

BST 140.778 Assignment 3

February 7, 2005

Fine print All computing assignments must be completed in the R statistical programming language. For non-computing assignments please appropriately typeset using \LaTeX . Bundle your R functions and \LaTeX files for each assignment in a zip or tar.gz file and email them to bcarvalh@jhsph.edu. Include *specific* instructions on how to run the code to answer the assignments in a README file. All code should be readable, formatted well, commented and clearly indicate the author and date. The general rule is: the more thought that has to go into understanding how to implement your code, the worse your grade will be. Please feel free to give each other small hints, but otherwise students must complete assignments individually.

y_i	275	50	110	104	21	8	41	7	30	243	129	38	22
n_i	8544	1032	4851	2064	480	399	513	198	1050	8259	2946	1053	405

y_i =number of births to women under 18, n_i =total number of births

Table 1: Pregnancy rates for women under 18 in 13 North Central Florida counties over the three year period 1989-1991 (Gainesville Sun, April 30, 1994).

1. Consider the data in Table 1. Suppose we assume that $y_i|p_i \sim \text{Bin}(n_i, p_i)$ and $p_i \sim \text{Beta}(\alpha, \beta)$. Write an EM algorithm that estimates $\hat{\alpha}$ and $\hat{\beta}$ treating the p_i as missing data. Note, this problem is notoriously sensitive to the starting values. Also, R's digamma (derivative of the log gamma) function is useful. This problem was assigned last year and it should go without saying that you cannot use or look at any code from last year's students.
2. Suppose that we have complete data model $g(x; \psi)$ and an observed data $y = y(x)$. Let $\mathcal{X}(y) = \{x|y(x) = y\}$. Recall that EM obtains the maximum of

$$l(\psi) = \log \int_{\mathcal{X}(y)} g(x; \psi) dx. \quad (1)$$

Suppose we instead would like to maximize

$$M(\psi) = l(\psi) + \log p(\psi). \quad (2)$$

Such a situation arises when obtaining a posterior mode (and p is the prior) or penalized ML estimates (and p is the penalty term). Given an initial estimate for ψ , let $\psi^{(n+1)}$ be the maximizer of

$$Q(\psi; \psi^{(n)}) + \log p(\psi)$$

where Q is the usual EM Q function for maximizing (1). Show that the sequence $\{\psi^{(n)}\}$ satisfies

$$M(\psi^{(n+1)}) \geq M(\psi^{(n)}).$$

That is, this EM algorithm for calculating posterior modes (or penalized MLEs) satisfies the ascent property.

3. We will use the following lemma when proving that EM converges. Let $\mathcal{L}(\psi)$ be a continuous likelihood for univariate ψ that we would like to maximize via a generalized EM algorithm. Let $\{\psi^{(k)}\}$ be our sequence of GEM iterates. Note this implies $\mathcal{L}(\psi^{(k-1)}) \leq \mathcal{L}(\psi^{(k)})$. Suppose a subsequence, $\{\psi^{(i_j)}\}$ converges to some point, say a . Prove that

$$\lim_{k \rightarrow \infty} \mathcal{L}(\psi^{(k)}) = \mathcal{L}(a).$$

That is, if a subsequence converges to something, then the whole sequence of likelihoods converge.