Statistics 221 – Assignment 2

Due: Wednesday, March 17, 2004

1. The Annual Report of the Pension Fund for 1952 reported that some 4,075 widows received pensions from the fund. The following table shows the number of children of these widows. A plausible model for these data is to assume that we are observing a mixture of two

# of children	0	1	2	3	4	5	6
# of widows	3062	587	284	103	33	4	2

   populations, population A which is always zero (observed with probability p), and population B, which follows a Poisson distribution with mean λ (observed with probability $1-p$). Use both Newton-Raphson and the method of scoring to estimate p and λ, and obtain their corresponding standard error estimates. Compare the results from these two methods. Report your convergence criterion and comment on why it was chosen.

2. Suppose W is a non-negative random variable having an exponential distribution with mean $\mu > 0$. Thus its probability density function is given by

$$f(w;\mu) = \frac{1}{\mu} e^{-w/\mu} I_{(0,\infty)}(w). \tag{1}$$

   In survival or reliability analyses, a study to observe a random sample $W_1, \ldots, W_n$ from (1) will generally be terminated in practice before all of these random variables are able to be observed. Let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T$, where $\mathbf{y}_j = (c_j, \delta_j)^T$ and $\delta_j = 0$ or 1 according to whether the observation W_j is censored or not at c_j. That is, if the observation W_j is uncensored, its realized value w_j is equal to c_j, whereas if it is censored at c_j, then w_j is some value greater than c_j.

   (a) Find the MLE $\hat{\mu}$ analytically.

   (b) Derive an EM algorithm for estimating μ

   (c) Show that the EM algorithm for this problem has a unique solution.

   (d) Find the rate of convergence of the EM sequence explicitly.

3. Suppose $y_1, \ldots, y_n$ are a random sample from a finite mixture model with density

$$L(\theta) = \sum_{k=1}^{K} \pi_k f(y_i; \mu_k, \sigma_k^2)$$

   where $f(y;\mu_k,\sigma_k^2)$ is the normal density with mean μ_k and variance σ_k^2. The $\pi_k, \mu_k, \sigma_k^2, k = 1, \ldots, K$ are unknown parameters, but assume that K is known (also note that $\sum_k \pi_k = 1$). One way to think of this is that each y_i is drawn from one of the normal populations with density $f(y;\mu_k,\sigma_k^2)$, but we don't know which one (and different y's may be drawn from

different populations). Finite mixture models also provide a convenient, flexible family that gives a good approximation to many non-normal distributions.

A natural augmented data model in this problem is to let $z_{ik} = 1$ if y_i is drawn from population k and $z_{ik} = 0$ otherwise. The z_{ik} are not observed, but if they were, the analysis would be simplified considerably.

(a) Give the joint distribution of $\{y_i, z_{i1}, \ldots, z_{iK}\}$.

(b) E-step: Give a formula for $Q(\theta|\theta_0)$, the expectation (at θ_0) of the complete data log likelihood conditional on the observed data. (θ represents all the parameters.)

(c) M-step: Give formulas for the values θ_1 of θ which maximizes $Q(\theta|\theta_0)$.

(d) Using the formulas derived above, write a program to implement the EM algorithm for arbitrary K and $(y_1, \ldots, y_n)$.

(e) The file leuk.txt (available on the Assignments page) contains data on the expression of 20 genes taken from 72 subjects with leukemia. It is believed that different forms of leukemia exist in this data set and that some of these genes can be used to examine the different forms of the disease. For $K = 1, 2, 3$, fit the model for the first 3 genes in the dataset (i.e. the first 3 columns). Also calculate the observed data log likelihood at the MLE ($\log L(\hat{\theta}_K)$) for each case.

(f) To decide the number of classes, the Bayesian Information Criterion (BIC) can be used. BIC chooses the model with the smallest value of

$$BIC(K) = -2 \log L(\hat{\theta}_K) + p(K) \log n$$

where $p(K)$ is the number of parameters in the mixture model with K classes.

For each of the 3 genes, what is the estimated number of classes?

(g) Calculate the observed information matrix at the MLE when $K = 2$ for the first gene, using the following methods:

i. Louis' method.

ii. SEM algorithm

Warning: Mixture models like this are known to be multimodal, so you need to be a bit careful, particularly if you try fitting the model with k larger than the true number of classes. Some of these modes will have a $\sigma_k = 0$, which is not an interesting solution. Think a bit about your starting values, try different ones, and watch whether you seem to be going to an "uninteresting" answer.

In this data set, there are actually three forms of leukemia included. The true form of leukemia for each sample is given in the last column of the data file. Its possible you might observe the problem described above with the first two genes, but probably not with the third. A question to think about (but don't hand in), why I am making this speculation? Simple graphical summaries might illustrate why.