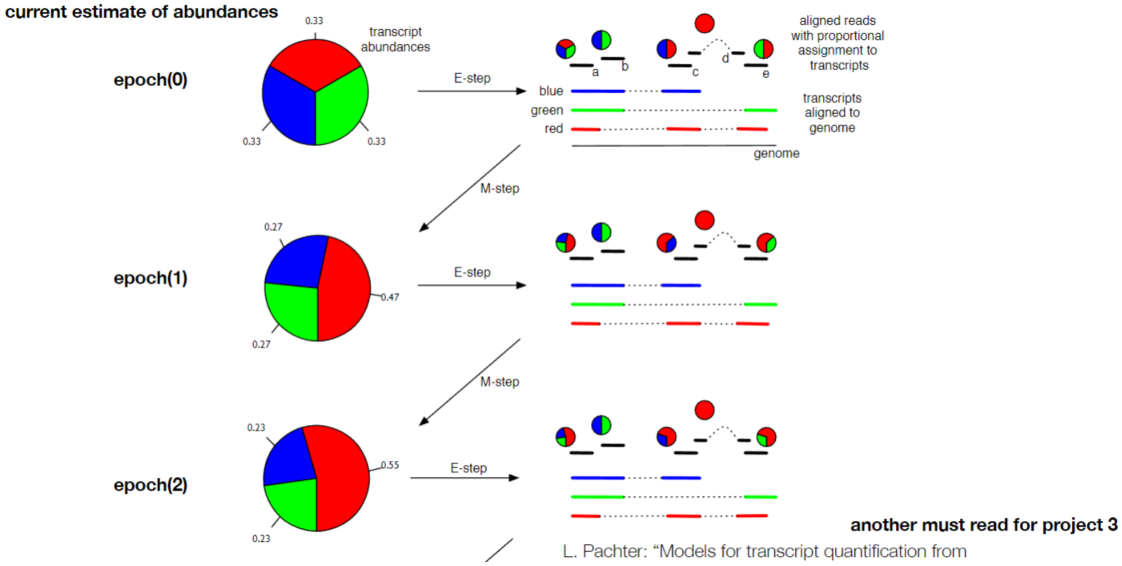
COM SCI C121 Week 6

Aidan Jan

May 7, 2024

EM Algorithm

As review:



another must read for project 3

L. Pachter: "Models for transcript quantification from

Essentially, we use the estimated transcript abundances to calculate the probability of the reads, then use the probabilities of the reads we attempt to maximize the probabilities by modifying the transcript abundances.

- The sections of the pie chart are denoted by θ_i, where i is a number or some sort of identifier for the section. In this case, we have three colors, so let θ_r represent the red abundance (beginning with 0.33), θ_g represent the green abundance, and θ_b represent the blue abundance.
- In the E-step, we take the ratios in the pie chart to give a probability to each of the reads.
- In the M-step, we take the ratios to what we actually see on the reads to recalculate the abundances. In the first step,
  - Red sections take up $2\frac{1}{3}$ of the 5 pie charts total.
  - Green sections take up $1\frac{1}{3}$ of the 5 pie charts.
  - Blue sections take up $1\frac{1}{3}$ of the 5 pie charts.
  - Then, we update the $\theta_r, \theta_g, \theta_b$ values.
  - $\theta_r = \frac{7}{3}/5 = \frac{7}{15} \approx 0.47$, $\theta_g = \frac{4}{3}/5 = \frac{4}{15} \approx 0.27$, $\theta_b = \frac{4}{3}/5 = \frac{4}{15} \approx 0.27$
- These steps are then repeated until the numbers stop changing.

In math terms:

$$L(\theta;x) = P(x|\theta) = \int P(x,z|\theta)\mathrm{d}z = \int P(x|\theta,z)P(z|\theta)\mathrm{d}z$$

$$\begin{aligned} \text{E-Step:} \quad & z|x,\theta^{(t)} = [\log(L(\theta;x,z))] \\ \text{M-Step:} \quad & \theta^{(t+1)} = \operatorname{argmax}_\theta \ Q(\theta|\theta^{(t)}) \end{aligned}$$

where

- z is the assignment (or true origin) of the sample.
- Q is the expectation of the log likelihood.

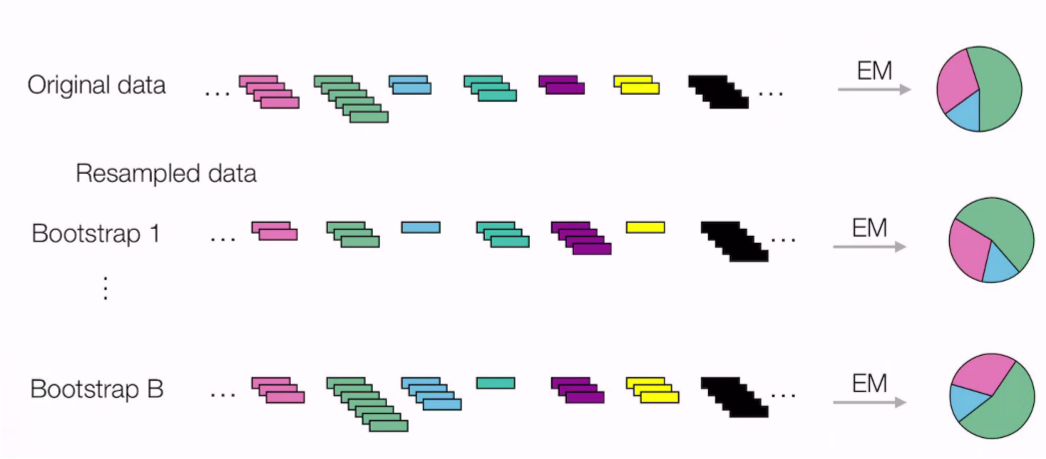
$$\begin{aligned} Q(\theta|\theta^{(t)}) &= E_{z|r,\theta^{(t)}}[\log P(r,z|\theta)] \\ &= \sum_{n,i,j} E_{z|r,\theta^{(t)}}[\mathbb{1}(z_{nij}=1)] \log\left(\frac{\theta_i}{l_i}P(r|z)\right) \end{aligned}$$

$$\begin{aligned} E_{z|r,\theta^{(t)}}[\mathbb{1}(z_{nij}=1)] &= P_{z|r,\theta^{(t)}}(z_{nij}=1|r,\theta^{(t)}) \\ &= \frac{P(r_n, z_{nij}|\theta^{(t)})}{P(r_n|\theta^{(t)})} \\ &= \frac{P(r_n|z_{nij}=1)P(O_n=i)P(S_n=j|O_n=i)}{\sum_k \sum_l P(r_n|z_{nkl}=1)P(O_n=k)P(S_n=l|O_n=k)} \\ &= \frac{c \cdot \theta_i^{(t)} \cdot \frac{1}{\tilde{l}_i}}{\sum_k \sum_l \frac{\theta_k^{(t)}}{\tilde{l}_k} \cdot 1} \end{aligned}$$

- c is a constant, where $c = P(r_n|z_{nij}=1)$. (in this example, we assume $c=1$.)
- For this point, assume $i = g$, referring to the colors of the pie chart (g = green)
  - The numerator of the last line of math represents the area of green on the chart, over the actual likelihood of green based on the probability. The denominator represents the sum of all slices on the pie chart (e.g., the number of pies)
- Note that this matches the intuition we had earlier about the M-step.

Where EM is Used

- Remember that when we know everything (like, the entire distribution), we can assume that $X_r \sim \text{Poisson}(\lambda)$
- However, we don't know everything. So we have to use EM to estimate the distribution.
- In most real life applications of sequencing, there are a lot of errors, a lot of isomers, and a lot of reads. EM can be used to determine the actual alignment more accurately than a naive bayes.
- EM is used in Kallisto:


Original data
...
...
EM
Resampled data
Bootstrap 1
...
...
EM
⋮
Bootstrap B
...
...
EM