Statistics 582, Problem Set 3

Wellner; 1/19/2011

Reading: Chapter 4, sections 5-7.

Due: Wednesday, January 26, 2011.

1. 1. Suppose, as in Example 4.3.10, that $\underline{X}_1, \ldots, \underline{X}_n$ are i.i.d. $\text{Mult}_k(1, \underline{p})$ so that $\underline{N}_n = \sum_{i=1}^n \underline{X}_i \sim \text{Mult}_k(n, \underline{p})$.
   (a) Use Jensen's inequality to show that the log-likelihood

$$l_n(\underline{p}|\underline{X}) = \sum_{j=1}^k N_j \log p_j + \sum_{i=1}^n \log\left(\frac{1!}{X_{i1}! \cdots X_{ik}!}\right)$$

   is maximized by $\hat{\underline{p}} = \underline{N}_n/n$. [Hint: write the first term of $l_n(\underline{p}|\underline{X})$ as $n \sum_{j=1}^k \hat{p}_j \log p_j$.]
   (b) Relate $l_n(\underline{p})$ to $K(\hat{\underline{p}}, \underline{p})$ and hence show again that the maximizing value of $\underline{p}$ is $\hat{p}$.

2. Consider nonparametric maximum likelihood estimation of F in the right-censored data problem considered in class, but extend the argument to include ties as follows:
   (a) When there are ties, let the distinct Z's be denoted by $T_1 < \ldots < T_k$.Let $m_1, \ldots, m_k$ and $n_1, \ldots, n_k$ be defined by $m_j \equiv \#$ of $Z_i\delta_i = T_j$, $n_j \equiv \#$ of $Z_i(1-\delta_i) = T_j$, and let $p_j \equiv \Delta F(T_j) = F(T_j) - F(T_j-)$, $j = 1, \ldots, k$, $p_{k+1} = 1 - F(T_k)$. Show that the likelihood (for F) is

$$L(F|\underline{Z}, \underline{\delta}) = \prod_{i=1}^k p_i^{m_i} (\sum_{j=i+1}^{k+1} p_j)^{n_i}.$$

   (b) By defining $\lambda_i = p_i / \sum_{j=i}^{k+1} p_j$ for $i = 1, \ldots, k$ and $\lambda_{k+1} = 1$, and rewriting the likelihood in terms of the λ_i's, show that the likelihood is maximized by

$$\hat{\lambda}_i = m_i / \sum_{j=i}^k (m_j + n_j) = \frac{n\Delta\mathbb{H}_n^{uc}(T_i)}{n(1 - \mathbb{H}_n(T_i-))}.$$

   and hence that the nonparametric MLE of F is (again) the Kaplan - Meier estimator

$$1 - \widehat{F}_n(t) = \prod_{s \le t} (1 - \Delta\widehat{\Lambda}_n(s)).$$

   (c) Compute $1 - \widehat{F}_n$ for the following data (length of time until complete remission in weeks for the "maintained group") from a study of the efficacy of chemotherapy for acute Myelogenous leukemia (AML):

$$9, \quad 13, \quad 13+, \quad 18, \quad 23, \quad 28+, \quad 31, \quad 31, \quad 34, \quad 45+, \quad 48, \quad 161+;$$

   here "+" indicates censoring ($\delta = 0$).

3. We showed in class that the nonparametric maximum likelihood estimator of F in the (right) censored data problem, possibly with ties, is the Kaplan-Meier (product limit) estimator $\widehat{\mathbb{F}}_n(t)$ given by

$$1 - \widehat{\mathbb{F}}_n(t) = \prod_{s \le t} (1 - \Delta \widehat{\Lambda}(s))$$

where $\widehat{\Lambda}_n(t)$ is the *Nelson-Aalen* estimator of

$$\Lambda(t) \equiv \Lambda_F(t) \equiv \int_0^t \frac{1}{1 - F_-} dF \,,$$

given by

$$\widehat{\Lambda}_n(t) = \int_0^t \frac{1}{1 - \mathbb{H}_n(s-)} d\mathbb{H}_n^{uc}(s) \qquad 0 \le s \le t \,.$$

Here

$$\mathbb{H}_n^{uc}(t) = \frac{1}{n} \sum_{i=1}^n \delta_i 1_{[Z_i \le t]} \,, \qquad \mathbb{H}_n(t) = \frac{1}{n} \sum_{i=1}^n 1_{[Z_i \le t]}$$

are the sub-empirical distribution function of the uncensored observations and the marginal empirical distribution of all the Z's uncensored or censored.
(a) Compute $1 - \widehat{\mathbb{F}}_n$ for the following data (time in days until vaginal cancer in rats, group 1; from Kalbfleisch and Prentice, 1980, page 2):

143, 164, 188, 188, 190, 192, 206, 209, 213, 216,
220, 227, 230, 234, 246, 265, 304, 216+, 244+

here + indicates censoring ($\delta = 0$).
(b) In class I gave a heuristic derivation of

$$\sqrt{n}(\widehat{\mathbb{F}}_n(t) - F(t)) \Rightarrow (1 - F(t)) B(C(t))$$

as a process uniformly in $t \in [0, \tau]$ for any $\tau < \tau_H$ (i.e. for any τ with $1 - H(\tau) = (1 - F(\tau))(1 - G(\tau)) > 0$, where B is a standard Brownian motion process and where

$$C(t) \equiv \int_0^t \frac{1}{(1 - H_-(s))^2} dH^{uc}(s) \,, \qquad 0 \le s \le t$$

Thus we have, for each fixed $t < \tau$,

$$\sqrt{n}(\widehat{\mathbb{F}}_n(t) - F(t)) \to_d N(0, (1 - F(t))^2 C(t))$$

Suggest an estimator of $C(t)$ and hence an estimator of $(1 - F(t))^2 C(t)$.
(c) Show that your estimator of $(1 - F(t))^2 C(t)$ is consistent.
(d) Use the estimator you suggest in (b) to obtain an approximate 90% confidence interval for $F(210)$ based for the data given in (a).

4. (Interval censored or current status data). Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables (survival times) with distribution function F as in Example 4.6.5. Suppose that $Y_1, \ldots, Y_n$ are i.i.d. random variables ("observation times") with a

distribution function G which are independent of the X_i's. Unfortunately, we cannot observe the X_i's directly but can only observe $(Y_i, 1_{[X_i \le Y_i]}) \equiv (Y_i, \delta_i)$, $i = 1, \ldots, n$.
(a) Consider the empirical functions

$$\mathbb{G}_n(t) = n^{-1} \sum_{i=1}^{n} 1\{Y_i \le t\} = \mathbb{P}_n 1\{Y \le t\},$$
$$\mathbb{V}_n(t) = n^{-1} \sum_{i=1}^{n} \delta_i 1\{Y_i \le t\} = \mathbb{P}_n \delta 1\{Y \le t\}.$$

Show that for each fixed t we have

$$\mathbb{G}_n(t) \to_{a.s.} G(t), \qquad \text{and} \qquad \mathbb{V}_n(t) \to_{a.s.} \int_0^t F dG \equiv V(t).$$

(b) Plot the cumulative sum diagram $\{(n\mathbb{G}_n(T_{(i)}), n\mathbb{V}_n(T_{(i)})) : i = 1, \ldots, n\}$ and the MLE $\hat{F}_n$ of F as described in example 4.6.5, page 38 of the notes, for the following data: $(3.5, 0)$, $(1.2, 1)$, $(5.7, 1)$, $(6.1, 0)$, $(4.2, 1)$.
(c) What would the MLE of F be (at $t = 4$) if we assumed that F is exponential θ distribution (with $1 - F_\theta(x) = \exp(-\theta x)$ for $x > 0$)? Compare with the value of the MLE $\hat{F}_n(2)$.

5. **Optional bonus problem:** Suppose that F and G are continuous distribution functions.
(a) Show that if $\mathbb{U}$ is a standard Brownian bridge process on $[0, 1]$ and $\mathbb{B}$ is a standard Brownian motion process on $[0, \infty)$, then $(1 + t)\mathbb{U}(t/(1 + t)) \stackrel{d}{=} \mathbb{B}(t)$ as processes on $[0, \infty)$.
(b) Use the result of (a) to show that the limit process for the Kaplan-Meier estimator $(1 - F(t))\mathbb{B}(C(t))$ satisfies

$$(1 - F(t))\mathbb{B}(C(t)) \stackrel{d}{=} \left(\frac{1 - F(t)}{1 - K(t)}\right) \mathbb{U}(K(t))$$

as processes on $[0, \tau]$ for any $\tau < \tau_H$ where $C(t) = \int_0^t (1 - H(s))^{-2} dH^{(uc)}(s)$ and $K(t) \equiv C(t)/(1 + C(t))$.
(c) Show that when there is no censoring (so $G \equiv 0$), $K(t) = F(t)$ for $t < \tau_H$.

6. **Optional bonus problem:** The random right censoring model introduced in example 4.6.2 involves censored, and hence missing data. (We wish we could have observed the X_i's themselves.) Efron (1967), *The two-sample problem with censored data*, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 4, pp. 831-853, gives a discussion of a "self-consistency" property which is related to an EM algorithm in this nonparametric setting. Find Efron's paper; state the self-consistency property of the Kaplan-Meier estimator; relate the self-consistency property to an EM-algorithm in this setting. (Note the discussion of a similar property in the case of case 1 interval censoring in Groeneboom's notes handed out on 1/14/2011.)