Statistica Sinica **34** (2024), 2219-2236
doi:https://doi.org/10.5705/ss.202020.0318

HYPOTHESIS TEST ON A MIXTURE FORWARD-INCUBATION-TIME EPIDEMIC MODEL WITH APPLICATION TO COVID-19 OUTBREAK

Chunlin Wang, Pengfei Li, Yukun Liu*, Xiao-Hua Zhou and Jing Qin

Xiamen University, University of Waterloo, East China Normal University, Peking University and National Institutes of Health

Abstract: The distribution of the incubation period of the novel coronavirus disease that emerged in 2019 (COVID-19) has crucial clinical implications for understanding this disease and devising effective disease-control measures. Our study is based on a cross-sectional and forward follow-up study that collected the duration times between a specific observation time and the onset of COVID-19 symptoms for individuals. The original study further proposed a mixture forward-incubation-time epidemic model, which is a mixture of an incubation-period distribution and a forward time distribution, to model the collected duration times and to estimate the incubation-period distribution of COVID-19. In this study, we provide sufficient conditions for the identifiability of the unknown parameters in the aforementioned epidemic model when the incubation period follows a two-parameter distribution. Under the same setup, we propose a likelihood ratio test (LRT) for testing the null hypothesis that the mixture forward-incubation-time epidemic model is a homogeneous exponential distribution. The testing problem is nonregular because a nuisance parameter is present only under the alternative. We establish the limiting distribution of the LRT and identify an explicit representation for it, and obtain the limiting distribution of the LRT under a sequence of local alternatives. Our simulation results indicate that the LRT exhibits desirable type-I errors and power. Lastly, we analyze a COVID-19 outbreak data set from China to illustrate the usefulness of the LRT.

Key words and phrases: Identifiability, likelihood ratio test, non-regularity.

1. Introduction

As the novel coronavirus disease that emerged in 2019 (COVID-19) spread rapidly worldwide, the World Health Organization (WHO) declared the COVID-19 outbreak a global pandemic on March 10, 2020. Currently, COVID-19 is still spreading around the world, posing a significant threat to global public health and affecting global economics and social development. As of January 7, 2022, the WHO had identified over 300 million confirmed cases of COVID-19, and observed more than 5 million deaths. Countries are fighting this pandemic by imposing measures such as isolation policies, travel restrictions, lockdowns,

*Corresponding author.

and social distancing. Of these measures (Cohen and Kupferschmidt (2020)), quarantining people who may have been exposed to COVID-19 seems to be the most effective way of preventing further disease transmission.

The incubation period of an infectious disease is the time between exposure to it and the first appearance of symptoms. Thus, an accurate estimation of the incubation-period distribution, or incubation distribution, is crucial (especially in regions where the epidemic is severe) for determining the length of appropriate quarantine periods for individuals who suspect they may have been exposed to the virus. In the literature, estimating incubation distributions has attracted much attention (Sartwell (1950); Kalbfleisch and Lawless (1989); Struthers and Farewell (1989); Kalbfleisch and Lawless (1991); Farewell et al. (2005); Wilkening (2008)), and studies on COVID-19 are still ongoing; see Backer, Klinkenberg and Wallinga (2020), Guan et al. (2020), Lauer et al. (2020), Li et al. (2020), Linton et al. (2020), Liu et al. (2021), Qin et al. (2020), Rahman et al. (2020), Wang et al. (2020b), and Liu, Ma and Jiang (2022), among others. The current results are based mostly on clinical experience or empirical statistical analysis of contact-tracing data. However, such data may be inaccurate because of the patient's recall bias or the interviewer's personal judgment on the possible date of exposure, rather than the actual date. For additional discussions, see Qin et al. (2020).

The lockdown of Wuhan, the capital city of Hubei province in China, provided an opportunity to estimate accurately the incubation distribution of COVID-19. Qin et al. (2020) designed a new cross-sectional and forward follow-up study, in which they collected the duration times between departing Wuhan and the onset of symptoms for 1,211 confirmed cases in people who left Wuhan before the lockdown with no COVID-19 symptoms, and then developed symptoms outside Wuhan; further details on the study and data collection can be found in Section 5. Using the theory of renewal processes, they proposed a *mixture forward-incubation-time epidemic model* to model the 1,211 observed duration times and to estimate the incubation distribution. This mixture model overcomes the issues of biased sampling, and considers the possibility that some patients may have been exposed to COVID-19 on their way out of Wuhan.

Herein, we follow the approach and model setup of Qin et al. (2020). Let Y be the *incubation period* with probability density function (pdf) $f(t)$. Consider a specific observation time that is either (i) the time of exposure to the disease, or (ii) some time thereafter, but before the onset of symptoms. However, whether the situation pertains to (i) or (ii) is unknown. For example, Qin et al. (2020) chose the observation time of an individual to be their departure time from Wuhan. Furthermore, let A be the time between the exposure time and the departure time, and V be the *forward time* calculated from the departure time to the symptom-onset time, given that the departure time is after the exposure time, but before the symptom-onset time, that is, $Y > A$. Treating this as a

renewal process that reaches equilibrium, it can be shown that the conditional pdf of V given $Y > A$ is given approximately by

$$g(t) = \frac{\int_t^\infty f(y)dy}{\int_0^\infty yf(y)dy} \quad \text{for } t > 0$$

(Linton et al. (2020); Qin (2017, Chap. 2)). See Section S1 of the Supplementary Material for a derivation of the form of $g(t)$. As Qin et al. (2020) point out, the study cohort may contain heterogeneous subpopulations: individuals who left Wuhan by train, bus, or plane were more likely to have come into contact with COVID-19, because they were in a crowded environment with possible human-to-human transmission of the virus. A similar argument pertains to the COVID-19 outbreak that occurred from late January to early February in 2020 onboard the Diamond Princess cruise ship (Verity et al. (2020)).

In the following, we use the duration-time data from Wuhan in Qin et al. (2020) to introduce the mixture forward-incubation-time epidemic model, in which the observation time of an individual is their departure time from Wuhan. Let T be the *duration time* between departure from Wuhan and the onset of symptoms. We consider two cases, namely, $A = 0$ and $A > 0$; the variable T satisfies $T = Y$ if $A = 0$, and $T = V$ if $A > 0$. Assuming that A and Y are independent, the conditional pdf of T given $A = 0$ is the pdf $f(x)$ of Y. Note that $A > 0$ is equivalent to $T = V = Y - A > 0$, and thus the conditional pdf of T given $A > 0$ is the conditional pdf of V given $Y > A$ or $g(t)$. Furthermore, denote p as the proportion of individuals who contracted COVID-19 as they left Wuhan, that is, $p = P(A = 0)$. Because we have no idea who contracted the disease before departure ($A > 0$) and who did so while departing ($A = 0$), T follows the *mixture forward-incubation-time epidemic model* (Qin et al. (2020))

$$h(t) = pf(t) + (1 - p)g(t), \quad t > 0. \tag{1.1}$$

Note that we can observe only T, and not Y or V. Let $t_1, \ldots, t_n$ be n observed duration times that are independent and identically distributed (i.i.d.) copies of T.

Note that there may exist a third portion of individuals who were infected outside Wuhan after departure. In this study, we assume that this portion of individuals does not exist, for two reasons. First, it is theoretically challenging to derive the pdf of the duration time for this portion of individuals. Additional work is required, and the results developed under model (1.1) can serve as a starting point for further research. Second, the goodness-of-fit test in Section S2 of the Supplementary Material seems to suggest that model (1.1) provides an adequate fit to the duration-time data from Wuhan.

Throughout this paper, we focus on model (1.1) with $f(t) = f(t; \lambda, \alpha)$, the pdf of a general two-parameter distribution. Then, the pdf of T becomes

$$h(t;\lambda,\alpha,p) = pf(t;\lambda,\alpha) + (1-p)g(t;\lambda,\alpha), \quad t>0, \tag{1.2}$$

and $t_1,\ldots,t_n$ are n i.i.d. observations from $h(t;\lambda,\alpha,p)$. Under the mixture model (1.2), Deng et al. (2021) discuss the asymptotic properties of the maximum likelihood estimators (MLEs) and the likelihood ratio statistic of the unknown parameters (λ,α,p), under the assumption that (λ,α,p) are identifiable. However, this assumption does not always hold. A counter example is the Weibull pdf $f(t;\lambda,\alpha) = \lambda\alpha(t\lambda)^{\alpha-1}\exp\{-(\lambda t)^{\alpha}\}I(t>0)$. It can be verified that $f(t;\lambda,\alpha) = g(t;\lambda,\alpha)$ when $\alpha = 1$. This implies that p is not identifiable in (1.2) when $f(t;\lambda,\alpha)$ is a Weibull pdf with $\alpha = 1$, and so we cannot use the asymptotic results in Deng et al. (2021) in such a situation. A similar conclusion holds when $f(t;\lambda,\alpha) = \{\Gamma(\alpha)\}^{-1}\lambda^{\alpha}t^{\alpha-1}\exp(-\lambda t)I(t>0)$, a gamma pdf.

In this study, we complement the work of Deng et al. (2021) in two ways. First, we provide sufficient conditions for the identifiability of (λ,α,p). Our results indicate the following: (i) (λ,α,p) is identifiable when $f(t;\lambda,\alpha)$ is a lognormal, Weibull or gamma pdf, but not when it is an exponential pdf; (ii) (λ,α) is identifiable, but p is not, when $f(t;\lambda,\alpha)$ is an exponential pdf. Second, we propose a likelihood ratio test (LRT) to test the null hypothesis that $f(t;\lambda,\alpha)$ is an exponential pdf. Under this null hypothesis, $h(t;\lambda,\alpha,p)$ also becomes an exponential pdf, so the proposed LRT also tests the homogeneity in model (1.2). Note that the nuisance parameter p disappears under the null model, and is only identified under the alternative hypothesis.

The problem of a nuisance parameter unidentified under the null hypothesis has long been recognized in the literature as a nonregular problem (Davies (1977, 1987)). Because of the partial identifiability of the nuisance parameter, classical inference methods such as the LRT may lose their usual statistical properties. The limiting distribution of the LRT often involves complex stochastic processes (Liu et al. (2020a)). The homogeneity testing problem under a two-component mixture model has been studied extensively in the literature; for example, see Liu and Shao (2003), Chen and Li (2009), and Chen, Li and Liu (2020), and the references therein. To the best of our knowledge, these papers assume that the two components come from the same distribution family, and do not share any underlying parameters. However, under model (1.2), the two components are not from the same distribution family and share the common parameters (λ,α). Thus, the existing results cannot be applied to the testing problem under model (1.2).

Despite the aforementioned challenges, we obtain the limiting distribution of the LRT for the nonregular testing problem, that is, testing the null hypothesis that $h(t;\lambda,\alpha,p)$ is the pdf of a homogeneous exponential distribution. We show that the asymptotic null distribution of the LRT is the supremum of a chi-squared process, and identify an explicit representation of the limiting distribution that can be used for rapid numerical calculation of the asymptotic critical values or

p-values of the proposed LRT. The results of finite-sample simulations show that the proposed LRT has tight control of type-I error rates and appreciable power in general. The proposed LRT is then used to analyze COVID-19 data from China. Following Qin et al. (2020), we choose $f(t;\lambda,\alpha)$ to be a Weibull pdf. Our analysis results indicate that the mixture forward-incubation-time model produces a better fit than that with a homogeneous exponential distribution.

Note that all of our results are based on the parametric model (1.2), and violating this model assumption may lead to invalid analysis results. This raises the goodness-of-fit test problem of model (1.2) in applications. We suggest using the goodness-of-fit test in Deng et al. (2021) to check the validity of model (1.2) based on $t_1,\ldots,t_n$; this test is reviewed briefly in the Supplementary Material.

The rest of this paper is organized as follows. In Section 2, we discuss sufficient conditions for the identifiability of (λ,α,p) in model (1.2), and apply the results to the case where $f(t;\lambda,\alpha)$ is a Weibull, gamma, or lognormal pdf. In Section 3, we establish the nonregular asymptotic distribution of the LRT for testing the null hypothesis that $h(t;\lambda,\alpha,p)$ is a homogeneous exponential distribution, and provide an explicit representation of this asymptotic distribution. Here we also derive the asymptotic distribution of the proposed LRT under a sequence of local alternatives. We report our simulation results in Section 4, and in Section 5, and we analyze real COVID-19 outbreak data from China. Finally, we conclude the paper with a discussion in Section 6. For convenience of presentation, all proofs are given in the Supplementary Material.

2. Identifiability of $(\boldsymbol{\lambda},\boldsymbol{\alpha},\boldsymbol{p})$

Identifiability is important in the application of the mixture forward-incubation-time epidemic model in (1.2). If some model parameters are not identifiable, then their point estimators cannot be consistent, and standard inferences for other parameters that are identifiable may be questionable. In this section, we establish the identifiability of (λ,α,p) in model (1.2) under the following conditions on $f(t;\lambda,\alpha)$. Let $F(t;\lambda,\alpha)$ be the cumulative distribution function corresponding to $f(t;\lambda,\alpha)$.

A1. Given (λ,α), $\lim_{t\to\infty} f(t;\lambda,\alpha)/\{1-F(t;\lambda,\alpha)\}$ exists and is either finite or ∞.

A2. When $(\lambda_1,\alpha_1)\neq(\lambda_2,\alpha_2)$, $\lim_{t\to\infty} f(t;\lambda_1,\alpha_1)/\{f(t;\lambda_2,\alpha_2)\}$ exists and is either zero or ∞.

A3. When $(\lambda_1,\alpha_1)\neq(\lambda_2,\alpha_2)$, both $\lim_{t\to\infty} f(t;\lambda_1,\alpha_1)/\{1-F(t;\lambda_2,\alpha_2)\}$ and $\lim_{t\to\infty} f(t;\lambda_2,\alpha_2)/\{1-F(t;\lambda_1,\alpha_1)\}$ exist and are either zero or ∞.

Theorem 1. *Assume model* (1.2) *and conditions* A1-A3. *Let*

$$A(\lambda, \alpha) = \lim_{t \to \infty} \frac{f(t; \lambda, \alpha)}{1 - F(t; \lambda, \alpha)}.$$

Suppose $h(t; \lambda_1, \alpha_1, p_1) = h(t; \lambda_2, \alpha_2, p_2)$ *for all* $t > 0$.

(a) If $A(\lambda_1, \alpha_1) = 0$ *or* ∞, *then* $(\lambda_1, \alpha_1, p_1) = (\lambda_2, \alpha_2, p_2)$.

(b) If $0 < A(\lambda_1, \alpha_1) < \infty$, *then* $(\lambda_1, \alpha_1) = (\lambda_2, \alpha_2)$. *Furthermore, if* $f(t; \lambda_1, \alpha_1) / \{1 - F(t; \lambda_1, \alpha_1)\}$ *is not a constant function of* t, *then* $p_1 = p_2$; *otherwise,* p_1 *and* p_2 *are not necessarily the same.*

After some calculus, it can be verified that conditions A1-A3 are satisfied by a Weibull, gamma, or lognormal distribution. We can further verify that $A(\lambda, \alpha) = 0$ for a lognormal distribution, $A(\lambda, \alpha) = \lambda$ for a gamma distribution, and $A(\lambda, \alpha) = 0$ or ∞ if $\alpha \neq 1$ and $A(\lambda, \alpha) = \lambda$ if $\alpha = 1$ for a Weibull distribution. Applying the results in Theorem 1 to Weibull, gamma, and lognormal distributions, we have the following identifiability results.

Corollary 1. *Under model* (1.2),

(a) (p, λ, α) *are identifiable when* $f(t; \lambda, \alpha)$ *is the pdf of a lognormal distribution;*

(b) (p, λ, α) *are identifiable when* $f(t; \lambda, \alpha)$ *is the pdf of a Weibull or gamma distribution, but not when it is the pdf of an exponential distribution;*

(c) (λ, α) *are identifiable, but* p *is not, when* $f(t; \lambda, \alpha)$ *is the pdf of an exponential distribution.*

Deng et al. (2021) mention the identifiability property of (λ, α, p), but do not give a formal proof. The results in Theorem 1 and Corollary 1 provide formal justifications, and further indicate when the results of Deng et al. (2021) are applicable and when they are not.

3. Testing Whether Incubation Distribution is Exponential

3.1. The LRT

Corollary 1 indicates that the parameter p is not identifiable when $f(t; \lambda, \alpha)$ is the pdf of an exponential distribution under model (1.2). Because of this, the asymptotic results in Deng et al. (2021) are not applicable in such a situation. In this section, we propose an LRT to check whether $f(t; \lambda, \alpha)$ is the pdf of an exponential distribution or, equivalently, whether $h(t; \lambda, \alpha, p)$ is the pdf of a homogeneous exponential distribution, based on n i.i.d. observations $t_1, \ldots, t_n$ from model (1.2).

Throughout this section, we assume that the following condition is satisfied.

C0. There exists a unique α_0 such that $f(t;\lambda,\alpha_0) = g(t;\lambda,\alpha_0)$, for all $t > 0$.

Condition C0 is satisfied by a Weibull or gamma distribution with $\alpha_0 = 1$, and condition C0 is satisfied if and only if $f(t;\lambda,\alpha_0)$ is the pdf of an exponential distribution. Under condition C0, testing the null hypothesis that $f(t;\lambda,\alpha)$ is the pdf of an exponential distribution is equivalent to testing

$$H_0 : \alpha = \alpha_0 \quad \text{versus} \quad H_1 : \alpha \neq \alpha_0. \tag{3.1}$$

Note that under model (1.2), the case of $\alpha = \alpha_0$ indicates that individuals in the cross-sectional and forward follow-up study are homogeneous, and that the duration time T defined in Section 1 follows an exponential distribution. When $\alpha \neq \alpha_0$, there are heterogeneous subgroups of individuals in the cross-sectional and forward follow-up study. In this case, we favor using the mixture model (1.2) to model the distribution of T. Theoretically, detecting the existence of such heterogeneous subpopulations is an important initial step before applying the mixture model (1.2). If we were to apply model (1.2) to homogenous duration times, then the MLE of (λ,α,p) would no longer have asymptotic normality. Consequently, the Wald-type confidence intervals for the quantiles of the incubation period may not have the nominal asymptotic coverage probabilities.

A natural solution to the testing problem (3.1) is one based on likelihood. Given the n observations $t_1,\ldots,t_n$ from model (1.2), the log-likelihood of (λ,α,p) is

$$\ell_n(\lambda,\alpha,p) = \sum_{i=1}^{n} \log\left\{pf(t_i;\lambda,\alpha) + (1-p)g(t_i;\lambda,\alpha)\right\}.$$

Let $(\hat\lambda,\hat\alpha,\hat p)$ be the MLE of (λ,α,p) under the full model, and let $\hat\lambda_0$ be the MLE of λ under the null model, that is,

$$(\hat\lambda,\hat\alpha,\hat p) = \operatorname*{argmax}_{\lambda,\alpha,p} \ell_n(\lambda,\alpha,p), \quad \hat\lambda_0 = \operatorname*{argmax}_{\lambda} \ell_n(\lambda,\alpha_0,1).$$

Note that under the null model, p does not appear, and λ is the only parameter to be estimated. We set $p = 1$ under the null model for convenience of presentation.

The LRT statistic for (3.1) is defined as

$$R_n = 2\left\{\sup_{\lambda,\alpha,p} \ell_n(\lambda,\alpha,p) - \sup_{\lambda} \ell_n(\lambda,\alpha_0,1)\right\} = 2\left\{\ell_n(\hat\lambda,\hat\alpha,\hat p) - \ell_n(\hat\lambda_0,\alpha_0,1)\right\}.$$

We reject the null hypothesis H_0 in (3.1) if the observed value of R_n exceeds some critical value determined by its limiting distribution, presented in Section 3.2.

3.2. Asymptotic null distribution of the LRT

We require some notation before presenting the asymptotic results of the LRT statistic R_n. Let (λ_0, α_0) be the true values of (λ, α) under the null model, and define

$$X_i = \frac{\partial f(t_i;\lambda_0,\alpha_0)/\partial\lambda}{f(t_i;\lambda_0,\alpha_0)}, \quad Y_{i1} = \frac{\partial f(t_i;\lambda_0,\alpha_0)/\partial\alpha}{f(t_i;\lambda_0,\alpha_0)}, \quad Y_{i2} = \frac{\partial g(t_i;\lambda_0,\alpha_0)/\partial\alpha}{g(t_i;\lambda_0,\alpha_0)}.$$

Note that under condition C0,

$$f(t_i;\lambda_0,\alpha_0) = g(t_i;\lambda_0,\alpha_0) \quad \text{and} \quad \frac{\partial g(t_i;\lambda_0,\alpha_0)/\partial\lambda}{g(t_i;\lambda_0,\alpha_0)} = X_i.$$

Define $\mathbf{b}_i = (X_i, Y_{i1}, Y_{i2})^\top$ and denote the variance-covariance matrix

$$\mathbf{B} = \mathbb{V}\mathrm{ar}(\mathbf{b}_i) = \begin{pmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{pmatrix}, \tag{3.2}$$

where the variance is taken with respect to the null model. Furthermore, define

$$\sigma_{11} = B_{33} - \frac{B_{13}^2}{B_{11}}, \quad \sigma_{12} = B_{23} - B_{33} - \frac{B_{12}B_{13}}{B_{11}} + \frac{B_{13}^2}{B_{11}},$$
$$\sigma_{22} = B_{22} + B_{33} - 2B_{23} - \frac{B_{12}^2}{B_{11}} - \frac{B_{13}^2}{B_{11}} + \frac{2B_{12}B_{13}}{B_{11}}.$$

For any $p_1, p_2 \in [0,1]$, let

$$\sigma(p_1,p_2) = p_1p_2\sigma_{22} + (p_1+p_2)\sigma_{12} + \sigma_{11}. \tag{3.3}$$

Our asymptotic results about R_n rely on conditions C1-C5, given in Section S3 of the Supplementary Material; they are typical regularity conditions in the literature on finite mixture models.

Theorem 2. *Suppose that conditions* C0 *and* C1-C5 *in the Supplementary Material are satisfied. Under model* (1.2) *and the null hypothesis in* (3.1)*, as* $n \to \infty$*,*

$$R_n \to R = \sup_{0\le p\le 1} Z^2(p)$$

in distribution, where $Z(p)$ *is a Gaussian process with zero mean, unit variance, and covariance function*

$$\mathbb{C}\mathrm{ov}\{Z(p_1), Z(p_2)\} = \frac{\sigma(p_1,p_2)}{\sqrt{\sigma(p_1,p_1)\sigma(p_2,p_2)}}, \quad 0 \le p_1, p_2 \le 1.$$

Theorem 2 shows that the LRT statistic R_n has a nonregular limiting distribution that is the supremum of a χ^2-process, and, in general, it does not have a closed form and is difficult to calculate numerically. Instead, we derive an equivalent representation of R that is simpler in form, and more convenient in terms of calculating the distribution function or quantiles of R using the Monte Carlo method.

We require some additional notation. Consider the following polar transformation: $(\cos\theta, \sin\theta) = \big(c_1(p), c_2(p)\big)$, where

$$c_1(p) = \frac{\sqrt{\sigma_{11} - \sigma_{12}^2/\sigma_{22}}}{\sqrt{\sigma(p,p)}} \quad \text{and} \quad c_2(p) = \frac{(p + \sigma_{12}/\sigma_{22})\sqrt{\sigma_{22}}}{\sqrt{\sigma(p,p)}}.$$

To find a simple representation for R, we require the following additional condition.

C6. There exist Δ_1 and Δ_2 such that $-\pi/2 < \Delta_1 < \Delta_2 < \pi/2$ and

$$\big\{\big(c_1(p), c_2(p)\big) : 0 \le p \le 1\big\} = \{(\cos\theta, \sin\theta) : \Delta_1 \le \theta \le \Delta_2\}.$$

Under condition C6, we define the three sets

$$\begin{aligned} A_1 &= \{\eta : \max_{\theta\in[\Delta_1,\Delta_2]} \cos^2(\theta - \eta) = 1\}, \\ A_2 &= \{\eta : \max_{\theta\in[\Delta_1,\Delta_2]} \cos^2(\theta - \eta) = \cos^2(\eta - \Delta_2)\}, \\ A_3 &= \{\eta : \max_{\theta\in[\Delta_1,\Delta_2]} \cos^2(\theta - \eta) = \cos^2(\eta - \Delta_1)\}. \end{aligned}$$

If both Δ_1 and Δ_2 are positive, then these sets have the following explicit forms:

$$\begin{aligned} A_1 &= [\Delta_1, \Delta_2] \cup [\Delta_1 - \pi, \Delta_2 - \pi], \\ A_2 &= \left[\Delta_2, \Delta + \frac{\pi}{2}\right] \cup \left[\Delta_2 - \pi, \Delta - \frac{\pi}{2}\right], \\ A_3 &= \left[\Delta + \frac{\pi}{2}, \pi\right] \cup [-\pi, \Delta_1 - \pi] \cup \left[\Delta - \frac{\pi}{2}, \Delta_1\right], \end{aligned} \tag{3.4}$$

where $\Delta = (\Delta_1 + \Delta_2)/2$. Figure 1 shows A_1-A_3 graphically when $f(t;\lambda,\alpha)$ is a Weibull pdf.

Theorem 3. *Assume the conditions of Theorem* 2 *and condition* C6 *hold. Furthermore, suppose that* ρ^2 *and* η *are two independent random variables that follow a* χ^2_2 *and a uniform distribution on* $[-\pi, \pi]$*, respectively. Then,* R *has the same distribution as*

$$T(\rho^2, \eta) = \rho^2\{I(\eta \in A_1) + I(\eta \in A_2)\cos^2(\eta - \Delta_2) + I(\eta \in A_3)\cos^2(\eta - \Delta_1)\}.$$

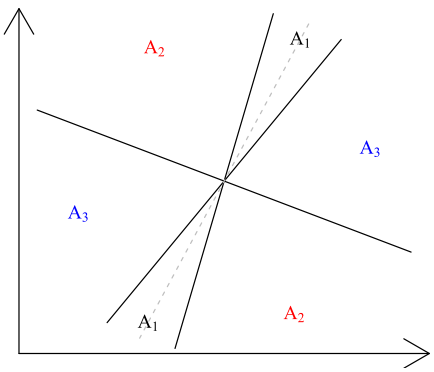

Figure 1. Graphical representation of sets A_1, A_2, and A_3 when $f(t;\lambda,\alpha)$ is a Weibull pdf.

Note that Δ_1, Δ_2, and A_1-A_3 may depend on λ_0. We can estimate λ_0 using $\hat{\lambda}_0$, the MLE of λ under the null model. Based on Theorem 3, we propose the following Monte Carlo procedure for approximating the distribution and quantiles of R. First, we generate a large number (e.g., $M = 10^8$) of independent copies of (ρ^2, η), denoted by (ρ_i^2, η_i) $(i = 1, \ldots, M)$. Then, we take the empirical distribution of $\{T(\rho_i^2, \eta_i), i = 1, \ldots, M\}$ to approximate the distribution of R. Accordingly, we can calculate the approximate p-value of the LRT or the approximate quantiles of R, which may serve as critical values of the proposed LRT.

The results in Theorems 2 and 3 rely on the forms of $\sigma(\cdot,\cdot)$ in (3.3) and (Δ_1, Δ_2) in condition C6. In the following, we identify two examples satisfying conditions C0-C6, and work out their $\sigma(\cdot,\cdot)$ and (Δ_1, Δ_2).

Example 1 (Weibull distribution). Recall that the pdf of a Weibull distribution is given as $f(t;\lambda,\alpha) = \lambda\alpha(t\lambda)^{\alpha-1}\exp\{-(\lambda t)^\alpha\}I(t>0)$. It can be shown that $\sigma(p_1,p_2) = p_1p_2(\pi^2/6-1) + (p_1+p_2)(2-\pi^2/6) + \pi^2/3 - 3$ and

$$\Delta_1 = \arccos\left(\sqrt{\frac{\pi^4 - 6\pi^2 - 36}{2\pi^4 - 30\pi^2 + 108}}\right), \quad \Delta_2 = \arccos\left(\sqrt{\frac{\pi^4 - 6\pi^2 - 36}{\pi^4 - 6\pi^2}}\right).$$

Because both Δ_1 and Δ_2 are positive, A_1-A_3 take the forms in (3.4).

Example 2 (Gamma distribution). Recall that the pdf of a gamma distribution is given as $f(t;\lambda,\alpha) = \{\Gamma(\alpha)\}^{-1}\lambda^\alpha t^{\alpha-1}\exp(-\lambda t)I(t>0)$. It can be shown that $\sigma(p_1,p_2) = p_1p_2\,(\pi^2/6 - 5/4) + (p_1+p_2)\,(7/4 - \pi^2/6) + \pi^2/3 - 13/4$ and

$$\Delta_1 = \arccos\left\{\sqrt{\frac{4\pi^4 - 54\pi^2 + 144}{(4\pi^2 - 39)(2\pi^2 - 15)}}\right\},$$

$$\Delta_2 = \arccos\left\{\sqrt{\frac{4\pi^4 - 54\pi^2 + 144}{(2\pi^2 - 12)(2\pi^2 - 15)}}\right\}.$$

Again, both Δ_1 and Δ_2 are positive, so A_1-A_3 take the forms in (3.4).

As we can see, $\sigma(\cdot,\cdot)$ and (Δ_1, Δ_2) for a Weibull or gamma distribution are independent of λ_0. Thus, there is no need to estimate λ_0 when using Theorem 3 for these two distributions.

3.3. Asymptotic power of the LRT

In this subsection, we study the asymptotic power of the proposed LRT. We consider the following sequence of local alternatives that are indexed by n:

$$H_a^n : \lambda = \lambda_0,\ p = p_0,\ \alpha = \alpha_0 + \delta n^{-1/2}, \tag{3.5}$$

where δ is a fixed constant, independent of n. The following theorem presents the asymptotic distribution of R_n under H_a^n.

Theorem 4. *Assume the conditions of Theorem* 2 *hold. Under the local alternative hypothesis* H_a^n *in* (3.5)*, as* $n \to \infty$*,*

$$R_n \to \sup_{0\le p\le 1} \{Z(p) + \omega(p, p_0)\}^2 \tag{3.6}$$

in distribution, where $\omega(p, p_0) = \delta\sigma(p, p_0)/\sqrt{\sigma(p, p)}$*, and* $Z(p)$ *is defined in Theorem* 2*.*

Note that the result in Theorem 4 has two important applications. First, it is useful for local power analysis for a potential alternative model with the model parameters (λ, α, p). We can insert this model into the local sequence and obtain $\delta = n^{1/2}(\alpha - \alpha_0)$. Then, we can assess the power of R_n for detecting this alternative model based on the limiting distribution under the local alternative. Second, the result in Theorem 4 provides insight on the power trend under different alternative models; for example, if $f(t; \lambda, \alpha)$ is the pdf of a Weibull distribution, then $|\omega(p, p_0)|$ increases as δ departs from zero or p_0 increases. This implies that the power of R_n increases as α departs from $\alpha_0 = 1$ and/or the value of p under the alternative model increases. This trend is confirmed in the following simulation study.

4. Simulations

In this section, we use simulations to check whether the limiting distribution of R_n provides an accurate approximation of its finite-sample distribution. We consider four sample sizes: $n = 100$, 200, 500, and 1000. Following Qin et al. (2020), we choose $f(t; \lambda, \alpha)$ to be a Weibull pdf, and set the true value of λ to

Table 1. Type-I error rates (in %) of R_n at a significance level of 10%, 5%, or 1%.

n	Significance level		
	10%	5%	1%
100	10.6	5.4	1.1
200	10.2	5.2	1.1
500	10.1	5.1	1.0
1,000	10.1	5.0	1.0

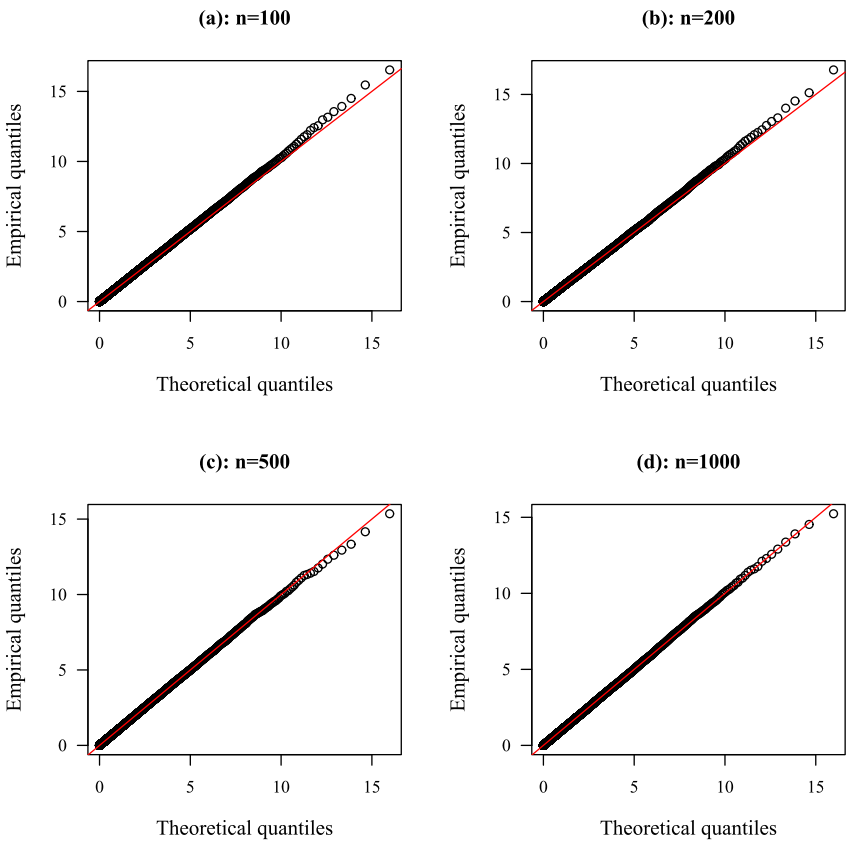

Figure 2. Quantile-quantile plots of R_n for different sample sizes.

one. Note that under H_0 in (3.1), the true value of α is one and p disappears. The simulated type-I errors of R_n based on 10^5 repetitions are summarized in Table 1. The simulation results show that the proposed LRT has tight control of type-I error rates for all combinations of sample size and significance level. Figure 2 shows the quantile-quantile plots of the LRT. As can be seen, the limiting null distribution of R_n provides an adequate approximation of its finite-sample distribution, even when the sample size is as small as 100.

Table 2. Power (in %) of R_n at a significance level of 10%, 5%, or 1%.

n	Significance level			Significance level		
	10%	5%	1%	10%	5%	1%
	$(p, \alpha) = (0.15, 1.35)$			$(p, \alpha) = (0.15, 1.65)$		
100	58.4	45.3	22.4	89.7	81.9	59.5
200	81.9	72.2	47.5	99.2	98.2	92.5
500	99.2	98.1	92.0	100.0	100.0	100.0
1,000	100.0	100.0	99.9	100.0	100.0	100.0
	$(p, \alpha) = (0.40, 1.35)$			$(p, \alpha) = (0.40, 1.65)$		
100	76.7	65.3	39.4	97.8	95.4	84.7
200	95.0	90.4	74.0	100.0	100.0	99.5
500	100.0	99.9	99.6	100.0	100.0	100.0
1,000	100.0	100.0	100.0	100.0	100.0	100.0
	$(p, \alpha) = (0.65, 1.35)$			$(p, \alpha) = (0.65, 1.65)$		
100	90.2	82.7	60.2	99.9	99.7	97.9
200	99.4	98.6	93.0	100.0	100.0	100.0
500	100.0	100.0	100.0	100.0	100.0	100.0
1,000	100.0	100.0	100.0	100.0	100.0	100.0

Next, we evaluate the power of the proposed LRT. We consider two true values of α, namely, 1.35 and 1.65, and three true values of p, namely, 0.15, 0.40, and 0.65. The simulated powers based on 10^4 repetitions are summarized in Table 2. We observe that the proposed LRT exhibits appreciable power in all the cases considered. Furthermore, its power increases as p or α increases. This trend agrees with the local power analysis after Theorem 4.

5. Application to COVID-19 Data

The outbreak of COVID-19 in Wuhan, China, in December 2019 attracted worldwide attention (Li et al. (2020); Wang et al. (2020a); Tu et al. (2020)). To prevent its spread, the Chinese government decided to lock down Wuhan on January 23, 2020. From public reports, many people left Wuhan before the lockdown, with no symptoms of COVID-19, but then later developed symptoms outside Wuhan.

Deng et al. (2021) provide data based on confirmed cases of COVID-19 reported in publicly available sources, such as provincial and municipal health commissions in China and the health authorities in other countries, as of February 15, 2020. The duration time for a patient was recorded as the time difference between leaving Wuhan and the earliest onset of symptoms (e.g., fever, cough). Our analysis involves a sample size of 1,211 cases and satisfies the design criteria of the mixture forward-incubation-time epidemic model (1.2). These criteria include the following. (1) The included cases were of people who left Wuhan

showing no COVID-19 symptoms, but then later developed symptoms elsewhere after traveling. Hence, cases of people whose first symptoms occurred before traveling were not included in the sample. (2) The date of leaving Wuhan had to be between January 19, 2020, and January 23, 2020, for the following reasons: (2a) before January 19, 2020, the public were as yet unaware of the severity of COVID-19, so there may have been a chance that a patient was actually infected outside Wuhan after they left; (2b) after January 23, 2020 (the date of the Wuhan lockdown), not many cases are available, resulting in an average follow-up time for the onset of symptoms as long as 25 days. This sample size of 1211 is relatively large compared with other incubation-period studies on COVID-19.

Following Qin et al. (2020), we use model (1.2) with $f(t;\lambda,\alpha)$ a Weibull pdf to analyze the 1,211 observed duration times. At the beginning of the outbreak, it was more likely to observe someone who had been infected closer to their departure date, because the number of infections grew exponentially. This may invalidate the assumptions for deriving the forward time distribution (Qin et al. (2020); Liu et al. (2020b)), so we are concerned about the validity of the model assumptions in (1.2) for the 1,211 observed duration times. To address this concern, Deng et al. (2021) performed a goodness-of-fit test for model (1.2). The asymptotic p-value of this test is found to be 0.37, which indicates that model (1.2) with $f(t;\lambda,\alpha)$ being a Weibull pdf provides a reasonable fit to the 1,211 observed duration times; see the Supplementary Material for more details. Next, we test for $\alpha = 1$ or, equivalently, whether the data come from a homogeneous exponential distribution, by using the proposed LRT when $f(t;\lambda,\alpha)$ is a Weibull pdf.

All observed duration times are integers between zero and 22 days, and, in theory, our proposed method may not be directly applicable. For illustration, we impute the value of the observed integer value i by using a random number from $\mathrm{U}(i, i+1)$, the uniform distribution on $(i, i+1)$; for example, the frequency for zero days is 82. Therefore, we generate 82 observations from $\mathrm{U}(0,1)$. After that, we apply the proposed testing procedure to the imputed data set. We repeat the procedure 1,000 times and obtain 1,000 estimates of (λ,α,p) and 1,000 LRT statistics R_n. Based on these 1,000 repetitions, the averages of the estimates for (λ,α,p) are $(0.655, 0.135, 1.645)$. The values of R_n range from 202.9 to 234.3, and because the p-value of any LRT statistic in $[202.9, 234.3]$ is almost zero, this provides overwhelming evidence for rejecting the null hypothesis of $\alpha = 1$.

We also analyze the data after adding 0.5 to each duration time, that is, any integer datum i is replaced with the midpoint of the interval $(i, i+1)$. The resulting R_n is around 230.7, with a p-value still almost zero. From both analyses, we conclude with statistical significance that the population distribution of the observed duration times cannot be modeled well enough by an exponential distribution.

The above analysis results indicate that the data contain heterogeneous subgroups. Unfortunately, we have no idea who in the cohort contracted the disease before, and who did so immediately upon departure, so it is more reasonable to use the mixture forward-incubation-time epidemic model (1.2) than using a homogeneous exponential distribution to model the observed duration times.

6. Conclusion

In this paper, we have provided sufficient conditions for the identifiability of the parameters in model (1.2) and applied the results to Weibull, gamma, and lognormal distributions. We have also proposed an LRT for testing the null hypothesis that $h(t;\lambda,\alpha,p)$ in (1.2) is the pdf of a homogeneous exponential distribution, and have derived the limiting distribution of the LRT under the null model and under a sequence of local alternatives. Our simulation results and an analysis of COVID-19 outbreak data demonstrate the usefulness of the LRT. These results strengthen the epidemiological application of the mixture forward-incubation-time epidemic model and enrich the literature on COVID-19 data analysis.

The proposed method relies on the model assumptions in (1.2). When analyzing different data sets for COVID-19 or for a new infectious virus, a goodness-of-fit test for the model assumptions in (1.2) is required before using the proposed LRT. We may also model the incubation-period distribution $f(t)$ nonparametrically in (1.1). However, (p, f) may not be identifiable under this setup. Some reasonable assumptions are required to ensure model identifiability, and we leave this as a future research topic.

Supplementary Material

The online Supplementary Material contains a derivation of the form of $g(t)$, a goodness-of-fit test for model (1.2), conditions C1–C5, and proofs of Theorems 1–4.

Acknowledgments

The authors are grateful to the co-editor, associate editor, and two referees for their constructive and insightful comments and suggestions. Wang's research was supported by the National Natural Science Foundation of China (12001454, 71988101) and the Open Research Fund of the Key Laboratory of Advanced Theory and Application in Statistics and Data Science (East China Normal University), Ministry of Education (KLATASDS2006). Li's research was supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2020-04964). Liu's research was supported by the National Natural

Science Foundation of China (12171157, 71931004, 32030063, 11771144) and the 111 project (B14019). Liu is the corresponding author. The first two authors contributed equally to this paper.

References

Backer, J. A., Klinkenberg, D. and Wallinga, J. (2020). Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. *Eurosurveillance* **25**, 2000062.

Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: The EM approach. *The Annals of Statistics* **37**, 2523–2542.

Chen, J., Li, P. and Liu, G. (2020). Homogeneity testing under finite location-scale mixtures. *The Canadian Journal of Statistics* **48**, 670–684.

Cohen, J. and Kupferschmidt, K. (2020). Countries test tactics in 'war' against COVID-19. *Science* **367**, 1287–1288.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64**, 247–254.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **74**, 33–43.

Deng, Y., You, C., Liu, Y., Qin, J. and Zhou, X.-H. (2021). Estimation of incubation period and generation time based on observed length-biased epidemic cohort with censoring for COVID-19 outbreak in China. *Biometrics* **77**, 929–941.

Farewell, V. T., Herzberg, A. M., James, K. W., Ho, L. M. and Leung, G. M. (2005). SARS incubation and quarantine times: When is an exposed individual known to be disease free? *Statistics in Medicine* **24**, 3431–3445.

Guan, W.-J., Ni, Z.-Y., Hu, Y., Liang, W.-H., Ou, C.-Q., He, J.-X. et al. (2020). Clinical characteristics of coronavirus disease 2019 in China. *New England Journal of Medicine* **382**, 1708–1720.

Kalbfleisch, J. D. and Lawless, J. F. (1989). Estimating the incubation time distribution and expected number of cases of transfusion-associated acquired immune deficiency syndrome. *Transfusion* **29**, 672–676.

Kalbfleisch, J. D. and Lawless, J. F. (1991). Regression models for right truncated data with applications to AIDS incubation times and reporting lags. *Statistica Sinica* **1**, 19–32.

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R. et al. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine* **172**, 577–582.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y. et al. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *New England Journal of Medicine* **382**, 1199–1207.

Linton, N. M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A. R., Jung, S.-M. et al. (2020). Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *Journal of Clinical Medicine* **9**, 538.

Liu, G., Li, P., Liu, Y. and Pu, X. (2020a). Hypothesis testing for quantitative trait locus effects in both location and scale in genetic backcross studies. *Scandinavian Journal of Statistics* **47**, 1064–1089.

Liu, X., He, Y., Ma, X. and Luo, L. (2020b). Analysis on the incubation and suspected period of COVID-19 based on 2172 confirmed cases outside Hubei province. *Acta Mathematicae*

Applicatae Sinica **43**, 278–294.

Liu, X., Ma, H. and Jiang, J. (2022). That Prasad-Rao is robust: Estimation of mean squared prediction error of observed best predictor under potential model misspecification. *Statistica Sinica* **32**, 2217–2240.

Liu, X. and Shao, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *The Annals of Statistics* **31**, 807–832.

Liu, X., Wang, L., Ma, X., Wang, J. and Wu, L. (2021). Modeling the effect of age on quantiles of the incubation period distribution of COVID-19. *BMC Public Health* **21**, 1762.

Qin, J. (2017). *Biased Sampling, Over-identified Parameter Problems and Beyond*. Springer, Singapore.

Qin, J., You, C., Lin, Q., Hu, T., Yu, S. and Zhou, X.-H. (2020). Estimation of incubation period distribution of COVID-19 using disease onset forward time: A novel cross-sectional and forward follow-up study. *Science Advances* **6**, eabc1202.

Rahman, J., Luo, S., Fan, Y. and Liu, X. (2020). Semiparametric efficient inferences for generalised partially linear models. *Journal of Nonparametric Statistics* **32**, 704–724.

Sartwell, P. E. (1950). The distribution of incubation periods of infectious disease. *American Journal of Epidemiology* **51**, 310–318.

Struthers, C. A. and Farewell, V. T. (1989). A mixture model for time to AIDS data with left truncation and an uncertain origin. *Biometrika* **76**, 814–817.

Tu, W., Tang, H., Chen, F., Wei, Y., Xu, T., Liao, K. et al. (2020). Epidemic update and risk assessment of 2019 novel coronavirus – China, January 28, 2020. *China CDC Weekly* **2**, 83–86.

Verity, R., Okell, L. C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N. et al. (2020). Estimates of the severity of coronavirus disease 2019: A model-based analysis. *The Lancet Infectious Diseases* **20**, 669–677.

Wang, C., Horby, P. W., Hayden, F. G. and Gao, G. F. (2020a). A novel coronavirus outbreak of global health concern. *The Lancet* **395**, 470–473.

Wang, L., Zhou, Y., He, J., Zhu, B., Wang, F., Tang, L. et al. (2020b). An epidemiological forecast model and software assessing interventions on the COVID-19 epidemic in China (with discussion). *Journal of Data Science* **18**, 409–432.

Wilkening, D. A. (2008). Modeling the incubation period of inhalational anthrax. *Medical Decision Making* **28**, 593–605.

Chunlin Wang

Department of Statistics, Xiamen University, Xiamen 361005, China.

E-mail: wangc@xmu.edu.cn

Pengfei Li

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada.

E-mail: pengfei.li@uwaterloo.ca

Yukun Liu

Key Laboratory of Advanced Theory and Application in Statistics and Data Science - MOE, and School of Statistics, East China Normal University, Shanghai 200241, China.

E-mail: ykliu@sfs.ecnu.edu.cn

Xiao-Hua Zhou
Department of Biostatistics, Peking University, Beijing 100871, China.
E-mail: azhou@math.pku.edu.cn
Jing Qin
National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, USA.
E-mail: jingqin@niaid.nih.gov

(Received August 2020; accepted March 2023)