Statistics 582, Problem Set 2

Wellner; 1/12/2011

Reading: Lehmann and Casella, TPE, Chapter 6, section 6.4,
especially pages 455 - 461 and 504-508.
Chapter 4, sections 5 - 6.

Due: Wednesday, January 19, 2011.

1. Compare the explanation of the EM algorithm in Lehmann and Casella TPE, pages 458-459 with the explanation given in Groeneboom's notes, pages 1 - 3 and 10 - 12. Correct the expressions given in (4.24) of TPE, page 459.

2. Suppose that the "complete data" X is given by three independent multinomial random vectors,

$$N(1) \equiv (N_{ij}(1) : i = 1, \ldots, r; j = 1, \ldots, s) \sim \text{Mult}_{rs}(n_1; p = (p_{ij}, i = 1, \ldots, r, j = 1, \ldots, s),$$

$$N(2) \equiv (N_{ij}(2) : i = 1, \ldots, r; j = 1, \ldots, s) \sim \text{Mult}_{rs}(n_2; p = (p_{ij}, i = 1, \ldots, r, j = 1, \ldots, s),$$

$$N(3) \equiv (N_{ij}(3) : i = 1, \ldots, r; j = 1, \ldots, s) \sim \text{Mult}_{rs}(n_3; p = (p_{ij}, i = 1, \ldots, r, j = 1, \ldots, s)\,.$$

Suppose that the "incomplete data" Y consists of $N(1), (N_{i\cdot}(2) : 1 \le i \le r), (N_{\cdot j}(3) : 1 \le j \le s)$. Thus $N(1)$ gives cell counts for a two-way table, while $(N_{i\cdot}(2) : 1 \le i \le r)$ and $(N_{\cdot j}(3) : 1 \le j \le s)$ give additional information on the marginal distributions of the table. (If n_2 and n_3 are very large relative to n_1, we might regard the marginal distributions as "known".)

A. What are the distributions of $N(1), (N_{i\cdot}(2) : 1 \le i \le r)$ and $(N_{\cdot j}(3) : 1 \le j \le s)$?

B. Find the conditional distribution(s) of X given Y.

C. Suggest an EM - algorithm for estimation of p.

Note: This problem was treated by Chen and Fienberg (1974), *Biometrics* **30**, 629-642.

3. Lehmann and Casella, TPE, Problem 4.16, page 506, modified as follows (It seems to me that ζ_i in the third line of the problem statement should be just ζ.)
We observe independent Bernoulli variables $X_1, \ldots, X_n$ which depend on unobservable variables Z_i distributed independently as $N(\zeta, \sigma^2)$ where

$$X_i = \begin{cases} 0, & \text{if } Z_i \le u_i, \\ 1, & \text{if } Z_i > u_i. \end{cases}$$

Assuming that $u_1, \ldots, u_n$ are known, we are interested in obtaining Maximum Likelihood estimates of ζ and σ^2.

(a) Show that the likelihood function is $\prod_{i=1}^n p_i^{X_i}(1-p_i)^{1-X_i}$ where $p_i = P(Z_i > u_i) = \Phi((\zeta - u_i)/\sigma)$, $i = 1, \ldots, n$. You will need to make further appropriate changes in Lehmann and Casella parts (c)-(e) as well. (I claim that ζ and σ^2 are not identifiable if u is a constant as stated the problem as given.)

4. Human beings can be classified into one of four blood groups (phenotypes) O, A, B, AB. The inheritance of blood groups is controlled by three genes, O, A, B, of which O is recessive to A and B. If r, p, q are the gene probabilities in the population of O, A, B respectively, then the probabilities of the six possible combinations (genotypes) in random mating (where two individuals drawn at random from the population contribute one gene each) are shown in the following table:

Phenotype	Genotype	probability
O	OO	r^2
A	AA	p^2
A	AO	$2rp$
B	BB	q^2
B	BO	$2rq$
AB	AB	$2pq$

We observe among N individuals the phenotype frequencies N_O, N_A, N_B, N_{AB}, and wish to estimate the gene probabilities from such data. A simple approach is to regard the observations as incomplete, the complete data set being the genotype frequencies N_{OO}, N_{AA}, N_{AO}, N_{BB}, N_{BO}, N_{AB}.

A. Derive the EM algorithm for estimation of (p, q, r).

B. Estimate (p, q, r) from $N_O = 176$, $N_A = 182$, $N_B = 60$, $N_{AB} = 17$.

C. Estimate the covariance matrix of the estimator $(\hat{p}, \hat{q}, \hat{r})$.

5. Suppose that $X_1, \ldots, X_n$ are i.i.d. exponential(θ) random variables (with density $p_\theta(x) = \theta \exp(-\theta x) 1_{(0,\infty)}(x)$ for $\theta > 0$). Suppose that $(u_1, v_1), \ldots, (u_n, v_n)$ are real numbers with $0 \leq u_j < v_j < \infty$ for $j = 1, \ldots, n$ and we observe $Y_1, \ldots, Y_n$ where $Y_i \equiv 1_{[u_i, v_i]}(X_i)$ for $i = 1, \ldots, n$.

(a) Derive the EM algorithm for estimation of θ based on observation of the Y_i's, including a careful treatment of the "E" step.

(b) Give an estimator of the information $I_Y(\theta)$.

6. **Optional bonus problem.** Lehmann and Casella, TPE, Problem 4.9, page 504.

7. **Optional bonus problem.** Find a "large scale" (i.e. a problem involving more than 10 parameters) statistical estimation problem in the literature involving which has been solved using an EM (or "EM like") algorithm. Give a precise reference to the paper and comment on whether or not alternative algorithms were proposed and compared with the EM algorithm.