Modern Computational Statistics, Problem Set 3

November 16, 2019 Due 12/02/2019

Problem 1.

Suppose a sample of size n is drawn from a mixture of two normal populations. Specifically, the density of the observed y_i is

$$p(y_i|w) = w\mathcal{N}(y_i|\mu_1, \sigma_1^2) + (1-w)\mathcal{N}(y_i|\mu_2, \sigma_2^2)$$

where $w \in [0, 1]$ is the mixture proportion.

(1) Assume μ_1, σ_1^2 and μ_2, σ_2^2 are all unknown. Introduce appropriate latent indicators and describe and implement an EM algorithm for computing the MLE of $(w, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$.
(2) Apply your program to the following data set, which records pH measurements of several water samples in various locations near Gittaz Lake in the French Alps (data taken from Houtot et al., 2002, Geophysics, 1048–1060)

$$y = (8.1, 8.2, 8.1, 8.2, 8.2, 7.4, 7.3, 7.4, 8.1, 8.1, 7.9, 7.8, 8.2, 7.9, 7.9, 8.1, 8.1).$$

Try different starting values. Does your EM algorithm always converge? Is there more than one mode?

Problem 2.

A total of n instruments are used to observe the same astronomical source. Suppose the number of photons recorded by instrument j can be modeled as $y_j \sim \text{Poisson}(x_j\theta + r_j)$ where $\theta \geq 0$ is the parameter of interest, and x_j and r_j are known positive constants. You may think of θ, x_j, r_j as the source intensity, the observation time, and the background intensity for instrument j, respectively. Assume the photon counts across different instruments are independent.

(1) Write down the likelihood function for θ.
(2) Introduce mutually independent latent variables $z_{j1} \sim \text{Poisson}(x_j\theta)$ and $z_{j2} \sim \text{Poisson}(r_j)$ and suppose we observe only $y_j \equiv z_{j1} + z_{j2}$. Under this formulation, derive an EM algorithm to find the MLE of θ.

Table 1: Data (x_j, r_j, y_j) for Problem 2

x	1.41	1.84	1.64	0.85	1.32	1.97	1.70	1.02	1.84	0.92
r	0.94	0.70	0.16	0.38	0.40	0.57	0.24	0.27	0.60	0.81
y	13	17	6	3	7	13	8	7	5	8

(3) Apply your EM algorithm to the data set given by Table 1. What is the MLE?
(4) For these data compute the observed Fisher information and the fraction of missing

information. (Recall the observed Fisher information is defined as the negative second derivative of the observed data log-likelihood evaluated at the MLE.)

Problem 3.
Consider the following scenario: you have an incomplete dataset consisting of 478 observations with 2 binary variables, Y_1 and Y_2. Y_1 and Y_2 are both observed for 300 observations, Y_1 is observed but Y_2 is missing for 88 observations, and Y_1 is missing but Y_2 is observed for 90 observations. Table 2 gives the counts for all cases. Assume the complete counts have a multinomial distribution with probability vector

$$\pi = [\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}]^\top,$$

where $\pi_{ij} = \text{Prob}(Y_1 = i, Y_2 = j)$. We wish to obtain the maximum likelihood estimate of π.

		Y_2		
		1	2	Missing
Y_1	1	100	50	28
	2	75	75	60
	Missing	30	60	

Table 2: A 2 × 2 Table with Supplemental Margins for Both Variables

(1) Write down the observed data log-likelihood. What assumptions are you making?
(2) Derive the E-step and M-step of an EM algorithm for calculating $\hat{\pi}$, the MLE of π.
(3) Apply the EM algorithm to numerically compute $\hat{\pi}$ for the data given in Table 2.
(4) Compare the ML estimate of the odds ratio $\pi_{11}\pi_{22}\pi_{12}^{-1}\pi_{21}^{-1}$ with the estimate from the complete cases (that is, using only those units with both Y_1 and Y_2 observed). Are they identically equal?