36-711 Statistical Computing
Homework 4
Due Nov XX, 2002

1. If the complete data X follows the exponential family, then the likelihood for X is of the form
$$f(X|\theta) = A(X)e^{B(\theta)+K(X)^T\gamma(\theta)}$$
Suppose $Y = t(X)$ is the observed data, and define
$$\begin{aligned} \mu(\theta) &= E[K(X)|\theta] && \text{complete-data expected suff. stats} \\ \mu_k(Y) &= E[K(X)|Y, \theta^k] && \text{missing-data expected suff. stats} \end{aligned}$$

   (a) Show that the E-M algorithm for estimating θ in this exponential family may be expressed as:
$$\begin{array}{rl} \text{Setup:} & \text{Compute } \mu(\theta) \text{ and an initial guess } \theta_0 \\ \text{For each step } k\text{:} & \\ \text{E step:} & \text{Compute } \mu_k(Y) \\ \text{M step:} & \theta^{k+1} = \text{argmax}_\theta B(\theta) + \mu_k(Y)^T\gamma(\theta) \end{array}$$

   (b) Use the fact that $B(\theta)$ is the log-normalizing constant for the model $f(X|\theta)$, to show that
$$\nabla B(\theta) = -\mu^T(\theta)\nabla\gamma(\theta)$$
Hence, show that setting the gradient equal to zero to find θ^{k+1} in the M step leads to the following restatement of the E-M algorithm:
$$\begin{array}{rl} \text{Setup:} & \text{Compute } \mu(\theta) \text{ and an initial guess } \theta_0 \\ \text{For each step } k\text{:} & \\ \text{E step:} & \text{Compute } \mu_k(Y) \\ \text{M step:} & \theta^{k+1} = \text{root}_\theta \left\{ \mu^T(\theta)\nabla\gamma(\theta) \;=\; \mu_k^T(Y)\nabla\gamma(\theta) \right\} \end{array}$$
If θ are the natural parameters for the model, then $\gamma(\theta) = \theta$. In that case, the M step is just setting $\theta^{k+1} = \text{root}_\theta \{ \mu(\theta) \;=\; \mu_k(Y) \}$. How does this compare with maximum likelihood in exponential families with no missing data?

2. In a Bayesian model for complete data X there will be a likelihood $f(X|\theta)$ and a prior distribution $\pi(\theta)$, and then the complete-data posterior distribution for θ will be of the form
$$\pi(\theta|X) \propto f(X|\theta)\pi(\theta)$$

   (a) Show that if $Y = t(X)$ is the observed data, then the observed-data posterior for θ is of the form
$$\pi(\theta|Y) \propto g(Y|\theta)\pi(\theta)$$
where $g(Y|\theta)$ is the complete-data likelihood.

(b) Define $Q_{post}(\theta, \theta^k) = E[\pi(X|\theta)|Y, \theta^k]$, and show (directly, or by analogy with or extension of the likelihood case considered in class) that

$$\log \pi(\theta|Y) - Q_{post}(\theta, \theta^k)$$

achieves its minimum at $\theta = \theta^k$. Hence the E-M algorithm

$$\begin{array}{ll} \text{E step:} & \text{Compute } Q_{post}(\theta, \theta^k) \\ \text{M step:} & \theta^{k+1} = \text{argmax}_\theta Q_{post}(\theta, \theta^k) \end{array}$$

finds a posterior mode (or saddle point) $\hat{\theta}_{post}$ of the observed-data posterior $\pi(\theta|Y)$.

(c) The quantities $-\nabla^2 \log \pi(\theta|Y)$ and $I_{post}(\theta) = E[-\nabla^2 \log \pi(\theta|Y)|Y]$ function as "observed" and "expected" posterior information functions and can be inverted to produce posterior standard errors for $\hat{\theta}_{post}$ [the proof of this involves the same sort of taylor expansion argument that leads to asymptotic normality of MLE's]. Show, by analogy with or extension of the likelihood case discussed in class, that the following Missing Information Principle holds:

$$-\nabla^2 \log \pi(\theta|Y) = E[-\nabla^2 \log \pi(\theta|X)|Y, \theta] - \text{Var}[\nabla \log \pi(\theta|X)|Y, \theta]$$

3. Let $t_1, \ldots, t_n$ be iid with $f(t) = \rho e^{-\rho t}$, the exponential distribution with parameter ρ. Suppose that we observe (y_i, δ_i) where $y_i = min(t_i, c_i)$ and $\delta_i = 1$ if $t_i < c_i$ (t_i is *uncensored*), and $\delta_i = 0$ otherwise (t_i is *censored*), for $i = 1, \ldots, n$. We want to use E-M to estimate ρ via maximum likelihood.

   (a) Show that the E-step computes

   $$Q(\rho, \rho^k) = n \log \rho - \rho \sum_{i-1}^{n} E[t_i|y_i, \delta_i, \rho^k]$$

   (b) For an uncensored observation, $t_i = y_i$. What is $E[t_i|y_i, \delta_i, \rho^k]$ for a censored observation?

   (c) What is the M-step for this problem?

   (d) Implement the E-M algorithm using the data given below (survivial time after transplant in the Stanford Heart Transplant study). How did you assess convergence? How many iterations were required? Try several different or randomly selected starting values. Is there evidence that the observed likelihood is multimodal?

   (e) Compute a standard error for the best $\hat{\rho}$ you can come up with, using the missing information principle.

   (f) The full data set has 184 observations. Compare your results for the first 45 observations (about 1/4 of the data) with your results for the full data set.

Stanford Heart Transplant Data

```
linux% cp ~/brian/F711/stan.R .
linux% Splus   # or R...
> source("stan.R")
> stan[1:5,]
  time status age
1   15      1  54
2    3      1  40
3   46      1  42
4  623      1  51
5  126      1  48
> dim(stan)
[1] 184   3
```

"time" is y_i, "status" (dead=1, alive=0) is δ_i.

4. Saying

```
> library(MASS)      # may skip this step in R
> data(faithful)     # skip this step in Splus
> attach(faithful)
```

in R or Splus defines a vector `waiting` of $n = 272$ waiting times between eruptions of the geyser Old Faithful at Yellowstone Park. [A related data set, `geyser` is discussed by Venables and Ripley (MASS 3rd ed, pp 252ff; 4th ed. pp. 127ff)].

(a) Make a histogram of the data in `waiting`. You will see that a plausible model for the waiting times $y_1, \ldots, y_n$ is a mixture of two normals:

$$y_i \mid \mu_1, \mu_2, \sigma_1, \sigma_2, q \overset{iid}{\sim} \Big\{ q \frac{1}{\sqrt{2\pi}\sigma_1} \exp[-(y_i - \mu_1)^2/2\sigma_1^2] + (1-q) \frac{1}{\sqrt{2\pi}\sigma_2} \exp[-(y_i - \mu_2)^2/2\sigma_2^2] \Big\} \qquad (*)$$

We wish to estimate the parameters μ_1, μ_2, σ_1, σ_2 and q. Explicitly write the likelihood $g(\underline{y}|\mu_1, \mu_2, \sigma_1, \sigma_2, q)$ for the observed data $\underline{y} = (y_1, \ldots, y_n)$.

(b) The likelihood is greatly simplified by using data augmentation. Consider the latent variable (a.k.a. missing data, or augmentation data)

$$z_i = \begin{cases} 1 & \text{if } y_i \text{ belongs to the left hump} \\ 0 & \text{if } y_i \text{ belongs to the right hump} \end{cases}, \quad i = 1, \ \ldots, \ n \ .$$

Show that if we assume

$$\begin{aligned} z_i \mid q &\overset{iid}{\sim} Bin(1, q) \\ y_i \mid \mu_1, \mu_2, \sigma_1, \sigma_2, z_i &\overset{indep}{\sim} \begin{cases} N(\mu_1, \sigma_1^2), & \text{if } z_i = 1 \\ N(\mu_2, \sigma_2^2), & \text{if } z_i = 0 \end{cases} \end{aligned}$$

then the distribution of the observed data $\underline{y} = (y_1, \ldots, y_n)$ is as in $(*)$.

Show, therefore, that we may take the complete-data likelihood for the data vectors $\underline{y}$ and $\underline{z}$ to be

$$f(\underline{y}, \underline{z}|\mu_1, \mu_2, \sigma_1, \sigma_2, q) = \prod_{i=1}^{n} \left\{ q \frac{1}{\sqrt{2\pi}\sigma_1} \exp[-(y_i - \mu_1)^2/2\sigma_1^2] \right\}^{z_i} \times \left\{ (1-q) \frac{1}{\sqrt{2\pi}\sigma_2} \exp[-(y_i - \mu_2)^2/2\sigma_2^2] \right\}^{1-z_i}$$

(c) Derive the function E-M objective function $Q(\theta, \theta^k)$ for this problem, where $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, q)$. Are there any simplifications to the E-step? The M-step?

Notes:

- $z_i \mid y_i, \mu_1, \mu_2, \sigma_1, \sigma_2, q$ are still independent Bernoulli's but now with different success probabilities; what are they?
- A modification of E-step that might be useful is to round the $E[z_i|y_i, \mu_1, \mu_2, \sigma_1, \sigma_2, q]$ to 0 or 1 rather than use the expected value itself. How does this simplify the M-step?

(d) Fit the model to the Old Faithful data, using E-M or some variation. Describe your findings, focusing on the following questions:

- What would be a good set of starting values?
- How do you assess convergence?
- Try from several different sets of staring values. Do you get different maxima? Which maximum or maxima are "best"? Why?

(e) Obtain standard errors for the parameters two ways: using `infomat.s`, and using the missing information principle.