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.

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]), and F_i would be the distribution function of the time of a possible contact with an infector.

It is clear that, without further assumptions, g and F_i are not identifiable. To remedy this, we assume, as in [1] (see also [4]), 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

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

where G is the incubation time distribution function.

Maximizing this log likelihood is an isotonic regression problem, which can be solved by specific isotonic methods, but we can also use the EM algorithm (see [3]). 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

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

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

$$\hat a = 3.03514, \qquad \hat b = 0.002619.$$

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

$$p_j' = p_j n^{-1} \sum_{i=1}^n 1_{\{X_j\in(S_i-E_i,S_i]\}} \Big/ \sum_{X_k\in(S_i-E_i,S_i]} p_k\,,$$

where the X_i are the possible points of mass for the incubation time distribution.

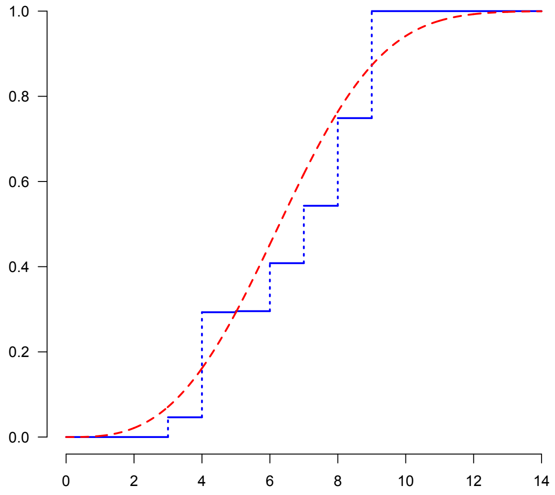

Fig 1: The nonparametric maximum likelihood estimate (MLE) $\hat{F}_n$ of the incubation time distribution function (blue), and the MLE using the Weibull distribution (red, dashed), for the data set analyzed in [1].

Maximizing the log likelihood is an isotonic regression problem. In the Wuhan data 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 becomes:

$$f(\boldsymbol{y}) = \sum_{0 \le i < j \le 7} N_{ij} \log\left(y_j - y_i\right), \tag{2}$$

where $y_i = G(i+2)$, $i = 0, \dots, 7$, $y_0 = 0$, $y_7 = 1$, 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 (2) under the restriction $0 < y_1 \le \cdots \le y_6 < 1$. Let $\boldsymbol{y} = (y_1, \dots, y_6)^T$. The (Fenchel) sufficient and necessary conditions for the solution are:

$$\sum_{j=i}^{6} \frac{\partial}{\partial y_j} f(\boldsymbol{y}) \le 0, \qquad i = 1, \dots, 6, \tag{3}$$

and

$$\sum_{i=1}^{6} y_i \frac{\partial}{\partial y_i} f(\boldsymbol{y}) = 0, \tag{4}$$

where f is defined by (2). Since the values y_i are strictly between 0 and 1, (4) can only hold if also

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

and we can therefore turn (5) into

$$\sum_{j=1}^{i}\frac{\partial}{\partial y_j}f(\boldsymbol{y})\ge 0,\qquad i=1,\dots,6. \tag{5}$$

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.

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

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

where

$$w_j=-\frac{\partial^2}{\partial y_j^2}f(\boldsymbol{y}).\qquad j=1,\dots,6. \tag{7}$$

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 [3].

As in [3], section 1.2, we can compute the smoothed maximum likelihood estimator (SMLE) and also an estimate of the density. The SMLE is defined by

$$\tilde F_{nh}(t)=\int \mathbb{K}((t-y)/h)\,d\hat F_n(y), \tag{8}$$

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

$$\mathbb{K}(x)=\int_{-\infty}^{x}K(u)\,du. \tag{9}$$

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

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

We estimate of the density by

$$\tilde f_{nh}(t)=h^{-1}\int K((t-y)/h)\,d\hat F_n(y). \tag{10}$$

For the present analysis we took $h=3$ in (8) and $h=4$ in (10) (as a side remark: generally, the bandwidth h has to be bigger in (10) than in (9)). The resulting estimates are shown in Figures 2 and 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 and Geurt Jongbloed. *Nonparametric Estimation under Shape Constraints.* Cambridge Univ. Press, Cambridge, 2014.

[4] 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.

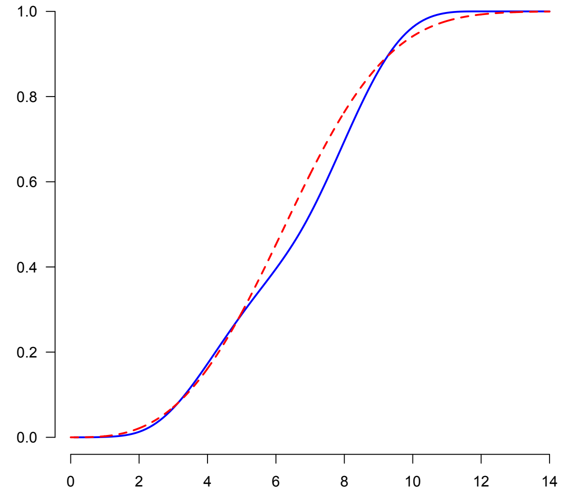

Fig 2: The smoothed nonparametric maximum likelihood estimate (SMLE) of the incubation time distribution function (blue), and the MLE using the Weibull distribution (red, dashed), for the data set analyzed in [1].

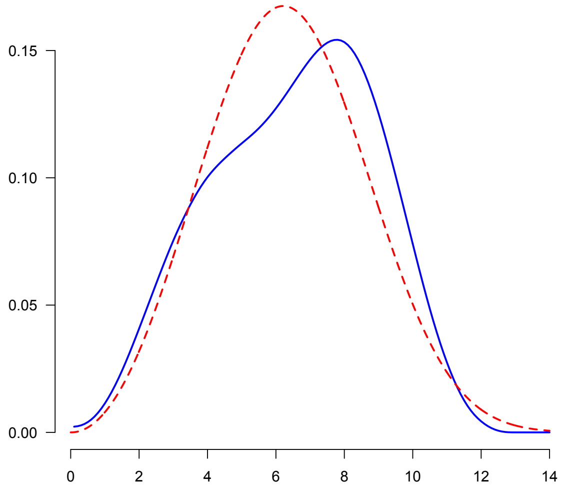

Fig 3: The smoothed nonparametric maximum likelihood estimate of the incubation time density function (blue), and the MLE of the density using the Weibull distribution (red, dashed), for the data set analyzed in [1].