Estimation of the incubation time distribution in the singly and doubly interval censored model

Groeneboom, Piet

DOI
10.1111/stan.12335

Publication date
2024

Document Version
Final published version

Published in
Statistica Neerlandica

Received: 6 October 2023 | Revised: 31 December 2023 | Accepted: 3 February 2024

DOI: 10.1111/stan.12335

ORIGINAL ARTICLE

WILEY

Estimation of the incubation time distribution in the singly and doubly interval censored model

Piet Groeneboom

Delft Institute of Applied Mathematics, Delft University of Technology, Delft, The Netherlands

Correspondence
Piet Groeneboom, Delft Institute of Applied Mathematics, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands.
Email: P.Groeneboom@tudelft.nl

We analyze nonparametric estimators for the distribution function of the incubation time in the singly and doubly interval censoring model. The classical approach is to use parametric families like Weibull, log-normal or gamma distributions in the estimation procedure. We propose nonparametric estimates for functions of the observations, which stay closer to the data than the classical parametric methods. We also give explicit limit distributions for discrete versions of the models and apply this to compute confidence intervals. The methods complement the analysis of the continuous model in Groeneboom (2021, 2023). R scripts for computation of the estimates are provided in Groeneboom (2020).

KEYWORDS

confidence intervals, deconvolution, double interval censoring, Fisher information, incubation time, single interval censoring, support reduction

MOS SUBJECT CLASSIFICATION

62G05; 62N01; 62-04

1 | INTRODUCTION

In the statistical analysis of the behavior of an infectious disease, one usually has to deal with events that are not directly observable. As an example, at the start of the Covid-19 pandemic, the

so-called effective reproductive number R_e played an important role ("Is it bigger or smaller than 1?"). The estimation of R_e faces the difficulty that infection events can usually not be observed and that therefore a deconvolution step is necessary, see, for example, Huisman et al. (2022).

For the estimation of the distribution of the incubation time one also faces the difficulty just mentioned. In this case one has an observation interval $[E_L, E_R]$ which is known to contain the time of infection I and a time S where the person becomes symptomatic. The incubation time is then given by $S - I$.

Following Britton and Scalia Tomba (2019) we set the left endpoint of the exposure interval $[E_L, E_R]$ equal to zero ("looking back"). Our observations then consist of the pair of the (lengths of the) exposure time E and the time of getting symptomatic S

$$(E, S),$$

where $S = I + U$, I is the infection time (also shifted for taking $E_L = 0$) and U the (length of the) incubation time (which does not have to be shifted). The times I and U are assumed independent, given E, and are not observable.

To make the distribution function F of the incubation time U identifiable, one has to make an assumption on the distribution function of the infection time. It is usually assumed that that the time till infection is uniformly distributed on the interval $[0, E]$, conditionally on the length of the exposure time E. The model is for example considered in Reich, Lessler, Cummings, and Brookmeyer (2009), Britton and Scalia Tomba (2019), Backer, Klinkenberg, and Wallinga (2020), and Groeneboom (2021). It is also possible to let the infection time have another distribution on $[0, E]$, but we have to make an assumption on this to make the distribution of the infection time identifiable. In the present paper we will assume that the distribution is uniform on $[0, E]$.

We define the (convolution) density q_F of (E, S) by

$$q_F(e, s) = e^{-1}\{F(s) - F(s - e)\} = e^{-1}\int_{u=(s-e)_+}^{s} dF(u), \qquad e > 0,\ s \geq 0. \tag{1}$$

w.r.t. μ, which is the product of the measure dF_E of the exposure time E and Lebesgue measure. The distribution function F satisfies $F(x) = 0$ for $x \leq 0$. So the underlying measure Q_F for (E, S) is defined by

$$dQ_F(e, s) = q_F(e, s)\, ds\ \ dF_E(e), \qquad e \in (0, M_2], \qquad s \geq 0, \tag{2}$$

where $M_2 < \infty$ is the upper bound of the support of E.

It seems reasonable to assume that the underlying distribution function F_0 of the incubation time is absolutely continuous, with density f_0, and that E has no mass on an interval $[0, \varepsilon)$, for some $\varepsilon > 0$. Observations of this type are shown schematically in Figure 1. Note that if $S < E$, one usually puts $S = E$, since then E is no longer relevant for the estimation.

In practice, the variables $S = I + U$ are usually rounded, taking the ceil of S, to integers (days), in which case the log likelihood for one observation becomes, conditionally on the values of E,

$$\log \int_{s=\lfloor S \rfloor}^{\lceil S \rceil} q_F(E, s)\, ds = \log \left\{ E^{-1} \int_{s=\lfloor S \rfloor}^{\lceil S \rceil} \{F(s) - F(s - E)\}\, ds \right\}, \tag{3}$$

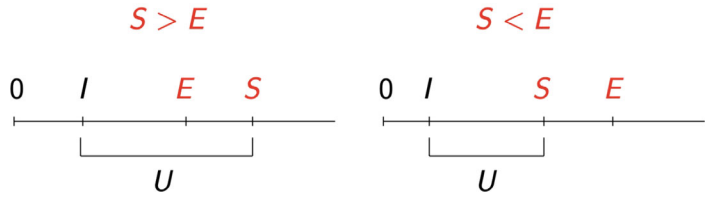

FIGURE 1 Singly interval censored data. E is the end of the exposure time, S the time of becoming symptomatic, I infection time and U the (length of the) incubation time. We only can observe E and S.

where $\lfloor S \rfloor$ and $\lceil S \rceil$ are the floor and ceil of S, respectively. Note that S itself is assumed to have a continuous distribution and that S therefore lies with probability one strictly between two consecutive integers, so we integrate over an interval of length 1. In this case, where we do not have more precise information on the values of the time of getting symptomatic, we call the model *doubly censored*: we only have an interval for the beginning of the incubation time (time of getting infected) and an interval (the 24 hr of a day) for the time of getting symptomatic. If we would have more precise information about the time of getting symptomatic, so that we can trat S as a continuous (observable) random variable, we would call the model *singly censored*. Theory for nonparametric analysis of this model is developed in Groeneboom (2023).

Usually the exposure times E are also only known as number of days, so represented by integers, as we will do in the sequel. In principle we can also consider a continuous exposure time E, but we will not do this to avoid an overcomplicated model. So instead of the parameters $F_0(S)$ and $F_0(S-E)$, we consider estimating the parameters

$$\overline{F}_0(\lceil S \rceil) = \int_{s=\lfloor S \rfloor}^{\lceil S \rceil} F_0(s)\, ds, \quad \text{and} \quad \overline{F}_0(\lceil S \rceil - E) = \int_{s=\lfloor S \rfloor}^{\lceil S \rceil} F_0(s-E)\, ds, \tag{4}$$

where $E \in \mathbb{N}$, and where we use the notation F_0 to denote the underlying distribution function.

For a sample $(E_1, S_1), \ldots, (E_n, S_n)$, where the S_i are integers, and represent the ceils in (4), we get the log likelihood

$$\ell(F) = \sum_{i=1}^{n} \log\{\overline{F}(S_i) - \overline{F}(S_i - E_i)\}, \tag{5}$$

for distribution functions F, where we assume that the E_i are also integers.

To give a specific example, we consider the following parametric model. The incubation time U has a truncated exponential distribution function F_a, defined by

$$F_a(x) = \frac{1-\exp(-x/a)}{1-\exp(-M_1/a)}\, 1_{[0,M_1]}(x) + 1_{(M_1,\infty)}(x), \qquad x \in \mathbb{R},$$

for some constants $M_1, a > 0$. There is no reason to use this distribution as a model for the incubation time, but neither is there a reason for using the Weibull, log-normal or gamma distributions, which are mostly used. The advantage of taking the exponential distribution in our example is that the integral in (3) has a simple form, in contrast with the integrals of the latter distribution functions.

Now, once we have fixed the upper bound for the incubation time M_1 (something that is needed in the nonparametric approach, but it also does not sound unrealistic to assume that such an upper bound exists), for example $M_1 = 15$, the parametric estimation of the distribution function of the incubation time boils down to the maximization of the function

$$\ell(F_a) = \sum_{i=1}^{n} \log \int_{s=\lfloor S_i \rfloor}^{\lceil S_i \rceil} \{F_a(s) - F_a(s - E_i)\}\, ds, \tag{6}$$

over $a > 0$, where

$$F_a(x) = \begin{cases} 0 & ,x < 0, \\ \left(1 - e^{-x/a}\right)/\left(1 - e^{-M_1/a}\right) & ,x \in [0, M_1], \\ 1 & ,x > M_1. \end{cases} \tag{7}$$

If $E_i < S_i$, the integrals $\int_{s=\lfloor S_i \rfloor}^{\lceil S_i \rceil} \{F_a(s) - F_a(s - E_i)\}\, ds$ are of the form

$$\begin{aligned} &\int_{s=k-1}^{k} \{F_a(s) - F_a(s - E_i)\}\, ds \\ &\quad = (1 - a\exp(-k/a)(\exp(1/a) - 1))/(1 - \exp(-M_1/a))) \\ &\qquad - (1 - a\exp(-(k-j)/a)(\exp(1/a) - 1))/(1 - \exp(-M_1/a)), \end{aligned} \tag{8}$$

(note that the E_i are assumed to be integers). If $E_i \geq S_i$ we get:

$$\begin{aligned} \int_{s=k-1}^{k} \{F_a(s) - F_a(s - E_i)\}\, ds &= \int_{s=k-1}^{k} F_a(s)\, ds \\ &= (1 - a\exp(-k/a)(\exp(1/a) - 1))/(1 - \exp(-M_1/a))). \end{aligned} \tag{9}$$

So maximization of $\ell(F_a)$ in (6) boils down to maximization of sums of logarithms of expressions of the form (8) and (9) as a function of the parameter a. We used the `R` package `nloptr` for this maximization, which is an interface to the nonlinear optimization package (Johnson, 2007) (written in `C`); the R script for our particular optimization problem can be found in Groeneboom (2020).

On the other hand, the nonparametric maximum likelihood estimator maximizes

$$\begin{aligned} \ell_2(\overline{F}) &= \sum_{i=1}^{n} \log \int_{s=\lfloor S_i \rfloor}^{\lceil S_i \rceil} \{F(s) - F(s - E_i)\}\, ds \\ &= \sum_{i=1}^{n} \log \left\{ \overline{F}(\lceil S_i \rceil) - \overline{F}(\lceil S_i \rceil - E_i) \right\}, \end{aligned} \tag{10}$$

over *all* distribution functions F, or, alternatively, all discrete distribution functions $\overline{F}$, only having jumps at the integers i. Since this is a maximization over a much wider class of functions, we cannot expect it to be as good as the parametric estimate for its own parametric model.

To get a feel for their relative performances, we show the box plots of the nonparametric and parametric estimates of the distribution function at the point 6 ("6 days") for this

particular parametric model. It is seen in Figure 2a that both estimates are well on target, but that the nonparametric estimate has a bigger variance. This is to be expected, since the nonparametric estimate does not presuppose this particular parametric model, as the parametric estimate in fact does. Figure 2b shows what happens if we take for the parametric estimate the value of $F_{\hat{a}}(6)$ instead of the integral $\int_5^6 F_{\hat{a}}(x)\,dx$, where $\hat{a}$ is the parametric estimate of the parameter a, resulting from the application of `nloptr`.

The doubly interval censored model, for which the intervals, containing the time of getting symptomatic, are longer than one day, is considered in Lauer et al. (2020). In this case there is again an interval $[E_L, E_R]$ for the infection time and an interval $[S_L, S_R]$ for the time of becoming symptomatic. One can, just as in Groeneboom (2021), shift the data in such a way that $E_L = 0$, which leaves us with three numbers: the time E ("length of Exposure time") and the times S_L and S_R, adapted for the shifting of E_L to zero. Denoting the (real) time of becoming symptomatic by S, we have that S is the sum of the the infection time I and the incubation time U. We also assume, conditionally on the exposure time E, that I and U are independent and that the time of becoming infected is uniformly distributed on the interval $[0, E]$ Typical schematic pictures of the doubly interval censored model are shown in Figure 3 for two different situations for the interval $[S_L, S_R]$, containing S.

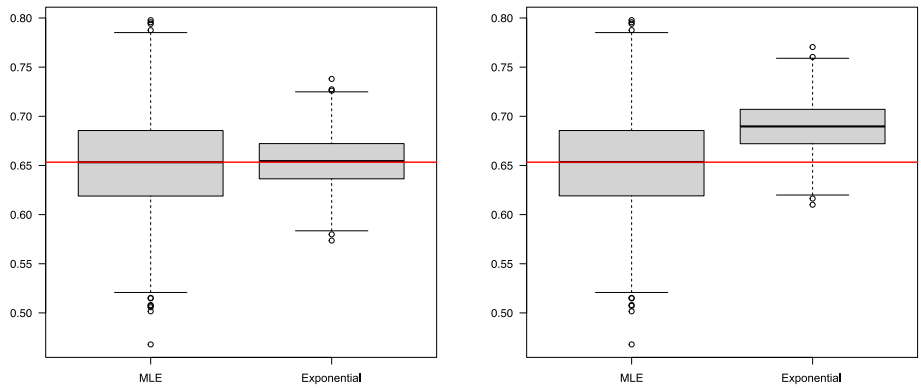

FIGURE 2 (a) Box plot of 1,000 nonparametric and parametric estimates of $\int_5^6 F_a(x)\,dx$, where F_a is the truncated exponential distribution function, defined by (7), with $a = 6$, and where the time variables of the sample are rounded to the nearest upper integer. (b) Box plot of 1,000 nonparametric estimates of $\int_5^6 F_a(x)\,dx$ and 1,000 parametric estimates of $F_a(6)$, where the time variables are rounded in the same way in the samples. In both cases, the red line segment shows the value of the real $\int_5^6 F_a(x)\,dx$.

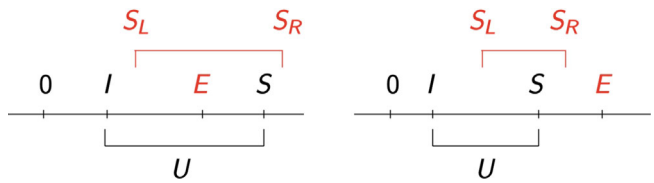

FIGURE 3 Doubly interval censored data. S the time of becoming symptomatic, I infection time and U the (length of the) incubation time. We can only observe E and the interval $[S_L, S_R]$, containing S.

In this case the log likelihood for one observation is of the form

$$\log\left[\mathbb{F}(S_R) - \mathbb{F}(S_L) - \mathbb{F}(S_R - E) + \mathbb{F}(S_L - E)\right], \tag{11}$$

neglecting parts not affecting the maximization problem, where $\mathbb{F}(u) = \int_0^u F(x)\,dx$, $u > 0$. Such an analysis if given in Lauer et al. (2020). The analysis in Lauer et al. (2020) is parametric, considering the log-normal, gamma, Weibull, and Erlang distributions as options for the incubation time distribution.

1.1 | Outline of the paper

One can restrict consideration to the estimation of parameters $\overline{F}_0(i)$ of type (4). We discusss this nonparametric model in Section 2. If one only considers a finite number of parameters of this type, one can specify the asymptotic (normal) distribution, using the Fisher information matrix The rate of convergence of the estimates is $\sqrt{n}$. The results are given in Theorems 1 and 2 of Section 2.

Computation of the nonparametric MLE is nontrivial. We discuss this in Section 3, where the support reduction is introduced for computing the MLE in both models. In Section 4 we discuss the construction of confidence intervals for both models. Proofs are given in the Appendix.

2 | THE NONPARAMETRIC MODEL

We consider the estimation the parameters $\overline{F}_0(i)$, defined by (4). The first question is whether the MLE is consistent for these parameters. The answer is affirmative, under some conditions on the distribution function F_E and F_0, as the following lemma shows. Here and the rest of the paper, we assume that the S_i and E_i are integers.

Lemma 1 (Consistency of $\hat{F}_n$ for $\overline{F}_0$). *Let $\overline{F}_0$ have strictly positive mass at all points $1, 2, \ldots, M_1$, where M_1 is a integer such that $\overline{F}_0(i) = 1$, for $i > M_1$. Moreover, let F_E have positive mass on all points $1, 2, \ldots, M_2$, where $M_2 > M_1/2$. Then $\hat{F}_n$ is a consistent estimate of $\overline{F}_0$.*

Proof. The proof parallels the proof for the continuous model in Groeneboom (2023).

The first observation is that $\hat{F}_n(i) = 1$ if $i > M_1$, since then $i > S_j - E_j$ for all pairs E_j, S_j for which $E_j > 0$, so the log likelihood becomes largest by putting $F(i) = 1$ (there is no compensating $-F(S_j - E_j)$, such that $i \leq S_j - E_j$).

Let $\ell(F)$ be defined by

$$\ell(F) = \int \log\{F(s) - F(s-e)\}\, d\mathbb{Q}_n(e,s),$$

for distribution functions F on $\mathbb{R}$, which satisfy $F(x) = 0$, $x \leq 0$, where $\mathbb{Q}_n$ is the empirical distribution of the pairs (E_i, S_i). Then we get:

$$\lim_{\varepsilon \downarrow 0} \varepsilon^{-1}\left\{\ell\left((1-\varepsilon)\hat{F}_n + \varepsilon F_0\right) - \ell\left(\hat{F}_n\right)\right\} \leq 0,$$

since $\hat{F}_n$ is the MLE. The limit exists because of the concavity of ℓ. Evaluating this limit, we get:

$$\int \frac{F_0(s) - F_0(s-e)}{\hat{F}_n(s) - \hat{F}_n(s-e)}\, d\mathbb{Q}_n(e,s) \le 1.$$

This gives for a limit point F of a subsequence $\hat{F}_{n_k}$, using Helly's compactness theorem,

$$\int \sum_{i=1}^{M_1+e} e^{-1} \frac{\{\overline{F}_0(i) - \overline{F}_0(i-e)\}^2}{F(i) - F(i-e)}\, dF_E(e) \le 1,$$

where F_E is the (discrete) distribution function of E.

We also have:

$$\int \sum_{i=1}^{M_1+e} e^{-1} \{F(i) - F(i-e)\}\, dF_E(e) = 1,$$

since $F(i) = 1$, $i > M_1$. Hence we get:

$$\int \sum_{i=1}^{M_1+e} e^{-1} \left\{ \frac{\{\overline{F}_0(i) - \overline{F}_0(i-e)\}^2}{F(i) - F(i-e)} + \{F(i) - F(i-e)\} \right\} dF_E(e) \le 2. \tag{12}$$

Let $j \in \{1, \dots, M_1\}$. Then, if $0 < j \le M_1/2$, there is an e with positive dF_E-measure, such that $j - e < 0$, and then

$$\frac{\{\overline{F}_0(j) - \overline{F}_0(j-e)\}^2}{F(j) - F(j-e)} + F(j) - F(j-e) = \frac{\{\overline{F}_0(j)\}^2}{F(j)} + F(j) > 2\overline{F}_0(j),$$

unless $F(j) = \overline{F}_0(j)$.

On the other hand, if $M_1/2 < j \le M_1$, there is an e with positive dF_E-measure, such that $j + e > M_1$, and then

$$\frac{\{\overline{F}_0(j+e) - \overline{F}_0(j)\}^2}{F(j+e) - F(j)} + F(j+e) - F(j) = \frac{\{1 - \overline{F}_0(j)\}^2}{1 - F(j)} + 1 - F(j) > 2\{1 - \overline{F}_0(j)\},$$

unless $F(j) = \overline{F}_0(j)$. This means that

$$\int \sum_{i=1}^{M_1+e} e^{-1} \left\{ \frac{\{\overline{F}_0(i) - \overline{F}_0(i-e)\}^2}{F(i) - F(i-e)} + \{F(i) - F(i-e)\} \right\} dF_E(e)$$
$$> 2 \int \sum_{i=1}^{M_1+e} e^{-1} \{\overline{F}_0(i) - \overline{F}_0(i-e)\}\, dF_E(e) = 2,$$

unless $F = F_0$. The assertion now follows from (12). ∎

We now get the following result where the intervals, containing the time of becoming symptomatic, consist of just one day.

Theorem 1. *Let F_E have support $\{1, 2, \ldots, M_2\}$, with positive mass at each i in this set. Let, given E, the infection time I have a (continuous) uniform distribution on $[0, E]$, and let the time of getting symptomatic S be the sum of the infection time I and the incubation time U, where I and U are independent, given E. Let U have an absolutely continuous distribution function F_0 on $\mathbb{R}$, such that $F_0(0) = 0$ and $F_0(M_1) = 1$, where $M_2 > M_1/2$, and let $\overline{F}_0(i)$ be defined by*

$$\overline{F}_0(i) = \int_{i-1}^{i} F_0(x)\, dx, \qquad i = 1, 2, \ldots . \tag{13}$$

Moreover, suppose $\mathcal{T} = \{0, 1, \ldots, m\}$, where $m = M_1 + 1$, and $p_0(i) \stackrel{\text{def}}{=} \overline{F}_0(i) - \overline{F}_0(i-1) > 0$, for each $i = 1, \ldots, m$, where the $p_0(i)$ satisfy

$$\sum_{i=1}^{m} p_0(i) = 1. \tag{14}$$

Furthermore, let $\mathcal{F}$ be the set of right-continuous discrete distribution functions only having jumps at the positive integers and let $\hat{F}_n \in \mathcal{F}$ maximize the log likelihood

$$\ell(F) = \sum_{i=1}^{n} \log\{F(S_i) - F(S_i - E_i)\}, \tag{15}$$

over $F \in \mathcal{F}$. Finally, let $\hat{p}_n(i) = \hat{F}_n(i) - \hat{F}_n(i-)$ be the corresponding point masses. Then:

(i)

$$n^{1/2}\big\{\big(\hat{p}_n(1), \ldots, \hat{p}_n(m-1)\big) - \big(p_0(1), \ldots, p_0(m-1)\big)\big\} \xrightarrow{D} N\big(\mathbf{0}, \boldsymbol{\Sigma}^{-1}\big), \tag{16}$$

where $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j=1,\ldots,m-1}$ is the Fisher information matrix with elements

$$\sigma_{ij} = \mathbb{E}\, \frac{\big(1_{(S-E,S]}(i) - 1_{(S-E,S]}(m)\big)\big(1_{(S-E,S]}(j) - 1_{(S-E,S]}(m)\big)}{\{\overline{F}_0(S) - \overline{F}_0(S-E)\}^2}, \tag{17}$$

for $i, j = 1, \ldots, m-1$, and where we assume that $\boldsymbol{\Sigma}$ is nonsingular.

(ii) Let the covariance matrix $\boldsymbol{\Sigma}$ be defined by (17) be nonsingular and let $\hat{F}_n$ maximize (15). Then:

$$n^{1/2}\Big\{\big(\hat{F}_n(1), \ldots, \hat{F}_n(m-1)\big) - \big(\overline{F}_0(1), \ldots, \overline{F}_0(m-1)\big)\Big\} \xrightarrow{D} N\big(\mathbf{0}, \boldsymbol{A}\boldsymbol{\Sigma}^{-1}\boldsymbol{A}^T\big), \tag{18}$$

where the matrix $\boldsymbol{A}$ has rows $\sum_{j=1}^{i} \boldsymbol{e}_j^T$, $i = 1, \ldots, m-1$, for the unit vectors $\boldsymbol{e}_j \in \mathbb{R}^{m-1}$.

Remark 1. Note that $p_0(m) = 1 - \sum_{i=1}^{m-1} p_0(i)$ and $\overline{F}_0(m) = 1$ and that we therefore have $m - 1$ free parameters, as in the case of the multinomial distribution.

Remark 2. The major difference with the conditions of theorem 4.1 in Groeneboom (2023) is that the maximum likelihood estimators have mass at points of a *a fixed finite set*. Also note that the estimation of only a fixed finite number of parameters pulls the rate of convergence from cube root n to square root n.

We can prove a similar result for the more general doubly interval censored case. This time the Fisher information matrix consists of the elements

$$\sigma_{ij} = \mathbb{E}\,\frac{\{\psi(E,S_L,S_R,i)-\psi(E,S_L,S_R,m)\}\{\psi(E,S_L,S_R,j)-\psi(E,S_L,S_R,m)\}}{\left\{\int \psi(E,S_L,S_R,t)\,d\overline{F}_0(t)\right\}^2},$$

for $i,j = 1,\ \ldots\ , m-1$, and where

$$\begin{aligned}\psi(e,s_L,s_E,t) = {} & (s_R - t)1_{\{t\in(0,s_R]\}} - (s_L - t)1_{\{t\in(0,s_L]\}} \\ & - (s_R - e - t)1_{\{t\in(0,s_R-e]\}} + (s_L - e - t)1_{\{t\in(0,s_L-e]\}},\end{aligned} \tag{19}$$

Note that

$$\int_{t\in(s_L,s_R]} \{F(t) - F(t-E)\}\,dt = \int \psi(e,s_L,s_E,t)\,dF(t),$$

As noticed in Section 3, exactly the same support reduction algorithm can be used, with the weights $w_i(j) = \psi(E_i, S_{L,i}, S_{R,i}, j)$ where ψ is defined by (19), instead of the weights $w_i(j) = 1_{(S_i-E_i,S_i]}(j)$ for the singly interval censored model (see (24)).

It is harder to formulate conditions under which the nonparametric MLE is consistent in this model. In the singly interval censored model we can use that we can identify values of $S_i - E_i$ for which $\hat{F}_n(S_i - E_i) = 0$ and values of S_i for which $\hat{F}_n(S_i) = 1$. But if we have and interval $[S_{i,L}, S_{i,R}]$ (with length larger than 1), containing the time point for getting symptomatic, this becomes more problematic. We therefore include the condition that $\hat{F}_n$ is consistent for $\overline{F}$ in our assumptions.

We get the following analogue of Theorem 1 for the doubly interval censored model.

Theorem 2. *Let F_0 and F_E be distributions with the properties defined in Theorem 1, and let the time of getting symptomatic S satisfy $S \in [S_L, S_R]$. Moreover, let $\overline{F}_0$ and p_0 satisfy the same conditions as in Theorem 1. Finally, let $\hat{F}_n$ maximize the log likelihood*

$$\ell(F) = \sum_{i=1}^{n} \log \int \psi(E, S_{L,i}, S_{R,i}, t)\,dF(t), \tag{20}$$

where ψ is defined by (19), over the set of discrete distribution functions $\mathcal{F}$, defined in Theorem 1, and let $\hat{p}_n(i) = \hat{F}_n(i) - \hat{F}_n(i-)$ be the corresponding point masses. Suppose $\hat{F}_n$ is consistent for $\overline{F}_0$. Then:

(i)

$$n^{1/2}\left\{\left(\hat{p}_n(1),\ \ldots\ ,\hat{p}_n(m-1)\right) - \left(p_0(1),\ \ldots\ ,p_0(m-1)\right)\right\} \xrightarrow{\mathcal{D}} N\left(\mathbf{0}, \boldsymbol{\Sigma}^{-1}\right), \tag{21}$$

where $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j,=1,\ldots,m-1}$ is the Fisher information matrix, assumed nonsingular, with elements

$$\sigma_{ij} = \mathbb{E}\,\frac{\{\psi(E,S_L,S_E,i)-\psi(E,S_L,S_E,m)\}\{\psi(E,S_L,S_E,j)-\psi(E,S_L,S_E,m)\}}{\left\{\int \psi(E,S_L,S_R,t)\,d\overline{F}_0(t)\right\}^2}, \tag{22}$$

for $i,j = 1,\ldots,m-1 = M_1$, *where* ψ *is defined by (19).*

(ii) Let the covariance matrix $\boldsymbol{\Sigma}$ *be defined by (17). Then*:

$$n^{1/2}\left\{\left(\hat{F}_n(1),\ldots,\hat{F}_n(m-1)\right)-\left(\overline{F}_0(1),\ldots,\overline{F}_0(m-1)\right)\right\} \xrightarrow{\mathcal{D}} N\left(\mathbf{0},\boldsymbol{A}\boldsymbol{\Sigma}^{-1}\boldsymbol{A}^T\right), \tag{23}$$

where the matrix $\boldsymbol{A}$ *has rows* $\sum_{j=1}^{i}\boldsymbol{e}_j^T$, $i = 1,\ldots,m-1$, *for the unit vectors* $\boldsymbol{e}_j \in \mathbb{R}^{m-1}$.

3 | COMPUTATION OF THE NONPARAMETRIC MAXIMUM LIKELIHOOD ESTIMATORS

In Groeneboom (2021) two methods were discussed to compute the nonparametric MLE in the singly interval censored model: the EM algorithm and the iterative convex minorant algorithm. The EM algorithm is excruciatingly slow for this model and for this reason the iterative convex minorant algorithm was used in the simulations. But in the case of doubly interval censored data it is less clear how the iterative convex minorant algorithm should be used, although we could think of ways to apply it in this situation too. However, we will turn to a third method of computing the nonparametric MLE, the *support reduction algorithm*, see Groeneboom, Jongbloed, and Wellner (2008). This method can be applied with equal ease to the two models.

We first discuss the support reduction algorithm for the singly interval censored model. The support reduction algorithm starts by specifying a grid of points $\mathcal{S} = \{1,\ldots,M_1+M_2\}$ which could be points of mass of the MLE. As an example, for the data set analyzed in Groeneboom (2021), one could take set of points $\mathcal{S} = \{1,2,\ldots,30\}$. We can also take the points S_i and $(S_i - E_i)1_{\{S_i-E_i>0\}}$ because these are the points appearing in the log likelihood.

The log likelihood, divided by n, for this set of points can be written

$$\ell(p_1,\ldots,p_M) = n^{-1}\sum_{i=1}^{n}\log\left\{\sum_{j=1}^{M}p_j 1_{\{j\in(S_i-E_i,S_i]\}}\right\},$$

where $M = M_1+M_2+1$, $p_j \geq 0$, and $\sum_{j=1}^{M}p_j = 1$, and where the subset of locations j of *strictly* positive mass p_j have to be estimated. Introducing the notation

$$w_i(j) = 1_{\{j\in(S_i-E_i,S_i]\}}, \tag{24}$$

we can write the log likelihood, divided by n, for $\{p_1,\ldots,p_M\}$ at $\{1,\ldots,M\}$

$$\ell(p_1,\ldots,p_M) = n^{-1}\sum_{i=1}^{n}\log\left\{\sum_{j=1}^{M}p_j w_i(j)\right\}. \tag{25}$$

Turning the maximization problem into a minimization problem on the cone $\mathbb{R}_+^M$, with a Lagrange term to ensure that the solution satisfies $\sum_{i=1}^{M}p_i = 1$, we get as our criterion function

$$\phi(p_1, \ldots, p_M) = -n^{-1}\sum_{i=1}^{n}\log\left\{\sum_{j=1}^{M}p_j w_i(j)\right\} + \sum_{i=1}^{M}p_i - 1. \tag{26}$$

For this function have the following lemma.

Lemma 2 (Fenchel duality conditions for minimization on a cone). *The function ϕ in (26) is minimized on $\mathbb{R}_+^M$ if and only if*

(i)

$$\frac{\partial}{\partial p_j}\phi(p_1, \ldots, p_M) \geq 0, \qquad j = 0, \ldots, M, \tag{27}$$

and

(ii)

$$\sum_{j=1}^{M}p_j\frac{\partial}{\partial p_j}\phi(p_1, \ldots, p_m) = 0. \tag{28}$$

The algorithms mentioned (EM, convex minorant algorithm and support reduction algorithm) run until the conditions of Lemma 1 are satisfied up to a certain tolerance, for which we take 10^{-10}. The EM algorithm tries to do this by simple iteration:

$$p_j' = p_j\left\{1 - \frac{\partial}{\partial p_j}\phi(p_1, \ldots, p_M)\right\}, \qquad j = 1, \ldots, M,$$

(see (7) in Groeneboom, 2021), the iterative convex minorant algorithm by introducing a quadratic approximation, parametrizing with the values of the distribution function instead of the point masses (see Groeneboom, 2021).

As in the iterative convex minorant algorithm, the support reduction algorithm employs quadratic approximation, but the parametrization uses the point masses. Expanding the first two terms of the log likelihood at a fixed vector $\boldsymbol{p}^{(0)} = (p_1^{(0)}, \ldots, p_M^{(0)})$, we get:

$$\begin{aligned}
&n^{-1}\sum_{i=1}^{n}\log\left\{\sum_{j=1}^{M}p_j w_i(j)\right\} - n^{-1}\sum_{i=1}^{n}\log\left\{\sum_{j=1}^{M}p_j^{(0)} w_i(j)\right\}\\
&\approx n^{-1}\sum_{i=1}^{M}\frac{\sum_{j=1}^{M}\left(p_j - p_j^{(0)}\right)w_i(j)}{\sum_{j=1}^{M}p_j^{(0)}w_i(j)} - \frac{1}{2}n^{-1}\sum_{i=1}^{n}\frac{\left\{\sum_{j=1}^{M}\left(p_j - p_j^{(0)}\right)w_i(j)\right\}^2}{\left\{\sum_{j=1}^{M}p_j^{(0)}w_i(j)\right\}^2}.
\end{aligned}$$

We now use iterative minimization. For fixed $m_0 < M$ and a subset $\{j_1, \ldots, j_{m_0}\} \subset \{1, \ldots, M\}$, we minimize

$$\frac{1}{2n}\sum_{i=1}^{n}\frac{\sum_{k=1}^{m_0}p_{j_k}^2 w_i(j_k)^2 + 2\sum_{k<\ell\leq m_0}p_{j_k}p_{j_\ell}w_i(j_k)w_i(j_\ell)}{\left\{\sum_{j=1}^{M}p_j^{(0)}w_i(j)\right\}^2} - \frac{2}{n}\sum_{i=1}^{n}\frac{\sum_{k=1}^{m_0}p_{j_k}w_i(j_k)}{\sum_{j=1}^{M}p_j^{(0)}w_i(j)} + \sum_{k=1}^{m_0}p_{j_k}. \tag{29}$$

as a function of $p_1, \ldots, p_{m_0}$, where, at the start of the iterations

$$m_0 = 1 \qquad \text{and} \qquad p_j^{(0)} = \frac{1}{M}, \qquad j = 1, \ldots, M,$$

and where the double sum in the first numerator vanishes if $m_0 = 1$.

Then we investigate if adding a point j_{m_0+1} not in the set $\{j_1, \ldots, j_{m_0}\}$ and minimizing (29) over $\{p_{j_1}, \ldots, p_{j_{m_0+1}}\}$, with m_0 replaced by $m_0 + 1$, leads to a smaller value of (2.1) with the extra point. This may lead to a solution with negative p_i's. In that case we remove the point k with the smallest value of $p_k < 0$ and solve the least squares minimization problem for (29) again. It can be proved that this procedure does not remove the point just added again (see, e.g., Groeneboom et al., 2008; Meyer, 2013). If this solution gives again values $p_i < 0$, we reduce the set further to a subset of $m_0 - 1$ points and solve the least squares minimization problem for (29) again with m_0 replaced by $m_0 - 1$, continuing until we find a solution with only positive p_i's.

Then we repeat the whole procedure, starting by investigating whether adding a point $j_{m_0'+1}$ not in the set $j_1, \ldots, j_{m_0'}\}$ leads to a smaller value of the criterion for the new subset $\{j_1, \ldots, j_{m_0'}\} \subset \{1, \ldots, M\}$. Continuing in this way we find a subset $\{j_1, \ldots, j_{m_0}\}$ and corresponding $p_{j_1}, \ldots, p_{j_{m_0}}$ which solves the least squares problem for all possible subsets $\{j_1, \ldots, j_{m_0}\} \subset \{1, \ldots, M\}$.

Next we change the values $p_j^{(0)}$ in the denominators of (29). Let

$$\boldsymbol{p} = (p_1, \ldots, p_M), \qquad \boldsymbol{p}^{(0)} = \left(p_1^{(0)}, \ldots, p_M^{(0)}\right),$$

where $\boldsymbol{p}$ consists of the values p_{j_k} found in the iterative least squares minimization procedure and zeroes for indices j_k not corresponding to indices of the subset $\{j_1, \ldots, j_{m_0}\}$. Using a line search procedure, for which we use Armijo's rule, we look for a convex combination

$$\boldsymbol{p}' = \alpha \boldsymbol{p} + (1 - \alpha)\boldsymbol{p}^{(0)}, \qquad \alpha \in (0, 1),$$

such that $\phi(\boldsymbol{p}') < \phi(\boldsymbol{p}^{(0)})$, where ϕ is defined by (26). Then we set $\boldsymbol{p}^{(0)} := \boldsymbol{p}'$, and repeat the iterative least squares minimization procedure, described above.

We repeat these inner and outer iterations until conditions (i) and (ii) of Lemma 2 are satisfied up to a certain tolerance, for which we took 10^{-10}.

As an example, the algorithm is applied to the data set, given in Groeneboom (2021), starting with $p_i^{(0)} = 1/M$ at the points $\{1, \ldots, M\}$, $M = 31$, and $p_1 = 1$ at $t_1 = 10$.

iteration	criterion	$\min_{j:p_j>0} \frac{\partial}{\partial p_j}\phi(\boldsymbol{p})$	$\lvert\langle \boldsymbol{p}, \phi'(\boldsymbol{p})\rangle\rvert$	$\#\{j : p_j > 0\}$
1	1.5042265478	−0.1222076701	0.0650411285	7
2	1.4607858577	−0.0245250080	0.0650411285	7
3	1.4528857033	−0.0016619636	0.0008604701	7
4	1.4523204985	−0.0000347676	0.0000174294	7
5	1.4522988585	−0.0000003963	0.0000001969	7
6	1.4522974627	−0.0000000040	0.0000000020	7
7	1.4522973319	−0.0000000000	0.0000000000	7

It is seen that after the first least squares iteration run, the algorithm has found 7 points of strictly positive mass (the points 3–9) and that this number does not change in the following outer iterations. It is also clear that the outer iteration are of (quadratic) Newton type. The end solution coincides in all 10 decimals with the result of the iterative convex minorant algorithm in Groeneboom (2021). It can be reproduced by running the R script for the support reduction algorithm in Groeneboom (2020).

The support reduction algorithm can be run in exactly the same way for the *doubly interval censored* data with more general intervals. The only change concerns the $w_j(1)$. This time $w_j(i)$ is defined by (19). The log likelihood is again given by (25), but with the new definition of the $w_j(i)$. The solution of the maximization problem on the set of points $S = \{1, \ldots, M\}$ is again characterized by Lemma 2.

4 | CONFIDENCE INTERVALS

To construct confidence intervals for the distribution function of the incubation time $\overline{F}_0$, discretized on the integers as in (4), based on the nonparametric MLE for the singly and doubly interval censoring models, we can use Theorems 1 and 2. Because the Weibull distribution is a popular tool for modeling the incubation time distribution in medical statistics, we use simulations from this distribution as our examples.

Since we have square root n convergence and asymptotic normality, we can also use bootstrap confidence intervals. We do not run into the inconsistency difficulties from which the classical nonparametric bootstrap suffers in the continuous model (see Groeneboom, 2021, 2023). Bootstrapping has the advantage that we do not have to estimate the asymptotic variances.

We start by considering asymptotic confidence intervals, using Theorems 1 and 2 for estimating the variance. If we want a 95% confidence intervals for the distribution function at a fixed point t, we can use the interval

$$[\hat{F}_n(t) - 1.96\ \hat{\sigma}_n(t)/\sqrt{n}, \hat{F}_n(t) + 1.96\ \hat{\sigma}_n(t)/\sqrt{n}], \tag{30}$$

where $\hat{F}_n$ is the nonparametric MLE and $\hat{\sigma}_n(t)$ is the square root of a diagonal element of the inverse *observed* Fisher information matrix, corresponding to the Fisher information matrix of Theorems 1 and 2.

The observed Fisher information matrix is defined by $\boldsymbol{F} = (f_{jk})$, where

$$f_{jk} = n^{-1} \sum_{i=1}^{n} \frac{\left(1_{(S_i - E_i, S_i]}(j) - 1_{(S_i - E_i, S_i]}(m)\right)\left(1_{(S_i - E_i, S_i]}(k) - 1_{(S_i - E_i, S_i]}(m)\right)}{\{\hat{F}_n(S_i) - \hat{F}_n(S_i - E_i)\}^2}, \tag{31}$$

and the j and k are points of mass of the MLE $\hat{F}_n$. If $1_1, \ldots, i_\ell$ are the indices of the points of mass, the $(\ell - 1) \times (\ell - 1)$ matrix $\boldsymbol{F}^{-1}$ is the inverse of the corresponding observed Fisher information matrix with elements $f_{jk}, j, k = i_1, \ldots, i_{\ell-1}$. The matrix $\boldsymbol{A}\boldsymbol{F}^{-1}\boldsymbol{A}^T$, where the matrix $\boldsymbol{A}$ has rows $\sum_{j=1}^{i} \boldsymbol{e}_j^T$, $i = 1, \ldots, \ell - 1$, for the unit vectors $\boldsymbol{e}_j \in \mathbb{R}^{\ell-1}$, is an estimator of the covariance matrix of $(\hat{F}_n(i_1), \ldots, \hat{F}_n(i_{\ell-1}))$. So the diagonal elements of this matrix are the estimates of the variances of $\hat{F}_n(i_j), j = 1, \ldots, \ell - 1$.

We use these diagonal elements D_{i_j} as estimates of all variances, by defining

$$D_k = \begin{cases} 0, & ,k < i_1, \\ D_{i_j} & ,i_j \le k < i_{j+1} \qquad ,1 \le j < \ell, \\ 0, & ,k \ge i_\ell. \end{cases} \tag{32}$$

The diagonal elements D_i in the extended definition (32), are used as estimates of the variances of $\hat{F}_n(1), \ldots, \hat{F}_n(M_1)$. The fit of the estimates with n times the actual variances of the $\hat{F}_n(i)$ over 1,000 samples is remarkably good for our simulation model, see Figure 4.

The resulting confidence intervals for a sample of size $n = 1{,}000$ is shown in Figure 5 at the points 3–10, where the MLE puts most of its mass. The coverage of these intervals is also shown. Here we generated 1,000 samples of size $n = 1{,}000$, and computed the fraction of times the parameters $\overline{F}_0(i) = \int_{i-1}^{i} F_0(t)\, dt$ were inside the intervals (30) at the points 3–10.

We can run a bootstrap experiment to generate confidence intervals of this type in the following way. We resample with replacement from the data (E_i, S_i) 1,000 samples of the same size n and compute for each of these bootstrap samples the MLE. This gives 1,000 bootstrap values $\hat{F}_n^*(t) - \hat{F}_n(t)$. For these bootstrap values of $\hat{F}_n^*(t) - \hat{F}_n(t)$ we compute the 0.025 and 0.975 quantiles $Q^*_{0.025}(t)$ and $Q^*_{0.975}(t)$, respectively. This gives the bootstrap 95% confidence intervals

$$[\hat{F}_n(t) - Q^*_{0.975}(t), \hat{F}_n(t) - Q^*_{0.025}(t)]. \tag{33}$$

The results are shown in Figure 6. It is seen that the results are similar to the results of the method, using the inverse observed Fisher information matrix for generating the confidence intervals.

We also simulated data for the doubly censored model. Here we took S_R discretely uniform on $\{\lceil S \rceil, \ldots, \lceil S \rceil + 3\}$ and S_L discretely uniform on $\{\lfloor S \rfloor, \ldots, \lfloor S \rfloor - 3\}$ (replacing $\lfloor S \rfloor - i$ by 0 if $\lfloor S \rfloor - i < 0$). This time the observed Fisher information matrix is defined by

$$f_{jk} = n^{-1} \sum_{i=1}^{n} \frac{\tilde{\psi}(E_i, S_{L,i}, S_{R,i}, j)\tilde{\psi}(E_i, S_{L,i}, S_{R,i}, k)}{\left\{ \int \psi(E_i, S_{L,i}, S_{R,i}, t)\, d\hat{F}_n(t) \right\}^2}, \tag{34}$$

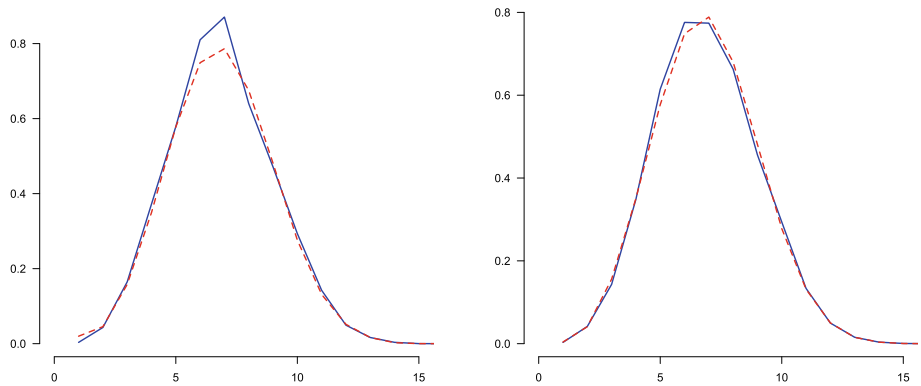

FIGURE 4 Estimates of the variances of the $\hat{F}_n(i)$ by n times the actual variances of 1,000 samples (blue, solid) and by means of the inverses of the observed Fisher information matrices over the 1,000 samples (dashed, red), (a) for sample size $n = 1{,}000$ and (b) for sample size $n = 10{,}000$. We used linear interpolation between the values at the points $1, 2, \ldots$.

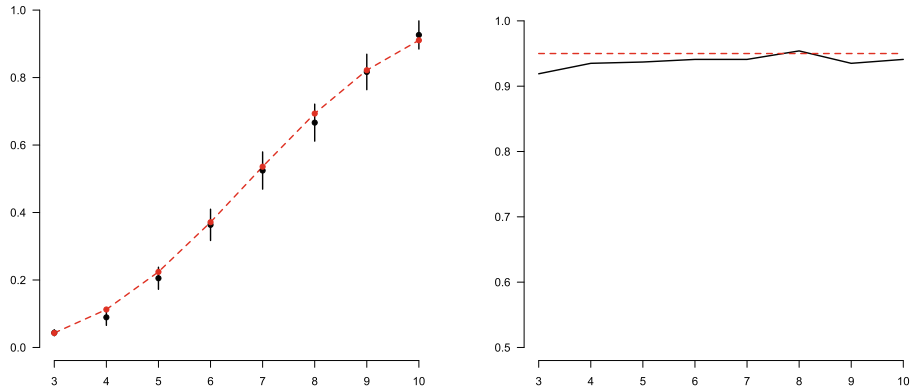

FIGURE 5 (a) 95% confidence intervals in the singly interval censored model, using (30), for the values of $\overline{F}_0(i) = \int_{i-1}^{i} F_0(x)\,dx$ (red dots and linearly interpolated dashed red curve) at the points $3, 4, \ldots, 10$ for a sample of size $n = 1{,}000$, where F_0 is the Weibull distribution function, with parameters $a = 3.035$ and $b = 0.0026$, truncated at $M_1 = 15$. The black dots are the values of $\hat{F}_n$ at these points. (b) Coverage percentages of the 95% confidence intervals at the points $3, 4, \ldots, 10$, using (30), for sample size $n = 1{,}000$.

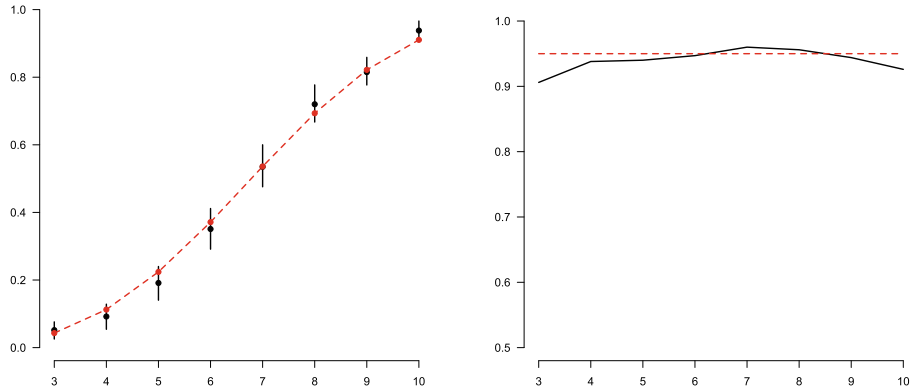

FIGURE 6 (a) 95% bootstrap confidence intervals in the singly interval censored model, using (33), for the values of $\overline{F}_0(i) = \int_{i-1}^{i} F_0(x)\,dx$ at the points $3, 4, \ldots, 10$ (red dots and linearly interpolated dashed red curve) for a sample of size $n = 1{,}000$, where F_0 is the Weibull distribution function, truncated at $M_1 = 15$. The black dots are the values of $\hat{F}_n$ at these points. (b) Coverage percentages of the bootstrap 95% confidence intervals at the points $3, 4, \ldots, 10$, using (33), for sample size $n = 1{,}000$.

where

$$\tilde{\psi}(E_i, S_{L,i}, S_{R,i}, t) = \psi(E_i, S_{L,i}, S_{R,i}, t) - \psi(E_i, S_{L,i}, S_{R,i}, m),$$

and where the j and k are points of mass of the MLE $\hat{F}_n$ and ψ is defined by (19). The diagonal elements D_{i_j} of the matrix $\boldsymbol{A}\boldsymbol{F}^{-1}\boldsymbol{A}^T$ are extended as in (32) and used as estimates of the variances of the $\hat{F}_n(i)$.

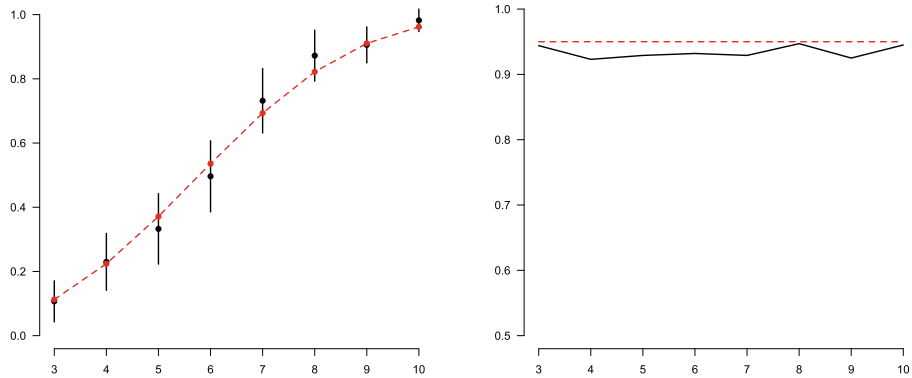

FIGURE 7 95% confidence intervals for a doubly interval censored simulation sample at the points 3, … , 10, using (30), with the variances estimated by the means of the diagonal of the inverse observed Fisher information matrices over 1,000 samples. (b) Coverage percentages for the same example over 1,000 samples, using the same estimates of the variances. Sample size is $n = 1{,}000$.

The coverages, based on the Fisher information matrix from one sample were not very good this time, but were rather satisfactory if we estimate the Fisher information matrix by the mean of these matrices over 1000 samples. See Figure 7. In practice we can take the mean over 1,000 bootstrap samples.

5 | CONCLUSION

We studied the nonparametric maximum likelihood estimator of the incubation time, which is such an important parameter in the Covid-19 pandemic. The incubation time in our model is part of the sum of the infection time I and the incubation time U. Usually the only really (partly) observable quantity is the time of becoming symptomatic, which is given by $S = I + U$. So our variable of interest U has to be pulled out of this sum by deconvolution. If S is observable, one speaks of the *singly interval censored model* (the beginning of the incubation time lies in an interval which represents the exposure time and this is the singly interval censored part). It seems most reasonable to assume that the time till infection and the incubation time have continuous distributions, and one can analyze this model under the assumption that S is exactly observable. In that case the MLE of the distribution function of the incubation time converges, under some conditions, at cube root n rate to Chernoff's distribution, see Groeneboom (2023).

However, most of the time S is not directly observable, but only an interval $[S_L, S_R]$ is available, which we know to contain S. Taking into account that the observations are usually rounded to days, one can restrict oneself to estimating the means over one day, represented by

$$\overline{F}_0(i) = \int_{i-1}^{i} F_0(t)\, dt, \tag{35}$$

if F_0 is the distribution function of the incubation time. Since this gives us a fixed bounded number of parameters, we can use classical theory, connecting maximum likelihood with the Fisher information, to derive asymptotic distribution theory for the maximum likelihood estimators of these parameters.

We applied the theory, developed in Section 2 to construct confidence intervals, either by using Theorems 1 and 2 directly, or by using a bootstrap method. In a sense, this purely nonparametric method lies at the other extreme of the parametric methods. If one wants to estimate the density, one will have to use some kind of smoothing, as was done in Groeneboom (2021), which is an intermediate method that is still nonparametric and avoids the need to choose between several parametric models.

The support reduction algorithm seems at present to be the most stable method to estimate the parameters. The computing of the MLE and the confidence intervals is implemented in Groeneboom (2020) and discussed in Section 3. R scripts for running the algorithms discussed in this paper are available at Groeneboom (2020).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in https://arxiv.org/abs/2310.04225 at https://arxiv.org/submit/5156251/preview.

ORCID

Piet Groeneboom https://orcid.org/0000-0001-8027-8114

REFERENCES

Backer, J. A., Klinkenberg, D., & Wallinga, J. (2020). Incubation period of 2019 novel coronavirus (2019-nCov) infections among travellers from Wuhan, China, 20-28 January 2020. *Euro Surveillance*, *25*, 2000062.

Britton, T., & Scalia Tomba, G. (2019). Estimation in emerging epidemics: Bases and remedies. *Journal of the Royal Society Interface*, *16*, 20180670.

Groeneboom, P. (2020). *Incubation time*. Retrieved from https://github.com/pietg/incubationtime

Groeneboom, P. (2021). Estimation of the incubation time distribution for COVID-19. *Statistica Neerlandica*, *75*, 161–179. https://doi.org/10.1111/stan.12231

Groeneboom, P. (2023). Nonparametric estimation of the incubation time distribution. https://arxiv.org/abs/2205.04399

Groeneboom, P., Jongbloed, G., & Wellner, J. A. (2008). The support reduction algorithm for computing non-parametric function estimates in mixture models. *Scandinavian Journal of Statistics*, *35*, 385–399. https://doi.org/10.1111/j.1467-9469.2007.00588.x

Huisman, J. S., Scire, J., Angst, D. C., Li, J., Neher, R. A., Maathuis, M. H., … Stadler, T. (2022). Estimation and worldwide monitoring of the effective reproductive number of SARS-CoV-2. *eLife*, *11*, e71345. https://doi.org/10.7554/eLife.71345

Johnson, S. G. (2007). *The NLopt nonlinear-optimization package*. Retrieved from https://github.com/stevengj/nlopt

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., … Lessler, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, *172*, 577–582.

Meyer, M. C. (2013). A simple new algorithm for quadratic programming with applications in statistics. *Communications in Statistics*, *42*, 1126–1139. https://doi.org/10.1080/03610918.2012.659820

Reich, N. G., Lessler, J., Cummings, D. A. T., & Brookmeyer, R. (2009). Estimating incubation period distributions with coarse data. *Statistics in Medicine*, *28*, 2769–2784. https://doi.org/10.1002/sim.3659

Rockafellar, R. T. (1970). *Convex analysis Princeton mathematical series* (Vol. *28*). Princeton, NJ: Princeton University Press.

How to cite this article: Groeneboom, P. (2024). Estimation of the incubation time distribution in the singly and doubly interval censored model. *Statistica Neerlandica*, 1–19. https://doi.org/10.1111/stan.12335

APPENDIX A. PROOFS

In contrast with the difficulties of the continuous model in Groeneboom (2023), the estimation of the parameters $\int_{i-1}^{i} F_0(t)\,dt$ seems rather standard, once we have figured out the relation between the continuous model and the discrete observations of the times of becoming symptomatic. We still have to deal with a deconvolution problem, which we do by using nonparametric maximum likelihood.

Proof of Lemma 2. The key step is to reduce the maximization on the simplex $\{\boldsymbol{p} = (p_1, \ldots, p_m) \in \mathbb{R}_+^m : \sum_{i=1}^m p_i = 1\}$ to maximization on the cone $\mathbb{R}_+^m$. One can check that minimizing minus the log likelihood (25) under the restriction $\sum_{i=1}^m p_i = 1$ is equivalent to minimizing (26), which is the criterion function + a Lagrange term with Lagrange multiplier $\lambda = 1$, on $\mathbb{R}_+^m$. Then the necessary and sufficient conditions for the minimum follow from Fenchel's duality theorem, see Rockafellar (1970), theorem 31.4. ∎

Proof of Theorem 1. The log likelihood is of the form

$$\ell(p_1, \ldots, p_m) = \sum_{i=1}^{n} \log \sum_{j=1}^{m} p_j 1_{((S_i - E_i)_+, S_i]}(j),$$

so we count the number of times the point of mass j belongs to an interval $(S_i - E_i)_+, S_i]$, where S_i and E_i are integers, and multiply this with the probability p_j. We have

$$\begin{aligned} &\frac{\partial}{\partial p_j} \ell\left(p_1, \ldots, 1 - \sum_{k=1}^{m-1} p_k\right) \\ &\quad = \sum_{i=1}^{n} \frac{1_{((S_i - E_i)_+, S_i]}(j) - 1_{((S_i - E_i)_+, S_i]}(m)}{\sum_{k=1}^{m} p_j 1_{((S_i - E_i)_+, S_i]}(k)}, \qquad j = 1, \ldots, m-1, \end{aligned} \tag{A1}$$

and

$$\begin{aligned} &\frac{\partial^2}{\partial p_j \partial p_l} \ell\left(p_1, \ldots, 1 - \sum_{k=1}^{m-1} p_k\right) \\ &\quad = -\sum_{i=1}^{n} \frac{\left\{1_{((S_i - E_i)_+, S_i]}(j) - 1_{((S_i - E_i)_+, S_i]}(m)\right\}\left\{1_{((S_i - E_i)_+, S_i]}(l) - 1_{((S_i - E_i)_+, S_i]}(m)\right\}}{\left\{\sum_{k=1}^{m} p_j 1_{((S_i - E_i)_+, S_i]}(k)\right\}^2}, \end{aligned}$$

for $j, l = 1, \ldots, m-1$, using the convention $0/0 = 0$.

By the assumptions on F_0 and F_E, the variables S_i and $(S_i - E_i)_+$ will be such that for large n, the score functions (A1) will be zero for $p_j = \hat{p}_j$, where $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_m)$ is the MLE of $\boldsymbol{p}_0 = (p_0(1), \ldots, p_0(m))$ (i.e., no isotonization is needed), and the result now follows from standard theory. ∎

Proof of Theorem 2. The proof is entirely similar to the proof of Theorem 1, but this time the score functions are given by

$$\begin{aligned}&\frac{\partial}{\partial p_j}\ell\left(p_1,\ \ldots\ ,1-\sum_{k=1}^{m-1}p_k\right)\\&=\sum_{i=1}^{n}\frac{\psi(E_i,S_{L,i},S_{R,i},j)-\psi(E_i,S_{L,i},S_{R,i},m)}{\sum_{k=1}^{m}\psi(E_i,S_{L,i},S_{R,i},t)\ p_k},\qquad j=1,\ \ldots\ ,m-1,\end{aligned}\tag{A2}$$

where ψ is defined by (19).

A key part of the treatment of the doubly censored case is the rewrite of the log likelihood for one observation, using the integration by parts

$$\int_{t\in(s_L,s_R]}\{F(t)-F(t-E)\}\ dt=\int\psi(e,s_L,s_E,t)\ dF(t),$$

where ψ is defined by (19). ■