ORIGINAL PAPER

Parameter estimation of the incubation period of COVID-19 based on the doubly interval-censored data model

Ming-Ze Yin · Qing-Wen Zhu · Xing Lü

Abstract With the spread of the novel coronavirus disease 2019 (COVID-19) around the world, the estimation of the incubation period of COVID-19 has become a hot issue. Based on the doubly interval-censored data model, we assume that the incubation period follows lognormal and Gamma distribution, and estimate the parameters of the incubation period of COVID-19 by adopting the maximum likelihood estimation, expectation maximization algorithm and a newly proposed algorithm (expectation mostly conditional maximization algorithm, referred as ECIMM). The main innovation of this paper lies in two aspects: Firstly, we regard the sample data of the incubation period as the doubly interval-censored data without unnecessary data simplification to improve the accuracy and credibility of the results; secondly, our new ECIMM algorithm enjoys better convergence and universality compared with others. With the framework of this paper, we conclude that 14-day quarantine period can largely interrupt the transmission of COVID-19, however, people who need specially monitoring should be isolated for about 20 days for the sake of safety. The results provide some suggestions for the prevention and control of COVID-19. The newly proposed ECIMM algorithm can also be used to deal with the doubly interval-censored data model appearing in various fields.

M.-Z. Yin · Q.-W. Zhu · X. Lü (✉)
Department of Mathematics, Beijing Jiaotong University,
Beijing 100044, China
e-mail: xinglv655@163.com; xlv@bjtu.edu.cn

1 Introduction

In the late 2019 and early 2020, a number of patients infected with the novel coronavirus disease 2019 (COVID-19) have been successively found in Wuhan, Hubei province, China [1,2]. This newly discovered virus causes severe acute respiratory disease. On March 11, 2020, the World Health Organization (WHO) announced the novel COVID-19 a pandemic [3]. As time goes by, COVID-19 epidemic has spread very rapidly all over the world. The distribution of COVID-19 cases by country in the world is shown in Fig. 1. The epidemic prevention work is gradually taken seriously by more and more countries [2]. In many countries, the drastic restrictive measures have not prevented the outbreak of new pandemic's waves [3,4].

Despite the worldwide manages to control the growth of COVID-19, the number of COVID-19 incidences is still rising at a reproduction rate of 3.77 [6,7]. The outbreak evolution for the current most affected countries is shown in Fig. 2. The mortality rate in cases infected with COVID-19 is 5.25% worldwide. This mortality rate is 7.60% in the European region, 2.24% in the Eastern Mediterranean region, 2.22% in the African region, 2.95% in the South-East Asia region, 5.07% in the region of Americas, and 3.55% in the region of Western Pacific [6,8]. The novel COVID-

Fig. 1 Global cases of
COVID-19 until May 12,
2021 [5]
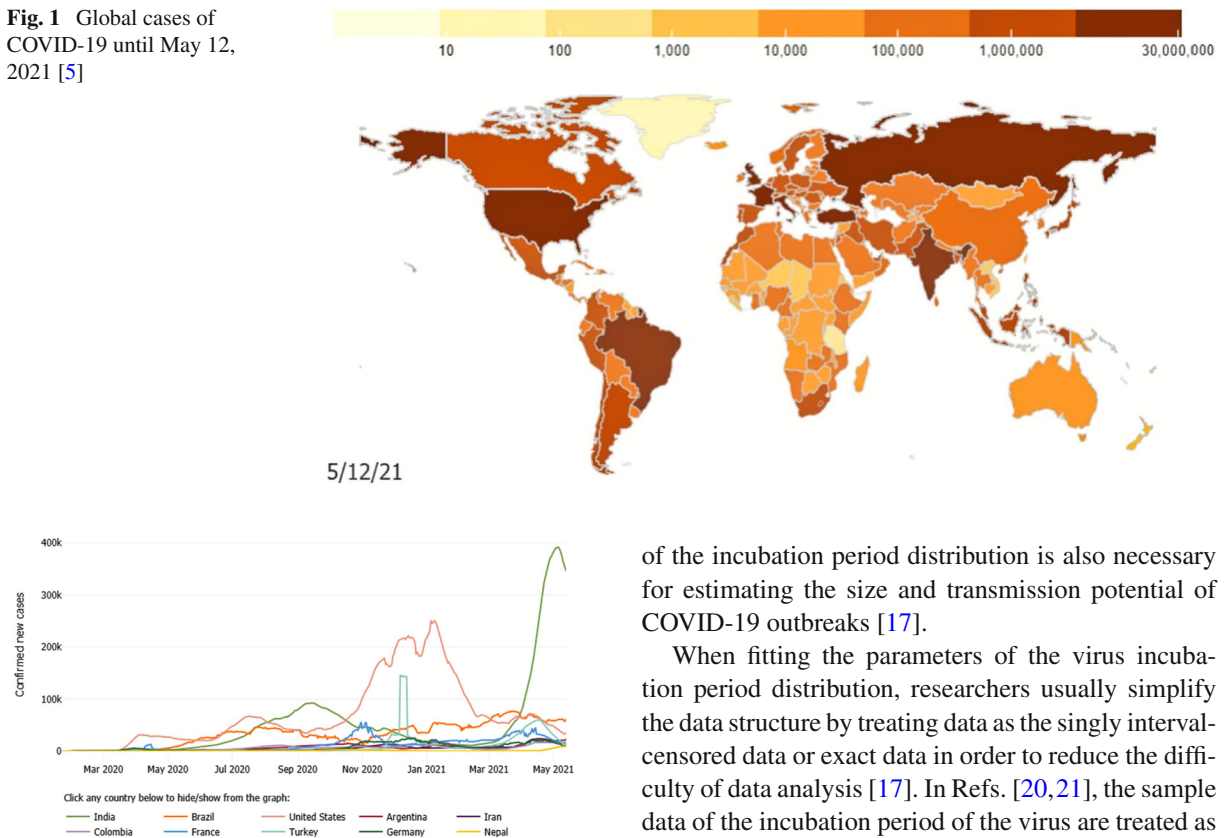
Fig. 2 Daily confirmed new cases of the current most affected
countries [13]

19 has become a worldwide pandemic affecting 219
countries with an estimate of more than 159 mil-
lion infected cases and over 3.3 million deaths (WHO
Coronavirus Disease [COVID-19] Dashboard, May 12,
2021) [11,12].

To protect against COVID-19 epidemic, we must
have a deep understanding of the basic characteristics
of COVID-19, among which one of the most important
features is the incubation period of the virus [2,14].
The incubation period of COVID-19 is the period from
infection to the earliest appearance of clinical symp-
toms of COVID-19 patients [15,16]. Estimation of
virus incubation period is of great significance for the
epidemiological investigation and the development of
epidemic prevention and controlling measures [17].
The incubation period is an aid for defining the time
period for which contact tracing is to be done [18].
It helps in active monitoring of people having higher
exposure and also in determining the length of active
monitoring so as to save resources [19]. Knowledge

of the incubation period distribution is also necessary
for estimating the size and transmission potential of
COVID-19 outbreaks [17].

When fitting the parameters of the virus incuba-
tion period distribution, researchers usually simplify
the data structure by treating data as the singly interval-
censored data or exact data in order to reduce the diffi-
culty of data analysis [17]. In Refs. [20,21], the sample
data of the incubation period of the virus are treated as
the doubly interval-censored data without unnecessary
data structure simplification to improve the accuracy of
the researches. In Refs. [17,22], maximum likelihood
estimation (MLE) and Bayesian estimation are widely
applied in the field of the doubly interval-censored data
model. It has been shown that the doubly interval-
censored data model makes the research results more
reliable. Fast and effective algorithm makes great con-
tribution to the data analysis and processing [9,10].
Expectation maximization (EM) algorithm optimizes
the process of maximizing the likelihood function to
get the parameter estimates through the iterative pro-
cedure [23,24]. Some extensions on the EM algorithm
have been proposed and widely applied as supplement
to the EM algorithm theory [25]. Expectation condi-
tional maximization (ECM) algorithm solves the prob-
lem of multi-parameter estimation by approaching the
optimal estimate values step by step [26], and expec-
tation mostly maximization (EMM) algorithm acceler-
ates the convergence speed of the iterative algorithm
by improving the expectation function in the doubly
interval-censored data model. In order to make the
EM algorithm more suitable for processing the dou-

bly interval-censored data, we propose a new algorithm named as expectation mostly conditional maximization (ECIMM) algorithm to estimate the parameters of the COVID-19 incubation period.

In this paper, the sample data of the incubation period of COVID-19 will be regarded as the doubly interval-censored data [20,21]. The incubation period of COVID-19 will be fitted by lognormal distribution and Gamma distribution based on the open data of COVID-19 incubation period collected so far [27]. In the field of statistical research, maximum likelihood estimation and the EM algorithm are mature parameter estimation methods [25,26,28]. The newly proposed ECIMM algorithm enjoys better universality and convergence compared with the related basic algorithm. We will use the maximum likelihood estimation, the EM algorithm and the ECIMM algorithm to estimate the parameters of the incubation period of COVID-19 [17,26], and propose some suggestions for the prevention and control of COVID-19 epidemic.

The rest of this paper is organized as follows. We will introduce the data background in Sect. 2, and the maximum likelihood estimation and the EM algorithm in Sect. 3. The ECIMM algorithm will be proposed in Sect. 4. In Sect. 5, we will use three methods to estimate the parameters and propose the suggestions for epidemic prevention. In Sect. 6, we will discuss our parameter estimation results with others. Finally, the conclusion and future work will be emphasized in Sect. 7.

2 Data background

Current common estimations of the incubation period of COVID-19 are mostly based on studies of accurate case data or simplified case data [17]. However, in the actual data acquisition, accurate data can not be obtained easily, and only the approximate intervals of the infection and onset time of patients can be investigated [17]. In order to obtain accurate estimation of COVID-19 incubation period, we will estimate parameters on the basis of the doubly interval-censored data model [17,20,21].

According to the doubly interval-censored data model [17,20,21], E and S represent the time when patients infected with COVID-19 are exposed to the novel coronavirus and the time when symptoms occur, respectively. The time of the incuba-

tion period is $T = S - E$. A typical observed value consists of four time points, namely $X = (E_L, E_R, S_L, S_R)$, where the subscripts L and R correspondingly represent the left and right endpoints of the interval of E and S, as shown in Fig. 3. When E and S are both intervals, the observed data is called the doubly interval-censored data. Accordingly, when one of E and S is an exact value and the other is an interval, the observed data is called the singly interval-censored data. When E and S are both exact values, the observed data is the exact data.

There are several distributions which are suitable for simulating the incubation period, e.g., the lognormal distribution and Gamma distribution [29–31]. In this paper, we will correspondingly assume that the incubation period follows the lognormal and Gamma distribution, and estimate the main parameters. As mean and quantiles play an important role in the research of the incubation period of COVID-19, we will pay attention to mean and quantiles of the novel coronavirus incubation period [32].

The data of this paper are from the online repository [27]. The repository consists of the information of 3397 patients infected with COVID-19, such as the location, the country, the gender and the age of each patient. The repository also includes the time when patients are exposed to the novel coronavirus and the time when symptoms occur, which are used to estimate the incubation period of the novel coronavirus in this paper.

3 Algorithm theory

3.1 Maximum likelihood estimation based on the doubly interval-censored data model

The density of T and E are recorded as $f_\theta(t)$ and $h_\lambda(v)$. In general, we can suppose that E and T are mutually independent and E follows uniform distribution. The
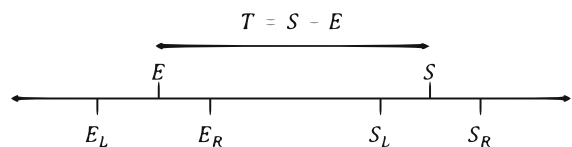

Fig. 3 The doubly interval-censored data

likelihood functions of the doubly interval-censored data, the singly interval-censored data and the exact data are as follows [33]:

$$L(\theta, \lambda; X) = \int_{E_L}^{E_R} \int_{S_L}^{S_R} h_\lambda(v) f_\theta(s - v) dv ds, \quad (1)$$

$$L(\theta; T_L, T_R) = \int_{T_L}^{T_R} f_\theta(t) dt, \quad (2)$$

$$L(\theta; T) = f_\theta(T). \quad (3)$$

An observed data of the sample may be the doubly interval-censored data, and can also be the singly interval-censored data or the exact data [17]. We introduce two indicative variables, called σ_i and ω_i. When $\sigma_i=1$, the observed data is the doubly interval-censored data. When $\omega_i=1$, the observed data is the singly interval-censored data. And $\sigma_i=\omega_i=0$ indicates that the observed data is the exact data. The likelihood function of this observed data is as follows [33]:

$$L(\theta, \lambda; X_i) = \left[\int_{E_{L_i}}^{E_{R_i}} \int_{S_{L_i}}^{S_{R_i}} h_\lambda(v) f_\theta(s - v) dv ds \right]^{\sigma_i}$$
$$\times \left[\int_{T_{L_i}}^{T_{R_i}} f_\theta(t) dt \right]^{\omega_i} \times [f_\theta(T_i)]^{1-\sigma_i-\omega_i}. \quad (4)$$

The likelihood function of the whole sample can be easily obtained by multiplying likelihood functions of all observed data:

$$L(\theta, \lambda; X) = \left[\int_{E_{L_i}}^{E_{R_i}} \int_{S_{L_i}}^{S_{R_i}} h_\lambda(v) f_\theta(s - v) dv ds \right]^{\sum_{i=1}^{n} \sigma_i}$$
$$\times \left[\int_{T_{L_i}}^{T_{R_i}} f_\theta(t) dt \right]^{\sum_{i=1}^{n} \omega_i} \times [f_\theta(T_i)]^{\sum_{i=1}^{n} 1-\sigma_i-\omega_i}. \quad (5)$$

According to the idea of maximum likelihood, the best estimates of the parameters are the values which maximize the likelihood function:

$$(\theta, \lambda) = \underset{(\theta, \lambda)}{\arg\max} \log L(\theta, \lambda). \quad (6)$$

3.2 EM algorithm based on the doubly interval-censored data model

The likelihood function of the whole sample is shown as Eq. 5.

In the rest of this paper, we transform $L(\theta, \lambda)$ to $p(x|\theta, \lambda)$ in order to express the algorithm more clearly and concisely.

EM algorithm is an iteration algorithm [23,24]. According to the theory of survival analysis, we introduce a latent variable z to advance the parameter estimation process. $\log p(x, z|\theta, \lambda)$ is the log-likelihood function of complete data. We calculate the mathematical expectation of the log-likelihood function, named as ELBO function. According to the idea of maximum likelihood, we get the best estimates of the parameters by finding the values which maximize the ELBO function:

E step:

$$\text{ELBO} = \int_z p(z|x, \theta^{(i)}, \lambda^{(i)}) \log p(x, z|\theta, \lambda) dz, \quad (7)$$

M step:

$$\underset{(\theta, \lambda)}{\arg\max} \int_z p(z|x, \theta^{(i)}, \lambda^{(i)}) \log p(x, z|\theta, \lambda) dz. \quad (8)$$

When

$$||\theta^{(i+1)} - \theta^{(i)}|| + ||\lambda^{(i+1)} - \lambda^{(i)}|| \quad (9)$$

meets the allowable error range for the specific problems, the iterative process should be stopped.

4 ECIMM algorithm

4.1 ECIMM algorithm based on the doubly interval-censored data model

The likelihood function of the whole sample is shown as Eq. 5.

According to the theory of survival analysis, we introduce a latent variable z with the density function $q(z)$ to transform the log-likelihood function:

$$\log p(x|\theta, \lambda) = \int_z q(z) \log \frac{p(x, z|\theta, \lambda)}{q(z)} dz$$
$$- \int_z q(z) \log \frac{p(z|x, \theta, \lambda)}{q(z)} dz. \quad (10)$$

We take the density function of the latent variable z as the undetermined function $q(z)$ while the EM algorithm treats the density function of the latent variable z as the posterior density function $p(z|x, \theta^{(i)}, \lambda^{(i)})$. We get the ELBO function of the ECIMM algorithm by

calculating the mathematical expectation of the log-likelihood function:

$$\text{ELBO} = \int_z p(z|x, \theta^{(i)}, \lambda^{(i)}) \log p(x, z|\theta, \lambda) dz$$
$$- \int_z \log q(z) dz. \tag{11}$$

The following algorithm steps are based on the optimized ELBO function. In step MM_1, we fix parameter values to obtain the optimal density estimate $\hat{q}(z)$ by maximizing ELBO function. In step MM_2, We fix the density function of z as $\hat{q}(z)$ to estimate the parameters by maximizing ELBO function.

MM_1 $step$: Fix θ, λ, and solve

$$\hat{q} = \arg\max_q \text{ELBO}, \tag{12}$$

MM_2 $step$: Fix $q = \hat{q}$, and accomplish the parameter estimation by the following steps:

$$(\theta^{(i)}, \lambda^{(i)}) \to (\theta^{(i+1)}, \lambda^{(i+1)}). \tag{13}$$

Theoretically, we can get the parameter estimates by computing the partial derivatives of the ELBO function and equating them to be zero. However, it is difficult to obtain those estimates when the ELBO function is multivariate in the doubly interval-censored data model. The ECIMM algorithm obtain parameter estimates by approaching the optimal values step by step.

CM_1 $step$: Fix $\theta = \theta^{(i)}$, and

$$\lambda^{(i+1)} = \arg\max_\lambda \text{ELBO}, \tag{14}$$

CM_2 $step$: Fix $\lambda = \lambda^{(i+1)}$, and

$$\theta^{(i+1)} = \arg\max_\theta \text{ELBO}. \tag{15}$$

When

$$||\theta^{(i+1)} - \theta^{(i)}|| + ||\lambda^{(i+1)} - \lambda^{(i)}|| \tag{16}$$

meets the allowable error range for the specific problems, the iterative process should be stopped.

Compared with the EM algorithm, which directly treats the density function of the latent variable z as the posterior conditional density, the ECIMM algorithm treats it as the undetermined variable $q(z)$ for research and analysis.

Since the density function $q(z) \leq 1$ and the integral value $\int \log q(z) dz \leq 0$, the ELBO function of the ECIMM algorithm is more suitable for parameter estimation based on the doubly interval-censored data model:

$$\int_z q(z) \log \frac{p(x, z|\theta, \lambda)}{q(z)} dz$$
$$\geq \int_z p(z|x, \theta, \lambda) \log p(x, z|\theta, \lambda) dz. \tag{17}$$

Because the ECIMM algorithm optimizes the ELBO function, the iterative value can be closer to the true value of the parameters. According to maximum likelihood theory, the ECIMM algorithm reduces the number of iteration steps and accelerates the convergence speed of the algorithm:

$$\left\| \theta^{(i+1)} - \theta* \right\| \leq \left\| \theta^{(i)} - \theta* \right\|, \tag{18}$$

$$\left\| \theta_{ECIMM}^{(i+1)} - \theta* \right\| \leq \left\| \theta_{EM}^{(i+1)} - \theta* \right\|. \tag{19}$$

At the same time, in the steps of maximizing the ELBO function to get the parameter estimates, the basic idea is to set the partial derivatives to be zero and solve the equations

$$\begin{cases} \frac{\partial \int_z p(z|x, \theta^{(i)}, \lambda^{(i)}) \log p(x, z|\theta, \lambda) dz}{\partial \theta} = 0, \\ \frac{\partial \int_z p(z|x, \theta^{(i)}, \lambda^{(i)}) \log p(x, z|\theta, \lambda) dz}{\partial \lambda} = 0. \end{cases} \tag{20}$$

However, since the ELBO function in the doubly interval-censored data model is always a multivariate function, it is hard to realize the algorithm due to the complexity of the calculation. ECIMM algorithm optimizes the algorithm by approaching the optimal parameter estimates step by step. It makes the algorithm much easier to implement so that it can be widely used to solve various problems:

$$\arg\max_\lambda \left[\int_z p(z|x, \theta^{(i)}, \lambda^{(i)}) \log p(x, z|\theta^{(i)}, \lambda) dz - \int_z \log \hat{q}(z) dz \right], \tag{21}$$

$$\arg\max_\theta \left[\int_z p(z|x, \theta^{(i)}, \lambda^{(i)}) \log p(x, z|\theta, \lambda^{(i+1)}) dz - \int_z \log \hat{q}(z) dz \right], \tag{22}$$

$$(\theta^{(i)}, \lambda^{(i)}) \to (\theta^{(i)}, \lambda^{(i+1)}) \to (\theta^{(i+1)}, \lambda^{(i+1)}). \tag{23}$$

Overall by improving the ELBO function in the doubly interval-censored data model and approaching the optimal estimate value step by step, we accelerate the convergence speed of the algorithm and improve the universality of the algorithm.

4.2 ECIMM algorithm based on the assumption of lognormal distribution

In the algorithm description process of fitting specific distribution, we record the ELBO function as Q function in order to intuitively reflect the values of parameters in the algorithm iteration. Assuming that E follows uniform distribution on (a, b) and T follows lognormal distribution, we apply the ECIMM algorithm to estimate the parameters of the novel coronavirus incubation period:

$$
\begin{aligned}
&Q[q(z), (\mu, \sigma; b-a)|(\mu^{(i)}, \sigma^{(i)}; (b-a)^{(i)})] \\
&= E\{\ln L(\mu, \sigma; b-a)|w_c, \mu^{(i)}, \sigma^{(i)}; (b-a)^{(i)}\} \\
&\quad - \int_z \log q(z) dz.
\end{aligned}
\tag{24}
$$

$MM_1\ step$: Fix $(\mu, \sigma; b-a)$, and solve

$$
\hat{q} = \arg\max_q Q[q(z), (\mu, \sigma; b-a)|(\mu^{(i)}, \sigma^{(i)}; (b-a)^{(i)})],
\tag{25}
$$

$MM_2\ step$: Fix $q = \hat{q}$, and accomplish the parameter estimation by the following steps:

$$
(\mu^{(i)}, \sigma^{(i)}, (b-a)^{(i)}) \rightarrow (\mu^{(i+1)}, \sigma^{(i+1)}, (b-a)^{(i+1)}).
\tag{26}
$$

$CM_1\ step$: Fix $(\mu, \sigma) = (\mu^{(i)}, \sigma^{(i)})$, and

$$
\widehat{b-a} = (b-a)^{(i+1)} = \arg\max_{b-a} Q[\hat{q}(z), (\mu^{(i)}, \sigma^{(i)}, (b-a))|(\mu^{(i)}, \sigma^{(i)}; (b-a)^{(i)})],
\tag{27}
$$

$CM_2\ step$: Fix $b-a = (b-a)^{(i+1)}$, $\mu = \mu^{(i)}$, and

$$
\hat{\sigma} = \sigma^{(i+1)} = \arg\max_{\sigma} Q[\hat{q}(z), (\mu^{(i)}, \sigma; (b-a)^{(i)+1})|(\mu^{(i)}, \sigma^{(i)}; (b-a)^{(i)})],
\tag{28}
$$

$CM_3\ step$: Fix $b-a = (b-a)^{(i+1)}$, $\sigma = \sigma^{(i+1)}$, and

$$
\hat{\mu} = \mu^{(i+1)} = \arg\max_{\mu} Q[(\hat{q}(z), \mu, \sigma^{(i+1)}; (b-a)^{(i+1)})|(\mu^{(i)}, \sigma^{(i)}; (b-a)^{(i)})].
\tag{29}
$$

It completes the process of

$$
(\mu^{(i)}, \sigma^{(i)}, (b-a)^{(i)}) \rightarrow (\mu^{(i+1)}, \sigma^{(i+1)}, (b-a)^{(i+1)}).
\tag{30}
$$

When

$$
||\mu^{(i+1)} - \mu^{(i)}|| + ||\sigma^{(i+1)} - \sigma^{(i)}|| + ||(b-a)^{(i+1)} - (b-a)^{(i)}||
\tag{31}
$$

meets the allowable error range for the specific problems, the iterative process should be stopped.

4.3 ECIMM algorithm based on the assumption of Gamma distribution

Assuming that E follows uniform distribution on (a, b) and T follows Gamma distribution, we apply the ECIMM algorithm to estimate the parameters of the novel coronavirus incubation period:

$$
\begin{aligned}
&Q[q(z), (k, \theta; b-a)|(k^{(i)}, \theta^{(i)}; (b-a)^{(i)})] \\
&= E\{\ln L(k, \theta; b-a)|w_c, k^{(i)}, \theta^{(i)}; (b-a)^{(i)}\} \\
&\quad - \int_z \log q(z) dz.
\end{aligned}
\tag{32}
$$

$MM_1\ step$: Fix $(k, \theta; b-a)$, and solve

$$
\hat{q} = \arg\max_q Q[q(z), (k, \theta; b-a)|(k^{(i)}, \theta^{(i)}; (b-a)^{(i)})],
\tag{33}
$$

$MM_2\ step$: Fix $q = \hat{q}$, and accomplish the parameter estimation by the following steps:

$$
(k^{(i)}, \theta^{(i)}, (b-a)^{(i)}) \rightarrow (k^{(i+1)}, \theta^{(i+1)}, (b-a)^{(i+1)}).
\tag{34}
$$

$CM_1\ step$: Fix $(k, \theta) = (k^{(i)}, \theta^{(i)})$, and

$$\widehat{b-a} = (b-a)^{(i+1)} = \arg\max_{b-a} Q[\hat{q}(z), (k^{(i)}, \theta^{(i)}, (b-a))|(k^{(i)}, \theta^{(i)}; (b-a)^{(i)})], \tag{35}$$

$CM_2\ step$: Fix $b-a = (b-a)^{(i+1)}$, $k = k^{(i)}$, and

$$\hat{\theta} = \theta^{(i+1)} = \arg\max_{\theta} Q[\hat{q}(z), (k^{(i)}, \theta; (b-a)^{(i+1)})|(k^{(i)}, \theta^{(i)}; (b-a)^{(i)})], \tag{36}$$

$CM_3\ step$: Fix $b-a = (b-a)^{(i+1)}$, $\theta = \theta^{(i+1)}$, and

$$\hat{k} = k^{(i+1)}$$
$$= \arg\max_{k} Q[(\hat{q}(z), k, \theta^{(i+1)}; (b-a)^{(i+1)})|(k^{(i)}, \theta^{(i)}; (b-a)^{(i)})]. \tag{37}$$

It completes the process of

$$(k^{(i)}, \theta^{(i)}, (b-a)^{(i)}) \rightarrow (k^{(i+1)}, \theta^{(i+1)}, (b-a)^{(i+1)}). \tag{38}$$

When

$$||k^{(i+1)} - k^{(i)}|| + ||\theta^{(i+1)} - \theta^{(i)}|| + ||(b-a)^{(i+1)} - (b-a)^{(i)}|| \tag{39}$$

meets the allowable error range for the specific problems, the iterative process should be stopped.

5 The parameter estimation of COVID-19 incubation period

We estimate the incubation period of COVID-19 using the doubly interval-censored data model with the sample size of 50, 200 and 500 [17,20,21]. We carry out the simulation through the maximum likelihood estimation method, the EM algorithm and the ECIMM algorithm based on the lognormal and Gamma distribution hypothesis, and the results are shown in Figs. 4, 5, 6, 7, 8, and 9. The parameter estimation results of the incubation period are obtained by using the above three parameter estimation methods based on the lognormal and Gamma distribution hypothesis, which are listed in Tables 1, 2, 3, 4, 5 and 6.

Figures 4, 5 and 6 show the simulation results of three estimation methods based on the assumption of lognormal distribution, with the sample size of 50,200 and 500, respectively. While Figs. 7, 8 and 9 show the simulation results of three estimation methods based on the assumption of Gamma distribution, with the sample size of 50,200 and 500, respectively. In each figure, the three different lines correspond to the three different estimation methods.
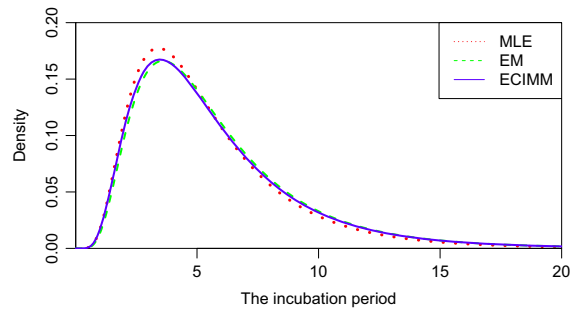
We can discover that in each figure, the simulation results obtained by three methods are close to each
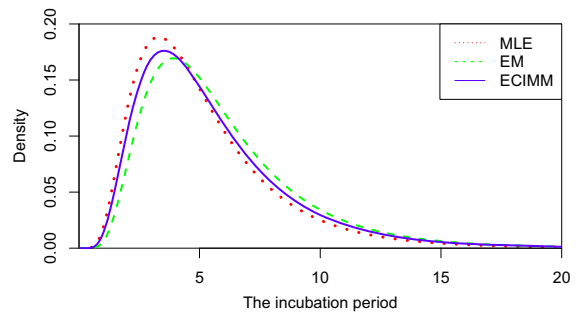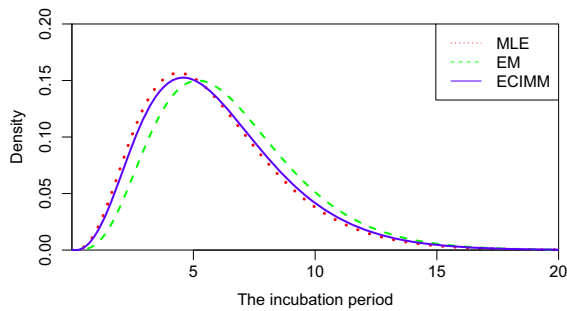

Fig. 4 The simulation results of three estimation methods based on the assumption of lognormal distribution ($n = 50$)
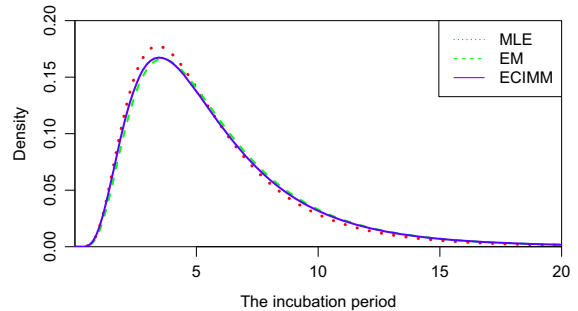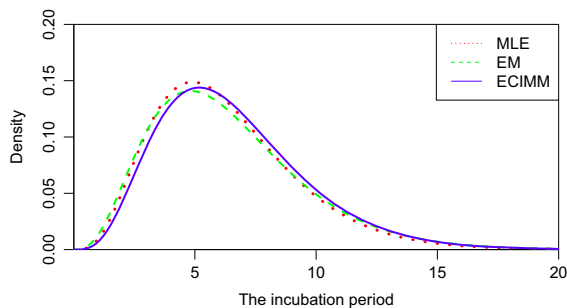

Fig. 5 The simulation results of three estimation methods based on the assumption of lognormal distribution ($n = 200$)
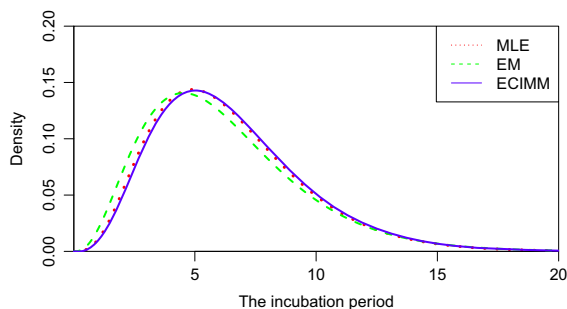

Fig. 6 The simulation results of three estimation methods based on the assumption of lognormal distribution ($n = 500$)

Fig. 7 The simulation results of three estimation methods based on the assumption of Gamma distribution ($n = 50$)

Fig. 8 The simulation results of three estimation methods based on the assumption of Gamma distribution ($n = 200$)

Fig. 9 The simulation results of three estimation methods based on the assumption of Gamma distribution ($n = 500$)

other, indicating that our new method is reasonable. As the sample size increases, the simulation results are more reliable. And we can find that the simulation results on the lognormal distribution assumption are different from the simulation results on the Gamma distribution assumption due to the different characteristics of two distributions.

As mean and quantiles play an important role in the research of the incubation period of COVID-19, we estimate mean and quantiles of the novel coronavirus incubation period. Tables 1, 2 and 3 list the estimation results of three estimation methods based on the assumption of lognormal distribution, with the sample size of 50,200 and 500, respectively. While Tables 4, 5 and 6 list the estimation results of three estimation methods based on the assumption of Gamma distribution, with the sample size of 50,200 and 500, respectively. We can find that the quantile values on the lognormal distribution assumption are larger than the quantile values on the Gamma distribution assumption, which is due to the different characteristics of two distributions.

According to Table 3, based on the lognormal distribution hypothesis, the average incubation period is about 6.8 days, and the probability of the incubation period not exceeding 15.31 days is 0.975. According to Table 6, based on the hypothesis of Gamma distribution, the average incubation period is about 6.5 days, and the probability of the incubation period not exceeding 13.84 days is 0.975. The results of the research show that the 14-day quarantine period can largely interrupt the transmission of COVID-19, which fits within the range for the incubation period of 0 to 14 days assumed by the WHO, and is consistent with current medical control measures [1].

With the improvement of epidemic prevention and control, the situation has stabilized in some areas [1]. People from high-risk areas need to be specially monitored. According to Table 3, the probability of the incubation period not exceeding 21.56 days is 0.995, based on the lognormal distribution hypothesis. According to Table 6, the probability of the incubation period not exceeding 17.19 days is 0.995, based on the assumption of Gamma distribution. The results suggest that for the sake of safety, people who need specially monitoring should be isolated for about 20 days. The probability that a patient infected with COVID-19 does not show disease symptoms during 20-day quarantine period is nearly less than 0.5%. Therefore the new quarantine period can effectively block the spread of COVID-19.

Remarks:

1. We estimate the parameters of COVID-19 incubation period based on the doubly interval-censored data model, which makes the research results more reliable.
2. Our new ECIMM algorithm enjoys good convergence and universality.

Table 1 The parameter estimation of lognormal distribution ($n = 50$)

The method of parameter estimation	MLE	EM Algorithm	ECIMM Algorithm
Mean	5.9	6.7	6.3
97.5% Quantile	13.17	14.40	14.09
99.5% Quantile	18.48	19.89	19.78

Table 2 The parameter estimation of lognormal distribution ($n = 200$)

The method of parameter estimation	MLE	EM Algorithm	ECIMM Algorithm
Mean	6.2	6.6	6.5
97.5% Quantile	14.11	15.13	15.15
99.5% Quantile	19.95	21.47	21.65

Table 3 The parameter estimation of lognormal distribution ($n = 500$)

The method of parameter estimation	MLE	EM Algorithm	ECIMM Algorithm
Mean	6.3	6.5	6.8
97.5% Quantile	14.89	15.24	15.31
99.5% Quantile	21.40	21.83	21.56

Table 4 The parameter estimation of Gamma distribution ($n = 50$)

The method of parameter estimation	MLE	EM Algorithm	ECIMM Algorithm
Mean	5.8	6.5	6.0
97.5% Quantile	12.57	13.34	12.92
99.5% Quantile	15.69	16.41	16.10

Table 5 The parameter estimation of Gamma distribution ($n = 200$)

The method of parameter estimation	MLE	EM Algorithm	ECIMM Algorithm
Mean	6.3	6.4	6.6
97.5% Quantile	13.31	13.96	13.81
99.5% Quantile	16.51	17.45	17.08

3. Our results can be regarded as a valuable supplement of COVID-19 prevention.

6 Discussion

We compare our results with others, as listed in Table 7. Backer uses Bayesian estimation method to estimate the main parameters with the sample size of 88 by assuming the incubation period follows lognormal, Gamma and Weibull distribution [19]. Qiu uses maximum likelihood estimation and Bayesian estimation method to estimate the main parameters with the sample size of 543 by assuming the incubation period follows lognormal, Gamma and Weibull distribution [17]. They estimate the mean and 97.5% quantile, while we estimate the mean, 97.5% quantile and 99.5% quantile. They estimate the parameters by using maximum likelihood estimation and Bayesian estimation method, while we use maximum likelihood estimation method, the EM algorithm and the ECIMM algorithm to estimate the main parameters.

Comparing the results, we find that the mean values are all between 6 and 7, in other words, they are consistent with each other. Based on the lognormal distribution assumption, the 97.5% quantile values obtained

Table 6 The parameter estimation of Gamma distribution ($n = 500$)

The method of parameter estimation	MLE	EM Algorithm	ECIMM Algorithm
Mean	6.4	6.2	6.5
97.5% Quantile	13.73	13.92	13.84
99.5% Quantile	17.09	17.54	17.19

Table 7 The comparison of results(NA indicates the result is not available)

Distribution	Method	Mean	97.5% Quantile	99.5% Quantile
Lognormal	MLE	6.3	14.9	21.4
Lognormal	EM	6.5	15.2	21.8
Lognormal	ECIMM	6.8	15.3	21.6
Gamma	MLE	6.4	13.7	17.1
Gamma	EM	6.2	13.9	17.5
Gamma	ECIMM	6.5	13.8	17.2
Lognormal [19]	Bayes	6.8	15.5	NA
Gamma [19]	Bayes	6.5	12.5	NA
Weibull [19]	Bayes	6.4	11.1	NA
Lognormal [17]	MLE	6.3	15.2	NA
Gamma [17]	MLE	6.3	13.8	NA
Weibull [17]	MLE	6.4	13.3	NA
Lognormal [17]	Bayes	6.4	15.4	NA
Gamma [17]	Bayes	6.3	13.8	NA

by three methods of our research are between 14.5 and 15.5, while the 97.5% quantile value obtained by Bayesian estimation method in Ref. [19] is 15.5, and the 97.5% quantile values obtained by maximum likelihood estimation and Bayesian estimation method in Ref. [17] are 15.2 and 15.4. Based on the Gamma distribution assumption, the 97.5% quantile values obtained by three methods of our research are between 13 and 14, while the 97.5% quantile value obtained by Bayesian estimation method in Ref. [19] is 12.5, and the 97.5% quantile values obtained by maximum likelihood estimation and Bayesian estimation method in Ref. [17] are both 13.8.

Comparing the results based on the Gamma distribution and lognormal distribution assumption, we find that the quantile values based on the lognormal distribution assumption are significantly greater than that based on the Gamma distribution assumption. It is due to the different characteristics of two distributions. Lognormal distribution has greater degree of dispersion than Gamma distribution. The quantile results based on the lognormal distribution assumption are more conservative.

We compare the results obtained by three methods in our research, and find that the results are similar, which can indicate that our new method is reasonable. However, the ECIMM algorithm shows fast convergence speed in dealing with the doubly interval-censored data of COVID-19 incubation period, and it can be widely used to deal with the doubly interval-censored data in various fields.

We hope that the research work of this paper can be regarded as the useful supplement to the related studies on the incubation period of COVID-19, and be helpful for the prevention and control of COVID-19.

7 Concluding remarks

In this paper, we have estimated the parameters of COVID-19 incubation period based on the doubly interval-censored data model. Statistical inference analysis has been conducted on lognormal distribution and Gamma distribution. The maximum likelihood estimation method, the EM algorithm and the ECIMM algorithm have been used for parameter estimation.

Each parameter estimation method of each distribution has been theoretically derived, which can be realized by mainstream computer programming software. We have obtained the estimates of mean, 97.5% quantile and 99.5% quantile, and suggested that 14-day quarantine period can largely interrupt the transmission of COVID-19, however, people who need specially monitoring should be isolated for about 20 days for the sake of safety. The research results can be regarded as a supplement of COVID-19 prevention.

Instead of simplifying the data structure, we regard the sample data of the incubation period as the doubly interval-censored data, which makes the results more accurate and reliable. Furthermore, we propose a new algorithm called ECIMM algorithm which has good convergence and universality. The ECIMM algorithm shows fast convergence speed when dealing with the doubly interval-censored data of COVID-19 incubation period, and it can be widely used to deal with the doubly interval-censored data in various fields.

In future studies, we will further extend the research results by estimating the parameters of the incubation period of COVID-19 based on other distribution assumptions. There are few researchers who have applied the ECIMM algorithm in current studies. We encourage further studies on its accuracy and convergence rate as the potential work. And we hope our work in this paper contributes to the prevention and control of COVID-19 and other epidemics.

Data availability statement Our manuscript has no associated data.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Alsiri, N.F., Alhadhoud, M.A., Palmer, S.: The impact of the coronavirus disease of 2019 on research. J. Clin. Epidemiol. **129**, 124–125 (2021)
2. Liu, C., Wu, X., Niu, R., Wu, X., Fan, R.: A new SAIR model on complex networks for analysing the 2019 novel coronavirus (COVID-19). Nonlinear Dyn. **101**(3), 1777–1787 (2020)
3. Soldi, G., Forti, N., Gaglione, D., Braca, P., Millefiori, L.M., Marano, S., Willett, P., Pattipati, K.: Quickest Detection and Forecast of Pandemic Outbreaks: Analysis of COVID-19 Waves. *IEEE Communications Magazine*, (2021). arXiv:2101.04620v2
4. Markovic, A., Muhlematter, C., Beaugrand, M., Camos, V., Kurth, S.: Severe effects of the COVID-19 confinement on young children's sleep: A longitudinal study identifying risk and protective factors. J. Sleep Res. https://doi.org/10.1111/jsr.13314
5. https://coronavirus.jhu.edu/data/animated-world-map
6. Mittal, H., Pandey, A.C., Pal, R., Tripathi, A.: A new clustering method for the diagnosis of CoVID19 using medical images. Appl. Intell. **51**, 2988–3011 (2021)
7. Rothan, H.A., Byrareddy, S.N.: The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. J. Autoimmun. **109**, 102433 (2020)
8. Altan, A., Karasu, S.: Recognition of COVID-19 disease from X-ray images by hybrid model consisting of 2D curvelet transform, chaotic salp swarm algorithm and deep learning technique. Chaos Solitons Fractals **140**, 110071 (2020)
9. Altan, A.: Performance of metaheuristic optimization algorithms based on swarm intelligence in attitude and altitude control of unmanned aerial vehicle for path following. *4th International Symposium on Multidisciplinary Studies and Innovative Technologies*, Turkey, (2020). https://doi.org/10.1109/ISMSIT50672.2020.9255181
10. Altan, A., Parlak, A., Adaptive control of a 3D printer using whale optimization algorithm for bio-printing of artificial tissues and organs. 2020 Innovations in Intelligent Systems and Applications Conference (ASYU). Turkey **2020**, (2020). https://doi.org/10.1109/ASYU50717.2020.9259820
11. Kalra, R.S., Kumar, V., Dhanjal, J.K., Garg, S., Li, X., Kaul, S.C., Sundar, D., Wadhwa, R.: COVID19-inhibitory activity of withanolides involves targeting of the host cell surface receptor ACE2: insights from computational and biochemical assays. J. Biomol. Struct. Dyn. (2021). https://doi.org/10.1080/07391102.2021.1902858
12. https://covid19.who.int/
13. https://coronavirus.jhu.edu/data/new-cases
14. Liu, X.H., He, Y., Ma, X.S., Luo, L.Q.: Statistical Data Analysis on the Incubation and Suspected Period of COVID-19 Based on 2172 Confirmed Cases Outside Hubei Province. Acta Mathematicae Applicatae Sinica **43**(02), 278–294 (2020)
15. Goel, K., Kumar, A., Nilam: Nonlinear dynamics of a time-delayed epidemic model with two explicit aware classes, saturated incidences, and treatment. Nonlinear Dyn. **101**(3), 1693–1715 (2020)
16. Wang, Y., Wei, Z., Cao, J.: Epidemic dynamics of influenza-like diseases spreading in complex networks. Nonlinear Dyn. **101**(3), 1801–1820 (2020)
17. Qiu, M.Y., Hu, T., Cui, H.J.: Parametric Estimation for the Incubation Period Distribution of COVID-19 under Doubly Interval Censoring. Acta Mathematicae Applicatae Sinica **43**(02), 200–210 (2020)
18. Rai, B., Shukla, A., Dwivedi, L.K.: Incubation period for COVID-19: a systematic review and meta-analysis.

J. Public Health (Berl.) (2021). https://doi.org/10.1007/s10389-021-01478-1

19. Backer, J.A., Klinkenberg, D., Wallinga, J.: Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20-28 January 2020. *Eurosurveillance*, 2020;25(5):pii=2000062. https://doi.org/10.2807/1560-7917.ES.2020.25.5.2000062

20. Gruttola, V.D., Lagakos, S.W.: Analysis of doubly-censored survival data, with application to AIDS. Biometrics **45**(1), 1–11 (1989)

21. Gomez, G.: Estimation of induction distributions with doubly censored data and application to AIDS. Theory Probabil. Appl. **37**(1), 32–39 (1993)

22. Komarek, A., Lesaffre, E., Harkanen, T., Declerck, D., Virtanen, J.I.: A Bayesian analysis of multivariate doubly-interval-censored dental data. Biostatistics **6**(1), 145–155 (2005)

23. Mclachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions, 2nd edn. Wiley InterScience (2007)

24. Hou, L.B.: Parameter Estimation of Exponential Distribution Under Fixed-time Interval Censoring. Stat. Decision **36**(05), 20–24 (2020) (in Chinese)

25. Li, L.X., Tang, M., Zeng, Y.Y., Shen, L.Y., Li, Z.H., Chen, H.L., Guo, S.N., Chen, J.B., Hou, Y.W., Chen, Z.: Statistical analysis methods and applications of the interval-censored survival data. Chin. J. Health Stat. **33**(03), 530–533 (2016) (in Chinese)

26. Wu, D.Y.: Statistical Inference of Several Distributions under Interval Censored Data and Empirical Analysis of Financial Data. Master's thesis, Beijing Technology and Business University (2016) (in Chinese)

27. Boston: Laboratory for the modeling of biological and socio-technical systems (MOBS). [Accessed 29 Jan 2020]. https://docs.google.com/spreadsheets/d/1jS24DjSPVWa4iuxuD4OAXrE3QeI8c9BC1hSlqr-NMiU/edit#gid=1449891965

28. Dejardin, D., Lesaffre, E.: Stochastic EM algorithm for doubly interval-censored data. Biostatistics **14**(4), 766–778 (2013)

29. Tojinbara, K., Sugiura, K., Yamada, A., Kakitani, I., Kwan, N.C.L.: Estimating the probability distribution of the incubation period for rabies using data from the 1948–1954 rabies epidemic in Tokyo. Prev. Vet. Med. **123**, 102–105 (2016)

30. Nishiura, H.: Determination of the appropriate quarantine period following smallpox exposure: An objective approach using the incubation period distribution. Int. J. Hyg. Environ. Health **212**(1), 97–104 (2009)

31. Linton, N.M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A., Jung, S., Yuan, B., Kinoshita, R., Nishiura, H.: Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data. J. Clin. Med. **9**(2), 538–546 (2020)

32. Lauer, S.A., Grantz, K.H., Bi, Q.: The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. Ann. Intern. Med. **172**(9), 577–582 (2020)

33. Gentleman, R., Geyer, C.J.: Maximum likelihood for interval censored data: consistency and computation. Biometrika **81**(3), 618–623 (1994)