Biostatistics 778: Advanced Statistical Computing

Homework 2

Due date: 2007-12-04

Problems

1. *Obtaining the observed information from EM*. Let $f(y \mid \theta)$ be the joint density of the observed data and let $g(y, z \mid \theta)$ be the joint density of the augmented or complete data. Let the observed score function be

$$S(y \mid \theta) = \frac{\partial}{\partial \theta} \log f(y \mid \theta)$$

and the complete data score function be

$$S(y, z \mid \theta) = \frac{\partial}{\partial \theta} \log g(y, z \mid \theta).$$

Finally, let the complete data information matrix be

$$I_{y,z}(\theta) = \frac{\partial}{\partial \theta \partial \theta'} \log g(y, z \mid \theta).$$

Prove the following statements:

$$S(y \mid \theta) = \mathbb{E}_{z|y} \left[S(y, z \mid \theta) \right]$$

and

$$\begin{aligned} I_y(\theta) &\triangleq \frac{\partial}{\partial \theta \partial \theta^T} \log f(y \mid \theta) \\ &= \mathbb{E}_{z|y} \left[I_{y,z}(\theta) \right] - \mathbb{E}_{z|y} \left[S(y, z \mid \theta) S(y, z \mid \theta)' \right] + S(y \mid \theta) S(y \mid \theta)'. \end{aligned}$$

In each case the expectation is taken with respect to the conditional distribution of the missing data z given the observed data y. Assume the necessary conditions so that integral and derivative can be interchanged.

2. *EM algorithm*. Hierarchical models are sometimes used to "combine evidence" in a meta-analysis or multi-site study. For example, when studying air pollution and health, it is common to estimate the association between daily changes in air pollution and some health outcome for many cities separately and then combine the estimates across cities via an hierarchical model. This approach simplifies things because you do not have to combine the data all together in a single analysis, but rather break the analysis into 2 separate parts.

Suppose we have estimates $\hat{\beta}_1, \ldots, \hat{\beta}_n$ and associated variance estimates $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_n^2$, i.e.

$$\text{Var}(\hat{\beta}_i) = \hat{\sigma}_i^2,$$

where i is an index for locations and $\hat{\beta}_i$ is the estimate of the log relative risk of air pollution on hospitalization for chronic obstructive pulmonary disease. A useful summary is the "national average effect" μ which tells us, on average for the entire United States, a unit increase in

air pollution is associated with a $(100 \times \mu)$% increase in hospital admissions. Assume the following hierarchical model:

$$\begin{aligned} \hat{\beta}_i &\sim \mathcal{N}(\beta_i, \hat{\sigma}_i^2) \\ \beta_i &\sim \mathcal{N}(\mu, \tau^2) \end{aligned}$$

where β_i is the "true" log relative risk for city i and $\hat{\beta}_i$ is our estimate of it. The parameter τ^2 is sometimes called the heterogeneity variance and describes the amount of variation in the "true" log relative risks (i.e. the β_is).

Use the EM algorithm to obtain estimates of μ and τ as well as empirical Bayes estimates of $\beta_1, \ldots \beta_n$. First derive the EM iterations and then write a program to compute the estimates from the data. Use Louis's method to obtain standard errors for μ and τ. Data for your program will be provided on the course website. Write up a brief summary describing your analysis.

3. *Rejection/Importance sampling*. Let $Y_i \sim$ Exponential(β) for $i = 1, \ldots, n$ (with mean $1/\beta$) and let β have a half-Normal (σ) prior distribution, i.e. the prior density of β is

$$\pi(\beta \mid \sigma) = \sqrt{\frac{2}{\pi\sigma^2}} \exp(-\beta^2/2\sigma^2)$$

for $\beta > 0$.

Write a function named `postsample` which takes an input vector `y`, a sample size `N`, and a value for the parameter σ and uses rejection sampling to simulate a sample of size `N` from the posterior distribution of $\beta \mid y_1, \ldots, y_n$. Specifically, produce a sample of size 1,000 for $\sigma = 0.5$ and the following ys:

```
20.100306  2.272066  3.796734  2.265275  3.480183
```

Write a function named `postmean` which takes as input

- a posterior sample from the distribution of $\beta \mid y_1, \ldots y_n$,
- a lower bound for σ, and
- an upper bound for σ,

and uses importance sampling/reweighting to plot the posterior mean of β as a function of σ.