Estimation of the incubation time distribution for COVID-19

Piet Groeneboom

Delft University of Technology, Building 28, Van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands
e-mail: P.Groeneboom@tudelft.nl

Abstract: We consider nonparametric estimation of the incubation time distribution of COVID-19.

1. Introduction

The Dutch Centre for Infectious Disease Control (Dutch: RIVM) analyzes in [1] 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 of an interval for the infection interval, for which we take (looking back) 0 as the left point for the ith individual (see [2]), $S_i \geq E_i$ is the time where the person becomes symptomatic (note that $S_i = E_i$ is possible), 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 [1] (see also [9]), 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

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

where G is the incubation time distribution function.

The algorithms we used for analyzing the data set can be found on [3]. We describe the data files given there. The original data file is `data_Wuhan_tsv`, which gives details on persons in the sample and which can be found in [1]. This was transformed into a data file `transformed_data_Wuhan.txt`, consisting of three columns, giving, respectively, the arrivals in and departures from Wuhan and the time the person became symptomatic. If the arrival time was unknown, 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 colums 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 $S_i - E_i = 0$.

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

$$G(x) = G_{a,b}(x) = 1 - \exp\{-bx^a\}, \qquad x > 0, \tag{2}$$

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

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

we get as our maximum likelihood estimaters of the parameters a and b:

$$\hat{a} = 3.03514, \qquad \hat{b} = 0.002619. \tag{3}$$

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 (3) were produced by the `R` script in [3], using the Hooke-Jeeves algorithm. For a convergence proof of the Hooke-Jeeves algorithm and interesting further discussion of the pattern search algorithms, see [8] and [11].

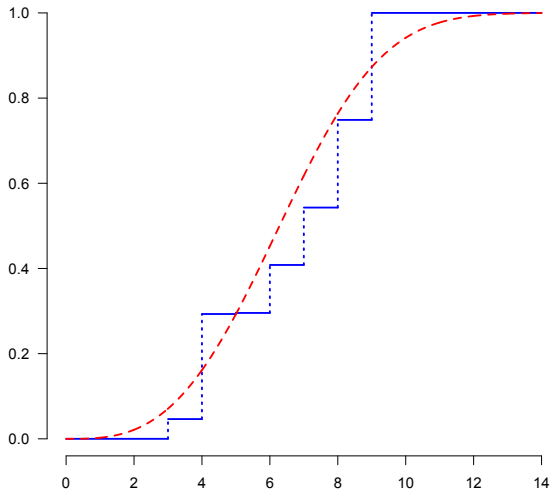


Fig 1: The nonparametric maximum likelihood estimate (MLE) $\hat{G}_n$ of the incubation time distribution function (blue), and the MLE using the Weibull distribution (red, dashed), for the data set analyzed in [1].

2. Algorithms for computing the nonparametric maximum likelihood estimator

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

$$p_j' = p_j n^{-1} \sum_{i=1}^{n} 1_{\{j \in (S_i - E_i, S_i]\}} \Big/ \sum_{k \in (S_i - E_i, S_i]} p_k \,, \tag{4}$$

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 [3].

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

Number of days	p_i
3	0.0463850922
4	0.2466837048
5	0.0024858945
6	0.1126655228
7	0.1347501680
8	0.2058210187
9	0.2512085991

TABLE 2
Probability masses of the nonparametric MLE.

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.

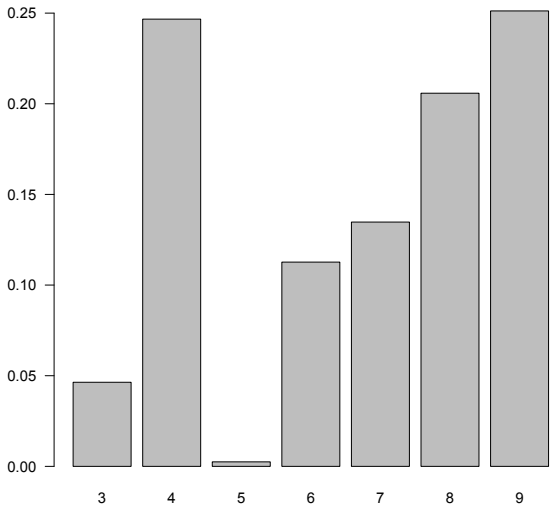


Fig 2: Bar chart of the probability masses of the nonparametric MLE

The iteration steps (4) folllow from the so-called self-consistency equations, which are derived by differentiating the criterion function

$$n^{-1}\sum_{i=1}^{n}\log\left\{\sum_{j\in(S_i-E_i,S_i]}p_j\right\}-\lambda\left\{\sum_{j=1}^{m}p_j-1\right\}, \tag{5}$$

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

$$\sum_{j=1}^{m}p_j=1. \tag{6}$$

This yields

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

and multiplying these relations with p_j and summing over j yields $\lambda = 1$, using the side condition (6). But the relations (7) only hold for the *active* (in this case 7) parameters $p_i > 0$ of the solution; in the iterations (4) the inactive parameters p_i will tend to zero. For more details, see, e.g., [7], 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 [7], Section 7.3.

The log likelihood is of type:

$$f(\boldsymbol{y}) = \sum_{i=1}^{m} k_i \log\left(G(U_i) - G(T_i)\right), \tag{8}$$

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

$$(T_i, U_i) = (0, V_i + W_i)\, 1_{\{V_i+W_i \le E_i\}} + (V_i + W_i - E_i, V_i + W_i)\, 1_{\{V_i+W_i > E_i\}} \qquad i = 1, \ldots, n, \tag{9}$$

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

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

For the Wuhan data set 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:

$$f(\boldsymbol{y}) = \sum_{0 \le i < j \le 7} N_{ij} \log\left(y_j - y_i\right), \tag{10}$$

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 (8) under the restriction $0 < y_1 \le \cdots \le y_6$; 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:

$$\sum_{j=i}^{6} \frac{\partial}{\partial y_j} f(\boldsymbol{y}) \le 0, \qquad i = 1, \ldots, 6, \tag{11}$$

and

$$\sum_{i=1}^{6} y_i \frac{\partial}{\partial y_i} f(\boldsymbol{y}) = 0, \tag{12}$$

where f is defined by (8). Since the values y_i are strictly between 0 and 1, (12) can only hold if also

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

and we can therefore turn (11) into

$$\sum_{j=1}^{i} \frac{\partial}{\partial y_j} f(\boldsymbol{y}) \geq 0, \qquad i = 1, \ldots, 6. \tag{13}$$

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

$$\sum_{j=1}^{i} \left(w_j, \frac{\partial}{\partial y_j} f(\boldsymbol{y}) + w_j y_j \right), \qquad i = 1, \ldots, 6, \tag{14}$$

where

$$w_j = -\frac{\partial^2}{\partial y_j^2} f(\boldsymbol{y}). \qquad j = 1, \ldots, 6. \tag{15}$$

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 (which does not happen in the present case). For more theory, see [7].

As in [7], section 1.2, we can compute the smoothed maximum likelihood estimator (SMLE) and also an estimate of the density. The SMLE is defined by

$$\tilde{G}_{nh}(t) = \int \mathbb{K}((t-y)/h)\, d\hat{G}_n(y), \tag{16}$$

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

$$\mathbb{K}(x) = \int_{-\infty}^{x} K(u)\, du. \tag{17}$$

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

$$K(u) = \frac{35}{32} \left(1 - u^2\right)^3 1_{[-1,1]}(u). \tag{18}$$

We estimate the density by

$$\tilde{g}_{nh}(t) = h^{-1} \int K((t-y)/h)\, d\hat{G}_n(y). \tag{19}$$

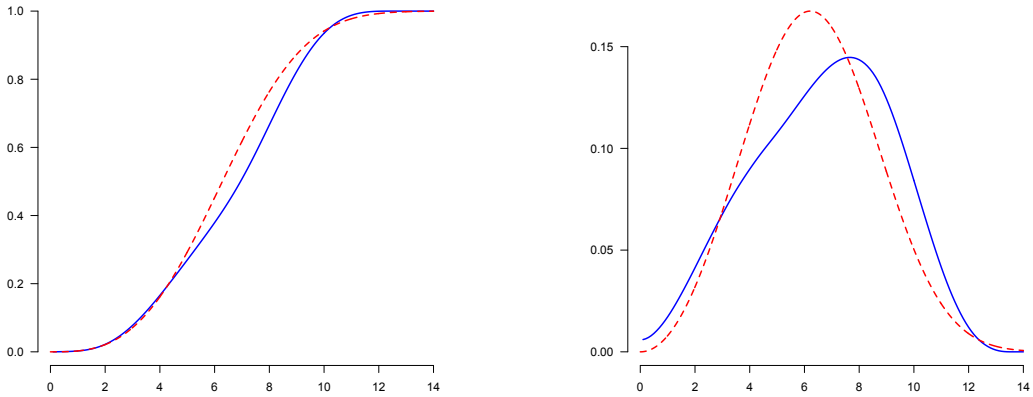
For the present analysis we took $h = 3.6$ in (16) and $h = 4.6$ in (19); these bandwidths were chosen by a bootstrap method, explained in Section 3. The resulting estimates are shown in Figure 3.

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 $[1, 42]$ represent the exit times ("length of stay in Wuhan"). 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 unkown) incubation time. Our observations are the pairs (9).

To determine the bandwidth h of our density estimator

$$\hat{g}_{nh}(t) = \int K_h(t-y)\, d\hat{G}_n(y), \tag{20}$$



(a) (b)

Fig 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 data set analyzed in [1] and (b): the smoothed nonparametric maximum likelihood estimate 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 [1].

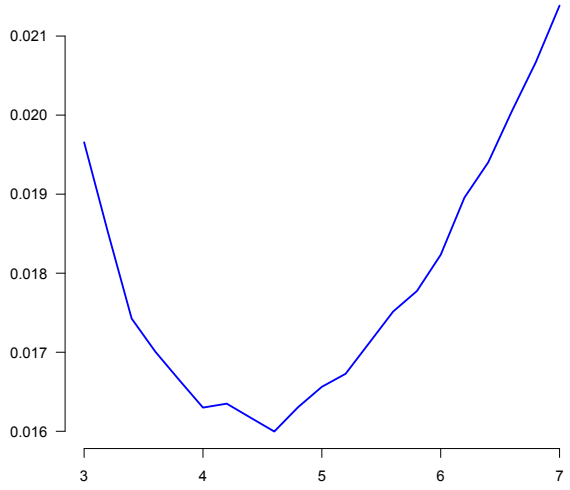


Fig 4: $\hat{\mathrm{MSE}}_g(h)$, given by (21), as function of h.

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 [10].

We take $B = 10,000$ bootstrap samples of observations (E_i, U_i^*), corresponding to the observations (E_i, U_i), where the E_i are the (actual) exit times if $E_i < S_i$, and where $E := S_i$ if the person got symptoms when still in Wuhan. The U_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 rejection sampling for a

fixed h_0, for which we took $h_0 = 4$ in the present case. Note that we keep the E_i the same as in the original sample, somewhat analogously to the procedure followed in [10], which relieves us from the duty to estimate the exit time distribution.

Next we computed

$$\hat{\mathrm{MSE}}_g(h) = B^{-1} \sum_{b=1}^{B} \int \{\hat{g}^*_{nh}(x) - \hat{g}_{nh_0}(x)\}^2 \, dx. \tag{21}$$

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 [10].

Similarly, we computed

$$\hat{\mathrm{MSE}}_G(h) = B^{-1} \sum_{b=1}^{B} \int \left\{\hat{G}^*_{nh}(x) - \hat{G}_{nh_0}(x)\right\}^2 \, dx, \tag{22}$$

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 on [3].

The method used here is called the "smoothed bootstrap". Another option is to use the ordinary bootstrap with smaller sample sizes (since otherwise for this method the bias is not estimated in the right way). This method was for example applied in [6] and implemented in the `R` package `curstatCI` ([5]) for the current status model. We tried this method out, but found its behavior not to be satisfactory in the present case, possibly because of the small sample size.

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 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 the indicator

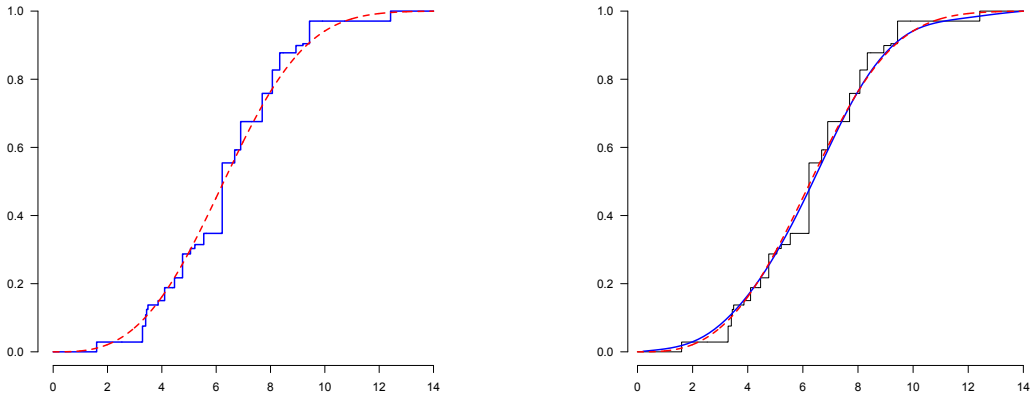
$$\Delta = \begin{cases} 1, & \text{if } S \leq E, \\ 0, & \text{if } S > E, \end{cases} \tag{23}$$

where E is again the exit time and S is the time of becoming symptomatic. A sample would contain triples (E_i, S_i, Δ_i), where Δ_i is defined by (23), with E_i and S_i instead of E and S, respectively.

We consider the following simulation experiment. E_i is a Uniform random variable on the interval $[0, 43]$, 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 Weibull (a, b), where a and b have the same values as the estimates $\hat{a}$ and $\hat{b}$ in (3), respectively. This means:

$$S_i = V_i + W_i,$$

and we can either have $S_i \leq E_i$ $(\Delta_i = 1)$ or $S_i > E_i$ $(\Delta_i = 0)$. The MLE of the incubation time, based on the triples $E_i, S_i, \Delta_i)$, where the continuous variable E_i and S_i are known, looks rather different from the MLE based on the discrete observations shown in Figure 1. An example of such an MLE is shown in Figure 5 for a sample of $n = 1000$. 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.



(a) (b)

Fig 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 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 Weibull distribution function (red, dashed). The bandwidyth of the SMLE is $h = 3$.

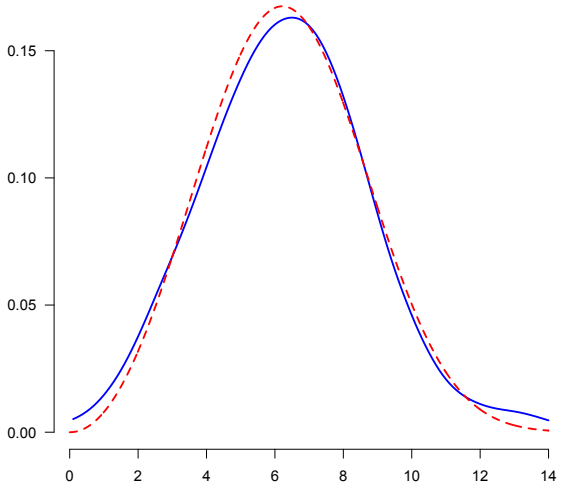


Fig 6: The estimate of the density of the incubation time (blue, solid), based on a sample of size $n = 1000$ and bandwidth $h = 3.6$. The red dashed curve is the Weibull density with parameters a and b of (3).

5. Concluding remarks

We offered an alternative nonparametric approach to the estimation of the incubation time distribution which was estimated by parametric methods in [1] for a data set 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 [1], 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 models where the observations are not discretized at days and discussed rates of convergence in that case. The present paper can be considered to be the technical companion of the column [4]. All numerical computations are given as `R` scripts in [3].

References

[1] Jantien A. Backer, Don Klinkenberg, and Jacco Wallinga. Incubation period of 2019 novel coronavirus (2019-nCov) infections among travellers from Wuhan, China, 20-28 january 2020. *Euro Surveill.*, 25, 2020. URL https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2020.25.5.2000062.

[2] Tom Britton and Gianpaolo Scalia Tomba. Estimation in emerging epidemics: bases and remedies. *J. R. Soc. Interface*, 16, 2019.

[3] Piet Groeneboom. Incubationtime. https://github.com/pietg/incubationtime, 2020.

[4] Piet Groeneboom. The Netherlands in Times of Corona (in Dutch). *Nieuw Archief voor Wiskunde*, 21:181–184, 2020. URL http://www.nieuwarchief.nl/serie5/pdf/naw5-2020-21-3-181.pdf.

[5] Piet Groeneboom and Kim Hendrickx. curstatCI. R package, 2017. URL https://cran.r-project.org/web/packages/curstatCI/index.html. Version 0.1.1.

[6] Piet Groeneboom and Kim Hendrickx. The nonparametric bootstrap for the current status model. *Electron. J. Stat.*, 11 (2):3446–3484, 2017. ISSN 1935-7524. .

[7] Piet Groeneboom and Geurt Jongbloed. *Nonparametric Estimation under Shape Constraints*. Cambridge Univ. Press, Cambridge, 2014.

[8] Tamara G. Kolda, Robert Michael Lewis, and Virginia Torczon. Optimization by direct search: new perspectives on some classical and modern methods. *SIAM Rev.*, 45(3):385–482, 2003. ISSN 0036-1445. . URL https://doi.org/10.1137/S003614450242889.

[9] Nicholas G. Reich, Justin Lessler, Derek A. T. Cummings, and Ron Brookmeyer. Estimating incubation period distributions with coarse data. *Stat. Med.*, 28(22):2769–2784, 2009. ISSN 0277-6715. . URL https://doi.org/10.1002/sim.3659.

[10] Bodhisattva Sen and Gongjun Xu. Model based bootstrap methods for interval censored data. *Comput. Statist. Data Anal.*, 81:121–129, 2015. ISSN 0167-9473. . URL http://dx.doi.org/10.1016/j.csda.2014.07.007.

[11] Virginia Torczon. On the convergence of pattern search algorithms. *SIAM J. Optim.*, 7(1):1–25, 1997. ISSN 1052-6234. . URL http://dx.doi.org/10.1137/S1052623493250780.