Heavy Random Truncation *

Shuyuan He
Dept of Prob. and Stat.
Peking University, Beijing

Grace L. Yang
Department of Mathematics
University of Maryland
College Park, MD 20742

July 8, 2006

1 Introduction

Consider an infinite sequence of independent random vectors

$$(X_m, Y_m), m = 1, 2, \cdots,$$

where the X_m have a common distribution function F and the Y_m have a common distribution function G. The components X_m and Y_m are also independent for each m.

Suppose both X_m and Y_m are observable only when $X_m \geq Y_m$. The observable pairs thus form a subsequence $\{j\}$ of the original sequence $\{m\}$. It is denoted by $\{(U_j, V_j), j = 1, 2, \cdots\}$.

*Research supported by National Natural Science Foundation of China(10231030)

Here the subsequence is labeled consecutively for simplicity. The limitation in observation induces dependence and the constraint $U_j \geq V_j$ in each pair j.

The vectors (U_j, V_j) remain iid. In describing the distributional properties of any pair we shall use (X, Y) to refer to any pair (X_m, Y_m), and (U, V) to (U_j, V_j).

The random truncation model is defined by the joint distribution $H(x, y)$ of (U, V). It is the conditional distribution of (X, Y) given $[X \geq Y]$,

$$(1) \quad H(x, y) = P[U \leq x, V \leq y] = P[X \leq x, Y \leq y | X \geq Y].$$

A problem of interest is to estimate the distribution function F of X based on a randomly truncated sample of n iid observations $(U_j, V_j), j = 1, \cdots, n$.

Truncated data occur in astronomy, economics, [e.g. Woodroofe (1985), Feigelson and Babu (1992)], epidemiology, biometry [e.g. Wang, et al. (1986), Tsai, et al. (1987), He & Yang (1994)], and possibly in other fields such as spike train data in neurophysiology.

The truncation event $[X \geq Y]$, among other things, affects the range of observation of the X. Only F_0 defined by

$$(2) \qquad F_0(x) = P[X \leq x | X \geq a_G]$$

is estimable from the truncated sample $(U_j, V_j), j = 1, \cdots, n$, where

$$a_G = \inf\{y : G(y) > 0\}$$

is the lower boundary of Y. We shall denote the upper boundary of Y by

$$b_G = \sup\{y : G(y) < 1\}. \tag{3}$$

Similar symbols, a_F, b_F, will be used for the boundaries of X.

Obviously, if $a_G \leq a_F$, $F_0 = F$. Analogously, define $G_0(y) = P[Y \leq y | Y \leq b_F]$. Thus if $b_F \geq b_G, G_0 = G$. Let $I[A]$ denote the indicator function of the event A. Let

(4)
$$F_n^*(s) = n^{-1} \Sigma_{i=1}^n I[U_i \leq s],$$

(5)
$$G_n^*(s) = n^{-1} \Sigma_{i=1}^n I[V_i \leq s],$$

(6)
(7)
$$R_n(s) = G_n^*(s) - F_n^*(s-)$$

(8)
$$= n^{-1} \Sigma_{i=1}^n I[V_i \leq s \leq U_i],$$

be the empirical processes of the data.

Here and in what follows, for any real function g, the left limit

$\lim_{y\uparrow s} g(y)$ is denoted by $g(s-)$ and the difference $g(s)-g(s-)$ by the curly brackets $g\{s\}$.

The nonparametric maximum likelihood estimates of F_0 and G_0 are given respectively by

(9) $$F_n(x) = 1 - \Pi_{s\leq x}\left[1 - \frac{F_n^*\{s\}}{R_n(s)}\right],$$

(10)

(11) $$G_n(x) = \Pi_{s>x}\left[1 - \frac{G_n^*\{s\}}{R_n(s)}\right],$$

where $x \in (-\infty, \infty)$ and an empty product is set equal to one.

One of the results obtained by Woodroofe (1985) is that for any continuous F and G,

$$\sup_x |F_n(x) - F_0(x)| \rightarrow 0 \ \ \textit{in probability as } n \rightarrow \infty.$$

If F and G are not continuous, the limit has to be modified. For arbitrary F and G, there are two kinds of limit F_0 and F_a where F_a is defined by

$$F_a(x) = P[X \leq x | X > a_G].$$

If condition $B1: \ a_F = a_G, \ G\{a_G\} = 0$ and $F\{a_F\} > 0$, holds, then

$$\sup_x |F_n(x) - F_a(x)| \rightarrow 0 \quad a.s,$$

and

$$\sqrt{n}(F_n(x) - F_a(x))$$

converges weakly to a function of Gaussian processes. Otherwise,

$$\sup_x |F_n(x) - F_0(x)| \to 0 \quad a.s.$$

and

$$\sqrt{n}(F_n(x) - F_0(x))$$

converges weakly to a function of Gaussian processes.

For the truncation probability $\alpha = P[X \geq Y]$, a proper estimate is not the sample proportion but $\alpha_n = \int G_n(s)dF_n(s)$ where F_n and G_n are product limit estimates of the distribution functions F and G of X and Y, respectively. Under some conditions , α_n is strongly consistent estimator of the truncation probability $\alpha = P[X \geq Y]$.

In this talk we will show if the truncation probability α changes with the data, the limit distribution will be a function of Poisson Processes.

For censoring case, similar work can be find in Wellner (1985).

2 Main Results

For each positive integer n, consider an infinite sequence of nonnegative independent random vectors $(X_{n,j}, Y_{n,j}), j = 1, 2, \cdots$, where the $X_{n,j}$ have a common right continuous distribution function F_n and the $Y_{n,j}$ have a common right continuous distribution function G_n with $G_n(0) = 0$. The components $X_{n,j}$ and $Y_{n,j}$ are also independent for each j.

Suppose both $X_{n,j}$ and $Y_{n,j}$ are observable only when $X_{n,j} \geq Y_{n,j}$ and the observation is denoted by $\{(U_{n,i}, V_{n,i}), i = 1, 2, \cdots, \}$. Here the subsequence is labeled consecutively for simplicity. The observational limitation induces the dependence and the constraint $U_{n,i} \geq V_{n,i}$ in each pair i.

However, the vectors $(U_{n,i}, V_{n,i})$ remain iid. Let $N(n) \leq N(n+1) \leq ...$ be an integer sequence such that $N(n) \to \infty$ as $n \to \infty$. Where and in what follows we use $I[A]$ or I_A for the indicator function of the event A.

A problem of interest is to estimate the distribution function F_n of $X_{n,j}$ based on the randomly truncated sample of $m(n)$ iid observations $(U_{n,i}, V_{n,i}), i = 1, \cdots, m(n)$ with $m_n = \Sigma_{j=1}^{N(n)} I\left[X_{n,j} \geq Y_{n,j}\right]$.

In what follows we suppose all the random variables are defined on probability space $(\Omega, \mathcal{A}, P)$. Let

$$\hat{H}_n(s) = \sum_{i=1}^{m(n)} I\left[U_{n,i} \leq s\right],$$

$$\hat{K}_n(s) = \sum_{i=1}^{m(n)} I\left[V_{n,i} \leq s\right],\ 0 \leq s < \infty,$$

$$\hat{R}_n(s) = \hat{K}_n(s) - \hat{H}_n(s-),\ 0 \leq s < \infty,$$

the empirical processes of the data. Then

$$\hat{H}_n(s) = \sum_{i=1}^{N(n)} I\left[X_{n,i} \leq s, X_{n,i} \geq Y_{n,i}\right],$$

$$\hat{K}_n(s) = \sum_{i=1}^{N(n)} I\left[Y_{n,i} \le s, X_{n,i} \ge Y_{n,i}\right].$$

Here and in what follows, for any real function g, the left limit $\lim_{y \uparrow s} g(y)$ is denoted by $g(s-)$ and the difference $g(s) - g(s-)$ by the curly brackets $g\{s\}$.

The nonparametric maximum likelihood estimate of F_n is given by

$$\hat{F}_n(x) = 1 - \prod_{s \le x} \left[1 - \frac{\hat{H}_n\{s\}}{\hat{R}_n(s)}\right],$$

where an empty product is set equal to one.

The cumulative hazard function (see Woodroofe (1985)) $\hat{\Lambda}_n$ of $\hat{F}_n$ is defined by

$$\hat{\Lambda}_n(x) = \int_0^x \frac{d\hat{F}_n(s)}{1 - \hat{F}_n(s-)} = \int_0^x \frac{d\hat{H}_n(s)}{\hat{R}_n(s)}.$$

Let $L = \{(x, y); x \ge y\}$ be the subset of $R_+^2 = [0, \infty) \times [0, \infty)$ and r the Euclidean metric. The following conditions will be used throughout.

Condition 1. For any $(x, y) \in L$, the limit

$$a(x, y) = \lim_n N(n) G_n(y)(1 - F_n(x))$$

exists and is continuous on L. $a(x,y) \to 0$ as $x \to \infty$.

Condition 2. For $t \geq 0$, the limit

$$\lim_n \int_0^t b(t) = N(n)(1 - F_n(s-))dG_n(s)$$

exists and is continuous.

Let $x \wedge y = \min(x,y)$. For $(x,y) \in R_+^2$ define

$$W_n(x,y) = N(n) \int_0^x \left(\int_0^y I_L dG_n \right) dF_n,$$

$$W(x,y) = b(x \wedge y) - a(x, x \wedge y)$$

respectively. Then W is continuous on R_+^2. Let $\mathcal{R}_+^2$ be the Borel σ-field of R_+^2 and

$$\begin{aligned} &\mu_n(D) = \int_D dW_n, \\ &\mu(D) = \int_D dW, \quad D \in \mathcal{R}_+^2, \quad n \geq 1. \end{aligned}$$

We have

$$\begin{aligned} &\lim_{n\to\infty} W_n(x,y) \\ =&\lim_n N(n) \int_0^{x \wedge y} (F_n(x) - F_n(s-))dG_n(s) \\ =&W(x,y). \end{aligned}$$

Hence for any continuous functions f: $R_+^2 \to R_+ = [0, \infty)$ with compact support, we get

$$\lim_n \int f d\mu_n = \int f d\mu. \tag{12}$$

Let $\mathcal{M}$ be the set of all locally finite measures on R_+^2 and ρ the metric of $\mathcal{M}$ which induces the topology. Then μ_n, $\mu \in \mathcal{M}$. For any metric space (S, d), let $C(S)$ denote the class of all bounded continuous function $S \to R_+$, and $C_1(S)$ the subclass of all functions in $C(S)$ with compact support. Let μ_n and μ be locally finite measures on S.

According to Kallenberg(1976), for finite μ_n and μ, $\mu_n \to \mu$ weakly if condition (12) is true for all $f \in C(S)$. For random elements ξ_n and ξ in (S, d), $\xi_n \to \xi$ weakly if $Ef(\xi_n) \to Ef(\xi)$ for all $f \in C_1(S)$ and $\xi_n \to \xi$ weakly if $Ef(\xi_n) \to Ef(\xi)$ for all $f \in C(S)$. It is clear that $\xi_n \to \xi$ weakly if and only if $P^{-1}\xi_n \to P^{-1}\xi$ weakly.

Lemma 2.1 *Let μ_n and μ be defined by (1). If $\mu_n(L) \to \mu(L) < \infty$, then $\mu_n \to \mu$ weakly.*

Lemma 2.2 *Let ξ_n and ξ be defined above.*

a). If $\mu_n(L) \to \mu(L) < \infty$, then as random elments in $(\mathcal{M}, \rho)$, $\xi_n \to \xi$ weakly.

b). For any $T \in (0, \infty)$, as random elements in $\mathcal{M}_T$, $\xi_n \to \xi$ weakly.

For $0 < T \leq \infty$ and $j = 1$ or 2, let $D_j[0,T)$ be the space of right continuous function f: $[0,T) \to R^j$ with left limits. Let d be the metric that induces the Skorohod topology on $D_j[0,T)$. Then $(D_j[0,T), d)$ is separable and complete.

Define a measurable mapping $\mathcal{M}_T \to D_2[0,T)$:

$$g(\beta) = (g_1(\beta), g_2(\beta)),\ \beta \in \mathcal{M}_T. \tag{13}$$

with

$$\begin{aligned} g_1(\beta)(t) &= \beta\left[B(t)\right], \\ g_2(\beta)(t) &= \beta\left[D(t)\right]. \end{aligned}$$

We have, for any $t \in [0,T)$

$$\begin{aligned} &g(\xi_n)(t) = (\hat{H}_n(t), \hat{K}_n(t)) \\ &= \left(\sum_{j=1}^{N(n)} I_{B(t)}(X_{n,j}, Y_{n,j}), \sum_{j=1}^{N(n)} I_{D(t)}(X_{n,j}, Y_{n,j}) \right), \\ &g(\xi)(t) = (H(t), K(t)) = (\xi\left[B(t)\right], \xi\left[D(t)\right]). \end{aligned}$$

$g(\xi)$, $g(\xi_n)$, $n = 1, 2, ...$ are random elements in $(D_2[0,T), d)$. H and K are Poisson processes with intensity function $\gamma_1(t) = \mu(B(t))$ and $\gamma_2(t) = \mu(D(t))$, respectively.

Lemma 2.3 *a). Let $T \in (0, \infty)$. As random elements in $(D_2[0,T), d)$*

$$(\hat{H}_n, \hat{K}_n) \to (H, K) \text{ weakly, as } n \to \infty. \tag{14}$$

b). If $\mu_n(L) \to \mu(L) < \infty$, a) is true for $T = \infty$.

Theorem 2.4 *a). Let $T \in (0, \infty)$. As random elements in $D_1[0,T)$*

$$\hat{\Lambda}_n(t) = \int_0^t \frac{d\hat{H}_n(s)}{\hat{K}_n(s) - \hat{H}_n(s-)} \to \Lambda(t) = \int_0^t \frac{dH(s)}{K(s) - H(s-)}$$

weakly.

b). If $\mu_n(L) \to \mu(L) < \infty$, a) is true for $T = \infty$.

c). As random elements in $\mathcal{M}_1$, $\hat{\lambda}_n \to \lambda$ weakly.

Theorem 2.5 *a). Let $T \in (0, \infty)$. As random elements in $D_1[0,T)$*

$$\hat{F}_n(t) = 1 - \prod_{s \le t} \left[1 - \frac{\hat{H}_n\{s\}}{\hat{R}_n(s)}\right] \to \tilde{F}(t) \equiv 1 - \prod_{s \le t} [1 - \Lambda\{s\}]$$

weakly.

b). If $\mu_n(L) \to \mu(L) < \infty$, a) is true for $T = \infty$.

c). As random elements in $\mathcal{M}_1$, $\hat{F}_n \to \tilde{F}$ weakly.

□

Remark: Since H and K are Poisson processes, with probability 1 the orbit of the limit processes Λ and $\tilde{F}$ are step functions. The condition $\mu_n(L) \to \mu(L) < \infty$ implies that $H(\infty) = \xi(L) < \infty$ a.s., hence with probability 1 the limit processes Λ and $\tilde{F}$ have only finite jumps.

REFERENCES

Ethier, S. N. and Kurtz, T. G. (1986). Markov Processes. John Wiley & Sons, New York.

Kallenberg, O. (1976). Random Measures. Academic Press, London.

Prohorov, Yu. V. (1965). Convergence of random processes and limit theorems in probability theorem. Theor. Probab. Appl. **1** 157-241.

Wellner, J. A. (1985). A heavy censoring limit theorem for the product limit estimator. Ann. Statist. **13** 150-162.

Woodroofe, M. (1985). *Estimating a distribution function with truncated data*, Ann. Statist. 13, 163-1177.