1 Grade of Membership Model for modeling methylation

Consider a bisulfite sequencing experiment that records the number of methylated and unmethylated sites per bin across the genome. The model can be formulated as follows for bin b in the genome and for sample n.

$$M_{nb} \sim Bin\left(Y_{nb} = M_{nb} + U_{nb}, p_{nb}\right)$$

where M_{nb} and U_{nb} denote the number of methylated and unmethylated sites in bin b and for sample n respectively. p_{nb} represents the probability of methylation, which under the Grade of Membership model assumption

$$p_{nb} = \sum_{k=1}^{K} \omega_{nk} g_{kb}$$

where ω_{nk} represent the grades of membership of the nth sample in the kth methylation profile and g_{kb} represents the probability of methylation in bin b for the kth methylation profile. Note that here we assume that the probability of methylation if fixed for all methylation sites in a particular bin for all the clusters.

Intuitively we assume that each bin comprises of methylations coming from one of the K methylation profiles or clusters in the grade of membership model.

Suppose for each CpG site s, we define a latent variable Z_{nks} to be an indicator variable for cluster/profile k for the site s in sample n

$$Pr(Z_{nks} = 1) = \frac{\omega_{nk} g_{k,b(s)}}{\sum_l \omega_{nl} g_{l,b(s)}} = p_{nk,b(s)}$$

where $b(s)$ denotes the bin that the site s belongs to.

Denoting Y_{nb} as the total number of sites in the bin b, we write

$$Y_{nb} = Y_{n1b} + Y_{n2b} + \cdots + Y_{nKb}$$

where we denote

$$Y_{nkb} = M_{nkb} + U_{nkb}$$

and

$$M_{nkb} | Y_{nkb} \sim Bin(Y_{nkb}, f_{kb})$$

$$\left(Y_{n1b}, Y_{n2b}, \cdots, Y_{nKb}\right) \sim Mult\left(Y_{nb}; \omega_{n1}, \omega_{n2}, \cdots, \omega_{nK}\right)$$

$$E\left(M_{nkb}|Y_{nb}\right) = E\left(\left(M_{nkb}|Y_{nkb}^{(t)}\right)|Y_{nb}\right) = E\left(Y_{nkb}g_{kb}^{(t)}|Y_{nb}\right) = Y_{nb}\omega_{nk}^{(t)}g_{kb}^{(t)}$$

But we would like to compute $E(M_{nkb}|M_{nb})$

$$\sum_k E(M_{nkb}|M_{nb}) = M_{nb}$$

$$E\left(M_{nb}|Y_{nb}\right) = \sum_{k=1}^{K} E\left(M_{nkb}|Y_{nb}\right) = Y_{nb}\sum_l \omega_{nl}^{(t)}g_{lb}^{(t)}$$

$$A_{nkb}^{(t)} = E\left(M_{nkb}|M_{nb}, Y_{nb}\right) = M_{nb}\frac{\omega_{nk}^{(t)}g_{kb}^{(t)}}{\sum_l \omega_{nl}^{(t)}g_{lb}^{(t)}}$$

Similarly one can show that

$$B_{nkb}^{(t)} = E\left(U_{nkb}|U_{nb}, Y_{nb}\right) = U_{nb}\frac{\omega_{nk}^{(t)}(1-g_{kb}^{(t)})}{\sum_l \omega_{nl}^{(t)}(1-g_{lb}^{(t)})}$$

Assume now M_{nkb} and U_{nkb} are the latent variables in the EM algorithm. Then the EM log-likelihood is given by

$$\begin{aligned}
E_{L|Data}\left[\log Pr(Data, L|Param)\right] &= \textstyle\sum_{n,b}\sum_k E_{U_{nkb},M_{nkb}|M_{nb},U_{nb},\omega,g}\left[\log Pr(U_{nkb}, M_{nkb}, M_{nb}, U_{nb}|\omega, g\right] && (1)\\
&\propto \textstyle\sum_{n,b}\sum_k A_{nkb}^{(t)} \times \log(\omega_{nk}g_{kb}) + B_{nkb}^{(t)} \times \log(\omega_{nk}(1-g_{kb})) && (2)\\
&\propto \textstyle\sum_{n,b}\sum_k \log(\omega_{nk})(A_{nkb}^{(t)} + B_{nkb}^{(t)}) + \log(g_{kb})A_{nkb}^{(t)} + \log(1-g_{kb})B_{nkb}^{(t)} && (3)\\
& && (4)
\end{aligned}$$

Optimizing for $\omega_{nk}^{(t+1)}$ under the constraint that $\sum_{k=1}^{K}\omega_{nk}^{(t+1)} = 1$, we get

$$\omega_{nk}^{(t+1)} = \frac{\sum_b(A_{nkb}^{(t)} + B_{nkb}^{(t)})}{\sum_l\sum_b(A_{nlb}^{(t)} + B_{nlb}^{(t)})} = \frac{1}{Y_{n+}}\sum_b(A_{nkb}^{(t)} + B_{nkb}^{(t)})$$

where Y_{n+} is the total number of sites for sample n.

Similarly, we can get the estimates for $g_{kb}^{(t+1)}$ as

$$g_{kb}^{(t+1)} = \frac{\sum_n A_{nkb}^{(t)}}{\sum_n(A_{nkb}^{(t)} + B_{nkb}^{(t)})}$$