MA 589 — Computational Statistics

Project 3

(Due: Friday, 10/26/18)

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,

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

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

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

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

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

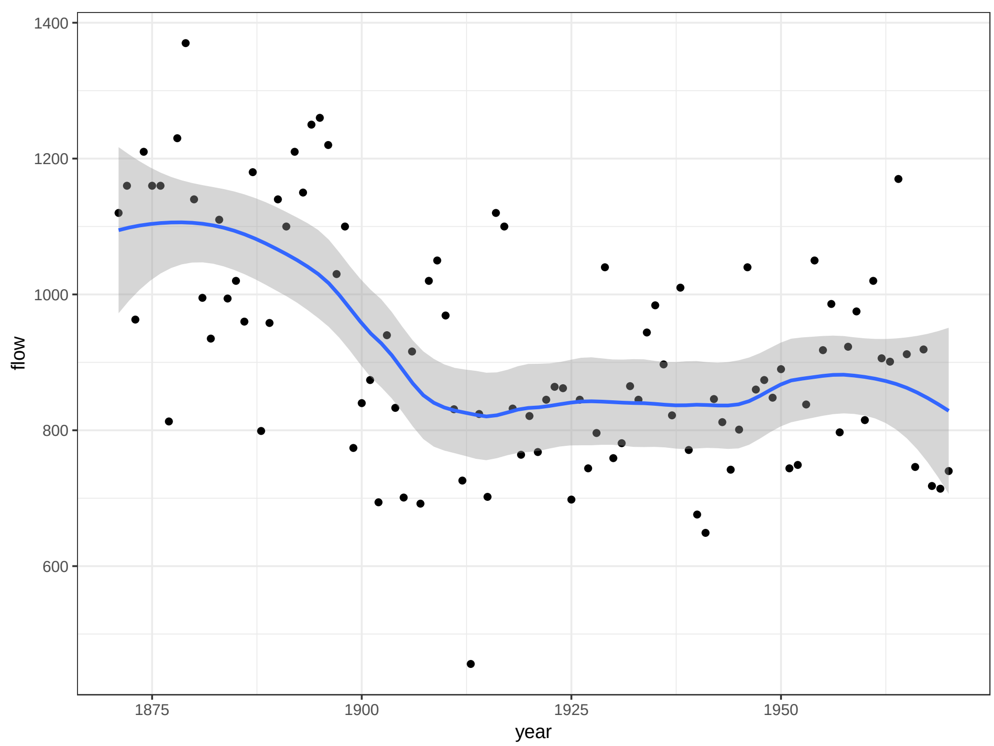
that is, the trimmed means of the data, run your EM algorithm to obtain estimates of the parameters. Take an absolute precision of 10^{-8} as a stopping criterion.

(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. The Nile river had an apparent change in its flow around the turn of the last century. Dataset "Nile", available in `R`[2], contains annual flows in 100 million m^3 from 1871 to 1970:



To assess if the river flow had a *change point*, assume that the time series data X of length n has a mixture normal distribution: $X \sim \sum_{i=1}^{n-1} N(\mu_{(i)}, \sigma^2 I_n)/(n-1)$, where $\mu_{(i)}$ is a mean vector with change point at i, $\mu_{(i)} = [\mu_1 I(j \le i) + \mu_2 I(j > i)]_{j=1,\ldots,n}$, that is,

$$\mu_{(1)} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_2 \\ \vdots \\ \mu_2 \\ \mu_2 \end{bmatrix}, \mu_{(2)} = \begin{bmatrix} \mu_1 \\ \mu_1 \\ \mu_2 \\ \vdots \\ \mu_2 \\ \mu_2 \end{bmatrix}, \ldots, \mu_{(n-1)} = \begin{bmatrix} \mu_1 \\ \mu_1 \\ \mu_1 \\ \vdots \\ \mu_1 \\ \mu_2 \end{bmatrix}.$$

[1] Only recommended if you know GLMs.

[2] Load it with `data(Nile)`.

Alternatively, we can represent the mixture using an indicator variable $Z \in \{1, \dots, n-1\}$, with $\mathbb{P}(Z = 1) = \cdots = \mathbb{P}(Z = n-1) = 1/(n-1)$, for the location of the change point:

$$X_j \mid Z \overset{\text{iid}}{\sim} N\Big(\mu_1 I(j \le Z) + \mu_2 I(j > Z), \sigma^2\Big), \qquad j = 1, \dots, n.$$

Thus, we need to estimate $\theta = (\mu_1, \mu_2, \sigma^2)$, and under the latent formulation we can use expectation-maximization.

(a) Write down the log-likelihood and show that, up to a constant,

$$\begin{aligned} Q(\theta; \theta^{(t)}) &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n-1} \mathbb{E}_{Z \mid X; \theta^{(t)}}\big[I(Z = i)\big](X - \mu_{(i)})^\top (X - \mu_{(i)}) \\ &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{j=1}^{n} \Big\{\mathbb{E}_{Z \mid X; \theta^{(t)}}\big[I(j \le Z)\big](X_j - \mu_1)^2 + \\ &\qquad\qquad \mathbb{E}_{Z \mid X; \theta^{(t)}}\big[I(j > Z)\big](X_j - \mu_2)^2\Big\} \end{aligned}$$

(b) Show that the E-step defines

$$\pi_i^{(t)} \doteq \mathbb{E}_{Z \mid X; \theta^{(t)}}\big[I(Z = i)\big] = \frac{\exp\{-(X - \mu_{(i)}^{(t)})^\top (X - \mu_{(i)}^{(t)})/(2\sigma^{2(t)})\}}{\sum_{k=1}^{n-1}\exp\{-(X - \mu_{(k)}^{(t)})^\top (X - \mu_{(k)}^{(t)})/(2\sigma^{2(t)})\}}$$

where $\mu_{(i)}^{(t)} = [\mu_1^{(t)} I(j \le i) + \mu_2^{(t)} I(j > i)]_{j=1,\dots,n}$, and thus $\mathbb{E}_{Z \mid X; \theta^{(t)}}\big[I(j \le Z)\big] = \sum_{i=j}^{n-1}\pi_i^{(t)}$ and $\mathbb{E}_{Z \mid X; \theta^{(t)}}\big[I(j > Z)\big] = \sum_{i=1}^{j-1}\pi_i^{(t)}$. Note that you need to compute $\log\pi_i^{(t)}$ all along to avoid underflows.

(c) Derive the M-step for the updates and show that

$$\mu_1^{(t+1)} = \frac{\sum_{j=1}^n \mathbb{E}_{Z \mid X; \theta^{(t)}}\big[I(j \le Z)\big]X_j}{\sum_{j=1}^n \mathbb{E}_{Z \mid X; \theta^{(t)}}\big[I(j \le Z)\big]}, \quad \mu_2^{(t+1)} = \frac{\sum_{j=1}^n \mathbb{E}_{Z \mid X; \theta^{(t)}}\big[I(j > Z)\big]X_j}{\sum_{j=1}^n \mathbb{E}_{Z \mid X; \theta^{(t)}}\big[I(j > Z)\big]},$$

$$\text{and} \quad \sigma^{2(t+1)} = \frac{1}{n}\sum_{i=1}^{n-1}\pi_i^{(t)}(X - \mu_{(i)}^{(t+1)})^\top (X - \mu_{(i)}^{(t+1)}).$$

(d) Now apply your EM algorithm to the Nile dataset. Start with $\mu_1^{(0)}$ as the mean of the first half of the series, $\mu_2^{(0)}$ as the mean for the second half, and $\sigma^{2(0)} = (\sum_{j=1}^{n/2}(X_j - \mu_1^{(0)})^2 + \sum_{j=n/2+1}^{n}(X_j - \mu_2^{(0)})^2)/n$ and iterate until convergence. Take a relative precision of 10^{-6} as stopping criterion.

What are your EM estimates $\widehat{\mu}_1$, $\widehat{\mu}_2$, and $\widehat{\sigma}^2$? To better visualize the estimates, make two plots: one with $\pi_i^{(T)}$ at the last iteration T over years, to assess the change point location; and other plot with the time series data and two horizontal lines for $\widehat{\mu}_{(i^*)}$, where i^* is the year that maximizes $\pi^{(T)}$ (the two horizontal lines correspond to the EM estimates for μ_1 up to i^* and μ_2 from i^* to the end of the series). For the last plot, also add confidence bands of $\pm\widehat{\sigma}$.