NONPARAMETRIC BAYES ESTIMATOR OF SURVIVAL FUNCTIONS FOR DOUBLY/INTERVAL CENSORED DATA

Mai Zhou

University of Kentucky

The non-parametric Bayes estimator with Dirichlet process prior of a survival function based on right censored data was considered by Susarla and Van Ryzin (1976) and many others. We obtain the non-parametric Bayes estimator of a survival function when data are right, left or interval censored. The resulting Bayes estimator with Dirichlet process prior has an explicit formula. In contrast, there is no explicit formula known for the non-parametric maximum likelihood estimator (NPMLE) with such data. In fact, we show that the NPMLE with doubly/interval censored data cannot, in general, equal to the limit of Bayes estimator for *any* sequence of priors. Several examples are given, showing that the NPMLE and the non-parametric Bayes estimator may or may not be the same, even when the prior is 'non-informative'.

AMS 1991 Subject Classification: Primary 60E15; secondary 60G30.

Key Words and Phrases: Dirichlet process prior, square error loss, non-informative prior, NPMLE.

1. Introduction, Notation and Preliminary

Suppose the original lifetimes are

$$X_1, \cdots, X_n \ . \tag{1.1}$$

They are non-negative, iid with a distribution $F(\cdot)$. However, these lifetimes are subject to censoring. In the case of right censoring, we only observe

$$Z_i = \begin{cases} X_i & \text{if } X_i \le C_i \\ C_i & \text{if } X_i > C_i \end{cases} \qquad \text{and} \qquad \Delta_i = \begin{cases} 1 & \text{if } X_i \le C_i \\ 0 & \text{if } X_i > C_i \end{cases} \tag{1.2}$$

where C_i are the (right) censoring times.

A generalization of right censoring is double censoring (Chang and Yang 1987). In the case of double censoring we only observe

$$Z_i = \begin{cases} X_i & \text{if } Y_i \le X_i \le C_i \\ C_i & \text{if } X_i > C_i \\ Y_i & \text{if } X_i < Y_i \end{cases} \qquad \text{and} \qquad \Delta_i = \begin{cases} 1 & \text{if } Y_i \le X_i \le C_i \\ 0 & \text{if } X_i > C_i \\ 2 & \text{if } X_i < Y_i \end{cases} \ . \tag{1.3}$$

In the above $(C_i, Y_i), i = 1, \cdots, n$ are the left and right censoring times with $C_i > Y_i$. Let the observations be arranged such that $Z_1, \cdots, Z_k$ are the uncensored observations, i.e. $\Delta_1 = 1, \cdots, \Delta_k = 1$. Notice $Z_1, \cdots, Z_k$ equal to $X_1, \cdots, X_k$ here. The rest of the sample $Z_{k+1}, \cdots, Z_n$ are the (either right or left) censored observations.

In the Bayesian estimation of $F(\cdot)$ we need not make assumptions about the distributions of the left and right censoring times C_i and Y_i. Further, for right censored observations, Y_i need not exist. For left censored observations, C_i need not exist. The calculations are conditioned on the observed censoring times. Thus the observations can be described in three parts $Z_1, \cdots, Z_k$ where $X_i = Z_i$; and $Z_{k+1}, \cdots, Z_m$ where $X_i > Z_i$; and $Z_{m+1}, \cdots, Z_n$, where $X_i < Z_i$.

Next we discuss interval censored data. The current status data or case 1 interval censored data consist of an observed "inspection" time T_i, and the information whether X_i is larger than or less than T_i (the status of X_i): (See Huang and Wellner (1996))

$$T_i \ , \qquad \Delta_i = \begin{cases} 0 & \text{if } X_i > T_i \\ 2 & \text{if } X_i < T_i \end{cases} \qquad .$$

Usually the "inspection" times T_i are assumed iid. Similar to the discussion above this iid assumption does not make a difference in the Bayesian analysis. Therefore, in the Bayesian analysis, the current status data is a special case of (1.3) where all the observations are either left or right censored, i.e. $k = 0$.

In the case 2 of interval censoring, we assume $X_1, \cdots, X_k$ are observed exactly (k non-random, and may be zero), and only the observations $X_{k+1}, \cdots, X_n$ are interval censored, and the following information is known:

(1) We observe $n - k$ intervals. With some abuse of notation, we denote the intervals by $[L_j, Z_j)$ for $j = k+1, \cdots, n$. (1.4)

(2) We know that $L_j \leq X_j < Z_j$.

This is called case 2 (or case k) of interval censored data by Huang and Wellner (1996). Again notice we do not need to make assumptions about the distribution of L_j or Z_j. Therefore this fits both case 2 and case k of interval censoring of Huang and Wellner. To make the notation consistent with the doubly censored case, we let $Z_1 = X_1, \cdots, Z_k = X_k$ for directly observable outcomes. The interval censored outcomes are $[L_j, \ Z_j)$ for $j = k+1, \cdots, n$. Notice when the interval $[L_j, Z_j) = [a, \infty)$, this reduces to the case of a right censored data, and when the interval $[L_j, Z_j) = [0, a)$, this reduces to the case of a left censored data. In that sense, interval censoring is more general in that it includes double censoring as a special case.

In Bayesian analysis, the probability $F(\cdot)$ is random. We assume in this paper that $F(\cdot)$ is distributed as a Dirichlet process with parameter α, a measure on the real line. Under the Dirichlet process prior assumption, the probability measure $P(A) = \int_A dF$ has the following property: given any partition of real line: $A_1, \cdots A_u$ the joint distribution of the random vector $(P(A_1), \cdots, P(A_u))$ has a Dirichlet distribution with parameter given by $\alpha(A_1), \cdots, \alpha(A_u)$. For more discussion and properties of Dirichlet process prior, please see Ferguson (1973), Susarla and Van Ryzin (1976) and Ferguson, Phadia and Tiwari (1993). Another possibility is to work with the cumulative hazard functions $H(t)$. A beta process prior on the space of the cumulative hazard function was introduced by Hjort (1990). While using a beta process prior for right censored data works well, it has no advantage over the Dirichlet process prior for doubly/interval censored data. The likelihood of the data do not simplify by using the hazard function with doubly censored data. Therefore, we will use the Dirichlet process prior in this paper.

Using a squared error loss, Susarla and Van Ryzin (1976) obtained the Bayes estimator for $F(\cdot)$ under Dirichlet process prior when data are only subject to right censoring. They also showed that when the weight parameter, α, of the Dirichlet process prior approaches zero, the non-parametric Bayes estimator reduces to the Kaplan-Meier estimator, the NPMLE. Some later papers studied the consistency of the Bayes estimator (Susarla and Van Ryzin 1978) and the posterior distribution (Ghosh and Ramamoorthi 1995). Huffer and Doss (1999) used Monte Carlo methods to compute the nonparametric Bayes estimator.

We obtain the Bayes estimator of $1 - F(\cdot)$ when data are subject to both right and left censoring or are subject to interval censoring. The large sample properties of this Bayes estimator are not discussed here. Though it is not unreasonable to expect that the Bayes estimators are consistent. However, we show that for *any* sequence of priors the nonparametric Bayes estimators under squared error loss *cannot* always converge to the corresponding NPMLE with doubly censored data. This is a bit surprising since in most cases, MLEs are limits of Bayes estimators.

The Bayes estimator we obtained is more complicated than those with only right censored data, especially when there are many left censored or interval censored observations. Nevertheless, it has an explicit formula that can easily be programmed with computer. In contrast, the nonparametric maximum likelihood estimator (NPMLE) in the case of doubly censored data or interval censored data does not have an explicit formula and requires iterative method to compute the estimator. See Turnbull (1974), Chen and Zhou (1999), and Fay (1999). Besides, (1) the Bayes estimator is always uniquely defined while the NPMLE is often only defined up to an equivalent class. This non-uniqueness of the NPMLE makes many important statistics like mean estimator difficult to define. The Bayes estimator do not have this problem. (2) the Bayes estimator is also smoother than the NPMLE.

To minimize the amount of new notation in addition to Susarla and Van Ryzin's (1976) paper, (here after SV) we use their convention (as SV did) that all observations are positive. Obviously we can extend this to the case where observations have support in $(-\infty, \infty)$ without much difficulty.

2. Bayes estimator with right, left/interval censored observations

The Bayes estimator of $1 - F(\cdot)$ under squared error loss of SV is the conditional expectation of $1 - F$ given all the observations. Similar to SV the conditional expectation is computed in two steps: first given all the uncensored observations we find the conditional *distribution* of $1 - F$. Second, given all the censored observations we compute the conditional *expectation*, where the distribution of those lifetimes before censoring are given by the result of the first step.

The following theorem specifies the conditional distribution of $F(\cdot)$ given all the uncensored observations, which accomplishes the first step.

Theorem 1 *The random probability measure P given all uncensored observations is distributed according to the Dirichlet process with parameter $\beta = \alpha + \sum_{uncensored} \delta_{Z_i}$, where δ_a is a unit measure on the point a.*

The proof of this theorem is similar to SV (1976) and Ferguson (1973), and is given in appendix.

Now, the conditional expectation of $1 - F(u) = P[u, \infty)$ is computed given the remaining $n - k - 1$ censored observations: either $Z_{k+1}, \Delta_{k+1}, \cdots, Z_n, \Delta_n$ in the doubly censored case; or $[L_{k+1},\ Z_{k+1}), \cdots, [L_n, Z_n)$ in the interval censored case. Notice the original $X_{k+1}, \cdots, X_n$ is now a random sample from a Dirichlet process with parameter β. Let E_β denote the expectation with respect to this Dirichlet process with parameter β.

2.1 One interval/left censored observation case

To fix idea and enhance readability, we first present in detail the Bayes estimator with many right censored observations but with only one interval censored observation, denoted by $[L_w, Z_w)$. (If $L_w = 0$ then this is left censored.) The general case with many interval/left censored observations will be given later.

By the similar argument as in SV Corollary 1, the conditional expectation, E_β, of $1 - F(u) = P[u, \infty) = P(X \geq u)$ given all the right censored data and one interval censored observation is

$$\hat{S}_D(u) = \frac{E_\beta\{P[u,\infty)P[L_w, Z_w)\prod_{right-censored} P[Z_i,\infty)\}}{E_\beta\{P[L_w, Z_w)\prod_{right-censored} P[Z_i,\infty)\}} \ . \tag{2.1}$$

This is also the desired Bayes estimator of $1 - F(u)$. We shall abbreviate the subscript of $right - censored$ to $r - c$ and $left - censored$ to $l - c$ and $interval - censored$ to $i - c$. Straightforward calculation yields

$$\hat{S}_D(u) = \frac{E_\beta P[u,\infty)\{P[L_w,\infty) - P[Z_w,\infty)\}\prod_{r-c} P[Z_i,\infty)}{E_\beta\{P[L_w,\infty) - P[Z_w,\infty)\}\prod_{r-c} P[Z_i,\infty)}$$

$$= \frac{E_\beta\{P[u,\infty)P[L_w,\infty)\prod_{r-c} P[Z_i,\infty)\} - E_\beta\{P[u,\infty)P[Z_w,\infty)\prod_{r-c} P[Z_i,\infty)\}}{E_\beta\{P[L_w,\infty)\prod_{r-c} P[Z_i,\infty)\} - E_\beta\{P[Z_w,\infty)\prod_{r-c} P[Z_i,\infty)\}}$$

$$= \frac{E_\beta \textcircled{1} - E_\beta \textcircled{2}}{E_\beta \textcircled{3} - E_\beta \textcircled{4}}. \tag{2.3}$$

The four expectations in (2.3) are all of a same type and can be computed explicitly by the Lemma below.

Given a set of positive numbers $0 < a_{k+1} < a_{k+2} < \cdots < a_m < \infty$, it partitions the nonnegative half line R^+ into intervals $[0, a_{k+1}), [a_{k+1}, a_{k+2}), \cdots,$ $[a_m, \infty)$. By Theorem 1 the random vector $P[0, a_{k+1}), P[a_{k+1}, a_{k+2}), \cdots P[a_m, \infty)$ has a Dirichlet distribution with parameter vector $(\beta_{k+1}, \cdots, \beta_{m+1})$ where $\beta_{k+1} = \beta[0, a_{k+1}), \cdots, \beta_{m+1} = \beta[a_m, \infty)$. The measure β is given as before by $\beta = \alpha + \sum_{uncensored} \delta_{Z_i}$.

Lemma 1 (Susarla and Van Ryzin) *With the notation above, we have*

$$E_\beta \prod_{i=k+1}^{m} P[a_i,\infty) = \prod_{i=0}^{m-k-1}\left(\frac{i + \sum_{j=0}^{i}\beta_{m+1-j}}{i+\beta(R^+)}\right) = \prod_{i=0}^{m-k-1}\left(\frac{i+\beta[a_{m-i},\infty)}{i+\beta(R^+)}\right) \ .$$

Proof: This is essentially SV (1976) Lemma 2 (a) with some extra simplifications. ♢

When $\alpha(R^+) = 0$ the expression on the right hand side of Lemma 1 is still well defined unless there are no uncensored observations in the sample. In the example 2 of section three later, there are no uncensored observation in the sample, we therefore do not discuss the limit of the Bayes estimator as $\alpha(R^+) \to 0$ there.

Remark: It is clear from the definition of β that when $\alpha(R^+) \to 0$, β is integer valued. This implies that the expectation in Lemma 1 has a rational number value (finite product of rational numbers) as $\alpha(R^+) \to 0$.

2.2 Many interval/left censored observations

When data contains many interval and many right censored observations, the Bayes estimator of $1 - F(u) = P(X \geq u)$ given all the data (censored or uncensored) is equal to

$$\hat{S}_D(u) = \frac{E_\beta\{P[u,\infty)\prod_{i-c} P[L_w, Z_w)\prod_{r-c} P[Z_i,\infty)\}}{E_\beta\{\prod_{i-c} P[L_w, Z_w)\prod_{r-c} P[Z_i,\infty)\}}. \tag{2.5}$$

When data contains many left and many right censored observations, the Bayes estimator of $1 - F(u)$ is

$$\hat{S}_D(u) = \frac{E_\beta\{P[u,\infty)\prod_{l-c}[1 - P[Z_w,\infty)]\prod_{r-c} P[Z_i,\infty)\}}{E_\beta\{\prod_{l-c}[1 - P[Z_w,\infty)]\prod_{r-c} P[Z_i,\infty)\}}\ . \tag{2.6}$$

Because left censored observation is a special case of interval censored observation as pointed out in the previous section, we only present in detail below the Bayes estimator with many interval/right censored observations.

Let us recall the identity

$$\prod_{i=1}^{m}(b_i - a_i) = \sum y_1 y_2 \cdots y_m \tag{2.7}$$

where $y_i =$ either b_i or $-a_i$ and the summation is over all possible 2^m choices. The integer m is defined as $m = \#\{i - c\} =$ number of interval censored observations.

By using (2.7), we can rewrite

$$\prod_{i-c} P[L_w, Z_w) = \prod_{i-c}\{P[L_w,\infty) - P[Z_w,\infty)\} = \sum P_1 P_2 \cdots P_m$$

where each $P_w =$ either $P[L_w,\infty)$ or $-P[Z_w,\infty)$ and the summation is over all 2^m different choices.

To make the expression more specific we introduce the follow notation. Define vectors $\xi = (\xi_1, \cdots, \xi_m)$ where each $\xi_i =$ either 0 or 1. There are 2^m such vectors.

Given m interval censored observations, $[L_i, Z_i)$, we define 2^m sets of numbers (each of size m) $\{c_i(\xi), i = 1, 2, \cdots, m\}$ where $c_i(\xi) = L_i$ if $\xi_i = 0$ otherwise $c_i(\xi) = Z_i$.

Associate with each set of $\{c_i(\xi), i = 1, \cdots m\}$ we also define a sign: if the set contains even number of Z_i's then the sign is positive, if the set contains odd number of Z_i's the sign is negative, i.e. $\pm = (-1)^{\sum \xi_i}$.

With these definition we can further rewrite

$$\prod_{i-c} P[L_w, Z_w) = \sum P_1 P_2 \cdots P_m = \sum_{\xi} \pm \prod_{i=1}^{m} P[c_i(\xi),\infty)$$

where the summation is over all 2^m different ξ's, and $\pm$ is the associated sign we defined above.

Finally, we define new sets of numbers by adding r ($r = \#\{r-c\}$) right censored observations $Z_1, \cdots, Z_r$ to $\{c_i(\xi), i = 1, \cdots, m\}$:

$$\{b_j(\xi), j = 1, \cdots, m+r\} = \{c_i(\xi), i = 1, \cdots, m\} \bigcup \{Z_1, \cdots, Z_r\}.$$

For any sets of real numbers $b_1, b_2, \cdots, b_k$, we denote by $b_{(-1)}, b_{(-2)} \cdots, b_{(-k)}$ the reversely ordered numbers (descending). So, $b^+_{(-i)}(\xi), i = 1, \cdots, m+r$ is a set of $m+r$ numbers ordered from largest to smallest.

With these sets of numbers defined, the denominator of (2.5) can be written as

$$\sum E_\beta \left(P_1 P_2 \cdots P_m \times \prod_{r-c} P[Z_i, \infty) \right) = \sum_\xi \left(\pm \prod_{i=1}^{m+r} \frac{i - 1 + \beta[b^+_{(-i)}(\xi), \infty)}{i - 1 + \beta(R^+)} \right)$$

where the summation is over 2^m different ξ's. We can similarly compute the numerator of (2.5) except there is one more term, $P[u, \infty)$, included with the right censored observations. Define

$$\{b^+_j(\xi)\} = \{c_i(\xi), i = 1, \cdots, m\} \bigcup \{Z_1, \cdots, Z_r, u\}.$$

Theorem 2 *The nonparametric Bayes estimator of the survival function $S(u) = 1 - F(u)$ with right censored and interval censored data under Dirichlet process prior is*

$$\hat{S}_D(u) = \frac{\sum E_\beta \left\{ P_1 P_2 \cdots P_m \times \prod_{r-c} P[Z_i, \infty) \times P[u, \infty) \right\}}{\sum E_\beta \left\{ P_1 P_2 \cdots P_m \times \prod_{r-c} P[Z_i, \infty) \right\}} ,$$

$$= \frac{\displaystyle\sum_\xi (-1)^{\sum \xi_s} \prod_{i=1}^{m+r+1} \frac{i - 1 + \beta[b^+_{(-i)}(\xi), \infty)}{i - 1 + \beta(R^+)}}{\displaystyle\sum_\xi (-1)^{\sum \xi_s} \prod_{i=1}^{m+r} \frac{i - 1 + \beta[b_{(-i)}(\xi), \infty)}{i - 1 + \beta(R^+)}} . \qquad (2.8)$$

The two sums in (2.8) are over all 2^m possible ξ's.

Admittedly the two summations above involves 2^m terms when there are m interval censored observations. Also, in the summations, there are both positive and negative terms that will cancel and diminishing significant digits. Rounding errors will be magnified if we use (2.8) directly. Our purpose here is to show that explicit formula exist for the Bayes estimators. Simplifications/alternative formula are desirable and will be pursued in the future.

Remark: From Lemma 1 and Theorem 2, we can infer that the limit of the Bayes estimator (2.8) when the α measure approaches zero is a step function, at least for $u <$

maximum observed value. This is because all the E_β involved will be step functions according to Lemma 1. We can also infer that when the α measure approaches zero, the Bayes estimator (2.8) is always rational number valued, since the E_β involved are all rational number valued.

3. Examples

The examples presented here are either hand-calculated or obtained by using software (Example 2 and the NPMLE in Example 1) we developed. We shall pay close attention to the limit of the Bayes estimator when $\alpha \to 0$ in the Dirichlet prior (non-informative prior), and compare the estimator with NPMLE. The software used here are packaged as **R** (http://www.r-project.org/) packages and can be found at

`http://www.ms.uky.edy/~mai/research/`

The software for computing NPMLE is also available at the above R site.

To minimize additional notation, we shall recycle the notation used by SV as much as possible. Assume $Z_{(k+1)}, \cdots, Z_{(m)}$ are the ordered, distinct censored (both right and left/interval) times among the sample (1.2). Assume there are no ties among the left/interval and right censored observations (but ties within right censored observations are allowed). At each censored observation $Z_{(i)}, k+1 \le i \le m$, let λ_i be the number of right censored observations that equal to $Z_{(i)}$. Thus if there are 2 right censored observations equal to $Z_{(i)}$ then $\lambda_i = 2$. If $Z_{(j)}$ is a left censored observation then $\lambda_j = 0$. To make the notation consistent, we define $Z_{(k)} = 0$ and $Z_{(m+1)} = \infty$.

Let $N(u)$ be the number of uncensored and right censored observations that are larger then or equal to u, i.e.

$$N(u) = \sum_{j:\ \Delta_j = 1} I_{[Z_j \ge u]} + \sum_{i=k+1}^{m} \lambda_i I_{[Z_{(i)} \ge u]} \tag{3.1}$$

and let $N^+(u) = N(u+)$.

We reproduce here SV's Bayes estimator (based only on the uncensored and right censored observations of the sample (1.2)) in a slightly modified form:

For $Z_{(l)} \le u < Z_{(l+1)}$ with $k \le l \le m+1$,

$$\hat{S}(u) = \frac{\alpha(u,\infty) + N^+(u)}{\alpha(R^+) + N^+(0)} \times \prod_{j=k+1}^{l} \left\{ \frac{\alpha[Z_{(j)},\infty) + N(Z_{(j)})}{\alpha[Z_{(j)},\infty) + N(Z_{(j)}) - \lambda_j} \right\} . \tag{3.2}$$

We changed two things: (1) we added the nodes $Z_{(j)}$ for left/interval censored observations, though with zero λ_j's. (2) replaced n by $N^+(0)$.

Example 1 Let us look at an example with one left censored observation and 4 right censored observations. These are the data used by SV (1976) but with an added left censored observation at $Z = 4$.

The ordered observations with their censoring indicators are listed below.

$Z_i's$:	0.8	1.0	2.7	3.1	4	5.4	7.0	9.2	12.1
Δ :	1	0	0	1	2	1	0	1	0

Table 1. Data with one left and four right censored observations.

Let the Bayes estimator of SV based only on uncensored and right censored observations be $\hat{S}(u)$, i.e. as defined in (3.2). Our estimator that takes into account one left censored observation can be written as follows.

(1) For $u > Z_{left} = 4$. After tedious but straight forward simplification we get

$$\hat{S}_D(u) = \hat{S}(u) \times \frac{\frac{\alpha[0,1)+1}{\alpha(R^+)+9} + \frac{\alpha[1,2.7)}{\alpha[1,\infty)+7} \times \frac{\alpha[1,\infty)+8}{\alpha(R^+)+9} + \frac{\alpha[2.7,4)+1}{\alpha[2.7,\infty)+6} \times \frac{\alpha[1,\infty)+8}{\alpha(R^+)+9} \times \frac{\alpha[2.7,\infty)+7}{\alpha[1,\infty)+7}}{\frac{\alpha[0,1)+1}{\alpha(R^+)+8} + \frac{\alpha[1,2.7)}{\alpha[1,\infty)+6} \times \frac{\alpha[1,\infty)+7}{\alpha(R^+)+8} + \frac{\alpha[2.7,4)+1}{\alpha[2.7,\infty)+5} \times \frac{\alpha[1,\infty)+7}{\alpha(R^+)+8} \times \frac{\alpha[2.7,\infty)+6}{\alpha[1,\infty)+6}} .$$

For u in other time intervals, the estimator can be similarly expressed as the product of $\hat{S}(u)$ and some other term, the details are omitted. The plot the Bayes estimator is given in Figure 1.

Next we compute the limit of the Bayes estimator. When $\alpha \to 0$, the Susarla and Van Ryzin estimator, $\hat{S}(u)$, has the limit of the Kaplan-Meier estimator S_{KM}. Our estimator has the limit as $\alpha \to 0$:

For $9.2 \le u < 12.1$, the limit equals to $S_{KM} \times \frac{70}{81} = \frac{7}{8} \times \frac{4}{5} \times \frac{3}{4} \times \frac{1}{2} \times \frac{70}{81} = 0.2268519$. For u in other intervals the limit can be computed similarly.

We plot the estimator with $\alpha(u, \infty) = B \exp(-\theta u)$. The plot shows estimators for $B = 8, \theta = 0.12$ and $B = 0.001, \theta = 0.12$. The latter is indistinguishable in appearance with the limit just calculated.

Computation of the NPMLE for doubly censored data can be done by EM type iteration (see Turnbull (1974), Chen and Zhou (1999)). For the data in the table 1 we obtain the following NPMLE by software:

t	0-0.8	0.8-3.1	3.1-5.4	5.4-9.2	9.2-12.1
NPMLE	1	.8457284	.6028477	.4521358	.2260679
Limit Bayes	1	.8425926	.6049383	.4537037	.2268519

Table 2. NPMLE and limit of Bayes estimator for data in table 1

The differences between the NPMLE and the limit of the nonparametric Bayes estimator are small but real. The likelihood of the distribution in table 2 is larger than those of the limit of the Bayes estimator. (3.70674×10^{-5} v.s. 3.704924×10^{-5}.)

Example 2: We took the first 10 observations from the breast cosmesis data with radiation of Finkelstein and Wolfe as reported by Fay (1999). Out of the 10, there are 4 right censored observations, 4 interval censored observations and 2 left censored observations (i.e. interval censored with left-ends = 0).

Data: $[45, \infty), [6, 10), [0, 7), [46, \infty), [46, \infty), [7, 16), [17, \infty), [7, 14), [37, 44), [0, 8)$.

We compute the nonparametric Bayes estimator with $\alpha(u, \infty) = B \exp(-\theta u)$. The resulting estimator with $B = 8$ and $\theta = 0.3$ is computed using the software we developed and plotted in Figure 2.

In the following two examples, the Bayes estimators are obtained with formula (2.8) and then we let $\alpha(R^+) \to 0$ to obtain the limit. The NPMLE's are also calculated not by software but analytically.

Example 3 Here we took a small example with one left and one right censored observation. The NPMLE and limit of non-parametric Bayes estimator turns out to be exactly the same.

$Z_i's:$	$Z_{(1)}$	$Z_{(2)}$	$Z_{(3)}$	$Z_{(4)}$	$Z_{(5)}$
$\Delta:$	1	0	2	1	1
Jump of $1 - \hat{F}(u)$	0.4	0	0	0.3	0.3

Table 4. Data with one left and one right censored observation.

Example 4 The order of two censoring indicators in the above table are switched and the NPMLE is different from the limit of Bayes estimator. The limit of Bayes estimator is not self-consistent either. To calculate the NPMLE, we first note for this data the NPMLE $F(\cdot)$ have only three jumps at $Z_{(1)}, Z_{(3)}$ and $Z_{(5)}$. Denote the jumps by p_1, p_2, p_3. By symmetry we must have $p_1 = p_3$. Using the constraint $\sum p_i = 1$, we can reduce the likelihood, $L = p_1(p_2 + p_3)p_2(p_1 + p_2)p_3$, to a function of only p_2. Straightforward calculation show $p_2 = \sqrt{5}/5$ maximizes the likelihood. Therefore $p_1 = p_3 = (5 - \sqrt{5})/10$ which is the entry 0.2763932 in the table.

$Z_i's:$	$Z_{(1)}$	$Z_{(2)}$	$Z_{(3)}$	$Z_{(4)}$	$Z_{(5)}$
$\Delta:$	1	0	1	2	1
Jump of limit Bayes	0.28000	0	0.44000	0	0.28000
Jump of NPMLE	0.2763932	0	0.4472136	0	0.2763932

Remark: This example showed that with positive probability, the NPMLE $1 - \hat{F}(\cdot)$ for doubly/interval censored data can take irrational number values. A closer look at the Example 4 provides some insight as why the two estimators are different, as described in next section.

Remark: Example 3 and 4 reveal two different situations. The difference is that left and right censored data overlap (right censored observation is smaller then the left censored observation) in example 4. The overlap in Example 3 is not real, since there is no probability mass inside the overlap.

4. Limit of Bayes and NPMLE

In this section we formally summarize some results concern the limit of Bayes and the NPMLE in doubly/interval censored data case. The argument below is valid for *any* prior, not just the Dirichlet process prior.

Theorem 3. *Suppose a sequence of prior, $\pi_v; v = 1, 2, \cdots$, is such that the (nonparametric) Bayes estimators $1 - \hat{F}_v(\cdot)$ under squared error loss, converge to the Kaplan-Meier estimator whenever the data has only right censoring. Then this same sequence of Bayes estimators cannot converge, in general, to the NPMLE for interval/doubly censored data.*

Proof: The Bayes estimator under squared error loss can be written as

$$1 - \hat{F}_v(u) = \frac{E_\pi P[u, \infty) L_F(data)}{E_\pi L_F(data)} \ ,$$

where $L_F(data)$ is the likelihood of the data when its distribution is F.

The assumption of the Theorem for right censored data says that as $v \to \infty$ we always have

$$\frac{E_\pi \{ P[u, \infty) \prod_{r-c} P[x_i, \infty) \prod_{uncensor} P(\{x_j\}) \}}{E_\pi \prod_{r-c} P[x_i, \infty) \prod_{uncensor} P(\{x_j\}) \}} \to 1 - F_{K-M}(u) \ . \tag{4.1}$$

Notice the Kaplan-Meier estimator, $F_{K-M}(u)$, is always rational number valued.

Now we look at a particular sample configuration with just one left censored observation, for example, the data as in Example 4. The Bayes estimator for this data can be written as

$$\frac{E_\pi P[u, \infty) P(\{Z_1\}) P[Z_2, \infty) P(\{Z_3\}) P[0, Z_4) P(\{Z_5\})}{E_\pi P(\{Z_1\}) P[Z_2, \infty) P(\{Z_3\}) P[0, Z_4) P(\{Z_5\})} \ .$$

Let us use the notation $P(Z) = P(\{Z\})$, and $P(Z^+) = P[Z, \infty)$. Write $P[0, Z_4) = 1 -$

$P[Z_4,\infty) = 1 - P(Z_4^+)$ and expand to get

$$= \frac{E_\pi P(Z_1)P(Z_3)P(Z_5)P(Z_2^+)P(u^+) - E_\pi P(Z_1)P(Z_3)P(Z_5)P(Z_2^+)P(Z_4^+)P(u^+)}{E_\pi P(Z_1)P(Z_3)P(Z_5)P(Z_2^+) - E_\pi P(Z_1)P(Z_3)P(Z_5)P(Z_2^+)P(Z_4^+)} \tag{4.2}$$

If we divide the numerator of (4.2) by $E_\pi P(Z_1)P(Z_3)P(Z_5)P(Z_2^+)P(u^+)$ then as $v \to \infty$, the numerator will converge to the limit: $1 - [1 - F^*_{K-M}(Z_4)]$ according to (4.1). Here the Kaplan-Meier estimator is based on three uncensored observations: Z_1, Z_3, Z_5 and two right censored observations: Z_2, u.

Similarly, if we divide the denominator of (4.2) by $EP(Z_1)P(Z_3)P(Z_5)P(Z_2^+)$ then it has the limit $1-[1-F^{**}_{K-M}(Z_4)]$, where the Kaplan-Meier estimator is based on three uncensored observations, Z_1, Z_3, Z_5 and one right censored observation Z_2.

In other words, multiply (4.2) by

$$\frac{E_\pi P(Z_1)P(Z_3)P(Z_5)P(Z_2^+)}{E_\pi P(Z_1)P(Z_3)P(Z_5)P(Z_2^+)P(u^+)} \tag{4.3}$$

then it will have a rational limit. This factor, (4.3), itself have a rational limit as $v \to \infty$ ($= [1 - F^{**}_{K-M}(u)]^{-1}$). This imply the limit of (4.2), as $v \to \infty$, is

$$\frac{F^*_{K-M}(Z_4)}{F^{**}_{K-M}(Z_4)} \times [1 - F^{**}_{K-M}(u)] \ .$$

But that cannot be the NPMLE as Example 4 show the NPMLE is irrational valued.

That also serve as the proof for the interval censored case since the left censored observation is just $[0, Z_{(4)})$ interval censored. ♢

Corollary 1. *There is no sequence of priors, π_v, such that the resulting sequence of Bayes estimators under squared error loss, $1 - F_v(\cdot)$, always converge to the NPMLE in the interval/doubly censored data case.*

Proof: Suppose, to the contrary, there is such a sequence of prior. Since right censoring is a special case of double/interval censoring (zero left censoring or all intervals are of the form $[a_i, \infty)$), this sequence of estimators must converge to the Kaplan-Meier estimator with such data. But by Theorem 3 such sequence cannot converge to the NPMLE for doubly/interval censored data case in general. ♢

5. Discussion

The formula (2.8) has 2^m (exponential order) terms when there are m interval censored observations. While we do not have a formal proof that the computation of the Bayes estimator cannot be reduced to polynomial order, it is not hard to see that the computation is equivalent to

$$\int \cdots \int \prod_j (\sum_{r=1}^{j} x_r) \prod_j (1 - \sum_{r=1}^{j} x_r)(1 - \sum x_j)^{\beta_m} \prod x_j^{\beta_j} \prod dx_j \ ;$$

on the region $x_j > 0$ and $\sum x_j \leq 1$. We believe it cannot be reduced to the polynomial order.

Remark: The irrational value of NPMLE with doubly censored data also imply that the EM algorithm, if started from the Kaplan-Meier estimator, cannot converge in finite steps in general.

I thank C. Srinivasan for many helpful discussions.

References

Chang, M. and Yang. G.L. (1987). Strong consistency of a nonparametric estimator of the survival function with doubly censored data. *Ann. Statist.* **15**, 1536-1547.

Chen, K. and Zhou, M. (1999). Testing hypothesis and confidence intervals with doubly censored data. Dept. of Statistics, Univ. of Kentucky Tech Report # 376. Accepted for publication.

Fay, M. (1999). Splus functions for nonparametric estimate and test for interval censored data. *http://lib.stat.cmu.edu/S/interval.tar.gz*

Ferguson, T. (1973). A Bayesian analysis of some nonparametric Problems. *Ann. Statist.* 209-230.

Ferguson, T., Phadia, E.G. and Tiwari, R.C. (1993). Bayesian nonparametric inference. *Current Issues in Statistical Inference: Essays in honor of D. Basu (edited by M. Ghosh and P.K. Pathak).* IMS Lecture Notes Monograph series **34**, 127-150.

Ghosh, J.K. and Ramamoorthi, R.V. (1995). Consistency of Bayesian inference for survival analysis with or without censoring. *Analysis of Censored Data (edited by Koul, H. and Deshpande, J.V.)* IMS Lecture Notes Monograph Series **27**, 95-103.

Hjort, N. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.* **18**, 1259-1294.

Huang, J. and Wellner, J. (1996). Interval censored survival data: a review of recent progress. *Proceedings of the First Seattle Symposium in Biostatistics.* Lin, D.Y. and Fleming, T. Editors. 123-170.

Huffer, F. and Doss, H. (1999). Software for Bayesian analysis of censored data using mixtures of Dirichlet priors. Preprint.

Kaplan, E. and Meier, P. (1958), Non-parametric estimator from incomplete observations. *J. Amer. Statist. Assoc.* 53, 457–481.

Susarla, V. and Van Ryzin, J. (1976), Nonparametric Bayesian estimation of survival curves from incomplete observations. *J. Amer. Statist. Assoc.* **71**, 897-902.

Susarla, V. and Van Ryzin, J. (1978), Large sample theory for a Bayesian nonparametric survival curves estimator based on censored samples. *Ann. Statist.* **6**, 755-768.

Turnbull, B. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *JASA*, 169-173.

Department of Statistics
University of Kentucky
Lexington, KY 40506-0027

Appendix: We only sketch the proof for the doubly censored case. The probability of $(\Delta = 1, X = u)$ is (see Chang 1990, (1))

$$(S_C(u) - S_Y(u))dP(X \leq u)$$

Recall the marginal distribution of X is $\frac{\alpha(u)}{\alpha(R^+)}$.

$$\int_{[\Delta=1,Z\in A]} (A.1) = \int_{[u\in A]} D(\cdot|\alpha(B_1) + d_u(B_1), \cdots, \alpha(B_l) + d_u(B_l))(S_C(u) - S_Y(u))d\frac{\alpha(u)}{\alpha(R^+)}$$

$$= \sum_{j=1}^{l} D(y_1, \cdots, y_l|\alpha_1^{(j)}, \cdots, \alpha_l^{(j)}) \int_{[u\in A\cap B_j]} (S_C(u) - S_Y(u))d\frac{\alpha(u)}{\alpha(R^+)}$$

On the other hand,

$$\mathcal{P}\{P(B_i) \leq y_i, i = 1, \cdots, l; \Delta = 1, Z \in A)\}$$

$$= \int_{u=0}^{\infty} \mathcal{P}\{P(B_i) \leq y_i, i = 1, \cdots, l; X \in [u, u + du) \cap A\}(S_C(u) - S_Y(u))$$

$$\sum_{j=1}^{l} \int \frac{\alpha(B_j \cap A \cap [u, u + du)}{\alpha(R^+)} (S_C(u) - S_Y(u))D(y_1, \cdots, y_l|\alpha_1^{(j)}, \cdots, \alpha_l^{(j)})$$

$$= \sum_{j=1}^{l} D(y_1, \cdots, y_l|\alpha_1^{(j)}, \cdots, \alpha_l^{(j)}) \int_{B_j\cap A} \frac{(S_C(u) - S_Y(u))}{\alpha(R^+)} d\alpha(u)$$

Figure 1: Plot for Example 1.

Figure 2: Plot for Example 2.