

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. 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 i th 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

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

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

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

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

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

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. The resulting MLE \hat{F}_n is shown in Figure 1, together with the MLE, assuming that the incubation time distribution is a Weibull

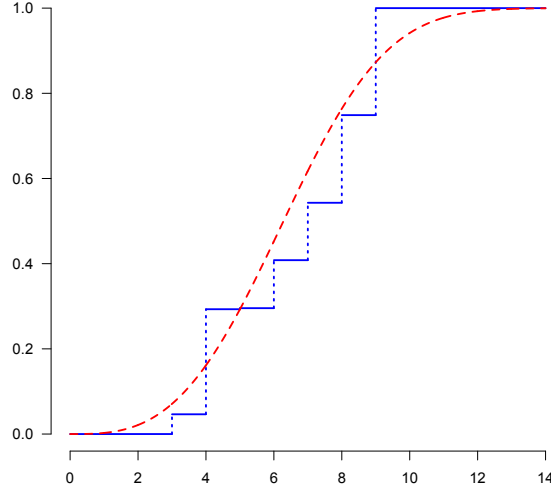


Fig 1: The nonparametric maximum likelihood estimate (MLE) \hat{F}_n of the incubation time distribution function (blue), and the MLE using the Weibull distribution (red, dashed), for the data set analyzed in [1].

distribution. The estimation of the first moment of the distribution of the incubation time is (in three decimals) equal to 6.338 if one assumes that the distribution is Weibull (see [1]), but equal to 6.665 based on the nonparametric MLE.

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

$$\tilde{F}_{nh}(t) = \int \mathbb{K}((x - y)/h) d\hat{F}_n(y), \quad (2)$$

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

$$\mathbb{K}(x) = \int_{-\infty}^x K(u) du. \quad (3)$$

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

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

We estimate of the density by

$$\tilde{f}_{nh}(t) = h^{-1} \int K((x - y)/h) d\hat{F}_n(y). \quad (4)$$

For the present analysis we took $h = 3$ in (2) and $h = 4$ in (4) (as a side remark: generally, the bandwidth h has to be bigger in (4) than in (3)). 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>.

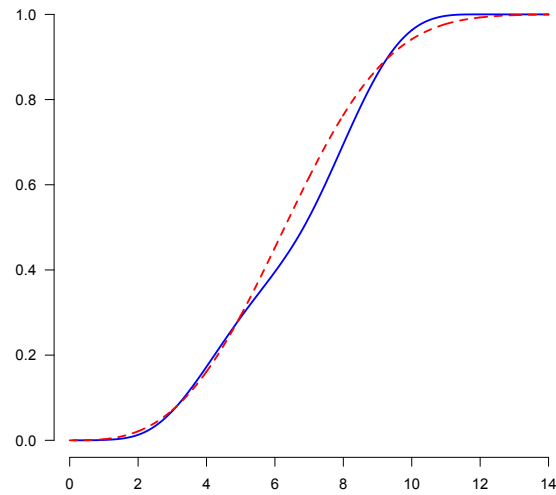


Fig 2: The smoothed nonparametric maximum likelihood estimate (SMLE) of the incubation time distribution function (blue), and the MLE using the Weibull distribution (red, dashed), for the data set analyzed in [1].

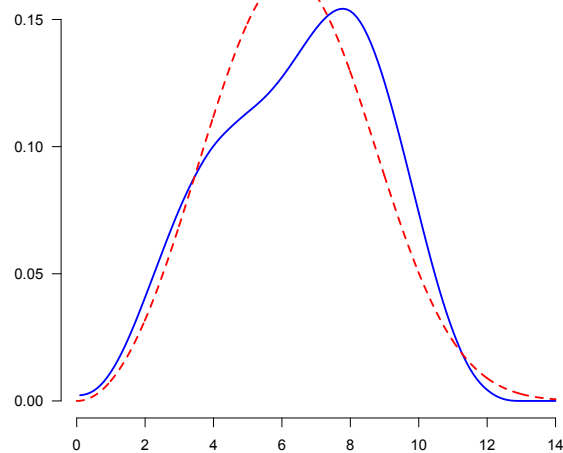


Fig 3: The smoothed nonparametric maximum likelihood estimate of the incubation time density function (blue), and the MLE of the density using the Weibull distribution (red, dashed), for the data set analyzed in [1].

- [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>.