Estimation of the incubation time

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.

1. Nomparametric estimation of the incubation time distribution

In [2] the probability of infection p and the incubation time distribution are simultaneously estimated from a part of the log likelihood in the model, which can be written

$$\sum_{i=1}^{n} \log \left\{ \sum_{j=0}^{N_i} p(1-p)^j g(S_i - E_{ij}) \right\},$$

(see (5.1) in [2]), where n is the size of the sample, N_i is geometrically distributed with parameter p, the E_{ij} are times of contact with possible infectors before symptom time S_i, where $E_{i0} = 0$, and g is the density of the incubation time, which is taken gamma or log normal for the simulation study in the supplementary material with [2]). Note that N_i, S_i and E_{ij} are all random variables.

We consider the situation where we drop the assumption that g is either gamma or log normal, and consider the possibility of estimating this nonparametrically. In this way we get a so-called semi-parametric model, where p is the probability of infection and the density g is the nonparametric part. As in [2] we try to estimate p and g simultaneously.

There are several ways to proceed, but we consider now a simple model, where we assume that g is a piecewise constant density on the points $S_i - E_{ij}$, $i = 1, \ldots, n$, $j = 0, \ldots, N_i$, on which we condition. This gives a density

$$g = \sum_{k=1}^{m} b_k 1_{(a_{k-1}, a_k]}, \qquad k = 1, \ldots, m,$$

where $a_0 = 0$, a_m is the largest point S_i and where

$$\sum_{k=1}^{m} b_k \, (a_k - a_{k-1}) = 1. \tag{1.1}$$

The side condition (1.1) can be incorporated by adding a Lagrange term. The criterion function then becomes:

$$\sum_{i=1}^{n} \log \left\{ \sum_{j=0}^{N_i} p(1-p)^j \sum_{k=1}^{m} b_j 1_{(a_{k-1}, a_k]} \, (S_i - E_{ij}) \right\} - \lambda \sum_{k=1}^{m} b_k \, (a_k - a_{k-1}) \,,$$

for p and a vector $\boldsymbol{b}$ of nonnegative parameters b_j, $i = 1, \ldots, m$. Differentiating w.r.t. b_k yields:

$$\frac{\partial}{\partial b_k} L(p, \boldsymbol{b}) = \sum_{i=0}^{n} \frac{\sum_{j=0}^{N_i} p(1-p)^j 1_{(a_{k-1}, a_k]} \, (S_i - E_{ij})}{\sum_{j=0}^{N_i} p(1-p)^j \sum_{k=1}^{m} b_j 1_{(a_{k-1}, a_k]} \, (S_i - E_{ij})} - \lambda \, (a_k - a_{k-1}) \,.$$

Putting this equal to zero, multiplying with b_k and summing over k we obtain

$$\sum_{k=1}^{m} b_k \frac{\partial}{\partial b_k} L(p, \boldsymbol{b}) = n - \lambda = 0,$$

so our criterion function becomes

$$L(p, \boldsymbol{b}) = \sum_{i=1}^{n} \log \left\{ \sum_{j=0}^{N_i} p(1-p)^j \sum_{k=1}^{m} b_j 1_{(a_{k-1}, a_k]} (S_i - E_{ij}) \right\} - n \sum_{k=1}^{m} b_k (a_k - a_{k-1}) .$$

The so-called score equations become:

$$\frac{\partial}{\partial p} L(p, \boldsymbol{b}) = \frac{n}{p} - \sum_{i=1}^{n} 1_{\{N_i > 0\}} \frac{\sum_{j=1}^{N_i} j(1-p)^{j-1} \sum_{k=1}^{m} b_k 1_{(a_{k-1}, a_k]} (S_i - E_{ij})}{\sum_{j=0}^{N_i} (1-p)^j \sum_{k=1}^{m} b_k 1_{(a_{k-1}, a_k]} (S_i - E_{ij})} = 0,$$

and

$$\frac{\partial}{\partial b_k} L(p, \boldsymbol{b}) = \sum_{i=1}^{n} \frac{\sum_{j=0}^{N_i} p(1-p)^j 1_{(a_{k-1}, a_k]} (S_i - E_{ij})}{\sum_{j=0}^{N_i} p(1-p)^j \sum_{l=1}^{m} b_l 1_{(a_{l-1}, a_l]} (S_i - E_{ij})} - n (a_k - a_{k-1}) = 0, \qquad k = 1, \ldots, m.$$

Under suitable conditions solving these equations will give a consistent estimate of the incubation time distribution. Note on the other hand that if in [2] the parametric distribution is misspecified (like a log normal distribution if the gamma distributions is the real underlying distribution), this will generally not be the case.

2. Computation

The self-consistency equations are:

$$p = n \Big/ \sum_{i=1}^{n} 1_{\{N_i > 0\}} \frac{\sum_{j=1}^{N_i} j(1-p)^{j-1} \sum_{k=1}^{m} b_k 1_{(a_{k-1}, a_k]} (S_i - E_{ij})}{\sum_{j=0}^{N_i} (1-p)^j \sum_{k=1}^{m} b_k 1_{(a_{k-1}, a_k]} (S_i - E_{ij})} ,$$

and

$$b_k = \frac{b_k}{n(a_k - a_{k-1})} \sum_{i=1}^{n} \frac{\sum_{j=0}^{N_i} (1-p)^j 1_{(a_{k-1}, a_k]} (S_i - E_{ij})}{\sum_{j=0}^{N_i} (1-p)^j \sum_{l=1}^{m} b_l 1_{(a_{l-1}, a_l]} (S_i - E_{ij})} , \qquad k = 1, \ldots, m.$$

It is clear that this leads to the iteration steps:

$$p' = n \Big/ \sum_{i=1}^{n} 1_{\{N_i > 0\}} \frac{\sum_{j=1}^{N_i} j(1-p)^{j-1} \sum_{k=1}^{m} b_k 1_{(a_{k-1}, a_k]} (S_i - E_{ij})}{\sum_{j=0}^{N_i} (1-p)^j \sum_{k=1}^{m} b_k 1_{(a_{k-1}, a_k]} (S_i - E_{ij})} ,$$

and

$$b'_k = \frac{b_k}{n(a_k - a_{k-1})} \sum_{i=1}^{n} \frac{\sum_{j=0}^{N_i} (1-p)^j 1_{(a_{k-1}, a_k]} (S_i - E_{ij})}{\sum_{j=0}^{N_i} (1-p)^j \sum_{l=1}^{m} b_l 1_{(a_{l-1}, a_l]} (S_i - E_{ij})} , \qquad k = 1, \ldots, m.$$

A picture of the estimate of the incubation time, obtained in this way for a sample of size $n = 1000$ is shown in Figure 1 below. The estimate of p (equal to $p = 0.5$ in the simulations) was $\hat{p} = 0.49371$. The points a_k were taken to be the equidistant points $3i$, $i = 1, \ldots, 9$, with a final interval, extending from 27 to the largent point S_i, which was in this case equal to 200.07. We started the iterations with a uniform density on $[0, a_{10}]$ and a value of p equal to 0.75. The latter starting value of p was also taken as a starting point of the maximum likelihood estimates in the supplementary material of [2].

3. First observation is known to be in an interval

If the only thing we know about the start of the incubation time is that it belongs to an interval, the log likelihood for one observation is:

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

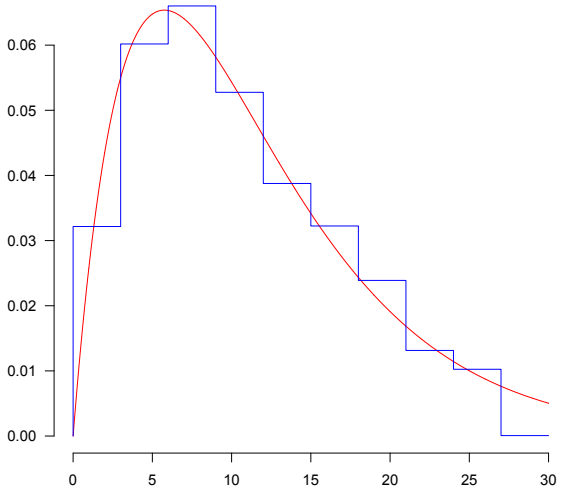

Fig 1: The estimate of the incubation time density (blue) for the simulation model in the supplementary material to [2], using the gamma density (red) for the incubation time density in the simulations and sample size $n = 1000$.

Here E_i would be the upper bound of an interval for E_{i1} in section 1, in the situation that there are no other possible starting points of the incubation interval, 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], that F_i is the uniform distribution on $[0, E_i]$. As before, conditioning on the E_i and S_i, 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

$$\sum_{i=1}^{n} \log \left\{ \int_{t=0}^{E_i} g(S_i - t)\, dt \right\} = \sum_{i=1}^{n} \log \{G(S_i) - G(S_i - E_i)\} \tag{3.1}$$

where G is the incubation time distribution function.

Maximizing this log likelihood is an isotonic regression problem. In the Wuhan data it can be checked that, without loss of generality, $G(i) = 0$, $i \leq 2$, and $G(i) = 1$, $i \geq 9$, 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:

$$\begin{aligned} f(y_1, \ldots, y_6) = {} & \log y_1 + 3 \log y_2 + 4 \log y_3 + 2 \log y_6 + 2 \log(y_2 - y_1) + \log(y_3 - y_1) + \log(y_4 - y_3) \\ & + \log(y_4 - y_2) + \log(y_5 - y_4) + \log(y_5 - y_2) + 2 \log(y_6 - y_3) + \log(y_6 - y_5) \\ & + 9 \log(1 - y_1) + 4 \log(1 - y_2) + 3 \log(1 - y_3) \\ & + 6 \log(1 - y_4) + 3 \log(1 - y_5) + 3 \log(1 - y_6), \end{aligned} \tag{3.2}$$

where $y_i = G(i + 2)$. We have to maximize (3.2) under the restriction $0 < y_1 \leq \cdots \leq y_6 < 1$. Let $\boldsymbol{y} = (y_1, \ldots, y_6)^T$. The (Fenchel) sufficient and necessary conditions for the solution are:

$$\sum_{i=j}^{6} \frac{\partial}{\partial y_i} f(\boldsymbol{y}) \leq 0, \qquad j = 1, \ldots, 6, \tag{3.3}$$

and

$$\sum_{i=1}^{6} y_i \frac{\partial}{\partial y_i} f(\boldsymbol{y}) = 0. \tag{3.4}$$

Since the values y_i are strictly between 0 and 1, (3.4) can only hold if also

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

and we can therefore turn (3.5) into

$$\sum_{i=1}^{j} \frac{\partial}{\partial y_i} f(\boldsymbol{y}) \geq 0, \qquad j = 1, \ldots, 6. \tag{3.5}$$

The 6 derivatives are given by:

$$\frac{\partial}{\partial y_1} f(\boldsymbol{y}) = \frac{1}{y_1} - \frac{9}{1-y_1} - \frac{2}{y_2 - y_1} - \frac{1}{y_3 - y_1},$$

$$\frac{\partial}{\partial y_2} f(\boldsymbol{y}) = \frac{3}{y_2} - \frac{4}{1-y_2} + \frac{2}{y_2 - y_1} - \frac{1}{y_4 - y_2} - \frac{1}{y_5 - y_2},$$

$$\frac{\partial}{\partial y_3} f(\boldsymbol{y}) = \frac{4}{y_3} - \frac{3}{1-y_3} + \frac{1}{y_3 - y_1} - \frac{1}{y_4 - y_3} - \frac{1}{y_6 - y_3},$$

$$\frac{\partial}{\partial y_4} f(\boldsymbol{y}) = \frac{1}{y_4 - y_2} - \frac{6}{1-y_4} + \frac{1}{y_4 - y_3} - \frac{1}{y_5 - y_4},$$

$$\frac{\partial}{\partial y_5} f(\boldsymbol{y}) = \frac{1}{y_5 - y_2} - \frac{3}{1-y_5} + \frac{1}{y_5 - y_4} - \frac{1}{y_6 - y_5},$$

and

$$\frac{\partial}{\partial y_6} f(\boldsymbol{y}) = \frac{2}{y_6} - \frac{3}{1-y_6} + \frac{2}{y_6 - y_3} + \frac{1}{y_6 - y_5}.$$

The optimization problem is equivalent with the optimization problem for the interval censoring, case 2, model, see [4], section 7.3 and can be solved by, e.g., the EM algorithm or the iterative convex minorant algorithm. R-scripts for both methods are given 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 Surveillance*, 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 and Geurt Jongbloed. *Nonparametric Estimation under Shape Constraints*. Cambridge Univ. Press, Cambridge, 2014.