14 On Censored and Truncated Data in Survival Analysis and Reliability Models

Catherine Huber, Valentin Solev, and Filia Vonta

CONTENTS

14.1 INTRODUCTION

A common feature of many failure time data in epidemiological studies or reliability studies is that they are simultaneously censored and truncated. There are various types of censored data with the most common type the right-censored data case which occurs when the failure time is not observed completely in the sense that it is only known to be larger than a censoring time which often is the end of the study. Censoring can also occur from the left, that is, when we only know that the time of interest happened before a censoring time. We may also have interval-censored data. This case usually occurs from grouped data or from the fact that patients or technical systems are examined at certain dates and the event of interest is only known to have occurred between two specific checking times.

Additionally, survival or reliability data could be truncated. There are also various types of truncation. For example right-truncated data occur in registers. An acquired

immune deficiency syndrome (AIDS) register only contains AIDS cases which have already been reported, which generates right-truncated samples of induction times. We also may have left truncated data or interval truncated data. In the latter case, observation of a process is for example not continuous in time and is done through a window of time (a time interval) which could exclude totally some subjects from the sample. This occurs in particular when the event of interest results in an irreversible change of state of the subject, that is, at time t_1, the subject is in state one, while at time t_2, it is in state two.

Both censoring and truncation create difficulties in the estimation of the parameters involved in the hypothesized models to describe these types of data. One could consider parametric, semi-parametric, or non-parametric models in order to describe a given dataset in the most efficient way. In this chapter we will concentrate on the non-parametric case only.

Turnbull (1976) and Frydman (1994) dealt with the non-parametric estimation of the distribution function F when the data are interval-censored and truncated. For the same case, non-parametric maximum likelihood estimators of the cumulative hazard function together with finite dimensional parameters associated to covariates are obtained in Alioum and Commenges (1996) for the Cox model and in Huber and Vonta (2004) for a generalization of the Cox model, that is, frailty or transformation models. Huang (1996) and Huang and Wellner (1995) examined the theoretical aspects related to the non-parametric maximum likelihood estimator (NPMLE) of the regression coefficient and the baseline distribution, in the case of the Cox model as well as in a class of semi-parametric models, with interval-censoring. Huber, Solev and Vonta (2009) give conditions on the involved distributions, the censoring, truncation, and failure distributions, all three of them assumed mutually independent, under which the consistency of the non-parametric maximum likelihood estimator of the density of the survival or reliability function, is established along with the rate of convergence.

In this work we review existing results for the case of censored and truncated data. In Section 14.2 we define a marginal non-parametric likelihood for interval-censored and interval-truncated data first introduced in Turnbull (1976) and later studied by Alioum and Commenges (1996) and Huber and Vonta (2004) in connection also to parametric and semi-parametric models. In Section 14.3 we formulate a complete non-parametric likelihood for interval-censored and interval-truncated data. Without restriction of the generality, we consider the case of right truncation. In order to derive this likelihood we define a common law of censoring and truncation. The censoring mechanism is represented as a denumerable partition of the total interval of observation time (a, b). A truncation is added to the censoring, conditioning the observations both of the survival and the censoring processes. As the section progresses, three distributions are successively studied, each one being conditional on fixed values which become random in the next subsection, leading finally to the joint law of censoring and truncation. In the last subsection of Section 14.3, based on the whole likelihood we discuss consistency in Hellinger distance of the density of survival or reliability under regularity conditions. We give assumptions on the set $\mathcal{F}$ of densities f of the survival time X which allow us to derive consistency of the NPMLE estimator of f along with the convergence rate (Huber, Solev and Vonta (2007) and (2009).

In Section 14.4 we provide an example where the joint law of the censoring and truncation can be explicitly computed, and which satisfies the conditions to get consistency and convergence rate of the NPMLE of the density f of the survival or reliability function.

14.2 FORMULATION – MARGINAL NON-PARAMETRIC LIKELIHOOD

We present here the general framework of the case of arbitrarily censored and truncated data for independent and identically distributed positive random variables following the formulation of Turnbull (1976), Frydman (1994) and Alioum and Commenges (1996). Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed positive random variables with survival function $S(x) = P(X > x)$. For every random variable X_i we have a pair of observations (A_i, B_i) where A_i is a set called the censoring set and B_i a set called the truncating set. The random variable X_i belongs to the sample only if X_i falls into the set B_i. Also, X_i is being censored by the set A_i in the sense that the only thing that we know about X_i is that it belongs to the set A_i where $A_i \subseteq B_i$. The sets A_i belong to a partition $\mathcal{P}_i$ of $[0, \infty)$ and we assume that B_i and $\mathcal{P}_i$ are independent of X_i and of the parameters of interest. We assume that the censoring sets A_i, $i = 1, \ldots, n$ can be expressed as a finite union of disjoint closed intervals, that is:

$$A_i = \bigcup_{j=1}^{k_i} [L_{ij}, R_{ij}]$$

where

$$0 \le L_{i1} \le R_{i1} < L_{i2} \le R_{i2} < \ldots < L_{ik_i} \le R_{ik_i} \le \infty \quad \text{for} \quad i = 1, \ldots, n,\ R_{i1} > 0,\ L_{ik_i} < \infty \quad .$$

Moreover, we assume that the truncating sets B_i can be expressed as a finite union of open intervals:

$$B_i = \bigcup_{j=1}^{n_i} \left(\mathcal{L}_{ij}, \mathcal{R}_{ij} \right)$$

where

$$0 \le \mathcal{L}_{i1} < \mathcal{R}_{i1} < \mathcal{L}_{i2} < \mathcal{R}_{i2} < \ldots < \mathcal{L}_{in_i} < \mathcal{R}_{in_i} \le \infty \quad \text{for} \quad i = 1, \ldots, n.$$

The likelihood of the n pairs of observations (A_i, B_i), $i = 1, 2, \ldots, n$ is proportional to:

$$l(S) = \prod_{i=1}^{n} l_i(S) = \prod_{i=1}^{n} \frac{P_S(A_i)}{P_S(B_i)} = \prod_{i=1}^{n} \frac{\sum_{j=1}^{k_i} \left\{ S(L_{ij}^-) - S(R_{ij}^+) \right\}}{\sum_{j=1}^{n_i} \left\{ S(L_{ij}^+) - S(R_{ij}^-) \right\}} \tag{14.1}$$

Let us define now the sets:

$$\tilde{L} = \{ L_{ij}, 1 \le j \le k_i, 1 \le i \le n \} \cup \{ \mathcal{R}_{ij}, 1 \le j \le n_i, 1 \le i \le n \} \cup \{0\}$$

and

$$\tilde{R} = \{R_{ij}, 1 \le j \le k_i, 1 \le i \le n\} \cup \{\mathcal{L}_{ij}, 1 \le j \le n_i, 1 \le i \le n\} \cup \{\infty\}.$$

Notice that the above likelihood is maximized when the values of $S(x)$ are as large as possible for $x \in \tilde{L}$ and as small as possible for $x \in \tilde{R}$. A set Q is defined uniquely as the union of disjoint closed intervals whose left endpoints lie in the set $\tilde{L}$ and right endpoints in the set $\tilde{R}$ respectively, and which contain no other members of $\tilde{L}$ or $\tilde{R}$. Thus:

$$Q = \bigcup_{j=1}^{v} \left[q'_j, p'_j\right]$$

where $0 = q'_1 \le p'_1 < q'_2 \le p'_2 < \ldots < q'_v \le p'_v = \infty$. Subsequently, we denote by C the union of intervals $\left[q'_j, p'_j\right]$ covered by at least one censoring set, W the union of intervals $\left[q'_j, p'_j\right]$ covered by at least one truncating set but not covered by any censoring set and $D = \overline{(\bigcup B_i)}$ the union of intervals $\left[q'_j, p'_j\right]$ not covered by any truncating set. D is actually included in the union of intervals $\left[q'_j, p'_j\right]$. That can be proved as follows. Let r be a point not covered by any truncating set and neither being a left nor a right endpoint of a truncating set. Then there exists l such that $r \in \left[q'_j, p'_j\right]$ as:

$$\mathcal{R}_{i_1 j_1} = \max_{i,j}\{\mathcal{R}_{ij} : \mathcal{L}_{ij} < r\} < r$$

$$\mathcal{L}_{i_2 j_2} = \min_{i,j}\{\mathcal{L}_{ij} : \mathcal{R}_{ij} > r\} > r$$

so that $r \in \left[q'_l, p'_l\right] \equiv \left[\mathcal{R}_{i_1 j_1}, \mathcal{L}_{i_2 j_2}\right]$.

Obviously, the set Q can be written as $Q = C \cup W \cup D$. Let us denote the set C as:

$$C = \bigcup_{i=1}^{m} \left[q_i, p_i\right]$$

where $q_1 \le p_1 < q_2 \le p_2 < \ldots < q_m \le p_m$. Let $s_j = S_{\overline{D}}(q_j^-) - S_{\overline{D}}(p_j^+)$ where $S_{\overline{D}}(x) = P\left(X > x \middle| X \in \overline{D}\right)$. The likelihood given in (1) can be written as a function of $s_1, s_2, \ldots, s_m$ that is:

$$l(s_1, \ldots, s_m) = \prod_{i=1}^{n} \frac{\sum_{j=1}^{m} \mu_{ij} s_j}{\sum_{j=1}^{m} \nu_{ij} s_j} \tag{14.2}$$

where $\mu_{ij} = I_{\left[\left[q_j, p_j\right] \subset A_i\right]}$ and $\nu_{ij} = I_{\left[\left[q_j, p_j\right] \subset B_i\right]}$, $i = 1, \ldots, n$ and $j = 1, \ldots, m$. The NPMLE of $S_{\overline{D}}$ was discussed by Turnbull (1976) and Frydman (1994). Turnbull (1976)

suggested a self-consistency algorithm in order to estimate the parameters $s_1,\ldots,s_m$. The algorithm is simple to implement and intuitively appealing.

Lemma 1. *Any survival function S which decreases outside the set $C \cup D$ cannot be the NPMLE of S.*

Lemma 2. *For fixed values of $S(q_j^-)$ and $S(p_j^+)$, for $1 \le j \le m$, the likelihood is independent of how the decrease actually occurs in the interval $[q_j,p_j]$, so that S is undefined within each interval $[q_j,p_j]$.*

For details in the proofs see Alioum and Commenges (1996) and Huber and Vonta (2004).

14.3 FORMULATION – COMPLETE NON-PARAMETRIC LIKELIHOOD

In the previous section we considered the likelihood of the observations A_i, B_i, $i = 1,\ldots,n$ which is not in fact a complete likelihood because it is defined conditionally on the censoring sets A_i and the truncating sets B_i. In this section we formulate the theory for a joint law of censoring and truncating mechanisms.

14.3.1 The Law of Censoring and Truncation

14.3.1.1 Random Covering

Let τ be a random partition defined on $(a; b)$, where usually a will be equal to 0 and b a finite strictly positive number but it could also be ∞:

$$\tau = \left\{ Y_0 = a < Y_1 < \ldots < Y_K < Y_{K+1} = b, \bigcup_{j=0}^{K} (Y_j, Y_{j+1}] = (a,b] \right\} \tag{14.3}$$

where K is a fixed number in $\{2, \ldots, K_0\}$ for some given K_0 such that $2 < K_0 < \infty$. The number K could of course be generalized to be random.

For each $x \in (a;b)$ we define:

$$k = k(x) = \inf\left\{ j: x \le Y_{j+1} \right\}. \tag{14.4}$$

$$\vartheta(x) = \left(Y_{k(x)}, Y_{k(x)+1}\right] := \left(L(x), R(x)\right] \; x \in (a,b). \tag{14.5}$$

where $L(x)$ and $R(x)$ may be thought of as the left and right values in partition τ that "bracket" (the survival) $X = x$.

Then it is clear that:

$$\vartheta(x) = \vartheta(y), \text{ or } \vartheta(x) \cap \vartheta(y) = \varnothing \tag{14.6}$$

and we call $\vartheta(x)$ a simple random covering of (a, b).

14.3.1.2 The Mechanism of Censoring and Truncation

The mechanism of censoring and truncating of a random variable X is defined as follows. Let X be a random variable, $\Delta=(Z_1,Z_2]$ be a random interval, $\vartheta(x)=\left(L(x),R(x)\right]$, $x\in(a,b)$ be a random covering, generated by a partition τ defined in Equation (14.3).

We suppose that the random covering $\vartheta(\cdot)$, the random variable X and the random interval Δ are independent, but we do not have complete observations. More precisely, we suppose that the random vector (X, Z_1, Z_2) is partly observable only in the case when $\left(L(X),R(X)\right]\subset\Delta$:

$$Z_1\le L(X)<R(X)\le Z_2.$$

In that case the available observations are the censoring interval $(L(X),R(X)]$ of the covering $\vartheta(\cdot)$, which contains X, and the random truncating interval $\Delta^*=\left(R(Z_1),L(Z_2)\right]$. When $\left(L(X),R(X)\right]\subset\left(Z_1,Z_2\right]$ we do not have any observation.

Let us define:

1) Conditionally on a fixed value t of τ the random interval Δ is taken from the conditional distribution:

$$\mathcal{P}_t\{A\}=P\left\{\Delta\in A\middle|\text{the interval }[Z_1,Z_2]\text{ contains at least two points of }t\right\}.$$

In other words, conditionally on fixed values of $\tau=t$ the random vector $Z=(Z_1, Z_2)$ is taken from the conditional distribution:

$$P_t\{B\}=P\left\{Z\in B\,\middle|\,\mathfrak{z}_1(t,Z_1)<\mathfrak{z}_2(t,Z_2)\right\}\equiv P\left\{Z\in B\middle|R\left(Z_1\right)<L\left(Z_2\right)\right\};$$

2) Conditionally on a fixed value of $\tau=t$ and $\Delta=\delta=(z_1,z_2]$, the random variable X is taken from the conditional distribution:

$$P_{t,\delta}\{C\}=P\left\{X\in C\middle|X\in\left(R(z_1),L(z_2)\right]\right\}.$$

In other words conditionally on fixed values of $\tau=t$ and $Z_1=z_1$, $Z_2=z_2$ the random variable X is taken from the conditional distribution:

$$P\left\{C\middle|t,z_1,z_2\right\}=P\left\{X\in C\middle|X\in\left(\mathfrak{z}_1(t,z_1),\mathfrak{z}_2(t,z_2)\right]\equiv\left(R(z_1),L(z_2)\right]\right\}. \qquad (14.7)$$

We now consider the simple case of right truncation where for a random variable Z the random truncating interval takes the form $\Delta=(a,Z]$, and we use the same notations as previously in this section. We denote by $\mathfrak{z}$ the random variable:

$$\mathfrak{z}=\mathfrak{z}(\tau,Z)=L(Z).$$

Recall that the random covering $\vartheta(\cdot)$, the survival X, and the truncating random variable Z are independent.

In this case, Equation (14.7) above becomes:

Conditionally on fixed values of $\tau=t$ and $Z=z$ the random variable X is taken from the conditional distribution:

$$P\{C \mid t,z\} = P\{X \in C \mid X \le \mathfrak{z}(t,z) \equiv L(z)\}.$$

14.3.1.3 The Distribution Associated with the Random Covering

Let $\vartheta(x) = \left(L(x), R(x)\right]$, $x \in (a,b)$, be a simple random covering. The distribution P_x of random vector $\upsilon(x) = \left(L(x), R(x)\right)$ will be called the distribution associated with the random covering $\vartheta(x)$.

We assume that for all x the distribution P_x has density with respect to Lebesgue measure λ^2 on the plane $\mathbb{R}^2$:

$$r_x(u,v) = \frac{dP_x}{d\lambda^2}.$$

It is easy to see that there exists a non-negative function $r(u, v)$ such that for all x:

$$r_x(u,v) = r(u,v)\mathbb{I}_{(u,v]}(x) \text{ (a.s.)}$$

The function $r(u, v)$ will be called the *baseline density of the simple random covering* $\vartheta(x)$. It is clear that the function $r(u, v)$ is the density of a σ-finite measure, but, for all x, the function $r(u,v)\mathbb{I}_{(u,v]}(x)$ is the density of a probability measure.

It is easy to see that the baseline density $r(u, v)$ depends only on the joint distributions of vectors (Y_j,Y_{j+1}).

Lemma 3. *The measure P_x is absolutely continuous with respect to the Lebesgue measure for all x and fixed K if and only if:*

(i) *for all j the distribution of the vector (Y_j,Y_{j+1}) has density $r^j(u, v)$ with respect to the Lebesgue measure,*

(ii) *the series $\sum_j r^j(u,v) < \infty$ (λ^2-a.s.) to a function $r(u, v)$,*

(iii) *the function r(u,v) satisfies the following condition: for all x:*

$$r_x(u,v) = r(u,v)\mathbb{I}_{(u,v]}(x).$$

14.3.1.4 The Distribution of the Random Vector ($L(x)$, $R(x)$, $L(z)$, $R(z)$)

From now on we concentrate on the case of right truncation. Due to censoring by the partition t, z is not observed. Instead we have $]L(z); R(z)] \ni z$, and only $L(z)$ is observed. Now for $x < z$ we denote by $P_{x,z}$ the distribution of the random vector ($L(x)$, $R(x)$, $L(z)$, $R(z)$).

Denote by λ^n the Lebesgue measure on $\mathbb{R}^n$. The distribution $P_{x,z}$ is not absolutely continuous with respect to the measure on λ^4. Denote by ν the measure, which is defined for continuous non-negative functions $\psi(s)=\psi(s_1,s_2,s_3,s_4)$ by the relation:

$$\begin{aligned}\iiiint \psi(s)d\nu = &\iint \psi(s_1,s_2,s_1,s_2)ds_1ds_2\\ &+\iiint \psi(s_1,s_2,s_2,s_4)ds_1ds_2ds_4\\ &+\iiiint \psi(s_1,s_2,s_3,s_4)ds_1ds_2ds_3ds_4.\end{aligned}$$

We suppose that the distribution $P_{x,z}$ is absolutely continuous with respect to the measure ν and denote its density by $q_{x,z}(s)$:

$$q_{x,z}(s)=q_{x,z}(s_1,s_2,s_3,s_4)=\frac{dP_{x,z}}{d\nu}$$

We suppose that for all $n>0$ the random vector $(Y_1, \ldots, Y_n)$ has a density with respect to the corresponding Lebesgue measure λ^n. For $i+1<j$ and K fixed let the function:

$r_{i,j}(y_1,y_2,y_3,y_4)$ be the density of the random vector $(Y_i, Y_{i+1}, Y_j, Y_{j+1})$,

$r_j(y_1,y_2,y_3)$ be the density of the random vector (Y_{j-1}, Y_j, Y_{j+1}),

and

$r^j(y_1,y_2)$ be the density of the random vector (Y_j, Y_{j+1}).

We assume that:

$$\vartheta_4(y_1,y_2,y_3,y_4)=\sum_{\substack{i,j:\\ i+1<j}} r_{ij}(y_1,y_2,y_3,y_4)<\infty \quad (\lambda^4\text{-a.s.}),$$

$$\vartheta_3(y_1,y_2,y_3)=\sum_j r_j(y_1,y_2,y_3)<\infty \quad (\lambda^3\text{-a.s.}),$$

and

$$\vartheta_2(y_1,y_2)=\sum_j r^j(y_1,y_2)<\infty \quad (\lambda^2\text{-a.s.}).$$

For a non-negative function $\psi(s)$, $s=\left(s_1,s_2,s_3,s_4\right)$ and $x<z$ we have

$$\begin{aligned}\mathbf{E}\psi(L(x),R(x),L(z),R(z)) &= \sum_{i,j}\mathbf{E}\psi(Y_i,Y_{i+1},Y_j,Y_{j+1})\mathbb{I}_{(Y_i,Y_{i+1}]}(x)\mathbb{I}_{(Y_j,Y_{j+1}]}(z)\\ &= \sum_{j}\mathbf{E}\psi(Y_j,Y_{j+1},Y_j,Y_{j+1})\mathbb{I}_{(Y_j,Y_{j+1}]}(x)\mathbb{I}_{(Y_j,Y_{j+1}]}(z)\\ &\quad+\sum_{j}\mathbf{E}\psi(Y_{j-1},Y_j,Y_j,Y_{j+1})\mathbb{I}_{(Y_{j-1},Y_j]}(x)\mathbb{I}_{(Y_j,Y_{j+1}]}(z)\\ &\quad+\sum_{\substack{i,j:\\ i+1<j}}\mathbf{E}\psi(Y_i,Y_{i+1},Y_j,Y_{j+1})\mathbb{I}_{(Y_i,Y_{i+1}]}(x)\mathbb{I}_{(Y_j,Y_{j+1}]}(z).\end{aligned}$$

Thus:

$$\begin{aligned}&\mathbf{E}\psi(L(x),R(x),L(z),R(z))\\ &= +\iint\psi(s_1,s_2,s_1,s_2)\mathfrak{d}_2(s_1,s_2)\mathbb{I}_{(s_1,s_2]}(x)\mathbb{I}_{(s_1,s_2]}(z)ds_1ds_2\\ &\quad+\iiint\psi(s_1,s_2,s_2,s_3)\mathfrak{d}_3(s_1,s_2,s_3)\mathbb{I}_{(s_1,s_2]}(x)\mathbb{I}_{(s_2,s_3]}(z)ds_1ds_2ds_3\\ &\quad+\iiiint\psi(s_1,s_2,s_3,s_4)\\ &\quad\times\mathfrak{d}_4(s_1,s_2,s_3,s_4)\mathbb{I}_{(s_1,s_2]}(x)\mathbb{I}_{(s_3,s_4]}(z)ds_1ds_2ds_3ds_4.\end{aligned}$$

If we define a ν-measurable function $\mathfrak{d}(s\,|\,x,z)$, $s=(s_1,s_2,s_3,s_4)$, by

$$\mathfrak{d}\left(s\,|\,x,z\right)=\mathbb{I}_{(s_1,s_2]}(x)\mathfrak{d}_*\left(s\,|\,z\right),$$

where

$$\mathfrak{d}_*\left(s\,|\,z\right)=\begin{cases}\mathfrak{d}_2(s_1,s_2)\mathbb{I}_{(s_1,s_2]}(z), & \text{if } s_1=s_3<s_2=s_4\\ \mathfrak{d}_3(s_1,s_2,s_4)\mathbb{I}_{(s_2,s_4]}(z), & \text{if } s_1<s_2=s_3<s_4\\ \mathfrak{d}_4(s_1,s_2,s_3,s_4)\mathbb{I}_{(s_3,s_4]}(z), & \text{if } s_1<s_2<s_3<s_4\\ 0, & \text{else}\end{cases}\tag{14.8}$$

then we obtain for $x<z$

$$\mathbf{E}\psi\left(L(x),R(x),L(z),R(z)\right)=\iiiint\psi(s)\mathfrak{d}\left(s\,|\,x,z\right)dv,$$

and therefore

$$q_{x,z}(s_1,s_2,s_3,s_4)=\mathbb{I}_{(s_1,s_2]}(x)\mathfrak{d}_*\left(s_1,s_2,s_3,s_4\,|\,z\right).\tag{14.9}$$

14.3.1.5 The Distribution of the Random Vector (*L*(*X*), *R*(*X*), *L*(*Z*), *R*(*Z*))

For the right-truncated density function $f(x)$ we shall use the following notation:

$$f_{b_s}(x) = \frac{f(x)}{\int_{u \le b_s} f(u)\, du} \mathbb{I}_{(a,b_s]}(x).$$

Now we suppose that for fixed z and fixed value of $\tau = t$, the random variable X is taken from the truncated distribution with density $f_{\mathfrak{z}}(x)$. Here $\mathfrak{z} = \mathfrak{z}(t,z) = L(z)$. It follows from Equation (14.9) that in that case the distribution P_z of random vector $(L(X), R(X), L(z), R(z))$ has density (with respect to the measure ν) $q(s_1, s_2, s_3, s_4 \mid z)$:

$$q\left(s_1, s_2, u, v \mid z\right) = \int q_{x,z}(s_1, s_2, u, v) f_u(x)\, dx,$$

and (see Equation (14.8))

$$q\left(s_1, s_2, u, v \mid z\right) = \int_{s_1}^{s_2} f_u(x)dx \times \mathfrak{d}_*\left(s_1, s_2, s_3, s_4 \mid z\right),$$

where for $s = (s_1, s_2, s_3, s_4)$

$$\mathfrak{d}_*\left(s \mid z\right) = \begin{cases} \mathfrak{d}_3(s_1, s_2, s_4)\mathbb{I}_{(s_2, s_4]}(z), & \text{if } s_1 < s_2 = s_3 < s_4 \\ \mathfrak{d}_4(s_1, s_2, s_3, s_4)\mathbb{I}_{(s_3, s_4]}(z), & \text{if } s_1 < s_2 < s_3 < s_4 \\ 0, & \text{else} \end{cases}$$

Therefore the distribution P_z is absolutely continuous with respect to the measure ν_*, which is defined for continuous non-negative functions $\psi(s)$ by the relation

$$\int\int\int\int \psi(s) d\nu_* = \int\int\int \psi(s_1, s_2, s_2, s_4)\, ds_1 ds_2 ds_4 + \int\int\int\int \psi(s_1, s_2, s_3, s_4)\, ds_1 ds_2 ds_3 ds_4,$$

and

$$\frac{dP_z}{d\nu_*} = q\left(s \mid z\right).$$

Now suppose that Z is a random variable with density g, which is independent from the random covering $\vartheta(\cdot)$. For fixed values $Z = z$ and $\tau = t$, the random variable X is taken from the conditional distribution with density $f_{\mathfrak{z}}(x)$, $\mathfrak{z} = \mathfrak{z}(t,z) = L(z)$. Denote by P_* the distribution of the random vector $(L(X), R(X), L(Z), R(Z))$. It is clear that the distribution P_* has density $q(s)$ with respect to the measure ν_*:

$$q(s_1,s_2,u,s_4)=\int_{s_1}^{s_2} f_u(x)\,dx\times\int \mathfrak{d}_*\left(s_1,s_2,u,s_4\,\middle|\,z\right)g(z)\,dz$$

$$=\int_{s_1}^{s_2} f_u(x)\,dx\times\mathfrak{d}(s_1,s_2,u,s_4).$$

Now consider the random vector $W=(L(X), R(X), L(Z))$. Let ν^{**} be the measure on $\mathbb{R}^3$, defined for continuous non-negative functions ψ by:

$$\iiint\psi(s_1,s_2,s_3)\,d\nu^{**}=\iint\psi(s_1,s_2,s_2)\,ds_1ds_2$$

$$+\iiint\psi(s_1,s_2,s_3)\,ds_1ds_2ds_3,$$

It is clear that the distribution P_W of random vector W is absolutely continuous with respect to the measure ν^{**} and:

$$p(y)=p(y_1,y_2,y_3)=\frac{dP_W}{d\nu^{**}}=\int q(y_1,y_2,y_3,u)\,du.$$

Therefore:

$$p(u,v,z)=\int_u^v f_z(x)\,dx\times r(u,v,z),$$

where

$$r(u,v,z)=\int\mathfrak{d}(u,v,z,x)\,dx.$$

14.3.2 Estimation of the Density of Survival or Reliability

In this part we present theoretical results about consistency under certain regularity conditions on the law of censoring, truncation, and survival of the density of the survival time X. The formulation is as follows. Let W, W_1, …, W_n be i.i.d. random vectors, $W=(L(X), R(X), L(Z))$, with unknown density:

$$p(u,v,w)=r(u,v,w)\times\frac{\int_u^v f(x)\,dx}{\int_{x\le w} f(x)\,dx} \tag{14.10}$$

We assume that the baseline density r and density f belong to given sets $\mathcal{G}$ and $\mathcal{F}$ correspondingly (see details in Huber, Solev and Vonta, 2009). We set:

$$\varphi(f;u,v,w)=\frac{\int_u^v f(x)\,dx}{\int_{x\le w} f(x)\,dx},$$

$$\mathcal{P}=\left\{p: p=r\,\varphi\,(f;\cdot),(r,f)\in\mathcal{G}\times\mathcal{F}\right\} \tag{14.11}$$

Definition 1. Let χ be a bounded set in $\mathbb{R}^d$, $0<\alpha\le 1$, $r^*\in\mathbb{N}$, and $\beta=r^*+\alpha$.

Then $\mathcal{C}^{\beta}_{M_0}$ is the set of all functions from χ onto $\mathbb{R}$ that possess uniformly bounded partial derivatives up to order r^* and whose highest partial derivatives are Lipschitz functions of order α. More precisely, for any $k=(k_1,\ldots,k_d)$:

$$\begin{aligned}\|g\|_{\beta}&=\max\sum_{k_i\le r^*}\sup_x\left|D^k g(x)\right|\\&+\max\sum_{k_i=r^*}\sup_{x,y}\frac{\left|D^k g(x)-D^k g(y)\right|}{\|x-y\|^{\alpha}}\le M_0\end{aligned} \tag{14.12}$$

where the supremum is taken over all x, y in the interior of χ with $x\neq y$.

Theorem 1. Suppose that the parameter of interest, that is, the true density f, with respect to Lebesgue measure, of the survival time X, belongs to the space:

$$\mathcal{F}=\left(f\ :\ f\in\mathcal{C}^{\beta}_{M_0}\text{ with compact support }\chi\text{ and }0<c_l\le f\le c_u<\infty\right) \tag{14.13}$$

with $\beta>1/2$. Also, suppose that the density r which describes the censoring and truncation mechanism is known and bounded by a constant $r_0>0$. Then the non-parametric maximum likelihood estimator $\hat{f}_n$ is consistent in the Hellinger distance for the density f, namely, for any $\varepsilon>0$:

$$\sup_{p=p^{r,f}\in\mathcal{P}} P_p\left\{h\left(f_n,f\right)>\varepsilon\right\}\to 0,\text{ as } n\to\infty.$$

More specifically, the rate of convergence is given by:

$$\sup_{p=p^{r,f}\in\mathcal{P}} P_p\left\{h\left(f_n,f\right)>n^{-\frac{\beta}{2\beta+1}}\right\}\le C\exp\left(-c_2\,n^{\frac{1}{2\beta+1}}\right).$$

The proofs use results of Stein (1993), Wong and Shen (1995) and van der Vaart and Wellner (2000) on non-parametric estimation.

14.4 EXAMPLE

We will provide here an example of the joint law of censoring and truncation in order to illustrate some of the theoretical results provided in the previous sections. We will

consider for simplicity the right truncation case. We also consider as the total interval of observation time the interval [0,1]. Since we do not want to have any boundary issues we consider the interval $[0+\varepsilon, 1-\varepsilon] \equiv [a.X, b.X]$ where ϵ is small. For a fixed m, we consider $m+1$ equidistant times $\{t_j, j=1,\ldots,m+1\}$. m is thus the number of sub-intervals of [$a.X$; $b.X$]:

$$t_j = a.X + (j-1)\frac{b.X - a.X}{m} \quad j = 1,\ldots,m+1.$$

The censoring time Y_j is assumed to occur uniformly inside the jth interval so that:

$$r^j(y_j, y_{j+1}) = \left(\frac{m}{(b.X - a.X)}\right)^2 1_{[t_j, t_{j+1}[}(y_j)\, 1_{[t_{j+1}, t_{j+2}[}(y_{j+1})$$

$$r_{i,j}(y_i, y_{i+1}, y_j, y_{j+1}) = \left(\frac{m}{(b.X - a.X)}\right)^4 1_{[t_i, t_{i+1}[}(y_i)\, 1_{[t_{i+1}, t_{i+2}[}(y_{i+1})$$

$$\times 1_{[t_j, t_{j+1}[}(y_j) 1_{[t_{j+1}, t_{j+2}[}(y_{j+1})$$

At the same time, the right-truncating variable Z follows an independent uniform ($a.Z$, $b.Z$) distribution where $b.Z > 1$ is fixed. So:

$$f_Z(z) = \frac{1}{b.Z - a.Z}, \quad \text{for} \quad a.Z \le z \le b.Z.$$

It is natural to assume that a certain proportion π of $Y's$, say for example $\pi = 3/4$ of the $Y's$ have occurred before the truncating variable truncates the sample.

Let therefore $a.Z = Y_{j.z}$ where $j.z = [\pi m]$. It is also reasonable to assume that some subjects share the same truncating value although in full generality each subject could have its own truncating value.

The survival time X follows a distribution with unknown density $f(x), x \in [a.X, b.X]$ with respect to the Lebesgue measure on $\mathbb{R}^+$, which we assume to belong to the space $\mathcal{F}$ defined in (13). What we observe actually is W_1, …, W_n, a sample of i.i.d. random vectors, where $W = (L(X), R(X), L(Z))$, with density:

$$p(u,v,w) = r(u,v,w) \times \frac{\int_u^v f(x)\,dx}{\int_{x \le w} f(x)\,dx}.$$

We would like to compute $r(u, v, w)$ for the above-described case of interval-censoring and right truncation.

Lemma 4. *For the observational scheme described above, the density $r(u, v, w)$ of the censoring and truncating mechanisms, has two parts, denoted by r_3 and r_2. The law r_3 is defined as:*

$$r_3(y_k, y_{k+1}, y_l)$$

$$= \begin{cases} r^k(y_k, y_{k+1}) & \text{if } 1 \le k < j.z-1 \\ r^k(y_k, y_{k+1}) \dfrac{(m-k-1)(b.X-a.X)/m + (b.Z-b.X)}{b.Z-a.Z} & \text{if } j.z-1 \le k \le m-1 \\ 0 & \text{if } k = m \end{cases} \tag{14.14}$$

for $a.X \le y_k < y_{k+1} < y_l \le b.X$

while the law r_2 is defined as:

$$r_2(y_k, y_{k+1}) = \begin{cases} 0 & \text{if } 1 \le k < j.z \\ r^k(y_k, y_{k+1}) \dfrac{(b.X-a.X)/m}{b.Z-a.Z} & \text{if } k \ge j.z \text{ and } k < m \\ r^k(y_k, y_{k+1}) \dfrac{b.Z-b.X}{b.Z-a.Z} & \text{if } k = m \end{cases} \tag{14.15}$$

for $a.X \le y_k < y_{k+1} \le b.X$.

Note that the above joint law of censoring and truncation satisfies the above general assumptions and in particular the boundedness assumption.

REFERENCES

A. Alioum and D. Commenges (1996). A proportional hazards model for arbitrarily censored and truncated data, *Biometrics*, 52(2), 512–524.

H. Frydman (1994). A note on nonparametric estimation of the distribution function from interval-censored and truncated observations, *Journal of the Royal Statistical Society, Series B*, 56(1), 71–74.

J. Huang (1996). Efficient estimation for the proportional hazards model with interval censoring, *The Annals of Statistics*, 24(2), 540–568.

J. Huang and J. A. Wellner (1995). Efficient estimation for the proportional model with "case 2" interval censoring. Technical Report, Department of Statistics, University of Washington.

C. Huber, V. Solev and F. Vonta (2007). Maximum likelihood estimators: Nonparametric approach, *Journal of Mathematical Sciences*, 147(4), 6975–6979.

C. Huber-Carol, V. Solev and F. Vonta (2009). Interval censored and truncated data: Rate of convergence of NPMLE of the density, *Journal of Statistical Planning and Inference*, 139(5), 1734–1749.

C. Huber and F. Vonta (2004). Frailty models for arbitrarily censored and truncated data, *Lifetime Data Analysis*, 10(4), 369–388.

C. M. Stein (1993). *Harmonic Analysis*, Princeton University Press.

B. W. Turnbull (1976). The empirical distribution function with arbitrary grouped, censored and truncated data, *Journal of the Royal Statistical Society*, 38, 290–295.

A. W. Van der Vaart and J. A. Wellner (2000). *Weak Convergence and Empirical Processes (With Applications to Statistics)*, Springer Series in Statistics, Springer.

W. H. Wong and N. X. Shen (1995). Probability inequalities for likelihood ratios and convergence rates of sieve mles, *The Annals of Statistics*, 23(2), 339–362.