Assignment 1

Ethan Holleman

October 26, 2021

1 Part 2

$$\ln P(X, C|\boldsymbol{\theta}) = \sum_i \sum_j C_{i,j} \ln \lambda_j + \sum_i \sum_j \sum_p \sum_k X_{i,j,p}[C_{i,j} \ln \psi^{(1)}_{p,k} + (1 - C_{i,j}) \ln \psi^{(0)}_{p,k}] \tag{1}$$

1.1 E step

$P(\boldsymbol{C}|\boldsymbol{X}, \boldsymbol{\theta})$ is effectively telling us for every motif for each sequence, was this motif likely to be a transcription factor binding site given all motifs and the parameters of the model. The probability of a motif being the transcription factor binding site is ultimately a function of the nucleotides that compose that motif and the position of those nucleotides within the motif.

$$P(\boldsymbol{C}|\boldsymbol{X}, \boldsymbol{\theta}) = \frac{P(\boldsymbol{X}, \boldsymbol{C}|\boldsymbol{\theta})}{P(\boldsymbol{X}|\boldsymbol{\theta})} \tag{2}$$

Product individual posteriors is proportional to posterior over all latent variables.

$$P(\boldsymbol{C}|\boldsymbol{X}, \boldsymbol{\theta}) \propto \prod_i \prod_j P(X_{i,j}, C_i|\boldsymbol{\theta}) \tag{3}$$

$$P(X_{i,j}, C_i|\boldsymbol{\theta}) = \prod_p \prod_k [P(X_{i,j,p=k}|C_i = j, \boldsymbol{\theta})^{C_{i,j}} P(X_{i,j,p=k}|C_i \neq j, \boldsymbol{\theta})^{1-C_{i,j}}]^{X_{i,j,p,k}} \tag{4}$$

$$P(X_{i,j}|\boldsymbol{\theta}) = \prod_j \lambda_j \psi^{(0)}_{p,k} + \prod_j \lambda_j \psi^{(1)}_{p,k} \tag{5}$$

$$P(\boldsymbol{C}|\boldsymbol{X},\boldsymbol{\theta}) = \frac{\prod_p \prod_k [P(X_{i,j,p=k}|C_i = j, \boldsymbol{\theta})^{C_{i,j}} P(X_{i,j,p=k}|C_i \neq j, \boldsymbol{\theta})^{1-C_{i,j}}]^{X_{i,j,p,k}}}{\prod_j \lambda_j \psi^{(0)}_{p,k} + \prod_j \lambda_j \psi^{(1)}_{p,k}} \tag{6}$$

1.2 M step

$$\boldsymbol{E}_q[\log P(\boldsymbol{X},\boldsymbol{C}|\boldsymbol{\theta})] = \sum_i \sum_j C_{i,j} \ln \lambda_j + \sum_i \sum_j \sum_p \sum_k X_{i,j,p}[C_{i,j} \ln \psi^{(1)}_{p,k} + (1-C_{i,j}) \ln \psi^{(0)}_{p,k}] \tag{7}$$

$$= \sum_i \sum_j \boldsymbol{E}_q[C_{i,j} \ln \lambda_j] + \sum_i \sum_j \sum_p \sum_k \boldsymbol{E}_q[X_{i,j,p}[C_{i,j} \ln \psi^{(1)}_{p,k} + (1-C_{i,j}) \ln \psi^{(0)}_{p,k}]] \tag{8}$$

$$= \sum_i \sum_j \boldsymbol{E}_q[C_{i,j}] \ln \lambda_j + \sum_i \sum_j \sum_p \sum_k \boldsymbol{E}_q X_{i,j,p}[C_{i,j} \ln \psi^{(1)}_{p,k} + (1-C_{i,j}) \ln \psi^{(0)}_{p,k}] \tag{9}$$

1.2.1 Derivative with respect to λ_j

Constraint equation defines bounds of λ_j that we must optimize $\boldsymbol{E}_q[\log P(\boldsymbol{X},\boldsymbol{C}|\boldsymbol{\theta})]$ within. λ_j is probability that a particular motif is the TFBS. There is only one TFBS per sequence so probability over all possible motifs must sum to 1.

$$g(\lambda_j) = \sum_j \lambda_j - 1 \tag{10}$$

$$h(\lambda_j) = \mathrm{ELBO}(q_t(C), \boldsymbol{\theta}) + \phi g(\lambda_j) \tag{11}$$

$$\frac{\partial h}{\partial \lambda_j} = \frac{\sum_i E_q[C_{i,j}]}{\lambda_j} - \phi \tag{12}$$

$$\lambda_j = \frac{\sum_i E_q[C_{i,j}]}{\phi} \tag{13}$$

Derivative with respect to ϕ (Lagrangian multiplier) is just the constraint equation.

$$\frac{\partial h}{\partial \phi} = g(\lambda_j) \tag{14}$$

Set equal to zero and plug in value for λ_j.

$$0 = \frac{\sum_i \sum_j E_q[C_{i,j}]}{\phi} - 1 \tag{15}$$

$$1 = \frac{\sum_i \sum_j E_q[C_{i,j}]}{\phi} \tag{16}$$

$$\phi = \sum_i \sum_j E_q[C_{i,j}] \tag{17}$$

Plug back into $\lambda_j = \frac{\sum_i E_q[C_{i,j}]}{\phi}$.

$$\lambda_j = \frac{\sum_i E_q[C_{i,j}]}{\sum_i \sum_j E_q[C_{i,j}]} \tag{18}$$

1.3 Derivative with respect to $\psi_{p,k}^{(0)}$

Define constraint equation to optimize $\psi_{p,k}^{(C_{i,j})}$ within.

$$g(\psi_{p,k}^{(0)}) = \sum_k \psi_{p,k}^{(0)} - 1 \tag{19}$$

$$h(\psi_{p,k}^{(0)}) = \mathrm{ELBO}(q_t(C), \boldsymbol{\theta}) + \phi g(\psi_{p,k}^{(0)}) \tag{20}$$

$$\frac{\partial h}{\partial \psi_{p,k}^{(0)}} = \frac{\sum_i \sum_j E_q X_{i,j,p,k} C_{i,j}}{\psi_{p,k}^{(0)}} - \phi \tag{21}$$

Set derivative equal to zero and solve for $\psi_{p,k}^{(0)}$.

$$\phi \psi_{p,k}^{(0)} = \sum_i \sum_j E_q X_{i,j,p,k} C_{i,j} \tag{22}$$

$$\psi_{p,k}^{(0)} = \frac{\sum_i \sum_j E_q X_{i,j,p,k} C_{i,j}}{\phi} \tag{23}$$

Take derivative of $h(\psi_{p,k}^{(0)})$ with respect to ϕ.

$$\frac{\partial h}{\partial \phi} = \sum_k \psi_{p,k}^{(0)} - 1 \tag{24}$$

$$= \frac{\sum_k \sum_i \sum_j E_q X_{i,j,p,k} C_{i,j}}{\phi} - 1 \tag{25}$$

$$\phi = \sum_k \sum_i \sum_j E_q X_{i,j,p,k} C_{i,j} \tag{26}$$

$$\psi_{p,k}^{(0)} = \frac{\sum_i \sum_j E_q X_{i,j,p,k} C_{i,j}}{\sum_k \sum_i \sum_j E_q X_{i,j,p,k} C_{i,j}} \tag{27}$$

1.4 Derivative with respect to $\psi_{p,k}^{(0)}$

Define constraint equation to optimize within. $\psi_{p,k}^{(C_{i,j})}$

$$g(\psi_{p,k}^{(1)}) = \sum_k \psi_{p,k}^{(1)} - 1 \tag{28}$$

$$h(\psi_{p,k}^{(1)}) = \mathrm{ELBO}(q_t(C), \boldsymbol{\theta}) + \phi g(\psi_{p,k}^{(1)}) \tag{29}$$

$$\frac{\partial h}{\partial \psi_{p,k}^{(1)}} = \frac{\sum_i \sum_j E_q X_{i,j,p,k}(1 - C_{i,j})}{\psi_{p,k}^{(1)}} - \phi \tag{30}$$

$$\psi_{p,k}^{(1)} = \frac{\sum_i \sum_j E_q X_{i,j,p,k}(1 - C_{i,j})}{\phi} \tag{31}$$

$$\frac{\partial h}{\partial \phi} = \sum_k \psi_{p,k}^{(1)} - 1 \tag{32}$$

$$= \frac{\sum_k \sum_i \sum_j E_q X_{i,j,p,k}(1 - C_{i,j})}{\phi} - 1 \tag{33}$$

$$\psi_{p,k}^{(1)} = \frac{\sum_i \sum_j E_q X_{i,j,p,k}(1 - C_{i,j})}{\sum_k \sum_i \sum_j E_q X_{i,j,p,k}(1 - C_{i,j})} \tag{34}$$