557: MATHEMATICAL STATISTICS II
NONPARAMETRIC MAXIMUM LIKELIHOOD

Suppose that $X_1, \ldots, X_n$ are a random sample from a distribution with cdf F_X that is not specified using a parametric model, that is, the whole function

$$F_X(x) = \Pr[X \leq x] \qquad -\infty < x < \infty$$

is the (infinite dimensional) parameter of the data-generating model. Denote by $\mathcal{F}$ the parameter space, that is, the set of distribution functions (non-decreasing right-continuous functions mapping $\mathbb{R} \to [0,1]$). Finally, denote the *probability measure* associated with F_X by P_{F_X}, so that

$$F_X(x) = P_{F_X}((-\infty, x])$$

Given observed data $\mathbf{X} = \mathbf{x} = (x_1, \ldots, x_n)$ we wish to estimate F_X.

The likelihood function for such data in this nonparametric setting takes the form

$$\mathscr{L}(\mathbf{x}; F_X) = \prod_{i=1}^{n} P_{F_X}(\{x_i\}) \qquad F_X \in \mathcal{F}$$

matching precisely the definition in the parametric setting. It is evident from this definition that $\mathscr{L}(\mathbf{x}; F_X) \geq 0$, and

$$\mathscr{L}(\mathbf{x}; F_X) = 0 \qquad \text{if} \qquad P_{F_X}(\{x_i\}) = 0, \text{ for some } i.$$

so to find the maximum likelihood estimate, we attempt to maximize over functions F_X for which $\mathscr{L}(\mathbf{x}; F_X) > 0$. Let $0 < c \leq 1$, and denote by $\mathcal{F}_c$ the subset of $\mathcal{F}$ whose elements satisfy

$$p_i = P_{F_X}(\{x_i\}) > 0 \qquad i = 1, 2, \ldots, n$$

such that

$$\sum_{i=1}^{n} p_i = c.$$

Note that $0 < c \leq 1$, as P_{F_X} assigns probabilities to sets in (the σ-algebra defined on) $\mathbb{R}$. To maximize $\mathscr{L}(\mathbf{x}; F_X)$ for $F_X \in \mathcal{F}_c$ subject to the constraint, consider the function

$$G(p_1, \ldots, p_n, \lambda) = \prod_{i=1}^{n} p_i + \lambda \left(\sum_{i=1}^{n} p_i - c \right)$$

where λ is a Lagrange multiplier. We have to solve the $n+1$ equations

$$\begin{aligned}
\frac{\partial G}{\partial p_j} &= \frac{\prod_{i=1}^{n} p_i}{p_j} + \lambda = 0 \qquad j = 1, \ldots, n \\
\frac{\partial G}{\partial \lambda} &= \sum_{i=1}^{n} p_i - c = 0
\end{aligned}$$

simultaneously for $p_1, \ldots, p_n, \lambda$. From the first set of equations, rearranging and summing over j we have

$$\frac{n}{\lambda} = -\frac{\sum_{i=1}^{n} p_i}{\prod_{i=1}^{n} p_i}$$

so that from the second equation

$$\frac{n}{\lambda} = -\frac{c}{\prod_{i=1}^{n} p_i} \qquad \therefore \qquad \lambda = -\frac{n \prod_{i=1}^{n} p_i}{c}$$

and hence for the jth equation in the first batch

$$\widehat{p}_j = -\frac{\prod_{i=1}^{n} \widehat{p}_i}{\widehat{\lambda}} = \frac{c}{n} \qquad j = 1, \ldots, n$$

yielding

$$\widehat{\lambda} = -\left(\frac{c}{n}\right)^{n-1}.$$

At this solution,

$$\mathscr{L}(\mathbf{x}; \widehat{F}_X) = \prod_{i=1}^{n} \widehat{p}_i = \left(\frac{c}{n}\right)^n$$

and it is easy to see (by the concavity of the $\log$ function) that for any probabilities $p_1, \ldots, p_n$ summing to c, as

$$\frac{1}{n}\sum_{i=1}^{n} \log p_i \leq \log\left(\frac{1}{n}\sum_{i=1}^{n} p_i\right) = \log\left(\frac{c}{n}\right)$$

it follows that

$$\prod_{i=1}^{n} p_i \leq \left(\frac{c}{n}\right)^n.$$

Thus we have a **global** maximum of $\mathscr{L}(\mathbf{x}; F_X)$ at the computed solution, that is,

$$\max_{F_X \in \mathcal{F}_c} \mathscr{L}(\mathbf{x}; F_X) = \left(\frac{c}{n}\right)^n$$

which is maximized when $c = 1$. Hence the maximum likelihood estimate of F_X, denoted $\widehat{F}_X$, in this nonparametric setting, is defined by the **discrete** probability measure

$$P_{\widehat{F}_X}(\{x\}) = \begin{cases} \widehat{p}_i & x = x_i,\ i = 1, \ldots, n \\ \\ 0 & \text{otherwise} \end{cases} \qquad = \qquad \begin{cases} \dfrac{1}{n} & x = x_i,\ i = 1, \ldots, n \\ \\ 0 & \text{otherwise} \end{cases}$$

which may be equivalently written

$$\Pr[X = x] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i\}}(x)$$

Thus, the nonparametric maximum likelihood estimate of F_X is the **empirical cdf**

$$\widehat{F}_X(x) \equiv \widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[x_i, \infty)}(x).$$

An alternative derivation used in failure time data allows for a similar construction for potentially censored data, that is, data for which for some i, only the event $X_i > x$ is observed. Suppose that data including values with censoring are collected; let $t_1 \leq ... \leq t_n$ denote the n independent failure/censoring times sorted into non-descending order, and let $(z_1, ..., z_n)$ denote the corresponding censoring variables, where $z_j = 1$ if failure is observed, and is zero otherwise. For completeness, define $t_0 = -\infty$, $t_{n+1} = \infty$.

Failure modelling for such data is achieved via functions such as the failure pmf f, survivor function S, hazard function h and cumulative hazard H, where, in discrete time,

$$f(j) = q_j = P[X = j] \qquad S(j) = S_j = P[X > j] \qquad h_j = \frac{q_j}{P[X \geq j]} = \frac{q_j}{S_{j-1}} \qquad H_j = \sum_{i=1}^{j} h_i$$

Nonparametric Likelihood: Define a partition of the observed data range into the disjoint, half-open intervals

$$(-\infty, t_1], (t_1, t_2], ..., (t_{n-1}, t_n], (t_n, \infty)$$

with corresponding interval probabilities $q_1, q_2, ..., q_{n-1}, q_n, q_{n+1}$,

$$q_j = F_X(t_j) - F_X(t_{j-1}) = S_X(t_{j-1}) - S(t_j)$$

and discrete hazards

$$h_1 = q_1 \qquad\qquad h_j = \frac{q_j}{1 - q_1 - q_2 - ... - q_{j-1}}$$

so that $q_1 = h_1$,

$$q_j = h_j \prod_{i=1}^{j-1} (1 - h_i) \qquad\qquad S_j = P[X > t_j] = 1 - \sum_{i=1}^{j} q_i = \prod_{i=1}^{j} (1 - h_i)$$

Suppose now that, for time point t_j, there are N_j observed failures/censorings, defined by binary indicators $\left(z_{j1}, ..., z_{jN_j}\right)$ (this generalizes the $N_j = 1$ case described in the first section, and allows for the possibility of ties). The likelihood for such observed data is

$$\mathscr{L}(\mathbf{t}, \mathbf{z}; \mathbf{q}) = \prod_{j=1}^{n} \left\{ \prod_{k=1}^{N_j} q_j^{z_{jk}} S_j^{(1-z_{jk})} \right\} \tag{1}$$

that will form the basis for inference.

For the data (t, z), the log likelihood from (1) is

$$\log \mathscr{L}(\mathbf{t}, \mathbf{z}; \mathbf{q}) = \sum_{j=1}^{n} \left\{ \sum_{k=1}^{N_j} z_{jk} \log q_j + \sum_{k=1}^{N_j} (1 - z_{jk}) \log S_j \right\}$$

which, in terms of the hazard parameterization yields

$$\begin{aligned} \log \mathscr{L}(\mathbf{t}, \mathbf{z}; \mathbf{h}) &= \sum_{j=1}^{n} \left\{ \sum_{k=1}^{N_j} z_{jk} \left[\log h_j + \sum_{i=1}^{j-1} \log(1 - h_i) \right] + \sum_{k=1}^{N_j} (1 - z_{jk}) \left[\sum_{i=1}^{j} \log(1 - h_i) \right] \right\} \\ &= \sum_{j=1}^{n} \sum_{k=1}^{N_j} z_{jk} \log h_j + \sum_{j=1}^{n} \sum_{k=1}^{N_j} \sum_{i=1}^{j-1} z_{jk} \log(1 - h_i) + \sum_{j=1}^{n} \sum_{k=1}^{N_j} \sum_{i=1}^{j} (1 - z_{jk}) \log(1 - h_i) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{n}\left\{\sum_{k=1}^{N_j} z_{jk}\right\}\log h_j + \sum_{i=1}^{n-1}\left\{\sum_{j=i+1}^{n}\sum_{k=1}^{N_j} z_{jk}\right\}\log(1-h_i) + \sum_{i=1}^{n}\left\{\sum_{j=i}^{n}\sum_{k=1}^{N_j}(1-z_{jk})\right\}\log(1-h_i) \\
&= \sum_{j=1}^{n}\left\{m_{1j}\log h_j + m_{2j}\log(1-h_j)\right\}
\end{aligned}
$$

where

$$
m_{1j} = \sum_{k=1}^{N_j} z_{jk} \qquad m_{2j} = \begin{cases} \left\{\sum_{i=j+1}^{n}\sum_{k=1}^{N_j} z_{ik}\right\} + \left\{\sum_{i=j}^{n}\sum_{k=1}^{N_j}(1-z_{ik})\right\} & 1 \le j \le n-1 \\ \sum_{k=1}^{N_n}(1-z_{nk}) & j = n \end{cases}
$$

In terms of the hazard parameters, the likelihood is the of the form of a *product binomial* expression. The expression for m_{2j} simplifies to be

$$
m_{2j} = \sum_{i=j+1}^{n}\sum_{k=1}^{N_j}\{z_{ik} + (1-z_{ik})\} + \sum_{k=1}^{N_i}(1-z_{jk}) = \sum_{i=j+1}^{n} N_i + N_j - \sum_{k=1}^{N_j} z_{jk} = \sum_{i=j}^{n} N_i - \sum_{k=1}^{N_j} z_{jk}
$$

The maximum likelihood estimates of the hazard probabilities are thus

$$
\widehat{h}_j = \frac{m_{1j}}{m_{1j}+m_{2j}} = \frac{\sum\limits_{k=1}^{N_j} z_{jk}}{\sum\limits_{i=j}^{n} N_i}
$$

and thus

$$
\widehat{q}_1 = \widehat{h}_1 \qquad \widehat{q}_j = \widehat{h}_j \prod_{i=1}^{j-1}\left(1-\widehat{h}_i\right) \qquad \widehat{S}_j = \prod_{i=1}^{j}\left(1-\widehat{h}_i\right) = \prod_{i=1}^{j}\left(1 - \frac{\sum\limits_{k=1}^{N_j} z_{ik}}{\sum\limits_{i=j}^{n} N_i}\right)
$$

If all $N_j = 1$

$$
\widehat{q}_j = \frac{z_j}{n-j+1}\prod_{i=1}^{j-1}\left(1 - \frac{z_i}{n-i+1}\right) \qquad \widehat{S}_j = \prod_{i=1}^{j}\left(1 - \frac{z_i}{n-i+1}\right)
$$

and if all $z_i = 1$ we obtain

$$
\begin{aligned}
q_1 &= \frac{1}{n} \\
q_2 &= \frac{1}{n-1}\left(1 - \frac{1}{n-1+1}\right) = \frac{1}{n} \\
q_3 &= \frac{1}{n-2}\left(1 - \frac{1}{n-1+1}\right)\left(1 - \frac{1}{n-2+1}\right) = \frac{1}{n}
\end{aligned}
$$

and so on, so that $q_i = 1/n$ for all i.