Submitted to the *Journal of Nonparametric Statistics*
Vol. 00, No. 00, Month 20XX, 1–6

Supplemental File of 'Truncation data analysis for the under-reporting probability in COVID-19 pandemic'

Wei Liang[a], Hongsheng Dai[b]* and Marialuisa Restaino[c]

[a] *Xiamen University, China;* [b] *University of Essex, UK;* [c] *University of Salerno, Italy*

(*Submitted October 2020*)

1. Notation and Assumption

Suppose there are K populations. For each $k = 1, 2, \cdots, K$, denote $(X^*_{i,k}, Y^*_{i,k}, C^*_{i,k})$, $i = 1,\ 2,\ \cdots$ as the continuous random variables from the kth population such that $X^*_{i,k}$ is independent of $(Y^*_{i,k}, C^*_{i,k})$, and $\mathbf{P}(C^*_{i,k} > Y^*_{i,k}) = 1$. The main survival time of interest is $X^*_{i,k}$, and $(Y^*_{i,k},\ C^*_{i,k})$ is left truncation time and right censoring time, respectively. This means we can only observe

$$(Y^*_{i,k},\ \min\{X^*_{i,k},\ C^*_{i,k}\},\ \delta^*_{i,k} = I_{[X^*_{i,k} \le C^*_{i,k}]})$$

given $X^*_{i,k} \ge Y^*_{i,k}$. Denote the observed biased sample for the kth population as

$$(Y_{1,k},\ \tilde{X}_{1,k},\ \delta_{1,k}),\ (Y_{2,k},\ \tilde{X}_{2,k},\ \delta_{2,k}),\ \cdots,\ (Y_{n_k,k},\ \tilde{X}_{n_k,k},\ \delta_{n_k,k}),$$

where $\tilde{X}_{i,k} = \min\{X_{i,k},\ C_{i,k}\}$ and $\delta_{i,k} = I[X_{i,k} \le C_{i,k}]$. Let $n = \sum_{k=1}^{K} n_k$.

Suppose $X^*_{i,k}$ follows the distribution $F_k(\cdot)$ with survival function $S_k(\cdot)$, and $(Y^*_{i,k},\ C^*_{i,k})$ follows the bivariate distribution $R_k(\cdot,\ \cdot)$. Let $G_k(\cdot)$ be the distribution of $Y_{i,k}*$, thus $G_k(\cdot) = R_k(\cdot,\ \infty)$. For any cumulative distribution F, denote

$$a_F = \inf\{x :\ F(x) > 0\}, \text{ and } \quad b_F = \sup\{x :\ F(x) < 1\}.$$

We need to assume $a_{G_k} < a_{F_k}$ and $b_{G_k} < b_{F_k}$ (the same as Condition 2.2), i.e. the minimum truncation time must be less than the minimum lifetime, and the maximum truncation time must be less than the maximum lifetime. This condition is a technical assumption for avoiding nonidentifiability problems regarding the estimation of S_k and G_k. The validity of this condition follows immediately by reducing the original problem to an estimation problem for a conditional distribution of G_k (Wang 1991) (equivalent to trimming a small amount of truncation data).

*Corresponding author. Email: hdaia@essex.ac.uk

2. NPMLE of S and G with Y^* and C^* correlated

For notational simplicity, we omit the subscript k in this section. From Wang (1991), the NPMLE of $S(\cdot)$ and $G(\cdot)$ is

$$\begin{aligned} \hat{S}(t) &= \prod_{\tilde{X}_{(i)} \le t} \left(1 - \frac{\#\{j : \tilde{X}_j = \tilde{X}_{(i)}\}}{\#\{j : Y_j \le \tilde{X}_{(i)} \le \tilde{X}_j\}} \right), \\ \hat{G}(t;\, \hat{S}) &= \left(\sum_j \frac{1}{\hat{S}(Y_j)} \right)^{-1} \sum_i \frac{1}{\hat{S}(Y_i)} I_{\{Y_i \le t\}}. \end{aligned} \tag{1}$$

According to the counting process, $\hat{S}$ can also be written as the product limit estimator

$$\hat{S}(t) = \prod_{s \in (0,t]} \left[1 - \mathrm{d}\hat{\Lambda}(s) \right] = \prod_{s \in (0,t]} \left[1 - \frac{\mathrm{d}N(s)}{\bar{H}(s)} \right],$$

where $N(t)$ is a counting process related to $\tilde{X}_i$, $\mathrm{d}N(t) = \sum_{i=1}^n I[t \le \tilde{X}_i < t + \mathrm{d}t,\, \delta_i = 1]$, and $\bar{H}(t) = \sum_{i=1}^n I[\tilde{X}_i \ge t > Y_i]$.

Define the truncation probability $\alpha = \mathbf{P}(X^* \ge Y^*)$, and

$$\begin{aligned} K(t) &= \mathbf{P}\left(\tilde{X} \le t,\, \delta = 1 \right) = \frac{1}{\alpha} \int_0^t \left(R(u,\, \infty) - R(u,\, u) \right) \mathrm{d}F(u), \\ L(t) &= \mathbf{P}\left(\tilde{X} \ge t \ge Y \right) = \frac{1}{\alpha} S(t) \left(R(t,\, \infty) - R(t,\, t) \right), \end{aligned}$$

then the Lemma 4.1 in Wang (1991) shows the asymptotic expansion of $\hat{S}$

$$\hat{S}(t) - S(t) = \frac{1}{n} \sum_{i=1}^n \xi(Y_i,\, \tilde{X}_i,\, \delta_i,\, t) + o_p(n^{-1/2}), \tag{2}$$

where

$$\xi(Y_i,\, \tilde{X}_i,\, \delta_i,\, t) = -S(t) \left[\frac{I_{\{\tilde{X}_i \le t,\, \delta_i = 1\}}}{L(t)} + \int_0^t \frac{I_{\{\tilde{X}_i \le u,\, \delta_i = 1\}}}{L^2(u)} \mathrm{d}L(u) - \int_0^t \frac{I_{\{Y_i \le u \le \tilde{X}_i\}}}{L^2(u)} \mathrm{d}K(u) \right].$$

From Lemma 4.2 and Lemma 4.3 in Wang (1991), we have

$$\hat{G}(t;\, \hat{S}) - G(t) = \frac{\alpha}{n} \sum_{i=1}^n \left(\psi(Y_i,\, \tilde{X}_i,\, \delta_i,\, t) + \eta(Y_i,\, t) \right) + o_p(n^{-1/2}), \tag{3}$$

where

$$\begin{aligned} \psi(Y_i,\, \tilde{X}_i,\, \delta_i,\, t) &= \frac{1}{\alpha} \int \frac{1}{S(u)} \xi(T_i,\, Y_i,\, \delta_i,\, u)(G(u) - I_{\{u \le t\}}) \,\mathrm{d}G(u), \\ \eta(Y_i,\, t) &= \frac{1}{S(Y_i)} \left(I_{\{Y_i \le t\}} - G(t) \right). \end{aligned}$$

Furthermore, for each pair $(t_1,\, t_2)$, Lemma 4.5 in Wang (1991) shows that

$$\mathrm{Cov}\left(\psi(Y_i,\, \tilde{X}_i,\, \delta_i,\, t_1),\; \eta(Y_i,\, t_2)\right) = 0.$$

This orthogonality will lead to simplification of the asymptotic covariance structure.

3. The large sample properties for the test statistic with Y^* and C^* correlated

With the result in the previous section, we now can derive the larger sample properties for the test statistic W, when truncation variable Y^* and censorig variable C^* are correlated.

Denote $\alpha_k = \mathbf{P}(X^*_{i,k} \geq Y^*_{i,k})$ as the truncation probability for group k, $k = 1,\, 2,\, \cdots,\, K$. We still consider to test the hypotheses

$$\mathrm{H}_0 : \alpha_1 = \cdots = \alpha_K \leftrightarrow \mathrm{H}_1 : \alpha_1 \geq \cdots \geq \alpha_K$$

with at least one $\geq$ to be strictly $>$.

For $k = 1,\, 2,\, \cdots,\, K$, denote

$$\hat{\alpha}_k = \int \hat{S}_k(u)\, \mathrm{d}\hat{G}_k(u;\, \hat{S}_k).$$

Then, the same as in the paper, the test statistic can be constructed as

$$W = \sum_{k=1}^{K-1} \left(\hat{\alpha}_k - \frac{\hat{\alpha}_{k+1} + \cdots + \hat{\alpha}_K}{K-k}\right) = \sum_{k=1}^{K} c_k \int \hat{S}_k(u)\, \mathrm{d}\hat{G}_k(u;\, \hat{S}_k),$$

where $c_1 = 1$, $c_k = 1 - \sum_{i=1}^{k-1}(K-i)^{-1}$, $k = 2,\, \cdots,\, K-1$ and $c_K = -\sum_{i=1}^{K-1}(K-i)^{-1}$.

Theorem 3.1 *Suppose for $k = 1, 2, \cdots, K$, $n_k/n \to p_k \in (0,\, 1)$, and assumptions hold. Assume that S_k and G_k are continuous. Under H_0, as $n \to \infty$, we have*

$$\sqrt{n}\, W \to N\left(0,\, \sigma_W^2\right).$$

□

Proof. Under H_0, we have $\sum_{k=1}^{K} c_k \alpha_k = 0$, then

$$\begin{aligned} W &= \sum_{k=1}^{K} c_k \left(\int \hat{S}_k(u)\, \mathrm{d}\hat{G}_k(u;\, \hat{S}_k) - \int S_k(u)\, \mathrm{d}G_k(u) \right) \\ &= \sum_{k=1}^{K} c_k \left(\int \hat{S}_k(u)\, \mathrm{d}\left(\hat{G}_k(u;\, \hat{S}_k) - G_k(u)\right) + \int \left(\hat{S}_k(u) - S_k(u)\right)\, \mathrm{d}G_k(u) \right) \\ &= \sum_{k=1}^{K} c_k \Big(\int (\hat{S}_k(u) - S_k(u))\, \mathrm{d}\left(\hat{G}_k(u;\, \hat{S}_k) - G_k(u)\right) \\ &\qquad + \int S_k(u)\, \mathrm{d}\left(\hat{G}_k(u;\, \hat{S}_k) - G_k(u)\right) + \int \left(\hat{S}_k(u) - S_k(u)\right)\, \mathrm{d}G_k(u) \Big) \\ &= \sum_{k=1}^{K} c_k \left(V_{k1} + V_{k2} + V_{k3}\right). \end{aligned}$$

Using the asymptotic expansion of $\hat{G}_k(u;\, \hat{S}_k)$ in equation (3), we get

$$\begin{aligned} V_{k2} &= \int S_k(u)\, \mathrm{d}\left(\hat{G}_k(u;\, \hat{S}_k) - G_k(u)\right) \\ &= \frac{1}{n_k} \sum_{i=1}^{n_k} \int S_k(u) \left(\frac{-1}{S_k(u)} \xi_k(Y_{i,k},\, \tilde{X}_{i,k},\, \delta_{i,k},\, u)\, \mathrm{d}G_k(u) + \frac{\alpha_k}{S_k(Y_{i,k})} \left(\mathrm{d}I_{\{Y_{i,k} \le u\}} - \mathrm{d}G_k(u) \right) \right) \\ &\quad + O_p(n_k^{-1/2}) \\ &= \frac{1}{n_k} \sum_{i=1}^{n_k} \left(\alpha_k - \frac{\alpha_k^2}{S_k(Y_{i,k})} - \int \xi_k(Y_{i,k},\, \tilde{X}_{i,k},\, \delta_{i,k},\, u)\, \mathrm{d}G_k(u) \right) + O_p(n_k^{-1/2}). \end{aligned}$$

From (2), we have

$$\begin{aligned} V_{k3} &= \int \left(\hat{S}_k(u) - S_k(u)\right)\, \mathrm{d}G_k(u) \\ &= \frac{1}{n_k} \sum_{i=1}^{n_k} \int \xi_k(Y_{i,k},\, \tilde{X}_{i,k},\, \delta_{i,k},\, u)\, \mathrm{d}G_k(u) + O_p(n_k^{-1/2}). \end{aligned}$$

Together with (2) and (3),

$$V_{k1} = O_p(n_k^{-1}).$$

Hence

$$\begin{aligned} \sqrt{n} W &= \sqrt{n} \sum_{k=1}^{K} c_k \left(V_{k1} + V_{k2} + V_{k3}\right) = \sum_{k=1}^{K} c_k p_k^{-1/2} \sqrt{n_k} \left(V_{k1} + V_{k2} + V_{k3}\right) \\ &= \sum_{k=1}^{K} c_k p_k^{-1/2} \left(n_k^{-1/2} \sum_{i=1}^{n_k} \left(\alpha_k - \frac{\alpha_k^2}{S_k(Y_{i,k})} \right) + o_p(1) \right) \to N(0,\, \sigma_V^2), \end{aligned}$$

where $\sigma_V^2 = \sum_{k=1}^K c_k p_k^{-1} \sigma_k^2$,

$$\sigma_k^2 = \mathbf{E}\left(\frac{\alpha_k^2}{S_k(Y_{i,k})}\right)^2 = \alpha_k^4 \int \frac{1}{S_k^2(y)} \frac{S_k(y)}{\alpha_k} \mathbf{d}G_k(y) = \alpha_k^3 \int \frac{\mathbf{d}G_k(y)}{S_k(y)}.$$

■

4. Comparison of the two different estimates of $\hat{G}$

If the assumption (truncation variable Y^* is independent of censoring variable C^*) holds, the estimate of $G(t)$ is given by

$$\hat{G}_1(t) = \prod_{s>t} \left[1 - \frac{\mathrm{d}N_G(s)}{\bar{H}(s)}\right],$$

where $N_G(t)$ is given by

$$\mathrm{d}N_G(t) = \sum_{i=1}^n I[t \leq Y_i < t + \mathrm{d}t].$$

This estimate is what we used in the main paper.

However, if the independent assumption doesn't hold, the above estimator is not consistent anymore. We need to use the estimator in (1) for G. To compare with $\hat{G}_1$ we re-denote it as $\hat{G}_2$, i.e.

$$\hat{G}_2(t) = \left(\sum_j \frac{1}{\hat{S}(Y_j)}\right)^{-1} \sum_i \frac{1}{\hat{S}(Y_i)} I_{\{Y_i \leq t\}}.$$

Estimator $\hat{G}_1$ depends on stronger condition (Y^*, C^* independent) and we expect that if the condition is satisfied then $\hat{G}_1$ will perform well. On the other hand, $\hat{G}_2$ depends on weaker condition (Y^*, C^* correlated). We therefore expect that $\hat{G}_2$ performs well in general, but when Y^* and C^* are indeed independent $\hat{G}_1$ shall have a better performance than $\hat{G}_2$ because the estimator $\hat{G}_1$ makes use of the stronger independence assumption.

In this section, we will compare the two estimates of the distribution of truncation variable $G(t)$ via numerical studies. We generate the pseudo data X_i^* and C_i^* from N(8.5, 2), and N(10, 1), respectively. The simulated data $\tilde{X}_i^* = \min\{X_i^*, C_i^*\}$, Y_i^* and $\delta_i^* = \min\{X_i^*, C_i^*\}$ satisfying $X_i^* > Y_i^*$ will be used in our study. This will give the censoring proportions about 15%. The truncation variable values Y_i^* are simulated from N(7, 1), which satisfies the independent assumption, and the truncation probability is about 0.8477. The sample sizes are chosen as 500. Figure 1 shows the simulation result of these two estimates. The bias of $\hat{G}_1(t)$ is much smaller than the bias of $\hat{G}_2(t)$. Indeed $\hat{G}_1$ performs much better than $\hat{G}_2$ when Y^* is independent of C^*.

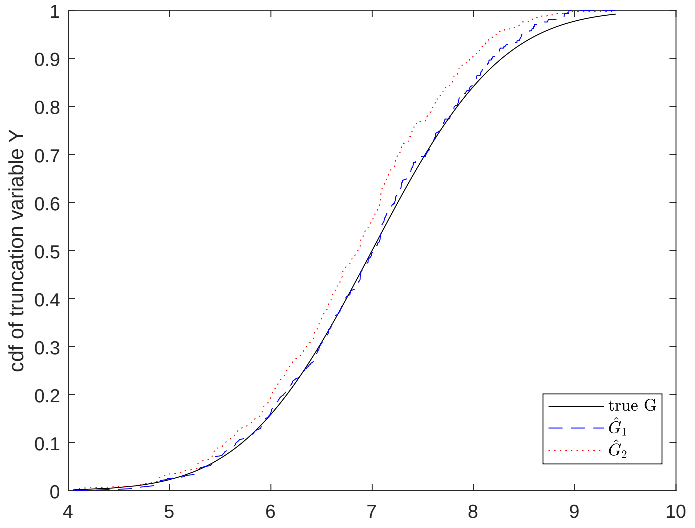

Figure 1. the comparison of estimates $\hat{G}_1$ and $\hat{G}_2$.

References

Mei-Cheng Wang (1991), Nonparametric Estimation from Cross-Sectional Survival Data, *Journal of the American Statistical Association*, 413, 130-143.