MAST30027: Modern Applied Statistics

Week 7 Lab

1. Simulate 100 samples from the following mixture model. For $i = 1 \ldots, 100$,

$$Z_i \sim \text{categorical } (0.7, 0.3),$$

$$X_i|Z_i = 1 \sim \text{Binomial}(20, p_1 = 0.2) \text{ and } X_i|Z_i = 2 \sim \text{Binomial}(20, p_2 = 0.8).$$

The binomial distribution has probability mass function

$$f(x; m, p) = \binom{m}{x} p^x (1-p)^{m-x}.$$

Please set a seed using 'set.seed(30027)'. Make a histogram of the simulated samples.

2. We pretend that we just observe the simulated samples. We will assume that the observed data follow a mixture of two binomial distributions. Specifically, for $i = 1 \ldots, 100$,

$$Z_i \sim \text{categorical } (\pi_1, 1 - \pi_1),$$

$$X_i|Z_i = 1 \sim \text{Binomial}(20, p_1) \text{ and } X_i|Z_i = 2 \sim \text{Binomial}(20, p_2).$$

We aim to obtain MLE of parameters $\theta = (\pi_1, p_1, p_2)$ using the EM algorithm.

a) Let $X = (X_1, \ldots, X_{100})$ and $Z = (Z_1, \ldots, Z_{100})$. Derive the expectation of the complete log-likelihood, $Q(\theta, \theta^0) = E_{Z|X,\theta^0}[\log(P(X, Z|\theta))]$.

b) Derive E-step and M-step of the EM algorithm.

c) Implement the EM algorithm and obtain MLE of the parameters by applying the implemented algorithm to the simulated data. Run the EM algorithm multiple times with different initial values and pick estimators with the highest incomplete log-likelihood. For each run, check that the incomplete log-likelihoods increases at each step by plotting them.