# Stats 270, Homework 4

**Due date:** **February 20**

1. In this exercise, you will statistically analyze the Wright-Fisher model with mutations. To simplify the analysis, assume that $\Pr(a \to A) = \Pr(A \to a) = u \in (0,1)$, so that transition probabilities of $\{X_n\}$ are

$$p_{ij} = \binom{2m}{j} p_i^j (1-p_i)^{2m-j},$$

where

$$p_i = \frac{i}{2m}(1-u) + \left(1 - \frac{i}{2m}\right) u.$$

   (a) Write a simulation routine to generate realizations from the Markov chain. Setting the mutation probability $u = 0.35$ and gene number $2m = 10$, generate 200 iterations of the chain starting from state 0.

   (b) Using your simulated data, compute the maximum likelihood estimate of the mutation probability $u$. I suggest doing this numerically.

   (c) Obtain a 95% confidence interval for $u$. You will need to estimate the stationary distribution.

   (d) Check your asymptotic-based answers by repeating the simulation and estimation 1000 times and reporting relevant summaries of the resulting empirical distribution of estimates of $u$.

   (e) Test the null hypothesis $\mathrm{H}_0 : u = 0.4$ against the alternative $\mathrm{H}_1 : u \neq 0.4$ using a likelihood ratio test.

2. Consider a Poisson mixture model

$$\Pr(y = l) = \alpha \frac{\lambda_1^l}{l!} e^{-\lambda_1} + (1-\alpha) \frac{\lambda_2^l}{l!} e^{-\lambda_2}.$$

   (a) Suppose we observe $y_1, \dots, y_n$ samples from the above distribution. Let us augment our data with missing indicators $x_1, \dots, x_n$ with $x_i \in \{1, 2\}$. We assume that $\Pr(x_1 = 1) = 1 - \Pr(x_1 = 2) = \alpha$ and that

$$\Pr(y_i = l \mid x_i = 1) = \frac{\lambda_1^l}{l!} e^{-\lambda_1},$$
$$\Pr(y_i = l \mid x_i = 2) = \frac{\lambda_2^l}{l!} e^{-\lambda_2}.$$

   Write down the log likelihood of the complete data, $(x_1, y_1), \dots, (x_n, y_n)$.

   (b) E-step. Show that to complete the E-step of the EM algorithm, it is sufficient to compute $\beta_{k,i} = \mathrm{E}\left(1_{\{x_i = 1\}} \mid \mathbf{y}, \alpha_k, \lambda_{k,1}, \lambda_{k,2}\right)$, where $k$ indexes EM algorithm iterations. Demonstrate that

$$\beta_{k,i} = \frac{\alpha_k \lambda_{k,1}^{y_i} e^{-\lambda_{k,1}}}{\alpha_k \lambda_{k,1}^{y_i} e^{-\lambda_{k,1}} + (1-\alpha_k) \lambda_{k,2}^{y_i} e^{-\lambda_{k,2}}}$$

(c) M-step. Show that maximizing the expected complete data log likelihood yields

$$\alpha_{k+1} = \frac{\sum_{i=1}^n \beta_{k,i}}{n},$$
$$\lambda_{k+1,1} = \frac{\sum_{i=1}^n \beta_{k,i} y_i}{\sum_{i=1}^n \beta_{k,i}},$$
$$\lambda_{k+1,2} = \frac{\sum_{i=1}^n (1-\beta_{k,i}) y_i}{\sum_{i=1}^n (1-\beta_{k,i})}.$$

(d) Simulate 300 observations from the Poisson mixture model with $\alpha = 0.3$, $\lambda_1 = 1.5$ and $\lambda_2 = 2.8$. Implement the EM algorithm and apply it to the simulated data. Report the parameter values from EM iterations using 3 different sets of initial parameter values.

3. Show that if $(\mathbf{x}, \mathbf{y})$ form a hidden Markov model with

$$\Pr(\mathbf{x}, \mathbf{y}) = \Pr(x_1) \prod_{t=2}^{n} \Pr(x_t \,|\, x_{t-1}) \prod_{t=1}^{n} \Pr(y_t \,|\, x_t), \tag{1}$$

then $\Pr(\mathbf{y}_{t+1:n} \mid x_t, y_t, x_{t-1}) = \Pr(\mathbf{y}_{t+1:n} \mid x_t)$ for $t = 2, \ldots, n-1$.