Statistics 582, Problem Set 3

Wellner; 1/21/98

Reading: Lehmann, TPE, Chapter 6, sections 6.1 and 6.2. Chapter 4, sections 3 - 6.
Due: Wednesday, January 28, 1998.

1. Suppose that $X, X_1, \ldots, X_n$ are i.i.d. Weibull(α_0, β_0) (if X has the Weibull(θ) distribution where $\theta = (\alpha, \beta)$, then $1 - F_\theta(x) = P_\theta(X > x) = \exp(-(x/\alpha)^\beta)$ for $x \ge 0$). Recall that the MLE $\widehat{\alpha}$ of α is given by

$$\widehat{\alpha} = \left\{ \frac{1}{n} \sum_{i=1}^n X_i^{\widehat{\beta}} \right\}^{1/\widehat{\beta}}$$

where $\widehat{\beta}$ is the MLE of β. As a simpler alternative to maximum likelihood, I propose to use the alternative estimator $\overline{\beta}$ of β obtained from the slope of an ordinary least squares fit of a Weibull Q-Q plot, and then estimate α by

$$\overline{\alpha} = \left\{ \frac{1}{n} \sum_{i=1}^n X_i^{\overline{\beta}} \right\}^{1/\overline{\beta}} .$$

A. Suppose that $\overline{\beta}_n \to_p \beta_0$ is known. Show that $\overline{\alpha}_n \to_p \alpha_0$. [Hint: use a uniform strong law of large numbers.]
B. Show that $\overline{\alpha}_n$ is a "pseudo-MLE" in the sense that $\overline{\alpha}_n$ maximizes $l_n(\alpha, \overline{\beta}_n)$.

2. 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 $(\hat{p}, \hat{q}, \hat{r})$.

3. 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 $N(1), (N_{i\cdot}(2) : 1 \leq i \leq r), (N_{\cdot j}(3) : 1 \leq j \leq s)$.
A. What are the distributions of $N(1), (N_{i\cdot}(2) : 1 \leq i \leq r)$ and $(N_{\cdot j}(3) : 1 \leq j \leq s)$?
B. Find the conditional distribution(s) of X given Y.
C. Suggest an EM - algorithm for estimation of p.

4. (Optional bonus problem: Profile likelihood). As in problem 1.3, consider the Weibull family of example 3.2.5: $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with $\Theta \subset R+2$ given by the (Lebesgue) densities

$$p_\theta(x) = \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1} \exp\left(-\left(\frac{x}{\alpha}\right)^\beta\right) 1_{[0,\infty)}(x)$$

where $\theta \equiv (\alpha, \beta) \in (0, \infty) \times (0, \infty) \subset R^2$.
A. For a sample of n observations from p_θ, we showed in the course of problem 1.3 that, for each fixed value of β the value of α which maximizes the likelihood as a function of α is

$$\widehat{\alpha}(\beta) = \left\{\frac{1}{n}\sum_{i=1}^n X_i^\beta\right\}^{1/\beta}.$$

Use this to compute the *profile likelihood* $l_{\text{profile}}(\beta) = l_{\text{profile}}(\beta|\underline{X})$ defined by

$$l_{\text{profile}}(\beta) = l(\widehat{\alpha}(\beta), beta) = l(\widehat{\alpha}(\beta), \beta|\underline{X}).$$

B. Use what we know from problem 10.3 to show that the profile likelihood is strictly concave and hence has a unique maximum. Show that maximizing the profile likelihood as a function of β yields the maximum likelihood estimate: i.e.

that $(\widehat{\alpha}, \widehat{\beta}) = (\widehat{\alpha}(\widehat{\beta}_{\text{profile}}), \widehat{\beta}_{\text{profile}})$.

C. What is the relationship of the score function for β from the profile likelihood, $\dot{l}_{\beta,\text{profile}}$ to the (efficient) score for β from the full likelihood? Prove or disprove my claim: the profile score for β (based on n observations) is asymptotically equivalent to the sum of efficient scores for β over the sample in the sense that their difference divided by $\sqrt{n}$ converges to 0 in probability.

D. What is the relationship of the observed information from the profile likelihood $-\ddot{l}_{\beta\beta,\text{profile}}$ to information quantities from the full likelihood?