Chapter 8: NPMLE, Censoring, and EM

8.1 Estimating an arbitrary F

(i) $X_1, \ldots, X_n$ **i.i.d** $\sim F$.

(ii) Problem: no dominating measure.
(One) solution: assume a dominating measure which is counting measure on the discrete values in $\{x_1, \ldots, x_n\}$.
Problem: dominating measure changes with $X^{(n)}$.

(iii) Then

$$L_{X^{(n)}}(F) = \prod_1^n p_i = \prod_1^k p_j^{n_j}$$

where $p_i = P_F(X = x_i)$, $n_j = \# x_i$ **equal to** x_j, **and (wlog)** $x_1, \ldots, x_k$ **are distinct.**

(iv) $\ell(F) = \Sigma_1^{k-1} n_j \log p_j + n_k \log(1 - \Sigma_1^{k-1} p_j)$, **gives** $\widehat{p_j} = n_j/n$.

(v) That is, $\widehat{F} = F_n$, **the empirical cdf. Despite (ii),** F_n **has "good" properties.**

(vi) Glivenko-Cantelli: $\sup_x |F_n(x) - F(x)| \to_{a.s.} 0$.

(vii) Donsker's Thm: $U_n \sim U(0,1)$ **with edf** G_n. $X_j \equiv F^{-1}(U_j) \sim F$. **Then**
$\sqrt{n}(G_n(u) - u)$ **and** $\sqrt{(}G_n^{-1}(u) - u)$ **each converges to Brownian Bridge process,** B. **And** $\sqrt{n}(F_n - F)$ **converges to the process** $B(F)$.
(i.e. have usual $\sqrt{n}$ **convergence, like parametric MLEs)**

8.2 Right-censoring and the Kaplan-Meier estimator

(i) (X_i, U_i) **i.i.d.** $X_i \sim f$, $U_i \sim G$.
We observe only $Y_i = \min(X_i, U_i)$ **and** $\delta_i = I(X_i \leq U_i)$.

(ii) We want to estimate F: G **normally not of interest. If we could observe all the** X_i, F_n **would be NPMLE of** F.

(iii) For simplicity, assume Y_i **distinct, and (notational convenience)** $y_1 < y_2 < \ldots < y_n < y_+$. **We construct an NPMLE putting mass only on** $\{y_1, \ldots, y_n, y_+\}$.

(iv) $\ell(F) \ = \ \Sigma_1^n(\delta_i \log f(y_i) \ + \ (1-\delta_i)\log(1-F(y_i)))$

(v) Let $k_1 < \ldots < k_{m+1}$ **be indices of uncensored** ($\delta_i = 1$), **obsv., with** $y_{k_{m+1}} = y_+$ **if** $\delta_n = 0$, **and** $y_{k_{m+1}} = y_n$ **if** $\delta_n = 1$.

(vi) Let $p_{k_j} = f(y_i)$, **and** $n_j = \#x_i$ **equal to** y_{k_j}. **Now do EM, with complete-data** $X_1, \ldots, X_n$:

$$\ell_c(F) \ = \ \sum_1^n \log f(x_i) \ = \ \sum_1^{m+1} n_j \log p_{k_j}$$

$$\textbf{Let } e_j \ = \ \mathrm{E}(n_j | Y^{(n)}, \delta^{(n)}) \ = \ \sum_1^n P(X_i = y_{k_j} | \delta_i) \quad \tilde{p_{k_j}} \ = \ e_j/n$$

(vii) Now with $F(t) = \Sigma_{j:y_{k_j} \leq t}\, p_{k_j}$, $F(t) \ = \ \mathrm{E}(F_n(t)|Y^{(n)}, \delta^{(n)})$

(viii) But now we find that a stationary point of EM is $\widehat{p_{k_i}} / \Sigma_{j=i}^{m+1} \widehat{p_{k_j}} \ = \ 1/(n-k_i+1) \quad i = 1, \ldots, m$.

(ix) Then $\widehat{F}(t) \ = \ \Sigma_{i:y_{k_i} \leq t}\, p_{k_i}$ **is NPMLE. This is Kaplan-Meier estimate, although not in usual form.**

(x) Consider

$$\prod_{i:y_{k_i} \leq t} \left(1 - \frac{1}{n-k_i+1}\right) \ = \ \ldots \ = \ \ldots \ = \ 1 - \widehat{F}(t)$$

$n - k_i + 1$ **i "population at risk" just before failure at** y_{k_i}.

8.3 Current status data

(i) As above, failure times X_i, but now we observe only times U_i, and $\delta_i = I(X_i \le U_i)$ (i alive/dead at U_i).

(ii) Again, (X_i, U_i) i.i.d, with X_i indep U_i, $X_i \sim F$, $U_i \sim G$. $L(F) \;=\; \Sigma_1^n(\delta_i \log F(u_i) \;+\; (1-\delta_i)\log(1 - F(u_i))$

(iii) Wlog, $u_1 < u_2 < \ldots < u_n < u_+ \equiv u_{n+1}$. We will put probability mass on (a subset of) $u_1, \ldots, u_n$ and maybe on u_+. Then need to find $p_k = P_F(X = u_k),\ k = 1, \ldots, n$.

(iv) Suppose at EM step m we have estimate $p_i^{(m)},\ i = 1, \ldots, n+1$, giving probs $Q_{ik}^{(m)} = P^{(m)}(X_i = u_k|\delta_i)$.

(v) Now $\ell_c(F) = \Sigma_1^n \log f(X_i)$ so

$$\mathrm{E}_m(\ell_c(F)|U^{(n)}, \delta^{(n)}) \;=\; \sum_{i=1}^{n} \mathrm{E}_m(\log f(X_i) \mid U^{(n)} = u, \delta^{(n)})$$

$$= \;\; \sum_{i=1}^{n}\sum_{k=1}^{n+1} \log p_k P^{(m)}(X_i = u_k|\delta_i) \;=\; \sum_{k=1}^{n+1}\left(\log p_k \left(\sum_{i=1}^{n} Q_{ik}^{(m)}\right)\right)$$

(vi) Maximizing (M-step): $p_k^{(m+1)} \;=\; n^{-1}\Sigma_{i=1}^n Q_{ik}^{(m)}$;
but $Q_{ik}^{(m)} = \delta_i p_k^{(m)}/F^{(m)}(u_i)$ if $u_i \ge u_k$, and
$(1-\delta_i)p_k^{(m)}/(1 - F^{(m)}(u_i))$ **if $u_i < u_k$.**

(viii) Thus $p_k^{(m+1)} \;=\; p_k^{(m)} S_{ik}^{(m)}$ where

$$S_{ik}^{(m)} \;=\; \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\delta_i I(u_i \ge u_k)}{F^{(m)}(u_i)} \;+\; \frac{(1-\delta_i)I(u_i < u_k)}{1 - F^{(m)}(u_i)}\right)$$

(ix) Either $p_k^{(m)} \to \widehat{p_k} \;>\; 0$; then $\widehat{S_{ik}} = 1$,
or $p_k^{(m)} \to 0$, and then $\widehat{S_{ik}} \le 1$.

8.4 The Cusum Diagram

(i) Define points in $\Re^2$**,** $P_0 = (0, 0)$**,** $P_k = (k, \Sigma_1^k \delta_i)$**.**

(ii) $F(u_i) = P($**failed by** $u_i) \approx (1/k)\,\Sigma_1^k \delta_i =$ **slope of** (P_0, P_k)**. BUT** F **must be non-decreasing. So we take largest convex fn** $\leq \{P_k\}$ **and let** $\widehat{F(u_i)}$ **be slope of this function at** i**.**

(iii) E.g. $\delta = (1, 0, 0, 1, 0, 1)$**, then** $\widehat{F(u_1)} = \widehat{F(u_2)} = \widehat{F(u_3)} = 1/3$**,** $\widehat{F(u_4)} = \widehat{F(u_5)} = 1/2$**,** $\widehat{F(u_6)} = 1$**.**

(iv) Note, change in slope at $k \Rightarrow \delta_{k+1} = 1$**, so we have prob mass at failure observation times.**

(v) Suppose slope changes are at $k_1 - 1, k_2 - 1, \ldots$**, with** $1 \leq k_1 < k_2 < \ldots < k_+ \leq n$**, so** $p_{k_i} > 0$**.**

(vi) If $k_j \leq l \leq k_{j+1} - 1$**,** $\widehat{F(u_l)} = \Delta_j/(k_{j+1} - k_j)$**, where** $\Delta_j = \Sigma_{i=k_j}^{k_{j+1}-1} \delta_i$**.**

(vii) If $k_j > m$**,** $\Sigma_{i=k_j}^{k_{j+1}-1} \delta_i/\widehat{F(u_i)} = (k_{j+1} - k_j)$**.**
If $k_{j+1} \leq m$**,** $\Sigma_{i=k_j}^{k_{j+1}-1}(1 - \delta_i)/(1 - \widehat{F(u_i)}) = (k_{j+1} - k_j)$**.**

(viii) If $k_{j+1} = m$**,** $S_m^* = n^{-1}(k_1 + (k_2 - k_1) + \ldots + (n - k_+)) = 1$**.**
If $k_j < m < k_{j+1}$**,**

$$\begin{aligned} S_m^* &= n^{-1}(k_1 + \ldots + (k_j - k_{j-1}) + G_{j,m} + (k_{j+2} - k_{j+1}) \\ &\qquad + \ldots + (n - k_+)) \\ &= n^{-1}(n + G_{j,m} - (k_{j+1} - k_j) \quad \textbf{where} \\ G_{j,m} &= \frac{k_{j+1} - k_j}{k_{j+1} - k_j - \Delta_j}(m - k_j + 1 - D_m) \\ &\quad + \frac{k_{j+1} - k_j}{\Delta_j}(\Delta_j - D_m) \leq k_{j+1} - k_j \end{aligned}$$

where $D_m = \Sigma_{i=k_j}^{m} \delta_i$**. So** $S_m^* \leq 1$**.**

QED!!!