Efficient Empirical Likelihood Inference for recovery rate of COVID19 under Double-Censoring

Jie Hu[a], Wei Liang[a,*], Hongheng Dai[b], Yanchun Bao[b]

[a] *School of Mathematical Science, Xiamen University, China.*
[b] *Department of Mathematical Sciences, University of Essex, UK.*

Abstract

Doubly censored data are very common in epidemiology studies. Ignoring censorship in the analysis may lead to biased parameter estimation. In this paper, we highlight that the publicly available COVID19 data may involve high percentage of double-censoring and point out the importance of dealing with such missing information in order to achieve better forecasting results. Existing statistical methods for doubly censored data may suffer from the convergence problems of the EM algorithms or may not be good enough for small sample sizes. This paper develops a new empirical likelihood method to analyse the recovery rate of COVID19 based on a doubly censored dataset. The efficient influence function of the parameter of interest is used to define the empirical likelihood (EL) ratio. We prove that $-2\log(\text{EL-ratio})$ asymptotically follows a standard χ^2 distribution. This new method does not require any scale parameter adjustment for the log-likelihood ratio and thus does not suffer from the convergence problems involved in traditional EM-type algorithms. Finite sample simulation results show that this method provides much less biased estimate than existing methods, when censoring percentage is large. The application to COVID19 data will help researchers in other field to achieve better estimates and forecasting results.

Keywords: COVID19, Doubly censored data, Efficient influence function, Empirical likelihood

1. Introduction

Doubly censored data, with both right and left censoring, occur when time-to-event data are censored either from above or below. Doubly-censored data are very common in studies of infectious disease with incubation period. The left censoring happens when the originating date of the incubation period is not fully observed due to practical sampling factors beyond experimental control. The date of the failure event is often right-censored. A particular doubly censored data on AIDS study can be found in [1]. Another example is time from symptom onset to recovery for people who get COVID19. For COVID19 studies [2], the incubation rate and recovery rate are the key factors for us to understand the epidemiology. In particular, in the current COVID19 outbreak, better understanding of the recovery rate will help governments to take the right intervention strategy at the right time. However, many existing research for COVID19 are based on published information from government or ministry of health websites and media reports [2]. Such data have high percentage of missing information, such as high percentage

*Corresponding author
Email address: wliang@xmu.edu.cn (Wei Liang)

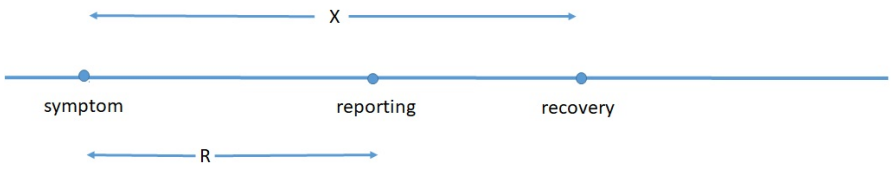

Figure 1: $\delta = 2$, right censoring and observing R only.

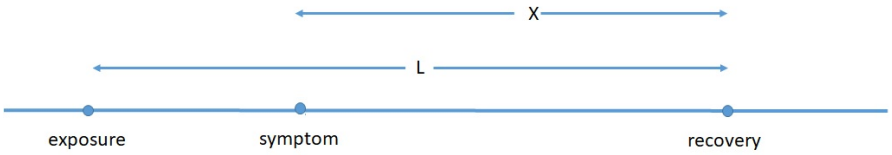

Figure 2: $\delta = 3$, left censoring and observing L only; exposure date observed and symptom date missing.

of left or right censoring. This may distort the estimation of recovery rate, which could further distort the epidemiology model forecasting, as we can see from [3] that different model parameters will give very different forecasting results.

The dataset used in [2] is from

https://github.com/mrc-ide/COVID19_CFR_submission

which has a large number of missing information on the symptom onset and on the date of recovery. Our main research interests here are to employ survival analysis techniques [4, 5] to study the recovery time, e.g. the time from symptom onset to recovery X, and to study the sensitivity of recovery rate on the epidemiology forecasting. The recovery times are clearly observed under right censoring because when the data were reported, recovery may not have happened to many patients. Therefore the right censoring time R is the time from the symptom onset date to the reporting date. See Figure 1 for scenarios when right censoring happens. Under right censoring, we will have no information about the left-censoring time L, the recorded exposure ending time to recovery. On the other hand, as we know that the symptom usually occur after exposure to the virus, when symptom onset date is missing and also reporting date is missing but the date of exposure to virus is available, we can impose the reasonable condition of $X < L$ on X, which gives the left censoring time L. So when left-censoring occurs the time L is from the date of exposure to the date of recovery. See Figure 2 for details of left-censoring. Under left censoring, we will have no information about R. When $X \in [L, R]$, we will observe X but cannot observe L and R. This is shown in Figure 3. In such cases we usually have that, $X = L$ (symptoms immediately occur after exposure; events recorded on the same day) or $L < X$ (the recorded exposure date means the ending time of an exposure period). We also have $X \leq R$ which means that recovery occurs before reporting.

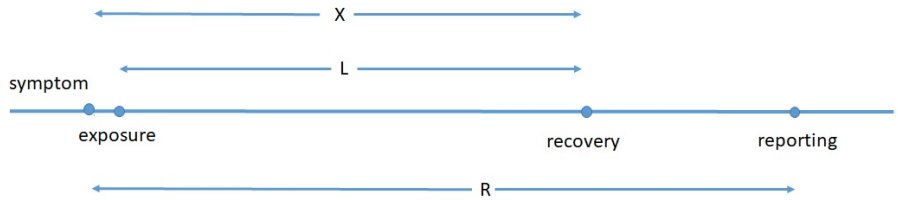

Figure 3: $\delta = 1$, X observed; exposure date means the ending date of exposure period.

In summary, under doubly censoring, the event time X is observed if $L \leq X \leq R$. We observe L in the case of left censoring with $X < L$, or observe R in the case of right censoring with $X > R$. Let $(X_i, L_i, R_i), i = 1, \cdots, n$, be n independent copies of (X, L, R), then observations under doubly censorship can be summarized as n independent pairs $(W_i, \delta_i), i = 1, \cdots, n$, where

$$W_i = \max(\min(X_i, R_i), L_i), \quad \text{and } \delta_i = \begin{cases} 1, & \text{if } L_i \leq X_i \leq R_i, \\ 2, & \text{if } X_i > R_i, \\ 3, & \text{if } X_i < L_i. \end{cases}$$

Usually, we assumed that the event time X is independent of the censoring vector (L, R).

Denote F as the cumulative distribution function of X. Suppose that we are interested in a parameter θ, defined by a functional $\theta = \theta(F)$. Many important parameters can be represented as this form, or sometimes we obtain θ via the corresponding estimating equation $g(X, \theta)$. For example, if we are interested in the expectation of a known function $m(X)$, then $\theta = \int m(x)\, \mathrm{d}F(x)$, and the corresponding estimating equation is $g(X, \theta) = m(X) - \theta$. Other examples include:

[1.] θ is the cumulative hazard function at given time t_0, i.e. $\theta = -\ln(1 - F(t_0))$, then the estimating equation is $g(X, \theta) = I_{\{X > t_0\}} - e^{-\theta}$;

[2.] θ is the mean residual life time at given time t_0, i.e.

$$\theta = E(X - t_0 | X > t_0) = \bar{F}^{-1}(t_0) \int_{t_0}^{\infty} (s - t_0)\, \mathrm{d}\, F(s),$$

where $\bar{F} = 1 - F$, then the estimating equation is $g(X, \theta) = (X - t_0 - \theta) I_{\{X \geq t_0\}}$.

To draw inference on the unknown parameter θ, a straightforward approach is to implement a distribution function estimation for F [6, 7, 8, 9]. Using the distribution function estimation, the asymptotic-normality based confidence interval for the parameter of interest θ can be constructed via the asymptotic variance estimator of the parameter estimate. But there are two main drawbacks associated with this method. First, the asymptotic variance usually takes a complicated form. Secondly, these confidence intervals based on asymptotic normal distribution do not always perform well for small samples. Other existing research about doubly-censored data may depend on specific model assumptions, such as (quantile) regression analysis [10, 11, 12] and two-sample tests [13]. In this paper, we will solve these estimation problems via empirical likelihood method [14], which is a very useful tool for constructing confidence interval for θ in nonparametric settings. Based on estimating equation $g(X, \theta)$, the original Empirical Likelihood (OEL) in [14] is defined as

$$\mathcal{R}^O(\theta) = \sup \left\{ \prod_{i=1}^{n} n p_i \Big| \sum_{i=1}^{n} p_i\, g(X_i, \theta) = 0, \sum_{i=1}^{n} p_i = 1, p_i \geq 0, i = 1, 2, \cdots, n \right\}.$$

It can be proved that

$$\mathcal{L}^O(\theta_0) = -2 \log \mathcal{R}^O(\theta_0) \to \chi^2(1), \text{ in dist.}$$

A very important work by [15] generalized the EL method to make inference for parameter defined by a general estimating equation. In general, the empirical likelihood approach has a number of advantages, such as the shape of the confidence region is determined automatically by the data. In many cases, the

log empirical likelihood ratio statistics has asymptotic χ^2 distribution, therefore the confidence interval for θ can be constructed without estimating asymptotic variance.

However, applying OEL methods to incomplete data will lead to a scaled χ^2 result. When the data is right censored, [16] utilized the Buckley-James estimator to define the estimating equation, and proved that the asymptotic distribution of the corresponding log-likelihood is a scaled χ^2 distribution. This limiting distribution can be used to construct the confidence interval for θ, if the scaled parameter is estimated. To avoid estimating the scaled parameter, [17] used the efficient influence function of the parameter under right censorship to define the log-likelihood ratio statistics and proved its asymptotic distribution is a χ^2 distribution. The confidence interval for θ based on this method is much more accurate. Under doubly censoring, [18] proposed Leveraged Bootstrap Empirical Likelihood (LBEL) by combining the EL method with the bootstrap. Since the asymptotic distribution of the log-likelihood based on LBEL method is a scaled χ^2 distribution, the scaled parameter as an adjustment coefficient needs to be estimated in practice. Besides, the LBEL method demands that the parameter of interest should be the linear functional of F.

Notice that the EL likelihood function $\prod_{i=1}^{n} p_i$ is not the real likelihood function for doubly censored data, [19] defined the likelihood function based on observations $\{(W_i,\, \delta_i)\}_{i=1}^{n}$

$$\mathcal{L}^{DC}(F) = \prod_{i=1}^{n} \Delta F(W_i)^{I_{\{\delta_i=1\}}} (\bar{F}(W_i))^{I_{\{\delta_i=2\}}} F(W_i)^{I_{\{\delta_i=3\}}}, \tag{1}$$

where DC is the abbreviation for Double Censoring, $\Delta F(t) = F(t) - F(t-)$ and $\bar{F} = 1 - F$ is the survival function. Using (1), [19] showed that this log-likelihood ratio subject to nonparametric moment constraints obeys the Wilks' phenomenon under some assumptions. This method avoids the scaled parameter, but is computationally difficult to find the nonparametric maximum likelihood. To solve this problem, [13] proposed an EM algorithm to calculate this log-likelihood ratio statistics. However, EM algorithm may suffer from the problem of convergence to a local maximum point. Different from [13], we investigate another approach in this paper. Inspired by [17], we develop the likelihood statistics defined by efficient score function for the parameter of interest θ. This method is called Efficient-EL method in our paper. Under this new approach, we demonstrate that the log empirical likelihood ratio converges to the standard χ^2 distribution without using any scale parameter adjustment, which means the confidence interval for different kinds of parameters θ can be obtained by a unified algorithm. In the mean time, it is computationally much more efficient than existing EL methods under doubly censoring.

The rest of the paper is organized as follows. The Efficient-EL inference for the differential functional parameter θ under doubly-censored data is given in Section 2, including the large sample properties and the computing algorithm. Simulation studies of the Efficient-EL and the EM-EL method proposed by [13] are provided in Section 3. We find that our approach performs much better for longer tail distributions, which usually lead to higher censoring proportions. In the mean time, the new method still performs as good as existing methods for lighter tail distributions which lead to lower censoring proportions. An application on COVID-19 study based on our proposed methodology is presented in Section 4. The paper concludes with a discussion in Section 5.

2. Efficient Empirical Likelihood Inference

Denote $G_L(t) = \mathrm{P}\{L \le t\}$ and $G_R(t) = \mathrm{P}\{R \le t\}$ as the distribution of L and R respectively. Suppose we are interested in the estimation problem for a parameter $\theta = \theta(F)$, and the corresponding estimating equation for θ is $g(X, \theta)$, that means $\mathrm{E}\, g(X, \theta) = 0$. Since X cannot be observed unless it falls in $[L, R]$, we define

$$g^{DC}(W, \delta; \theta) = I_{\{\delta=1\}} \frac{g(W, \theta)}{G_R(W) - G_L(W)} + I_{\{\delta=2\}} \frac{g(W, \theta)}{1 - G_R(W)} + I_{\{\delta=3\}} \frac{g(W, \theta)}{G_L(W)}.$$

It is easy to see that, given the distribution F, G_L, G_R, we have $\mathrm{E}\, g^{DC}(W, \delta; \theta) = 0$ which gives an estimating equation for θ. Then, the EL ratio can be defined by

$$\mathcal{R}^{DC}(\theta) = \sup \left\{ \prod_{i=1}^{n} n p_i \Big| \sum_{i=1}^{n} p_i\, g^{DC}(W_i, \delta_i; \theta) = 0, \sum_{i=1}^{n} p_i = 1, p_i \ge 0, i = 1, \cdots, n \right\}.$$

Substituting the unknown G_L, G_R with its consistent estimators will lead to a scaled asymptotic χ^2 distribution. [19] used the likelihood function (1) to solve the problem. Different from their idea, we will try to reconsider the estimating equation to overcome the scaled χ^2 asymptotic distribution problem.

2.1. The main theorems

Assume $[\alpha, \beta] \subset [0, \infty)$ be the support of F, and the following assumptions hold.

$$G_L(x) - G_R(x-) > 0 \text{ on } x \in [\alpha, \beta], \tag{A1}$$

$$F, G_L \text{ and } G_R \text{ are continuous with } G_L(\beta) = 1, G_R(\alpha) = 0. \tag{A2}$$

Define $\mathrm{BV}[\alpha, \beta] = \{h : [\alpha, \beta] \to \mathbb{R}, h \text{ is bounded and of bounded variation}\}$ and $H_F = \{h \in \mathrm{BV}[\alpha, \beta] : \int h\, \mathrm{d}F = 0\}$. The following Lemma provides the efficient influence function for θ.

Lemma 2.1. *Let $dF_t(x) = (1 + t\, h(x))\, dF(x)$ be a submodel of $F(x)$, which approaches F at direction $h \in H_F$. Assume (A1) and (A2) hold and the Hadamard derivative of $\theta(F_t)$ exists, denoted by $\dot{\theta}_0$. Then the efficient influence function for θ is*

$$\psi(w, \delta; \theta) = \ell_F (\ell^* \ell_F)^{-1} \dot{\theta}_0,$$

where ℓ_F is the score operator

$$(\ell_F h)(w, \delta) = I_{\{\delta=1\}} h(w) + I_{\{\delta=2\}} \frac{\int_{(w, \infty)} h\, dF}{1 - F(w)} + I_{\{\delta=3\}} \frac{\int_{[0, w]} h\, dF}{F(w)},$$

and ℓ^* is its corresponding adjoint operator

$$(\ell^* g)(s) = g(s, 1)\Big(G_L(s) - G_R(s-)\Big) + \int_{[0, s)} g(u, 2)\, dG_R(u) + \int_{[s, \infty)} g(u, 3)\, dG_L(u).$$

Proof. See Appendix . □

The assumptions (A1) and (A2) guarantee the operator $\ell^* \ell_F : \mathrm{BV}[\alpha, \beta] \to \mathrm{BV}[\alpha, \beta]$ is invertible. The following are some examples of derivatives $\dot{\theta}_0$ (in all of the examples we let t_0 be fixed).

[1.] For mean $\theta = \mathrm{E}\, X$, we have $\dot{\theta}_0 = x - \theta$.

[2.] For the kth moments $\theta = \mathrm{E}\, X^k$, we have $\dot{\theta}_0 = x^k - \theta$.

[3.] For cumulative distribution function $\theta = F(t_0)$, we have $\dot{\theta}_0 = \mathrm{I}_{\{x \le t_0\}} - \theta$.

[4.] For cumulative hazard function $\theta = -\ln(1 - F(t_0))$, we have

$$\dot{\theta}_0 = 1 - \mathrm{e}^{\theta}\, \mathrm{I}_{\{x > t_0\}}.$$

Since the operators ℓ^* and ℓ_F dependent on $(F,\, G_L,\, G_R)$, we should write $\psi = \psi(w,\, \delta;\, \theta,\, F,\, G_L,\, G_R)$ more precisely. Let $\xi = (F,\, G_L,\, G_R)$, the efficient influence function can be denote as $\psi(W,\, \delta;\, \theta,\, \xi)$, hence

$$\mathrm{E}\, \psi(W,\, \delta;\, \theta,\, \xi) = 0.$$

Notice that the nuisance parameter ξ is unknown, we need to estimate it firstly.

For $j = 1,\, 2,\, 3$, define

$$\hat{H}_k(t) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}_{\{W_i \le t,\, \delta_i = k\}} \quad \text{and} \quad \hat{H}(t) = \sum_{k=1}^{3} \hat{H}_k(t).$$

[8] gave the self-consistent estimators $\hat{F}$, $\hat{G}_L$, $\hat{G}_R$ of F, G_L, G_R by solving the following equations:

$$\hat{H}(t) = (1 - \hat{F}(t))\hat{G}_R(t) + \hat{F}(t)\hat{G}_L(t), \tag{2}$$

$$\hat{G}_R(t) = \int_0^t \frac{\mathrm{d}\hat{H}_2(u)}{1 - \hat{F}(u)}, \tag{3}$$

$$\hat{G}_L(t) = 1 - \int_t^{\infty} \frac{\mathrm{d}\hat{H}_3(u)}{\hat{F}(u)}. \tag{4}$$

Based on equation (2), a naive and simple iterative algorithm can be used to get $\hat{F}$, and then $\hat{G}_L$, $\hat{G}_R$ can be calculated by equations(3) and (4). In order to guarantee the asymptotic consistency and normality of $\hat{F}$, $\hat{G}_L$ and $\hat{G}_R$, we assume F, G_L and G_R satisfy conditions (A1)–(A6) in [9] throughout this paper.

Define $\hat{\xi} = (\hat{F},\, \hat{G}_L,\, \hat{G}_R)$, then the efficient influence function $\psi(W_i,\, \delta_i;\, \theta,\, \xi)$ for θ can be estimated by $\psi(W_i,\, \delta_i;\, \theta,\, \hat{\xi})$. For simplicity of notations, denote $\psi_i(\theta) = \psi(W_i,\, \delta_i;\, \theta,\, \xi)$ and $\hat{\psi}_i(\theta) = \psi(W_i,\, \delta_i;\, \theta,\, \hat{\xi})$, then the corresponding Efficient EL ratio is defined as $\hat{\mathcal{R}}^{eDC}(\theta) \doteq \hat{\mathcal{R}}^{eDC}(\theta,\, \hat{\xi}) =$

$$\sup \left\{ \prod_{i=1}^{n} n p_i \Big| \sum_{i=1}^{n} p_i\, \hat{\psi}_i(\theta) = 0,\, \sum_{i=1}^{n} p_i = 1,\, p_i \ge 0,\, i = 1,\, 2,\, \cdots,\, n \right\}. \tag{5}$$

Using Lagrangian multipliers, $p_i = n^{-1}(1 + \lambda\, \hat{\psi}_i(\theta))^{-1}$, we further have

$$\hat{\mathcal{R}}^{eDC}(\theta) = \prod_{i=1}^{n} \frac{1}{1 + \lambda\, \hat{\psi}_i(\theta)},$$

where λ is the solution of the following equation

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\hat{\psi}_i(\theta)}{1 + \lambda\, \hat{\psi}_i(\theta)} = 0$$

and the following asymptotic results.

Theorem 2.1. *Suppose the assumptions in Lemma 2.1 hold, θ_0 is the true value of the parameter of interest, and $\boldsymbol{E}\psi^2(W, \delta; \theta_0)$ exists, then we have*

$$\hat{\mathcal{L}}^{eDC}(\theta_0) \equiv -2\log\hat{\mathcal{R}}^{eDC}(\theta_0) \to \chi^2(1), \quad \textit{in dist.}$$

Proof. Under the Lemma Appendix B.1 and Lemma Appendix B.2 in Appendix, this proof is similar to the proof of Original EL and therefore it is omitted. □

Theorem 2.1 shows that the estimated log empirical likelihood ratio converges to the standard χ^2 distribution without adjustment, which means the confidence interval for different kinds of parameters θ can be obtained by an unified algorithm. Hence, a confidence region for the parameter θ with asymptotic coverage probability $1-\alpha$ can be define as

$$CI = \left\{\theta:\ 2\sum_{i=1}^{n}\log\left(1+\lambda\,\hat{\psi}_i(\theta)\right) \le \chi^2_\alpha(1)\right\}. \tag{6}$$

By recalling the definition of the efficient influence function for θ

$$\psi(w, \delta; \theta) = \ell_F(\ell^*\ell_F)^{-1}\dot{\theta}_0,$$

in the following subsection we present an algorithm for the calculation of the numerical solution of $\hat{\psi}_i$ and the confidence region CI.

2.2. Algorithm for Efficient-EL Method

Before presenting the algorithm, we need to introduce the following notations,

$$\hat{K}_1(t) = \begin{cases} n^{-1}\sum_{i=1}^{n}\left(1-\hat{F}(W_i)\right)^{-2} I_{\{\delta_i=2, W_i<t\}}, & \text{if} \quad t < B_n, \\ \hat{K}_1(B_n-), & \text{if} \quad t \ge B_n, \end{cases}$$

$$\hat{K}_2(t) = \begin{cases} n^{-1}\sum_{i=1}^{n}\hat{F}^{-2}(W_i) I_{\{\delta_i=3, W_i\ge t\}}, & \text{if} \quad t \ge A_n, \\ \hat{K}_2(A_n), & \text{if} \quad t < A_n, \end{cases}$$

where $A_n = \min\left\{W_i:\ \hat{F}(W_i) > 0\right\}$, $B_n = \max\left\{W_i:\ \hat{F}(W_i-) < 1\right\}$, and

$$K_{ij} = \frac{1}{n}\frac{\hat{K}_1(W_i\wedge W_j) + \hat{K}_2(W_i\vee W_j)}{\hat{G}_L(W_j) - \hat{G}_R(W_j-)}\, I_{\{\delta_j=1\}}.$$

For a given θ, define the least favorable direction $h_\theta(x) = (\ell^*\ell_F)^{-1}\dot{\theta}_0(x;\theta)$, then the efficient influence function is $\psi(w, \delta; \theta, \xi) = \ell_F h_\theta(x)$. Notice that only the values of $\psi(w, \delta; \theta, \xi)$ at the sample points (W_i, δ_i) are needed, therefore we can just calculate $h_\theta(W_1)$, $h_\theta(W_2)$, $\cdots$, $h_\theta(W_n)$. The following Corollary 2.1 shows a key equation for $\hat{h}_\theta(W_i)$ which will be used in the Efficient-EL algorithm.

Corollary 2.1. *The estimator $\hat{h}_\theta(W_i)$ satisfies the equation*

$$\begin{pmatrix} \dot{\theta}_0(W_1;\theta) \\ \vdots \\ \dot{\theta}_0(W_n;\theta) \end{pmatrix} = \begin{pmatrix} \Delta G_1 + K_{11} & K_{12} & \cdots & K_{1n} \\ K_{21} & \Delta G_2 + K_{22} & \cdots & K_{2n} \\ \cdots & & & \\ K_{n1} & K_{n2} & \cdots & \Delta G_n + K_{nn} \end{pmatrix} \begin{pmatrix} \hat{h}_\theta(W_1) \\ \vdots \\ \hat{h}_\theta(W_n) \end{pmatrix}, \tag{7}$$

where $\Delta G_i := \hat{G}_L(W_i) - \hat{G}_R(W_i-)$.

The Efficient-EL ratio $\hat{\mathcal{R}}^{eDC}(\theta)$ can then be calculated by the following algorithm. Hence, the confidence interval for θ, CI in (6), can be constructed using the output $\hat{\psi}_i(\theta)$ by this algorithm.

Algorithm 1 Efficient-EL Algorithm

Solving (2)-(4) to get the self-consistent estimators $\hat{F}$, $\hat{G}_L$ and $\hat{G}_R$ of F, G_L and G_R.
for $i = 1$ to n, **do**
  Calculate $\dot{\theta}_0(W_i; \theta)$ and $\Delta G_i = \hat{G}_L(W_i) - \hat{G}_R(W_i-)$,
  for $j = 1$ to n, **do**
    Calculate K_{ij}.
  end for
end for
Solve the equation (7) and get $\hat{h}_\theta(W_1)$, $\hat{h}_\theta(W_2)$, $\cdots$, $\hat{h}_\theta(W_n)$.
for $i = 1$ to n, **do**
  Calculate $\hat{\psi}_i(\theta) = \psi(W_i, \delta_i; \theta, \hat{\xi})$,
  if $A_n \le W_i < B_n$ **then**

$$\begin{aligned}\hat{\psi}_i(\theta) \quad = \quad & I_{\{\delta_i=1\}}\hat{h}_\theta(W_i) + \frac{I_{\{\delta_i=2\}}}{n(1-\hat{F}(W_i))}\sum_{k=1}^{n}\frac{I_{\{\delta_k=1,\, W_k>W_i\}}\hat{h}_\theta(W_k)}{\Delta G_k} \\ & + \frac{I_{\{\delta_i=3\}}}{n\hat{F}(W_i)}\sum_{k=1}^{n}\frac{I_{\{\delta_k=1,\, W_k\le W_i\}}\hat{h}_\theta(W_k)}{\Delta G_k},\end{aligned}$$

  else if $W_i < A_n$ **then**
    $\hat{\psi}_i(\theta) = \psi(A_n, \delta_i; \theta, \hat{\xi})$,
  else
    $\hat{\psi}_i(\theta) = \psi(B_n-, \delta_i; \theta, \hat{\xi})$.
  end if
end for
Output $\hat{\psi}_i(\theta)$.

3. Simulation Studies

In this section, we will implement simulation studies to study recovery time distribution, which is very important for the analysis of Susceptible-Exposed-Infectious-Resistant (SEIR) epidemiology model. SEIR model in epidemic disease studies involves four states: susceptible (S), exposed (E), infected (I), and resistant (R) via

$$\frac{\mathrm{d}\,S}{\mathrm{d}\,t} = -\frac{\beta}{N}SI, \qquad \frac{\mathrm{d}\,E}{\mathrm{d}\,t} = \frac{\beta}{N}SI - \sigma E,$$
$$\frac{\mathrm{d}\,I}{\mathrm{d}\,t} = \sigma E - \gamma I, \qquad \frac{\mathrm{d}\,R}{\mathrm{d}\,t} = \gamma I.$$

In this SEIR model, the infectious rate β controls the rate of spread which represents the probability of transmitting disease between a susceptible and an infectious individual. The incubation rate σ is the rate of latent individuals becoming infectious (average duration of incubation is $1/\sigma$). Recovery rate γ is determined by the average duration of infection. $N = S + E + I + R$ is the total population. The basic reproductive number, $R_0 = \beta/\gamma$, does not change in this model.

Here we focus on using the proposed double censoring model to estimate the recovery time, because the infection time only involves right censoring in the data and therefore they can be estimated using standard right censoring techniques [4]. Therefore, our simulation focus will be on the mean recovery time and model forecasting to illustrate the importance of recovery time estimation on the forecasting accuracy. We will also study the mean residual recovery time, which is also very important to forecast the expected additional recovery time given that the patient has not recovered at a certain time.

3.1. Simulation studies for recovery time

In this subsection, we will illustrate the performance of our method via different simulation scenarios. We denote Uniform(a, b) as the uniform distribution on $[a, b]$, Exp(λ) as the exponential distribution with mean λ and LogNormal(μ, σ^2) as the Log-Normal distribution with parameters μ and σ^2.

There are two parameters of our interests. The first is the mean of X, denoted by θ_1, and its corresponding estimating equation is $g_1(X, \theta_1) = X - \theta_1$. Note that θ_1 is the inverse of the mean recovery rate parameter γ. The second is the Mean Residual Lifetime (MRL) of X given t_0, denoted by $\theta_2(t_0)$ or MRL(t_0), and its corresponding estimating equation is $g_2(X, \theta_2) = (X - t_0 - \theta_2)I_{\{X \geq t_0\}}$. MRL stands for the remaining mean time needed for an infected patient to recover.

Based on the simulated data, we use all complete data X_i to construct the benchmark confidence interval, named as complete data EL (or complete-EL) result. We will compare the Efficient-EL confidence interval proposed in the previous section and EM-EL confidence interval given in [13], with the benchmark complete-EL results.

3.1.1. Simulation Results for Mean and Mean Residual Lifetime

Uniform$(0, 3)$ is considered as the underlying lifetime distribution F in this subsection. The left censoring time L and censoring interval length $R - L$ are uniformly distributed on interval $[c_1, c_2]$ and $[c_3, c_4]$. We set c_i and μ_i to be different values to achieve 10%, 20%, 30% left censoring proportions and 10%, 20%, 30% right censoring proportions respectively. Based on 5000 simulated data sets, we construct Efficient-EL confidence intervals, EM-EL confidence intervals and Complete-EL confidence intervals. The coverage probabilities for mean and MRL(t_0) are summarized in Table 1.

From these results, we notice that as the sample size n increases, all coverage probabilities converge to the nominal levels. When n is fixed, coverage probabilities of Efficient-EL confidence intervals and EM-EL confidence intervals decrease as the censoring proportion increases. The coverage probabilities of the confidence intervals for parameter MRL(t_0) decrease when t_0 increases. In all cases, the performance of Efficient-EL and EM-EL methods are close to that of Complete-EL method when censoring proportion is not large.

In the top half of the Table 1, Efficient-EL and EM-EL methods perform similarly. The difference among these two methods and Complete-EL method is small, especially for small censoring proportion or large sample size. However, the performance of these methods for the parameter MRL is different (see the bottom half of Table 1). The coverage probabilities of Efficient-EL confidence intervals performs better than that of EM-EL for almost all scenarios when $t_0 = 10\%$ quantile of F. Meanwhile, Efficient-EL method performs as good as EM-EL when $t_0 = 50\%$ quantile of F, for most cases.

Table 1: Coverage probabilities for Mean and MRL with 10%, 50% quantile under Uniform(0, 3) distribution. Two percentages in each column stand for left censoring proportion and right censoring proportion. Efficient-EL results with better performances than EM-EL highlighted in bold.

		Nominal Level = 0.90			Nominal Level = 0.95		
Mean		10%+10%	20%+20%	30%+30%	10%+10%	20%+20%	30%+30%
n=50	Complete-EL	0.898	0.894	0.896	0.944	0.943	0.945
	Efficient-EL	**0.897**	**0.888**	**0.877**	**0.944**	**0.935**	**0.932**
	EM-EL	0.896	0.878	0.873	0.944	0.932	0.930
n=80	Complete-EL	0.906	0.899	0.897	0.955	0.949	0.948
	Efficient-EL	0.903	**0.892**	**0.887**	**0.950**	0.942	**0.940**
	EM-EL	0.904	0.891	0.882	0.951	0.943	0.940
n=100	Complete-EL	0.909	0.896	0.903	0.953	0.948	0.954
	Efficient-EL	0.906	**0.898**	**0.899**	**0.951**	**0.946**	**0.944**
	EM-EL	0.905	0.888	0.891	0.952	0.944	0.942
MRL(t_0 = 10% quantile)		10%+10%	20%+20%	30%+30%	10%+10%	20%+20%	30%+30%
n=50	Complete-EL	0.907	0.898	0.906	0.954	0.943	0.953
	Efficient-EL	0.904	**0.874**	**0.854**	**0.949**	**0.929**	**0.910**
	EM-EL	0.898	0.851	0.831	0.947	0.914	0.894
n=80	Complete-EL	0.895	0.894	0.892	0.949	0.948	0.945
	Efficient-EL	**0.892**	**0.880**	**0.866**	**0.941**	**0.931**	**0.921**
	EM-EL	0.886	0.859	0.835	0.936	0.921	0.907
n=100	Complete-EL	0.896	0.889	0.897	0.949	0.948	0.947
	Efficient-EL	0.896	**0.884**	**0.868**	**0.949**	**0.936**	**0.923**
	EM-EL	0.898	0.859	0.838	0.946	0.925	0.907
MRL(t_0 = 50% quantile)		10%+10%	20%+20%	30%+30%	10%+10%	20%+20%	30%+30%
n=50	Complete-EL	0.897	0.891	0.896	0.945	0.943	0.950
	Efficient-EL	0.871	**0.838**	0.819	0.928	0.896	0.877
	EM-EL	0.888	0.833	0.831	0.938	0.897	0.895
n=80	Complete-EL	0.895	0.901	0.891	0.949	0.950	0.841
	Efficient-EL	0.885	**0.856**	0.844	0.935	**0.912**	0.909
	EM-EL	0.889	0.848	0.846	0.941	0.909	0.915
n=100	Complete-EL	0.893	0.901	0.897	0.949	0.948	0.947
	Efficient-EL	0.892	**0.873**	0.871	0.942	**0.928**	0.923
	EM-EL	0.894	0.857	0.864	0.945	0.921	0.929

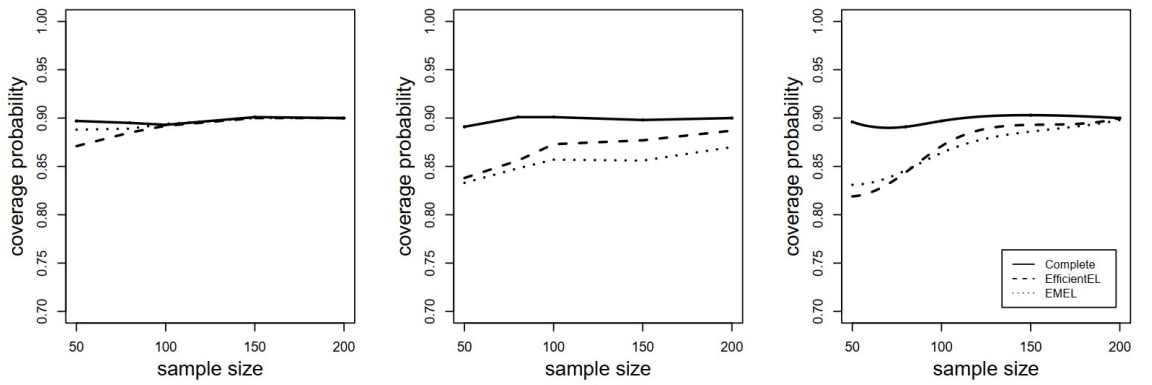

Figure 4: The coverage probabilities for MRL($t_0 = 0.5$) under Unifrom(0, 3) distribution when nominal level is 90%. The figures from left to right show the results for different censoring percentages: left plot 10% left-censoring and 10% right-censoring; middle plot 20% left-censoring and 20% right-censoring; right plot: 30% left-censoring and 30% right-censoring.

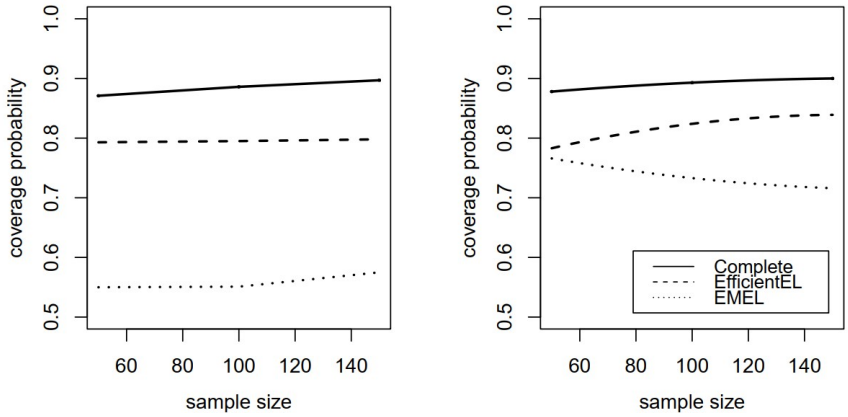

Figure 5: The coverage probabilities for MRL($t_0 = 0.3$) under Exp(1) distribution when nominal level is 90%. The left figure shows the results for 20% left and 40% right censoring proportion, while the right figure shows the result for 40% left and 20% right censoring proportion.

We also plot the results of Table 1 and draw the coverage probability curves of different methods in Figure 4. Comparing to EM-EL method, Efficient-EL method shows a much better convergence pattern, converging faster to the Complete-EL results.

3.1.2. The impact of different censoring proportions and different distributions

In this section, we investigate the impact of different censoring proportions. Here we use Exponential distribution and Log-Normal distribution as the underlying distributions and consider the complicated parameter MRL(t_0), where t_0 is the 30% quantile of the underlying distributions. For exponential distribution, we set the left censoring time L as Exp(c_1) and censoring interval length $R - L$ as Exp(c_2). For LogNorm(0, 0.64), the left censoring time L follows Exp(c_1), and censoring interval length $R - L$ follows LogNorm(c_2, 0.25). Let c_i to be different values to achieve 20% left censoring and 40% right censoring, and 40% left censoring and 20% right censoring, respectively. Based on 5000 simulated data sets, the coverage probabilities are summarized in Figure 5 and Figure 6.

We can see that higher right censoring proportion leads to lower coverage probabilities. The coverage

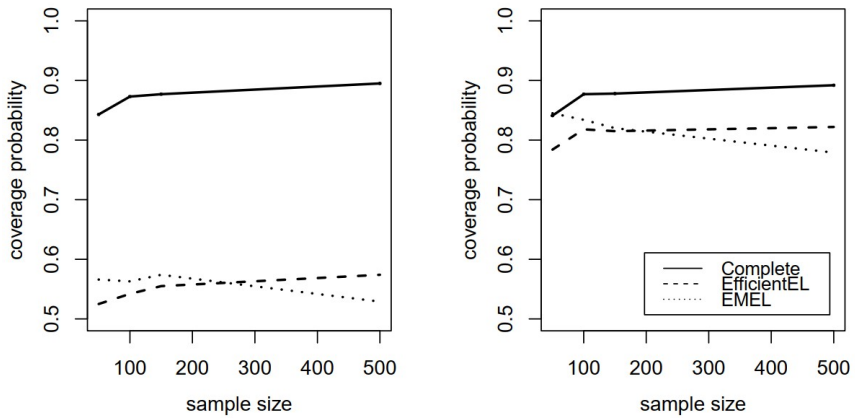

Figure 6: The coverage probabilities for MRL($t_0 = 0.3$) under LogNorm(0, 0.64) distribution when nominal level is 90%. The left figure is the coverage probability curve under 20% left and 40% right censoring proportion setting, while the right figure shows the result under 40% left and 20% right censoring proportion setting.

probabilities of confidence intervals constructed by the proposed Efficient-EL approach is much better than EM-EL methods under Exponential distribution. In Figure 5, the left plot with 20% left censoring and 40% right censoring shows that Efficient-EL has coverage probability 0.80 which is much closer to the bench mark (about 0.90), while EM-EL only has coverage probability less than 0.60. The right plot with 40% left censoring and 20% right censoring also shows Efficient-EL is better. In particular, EM-EL seems to have the problem not converging to the bench mark 0.90 as sample size increases.

Under the Log-Normal distribution, EM-EL appears to perform similarly as Efficient-EL, but EM-EL does not show a clear pattern of convergence (see Figure 6). In other words, as sample size increases the coverage probabilities of Efficient-EL based confidence intervals steadily increase, while the coverage probabilities of EM-EL seem not to have a clear increasing pattern (coverage probabilities of EM-EL may not converge to the nominal level as sample size becomes larger). Taking censoring proportion 40%+20% as a specific example, as the sample size increase from 50 to 150, the coverage probabilities of Efficient-EL increase from 0.784 to 0.815, while EM-EL decrease from 0.845 to 0.820. See Figure 6 for details.

In summary, under both exponential distribution and log-normal distribution, the new Efficient-EL approach is more reliable for highly-censored data.

3.2. SEIR model forecasting

In this subsection, we will present how the SEIR forecasting results are affected by choosing different recovery rate parameter γ, i.e. the sensitivity of γ. In our simulation we set population $N = 10^6$ and discrete time steps of size 0.1 of a simulated day. We consider the reproduction number $R_0 = 2$ and 3, then let the average duration of infection be 10, 20 and 30 days (the corresponding recovery rate γ is 1/10, 1/20, 1/30), respectively. These values are chosen according to the data analysis result in Section 4. The incubation period($= 1/\sigma$) is chosen between 2 and 10 days, which mimic the real COVID data analysis results [20]. From the summarized results presented in Figure 7, we can see that under different R_0 and σ values, the total number of infections will be highly affected by the recovery rate γ. The maximum number of infection can be different in the scale of 20,000 to 100,000 in a population of

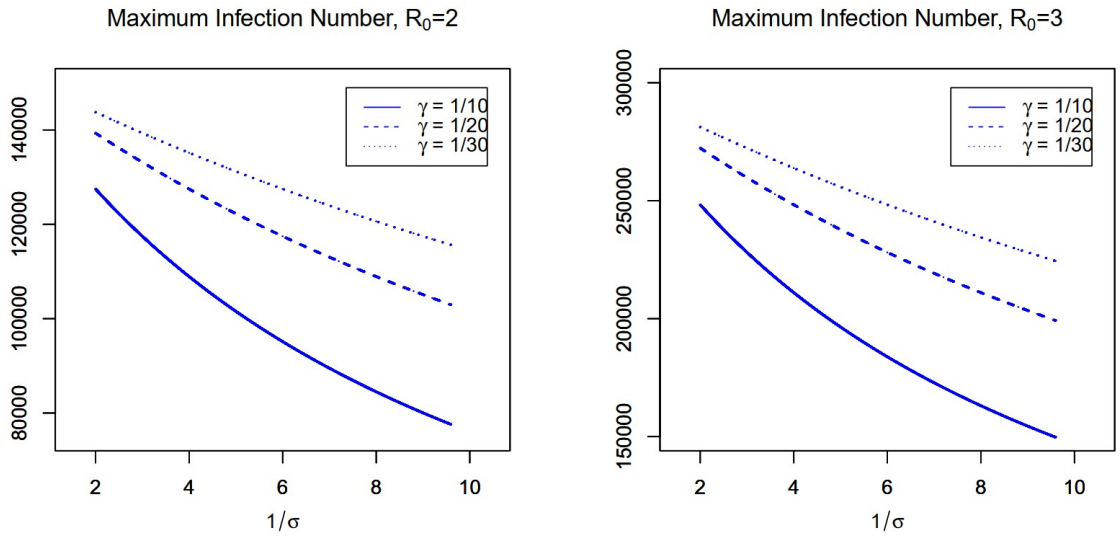

Figure 7: Maximum number of infections curves under different quarantine protocols. Three different sets of curves represent different recovery period.

1,000,000. Therefore, even if the confidence interval of recovery rate was estimated wrongly at a very small scale, the final forecasting results will be very different.

4. Analysis of COVID19 Data

4.1. Recovery time analysis

There has already been a vast literature on COVID19 research about Susceptible-Exposed-Infectious-Resistant (SEIR) epidemiology model, based on which the UK government's lock-down strategy were made [3]. Recovery time is a very important factor in such SEIR model. However publicly available data could have a large proportion of missing information, to stop us achieving a proper analysis for it. For example, the dataset from

https://github.com/mrc-ide/COVID19_CFR_submission

has a large number of missing information on the symptom onset and on the date of recovery. It actually gave a double censoring dataset for the recovery time. The event time X of interest is time length from symptom onset to recovery. The right censoring variable R is from symptom onset to the reporting date. The left censoring variable L is from the date of exposure (or the ending date of exposure period) to recovery. The total number of observations used in our analysis is $n = 547$ and the data are collected from 20th January 2020 to 28th February 2020.

Firstly, we list the censoring proportions of this dataset under different groups in Table 2. Using the Efficient-EL and EM-EL methods, the confidence intervals of recovery time for different groups can be calculated, respectively. These results are listed in Table 2. From Table 2, we can see that the elder groups have longer average recovery period, but there is no significant different between male and female. The confidence intervals based on EM-EL method seem to be shorter than that of Efficient-EL. This corresponds to the simulation results where EM-EL has worse coverage probability in most cases.

Table 2: The analysis of COVID19 data for different groups

	proportion			sample	Efficient-EL		EM-EL		
Group	left	observed	right	size	CI Lower	CI Upper	CI Lower	CI Upper	Mean
Male	0.052	0.185	0.763	323	17.370	22.153	18.482	20.827	19.842
Female	0.063	0.184	0.753	218	18.171	21.567	18.541	22.186	20.243
Age under 30	0.140	0.215	0.645	85	9.527	22.411	15.596	20.014	17.759
Age 30-50	0.082	0.212	0.707	186	17.651	21.172	17.853	21.528	19.605
Age 50-60	0.066	0.168	0.766	115	18.947	23.813	19.856	23.955	21.731
Age 60-70	0.076	0.124	0.800	83	17.902	24.627	19.680	24.786	22.041
Age over 70	0.089	0.089	0.822	68	18.951	25.614	19.935	23.970	22.173
Overall	0.059	0.170	0.771	547	18.784	20.928	18.804	20.837	20.013

4.2. SEIR model forecasting

We also carry out a simulation study similar to [3] to compare the forecasting results based on different model parameter values, in order to address the importance of parameter estimation for such forecasting analysis. We set $\sigma = 1/5.1$, according to [3]. Since SEIR model dose not include mortality, we classify death and recovered as one group, re-estimate the recovery time and get the 95% confidence interval $[18.784, 20.928]$ and mean 20.013. Hence, three different recovery periods: short duration 15 days ($\gamma = 1/15$, corresponding to results without using double censoring analysis, no right censoring, over estimation of recovery rate), medium duration 20 days ($\gamma = 1/20$, corresponding to our result based on double censoring) and long duration 25 days ($\gamma = 1/25$, corresponding to results without using double censoring analysis, no left censoring, under estimation of recovery rate) are considered in our simulation.

We also consider two different quarantine protocols: no government interventions $R_0 = 2.4$ following [3] and with mild government interventions $R_0 = 1.5$, which lead to the parameter value $\beta = R_0\gamma$ in our simulation. All of our simulation are carried out via the R package deSolve of SEIR model. The daily new cases are plotted in Figure 8, where for the curves from left to right, the dashed line means 15-day recovery period, the solid line means 20-day recovery period and the dotted line means 25-day recovery period. For both $R_0 = 2.4$ and $R_0 = 1.5$ we can see that with a shorter recovery time, the COVID19 outbreak will end much quicker. Also the mode of daily infected cases will be much smaller under the scenario of shorter recovery time.

To achieve the herd immunity proposed by the UK government requires a proportion of the UK population being immune to the virus to stop it from spreading. It is well-known that such herd immunity can be stimulated by vaccination or recovery following infection. Based our result using a sophisticated double censoring statistical model, we can see clearly that the recovery period should be much shorter than the estimated figures proposed by other existing works. With $R_0 = 1.5$, the peak of the curve with recovery rate 1/20 is will occur on day 592 (95% confidence interval $[527, 761]$), the peak with recovery rate 1/15 will occur on day 479 (95% confidence interval $[422, 609]$) and the peak with recover rate 1/25 will occur on day 705 (95% confidence interval $[621, 883]$). Therefore, with a slight over or under estimation for the recovery rate, the forecasting peak date will be different at a scale of about 110 days. This would imply that the outbreak could end about four months earlier than people expected.

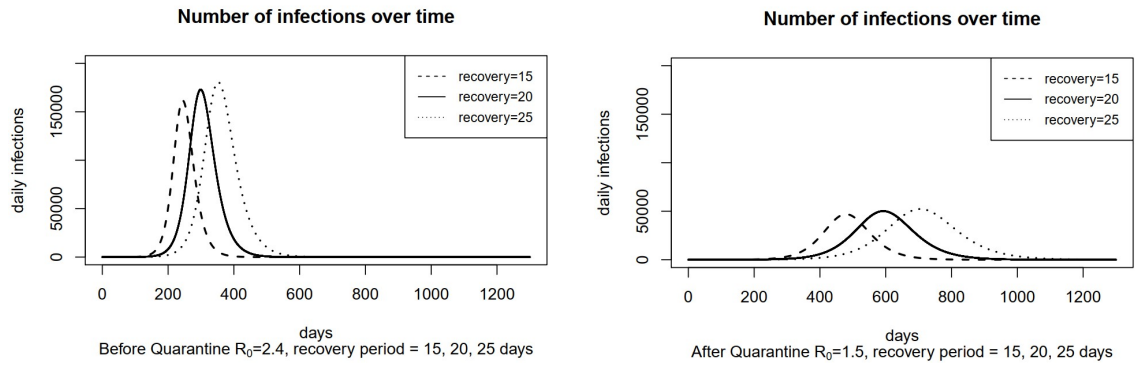

Figure 8: Increased infections curves before and after quarantine. Three different sets of curves represent different recovery period from left to right.

5. Conclusions

Through our COVID19 forecasting analysis and [3], we can see that correct estimation of SEIR model parameters may change the final forecasting results significantly, for example the peak date estimation may be different at the scale of months. For such a rapid spread disease, it will be extremely challenging to carry out a real-time monitoring task for the pandemic [21]. The data collected in real-time will certainly involve different kind of censoring. This paper highlighted the importance of dealing with the censored data and presented a efficient new statistical estimation approach. By utilizing the efficient influence function of the parameter of interest as an estimating equation, a new method of constructing EL confidence interval for doubly censored data is proposed in this paper. This new Efficient-EL method is easy to calculate since it does not need to estimate scale parameter. Simulation studies show that the new method performances better than the EM-EL method in terms of coverage probabilities.

Comparing model predictions with our estimated recovery rate parameter and existing parameter values used in other research works, we found that the peak of the epidemic predicted could be months different from each other. This could lead to wrong health policy decisions, for example taking or removing lock-down decisions at the wrong time points, which may lead to a second peak of outbreak or making the lock-down period too long to cause severe economic damage and mental health problems for more people. Our analysis highlights the importance of doing such sophisticated survival analysis will provide better estimation for the parameters in the SEIR models.

To our knowledge, this is the first work which considered using censoring techniques in survival analysis to carry out parameter estimation for COVID19 data. Most existing COVID19 research such as [22] and [3] did not address the issues of highly contaminated data due to censoring or simply use prespecified model parameters. Although only a relatively small data set is used, the methodology can be used by other researcher who have the access to larger COVID19 dataset with individual information. It will help interdisciplinary collaboration between statisticians and epidemiologists and help policy makers on public health policy making.

6. Acknowledgments

This work is supported by the National Natural Science Foundation of China, grant no.: 11701484 and the Fundamental Research Funds for the Central Universities in China, grant no.:20720190067.

Appendix A. Proof of Lemma 2.1

Proof. For any $h \in H_F$, define $\mathrm{d}F_t = (1+th)\,\mathrm{d}F$, then the likelihood of doubly censored random variable is

$$L_t(w,\,\delta) = \Delta F_t(w)^{I_{\{\delta=1\}}}(\bar{F}_t(w))^{I_{\{\delta=2\}}}F_t(w)^{I_{\{\delta=3\}}}.$$

Let P_F be the distribution of doubly censored random variable $(W,\,\delta)$, then the score operator $\ell_F : H_F \to \mathbb{L}^2(P_F)$ is

$$\begin{aligned}(\ell_F h)(w,\,\delta) &= \frac{\partial}{\partial t}\Big|_{t=0}\ln L_t(w,\,\delta)\\ &= I_{\{\delta=1\}}h(w) + I_{\{\delta=2\}}\frac{\int_{(w,\,\infty)} h\,\mathrm{d}F}{1-F(w)} + I_{\{\delta=3\}}\frac{\int_{[0,\,w]} h\,\mathrm{d}F}{F(w)}.\end{aligned}$$

By the definition of the adjoint operator $\ell^* : \mathbb{L}^2(P_F) \to H_F$, for any $h_1,\,h_2 \in H_F$,

$$< \ell_F h_1,\,\ell_F h_2 >_{P_F} = < h_1,\,\ell^*\ell_F h_2 >_F .$$

Using Fubini's theorem,

$$\begin{aligned}&< \ell_F h_1,\,\ell_F h_2 >_{P_F} = \int (\ell_F h_1\,\ell_F h_2)\,\mathrm{d}P_F\\ =\;& \int h_1(x)h_2(x)\Big(G_L(x) - G_R(x-)\Big)\mathrm{d}F + \int \frac{\int_{(r,\,\infty)} h_1\mathrm{d}F\int_{(r,\,\infty)} h_2\mathrm{d}F}{1-F(r)}\mathrm{d}G_R(r)\\ &\qquad + \int \frac{\int_{[0,\,l]} h_1\mathrm{d}F\int_{[0,\,l]} h_2\mathrm{d}F}{F(l)}\mathrm{d}G_L(r)\\ =\;& \int h_1(x)\left(h_2(x)\Big(G_L(x) - G_R(x-)\Big) + \int_{[0,x)}\frac{\int_{(r,\,\infty)} h_2\mathrm{d}F}{1-F(r)}\mathrm{d}G_R(r) + \int_{[x,\,\infty)}\frac{\int_{[0,\,l]} h_2\mathrm{d}F}{F(l)}\mathrm{d}G_L(l)\right)\end{aligned}$$

we get

$$(\ell^*\ell_F h)(x) = \Big(G_L(x) - G_R(x-)\Big)h(x) + \int\left(\int_{[x\vee s,\,\infty)}\frac{\mathrm{d}\,G_L}{F} + \int_{[0,\,x\wedge s)}\frac{\mathrm{d}\,G_R}{1-F}\right)h(s)\,\mathrm{d}\,F(s).$$

According to Lemma A.2 (i) in [19], under the assumptions, the operator $\ell^*\ell_F$ is one to one, onto and continuously invertible. By the definition of $\dot{\theta}_0$, for any $h \in H_F$,

$$\frac{\partial}{\partial t}\Big|_{t=0}\theta(F_t) = \int \dot{\theta}_0\,h\,\mathrm{d}F = < \dot{\theta}_0,\,h >_F = < \ell^*\ell_F(\ell^*\ell_F)^{-1}\dot{\theta}_0,\,h >_F = < \ell_F(\ell^*\ell_F)^{-1}\dot{\theta}_0,\,\ell_F h >_{P_F} .$$

According to the definition in [23], $\psi(w,\,\delta;\,\theta) = \ell_F(\ell^*\ell_F)^{-1}\dot{\theta}_0$ is the efficient influence function.

□

Appendix B. Lemma for Theorem 2.1

To prove the theorem 2.1, we need the following two Lemmas. Define

$$h_0 = (\ell^* \ell_F)^{-1} \dot{\theta}_0(x;\, \theta_0), \qquad \hat{h}_0 = (\ell^* \ell_{\hat{F}})^{-1} \dot{\theta}_0(x;\, \theta_0).$$

Lemma Appendix B.1. *Under the assumptions of Theorem 2.1, we have*

$$\left|\left|\hat{h}_0 - h_0\right|\right|_\infty \to 0.$$

Proof. From

$$\left|\left|\hat{h}_0 - h_0\right|\right|_\infty \le \left|\left|(\ell^* \ell_{\hat{F}})^{-1}((\ell^* \ell_F) h_0 - (\ell^* \ell_{\hat{F}}) h_0)\right|\right|_\infty$$

we just need to prove

$$\left|\left|\ell^* \ell_F - \ell_* \ell_{\hat{F}}\right|\right|_\infty = \sup_{h \in \mathrm{BV}[\alpha,\, \beta]} \left|\left|\ell^* \ell_F h(x) - \ell^* \ell_{\hat{F}} h(x)\right|\right|_\infty \to 0, \tag{B.1}$$

and to prove the result $\left|\left|(\ell^* \ell_{\hat{F}})^{-1}\right|\right|_\infty$ is bounded.

[1.] Since

$$\ell_F h(w,\, \delta) = I_{\{\delta=1\}} h(w) + I_{\{\delta=2\}} \frac{\int_{(w,\, \infty)} h \,\mathrm{d}F}{1 - F(w)} + I_{\{\delta=3\}} \frac{\int_{[0,\, w]} h \,\mathrm{d}F}{F(w)},$$

so

$$\ell_F h(w,\, 2) = \frac{\int_{(w,\, \infty)} h \,\mathrm{d}F}{1 - F(w)}, \qquad \ell_F h(w,\, 3) = \frac{\int_{[0,\, w]} h \,\mathrm{d}F}{F(w)}.$$

Hence

$$\ell^* \ell_F h(x) = h(x)\Big(G_L(x) - G_R(x-)\Big) + \int_{[0,x)} \ell_F h(r,\, 2) \,\mathrm{d}G_R(r) + \int_{[x,\, \infty)} \ell_F h(l,\, 3) \,\mathrm{d}G_L(l),$$

and

$$\begin{aligned}
&\left|\left|\ell^* \ell_F h(x) - \ell^* \ell_{\hat{F}} h(x)\right|\right|_\infty = \Delta_1 + \Delta_2 + \Delta_3, \\
&\Delta_1 = \sup_{x \in [\alpha,\, \beta]} \left| \Big((G_L(x) - G_R(x-)) - (\hat{G}_L(x) - \hat{G}_R(x-))\Big)\, h(x) \right|, \\
&\Delta_2 = \sup_{x \in [\alpha,\, \beta]} \left| \int_{[0,x)} \ell_F h(r,\, 2) \,\mathrm{d}G_R(r) - \int_{[0,x)} \ell_{\hat{F}} h(r,\, 2) \,\mathrm{d}\hat{G}_R(r) \right|, \\
&\Delta_3 = \sup_{x \in [\alpha,\, \beta]} \left| \int_{[x,\, \infty)} \ell_F h(l,\, 3) \,\mathrm{d}G_L(l) - \int_{[x,\, \infty)} \ell_{\hat{F}} h(l,\, 3) \,\mathrm{d}\hat{G}_L(l) \right|.
\end{aligned}$$

From [8], we have

$$\Delta_1 \le \sup_{x \in [\alpha,\, \beta]} \left| \Big(\hat{G}_L - G_L\Big)(x) h(x) \right| + \sup_{x \in [\alpha,\, \beta]} \left| \Big(\hat{G}_R - G_R\Big)(x-) h(x) \right| \to 0.$$

We also have

$$\begin{aligned}
\Delta_2 &= \sup_{x \in [\alpha, \beta]} \left| \int_{[0,x)} (\ell_{\hat{F}} h)(u, 2) \,\mathrm{d}\hat{G}_R(u) - \int_{[0,x)} (\ell_F h)(u, 2) \,\mathrm{d}G_R(u) \right| \\
&\le \sup_x \left| \int_{[0,x)} (\ell_{\hat{F}} h)(u, 2) \,\mathrm{d}\left(\hat{G}_R(u) - G_R(u)\right) \right| + \sup_x \int_{[0,x)} \left| (\ell_{\hat{F}} h)(u, 2) - (\ell_F h)(u, 2) \right| \mathrm{d}G_R(u) \\
&\le \sup_x \left| \int_{[0,x)} (\ell_{\hat{F}} h)(u, 2) \,\mathrm{d}\left(\hat{G}_R(u) - G_R(u)\right) \right| + \int \left| (\ell_{\hat{F}} h)(u, 2) - (\ell_F h)(u, 2) \right| \mathrm{d}G_R(u),
\end{aligned}$$

Lemma 3.1 in [9] shows that the first part of above equation is $o(1)$, and the proof of Lemma A.2(ii) in [19] shows that the second part is also $o(1)$. Therefore $\Delta_2 \to 0$. Similarly, we get $\Delta_3 \to 0$. Hence, for any $h \in \mathrm{BV}[\alpha, \beta]$,

$$\left|\left|\ell^* \ell_F h(x) - \ell^* \ell_{\hat{F}} h(x)\right|\right|_\infty \to 0.$$

[2.] Now we will prove that $\left|\left|(\ell^* \ell_{\hat{F}})^{-1}\right|\right|_\infty$ is bounded. For the convenience of proof, we denote $\mathcal{S} = \ell^* \ell_F$ and $\hat{\mathcal{S}} = \ell^* \ell_{\hat{F}}$. Next, we define

$$\hat{\mathcal{T}} = \mathcal{S}^{-1}(\mathcal{S} - \hat{\mathcal{S}}), \qquad \hat{\mathcal{U}} = \sum_{k=0}^{\infty} \hat{\mathcal{T}}^k.$$

It is easy to verify that

$$\hat{\mathcal{S}}^{-1} = \hat{\mathcal{U}}\, \mathcal{S}^{-1} \qquad \text{and} \qquad \hat{\mathcal{U}}^{-1} = I - \hat{\mathcal{T}}.$$

Since $\left|\left|\hat{\mathcal{T}}\right|\right|_\infty = \left|\left|\mathcal{S}^{-1}(\mathcal{S} - \hat{\mathcal{S}})\right|\right|_\infty \le \left|\left|\mathcal{S}^{-1}\right|\right|_\infty \left|\left|\mathcal{S} - \hat{\mathcal{S}}\right|\right|_\infty \to 0$, we have

$$\left|\left|\hat{\mathcal{U}}\right|\right|_\infty = \left|\left|\sum_{k=0}^{\infty} \hat{\mathcal{T}}^k\right|\right|_\infty \le \sum_{k=0}^{\infty} \left|\left|\hat{\mathcal{T}}\right|\right|_\infty^k = \frac{1}{1 - \left|\left|\hat{\mathcal{T}}\right|\right|_\infty} \to 1.$$

Therefore

$$\left|\left|\hat{\mathcal{S}}^{-1}\right|\right|_\infty = \left|\left|\hat{\mathcal{U}}\, \mathcal{S}^{-1}\right|\right|_\infty \le \left|\left|\hat{\mathcal{U}}\right|\right|_\infty \left|\left|\mathcal{S}^{-1}\right|\right|_\infty$$

285 is bounded.

□

Lemma Appendix B.2. *Define $\psi_{i0} = \psi_i(\theta_0)$, $\hat{\psi}_{i0} = \hat{\psi}_i(\theta_0)$ and $\sigma^2 = \boldsymbol{E}\psi^2(W, \delta; \theta_0)$. Under the assumptions of Theorem 2.1, we have*

$$\max_{1 \le i \le n} \left|\hat{\psi}_{i0}\right| = o_p(n^{1/2}), \tag{1}$$

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{\psi}_{i0}\right) \to N(0, \sigma^2), \tag{2}$$

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\psi}_{i0}^2 = \sigma^2 + o_p(1). \tag{3}$$

Proof. [1.] Since

$$\begin{aligned}\max_{1 \le i \le n} \left|\hat{\psi}_{i0}\right| &\le \max_{1 \le i \le n} \left|\hat{\psi}_{i0} - \psi_{i0}\right| + \max_{1 \le i \le n} |\psi_{i0}| = \Big(\max_{1 \le i \le n} \left|\hat{\psi}_{i0} - \psi_{i0}\right|^2\Big)^{1/2} + \max_{1 \le i \le n} |\psi_{i0}| \\ &\le n^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n} \left|\hat{\psi}_{i0} - \psi_{i0}\right|^2\right)^{1/2} + o_p(n^{1/2}),\end{aligned}$$

we only need to prove $n^{-1}\sum_{i=1}^{n} \left|\hat{\psi}_{i0} - \psi_{i0}\right|^2 = o_p(1)$. Note that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \left|\hat{\psi}_{i0} - \psi_{i0}\right|^2 &= \sum_{k=1}^{3}\left(\frac{1}{n}\sum_{i=1}^{n} \left|\psi(W_i, k; \theta_0, \hat{\xi}) - \psi(W_i, k; \theta_0, \xi)\right|^2 I_{\{\delta_i = k\}}\right) \\ &= \sum_{k=1}^{3} \int \Big(\psi(w, k; \theta_0, \hat{\xi}) - \psi(w, k; \theta_0, \xi)\Big)^2 \mathrm{d}\hat{H}_k(w) = \sum_{k=1}^{3} \Gamma_k,\end{aligned}$$

where

$$\Gamma_1 = \int \left(\psi(w,\, 1;\, \theta_0,\, \hat{\xi}) - \psi(w,\, 1;\, \theta_0,\, \xi)\right)^2 \mathrm{d}\hat{H}_1(w),$$
$$\Gamma_2 = \int \left(\psi(w,\, 2;\, \theta_0,\, \hat{\xi}) - \psi(w,\, 2;\, \theta_0,\, \xi)\right)^2 \left(1 - \hat{F}(w)\right) \mathrm{d}\hat{G}_R(w),$$
$$\Gamma_3 = \int \left(\psi(w,\, 3;\, \theta_0,\, \hat{\xi}) - \psi(w,\, 3;\, \theta_0,\, \xi)\right)^2 \hat{F}(w)\, \mathrm{d}\hat{G}_L(w).$$

Using Lemma Appendix B.1, we have

$$\begin{aligned}\Gamma_1 &= \int \left(\ell_{\hat{F}}\hat{h}_0(w,\, 1) - \ell_F h_0(w,\, 1)\right)^2 \mathrm{d}\hat{H}_1(w) = \int \left(\hat{h}_0(w) - h_0(w)\right)^2 \mathrm{d}\hat{H}_1(w) \\ &\leq \left|\left|\hat{h}_0 - h_0\right|\right|_\infty^2 \to 0.\end{aligned}$$

For the second part,

$$\begin{aligned}\Gamma_2 = &\int \left(\ell_{\hat{F}}\hat{h}_0(w,\, 2) - \ell_F h_0(w,\, 2)\right)^2 \left(1 - \hat{F}(w)\right) \mathrm{d}\left(\hat{G}_R(w) - G_R(w)\right) \\ &+ \int \left(\ell_{\hat{F}}\hat{h}_0(w,\, 2) - \ell_F h_0(w,\, 2)\right)^2 \left(1 - \hat{F}(w)\right) \mathrm{d}G_R(w).\end{aligned}$$

From Lemma 3.1 in [9], we know that the first part of above equation is $o(1)$. Since $\ell_{\hat{F}}\hat{h}_0(w,\, \delta) - \ell_F h_0(w,\, \delta) \to 0$, together with dominated convergence theorem, the second part of above equation is $o(1)$. Therefore $\Gamma_2 \to 0$, and Γ_3 converges to 0 can be proved similarly. Hence part (1) is proved.

[2.] Using the equations (2) - (4), we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \hat{\psi}_{i0} &= \sum_{k=1}^3 \int \psi(w,\, k;\, \theta_0,\, \hat{\xi})\, \mathrm{d}\hat{H}_k(w) \\ &= \int \hat{h}_0 \left(\hat{G}_R - \hat{G}_L\right) \mathrm{d}\hat{F} + \int \left(1 - \hat{G}_R\right) \hat{h}_0\, \mathrm{d}\hat{F} + \int \hat{G}_L \hat{h}_0\, \mathrm{d}\hat{F} \\ &= \int \hat{h}_0(w)\, \mathrm{d}\hat{F}(w) = \int \dot{\theta}_0\, \mathrm{d}\hat{F} = \int \dot{\theta}_0\, \mathrm{d}\left(\hat{F} - F\right).\end{aligned}$$

Due to the definition of efficient influence function and the proof of Lemma A.3 in [19], we have

$$\int \dot{\theta}_0\, \mathrm{d}(\hat{F} - F) = \frac{1}{n}\sum_{i=1}^n \ell_F h_0(W_i,\, \delta_i) + o_p(n^{-1/2}) = \frac{1}{n}\sum_{i=1}^n \psi_{i0} + o_p(n^{-1/2}).$$

Therefore

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \hat{\psi}_{i0}\right) \to N(0,\, \sigma^2).$$

[3.] Since

$$\frac{1}{n}\sum_{i=1}^n \hat{\psi}_{i0}^2 = \frac{1}{n}\sum_{i=1}^n \left(\hat{\psi}_{i0} - \psi_{i0}\right)^2 + \frac{2}{n}\sum_{i=1}^n \left(\hat{\psi}_{i0} - \psi_{i0}\right)\psi_{i0} + \frac{1}{n}\sum_{i=1}^n \psi_{i0}^2,$$

and

$$\left|\frac{1}{n}\sum_{i=1}^n \left(\hat{\psi}_{i0} - \psi_{i0}\right)\psi_{i0}\right| \leq \left|\frac{1}{n}\sum_{i=1}^n \left(\hat{\psi}_{i0} - \psi_{i0}\right)^2\right|^{1/2} \left|\frac{1}{n}\sum_{i=1}^n \psi_{i0}^2\right|^{1/2} = o_p(1),$$

we get

$$\frac{1}{n}\sum_{i=1}^n \hat{\psi}_{i0}^2 = \frac{1}{n}\sum_{i=1}^n \psi_{i0}^2 + o_p(1) = \sigma^2 + o_p(1).$$

□

Appendix C. Proof of Corollary 2.1

Proof. From the definition, the least favorable direction $h_\theta = (\ell^*\ell_F)^{-1}\dot{\theta}_0(x;\,\theta)$ satisfies the equation $(\ell^*\ell_F)h_\theta = \dot{\theta}_0(x;\,\theta)$, that is

$$\Big(G_L(x) - G_R(x-)\Big)h_\theta(x) + \int \left(\int_{[x\vee s,\,\infty)} \frac{\mathrm{d}\,G_L}{F} + \int_{[0,\,x\wedge s)} \frac{\mathrm{d}\,G_R}{1-F}\right) h_\theta(s)\,\mathrm{d}\,F(s) = \dot{\theta}_0(x;\,\theta). \tag{C.1}$$

Since

$$\mathrm{d}\,F(s) = \frac{\mathrm{d}\,H_1(s)}{G_L(s) - G_R(s-)}, \quad \mathrm{d}\,G_R(s) = \frac{\mathrm{d}\,H_2(s)}{1-F(s)}, \quad \mathrm{d}\,G_L(s) = \frac{\mathrm{d}\,H_3(s)}{F(s)}.$$

therefore

$$\begin{aligned}
&\int \left(\int_{[x\vee s,\,\infty)} \frac{\mathrm{d}\,G_L}{F} + \int_{[0,\,x\wedge s)} \frac{\mathrm{d}\,G_R}{1-F}\right) h_\theta(s)\,\mathrm{d}\,F(s) \\
&= \int \frac{1}{G_L(s) - G_R(s-)} \left(\int_{[0,\,x\wedge s)} \frac{\mathrm{d}\,H_2(u)}{(1-F(u))^2} + \int_{[x\vee s,\,\infty)} \frac{\mathrm{d}\,H_3(u)}{F^2(u)}\right) h_\theta(s)\,\mathrm{d}\,H_1(s) \\
&= \int \frac{1}{G_L(s) - G_R(s-)} \Big(K_1(x\wedge s) + K_2(x\vee s)\Big) h_\theta(s)\,\mathrm{d}\,H_1(s),
\end{aligned}$$

where

$$K_1(t) = \int_{[0,\,t)} \frac{\mathrm{d}\,H_2(u)}{(1-F(u))^2}, \quad K_2(t) = \int_{[t,\,\infty)} \frac{\mathrm{d}\,H_3(u)}{F^2(u)}.$$

Hence, equation (C.1) can be rewritten as

$$\Big(G_L(x) - G_R(x-)\Big)h_\theta(x) + \int \frac{\Big(K_1(x\wedge s) + K_2(x\vee s)\Big)}{G_L(s) - G_R(s-)} h_\theta(s)\,\mathrm{d}\,H_1(s) = \dot{\theta}_0(x;\,\theta). \tag{C.2}$$

Substitute $\hat{F}$, $\hat{G}_R$ and $\hat{G}_L$ into K_1, K_2 and (C.2), we get $\hat{K}_1(t)$, $\hat{K}_2(t)$ and

$$\begin{aligned}
\dot{\theta}_0(x;\,\theta) \;=\; &\Big(\hat{G}_L(x) - \hat{G}_R(x-)\Big)h_\theta(x) \\
&+ \frac{1}{n}\sum_{j=1}^{n} \frac{\hat{K}_1(x\wedge W_j) + \hat{K}_2(x\vee W_j)}{\hat{G}_L(W_j) - \hat{G}_R(W_j-)}\, h_\theta(W_j)\, I_{\{\delta_j=1\}}.
\end{aligned} \tag{C.3}$$

Set $x = W_i$ $(i = 1,\,2,\,\cdots,\,n)$ in equation (C.3), we get the equation (7).

□

References

[1] V. De Gruttola, S. W. Lagakos, Analysis of doubly-censored survival data, with application to aids, Biometrics (1989) 1–11.

[2] R. Verity, L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P. G. Walker, H. Fu, et al., Estimates of the severity of coronavirus disease 2019: a model-based analysis, The Lancet infectious diseases 20 (6) (2020) 669–677.

[3] N. M. Ferguson, D. Laydon, G. Nedjati-Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunubá, G. Cuomo-Dannenburg, et al., Impact of non-pharmaceutical interventions (npis) to reduce covid-19 mortality and healthcare demand (2020).

[4] E. Kenah, Non-parametric survival analysis of infectious disease data, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75 (2) (2013) 277–303.

[5] W. R. KhudaBukhsh, B. Choi, E. Kenah, G. A. Rempała, Survival dynamical systems: individual-level survival analysis from population-level epidemic models, Interface Focus 10 (1) (2020) 20190048.

[6] B. W. Turnbull, Nonparametric estimation of a survivorship function with doubly censored data, Journal of the American statistical association 69 (345) (1974) 169–173.

[7] W.-Y. Tsai, J. Crowley, A large sample study of generalized maximum likelihood estimators from incomplete data via self-consistency, The Annals of Statistics (1985) 1317–1334.

[8] M. N. Chang, G. L. Yang, Strong consistency of a nonparametric estimator of the survival function with doubly censored data, The Annals of Statistics (1987) 1536–1547.

[9] M. N. Chang, Weak convergence of a self-consistent estimator of the survival function with doubly censored data, The Annals of Statistics 18 (1) (1990) 391–404.

[10] C.-H. Zhang, X. Li, Linear regression with doubly censored data, The Annals of Statistics 24 (6) (1996) 2720–2743.

[11] J.-J. Ren, M. Gu, Regression m-estimators with doubly censored data, The Annals of Statistics 25 (6) (1997) 2638–2664.

[12] S. Ji, L. Peng, Y. Cheng, H. Lai, Quantile regression for doubly censored data, Biometrics 68 (1) (2012) 101–112.

[13] J. Shen, K. C. Yuen, C. Liu, Empirical likelihood confidence regions for one-or two-samples with doubly censored data, Computational Statistics & Data Analysis 93 (2016) 285–293.

[14] A. B. Owen, Empirical likelihood ratio confidence intervals for a single functional, Biometrika 75 (2) (1988) 237–249.

[15] J. Qin, J. Lawless, Empirical likelihood and general estimating equations, the Annals of Statistics 22 (1) (1994) 300–325.

[16] Q.-H. Wang, B.-Y. Jing, Empirical likelihood for a class of functionals of survival distribution with censored data, Annals of the Institute of Statistical Mathematics 53 (3) (2001) 517–527.

[17] S. He, W. Liang, J. Shen, G. Yang, Empirical likelihood for right censored lifetime data, Journal of the American Statistical Association 111 (514) (2016) 646–655.

[18] J.-J. Ren, Weighted empirical likelihood ratio confidence intervals for the mean with censored data, Annals of the Institute of Statistical Mathematics 53 (3) (2001) 498–516.

[19] S. A. Murphy, A. W. van der Vaart, Semiparametric likelihood ratio inference, The Annals of Statistics 25 (4) (1997) 1471–1509.

[20] W. H. Organization, Novel coronavirus (2019-ncov): situation report, 7, Technical documents (2020-01-27).

[21] P. J. Birrell, L. Wernisch, B. D. Tom, L. Held, G. O. Roberts, R. G. Pebody, D. De Angelis, Efficient real-time monitoring of an emerging influenza pandemic: How feasible?, The annals of applied statistics 14 (1) (2020) 74–93.

[22] A. J. Kucharski, T. W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R. M. Eggo, F. Sun, M. Jit, J. D. Munday, et al., Early dynamics of transmission and control of covid-19: a mathematical modelling study, The lancet infectious diseases 20 (5) (2020) 553–558.

[23] A. Tsiatis, Semiparametric Theory and Missing Data, Springer Science & Business Media, 2007.