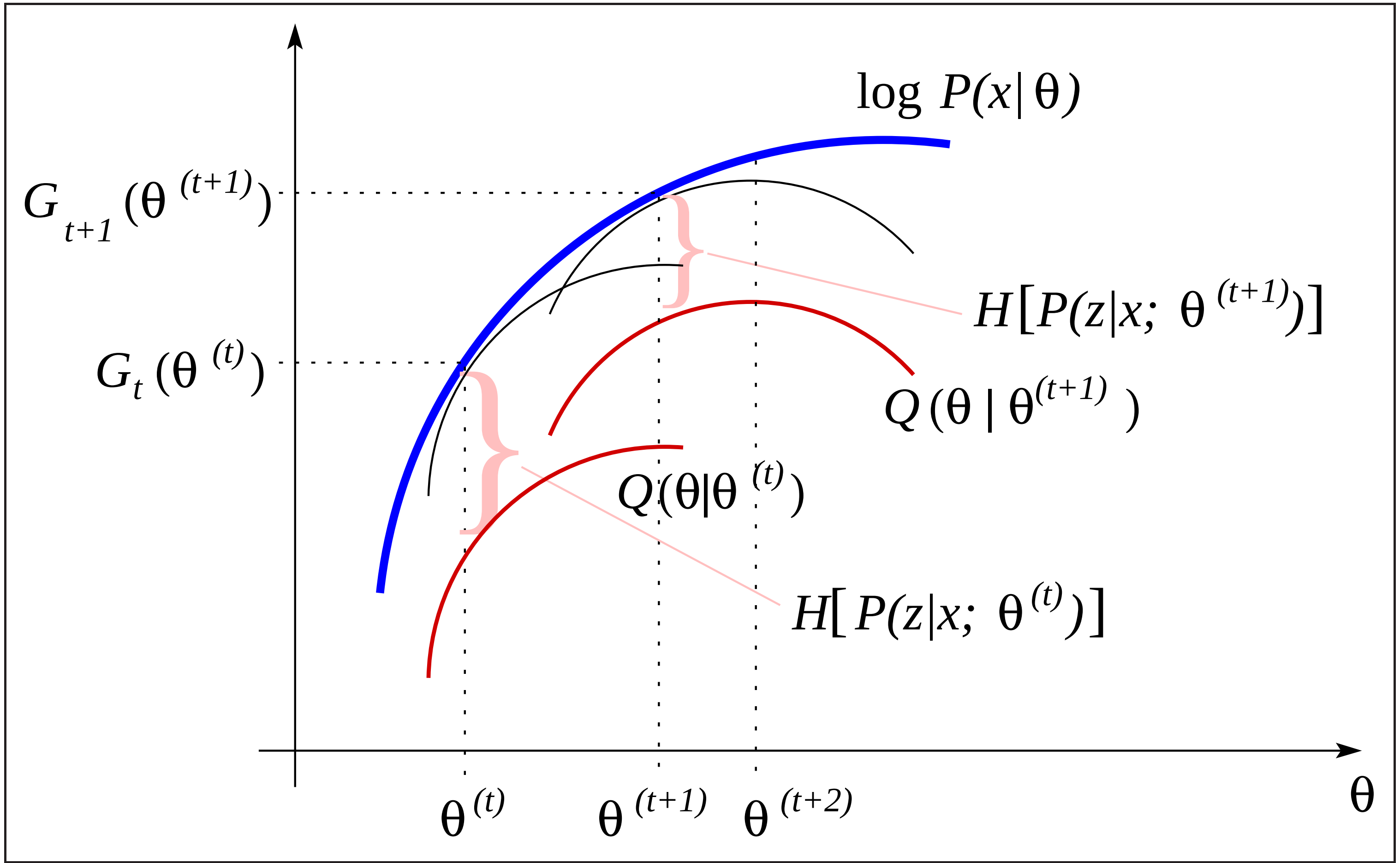
Demonstraţia relaţiei (2), cont.

Scăzând membru cu membru ultimele două egalităţi, obţinem:

$$\log P(x \mid \theta) - \log P(x \mid \theta^{(t)}) = Q(\theta \mid \theta^{(t)}) - Q(\theta^{(t)} \mid \theta^{(t)}) + CH(\theta \mid \theta^{(t)}) - CH(\theta^{(t)} \mid \theta^{(t)})$$

Conform inegalităţii lui Gibbs, avem $CH(\theta \mid \theta^{(t)}) \geq CH(\theta^{(t)} \mid \theta^{(t)})$**, deci în final rezultă:**

$$\log P(x \mid \theta) - \log P(x \mid \theta^{(t)}) \geq Q(\theta \mid \theta^{(t)}) - Q(\theta^{(t)} \mid \theta^{(t)})$$



log $P(x|\theta)$
$G_{t+1}(\theta^{(t+1)})$
$G_t(\theta^{(t)})$
$H[P(z|x;\ \theta^{(t+1)})]$
$Q(\theta\ |\ \theta^{(t+1)})$
$Q(\theta|\theta^{(t)})$
$H[P(z|x;\ \theta^{(t)})]$
$\theta^{(t)}$
$\theta^{(t+1)}$
$\theta^{(t+2)}$
θ

Using the EM algorithm for *learning a categorical distribution* [and implicitly a multinomial distribution]

Application to the ABO blood model

Liviu Ciortuz, following

Anirban DasGupta, *Probability for Statistics and Machine Learning*, Springer, 2011, Ex. 20.10

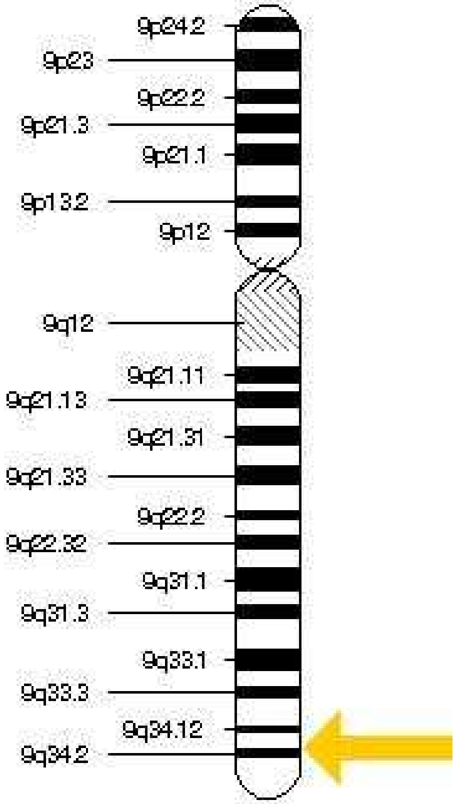
Grupele sangvine ale oamenilor sunt determinate de variantele („alelele“) unei gene situate pe cromozomul 9 (mai exact, la poziţia 9q34.2), numită gena *ABO*. Se ştie că fiecare dintre noi dispunem de câte o pereche de astfel de cromozomi, deci de câte două copii ale genei *ABO*, şi anume una moştenită de la tată şi una moştenită de la mamă.

Se notează cu A, B şi O cele trei tipuri de alele ale genei *ABO*. Alelele A şi B sunt *dominante* în raport cu alela O. (Altfel spus, alela O este recesivă în raport cu fiecare dintre alelele A şi B.) Alelele A şi B sunt *codominante*. Prin urmare, grupele sangvine pot fi specificate conform tabelului alăturat.

grupe sangvine	alele moştenite
A	AA AO
B	BB BO
AB	AB
O	OO

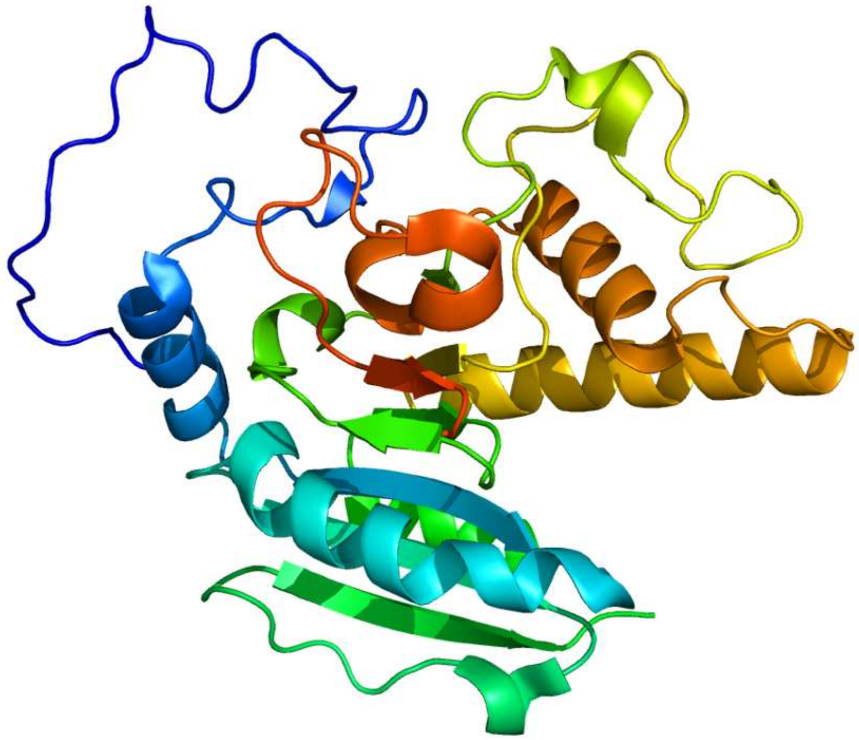
Presupunem că, într-o populaţie oarecare, fiecare dintre alelele A, B şi O are la bărbaţi aceeaşi frecvenţă ca şi la femei. În cele ce urmează, aceste frecvenţe (notate respectiv cu p_A, p_B şi p_O) vor fi considerate a priori necunoscute, dar urmează să le determinăm.

Location of the ABO gene on chromosome 9

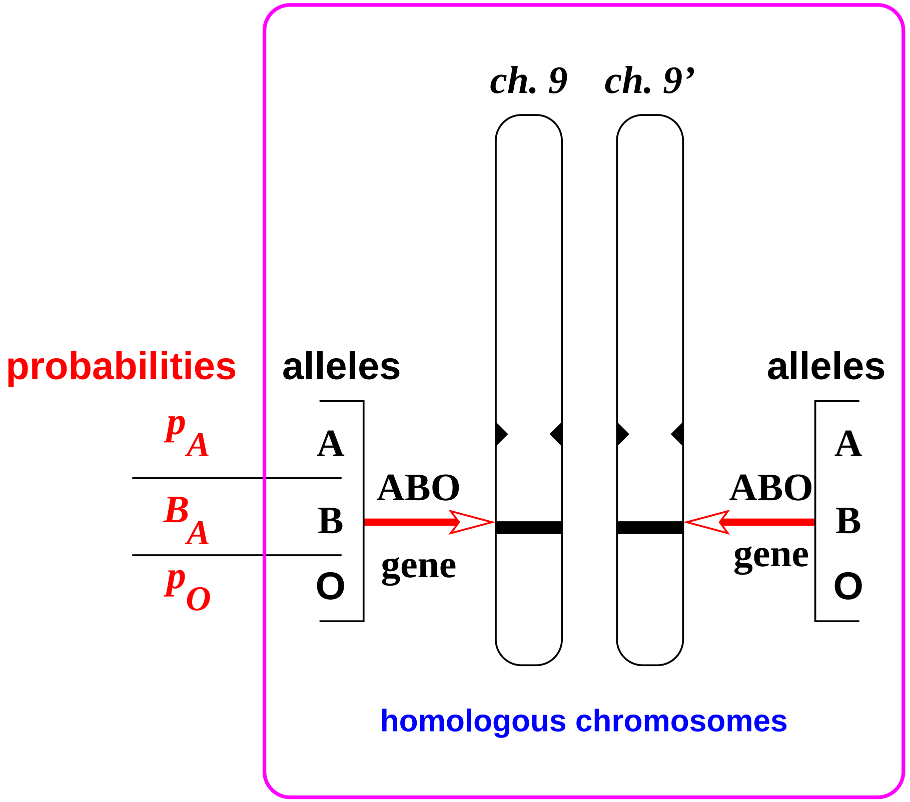


Source: wiki

The ABO protein



Source: wiki



n_{AA} → AA

n_{BB} → BB

pheno–type	blood type	*Probability*	*counts*
AA *AO*	*A*	$p_A^2 +$ $2\ p_A p_O$	n_A
BB *BO*	*B*	$p_B^2 +$ $2\ p_B p_O$	n_B
AB	*AB*	$2\ p_A p_B$	n_{AB}
O	*O*	p_O^2	n_O

parameters — unobservable data — observable data

Output ← EM algorithm ← Input

a. Considerăm că într-un eşantion de populaţie format din n persoane sunt n_A persoane cu grupa sangvină A, n_B persoane cu grupa sangvină B, n_{AB} persoane cu grupa sangvină AB şi n_O persoane cu grupa sangvină O. (Evident, $n = n_a + n_B + n_{AB} + n_O$.)

Pornind de la aceste date ,,observabile“, să se deriveze *algoritmul EM* pentru determinarea probabilităţilor p_A, p_B şi p_O. (Evident, întrucât suma lor este 1, va fi suficient să se determine doar două dintre ele.)

Indicaţii

1. Presupunând că procesul de asociere a alelelor moştenite de către un individ de la părinţii lui respectă proprietăţile specifice evenimentelor aleatoare independente, rezultă că probabilităţile de realizare a combinaţiilor (în termeni genetici: ,,fenotipurile“) AA, AO, BB, BO, AB **şi** OO **într-o populaţie oarecare sunt** p_A^2, $2p_Ap_O$, p_B^2, $2p_Bp_O$, $2p_Ap_B$, **şi respectiv** p_O^2.

2. Considerând $n_A = n_{AA}+n_{AO}$ **şi** $n_B = n_{BB}+n_{BO}$, **unde semnificaţiile numerelor** n_{AA}, n_{AO}, n_{BB} **şi** n_{BO} **sunt similare cu semnificaţiile numerelor** n_A, n_B, n_{AB}, **şi** n_O **care au fost precizate mai sus, este natural ca în formularea algoritmului EM datele** n_{AA} **şi** n_{BB} **să fie considerate ,,neobservabile“. Ca *parametri* ai modelului, se vor considera probabilităţile** p_A, p_B **şi** p_O.

3. La pasul E al algoritmului EM veţi calcula $\hat{n}_{AA}$ şi $\hat{n}_{BB}$, care reprezintă respectiv numărul „aşteptat“ de apariţii ale combinaţiei de alele AA şi numărul „aşteptat“ de apariţii ale combinaţiei de alele BB în populaţia dată.

Veţi scrie apoi funcţia de log-verosimilitate a datelor complete („observabile“ şi „neobservabile“), exprimată cu ajutorul distribuţiei $Multinomial(n; p_A^2, 2p_Ap_O, p_B^2, 2p_Bp_O, 2p_Ap_B, p_O^2)$.

4. La pasul M al algoritmului EM, pornind de la media funcţiei de log-verosimilitate care a fost calculată la pasul E, veţi stabili regulile de actualizare pentru probabilităţile p_A, p_B şi p_O.

Atenţie: suma acestor probabilităţi fiind 1, problema de optimizare pe care va trebui să o rezolvaţi la pasul M este una cu restricţii. În acest sens, metoda multiplicatorilor lui Lagrange vă poate fi de folos.

Solution

Notations:

- **parameters:** $p = \{p_A, p_B, p_O\}$
  (in fact, it will be enough to estimate p_A **and** p_B**, since** $p_A + p_B + p_O = 1$**)**
- **observable data:** $n_{obs} = \{n_A, n_B, n_{AB}, n_O\}$
- **unobservable data:** $n_{unobs} = \{n_{AA}, n_{AO}, n_{BB}, n_{BO}\}$
  (in fact, we will restrict n_{unobs} **to** $\{n_{AA}, n_{BB}\}$**)**
- **complete data:** $n_{compl} = n_{obs} \cup n_{unobs}$
- $n = n_A + n_B + n_{AB} + n_O$**,** $\quad n_A = n_{AA} + n_{AO}$**,** $\quad n_B = n_{BB} + n_{BO}$**.**

The likelihood of complete data:

$$Lp) \stackrel{not.}{=} P(n_{compl}|p) =$$

$$\frac{n!}{n_{AA}!\, n_{AO}!\, n_{BB}!\, n_{BO}!\, n_{AB}!\, n_O!} \cdot (p_A^2)^{n_{AA}} \cdot (2\, p_A\, p_O)^{n_{AO}} \cdot (p_B^2)^{n_{BB}} \cdot (2\, p_B\, p_O)^{n_{BO}} \cdot (2\, p_A\, p_B)^{n_{AB}} \cdot (p_O^2)^{n_O}$$

The likelihood function:

$$
\begin{aligned}
\ell(p) &\stackrel{def.}{=} \ln L(p) \\
&= \ln c + n_{AA} \ln(p_A^2) + n_{AO} \ln(2\, p_A\, p_O) + n_{BB} \ln(p_B^2) + n_{BO} \ln(2\, p_B\, p_O) + n_{AB} \ln(2\, p_A\, p_B) + n_O \ln(p_O^2) \\
&= \ln c' + 2\, n_{AA} \ln p_A + n_{AO}(\ln p_A + \ln p_O) + \\
&\quad 2\, n_{BB} \ln p_B + n_{BO}(\ln p_B + \ln p_O) + n_{AB}(\ln p_A + \ln p_B) + 2\, n_O \ln p_O \\
&= \ln c' + 2\, n_{AA} \ln p_A + (n_A - n_{AA})(\ln p_A + \ln p_O) + \\
&\quad 2\, n_{BB} \ln p_B + (n_B - n_{BB})(\ln p_B + \ln p_O) + n_{AB}(\ln p_A + \ln p_B) + 2\, n_O \ln p_O,
\end{aligned}
$$

where c and c' are constants which do not depend on the p parameter.
The "auxiliary" function:

$$
\begin{aligned}
Q(p|p^{(t)}) &\stackrel{def.}{=} E[\ell(p)|n_{obs}\,; p^{(t)}] \\
&= \ln c + 2\, \hat{n}_{AA} \ln p_A + (n_A - \hat{n}_{AA})(\ln p_A + \ln p_O) + \\
&\quad 2\, \hat{n}_{BB} \ln p_B + (n_B - \hat{n}_{BB})(\ln p_B + \ln p_O) + n_{AB}(\ln p_A + \ln p_B) + 2\, n_O \ln p_O,
\end{aligned}
$$

where

$$
\begin{aligned}
\hat{n}_{AA} &\stackrel{not.}{=} E[n_{AA}|n_{obs}\,; p^{(t)}] = E[n_{AA}|n_A, n_B, n_{AB}, n_O; p_A^{(t)}, p_B^{(t)}, p_O^{(t)}] \\
\hat{n}_{BB} &\stackrel{not.}{=} E[n_{BB}|n_{obs}\,; p^{(t)}] = E[n_{BB}|n_A, n_B, n_{AB}, n_O; p_A^{(t)}, p_B^{(t)}, p_O^{(t)}]
\end{aligned}
$$

E step: **As indicated by the expression of Q, here we have to compute $\hat{n}_{AA}$ and $\hat{n}_{BB}$ the *expected number* of individuals having the *phenotype* (aka, the allele pair) AA, and respectively BB.**

Firstly, taking into account that n_{AA} (or, equivalently, the phenotype AA), when seen as a random variable, follows a *binomial distribution* of parameters n_A and $\dfrac{(p_A^{(t)})^2}{(p_A^{(t)})^2 + 2\, p_A^{(t)}\, p_O^{(t)}}$, its expectation will be:

$$
\begin{aligned}
\hat{n}_{AA} &\stackrel{not.}{=} E[n_{AA} | n_A, n_B, n_{AB}, n_O; p_A^{(t)}, p_B^{(t)}, p_O^{(t)}] \\
&= \frac{(p_A^{(t)})^2}{(p_A^{(t)})^2 + 2\, p_A^{(t)}\, p_O^{(t)}} \cdot n_A
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\hat{n}_{BB} &\stackrel{not.}{=} E[n_{BB} | n_A, n_B, n_{AB}, n_O; p_A^{(t)}, p_B^{(t)}, p_O^{(t)}] \\
&= \frac{(p_B^{(t)})^2}{(p_B^{(t)})^2 + 2\, p_B^{(t)}\, p_O^{(t)}} \cdot n_B
\end{aligned}
$$

M step:

Given the constraint $p_A + p_B + p_O = 1$, we will introduce the Lagrange variable/"multiplier" λ and solve for the optimisation problem

$$
\begin{aligned}
p^{(t+1)} &\stackrel{not.}{=} (p_A^{(t+1)}, p_B^{(t+1)}, p_O^{(t+1)}) \\
&= \underset{p_A,\, p_B,\, p_O}{\operatorname{argmax}} \left[Q(p_A, p_B, p_O | p_A^{(t)}, p_B^{(t)}, p_O^{(t)}) + \lambda(1 - (p_A + p_B + p_O))\right] \\
&= \underset{p_A,\, p_B,\, p_O}{\operatorname{argmax}} [\ln c' + 2\, \hat{n}_{AA} \ln p_A + (n_A - \hat{n}_{AA})(\ln p_A + \ln p_O) + 2\, \hat{n}_{BB} \ln p_B + \\
&\qquad (n_B - \hat{n}_{BB})(\ln p_B + \ln p_O) + n_{AB}(\ln p_A + \ln p_B) + 2\, n_O \ln p_O + \\
&\qquad \lambda(1 - (p_A + p_B + p_O))]
\end{aligned}
$$

Taking the partial derivatives of the objective function w.r.t. p_A, p_B and p_O and solving them leads to:

$$\frac{1}{p_A}(2\, \hat{n}_{AA} + n_A - \hat{n}_{AA} + n_{AB}) - \lambda = 0 \Rightarrow \hat{p}_A = \frac{1}{\lambda}(\hat{n}_{AA} + n_A + n_{AB})$$

$$\frac{1}{p_B}(2\, \hat{n}_{BB} + n_B - \hat{n}_{BB} + n_{AB}) - \lambda = 0 \Rightarrow \hat{p}_B = \frac{1}{\lambda}(\hat{n}_{BB} + n_B + n_{AB})$$

$$\frac{1}{p_O}(n_A - \hat{n}_{AA} + n_B - \hat{n}_{BB} + 2\, n_O) - \lambda = 0 \Rightarrow \hat{p}_O = \frac{1}{\lambda}(n_A - \hat{n}_{AA} + n_B - \hat{n}_{BB} + 2n_O)$$

Now, enforcing the constraint $\hat{p}_A + \hat{p}_B + \hat{p}_O = 1$ gives

$$\frac{1}{\lambda}(\hat{n}_{AA} + n_A + n_{AB} + \hat{n}_{BB} + n_B + n_{AB} + n_A - \hat{n}_{AA} + n_B - \hat{n}_{BB} + 2n_O) = 1 \Leftrightarrow$$
$$\frac{2}{\lambda}(n_A + n_B + n_{AB} + n_O) = 1$$

Since $n = n_A + n_B + n_{AB} + n_O$, it folows immediately that $\frac{1}{\lambda}2n = 1$. Therefore $\lambda = 2n$.

Together with the results obtained on the previous slide, this leads to the following *updating relations*:

$$\hat{p}_A^{(t+1)} = \frac{1}{2n}(\hat{n}_{AA} + n_A + n_{AB})$$
$$\hat{p}_B^{(t+1)} = \frac{1}{2n}(\hat{n}_{BB} + n_B + n_{AB})$$
$$\hat{p}_O^{(t+1)} = \frac{1}{2n}(n_A - \hat{n}_{AA} + n_B - \hat{n}_{BB} + 2n_O) = \frac{1}{2n}(n_{AO} + n_{BO} + 2n_O)$$

b. Implementaţi algoritmul EM pe care l-aţi conceput la punctul a **şi rulaţi-l pentru input-ul** $n_A = 186$**,** $n_B = 38$**,** $n_{AB} = 13$**,** $n_O = 284$ **(**$n = 521$**).**

Ca *valori iniţiale* pentru parametri, veţi lucra mai întâi cu $p_A = p_B = p_O = \frac{1}{3}$**, iar apoi cu** $p_A = p_O = 0.1$ **şi** $p_B = 0.98$**.**

Pentru *oprire*, veţi cere ca $p_A^{(t)}$**,** $p_B^{(t)}$ **şi** $p_O^{(t)}$ **să nu difere (fiecare în parte) cu mai mult de** 10^{-4} **faţă de valorile calculate la iteraţia precedentă. Comparaţi rezultatele obţinute pentru fiecare din cele două iniţializări.**

Starting values:
$p_A = p_B = p_O = 1/3$.

Iterations:

t	p_A	p_B	p_C
1	0.2505	0.0611	0.6884
2	0.2185	0.0505	0.7311
3	0.2142	0.0502	0.7357
4	0.2137	0.0501	0.7362
5	0.2136	0.0501	0.7363
6	0.2136	0.0501	0.7363

Starting values:
$p_A = p_O = 0.1$, $p_B = 0.98$.

Iterations:

t	p_A	p_B	p_C
1	0.2505	0.0847	0.6648
2	0.2193	0.0511	0.7296
3	0.2143	0.0502	0.7355
4	0.2137	0.0501	0.7362
5	0.2136	0.0501	0.7363
6	0.2136	0.0501	0.7363

Note: The final results are the same.

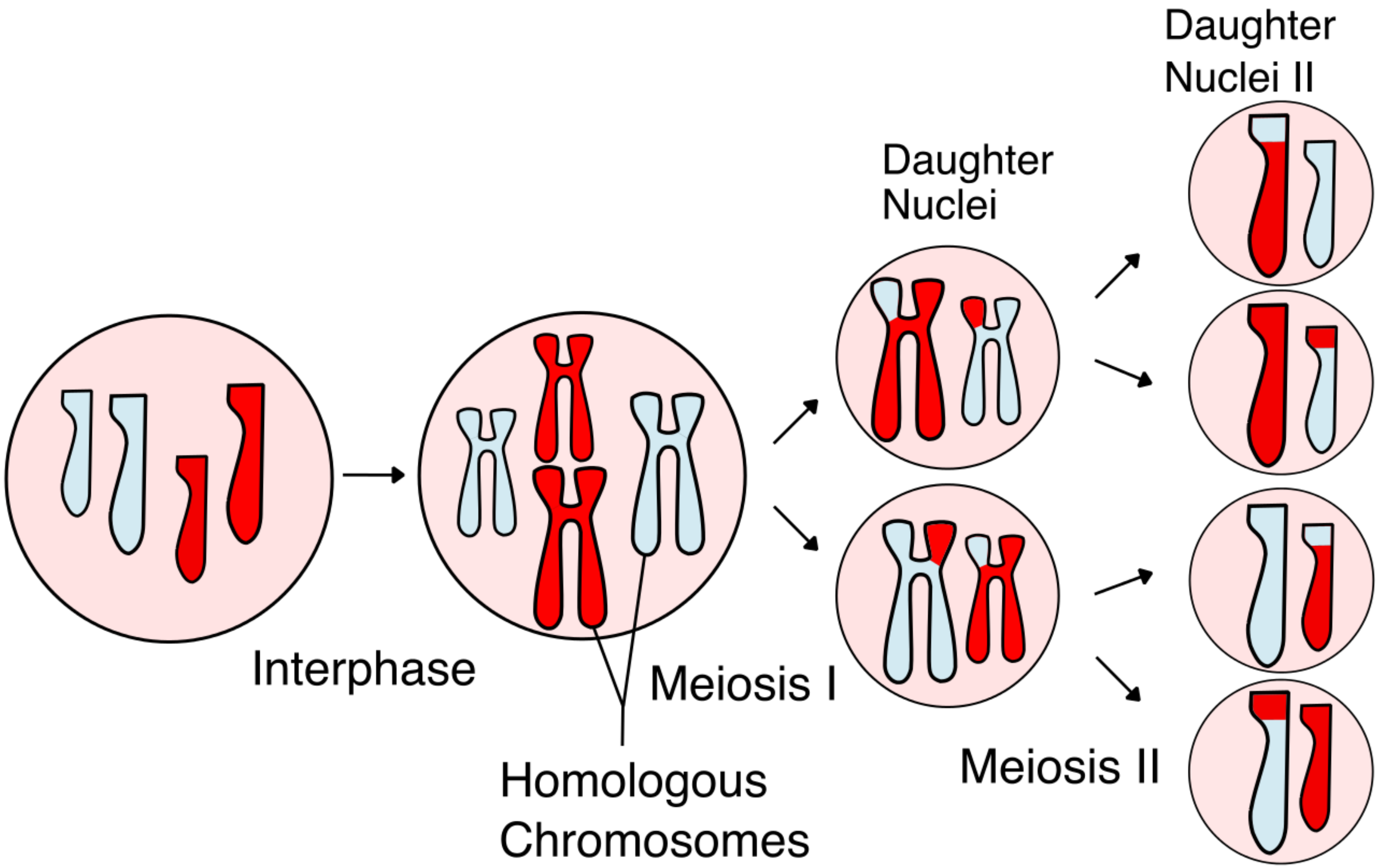
Using the EM algorithm for *learning a multinomial distribution* which depends on a single parameter

Application to a bioinformatics task: computing probabilities for two linked bi-allelic loci in haploid cells

Liviu Ciortuz, following

Brani Vidakovic, Georgia Institute of Technology, Bayesian Statistics course (ISyE 8843A), 2004, Handout 12, sec. 1.2.1

Biological background: Meiosis

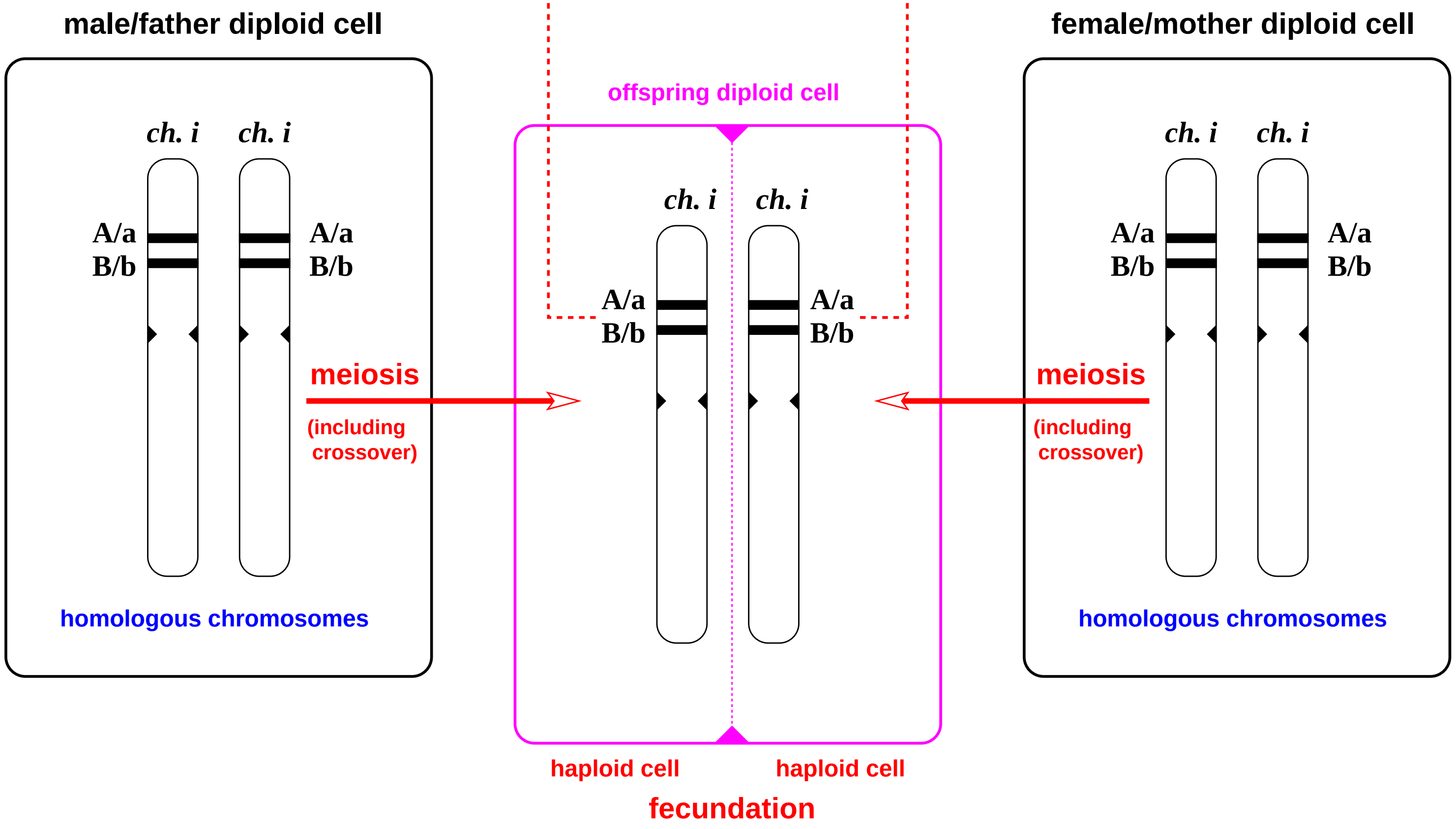


Source: wiki

Biological background (cont'd)

	P
AB	*(1−p)/2*
Ab	*p/2*
aB	*p/2*
ab	*(1−p)/2*

	P
AB	*(1−p')/2*
Ab	*p'/2*
aB	*p'/2*
ab	*(1−p')/2*



Biological background (cont'd)

Note: Alleles A and B are dominant w.r.t. a and respectively b.

Genotype:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
AA BB	Aa BB	aA BB	AA Bb	Aa Bb	aA Bb	AA bB	Aa bB	aA bB	AA bb	Aa bb	aA bb	aa BB	aa Bb	aa bB	aa bb

Genotypes 1–9	Genotypes 10–12	Genotypes 13–15	Genotype 16
AB	**Ab**	**aB**	**ab**
n_{AB}	n_{Ab}	n_{aB}	n_{ab}
$\frac{2+\psi}{4}$	$\frac{1-\psi}{4}$	$\frac{1-\psi}{4}$	$\frac{\psi}{4}$

with $\psi \stackrel{not.}{=} (1-p)(1-p')$**.**
We have designated in magenta colour the offspring phenotype, in blue the counts, i.e. number of individuals of having a certain phenotype in a given population, and in read the corresponding probabilities.

Note that

$$P(ab) = \frac{(1-p)(1-p')}{4} \text{ and}$$

$$P(Ab) = P(aB) = \frac{p}{2}\cdot\frac{p'}{2} + \frac{p}{2}\cdot\frac{1-p'}{2} + \frac{1-p'}{2}\cdot\frac{p}{2} = \frac{1-(1-p)(1-p')}{4}.$$

Fie o variabilă aleatoare X care ia valori în mulţimea $\{v_1, v_2, v_3, v_4\}$ şi urmează distribuţia $\mathit{Multinomial}\left(n; \frac{2+\psi}{4}, \frac{1-\psi}{4}, \frac{1-\psi}{4}, \frac{\psi}{4}\right)$, unde ψ este un parametru cu valori în intervalul $(0, 1)$. Cele patru probabilităţi listate în definiţia acestei distribuţii multinomiale sunt în corespondenţă directă cu valorile v_1, v_2, v_3 şi v_4.

a. Presupunem că n_1, n_2, n_3 şi n_4 sunt numărul de „realizări“ ale valorilor v_1, v_2, v_3 şi v_4 în totalul celor n „observaţii“. Pentru fixarea ideilor, vom considera $n_1 = 125$, $n_2 = 18$, $n_3 = 20$ şi $n_4 = 34$.

Calculaţi estimarea de verosimilitate maximă (MLE) a parametrului ψ.

Solution

By denoting $n = n_1 + n_2 + n_3 + n_4$**,**
and kowing that $n \sim \text{Multinomial}\left(n; \frac{2+\psi}{4}, \frac{1-\psi}{4}, \frac{1-\psi}{4}, \frac{\psi}{4}\right)$**,**
it follows that the verosimility function will be

$$L(\psi) \stackrel{def.}{=} P(D|\psi) = \frac{n!}{n_1!\, n_2!\, n_3!\, n_4!} \cdot \left(\frac{2+\psi}{4}\right)^{n_1} \cdot \left(\frac{1-\psi}{4}\right)^{n_2} \cdot \left(\frac{1-\psi}{4}\right)^{n_3} \cdot \left(\frac{\psi}{4}\right)^{n_4},$$

while

$$\ell(\psi) \stackrel{def.}{=} \ln L(\psi) = \ln c + n_1 \ln(2+\psi) + (n_2 + n_3) \ln(1-\psi) + n_4 \ln(\psi),$$

where c **is constant w.r.t.** ψ**.**

$$
\begin{aligned}
\frac{\partial}{\partial\psi}\ell(\psi) &= \frac{n_1}{2+\psi} - \frac{n_2+n_3}{1-\psi} + \frac{n_4}{\psi} = \frac{n_1\cdot\psi(1-\psi) - (n_2+n_3)\cdot\psi(2+\psi) + n_4\cdot(2+\psi)(1-\psi)}{(2+\psi)(1-\psi)\psi} \\
&= \frac{n_1\psi - n_1\psi^2 - 2n_2\psi - n_2\psi^2 - 2n_3\psi - n_3\psi^2 + 2n_4 - 2n_4\psi + n_4\psi - n_4\psi^2}{(2+\psi)(1-\psi)\psi} \\
&= \frac{-\psi^2(n_1+n_2+n_3+n_4) + \psi(n_1 - 2n_2 - 2n_3 - n_4) + 2n_4}{(2+\psi)(1-\psi)\psi} \\
&= \frac{-n\psi^2 + \psi(n_1 - 2n_2 - 2n_3 - n_4) + 2n_4}{(2+\psi)(1-\psi)\psi}
\end{aligned}
$$

Note that the denominator $(2+\psi)(1-\psi)\psi$ is always positive.
By substituting n_1, n_2, n_3 n_4 for 125, 18, 20 and respectively 34, the nominator becomes $-197\psi^2 + 15\psi + 68$.
By analysing the sign of this expression, it is easy to see that for our data the function $\ell(\psi)$ reaches its maximum in the interval $(0,1)$ for

$$
\hat{\psi} = \frac{-15 + \sqrt{15^2 + 4\cdot 197\cdot 68}}{-2\cdot 197} = \frac{-15+\sqrt{225+53809}}{-394} = \frac{-15+231.967}{-394} = 0.626769036.
$$

b. Fie acum variabila aleatoare $X' \sim Multinomial\left(n; \frac{1}{2}, \frac{\psi}{4}, \frac{1-\psi}{4}, \frac{1-\psi}{4}, \frac{\psi}{4}\right)$**. Valorile luate de variabila** X'**, corespunzător acestor cinci probabilităţi, sunt (în ordine)** v_1, v_1, v_2, v_3 **şi** v_4**.**

Observaţi că, în raport cu distribuţia multinomială de la punctul a**, am ,,descompus" probabilitatea** $\frac{2+\psi}{4}$ **în două probabilităţi,** $\frac{1}{2}$ **şi** $\frac{\psi}{4}$**, ca şi cum ele ar corespunde unor evenimente disjuncte.**

Vom considera n_{11}**,** n_{12}**,** n_2**,** n_3 **şi respectiv** n_4 **numărul de ,,realizări" ale acestor valori, însă de data aceasta** n_{11} **şi** n_{12} **vor fi neobservabile. În schimb, vom furniza suma** $n_1 = n_{11} + n_{12}$ **ca dată ,,observabilă", alături de celelalte date observabile,** n_2**,** n_3 **şi** n_4**.**

Să se estimeze parametrul ψ **folosind algoritmul EM. Cum este această estimare faţă de valoarea obţinută la punctul** a**?**

***Indicaţie*: Veţi elabora formulele corespunzătoare pasului E şi pasului M. Apoi veţi face o implementare şi veţi rula algoritmul EM (pornind, de exemplu, cu valoarea iniţială** 0.5 **pentru** ψ**) până când valorile acestui parametru până la cea de-a şasea zecimală nu se mai modifică.**

Solution: E-step

The verosimility function is now

$$
\begin{aligned}
P(D|\psi) &= \frac{n!}{n_{11}!\, n_{12}!\, n_2!\, n_3!\, n_4!} \cdot \left(\frac{1}{2}\right)^{n_{11}} \cdot \left(\frac{\psi}{4}\right)^{n_{12}} \cdot \left(\frac{1-\psi}{4}\right)^{n_2} \cdot \left(\frac{1-\psi}{4}\right)^{n_3} \cdot \left(\frac{\psi}{4}\right)^{n_4} \\
&= \frac{n!}{n_{11}!\, n_{12}!\, n_2!\, n_3!\, n_4!} \cdot \left(\frac{1}{2}\right)^{n_{11}} \cdot \left(\frac{1-\psi}{4}\right)^{n_2+n_3} \cdot \left(\frac{\psi}{4}\right)^{n_{12}+n_4},
\end{aligned}
$$

while the log-verosimility function can be written as

$$
\ln P(D|\psi) = \ln d + (n_2 + n_3)\ln(1-\psi) + (n_{12} + n_4)\ln(\psi),
$$

where d is constant w.r.t. ψ.
Therefore, the „auxiliary“ function will be

$$
\begin{aligned}
Q(\psi|\psi^{(t+1)}) &= E_{n_{12}|n_1,n_2,n_3,n_4,\psi^{(t)}}[\ln P(D|\psi)] \\
&= \ln d + (n_2 + n_3)\ln(1-\psi) + (\hat{n}_{12} + n_4)\ln(\psi),
\end{aligned}
$$

where

$$
\hat{n}_{12} \stackrel{not.}{=} E[n_{12}|n_1, n_2, n_3, n_4, \psi^{(t)}] = \frac{\frac{1}{2}}{\frac{1}{2} + \frac{\psi}{4}} = \frac{\psi^{(t)}}{2 + \psi^{(t)}} \cdot n_1.
$$

M-step

$$\frac{\partial}{\partial \psi} E[\ln P(D|\psi)] = -(n_2+n_3)\frac{1}{1-\psi} + (\hat{n}_{12}+n_4)$$

$$\Rightarrow \frac{\partial^2}{\partial \psi^2} E[\ln P(D|\psi)] = (n_1+n_3)\frac{-1}{n_1+n_3} - (\hat{n}_{12}+n_4)\frac{1}{\psi^2} \leq 0$$

$\Rightarrow$ **the auxiliary function is concave.**

$$\frac{\partial}{\partial \psi} E[\ln P(D|\psi)] = 0 \Leftrightarrow$$

$$(\hat{n}_{12}+n_4)(1-\psi) = \psi(n_2+n_3) \Leftrightarrow$$

$$\psi(\hat{n}_{12}+n_2+n_3+n_4) = \hat{n}_{12}+n_4 \Leftrightarrow$$

$$\psi^{(t+1)} = \frac{\hat{n}_{12}+n_4}{n-\hat{n}_{11}}$$

where

$$\hat{n}_{11} \stackrel{not.}{=} E[n_{11}|n_1,n_2,n_3,n_4,\psi^{(t)}] = n_1 - \hat{n}_{12} = \frac{2}{2+\psi^{(t)}} \cdot n_1.$$

A Matlab implementation of this EM algorithm produces $\psi = 0.62682139$**.**

The EM algorithm for solving a Bernoulli mixture model

CMU, 2008 fall, Eric Xing, HW4, pr. 1.4-7

Suppose I have two unfair coins. The first lands on heads with probability p, and the second lands on heads with probability q.

Imagine n tosses, where for each toss I choose to use the first coin with probability π and choose to use the second with probability $1 - \pi$. The outcome of each toss i is $x_i \in \{0, 1\}$.

Suppose I tell you the outcomes of the n tosses, $x \stackrel{not.}{=} \{x_1, x_2, \ldots, x_n\}$, but I don't tell you which coins I used on which toss.

Given only the outcomes, x, your job is to compute estimates for θ which is the set of all parameters, $\theta = \{p, q, \pi\}$ using the EM algorithm.

To compute these estimates, we will create a latent variable Z, where $z_i \in \{0, 1\}$ indicates the coin used for the n^{th} toss. For example $z_2 = 1$ indicates the first coin was used on the second toss.

We define the "incomplete" data log-likelihood as $\log P(x|\theta)$ and the "complete" data log-likelihood as log $P(x, z|\theta)$.

a. Show that $E[z_i|x_i, \theta] = P(z_i = 1|x_i, \theta)$.

b. Use Bayes rule to compute $P(z_i = 1|x_i, \theta^{(t)})$, where $\theta^{(t)}$ denotes the parameters at iteration t.

c. Write down the complete log-likelihood, $\log P(x, z|\theta)$.

d. *E-Step*: Show that the expected log-likelihood of the complete data $Q(\theta \mid \theta^{(t)}) \stackrel{not.}{=} E_{P(z|x,\theta^{(t)})}[\log P(x, z \mid \theta)]$ is given by

$$
\begin{aligned}
Q(\theta \mid \theta^{(t)}) \;\; = \;\; & \sum_{i=1}^{n} E[z_i \mid x_i, \theta^{(t)}] \cdot (\log \pi + x_i \log p + (1 - x_i) \log(1 - p)) + \\
& + \left(1 - E[z_i \mid x_i, \theta^{(t)}]\right) \cdot (\log(1 - \pi) + x_i \log q + (1 - x_i) \log(1 - q))
\end{aligned}
$$

e. *M-Step*: Describe the process you would use to obtain the update equations for $p^{(t+1)}$, $q^{(t+1)}$ şi $\pi^{(t+1)}$.

Solution

a.

$$E[z_i \mid x_i, \theta] = \sum_{z \in \{0,1\}} z_i P(z_i \mid x_i, \theta) = 0 \cdot P(z_i = 0 \mid x_i, \theta) + 1 \cdot P(z_i = 1 \mid x_i, \theta)$$

$$\Rightarrow E[z_i \mid x_i, \theta] = P(z_i = 1 \mid x_i, \theta)$$

b.

$$\begin{aligned} & P(z_i = 1 \mid x_i, \theta) \\ = \; & \frac{P(x_i \mid z_i = 1, \theta) P(z_i = 1 \mid \theta)}{P(x_i \mid z_i = 1, \theta) P(z_i = 1 \mid \theta) + P(x_i \mid z_i = 0, \theta) P(z_i = 0 \mid \theta)} \\ = \; & \frac{p^{x_i} \cdot (1-p)^{1-x_i} \cdot \pi}{p^{x_i} \cdot (1-p)^{1-x_i} \cdot \pi + q^{x_i} \cdot (1-q)^{1-x_i} \cdot (1-\pi)} \end{aligned}$$

c.

***Note* (in Romanian): Din punct de vedere metodologic, pentru a calcula $P(x, z|\theta)$ ne putem inspira de la punctul precedent: observăm că la numitorul fracţiei care ne dă valoarea lui $P(z_i|x_i, \theta)$ avem $P(x_i, z_i = 1|\theta)$ şi $P(x_i, z_i = 0|\theta)$. Pornind de la aceste două expresii, ne vom propune să le exprimăm în mod unitar, adică sub forma unei singure expresii.**

$$
\begin{aligned}
\log P(x, z \mid \theta) &\stackrel{i.i.d.}{=} \log \prod_{i=1}^{n} P(x_i, z_i \mid \theta) = \log \prod_{i=1}^{n} P(x_i \mid z_i, \theta) \cdot P(z_i \mid \theta) \\
&= \log \prod_{i=1}^{n} \left(p^{x_i}(1-p)^{1-x_i}\pi\right)^{z_i} \left(q^{x_i}(1-q)^{1-x_i}(1-\pi)\right)^{1-z_i} \\
&= \sum_{i=1}^{n} \log \left(\left(p^{x_i}(1-p)^{1-x_i}\pi\right)^{z_i} \left(q^{x_i}(1-q)^{1-x_i}(1-\pi)\right)^{1-z_i}\right) \\
&= \sum_{i=1}^{n} \left[z_i \log\left(p^{x_i}(1-p)^{1-x_i}\pi\right) + (1-z_i)\log\left(q^{x_i}(1-q)^{1-x_i}(1-\pi)\right)\right]
\end{aligned}
$$

d.

$$
\begin{aligned}
Q(\theta \mid \theta^{(t)}) &\stackrel{not.}{=} E_{P(z|x,\theta^{(t)})}[\log P(x, z \mid \theta)] \\
&= E_{P(z|x,\theta^{(t)})}\Big[\sum_{i=1}^{n} \big[z_i \log \big(p^{x_i}(1-p)^{1-x_i}\pi\big) + \\
&\qquad\qquad (1-z_i) \log \big(q^{x_i}(1-q)^{1-x_i}(1-\pi)\big)\big]\Big] \\
&= \sum_{i=1}^{n} \big[E[z_i \mid x_i, \theta^{(t)}] \cdot \log p^{x_i}(1-p)^{1-x_i}\pi + \\
&\qquad + \big(1 - E[z_i \mid x_i, \theta^{(t)}]\big) \cdot q^{x_i}(1-q)^{1-x_i}(1-\pi)\big] \\
&= \sum_{i=1}^{n} \big[E[z_i \mid x_i, \theta^{(t)}] \cdot (\log \pi + x_i \log p + (1-x_i)\log(1-p)) + \\
&\qquad + \big(1 - E[z_i \mid x_i, \theta^{(t)}]\big) \cdot (\log(1-\pi) + x_i \log q + (1-x_i)\log(1-q))\big]
\end{aligned}
$$

e.

$$\mu_i^{(t)} \stackrel{not.}{=} E[z_i \mid x_i, \theta^{(t)}] = \frac{(p^{(t)})^{x_i} \cdot (1-p^{(t)})^{1-x_i} \cdot \pi^{(t)}}{(p^{(t)})^{x_i} \cdot (1-p^{(t)})^{1-x_i} \cdot \pi^{(t)} + (q^{(t)})^{x_i} \cdot (1-q^{(t)})^{1-x_i} \cdot (1-\pi^{(t)})}$$

$$\begin{aligned} \Rightarrow Q(\theta \mid \theta^{(t)}) \;&=\; \sum_{i=1}^{n} \Big[\mu_i^{(t)} \left(\log \pi + x_i \log p + (1-x_i) \log(1-p)\right) + \\ &+ \left(1-\mu_i^{(t)}\right) \cdot \left(\log(1-\pi) + x_i \log q + (1-x_i) \log(1-q)\right)\Big] \end{aligned}$$

$$\frac{\partial Q(\theta \mid \theta^{(t)})}{\partial p} = 0 \Leftrightarrow \sum_{i=1}^{n} \mu_i^{(t)} \left(\frac{x_i}{p} - \frac{1-x_i}{1-p}\right) = 0 \Leftrightarrow \frac{1}{p} \sum_{i=1}^{n} \mu_i^{(t)} x_i = \frac{1}{1-p} \sum_{i=1}^{n} \mu_i^{(t)} (1-x_i)$$

$$\Leftrightarrow \quad (1-p) \sum_{i=1}^{n} \mu_i^{(t)} x_i = p \sum_{i=1}^{n} \mu_i^{(t)} (1-x_i) \Leftrightarrow \sum_{i=1}^{n} \mu_i^{(t)} x_i = p \left(\sum_{i=1}^{n} \mu_i^{(t)} (1-x_i) + \sum_{i=1}^{n} \mu_i^{(t)} x_i\right)$$

$$\Leftrightarrow \quad \sum_{i=1}^{n} \mu_i^{(t)} x_i = p \sum_{i=1}^{n} \mu_i^{(t)} \qquad \Rightarrow \qquad p^{(t+1)} = \frac{\sum_{i=1}^{n} \mu_i^{(t)} x_i}{\sum_{i=1}^{n} \mu_i^{(t)}} \in [0,1]$$

$$\frac{\partial Q(\theta \mid \theta^{(t)})}{\partial q} = 0 \Leftrightarrow \sum_{i=1}^{n} (1-\mu_i^{(t)}) \left(\frac{x_i}{q} - \frac{1-x_i}{1-q} \right) = 0 \Rightarrow q^{(t+1)} = \frac{\sum_{i=1}^{n} (1-\mu_i^{(t)}) x_i}{\sum_{i=1}^{n} (1-\mu_i^{(t)})} \in [0,1]$$

$$\frac{\partial Q(\theta \mid \theta^{(t)})}{\partial \pi} = 0 \Leftrightarrow \sum_{i=1}^{n} \left(\frac{\mu_i^{(t)}}{\pi} - \frac{1-\mu_i^{(t)}}{1-\pi} \right) = 0 \Leftrightarrow \frac{1}{\pi} \sum_{i=1}^{n} \mu_i^{(t)} = \frac{1}{1-\pi} \sum_{i=1}^{n} (1-\mu_i^{(t)})$$

$$\Leftrightarrow \quad (1-\pi) \sum_{i=1}^{n} \mu_i^{(t)} = \pi \sum_{i=1}^{n} (1-\mu_i^{(t)}) \Leftrightarrow \sum_{i=1}^{n} \mu_i^{(t)} = \pi \left(\sum_{i=1}^{n} (1-\mu_i^{(t)}) + \sum_{i=1}^{n} \mu_i^{(t)} \right)$$

$$\Leftrightarrow \quad \sum_{i=1}^{n} \mu_i^{(t)} = \pi \sum_{i=1}^{n} 1 \Leftrightarrow \sum_{i=1}^{n} \mu_i^{(t)} = n\pi \qquad \Rightarrow \qquad \pi^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \mu_i^{(t)} \in [0,1]$$

Note: **It can be easily shown that *the second derivatives* ($\frac{\partial^2}{\partial p^2}$, $\frac{\partial^2}{\partial q^2}$ and $\frac{\partial^2}{\partial \pi^2}$) are all *negative* on the respective domains of definition. Therefore, the above solutions ($p^{(t+1)}$, $q^{(t+1)}$ and $\pi^{(t+1)}$) maximize the Q functional.**

Estimarea parametrilor
unei mixturi de distribuţii [de tip] Bernoulli

A. când toate variabilele sunt observabile: MLE;
B. când unele variabile sunt neobservabile: algoritmul EM

prelucrare de Liviu Ciortuz, după

***"What is the expectation maximization algorithm?"*,**
Chuong B. Do, Serafim Batzoglou,
Nature Biotechnology, vol. 26, no. 8, 2008, pag. 897-899

Fie următorul *experiment probabilist*:

Dispunem de două monede, A şi B.
Efectuăm 5 serii de operaţiuni de tipul următor:

- **Alegem în mod aleatoriu una dintre monedele A şi B, cu probabilitate egală (1/2);**
- **Aruncăm de 10 ori moneda care tocmai a fost aleasă (Z) şi notăm rezultatul, sumarizat ca număr (X) de feţe 'head' (ro. 'stemă') obţinute în urma aruncării.**

A. La acest punct vom considera că s-a obţinut următorul rezultat pentru experimentul nostru:

i	Z_i	X_i
1	B	$5H \quad (5T)$
2	A	$9H \quad (1T)$
3	A	$8H \quad (2T)$
4	B	$4H \quad (6T)$
5	A	$7H \quad (3T)$

Semnificaţia variabilelor aleatoare Z_i şi X_i pentru $i = 1, \ldots, 5$ din tabelul de mai sus este imediată.

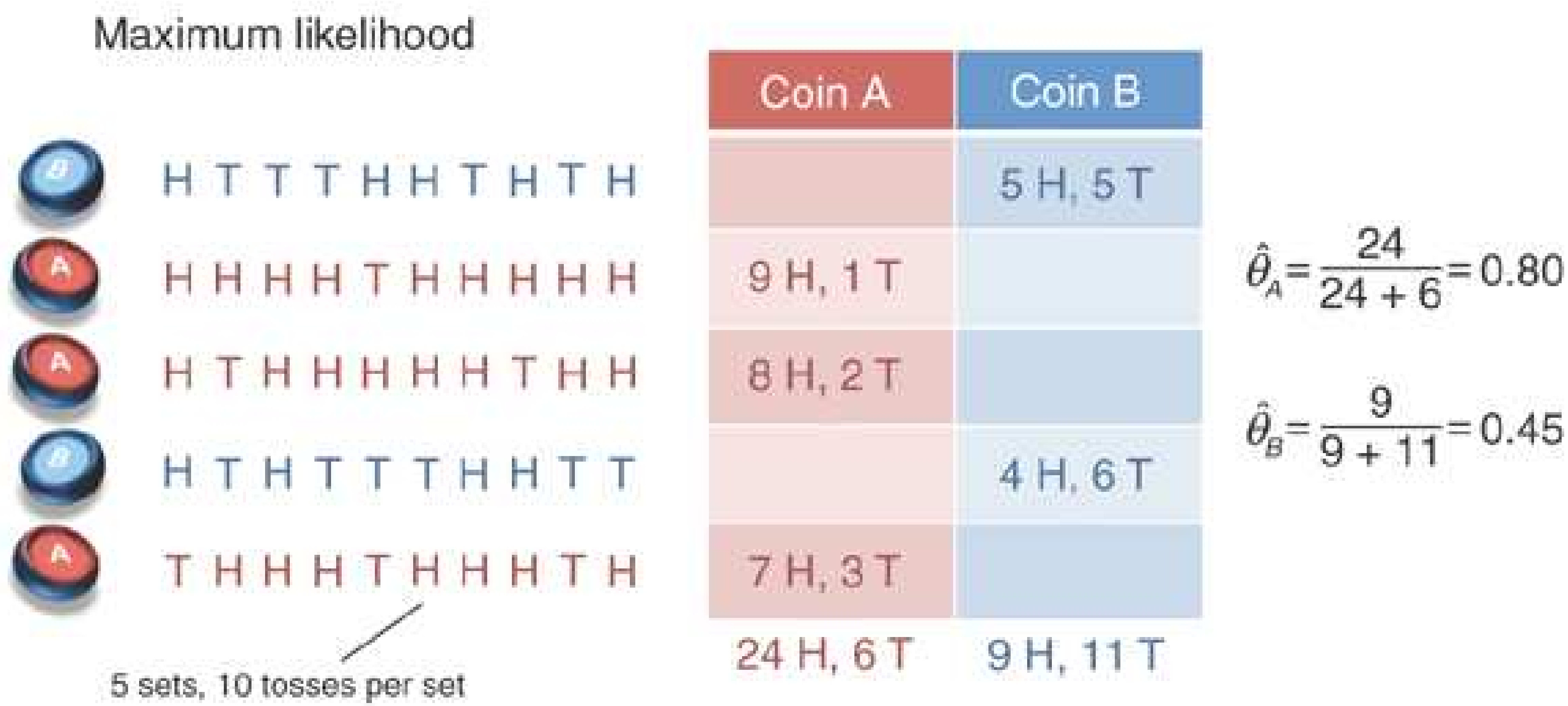
i. **Calculaţi $\hat{\theta}_A$ şi $\hat{\theta}_B$, probabilităţile de apariţie a feţei 'head'/stemă pentru cele două monede, folosind *definiţia clasică pentru probabilitatea evenimentelor aleatoare*, şi anume raportul dintre numărul de cazuri favorabile şi numărul de cazuri posibile, relativ la întregul experiment.**

Răspuns:

Analizând datele din tabelul din enunţ, rezultă imediat

$$\hat{\theta}_A = \frac{24}{24+6} = 0.8 \textbf{ şi } \hat{\theta}_B = \frac{9}{9+11} = 0.45.$$

De observat că termenii 6 şi respectiv 11 de la numitorii acestor fracţii reprezintă numărul de feţe 'tail'/ban care au fost obţinute la aruncarea monedei A şi respectiv B: $6T = 1T + 2T + 3T$, $11T = 5T + 6T$.


Maximum likelihood
Coin A
Coin B
H T T T H H T H T H
H H H H T H H H H H
H T H H H H H T H H
H T H T T T H H T T
T H H H T H H H T H
5 H, 5 T
9 H, 1 T
8 H, 2 T
4 H, 6 T
7 H, 3 T
24 H, 6 T
9 H, 11 T
5 sets, 10 tosses per set
$\hat{\theta}_A = \frac{24}{24+6} = 0.80$
$\hat{\theta}_B = \frac{9}{9+11} = 0.45$

Observaţie

Dacă în locul variabilelor binare $Z_i \in \{A, B\}$ pentru $i = 1, \ldots, 5$ introducem în mod natural variabilele-indicator $Z_{i,A} \in \{0, 1\}$ şi $Z_{i,B} \in \{0, 1\}$ tot pentru $i = 1, \ldots, 5$, definite prin $Z_{i,A} = 1$ iff $Z_i = A$, şi $Z_{i,A} = 0$ iff $Z_{i,B} = 0$, atunci procesările necesare pentru calculul probabilităţilor/parametrilor $\hat{\theta}_A$ şi $\hat{\theta}_B$ pot fi prezentate în mod sintetizat ca în tabelul de mai jos.[a]

i	$Z_{i,A}$	$Z_{i,B}$	X_i
1	0	1	$5H$
2	1	0	$9H$
3	1	0	$8H$
4	0	1	$4H$
5	1	0	$7H$

$\Rightarrow$

[a]Prezentăm acest "artificiu" ca pregătire pentru rezolvarea (ulterioară a) punctului B al prezentei probleme.

$$\Rightarrow \quad \begin{array}{r|l} X_i \cdot Z_{i,A} & X_i \cdot Z_{i,B} \\ \hline 0H \quad (0T) & \quad 5H \quad (5T) \\ 9H \quad (1T) & \quad 0H \quad (0T) \\ 8H \quad (2T) & \quad 0H \quad (0T) \\ 0H \quad (0T) & \quad 4H \quad (6T) \\ 7H \quad (3T) & \quad 0H \quad (0T) \\ \hline \sum_{i=1}^{5} X_i \cdot Z_{i,A} = 24H & \sum_{i=1}^{5} X_i \cdot Z_{i,B} = 9H \\ \sum_{i=1}^{5} (10 - X_i) \cdot Z_{i,A} = 6T & \sum_{i=1}^{5} (10 - X_i) \cdot Z_{i,B} = 11T \end{array} \quad \Rightarrow$$

$$\Rightarrow \quad \begin{cases} \hat{\theta}_A = \dfrac{24}{24+6} = 0.8 \\[2ex] \hat{\theta}_B = \dfrac{9}{9+11} = 0.45 \end{cases}$$

ii. **Calculaţi** $L_1(\theta_A, \theta_B) \stackrel{not.}{=} P(X, Z \mid \theta_A, \theta_B)$**, funcţia de verosimilitate a datelor** $X \stackrel{not.}{=} < X_1, \ldots, X_5 >$ **şi** $Z \stackrel{not.}{=} < Z_1, \ldots, Z_5 >$**, în raport cu parametrii** θ_A **şi respectiv** θ_B **ai distribuţiilor *Bernoulli* care modelează aruncarea celor două monede.**

***Observaţie*: Facem *presupunerea* că am reţinut succesiunea [tuturor] rezultatelor obţinute la aruncările celor două monede. Precizarea *de facto* a acestei succesiuni nu este esenţială. Dacă nu s-ar face această *presupunere*, ar trebui să să lucrăm cu *distribuţia binomială*.**

Răspuns:

Calculul verosimilităţii datelor:

$$
\begin{aligned}
L_1(\theta_A, \theta_B) & \stackrel{def.}{=} P(X, Z_A, Z_B \mid \theta_A, \theta_B) \stackrel{indep.\ cdt.}{=} \prod_{i=1}^{5} P(X_i, Z_{i,A}, Z_{i,B} \mid \theta_A, \theta_B) \\
& = \prod_{i=1}^{5} P(X_i \mid Z_{i,A}, Z_{i,B}; \theta_A, \theta_B) \cdot P(Z_{i,A}, Z_{i,B} \mid \theta_A, \theta_B) \\
& = P(X_1 \mid Z_{B,1} = 1, \theta_B) \cdot 1/2 \cdot \\
& \quad\; P(X_2 \mid Z_{A,2} = 1, \theta_A) \cdot 1/2 \cdot \\
& \quad\; P(X_3 \mid Z_{A,3} = 1, \theta_A) \cdot 1/2 \cdot \\
& \quad\; P(X_4 \mid Z_{B,4} = 1, \theta_B) \cdot 1/2 \cdot \\
& \quad\; P(X_5 \mid Z_{A,5} = 1, \theta_A) \cdot 1/2 \\
& = \theta_B^5 (1-\theta_B)^5 \cdot \theta_A^9 (1-\theta_A) \cdot \theta_A^8 (1-\theta_A)^2 \cdot \theta_B^4 (1-\theta_B)^6 \cdot \theta_A^7 (1-\theta_A)^3 \cdot \frac{1}{2^5} \\
& = \frac{1}{2^5}\, \theta_A^{24} (1-\theta_A)^6\, \theta_B^9 (1-\theta_B)^{11}
\end{aligned}
$$

iii. **Calculaţi** $\hat{\theta}_A \stackrel{not.}{=} \arg\max_{\theta_A} \log L_1(\theta_A, \theta_B)$ **şi** $\hat{\theta}_B \stackrel{not.}{=} \arg\max_{\theta_B} \log L_1(\theta_A, \theta_B)$ **folosind derivatele parţiale de ordinul întâi.**

Observaţii:

1. Baza logaritmului, fixată dar lăsată mai sus nespecificată, se va considera supraunitară (de exemplu 2, e **sau 10).**

2. Lucrând corect, veţi obţine acelaşi rezultat ca la punctul ***i.***

Răspuns:

Funcţia de log-verosimilitate a datelor complete se exprimă astfel:

$$\log L_1(\theta_A, \theta_B) = -5\log 2 + 24\log\theta_A + 6\log(1-\theta_A) + 9\log\theta_B + 11\log(1-\theta_B)$$

Prin urmare, maximul acestei funcţii în raport cu parametrul θ_A se calculează astfel:

$$\frac{\partial \log L_1(\theta_A, \theta_B)}{\partial \theta_A} = 0 \Leftrightarrow \frac{24}{\theta_A} - \frac{6}{1-\theta_A} = 0 \Leftrightarrow \frac{4}{\theta_A} = \frac{1}{1-\theta_A} \Leftrightarrow 4 - 4\,\theta_A = \theta_A \Leftrightarrow \hat{\theta}_A = 0.8$$

Similar, se face calculul şi pentru $\dfrac{\partial \log L_1(\theta_A, \theta_B)}{\partial \theta_B}$ şi se obţine $\hat{\theta}_B = 0.45$.[a]

Cele două valori obţinute, $\hat{\theta}_A$ şi $\hat{\theta}_B$, reprezintă estimarea de verosimilitate maximă (MLE) a probabilităţilor de apariţie a feţei 'head' ('stema') pentru moneda A şi respectiv moneda B.

---

[a] Se verifică uşor faptul că într-adevăr rădăcinile derivatelor parţiale de ordinul întâi pentru funcţia de log-verosimilitate reprezintă puncte de maxim. Pentru aceasta se studiază semnele acestor derivate.

Observaţii

1. Am arătat pe acest caz particular că metoda de calculare a probabilităţilor ($\hat{\theta}_A$ şi $\hat{\theta}_B$) direct din datele observate (aşa cum o ştim din liceu) corespunde de fapt metodei de estimare în sensul verosimilităţii maxime (MLE).

2. La punctul B vom arăta cum anume se poate face estimarea aceloraşi parametri θ_A şi θ_B în cazul in care o parte din date, şi anume variabilele Z_i (pentru $i = 1, \ldots, 5$) sunt neobservabile.

B. La acest punct se va relua experimentul de la punctul A, însă de data aceasta vom considera că valorile variabilelor Z_i nu sunt cunoscute.

i	Z_i	X_i
1	?	$5H \quad (5T)$
2	?	$9H \quad (1T)$
3	?	$8H \quad (2T)$
4	?	$4H \quad (6T)$
5	?	$7H \quad (3T)$

iv. **Pentru convenienţă, în locul variabilelor „neobservabile“ Z_i pentru $i = 1, \ldots, 5$ vom considera variabilele-indicator (de asemenea neobservabile) $Z_{i,A}, Z_{i,B} \in \{0, 1\}$, cu $Z_{i,A} = 1$ iff $Z_{i,B} = 0$ şi $Z_{i,B} = 1$ iff $Z_{i,A} = 0$. Evident, întrucât variabilele Z_i sunt aleatoare, rezultă că şi variabilele $Z_{i,A}$ şi $Z_{i,B}$ sunt aleatoare.**

Folosind teorema lui Bayes, calculaţi mediile variabilelor neobservabile $Z_{i,A}$ şi $Z_{i,B}$ condiţionate de variabilele observabile X_i. Veţi considera că parametrii acestor distribuţii Bernoulli care modelează aruncarea monedelor A şi B au valorile $\theta_A^{(0)} = 0.6$ şi respectiv $\theta_B^{(0)} = 0.5$.

Aşadar, se cer: $E[Z_{i,A} \mid X_i, \theta_A^{(0)}, \theta_B^{(0)}]$ şi $E[Z_{i,B} \mid X_i, \theta_A^{(0)}, \theta_B^{(0)}]$ pentru $i = 1, \ldots, 5$. Ca şi mai înainte, probabilităţile a priori $P(Z_{i,A} = 1)$ şi $P(Z_{i,B} = 1)$ se vor considera egale cu 1/2.

Notă

Algoritmul EM ne permite să facem în mod iterativ estimarea parametrilor θ_A şi θ_B în funcţie de valorile variabilelor observabile, X_i, şi de valorile inţiale atribuite parametrilor (în cazul nostru, $\theta_A^{(0)} = 0.6$ şi $\theta_B^{(0)} = 0.5$).

Vom face o sinteză a calculelor de la prima iteraţie a algoritmului EM — detaliate la punctele *iv*, v şi *vi* de mai jos — sub forma următoare, care seamănă într-o anumită măsură cu tabelele de la punctul A:

i	$Z_{i,A}$	$Z_{i,B}$	X_i
1	–	–	$5H$
2	–	–	$9H$
3	–	–	$8H$
4	–	–	$4H$
5	–	–	$7H$

$\stackrel{E}{\Longrightarrow}$

$E[Z_{i,A}]$	$E[Z_{i,B}]$
$0.45H$	$0.55H$
$0.80H$	$0.20H$
$0.73H$	$0.27H$
$0.35H$	$0.65H$
$0.65H$	$0.35H$

$\stackrel{M}{\Longrightarrow}$

Notă (cont.)

$$
\stackrel{M}{\Rightarrow}
\begin{array}{r|l}
X_i \cdot E[Z_{i,A}] & X_i \cdot E[Z_{i,B}] \\
\hline
2.2H \quad (2.2T) & \quad 2.8H \quad (2.8T) \\
7.2H \quad (0.8T) & \quad 1.8H \quad (0.2T) \\
5.9H \quad (1.5T) & \quad 2.1H \quad (0.5T) \\
1.4H \quad (2.1T) & \quad 2.6H \quad (3.9T) \\
4.5H \quad (1.9T) & \quad 2.5H \quad (1.1T) \\
\hline
\sum_{i=1}^{5} X_i \cdot E[Z_{i,A}] = 21.3H & \sum_{i=1}^{5} X_i \cdot E[Z_{i,B}] = 11.7H \\
\sum_{i=1}^{5}(10 - X_i) \cdot E[Z_{i,A}] = 8.7T & \sum_{i=1}^{5}(10 - X_i) \cdot E[Z_{i,B}] = 8.3T
\end{array}
\Rightarrow
$$

$$
\Rightarrow \left\{
\begin{array}{l}
\hat{\theta}_A^{(1)} = \dfrac{21.3}{21.3 + 8.7} \approx 0.71 \\[2ex]
\hat{\theta}_B^{(1)} = \dfrac{11.7}{11.7 + 8.3} \approx 0.58
\end{array}
\right.
$$

Notă (cont.)

Vom arăta că:

- **într-adevăr, este posibilă calcularea mediilor variabilelor neobservabile $Z_{i,A}$ şi $Z_{i,B}$, condiţionate de variabilele observabile X_i şi în funcţie de valorile asignate iniţial (0.6 şi 0.5 în enunţ, dar în general ele se pot asigna în mod aleatoriu) pentru parametrii θ_A şi θ_B;**

- **faţă de tabloul de sinteză de la punctul precedent, când toate variabilele erau observabile şi se calculau produsele $X_i \cdot Z_{i,A}$ şi $X_i \cdot Z_{i,B}$, aici se înlocuiesc variabilele $Z_{i,A}$ şi $Z_{i,B}$ cu mediile $E[Z_{i,A}]$ şi $E[Z_{i,B}]$ în produsele respective. De fapt, în loc să se calculeze $\sum_i X_i \cdot Z_{i,A}$ se calculează media $E[\sum_i X_i \cdot Z_{i,A}]$, şi similar pentru B.**

***Notaţie*: Pentru simplitate, în cele de mai sus (inclusiv în tabelele precedente), prin $E[Z_{i,A}]$ am notat $E[Z_{i,A} \mid X_i, \theta^{(0)}]$, iar prin $E[Z_{i,B}]$ am notat $E[Z_{i,B} \mid X_i, \theta^{(0)}]$, unde $\theta^{(0)} \stackrel{not.}{=} (\theta_A^{(0)}, \theta_B^{(0)})$.**

Răspuns (iv)

Întrucât variabilele $Z_{i,A}$ au valori boolene (0 sau 1), rezultă că

$$\begin{aligned} E[Z_{i,A} \,|\, X_i, \theta^{(0)}] &= 0 \cdot P(Z_{i,A} = 0 \,|\, X_i, \theta^{(0)}) + 1 \cdot P(Z_{i,A} = 1 \,|\, X_i, \theta^{(0)}) \\ &= P(Z_{i,A} = 1 \,|\, X_i, \theta^{(0)}) \end{aligned}$$

Probabilităţile $P(Z_{i,A} = 1 \mid X, \theta^{(0)}) = P(Z_{i,A} = 1 \mid X_i, \theta^{(0)})$, pentru $i = 1, \ldots, 5$, se pot calcula folosind teorema lui Bayes:

$$\begin{aligned} P(Z_{i,A} = 1 \mid X_i, \theta^{(0)}) &= \frac{P(X_i \,|\, Z_{i,A} = 1, \theta^{(0)}) \cdot P(Z_{i,A} = 1 \,|\, \theta^{(0)})}{\sum_{j \in \{0,1\}} P(X_i \,|\, Z_{i,A} = j, \theta^{(0)}) \cdot P(Z_{i,A} = j \,|\, \theta^{(0)})} \\ &= \frac{P(X_i \mid Z_{i,A} = 1, \theta_A^{(0)})}{P(X_i \mid Z_{i,A} = 1, \theta_A^{(0)}) + P(X_i \mid Z_{i,B} = 1, \theta_B^{(0)})} \end{aligned}$$

S-a ţinut cont că $P(Z_{i,A} = 1 \mid \theta^{(0)}) = P(Z_{i,B} = 1 \mid \theta^{(0)}) = 1/2$ (a se vedea enunţul).

De exemplu, pentru $i = 1$ vom avea:

$$E[Z_{A,1} \mid X_1, \theta^{(0)}] \;=\; \frac{0.6^5(1-0.6)^5}{0.6^5(1-0.6)^5 + 0.5^5(1-0.5)^5} = \frac{1}{1 + \left(\dfrac{0.25}{0.24}\right)^5} \approx 0.45$$

Similar cu $E[Z_{A,1} \mid X_1, \theta^{(0)}]$ se calculează şi celelalte medii $E[Z_{i,A} \mid X_i, \theta^{(0)}]$ pentru $i = 2, \ldots, 5$ şi $E[Z_{i,B} \mid X_i, \theta^{(0)}]$ pentru $i = 1, \ldots, 5$.
Am înregistrat aceste valori/medii în cel de-al doilea tabel din *Nota* de mai sus.

***Observaţie*: Se poate ţine cont că, de îndată ce s-a calculat $E[Z_{i,A} \mid X_i, \theta^{(0)}]$, se poate obţine imediat şi $E[Z_{i,B} \mid X_i, \theta^{(0)}] = 1 - E[Z_{i,A} \mid X_i, \theta^{(0)}]$, fiindcă $Z_{i,A} + Z_{i,B} = 1$.**

v. **Calculaţi media funcţiei de log-verosimilitate a datelor complete, X (observabile) şi Z (neobservabile):**

$$L_2(\theta_A, \theta_B) \stackrel{def.}{=} E_{P(Z|X,\theta^{(0)})}[\log P(X, Z \mid \theta)],$$

unde $\theta \stackrel{not.}{=} (\theta_A, \theta_B)$ şi $\theta^{(0)} \stackrel{not.}{=} (\theta_A^{(0)}, \theta_B^{(0)})$.

Semnificaţia **notaţiei de mai sus este următoarea:**

funcţia $L_2(\theta_A, \theta_B)$ este o medie a variabilei aleatoare reprezentată de log-verosimilitatea datelor complete (observabile şi, respectiv, neobservabile), iar această medie se calculează în raport cu distribuţia probabilistă condiţională a datelor neobservabile, $P(Z \mid X, \theta^{(0)})$.

Observaţie: **La elaborarea calculului, veţi folosi mai întâi proprietatea de liniaritate a mediilor variabilelor aleatoare, şi apoi rezultatele de la punctul *iv.***

Răspuns:

Media funcţiei de log-verosimilitate a datelor complete, $L_2(\theta_A, \theta_B)$, se calculează astfel:

$$
\begin{aligned}
L_2(\theta_A, \theta_B) &\stackrel{def.}{=} E_{P(Z|X,\theta^{(0)})}[\log P(X, Z \mid \theta)] \\
&\stackrel{indep.\ cdt.}{=} E_{P(Z|X,\theta^{(0)})}[\log \prod_{i=1}^{5} P(X_i, Z_{i,A}, Z_{i,B} \mid \theta_A, \theta_B)] \\
&\stackrel{reg.\ de\ mult.}{=} E_{P(Z|X,\theta^{(0)})}[\log \prod_{i=1}^{5} P(X_i \mid Z_{i,A}, Z_{i,B}; \theta_A, \theta_B) \cdot P(Z_{i,A}, Z_{i,B} \mid \theta_A, \theta_B)]
\end{aligned}
$$

În continuare, omiţând din nou distribuţia probabilistă în raport cu care se calculează media aceasta întrucât ea poate fi subînţeleasă, vom scrie:

$$
\begin{aligned}
&L_2(\theta_A, \theta_B) \\
&\quad = E[\log \prod_{i=1}^{5} \cdot(\theta_A^{Z_{i,A}})^{X_i} \cdot [(1-\theta_A)^{Z_{i,A}}]^{10-X_i} \cdot (\theta_B^{Z_{i,B}})^{X_i} \cdot [(1-\theta_B)^{Z_{i,B}}]^{10-X_i} \cdot \frac{1}{2}\,] \\
&\quad = E[\sum_{i=1}^{5} [X_i \cdot Z_{i,A} \cdot \log\theta_A + (10 - X_i) \cdot Z_{i,A} \cdot \log(1-\theta_A) + \\
&\qquad\qquad X_i \cdot Z_{i,B} \cdot \log\theta_B + (10 - X_i) \cdot Z_{i,B} \cdot \log(1-\theta_B) - \log 2]\,] \\
&\quad \overset{lin.\ med.}{=} \sum_{i=1}^{5} [X_i \cdot E[Z_{i,A}] \cdot \log\theta_A + (10 - X_i) \cdot E[Z_{i,A}] \cdot \log(1-\theta_A) + \\
&\qquad\qquad X_i \cdot E[Z_{i,B}] \cdot \log\theta_B + (10 - X_i) \cdot E[Z_{i,B}] \cdot \log(1-\theta_B) - \log 2] \\
&\quad = \sum_{i=1}^{5} \log[\theta_A^{X_i \cdot E[Z_{i,A}]} \cdot (1-\theta_A)^{(10-X_i)\cdot E[Z_{i,A}]} \cdot \theta_B^{X_i \cdot E[Z_{i,B}]} \cdot (1-\theta_B)^{(10-X_i)\cdot E[Z_{i,B}]} \cdot \frac{1}{2}] \\
&\quad = \log(\theta_A^{2.2} \cdot (1-\theta_A)^{2.2} \cdot \theta_B^{2.8} \cdot (1-\theta_B)^{2.8} \cdot \ldots \cdot \theta_A^{4.5} \cdot (1-\theta_A)^{1.9} \cdot \theta_B^{2.5} \cdot (1-\theta_B)^{1.1} \cdot \frac{1}{2}).
\end{aligned}
$$

La ultima egalitate de mai sus, cantităţile fracţionare provin din calculele simple $X_1 \cdot E[Z_{1,A} \mid X_1, \theta] \approx 2.2$**,** $X_1 \cdot E[Z_{1,B} \mid X_1, \theta] \approx 2.8$**,** $\ldots$**,** $X_5 \cdot E[Z_{1,A} \mid X_5, \theta] \approx 4.5$**,** $X_5 \cdot E[Z_{1,B} \mid X_5, \theta] \approx 2.5$ **(a se vedea tabelele din cadrul *Notei* precedente).**

***vi.* Calculaţi $\theta_A^{(1)} \stackrel{not.}{=} \arg\max_{\theta_A} L_2(\theta_A, \theta_B)$ şi $\theta_B^{(1)} \stackrel{not.}{=} \arg\max_{\theta_B} L_2(\theta_A, \theta_B)$.**

Răspuns:

Valorile parametrilor θ_A şi θ_B pentru care se atinge maximul mediei funcţiei de log-verosimilitate a datelor complete se obţin cu ajutorul derivatelor parţiale de ordinul întâi:[a]

$$
\begin{aligned}
& \frac{\partial L_2(\theta_A, \theta_B)}{\partial \theta_A} = 0 \\
\Rightarrow\ & \frac{\partial}{\partial \theta_A}(2.2 \log \theta_A + 2.2 \log(1-\theta_A) + \ldots + 4.5 \log \theta_A + 1.9 \log(1-\theta_A)) = 0 \\
\Rightarrow\ & \frac{2.2}{\theta_A} - \frac{2.2}{1-\theta_A} + \ldots + \frac{4.5}{\theta_A} - \frac{1.9}{1-\theta_A} = 0 \Rightarrow \ldots \Rightarrow \theta_A^{(1)} \approx 0.71.
\end{aligned}
$$

Similar, vom obţine $\theta_B^{(1)} \approx 0.58$.

[a] Este imediat că derivatele de ordinul al doilea au valori negative pe tot domeniul de definiţie.

C. Formalizaţi paşii E şi M ai algoritmului EM pentru estimarea parametrilor θ_A şi θ_B în condiţiile de la punctul B.

Răspuns:

Formulele care se folosesc în cadrul algoritmului EM pentru rezolvarea problemei date — i.e. estimarea parametrilor θ_A şi θ_B atunci când variabilele Z_i sunt neobservabile —, se elaborează/deduc astfel:

Pasul E:

$$E[Z_{i,A} \mid X,\theta] = P(Z_{i,A}=1 \mid X,\theta) = P(Z_{i,A}=1 \mid X_i,\theta)$$

$$\stackrel{T.\ B.}{=} \frac{P(X_i \mid Z_{i,A}=1;\theta)\cdot \overbrace{P(Z_{i,A}=1 \mid \theta)}^{1/2}}{P(X_i \mid Z_{i,A}=1;\theta)\cdot P(Z_{i,A}=1 \mid \theta) + P(X_i \mid Z_{i,B}=1;\theta)\cdot \underbrace{P(Z_{i,B}=1 \mid \theta)}_{1/2}}$$

$$= \frac{P(X_i \mid Z_{i,A}=1;\theta)}{P(X_i \mid Z_{i,A}=1;\theta) + P(X_i \mid Z_{i,B}=1;\theta)}$$

$$= \frac{\theta_A^{X_i}(1-\theta_A)^{10-X_i}}{\theta_A^{X_i}(1-\theta_A)^{10-X_i} + \theta_B^{X_i}(1-\theta_B)^{10-X_i}}$$

Similar, vom obţine:

$$E[Z_{i,B} \mid X,\theta] = \frac{\theta_B^{X_i}(1-\theta_B)^{10-X_i}}{\theta_A^{X_i}(1-\theta_A)^{10-X_i} + \theta_B^{X_i}(1-\theta_B)^{10-X_i}}$$

Notând cu

- x_i **valoarea variabilei** X_i,
- $\theta_A^{(t)}$ **şi respectiv** $\theta_B^{(t)}$ **estimările parametrilor** θ_A **şi** θ_B **la iteraţia** t **a algoritmului EM,**
- $p_{i,A}^{(t+1)}$ **şi respectiv** $p_{i,B}^{(t+1)}$, **mediile** $E[Z_{i,A} \mid X_i, \theta_A^{(t)}]$ **şi** $E[Z_{i,B} \mid X_i, \theta_B^{(t)}]$,

vom avea:

$$
\begin{aligned}
p_{i,A}^{(t+1)} &= \frac{(\theta_A^{(t)})^{x_i}(10-\theta_A^{(t)})^{10-x_i}}{(\theta_A^{(t)})^{x_i}(10-\theta_A^{(t)})^{10-x_i}+(\theta_B^{(t)})^{x_i}(10-\theta_B^{(t)})^{10-x_i}} \\
p_{i,B}^{(t+1)} &= \frac{(\theta_B^{(t)})^{x_i}(10-\theta_B^{(t)})^{10-x_i}}{(\theta_A^{(t)})^{x_i}(10-\theta_A^{(t)})^{10-x_i}+(\theta_B^{(t)})^{x_i}(10-\theta_B^{(t)})^{10-x_i}}
\end{aligned}
$$

Pasul M:

Ca şi mai înainte, în formulele de mai jos vom folosi notaţiile simplificate $E[Z_{i,A}] \stackrel{not.}{=} E[Z_{i,A} \mid X_i, \theta^{(t)}]$ şi $E[Z_{i,B}] \stackrel{not.}{=} E[Z_{i,B} \mid X_i, \theta^{(t)}]$.

Cu aceste notaţii, procedând similar cu calculul de la partea B, punctul v, vom avea:

$$L_2(\theta_A, \theta_B) = \log \prod_{i=1}^{5} \theta_A^{x_i E[Z_{i,A}]} (1-\theta_A)^{(10-x_i)E[Z_{i,A}]} \; \theta_B^{x_i E[Z_{i,B}]} (1-\theta_B)^{(10-x_i)E[Z_{i,B}]}$$

Prin urmare,

$$\frac{\partial}{\partial \theta_A} L_2(\theta_A, \theta_B) = 0 \Rightarrow \frac{1}{\theta_A} \sum_{i=1}^{5} x_i E[Z_{i,A}] = \frac{1}{1-\theta_A} \sum_{i=1}^{5} (10 - x_i) E[Z_{i,A}]$$

$$\Rightarrow \quad (1-\theta_A) \sum_{i=1}^{5} x_i E[Z_{i,A}] = \theta_A \sum_{i=1}^{5} (10 - x_i) E[Z_{i,A}]$$

$$\Rightarrow \quad \sum_{i=1}^{5} x_i E[Z_{i,A}] = 10\, \theta_A \sum_{i=1}^{5} E[Z_{i,A}]$$

$$\Rightarrow \quad \theta_A = \frac{\sum_{i=1}^{5} x_i \; E[Z_{i,A}]}{10 \sum_{i=1}^{5} E[Z_{i,A}]} \quad \textbf{şi, similar,} \quad \theta_B = \frac{\sum_{i=1}^{5} x_i \; E[Z_{i,B}]}{10 \sum_{i=1}^{5} E[Z_{i,B}]}$$

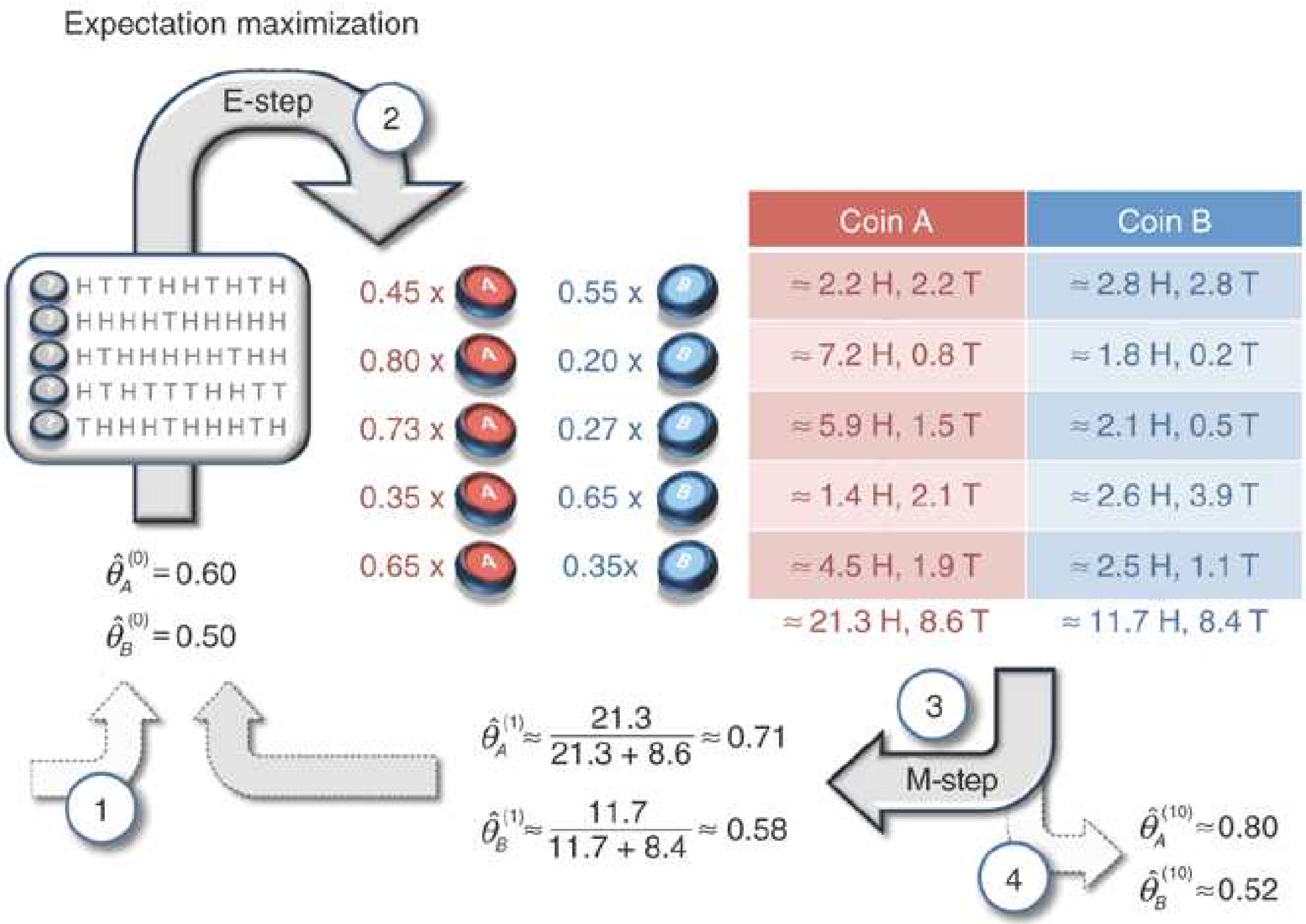
Aşadar, la pasul M al algoritmului EM vom avea:

$$\theta_A^{(t+1)} = \frac{\sum_{i=1}^{5} x_i \; p_{i,A}^{(t+1)}}{10 \sum_{i=1}^{5} p_{i,A}^{(t+1)}} \quad \textbf{şi} \quad \theta_B^{(t+1)} = \frac{\sum_{i=1}^{5} x_i \; p_{i,B}^{(t+1)}}{10 \sum_{i=1}^{5} p_{i,B}^{(t+1)}}$$

Observaţie

Implementând algoritmul EM cu relaţiile obţinute pentru pasul E şi pasul M, după execuţia a 10 iteraţii se vor obţine valorile $\theta_A^{(10)} \approx 0.80$ şi $\theta_B^{(10)} \approx 0.52$.

Este interesant de observat că estimarea obţinută pentru parametrul θ_A este acum la acelaşi nivel cu cea obţinută prin metoda verosimilităţii maxime (MLE) în cazul observării tuturor variabilelor (0.80, vezi rezolvarea de la partea A, punctul i), iar estimarea obţinută pentru parametrul θ_B a coborât de la valorea 0.58 care a fost obţinută la prima iteraţie a algoritmului EM la o valoare (0.52) care este considerabil mai apropiată de estimarea prin metoda MLE (0.45).


Expectation maximization
E-step
2
H T T T H H T H T H
H H H H T H H H H H
H T H H H H H T H H
H T H T T T H H T T
T H H H T H H H T H
0.45 x
0.80 x
0.73 x
0.35 x
0.65 x
0.55 x
0.20 x
0.27 x
0.65 x
0.35x
Coin A
Coin B
≈ 2.2 H, 2.2 T
≈ 2.8 H, 2.8 T
≈ 7.2 H, 0.8 T
≈ 1.8 H, 0.2 T
≈ 5.9 H, 1.5 T
≈ 2.1 H, 0.5 T
≈ 1.4 H, 2.1 T
≈ 2.6 H, 3.9 T
≈ 4.5 H, 1.9 T
≈ 2.5 H, 1.1 T
≈ 21.3 H, 8.6 T
≈ 11.7 H, 8.4 T
$\hat{\theta}_A^{(0)} = 0.60$
$\hat{\theta}_B^{(0)} = 0.50$
1
$\hat{\theta}_A^{(1)} \approx \frac{21.3}{21.3 + 8.6} \approx 0.71$
$\hat{\theta}_B^{(1)} \approx \frac{11.7}{11.7 + 8.4} \approx 0.58$
3
M-step
4
$\hat{\theta}_A^{(10)} \approx 0.80$
$\hat{\theta}_B^{(10)} \approx 0.52$

The EM algorithm for solving a mixture of K categorical distributions

CMU, 2015 spring, Tom Mitchell, Nina Balcan, HW6, pr. 1

Mixture models **are helpful for modelling unknown subpopulations in data. If we have a collection of data points** $X = \{X_1, \ldots, X_n\}$**, where each** X_i **is [independently] drawn from one of** K **possible distributions, we can introduce a discrete-valued random variable** $Z_i \in \{1, \ldots, K\}$ **that indicates which distribution** X_i **is drawn from.**

This exercise deals with the ***categorical mixture model*****, where each observation** X_i **is a discrete value drawn from a categorical distribution. The** ***parameter*** **for a categorical distribution is a** K**-dimensional vector** π **that lists the probability of each of** K **possible values (therefore,** $\sum_k \pi_k = 1$**).**

For ***example*****, suppose our categorical mixture model has 3 underlying distributions. Then, each** Z_i **could take on one of three values,** $\{1, 2, 3\}$**, with respective probabilities** π_1, π_2, π_3**; equivalently,** $Z_i \sim \mathit{Categorical}(\pi)$**, where** $\pi = (\pi_1, \pi_2, \pi_3) \in \mathbb{R}^3_+$**, with** $\pi_1 + \pi_2 + \pi_3 = 1$**. The** ***observation*** X_i **is then generated from another categorical distribution, depending on the value of** Z_i**.**

The ***generative process*** **for a** ***categorical mixture model*** **is summarized as follows:**

$$Z_i \sim \mathit{Categorical}(\pi)$$
$$X_i \sim \mathit{Categorical}(\theta_{Z_i})$$

For this model, where we observe X but not Z, we want to *learn* the *parameters* of the K categorical components $\Theta = \{\pi, \theta_1, \ldots, \theta_K\}$, where each $\theta_k \in \mathbb{R}^M$ is the parameter for the categorical distribution associated with the k-th mixture component. (It implies that each X_i can take on one of M possible values). We will use the EM algorithm to accomplish this.

A *note* on notation and a *hint*:

When working with categorical distributionsit is helpful to use *indicator functions*. The indicator function $1_{\{x=j\}}$ has the value 1 when $x = j$ and 0 otherwise. With this notation, we can express the probability that a random variable drawn from a categorical distribution (e.g., $Y \sim \mathit{Categorical}(\phi)$, where $\phi \in \mathbb{R}^N$) takes on a particular value as

$$P(Y) = \prod_{i=1}^{N} \phi_i^{1_{\{Y=i\}}}.$$

a. What is the *joint distribution* $P(X, Z; \Theta)$?

b. What is the *posterior distribution* of the *latent variables*, $P(Z|X; \Theta)$?

c. Compute the *expectation of the log-likelihood*,

$$Q(\Theta|\Theta') = E_{Z|X;\Theta'}[\log P(X, Z; \Theta')]$$

d. What is the update step for Θ? That is, what Θ maximizes $Q(\Theta|\Theta')$?

***Hint*: Make sure your solution enforces the constraint that parameters to categorical distributions must sum to 1. Lagrange multipliers are a great way to solve constrained optimization problems.**

Answer:

a.

$$
\begin{aligned}
P(X, Z; \Theta) &= \prod_{i=1}^{n} P(X_i, Z_i; \Theta) = \prod_{i=1}^{n} P(X_i | Z_i; \Theta) P(Z_i; \Theta) \\
&= \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\pi_k \prod_{j=1}^{M} \theta_{kj}^{1_{\{X_i = v_j\}}} \right]^{1_{\{Z_i = k\}}}
\end{aligned}
$$

b.

$$
P(Z_i = k | X_i; \Theta) \overset{Bayes\ F.}{=} \frac{P(X_i | Z_i = k; \Theta) P(Z_i = k; \Theta)}{P(X_i; \Theta)}
$$

$$
= \frac{P(X_i | Z_i = k; \Theta) P(Z_i = k; \Theta)}{\sum_{l=1}^{K} P(Z_i = l; \Theta) P(X_i | Z_i = l; \Theta)} = \frac{\prod_{k=1}^{K} \left[\pi_k \prod_{j=1}^{M} \theta_{kj}^{1_{\{X_i = v_j\}}} \right]^{1_{\{Z_i = k\}}}}{\sum_{l=1}^{K} \pi_l \prod_{j=1}^{M} \theta_{lj}^{1_{\{X_i = v_j\}}}}
$$

c.

$$
\begin{aligned}
Q(\Theta|\Theta') &\overset{def.}{=} E_{Z|X;\Theta'}[\log P(X,Z;\Theta)] \\
&\overset{a.}{=} E_{Z|X;\Theta'}\left[\log \prod_{i=1}^{n}\prod_{k=1}^{K}\left[\pi_k \prod_{j=1}^{M} \theta_{kj}^{1_{\{X_i=v_j\}}}\right]^{1_{\{Z_i=k\}}}\right] \\
&= E_{Z|X;\Theta'}\left[\sum_{i=1}^{n}\sum_{k=1}^{K} 1_{\{Z_i=k\}}\left(\log \pi_k + \sum_{j=1}^{M} 1_{\{X_i=v_j\}} \log \theta_{kj}\right)\right] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \underbrace{E_{Z|X;\Theta'}[1_{\{Z_i=k\}}]}_{P(Z_i=k|X_i;\Theta)} \left(\log \pi_k + \sum_{j=1}^{M} 1_{\{X_i=v_j\}} \log \theta_{kj}\right)
\end{aligned}
$$

Using

$$
\gamma_{ik} \overset{not.}{=} E_{Z|X;\Theta'}[1_{\{Z_i=k\}}] = P(Z_i=k|X_i;\Theta') \overset{b.}{=} \frac{\pi'_k \prod_{j=1}^{M} (\theta'_{kj})^{1_{\{X_i=v_j\}}}}{\sum_{l=1}^{K} \pi'_l \prod_{j=1}^{M} (\theta'_{lj})^{1_{\{X_i=v_j\}}}},
$$

leads to

$$
Q(\Theta|\Theta') = \sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_{ik} \left(\log \pi_k + \sum_{j=1}^{M} 1_{\{X_i=v_j\}} \log \theta_{kj}\right)
$$

d. To find the updates rules for the parameters, we solve the following *constrained optimization problem*:

$$\begin{array}{rl} & \max_{\pi,\theta_{1:K}} Q(\Theta|\Theta') \\ \text{subject to} & \sum_{k=1}^{K} \pi_k = 1 \\ & \sum_{j=1}^{M} \theta_{kj} = 1, \forall k = 1, \ldots, K \end{array}$$

We introduce a *Lagrange multiplier* for each constraint associated to this problem and form the *Lagrangian functional*:

$$\mathcal{L} = Q(\Theta|\Theta') + \lambda_\pi \left(1 - \sum_{k=1}^{K} \pi_k\right) + \sum_{k=1}^{K} \lambda_{\theta_k} \left(1 - \sum_{j=1}^{M} \theta_{kj}\right)$$

Now we will differentiate the Lagrangian functional with respect to each parameter we want to optimize, set the derivative to zero, and solve for the optimal value.

We'll start with optimizing for π, the parameter to the categorical distribution for the latent variables Z.

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = 0 \Leftrightarrow \sum_{i=1}^{n} \gamma_{ik} \cdot \frac{1}{\pi_k} - \lambda_\pi = 0 \Leftrightarrow \pi_k^* = \frac{1}{\lambda_\pi} \sum_{i=1}^{n} \gamma_{ik}$$

We can find the value of the Lagrange multiplier λ_π by enforcing the constraint:

$$\sum_{k=1}^{K} \pi_k^* = 1 \Leftrightarrow \sum_{k=1}^{K} \frac{1}{\lambda_\pi} \sum_{i=1}^{n} \gamma_{ik} = 1 \Leftrightarrow \frac{1}{\lambda_\pi} \sum_{k=1}^{K} \sum_{i=1}^{n} \gamma_{ik} = 1 \Leftrightarrow$$

$$\lambda_\pi = \sum_{k=1}^{K} \sum_{i=1}^{n} \gamma_{ik} \Leftrightarrow \lambda_\pi = \sum_{i=1}^{n} \underbrace{\sum_{k=1}^{K} \gamma_{ik}}_{1} \Leftrightarrow \lambda_\pi = n$$

Substituting this value for λ_π back into the expression of π_k^* gives the answer:

$$\pi_k^* = \frac{1}{n} \sum_{i=1}^{n} \gamma_{ik}$$

The process for optimizing the observation parameters θ_{kj} is similar. First, we differentiate the Lagrangian $\mathcal{L}$ with respect to θ_{kj}, set to zero, and [finally] solve.

$$\frac{\partial \mathcal{L}}{\partial \theta_{kj}} = 0 \Leftrightarrow \sum_{i=1}^{n} \gamma_{ik} \; 1_{\{X_i = v_j\}} \cdot \frac{1}{\theta_{kj}} - \lambda_{\theta_k} = 0$$

$$\Rightarrow \theta^*_{kj} = \frac{1}{\lambda_{\theta_k}} \sum_{i=1}^{n} \gamma_{ik} \; 1_{\{X_i = v_j\}}$$

Again, we can find the value of the Lagrange multiplier λ_{θ_k} by enforcing the summation constraint:

$$\sum_{j=1}^{M} \theta^*_{kj} = 1 \Leftrightarrow \sum_{j=1}^{M} \frac{1}{\lambda_{\theta_k}} \sum_{i=1}^{n} \gamma_{ik} \; 1_{\{X_i = v_j\}} = 1 \Leftrightarrow \frac{1}{\lambda_{\theta_k}} \sum_{j=1}^{M} \sum_{i=1}^{n} \gamma_{ik} \; 1_{\{X_i = v_j\}} = 1$$

$$\Leftrightarrow \lambda_{\theta_k} = \sum_{j=1}^{M} \sum_{i=1}^{n} \gamma_{ik} \; 1_{\{X_i = v_j\}}$$

Substituting this value for λ_{θ_k} back into the expression of θ^*_{kj} gives the answer:

$$\theta^*_{kj} = \frac{\sum_{i=1}^{n} \gamma_{ik} \; 1_{\{X_i = v_j\}}}{\sum_{l=1}^{M} \sum_{i=1}^{n} \gamma_{ik} \; 1_{\{X_i = v_l\}}} = \frac{\sum_{i=1}^{n} \gamma_{ik} \; 1_{\{X_i = v_j\}}}{\sum_{i=1}^{n} \sum_{l=1}^{M} \gamma_{ik} \; 1_{\{X_i = v_l\}}}$$

The EM algorithm for solving a mixture of K categorical distributions, applied to the problem of Word Sense Disambiguation,

i.e. identifying the semantic domains associated to words in a text document

CMU, 2012 fall, Eric Xing, Aarti Singh, HW3, pr. 3

The objective of this exercise is to derive the update equations of the EM algorithm for optimizing the latent variables [designating the semantic domains] involved in generating a text document.
Each *word* will be seen as a random variable w that can take values $1, \ldots, V$ from the *vocabulary* of words. **In fact, we will denote each w by an array of V components such that $w(i) = 1$ if w takes the value of the i-th word in the vocabulary. Hence, $\sum_{i=1}^{V} w(i) = 1$.**
Given a *document* containing words w_j, $j = 1, \ldots, N$, where N is the length of the document, we will assume that these words are generated from a mixture of K discrete *topics*:

$$P(w) \ = \ \sum_{m=1}^{K} \pi_m \, P(w|\mu_m) \text{ and } P(w|\mu_m) = \prod_{i=1}^{V} \mu_m(i)^{w(i)},$$

where
π_m denotes the prior [probability] for the latent topic variable $t = m$,
$\mu_k \stackrel{not.}{=} (\mu_k(1), \ldots, \mu_k(i), \ldots, \mu_k(V))$, with $\mu_k(i) \geq 0$ for $i = 1, \ldots, V$ and $\sum_{i=1}^{V} \mu_m(i) = 1$ for each $k = 1, \ldots, K$, and
$\mu_m(i) \stackrel{not.}{=} P(w(i) = 1 | t = m)$.

a. In the expectation step, for each word w_j, compute

$$F_j(t) \stackrel{not.}{=} P(t|w_j;\theta),$$

the probability that w_j belongs to each of the K topics, where θ is the set of parameters of this mixture model.

Answer:

$$\begin{aligned}
F_j(t_j = m) &\stackrel{not.}{=} P(t_j = m|w_j;\theta) \stackrel{Bayes\ T.}{=} \frac{P(w_j|t_j = m;\theta)\ P(t_j = m|\theta)}{P(w_j|\theta)} \\
&= \frac{\pi_m\ P(w_j|\mu_m)}{\sum_{m'=1}^{K} \pi_{m'}\ P(w_j|\mu_{m'})} = \frac{\pi_m \prod_{l=1}^{V} \mu_m(l)^{w_j(l)}}{\sum_{m'=1}^{K} \pi_{m'} \prod_{l=1}^{V} \mu_{m'}(l)^{w_j(l)}}
\end{aligned}$$

b. In the maximization step, compute θ that maximizes the log-likelihood of the data

$$l(w;\theta) \stackrel{not.}{=} \log \prod_{j=1}^{N} P(w_j;\theta)$$

***Hint*: Suming over the latent topic variable, we can write $l(w;\theta)$ as**

$$l(w;\theta) = \sum_{j=1}^{N} \log \sum_{t} P(w_j, t;\theta) = \sum_{j=1}^{N} \log F_j(t) \frac{\sum_t P(w_j, t;\theta)}{F_j(t)}$$

Further on, using Jensen's inequality (see pr. EM-2) we get:

$$l(w;\theta) \quad \geq \quad \sum_{j=1}^{N} \sum_{t} F_j(t) \log \frac{P(w_j, t;\theta)}{F_j(t)} = \sum_{j=1}^{N} \sum_{t} F_j(t) \log P(w_j, t;\theta) + H(F_j)$$

Hence compute θ as:

$$\operatorname*{argmax}_{\theta} \sum_{j=1}^{N} \sum_{t} F_j(t) \log P(w_j, t;\theta)$$

Answer:

$$
\begin{aligned}
\theta &= \operatorname*{argmax}_{\theta} \sum_{j=1}^{N} \sum_{m=1}^{K} F_j(t_j = m) \log P(w_j, t_j = m|\theta) \\
&= \operatorname*{argmax}_{\theta} \sum_{j=1}^{N} \sum_{m=1}^{K} F_j(t_j = m) \log P(w_j|t_j = m; \theta)\ P(t_j = m|\theta) \\
&= \operatorname*{argmax}_{\theta} \sum_{j=1}^{N} \sum_{m=1}^{K} F_j(t_j = m) \log \left(\pi_m \prod_{l=1}^{V} \mu_m(l)^{w_j(l)} \right) \\
&= \operatorname*{argmax}_{\theta} \sum_{j=1}^{N} \sum_{m=1}^{K} [F_j(t_j = m) \log \pi_m + F_j(t_j = m) \sum_{l=1}^{V} \log \mu_m(l)^{w_j(l)}] \\
&= \operatorname*{argmax}_{\theta} \sum_{j=1}^{N} \sum_{m=1}^{K} [F_j(t_j = m) \log \pi_m + F_j(t_j = m) \sum_{l=1}^{V} w_j(l) \log \mu_m(l)] \qquad (3)
\end{aligned}
$$

To optimize $\mu_m(l)$:

After eliminating from (3) the terms which are constant with respect to μ_m we get:

$$\sum_{j=1}^{N} F_j(t_j = m) \sum_{l=1}^{V} w_j(l) \log \mu_m(l)$$

We will use a Lagrangian to constrain μ_m to be a probability distribution:

$$\mathcal{L}(\mu_m(l)) = \sum_{j=1}^{N} F_j(t_j = m) \sum_{l=1}^{V} w_j(l) \log \mu_m(l) + \beta \left(\sum_{l=1}^{V} \mu_m(l) - 1 \right)$$

Then, solving for $\mu_m(l)$:

$$\frac{\partial}{\partial \mu_m(l)} \mathcal{L}(\mu_m(l)) = 0 \Leftrightarrow \sum_{j=1}^{N} F_j(t_j = m) \frac{w_j(l)}{\mu_m(l)} + \beta = 0$$

$$\Leftrightarrow \quad \frac{1}{\mu_m(l)} \sum_{j=1}^{N} F_j(t_j = m)\, w_j(l) + \beta = 0 \Leftrightarrow \frac{1}{\mu_m(l)} = \frac{-\beta}{\sum_{j=1}^{N} F_j(t_j = m)\, w_j(l)}$$

$$\Leftrightarrow \quad \mu_m(l) = \frac{\sum_{j=1}^{N} F_j(t_j = m)\, w_j(l)}{-\beta} \qquad (4)$$

Knowing that $\sum_{l=1}^{V} \mu_m(l) = 1$, we have:

$$\sum_{l=1}^{V} \frac{\sum_{j=1}^{N} F_j(t_j = m)\, w_j(l)}{-\beta} = 1 \quad \Leftrightarrow \quad -\beta = \sum_{l=1}^{V} \sum_{j=1}^{N} F_j(t_j = m)\, w_j(l)$$

Hence, substituting for $-\beta$ in (4), we get:

$$\begin{aligned}
\mu_m(l) &= \frac{\sum_{j=1}^{N} F_j(t_j = m)\, w_j(l)}{\sum_{l=1}^{V} \sum_{j=1}^{N} F_j(t_j = m)\, w_j(l)} = \frac{\sum_{j=1}^{N} F_j(t_j = m)\, w_j(l)}{\sum_{j=1}^{N} \sum_{l=1}^{V} F_j(t_j = m)\, w_j(l)} \\
&= \frac{\sum_{j=1}^{N} F_j(t_j = m)\, w_j(l)}{\sum_{j=1}^{N} F_j(t_j = m) \underbrace{\sum_{l=1}^{V} w_j(l)}_{1}} = \frac{\sum_{j=1}^{N} F_j(t_j = m)\, w_j(l)}{\sum_{j=1}^{N} F_j(t_j = m)}
\end{aligned}$$

Note: Intuitively the last expression can be interpreted as the portion [of word[occurrence]s] that had $w(l) = 1$ among all words [in the given document] which are deemed to belong to cluster m.

To optimize π_m, we proceed similarly:

We begin by removing from (3) the terms that are constant with respect to π_m, thus getting:

$$\sum_{j=1}^{N} F_j(t_j = m) \log \pi_m$$

So, using the Lagrangian with the constraint that $\sum_{m=1}^{K} \pi_m = 1$

$$\mathcal{L}(\pi_m) = \sum_{j=1}^{N} F_j(t_j = m) \log \pi_m + \beta \left(\sum_{m=1}^{K} \pi_m - 1 \right),$$

and solving for π_m, we have:

$$\frac{\partial}{\partial \pi_m} \mathcal{L}(\pi_m) = 0 \Leftrightarrow \sum_{j=1}^{N} \frac{F_j(t_j = m)}{\pi_m} + \beta = 0 \Leftrightarrow \frac{1}{\pi_m} \sum_{j=1}^{N} F_j(t_j = m) = -\beta$$

$$\Leftrightarrow \pi_m = \frac{\sum_{j=1}^{N} F_j(t_j = m)}{-\beta} \tag{5}$$

Since $\sum_{m=1}^K \pi_m = 1$, it gives us:

$$\sum_{m=1}^{K} \frac{\sum_{j=1}^N F_j(t_j = m)}{-\beta} = 1 \;\Leftrightarrow\; \frac{1}{-\beta} \sum_{m=1}^{K} \sum_{j=1}^{N} F_j(t_j = m) = 1$$

$$\Leftrightarrow\; -\beta = \sum_{m=1}^{K} \sum_{j=1}^{N} F_j(t_j = m)$$

Substituting for $-\beta$ in (5), we get:

$$\pi_m = \frac{\sum_{j=1}^N F_j(t_j = m)}{\sum_{m=1}^K \sum_{j=1}^N F_j(t_j = m)} = \frac{\sum_{j=1}^N F_j(t_j = m)}{\sum_{j=1}^N \underbrace{\sum_{m=1}^{K} F_j(t_j = m)}_{1}} = \frac{\sum_{j=1}^N F_j(t_j = m)}{N}$$

Note: Intuitively the last expression can be interpreted as the portion [of word[occurrence]s] that belongs to cluster m among the total of N words.

To summarize:

E step:

$$F_j(t_j = m) = \frac{\pi_m \prod_{l=1}^{V} \mu_m(l)^{w_j(l)}}{\sum_{m'=1}^{K} \pi_{m'} \prod_{l=1}^{V} \mu_{m'}(l)^{w_j(l)}} \text{ for } j = 1, \ldots, N \text{ and } m = 1, \ldots, K$$

M step:

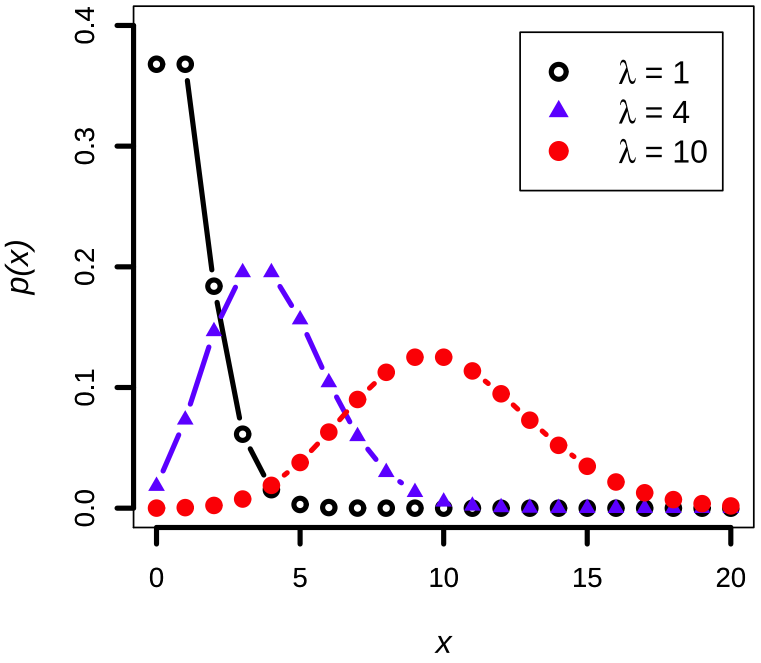
$$\mu_m(l) = \frac{\sum_{j=1}^{N} F_j(t_j = m)\, w_j(l)}{\sum_{j=1}^{N} F_j(t_j = m)} \text{ for } m = 1, \ldots, K \text{ and } l = 1, \ldots, V$$

$$\pi_m = \frac{\sum_{j=1}^{N} F_j(t_j = m)}{N} \text{ for } m = 1, \ldots, K$$

The EM algorithm for solving mixtures of Poisson distributions

University of Utah, 2008 fall, Hal Daumé III, HW9, pr. 1

The Poisson distribution is a distribution over positive count values; for a count x with parameter λ, the Poisson has the form $Poisson(x|\lambda) = \frac{1}{e^{\lambda}} \cdot \frac{\lambda^x}{x!}$.



It can be proven that the maximum likelihood estimation for λ given a sequence of counts $x_1, \ldots, x_n$ was simply $\frac{1}{n}\sum_{i=1}^{n} x_i$ – the mean of the counts.

Let's consider a generalization of this: *the Poisson mixture model*. This is actually used in *web server monitoring*. The number of accesses to a web server in a minute typically follows a Poisson distribution.

Suppose we have n web servers we are monitoring and we monitor each for M minutes. Thus, we have $n \times M$ counts; call $x_{i,m}$ the number of hits to web server i in minute m. Our goal is to *cluster* the web servers according to their hit frequency.

Using the EM algorithm, construct a Poisson mixture model for this problem and compute the expectations and maximization steps for this model.

Hint: **Suppose we want K clusters; let z_i be the latent variable telling us which cluster the web server i belongs to (from 1 to K). Let λ_k denote the parameter for the Poisson distribution for cluster k, and π_k the . Then, the complete data likelihood is:**

$$
\begin{aligned}
L(\bar{\lambda}, \bar{\pi}) &\stackrel{def.}{=} p(\bar{x}, \bar{z} | \bar{\lambda}, \bar{\pi}) \stackrel{i.i.d.}{=} \prod_{i=1}^{n} p(x_i, z_i | \bar{\lambda}, \bar{\pi}) = \prod_{i=1}^{n} p(x_i | z_i, \bar{\lambda}, \bar{\pi}) \cdot \underbrace{p(\underbrace{z_i}_{k} | \bar{\lambda}, \bar{\pi})}_{\pi_k} \\
&= \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\pi_k \prod_{m=1}^{M} Poisson(x_{i,m} | \lambda_k) \right]^{1_{\{z_i = k\}}}
\end{aligned}
$$

where

$\bar{x} \stackrel{not.}{=} (x_1, \ldots, x_n)$, **with** $x_i \stackrel{not.}{=} (x_{i,1}, \ldots, x_{i,M})$ **for** $i = 1, \ldots, n$,

$\bar{z} \stackrel{not.}{=} (z_1, \ldots, z_K)$, $\bar{\pi} \stackrel{not.}{=} (\pi_1, \ldots, \pi_K)$; $\bar{\lambda} \stackrel{not.}{=} (\lambda_1, \ldots, \lambda_K)$, **and**

$1_{\{z_i = k\}}$ **is one if** $z_i = k$ **and zero otherwise.**

Solution

The likelihood function:

$$L(\bar{\lambda}, \bar{\pi}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\pi_k \prod_{m=1}^{M} \frac{1}{e^{\lambda_k}} \frac{\lambda_k^{x_{i,m}}}{x_{i,m}!} \right]^{1_{\{z_i=k\}}}$$

The likelihood function:

$$\begin{aligned}
\ell(\bar{\lambda}, \bar{\pi}) &\stackrel{def.}{=} \ln L(\bar{\lambda}, \bar{\pi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} 1_{\{z_i=k\}} \left[\ln \pi_k + \sum_{m=1}^{M} \ln \frac{1}{e^{\lambda_k}} \frac{\lambda_k^{x_{i,m}}}{x_{i,m}!} \right] \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} 1_{\{z_i=k\}} \left[\ln \pi_k + \sum_{m=1}^{M} (-\lambda_k + x_{i,m} \ln \lambda_k - \ln(x_{i,m})!) \right]
\end{aligned}$$

The auxiliary function:

$$\begin{aligned}
& Q(\bar{\lambda}, \bar{\pi} | \bar{\lambda}^{(t)}, \bar{\pi}^{(t)}) \stackrel{def.}{=} E_{P(z_i=k|x_i, \bar{\lambda}^{(t)}, \bar{\pi}^{(t)})} [\ell(\bar{\lambda}, \bar{\pi})] \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} P(z_i = k | x_i, \bar{\lambda}^{(t)}, \bar{\pi}^{(t)}) \left[\ln \pi_k + \sum_{m=1}^{M} (-\lambda_k + x_{i,m} \ln \lambda_k - \ln(x_{i,m})!) \right]
\end{aligned}$$

E-step:

$$p_{ik}^{(t)} \stackrel{not.}{=} P(z_i = k|x_i, \bar{\lambda}^{(t)}, \bar{\pi}^{(t)})$$

$$\stackrel{Bayes\ F.}{=} \frac{P(x_i|z_i = k, \bar{\lambda}^{(t)}, \bar{\pi}^{(t)}) \cdot P(z_i = k|\bar{\lambda}^{(t)}, \bar{\pi}^{(t)})}{\sum_{k'=1}^{K} P(x_i|z_i = k', \bar{\lambda}^{(t)}, \bar{\pi}^{(t)}) \cdot P(z_i = k'|\bar{\lambda}^{(t)}, \bar{\pi}^{(t)})}$$

$$= \frac{\pi_k^{(t)} \cdot P(x_i|z_i = k, \bar{\lambda}^{(t)}, \bar{\pi}^{(t)})}{\sum_{k'=1}^{K} \pi_{k'}^{(t)} \cdot P(x_i|z_i = k', \bar{\lambda}^{(t)}, \bar{\pi}^{(t)})}$$

$$= \frac{\pi_k^{(t)} \cdot \prod_{m=1}^{M} Poisson(x_{i,m}|\lambda_k)}{\sum_{k'=1}^{K}(\pi_{k'}^{(t)} \cdot \prod_{m=1}^{M} Poisson(x_{i,m}|\lambda_{k'}))}$$

M-step:

Since $\sum_{k=1}^{K} \pi_k = 1$, we have to introduce the Lagrange multiplier λ:

$$\frac{\partial}{\partial \pi_k}\left[Q(\bar{\lambda}, \bar{\pi} | \bar{\lambda}^{(t)}, \bar{\pi}^{(t)}) + \lambda(1 - \sum_{k=1}^{K} \pi_k)\right] = 0 \Leftrightarrow$$

$$\sum_{i=1}^{n} \frac{p_{ik}^{(t)}}{\pi_k} - \lambda = 0 \Leftrightarrow \pi_k^{(t+1)} = \frac{1}{\lambda} \sum_{i=1}^{n} p_{ik}^{(t)}$$

It is easy to see that $\pi_k^{(t+1)}$ is indeed a maximum point for Q w.r.t. the parameter π_k.

Imposing the constraint $\sum_{k=1}^{K} \pi_k^{(t+1)} = 1$ leads to:

$$\frac{1}{\lambda} \sum_{k=1}^{K} \sum_{i=1}^{n} p_{ik}^{(t)} = 1 \Leftrightarrow \frac{1}{\lambda} \sum_{i=1}^{n} \underbrace{\sum_{k=1}^{K} p_{ik}^{(t)}}_{1} = 1 \Leftrightarrow \frac{n}{\lambda} = 1 \Leftrightarrow \lambda = n.$$

Therefore, $\pi_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} p_{ik}^{(t)}$, which is alyways a non-negative quantity.

Now, regarding the parameters λ_k:

$$\frac{\partial}{\partial \lambda_k} Q(\bar\lambda, \bar\pi | \bar\lambda^{(t)}, \bar\pi^{(t)}) = 0 \Leftrightarrow \sum_{i=1}^{n} p_{ik} \left(\sum_{m=1}^{M} \left(-1 + x_{i,m} \cdot \frac{1}{\lambda_k} \right) \right) = 0 \Leftrightarrow$$

$$\sum_{i=1}^{n} p_{ik} \left(-M + \frac{1}{\lambda_k} \sum_{m=1}^{M} x_{i,m} \right) = 0 \Leftrightarrow \frac{1}{\lambda_k} \sum_{i=1}^{n} p_{ik} \left(\sum_{m=1}^{M} x_{i,m} \right) = M \sum_{i=1}^{n} p_{ik} \Leftrightarrow$$

$$\lambda_k^{(t+1)} = \frac{\sum_{i=1}^{n} p_{ik} \left(\sum_{m=1}^{M} x_{i,m} \right)}{M \sum_{i=1}^{n} p_{ik}}$$

It can be easily shown that for each $k \in \{1, \ldots, K\}$, the value $\lambda_k^{(t+1)}$ is positive, and it maximizes Q w.r.t. the parameter λ_k.

The EM algorithm for solving a Gamma Mixture Model

G. Vegas-Sánchez-Ferrero, M. Martín-Fernández, J. Miguel Sanches

***A Gamma Mixture Model for IVUS Imaging*, 2014**

[**adapted by Liviu Ciortuz**]

The Gamma Mixture Model

$$p(x|\theta) = \sum_{k=1}^{K} \pi_k \, Gamma(x|\theta_k),$$

with $r \stackrel{not.}{=} (r_1, \ldots, r_K)$, $\beta \stackrel{not.}{=} (\beta_1, \ldots, \beta_K)$, $\theta = (r, \beta)$, **and** $\theta_k = (r_k, \beta_k)$ **for** $k = 1, \ldots, K$, **and** $\pi_k > 0$ **for** $k = 1, \ldots, K$ **and** $\sum_{k=1}^{K} \pi_k = 1$.

Consider $x_1, \ldots, x_n \in \mathbb{R}^+$, **each** x_i **being generated by one component of the above mixture, denoted by** $z_i \in \{1, \ldots, K\}$.

Find the maximum likelihood estimation of the parameters θ **by using the EM algorithm.**

Solution

- **The verosimility of the "complete" data** (x, z), **with** $x \stackrel{not.}{=} (x_1, \ldots, x_n)$, **and** $z \stackrel{not.}{=} (z_1, \ldots, z_n)$, **can be written as follows:**

$$p(x, z|\theta) \stackrel{i.i.d.}{=} \prod_{i=1}^{n} p(x_i|z_i, \theta) = \prod_{i=1}^{n} \prod_{k=1}^{K} [p(x_i|z_i, \theta) \cdot \underbrace{p(z_i|\theta)}_{\pi_k}]^{1_{\{z_i=k\}}} = \prod_{i=1}^{n} \prod_{k=1}^{K} [\underbrace{p(x_i|z_i, \theta)}_{Gamma(x_i|\theta_k)} \cdot \pi_k]^{1_{\{z_i=k\}}}$$

where $1_{\{z_i=k\}}$ **is the indicator function.**

- **The log-verosimility function:**

$$\ell(\theta) \stackrel{def.}{=} \ln p(x,z|\theta) = \sum_{i=1}^{n}\sum_{k=1}^{K} 1_{\{z_i=k\}} \left[\ln \pi_k + \ln Gamma(x_i|\theta_k)\right]$$

$$\stackrel{\theta_k=(r_k,\beta_k)}{=} \sum_{i=1}^{n}\sum_{k=1}^{K} 1_{\{z_i=k\}} \left[\ln \pi_k - r_k \ln \beta_k - \ln \Gamma(r_k) + (r_k-1)\ln x_i - \frac{x_i}{\beta_k}\right]$$

- **The auxiliary function:**

$$Q(\theta|\theta^{(t)}) \stackrel{\theta^{(t)}=(r^{(t)},\beta^{(t)})}{=} E_{P(Z|X,\theta^{(t)})}\left[\sum_{i=1}^{n}\sum_{k=1}^{K} 1_{\{z_i=k\}} \left[\ln \pi_k - r_k \ln \beta_k - \ln \Gamma(r_k) + (r_k-1)\ln x_i - \frac{x_i}{\beta_k}\right]\right]$$

$$\stackrel{lin.\ of\ exp.}{=} \sum_{i=1}^{n}\sum_{k=1}^{K} \underbrace{P(z_i=k|x_i,\theta^{(t)})}_{not.:\ \gamma_{ik}} \left[\ln \pi_k - r_k \ln \beta_k - \ln \Gamma(r_k) + (r_k-1)\ln x_i - \frac{x_i}{\beta_k}\right]$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_{ik} \left[\ln \pi_k - r_k \ln \beta_k - \ln \Gamma(r_k) + (r_k-1)\ln x_i - \frac{x_i}{\beta_k}\right]$$

E step

$$\gamma_{ik}^{(t+1)} \stackrel{not.}{=} P(z_i = k | x_i, \theta^{(t)}) \stackrel{Bayes\ F.}{=} \frac{Gamma(x_i | z_i = k, \theta) \cdot \overbrace{P(z_i = k|\theta)}^{\pi_k}}{\sum_{k'=1}^{K} Gamma(x_i | z_i = k', \theta) \cdot \underbrace{P(z_i = k'|\theta)}_{\pi_{k'}}}$$

$$= \frac{Gamma(x_i | \theta_k) \cdot \pi_k}{\sum_{k'=1}^{K} Gamma(x_i | \theta_{k'}) \cdot \pi_{k'}}$$

Note that $\gamma_{ik}^{(t+1)} > 0$ **for** $k = 1, \ldots, K$ **because** $x_i > 0$ **for** $i = 1, \ldots, n$.

M step: π_k

Because $\sum_{k=1}^{K} \pi_k = 1$, we will use a Lagrange multiplier λ, and solve as usually:

$$\frac{\partial}{\partial \pi_k}\left[Q(\theta|\theta^{(t)}) + \lambda\left(1 - \sum_{k=1}^{K} \pi_k\right)\right] = 0 \Leftrightarrow \sum_{i=1}^{n} \gamma_{ik}\frac{1}{\pi_k} - \lambda = 0 \Leftrightarrow \frac{1}{\pi_k}\sum_{i=1}^{n} \gamma_{ik} = \lambda \Leftrightarrow$$

$$\pi_k^{(t+1)} = \frac{1}{\lambda}\sum_{i=1}^{n} \gamma_{ik}$$

By substituting this expression into the constraint $\sum_{k=1}^{K} \pi_k = 1$, we will get:

$$\frac{1}{\lambda}\sum_{k=1}^{K}\sum_{i=1}^{n} \gamma_{ik} = 1 \Leftrightarrow \frac{1}{\lambda}\sum_{i=1}^{n}\underbrace{\sum_{k=1}^{K} \gamma_{ik}}_{1} = 1 \Leftrightarrow \frac{1}{\lambda} n = 1 \Leftrightarrow \lambda = n$$

Therefore,

$$\pi_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \gamma_{ik} k^{(t+1)}$$

Note that $\pi_k^{(t+1)} > 0$ for $k = 1, \ldots, K$.

M step: β_k

$$\frac{\partial}{\partial \beta_k} Q(\theta|\theta^{(t)}) = 0 \Leftrightarrow \sum_{i=1}^{n} \gamma_{ik} \left(-\frac{r_k}{\beta_k} + \frac{x_i}{\beta_k^2} \right) = 0 \Leftrightarrow \frac{1}{\beta_k^2} \left[\sum_{i=1}^{n} \gamma_{ik} (x_i - r_k \beta_k) \right] = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} \gamma_{ik} (x_i - r_k \beta_k) = 0 \Leftrightarrow \sum_{i=1}^{n} \gamma_{ik} r_k \beta_k = \sum_{i=1}^{n} \gamma_{ik} x_i \Leftrightarrow r_k \beta_k \sum_{i=1}^{n} \gamma_{ik} = \sum_{i=1}^{n} \gamma_{ik} x_i$$

Therefore,

$$\beta_k^{(t+1)} = \frac{\sum_{i=1}^{n} \gamma_{ik}^{(t+1)} x_i}{r_k \sum_{i=1}^{n} \gamma_{ik}}^{(t+1)} \quad \textbf{for } k = 1, \ldots, K.$$

It can be easily shown that $\beta_k^{(t+1)}$ is always positive, and it maximizes $Q(\theta|\theta^{(t)})$ w.r.t. β_k.

M step: r_k

By substituting [the expression of] $\beta_k^{(t+1)}$ into $Q(\theta|\theta^{(t)})$, we will get:

$$Q(r, \beta^{(t+1)}|\theta^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik} \left[\ln \pi_k - r_k \ln \beta^{(t+1)} - \ln \Gamma(r_k) + (r_k - 1) \ln x_i - \frac{x_i}{\beta_k^{(t+1)}} \right]$$

Then, solving for the partial derivative of $Q(r, \beta^{(t+1)}|\theta^{(t)})$ w.r.t. r_k, amounts to:

$$\frac{\partial}{\partial r_k} Q(r, \beta^{(t+1)}|\theta^{(t)}) = 0 \Leftrightarrow$$

$$\frac{\partial}{\partial r_k} \sum_{i=1}^{n} \gamma_{ik}^{(t+1)} \left[\ln \pi_k - r_k \ln \beta_k^{(t+1)} - \ln \Gamma(r_k) + (r_k - 1) \ln x_i - \frac{x_i}{\beta_k^{(t+1)}} \right] = 0 \Leftrightarrow$$

$$\sum_{i=1}^{n} \gamma_{ik}^{(t+1)} \left[- \ln \beta_k^{(t+1)} - r_k \frac{\partial}{\partial r_k} \ln \beta_k^{(t+1)} - \frac{\Gamma'(r_k)}{\Gamma(r_k)} + \ln x_i - x_i \frac{\partial}{\partial r_k} \frac{1}{\beta_k^{(t+1)}} \right] = 0$$

Since

$$\frac{\partial}{\partial r_k} \ln \beta_k^{(t+1)} = \frac{\partial}{\partial r_k} \ln \frac{\sum_{i=1}^n \gamma_{ik}^{(t+1)} x_i}{r_k \sum_{i=1}^n \gamma_{ik}^{(t+1)}} = \frac{\partial}{\partial r_k} \left(- \ln r_k + \ln \frac{\sum_{i=1}^n \gamma_{ik}^{(t+1)} x_i}{\sum_{i=1}^n \gamma_{ik}^{(t+1)}} \right) = - \frac{1}{r_k}$$

and

$$\frac{\partial}{\partial r_k} \frac{1}{\beta_k^{(t+1)}} = \frac{\partial}{\partial r_k} \frac{r_k \sum_{i=1}^n \gamma_{ik}^{(t+1)}}{\sum_{i=1}^n \gamma_{ik}^{(t+1)} x_i} = \frac{\sum_{i=1}^n \gamma_{ik}^{(t+1)}}{\sum_{i=1}^n \gamma_{ik}^{(t+1)} x_i},$$

it follows that

$$\frac{\partial}{\partial r_k} Q(r, \beta^{(t+1)} | \theta^{(t)}) = 0 \Leftrightarrow$$

$$\sum_{i=1}^n \gamma_{ik}^{(t+1)} \left[\ln r_k - \ln \frac{\sum_{i'=1}^n \gamma_{i'k}^{(t+1)} x_{i'}}{\sum_{i'=1}^n \gamma_{i'k}^{(t+1)}} + 1 - \frac{\Gamma'(r_k)}{\Gamma(r_k)} + \ln x_i - x_i \frac{\sum_{i'=1}^n \gamma_{i'k}^{(t+1)}}{\sum_{i'=1}^n \gamma_{i'k}^{(t+1)} x_{i'}} \right] = 0$$

Therefore,

$$\left(\ln r_k - \frac{\Gamma'(r_k)}{\Gamma(r_k)}\right) \sum_{i=1}^n \gamma_{ik}^{(t+1)} =$$

$$\sum_{i=1}^n \gamma_{ik}^{(t+1)} \ln \frac{\sum_{i'=1}^n \gamma_{i'k}^{(t+1)} x_{i'}}{\sum_{i'=1}^n \gamma_{i'k}^{(t+1)}} - \sum_{i=1}^n \gamma_{ik}^{(t+1)} - \sum_{i=1}^n \gamma_{ik}^{(t+1)} \ln x_i + \sum_{i=1}^n \gamma_{ik}^{(t+1)} x_i \cdot \frac{\sum_{i'=1}^n \gamma_{i'k}^{(t+1)}}{\sum_{i'=1}^n \gamma_{i'k}^{(t+1)} x_{i'}} \Leftrightarrow$$

$$\left(\ln r_k - \frac{\Gamma'(r_k)}{\Gamma(r_k)}\right) \sum_{i=1}^n \gamma_{ik}^{(t+1)} = \sum_{i=1}^n \gamma_{ik}^{(t+1)} \ln \frac{\sum_{i'=1}^n \gamma_{i'k}^{(t+1)} x_{i'}}{\sum_{i'=1}^n \gamma_{i'k}^{(t+1)}} - \sum_{i=1}^n \gamma_{ik}^{(t+1)} \ln x_i$$

So,

$$\ln r_{ik} - \frac{\Gamma'(r_k)}{\Gamma(r_k)} = \ln \frac{\sum_{i=1}^n \gamma_{ik}^{(t+1)} x_i}{\sum_{i=1}^n \gamma_{ik}^{(t+1)}} - \frac{\sum_{i=1}^n \gamma_{ik}^{(t+1)} \ln x_i}{\sum_{i=1}^n \gamma_{ik}^{(t+1)}}$$

Unfortunately, there is no nkown close form of r_k satisfying this equation. However, it can be shown that the function $\ln r_{ik} - \dfrac{\Gamma'(r_k)}{\Gamma(r_k)}$ is well-behaved, yielding a unique solution to the above equation, which will be denoted $r_{ik}^{(t+1)}$; it can be found through certain numerical methods.

***Note*: The function $\psi(r) \stackrel{not}{=} \ln r_{ik} - \dfrac{\Gamma'(r_k)}{\Gamma(r_k)}$ is the so-called di-gamma function.**

Note

It is interesting to see that
if in the *update equations* of this EM algorithm for solving the Gamma Mixture Model we substitute $r = r_k$ and $\gamma_k = \frac{1}{k}$ for $k = 1, \ldots, K$, we obtain exactly the Maximum Likeliood estimations for the parameters r and β of the Gamma distribution.

Using the EM algorithm for estimating the *selection probability* for a mixture of two (arbitrary) distributions

CMU, 2006 spring, Carlos Guestrin, final exam, pr. 8

CMU, 2004 fall, Carlos Guestrin, HW2, pr. 2.1

We want to derive an EM algorithm for estimating the mixing parameter for a mixture of arbitrary probability densities f_1 and f_2.

For *example*, $f_1(x)$ could be a standard normal distribution centered at 0, and $f_2(x)$ could be the uniform distribution between $[0, 1]$. You can think about such mixtures in the following way: First, you flip a coin. With probability λ (i.e., the coin comes up *heads*), you will sample x from density f_1, and with probability $(1 - \lambda)$ you sample from density f_2.

More formally, let $f_\lambda(x) = \lambda f_1(x) + (1 - \lambda) f_2(x)$, where f_1 and f_2 are arbitrary probability density functions on $\mathbb{R}$, and $\lambda \in [0, 1]$ is an unknown mixture parameter.

a.
Given a data point x, and a value for the mixture parameter λ, compute the probability that x was generated from density f_1.

b.
Now, suppose you are given a data set $\{x_1, \ldots, x_n\}$ drawn i.i.d. from the mixture density, and a set of coin flips $\{z_1, z_2, \ldots, z_n\}$, such that $z_i = 1$ means that x_i is a sample from f_1, and $z_i = 0$ means that x_i was generated from density f_2.
For a fixed parameter λ, compute the complete log-likelihood of the data, i.e., $\ln P(x_1, z_1, x_2, z_2, \ldots, x_n, z_n|\lambda)$.

c.
Now, suppose you are given only a sample $\{x_1, \ldots, x_n\}$ drawn i.i.d. from the mixture density, without the knowledge about which component the samples were drawn from (i.e., the z_i are unknown).
Using your derivations from part a and b, derive the E- and M-steps for an EM algorithm to compute the maximum likelihood estimate (MLE) of the mixture parameter λ.

Solution

a. $P(Z=1|X=x) = \dfrac{P(X=x|Z=1)\cdot P(Z=1)}{P(X=x)} = \dfrac{\lambda f_1(x)}{f_\lambda(x)}$.

b. $P(x_1, z_1, x_2, z_2, \ldots, x_n, z_n|\lambda) = \prod_{i=1}^n P(x_i, z_i|\lambda) = \prod_{i=1}^n (P(x_i|z_i,\lambda)\cdot P(z_i|\lambda))$.

$$\begin{aligned}
P(x_i|z_i,\lambda)\cdot P(z_i|\lambda) &= \begin{cases} \lambda f_1(x_i) & \textbf{if } z_i = 1 \\ (1-\lambda) f_2(x_i) & \textbf{if } z_i = 0 \end{cases} \\
&= \lambda^{z_i} f_1(x_i)^{z_i} (1-\lambda)^{1-z_i} f_2(x_i)^{1-z_i}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\ln P(x_1, z_1, x_2, z_2, \ldots, x_n, z_n|\lambda) &= \sum_{i=1}^n \ln P(x_i, z_i|\lambda) \\
&= \sum_{i=1}^n \ln\left(\lambda^{z_i} f_1(x_i)^{z_i} (1-\lambda)^{1-z_i} f_2(x_i)^{1-z_i}\right) \\
&= \sum_{i=1}^n z_i((\ln\lambda + \ln f_1(x_i)) + (1-z_i)(\ln(1-\lambda) + \ln f_2(x_i)))
\end{aligned}$$

c.

E-step: $q(z_i) \stackrel{not.}{=} P(z_i = 1|x_i, \lambda^{(t)}) \stackrel{B.Th.}{=} \dfrac{\lambda^{(t)} f_1(x_i)}{f_{\lambda^{(t)}}(x_i)}$

M-step:

$$
\begin{aligned}
\lambda^{(t+1)} &= \operatorname*{argmax}_{\lambda} E_q[\ln \prod_{i=1}^{n} P(x_1, z_1, x_2, z_2, \ldots, x_n, z_n|\lambda)] \\
&= \operatorname*{argmax}_{\lambda} (\ln \lambda \cdot \underbrace{\sum_{i=1}^{n} q(z_i)}_{c} + \ln(1-\lambda) \cdot \underbrace{\sum_{i=1}^{n} (1 - q(z_i))}_{n-c})
\end{aligned}
$$

$$
\begin{aligned}
& \frac{\partial}{\partial \lambda}(c \ln \lambda + (n-c)\ln(1-\lambda)) = 0 \Leftrightarrow \frac{c}{\lambda} = \frac{n-c}{1-\lambda} \\
\Leftrightarrow \quad & c(1-\lambda) = \lambda(n-c) \Leftrightarrow c = n\lambda \Leftrightarrow \lambda = \frac{c}{n} = \frac{\sum_{i=1}^{n} q(z_i)}{n} \\
\Rightarrow \quad & \lambda^{(t+1)} = \frac{\lambda^{(t)}}{n} \sum_{i=1}^{n} \frac{f_1(x_i)}{f_{\lambda^{(t)}}(x_i)}
\end{aligned}
$$

***Note*:** $\dfrac{\partial^2}{\partial \lambda^2}(c \ln \lambda + (n-c)\ln(1-\lambda)) = -\dfrac{c}{\lambda} - (n-c)\dfrac{1}{(1-\lambda)^2} \leq 0 \Rightarrow \lambda^{(t+1)}$ **corresponds to a maximum.**

Note

The EM algorithm can [be naturally extened so as to] handle mixtures of an arbitrary (although fixed) number of probabilistic distributions.

Exemplification [for a mixture of 3 densities]

Source: Brani Vidakovic, *EM Algorithm and Mixtures* [Handout 12]

A sample of size 150 is generated from the mixture

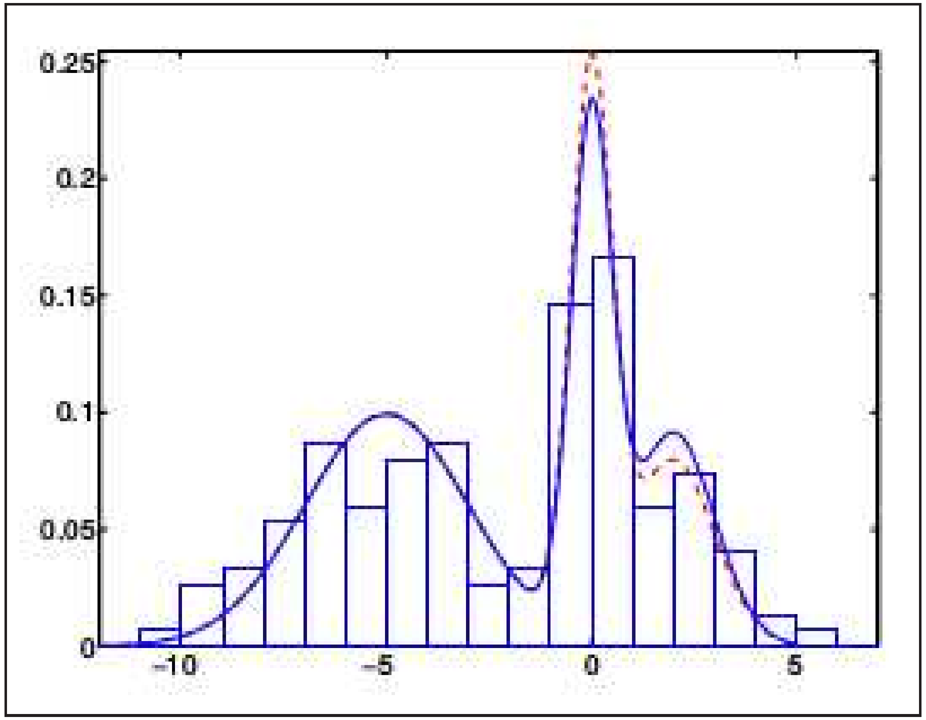
$$f(x) = 0.5\mathcal{N}(-5, 2^2) + 0.3\mathcal{N}(0, 0.5^2) + 0.2\mathcal{N}(2, 1),$$

where $\mathcal{N}$ denotes the normal distribution.

The mixing weights are estimated by the EM algorithm.

$M = 20$ **iterations of the EM algorithm yielded the weights** $(0.4977$, 0.2732 **and** 0.2290.

The nearby figure shows the histogram of data, the theoretical mixture (dotted line) and the EM estimate (solid line).



Deriving the EM algorithm for estimating the parameters of two random variables of exponential distribution, given a set of instances generated by their sum

CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW2, pr. 2.2

Suppose that $Z_1 \sim \exp(1/\lambda_1)$ and $Z_2 \sim \exp(1/\lambda 2)$, and $\lambda_1 \neq \lambda_2$. Z_1 and Z_2 are independent. Let $X = Z_1 + Z_2$ denote the sum of Z_1 and Z_2. Suppose that $\{x_1, x_2, \ldots, x_n\}$ are i.i.d. samples from the distribution of X.

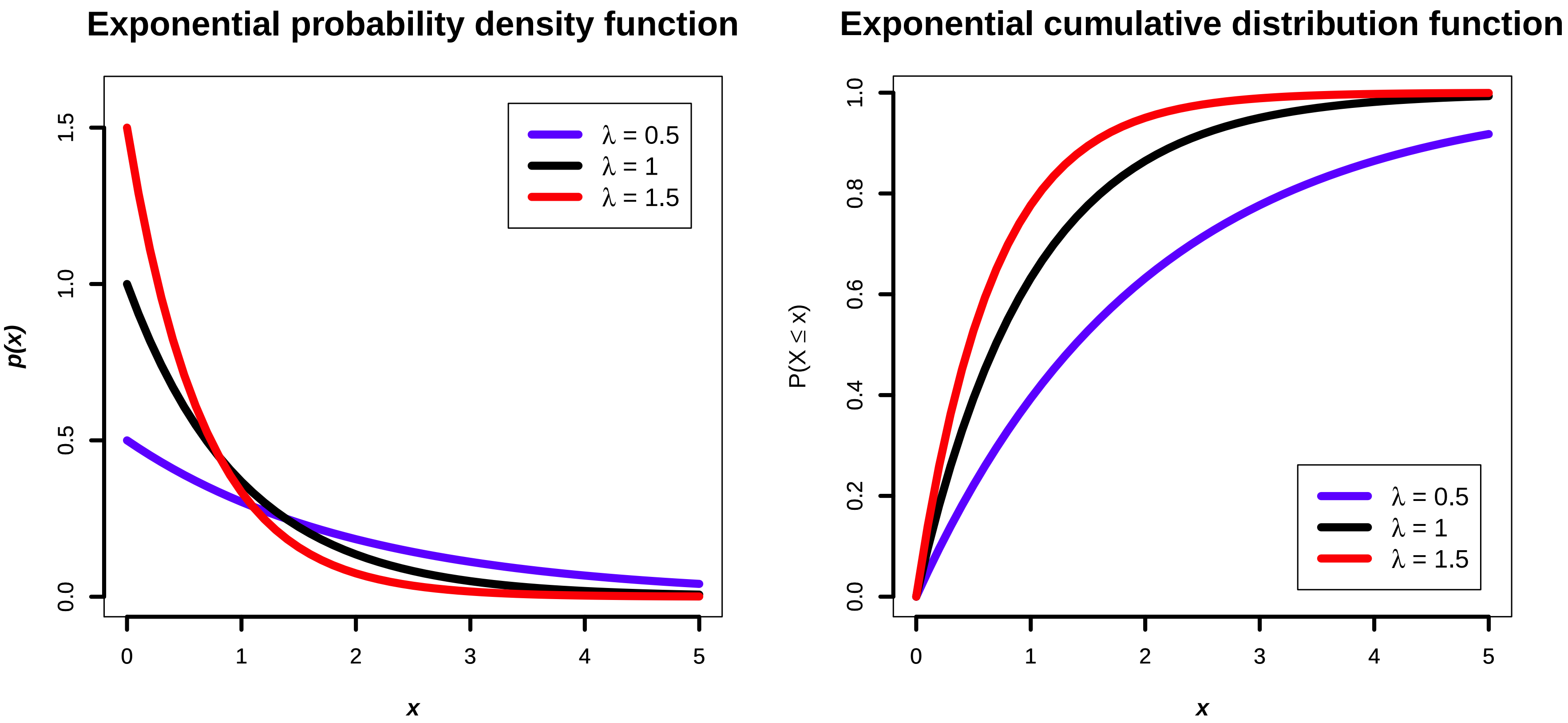
a. Derive an expression for the density of X in terms of λ_1 and λ_2.

***Hint* 1: The density of Z_1 is $f_{\lambda_1}(z) = \lambda_1 \cdot e^{-\lambda_1 z}$ for $z \geq 0$, and 0 otherwise. The density of Z_2 is defined in a similar way.**

***Hint* 2: You could first derive c.d.f. (cumulative distribution function) of X, which is defined as $F(x) = P(Z_1 + Z_2 < x) = \int_0^x \int_0^{x-z_1} f_{\lambda_1}(z_1) \cdot f_{\lambda_2}(z_2)\, dz_2\, dz_1$.**

***Explanation*:**

$$
\begin{aligned}
F(x) &\stackrel{def.}{=} P(Z_1 + Z_2 < x) \\
&= \int_0^x \left(\int_0^{x-z_2} f_{\lambda_1}(z_1)\, dz_1 \right) \cdot f_{\lambda_2}(z_2)\, dz_2 \\
&= \int_0^x \left(\int_0^{x-z_1} f_{\lambda_2}(z_2)\, dz_2 \right) \cdot f_{\lambda_1}(z_1)\, dz_1 = \int_0^x \int_0^{x-z_1} f_{\lambda_1}(z_1) \cdot f_{\lambda_2}(z_2)\, dz_2\, dz_1.
\end{aligned}
$$


Exponential probability density function
p(x)
0.0
0.5
1.0
1.5
0
1
2
3
4
5
x
λ = 0.5
λ = 1
λ = 1.5
Exponential cumulative distribution function
P(X ≤ x)
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
x
λ = 0.5
λ = 1
λ = 1.5

Answer

The c.d.f. of X:

$$F(x) \stackrel{def.}{=} P(Z_1 + Z_2 < x) = \int_0^x \int_0^{x-z_1} f_{\lambda_1}(z_1) \cdot f_{\lambda_2}(z_2)\, dz_2\, dz_1 = \int_0^x \int_0^{x-z_1} \lambda_1 e^{-\lambda_1 z_1} \cdot \lambda_2 e^{-\lambda_2 z_2}\, dz_2\, dz_1$$

$$= \int_0^x (-\lambda_1) e^{-\lambda_1 z_1} \left(\int_0^{x-z_1} (-\lambda_2) e^{-\lambda_2 z_2}\, dz_2 \right) dz_1 = \int_0^x (-\lambda_1) e^{-\lambda_1 z_1} \left(e^{-\lambda_2 z_2} \Big|_0^{x-z_1} \right) dz_1$$

$$= \int_0^x (-\lambda_1) e^{-\lambda_1 z_1} \left(e^{-\lambda_2 (x-z_1)} - \underbrace{e^{-\lambda_2 \cdot 0}}_{1} \right) dz_1$$

$$= \int_0^x (-\lambda_1) e^{-\lambda_1 z_1} e^{-\lambda_2 (x-z_1)} dz_1 - \int_0^x (-\lambda_1) e^{-\lambda_1 z_1} dz_1$$

$$= \frac{-\lambda_1}{\lambda_2 - \lambda_1} e^{-\lambda_2 x} \int_0^x (\lambda_2 - \lambda_1) e^{(\lambda_2 - \lambda_1) z_1} dz_1 - \int_0^x (-\lambda_1) e^{-\lambda_1 z_1} dz_1$$

$$= -\frac{\lambda_1}{\lambda_2 - \lambda_1} e^{-\lambda_2 x} \left(e^{(\lambda_2 - \lambda_1) z_1} \Big|_0^x \right) - e^{-\lambda_1 z_1} \Big|_0^x$$

$$= -\frac{\lambda_1}{\lambda_2 - \lambda_1} e^{-\lambda_2 x} \left(e^{(\lambda_2 - \lambda_1) x} - 1 \right) - \left(e^{-\lambda_1 x} - 1 \right) = -\frac{\lambda_1}{\lambda_2 - \lambda_1} e^{-\lambda_1 x} + \frac{\lambda_1}{\lambda_2 - \lambda_1} e^{-\lambda_2 x} - e^{-\lambda_1 x} + 1$$

$$= 1 - \frac{1}{\lambda_2 - \lambda_1} \left(\lambda_1 e^{-\lambda_1 x} - \lambda_1 e^{-\lambda_2 x} + \lambda_2 e^{-\lambda_1 x} - \lambda_1 e^{-\lambda_1 x} \right) = 1 - \frac{\lambda_2 e^{-\lambda_1 x} - \lambda_1 e^{-\lambda_2 x}}{\lambda_2 - \lambda_1}$$

The p.d.f. of X:

$$p(x) = \frac{\partial F(x)}{\partial x} = -\frac{1}{\lambda_2 - \lambda_1}\left(-\lambda_1\lambda_2 e^{-\lambda_1 x} + \lambda_1\lambda_2 e^{-\lambda_2 x}\right) = \frac{\lambda_1\lambda_2}{\lambda_2 - \lambda_1}\left(e^{-\lambda_1 x} - e^{-\lambda_2 x}\right)$$

b. Derive the E-step and M-step of the EM algorithm, and give explicit expressions for the parameter updates in the EM process for computing the MLE of λ_1 and λ_2.

***Pasul E (Expectation)*: Se calculează funcţia ,,auxiliară“ (media log-verosimilităţii datelor complete) pentru iteraţia t:**

$$Q(\lambda \mid \lambda^{(t)}) = E_{P(Z|X,\lambda^{(t)})}[\log P(X, Z \mid \lambda)]$$

***Pasul M (Maximization)*: Se maximizează media log-verosimilităţii datelor complete, calculată la pasul E, în raport cu λ:**

$$\lambda^{(t+1)} = \operatorname*{argmax}_{\lambda} Q(\lambda \mid \lambda^{(t)})$$

***Notaţiile* de mai sus au următoarele semnificaţii:**

$$\begin{array}{rcl}
\lambda & = & (\lambda_1, \lambda_2) \\
\lambda^{(t)} & = & \textbf{valoarea parametrului } \lambda \textbf{ la iteraţia } t \\
X & = & \textbf{variabila observabilă, cu instanţele } x_1, x_2, \ldots, x_n \\
Z & = & (Z_1, Z_2) \textbf{ variabilele ascunse/neobservabile} \\
 & & \textbf{având valorile } z_{1j}, z_{2j}, \ldots, z_{nj} \textbf{ cu } j \in \{1, 2\}, \\
 & & \textbf{aşa încât } x_i = z_{i1} + z_{i2}.
\end{array}$$

Solution: E step; computation of the "auxiliary" function

$$
\begin{aligned}
Q(\lambda \mid \lambda^{(t)}) &= E_{P(Z|X,\lambda^{(t)})}\left[\log \prod_{i=1}^{n} p(x_i, z_{i1}, z_{i2} \mid \lambda)\right] = E_{P(Z|X,\lambda^{(t)})}\left[\log \prod_{i=1}^{n} f_{\lambda_1}(z_{i1}) \cdot f_{\lambda_2}(z_{i2})\right] \\
&= E_{P(Z|X,\lambda^{(t)})}\left[\sum_{i=1}^{n} \log\left(\lambda_1 e^{-\lambda_1 z_{i1}} \cdot \lambda_2 e^{-\lambda_2 z_{i2}}\right)\right] \\
&= E_{P(Z|X,\lambda^{(t)})}\left[\sum_{i=1}^{n} (\log \lambda_1 - \lambda_1 z_{i1} + \log \lambda_2 - \lambda_2 z_{i2})\right] \\
&= E_{P(Z|X,\lambda^{(t)})}\left[n \log \lambda_1 + n \log \lambda_2 - \sum_{i=1}^{n} (\lambda_1 z_{i1} + \lambda_2 z_{i2})\right] \\
&= n \log \lambda_1 + n \log \lambda_2 - \lambda_1 \sum_{i=1}^{n} E_{p(z_{i1}|x_i,\lambda^{(t)})}[z_{i1}] - \lambda_2 \sum_{i=1}^{n} E_{p(z_{i2}|x_i,\lambda^{(t)})}[z_{i2}]
\end{aligned}
$$

Solution: E step; computation of expectations

$$
\begin{aligned}
E_{p(z_{i1}|x_i,\lambda^{(t)})}[z_{i1}] &\stackrel{def.}{=} \int_0^{x_i} z_{i1}\cdot p(z_{i1}\mid x_i,\lambda^{(t)})\,dz_{i1} \\
p(z_{i1}\mid x_i,\lambda^{(t)}) &\stackrel{def.}{=} \frac{p(z_{i1},x_i\mid\lambda^{(t)})}{p(x_i\mid\lambda^{(t)})} \\
&= \frac{p(z_{i1},z_{i2}|\lambda^{(t)})}{p(x_i\mid\lambda^{(t)})} = \frac{p(z_{i1}|\lambda_1^{(t)})\cdot p(z_{i2}|\lambda_2^{(t)})}{p(x_i\mid\lambda^{(t)})} \stackrel{a.}{=} \frac{f_{\lambda_1^{(t)}}(z_{i1})\cdot f_{\lambda_2^{(t)}}(x_i-z_{i1})}{\dfrac{\lambda_1^{(t)}\lambda_2^{(t)}}{\lambda_2^{(t)}-\lambda_1^{(t)}}\left(e^{-\lambda_1^{(t)}x_i}-e^{-\lambda_2^{(t)}x_i}\right)} \\
&= (\lambda_2^{(t)}-\lambda_1^{(t)})\cdot\frac{\lambda_1^{(t)}e^{-\lambda_1^{(t)}z_{i1}}\cdot\lambda_2^{(t)}e^{-\lambda_2^{(t)}(x_i-z_{i1})}}{\lambda_1^{(t)}\lambda_2^{(t)}\left(e^{-\lambda_1^{(t)}x_i}-e^{-\lambda_2^{(t)}x_i}\right)} \\
&= (\lambda_2^{(t)}-\lambda_1^{(t)})\cdot\frac{e^{(\lambda_2^{(t)}-\lambda_1^{(t)})z_{i1}}}{\left(e^{-\lambda_1^{(t)}x_i}-e^{-\lambda_2^{(t)}x_i}\right)\cdot e^{\lambda_2^{(t)}x_i}} \\
&= (\lambda_2^{(t)}-\lambda_1^{(t)})\cdot\frac{e^{(\lambda_2^{(t)}-\lambda_1^{(t)})z_{i1}}}{e^{(\lambda_2^{(t)}-\lambda_1^{(t)})x_i}-1}
\end{aligned}
$$

$$\Rightarrow E_{p(z_{i1}|x_i,\lambda^{(t)})}[z_{i1}] \stackrel{def.}{=} \int_0^{x_i} z_{i1} \cdot p(z_{i1} \mid x_i, \lambda^{(t)})\, dz_{i1} = \int_0^{x_i} z_{i1} \cdot (\lambda_2^{(t)} - \lambda_1^{(t)}) \cdot \frac{e^{(\lambda_2^{(t)}-\lambda_1^{(t)})z_{i1}}}{e^{(\lambda_2^{(t)}-\lambda_1^{(t)})x_i} - 1}\, dz_{i1}$$

$$= \frac{1}{e^{(\lambda_2^{(t)}-\lambda_1^{(t)})x_i} - 1} \int_0^{x_i} z_{i1} \cdot (\lambda_2^{(t)} - \lambda_1^{(t)}) \cdot e^{(\lambda_2^{(t)}-\lambda_1^{(t)})z_{i1}}\, dz_{i1}$$

Vom rezolva ultima integrală de mai sus utilizând formula de integrare prin părţi.

$$\int_0^{x_i} z_{i1} \cdot (\lambda_2^{(t)} - \lambda_1^{(t)}) \cdot e^{(\lambda_2^{(t)}-\lambda_1^{(t)})z_{i1}}\, dz_{i1} = \int_0^{x_i} z_{i1} \cdot \frac{\partial}{\partial z_{i1}} \left(e^{(\lambda_2^{(t)}-\lambda_1^{(t)})z_{i1}} \right) dz_{i1}$$

$$= \left(z_{i1} \cdot e^{(\lambda_2^{(t)}-\lambda_1^{(t)})z_{i1}} \right)\Big|_0^{x_i} - \int_0^{x_i} e^{(\lambda_2^{(t)}-\lambda_1^{(t)})z_{i1}}\, dz_{i1}$$

$$= \left(x_i \cdot e^{(\lambda_2^{(t)}-\lambda_1^{(t)})x_i} - 0 \right) - \frac{1}{\lambda_2^{(t)} - \lambda_1^{(t)}} \cdot e^{(\lambda_2^{(t)}-\lambda_1^{(t)})z_{i1}}\Big|_0^{x_i} = x_i \cdot e^{(\lambda_2^{(t)}-\lambda_1^{(t)})x_i} - \frac{e^{(\lambda_2^{(t)}-\lambda_1^{(t)})x_i} - 1}{\lambda_2^{(t)} - \lambda_1^{(t)}}$$

Prin urmare,

$$E_{p(z_{i1}|x_i,\lambda^{(t)})}[z_{i1}] = \frac{1}{e^{(\lambda_2^{(t)}-\lambda_1^{(t)})x_i} - 1} \cdot \left(x_i \cdot e^{(\lambda_2^{(t)}-\lambda_1^{(t)})x_i} - \frac{e^{(\lambda_2^{(t)}-\lambda_1^{(t)})x_i} - 1}{\lambda_2^{(t)} - \lambda_1^{(t)}} \right)$$

$$= \frac{x_i \cdot e^{(\lambda_2^{(t)}-\lambda_1^{(t)})x_i}}{e^{(\lambda_2^{(t)}-\lambda_1^{(t)})x_i} - 1} - \frac{1}{\lambda_2^{(t)} - \lambda_1^{(t)}}$$

$$
\begin{aligned}
x_i &= z_{i1} + z_{i2} \\
\Rightarrow E_{p(z_{i2}|x_i,\lambda^{(t)})}[z_{i2}] &= E_{p(z_{i2}|x_i,\lambda^{(t)})}[x_i - z_{i1}] = x_i - E_{p(z_{i2}|x_i,\lambda^{(t)})}[z_{i1}] \\
&= x_i - E_{p(z|x_i,\lambda^{(t)})}[z_{i1}] = x_i - E_{p(z_{i1}|x_i,\lambda^{(t)})}[z_{i1}] \\
&= x_i - \frac{x_i \cdot e^{(\lambda_2^{(t)} - \lambda_1^{(t)})x_i}}{e^{(\lambda_2^{(t)} - \lambda_1^{(t)})x_i} - 1} + \frac{1}{\lambda_2^{(t)} - \lambda_1^{(t)}} \\
&= \frac{1}{\lambda_2^{(t)} - \lambda_1^{(t)}} - \frac{x_i}{e^{(\lambda_2^{(t)} - \lambda_1^{(t)})x_i} - 1}
\end{aligned}
$$

Solution: M step

$$
\begin{array}{rcl}
\lambda^{(t+1)} & = & \underset{\lambda}{\operatorname{argmax}}\, Q(\lambda \mid \lambda^{(t)}) \\
 & = & \underset{\lambda_1>0,\lambda_2>0}{\operatorname{argmax}} \left(n\log\lambda_1 + n\log\lambda_2 - \lambda_1 \sum_{i=1}^{n} E_{p(z_{i1}|x_i,\lambda^{(t)})}[z_{i1}] - \; -\lambda_2 \sum_{i=1}^{n} E_{p(z_{i2}|x_i,\lambda^{(t)})}[z_{i2}] \right)
\end{array}
$$

$$
\begin{array}{rcl}
\dfrac{\partial}{\partial\lambda_1} Q(\lambda \mid \lambda^{(t)}) = 0 & \Leftrightarrow & \dfrac{n}{\lambda_1} = \displaystyle\sum_{i=1}^{n} E_{p(z_{i1}|x_i,\lambda^{(t)})}[z_{i1}] \\
 & \Rightarrow & \lambda_1^{(t+1)} = \dfrac{n}{\sum_{i=1}^{n} E_{p(z_{i1}|x_i,\lambda^{(t)})}[z_{i1}]} > 0 \\
\dfrac{\partial}{\partial\lambda_2} Q(\lambda \mid \lambda^{(t)}) = 0 & \Leftrightarrow & \dfrac{n}{\lambda_2} = \displaystyle\sum_{i=1}^{n} E_{p(z_{i2}|x_i,\lambda^{(t)})}[z_{i2}] \\
 & \Rightarrow & \lambda_2^{(t+1)} = \dfrac{n}{\sum_{i=1}^{n} E_{p(z_{i2}|x_i,\lambda^{(t)})}[z_{i2}]} > 0
\end{array}
$$

Observaţie:
Se verifică imediat că aceste soluţii într-adevăr maximizează $Q(\lambda \mid \lambda^{(t)})$.

Using the EM algorithm for learning a Poisson distribution *when some data are missing*

Liviu Ciortuz, following

Anirban DasGupta, *Probability for Statistics and Machine Learning*, Springer, 2011, Ex. 20.8

Presupunem că vrem să modelăm statistic numărul de accidente uşoare care s-au produs în n locaţii într-un anumit interval de timp, să zicem o săptămână. În acest scop, vom folosi o distribuţie Poisson de parametru λ. La sârşitul perioadei de timp respective, ni se transmite de la m din cele n locaţii câte o ,,înregistrare" (notată cu x_i), reprezentând numărul de accidente uşoare produse în locaţia i.

În mod implicit, ar trebui să considerăm că în cele $n-m$ locaţii de la care n-am primit înregistrări nu s-a produs niciun accident. Însă, în urma ,,inspectării" datelor suntem determinaţi să luăm în considerare *presupunerea* că în unele din aceste k locaţii s-a produs câte un [singur] accident uşor, care a fost ,,trecut sub tăcere" la raportare.

Aşadar, formalizând, vom considera datele ,,neobservabile" $z_1 = 0, \ldots, z_{n_0} = 0, z_{n_0+1} = 1, \ldots, z_{n_0+n_1} = 1$, cu $n_0 + n_1 = n - m$, alături de datele observabile $x_1, \ldots, x_m \in \mathbb{N}$. Toate aceste date sunt produse de variabile aleatoare urmând distribuţia $Poisson$ de [acelaşi] parametru λ.

Elaboraţi algoritmul EM (în speţă pasul E şi pasul M) pentru estimarea parametrului λ.

***Sugestie*: În loc să se lucreze cu z_j, cu $j = 1, \ldots, m$, va fi suficient să consideraţi ca date neobservabile n_0 şi n_1. Mai mult, considerând numerele n şi m cunoscute, va fi suficient să lucraţi doar cu n_1 ca dată neobservabilă (bineînţeles, pe lângă datele observabile x_i).**

Answer

$z_1, \ldots, z_{n+0}, z_{n_0+1}, \ldots, z_{n_0+n_1}, x_1, \ldots, x_m \sim Poisson(\lambda)$

Poisson p.m.f.: $P(x|\lambda) = \frac{1}{e^\lambda} \cdot \frac{\lambda^x}{x!}$

The verosimility of complete data:

$$\begin{aligned}
L(\lambda) &\stackrel{def.}{=} P(z_1, \ldots, z_{n+0}, z_{n_0+1}, \ldots, z_{n_0+n_1}, x_1, \ldots, x_m|\lambda) \\
&\stackrel{i.i.d.}{=} \prod_{j=1}^{n_0+n_1} P(z_i|\lambda) \cdot \prod_{i=1}^{m} P(x_i|\lambda) = \prod_{j=1}^{n_0} \frac{1}{e^\lambda}\frac{\lambda^0}{0!} \cdot \prod_{j=n_0+1}^{n_1} \frac{1}{e^\lambda}\frac{\lambda}{1!} \cdot \prod_{i=1}^{m} \frac{1}{e^\lambda}\frac{\lambda^{x_i}}{x_i!} \\
&= \frac{1}{(e^\lambda)^{n_0+n_1+m}} \cdot \lambda^{n_1} \cdot \prod_{i=1}^{m} \frac{\lambda^{x_i}}{x_i!} = \frac{1}{(e^\lambda)^n} \cdot \lambda^{n_1} \cdot \prod_{i=1}^{m} \frac{\lambda^{x_i}}{x_i!} \\
&= e^{-n\lambda} \cdot \lambda^{n_1} \cdot \lambda^{\sum_{i=1}^m x_i} \cdot \frac{1}{\prod_{i=1}^m x_i!} = e^{-n\lambda} \cdot \lambda^{n_1+\sum_{i=1}^m x_i} \cdot \frac{1}{\prod_{i=1}^m x_i!}
\end{aligned}$$

The log-verosimility of complete data:

$$\ell(\lambda) \stackrel{def.}{=} \ln L(\lambda) = -n\lambda + \left(n_1 + \sum_{i=1}^m x_i\right) \ln\lambda - \sum_{i=1}^m \ln x_i!$$

The auxiliary function:

$$Q(\lambda|\lambda^{(t)}) \stackrel{def.}{=} E_{n_1|x_i,\lambda^{(t)}}[\ell(\lambda)] = -n\lambda + \left(E[n_1|x_i,\lambda^{(t)}] + \sum_{i=1}^{m} x_i\right)\ln\lambda - \sum_{i=1}^{m}\ln x_i!.$$

E-step:

$E_{n_1|x_i,\lambda^{(t)}}[\ell(\lambda)]$ **is the expected number of unobservable instances** z_j **having value** 1**, out of the total of** $n-m$ **unobservable instances.**

The definition of the density function for the Poisson distribution implies

$P(x=1|\lambda) = \frac{1}{e^\lambda}\cdot\frac{\lambda^1}{1!} = \frac{1}{e^\lambda}\cdot\lambda$ **and** $P(x=0|\lambda) = \frac{1}{e^\lambda}\cdot\frac{\lambda^0}{0!} = \frac{1}{e^\lambda}$**.**

Therefore $E[n_1|x_i,\lambda^{(t)}]$ **is exactly the expectation of the binomial distribution** $Bin\left(n-m; \frac{\lambda^{(t)}}{1+\lambda^{(t)}}\right)$**. So,**

$$E[n_1|x_i,\lambda^{(t)}] = (n-m)\frac{\lambda^{(t)}}{1+\lambda^{(t)}}.$$

M-step:

Now we can write

$$Q(\lambda|\lambda^{(t)}) = -n\lambda + \left[(n-m)\frac{\lambda^{(t)}}{1+\lambda^{(t)}} + \sum_{i=1}^{m} x_i\right]\ln\lambda - \sum_{i=1}^{m}\ln x_i!.$$

The first two derivatives of this function w.r.t. λ are:

$$\begin{aligned}
\frac{\partial}{\partial\lambda}Q(\lambda|\lambda^{(t)}) &= -n + \left[(n-m)\frac{\lambda^{(t)}}{1+\lambda^{(t)}} + \sum_{i=1}^{m} x_i\right]\frac{1}{\lambda} \\
\frac{\partial^2}{\partial\lambda^2}Q(\lambda|\lambda^{(t)}) &= -\left[(n-m)\frac{\lambda^{(t)}}{1+\lambda^{(t)}} + \sum_{i=1}^{m} x_i\right]\frac{1}{\lambda^2}
\end{aligned}$$

From the problem's statement we know that $n > m$ and $\sum_{i=1}^{m} x_i \geq 0$.
At the initialization step of the EM algorithm the parameter λ is asigned a positive value ($\lambda^{(0)}$), and we will soon see that $\lambda^{(t)} > 0$ at each iteration $t > 0$.

Therefore, the second derivative of Q is always negative, meaning that Q is concave and therefore it has a maximum.

In order to find this maximum we equate the first derivative to 0, which gives us the *update equation*:

$$\lambda^{(t+1)} = \frac{1}{n} \left[(n-m) \frac{\lambda^{(t)}}{1+\lambda^{(t)}} + \sum_{i=1}^{m} x_i \right]$$

One can easily see now (ny induction) that $\lambda^{(t+1)} > 0$.