

# Homework 3.1 - Stat 215A, Fall 2017

Due: Tuesday, October 24, 11:00 AM

## 1 EM Algorithm

Suppose  $X_1, \dots, X_n$  are i.i.d. observations from a mixture of two Poisson distributions:

$$\begin{aligned}P_0(X) &= \frac{\mu_0^X e^{-\mu_0}}{X!} \\P_1(X) &= \frac{\mu_1^X e^{-\mu_1}}{X!}\end{aligned}$$

with mixing probabilities of  $\pi$  and  $1 - \pi$  (i.e. there is an initial probability that an observation  $X_i$  is drawn from  $P_0$  and a probability  $1 - \pi$  from  $P_1$ ).

1. Define the complete data vector and the distribution of the missing random variable.
2. Write down the E and M steps for estimating  $\mu_0, \mu_1, \pi$ .
3. Give an initial estimator to start the EM algorithm.
4. Write down the E and M steps if the second distribution is actually Bernoulli( $p$ )

$$P_1(X) = p^X (1 - p)^{1-X}$$

5. Write R code to implement the E and M steps. Run it on some simulated data where you know the true parameters. Show the accuracy of clustering as you vary the values of  $\mu_0$  and  $\mu_1$ .
6. How would you create confidence intervals for the inferred parameters  $\mu_0$  and  $\mu_1$ ? Can you use an asymptotic normality argument?

## 2 The Linear Model in the Neyman-Rubin framework

Recall the Neyman-Rubin model of potential outcomes:

$$Y_i = T_i a_i + (1 - T_i) b_i$$

where  $T_i \in \{0, 1\}$  is a random treatment indicator for subject  $i$ , and  $a_i, b_i$  are the potential outcomes.

Say there is a covariate  $z_i$  which is measured for each subject. Define  $Q_a$  to be constant used for adjustment of the  $a$  potential outcomes, define  $Q_b$  to be the constant used for adjustment of the  $b$  potential outcomes. Define  $\bar{a}, \bar{b}, \bar{z}$  to be the population means of  $a_i, b_i, z_i$ , respectively.

By simply adding and subtracting some terms, we can write the Neyman - Rubin model as

$$Y_i = \bar{a}T_i + \bar{b}(1 - T_i) + Q_a(z_i - \bar{z})T_i + Q_b(z_i - \bar{z})(1 - T_i) + \epsilon_i$$

where

$$\epsilon_i := T_i(a_i - (\bar{a} + Q_a(z_i - \bar{z}))) + (1 - T_i)(b_i - (\bar{b} + Q_b(z_i - \bar{z})))$$

$$\begin{aligned}E(\epsilon_i) &= E(T_i(a_i - (\bar{a} + Q_a(z_i - \bar{z}))) + (1 - T_i)(b_i - (\bar{b} + Q_b(z_i - \bar{z})))) \\&= p_A(a_i - (\bar{a} + Q_a(z_i - \bar{z}))) + (1 - p_A)(b_i - (\bar{b} + Q_b(z_i - \bar{z})))\end{aligned}$$

1. In your own words, give an interpretation of each term in this decomposition. What are the predictors in the model? What is the response? What are the error terms, and are they i.i.d.?
2. How should you choose  $Q_a$  and  $Q_b$ ? Justify your choice.