Ha Vu
FEAST: fast expectation-maximization for microbial source tracking
August 14, 2023

Derivations of FEAST EM updates

1. What is FEAST

FEAST [1] is an EM-based algorithm for the problem of source-tracking in microbiome data. In particular, given the read counts of a sink (e.g. infant skin) $\mathbf{X}$ with dimension $N * 1$ where N is the number of taxa, we want to estimate the proportion of the observed sink microbiome profile is from different sources, e.g. adult skin, sheet, etc.. The source microbiome data is given also in units of read counts. We denote the source profile as $\mathbf{Y}$ of dimension $K * N$ where K is the number of reference sources, and N is number of taxa (the same taxa that we observed in the sink). We want to estimate a vector α of dimension $K+1$, with each entry shows the proportion of the sink microbiome is from each of the sources, and the $K+1$-th entry corresponds to the unknown source. The constraint for α is $\sum_{i=1}^{K+1} \alpha_i = 1$.

2. Why do we redo FEAST derivation?

I found the derivations in the paper a bit hand-wavy. I want to redo it for my own understanding.

3. Steps in EM

You can skip this section if it's too confusing, working with an example and going back to this step-by-step reference is best. I am aware this whole section may be not easy to understand for readers that are not myself. The EM algorithm composes of the following steps:

- Design the systems: observed variables $\mathbf{X}, \mathbf{Y}$, hidden variables $\mathbf{Z}$, and relationship between variables, and how those relationships are parameterized. The goal of EM is to learn the values of parameters. Note: In EM, we do not assume any prior distribution of parameters (frequentist approach). I found that EM can be applied to cases where the hidden variables is an indicator variable (e.g. indicating membership for each datapoint).
- Write the complete data likelihood and log-likelihood (complete means both observed and hidden variables, $P(\mathbf{X}, \mathbf{Y}, \mathbf{Z}|\mathbf{\Theta})$ where $\mathbf{\Theta}$ denotes all the model parameters).
- E-step: Derive the expectation of the hidden variables, given the observed variable values and the current model parameters $\mathbf{\Theta}^{(t)}$, i.e $E(\mathbf{Z}|\mathbf{X}, \mathbf{Y}, \mathbf{\Theta}^{(t)})$
- M-step: Plug the expected values of hidden variables ($\mathbf{\Theta}^{(t)}$, i.e $E(\mathbf{Z}|\mathbf{X}, \mathbf{Y}, \mathbf{\Theta}^{(t)})$) from previous step into the complete log-likelihood (or likelihood, whichever makes the derivation easier), denote the plugged-in log-likelihood as $Q(\mathbf{\Theta}|\mathbf{X}, \mathbf{Y})$. Then, derive the formulas for $\mathbf{\Theta}(t+1)$ that will maximize the "plugged-in" complete likelihood Q. Note: In cases where we impose containts on the parameters, we will optimize the loss function formulated with Langrange multiplier.

4. FEAST formulation

Below, I formulate the FEAST problem in a way that is both faithful to the paper's formulation and notations, but also include notions of hidden variables that will help us derive the EM steps most **rigorously**.

- C: total number of reads in the sink
- $\mathbf{X}$, size $C * N$, each binary entry indicates the taxa (out of N) of each read (out of C).
- $\mathbf{Z}$, size $C * (K+1)$, each binary entry indicates the source (out of $K+1$ sources) of each read (C reads).
- $\widetilde{\mathbf{X}}$, size N, each entry shows the read counts from each taxa (N). $\widetilde{\mathbf{X}}_j = \sum_{c=1}^{C} X_{cj}$.
- α, size $K+1$ with constraint $\sum_{i=1}^{K+1} \alpha_i = 1$, each entry shows the proportion of the sink that is from each source.
- $\mathbf{Y}$, size $K * N$, each entry shows the read counts for each taxa (out of N) in each reference source (K).
- C_i denotes the sum of read counts in source i, $C_i = \sum_{j=1}^{N} Y_{ij}$.

- γ_i, size N, denotes the relative abundance of each taxa (out of N in the source i
- $Y_i \sim Multinom(C_i, (\gamma_{i1}, ..., \gamma_{iN}))$ The read counts in each reference source (size N entries) is observed from a multinomial distribution.
- $Z_c \sim Multinom(1, (\alpha_1, \alpha_2, ..., \alpha_{K+1}))$. Here, Z_c is a binary vector of size $K+1$ with one entry of 1.
- $X_c|Z_c = i \sim Multinom(1, (\gamma_{i1}, ..., \gamma_{iN})$. Here Z_c notation is abused a little because Z_c is an integer variable indicating the source of read c, instead of being the one-hot encoding vector of length $K+1$ as used in previous point.
- Therefore, $P(X_c = j) = \sum_{i=1}^{K+1} P(Z_c = i)P(X_c = j|Z_c = i) = \sum_{i=1}^{K+1} \alpha_i \gamma_{ij} := \beta_j$. Note $\beta = (\beta_1, ..., \beta_N)$ denotes the marginal probabilities of taxa $1, .., N$ distribution in sink. Therefore, $X_c \sim Multinom(1, (\beta_1, ..., \beta_N))$. Or, $\widetilde{\mathbf{X}} \sim Multinom(C, (\beta_1, ..., \beta_N))$.
- Note that c indexes reads in the sink $(1, ..., C)$, i indexes sources $(1, ..., K+1)$, and j indexes taxa $(1, ..., N)$

5. EM derivation for FEAST

First, write complete likelihood and log-likelihood.

$$P(\mathbf{X}, \mathbf{Y}, \mathbf{Z} \mid \alpha, \gamma) = P(\mathbf{Y}|\gamma)P(\mathbf{Z}|\alpha)P(\mathbf{X}|\mathbf{Z}, \alpha, \gamma) \tag{1}$$

$$= \prod_{i=1}^{K} \binom{C_i}{y_{i1}, ..., y_{iN}} \prod_{j=1}^{N} \gamma_{ij}^{y_{ij}} \cdot \prod_{c=1}^{C} \prod_{i=1}^{K+1} \alpha_i^{z_{ci}} \cdot \prod_{c=1}^{C} \prod_{i=1}^{K+1} \prod_{j=1}^{N} \gamma_{ij}^{z_{ci} x_{cj}} \tag{2}$$

Therefore,

$$log P(\mathbf{X}, \mathbf{Y}, \mathbf{Z} \mid \alpha, \gamma) = constant + \sum_{i=1}^{K} \sum_{j=1}^{N} y_{ij} \cdot log(\gamma_{ij}) + \sum_{c=1}^{C} \sum_{i=1}^{K+1} z_{ci} \cdot log(\alpha_i) + \sum_{c=1}^{C} \sum_{i=1}^{K+1} \sum_{j=1}^{N} z_{ci} x_{cj} \cdot log(\gamma_{ij}) \tag{3}$$

E-step: calculate the expected values hidden variables. Here, the hidden variables are Z_{ci}.

$$E(z_{ci}|\mathbf{X}, \mathbf{Y}, \alpha^{(t)}, \gamma^{(t)}) = P(z_{ci} = 1|\mathbf{X}, \mathbf{Y}, \alpha^{(t)}, \gamma^{(t)}) \tag{4}$$

$$= \frac{P(z_c = i, \mathbf{X}, \mathbf{Y}, \alpha^{(t)}, \gamma^{(t)})}{\sum_{k=1}^{K+1} P(z_c = k, \mathbf{X}, \mathbf{Y}, \alpha^{(t)}, \gamma^{(t)})} \tag{5}$$

$$= \frac{P(z_c = i, X_c) P(\mathbf{X}_{\neq \mathbf{c}}, \mathbf{Y}, \alpha^{(t)}, \gamma^{(t)})}{\sum_{k=1}^{K+1} P(z_c = k, X_c) P(\mathbf{X}_{\neq \mathbf{c}}, \mathbf{Y}, \alpha^{(t)}, \gamma^{(t)})} \tag{6}$$

$$= \frac{\alpha_i^{(t)} \gamma_{i,x_c}^{(t)}}{\sum_{k=1}^{K+1} \alpha_k^{(t)} \gamma_{i,x_c}^{(t)}} \tag{7}$$

$$:= w_{ci}^{(t)} \tag{8}$$

M-step: calculate the updates on model parameters by maximizing the log-likelihood given constraints and with the plugged-in expected values of hidden variables.
First, plug in $w_{ci}^{(t)}$ from E-step into the log-likelihood

$$log P(\mathbf{X}, \mathbf{Y}, \mathbf{Z} \mid \alpha, \gamma) = constant + \sum_{i=1}^{K} \sum_{j=1}^{N} y_{ij} \cdot log(\gamma_{ij}) + \sum_{c=1}^{C} \sum_{i=1}^{K+1} w_{ci}^{(t)} \cdot log(\alpha_i) + \sum_{c=1}^{C} \sum_{i=1}^{K+1} \sum_{j=1}^{N} w_{ci}^{(t)} x_{cj} \cdot log(\gamma_{ij}) \tag{9}$$

Second, given the constraints on parameters, we define the loss function using Langrange multiplier

$$L(\alpha,\gamma) = logP(\mathbf{X},\mathbf{Y},\mathbf{Z}\mid\alpha,\gamma) - constraints \tag{10}$$

$$= logP(\mathbf{X},\mathbf{Y},\mathbf{Z}\mid\alpha,\gamma) - ((\sum_{i=1}^{K+1}\alpha_i)-1) - \sum_{i=1}^{K+1}((\sum_{j=1}^{N}\gamma_{ij})-1) \tag{11}$$

$$= \sum_{i=1}^{K}\sum_{j=1}^{N} y_{ij}\cdot log(\gamma_{ij}) + \sum_{c=1}^{C}\sum_{i=1}^{K+1} w_{ci}^{(t)}\cdot log(\alpha_i) + \sum_{c=1}^{C}\sum_{i=1}^{K+1}\sum_{j=1}^{N} w_{ci}^{(t)} x_{cj}\cdot log(\gamma_{ij})$$

$$- ((\sum_{i=1}^{K+1}\alpha_i)-1) - \sum_{i=1}^{K+1}((\sum_{j=1}^{N}\gamma_{ij})-1) \tag{12}$$

Now, we want to find the solution for α_i and γ_{ij} that will optimize the loss function. To do that, we need:

- Set $\frac{\partial L(\alpha,\gamma)}{\partial \alpha_i} = 0$ and solve for α_i
- Set $\frac{\partial L(\alpha,\gamma)}{\partial \gamma_{ij}} = 0$ and solve for γ_{ij}

Let's solve for α first

$$\frac{\partial L(\alpha,\gamma)}{\partial \alpha_i} = 0 \iff \frac{\sum_{c=1}^{C} w_{ci}^{(t)}}{\alpha_i} - 1 = 0 \tag{13}$$

$$\iff \alpha_i = \sum_{c=1}^{C} w_{ci}^{(t)} = \sum_{c=1}^{C}\frac{\alpha_i^{(t)}\gamma_{i,x_c}^{(t)}}{\sum_{k=1}^{K+1}\alpha_k^{(t)}\gamma_{i,x_c}^{(t)}} = \frac{\alpha_i^{(t)}\gamma_{i,x_1}^{(t)}}{\sum_{k=1}^{K+1}\alpha_k^{(t)}\gamma_{i,x_1}^{(t)}} + ... + \frac{\alpha_i^{(t)}\gamma_{i,x_C}^{(t)}}{\sum_{k=1}^{K+1}\alpha_k^{(t)}\gamma_{i,x_C}^{(t)}} \tag{14}$$

Remember that if we write $\widetilde{\mathbf{X}}_j = \sum_{c=1}^{C} X_{cj}$, then there are $\widetilde{X}_j$ terms of values $\frac{\alpha_i^{(t)}\gamma_{ij}^{(t)}}{\sum_{k=1}^{K+1}\alpha_k^{(t)}\gamma_{kj}^{(t)}}$ in the sum above, where $j \in \{1,...,N\}$. Hence:

$$\alpha_i = \sum_{j=1}^{N}\frac{\widetilde{X}_j\cdot\alpha_i^{(t)}\gamma_{ij}^{(t)}}{\sum_{k=1}^{K+1}\alpha_i^{(t)}\gamma_{kj}^{(t)}} \tag{15}$$

But, since we still have the constraint $\sum_{i=1}^{K+1}\alpha_i = 1$, therefore:

$$\alpha_i^{(t+1)} = \frac{\frac{\sum_{j=1}^{N}\widetilde{X}_j\cdot\alpha_i^{(t)}\gamma_{ij}^{(t)}}{\sum_{k=1}^{K+1}\alpha_k^{(t)}\gamma_{kj}^{(t)}}}{\sum_{h=1}^{K+1}\sum_{j=1}^{N}\frac{\widetilde{X}_j^{(t)}\alpha_h^{(t)}\gamma_{hj(t)}}{\sum_{k=1}^{K+1}\alpha_k^{(t)}\gamma_{kj}^{(t)}}} \tag{16}$$

$$= \frac{\sum_{j=1}^{N}\frac{\widetilde{X}_j\cdot\alpha_i^{(t)}\gamma_{ij}^{(t)}}{\sum_{k=1}^{K+1}\alpha_k^{(t)}\gamma_{kj}^{(t)}}}{\sum_{j=1}^{N}\sum_{h=1}^{K+1}\frac{\widetilde{X}_j^{(t)}\alpha_h^{(t)}\gamma_{hj}^{(t)}}{\sum_{k=1}^{K+1}\alpha_k^{(t)}\gamma_{kj}^{(t)}}} \tag{17}$$

$$= \frac{\sum_{j=1}^{N}\widetilde{X}_j\cdot\frac{\alpha_i^{(t)}\gamma_{ij}^{(t)}}{\sum_{k=1}^{K+1}\alpha_k^{(t)}\gamma_{kj}^{(t)}}}{\sum_{j=1}^{N}\widetilde{X}_j\frac{\sum_{h=1}^{K+1}\alpha_h^{(t)}\gamma_{hj}^{(t)}}{\sum_{k=1}^{K+1}\alpha_k^{(t)}\gamma_{kj}^{(t)}}} \tag{18}$$

$$= \frac{\sum_{j=1}^{N}\widetilde{X}_j\cdot\frac{\alpha_i^{(t)}\gamma_{ij}^{(t)}}{\sum_{k=1}^{K+1}\alpha_k^{(t)}\gamma_{kj}^{(t)}}}{\sum_{j=1}^{N}\widetilde{X}_j} \tag{19}$$

$$= \frac{\sum_{j=1}^{N}\widetilde{X}_j\cdot\frac{\alpha_i^{(t)}\gamma_{ij}^{(t)}}{\sum_{k=1}^{K+1}\alpha_k^{(t)}\gamma_{kj}^{(t)}}}{C} \tag{20}$$

$$= \sum_{j=1}^{N}\frac{\widetilde{X}_j}{C}\cdot\frac{\alpha_i^{(t)}\gamma_{ij}^{(t)}}{\sum_{k=1}^{K+1}\alpha_k^{(t)}\gamma_{kj}^{(t)}} \tag{21}$$

Now, let's solve for γ_{ij}, where $1 \leq i \leq K$

$$\frac{\partial L(\alpha,\gamma)}{\partial \gamma_{ij}} = 0 \iff \frac{y_{ij}}{\gamma_{ij}} + \sum_{c=1}^{C} \frac{w_{ci}^{(t)} x_{cj}}{\gamma_{ij}} - 1 = 0 \tag{22}$$

$$\iff \gamma_{ij} = y_{ij} + \sum_{c=1}^{C} w_{ci} x_{cj} = y_{ij} + \sum_{c=1}^{C} \frac{\alpha_i^{(t)} \gamma_{i,x_c}^{(t)}}{\sum_{k=1}^{K+1} \alpha_k^{(t)} \gamma_{k,x_c}^{(t)}} \cdot x_{cj} = y_{ij} + \widetilde{X}_j \frac{\alpha_i^{(t)} \gamma_{ij}^{(t)}}{\sum_{k=1}^{K+1} \alpha_k^{(t)} \gamma_{kj}^{(t)}} \tag{23}$$

Since we have the constraint $\sum_{j=1}^{N} \gamma_{ij} = 1$, therefore:

$$\gamma_{ij}^{(t)} = \frac{y_{ij} + \widetilde{X}_j \frac{\alpha_i^{(t)} \gamma_{ij}^{(t)}}{\sum_{k=1}^{K+1} \alpha_k^{(t)} \gamma_{kj}^{(t)}}}{\sum_{h=1}^{N} y_{ih} + \widetilde{X}_h \frac{\alpha_i^{(t)} \gamma_{ih}^{(t)}}{\sum_{k=1}^{K+1} \alpha_k^{(t)} \gamma_{kh}^{(t)}}} \tag{24}$$

Now, let's solve for γ_{ij}, where $i = K+1$

$$\frac{\partial L(\alpha,\gamma)}{\partial \gamma_{K+1,j}} = 0 \iff \sum_{c=1}^{C} \frac{w_{ci}^{(t)} x_{cj}}{\gamma_{K+1,j}} - 1 = 0 \iff \gamma_{K+1,j} = \sum_{c=1}^{C} w_{ci}^{(t)} x_{cj} = \widetilde{X}_j \frac{\alpha_i^{(t)} \gamma_{ij}^{(t)}}{\sum_{k=1}^{K+1} \alpha_k^{(t)} \gamma_{kj}^{(t)}} \tag{25}$$

Since we need $\sum_{j=1}^{N} \gamma_{K+1,j} = 1$, then:

$$\gamma_{K+1,j}^{(t+1)} = \frac{\widetilde{X}_j \frac{\alpha_i^{(t)} \gamma_{ij}^{(t)}}{\sum_{k=1}^{K+1} \alpha_k^{(t)} \gamma_{kj}^{(t)}}}{\sum_{h=1}^{N} \widetilde{X}_h \frac{\alpha_i^{(t)} \gamma_{ih}^{(t)}}{\sum_{k=1}^{K+1} \alpha_k^{(t)} \gamma_{kh}^{(t)}}} \tag{26}$$

6. FEAST algorithm

As you can see, in the E-step above, we calculated the expected value of source membership for **each read** in the sink z_{ci} for $1 \leq c \leq C$. In real implementation, we will do **NOT** do that because it requires unnecessary computational cost. The EM-algorithm can be shortened by the observation that $\widetilde{X}_j \frac{\alpha_i^{(t)} \gamma_{ij}^{(t)}}{\sum_{k=1}^{K+1} \alpha_k^{(t)} \gamma_{kj}^{(t)}}$ appears repeatedly in the update rules for α_i and γ_{ij} (in the M-step).

Algorithm 1 FEAST algorithm

Require: $\alpha_i \geq 0$ and $\gamma_{ij} \geq 0$
Ensure: $\sum_{i=1}^{K+1} \alpha_i = 1$ and $\sum_{j=1}^{N} \gamma_{ij} = 1$
  $\alpha_i \leftarrow \alpha_i^{(0)}$
  $\gamma_{ij} \leftarrow \gamma_{ij}^{(0)}$
  while $L(\alpha,\gamma)$ **do** does not converge
    $\widetilde{W}_{ij} \leftarrow \widetilde{X}_j \frac{\alpha_i^{(t)} \gamma_{ij}^{(t)}}{\sum_{k=1}^{K+1} \alpha_k^{(t)} \gamma_{kj}^{(t)}}$ ▷ E-step: expected number of reads of taxa j in sink from source i
    $\alpha_i^{(t+1)} \leftarrow \sum_{j=1}^{N} \frac{\widetilde{W}_{ij}}{C}$ ▷ M-step: Update on α_i
    $\gamma_{ij}^{(t+1)} \leftarrow \frac{y_{ij} + \widetilde{W}_{ij}}{\sum_{h=1}^{N} y_{ih} + \widetilde{W}_{ih}}$ for $1 \leq i \leq K$ ▷ M-step: Update on γ_{ij}
    $\gamma_{ij}^{(t+1)} \leftarrow \frac{\widetilde{W}_{K+1,j}}{\sum_{h=1}^{N} \widetilde{W}_{K+1,h}}$ ▷ M-step: Update on $\gamma_{K+1,j}$
  end while

References

[1] Liat Shenhav, Mike Thompson, Tyler A. Joseph, Leah Briscoe, Leah Furman, David Bogumil, Itzhak Mizrahi, Itsik Pe'er, and Eran Halperin. Feast: fast expectation-maximization for microbial source tracking. *Nature Methods*, 2019.

Los Angeles CA