MA 589 — Computational Statistics

Project 3

(Due: Tuesday, 10/25/16)

1. A traffic engineer requests your help in identifying "black spots" in his city. He has data on the number of accidents X in one year at $n = 20$ traffic intersections:

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
X_i	2	0	0	1	3	0	1	6	2	0	1	0	2	0	8	0	1	3	2	0

After discussing with him, you both agree to model the number of accidents X_i at intersection i using a *mixture* of Poisson distributions,

$$X_i \mid Z_i \overset{\text{iid}}{\sim} \mathsf{Po}(Z_i\lambda_d + (1 - Z_i)\lambda_c)$$
$$Z_i \overset{\text{iid}}{\sim} \mathsf{Bern}(\pi),$$

where $Z_i = 1$ identifies the i-th intersection as being "dangerous" with a higher rate of accidents per year λ_d and $Z_i = 0$ codes for the intersection being "calm", with a smaller rate λ_c. Your task is to exploit the latent variable (Z) formulation above and estimate π, λ_c, and λ_d using expectation-maximization.

(a) Derive E-step of your EM algorithm: write the complete data log likelihood, and then the expected log likelihood Q by showing that

$$\alpha_i^{(t)} \doteq \mathbb{E}_{Z \mid X; \pi^{(t)}, \lambda_c^{(t)}, \lambda_d^{(t)}}[Z_i] = \mathbb{P}(Z_i = 1 \mid X_i; \pi^{(t)}, \lambda_c^{(t)}, \lambda_d^{(t)})$$
$$= \frac{\pi^{(t)} p(X_i; \lambda_d^{(t)})}{\pi^{(t)} p(X_i; \lambda_d^{(t)}) + (1 - \pi^{(t)}) p(X_i; \lambda_c^{(t)})},$$

where $p(X_i; \lambda)$ is the Poisson pmf with rate λ evaluated at X_i.

(b) Now, for the M-step, differentiate Q to obtain the update equations:

$$\pi^{(t+1)} = \frac{\sum_{i=1}^n \alpha_i^{(t)}}{n}, \quad \lambda_c^{(t+1)} = \frac{\sum_{i=1}^n (1 - \alpha_i^{(t)}) X_i}{\sum_{i=1}^n (1 - \alpha_i^{(t)})}, \quad \lambda_d^{(t+1)} = \frac{\sum_{i=1}^n \alpha_i^{(t)} X_i}{\sum_{i=1}^n \alpha_i^{(t)}}.$$

(c) Starting at $\pi^{(0)} = 0.5$,

$$\lambda_c^{(0)} = \frac{\sum_{i=1}^n X_i I(X_i < \overline{X})}{\sum_{i=1}^n I(X_i < \overline{X})} \quad \text{and} \quad \lambda_d^{(0)} = \frac{\sum_{i=1}^n X_i I(X_i > \overline{X})}{\sum_{i=1}^n I(X_i > \overline{X})}, \tag{1}$$

that is, the trimmed means of the data, run your EM algorithm to obtain estimates of the parameters. Take an absolute precision of 10^{-8} as a stopping criterion.

(d) Based on your EM estimates, what is the probability of the first intersection being dangerous given X_1? What about the fifth intersection? Which intersections would you flag as black spots?

(e) Run your EM algorithm again, but *swapping* the starting values for λ_c and λ_d at Equation 1. Compare your estimates now to the previous values; how can you explain these results?

(f) $*$[1] Rewrite Q to show that, regarding $\alpha^{(t)}$ as data, we can obtain estimates for π, λ_c, and λ_d by assuming $\alpha_i^{(t)} \sim \mathsf{QuasiBinom}(1, \pi)$, $\alpha_i^{(t)} X_i \sim \mathsf{QuasiPo}(\alpha_i^{(t)}\lambda_c)$, and $(1-\alpha_i^{(t)})X_i \sim \mathsf{QuasiPo}((1-\alpha_i^{(t)})\lambda_d)$, and so the update equations in (b) can be computed using R's `glm` (with one step update).

2. You work at a light bulb factory that is experimenting a bulb prototype that is known, by design, to have a lifetime that follows an *exponential* distribution with rate λ failures per month. The factory wishes to estimate the average lifetime of the new bulb—or equivalently, λ—and to this end they give you $n = 100$ bulbs.

   You put the 100 prototypes in a room, light them up, and lock the door; after one month you return and realize that 40 bulbs have failed[2]. You close the door again, wait another month, and now 29 more bulbs have failed. You repeat the process once again for another month to see that 19 bulbs had a lifetime between two and three months. After the third month you finish the experiment with 12 bulbs still working.

   If $Z_i \overset{\text{iid}}{\sim} \mathsf{Exp}(\lambda)$, $i = 1, \ldots, 100$, are the missing lifetimes of the light bulbs in months, the data you actually observe can be coded as

$$X_i = \begin{cases} 0, & \text{if } 0 \le Z_i < 1 \\ 1, & \text{if } 1 \le Z_i < 2 \\ 2, & \text{if } 2 \le Z_i < 3 \\ 3, & \text{if } 3 \le Z_i < \infty \end{cases}$$

   Given that you observe a censored version of Z, you will be developing an EM procedure to estimate λ.

   (a) If $Y \sim \mathsf{Exp}(\lambda)$ then

$$\mathbb{E}[Y \,|\, a \le Y < b] = \frac{1}{\lambda} + \frac{ae^{-\lambda a} - be^{-\lambda b}}{e^{-\lambda a} - e^{-\lambda b}}.$$

   Use this fact to derive the E-step of your algorithm: write the complete data log likelihood and then define the expected log likelihood Q by finding first $\alpha_j^{(t)} \doteq \mathbb{E}[Z_i \,|\, X_i = j; \lambda^{(t)}]$ for all "bands" $j = 0, 1, 2, 3$.

   (b) Differentiate Q with respect to λ to obtain the update equation for the M-step:

$$\frac{1}{\lambda^{(t+1)}} = \frac{\sum_{j=0}^{3} \alpha_j^{(t)} n_j}{\sum_{j=0}^{3} n_j},$$

   where $n_j = \sum_{i=1}^{n} I(X_i = j)$ is the number of failed bulbs in band j.

[1] Only recommended if you know GLMs.

[2] I know, it's a lame bulb... But it could be *really* cheap!

(c) Now that you have both steps, implement and run your EM procedure to obtain $\widehat{\lambda}$ checking for convergence to within 10^{-8} in absolute precision. Choose a reasonable starting value (explain your choice.) What is your estimate for the average lifetime of the new light bulb?

(d) Suppose you only know that $X = 40$ bulbs have failed within a month. Use the update equation based on an expected log likelihood Q for this case to show that $\widehat{\lambda} = -\log(1 - X/n)$. (Hint: explicitly write $\alpha_j^{(t)}$ in Q as a function of $\lambda^{(t)}$ and assume convergence of the update equation[3].)

[3] Can you also show that this is the MLE for λ?