Ethan Weiland 11/09/22

Stat 610 Homework 7

Thursday, November 10, 11:59pm

Assignment

This assignment is (slightly) adapted from Lange, "Numerical Analysis for Statisticians."

Consider the data from The London Times during the years 1910-1912 given in the table below.

Deaths i	Frequency n_i
0	162
1	267
2	271
3	185
4	111
5	61
6	27
7	8
8	3
9	1

The column labeled "Deaths i" refers to the number of deaths to women 80 years and older reported by day. The column labeled "Frequency n_i" refers to the number of days with i deaths. A Poisson distribution gives a poor fit to these data, possibly because of different patterns of deaths in winter and summer. A mixture of two Poissons provides a much better fit. Under the Poisson admixture model, the likelihood of the observed data is

$$\prod_{i=0}^{9}\left[\alpha e^{-\mu_1}\frac{\mu_1^i}{i!}+(1-\alpha)e^{-\mu_2}\frac{\mu_2^i}{i!}\right]^{n_i}$$

where $\alpha$is the admixture parameter and μ_1 and μ_2 are the means of the two Poisson distributions.

Implement an EM algorithm for this model. Let $\theta=(\alpha,\mu_1,\mu_2)^T$ and

$$z_i(\theta)=\frac{\alpha e^{-\mu_1}\mu_1^i}{\alpha e^{-\mu_1}\mu_1^i+(1-\alpha)e^{-\mu_2}\mu_2^i}$$

be the posterior probability that a day with i deaths belongs to Poisson population 1.

Check that the EM algorithm is given by

$$\begin{aligned}\alpha_{m+1}&=\frac{\sum_i n_i z_i(\theta_m)}{\sum_i n_i}\\ \mu_{m+1,1}&=\frac{\sum_i n_i i z_i(\theta_m)}{\sum_i n_i z_i(\theta_m)}\\ \mu_{m+1,2}&=\frac{\sum_i n_i i[1-z_i(\theta_m)]}{\sum_i n_i[1-z_i(\theta_m)]}\end{aligned}$$

From the initial estimates $\alpha_0 = .3$, $\mu_{0,1} = 1$, and $\mu_{0,2} = 2.5$, compute via the EM algorithm the maximum likelihood estimates. Note how slowly the EM algorithm converges in this example.

Submission parameters

Submit two files:

- A pdf writeup containing an explanation of the update formulas and the parameter estimates after each iteration of the algorithm.
- A file containing the code you used.

Explanation of the update formulas:

1. In EM, we would like to maximize the observed data likelihood (which we are given)
2. We start with an initial estimate of the parameters
3. We then compute the expected value of the complete data likelihood given the observed data and our guess at the parameters
4. The update formulas are formulas that produce the maximum of this complete data likelihood

The parameter estimates after 1, 2, 3, 4, 5, and 100,000 iterations are printed in the R script. The parameter estimates after 100,000 iterations converge as they should to those listed in the textbook.