



Semiparametric regression analysis of doubly-censored data with applications to incubation period estimation

Kin Yau Wong¹ · Qingning Zhou² · Tao Hu³

Received: 24 October 2021 / Accepted: 28 June 2022 / Published online: 13 July 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

The incubation period is a key characteristic of an infectious disease. In the outbreak of a novel infectious disease, accurate evaluation of the incubation period distribution is critical for designing effective prevention and control measures. Estimation of the incubation period distribution based on limited information from retrospective inspection of infected cases is highly challenging due to censoring and truncation. In this paper, we consider a semiparametric regression model for the incubation period and propose a sieve maximum likelihood approach for estimation based on the symptom onset time, travel history, and basic demographics of reported cases. The approach properly accounts for the pandemic growth and selection bias in data collection. We also develop an efficient computation method and establish the asymptotic properties of the proposed estimators. We demonstrate the feasibility and advantages of the proposed methods through extensive simulation studies and provide an application to a dataset on the outbreak of COVID-19.

Keywords COVID-19 · Cox proportional hazards model · Sieve estimation · Survival analysis · Truncated data

1 Introduction

The incubation period is an important epidemiological feature of infectious diseases. It is defined as the time elapsed from infection of the disease to the onset of symptoms. Accurate evaluation of the incubation period distribution of a novel infectious disease

✉ Qingning Zhou
qzhou8@uncc.edu

¹ Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, People's Republic of China

² Department of Mathematics and Statistics, The University of North Carolina at Charlotte, Fretwell 335L, 9201 University City Blvd., Charlotte, NC 28223, USA

³ School of Mathematical Sciences, Capital Normal University, Beijing, People's Republic of China

can inform effective prevention and control measures, such as the appropriate duration of quarantine and the time to reopen workplaces, schools, and restaurants after a lockdown. Also, the distribution of the incubation period can be used to estimate the reproduction number R_0 and to predict transmission dynamics and epidemic trends. Typically, early in a pandemic, only limited information, such as the travel history, contact history, symptom onset date, or hospitalization date of infected cases, is available for analyzing the incubation period. As a result, estimation of the incubation period distribution is highly challenging.

This study is motivated by the outbreak of coronavirus disease 2019 (COVID-19), which first occurred in early December 2019 in Wuhan, China. Our study is based on the data collected by Zhao et al. (2021), which contain information typically available early in disease outbreaks. The dataset consists of confirmed COVID-19 cases reported by local health agencies at 14 locations, including 8 cities/provinces in mainland China and 6 countries/regions in East Asia, collected in mid to late February 2020. These cases had visited or were residents of Wuhan before Wuhan's lockdown on 23 January 2020 and were diagnosed after arrival at the 14 locations. Because Wuhan was the first center of epidemic outbreak, the cases were plausibly infected during their stay at Wuhan. Available information of the cases includes basic demographics, date of arrival at Wuhan (for travellers), date of departure from Wuhan, and date of symptom onset.

Estimation of the incubation period distribution is complicated by censoring and truncation. First, the start of the incubation period, that is, the infection time, is not observed but is only known to fall within an interval (e.g., the stay at Wuhan in our example), which we refer to as the exposure period. Also, a subject may be tested to be infected before symptom onset and is not followed thereafter, in which case the symptom onset time is right-censored. Since both the infection and symptom onset times may be censored, we say that the data are doubly-censored. In addition, data were collected retrospectively, so only subjects who were infected during the exposure period could be included in the dataset, and the infection time is thus right-truncated.

Several investigators have studied the analysis of doubly-censored data for the estimation of incubation period distribution. Goggins et al. (1999) considered a full-likelihood approach that includes both the infection time and incubation period distributions in the likelihood and proposed a Monte-Carlo EM algorithm for computation. Sun et al. (1999, 2004) proposed a two-step approach, where the infection time distribution is estimated using the intervals containing the infection times, and the incubation period distribution is estimated using estimating equations based on the estimated infection time distribution. Pan (2001) proposed a multiple-imputation approach to impute the infection time and applied conventional estimation methods to the imputed datasets. Reich et al. (2009) considered interval-censored infection and symptom onset times, that is, doubly interval-censored data, and studied maximum likelihood estimation under a parametric model without covariates. Dejaridin and Lesaffre (2013) considered a full-likelihood approach and developed a stochastic EM algorithm for doubly interval-censored data. Li and Owzar (2016) considered doubly interval-censored data and allowed the infection time and incubation period to be associated. Li et al. (2020) considered interval-censored infection times and a mixture of exactly-observed and interval-censored symptom onset times.

Although our data structure bears superficial resemblance with those considered in the literature on the analysis of doubly-censored data, there are important differences that make the current analysis substantially more difficult. In particular, information on the infection time is very limited in the current study. First, there is essentially only one “observation time” for the infection (e.g., the time of departure from Wuhan in our example), so the infection time is subject to case-1 interval censoring instead of general interval censoring, yielding much less information. Second, subjects with infection time larger than the observation time could not be observed, so there is a severe truncation problem. In fact, the observation times alone contain no information about the infection time distribution. Thus, imputation or estimating-equation approaches of Sun et al. (1999, 2004), Pan (2001), and Li et al. (2020), which require a consistent estimator of the infection time distribution, are not applicable.

We propose a likelihood approach to estimate the incubation period distribution. In particular, we jointly model the infection time and time to symptom onset, while taking into account the censoring and truncation issues and the growth of the pandemic over time. We assume a semiparametric Cox proportional hazards model for the incubation period, propose a sieve maximum likelihood approach for estimation, and devise an efficient computation method. We also establish the consistency, rate of convergence, and asymptotic normality of the proposed estimators.

The current study differs from most existing work on incubation period analysis for COVID-19 in two important aspects. First, most existing studies are based on parametric models (Backer et al. 2020; Lauer et al. 2020; Linton et al. 2020; Deng et al. 2020; Qin et al. 2020), whereas we consider a flexible semiparametric model with a nonparametric baseline hazard function. In particular, based on infected subjects’ time to symptom onset since their departure from Wuhan, Deng et al. (2020) and Qin et al. (2020) estimated the (parametric) incubation period distribution using renewal process theory with the incubation period treated as a renewal and the duration between departure and symptom onset as a forward recurrence time. Second, we allow the incubation period distribution to depend on covariates, such as sex and age. While most existing studies focused on the overall incubation period, some studies suggested that the incubation period distribution may differ across patients from different subpopulations. For example, Dai et al. (2020) and Tan et al. (2020) found that older subjects have longer median incubation period than younger subjects, and Xiao et al. (2021) suggested that the incubation period distribution is associated with the meteorological temperature of the area. A regression framework allows us to more precisely evaluate the incubation period distribution for given subpopulations. Such evaluations are important for designing targeted quarantine and isolation policies.

The rest of this paper is structured as follows. In Sect. 2, we formulate the model, define the sieve maximum likelihood estimator (MLE), and describe the numerical implementation of the proposed methods. In Sect. 3, we present asymptotic properties of the proposed estimators. In Sect. 4, we report the results from simulation studies, and in Sect. 5, we present the analysis results of the COVID-19 dataset. We make some concluding remarks in Sect. 6. Technical details are relegated to the Appendix.

2 Methods

For a generic subject, let τ be the exposure start time, measured from a fixed initial calendar time. Let S be the time to infection since the exposure start, \tilde{E} be the time to symptom onset since the exposure start, and \mathbf{X} be a p -vector of covariates. Assume that $\kappa(t; \alpha, r, \tau) \equiv \alpha e^{r(t+\tau)}$ is the hazard of infection at time t (since the exposure start), where α and r are positive parameters, α represents the hazard at the initial calendar time, and r characterizes the growth of the pandemic. This model assumes that the hazard of infection grows exponentially with rate r since the initial calendar time. When $\alpha \rightarrow 0$, the model reduces to the model for the time to infection considered in Zhao et al. (2021). Assume that the incubation period $(\tilde{E} - S)$ follows the proportional hazards model with covariates \mathbf{X} , so the hazard function is $\lambda(t; \mathbf{X}) = \lambda(t)e^{\mathbf{X}^T \boldsymbol{\beta}}$, where λ is an unspecified baseline hazard function, and $\boldsymbol{\beta}$ is a vector of regression parameters. Assume that S and $(\tilde{E} - S)$ are independent.

Suppose that each subject is exposed to the risk of infection until time U (measured from the exposure start), and a subject can be observed only if the infection occurs before U . Assume that U is independent of (S, \tilde{E}) . The conditional density of (S, \tilde{E}) given \mathbf{X} , U , and the event $\{S < U\}$ is

$$\begin{aligned} f(S, \tilde{E} \mid \mathbf{X}, U, S < U) &= f_T(\tilde{E} - S \mid \mathbf{X}, S) f_S(S \mid U, S < U) \\ &= \lambda(\tilde{E} - S) e^{\mathbf{X}^T \boldsymbol{\beta}} \exp \left\{ -\Lambda(\tilde{E} - S) e^{\mathbf{X}^T \boldsymbol{\beta}} \right\} \frac{\kappa(S; \alpha, r, \tau) e^{-K(S; \alpha, r, \tau)}}{1 - e^{-K(U; \alpha, r, \tau)}}, \end{aligned}$$

where $\Lambda(t) = \int_0^t \lambda(s) ds$, $K(t; \alpha, r, \tau) = \int_0^t \kappa(s; \alpha, r, \tau) ds$, and f_T and f_S are (conditional) densities of $(\tilde{E} - S)$ and S , respectively.

We allow the symptom onset time to be right-censored. Let C be the censoring time (measured from the exposure start), $E \equiv \min(\tilde{E}, C)$ be the observed event or censoring time, and $\Delta \equiv I(\tilde{E} \leq C)$ be the event indicator. Also, S is unobserved but is only known to fall within $(0, U)$. For a sample of size n , the observed data consist of $(U_i, \tau_i, E_i, \Delta_i, \mathbf{X}_i)$ for $i = 1, \dots, n$. The conditional likelihood function for $(\boldsymbol{\beta}, \alpha, r, \lambda)$ given the event that infection occurs before the end of exposure is

$$\begin{aligned} L_n(\boldsymbol{\beta}, \alpha, r, \lambda) &= \prod_{i=1}^n \frac{e^{\Delta_i \mathbf{X}_i^T \boldsymbol{\beta}}}{1 - e^{-K(U_i; \alpha, r, \tau_i)}} \int_0^{\min(E_i, U_i)} \lambda(E_i - s)^{\Delta_i} \exp \left\{ -\Lambda(E_i - s) e^{\mathbf{X}_i^T \boldsymbol{\beta}} \right\} \\ &\quad \times \kappa(s; \alpha, r, \tau_i) e^{-K(s; \alpha, r, \tau_i)} ds. \end{aligned}$$

The conditional likelihood incorporates the infection time distribution and thereby accounts for the pandemic growth over time. Also, because only subjects who were infected during the exposure period could be observed, there is selection bias induced by truncation of infection times. To account for such selection bias, we consider the conditional likelihood given $S < U$.

Because the (conditional) likelihood function involves the nonparametric function λ , maximum likelihood estimation is highly challenging or even infeasible. We propose a sieve maximum likelihood approach and approximate λ by B-spline functions. In particular, let b_1, \dots, b_{m_n} be a set of B-spline functions of order l over a knot sequence $0 = t_1 = \dots = t_l < t_{l+1} < \dots < t_{m_n} < t_{m_n+1} = \dots = t_{m_n+l} = \zeta$ for some fixed time point ζ . In practice, we can set $\zeta = \max_i E_i$. Define the sieve space

$$\mathcal{B}_n = \left\{ \sum_{j=1}^{m_n} \gamma_j b_j : M_n^{-1} \leq \gamma_j \leq M_n \text{ for } j = 1, \dots, m_n \right\}$$

for some diverging sequence M_n . Let \mathcal{A} be a prespecified compact set in $\mathbb{R}^p \times \mathbb{R}^+ \times \mathbb{R}^+$ that denotes the parameter space for the Euclidean parameters. The sieve MLE is $(\hat{\beta}_n, \hat{\alpha}_n, \hat{r}_n, \hat{\lambda}_n) = \arg \max_{(\beta, \alpha, r) \in \mathcal{A}, \lambda \in \mathcal{B}_n} L_n(\beta, \alpha, r, \lambda)$.

We propose to log-transform the positive parameters and use a hybrid of gradient descent and the Newton–Raphson method to compute the sieve MLE. In particular, in early iterations, we update the parameter estimates using the gradient descent method on the log-likelihood function. Let U_{ϑ} be the vector of first derivatives of $\log L_n(\beta, \alpha, r, \sum_j \gamma_j b_j)$ with respect to $\vartheta \equiv (\beta^T, \log \alpha, \log r, \log \gamma^T)^T$. We initialize the parameter vector at $\hat{\vartheta}^{(0)}$, and at the j th iteration, we use the gradient descent method to update ϑ by

$$\hat{\vartheta}_n^{(j)} = \hat{\vartheta}_n^{(j-1)} - a_j U_{\vartheta} \Big|_{\vartheta = \hat{\vartheta}_n^{(j-1)}},$$

where $\hat{\vartheta}_n^{(j-1)}$ is the current parameter estimates, and a_j is some negative number determined by the Armijo backtracking method. We first set a_j to be the Barzilai–Borwein step size (Sun and Yuan 2006, pp. 126–127). Specifically, let $s_j = \hat{\vartheta}_n^{(j-1)} - \hat{\vartheta}_n^{(j-2)}$ and $y_j = U_{\vartheta}|_{\vartheta = \hat{\vartheta}_n^{(j-1)}} - U_{\vartheta}|_{\vartheta = \hat{\vartheta}_n^{(j-2)}}$. We initially set $a_j = s_j^T s_j / s_j^T y_j$ for $j \geq 2$ and $a_j = -0.01$ for $j = 1$. If the updated log-likelihood value does not pass the Armijo–Goldstein condition, that is, if it is smaller than $\log L_n(\hat{\vartheta}_n^{(j-1)}) + 10^{-4} a_j \|U_{\vartheta}|_{\vartheta = \hat{\vartheta}_n^{(j-1)}}\|^2$ under the current value of a_j , then we recalculate the parameter estimates with a_j discounted by a multiplicative factor of 0.8 until the resulting log-likelihood value passes the condition. When the norm of the score statistic is smaller than a certain threshold, we perform the Newton–Raphson algorithm until convergence. In particular, at the j th iteration, the Newton–Raphson algorithm updates ϑ by

$$\hat{\vartheta}_n^{(j)} = \hat{\vartheta}_n^{(j-1)} - I_{\vartheta}^{-1} U_{\vartheta} \Big|_{\vartheta = \hat{\vartheta}_n^{(j-1)}},$$

where I_{ϑ} is the Hessian matrix for $\log L_n$ with respect to ϑ .

When the parameter estimates are not close enough to the maximum, the Hessian matrix I_{ϑ} may not be negative definite, and there is a direction in the parameter space

along which the log-likelihood function is locally convex. In this case, the Newton–Raphson step should not be performed, and instead we update the parameter estimates along this direction of convexity. Let \mathbf{g} denote the eigenvector that corresponds to the largest (positive) eigenvalue of \mathbf{I}_ϑ . We update ϑ by $\widehat{\vartheta}_n^{(j)} = \widehat{\vartheta}_n^{(j-1)} - a_j \mathbf{g}$, where a_j is selected similarly as in gradient descent.

To compute the log-likelihood function and its derivatives, we use the Legendre–Gauss quadrature to approximate the integrations involved. Because the log-likelihood may possess multiple local maxima, we set the initial parameter values based on a grid search to increase the chance of obtaining the global maximum. In particular, we define a grid for (α, r) . At each grid point (α^*, r^*) , we compute the maximizer of the log-likelihood with respect to (β, γ) , denoted by $(\beta(\alpha^*, r^*), \gamma(\alpha^*, r^*))$, at $(\alpha, r) = (\alpha^*, r^*)$. The parameter values $(\beta(\alpha^*, r^*), \alpha^*, r^*, \gamma(\alpha^*, r^*))$ that yield the largest likelihood value is set to be the initial values for the gradient methods.

When the sample size is small, the likelihood may be maximized at the boundary value of 0 for some positive parameters. In this case, the maximum does not exist for the log-transformed parameters, and the algorithm may return error values at extreme parameter values. To avoid these problems, we propose to set a (small) lower boundary for the positive parameters. When the proposed updated parameter values exceed the boundary, then we update the parameters along the proposed direction but with a step size that keeps all parameters within the boundary.

To estimate the standard error of the Euclidean parameter estimators, we treat the model as fully parametric, with parameters $(\beta, \alpha, r, \gamma)$. Let $\widehat{\mathbf{I}}$ be the observed information matrix under this parametric model evaluated at the sieve MLE. We estimate the variance of $(\widehat{\beta}, \widehat{\alpha}, \widehat{r})$ by the corresponding elements of $\widehat{\mathbf{I}}^{-1}$.

3 Asymptotic properties of the sieve MLE

Let $\theta = (\xi, \lambda) = (\beta, \alpha, r, \lambda)$ denote the collection of all parameters, $\theta_0 = (\xi_0, \lambda_0)$ the true value of θ , and $\widehat{\theta}_n = (\widehat{\xi}_n, \widehat{\lambda}_n)$ the sieve MLE of θ . The regularity conditions below are needed for the forthcoming theorems.

- (C1) The true parameter value ξ_0 is an interior point of \mathcal{A} . The function $\lambda_0(\cdot)$ has a positive lower bound and bounded q th derivative on $[0, \zeta]$ for some $q \geq 2$ and positive constant ζ .
- (C2) The distribution of \mathbf{X} has a bounded support in \mathbb{R}^p . If $\mathbf{a}^T \mathbf{X} = b$ with probability one for some constants $\mathbf{a} \in \mathbb{R}^p$ and $b \in \mathbb{R}$, then $\mathbf{a} = \mathbf{0}$ and $b = 0$.
- (C3) Conditional on \mathbf{X} , S and $(\widetilde{E} - S)$ are independent, and they are independent of C and U . Also, $0 \leq U < C \leq \zeta$, and the support of U includes $[0, \zeta_0]$ for some positive constant $\zeta_0 \leq \zeta$. The joint density function of C and U is continuously differentiable on its support. In addition, $Pr(C = \zeta \mid \mathbf{X}) \geq c_0$ almost surely for some positive constant c_0 .
- (C4) The knots (t_1, \dots, t_{m_n+l}) satisfy

$$\frac{\max_{j=l+1, \dots, m_n+1} |t_j - t_{j-1}|}{\min_{j=l+1, \dots, m_n+1} |t_j - t_{j-1}|} < K$$

- for some positive constant K , and the number of knots satisfies $m_n = O(n^\nu)$ for some $\nu \in (0, 1/3)$, and $M_n = O(\log \log n)$.
- (C5) The exposure start time is uniformly bounded, that is, there exists a positive constant K such that $Pr(\tau \in [0, K]) = 1$.

Remark 1 Condition (C1) is typical for semiparametric regression models, and condition (C2) is necessary for the identifiability of the regression parameters. Condition (C3) ensures that the supports of C and U are wide enough to allow the identification of λ over the interval $[0, \zeta]$. Condition (C4) controls the rates at which the number of B-spline functions and the maximum spline coefficient value diverge to infinity. Condition (C5) requires that the exposure start time is uniformly bounded.

Define the distance $d(\theta_1, \theta_2) = (\|\xi_1 - \xi_2\|^2 + \|\lambda_1 - \lambda_2\|_{L_2[0, \zeta]}^2)^{1/2}$, where $\|\cdot\|$ denotes the Euclidean norm and $\|\cdot\|_{L_2[0, \zeta]}$ denotes the L_2 -norm over $[0, \zeta]$. We have the following results.

Theorem 1 Under conditions (C1)–(C5), $d(\hat{\theta}_n, \theta_0) \rightarrow 0$ almost surely, and

$$d(\hat{\theta}_n, \theta_0) = O_p[n^{-\min\{q\nu, (1-\nu)/2\}}]$$

as $n \rightarrow \infty$, where q is given in (C1).

Theorem 2 Assume that conditions (C1)–(C5) hold, $q > 2$, and $1/\{2(q+1)\} < \nu < 1/5$. We have

$$\sqrt{n}(\hat{\xi}_n - \xi_0) \rightarrow_d N\{0, I^{-1}(\xi_0)\},$$

where $I(\xi_0)$ is the information matrix for ξ defined in the proof of this theorem.

The proofs of Theorems 1 and 2 are given in the [Appendix](#).

4 Simulation studies

We set $U_i \sim \text{Uniform}(0, 10)$ and $\tau_i \sim \text{Uniform}(0, 10)$ for $i = 1, \dots, n$, where U_i and τ_i are independent. We set $\beta = (0.5, -0.5)^T$, $\lambda(t) = 0.05 + 0.015t$, $r = 0.2$ or 0.3 , and $\alpha = 0.01$ or 0.1 . The (nontruncated) means of time to infection and the truncation rates under different values of α and r are given in Table 1. We generated $X \equiv (X_1, X_2)^T$ with $X_1 \sim N(0, 1)$ and $X_2 \sim \text{Bernoulli}(0.5)$, where X_1 and X_2 are independent. The mean incubation period is about 9.4. We set a universal right censoring time for \tilde{E} at 30, resulting in a censoring proportion of around 2–3%. This low censoring rate is to match the real data, where most subjects were identified to be infected at symptom onset, and only subjects who underwent the disease test before symptom onset would be censored. We set the sample size to be $n = 400$ or 800 .

We set the degree of B-spline functions b_j 's to be 2 and a single (inner) grid point to be 8, which is approximately the median of $(\tilde{E} - S)$. The grid for the initial values of (α, r) is $\{0.001, 0.005, 0.01, 0.05, 0.1\} \times \{0.1, 0.2, 0.3, 0.4, 0.5\}$. We set the

Table 1 Mean times to infection and truncation rates in the simulation studies

α	r	Mean time to infection	Truncation Rate
10^{-10}	0.2	99.2	≈ 1
	0.3	65.8	≈ 1
0.01	0.2	9.0	0.75
	0.3	5.9	0.56
0.1	0.2	2.6	0.25
	0.3	1.7	0.17

Table 2 Simulation results under the correct model with $n = 400$

Setting	Parameter	True	Bias	ESD	SEE	Cov
1	β_1	0.500	0.007	0.062	0.060	0.94
	β_2	-0.500	-0.009	0.111	0.110	0.95
	α	0.010	0.004	0.031	0.019	0.94
	r	0.200	0.013	0.104	0.107	0.91
2	β_1	0.500	0.006	0.060	0.060	0.94
	β_2	-0.500	-0.010	0.108	0.109	0.96
	α	0.010	0.005	0.022	0.016	0.93
	r	0.300	0.019	0.110	0.100	0.93
3	β_1	0.500	0.005	0.059	0.058	0.95
	β_2	-0.500	-0.007	0.109	0.107	0.94
	α	0.100	0.018	0.144	0.095	0.95
	r^*	0.200	0.051	0.149	0.111	0.91
4	β_1	0.500	0.006	0.060	0.058	0.94
	β_2	-0.500	-0.007	0.109	0.106	0.95
	α	0.100	0.017	0.168	0.113	0.96
	r^{**}	0.300	0.074	0.191	0.154	0.90

* The true value and bias for $\log r$ are -1.609 and 0.004

** The true value and bias for $\log r$ are -1.204 and 0.041

“True” stands for the true parameter value, “ESD” stands for the empirical standard deviation of the estimated values, “SEE” stands for the mean standard error estimates, and “Cov” stands for the empirical coverage of the 95% confidence interval. The number of replicates in which one or more parameter estimates are at the boundary for Settings 1–4 are 50, 18, 4, and 3, respectively. Replicates with estimates at the boundary are discarded

lower boundary for the positive parameters to be 10^{-8} . We present the bias, empirical standard deviation, mean standard error estimates, and empirical coverage of 95% confidence interval for the Euclidean parameters in Tables 2 and 3; the results are based on 1000 replicates. For α and r , the coverages are calculated under the logarithm scale. We also plot the averaged estimated cumulative baseline hazard functions in Fig. 1 for $n = 400$; the results for $n = 800$ are visually identical and thus are omitted.

For β_1 and β_2 , the estimators are virtually unbiased in all settings, the standard error estimates closely resemble the empirical standard deviations, and the empirical coverages are close to the nominal 95% level. The mean estimated cumulative base-

Table 3 Simulation results under the correct model with $n = 800$

Setting	Parameter	True	Bias	ESD	SEE	Cov
1	β_1	0.500	0.003	0.043	0.042	0.96
	β_2	-0.500	-0.005	0.078	0.078	0.94
	α	0.010	0.001	0.012	0.010	0.95
	r	0.200	0.002	0.077	0.079	0.93
2	β_1	0.500	0.002	0.042	0.042	0.95
	β_2	-0.500	-0.004	0.077	0.077	0.96
	α	0.010	0.003	0.013	0.009	0.94
	r	0.300	0.008	0.069	0.068	0.95
3	β_1	0.500	0.001	0.041	0.041	0.95
	β_2	-0.500	-0.006	0.076	0.076	0.95
	α	0.100	0.010	0.066	0.058	0.95
	r^*	0.200	0.018	0.079	0.070	0.93
4	β_1	0.500	0.002	0.040	0.040	0.95
	β_2	-0.500	-0.005	0.075	0.075	0.96
	α	0.100	0.006	0.062	0.057	0.97
	r^{**}	0.300	0.036	0.127	0.105	0.93

* The true value and bias for $\log r$ are -1.609 and 0.017

** The true value and bias for $\log r$ are -1.204 and 0.050

See NOTE to Table 2. The number of replicates in which one or more parameter estimates are at the boundary for Settings 1–4 are 17, 1, 0, and 0, respectively. Replicates with estimates at the boundary are discarded

line hazard functions closely resemble the true functions. Under $n = 400$, the average values of $\hat{\alpha}$ and \hat{r} may not be close to the true values due to some extreme estimates, and the standard error estimates may be smaller than the empirical standard deviations, yielding under-coverage of the confidence intervals. The performance of the estimators is much better under $n = 800$, suggesting that the suboptimal performance in some cases is due to insufficient sample size. After all, the actual time to infection is never observed and is always mixed with the incubation period, leading to limited information for the estimation of the infection time distribution. When the true values of α and r are small, $\hat{\alpha}$ and $\hat{\gamma}_1$ (the intercept of $\hat{\lambda}$) are at the boundary for a number of replicates, and the number is substantially smaller under a larger sample size. In some cases, \hat{r} has a substantial bias. This is due to skewness of the distribution of \hat{r} , and the bias of $\log \hat{r}$ can be much smaller.

We investigate the performance of the proposed methods under a substantially larger right-censoring proportion. This corresponds to the scenario where many subjects were tested to be infected before symptom onset (and were not followed thereafter). In particular, we generated the data as described above but set a universal censoring time of 15, resulting in a censoring proportion of about 20–35%. Because the information available for the baseline hazard function is much less due to early censoring, we set the degree of B-spline functions to be 1 instead of 2. The results for the Euclidean parameters, based on $n = 400$ and 1000 replicates, are presented in Table 4. The

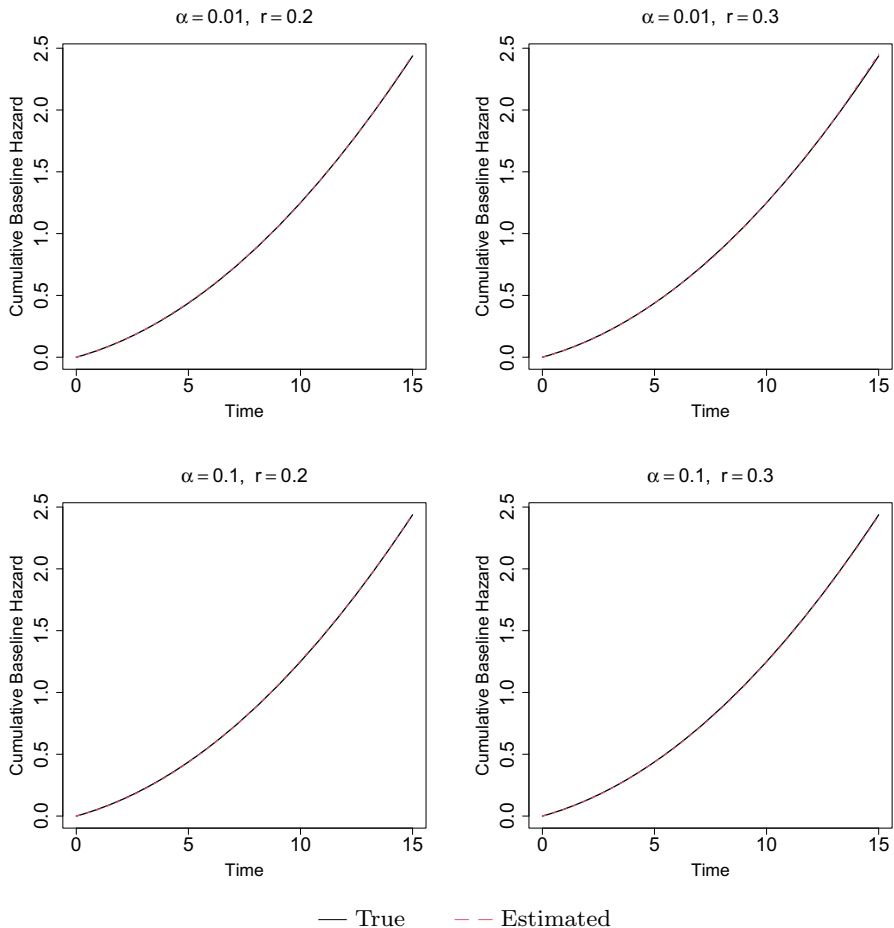


Fig. 1 Estimated cumulative baseline hazard functions under the correct model with $n = 400$

estimated cumulative baseline hazard functions are virtually indistinguishable from the true function and are not presented. Under a larger censoring proportion, the standard error of the estimators are generally larger. The performance of the point and interval estimators has similar pattern as that under a small censoring rate.

To investigate the robustness of the methods, we considered misspecified infection time distributions. In particular, we considered S from the Weibull distribution with shape parameter equal to 2 and scale parameter equal to $5e^{-0.1\tau}$ and also $S = \exp(Z)$, where Z is normal with mean $1.2e^{-0.2\tau}$ and variance 0.64. Other variables were generated according to the original setting. We set the degree of B-spline functions to be 2, $n = 800$, and number of replicates to be 1000. We present the summaries for the estimators of β_1 and β_2 in Table 5 and the averaged estimated cumulative baseline hazard functions in Fig. 2. Under misspecified infection time distributions, the estimation and inference of β are still highly satisfactory. As expected, the estimated cumulative hazard function departs from the true function, but the overall bias is tiny.

Table 4 Simulation results under the correct model with $n = 400$ and a large censoring proportion

Setting	Parameter	True	Bias	ESD	SEE	Cov
1	β_1	0.500	0.007	0.072	0.071	0.95
	β_2	-0.500	-0.008	0.134	0.133	0.95
	α	0.010	0.004	0.025	0.019	0.95
	r	0.200	0.008	0.104	0.105	0.92
2	β_1	0.500	0.006	0.068	0.070	0.96
	β_2	-0.500	-0.009	0.131	0.130	0.95
	α	0.010	0.006	0.026	0.017	0.93
	r	0.300	0.018	0.109	0.100	0.93
3	β_1	0.500	0.005	0.067	0.065	0.94
	β_2	-0.500	-0.005	0.124	0.122	0.96
	α	0.100	0.023	0.214	0.108	0.95
	r^*	0.200	0.050	0.148	0.112	0.91
4	β_1	0.500	0.005	0.066	0.063	0.94
	β_2	-0.500	-0.004	0.121	0.118	0.94
	α	0.100	0.016	0.151	0.099	0.96
	r^{**}	0.300	0.061	0.181	0.151	0.91

* The true value and bias for $\log r$ are -1.609 and -0.012

** The true value and bias for $\log r$ are -1.204 and 0.011

See NOTE to Table 2. The number of replicates in which one or more parameter estimates are at the boundary for Settings 1–4 are 27, 4, 3, and 0, respectively. Replicates with estimates at the boundary are discarded

Table 5 Simulation results under misspecified infection time distributions

Setting	Parameter	True	Bias	ESD	SEE	Cov
Weibull	β_1	0.500	0.008	0.041	0.042	0.95
	β_2	-0.500	-0.011	0.079	0.076	0.94
Lognormal	β_1	0.500	0.001	0.040	0.041	0.96
	β_2	-0.500	-0.005	0.077	0.075	0.94

See NOTE to Table 2. The number of replicates in which in which one or more parameter estimates are at the boundary for Setting Weibull and Setting Lognormal are 154 and 40, respectively. Replicates with estimates at the boundary are discarded

We compared the proposed methods with two simple methods, namely midpoint imputation and interval-censored data regression. For midpoint imputation, if $E_i > U_i$, then we set the time to infection to be $U_i/2$, such that the incubation period is (or is censored at) $E_i - U_i/2$. If $E_i \leq U_i$, then we set the time to infection to be $E_i/2$, such that the incubation period is $E_i/2$. We then apply the standard maximum partial-likelihood estimation method on the imputed dataset. For interval-censored data regression, we treat the incubation period as interval-censored within the smallest interval known to contain the incubation period. In particular, if $E_i \geq U_i$, then we set the incubation period to be interval-censored between $(E_i - U_i, E_i)$ for $\Delta_i = 1$ and

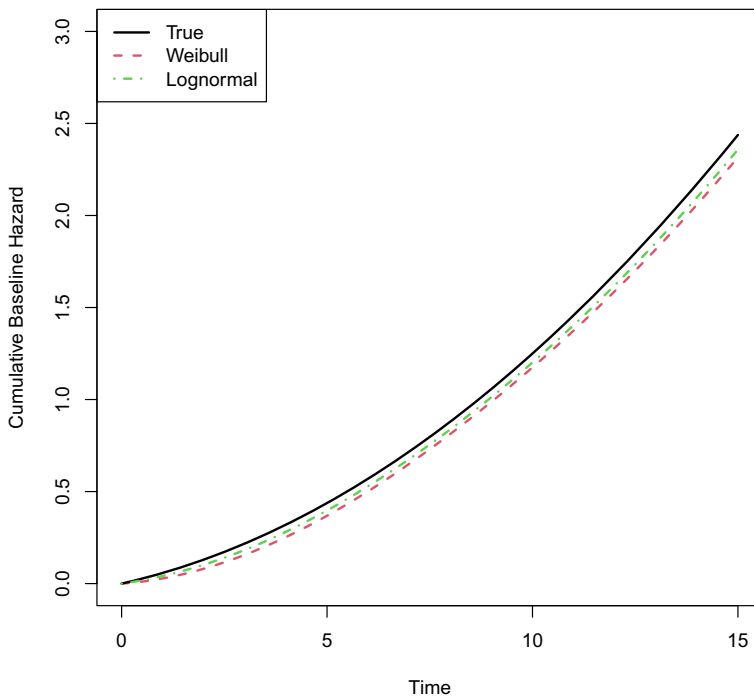


Fig. 2 Estimated cumulative baseline hazard functions under misspecified infection time distributions

right-censored at $E_i - U_i$ for $\Delta_i = 0$. If $E_i < U_i$, then we set the incubation period to be interval-censored between $(0, E_i)$; right censoring is assumed to be impossible under $E_i < U_i$. Then, we approximate the baseline hazard function using the same spline functions as in the proposed methods and estimate the parameters using (sieve) MLE. This approach is similar to the “interval-reduced” method of Reich et al. (2009), who considered a parametric model with no covariates. Note that this approach is based on the likelihood

$$\prod_{i=1}^n \left\{ e^{-\Lambda((E_i - U_i)_+) e^{\mathbf{X}_i^T \boldsymbol{\beta}}} - e^{-\Lambda(E_i) e^{\mathbf{X}_i^T \boldsymbol{\beta}}} \right\}^{\Delta_i} \left\{ e^{-\Lambda(E_i - U_i) e^{\mathbf{X}_i^T \boldsymbol{\beta}}} \right\}^{1 - \Delta_i},$$

where a_+ denotes $\max(a, 0)$. If $\Delta_i = 1$ for $i = 1, \dots, n$ and $\kappa(s; \alpha, r, \tau_i)$ does not depend on s , then this likelihood is proportional to the proposed conditional likelihood, but this likelihood is in general incorrect.

The summary results, based on $n = 800$ and 1000 replicates, for the estimators of β_1 and β_2 are presented in Table 6, and the averaged estimated cumulative baseline hazard functions are presented in Fig. 3. Because they do not consider the pandemic growth over time, both methods yield biased estimation of the cumulative baseline hazard function. Midpoint imputation yields biased point estimation and interval estimation with under-coverage. The bias under the interval-censored data regression is small,

Table 6 Simulation results under alternative simple methods

Setting			Midpoint Imputation				Interval-Censored Data Regression				
α	r	Parameter	True	Bias	ESD	SEE	Cov	Bias	ESD	SEE	Cov
0.01	0.2	β_1	0.500	-0.026	0.04	0.039	0.89	0.003	0.047	0.043	0.93
		β_2	-0.500	0.025	0.073	0.073	0.94	-0.004	0.085	0.078	0.93
0.01	0.3	β_1	0.500	-0.026	0.039	0.039	0.90	0.001	0.045	0.043	0.93
		β_2	-0.500	0.024	0.074	0.073	0.93	-0.003	0.082	0.078	0.94
0.1	0.2	β_1	0.500	-0.021	0.038	0.039	0.92	-0.004	0.045	0.042	0.93
		β_2	-0.500	0.016	0.073	0.073	0.94	0.002	0.083	0.077	0.93
0.1	0.3	β_1	0.500	-0.017	0.038	0.039	0.93	-0.002	0.044	0.041	0.92
		β_2	-0.500	0.014	0.074	0.073	0.95	0.003	0.082	0.076	0.94

See NOTE to Table 2

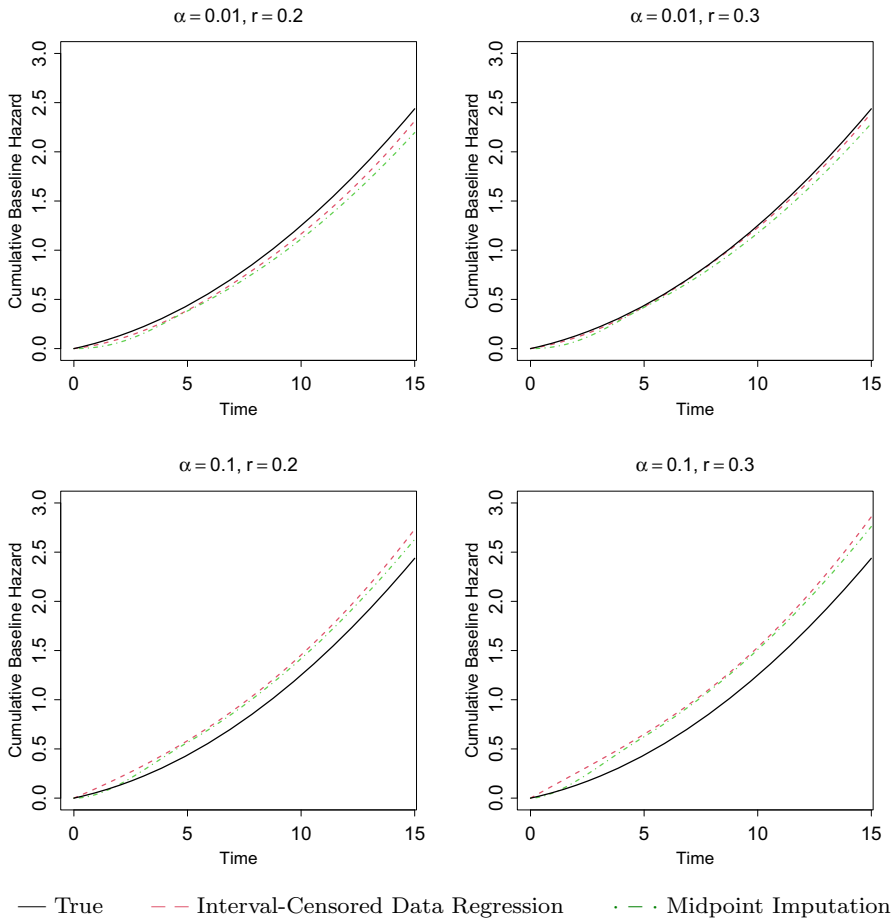


Fig. 3 Estimated cumulative baseline hazard functions using alternative simple methods

but the standard error estimates tend to be smaller than the corresponding standard deviation.

When the initial hazard of infection α is small, accurate estimation of the parameter is difficult. In fact, as $\alpha \rightarrow 0$, the density of S becomes proportional to e^{rS} , and the conditional density of S given $S < U$ does not involve α . To investigate the effect of inaccurate estimation of α when the true value of α is small, we consider an extra set of simulation studies as follows. We generated the data according to the original setting but with $\alpha = 10^{-10}$; the mean incubation times are given in Table 1. Then, we perform estimation with α (incorrectly) fixed to be 10^{-8} . This mimics the situation where the true value of α is beyond the lower boundary of 10^{-8} . We considered sample sizes of $n = 400$ and 800 . The point estimates, standard error estimates, and coverages of the Euclidean parameters based on 1000 replicates are presented in Table 7. In the calculation of the standard error estimates, we remove the row and

Table 7 Simulation results under $\alpha = 10^{-10}$

Setting	n	Parameter	True	Bias	ESD	SEE	Cov
1	400	β_1	0.500	0.005	0.060	0.060	0.96
		β_2	-0.500	-0.012	0.112	0.110	0.95
		r	0.200	0.007	0.093	0.091	0.91
2	400	β_1	0.500	0.006	0.060	0.060	0.95
		β_2	-0.500	-0.012	0.113	0.110	0.94
		r	0.300	0.007	0.098	0.096	0.92
3	800	β_1	0.500	0.003	0.042	0.042	0.95
		β_2	-0.500	-0.006	0.079	0.078	0.95
		r	0.200	0.000	0.067	0.066	0.94
4	800	β_1	0.500	0.004	0.043	0.042	0.96
		β_2	-0.500	-0.007	0.077	0.078	0.97
		r	0.300	0.000	0.067	0.070	0.95

See NOTE to Table 2. The number of replicates in which one or more parameter estimates are at the boundary for Settings 1–4 are 45, 57, 7, and 9, respectively. Replicates with estimates at the boundary are discarded

column corresponding to α in the information matrix, essentially treating α as known (at 10^{-8}).

In all cases, the biases of all parameters are small, the standard error estimates resemble the corresponding standard deviations, and the empirical coverages are close to the nominal 95% level. Similar to the results presented in Fig. 1, the estimated cumulative baseline hazard functions closely resembles the true function and are not presented. Due to misspecification of the value of α , some positive parameters are estimated at the boundary, and the number of replicates at the boundary is substantially smaller under a larger sample size. Note that in the calculation of the standard error estimates, we essentially treat α as known (at an incorrect value), but the standard error estimates and confidence intervals of the other parameters are still reliable. We suggest that in general, when parameters are estimated at the boundary, we can treat them as known and only evaluate the standard error estimates of the remaining parameters. In fact, in all simulation studies, if we kept the replicates with parameter estimates at the boundary and estimated the standard error of the remaining parameters in this way, the summary results would still be highly similar to the presented results.

5 Real data analysis

We analyzed the data set processed by Zhao et al. (2021) mentioned in Sect. 1. The data recorded confirmed COVID-19 cases from 14 locations, and the subjects were from or had visited Wuhan between 1 December 2019 and 23 January 2020; the data set was based on information from press releases of official health agencies. After removing subjects whose source of infection was not clear and those who miss information on the exposure/symptoms onset, 462 subjects with information on the period of stay at Wuhan, symptoms onset date, sex, and age are available. We set the initial calendar

time (from which the exposure start time is measured) to be 1 December 2019 and $\tau = 0$ for subjects who were from Wuhan. The symptoms onset dates of 6 subjects are right-censored. In the sample, 59.7% of the subjects are of age 50 or younger, and 53.9% are male.

We fit the proposed model with covariates age and sex, where age is dichotomized into 50 or younger versus older than 50. We considered B-spline functions with degrees 1, 2, and 3 and number of (inner) grid points of 1 to 5 and selected the model using AIC. The location of the inner grid points were chosen to be equi-distant quantiles of Gamma(1.8075, 0.3305); this is the estimated distribution of the incubation period in Zhao et al. (2021). The AIC selected degree 1 with 2 grid points. Same as the simulation studies, we set a lower boundary of 10^{-8} for the positive parameters.

The estimated values of α and γ_1 are at the boundary of 10^{-8} . The estimated coefficient for age is 0.2502, with standard error 0.1067, such that older subjects tend to have a shorter incubation period; this effect is significant at 5% significance level. This result is in disagreement with some findings in the literature (Dai et al. 2020; Tan et al. 2020). Nevertheless, as reported in Fig. 4 of Kong (2020), there is no consistent trend between the duration of the incubation period and age over different studies. The difference between the current results and those reported in the literature may be due to differences in patient samples. The estimated coefficient for sex is -0.0827 , with standard error 0.1042, such that females tend to have a shorter incubation period than males. The estimated value of r is 0.3044, which is close to the estimated values of 0.28–0.41 reported in Zhao et al. (2021). This corresponds to a daily increase of hazard of infection by a multiplicative factor of 1.36. Fig. 4 shows the estimated survival functions of different subgroups.

6 Discussion

In this paper, we consider the estimation of the incubation period distribution, where the time to infection is subject to interval censoring and truncation, and the time to symptom onset is subject to (potential) right censoring. We consider a flexible semiparametric regression model for the incubation period distribution and propose a sieve MLE. We establish the consistency and rate of convergence of all estimators and the asymptotic normality of the Euclidean parameter estimators.

One major challenge in model estimation is that information about the infection time is very limited. The infection time S is never observed and is mixed with the incubation period, and infection is only known to have occurred before U . In addition, only subjects with $S \leq U$ could be observed. Therefore, the likelihood contribution from U on the distribution of S is $P(S \leq U \mid S \leq U) = 1$, that is, U alone does not contain any information on the distribution of S . As a result, existing multiple imputation approaches, such as Sun et al. (1999, 2004), are not feasible, because these approaches require estimation of the infection time distribution solely using the intervals containing the infection times.

We propose to use grid search to select the initial values for the estimation. This step is essential, because the likelihood may exhibit multiple local maxima especially over different values of r . To illustrate how the local maxima may arise, we plot the

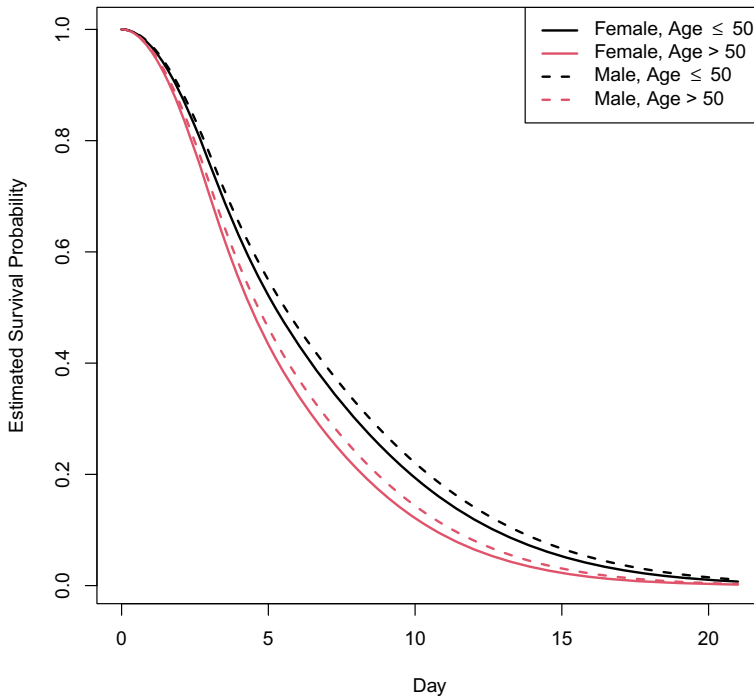


Fig. 4 Estimated survival functions for different subgroups

conditional cumulative distribution functions (CDF) of S given $S < U$ in Fig. 5, where $U = 10$, $\tau = 0$, $\alpha = 0.5$ or 0.001 , and $r = -0.5, -0.25, 0, 0.25$, or 0.5 . For $\alpha = 0.001$, the CDF at different values of r are well-separated, and the likelihood function usually does not exhibit multiple local maxima. By contrast, for $\alpha = 0.5$, the CDF at positive and negative values of r may be closer to each other than they are to the CDF at $r = 0$. In this case, when the true value of r is 0.25 , the likelihood may exhibit a local maximum at around 0.25 and at a negative value of r (or at zero if r is restricted to be nonnegative). As a result, when the initial value of r is very close to zero or negative, the gradient methods may not be able to update r towards the positive direction.

For satisfactory estimation of all model parameters, the dataset should contain both large and small values of U . A smaller value of U allows the infection time to be more precisely determined, in turn providing more information to the estimation of the incubation period distribution. In fact, in the regularity conditions, we assume that the support of U includes a small interval $[0, \zeta_0]$ in order to guarantee the identifiability of the incubation period distribution. On the other hand, a larger value of U provides more information to the estimation of the infection time distribution.

We assume a simple distribution for the time to infection, where the hazard increases exponentially at a constant rate over time, and the distribution is equal for all subjects. A possible extension is to consider more flexible models for the time to infection. For example, if a major pandemic-related event, such as a lock-down,

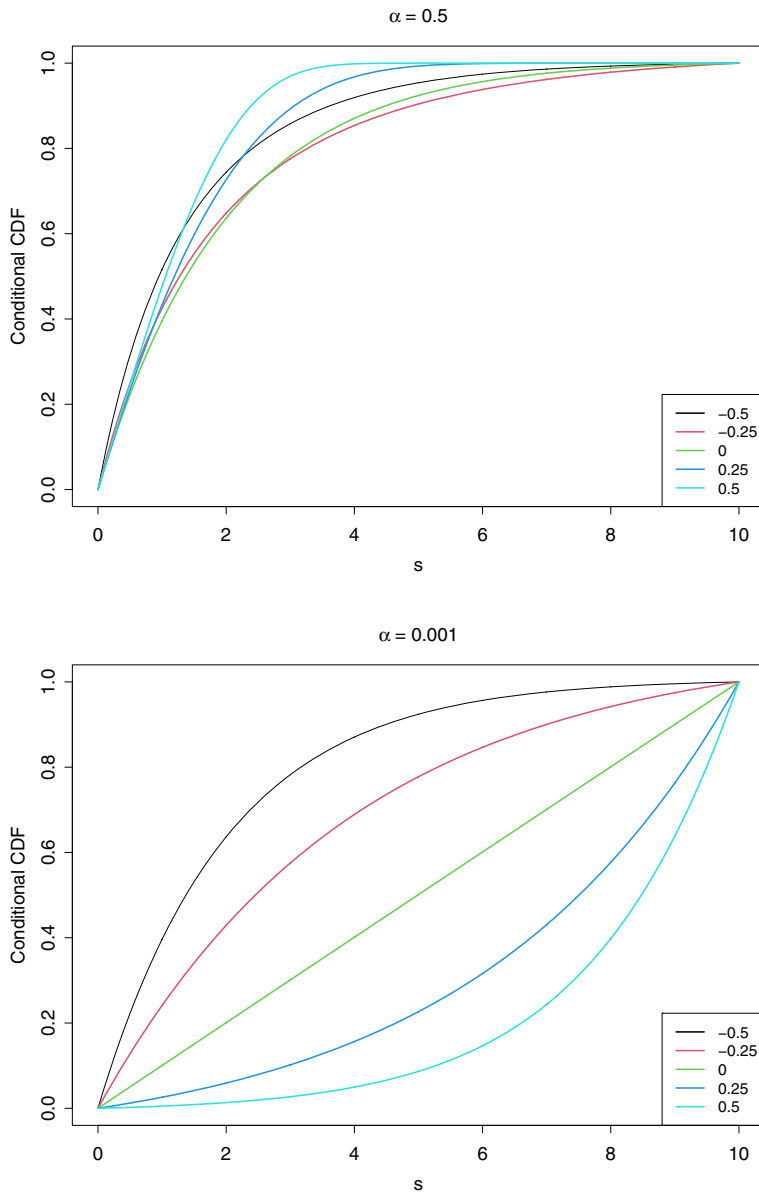


Fig. 5 Conditional CDF of S

occurs at time τ' , then we may allow the hazard growth rate to change after τ' , with $\kappa(t) = \alpha e^{r_1(t+\tau)+r_2(t+\tau-\tau')_+}$ for some parameters r_1 and r_2 , where $a_+ = \max(0, a)$. Also, we may consider a regression model for S that depends on X or some external covariates. In general, we can assume a nonparametric or semiparametric regression model for S and estimate the model by sieve estimation. However, because only

severely-limited information is available for the estimation of the distribution of S , such flexible models should be considered only when the sample size is large.

We assumed the Cox proportional hazards model for the incubation period. Although the Cox model is flexible in that the baseline hazard function is nonparametric, it assumes a particular structure for the covariates effect. It would be of interest to develop model checking techniques for the incubation period distribution. However, this would be highly challenging, as the incubation period is mixed with the infection time distribution in the likelihood, and existing methods are not applicable.

Acknowledgements The authors thank the Editor, Associate Editor and reviewers for their valuable comments and suggestions that have greatly improved the paper. Dr. Qingning Zhou's research was partially supported by the National Science Foundation grant DMS1916170.

Declarations

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

Appendix: Technical proofs

This appendix includes the proofs of Theorems 1 and 2. We assume that $\tau = 0$ for simplicity of presentation. Let $L(\theta)$ denote the likelihood function for a generic subject:

$$L(\theta) = \frac{e^{\Delta X^T \beta}}{1 - \exp \left\{ - \int_0^U \alpha e^{rt} dt \right\}} \int_0^{\min(E, U)} \lambda(E-s)^\Delta \exp \left\{ -\Lambda(E-s) e^{X^T \beta} \right\} \times \alpha e^{rs} \exp \left\{ - \int_0^s \alpha e^{rt} dt \right\} ds. \quad (S1)$$

Let \mathbb{P} and \mathbb{P}_n denote the true and empirical measures, respectively. Before proving Theorems 1 and 2, we present and prove the following lemma.

Lemma 1 *Assume that conditions (C1)–(C3) hold. If there exists a set of parameters $\tilde{\theta}$ such that $L(\tilde{\theta}) = L(\theta_0)$ with probability 1, then $\tilde{\theta} = \theta_0$.*

Proof of Lemma 1 Suppose that $L(\tilde{\theta}) = L(\theta_0)$ almost surely. Set $\Delta = 1$ and let $U \rightarrow 0$. By the L'Hospital's rule and taking derivatives with respect to U on both the numerator and denominator of L , we obtain

$$\lambda(E) e^{X^T \beta} \exp \left\{ -\Lambda(E) e^{X^T \beta} \right\} = \lambda_0(E) e^{X^T \beta_0} \exp \left\{ -\Lambda_0(E) e^{X^T \beta_0} \right\},$$

and further by taking integration, $\exp \{ -\Lambda(E) e^{X^T \beta} \} = \exp \{ -\Lambda_0(E) e^{X^T \beta_0} \}$. Thus, under conditions (C1)–(C3), we have $\beta = \beta_0$ and $\lambda(t) = \lambda_0(t)$ for $t \in [0, \zeta]$.

Next we show that $\alpha = \alpha_0$ and $r = r_0$. Setting $\Delta = 1$, we have

$$\begin{aligned} L(\theta) &= \frac{1}{F_S(U; \alpha, r)} \int_0^{\min(E, U)} f_T(E - s; \beta, \lambda) f_S(s; \alpha, r) ds \\ &= \frac{1}{F_S(U; \alpha, r)} \int_0^{\min(E, U)} f_T(t; \beta, \lambda) f_S(E - t; \alpha, r) dt, \end{aligned}$$

where $f_T(t; \beta, \lambda) = \lambda(t)e^{X^T \beta} \exp\{-\Lambda(t)e^{X^T \beta}\}$ is the density function of the incubation time $(\tilde{E} - S)$, $f_S(s; \alpha, r) = \alpha e^{rs} \exp\{-\int_0^s \alpha e^{rt} dt\}$ is the density function of the infection time S , and $F_S(t; \alpha, r) = \int_0^t f_S(s; \alpha, r) ds$ is the distribution function of S . Taking the first derivative of L with respect to E and setting $E \rightarrow 0$, we obtain

$$\frac{1}{F_S(U; \alpha, r)} f_T(0; \beta, \lambda) f_S(0; \alpha, r) = \frac{1}{F_S(U; \alpha_0, r_0)} f_T(0; \beta_0, \lambda_0) f_S(0; \alpha_0, r_0). \quad (\text{S2})$$

Taking the second derivative of L with respect to E and let $E \rightarrow 0$, we have

$$\begin{aligned} &\frac{1}{F_S(U; \alpha, r)} \left\{ f'_T(0; \beta, \lambda) f_S(0; \alpha, r) + f_T(0; \beta, \lambda) f'_S(0; \alpha, r) \right\} \\ &= \frac{1}{F_S(U; \alpha_0, r_0)} \left\{ f'_T(0; \beta_0, \lambda_0) f_S(0; \alpha_0, r_0) + f_T(0; \beta_0, \lambda_0) f'_S(0; \alpha_0, r_0) \right\}, \end{aligned} \quad (\text{S3})$$

where $f'_T(\cdot; \beta, \lambda)$ and $f'_S(\cdot; \alpha, r)$ are the derivatives of $f_T(\cdot; \beta, \lambda)$ and $f_S(\cdot; \alpha, r)$, respectively. Note that $\beta = \beta_0$, $\lambda = \lambda_0$, $f_S(0; \alpha, r) = \alpha$ and $f'_S(0; \alpha, r) = \alpha r$. From (S2) and (S3), it is easy to obtain that $\alpha = \alpha_0$ and $r = r_0$. \square

Proof of Theorem 1 Define $l(\theta) = \log L(\theta)$, $\mathcal{L}_n = \{l(\theta) : \theta \in \Theta_n\}$, and $\Theta_n = \mathcal{A} \times \mathcal{B}_n$, where \mathcal{A} and \mathcal{B}_n are defined in Sect. 2. Note that $M_n = O(\log \log n)$ and $\sum_{k=1}^{m_n} b_k(t) = 1$ for any t . By the mean-value theorem for integrals, for any $\theta \in \Theta_n$,

$$\frac{\min(E, U)/\{\log n (\log \log n)\}}{1 - \exp\{-\int_0^U \alpha e^{rt} dt\}} \lesssim L(\theta) \lesssim \frac{\min(E, U) \log \log n}{1 - \exp\{-\int_0^U \alpha e^{rt} dt\}},$$

where $A \lesssim B$ means that $A \leq cB$ for some positive constant c .

By the mean-value theorem, one can easily show that for any $\theta_1 \equiv (\beta_1, \alpha_1, r_1, \lambda_1)$ and $\theta_2 \equiv (\beta_2, \alpha_2, r_2, \lambda_2)$ in Θ_n ,

$$|l(\theta_1) - l(\theta_2)| \lesssim \log n (\log \log n)^3 (\|\beta_1 - \beta_2\| + |\alpha_1 - \alpha_2| + |r_1 - r_2| + \|\gamma_1 - \gamma_2\|_\infty),$$

where $\gamma_j = (\gamma_{j1}, \dots, \gamma_{j, m_n})^T$ such that $\lambda_j(t) = \sum_{k=1}^{m_n} \gamma_{jk} b_k(t)$ for $j = 1, 2$. Thus, by Lemma 2.5 of van de Geer (2000), the covering number of \mathcal{L}_n satisfies for any $\epsilon > 0$,

$$N\{\epsilon, \mathcal{L}_n, L_1(\mathbb{P}_n)\} \lesssim \{\log n (\log \log n)^3\}^{p+2+m_n} M_n^{m_n} \epsilon^{-(p+2+m_n)}.$$

In addition, note that $\mathbb{P}l(\boldsymbol{\theta})^2 \lesssim (\log \log n)^2$ for $\boldsymbol{\theta} \in \Theta_n$. Following the proof of Lemma 2 of Zhou et al. (2017), we can show that

$$\sup_{\boldsymbol{\theta} \in \Theta_n} |\mathbb{P}_n l(\boldsymbol{\theta}) - \mathbb{P}l(\boldsymbol{\theta})| \rightarrow 0 \quad (\text{S4})$$

almost surely.

By condition (C1) and Jackson's Theorem (de Boor 2001, p. 149), there exists $\lambda_n \in \mathcal{B}_n$ such that $\|\lambda_n - \lambda_0\|_\infty \equiv \sup_{t \in [0, \zeta]} |\lambda_n(t) - \lambda_0(t)| = O(n^{-q\nu})$. We have $d(\boldsymbol{\theta}_n, \boldsymbol{\theta}_0) = O(n^{-q\nu})$, where $\boldsymbol{\theta}_n = (\boldsymbol{\xi}_0, \lambda_n)$. By the definition of the sieve MLE, $\mathbb{P}_n l(\hat{\boldsymbol{\theta}}_n) \geq \mathbb{P}_n l(\boldsymbol{\theta}_n)$, so

$$\mathbb{P}l(\boldsymbol{\theta}_0) - \mathbb{P}l(\hat{\boldsymbol{\theta}}_n) \leq |\mathbb{P}l(\boldsymbol{\theta}_0) - \mathbb{P}l(\boldsymbol{\theta}_n)| + |(\mathbb{P}_n - \mathbb{P})l(\boldsymbol{\theta}_n)| + |(\mathbb{P}_n - \mathbb{P})l(\hat{\boldsymbol{\theta}}_n)|. \quad (\text{S5})$$

By (S4), the last two terms on the right-hand side of (S5) converge to zero almost surely. By the mean-value theorem, the first term on the right-hand side of (S5) converges to zero. The left-hand side of (S5) is non-negative by the property of Kullback–Leibler divergence, and thus it converges to zero almost surely.

Define $K_\epsilon = \{\boldsymbol{\theta} \in \Theta_n : d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \geq \epsilon\}$ for $\epsilon > 0$. If $\hat{\boldsymbol{\theta}}_n \in K_\epsilon$, then $\zeta_n \equiv \mathbb{P}l(\boldsymbol{\theta}_0) - \mathbb{P}l(\hat{\boldsymbol{\theta}}_n) \geq \mathbb{P}l(\boldsymbol{\theta}_0) - \sup_{\boldsymbol{\theta} \in K_\epsilon} \mathbb{P}l(\boldsymbol{\theta}) \equiv \delta_\epsilon$. That is, $\{\hat{\boldsymbol{\theta}}_n \in K_\epsilon\} \subseteq \{\zeta_n \geq \delta_\epsilon\}$. By Lemma 1 and the property of Kullback–Leibler divergence, we have $\delta_\epsilon > 0$ for any $\epsilon > 0$. By (S5), $\zeta_n \rightarrow 0$ almost surely. Thus, $\cap_{k=1}^\infty \cup_{n=k}^\infty \{\hat{\boldsymbol{\theta}}_n \in K_\epsilon\} \subseteq \cap_{k=1}^\infty \cup_{n=k}^\infty \{\zeta_n \geq \delta_\epsilon\}$ has probability zero asymptotically for any $\epsilon > 0$, which implies $d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) \rightarrow 0$ almost surely.

To prove the rate of convergence of $\hat{\boldsymbol{\theta}}_n$, we first use Theorem 3.4.1 of van der Vaart and Wellner (1996) to establish a suboptimal convergence rate. Then, based on this convergence rate, we show that the sieve MLE is uniformly consistent. Finally, based on this uniform consistency, we improve the convergence rate to the desired rate using Theorem 3.4.1 of van der Vaart and Wellner (1996) again. Let $\delta_n = K_1(\log \log n)^{1/2} n^{-q\nu}$ for some positive constant K_1 . For $\delta > \delta_n$, define $\mathcal{L}_{n,\delta} = \{l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_n) : \boldsymbol{\theta} \in \Theta_n, \delta/2 < d_n(\boldsymbol{\theta}, \boldsymbol{\theta}_n) < \delta\}$, where $d_n(\boldsymbol{\theta}, \boldsymbol{\theta}_n) = d(\boldsymbol{\theta}, \boldsymbol{\theta}_n)/(\log \log n)^{1/2}$. Let $H(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \equiv [\int \{L(\boldsymbol{\theta}_1)^{1/2} - L(\boldsymbol{\theta}_2)^{1/2}\}^2 d\mu]^{1/2}$ be the Hellinger distance between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, where μ is the dominating measure. Note that $\log(x) \leq 2(\sqrt{x} - 1)$ for $x \geq 0$, and thus

$$\mathbb{P}\{l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_0)\} \lesssim -H(\boldsymbol{\theta}, \boldsymbol{\theta}_0)^2 \lesssim -(\log \log n)^{-1} P\{L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}_0)\}^2. \quad (\text{S6})$$

By the mean-value theorem,

$$\mathbb{P}\{L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}_0)\}^2 \geq \mathbb{P}\{\dot{L}_\xi(\boldsymbol{\theta}_0)^T(\boldsymbol{\xi} - \boldsymbol{\xi}_0) + \dot{L}_\lambda(\boldsymbol{\theta}_0)[\lambda - \lambda_0]\}^2 + O\{d(\boldsymbol{\theta}, \boldsymbol{\theta}_0)^4\},$$

where $\dot{L}_\xi(\boldsymbol{\theta})$ is the derivative of $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\xi}$, and $\dot{L}_\lambda(\boldsymbol{\theta})[h]$ is the derivative of $L(\boldsymbol{\theta})$ with respect to λ along the submodel $\lambda_\epsilon = \lambda + \epsilon h$.

We define a norm space $\mathcal{G} = \mathbb{R}^{p+2} \times L_2[0, \zeta]$ with a norm $\|\mathbf{g}\|_1 = (\|\mathbf{g}_1\|^2 + \|\mathbf{g}_2\|_{L_2[0, \zeta]}^2)^{1/2}$, where $\mathbf{g} = (\mathbf{g}_1, \mathbf{g}_2) \in \mathcal{G}$ with $\mathbf{g}_1 = (\mathbf{g}_\beta^T, g_\alpha, g_r)^T \in \mathbb{R}^{p+2}$ and

$g_2 \in L_2[0, \zeta]$. In addition, define a seminorm $\|\cdot\|_2$ on \mathcal{G} , with $\|\mathbf{g}\|_2^2 = \mathbb{P}\{\dot{L}_\xi(\boldsymbol{\theta}_0)^T \mathbf{g}_1 + \dot{L}_\lambda(\boldsymbol{\theta}_0)[g_2]\}^2$. By the Minkowski inequality, this seminorm is well-defined. If we can show that $\|\mathbf{g}\|_2 = 0$ implies $\mathbf{g} = \mathbf{0}$, then $\|\cdot\|_2$ is a norm. In fact, if $\|\mathbf{g}\|_2 = 0$ for some $\mathbf{g} = (\mathbf{g}_1, g_2) = (\mathbf{g}_\beta, g_\alpha, g_r, g_2) \in \mathcal{G}$, then $\dot{L}_\xi(\boldsymbol{\theta}_0)^T \mathbf{g}_1 + \dot{L}_\lambda(\boldsymbol{\theta}_0)[g_2] = 0$ almost surely. As in the proof of Lemma 1, we let $U \rightarrow 0$ and obtain

$$\frac{\partial \tilde{L}}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}, \lambda_0)^T \mathbf{g}_\beta \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} + \frac{\partial \tilde{L}}{\partial \epsilon}(\boldsymbol{\beta}_0, \lambda + \epsilon g_2) \Big|_{\epsilon=0} = 0 \quad (\text{S7})$$

almost surely, where $\tilde{L}(\boldsymbol{\beta}, \lambda) = \lambda(E)^{\Delta} e^{\mathbf{X}^T \boldsymbol{\beta} \Delta} \exp\{-\Lambda(E)e^{\mathbf{X}^T \boldsymbol{\beta}}\}$. Letting $\Delta = 1$ in (S7) and taking integration of \tilde{L} with respect to E , we can derive that $\int_0^E \{\lambda_0(u) \mathbf{X}^T \mathbf{g}_\beta + g_2(u)\} du = 0$. Therefore, $\int_0^t \{\lambda_0(u) \mathbf{X}^T \mathbf{g}_\beta + g_2(u)\} du = 0$ for any $t \in [0, \zeta]$, which implies $\mathbf{X}^T \mathbf{g}_\beta + g_2(t)/\lambda_0(t) = 0$. By condition (C2), we have $\mathbf{g}_\beta = \mathbf{0}$ and $g_2(t) = 0$ for $t \in [0, \zeta]$. Similarly as in the proof of Lemma 1, we can show that $g_\alpha = g_r = 0$. Thus, $\mathbf{g} = \mathbf{0}$, and $\|\cdot\|_2$ is indeed a norm on \mathcal{G} .

It is easy to see that $\|\mathbf{g}\|_2 \leq K_2 \|\mathbf{g}\|_1$ for some positive constant K_2 . By the bounded inverse theorem in the Banach space, we have $\|\mathbf{g}\|_2 \geq K_3 \|\mathbf{g}\|_1$ for some $K_3 > 0$. Hence, by (S6), $\mathbb{P}\{l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_0)\} \lesssim -(\log \log n)^{-1} [d(\boldsymbol{\theta}, \boldsymbol{\theta}_0)^2 + O\{d(\boldsymbol{\theta}, \boldsymbol{\theta}_0)^4\}] \lesssim -d_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0)^2$ for small enough $d_n(\boldsymbol{\theta}, \boldsymbol{\theta}_n)$. For any $l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_n) \in \mathcal{L}_{n,\delta}$, by the mean-value theorem, we have

$$\begin{aligned} \mathbb{P}\{l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_n)\} &= \mathbb{P}\{l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_0)\} + \mathbb{P}\{l(\boldsymbol{\theta}_0) - l(\boldsymbol{\theta}_n)\} \\ &\leq -K_4 d_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0)^2 + K_5 d(\boldsymbol{\theta}_n, \boldsymbol{\theta}_0)^2 \\ &\leq -K_4 d_n(\boldsymbol{\theta}, \boldsymbol{\theta}_n)^2 + K_4 d_n(\boldsymbol{\theta}_n, \boldsymbol{\theta}_0)^2 + K_5 n^{-2qv} \\ &\leq -K_4 \delta^2/4 + K_4 K_5 (\log \log n)^{-1} n^{-2qv} + K_5 n^{-2qv}, \end{aligned}$$

where K_4 and K_5 are positive constants. Therefore, for large enough K_1 , we have

$$\sup_{f \in \mathcal{L}_{n,\delta}} \mathbb{P} f \leq -K_6 \delta^2 \quad (\text{S8})$$

for all $\delta > \delta_n$ and some positive constant K_6 .

By the mean-value theorem, $|l(\boldsymbol{\theta}_1) - l(\boldsymbol{\theta}_2)| \leq K_7 \log n (\log \log n)^3 (\|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|_\infty + \|\lambda_1 - \lambda_2\|_\infty)$ for some $K_7 > 0$. Let $\Psi_n = \{\boldsymbol{\theta} \in \Theta_n : d(\boldsymbol{\theta}, \boldsymbol{\theta}_n) < \delta\}$. We have

$$\log N_{[]} \{\epsilon, \mathcal{L}_{n,\delta}, L_2(\mathbb{P})\} \leq \log N \{K_7^{-1} (\log n)^{-1} (\log \log n)^{-7/2} \epsilon/2, \Psi_n, \|\cdot\|_\infty\}.$$

Let $\Phi_n = \{\boldsymbol{\gamma} \in \mathbb{R}^{m_n} : \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_n\| < m_n^{1/2} \delta\}$, where $\boldsymbol{\gamma}_n$ is the coefficient vector for λ_n . By Theorem 5.2 of de Boer (1976), we have $N(\epsilon, \Psi_n, \|\cdot\|_\infty) \leq N(K_8^{-1} \epsilon, \Phi_n, \|\cdot\|_\infty)$ for some $K_8 > 0$. Furthermore, by Lemma 0.4 of Wu and Zhang (2012), we obtain $\log N(\epsilon, \Phi_n, \|\cdot\|_\infty) \lesssim m_n \log(\delta/\epsilon)$. Combining the above results, we have for $0 < \epsilon < \delta$,

$$\log N_{[]} \{\epsilon, \mathcal{L}_{n,\delta}, L_2(\mathbb{P})\} \leq K_9 m_n \log \{ \log n (\log \log n)^{7/2} \delta / \epsilon \}$$

for some $K_9 > 0$. The bracketing integral of $\mathcal{L}_{n,\delta}$ satisfies

$$\begin{aligned}\tilde{J}_{[]} \{\delta, \mathcal{L}_{n,\delta}, L_2(\mathbb{P})\} &= \int_0^\delta \sqrt{1 + \log N_{[]} \{\epsilon, \mathcal{L}_{n,\delta}, L_2(\mathbb{P})\}} d\epsilon \\ &\leq \int_0^\delta \sqrt{1 + K_9 m_n \log \{ \log n (\log \log n)^{7/2} \delta / \epsilon \}} d\epsilon \\ &\lesssim \int_0^\delta m_n^{1/2} \{ \log n (\log \log n)^{7/2} \delta / \epsilon \}^{1/2} d\epsilon \\ &= m_n^{1/2} (\log n)^{1/2} (\log \log n)^{7/4} \delta.\end{aligned}$$

As shown in the proof of consistency, $\sup_{f \in \mathcal{L}_{n,\delta}} \|f\|_\infty = O(\log \log n)$, and for $l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_n) \in \mathcal{L}_{n,\delta}$, $\mathbb{P}\{l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_n)\}^2 \leq K_{10}(\log n)^2 (\log \log n)^6 d(\boldsymbol{\theta}, \boldsymbol{\theta}_n)^2 \leq K_{10}(\log n)^2 (\log \log n)^7 \delta^2$ for some $K_{10} > 0$. Then, by Lemma 3.4.2 of van der Vaart and Wellner (1996), we have

$$\begin{aligned}\mathbb{E} \|\sqrt{n}(\mathbb{P}_n - \mathbb{P})\|_{\mathcal{L}_{n,\delta}} &\lesssim \tilde{J}_{[]} \{K_{10} \log n (\log \log n)^{7/2} \delta, \mathcal{L}_{n,\delta}, L_2(\mathbb{P})\} \\ &\quad \times \left[1 + \frac{\tilde{J}_{[]} \{K_{10} \log n (\log \log n)^{7/2} \delta, \mathcal{L}_{n,\delta}, L_2(\mathbb{P})\}}{(\log n)^2 (\log \log n)^7 \delta^2 \sqrt{n}} \log \log n \right] \\ &\lesssim \phi_n(\delta),\end{aligned}\tag{S9}$$

where $\phi_n(\delta) = m_n^{1/2} (\log n)^{3/2} (\log \log n)^{21/4} \delta + m_n n^{-1/2} \log n (\log \log n)^{9/2}$. One can easily see that $\phi_n(\delta)/\delta$ is decreasing in δ .

Let $r_n = n^{\min\{qv, (1-v)/2\}} / \{(\log n)^{3/2} (\log \log n)^{21/4}\}$. Then, we have $r_n \lesssim \delta_n^{-1}$ and $r_n^2 \phi_n(1/r_n) \lesssim n^{1/2}$. Because $\mathbb{P}_n l(\hat{\boldsymbol{\theta}}_n) \geq \mathbb{P}_n l(\boldsymbol{\theta}_n)$, $d_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_n) \leq d_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) + d_n(\boldsymbol{\theta}_n, \boldsymbol{\theta}_0) \rightarrow 0$ in probability, and that (S8) and (S9) hold, we conclude that $r_n d_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_n) = O_p(1)$ by Theorem 3.4.1 of van der Vaart and Wellner (1996). Thus, $r_n d_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) \leq r_n d_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_n) + r_n d_n(\boldsymbol{\theta}_n, \boldsymbol{\theta}_0) \leq O_p(1) + o(1) = O_p(1)$.

To show that $\hat{\boldsymbol{\theta}}_n$ is consistent under the supremum norm, note that

$$\|\hat{\lambda}_n - \lambda_n\|_\infty = \left\| \sum_{k=1}^{m_n} (\hat{\gamma}_k - \gamma_{nk}) b_k(t) \right\|_\infty \leq \sum_{k=1}^{m_n} |\hat{\gamma}_k - \gamma_{nk}| \|b_k\|_\infty \leq m_n^{1/2} \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_n\|,$$

where $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \dots, \hat{\gamma}_{m_n})^T$ and $\boldsymbol{\gamma}_n = (\gamma_{n1}, \dots, \gamma_{n,m_n})^T$. By Theorem 5.2 of de Boor (1976), $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_n\| \lesssim m_n^{1/2} \|\hat{\lambda}_n - \lambda_n\|_{L_2[0,\xi]} = O_p\{n^{v/2} (\log \log n)^{1/2} r_n^{-1}\}$. Therefore, $\|\hat{\lambda}_n - \lambda_0\|_\infty \leq \|\hat{\lambda}_n - \lambda_n\|_\infty + \|\lambda_n - \lambda_0\|_\infty = O_p\{n^v (\log \log n)^{1/2} r_n^{-1}\} + O(n^{-qv}) = o_p(1)$, since $v < 1/3$ and $q \geq 2$.

Let $\hat{\boldsymbol{\theta}}_n^* = \arg \max_{\boldsymbol{\theta} \in \Theta_n^*} L_n(\boldsymbol{\theta})$, where $\Theta_n^* = \mathcal{A} \times \mathcal{B}_n^*$, and

$$\mathcal{B}_n^* = \left\{ \sum_{j=1}^{m_n} \gamma_j b_j : M^{-1} \leq \gamma_j \leq M \text{ for } j = 1, \dots, m_n \right\}$$

for some large enough positive constant M . Following similar arguments as the above, we can show that $\rho_n d(\widehat{\boldsymbol{\theta}}_n^*, \boldsymbol{\theta}_0) = O_p(1)$ with $\rho_n = n^{\min\{qv, (1-v)/2\}}$. Because $\|\widehat{\lambda}_n - \lambda_0\|_\infty = o_p(1)$, $\widehat{\boldsymbol{\theta}}_n$ and $\widehat{\boldsymbol{\theta}}_n^*$ are asymptotically equivalent, that is, $Pr(\widehat{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_n^*) \rightarrow 1$ as $n \rightarrow \infty$. We conclude that $\rho_n d(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = O_p(1)$, which completes the proof. \square

Proof of Theorem 2 Let $f_S(t; \alpha, r) = \alpha e^{rt} \exp\{-\int_0^t \alpha e^{rs} ds\}$ denote the density function of the infection time S , and $F_S(t; \alpha, r) = \int_0^t f_S(s; \alpha, r) ds = 1 - \exp\{-\int_0^t \alpha e^{rs} dr\}$ denote the distribution function of S . The score function for $\boldsymbol{\xi}$ is

$$\boldsymbol{l}_\xi(\boldsymbol{\xi}, \lambda) = \begin{pmatrix} \boldsymbol{l}_\beta(\boldsymbol{\xi}, \lambda) \\ l_\alpha(\boldsymbol{\xi}, \lambda) \\ l_r(\boldsymbol{\xi}, \lambda) \end{pmatrix}$$

with

$$\begin{aligned} \boldsymbol{l}_\beta(\boldsymbol{\xi}, \lambda) &= \frac{1}{L(\boldsymbol{\theta})} \left[\frac{1}{F_S(U; \alpha, r)} \int_0^{\min(E, U)} \exp\{\boldsymbol{X}^T \boldsymbol{\beta} - \Lambda(E-s)e^{X^T \boldsymbol{\beta}}\} \right. \\ &\quad \left. [\Delta \lambda(E-s) - \lambda(E-s)^\Delta e^{X^T \boldsymbol{\beta} \Delta} \Lambda(E-s)] f_S(s; \alpha, r) ds \right] \boldsymbol{X} \\ l_\alpha(\boldsymbol{\xi}, \lambda) &= \frac{1}{L(\boldsymbol{\theta})} \left[\int_0^{\min(E, U)} \lambda(E-s)^\Delta e^{X^T \boldsymbol{\beta} \Delta} \exp\{-\Lambda(E-s)e^{X^T \boldsymbol{\beta}}\} \frac{\partial}{\partial \alpha} \left\{ \frac{f_S(s; \alpha, r)}{F_S(U; \alpha, r)} \right\} ds \right] \\ l_r(\boldsymbol{\xi}, \lambda) &= \frac{1}{L(\boldsymbol{\theta})} \left[\int_0^{\min(E, U)} \lambda(E-s)^\Delta e^{X^T \boldsymbol{\beta} \Delta} \exp\{-\Lambda(E-s)e^{X^T \boldsymbol{\beta}}\} \frac{\partial}{\partial r} \left\{ \frac{f_S(s; \alpha, r)}{F_S(U; \alpha, r)} \right\} ds \right]. \end{aligned}$$

In addition, for $h \in L_2[0, \zeta]$, the score function for λ along the submodel $\lambda_{\epsilon, h} = \lambda + \epsilon h$ is

$$\begin{aligned} l_\lambda(\boldsymbol{\xi}, \lambda)[h] &= \frac{1}{L(\boldsymbol{\theta})} \left[\frac{1}{F_S(U; \alpha, r)} \int_0^{\min(E, U)} \exp\{\boldsymbol{X}^T \boldsymbol{\beta} - \Lambda(E-s)e^{X^T \boldsymbol{\beta}}\} \right. \\ &\quad \left. \left\{ \Delta h(E-s) - \lambda(E-s)^\Delta e^{X^T \boldsymbol{\beta} \Delta} \int_0^{E-s} h(u) du \right\} f_S(s; \alpha, r) ds \right]. \end{aligned}$$

For a vector of functions $\boldsymbol{h} = (h_1, \dots, h_{p+2})^T$, $\boldsymbol{l}_\lambda(\boldsymbol{\xi}, \lambda)[\boldsymbol{h}]$ denotes $(l_\lambda(\boldsymbol{\xi}, \lambda)[h_1], \dots, l_\lambda(\boldsymbol{\xi}, \lambda)[h_{p+2}])^T$. Let \boldsymbol{h}^* be the least-favorable direction, a $(p+2)$ -dimensional vector with components in $L_2[0, \zeta]$, such that

$$\boldsymbol{l}_\lambda^* \boldsymbol{l}_\lambda[\boldsymbol{h}^*] = \boldsymbol{l}_\lambda^* \boldsymbol{l}_\xi, \quad (\text{S10})$$

where \boldsymbol{l}_ξ and $\boldsymbol{l}_\lambda[\boldsymbol{h}^*]$ are evaluated at $(\boldsymbol{\xi}_0, \lambda_0)$, and \boldsymbol{l}_λ^* is the adjoint operator of \boldsymbol{l}_λ . We will show the existence of \boldsymbol{h}^* later. From (S10), we can see that each component of \boldsymbol{h}^* is continuously differentiable under conditions (C1) and (C3). Therefore, there exists $\boldsymbol{h}_n^* \in \mathcal{B}_n^{p+2}$ such that $\|\boldsymbol{h}_n^* - \boldsymbol{h}^*\|_\infty = O(n^{-\nu})$. Since $\mathbb{P}_n \boldsymbol{l}_\lambda(\widehat{\boldsymbol{\xi}}_n, \widehat{\lambda}_n)[\boldsymbol{h}_n^*] = \mathbf{0}$ and $\mathbb{P}_n \boldsymbol{l}_\lambda(\boldsymbol{\xi}_0, \lambda_0)[\boldsymbol{h}^* - \boldsymbol{h}_n^*] = \mathbf{0}$, we have

$$\begin{aligned}
& \mathbb{P}_n \mathbf{I}_\lambda(\widehat{\boldsymbol{\xi}}_n, \widehat{\lambda}_n)[\mathbf{h}^*] \\
&= \mathbb{P}_n \mathbf{I}_\lambda(\widehat{\boldsymbol{\xi}}_n, \widehat{\lambda}_n)[\mathbf{h}^*] - \mathbb{P}_n \mathbf{I}_\lambda(\widehat{\boldsymbol{\xi}}_n, \widehat{\lambda}_n)[\mathbf{h}_n^*] \\
&= (\mathbb{P}_n - \mathbb{P}) \mathbf{I}_\lambda(\widehat{\boldsymbol{\xi}}_n, \widehat{\lambda}_n)[\mathbf{h}^* - \mathbf{h}_n^*] + \mathbb{P} \{ \mathbf{I}_\lambda(\widehat{\boldsymbol{\xi}}_n, \widehat{\lambda}_n)[\mathbf{h}^* - \mathbf{h}_n^*] - \mathbf{I}_\lambda(\boldsymbol{\xi}_0, \lambda_0)[\mathbf{h}^* - \mathbf{h}_n^*] \}.
\end{aligned}$$

By the mean-value theorem and that $\nu > 1/\{2(q+1)\}$, the second term on the right-hand side above is $O_p[n^{-\nu} n^{-\min\{q\nu, (1-\nu)/2\}}] = o_p(n^{-1/2})$. We will show later that the first term is $o_p(n^{-1/2})$. Thus, $\mathbb{P}_n \mathbf{I}_\lambda(\widehat{\boldsymbol{\xi}}_n, \widehat{\lambda}_n)[\mathbf{h}^*] = o_p(n^{-1/2})$. Since $\mathbb{P}_n \mathbf{I}_\xi(\widehat{\boldsymbol{\xi}}_n, \widehat{\lambda}_n) = \mathbb{P} \mathbf{I}_\xi(\boldsymbol{\xi}_0, \lambda_0) = \mathbb{P} \mathbf{I}_\lambda(\boldsymbol{\xi}_0, \lambda_0)[\mathbf{h}^*] = \mathbf{0}$, we have

$$\begin{aligned}
(\mathbb{P}_n - \mathbb{P}) \mathbf{I}_\xi(\widehat{\boldsymbol{\xi}}_n, \widehat{\lambda}_n) &= -\{ \mathbb{P} \mathbf{I}_\xi(\widehat{\boldsymbol{\xi}}_n, \widehat{\lambda}_n) - \mathbb{P} \mathbf{I}_\xi(\boldsymbol{\xi}_0, \lambda_0) \} \\
(\mathbb{P}_n - \mathbb{P}) \mathbf{I}_\lambda(\widehat{\boldsymbol{\xi}}_n, \widehat{\lambda}_n)[\mathbf{h}^*] &= -\{ \mathbb{P} \mathbf{I}_\lambda(\widehat{\boldsymbol{\xi}}_n, \widehat{\lambda}_n)[\mathbf{h}^*] - \mathbb{P} \mathbf{I}_\lambda(\boldsymbol{\xi}_0, \lambda_0)[\mathbf{h}^*] \} + o_p(n^{-1/2}).
\end{aligned}$$

Further, note that $d(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0)^2 = o_p(n^{-1/2})$ if $1/(4q) < \nu < 1/2$. We apply the Taylor series expansion about $(\boldsymbol{\xi}_0, \lambda_0)$ to the right-hand sides of the above two equations and obtain

$$(\mathbb{P}_n - \mathbb{P}) \mathbf{I}_\xi(\widehat{\boldsymbol{\xi}}_n, \widehat{\lambda}_n) = -\mathbb{P} \mathbf{I}_{\xi\xi}(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_0) - \mathbb{P} \mathbf{I}_{\xi\lambda}(\widehat{\lambda}_n - \lambda_0) + o_p(n^{-1/2}) \quad (\text{S11})$$

$$(\mathbb{P}_n - \mathbb{P}) \mathbf{I}_\lambda(\widehat{\boldsymbol{\xi}}_n, \widehat{\lambda}_n)[\mathbf{h}^*] = -\mathbb{P} \mathbf{I}_{\lambda\xi}(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_0)[\mathbf{h}^*] - \mathbb{P} \mathbf{I}_{\lambda\lambda}(\widehat{\lambda}_n - \lambda_0)[\mathbf{h}^*] + o_p(n^{-1/2}), \quad (\text{S12})$$

where $\mathbf{I}_{\xi\xi}$ is the derivative of \mathbf{I}_ξ with respect to $\boldsymbol{\xi}$, $\mathbf{I}_{\xi\lambda}[h]$ is the derivative of \mathbf{I}_ξ along the submodel $\lambda_{\epsilon, h} = \lambda + \epsilon h$, $\mathbf{I}_{\lambda\xi}$ is the derivative of $\mathbf{I}_\lambda[h]$ with respect to $\boldsymbol{\xi}$, and $\mathbf{I}_{\lambda\lambda}[\mathbf{h}_1, \mathbf{h}_2]$ is the derivative of $\mathbf{I}_\lambda[\mathbf{h}_1]$ along the submodel $\lambda_{\epsilon, h_2} = \lambda + \epsilon h_2$. All of these derivatives are evaluated at $(\boldsymbol{\xi}_0, \lambda_0)$. From (S10), we have $\mathbb{P} \mathbf{I}_{\lambda\lambda}[\mathbf{h}^*, \widehat{\lambda}_n - \lambda_0] = \mathbb{P} \mathbf{I}_{\xi\lambda}[\widehat{\lambda}_n - \lambda_0]$.

Thus, taking difference of (S11) and (S12) yields

$$\begin{aligned}
& \sqrt{n}(\mathbb{P}_n - \mathbb{P}) \{ \mathbf{I}_\xi(\widehat{\boldsymbol{\xi}}_n, \widehat{\lambda}_n) - \mathbf{I}_\lambda(\widehat{\boldsymbol{\xi}}_n, \widehat{\lambda}_n)[\mathbf{h}^*] \} \\
&= \mathbb{P} \{ (\mathbf{I}_\xi - \mathbf{I}_\lambda[\mathbf{h}^*]) (\mathbf{I}_\xi - \mathbf{I}_\lambda[\mathbf{h}^*])^T \} \sqrt{n}(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_0) + o_p(1),
\end{aligned}$$

where $\mathbf{I}_\xi - \mathbf{I}_\lambda[\mathbf{h}^*]$ is evaluated at $(\boldsymbol{\xi}_0, \lambda_0)$. We will show that the matrix $\mathbf{I}(\boldsymbol{\xi}_0) \equiv \mathbb{P} \{ (\mathbf{I}_\xi - \mathbf{I}_\lambda[\mathbf{h}^*]) (\mathbf{I}_\xi - \mathbf{I}_\lambda[\mathbf{h}^*])^T \}$ is invertible, and $\mathbf{I}_\xi(\widehat{\boldsymbol{\xi}}_n, \widehat{\lambda}_n) - \mathbf{I}_\lambda(\widehat{\boldsymbol{\xi}}_n, \widehat{\lambda}_n)[\mathbf{h}^*]$ belongs to a Donsker class and converges in the $L_2(\mathbb{P})$ -norm to $\mathbf{I}_\xi - \mathbf{I}_\lambda[\mathbf{h}^*]$. It then follows that $\sqrt{n} \|\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_0\| = O_p(1)$ and

$$\sqrt{n}(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_0) = \mathbf{I}(\boldsymbol{\xi}_0)^{-1} \sqrt{n}(\mathbb{P}_n - \mathbb{P})(\mathbf{I}_\xi - \mathbf{I}_\lambda[\mathbf{h}^*]) + o_p(1),$$

which yields the desired result by the central limit theorem.

It remains to show that (i) the least favorable direction \mathbf{h}^* satisfying (S10) exists; (ii) the matrix $\mathbf{I}(\boldsymbol{\xi}_0) \equiv \mathbb{P} \{ (\mathbf{I}_\xi - \mathbf{I}_\lambda[\mathbf{h}^*]) (\mathbf{I}_\xi - \mathbf{I}_\lambda[\mathbf{h}^*])^T \}$ is invertible; (iii) $\mathbf{I}_\xi(\widehat{\boldsymbol{\xi}}_n, \widehat{\lambda}_n) - \mathbf{I}_\lambda(\widehat{\boldsymbol{\xi}}_n, \widehat{\lambda}_n)[\mathbf{h}^*]$ belongs to a Donsker class and converges in the $L_2(\mathbb{P})$ -norm to $\mathbf{I}_\xi - \mathbf{I}_\lambda[\mathbf{h}^*]$; and (iv) $(\mathbb{P}_n - \mathbb{P}) \mathbf{I}_\lambda(\widehat{\boldsymbol{\xi}}_n, \widehat{\lambda}_n)[\mathbf{h}^* - \mathbf{h}_n^*] = o_p(n^{-1/2})$, where $\mathbf{h}_n^* \in \mathcal{B}_n^{p+2}$ such that $\|\mathbf{h}_n^* - \mathbf{h}^*\|_\infty = O(n^{-\nu})$.

To show (i), we consider the linear operator $\Gamma = l_\lambda^* l_\lambda$ from $L_2[0, \zeta]$ to itself. We equip $L_2[0, \zeta]$ with an inner product $\langle h_1, h_2 \rangle = \int_0^\zeta h_1(t)h_2(t) dt$. In addition, we define a seminorm on $L_2[0, \zeta]$ as $\|h\|_\Gamma = \langle \Gamma(h), h \rangle^{1/2}$. We first show that $\|\cdot\|_\Gamma$ is in fact a norm. If $\|h\|_\Gamma = 0$, then $\mathbb{P}\{l_\lambda[h]^2\} = \langle \Gamma(h), h \rangle = 0$. It follows that $l_\lambda[h] = 0$ almost surely, which yields

$$\frac{1}{F_S(U; \alpha_0, r_0)} \int_0^{\min(E, U)} \exp\{X^T \beta_0 - \Lambda_0(E-s)e^{X^T \beta_0}\} \\ \times \left\{ \Delta h(E-s) - \lambda_0(E-s)^\Delta e^{X^T \beta_0 \Delta} \int_0^{E-s} h(u) du \right\} f_S(s; \alpha_0, r_0) ds = 0.$$

Setting $U \rightarrow 0$, we have

$$\exp\{X^T \beta_0 - \Lambda_0(E)e^{X^T \beta_0}\} \left\{ \Delta h(E) - \lambda_0(E)^\Delta e^{X^T \beta_0 \Delta} \int_0^E h(u) du \right\} = 0.$$

Setting $\Delta = 1$ and taking integration with respect to E , we obtain $\int_0^E h(u) du = 0$ almost surely, which implies $h(t) = 0$ for $t \in [0, \zeta]$. Therefore, $\|\cdot\|_\Gamma$ is a norm. It is easy to see that $\|h\|_\Gamma \leq K_1 \langle h, h \rangle^{1/2}$ for some $K_1 > 0$. By the bounded inverse theorem in the Banach space, we have $\langle h, h \rangle^{1/2} \leq K_2 \|h\|_\Gamma$ for some $K_2 > 0$, that is, $\langle \Gamma(h), h \rangle \geq K_2 \langle h, h \rangle$. By the Lax-Milgram theorem, the solution \mathbf{h}^* to (S10) exists.

Now we prove (ii). If the matrix $\mathbf{I}(\xi_0)$ is singular, then there exists a nonzero vector $\mathbf{w} = (\mathbf{w}_\beta^T, w_\alpha, w_r)^T$ such that $\mathbf{w}^T \mathbb{P}\{(\mathbf{I}_\xi - \mathbf{I}_\lambda[\mathbf{h}^*])(\mathbf{I}_\xi - \mathbf{I}_\lambda[\mathbf{h}^*])^T\} \mathbf{w} = 0$, which implies that the score function along the submodel $\xi_{\epsilon, \mathbf{w}} = \xi_0 + \epsilon \mathbf{w}$ and $\lambda_{\epsilon, \mathbf{w}^T \mathbf{h}^*} = \lambda_0 + \epsilon \mathbf{w}^T \mathbf{h}^*$ is zero almost surely. Decompose $\mathbf{h}^* = (\mathbf{h}_\beta^{*T}, h_\alpha^*, h_r^*)^T$. Setting $U \rightarrow 0$, we have

$$\exp\{X^T \beta_0 - \Lambda_0(E)e^{X^T \beta_0}\} \left\{ \Delta \tilde{h}(E) - \lambda_0(E)^\Delta e^{X^T \beta_0 \Delta} \int_0^E \tilde{h}(u) du \right\} = 0,$$

where $\tilde{h}(t) = \mathbf{w}_\beta^T \{\lambda_0(t) \mathbf{X} + \mathbf{h}_\beta^*(t)\}$. Set $\Delta = 1$, and take integration with respect to E . We obtain $\int_0^E \tilde{h}(u) du = 0$ with probability one. Therefore, $\int_0^t \tilde{h}(u) du = 0$ for $t \in [0, \zeta]$, which implies that $\mathbf{w}_\beta^T \mathbf{X} + \mathbf{w}_\beta^T \mathbf{h}_\beta^*(t)/\lambda_0(t) = 0$. By condition (C2), we have $\mathbf{w}_\beta = \mathbf{0}$. Similarly as in the proof of Lemma 1, we can show that $w_\alpha = w_r = 0$. Thus, $\mathbf{I}(\xi_0)$ is invertible.

To show (iii), we consider the classes $\mathcal{G}_\xi = \{\mathbf{I}_\xi(\xi, \lambda) : \xi \in \mathcal{A}, \lambda \in \Psi\}$ and $\mathcal{G}_\lambda = \{\mathbf{I}_\lambda(\xi, \lambda)[h] : \xi \in \mathcal{A}, \lambda \in \Psi, h \in \text{BV}_\omega[0, \zeta]\}$, where $\Psi = \{\lambda \in L_2[0, \zeta] : 1/M \leq \lambda \leq M \text{ and } \|\lambda\|_{\text{TV}} \leq C\}$, $\|\cdot\|_{\text{TV}}$ denotes the total variation norm, $\text{BV}_\omega[0, \zeta]$ denotes the set of all functions on $[0, \zeta]$ that are of total variation bounded by ω , and M, C and ω are positive constants. In the proof of convergence rate, we have shown that $\|\hat{\lambda}_n - \lambda_0\|_\infty = o_p(1)$. Further note that

$$\|\hat{\lambda}_n - \lambda_0\|_{\text{TV}} \leq \|\hat{\lambda}_n - \lambda_n\|_{\text{TV}} + \|\lambda_n - \lambda_0\|_{\text{TV}}$$

$$\begin{aligned} &\leq \sum_{k=1}^{m_n} |\hat{\gamma}_k - \gamma_{0k}| \|b'_k\|_\infty + o(1) \\ &\lesssim m_n^{3/2} \|\hat{\gamma} - \gamma_0\| + o(1), \end{aligned}$$

where b'_k is the derivative of b_k . If $q > 2$ and $\nu < 1/5$, then $m_n^{3/2} \|\hat{\gamma} - \gamma_0\| = O_p[n^{2\nu - \min\{q\nu, (1-\nu)/2\}}] = o_p(1)$. Thus, $\Pr(\hat{\lambda}_n \in \Psi) \rightarrow 1$ as $n \rightarrow \infty$. From (S10), we can see that each component of \mathbf{h}^* is continuously differentiable and thus belongs to $BV_\omega[0, \zeta]$. If we can show that \mathcal{G}_ξ and \mathcal{G}_λ are Donsker, then the first part of (iii) follows. To show that \mathcal{G}_ξ is Donsker, we derive its bracketing number. By the mean-value theorem, we can obtain that for any $(\xi_1, \lambda_1) \equiv (\beta_1, \alpha_1, r_1, \lambda_1)$ and $(\xi_2, \lambda_2) \equiv (\beta_2, \alpha_2, r_2, \lambda_2)$ with $\xi_1, \xi_2 \in \mathcal{A}$ and $\lambda_1, \lambda_2 \in \Psi$,

$$|I_\xi(\xi_1, \lambda_1) - I_\xi(\xi_2, \lambda_2)| \lesssim \|\beta_1 - \beta_2\| + |\alpha_1 - \alpha_2| + |r_1 - r_2| + \|\lambda_1 - \lambda_2\|_{L_2[0, \zeta]}.$$

Note that the compact set \mathcal{A} can be covered by $K_3\epsilon^{-(p+2)}$ number of ϵ -brackets for some $K_3 > 0$. Also, by Example 19.11 of van der Vaart (1998), we have $N_{[]}(\epsilon, \Psi, L_2[0, \zeta]) \leq e^{K_4/\epsilon}$ for some $K_4 > 0$. Therefore, $N_{[]}(\epsilon, \mathcal{G}_\xi, L_2(\mathbb{P})) \lesssim \epsilon^{-(p+2)} e^{K_5/\epsilon}$ for some $K_5 > 0$, so $\log N_{[]}(\epsilon, \mathcal{G}_\xi, L_2(\mathbb{P})) \lesssim \epsilon^{-1}$, and further

$$J_{[]} \{1, \mathcal{G}_\xi, L_2(\mathbb{P})\} = \int_0^1 \sqrt{\log N_{[]}(\epsilon, \mathcal{G}_\xi, L_2(\mathbb{P}))} d\epsilon \lesssim \int_0^1 \epsilon^{-1/2} d\epsilon < \infty.$$

It follows from Theorem 19.5 of van der Vaart (1998) that \mathcal{G}_ξ is Donsker. Similarly, we can show that \mathcal{G}_λ is also Donsker. The $L_2(\mathbb{P})$ convergence of $I_\xi(\hat{\xi}_n, \hat{\lambda}_n) - I_\lambda(\hat{\xi}_n, \hat{\lambda}_n)[\mathbf{h}^*]$ follows from the mean-value theorem. Thus, (iii) is proved.

Lastly, (iv) follows from the Donsker property of \mathcal{G}_λ by Corollary 2.3.12 of van der Vaart and Wellner (1996). This completes the proof. \square

References

- Backer JA, Klinkenberg D, Wallinga J (2020) Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. *Eurosurveillance* 25(5):2000062
- Dai J, Yang L, Zhao J (2020) Probable longer incubation period for elderly COVID-19 cases: analysis of 180 contact tracing data in Hubei Province, China. *Risk Management and Healthcare Policy* 13:1111–1117
- de Boor C (1976) Approximation Theory II. In: Lorentz GG, Chui CK, Schumaker LL (eds) *Splines as linear combinations of B-splines. A survey*. Academic Press, New York, pp 1–47
- de Boor C (2001) *A Practical Guide to Splines*, Revised Ed. Springer, Berlin
- Dejardin D, Lesaffre E (2013) Stochastic EM algorithm for doubly interval-censored data. *Biostatistics* 14(4):766–778
- Deng Y, You C, Liu Y, Qin J, Zhou X-H (2020). Estimation of incubation period and generation time based on observed length-biased epidemic cohort with censoring for COVID-19 outbreak in China. *Biometrics*, [online]. <https://doi.org/10.1111/biom.13325>.
- Goggins WB, Finkelstein DM, Zaslavsky AM (1999) Applying the Cox proportional hazards model for analysis of latency data with interval censoring. *Stat Med* 18(20):2737–2747
- Kong T-K (2020) Longer incubation period of coronavirus disease 2019 (COVID-19) in older adults. *Aging Medicine* 3(2):102–109

- Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, Azman AS, Reich NG, Lessler J (2020) The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Ann Intern Med* 172(9):577–582
- Li S, Sun J, Tian T, Cui X (2020) Semiparametric regression analysis of doubly censored failure time data from cohort studies. *Lifetime Data Anal* 26(2):315–338
- Li Z, Owzar K (2016) Fitting Cox models with doubly censored data using spline-based sieve marginal likelihood. *Scand J Stat* 43(2):476–486
- Linton NM, Kobayashi T, Yang Y, Hayashi K, Akhmetzhanov AR, Jung S-M, Yuan B, Kinoshita R, Nishiura H (2020) 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
- Pan W (2001) A multiple imputation approach to regression analysis for doubly censored data with application to AIDS studies. *Biometrics* 57(4):1245–1250
- Qin J, You C, Lin Q, Hu T, Yu S, 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. *Sci Adv* 6(33):eabc1202
- Reich NG, Lessler J, Cummings DAT, Brookmeyer R (2009) Estimating incubation period distributions with coarse data. *Stat Med* 28(22):2769–2784
- Sun J, Liao Q, Pagano M (1999) Regression analysis of doubly censored failure time data with applications to AIDS studies. *Biometrics* 55(3):909–914
- Sun L, Kim Y-J, Sun J (2004) Regression analysis of doubly censored failure time data using the additive hazards model. *Biometrics* 60(3):637–643
- Sun W, Yuan Y-X (2006) *Optimization Theory and Methods: Nonlinear Programming*, vol 1. Springer, New York
- Tan WYT, Wong LY, Leo YS, Toh MPHS (2020) Does incubation period of COVID-19 vary with age? A study of epidemiologically linked cases in singapore. *Epidemiology & Infection* 148:e197
- van de Geer SA (2000) *Applications of Empirical Process Theory*. Cambridge University Press, UK
- van der Vaart AW (1998) *Asymptotic Statistics*. Cambridge University Press, UK
- van der Vaart AW, Wellner JA (1996) *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York
- Wu Y, Zhang Y (2012) Partially monotone tensor spline estimation of the joint distribution function with bivariate current status data. *Ann Stat* 40(3):1609–1636
- Xiao Z, Guo W, Luo Z, Liao J, Wen F, Lin Y (2021) Examining geographical disparities in the incubation period of the COVID-19 infected cases in Shenzhen and Hefei. China. *Environmental Health and Preventive Medicine* 26(1):10
- Zhao Q, Ju N, Bacallado S, Shah RD (2021) Bets: The dangers of selection bias in early analyses of the coronavirus disease (COVID-19) pandemic. *The Annals of Applied Statistics* 15(1):363–390
- Zhou Q, Hu T, Sun J (2017) A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data. *J Am Stat Assoc* 112(518):664–672

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.