Estimation of the incubation time distribution for COVID-19

Groeneboom, Piet

DOI
10.1111/stan.12231

Publication date
2020

Document Version
Final published version

Published in
Statistica Neerlandica

Received: 23 September 2020 | Accepted: 12 November 2020

DOI: 10.1111/stan.12231

ORIGINAL ARTICLE

WILEY

Estimation of the incubation time distribution for COVID-19

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, Delft, The Netherlands.
Email: P.Groeneboom@tudelft.nl

We consider smooth nonparametric estimation of the incubation time distribution of COVID-19, in connection with the investigation of researchers from the National Institute for Public Health and the Environment (Dutch: RIVM) of 88 travelers from Wuhan: Backer et al. (2020). The advantages of the smooth nonparametric approach with respect to the parametric approach, using three parametric distributions (Weibull, log-normal and gamma) in Backer et al. (2020) is discussed. It is shown that the typical rate of convergence of the smooth estimate of the density is $n^{2/7}$ in a continuous version of the model, where n is the sample size. The (nonsmoothed) nonparametric maximum likelihood estimator itself is computed by the iterative convex minorant algorithm (Groeneboom and Jongbloed (2014)). All computations are available as `R` scripts in Groeneboom (2020a).

KEYWORDS

incubation time, iterative convex minorant algorithm, nonparametric MLE, smooth nonparametric density estimation, Weibull distribution

1 | INTRODUCTION

Researchers from the Centre for Infectious Disease Control and Prevention of the National Institute for Public Health and the Environment (Dutch: RIVM) analyze in Backer, Klinkenberg, and

Wallinga (2020) a data set of 88 travelers who are assumed to have picked up the COVID-19 virus in Wuhan. The distribution of their incubation times is estimated using certain simple distributions, like Weibull, log-normal and gamma. If the only thing we know about the start of the incubation time is that it belongs to an interval $[0, E_i]$, the log likelihood for one observation is:

$$\log \int_{t\in[0,E_i]} g(S_i - t)\, dF_i(t).$$

Here E_i would be the upper bound for the exposure interval, for which we take (looking back) 0 as the left point for the ith individual (see Britton & Scalia Tomba, 2019), S_i is the time where the person becomes symptomatic (note that both $S_i \le E_i$ and $S_i > E_i$ can occur), and F_i would be the distribution function of the time of a possible contact with an infector. The exit times and times of becoming symptomatic of the 88 Wuhan travelers are shown in Table 1.

It is clear that, without further assumptions, g and F_i are not identifiable. To remedy this, we assume, as in Backer et al. (2020) (see also Reich, Lessler, Cummings, & Brookmeyer, 2009), that F_i is the uniform distribution on $[0, E_i]$. If we want to use maximum likelihood, we have to maximize

$$\sum_{i=1}^{n} \log \left\{ \int_{t=0}^{E_i} g(S_i - t)\, dt/E_i \right\},$$

and since the E_i do not matter in the maximization problem, we end up with the problem of maximizing

$$\sum_{i=1}^{n} \log \left\{ \int_{t=0}^{E_i} g(S_i - t)\, dt \right\}, \tag{1}$$

where g is the density of the incubation time.

So we deal with the following model. We have an exit time E_i for the exposure interval, an infection time V_i and an incubation time W_i. The time of becoming symptomatic is denoted by S_i, and S_i is, conditionally on the exit time, assumed to be the independent sum of V_i and W_i. Our observations are

$$(E_i, S_i, \Delta_i), \quad i = 1, \ldots, n, \tag{2}$$

where n is the sample size and where the indicator Δ_i is defined by

$$\Delta_i = 1_{\{S_i \le E_i\}}, \quad i = 1, \ldots, n. \tag{3}$$

Using the present notation, the log likelihood for the incubation time distribution function G becomes

$$\ell(G) = \sum_{i=1}^{n} [\Delta_i \log G(S_i) + (1 - \Delta_i) \log\{G(S_i) - G(S_i - E_i)\}]. \tag{4}$$

Note that the time of becoming symptomatic is still in Wuhan if $\Delta_i = 1$.

The algorithms we used for analyzing the data set can be found on Groeneboom (2020a). We describe the data files given there. The original data file is `data_Wuhan_tsv`, which gives

TABLE 1 Exit times and times of becoming symptomatic of the 88 Wuhan travelers after shifting the entrance times to 0

i	E_i	S_i	i	E_i	S_i
1	5	5	45	39	40
2	30	33	46	35	42
3	21	22	47	2	6
4	1	4	48	36	37
5	1	6	49	38	39
6	8	8	50	1	8
7	4	4	51	38	41
8	3	3	52	38	41
9	33	34	53	38	39
10	33	34	54	11	11
11	8	8	55	36	39
12	1	4	56	11	11
13	20	21	57	40	41
14	20	28	58	36	37
15	30	32	59	36	41
16	35	38	60	36	39
17	3	7	61	27	31
18	35	37	62	38	40
19	36	38	63	36	42
20	31	38	64	40	43
21	34	35	65	41	43
22	29	31	66	37	43
23	36	37	67	1	7
24	3	8	68	40	42
25	7	9	69	40	42
26	38	39	70	31	39
27	30	36	71	40	41
28	28	36	72	40	41
29	35	36	73	41	42
30	33	34	74	41	43
31	3	8	75	4	5
32	2	4	76	4	5
33	2	5	77	40	41
34	5	5	78	36	40
35	36	37	79	36	40
36	31	35	80	40	42
37	41	42	81	36	42
38	41	42	82	38	43
39	3	4	83	2	9
40	38	39	84	38	43
41	39	41	85	37	43
42	39	41	86	41	42
43	39	41	87	40	43
44	33	39	88	40	43

details on the persons in the sample and which can be found in Backer et al. (2020). This was transformed into a data file `transformed_data_Wuhan.txt`, consisting of three columns, giving, respectively, the arrivals in (if available) and departures from Wuhan and the time the person became symptomatic. If the arrival time was not available (possibly because the person was a Wuhan resident), this time was set to −18, which means 18 days before December 31, 2019, which is the zero on the time scale. For traveler number 67, who apparently had a connecting flight, the duration of stay in Wuhan was changed from 0 to 1 day. This, in turn, was transformed into the input file `inputdata_Wuhan.txt`, where the time, spent in Wuhan, was shifted making the left point equal to zero, and consists of two columns: the first columns contains the data $S_i - E_i$ (time of becoming symptomatic minus exit time from Wuhan) and S_i, time of becoming symptomatic, where all times are shifted to have entrance time zero. If the person became symptomatic in Wuhan we put E_i equal to S_i, so $S_i - E_i = 0$.

Assuming that the distribution of the possible time of infection is uniform on the exposure interval, and estimating the distribution function G by the Weibull distribution, parametrized as

$$G(x) = G_{a,b}(x) = 1 - \exp\{-bx^a\}, \quad x > 0, \tag{5}$$

we get as our maximum likelihood estimators (MLE) of the parameters a and b:

$$\hat{a} = 3.03514, \quad \hat{b} = 0.002619. \tag{6}$$

Using the Weibull maximum likelihood method, the estimate was computed by two methods. One is a very simple method using `Weibull.cpp`, which is used in `analysis_EM.R` and `analysis_ICM.R`, where also the nonparametric estimate to be discussed in the next sections is computed. For this "pattern search" algorithm for looking for the parameters of the Weibull distribution one does not have to compute the derivatives of the log likelihood. It is based on the Hooke–Jeeves algorithm. The other one can be found in `R_Weibull_Estimation.R`, where we use the `R` package `lbfgs`, and where the gradient (derivatives of the log likelihood) has to be provided.

The results obtained for the Weibull distribution approach of the two algorithms are remarkably similar. The values in (6) were produced by the `R` script in Groeneboom (2020a), using the Hooke–Jeeves algorithm. For a convergence proof of the Hooke–Jeeves algorithm and interesting further discussion of the pattern search algorithms, see Kolda, Lewis, and Torczon (2003) and Torczon (1997).

The aim of the present paper, however, is to draw attention to the nonparametric MLE of the incubation time distribution, which is often also denoted by NPMLE (Nonparametric MLE). This is the distribution function $\hat{G}_n$, maximizing (4) over all distribution functions G. The problem of maximizing (4) over all distribution functions G instead of just Weibull, log-normal or gamma distribution functions is nontrivial and discussed in Section 2. We also discuss the smooth estimators based on the MLE, the so-called SMLE (Smoothed MLE) and the nonparametric density estimator, based on the MLE.

When we want to get an idea of properties of the incubation time distribution, there are (at least) three approaches.

1. We "fit" the data with a parametric distribution from a well-known family of distributions like the Weibull, log-normal or gamma distributions. The big disadvantage of this approach is that

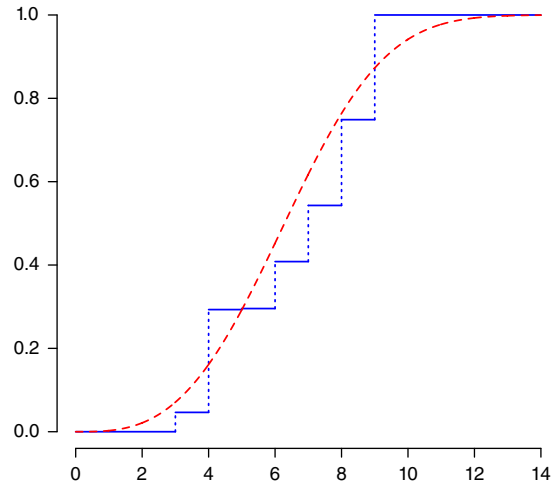

FIGURE 1 The nonparametric maximum likelihood estimate (MLE) $\hat{G}_n$ of the incubation time distribution function (blue), and the MLE using the Weibull distribution (red, dashed), for the dataset analyzed in Backer et al. (2020)

TABLE 2 Probability masses of the nonparametric maximum likelihood estimator

Number of days	p_i
3	.0463850922
4	.2466837048
5	.0024858945
6	.1126655228
7	.1347501680
8	.2058210187
9	.2512085991

one usually does not have a good argument for choosing one of these distributions and that important aspects of the data might be completely hidden by the choice of such a distribution.

Convincing examples of this situation are given in chapter 1 of Silverman (1986). If one fits the multimodal distribution of the eruptions of the Old Faithful Geyser in Yellowstone Park, Wyoming, by a unimodal distribution, one will only see one mode instead of the multiple modes that really are there. In that chapter also other interesting examples of how special aspects of the data are revealed by nonparametric density estimation are given.

In fact, estimates of the simple parametric type such as the Weibull, etc. will usually be *inconsistent*: no matter how many observations one has, there will not be convergence to the right distribution. The ubiquitous appearance of the normal distribution has a completely different origin: the central limit theorem. But this reasoning will generally not apply in the same way for fitting with the Weibull, etc. distribution.

Another disadvantage which clearly shows up if people use this method (as in Backer et al., 2020) is that one usually has to introduce several families of distributions (gamma, log-normal, Weibul …), because there is no compelling reason to pick one of these.

2. We compute the nonparametric MLE. The result for the Wuhan data is shown in Figure 1 and the bar chart of the point masses of the MLE is shown in Figure 2 (the values of the point masses are shown in Table 2).

This is what one gets if one makes no assumptions at all about the distribution function and this is the "antipode" of the fitting with the Weibull etc. distribution. Figure 2 clearly

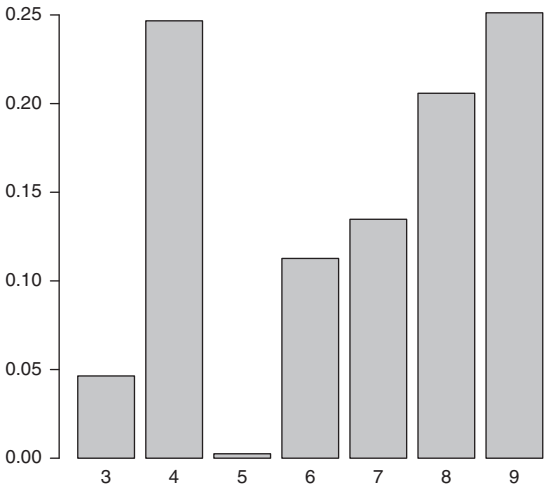

FIGURE 2 Bar chart of the probability masses of the nonparametric MLE

shows a bimodal discrete density, but one wonders: is this bimodality due to chance fluctuations or is it real? Note that this discrete density is rather different from the density estimation of Silverman (1986), mentioned in point 1. In the latter case one assumes the existence of a (continuous) density with respect to Lebesgue measure instead of a discrete density.

How do we view the distribution of the incubation time? My own inclination is to assume the existence of a continuous density with respect to Lebesgue measure for the incubation time distribution and to use methods as in Silverman (1986), which entails smoothing. Which takes us to:

3. We estimate the density of the incubation time with respect to Lebesgue measure in a nonparametric way. In this case we also need an extra parameter, the *smoothing parameter* or *bandwidth*. Now one could argue (as has been done): "Ah, you objected in point 1 to the use of parametric distributions such as for example the Weibull distribution, but now you introduce a parameter again, the bandwidth!". Fair enough, but: *"The bandwidth is a parameter of a totally different nature than the parameters of the Weibull distribution!"*. With the bandwidth one tries to mediate between the noise and the bias, something we cannot do with the nonparametric estimate, introduced in point 2. Moreover, we can do this in a data-adaptive way, to create independence of a priori assumptions, a type of independence we cannot achieve with the estimates in point 1 above.

We must add, however, that the density estimation problem here is considerably more difficult than the density estimation problems considered in Silverman (1986). This is caused by the fact that our observations are indirect; we assume that the infection took place during the stay in Wuhan, but we do not know when. We only have an interval for this infection time. For this situation we have to use the so-called *interval censoring model*, which is for example discussed in Groeneboom and Jongbloed (2014). In fact, we have to deal with a combination of interval censoring (the infection time is contained in an interval, we cannot observe it directly) and *deconvolution*, since we have to extract the information from the sum of the infection time and the incubation time. For this reason we get slower rates of convergence of the density estimate: $n^{2/7}$ instead of the usual rate of convergence in density estimation, which is $n^{2/5}$ (see Silverman, 1986 for the latter rate). An additional complication is that the observations are usually discretized, but we analyze in the sequel both the continuous model just described in Section 4 and the discretized model for which we cannot hope to achieve rate $n^{2/7}$ at each point.

Similar considerations hold for the SMLE, estimating the distribution function. In this case we also need a bandwidth (smaller than the bandwidth for the density estimate) and the rate will be $n^{2/5}$, which is the rate in ordinary density estimation. So in this sense the SMLE is comparable to an ordinary density estimate and the density estimate for the incubation time distribution is comparable to the ordinary estimate of the derivative of a density.

In this paper we focus on the method, described under point 3 above and give algorithms for computing the estimators. `R` scripts for all these methods are given in Groeneboom (2020a).

It should be noted that the asymptotic distribution of the MLE itself is unknown. In the continuous (not discretized) model it is expected to have the Chernoff limit distribution (location of the maximum of two-sided Brownian motion minus a parabola), but at present this is unknown, as it also is for the related limit distribution of the MLE in the so-called interval censoring, case 2, model (see Groeneboom & Jongbloed, 2014).

But we do not need the limit distribution of the MLE itself for deriving the (normal) limit distributions of the SMLE and density estimate, based on the MLE. As an example, we give the derivation for the limit distribution of the density estimate in the simulation model discussed in Section 4 in Appendix (Section A1). The fit of the variances, predicted by the asymptotic theory and the variances coming from the simulation study is remarkably good, see Table A1 and Figure A2.

2 | ALGORITHMS FOR COMPUTING THE NONPARAMETRIC MLE

The EM iterations for the MLE maximizing (1), without making this parametric restriction, are in this case given by:

$$p_j' = p_j n^{-1} \sum_{i=1}^{n} 1_{\{j \in (S_i - E_i, S_i]\}} \Big/ \sum_{k \in (S_i - E_i, S_i]} p_k \,, \tag{7}$$

where the ratios are zero if the denominators are zero. The implementation of this algorithm for the present situation can be found in `analysis_EM.R` in Groeneboom (2020a).

The EM iterations were started with the discrete uniform distribution on the 43 points $1, \ldots, 43$, which corresponds to the range of values (days) in Table 1, but withdrew its mass after 10,000 iterations to the 7 points $3, \ldots, 9$, which leads to the discrete distribution function, shown in Figure 1. A bar chart of the corresponding probability masses is shown in Figure 2. It is seen that this is a bimodal discrete probability distribution with modes at resp. 4 and 9 days, with the highest value at the second mode. This discrete probability distribution is also given in Table 2.

The iteration steps (7) follow from the so-called self-consistency equations, which are derived by differentiating the criterion function

$$n^{-1} \sum_{i=1}^{n} \log \left\{ \sum_{j \in (S_i - E_i, S_i]} p_j \right\} - \lambda \left\{ \sum_{j=1}^{m} p_j - 1 \right\}, \tag{8}$$

with respect to (w.r.t.) p_i, where in this case $m = 43$, and λ is a nonnegative Lagrange multiplier, chosen in such a way that

$$\sum_{j=1}^{m} p_j = 1. \tag{9}$$

This yields

$$n^{-1} \sum_{i=1}^{n} 1_{\{j \in (S_i - E_i, S_i]\}} \Big/ \sum_{k \in (S_i - E_i, S_i]} p_k = \lambda, \quad j = 1, \ldots m, \tag{10}$$

and multiplying these relations with p_j and summing over j yields $\lambda = 1$, using the side condition (9). But the relations (10) only hold for the *active* (in this case 7) parameters $p_i > 0$ of the solution; in the iterations (7) the inactive parameters p_i will tend to zero. For more details, see, for example, Groeneboom and Jongbloed (2014), section 7.2.

Because of the monotonicity of the distribution function G, maximizing the log likelihood over all distribution functions G is an isotonic regression problem, which can be solved by specific isotonic methods. In the present case we can apply the *iterative convex minorant algorithm*, discussed in Groeneboom and Jongbloed (2014), section 7.3.

As discussed in Section 1, the log likelihood is of type:

$$f(\boldsymbol{y}) = \sum_{i=1}^{m} k_i \log(G(U_i) - G(T_i)), \tag{11}$$

where k_i is the number of observations (T_i, U_i), and where

$$(T_i, U_i) = (0, V_i + W_i) 1_{\{V_i + W_i \le E_i\}} + (V_i + W_i - E_i, V_i + W_i) 1_{\{V_i + W_i > E_i\}} \quad i = 1, \ldots, n, \tag{12}$$

where $n = 88$, and where V_i is the infection time, W_i the incubation time and, as before, E_i the exit time of the travelers from Wuhan, where all observations are centred by subtracting the entrance time.

We first make the so-called preliminary reduction to reduce the problem to a maximization problem in the interior of a convex cone of type

$$\{\boldsymbol{y} = (y_1, \ldots, y_m)^T : 0 < y_1 \le \ldots \le y_m\}.$$

For the Wuhan dataset it can be checked that, without loss of generality, $G(i) = 0$, $i \le 2$, and $G(i) = 1$, $i \ge 9$, since in this case values strictly between 0 and 1 can only make the likelihood smaller. If we make this preliminary reduction, the log likelihood for the ordered parameters y_i, representing the values of the distribution function G at the observation points, becomes:

$$f(\boldsymbol{y}) = \sum_{0 \le i < j \le 7} N_{ij} \log(y_j - y_i), \tag{13}$$

where $y_i = G(i+2)$, $i = 0, \ldots, 7$, $y_0 = 0$, $y_7 = 1$, and where the triangular array (N_{ij}), $0 \le i < j \le 7$, is given by:

$$\begin{array}{ccccccc} 1 & 3 & 4 & 0 & 0 & 2 & 0 \\ & 2 & 1 & 0 & 0 & 0 & 9 \\ & & 0 & 1 & 1 & 0 & 4 \\ & & & 1 & 0 & 2 & 3 \\ & & & & 1 & 0 & 6 \\ & & & & & 1 & 3 \\ & & & & & & 3 \end{array}$$

We have to maximize (11) under the restriction $0 < y_1 \leq \cdots \leq y_6$; by the preliminary reduction, we lost the additional condition $y_6 < 1$. Let $\boldsymbol{y} = (y_1, \ldots, y_6)^T$. The (Fenchel) sufficient and necessary conditions for the solution are:

$$\sum_{j=i}^{6} \frac{\partial}{\partial y_j} f(\boldsymbol{y}) \leq 0, \quad i = 1, \ldots, 6, \tag{14}$$

and

$$\sum_{i=1}^{6} y_i \frac{\partial}{\partial y_i} f(\boldsymbol{y}) = 0, \tag{15}$$

where f is defined by (11). Since the values y_i are strictly between 0 and 1, (15) can only hold if also

$$\sum_{i=1}^{6} \frac{\partial}{\partial y_i} f(\boldsymbol{y}) = 0,$$

and we can therefore turn (14) into

$$\sum_{j=1}^{i} \frac{\partial}{\partial y_j} f(\boldsymbol{y}) \geq 0, \quad i = 1, \ldots, 6. \tag{16}$$

The resulting (nonparametric) MLE $\hat{F}_n$ is shown in Figure 1, together with the MLE assuming that G is a Weibull distribution. The EM algorithm and the iterative convex minorant (ICM) algorithm give exactly the same solutions, but the ICM algorithm needs less iterations (106 in this case; the EM algorithm needs between 1000 and 10,000 iterations).

To compute the MLE via the iterative convex minorant algorithm, we have to construct so-called cusum (cumulative sum) diagrams. The cusum diagram consists of the point $(0, 0)$ and the points

$$\sum_{j=1}^{i} \left(w_j, \frac{\partial}{\partial y_j} f(\boldsymbol{y}) + w_j y_j \right), \quad i = 1, \ldots, 6, \tag{17}$$

where

$$w_j = -\frac{\partial^2}{\partial y_j^2} f(\boldsymbol{y}). \quad j = 1, \ldots, 6. \tag{18}$$

At each iteration step the left derivative vector $\boldsymbol{y}'$ of the greatest convex minorant of the cusum diagram is computed on the basis of the current value $\boldsymbol{y}$, and the stationary point of this iteration is the solution of the optimization problem. We perform line search in case the full step to $\boldsymbol{y}'$ would not lead to improvement or would go out of bounds. For more theory, see Groeneboom and Jongbloed (2014).

As in Groeneboom and Jongbloed (2014), section 1.2, we can compute the SMLE and also an estimate of the density. The SMLE is defined by

$$\tilde{G}_{nh}(t) = \int \mathbb{K}((t-y)/h)\; d\hat{G}_n(y), \tag{19}$$

where $h > 0$ and $\mathbb{K}$ is an integrated kernel

$$\mathbb{K}(x) = \int_{-\infty}^{x} K(u)\; du. \tag{20}$$

Here K is a symmetric kernel with support $[-1, 1]$, for example the triweight kernel

$$K(u) = \frac{35}{32}(1-u^2)^3 1_{[-1,1]}(u). \tag{21}$$

We estimate the density by

$$\tilde{g}_{nh}(t) = h^{-1} \int K((t-y)/h)\; d\hat{G}_n(y). \tag{22}$$

For the present analysis we took $h = 3.6$ in (19) and $h = 4.6$ in (22); these bandwidths were chosen by a bootstrap method, explained in Section 3. The resulting estimates are shown in Figure 3.

3 | DATA-ADAPTIVE BANDWIDTH CHOICE FOR THE DENSITY ESTIMATE AND THE SMLE

Let the random variables E_i with values on the integers ("days") on the interval $[1, 43]$ represent the exit times. Furthermore, let V_i denote the (unknown) infection time, which we take, conditionally on E_i, to be uniform on $[0, E_i]$, and let W_i denote the (again unknown) incubation time. Our observations are the triples (E_i, S_i, Δ_i), given by (2).

To determine the bandwidth h of our density estimator

$$\hat{g}_{nh}(t) = \int K_h(t-y)\; d\hat{G}_n(y), \tag{23}$$

where $\hat{F}_n$ is the MLE of the distribution function F of the incubation time, we follow a method somewhat similar to the method used in Sen and Xu (2015).

We take $B = 10,000$ bootstrap samples of observations $\left(E_i, S_i^*, \Delta_i^*\right)$, corresponding to the observations (E_i, S_i, Δ_i). The S_i^* are generated as the sums (rounded to the nearest integer) of a Uniform$(0, E_i)$ random variable V_i^* and a random variable W_i^*, generated from the density $\hat{f}_{nh_0}$ by

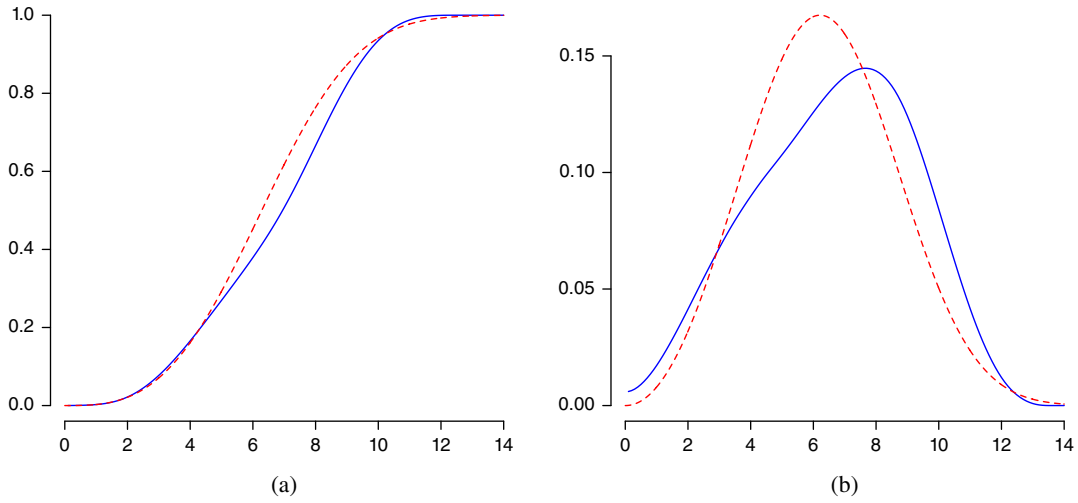

FIGURE 3 (a): The smoothed nonparametric maximum likelihood estimate (SMLE) of the incubation time distribution function (blue), and the MLE using the Weibull distribution (red, dashed), for the dataset analyzed in Backer et al. (2020) and (b): the SMLE of the incubation time density function (blue), and the MLE of the density using the Weibull distribution (red, dashed), for the data set analyzed in Backer et al. (2020)

rejection sampling for a fixed h_0, for which we took $h_0 = 4$ in the present case. The Δ_i^* are given by

$$\Delta_i^* = 1_{\{V_i^* + W_i^* \le E_i\}} .$$

Note that we keep the E_i the same as in the original sample, somewhat analogously to the procedure followed in Sen and Xu (2015), which relieves us from the duty to estimate the exit time distribution.

Next we computed

$$\hat{\mathrm{MSE}}_g(h) = B^{-1} \sum_{b=1}^{B} \int \{\hat{g}_{nh}^*(x) - \hat{g}_{nh_0}(x)\}^2 \, dx. \tag{24}$$

The resulting loss function $\hat{\mathrm{MSE}}_f(h)$ is shown in Figure 4, which gave as the minimizing bandwidth $\hat{h} \approx 4.6$. Taking $h_0 = 3$ in our function of reference $\hat{g}_{nh_0}$ gave the same minimizing value. The (approximate) independence of the starting value h_0 was also observed for the analogous bandwidth selection procedure in Sen and Xu (2015).

Similarly, we computed

$$\hat{\mathrm{MSE}}_G(h) = B^{-1} \sum_{b=1}^{B} \int \left\{\hat{G}_{nh}^*(x) - \hat{G}_{nh_0}(x)\right\}^2 \, dx, \tag{25}$$

as a function of h by the same bootstrap procedure, where $\hat{G}_{nh}^*$ was computed for the bootstrap samples. The integrals were approximated by Riemann sums with step size 0.1 on the interval $[0, 14]$. The R scripts for this procedure can again be found in Groeneboom (2020a). The method used here is called the "smoothed bootstrap", because we generate the bootstrap samples from the smooth estimate $\hat{g}_{nh_0}$ of the density of the incubation time (added to a uniform$[0, E_i]$ random

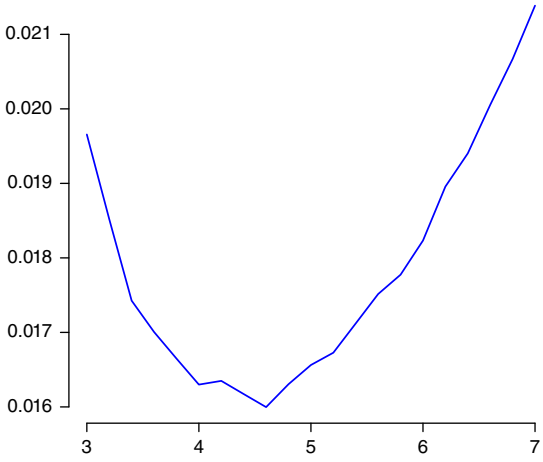

FIGURE 4 $\widehat{\mathrm{MSE}}_g(h)$, given by (24), as function of h

variable) instead of just resampling with replacement from the data (E_i, S_i, Δ_i), as one would do in the ordinary bootstrap.

A perhaps slightly unorthodox variant of the present method is the smooth bootstrap where we do not round the sums of V_i^* and W_i^* to the nearest integer, but just use them as continuous variables (for more information on the continuous model see the next session). The unorthodox aspect is that, in our bootstrap experiment, we do not recreate exactly the same situation as in our original setting, where the data are integers. In fact, we create data for the continuous model, where we can easier compare bias and variance. We tried this out for the density estimates, and it actually gave exactly the same minimizing bandwidth $h = 4.6$ for the least squares criterion. More research on this method is necessary, though.

4 | THE CONTINUOUS MODEL

Applying the method of the preceding section to the discrete data, where one only uses days on the time axis, is somewhat dubious, since, in fact, we do not have information on a finer scale, which would allow us to let the bandwidth (and therefore the bias) tend to zero. It is conceivable that we have information on a finer scale, for example, the time of the outgoing flight or the time of day of becoming symptomatic. Presently both times are interval censored (where one day is the interval). We could therefore introduce another assumption, for example that the time of becoming symptomatic is uniformly distributed over a day. In any case, there seems to be enough reason to study the continuous model, where one would have (approximately) continuous observations, and to analyze what can be expected in this case.

We define as before the indicator Δ by

$$\Delta = 1_{\{S \le E\}}, \tag{26}$$

where E is again the exit time and S is the time of becoming symptomatic, and consider the following simulation experiment. E_i is uniform$[0, M]$, the time of infection V_i is a Uniform random variable on $[0, E_i]$, conditionally on E_i, and the incubation time W_i is a truncated Weibull(a, b) distribution, where a and b have the same values as the estimates $\hat{a}$ and $\hat{b}$ in (6), respectively, and where the truncation interval $[0, M_1]$ is contained in the interval $[0, M]$. In the present

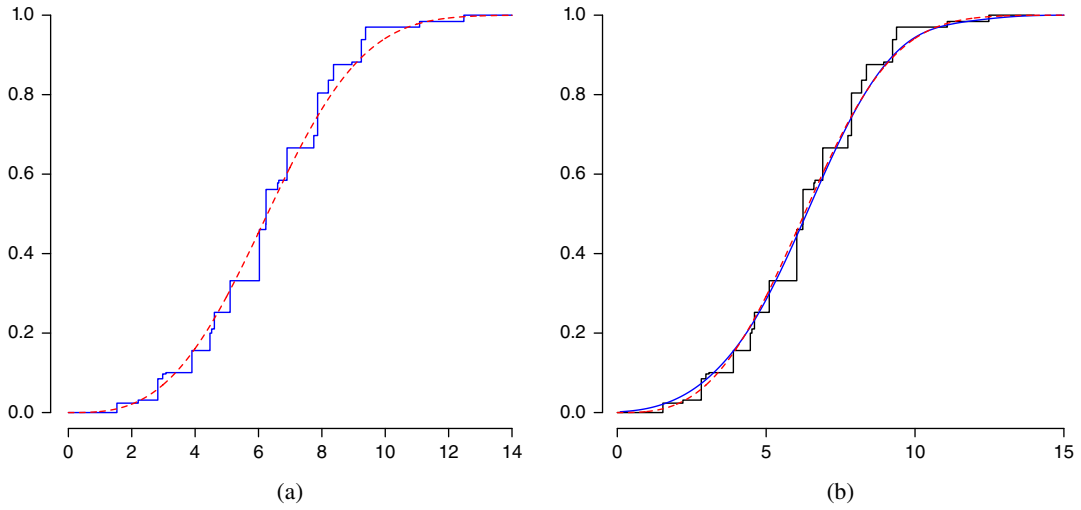

FIGURE 5 (a) The nonparametric maximum likelihood estimate (MLE) $\hat{G}_n$ of the incubation time distribution function (blue) for a sample of size $n = 1000$, and the truncated Weibull distribution function (red, dashed) with parameters a and b, in the simulation model where the variables are not discretized. (b) The MLE (black) and the SMLE (blue), for the same sample, and the truncated (on $[0, M_1]$) Weibull distribution function (red, dashed). The bandwidth of the SMLE is $h = 3$

simulation, we took $M_1 = 20$ and $M = 30$. In this way the upper bound for the observations S_i is equal to 50, which is somewhat comparable with the upper bound 43 of the observations S_i for the Wuhan travelers. This means that $S_i = V_i + W_i$, where we assume that V_i and W_i are independent, and that our observations are the triples (E_i, S_i, Δ_i).

The MLE of the incubation time, where E_i and S_i are known, looks rather different from the MLE based on the discretized observations shown in Figure 1. An example of such an MLE is shown in Figure 5 for a sample of $n = 1,000$. Since in this case the MLE can have more jumps, it has the possibility to be much closer to the continuous distribution function. It maximizes again expression (1), but this time the variables E_i and S_i are not discretized.

In this setup, the SMLE will, in the interior of the interval $[0, M_1]$, pointwise have the $n^{2/5}$ rate and the corresponding nonparametric density estimate the $n^{2/7}$ rate of convergence, and the pointwise limit distributions will be normal in both cases (see Section A1 of the present paper and Groeneboom and Jongbloed (2014), section 11.4). For the density estimate in the present simulation model we get the following result.

Theorem 1. *Let $\tilde{g}_{n,h_n}$ be the estimate of the density, defined by*

$$\tilde{g}_{n,h_n}(t) = h^{-1} \int K((t-y)/h_n)\, d\hat{G}_n(y) = \int K_{h_n}(t-y)\, d\hat{G}_n(y),$$

where $h_n \sim cn^{-1/7}$, for some $c > 0$. Let the score function $\theta_{t,h,G}$ be defined by

$$\theta_{t,h,G}(e,s,\delta) = \delta \frac{\phi(s)}{G(s)} + (1-\delta)\frac{\phi(s) - \phi(s-e)}{G(s) - G(s-e)\}}, \tag{27}$$

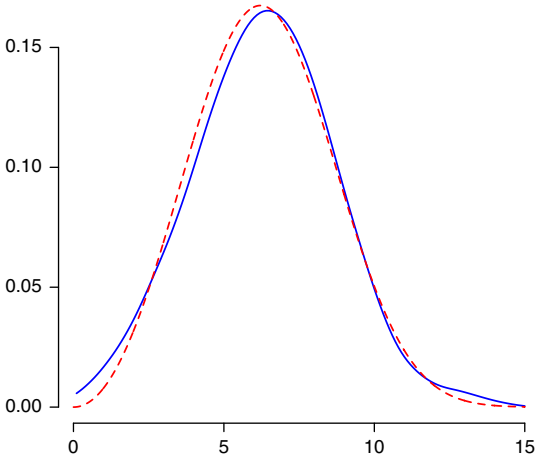

FIGURE 6 The nonparametric estimate of the density of the incubation time (blue, solid), based on a sample of size $n = 1,000$, based on the truncated Weibull distribution, where we use bandwidth $h = 3.4$. The red dashed curve is the truncated Weibull density with parameters a and b of (6)

where δ is the indicator $\delta = 1_{\{s \le e\}}$ and where ϕ solves the integral equation

$$
\begin{aligned}
&-\frac{\phi(w)}{MG(w)}\log(M/w) + \frac{1}{M}\int_{e=0}^{w}\frac{1}{e}\left\{\frac{\phi(w+e)-\phi(w)}{G(w+e)-G(w)} - \frac{\phi(w)-\phi(w-e)}{G(w)-G(w-e)}\right\}de \\
&\quad + \frac{1}{M}\int_{e=w}^{M}\frac{1}{e}\frac{\phi(w+e)-\phi(w)}{G(w+e)-G(w)}\,de \\
&\quad = \frac{\partial}{\partial w}K_h(w-t),
\end{aligned}
\tag{28}
$$

defining $0/0 = 0$. Let $\mathbb{P}_n$ be the empirical probability measure of a sample $(E_1, S_1, \Delta_1), \ldots, (E_n, S_n, \Delta_n)$. Then we have, taking $h = h_n \sim cn^{-1/7}$, for a $c > 0$, and $G = G_0$ (the underlying incubation time distribution) in (27) and (28),

$$
n^{2/7}\left\{\tilde{g}_{n,h_n}(t) - \int K_{h_n}(t-y)\,dG_0(y)\right\} = n^{2/7}\int K_{h_n}(t-y)\,d(\hat{G}_n - G_0)(y) \xrightarrow{\mathcal{D}} N(0,\sigma^2), \tag{29}
$$

where $N(0,\sigma^2)$ is a normal distribution with mean zero and variance σ^2 given by:

$$
\sigma^2 = \lim_{n\to\infty}\mathrm{var}\left(n^{2/7}\int \theta_{t,h_n,G_0}(e,s,\delta)\,d\mathbb{P}_n(e,s,\delta)\right).
$$

A sketch of the proof is given in Appendix and the rather good fit of the simulated variance and the variances predicted by this asymptotic result is shown in Table A1 and Figure A2. We do not have an explicit expression for the function ϕ, but could solve the integral equation numerically. In the present simulation study, G_0 is given by the truncated Weibull distribution function with parameters given by (6) (Figure 6).

This means that we can apply the same techniques as in Groeneboom and Hendrickx (2017b) and the R-package Groeneboom and Hendrickx (2017a), and for example compute pointwise bootstrap confidence intervals for the density. The bandwidth was determined by taking bootstrap samples of size $m = 50$, using bandwidths of size $cm^{-1/7}$ and using the optimal constant $\hat{c}$ over the east squares criterion in the bandwidth $\hat{c}n^{-1/7} = 3.51991$, where $n = 1,000$, for the density in the original sample, where we compare with the density estimate with bandwidth $h = 3$ in the

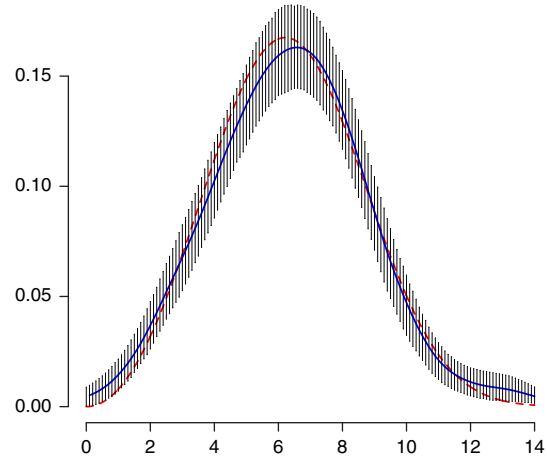

FIGURE 7 Density estimate (blue) and pointwise bootstrap 95% confidence intervals for the density of the incubation time distribution for a sample of size $n = 1,000$ (same sample as in Figures 5 and 6). The truncated Weibull density is given by the red dashed curve

original sample. This follows the procedure shown in the vignette of the R-package Groeneboom and Hendrickx (2017a). For the motivation for taking bootstrap samples of a smaller sample size, see Groeneboom and Hendrickx (2017b). The method goes back to Hall (1990). Since we have a simulation model here, we can also compute the real minimizing h, in a comparison with the truncated Weibull density. This yielded $h = 3.4$ in the present case, which is a value not far from the bandwidth found by the bootstrap sampling. In the pictures of this section, we took $h = 3.4$.

The bootstrap 95% confidence intervals for the density are shown for a sample of size $n = 1,000$ in Figure 7. These computations can again be checked on Groeneboom (2020a). For these intervals just 1,000 bootstrap samples were taken, resampling with replacement from the original sample of triples (E_i, S_i, Δ_i), computing the density estimate again in the bootstrap samples and determining the 2.5% and 97.5% percentiles of the values of the density estimates in the 1,000 bootstrap samples. To get really good intervals it is probably necessary to use an asymptotic pivot though, based on Theorem 1. This matter is subject to further investigation.

5 | CONCLUDING REMARKS

We offered an alternative nonparametric approach to the estimation of the incubation time distribution which was estimated by parametric methods in Backer et al. (2020) for a dataset of travelers from Wuhan. In this way we do not have to choose a parametric distribution, like the Weibull, log-normal or gamma, as in Backer et al. (2020), but compute a nonparametric maximum likelihood estimate instead which does not need the arbitrary choice of parameters at all.

However, to give a smooth estimate of the distribution function and (continuous) density, we have to choose a bandwidth parameter. For this choice a smoothed bootstrap approach was suggested. We also considered the model where the observations are not discretized and discussed rates of convergence, bootstrap confidence intervals and a limit theorem in that case. The present paper can be considered to be the technical companion of the column Groeneboom (2020b). All numerical computations are given as R scripts in Groeneboom (2020a).

ACKNOWLEDGEMENTS

I want to thank Guus Balkema, Ronald Geskus, and Siem Heisterkamp and a referee for their comments.

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. *Eurosurveillance*, *25*(5), 2000062.

Britton, T., & Scalia Tomba, G. (2019). Estimation in emerging epidemics: Bases and remedies. *Journal of the Royal Society Interface*, *16*. https://doi.org/10.1098/rsif.2018.0670.

Groeneboom, P. (2020a). Incubationtime. Retrieved from https://github.com/pietg/incubationtime

Groeneboom, P. (2020b). The Netherlands in times of corona (in Dutch). *Nieuw Archief voor Wiskunde*, *21*, 181–184.

Groeneboom, P., & Hendrickx, K. (2017a). curstatCI. R package. Version 0.1.1.

Groeneboom, P., & Hendrickx, K. (2017b). The nonparametric bootstrap for the current status model. *Electronic Journal of Statistics*, *11*(2), 3446–3484.

Groeneboom, P., & Jongbloed, G. (2014). *Nonparametric estimation under shape constraints*. Cambridge, MA: Cambridge University Press.

Hall, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *Journal of Multivariate Analysis*, *32*, 177–203.

Kolda, T. G., Lewis, R. M., & Torczon, V. (2003). Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Review*, *45*(3), 385–482.

Reich, N. G., Lessler, J., Cummings, D. A. T., & Brookmeyer, R. (2009). Estimating incubation period distributions with coarse data. *Statistics in Medicine*, *28*(22), 2769–2784.

Sen, B., & Xu, G. (2015). Model based bootstrap methods for interval censored data. *Computational Statistics & Data Analysis*, *81*, 121–129.

Silverman, B. (1986). *Density estimation for statistics and data analysis* (Vol. *26*). Boca Raton, FL: CRC Press.

Torczon, V. (1997). On the convergence of pattern search algorithms. *SIAM Journal on Optimization*, *7*(1), 1–25.

How to cite this article: Groeneboom P. Estimation of the incubation time distribution for COVID-19. *Statistica Neerlandica*. 2020;1–19. https://doi.org/10.1111/stan.12231

APPENDIX

Using the notation of p. 330 of Groeneboom and Jongbloed (2014), we define the score function $\theta_{t,h,G}$ by:

$$\begin{aligned}\theta_{t,h,G}(e,s,\delta) &= E[a(W)|(E,S,\Delta)=(e,s,\delta)]\\ &= \delta\frac{\int_{w\le s}a(w)\,dG(w)}{G(s)} + (1-\delta)\frac{\int_{w\in(s-e,s]}a(w)\,dG(w)}{G(s)-G(s-e)},\end{aligned} \tag{A1}$$

where $\delta = 1_{\{s\le e\}}$. We assume $G(M_1)=1$, where G is the distribution function of the incubation time and M_1 is the upper bound of the support of the distribution (taken to be $M_1=20$ in the simulations).

Defining, as in for example the interval censoring model,

$$\phi(u) = \int_{y\le u} a(y)\,dG(y),$$

TABLE A1 A comparison of variances, given by a simulation of 1,000 samples of size $n = 1000$ and the right-hand side of (A7). The bandwidth $h = 3.4$ and G is the distribution function of the Weibull distribution, truncated on the interval $[0, 20]$

t	Simulation variances	$n^{-3/7}E\,\theta_{t,h,G}(E,S,\Delta)^2$
2	0.001524376	0.001528899
3	0.002652881	0.002551415
4	0.003535091	0.003457335
5	0.004193696	0.004037131
6	0.004351735	0.004275926
7	0.004238654	0.004226677
8	0.004073332	0.003842444
9	0.003385165	0.003076003
10	0.002352065	0.002082613
11	0.001402003	0.001165108

we get:

$$\theta_{t,h,G}(e,s,\delta) = \delta\frac{\phi(s)}{G(s)} + (1-\delta)\frac{\phi(s)-\phi(s-e)}{G(s)-G(s-e)\}}, \tag{A2}$$

where we define $0/0 = 0$. Note that ϕ is absolutely continuous w.r.t. G and that $\phi(s) = 0$, $s \geq M_1$, since we assume, as usual, $a \in L_2^0(G)$, where $L_2^0(G)$ is the space of square integrable functions f w.r.t. dG, with the property $\int f(x)\,dG(x) = 0$.

In the present model, the infection time is uniform on $[0, E]$ and E is Uniform$[0, M]$. So we get the following integral equation for the estimation of the density if $w \in [0, M)$,

$$\begin{aligned}
&E[\theta_{t,h,G}(E,S,\Delta)|W=w]\\
&\quad= \int_{e\in[w,M]}\frac{1}{Me}\left\{\int_{s\in[w,e]}\frac{\phi(s)}{G(s)}\,ds\right\}\,de + \int_{e\in[w,M]}\frac{1}{Me}\left\{\int_{s\in[e,w+e]}\frac{\phi(s)-\phi(s-e)}{G(s)-G(s-e)}\,ds\right\}\,de\\
&\qquad+\int_{e\in(0,w]}\frac{1}{Me}\left\{\int_{s\in[w,w+e]}\frac{\phi(s)-\phi(s-e)}{G(s)-G(s-e)}\,ds\right\}\,de\\
&\quad= K_h(w-t).
\end{aligned} \tag{A3}$$

Differentiation w.r.t. w yields for the density estimate:

$$\begin{aligned}
&-\frac{\phi(w)}{MG(w)}\log(M/w) + \frac{1}{M}\int_{e=0}^{w}\frac{1}{e}\left\{\frac{\phi(w+e)-\phi(w)}{G(w+e)-G(w)} - \frac{\phi(w)-\phi(w-e)}{G(w)-G(w-e)}\right\}\,de\\
&\qquad+\frac{1}{M}\int_{e=w}^{M}\frac{1}{e}\frac{\phi(w+e)-\phi(w)}{G(w+e)-G(w)}\,de\\
&\quad= \frac{\partial}{\partial w}K_h(w-t),
\end{aligned} \tag{A4}$$

So we get the representation

$$\int K_h(t-y)\,d(\hat{G}_n - G_0)(y) = \int \theta_{t,h,\hat{G}_n}(e,s,\delta)\,dP_0(e,s,\delta),$$

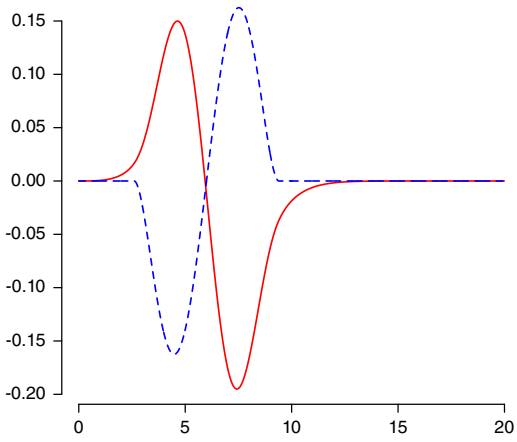

FIGURE A1 The function ϕ (red, solid), for $h = 3.4$, the triweight kernel K, and $t = 6$. The blue dashed curve is the function $w \mapsto \frac{\partial}{\partial w} K_h(w - t)$

where $\hat{G}_n$ is the MLE and ϕ solves (A3) for $G = \hat{G}_n$ (compare to (11.44), p. 331 of Groeneboom & Jongbloed, 2014).

This leads to

$$n^{2/7} \int K_h(t-y)\, d(\hat{G}_n - G_0)(y) \sim n^{2/7} \int \theta_{t,h,G_0}(e,s,\delta)(\mathbb{P}_n - P_0)(e,s,\delta), \tag{A5}$$

where θ_{t,h,G_0} is defined by (A2), where $G = G_0$, the underlying distribution function of the incubation time, and ϕ is the solution of the equation (A3) and satisfies $\phi(M_1) = 0$. Moreover, (A5) would imply:

$$n^{2/7} \int K_{h_n}(t-y)\, d(\hat{G}_n - G_0)(y) \xrightarrow{\mathcal{D}} N(0, \sigma^2), \tag{A6}$$

where

$$\sigma^2 = \lim_{n\to\infty} \mathrm{var} \left(n^{2/7} \int \theta_{t,h_n,G_0}(e,s,\delta)\, d\mathbb{P}_n(e,s,\delta) \right),$$

and $n^{1/7} h_n \to c > 0$. A picture of the function ϕ, solving (A3), is shown in Figure A1. This can be found by a simple iteration procedure for the integral equation (A4) or a matrix equation after discretization, which can also be found in Groeneboom (2020a).

Note that, letting $\Phi(s) = \int_0^s \phi(u)\, du$, and defining $0/0 = 0$.

$$\begin{aligned}
&E\, \theta_{t,h,G}(E,S,\Delta) \\
&= \frac{1}{M} \int_{s\le e} \frac{1}{e} \frac{\phi(s)}{G(s)} G(s)\, de\, ds + \frac{1}{M} \int_{e<s} \frac{1}{e} \left\{ \frac{\phi(s) - \phi(s-e)}{G(s) - G(s-e)} \right\} \{G(s) - G(s-e)\}\, de\, ds \\
&= \frac{1}{M} \int_{s\le e} \frac{1}{e} \phi(s)\, de\, ds + \frac{1}{M} \int_{e<s} \frac{1}{e} \{\phi(s) - \phi(s-e)\}\, de\, ds \\
&= \frac{1}{M} \int_{e=0}^{M} \frac{1}{e} \Phi(e)\, de + \frac{1}{M} \int_{e=0}^{M} \frac{1}{e} \{\Phi(M_1) - \Phi(e) - \Phi(M_1) + \Phi(0)\}\, de \\
&= \frac{1}{M} \int_{e=0}^{M} \frac{1}{e} \Phi(e)\, de - \frac{1}{M} \int_{e=0}^{M} \frac{1}{e} \Phi(e)\, de = 0,
\end{aligned}$$

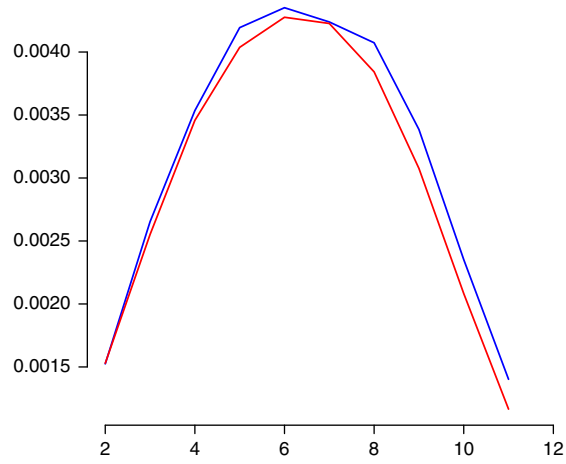

FIGURE A2 Plot of Table A1. The variances in the simulation are given by the blue curve and the red curve gives the values $n^{-3/7}E\,\theta_{t,h,G}(E,S,\Delta)^2$, for $t=2,3,\ldots,11$, $h=3.4$, where G is the truncated Weibull distribution function

using $\phi(s)=0$, $s\geq M_1$, were M_1 is the upper bound of the support of the density of the incubation time. Note that we use $M\geq M_1$, where $[0,M]$ is the interval containing the exit times (assumed to be uniformly distributed on $[0,M]$ in the simulation experiment). In Figure A1 we have $M_1=20$ and $M=30$. For the asymptotic variance, we get:

$$\begin{aligned}
&E\,\theta_{t,h,G}(E,S,\Delta)^2\\
&\quad=\frac{1}{M}\int_{s\leq e}\frac{1}{e}\frac{\phi(s)^2}{G(s)^2}G(s)\,de\,ds+\frac{1}{M}\int_{s>e}\frac{1}{e}\left\{\frac{\phi(s)-\phi(s-e)}{G(s)-G(s-e)}\right\}^2\{G(s)-G(s-e)\}\,de\,ds\\
&\quad=\frac{1}{M}\int_{s\leq e}\frac{\phi(s)^2}{e\,G(s)}\,de\,ds+\frac{1}{M}\int_{s>e}\frac{\{\phi(s)-\phi(s-e)\}^2}{e\{G(s)-G(s-e)\}}\,de\,ds.
\end{aligned}\tag{A7}$$

Note that:

$$\begin{aligned}
&\mathrm{var}\left(n^{2/7}\int K_h(t-y)\,d\hat{G}_n(y)\right)\sim\mathrm{var}\left(n^{2/7}\int\theta_{t,h,G_0}(e,s,\delta)\,d\mathbb{P}_n(e,s,\delta)\right)\\
&\quad=n^{-3/7}E\,\theta_{t,h,G}(E,S,\Delta)^2.
\end{aligned}$$

A table for the variances of the density estimates at $t=2,3,\ldots,11$, as computed from 1,000 samples of size $n=1,000$ and from $n^{-3/7}E\,\theta_{t,h,G}(E,S,\Delta)^2$, as given by (A7). The table is given graphically in Figure A2.