STA2311 (Fall 2023) - Practice Problems for Class 3 (The EM Algorithm)

1. Consider again the mining town example from Class 2, where we assume the observed data is generated by a zero-inflated Poisson model.

   (a) By defining the indicators
   $$Z_i = \begin{cases} 1, & \text{observation } i \text{ comes from a subpopulation of families with children} \\ 0, & \text{observation } i \text{ comes from a subpopulation of families without children} \end{cases},$$
   formulate this as a missing data problem.

   (b) Devise an EM algorithm for estimating (λ, ξ).

2. Consider again the allele example from Class 2, in which we seek to estimate the true frequencies (p_a, p_b, p_o) of alleles a, b, and o in the population based on observed blood type counts n_A, n_B, n_{AB}, and n_O out of a total sample of size n. Let N_{xy} be the *number* of samples with genotype (i.e., allele pair) xy, for $x, y \in \{a, b, o\}$. Then $n_A = N_{aa} + N_{ao}$ and $n_B = N_{bb} + N_{bo}$, where $N_{aa}, N_{ao}, N_{bb}, N_{bo}$ are unknown.

   (a) By defining the indicators
   $$Z_i = \begin{cases} 1, & \text{subject } i \text{ has genotype } aa \\ 0, & \text{subject } i \text{ has genotype } ao \end{cases} \quad \text{and} \quad W_j = \begin{cases} 1, & \text{subject } j \text{ has genotype } bb \\ 0, & \text{subject } j \text{ has genotype } bo \end{cases}$$
   for appropriate i and j, formulate this as a missing data problem.

   (b) Devise an EM algorithm for estimating (p_a, p_b, p_o).

3. Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Exp}(\lambda)$ so that $P(X_i \leq t) = 1 - e^{-\lambda t}$ for all $1 \leq i \leq n$. Suppose we do not observe the X_i values, but only observe whether they fall within three intervals. Let $Z_{1i} = \mathbb{1}_{X_i < a}$, $Z_{2i} = \mathbb{1}_{a \leq X_i < b}$, and $Z_{3i} = \mathbb{1}_{b \leq X_i}$. Based on observed data $\{(Z_{1i}, Z_{2i}, Z_{3i}) : 1 \leq i \leq n\}$ devise an EM algorithm for estimating λ.

4. The file `em-regress.txt` contains measurements on $n = 50$ units. Each unit provides a response Y_i and two covariate values, X_{1i} and X_{2i}. For ten of the units, the response variable has been lost so only the covariate values are available.

   Assume a linear regression model in which
   $$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i, \quad 1 \leq i \leq n$$
   and $\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

   (a) Find the MLE for β_0, β_1, β_2 and σ using *all* of the available data.

   (b) Find the variance of the MLE using an EM-related method of your choice.

5. Let $S(\boldsymbol{\theta} \mid \boldsymbol{y}_{\text{obs}})$ be the score function of the observed-data log-likelihood, and let $S_c(\boldsymbol{\theta} \mid \boldsymbol{y}_{\text{com}})$ be the score function of the complete-data log-likelihood. Assuming the operations of integration and differentiation can be swapped, prove that
   $$S(\boldsymbol{\theta} \mid \boldsymbol{y}_{\text{obs}}) = \mathbb{E}_{\boldsymbol{\theta}}[S_c(\boldsymbol{\theta} \mid \boldsymbol{Y}_{\text{com}}) \mid \boldsymbol{Y}_{\text{obs}} = \boldsymbol{y}_{\text{obs}}].$$

6. Assuming the operations of integration and differentiation can be swapped, prove the identity shown on the bottom of Slide 33:

$$\boldsymbol{\mathcal{I}}_{\text{mis}}(\theta^*) = \mathbb{E}_{\theta^*}\left[\left(\frac{\partial}{\partial\theta}\log\Big(f(\tilde{\boldsymbol{Y}}_{\text{obs}}, \tilde{\boldsymbol{Y}}_{\text{mis}} \mid \theta)\Big)\right)^2\Bigg|_{\theta=\theta^*} \mid \tilde{\boldsymbol{Y}}_{\text{obs}}\right].$$