Biometrics 000, 000–000 DOI: 000
000 0000

Supporting information for "Determination and estimation of optimal quarantine duration for infectious diseases with application to data analysis of COVID-19"

Ruoyu Wang[1,2], and Qihua Wang[1,2,*]

[1]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.

[2]University of Chinese Academy of Sciences, Beijing 100049, China.

[*]*email:* qhwang@amss.ac.cn

This paper has been submitted for consideration for publication in *Biometrics*

1. Proof of Theorem 1

Proof. By Condition 1, we have $\lim_{y\to\infty} f_1(y \mid x) = 0$ for any $x \in \mathcal{X}$ and hence $t_c(x)$ is continuous and strictly monontonous with respect to (w.r.t.) c for $0 < c \leq c^*$ and $f_1(y \mid x)$ is continuous and strictly monontonous with respect to y. Thus $\mathbb{E}_1[F_1(t_c(X) \mid X)]$ is continuous and strictly monontonous with respect to c for $0 < c \leq c^*$ and $\lim_{c\to 0} \mathbb{E}_1[F_1(t_c(X) \mid X)] = 1$ by dominated convergence theorem. Since $\mathbb{E}_1[F_1(t_{c^*}(X) \mid X)] \leq 1 - \epsilon$, by intermediate value theorem, there is a constant $0 < c_0 \leq c^*$ such that $\mathbb{E}_1[F_1(t_{c_0}(X) \mid X)] = 1 - \epsilon$. For any rule function $t(\cdot)$, define the lagrange problem

$$\begin{aligned}
&\mathbb{E}_0 t(X) - \frac{1}{c_0}\mathbb{P}_1(Y \leq t(X)) \\
&= \mathbb{E}_0 t(X) - \frac{1}{c_0}\mathbb{E}_1[F_1(t(X) \mid X)] \\
&= \int t(x) f_0(x) d\mu(x) - \frac{1}{c_0}\int F_1(t(x) \mid x) f_1(x) d\mu(x). \qquad \text{(S1)}
\end{aligned}$$

Suppose $t_0(\cdot)$ is the minimum point of variation problem (S1). Then for any $x \in \mathcal{X}$, $t_0(x)$ satisfies the Euler's equation (Gelfand and Fomin, 1963)

$$f_0(x) - \frac{1}{c_0} f_1(t_0(x) \mid x) f_1(x) = 0$$

or equivalently

$$f_1(t_0(x) \mid x)\frac{f_1(x)}{f_0(x)} = c_0.$$

Because $0 < c_0 \leq c^*$ and $f_1(y \mid x)$ is either strictly monontonous or unimodal and piecewise strictly monontonous with respect to y, the set $C_x = \{y : f_1(y \mid x)\frac{f_1(x)}{f_0(x)} \geq c_0\}$ is either a single point or a closed interval. For any given $x \in \mathcal{X}$, let $t_+(x) = \sup C_x$ and $t_-(x) = \inf C_x$. Then $t_+(x) = t_{c_0}(x)$ and any solution of $f_1(y \mid x) f_1(x)/f_0(x) = c_0$ equals to $t_-(x)$ or $t_+(x)$. Hence for any $x \in \mathcal{X}$, $t_0(x) = t_-(x)$ or $t_0(x) = t_+(x)$. Define $\mathcal{X}_+ = \{x : t_0(x) = t_+(x)\}$ and

$\mathcal{X}_- = \{x : t_0(x) = t_-(x)\}$. Then

$$\begin{aligned} &\int t_0(x) f_0(x) d\mu(x) - \frac{1}{c_0} \int F_1(t_0(x) \mid x) f_1(x) d\mu(x) - \\ &\int t_+(x) f_0(x) d\mu(x) + \frac{1}{c_0} \int F_1(t_+(x) \mid x) f_1(x) d\mu(x) \\ = &\int_{\mathcal{X}_-} \Big[(t_0(x) - t_+(x)) f_0(x) + \frac{1}{c_0} \int_{t_0(x)}^{t_+(x)} f_1(y \mid x) f_1(x) dy \Big] d\mu(x) \\ = &\iint_H \frac{1}{c_0} (f_1(y \mid x) f_1(x) - f_0(x)) dy d\mu(x) \leq 0, \end{aligned}$$

where the last inequality follows from the fact that $t_0(\cdot)$ is a minimum point of problem (S1) and $H = \{(x, y) : x \in \mathcal{X}_-, t_0(x) < y < t_+(x)\}$.

On the other hand,

$$\frac{1}{c_0} f_1(y \mid x) f_1(x) - f_0(x) > 0$$

on H. This implies H is a null set and $t_0(\cdot) = t_+(\cdot) = t_{c_0}(\cdot)$ with probability one. Thus $t_{c_0}(\cdot)$ is the unique minimum point of problem (S1) and satisfies $\int F_1(t_{c_0}(x) \mid x) f_1(x) d\mu(x) = \mathbb{E}_1[F_1(t_{c_0}(X) \mid X)] = 1 - \epsilon$. Next, we show that $t_{c_0}(\cdot)$ is the unique minimum point of the primal problem

$$\min_t \int t(x) f_0(x) d\mu(x) \quad s.t. \quad \int F_1(t(x) \mid x) f_1(x) d\mu(x) \geq 1 - \epsilon. \tag{S2}$$

If $\widetilde{t}(\cdot)$ is a minimum point of the problem (S2), then $\int \widetilde{t}(x) f_0(x) d\mu(x) \leq \int t_{c_0}(x) f_0(x) d\mu(x)$ and $\int F_1(\widetilde{t}(x) \mid x) f_1(x) d\mu(x) \geq \int F_1(t_{c_0}(x) \mid x) f_1(x) d\mu(x) = 1 - \epsilon$. Thus

$$\begin{aligned} &\int \widetilde{t}(x) f_0(x) d\mu(x) - \frac{1}{c_0} \int F_1(\widetilde{t}(x) \mid x) f_1(x) d\mu(x) \\ \leq &\int t_{c_0}(x) f_0(x) d\mu(x) - \frac{1}{c_0} \int F_1(t_{c_0}(x) \mid x) f_1(x) d\mu(x). \end{aligned}$$

This implies $\widetilde{t}(\cdot)$ is a minimum point of problem (S1) and hence $\widetilde{t}(\cdot) = t_{c_0}(\cdot)$ with probability one since $t_{c_0}(\cdot)$ is the unique minimum point of problem (S1). This proves that $t_{c_0}(\cdot)$ is the unique minimum point of problem (S2) and hence problem (1) in the main part of this paper.

2. Convergence rate

2.1 *Conditions and results*

CONDITION 1: $f_0(x)$ is bounded away from zero and $f_1(y \mid x)$, $f_1(x)$, $f_0(x)$ are bounded from above.

CONDITION 2: There are some δ, M_1, M_2, $M_3 > 0$ such that (i) $\forall x \in \mathcal{X}$, $-M_1 \le f_1'(y \mid x) f_1(x)/f_0(x) \le -M_2$; (ii) $c_0 + M_2\delta \le c^*$; (iii) and $f_1(y \mid x) \ge M_3$ for all $y \in (t_{c_0}(x) - \delta, t_{c_0}(x) + \delta)$.

Let $e_{1n} = \sup_x |\widehat{f}_1(x) - f_1(x)|$, $e_{2n} = \sup_x |\widehat{f}_0(x) - f_0(x)|$, $e_{3n} = \sup_{y,x} |\widehat{f}_1(y \mid x) - f_1(y \mid x)|$ and $e_{4n} = \sup_{y,x} |\widehat{F}_1(y \mid x) - F_1(y \mid x)|$. The convergence rates of e_{jn}, $j = 1, 2, 3, 4$, are available in many statistic literatures (Hansen, 2008; van der Vaart, 1998). We establish the relationship among these convergence rates and the convergence rate of the resultant estimated optimal quarantine rule in the next theorem.

THEOREM 1: *Suppose that* $\max\{e_{1n}, e_{2n}, e_{3n}, e_{4n}, n^{-1/2}\} = O_P(r_n)$ *where* r_n *is a sequence of positive numbers that converges to zero, under the conditions of Theorem 1 and Conditions 1 and 2, we have*

$$\sup_x |\widehat{t}_{\rm opt}(x) - t_{c_0}(x)| = O_P(r_n).$$

The convergence rates of e_{jn}, for $j = 1, 2, 3, 4$, are often slower than or of the same order as $n^{-1/2}$. In these cases, the result of Theorem 1 demonstrate that the uniform convergence rate among $\widehat{t}_{\rm opt}(x)$ is the same as the slowest convergence rate of e_{1n}, e_{2n}, e_{3n} and e_{4n}. Thus in order to get an accurate estimation of the optimal quarantine rule, we only need to estimate $f_1(x)$, $f_0(x)$, $f_1(y \mid x)$ and $F_1(y \mid x)$ accurately.

2.2 *Proof of Theorem 1*

Proof. First, note that

$$\begin{aligned}
&\left|\frac{\widehat{f}_1(y \mid x)\widehat{f}_1(x)}{\widehat{f}_0(x)} - \frac{f_1(y \mid x)f_1(x)}{f_0(x)}\right| \\
&\le \widehat{f}_1(y \mid x)\left|\frac{\widehat{f}_1(x)}{\widehat{f}_0(x)} - \frac{f_1(x)}{f_0(x)}\right| + \frac{f_1(x)}{f_0(x)}|\widehat{f}_1(y \mid x) - f_1(y \mid x)| \\
&\le \widehat{f}_1(y \mid x)\frac{1}{\widehat{f}_0(x)f_0(x)}(\widehat{f}_1(x)|\widehat{f}_0(x) - f_0(x)| + f_0(x)|\widehat{f}_1(x) - f_1(x)|) \\
&+ \frac{f_1(x)}{f_0(x)}|\widehat{f}_1(y \mid x) - f_1(y \mid x)|.
\end{aligned}$$

By Condition 1 and the convergence rate of e_{jn} for $j = 1, 2, 3, 4$, we have

$$\sup_{x,y}\left|\frac{\widehat{f}_1(y \mid x)\widehat{f}_1(x)}{\widehat{f}_0(x)} - \frac{f_1(y \mid x)f_1(x)}{f_0(x)}\right| = O_P(r_n).$$

By Condition 2 (i) and (ii),

$$\frac{1}{M_1}(c - c') \le t_c(x) - t_{c'}(x) \le \frac{1}{M_2}(c - c') \tag{1}$$

for any $c, c' \in (c_0 - M_2\delta, c_0 + M_2\delta)$ such that $c > c'$. Then for any $c \in (c_0 - M_2\delta, c_0 + M_2\delta)$

$$\begin{aligned}
\frac{\widehat{f}_1(t_c(x) + a_n r_n \mid x)\widehat{f}_1(x)}{\widehat{f}_0(x)} \le& \frac{f_1(t_c(x) + a_n r_n \mid x)}{f_0(x)} + \sup_{y,x}\left|\frac{\widehat{f}_1(y \mid x)\widehat{f}_1(x)}{\widehat{f}_0(x)} - \frac{f_1(y \mid x)f_1(x)}{f_0(x)}\right| \\
\le& c - M_2 a_n r_n + O_P(r_n)
\end{aligned}$$

for sufficiently large n, where $\{a_n\}_{n=1}^{\infty}$ is a sequence of positive numbers such that $a_n r_n \to 0$ and the O_P is uniform in c. Thus

$$\frac{\widehat{f}_1(t_c(x) + a_n r_n \mid x)\widehat{f}_1(x)}{\widehat{f}_0(x)} < c \tag{2}$$

with probability approaching 1 for any $\{a_n\}_{n=1}^{\infty}$ such that $a_n \to \infty$, $a_n r_n \to 0$. Similarly,

$$\frac{\widehat{f}_1(t_c(x) - a_n r_n \mid x)\widehat{f}_1(x)}{\widehat{f}_0(x)} \ge c + c_2 a_n r_n + O_P(r_n).$$

for the same $\{a_n\}_{n=1}^{\infty}$ and

$$\frac{\widehat{f}_1(t_c(x) - a_n r_n \mid x)\widehat{f}_1(x)}{\widehat{f}_0(x)} > c$$

with probability approaching 1. Hence

$$\sup_{c\in(c_0-M_2\delta,c_0+M_2\delta)}\left|\widehat{t}_c(x)-t_c(x)\right|\le a_n r_n$$

with probability approaching 1 for any a_n converging to infinity slowly. Thus

$$\sup_{c\in(c_0-M_2\delta,c_0+M_2\delta)}\left|\widehat{t}_c(x)-t_c(x)\right|=O_P(r_n).$$

Note that according to Condition 1

$$\begin{aligned}&\sup_{c\in(c_0-M_2\delta,c_0+M_2\delta)}\left|\frac{1}{n_1}\sum_{I_i=1}F_1(\widehat{t}_c(X_i)\mid X_i)-\frac{1}{n_1}\sum_{I_i=1}F_1(t_c(X_i)\mid X_i)\right|\\&\le L\sup_{c\in(c_0-M_2\delta,c_0+M_2\delta)}\left|\widehat{t}_c(x)-t_c(x)\right|=O_P(r_n),\end{aligned}\tag{3}$$

where $L=\sup_{x,y}f_1(y\mid x)<\infty$. According to Example 19.11 in van der Vaart (1998), the function class $\{F_1(t_c(\cdot)\mid\cdot\,):c\in(c_0-M_2\delta,c_0+M_2\delta)\}$ is a Donsker class. Thus

$$\sup_{c\in(c_0-M_2\delta,c_0+M_2\delta)}\left|\frac{1}{n_1}\sum_{I_i=1}F_1(t_c(X_i)\mid X_i)-\mathbb{E}_1F_1(t_c(X)\mid X)\right|=O_P\left(\frac{1}{\sqrt{n}}\right).\tag{4}$$

Note that

$$\sup_{c\in(c_0-M_2\delta,c_0+M_2\delta}\left|\frac{1}{n_1}\sum_{I_i=1}\widehat{F}_1(\widehat{t}_c(X_i)\mid X_i)-\frac{1}{n_1}\sum_{I_i=1}F_1(\widehat{t}_c(X_i)\mid X_i)\right|\le e_{4n}=O_P(r_n).$$

Then combining this with (3) and (4), we have

$$\sup_{c\in(c_0-M_2\delta,c_0+M_2\delta)}\left|\frac{1}{n_1}\sum_{I_i=1}\widehat{F}_1(\widehat{t}_c(X_i)\mid X_i)-\mathbb{E}_1F_1(t_c(X)\mid X)\right|=O_P(r_n).$$

Because $1-\mathbb{E}_1F_1(t_{c_0}(X)\mid X)=\epsilon$, according to Conditions 2 (iii) and (1), we have

$$|1-\mathbb{E}_1F_1(t_c(X)\mid X)-\epsilon|=|\mathbb{E}_1F_1(t_c(X)\mid X)-\mathbb{E}_1F_1(t_{c_0}(X)\mid X)|\ge\frac{M_3}{M_1}|c-c_0|.$$

Then with the same arguments we used to show (2), we get

$$1-\frac{1}{n_1}\sum_{I_i=1}\widehat{F}_1(\widehat{t}_{c_0+a_nr_n}(X_i)\mid X_i)<\epsilon$$

and

$$1-\frac{1}{n_1}\sum_{I_i=1}\widehat{F}_1(\widehat{t}_{c_0-a_nr_n}(X_i)\mid X_i)>\epsilon$$

with probability approaching 1. By monotonicity of $1-\frac{1}{n_1}\sum_{I_i=1}\widehat{F}_1(\widehat{t}_c(X_i)\mid X_i)$ with respect to c, we have $|\widehat{c}_0-c_0|<a_nr_n$ with probability approaching 1. Hence $|\widehat{c}_0-c_0|=O_P(r_n)$. Again by (1),

$$\sup_x|\widehat{t}_{\text{opt}}(x)-t_{c_0}(x)|=O_P(r_n).$$

3. Simulation results with $\epsilon=0.01$

In this section we provide some simulation results with the choice $\epsilon=0.01$. Here we consider the same data generation processes as in Section 3 in the main text. Quarantine durations for people with different feature values obtained by the proposed method and the two quantile methods under the four scenarios are plotted in Fig. 1. All the results are averaged over 200 simulation datasets.

[Figure 1 about here.]

The average quarantine duration (AQD) of uninfected people and the escape probability (EP) are summarized in the following table. Because non-integer quarantine duration is not practical, the quarantine duration is rounded to the nearest integer in calculation. All the results are averaged over 200 simulation datasets.

[Table 1 about here.]

Table 1 shows that the proposed optimal quarantine rule still performs well with the choice $\epsilon=0.01$.

4. Evaluation of the Weibull model

In this section, we assess how well the Weibull conditional density model that assumed in Section 4.1 fits the data. A parametric model is useful as long as it can approximate the true data generation process well, even though it may not be correct. Hence, instead of performing a goodness of fit test, we estimate the following distance between the true distribution and our assumed model with least false parameters

$$D = \iint (F_1(y \mid x) - F_1(y \mid x, \alpha^*, \gamma^*))^2 dF_1(x) dG_1(y)$$

where α^*, γ^* are least false parameters that our estimators converge to, $F_1(y \mid x, \alpha, \gamma) = 1 - \exp(-(y/\gamma^{\mathrm{T}} v(x))^{\alpha})$, and $F_1(x)$, $G_1(y)$ are the marginal distribution functions of X and Y conditional on $I = 1$, respectively. Remind that $F_1(y \mid x)$ is the true distribution function of Y conditional on $X = x$ and $I = 1$. Thus D is a metric ranging from 0 to 1 that can describe how well our model can approximate the true distribution. We estimate $F_1(y \mid x)$ by kernel method with a Gaussian associate kernel, estimate $F_1(y \mid x, \alpha^*, \gamma^*)$ by $F_1(y \mid x, \widehat{\alpha}, \widehat{\gamma})$ and estimate $F_1(x)$ and $G_1(y)$ by their empirical version, respectively. Then by plugging in these estimations we get an estimate of D. The estimate is 0.0006, which is extremely small. Thus the assumed Weibull conditional density model can approximate the true data generation process well, and we can expect it to work well in practice.

5. Countries at different risk levels

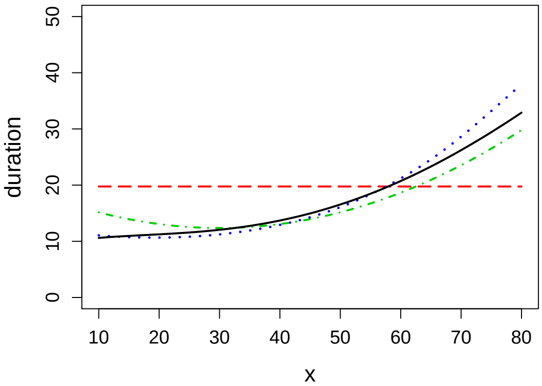
[Table 2 about here.]

References

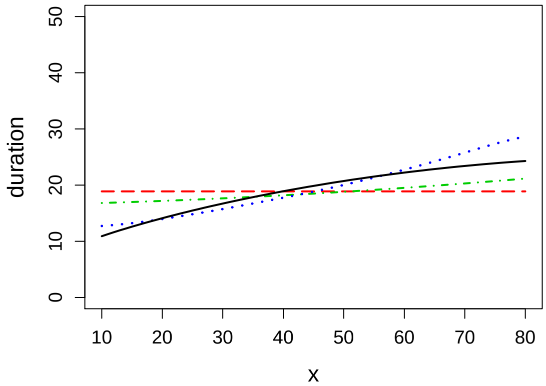
Gelfand, I. M. and Fomin, S. (1963). *Calculus of variations.* New Jersey, NJ: Prentice Hall, INC.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* **24,** 726–748.

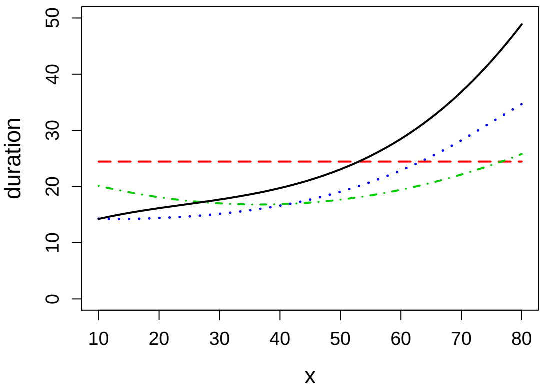
van der Vaart, A. W. (1998). *Asymptotic Statistics.* New York, NY: Cambridge University Press.



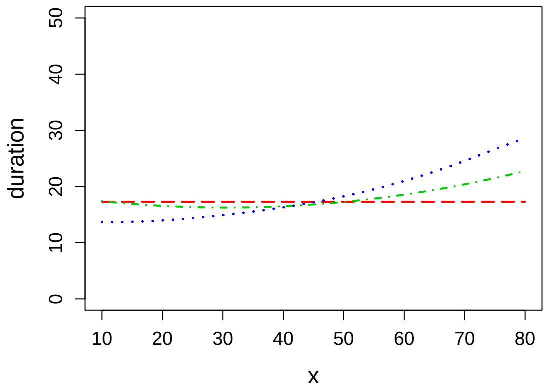
(a) Scenario 1.



(b) Scenario 2.



(c) Scenario 3.



(d) Scenario 4.

Figure 1: Quarantine duration for people with different feature values: 0.99 quantile, red dashed line; 0.99 conditional quantile, green dashes dotted line; optimal duration, blue dotted line; theoretical optimal duration, black solid line.

Table 1: Average quarantine duration of uninfected people and escape probability associated with the three quarantine rules under different scenarios with $\epsilon = 0.01$

Scenario	Method	AQD	EP
1	0.95 quantile	19.76	1.2%
	0.95 conditional quantile	13.90	1.1%
	optimal quarantine rule	12.97	1.1%
2	0.95 quantile	18.91	1.2%
	0.95 conditional quantile	17.88	1.1%
	optimal quarantine rule	16.54	1.1%
3	0.95 quantile	24.47	1.2%
	0.95 conditional quantile	17.88	2.1%
	optimal quarantine rule	16.38	1.8%
4	0.95 quantile	17.28	1.1%
	0.95 conditional quantile	16.83	0.6%
	optimal quarantine rule	15.80	0.5%

Table 2: Countries at different risk levels

risk group	countries
high risk	Amenria, Belgium, Brazil, Cabo Verde, Canada, Chile, Gabon, Kuwait, Panama, Peru, Qatar, Singapore, Spain, United Arab Emirates
medium risk	Afghanistan, Algeria, Argentina, Azerbaijan, Bulgaria, Colombia, Equatorial Guinea, Eswatini, Finland, France, Germany, Guinea, Mexico, Netherlands, North Macedonia, Oakistan, Paraguay, Portugal, Romania, Senegal, Serbia, South Africa, Switzerland, United States
low risk	Angola, Austrilia, Bahamas, Benin, Burkina Faso, Cameroon, Central African Republic, China, Cuba, Estonia, Ethiopia, Gambia, Greece, Guatemala, India, Japan, Lebanon, Liberia, Lithuania, Madagascar, Mali, Mauritania, Mozambique, Nepal, Phillippines, Rwanda, Slovakia, Sri Lanka, Sudan, Thailand, Togo, Tunisia, Uganda