*Bernoulli* **11**(5), 2005, 863–892

# An asymptotic theory for the nonparametric maximum likelihood estimator in the Cox gene model

I-SHOU CHANG[1], CHAO AGNES HSUING[2*], MEI-CHUAN WANG[3] and CHI-CHUNG WEN[2**]

[1]*President's Laboratory, National Health Research Institutes, 35 Keyan Road, Zhunan Town, Miaoli County 350, Taiwan. E-mail: ischang@nhri.org.tw*
[2]*Division of Biostatistics and Bioinformatics, National Health Research Institutes, 35 Keyan Road, Zhunan Town, Miaoli County 350, Taiwan. E-Mail: *hsiung@nhri.org.tw; **wenqz@nhri.org.tw*
[3]*Department of Mathematics and Science Education, Taipei Municipal Teachers College, Taipei, Taiwan. E-mail: wmei@mail1.tmtc.edu.tw*

The Cox model with a gene effect for age at onset was introduced and studied by Li, Thompson and Wijsman. We study the nonparametric maximum likelihood estimation of the gene effect and the regression coefficient in this model. We indicate conditions under which the parameters are identifiable and the nonparametric maximum likelihood estimate is consistent and asymptotically normal. We also apply the theory of observed profile information to obtain a consistent estimate of the asymptotic variance. Besides providing theoretical support for Li *et al*., our work provides an alternative approach to the numerical methods in this model.

*Keywords:* age at onset; asymptotic normality; Cox gene model; discrete frailty model; identifiability; nonparametric maximum likelihood estimate; profile likelihood information

## 1. Introduction

Analysis of age at onset is an important approach to segregation and linkage studies of some complex genetic diseases. Among the many interesting models proposed for this purpose was a class of multivariate proportional hazards models in which the dependence of age at onset among family members is due to the segregation of an (unobserved) dominant gene in the family (Li *et al*. 1998; Siegmund and McKnight 1998). This paper provides an asymptotic theory for one of these models.

The following notation and model assumptions are used throughout the paper. Let $\mathrm{T}_{ik}$, $C_{ik}$, and $Z_{ik}$ denote respectively the age at onset, the censoring time, and the covariate of the $i$th individual in the $k$th family. Here $i = 1, \ldots, m$, $k = 1, \ldots, K$, and $m > 2$. Let $g_{ik}$ be the genotype of the $i$th individual in the $k$th family at a certain locus. Assume that there are two alleles at this locus and denote them by $A$ and $a$. Thus, the genotype takes one of

1350–7265 

the three values, *aa*, *aA*, or *AA*. Let $S_{ik} = I_{[g_{ik}=aA \text{ or } AA]}$ denote the susceptibility type. Denote by $q_0$ the population allele frequency of $A$.

Let $T_k = (T_{1k}, \ldots, T_{mk})$, $C_k = (C_{1k}, \ldots, C_{mk})$, $Z_k = (Z_{1k}, \ldots, Z_{mk})$, and $S_k = (S_{1k}, \ldots, S_{mk})$. Assume that, given $Z_k$ and $S_k$, $T_{1k}, \ldots, T_{mk}$, $C_{1k}, \ldots, C_{mk}$ are conditionally independent, and for an individual having covariate $z$ and susceptibility type $s$, the hazard of disease onset at age $t$ is

$$\lambda_0(t)\exp(\beta_0^{\mathrm{T}} z + \mu_0 s). \tag{1.1}$$

Here, $\lambda_0(\cdot)$ is a non-negative deterministic baseline function; $\beta_0 \in \mathfrak{R}^D$, $Z_{ik} \in \mathfrak{R}^D$, and $\mu_0 \in \mathfrak{R}$. Assume further that $(C_k, Z_k)$ and $S_k$ are independent. Let $X_{ik} = T_{ik} \wedge C_{ik}$, the minimum of $T_{ik}$ and $C_{ik}$, and $\delta_{ik} = I_{[T_{ik} \leqslant C_{ik}]}$.

Assume that $(T_k, C_k, Z_k, S_k)$ are independent and identically distributed for $k = 1, \ldots, K$. The purpose of this paper is to provide an asymptotic theory for the maximum likelihood estimation of $(\Lambda_0, \beta_0, \mu_0, q_0)$ based on $\{(X_k, \delta_k, Z_k) | k = 1, \ldots, K\}$, where $X_k = (X_{1k}, \ldots, X_{mk})$, $\delta_k = (\delta_{1k}, \ldots, \delta_{mk})$, and $\Lambda_0(t) = \int_0^t \lambda_0(u)\mathrm{d}u$.

We note that $\beta_0$ specifies the log of the risk ratio associated with the covariates $Z_{ik}$. If $\mu_0 > 0$, this model says that there is a single Mendelian diallelic locus governing the age-specific disease rate in a dominant way, and the gene carriers have their risk of onset increased by a factor of $\mathrm{e}^{\mu_0}$, compared to the non-carriers.

The assumption that $(C_k, Z_k)$ and $S_k$ are independent means that the gene in question has no effect on the covariates and has no correlation with the censoring variable. This is a desirable situation in statistical modelling and a reasonable assumption in genetics. Furthermore, it helps to simplify the theory.

We note that the assumption that $(T_k, C_k, Z_k, S_k)$ is an independent and identically distributed sequence is imposed so that this model is amenable to empirical process theory and standard semiparametric theory, and the asymptotic theory can be established. It would be interesting to extend the theory to other settings. An interesting extension would be to treat the covariate values of the subjects as being chosen by the experimenter and assume that $(T_k, C_k, S_k)$ is an independent, not necessarily identically distributed, sequence. Although desirable, this extension is beyond the scope of the present paper.

The above model was studied by Li *et al*. (1998) and was referred to as the Cox gene model. They proposed a generalized maximum likelihood estimation using a Monte Carlo EM algorithm, and indicated that their estimation procedure performs well in a simulation study and in the analysis of a real data set. In fact, they paid particular attention to the computational aspects of the problem. Readers are encouraged to consult Li *et al*. (1998) and Li and Thompson (1997) for additional background information on human genetics and for references regarding other approaches.

We note that if $S_k$ is observable, then (1.1) is the classical Cox model if the onset times are independent given covariates. Readers are referred to Andersen *et al*. (1993) for methods regarding this model.

In this paper, we give an asymptotic theory for the estimators of Li *et al*. (1998), including consistency, asymptotic normality, and an asymptotic theory for the calculation of the asymptotic variance. In view of the proportional hazards univariate frailty regression models (Kosorok *et al*. 2004 and references therein), the Cox model with shared gamma frailty (Murphy 1994; 1995), and the Cox model with correlated gamma frailty (Parner 1998), our work provides an asymptotic theory for the Cox's regression model with binary frailty allowing for shared frailty within families.

All the results of this paper rest on an identifiability assumption. We provide a method to check this identifiability assumption in the Appendix, and illustrate this method by showing that nuclear families with three siblings satisfy the assumption. Mack *et al*. (1990) found from simulation studies that simultaneous estimation of allele frequency and genetic relative risk is not possible based on sib-pair data, and suggested using more complex family structures as a means for obtaining more information. We also note that the simulation study of Li *et al*. (1998) is based on 10 four-generation pedigrees with 46 individuals per pedigree.

For the consistency, we first present an integral equation for the nonparametric maximum likelihood estimate (NPMLE). This equation gives a precise characterization of the score functions. Based on this, we apply empirical process theory to establish consistency, following the approaches and techniques developed in Murphy (1994), Murphy *et al*. (1997), Parner (1998), and Kosorok *et al*. (2004), among others. With the consistency, we then treat the NPMLE as a solution to certain estimating equations and prove its asymptotic normality by applying Theorem 3.3.1 in van der Vaart and Wellner (1996). We note that this approach to asymptotic normality was taken in van der Vaart (1994; 1995; 1996), Murphy (1995), and Parner (1998), among others. See also van der Vaart (1998) and van de Geer (2000) for relevant empirical process theory and its applications in the study of semiparametric models.

In the case where the regression parameter $\beta$ and the genetic parameters $\mu$ and $q$ are the parameters of interest, we apply the theory of observed profile information developed in Murphy and van der Vaart (1999) to provide an estimate of the asymptotic variance.

Although the framework of this paper is outlined in the previous references, additional work is needed to verify their general conditions to obtain these results (Murphy and van der Vaart 2000; Bickel and Ritov 2000).

Besides providing theoretical support for the statistical methods developed in Li *et al*. (1998), our work offers an alternative approach to the numerical methods of Li *et al*. (1998) and Siegmund and McKnight (1998). For example, the integral equation may be used to find approximations to the NPMLE, and the difficult problem of approximating the asymptotic variance of the NPMLE may be approached in terms of the observed profile information. These and other related computational issues are discussed in Chang *et al*. (2004b).

The organization of this paper is as follows. Section 2 presents the likelihood function, the identifiability assumption, the score functions, and the integral equation to be satisfied by the NPMLE. Sections 3 and 4 establish the consistency and asymptotic normality, respectively. Section 5 studies the problem of estimating $\beta$, $\mu$, and $q$, and presents the asymptotic variance in terms of the observed profile information. Section 6 is a discussion indicating possible extensions and future opportunities. Finally, the Appendix contains the technical proofs and quotes important theorems used in this paper.

## 2. Nonparametric maximum likelihood estimate

The parameter space we consider is $\Theta = \{(\Lambda, \beta, \mu, q) | \Lambda \in \mathcal{L}, \beta \in \mathcal{B}, \mu \in \mathcal{U}, q \in \mathcal{Q}\}$. Here, $\mathcal{L} = \{\Lambda : [0, \tau] \to [0, \infty) | \Lambda(0) = 0, \ \Lambda$ is non-decreasing and right continuous$\}$ for some positive real number $\tau$, $\mathcal{B}$ and $\mathcal{U}$ are compact subsets of $\mathfrak{R}^D$ and $(0, \infty)$, respectively, and $\mathcal{Q}$ is a closed subinterval of (0, 1). The analysis in this paper is restricted to the interval $[0, \tau]$; in order to obtain a reasonably simplified theory, an ideal $\tau$ should be large enough but still satisfy the property that there are subjects with ages at onset bigger than $\tau$. Elements of $\mathcal{L}$ are considered as the restrictions to $[0, \tau]$ of cumulative hazard functions. The true parameters $\beta_0$, $\mu_0$, and $q_0$ are assumed to be interior points of $\mathcal{B}$, $\mathcal{U}$, and $\mathcal{Q}$, respectively; and the true parameter $\Lambda_0$ has a positive and bounded derivative on $[0, \tau]$. We note that the assumption that $\Lambda_0$ has a positive derivative on $[0, \tau]$ is made to simplify the presentation for the proof of identifiability, and some extension in this regard is possible. We assume, for some $i = 1, \ldots, m$, that $P(T_{ik} > \tau | Z_{1k}, \ldots, Z_{mk}) > 0$ almost surely and $P(C_{ik} \geqslant \tau | Z_{1k}, \ldots, Z_{mk}) > 0$ a.s. Here these conditional probabilities are under the true model. We further assume that the support of $Z_{ik}$ is bounded and the linear span of the support of $\sum_{i=1}^m Z_{i1}$ has dimension $D$. We note that $Z_{ik}$ can be discrete or continuous. Because we are interested in estimating $\Lambda_0$ on $[0, \tau]$ using the likelihood, we assume $P(C_{ik} \leqslant \tau | Z_{1k}, \ldots, Z_{mk}) = 1$ a.s. for every $i = 1, \ldots, m$, to simplify the presentation without loss of generality.

While the assumptions in Section 1 are motivated by practical problems and make good intuitive sense, the above detailed description of the parameter space and related regularity conditions helps facilitate the development of a rigorous asymptotic theory. In particular, we will provide conditions under which the likelihood for a single family, considered as functions, on the sample space, indexed by the parameter values, defines different functions on the sample space by different parameter values. This property is referred to as identifiability of parameters in this paper and is needed in the study of consistency (see, for example, van der Vaart 1998, p. 62).

It follows that the likelihood for the $k$th family is

$$\tilde{L}_{(k),\Lambda,\beta,\mu,q} \equiv \sum_{s \in S} p(s, q) \left\{ \prod_{i=1}^m \left[ \lambda(X_{ik}) \mathrm{e}^{\beta^{\mathrm{T}} Z_{ik} + \mu s_i} \right]^{\delta_{ik}} \exp\left[ -\Lambda(X_{ik}) \mathrm{e}^{\beta^{\mathrm{T}} Z_{ik} + \mu s_i} \right] \right\}, \tag{2.1}$$

if the derivative of $\Lambda$ exists. Here $p(s, q) \equiv p(s_1, \ldots, s_m, q)$ is the probability that the susceptibility vector takes the value $(s_1, \ldots, s_m)$ when the dominant allele $A$ has frequency $q$. $S$ is the set of all possible values of $S_k = (S_{1k}, \ldots, S_{mk})$, and $\lambda(t) = \mathrm{d}\Lambda/\mathrm{d}t(t)$. We know that $p(s, q)$ depends on the family structure, and we assume that $p(s, q)$ is a smooth function and has a known functional form in this paper.

In this paper we need the following assumptions.

***Assumption I*** *Identifiability. For each* $\Lambda$ *that is absolutely continuous with respect to* $\Lambda_0$, $\tilde{L}_{(1),\Lambda,\beta,\mu,q} = \tilde{L}_{(1),\Lambda_0,\beta_0,\mu_0,q_0}$ *a.s. implies* $\Lambda = \Lambda_0$ *on* $[0, \tau]$, $\beta = \beta_0$, $\mu = \mu_0$, $q = q_0$.

***Assumption II.*** *There exists $t^*$ in the support of the conditional distribution of $C_{i1}$ given $Z_{i1}$ for every $i = 1, \ldots, m$ such that if*

$$\left( a_1 \frac{\partial}{\partial y_0} + a_2 \frac{\partial}{\partial y_1} + a_3 \frac{\partial}{\partial \mu} + a_4 \frac{\partial}{\partial q} \right) \Big|_{(y_0, y_1, \mu, q) = (1, \Lambda_0(t^*), \mu_0, q_0)}$$

$$\log \sum_s p(s, q) \left\{ \prod_{i=1}^m (y_0 e^{\mu s_i})^{\delta_{i1}} \exp(-y_1 e^{\beta_0^{\mathrm{T}} Z_{i1} + \mu s_i}) \right\} = 0,$$

for every possible value of $(\delta_{11}, \ldots, \delta_{m1}, Z_{11}, \ldots, Z_{m1})$, then $a_1 = a_2 = a_3 = a_4 = 0$.

***Remark 1.*** Assumption I formalizes the identifiability of the parameters in our model. Assumption II puts a restriction on the log-likelihood of our model, which includes the expression $p(s, q)$. In Appendix A.1, we show that Assumption II implies Assumption I. Appendix A.1 also includes a method to check the validity of Assumption II under the condition of random mating and Mendelian segregation. This method is illustrated by showing that a family consisting of three siblings satisfies Assumption II, and that if $\beta_0 \neq 0$ and $t^* \in (0, \tau]$, then Assumption II is valid even with $m = 1$. We note that identifiability of the mixed proportional hazards model has been studied by Elbers and Ridder (1982), Heckman and Singer (1984), Heckman and Taber (1994), Kortram *et al*. (1995), and Kosorok *et al*. (2004), among others.

***Remark 2.*** Assumption II is concerned with the behaviour of the model at time $t^*$. This assumption offers a way to study the invertibility of the information operator as well as the identifiability of the parameters in this model.

Because (2.1) could become arbitrarily large within the class of absolutely continuous $\Lambda$, we consider in the following the likelihood

$$L_K(\Lambda, \beta, \mu, q) \equiv \prod_{k=1}^K \sum_{s \in S} p(s, q) \left\{ \prod_{i=1}^m (\Delta\Lambda(X_{ik}) e^{\beta^{\mathrm{T}} Z_{ik} + \mu s_i})^{\delta_{ik}} \exp(-\Lambda(X_{ik}) e^{\beta^{\mathrm{T}} Z_{ik} + \mu s_i}) \right\}. \quad (2.2)$$

Here $\Delta\Lambda(t) = \Lambda(t) - \Lambda(t-)$.

The NPMLE $(\hat{\Lambda}_K, \hat{\beta}_K, \hat{\mu}_K, \hat{q}_K)$ we propose is the maximizer of (2.2) over $\mathcal{L}_* \times \mathcal{B} \times \mathcal{U} \times \mathcal{Q}$, where $\mathcal{L}_* \subset \mathcal{L}$ comprises step functions. In fact, $\hat{\Lambda}_K$ has positive jumps precisely at all $X_{ik}$ with $\delta_{ik} = 1$.

It is interesting to see that this NPMLE is also the NPMLE in the sense of Kiefer and Wolfowitz (1956), as is the partial likelihood estimate in Cox's regression model (Johansen 1983). In fact, we can follow the idea of Johansen (1983) to construct a model for jump processes depending on a non-negative increasing function $\Lambda$ with the property that if $\Lambda$ is absolutely continuous, it reduces to the Cox gene model (2.1). The likelihood for this model is

$$\prod_{k=1}^{K}\left[\sum_{s} p(s,\, q)\prod_{i=1}^{m}\prod_{u\leqslant\tau}\left\{\frac{[\mathrm{e}^{\beta^{\mathrm{T}}z_{ik}+\mu s_i}I_{(0,\mathrm{T}_{ik}\wedge C_{ik}]}(u)\Lambda(\mathrm{d}u)]^{\tilde{X}_{ik}(\mathrm{d}u)}}{\tilde{X}_{ik}(\mathrm{d}u)!}\cdot \mathrm{e}^{-\int_0^\tau \mathrm{e}^{\beta^{\mathrm{T}}z_{ik}+\mu s_i}}I_{(0,\mathrm{T}_{ik}\wedge C_{ik}]}(u)\Lambda(\mathrm{d}u)\right\}\right],$$

where $\tilde{X}_{ik}(\cdot) = I_{[\mathrm{T}_{ik},\infty)}(\cdot \wedge C_{ik})$, which has compensator $\int_0^t \mathrm{e}^{\beta^{\mathrm{T}}z_{ik}+\mu s_{ik}}I_{(0,\mathrm{T}_{ik}\wedge C_{ik}]}(u)\Lambda(\mathrm{d}u)$, is one of the jump processes in question. Considering the last part of this likelihood, we know it is maximized only when $\Lambda$ is a step function with possible jumps at $X_{ik}$ with $\delta_{ik} = 1$; this shows that the NPMLE based on (2.2) is also the NPMLE in the sense of Kiefer and Wolfowitz (1956).

We assume that all the random variables are defined on a sample space $\Omega$ with a specific $\sigma$-field. Let $\omega \in \Omega$ and $K$ be fixed. Observe from (2.2) that

$$|L_K(\Lambda,\, \beta,\, \mu,\, q)|$$
$$\leqslant \prod_{k=1}^{K}\sum_{s\in S} p(s,\, q)\left[\prod_{i=1}^{m}\left\{\Lambda(X_{ik})\mathrm{e}^{\beta^{\mathrm{T}}Z_{ik}+\mu s_i}\mathrm{e}^{-\Lambda(X_{ik})\mathrm{e}^{\beta^{\mathrm{T}}Z_{ik}+\mu s_i}}\right\}^{\delta_{ik}}\left\{\mathrm{e}^{-\Lambda(X_{ik})\mathrm{e}^{\beta^{\mathrm{T}}Z_{ik}+\mu s_i}}\right\}^{1-\delta_{ik}}\right].$$

Using the compactness of $\mathcal{B}\times\mathcal{U}\times\mathcal{Q}$ and the fact that $\lim_{y\to\infty} y\mathrm{e}^{-y} = 0$, we conclude that:

**Theorem 2.1.** *$\hat{\Lambda}_K(\tau) < M_K$ for some $M_K > 0$, and the NPMLE $(\hat{\Lambda}_K,\, \hat{\beta}_K,\, \hat{\mu}_K,\, \hat{q}_K)$ exists.*

In the rest of this section, we present the score functions and some integral equations to be satisfied by the NPMLE.

Let $BV[0,\, \tau]$ denote the set of all real-valued functions on $[0,\, \tau]$ with finite variation. For $h_1 \in BV[0,\, \tau]$, $h_2 \in \mathfrak{R}^D$, $h_3 \in \mathfrak{R}^1$, $h_4 \in \mathfrak{R}^1$, we define the score functions as follows. Let $\Lambda_\varepsilon(t) = \int_0^t (1+\varepsilon h_1(u))\mathrm{d}\Lambda(u)$. We define

$$\ell_{1,(\Lambda,\beta,\mu,q)}[h_1](X_1,\, \delta_1,\, Z_1) = \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\log L_1(\Lambda_\varepsilon,\, \beta,\, \mu,\, q)\bigg|_{\varepsilon=0},$$

$$\ell_{2,(\Lambda,\beta,\mu,q)}[h_2](X_1,\, \delta_1,\, Z_1) = \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\log L_1(\Lambda,\, \beta+\varepsilon h_2,\, \mu,\, q)\bigg|_{\varepsilon=0},$$

$$\ell_{3,(\Lambda,\beta,\mu,q)}[h_3](X_1,\, \delta_1,\, Z_1) = \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\log L_1(\Lambda,\, \beta,\, \mu+\varepsilon h_3,\, q)\bigg|_{\varepsilon=0},$$

$$\ell_{4,(\Lambda,\beta,\mu,q)}[h_4](X_1,\, \delta_1,\, Z_1) = \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\log L_1(\Lambda,\, \beta,\, \mu,\, q+\varepsilon h_4)\bigg|_{\varepsilon=0}.$$

Let

$$f_k(\Lambda,\, \beta,\, \mu,\, q,\, s) = p(s,\, q)\left(\prod_{i=1}^{m}[\exp(\mu s_i)]^{\delta_{ik}}\exp\left[-\Lambda(X_{ik})\exp(\beta^{\mathrm{T}}Z_{ik}+\mu s_i)\right]\right),$$

which is random, although the randomness is not indicated explicitly in the left-hand notation. Then

$$\ell_{1,(\Lambda,\beta,\mu,q)}[h_1](X_1,\, \delta_1,\, Z_1)$$

$$= \sum_{i=1}^{m} \delta_{i1} h_1(X_{i1}) - \frac{\sum_s f_1(\Lambda,\, \beta,\, \mu,\, q,\, s)\left(\sum_{i=1}^{m} \exp(\beta^{\mathrm{T}} Z_{i1} + \mu s_i)\int_0^{X_{i1}} h_1(t)\mathrm{d}\Lambda(t)\right)}{\sum_s f_1(\Lambda,\, \beta,\, \mu,\, q,\, s)}, \tag{2.3}$$

$$\ell_{2,(\Lambda,\beta,\mu,q)}[h_2](X_1,\, \delta_1,\, Z_1)$$

$$= h_2^{\mathrm{T}}\left(\sum_{i=1}^{m} \delta_{i1} Z_{i1} - \frac{\sum_s f_1(\Lambda,\, \beta,\, \mu,\, q,\, s)\left(\sum_{i=1}^{m} Z_{i1}\Lambda(X_{i1})\exp(\beta^{\mathrm{T}} Z_{i1} + \mu s_i)\right)}{\sum_s f_1(\Lambda,\, \beta,\, \mu,\, q,\, s)}\right), \tag{2.4}$$

$$\ell_{3,(\Lambda,\beta,\mu,q)}[h_3](X_1,\, \delta_1,\, Z_1)$$

$$= h_3 \frac{\sum_s f_1(\Lambda,\, \beta,\, \mu,\, q,\, s)\left(\sum_{i=1}^{m} \delta_{i1} s_i - \sum_{i=1}^{m} s_i \Lambda(X_{i1})\exp(\beta^{\mathrm{T}} Z_{i1} + \mu s_i)\right)}{\sum_s f_1(\Lambda,\, \beta,\, \mu,\, q,\, s)}, \tag{2.5}$$

$$\ell_{4,(\Lambda,\beta,\mu,q)}[h_4](X_1,\, \delta_1,\, Z_1) = h_4 \frac{\sum_s \partial f_1(\Lambda,\, \beta,\, \mu,\, q,\, s)/\partial_q}{\sum_s f_1(\Lambda,\, \beta,\, \mu,\, q,\, s)}, \tag{2.6}$$

where $\sum_s$ denotes summation over $s \in S$.

Derivations for (2.3), (2.4), (2.5), and (2.6) are straightforward and hence omitted.

By Theorem 2.1, it is clear that a necessary condition for $(\hat{\Lambda}_K,\, \hat{\beta}_K,\, \hat{\mu}_K,\, \hat{q}_K)$ to be the NPMLE is $\mathbb{P}_K \ell_{1,(\hat{\Lambda}_K,\hat{\beta}_K,\hat{\mu}_K,\hat{q}_K)}[h_1] = 0$. Here, $\mathbb{P}_K$ means taking expectation relative to the empirical distribution for the data $\{(X_k,\, \delta_k,\, Z_k) | k = 1, \ldots, K\}$; that is, $\mathbb{P}_K\, g \equiv K^{-1}\sum_{k=1}^{K} g(X_k,\, \delta_k,\, Z_k)$, for a function $g$ defined on the range of $(X_k,\, \delta_k,\, Z_k)$. In fact, we will show that the NPMLE converges to the true value almost surely, hence $\hat{\beta}_K$, $\hat{\mu}_K$, and $\hat{q}_K$ are interior point of $\mathcal{B}$, $\mathcal{U}$, and $\mathcal{Q}$ respectively for large $K$. This shows $\mathbb{P}_K \ell_{j,(\hat{\Lambda}_K,\hat{\beta}_K,\hat{\mu}_K,\hat{q}_K)}[h_j] = 0$ for all large $K$ and $j = 1, 2, 3, 4$.

The following lemma gives an integral equation for $\hat{\Lambda}_K$, which is useful in establishing the consistency, in Section 3, and the numerical method, in Chang *et al.* (2004b). Let

$$W_K(\Lambda,\, \beta,\, \mu,\, q;\, u) = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{m} \frac{\sum_s f_k(\Lambda,\, \beta,\, \mu,\, q,\, s)\exp(\beta^{\mathrm{T}} Z_{ik} + \mu s_i)}{\sum_s f_k(\Lambda,\, \beta,\, \mu,\, q,\, s)} I_{(0,X_{ik}]}(u),$$

$$G_K(u) = \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{m} I_{[\mathrm{T}_{ik},\infty)}(u \wedge C_{ik}),$$

$$W(\Lambda, \beta, \mu, q; u) = \mathrm{E}\, W_1(\Lambda, \beta, \mu, q; u),$$

$$G(u) = \mathrm{E}\, G_1(u).$$

**Lemma 2.1.**

$$\hat{\Lambda}_K(t) = \int_0^t \frac{1}{W_K(\hat{\Lambda}_K, \hat{\beta}_K, \hat{\mu}_K, \hat{q}_K; u)} \mathrm{d}G_K(u), \tag{2.7}$$

$$\Lambda_0(t) = \int_0^t \frac{1}{W(\Lambda_0, \beta_0, \mu_0, q_0; u)} \mathrm{d}G(u). \tag{2.8}$$

***Proof.*** Since $\mathbb{P}_K \ell_{1,(\hat{\Lambda}_K,\hat{\beta}_K,\hat{\mu}_K,\hat{q}_K)}[h_1] = 0$ for every $h_1 \in BV[0, \tau]$, we set $h_1(t) = I_{(0,u]}(t)$ in (2.3) and obtain, for every $u$,

$$\sum_{k=1}^{K} \sum_{i=1}^{m} \delta_{ik} I_{(0,u]}(X_{ik}) = \sum_{k=1}^{K} \frac{\sum_s f_k(\hat{\Lambda}_K, \hat{\beta}_K, \hat{\mu}_K, \hat{q}_K, s) \left( \sum_{i=1}^{m} \mathrm{e}^{\hat{\beta}_K^{\mathrm{T}} Z_{ik} + \hat{\mu}_K s_i} \int_0^u I_{(0,X_{ik}]}(t) \mathrm{d}\hat{\Lambda}_K(t) \right)}{\sum_s f_k(\hat{\Lambda}_K, \hat{\beta}_K, \hat{\mu}_K, \hat{q}_K, s)}. \tag{2.9}$$

Rewriting (2.9), we obtain

$$\sum_{k=1}^{K} \sum_{i=1}^{m} I_{[\mathrm{T}_{ik},\infty)}(u \wedge C_{ik}) = \int_0^u \sum_{k=1}^{K} \sum_{i=1}^{m} \frac{\sum_s f_k(\hat{\Lambda}_K, \hat{\beta}_K, \hat{\mu}_K, \hat{q}_K, s) \mathrm{e}^{\hat{\beta}_K^{\mathrm{T}} Z_{ik} + \hat{\mu}_K s_i}}{\sum_s f_k(\hat{\Lambda}_K, \hat{\beta}_K, \hat{\mu}_K, \hat{q}_K, s)} I_{(0,X_{ik}]}(t) \mathrm{d}\hat{\Lambda}_K(t), \tag{2.10}$$

which immediately gives (2.7).

The proof of (2.8) proceeds in the same way as that for (2.7) by noting that

$$\mathrm{E}\, \ell_{1,(\Lambda_0, \beta_0, \mu_0, q_0)}[h_1](X_1, \delta_1, Z_1) = 0$$

for every $h_1 \in BV[0, \tau]$. Therefore, the details of the proof are omitted. □

# 3. Consistency of NPMLE

The purpose of this section is to prove the following theorem.

**Theorem 3.1.** $\sup_{t\in[0,\tau]} |\hat{\Lambda}_K(t) - \Lambda_0(t)|$, $\|\hat{\beta}_K - \beta_0\|_D$, *$\hat{\mu}_K - \mu_0$, and $\hat{q}_K - q_0$ converge to 0 almost surely, as $K$ tends to infinity.*

Here and in the following, $\|\cdot\|_D$ is the Euclidean norm on $\Re^D$. We need a few lemmas before presenting the proof.

**Lemma 3.1.**

$$\sup_{\substack{u\in[0,\tau]\\ \Lambda\in\mathcal{L}_M,\beta\in\mathcal{B},\mu\in\mathcal{U},q\in\mathcal{Q}}} |W_K(\Lambda, \beta, \mu, q; u) - W(\Lambda, \beta, \mu, q; u)|$$

*converges to 0 almost surely, as K goes to infinity, where* $\mathcal{L}_M = \{\Lambda \in \mathcal{L} | \Lambda(\tau) \leqslant M\}$ *with* $0 < M < \infty$.

**Lemma 3.2.** *Let* $\Lambda \in \mathcal{L}_M$, $\beta \in \mathcal{B}$, $\mu \in \mathcal{U}$, *and* $q \in \mathcal{Q}$ *be given. Then*

$$\sup_{t\in[0,\tau]} \left| \int_0^t \frac{1}{W_K(\Lambda, \beta, \mu, q; u)} \mathrm{d}(G_K(u) - G(u)) \right|$$

*converges to 0 almost surely, as K goes to infinity.*

Lemma 3.1 and Lemma 3.2 are proved in Appendix A.2 and A.3, respectively, as applications of empirical process theory. Using Lemmas 3.1 and 3.2, we show that:

**Lemma 3.3.**

$$\sup_{t\in[0,\tau]} \left| \int_0^t \frac{1}{W_K(\Lambda, \beta, \mu, q; u)} \mathrm{d}G_K(u) - \int_0^t \frac{1}{W(\Lambda, \beta, \mu, q; u)} \mathrm{d}G(u) \right|$$

*converges to 0 almost surely.*

***Proof.*** Consider

$$\begin{aligned}
&\sup_{t\in[0,\tau]} \left| \int_0^t \frac{1}{W_K(\Lambda, \beta, \mu, q; u)} \mathrm{d}G_K(u) - \int_0^t \frac{1}{W(\Lambda, \beta, \mu, q; u)} \mathrm{d}G(u) \right| \\
&\quad \leqslant \sup_{t\in[0,\tau]} \left| \int_0^t \frac{1}{W_K(\Lambda, \beta, \mu, q; u)} \mathrm{d}(G_K - G)(u) \right| \\
&\qquad + \sup_{t\in[0,\tau]} \left| \int_0^t \left[ \frac{1}{W_K(\Lambda, \beta, \mu, q; u)} - \frac{1}{W(\Lambda, \beta, \mu, q; u)} \right] \mathrm{d}G(u) \right|. \qquad (3.1)
\end{aligned}$$

From the definition of $W_K$, there exists $c_2 > 0$ such that

$$W_K(\Lambda, \beta, \mu, q; u) \geqslant \frac{c_2}{K} \sum_{k=1}^K \sum_{i=1}^m I_{(0, X_{ik}]}(\tau),$$

for $(\Lambda, \beta, \mu, q)$ in $\Theta$ and $u$ in $[0, \tau]$; we know from the law of large numbers that $W_K$ is bounded away from 0 almost surely for all large $K$. Combining this with Lemmas 3.1, 3.2 and (3.1), we immediately obtain Lemma 3.3. □

**Lemma 3.4.** $\limsup_{K\to\infty} \hat{\Lambda}_K(\tau) < \infty$ *a.s.*

***Proof.*** Let

$$\tilde{\Lambda}_K(t) = \int_0^t \frac{1}{W_K(\Lambda_0, \beta_0, \mu_0, q_0; u)} \mathrm{d}G_K(u). \tag{3.2}$$

It follows from Lemmas 3.3 and 2.1 that $\sup_{t\in[0,\tau]}|\tilde{\Lambda}_K(t) - \Lambda_0(t)|$ converges to 0 almost surely. Let $A_k = [X_{ik} = \tau$ for some $i = 1, \ldots, m]$. Since $\sum_k P(A_k) = \infty$ and the events $\{A_k\}$ are independent, we have $P(A_k \text{ i.o.}) = 1$ by the Borel–Cantelli lemma. Here, the abbreviation i.o. stands for 'infinitely often'.

Suppose the conclusion of this lemma does not hold. Then there exists an $\omega \in [A_k \text{ i.o.}]$ satisfying $\sup_{t\in[0,\tau]}|\tilde{\Lambda}_K(t) - \Lambda_0(t)| \to 0$ such that for $\hat{\Lambda}_K(\equiv \hat{\Lambda}_K\{\omega\})$, $\limsup_{K\to\infty}\hat{\Lambda}_K(\tau) = \infty$. Since $\mathcal{B}\times\mathcal{U}\times\mathcal{Q}$ is compact, we can obtain a subsequence $\{K_j\}$ such that $(\hat{\beta}_{K_j}, \hat{\mu}_{K_j}, \hat{q}_{K_j}) \to (\tilde{\beta}, \tilde{\mu}, \tilde{q})$ and $\hat{\Lambda}_{K_j}(\tau) \to \infty$.

Because $(\hat{\Lambda}_{K_j}, \hat{\beta}_{K_j}, \hat{\mu}_{K_j}, \hat{q}_{K_j})$ maximizes $L_{K_j}$, we know that

$$\begin{aligned}
0 &\leqslant \frac{1}{K_j}\log L_{K_j}(\hat{\Lambda}_{K_j}, \hat{\beta}_{K_j}, \hat{\mu}_{K_j}, \hat{q}_{K_j}) - \frac{1}{K_j}\log L_{K_j}(\tilde{\Lambda}_{K_j}, \beta_0, \mu_0, q_0) \qquad (3.3)\\
&= \frac{1}{K_j}\sum_{k=1}^{K_j} \log\left\{\frac{\sum_s p(s, \hat{q}_{K_j})\prod_{i=1}^m \left[\Delta\hat{\Lambda}_{K_j}(X_{ik})\mathrm{e}^{\hat{\beta}_{K_j}^{\mathrm{T}} Z_{ik}+\hat{\mu}_{K_j}s_i}\right]^{\delta_{ik}} \mathrm{e}^{-\hat{\Lambda}_{K_j}(X_{ik})\mathrm{e}^{\hat{\beta}_{K_j}^{\mathrm{T}} Z_{ik}+\hat{\mu}_{K_j}s_i}}}{\sum_s p(s, q_0)\prod_{i=1}^m \left[\Delta\tilde{\Lambda}_{K_j}(X_{ik})\mathrm{e}^{\beta_0^{\mathrm{T}} Z_{ik}+\mu_0 s_i}\right]^{\delta_{ik}} \mathrm{e}^{-\tilde{\Lambda}_{K_j}(X_{ik})\mathrm{e}^{\beta_0^{\mathrm{T}} Z_{ik}+\mu_0 s_i}}}\right\}(I_{A_k} + I_{A_k^c})\\
&\leqslant \log\left\{\frac{1}{K_j}\sum_{k=1}^{K_j}\frac{\sum_s p(s, \hat{q}_{K_j})\prod_{i=1}^m \left[\Delta\hat{\Lambda}_{K_j}(X_{ik})\mathrm{e}^{\hat{\beta}_{K_j}^{\mathrm{T}} Z_{ik}+\hat{\mu}_{K_j}s_i}\right]^{\delta_{ik}} \mathrm{e}^{-\hat{\Lambda}_{K_j}(X_{ik})\mathrm{e}^{\hat{\beta}_{K_j}^{\mathrm{T}} Z_{ik}+\hat{\mu}_{K_j}s_i}}}{\sum_s p(s, q_0)\prod_{i=1}^m \left[\Delta\tilde{\Lambda}_{K_j}(X_{ik})\mathrm{e}^{\beta_0^{\mathrm{T}} Z_{ik}+\mu_0 s_i}\right]^{\delta_{ik}} \mathrm{e}^{-\tilde{\Lambda}_{K_j}(X_{ik})\mathrm{e}^{\beta_0^{\mathrm{T}} Z_{ik}+\mu_0 s_i}}} I_{A_k}\right\}\\
&\quad + \log\left\{\frac{1}{K_j}\sum_{k=1}^{K_j}\frac{\sum_s p(s, \hat{q}_{K_j})\prod_{i=1}^m \left[\Delta\hat{\Lambda}_{K_j}(X_{ik})\mathrm{e}^{\hat{\beta}_{K_j}^{\mathrm{T}} Z_{ik}+\hat{\mu}_{K_j}s_i}\right]^{\delta_{ik}} \mathrm{e}^{-\hat{\Lambda}_{K_j}(X_{ik})\mathrm{e}^{\hat{\beta}_{K_j}^{\mathrm{T}} Z_{ik}+\hat{\mu}_{K_j}s_i}}}{\sum_s p(s, q_0)\prod_{i=1}^m \left[\Delta\tilde{\Lambda}_{K_j}(X_{ik})\mathrm{e}^{\beta_0^{\mathrm{T}} Z_{ik}+\mu_0 s_i}\right]^{\delta_{ik}} \mathrm{e}^{-\tilde{\Lambda}_{K_j}(X_{ik})\mathrm{e}^{\beta_0^{\mathrm{T}} Z_{ik}+\mu_0 s_i}}} I_{A_k^c}\right\}.
\end{aligned}$$

The last inequality of (3.3) follows from Jensen's inequality.

It follows from the definition of $W_K$ that there exist $c_1$ and $c_2$ in $(0, \infty)$ such that

$$\frac{c_2}{K}\sum_{k=1}^K\sum_{i=1}^m I_{(0,X_{ik}]}(u) \leqslant W_K(\Lambda, \beta, \mu, q; u) \leqslant \frac{c_1}{K}\sum_{k=1}^K\sum_{i=1}^m I_{(0,X_{ik}]}(u), \tag{3.4}$$

for every $K$ and for every $(\Lambda, \beta, \mu, q, u)$ in its domain $\mathcal{L}\times\mathcal{B}\times\mathcal{U}\times\mathcal{Q}\times[0, \tau]$.

Using (2.7) and (3.2), we know that

$$\frac{\Delta\hat{\Lambda}_{K_j}(X_{ik})}{\Delta\tilde{\Lambda}_{K_j}(X_{ik})} = \frac{W_{K_j}(\Lambda_0, \beta_0, \mu_0, q_0; X_{ik})}{W_{K_j}(\hat{\Lambda}_{K_j}, \hat{\beta}_{K_j}, \hat{\mu}_{K_j}, \hat{q}_{K_j}; X_{ik})}. \tag{3.5}$$

It follows from (3.4) and (3.5) that

$$0 < \frac{c_2}{c_1} \leqslant \underline{\lim}_{j\to\infty} \frac{\Delta\hat{\Lambda}_{K_j}(X_{ik})}{\Delta\tilde{\Lambda}_{K_j}(X_{ik})} \leqslant \overline{\lim}_{j\to\infty} \frac{\Delta\hat{\Lambda}_{K_j}(X_{ik})}{\Delta\tilde{\Lambda}_{K_j}(X_{ik})} \leqslant \frac{c_1}{c_2}. \tag{3.6}$$

Using (3.6), the fact that $(\hat{\beta}_{K_j}, \hat{\mu}_{K_j}, \hat{q}_{K_j}) \to (\tilde{\beta}, \tilde{\mu}, \tilde{q})$, that $\tilde{\Lambda}_{K_j}$ converges to $\Lambda_0$ on $[0, \tau]$, and that $K_j^{-1}\sum_{k=1}^{K_j}\prod_{i=1}^{m} \mathrm{e}^{-\hat{\Lambda}_{K_j}(X_{ik})} I_{A_k} \to 0$ as $j \to \infty$, we can show that the difference of log-likelihood functions in (3.3) goes to $-\infty$ as $j \to \infty$. This leads to a contradiction, and the proof is thus complete. □

***Proof of Theorem 3.1.*** Consider $\omega \in \Omega$ for which $\limsup_{K\to\infty} \hat{\Lambda}_K(\tau) < \infty$, $\sup_{t\in[0,\tau]}|G_K(t) - G(t)| \to 0$, and $\sup_{t\in[0,\tau]}|\tilde{\Lambda}_K(t) - \Lambda_0(t)| \to 0$. Here $\tilde{\Lambda}_K$ is given by (3.2). Observing from (2.7) that

$$|\hat{\Lambda}_K(s) - \hat{\Lambda}_K(t)| \leqslant O(1)|G_K(s) - G_K(t)| \leqslant O(1)|G(s) - G(t)| + o(1),$$

for $s, t \in [0, \tau]$, we can make use of the compactness of $\mathcal{B} \times \mathcal{U} \times \mathcal{Q}$ and the arguments in proving the Arzelà-Ascoli theorem (see, for example, Rudin 1976, Theorem 7.25) to show that there exists a subsequence $\{K_j\}$ for which $(\hat{\Lambda}_{K_j}, \hat{\beta}_{K_j}, \hat{\mu}_{K_j}, \hat{q}_{K_j})$ converges uniformly to some $(\Lambda^*, \beta^*, \mu^*, q^*)$. We will show that $(\Lambda^*, \beta^*, \mu^*, q^*) = (\Lambda_0, \beta_0, \mu_0, q_0)$.

We now explain that

$$\begin{aligned}\hat{\Lambda}_{K_j}(t) &= \int_0^t \frac{1}{W_{K_j}(\hat{\Lambda}_{K_j}, \hat{\beta}_{K_j}, \hat{\mu}_{K_j}, \hat{q}_{K_j}; u)} \mathrm{d}G_{K_j}(u) \\ &= \int_0^t \frac{1}{W_{K_j}(\hat{\Lambda}_{K_j}, \hat{\beta}_{K_j}, \hat{\mu}_{K_j}, \hat{q}_{K_j}; u)} \mathrm{d}G(u) + o(1) \\ &= \int_0^t \frac{1}{W(\hat{\Lambda}_{K_j}, \hat{\beta}_{K_j}, \hat{\mu}_{K_j}, \hat{q}_{K_j}; u)} \mathrm{d}G(u) + o(1) \\ &= \int_0^t \frac{1}{W(\Lambda^*, \beta^*, \mu^*, q^*; u)} \mathrm{d}G(u) + o(1).\end{aligned}$$

The first equality is Lemma 2.1, the second equality can be proved by using Lemma 3.1 and the arguments for proving Lemma 3.2, the third equality follows from Lemma 3.1, and the last equality follows from the uniform convergence of $(\hat{\Lambda}_{K_j}, \hat{\beta}_{K_j}, \hat{\mu}_{K_j}, \hat{q}_{K_j})$ and the Lebesgue dominated convergence theorem. Consequently,

$$\Lambda^*(t) = \int_0^t \frac{1}{W(\Lambda^*, \beta^*, \mu^*, q^*; u)} \mathrm{d}G(u).$$

This, together with Lemmas 2.1 and 3.1, implies, uniformly in $t$, that

$$\frac{\mathrm{d}\hat{\Lambda}_{K_j}}{\mathrm{d}\tilde{\Lambda}_{K_j}}(t) = \frac{W_{K_j}(\Lambda_0, \beta_0, \mu_0, q_0;\ t)}{W_{K_j}(\hat{\Lambda}_{K_j}, \hat{\beta}_{K_j}, \hat{\mu}_{K_j}, \hat{q}_{K_j};\ t)} \to \frac{W(\Lambda_0, \beta_0, \mu_0, q_0;\ t)}{W(\Lambda^*, \beta^*, \mu^*, q^*;\ t)} = \frac{\mathrm{d}\Lambda^*}{\mathrm{d}\Lambda_0}(t),$$

which is bounded and bounded away from 0 by (3.4).

Considering $[\mathrm{d}\hat{\Lambda}_{K_j}/\mathrm{d}\tilde{\Lambda}_{K_j}(x)]^{[\delta=1]}$ as a function in $(x, \delta)$ with $\omega$ fixed, we can use the argument in the proof of Lemma 3.2 and properties concerning random functions and Glivenko–Cantelli class (see, for example, van der Vaart 1998, p. 279) to obtain

$$\begin{aligned} &\frac{1}{K_j}\sum_{k=1}^{K_j} \log \prod_{i=1}^{m}\left[\frac{\mathrm{d}\hat{\Lambda}_{K_j}}{\mathrm{d}\tilde{\Lambda}_{K_j}}(X_{ik})\right]^{\delta_{ik}} - \mathrm{E}\log \prod_{i=1}^{m}\left[\frac{\mathrm{d}\hat{\Lambda}_{K_j}}{\mathrm{d}\tilde{\Lambda}_{K_j}}(X_{i1})\right]^{\delta_{i1}} \\ &= \int_0^{\tau} \log \frac{\mathrm{d}\hat{\Lambda}_{K_j}}{\mathrm{d}\tilde{\Lambda}_{K_j}}(t)\mathrm{d}(G_{K_j}(t) - G(t)) \\ &= o(1) \text{ a.s.} \end{aligned} \tag{3.7}$$

In fact, the expectation in (3.7) is taken only over $(X_{i1}, \delta_{i1})$ with parameter estimator $\hat{\Lambda}_{K_j}$ substituted after taking the expectation.

Because the set of functions

$$\log \sum_s p(s, q)\left(\prod_{i=1}^{m}[\mathrm{e}^{\beta^{\mathrm{T}} Z_{i1}+\mu s_i}]^{\delta_{i1}} \exp[-\Lambda(X_{i1})\mathrm{e}^{\beta^{\mathrm{T}} Z_{i1}+\mu s_i}]\right),$$

indexed by $(\Lambda, \beta, \mu, q) \in \mathcal{L}_M \times \mathcal{B} \times \mathcal{U} \times \mathcal{Q}$, is Glivenko–Cantelli, we can conclude that

$$\begin{aligned} &\left| \frac{1}{K_j}\sum_{k=1}^{K_j} \log \frac{\sum_s p(s, \hat{q}_{K_j})\left(\prod_{i=1}^{m}\left[\mathrm{e}^{\hat{\beta}_{K_j}^{\mathrm{T}} Z_{ik}+\hat{\mu}_{K_j}s_i}\right]^{\delta_{ik}} \mathrm{e}^{-\hat{\Lambda}_{K_j}(X_{ik})\mathrm{e}^{\hat{\beta}_{K_j}^{\mathrm{T}} Z_{ik}+\hat{\mu}_{K_j}s_i}}\right)}{\sum_s p(s, q_0)\left(\prod_{i=1}^{m}\left[\mathrm{e}^{\beta_0^{\mathrm{T}} Z_{ik}+\mu_0 s_i}\right]^{\delta_{ik}} \mathrm{e}^{-\tilde{\Lambda}_{K_j}(X_{ik})\mathrm{e}^{\beta_0^{\mathrm{T}} Z_{ik}+\mu_0 s_i}}\right)} \right. \\ &\left. - \mathrm{E}\log \frac{\sum_s p(s, \hat{q}_{K_j})\left(\prod_{i=1}^{m}\left[\mathrm{e}^{\hat{\beta}_{K_j}^{\mathrm{T}} Z_{i1}+\hat{\mu}_{K_j}s_i}\right]^{\delta_{i1}} \mathrm{e}^{-\hat{\Lambda}_{K_j}(X_{i1})\mathrm{e}^{\hat{\beta}_{K_j}^{\mathrm{T}} Z_{i1}+\hat{\mu}_{K_j}s_i}}\right)}{\sum_s p(s, q_0)\left(\prod_{i=1}^{m}\left[\mathrm{e}^{\beta_0^{\mathrm{T}} Z_{i1}+\mu_0 s_i}\right]^{\delta_{i1}} \mathrm{e}^{-\tilde{\Lambda}_{K_j}(X_{i1})\mathrm{e}^{\beta_0^{\mathrm{T}} Z_{i1}+\mu_0 s_i}}\right)} \right| = o(1) \text{ a.s.} \end{aligned} \tag{3.8}$$

Here we use the same arguments concerning random functions in deriving (3.7). In particular, the expectation in (3.8) is taken only over $(X_{i1}, \delta_{i1}, Z_{i1})$ with parameter estimators $(\hat{\Lambda}_{K_j}, \hat{\beta}_{K_j}, \hat{\mu}_{K_j}, \hat{q}_{K_j})$ substituted after taking the expectation. The expectation in (3.9) is to be understood in the same manner.

Using (3.7) and (3.8), we obtain

$$
\begin{aligned}
0 &\leqslant \frac{1}{K_j}\log L_{K_j}(\hat{\Lambda}_{K_j}, \hat{\beta}_{K_j}, \hat{\mu}_{K_j}, \hat{q}_{K_j}) - \frac{1}{K_j}\log L_{K_j}(\tilde{\Lambda}_{K_j}, \beta_0, \mu_0, q_0) \\
&= \frac{1}{K_j}\sum_{k=1}^{K_j}\log\prod_{i=1}^{m}\left[\frac{\mathrm{d}\hat{\Lambda}_{K_j}}{\mathrm{d}\tilde{\Lambda}_{K_j}}(X_{ik})\right]^{\delta_{ik}} \\
&\quad + \frac{1}{K_j}\sum_{k=1}^{K_j}\log\frac{\sum_s p(s, \hat{q}_{K_j})\left(\prod_{i=1}^{m}\left[\mathrm{e}^{\hat{\beta}_{K_j}^{\mathrm{T}} Z_{ik}+\hat{\mu}_{K_j}s_i}\right]^{\delta_{ik}}\exp\left[-\hat{\Lambda}_{K_j}(X_{ik})\mathrm{e}^{\hat{\beta}_{K_j}^{\mathrm{T}} Z_{ik}+\hat{\mu}_{K_j}s_i}\right]\right)}{\sum_s p(s, q_0)\left(\prod_{i=1}^{m}\left[\mathrm{e}^{\beta_0^{\mathrm{T}} Z_{ik}+\mu_0 s_i}\right]^{\delta_{ik}}\exp\left[-\tilde{\Lambda}_{K_j}(X_{ik})\mathrm{e}^{\beta_0^{\mathrm{T}} Z_{ik}+\mu_0 s_i}\right]\right)} \\
&= \mathrm{E}\log\frac{\tilde{L}_{(1),\hat{\Lambda}_{K_j},\hat{\beta}_{K_j},\hat{\mu}_{K_j},\hat{q}_{K_j}}}{\tilde{L}_{(1),\tilde{\Lambda}_{K_j},\beta_0,\mu_0,q_0}} + o(1) \text{ a.s.}
\end{aligned} \tag{3.9}
$$

Considering the $\omega$ chosen at the beginning of the proof for which (3.9) is also satisfied, using Jensen's inequality and the Kullback–Leibler divergence theorem (see, for example, van der Vaart 1998, p. 62), we obtain $\tilde{L}_{(1),\Lambda^*,\beta^*,\mu^*,q^*} = \tilde{L}_{(1),\Lambda_0,\beta_0,\mu_0,q_0}$. Since $\Lambda^*$ is absolutely continuous relative to $G$ and thus also to $\Lambda_0$ by Lemma 2.1, we can use identifiability in Assumption I to conclude that $\Lambda^* = \Lambda_0$ on $[0, \tau]$, $\beta^* = \beta_0$, $\mu^* = \mu_0$, $q^* = q_0$. This completes the proof. □

# 4. Asymptotic normality of NPMLE

We will prove the asymptotic normality by verifying the conditions in Theorem 3.3.1 and Lemma 3.3.5 of van der Vaart and Wellner (1996), both stated in Appendix A.4 for reference. For this purpose, a few lemmas are needed. Let $\mathcal{H} = BV[0, \tau] \times \mathfrak{R}^D \times \mathfrak{R}^1 \times \mathfrak{R}^1$. For $(h_1, h_2, h_3, h_4) \in \mathcal{H}$, we introduce the norm $\|(h_1, h_2, h_3, h_4)\|_{\mathcal{H}} = \|h_1\|_V + \|h_2\|_D + |h_3| + |h_4|$, where $\|h_1\|_V$ is the sum of the absolute value of $h_1(0)$ and the total variation of $h_1$ on $[0, \tau]$. Let $H_p$ be the subset of $\mathcal{H}$ with $\|(h_1, h_2, h_3, h_4)\|_{\mathcal{H}} \leqslant p$ if $p < \infty$. If $p = \infty$, then the previous inequality is strict. Define $(\Lambda, \beta, \mu, q)(\mathbf{h}) = \int_0^\tau h_1\mathrm{d}\Lambda + h_2^{\mathrm{T}}\beta + h_3\mu + h_4 q$, where $\mathbf{h} = (h_1, h_2, h_3, h_4)$, and consider the parameter space $\Theta$ a subset of $\ell^\infty(H_p)$, the space of all bounded real-valued functions on $H_p$ under the supremum norm $\|\theta\|_{\ell^\infty(H_p)} = \sup_{\mathbf{h}\in H_p}|\theta(\mathbf{h})|$. We note that

$$
\begin{aligned}
(p/\sqrt{D})(\|\Lambda - \Lambda_0\|_* \vee \|\beta - \beta_0\|_D \vee |\mu - \mu_0| \vee |q - q_0|) &\leqslant \|\theta - \theta_0\|_{\ell^\infty(H_p)} \\
&\leqslant 4p(\|\Lambda - \Lambda_0\|_* \vee \|\beta - \beta_0\|_D \vee |\mu - \mu_0| \vee |q - q_0|),
\end{aligned}
$$

where $\|\Lambda - \Lambda_0\|_* = \sup_{\|h_1\|_V\leqslant 1}|\int_0^\tau h_1\,\mathrm{d}(\Lambda - \Lambda_0)|$ is the natural norm for a bounded linear operator on the normed space $BV[0, \tau]$.

Define $\Psi_K : \Theta \to \ell^\infty(H_p)$ by

$$\Psi_K(\Lambda, \beta, \mu, q)(h_1, h_2, h_3, h_4) = \mathbb{P}_K \sum_{j=1}^{4} \ell_{j,(\Lambda,\beta,\mu,q)}[h_j]$$

$$= \frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{4} \ell_{j,(\Lambda,\beta,\mu,q)}[h_j](X_k, \delta_k, Z_k).$$

To simplify the notation, let

$$\phi_{\theta,\mathbf{h}} = \sum_{j=1}^{4} \ell_{j,\theta}[h_j],$$

for $\theta = (\Lambda, \beta, \mu, q)$ and $\mathbf{h} = (h_1, h_2, h_3, h_4)$.

Let $\Psi : \Theta \to \ell^\infty(H_p)$ be defined by $\Psi(\theta)(h_1, h_2, h_3, h_4) = \mathrm{E}\Psi_1(\theta)(h_1, h_2, h_3, h_4)$.

**Lemma 4.1.** $\sqrt{K}(\Psi_K(\theta_0) - \Psi(\theta_0))$ *converges weakly to a Gaussian process* $\mathcal{W}$ in $\ell^\infty(H_p)$ *for every* $0 < p < \infty$.

***Proof.*** According to empirical process theory, it is sufficient to show that $\{\phi_{\theta_0,\mathbf{h}} | \mathbf{h} \in H_p\}$ is a Donsker class. Since $\phi_{\theta_0,\mathbf{h}}$ depends on $h_2$, $h_3$, and $h_4$ linearly, it follows from Theorem 2.10.6 in van der Vaart and Wellner (1996) (Appendix A.4) that we only need to show that both $\{h_1(\cdot) | h_1 \in BV[0, \tau],\ \|h_1\|_V < p\}$ and $\{\int_0^{\cdot} h_1(t)\mathrm{d}\Lambda_0(t) | h_1 \in BV[0, \tau],\ \|h_1\|_V < p\}$ are Donsker classes. Due to the fact that the class of functions with a common upper bound of their total variations is Donsker (see, for example, van der Vaart 1998, Example 19.11), they are both indeed Donsker. This completes the proof. □

**Lemma 4.2.** $\{\phi_{\theta,\mathbf{h}} - \phi_{\theta_0,\mathbf{h}} |\ \|\theta - \theta_0\|_{\ell^\infty(H_p)} < \delta,\ \mathbf{h} \in H_p\}$ *is Donsker.*

We omit the proof of this lemma, because it only involves arguments similar to, though more complicated than, those used in the proof of Lemma 4.1.

**Lemma 4.3.** $\lim_{\theta\to\theta_0} \sup_{\mathbf{h}\in H_p} \mathrm{E}(\phi_{\theta,\mathbf{h}} - \phi_{\theta_0,\mathbf{h}})^2 = 0.$

The proof of Lemma 4.3 is also omitted, because it is straightforward.

Let $\mathrm{lin}\Theta$ denote the set of all finite linear combinations of $\theta - \theta_0$, for $\theta \in \Theta$.

**Lemma 4.4.** *Let* $p < \infty$. *There is a continuous linear map* $\dot{\Psi}_{\theta_0} : \mathrm{lin}\Theta \to \ell^\infty(H_p)$ *satisfying*

$$\|\Psi(\theta) - \Psi(\theta_0) - \dot{\Psi}_{\theta_0}(\theta - \theta_0)\|_{\ell^\infty(H_p)} = \circ(\|\theta - \theta_0\|_{\ell^\infty(H_p)}).$$

*In addition,* $\dot{\Psi}_{\theta_0}$ *has a continuous inverse on its range.*

***Proof.*** (i) *Existence of* $\dot{\Psi}_{\theta_0}$. Let $\theta = (\Lambda, \beta, \mu, q)$. Using (2.3)–(2.6), we obtain the following first-order Taylor expansion for $\phi_{\theta,\mathbf{h}}(X_1, \delta_1, Z_1)$ as a function defined on $\Theta \subset \ell^\infty(H_p)$:

$$\phi_{\theta,\mathbf{h}}(X_1, \delta_1, Z_1) - \phi_{\theta_0,\mathbf{h}}(X_1, \delta_1, Z_1) \tag{4.1}$$

$$= \eta\bigg(\Lambda(X_{11}) - \Lambda_0(X_{11}), \ldots, \Lambda(X_{m1}) - \Lambda_0(X_{m1}), \int_0^{X_{11}} h_1(t)\mathrm{d}(\Lambda(t) - \Lambda_0(t)), \ldots,$$

$$\int_0^{X_{m1}} h_1(t)\mathrm{d}(\Lambda(t) - \Lambda_0(t)), \beta - \beta_0, \mu - \mu_0, q - q_0, h_2, h_3, h_4\bigg) + \circ(\|\theta - \theta_0\|_{\ell^\infty(H_p)})$$

for some linear function $\eta$, as $\theta$ gets close to $\theta_0$.

Examining (4.1) and (2.3)–(2.6) and considering the weaker form

$$|\Psi(\theta)(\mathbf{h}) - \Psi(\theta_0)(\mathbf{h}) - \dot{\Psi}_{\theta_0}(\theta - \theta_0)(\mathbf{h})| = \circ(|(\theta - \theta_0)(\mathbf{h})|)$$

of Lemma 4.4, we can see that

$$\Psi(\theta)(\mathbf{h}) - \Psi(\theta_0)(\mathbf{h})$$

$$= \mathrm{E}[\phi_{\theta,\mathbf{h}}(X_1, \delta_1, Z_1) - \phi_{\theta_0,\mathbf{h}}(X_1, \delta_1, Z_1)]$$

$$= -\bigg[\int_0^\tau \sigma_1(\mathbf{h})(t)\mathrm{d}(\Lambda(t) - \Lambda_0(t)) + \sigma_2(\mathbf{h})^{\mathrm{T}}(\beta - \beta_0) + \sigma_3(\mathbf{h})(\mu - \mu_0) + \sigma_4(\mathbf{h})(q - q_0)\bigg]$$

$$+ R(\theta)(\mathbf{h}), \tag{4.2}$$

for some continuous linear operator $\sigma = (\sigma_1, \sigma_2, \sigma_3, \sigma_4)$ from $H_\infty$ to $H_\infty$, and remainder term $R(\theta)$ satisfying

$$\lim_{\theta\to\theta_0} \frac{\sup_{\mathbf{h}\in H_p} |R(\theta)(\mathbf{h})|}{\|\theta - \theta_0\|_{\ell^\infty(H_p)}} = 0.$$

In fact, the above weaker form suggests

$$\dot{\Psi}_{\theta_0}(\theta - \theta_0)(\mathbf{h}) = \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\bigg|_{\varepsilon=0} \Psi(\theta_0 + \varepsilon(\theta - \theta_0))(\mathbf{h}),$$

which can be computed easily and gives

$$\dot{\Psi}_{\theta_0}(\theta - \theta_0)(\mathbf{h}) \tag{4.3}$$

$$= -\bigg[\int_0^\tau \sigma_1(\mathbf{h})(t)\mathrm{d}(\Lambda(t) - \Lambda_0(t)) + \sigma_2(\mathbf{h})^{\mathrm{T}}(\beta - \beta_0) + \sigma_3(\mathbf{h})(\mu - \mu_0) + \sigma_4(\mathbf{h})(q - q_0)\bigg].$$

(ii) *Invertibility of* $\dot{\Psi}_{\theta_0}$. In order to show that $\dot{\Psi}_{\theta_0}$ has a continuous inverse on its range, it is sufficient to show that there is an $\varepsilon > 0$ such that

$$\inf_{0\neq\theta-\theta_0\in\mathrm{lin}\Theta} \frac{\|\dot{\Psi}_{\theta_0}(\theta - \theta_0)\|_{\ell^\infty(H_p)}}{\|\theta - \theta_0\|_{\ell^\infty(H_p)}} \geqslant \varepsilon.$$

According to Lemmas 4.5 and 4.6, we know that $\sigma : H_\infty \to H_\infty$ is continuously invertible. This implies that there exists $r > 0$ such that $\sigma^{-1}(H_r) \subset H_p$. Hence, using (4.3), we know that

$$\begin{aligned}
&\|\dot{\Psi}_{\theta_0}(\theta - \theta_0)\|_{\ell^\infty(H_p)} \\
&\geqslant \sup_{\mathbf{h}\in H_r} \left| \int_0^\tau h_1(u)\mathrm{d}(\Lambda - \Lambda_0) + h_2^{\mathrm{T}}(\beta - \beta_0) + h_3(\mu - \mu_0) + h_4(q - q_0) \right| \\
&= \|\theta - \theta_0\|_{\ell^\infty(H_r)}.
\end{aligned}$$

This shows that

$$\inf_{0\neq\theta-\theta_0\in\mathrm{lin}\Theta} \frac{\|\dot{\Psi}_{\theta_0}(\theta - \theta_0)\|_{\ell^\infty(H_p)}}{\|\theta - \theta_0\|_{\ell^\infty(H_p)}} \geqslant \frac{r}{4p\sqrt{D}}.$$

This completes the proof. □

Considering (4.3) and the negative second directional derivative of the log-likelihood of the parametric submodel

$$(\epsilon_1, \epsilon_2) \mapsto \left(\Lambda_0 + \epsilon_1 \int_0^{\cdot} h_1\,\mathrm{d}\Lambda_0 + \epsilon_2 \int_0^{\cdot} h_1^*\,\mathrm{d}\Lambda_0, \beta_0 + \epsilon_1 h_2 + \epsilon_2 h_2^*, \mu_0 + \epsilon_1 h_3 + \epsilon_2 h_3^*, q_0 + \epsilon_1 h_4 + \epsilon_2 h_4^*\right)$$

for $\epsilon_1$, $\epsilon_2$ near 0, we obtain the following equation connecting the information and the score:

$$\begin{aligned}
&\int_0^\tau \sigma_1(\mathbf{h})(u)h_1^*(u)\mathrm{d}\Lambda_0(u) + h_2^{*\mathrm{T}}\sigma_2(\mathbf{h}) + h_3^*\sigma_3(\mathbf{h}) + h_4^*\sigma_4(\mathbf{h}) \\
&\quad = \mathrm{E}\left[\sum_{j=1}^4 \ell_{j,\theta_0}[h_j](X_1, \delta_1, Z_1)\right]\left[\sum_{j=1}^4 \ell_{j,\theta_0}[h_j^*](X_1, \delta_1, Z_1)\right],
\end{aligned} \tag{4.4}$$

for $(h_1, h_2, h_3, h_4)$ and $(h_1^*, h_2^*, h_3^*, h_4^*)$ in $H_\infty$.

**Lemma 4.5.** *$\sigma$ is one to one.*

***Proof.*** Assume $\sigma(\mathbf{h}) = 0$. Using (4.4), we know that

$$\sum_{j=1}^4 \ell_{j,\theta_0}[h_j](X_1, \delta_1, Z_1) = 0 \text{ a.s.} \tag{4.5}$$

Considering $X_{i1}$ near 0 from the right and $\delta_{i1} = 1$ for every $i = 1, \ldots, m$ in (4.5) and using (2.3)–(2.6) and a similar argument in Appendix A.1, we obtain

$$\sum_{i=1}^{m} h_1(0) + h_2^{\mathrm{T}} \sum_{i=1}^{m} Z_{i1} + h_3 \frac{\sum_s p(s,\, q_0) \left( \prod_{i=1}^{m} \mathrm{e}^{\mu_0 s_i} \right) \left( \sum_{i=1}^{m} s_i \right)}{\sum_s p(s,\, q_0) \left( \prod_{i=1}^{m} \mathrm{e}^{\mu_0 s_i} \right)}$$

$$+ h_4 \frac{\sum_s \partial p(s,\, q)/\partial q|_{q=q_0} \left( \prod_{i=1}^{m} \mathrm{e}^{\mu_0 s_i} \right)}{\sum_s p(s,\, q_0) \left( \prod_{i=1}^{m} \mathrm{e}^{\mu_0 s_i} \right)} = 0.$$

Using the assumption that the linear span of the support of $\sum_{i=1}^{m} Z_{i1}$ has dimension $D$, we know that $h_2 = 0$. Putting $h_2 = 0$ in (4.5) and considering $X_{i1}$ near $t^*$ from the right for $i = 1, \ldots, m$, we can similarly obtain

$$h_1(t^*) \left[ \sum_{i=1}^{m} \delta_{i1} \right] - \int_0^{t^*} h_1 \, \mathrm{d}\Lambda_0 \left[ \frac{\sum_s f_0(\Lambda_0(t^*),\, \beta_0,\, \mu_0,\, q_0,\, s) \left( \sum_{i=1}^{m} \mathrm{e}^{\beta_0^{\mathrm{T}} Z_{i1} + \mu_0 s_i} \right)}{\sum_s f_0(\Lambda_0(t^*),\, \beta_0,\, \mu_0,\, q_0,\, s)} \right]$$

$$+ h_3 \left[ \frac{\sum_s f_0(\Lambda_0(t^*),\, \beta_0,\, \mu_0,\, q_0,\, s) \left( \sum_{i=1}^{m} \delta_{i1} s_i - \sum_{i=1}^{m} s_i \Lambda_0(t^*) \mathrm{e}^{\beta_0^{\mathrm{T}} Z_{i1} + \mu_0 s_i} \right)}{\sum_s f_0(\Lambda_0(t^*),\, \beta_0,\, \mu_0,\, q_0,\, s)} \right]$$

$$+ h_4 \left[ \frac{\sum_s \partial f_0(\Lambda_0(t^*),\, \beta_0,\, \mu_0,\, q,\, s)/\partial q|_{q=q_0}}{\sum_s f_0(\Lambda_0(t^*),\, \beta_0,\, \mu_0,\, q_0,\, s)} \right] = 0,$$

where $f_0(y,\, \beta,\, \mu,\, q,\, s) = p(s,\, q)(\prod_{i=1}^{m} [\mathrm{e}^{\mu s_i}]^{\delta_{i1}} \exp[-y \mathrm{e}^{\beta^{\mathrm{T}} Z_{i1} + \mu s_i}])$. Using Assumption II, we know that $h_1(t^*) = \int_0^{t^*} h_1 \, \mathrm{d}\Lambda_0 = h_3 = h_4 = 0$.

Putting $h_1(t^*) = \int_0^{t^*} h_1 \, \mathrm{d}\Lambda_0 = h_2 = h_3 = h_4 = 0$ and considering $\delta_{11} = 1$, $\delta_{i1} = 0$ for $i = 2, \ldots, m$, and $X_{i1}$ near $t^*$ from the right for $i = 2, \ldots, m$ in (4.5), we obtain $h_1(X_{11}) = B(X_{11},\, Z_{11},\, \ldots,\, Z_{m1}) \int_0^{X_{11}} h_1 \, \mathrm{d}\Lambda_0$ for almost every $(X_{11},\, Z_{11},\, \ldots,\, Z_{m1})$, where

$$B(X_{11}, Z_{11}, \ldots, Z_{m1})$$

$$= \frac{\sum_s p(s, q_0) \mathrm{e}^{2\mu_0 s_1 + \beta_0^{\mathrm{T}} Z_{11}} \exp\left[-\Lambda_0(X_{11}) \mathrm{e}^{\beta_0^{\mathrm{T}} Z_{11} + \mu_0 s_1}\right] \prod_{i=2}^m \exp\left[-\Lambda_0(t^*) \mathrm{e}^{\beta_0^{\mathrm{T}} Z_{i1} + \mu_0 s_i}\right]}{\sum_s p(s, q_0) \mathrm{e}^{\mu_0 s_1} \exp\left[-\Lambda_0(X_{11}) \mathrm{e}^{\beta_0^{\mathrm{T}} Z_{11} + \mu_0 s_1}\right] \prod_{i=2}^m \exp\left[-\Lambda_0(t^*) \mathrm{e}^{\beta_0^{\mathrm{T}} Z_{i1} + \mu_0 s_i}\right]}.$$

Let $z_{i1}$ be any point in the support of the distribution of $Z_{i1}$. Define

$$g(t) = B(t, z_{11}, \ldots, z_{m1}) \int_0^t h_1 \, \mathrm{d}\Lambda_0.$$

Then $g = h_1$ a.s. $[\Lambda_0]$. Since $h_1$ is in $BV[0, \tau]$, $\Lambda_0$ has a positive and bounded derivative on $[0, \tau]$, and $B(\cdot, z_{11}, \ldots, z_{m1})$ is continuous, we know from $g(\cdot) = B(\cdot, z_{11}, \ldots, z_{m1}) \int_0^{\cdot} h_1 \, \mathrm{d}\Lambda_0$ that $g$ is continuous. From this, we can use the mean-value theorem for Riemann–Stieltjes integrals (see, for example, Wheeden and Zygmund 1997, p. 28) to show that $\int_0^{\cdot} g \, \mathrm{d}\Lambda_0$ is a differentiable function and has derivative $g(\cdot)\lambda_0(\cdot)$. Since the derivative of $\int_0^{\cdot} g(s) \mathrm{d}\Lambda_0(s) / \exp(\int_0^{\cdot} B(s, z_{11}, \ldots, z_{m1}) \mathrm{d}\Lambda_0(s))$ is 0, we obtain $g(\cdot) = bB(\cdot, z_{11}, \ldots, z_{m1}) \mathrm{e}^{\int_0^{\cdot} B(s, z_{11}, \ldots, z_{m1}) \mathrm{d}\Lambda_0(s)}$ for some constant $b$. By $g(0) = 0$, we obtain $g = 0$ identically and hence $h_1 = 0$ a.s. $[\Lambda_0]$. Putting $h_1 = 0$ a.s. $[\Lambda_0]$ and $h_2 = h_3 = h_4 = 0$ in $\sigma_1(\mathbf{h}) = 0$, we obtain from (4.6) below that $h_1(u) W(\theta_0; u) = 0$ for $u \in [0, \tau]$. Since $W(\theta_0, \cdot)$ is uniformly bounded away from 0 on $[0, \tau]$, $h_1 = 0$ identically. This completes the proof. $\square$

**Lemma 4.6.** $\sigma$ *is continuously invertible.*

***Proof.*** Define $A : H_\infty \to H_\infty$ by $A(h_1, h_2, h_3, h_4) = (h_1(\cdot) W(\theta_0; \cdot), h_2, h_3, h_4)$. Since $W(\theta_0, \cdot)$ is uniformly bounded away from 0 on $[0, \tau]$, $A$ is continuous, linear, and invertible.

Let $\mathcal{K} = \sigma - A$. Hence $\sigma = A(I + A^{-1}\mathcal{K})$. It is sufficient to show that $I + A^{-1}\mathcal{K}$ is continuously invertible. According to Lemma 4.5, $I + A^{-1}\mathcal{K}$ is one to one. This, together with Theorem 4.25 in Rudin (1973), implies that $I + A^{-1}\mathcal{K}$ is invertible if $A^{-1}\mathcal{K}$ is compact. For this, we need to show that $\mathcal{K}$ is compact.

Let $\mathcal{K} = (\mathcal{K}_1, \mathcal{K}_2, \mathcal{K}_3, \mathcal{K}_4)$. We only show $\mathcal{K}_1$ is compact, since the compactness of $\mathcal{K}_2$, $\mathcal{K}_3$, and $\mathcal{K}_4$ can be shown similarly. Since a bounded linear operator with finite-dimensional range is compact, we need only show that $\mathcal{K}_1$ is compact on $\{(h_1, 0, 0, 0) | h_1 \in BV[0, \tau]\}$.

Observing from (4.3) that

$$-\int_0^\tau \sigma_1(h_1, 0, 0, 0) \mathrm{d}(\Lambda - \Lambda_0) = \dot{\Psi}_{\theta_0}(\Lambda - \Lambda_0, 0, 0, 0)(h_1, 0, 0, 0)$$

$$= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\bigg|_{\varepsilon=0} \Psi(\Lambda_0 + \varepsilon(\Lambda - \Lambda_0), \beta_0, \mu_0, q_0)(h_1, 0, 0, 0),$$

we can show that

$$
\begin{aligned}
&\sigma_1(h_1,\,0,\,0,\,0)(u) \qquad (4.6)\\
&\quad = h_1(u)W(\theta_0;\,u)\\
&\qquad - \mathrm{E}\left\{\frac{\sum_s\left[f_1(\theta_0,\,s)\sum_{i=1}^m \mathrm{e}^{\beta_0^{\mathrm{T}}Z_{i1}+\mu_0 s_i}I_{[X_{i1}>u]}\sum_{i=1}^m \mathrm{e}^{\beta_0^{\mathrm{T}}Z_{i1}+\mu_0 s_i}\int_0^{X_{i1}} h_1\,\mathrm{d}\Lambda_0\right]}{\sum_s f_1(\theta_0,\,s)}\right.\\
&\qquad \left. - \frac{\left(\sum_s\left[f_1(\theta_0,\,s)\sum_{i=1}^m \mathrm{e}^{\beta_0^{\mathrm{T}}Z_{i1}+\mu_0 s_i}I_{[X_{i1}>u]}\right]\right)\left(\sum_s\left[f_1(\theta_0,\,s)\sum_{i=1}^m \mathrm{e}^{\beta_0^{\mathrm{T}}Z_{i1}+\mu_0 s_i}\int_0^{X_{i1}} h_1\,\mathrm{d}\Lambda_0\right]\right)}{\left(\sum_s f_1(\theta_0,\,s)\right)^2}\right\}.
\end{aligned}
$$

Hence

$$
\begin{aligned}
&\mathcal{K}_1(h_1,\,0,\,0,\,0)(u)\\
&\quad = -\mathrm{E}\left\{\frac{\sum_s\left[f_1(\theta_0,\,s)\left(\sum_{i=1}^m \mathrm{e}^{\beta_0^{\mathrm{T}}Z_{i1}+\mu_0 s_i}I_{[X_{i1}>u]}\right)\left(\sum_{i=1}^m \mathrm{e}^{\beta_0^{\mathrm{T}}Z_{i1}+\mu_0 s_i}\int_0^{X_{i1}} h_1\,\mathrm{d}\Lambda_0\right)\right]}{\sum_s f_1(\theta_0,\,s)}\right.\\
&\qquad \left. - \frac{\left(\sum_s\left[f_1(\theta_0,\,s)\sum_{i=1}^m \mathrm{e}^{\beta_0^{\mathrm{T}}Z_{i1}+\mu_0 s_i}I_{[X_{i1}>u]}\right]\right)\left(\sum_s\left[f_1(\theta_0,\,s)\sum_{i=1}^m \mathrm{e}^{\beta_0^{\mathrm{T}}Z_{i1}+\mu_0 s_i}\int_0^{X_{i1}} h_1\,\mathrm{d}\Lambda_0\right]\right)}{\left(\sum_s f_1(\theta_0,\,s)\right)^2}\right\},
\end{aligned}
$$

which immediately implies that

$$
\|\mathcal{K}_1(h_1,\,0,\,0,\,0)\|_V \leqslant c\int_0^\tau \|h_1\|_V\,\mathrm{d}\Lambda_0(t), \tag{4.7}
$$

for every $h_1 \in BV[0,\,\tau]$ and for some constant $c$. This shows $\mathcal{K}_1$ is compact on $\{(h_1,\,0,\,0,\,0)|h_1 \in BV[0,\,\tau]\}$ by Helly's lemma. This completes the proof. □

***Remark 3.*** $\sigma$ is called the information operator. It has a very complicated form in our calculations. We give a complete proof of the asymptotic normality of $\hat{\theta}_K$, without presenting an explicit formula for $\sigma$.

**Theorem 4.1.** $\sqrt{K}((\hat{\Lambda}_K,\,\hat{\beta}_K,\,\hat{\mu}_K,\,\hat{q}_K) - (\Lambda_0,\,\beta_0,\,\mu_0,\,q_0))$ *converges weakly to a tight Gaussian process* $\mathcal{G} \equiv -\dot{\Psi}^{-1}_{(\Lambda_0,\beta_0,\mu_0,q_0)}\mathcal{W}$ *on* $\ell^\infty(H_p)$ *with mean zero and covariance process*

$$\operatorname{cov}\big(\mathcal{G}(\mathbf{h}), \mathcal{G}(\tilde{\mathbf{h}})\big) = \int_0^\tau h_1 \tilde{\sigma}_1(\tilde{\mathbf{h}}) \mathrm{d}\Lambda_0 + h_2^{\mathrm{T}} \tilde{\sigma}_2(\tilde{\mathbf{h}}) + h_3 \tilde{\sigma}_3(\tilde{\mathbf{h}}) + h_4 \tilde{\sigma}_4(\tilde{\mathbf{h}}), \tag{4.8}$$

*where* $(\tilde{\sigma}_1, \tilde{\sigma}_2, \tilde{\sigma}_3, \tilde{\sigma}_4) = \tilde{\sigma} : H_\infty \mapsto H_\infty$ *is the inverse of* $\sigma = (\sigma_1, \sigma_2, \sigma_3, \sigma_4)$ *given in* (4.2).

***Proof.*** Since Lemmas 4.1–4.4 combined indicate that the conditions in the Theorem 3.3.1 and Lemma 3.3.5 in van der Vaart and Wellner (1996) are satisfied, we obtain the weak convergence of

$$\sqrt{K}\Big((\hat{\Lambda}_K, \hat{\beta}_K, \hat{\mu}_K, \hat{q}_K) - (\Lambda_0, \beta_0, \mu_0, q_0)\Big).$$

We will now calculate its asymptotic variance. It follows from $\Psi(\theta_0) = 0$, Lemma 4.1, and (4.4) that $\sqrt{K}\Psi_K(\Lambda_0, \beta_0, \mu_0, q_0)$ converges weakly to a tight Gaussian process $\mathcal{W}$ in $\ell^\infty(H_p)$ with

$$\operatorname{var}(\mathcal{W}(\mathbf{h})) = \int_0^\tau \sigma_1(\mathbf{h})(u) h_1(u) \mathrm{d}\Lambda_0(u) + h_2^{\mathrm{T}} \sigma_2(\mathbf{h}) + h_3 \sigma_3(\mathbf{h}) + h_4 \sigma_4(\mathbf{h}). \tag{4.9}$$

Let $\theta_0 = (\Lambda_0, \beta_0, \mu_0, q_0)$. It follows from (4.3) and Theorem 3.3.1 of van der Vaart and Wellner (1996) that

$$\begin{aligned}
&\int_0^\tau \sigma_1(\mathbf{h}) \mathrm{d}\Big(\sqrt{K}(\hat{\Lambda}_K - \Lambda_0)\Big) + \sigma_2^{\mathrm{T}}(\mathbf{h})\Big(\sqrt{K}(\hat{\beta}_K - \beta_0)\Big) \\
&\qquad + \sigma_3(\mathbf{h})\Big(\sqrt{K}(\hat{\mu}_K - \mu_0)\Big) + \sigma_4(\mathbf{h})\Big(\sqrt{K}(\hat{q}_K - q_0)\Big) \\
&\qquad = -\dot{\Psi}_{\theta_0}\Big(\sqrt{K}\Big((\hat{\Lambda}_K, \hat{\beta}_K, \hat{\mu}_K, \hat{q}_K) - (\Lambda_0, \beta_0, \mu_0, q_0)\Big)\Big)(\mathbf{h}) \\
&\qquad = \sqrt{K}(\Psi_K(\Lambda_0, \beta_0, \mu_0, q_0) - \Psi(\Lambda_0, \beta_0, \mu_0, q_0))(\mathbf{h}) + o_{p^*}(1).
\end{aligned} \tag{4.10}$$

Setting $\mathbf{g} = \sigma(\mathbf{h})$ in (4.10) and using (4.9), we know that

$$\int_0^\tau g_1 \mathrm{d}\Big(\sqrt{K}(\hat{\Lambda}_K - \Lambda_0)\Big) + g_2^{\mathrm{T}}\Big(\sqrt{K}(\hat{\beta}_K - \beta_0)\Big) + g_3\Big(\sqrt{K}(\hat{\mu}_K - \mu_0)\Big) + g_4\Big(\sqrt{K}(\hat{q}_K - q_0)\Big)$$

is asymptotically normal with mean 0 and variance

$$\int_0^\tau g_1 \tilde{\sigma}_1(\mathbf{g}) \mathrm{d}\Lambda_0 + g_2^{\mathrm{T}} \tilde{\sigma}_2(\mathbf{g}) + g_3 \tilde{\sigma}_3(\mathbf{g}) + g_4 \tilde{\sigma}_4(\mathbf{g}). \tag{4.11}$$

Using (4.11), we immediately obtain (4.8). □

# 5. Observed profile information

In this section we focus our attention on the estimation of $\beta$, $\mu$, and $q$, and present the efficient score function, the efficient Fisher information, and the asymptotic normality of

$(\hat{\beta}_K, \hat{\mu}_K, \hat{q}_K)$. We also apply the theory of observed profile information, developed in Murphy and van der Vaart (1999), to calculate the asymptotic variance.

For $\mathbf{h} \in H_\infty$, let $\tilde{\sigma}_{234}(\mathbf{h}) = (\tilde{\sigma}_2(\mathbf{h})^{\mathrm{T}}, \tilde{\sigma}_3(\mathbf{h}), \tilde{\sigma}_4(\mathbf{h}))^{\mathrm{T}}$. Let $e_i$ be the $(D+2)$-dimensional row vector with a 1 in the $i$th position and zeros elsewhere, for every $i = 1, \ldots, D+2$. Define the $(D+2) \times (D+2)$ matrix $\Sigma$ by $\Sigma^{-1} = (\tilde{\sigma}_{234}(0, \mathrm{e}_1), \ldots, \tilde{\sigma}_{234}(0, \mathrm{e}_{D+2}))^{\mathrm{T}}$. We note that $\Sigma$ is positive definite and symmetric.

We define

$$\ell_{234,\theta}[(h_2^{\mathrm{T}}, h_3, h_4)](X_1, \delta_1, Z_1) = \ell_{2,\theta}[h_2](X_1, \delta_1, Z_1) + \ell_{3,\theta}[h_3](X_1, \delta_1, Z_1) + \ell_{4,\theta}[h_4](X_1, \delta_1, Z_1),$$

for $h_2 \in \mathfrak{R}^D$, $h_3 \in \mathfrak{R}$, and $h_4 \in \mathfrak{R}$. Viewing $(h_2^{\mathrm{T}}, h_3, h_4)$ as a $(D+2)$-dimensional row vector, we can consider $\ell_{234,\theta}[\cdot](X_1, \delta_1, Z_1)$ as a $(D+2)$-dimensional column vector; $\ell_{234,\theta}[\cdot](X_1, \delta_1, Z_1)$ will be abbreviated as $\ell_{234,\theta}$.

We also define

$$g^* = -\Sigma \begin{pmatrix} \tilde{\sigma}_1(0, \mathrm{e}_1) \\ \vdots \\ \tilde{\sigma}_1(0, \mathrm{e}_{D+2}) \end{pmatrix}.$$

Then we have the following lemmas concerning the efficient score function and the efficient Fisher information. These concepts for semiparametric models were studied by Bickel *et al.* (1993) and van der Vaart (1998), among others.

**Lemma 5.1.** *The efficient score function for the estimation of* $(\beta, \mu, q)$ *is*

$$\tilde{\ell}_0 = \ell_{234,\theta_0} - \ell_{1,\theta_0}[g^*].$$

***Proof.*** We need to show that $\tilde{\ell}_0$ is orthogonal to $\ell_{1,\theta_0}[g_1]$ for every $g_1 \in BV[0, \tau]$. We observe that

$$\begin{aligned}
&e_i\Sigma^{-1}\mathrm{E}\,\tilde{\ell}_0\,\ell_{1,\theta_0}[g_1](X_1, \delta_1, Z_1) \\
&\quad = \mathrm{E}\big(\ell_{234,\theta_0}[e_i\Sigma^{-1}](X_1, \delta_1, Z_1) - \ell_{1,\theta_0}[e_i\Sigma^{-1}g^*](X_1, \delta_1, Z_1)\big)\ell_{1,\theta_0}[g_1](X_1, \delta_1, Z_1) \\
&\quad = \mathrm{E}(\ell_{2,\theta_0}[\tilde{\sigma}_2(0, e_i)](X_1, \delta_1, Z_1) + \ell_{3,\theta_0}[\tilde{\sigma}_3(0, e_i)](X_1, \delta_1, Z_1) \\
&\qquad + \ell_{4,\theta_0}[\tilde{\sigma}_4(0, e_i)](X_1, \delta_1, Z_1) + \ell_{1,\theta_0}[\tilde{\sigma}_1(0, e_i)](X_1, \delta_1, Z_1))\ell_{1,\theta_0}[g_1](X_1, \delta_1, Z_1) \\
&\quad = \int \sigma_1(\tilde{\sigma}(0, e_i))\,g_1\,\mathrm{d}\Lambda_0 \\
&\quad = 0,
\end{aligned}$$

by (4.4). This completes the proof. □

**Lemma 5.2.** $\Sigma = \mathrm{E}\,\tilde{\ell}_0\,\tilde{\ell}_0{}^{\mathrm{T}}$.

We omit the proof of Lemma 5.2 because it is a straightforward application of (4.4). Setting $\mathbf{h} = \tilde{\sigma}(0, e_i)$ in (4.10), we obtain

$$e_i\sqrt{K}\begin{pmatrix}\hat{\beta}_K - \beta_0\\ \hat{\mu}_K - \mu_0\\ \hat{q}_K - q_0\end{pmatrix} = \sqrt{K}(\mathbb{P}_K - P_0)\left(\sum_{j=1}^{4}\ell_{j,\theta_0}[\tilde{\sigma}_j(0, e_i)]\right) + o_{P^*}(1)$$

$$= \sqrt{K}(\mathbb{P}_K - P_0)\big(e_i\Sigma^{-1}\tilde{\ell}_0\big) + o_{P^*}(1),$$

which means

$$\sqrt{K}\begin{pmatrix}\hat{\beta}_K - \beta_0\\ \hat{\mu}_K - \mu_0\\ \hat{q}_K - q_0\end{pmatrix} = \Sigma^{-1}\sqrt{K}(\mathbb{P}_K - P_0)\tilde{\ell}_0 + o_{P^*}(1). \tag{5.1}$$

It follows from (5.1) that $\sqrt{K}((\hat{\beta}_K - \beta_0)^{\mathrm{T}}, \hat{\mu}_K - \mu_0, \hat{q}_K - q_0)^{\mathrm{T}}$ has asymptotic variance $\Sigma^{-1}$. $\Sigma$ is called the efficient Fisher information matrix.

Let

$$M_K\big((\beta^{\mathrm{T}}, \mu, q)^{\mathrm{T}}\big) = \sup_{\Lambda\in\mathcal{L}_M}\frac{1}{K}\log L_K(\Lambda, \beta, \mu, q),$$

which is called the profile likelihood function for estimating $(\beta, \mu, q)$. We recall that $\mathcal{L}_M$ was given in Lemma 3.1. Then $\Sigma$ can be approximated by using the following theorem.

**Theorem 5.1.** *Let $\gamma_1, \gamma_2, \ldots$ be a sequence of random variables such that $\gamma_K$ converges to 0 in probability and $(\sqrt{K}\gamma_K)^{-1} = O_p(1)$. Then*

$$-2\frac{M_K\Big((\hat{\beta}_K^{\mathrm{T}}, \hat{\mu}_K, \hat{q}_K)^{\mathrm{T}} + \gamma_K v_K\Big) - M_K\Big((\hat{\beta}_K^{\mathrm{T}}, \hat{\mu}_K, \hat{q}_K)^{\mathrm{T}}\Big)}{\gamma_K^2} \tag{5.2}$$

*converges in probability to $v^{\mathrm{T}}\Sigma v$, for every sequence $v_K$ in $\Re^{D+2}$ converging in probability to $v$.*

*Using* (5.2)*, we can show that*

$$-\Big[M_K\Big((\hat{\beta}_K^{\mathrm{T}}, \hat{\mu}_K, \hat{q}_K)^{\mathrm{T}} + \gamma_K e_i + \gamma_K \mathrm{e}_j\Big) - M_K\Big((\hat{\beta}_K^{\mathrm{T}}, \hat{\mu}_K, \hat{q}_K)^{\mathrm{T}} + \gamma_K e_i\Big)$$

$$- M_K\Big((\hat{\beta}_K^{\mathrm{T}}, \hat{\mu}_K, \hat{q}_K)^{\mathrm{T}} + \gamma_K \mathrm{e}_j\Big) + M_K\Big((\hat{\beta}_K^{\mathrm{T}}, \hat{\mu}_K, \hat{q}_K)^{\mathrm{T}}\Big)\Big]\Big/\gamma_K^2$$

*converges in probability to the* $(i, j)$*th entry of* $\Sigma$.

The proof of Theorem 5.1 is essentially a verification of the Theorem 2.1 in Murphy and van der Vaart (1999) concerning submodels. Because it is tedious and makes use of arguments appearing earlier already in this paper, we omit it and refer readers to Chang *et al*. (2004a) for details.

# 6. Discussion

We have established the consistency and the asymptotic normality of the NPMLE for the Cox gene model. In particular, we have obtained a consistent estimate of the asymptotic variance for the estimates of the finite-dimensional parameter of interest. These provide justifications for the approach proposed by Li *et al*. (1998). We have also provided a method to study the identifiability of the parameters and shown that, for a suitable choice of the parameter space, these parameters are identifiable for pedigrees consisting of three siblings.

Although we have only looked at the dominant model, our approach can be extended to study recessive or more general models, where different genetic risks are associated with different genotypes *aa*, *aA*, and *AA*.

It is desirable to propose numerical methods for NPMLE based on the theory in this paper. In fact, Chang *et al*. (2004b) makes use of the integral equations in Lemma 2.1 and the self-consistency equations derived from the score functions for $\beta$, $\mu$, and $q$ to approximate the estimates $\hat{\Lambda}_K$, $\hat{\beta}_K$, $\hat{\mu}_K$, and $\hat{q}_K$ respectively, and utilizes Theorem 5.1 on profile likelihood to approximate the variances of $\hat{\beta}_K$, $\hat{\mu}_K$, and $\hat{q}_K$. Simulations in Chang *et al*. (2004b) indicate that these numerical methods are quite satisfactory in terms of both speed and accuracy.

As we mentioned in Introduction, it would be interesting to extend the present theory to the situation where covariate values of the subjects are chosen by the experimenter and the sequence $(\mathrm{T}_k, C_k, S_k)$ is assumed independent.

In addition, it is desirable to study the hypothesis that there is some gene effect on the age at onset of a given disease. Because the null hypothesis that there is no gene effect may be represented by $\mu = 0$ or $q = 0$, which is not in the parameter space studied in this paper, further study of the likelihood is needed.

As pointed out in Li *et al*. (1998), an important extension of the Cox gene model would incorporate the data on a known or putatively linked marker. This allows the possibility of combined segregation and linkage analysis (Guo and Thompson 1992). We are keenly interested in developing an asymptotic theory for these extensions, and the relevant numerical methods.

# Appendix

## A.1. Identifiability

**Proposition A.1.** *Assumption II implies Assumption I.*

***Proof.*** Assuming $\tilde{L}_{(1),\Lambda,\beta,\mu,q} = \tilde{L}_{(1),\Lambda_0,\beta_0,\mu_0,q_0}$ a.s. with $\Lambda$ being absolutely continuous relative to $\Lambda_0$, we know that

$$\prod_{i=1}^{m}\left(\frac{\mathrm{d}\Lambda}{\mathrm{d}\Lambda_0}(X_{i1})\right)^{\delta_{i1}} = \frac{\sum_s p(s,\, q_0)\left(\prod_{i=1}^{m}\left[\mathrm{e}^{\beta_0^{\mathrm{T}} Z_{i1}+\mu_0 s_i}\right]^{\delta_{i1}} \exp\left[-\Lambda_0(X_{i1})\mathrm{e}^{\beta_0^{\mathrm{T}} Z_{i1}+\mu_0 s_i}\right]\right)}{\sum_s p(s,\, q)\left(\prod_{i=1}^{m}\left[\mathrm{e}^{\beta^{\mathrm{T}} Z_{i1}+\mu s_i}\right]^{\delta_{i1}} \exp\left[-\Lambda(X_{i1})\mathrm{e}^{\beta^{\mathrm{T}} Z_{i1}+\mu s_i}\right]\right)} \quad \text{a.s.} \quad (\text{A}.1)$$

Let $z_{i1}$ be any point in the support of the distribution $Z_{i1}$. Let

$$\Omega_n = \begin{bmatrix} \delta_{11}=1,\, \delta_{21}=\ldots=\delta_{m1}=0;\ Z_{i1}\in\left(z_{i1}-\frac{1}{c(n)},\, z_{i1}+\frac{1}{c(n)}\right) \text{ for } i=1,\,\ldots,\, m; \\ X_{11}\in\left[t^*,\, t^*+\frac{1}{c(n)}\right),\ X_{i1}\in\left(t^*-\frac{1}{c(n)},\, t^*+\frac{1}{c(n)}\right) \text{ for } i=2,\,\ldots,\, m \end{bmatrix},$$

where $c(n)$ is a sequence tending to infinity. If $m=1$, then $\Omega_n$ is defined to be

$$\left[\delta_{11}=1,\, X_{11}\in\left[t^*,\, t^*+\frac{1}{c(n)}\right),\, Z_{11}\in\left(z_{11}-\frac{1}{c(n)},\, z_{11}+\frac{1}{c(n)}\right)\right].$$

Since $t^*$ is in the support of the conditional distribution of $C_{i1}$ given $Z_{i1}$ and $\Lambda_0$ has a positive derivative on $[0,\, \tau]$, we know from the conditional independence of $T_{11},\, \ldots,\, T_{m1},\, C_{11},\, \ldots,\, C_{m1}$ given $Z_{11},\, \ldots,\, Z_{m1},\, S_{11},\, \ldots,\, S_{m1}$ that $P(\Omega_n)>0$, and the closure of $\{X_{11}(\omega)|\omega\in\Omega_n\}$ contains $[t^*,\, t^*+1/c(n))$ for large positive integers $n$. Using these, we can conclude from (A.1) that $\lim_{t\downarrow t^*}\mathrm{d}\Lambda(t)/\mathrm{d}\Lambda_0$ exists. We denote this limit by $y^*$.

Without loss of generality, we assume $P(C_{11}\geqslant\tau|Z_{11},\, \ldots,\, Z_{m1})>0$. Here the conditional probability refers to the one under the true model parameter. If $m\geqslant 2$, we can use the same arguments in the previous paragraph to consider

$$\Omega_n' = \begin{bmatrix} \delta_{11}=1,\, \delta_{21}=\ldots=\delta_{m1}=0;\ X_{i1}\in\left(t^*-\frac{1}{c(n)},\, t^*+\frac{1}{c(n)}\right) \text{ for } i=2,\,\ldots,\, m; \\ Z_{i1}\in\left(z_{i1}-\frac{1}{c(n)},\, z_{i1}+\frac{1}{c(n)}\right) \text{ for } i=1,\,\ldots,\, m \end{bmatrix},$$

and conclude from (A.1) that

$$\sum_s p(s,\, q)\left[\frac{\mathrm{d}\Lambda}{\mathrm{d}\Lambda_0}(X_{11})\mathrm{e}^{\beta^{\mathrm{T}} z_{11}+\mu s_1}\exp[-\Lambda(X_{11})\mathrm{e}^{\beta^{\mathrm{T}} z_{11}+\mu s_1}]\right]\left[\prod_{i=2}^{m}\exp[-\Lambda(t^*)\mathrm{e}^{\beta^{\mathrm{T}} z_{i1}+\mu s_i}]\right]$$
$$= \sum_s p(s,\, q_0)\left[\mathrm{e}^{\beta_0^{\mathrm{T}} z_{11}+\mu_0 s_1}\exp[-\Lambda_0(X_{11})\mathrm{e}^{\beta_0^{\mathrm{T}} z_{11}+\mu_0 s_1}]\right]\left[\prod_{i=2}^{m}\exp[-\Lambda_0(t^*)\mathrm{e}^{\beta_0^{\mathrm{T}} z_{i1}+\mu_0 s_i}]\right], \quad (\text{A}.2)$$

for almost every $X_{11}\in[0,\, \tau]$. If $m=1$, then (A.2) becomes

$$\sum_s p(s,\, q)\left[\frac{\mathrm{d}\Lambda}{\mathrm{d}\Lambda_0}(X_{11})\mathrm{e}^{\beta^{\mathrm{T}} z_{11}+\mu s_1}\exp[-\Lambda(X_{11})\mathrm{e}^{\beta^{\mathrm{T}} z_{11}+\mu s_1}]\right]$$
$$= \sum_s p(s,\, q_0)\left[\mathrm{e}^{\beta_0^{\mathrm{T}} z_{11}+\mu_0 s_1}\exp[-\Lambda_0(X_{11})\mathrm{e}^{\beta_0^{\mathrm{T}} z_{11}+\mu_0 s_1}]\right].$$

Making use of differential calculus, we know from (A.2) that it is sufficient to show that $(y^*, \Lambda(t^*), \beta, \mu, q) = (1, \Lambda_0(t^*), \beta_0, \mu_0, q_0)$ to establish the identifiability.

Considering $X_{i1}$ near 0 from the right for $i = 1, \ldots, m$ in (A.1), we can use the same arguments in the lines following (A.1) to show $\lim_{t\downarrow 0} \mathrm{d}\Lambda(t)/\mathrm{d}\Lambda_0$ exists and

$$\prod_{i=1}^{m} \mathrm{e}^{(\beta_0-\beta)^{\mathrm{T}} Z_{i1}} = \frac{\tilde{y}^m \sum_s p(s, q) \prod_{i=1}^{m} \mathrm{e}^{\mu s_i}}{\sum_s p(s, q_0) \prod_{i=1}^{m} \mathrm{e}^{\mu_0 s_i}}, \tag{A.3}$$

where $\tilde{y} = \lim_{t\downarrow 0} \mathrm{d}\Lambda(t)/\mathrm{d}\Lambda_0$. Since the linear span of $\sum_{i=1}^m Z_{i1}$ has dimension $D$, we know from (A.3) that $\beta = \beta_0$.

Setting $\beta = \beta_0$ in (A.1) and considering $X_{i1}$ near $t^*$ from the right for $i = 1, \ldots, m$, we obtain

$$\log \sum_s p(s, q)\left(\prod_{i=1}^{m} \left[y^* \mathrm{e}^{\mu s_i}\right]^{\delta_{i1}} \exp\left[-\Lambda(t^*)\mathrm{e}^{\beta_0^{\mathrm{T}} Z_{i1}+\mu s_i}\right]\right)$$

$$= \log \sum_s p(s, q_0)\left(\prod_{i=1}^{m} [\mathrm{e}^{\mu_0 s_i}]^{\delta_{i1}} \exp\left[-\Lambda_0(t^*)\mathrm{e}^{\beta_0^{\mathrm{T}} Z_{i1}+\mu_0 s_i}\right]\right). \tag{A.4}$$

Since each $\delta_{i1}$ and $Z_{i1}$ in (A.4) can take many different values, (A.4) defines many equations. Noting that the left-hand side of (A.4) is a function of $(y^*, \Lambda(t^*), \mu, q)$, we can apply the inverse function theorem (see, for example, Rudin 1976, p. 221) to check whether $(y^*, \Lambda(t^*), \mu, q) = (1, \Lambda_0(t^*), \mu_0, q_0)$ is the unique solution in a neighbourhood of $(1, \Lambda_0(t^*), \mu_0, q_0)$ in $\mathfrak{R}^4$. This amounts to choosing four of these equations and checking if the determinant of the Jacobian of the function in question is zero. This is guaranteed by Assumption II. This completes the proof. □

**Proposition A.2.** *If random mating and Mendelian segregation are assumed, then there exist parameter values so that Assumption II is valid for families consisting of three siblings; if* $\beta_0 \neq 0$ *and* $t^* \in (0, \tau]$, *it is also valid when there is only one member in each family.*

***Proof.*** Assume that the pedigree consists of three siblings. Fixing $Z_{11}$, $Z_{21}$, and $Z_{31}$ and considering $(\delta_{11}, \delta_{21}, \delta_{31}) = (0, 0, 0)$, $(1, 0, 0)$, $(1, 1, 0)$, $(1, 1, 1)$ in the mapping

$$(y_0, y_1, \mu, q) \mapsto \log \sum_s p(s, q)\left(\prod_{i=1}^{3} [y_0 \mathrm{e}^{\mu s_i}]^{\delta_{i1}} \exp\left[-y_1 \mathrm{e}^{\beta_0^{\mathrm{T}} Z_{i1}+\mu s_i}\right]\right),$$

we obtain a function $F$ from $\mathfrak{R}^4$ to $\mathfrak{R}^4$. A straightforward calculation using the assumption of random mating and Mendelian segregation shows that $p(s, q)$ is a fourth-order polynomial in $q$. For example, $p(1, 0, 0, q) = \frac{1}{2}q(1-q)^3 + \frac{3}{16}q^2(1-q)^2$. It is easy to see that the determinant of the Jacobian of $F$ is a real analytic function, and that this determinant is not identically zero, because it is not zero for $y_0 = 1$, $y_1 = 1$, $q = \frac{1}{2}$, and an appropriate $\mu$.

We note that a real analytic function is identically zero on a connected region only if its value is zero on any open subset in it. This implies that the determinant can be zero only on a nowhere dense closed subset of $\mathfrak{R}^4$. Therefore, as long as the determinant is not zero at $(1, \Lambda_0(t^*), q_0, \mu_0)$, Assumption II is satisfied on a neighbourhood of it. This completes the proof for the case each family consisting of three siblings.

Assuming there is only one member in each family and considering the function $F$ specified by $(\delta_{11}, Z_{11}) = (0, z'), (0, z''), (1, z'), (1, z'')$ with $z' \neq z''$ in the range of $Z_{11}$, we can establish the statement using the same arguments for the case that each family consists of three siblings. This completes the proof. $\square$

## A.2. Proof of Lemma 3.1

Let

$$g(\Lambda, \beta, \mu, q, u, X_k, \delta_k, Z_k) = \sum_{i=1}^{m} \frac{\sum_s f(\Lambda, \beta, \mu, q, s, X_k, \delta_k, Z_k) \exp(\beta^{\mathrm{T}} Z_{ik} + \mu s_i)}{\sum_s f(\Lambda, \beta, \mu, q, s, X_k, \delta_k, Z_k)} I_{(0, X_{ik}]}(u),$$

where $f(\Lambda, \beta, \mu, q, s, X_k, \delta_k, Z_k) = f_k(\Lambda, \beta, \mu, q, s)$. Then

$$\begin{aligned} &W_K(\Lambda, \beta, \mu, q; u) - W(\Lambda, \beta, \mu, q; u) \\ &= \frac{1}{K} \sum_{k=1}^{K} (g(\Lambda, \beta, \mu, q, u, X_k, \delta_k, Z_k) - \mathrm{E}g(\Lambda, \beta, \mu, q, u, X_k, \delta_k, Z_k)) \\ &= \mathbb{P}_K\, g - \mathrm{E}g(X_1, \delta_1, Z_1). \end{aligned}$$

For every fixed $s$, we know that

$$\begin{aligned} &f(\Lambda, \beta, \mu, q, s, X_k, \delta_k, Z_k) \\ &\quad = \phi_s(\Lambda(X_{1k}), \beta^{\mathrm{T}} Z_{1k} + \mu s_1, (\exp(\mu s_1))^{\delta_{1k}}, \ldots, \Lambda(X_{mk}), \beta^{\mathrm{T}} Z_{mk} + \mu s_m, (\exp(\mu s_m))^{\delta_{mk}}), \end{aligned}$$

for some smooth function $\phi_s$. Using this and the fact that the set of functions mapping $Z_{ik}$ to $\beta^{\mathrm{T}} Z_{ik} + \mu s_i$, indexed by $\beta \in \mathcal{B}$ and $\mu \in \mathcal{U}$, is Donsker (van der Vaart and Wellner 1996, Theorem 2.7.1) and that the set of functions mapping $X_{ik}$ to $\Lambda(X_{ik})$, indexed by $\Lambda \in \mathcal{L}_M$, is also Donsker (van der Vaart and Wellner 1996, Example 2.10.4), we can conclude that the Donsker property is preserved by $\phi_s$ (van der Vaart and Wellner 1996, Theorem 2.10.6 and related results). Since there are only finitely many possible values of $s = (s_1, \ldots, s_m)$, we can apply the above arguments concerning permanence of the Donsker property to conclude that $\mathcal{G} = \{g(\Lambda, \beta, \mu, q, u, \cdot, \cdot, \cdot) | \Lambda \in \mathcal{L}_M, \beta \in \mathcal{B}, \mu \in \mathcal{U}, q \in \mathcal{Q}, u \in [0, \tau]\}$ is also a Donsker class. This implies that $\sqrt{K}(\mathbb{P}_K\, g - \mathrm{E}g(X_1, \delta_1, Z_1))$ converges weakly to a tight Borel measurable Gaussian element in $\ell^\infty(\mathcal{G})$, as $K$ goes to infinity. Thus $\mathbb{P}_K\, g - \mathrm{E}g(X_1, \delta_1, Z_1)$ converges weakly to zero as a random element in $\ell^\infty(\mathcal{G})$. This completes the proof.

## A.3. Proof of Lemma 3.2

It follows from the permanence of the Donsker property and the fact that the class of non-negative increasing functions with a common upper bound is Donsker that the class of functions $\sum_{i=1}^{m} I_{[T_{ik},\infty)}(u \wedge C_{ik}) = \sum_{i=1}^{m} I_{(0,u]}(X_{ik}) \cdot \delta_{ik}$, indexed by $u$, is Donsker and hence Glivenko–Cantelli (van der Vaart and Wellner 1996, Examples 2.10.4, 2.10.7, and 2.10.8). This shows that $\sup_{t\in[0,\tau]}|G_K(t) - G(t)|$ goes to zero almost surely. Using (3.4), the strong law of large numbers, and the condition $P(X_{ik} \geqslant \tau) > 0$, we have that $W_K(\Lambda, \beta, \mu, q; u)$ is bounded and bounded away from 0 on $[0, \tau]$ for all large $K$. Combining this, the uniform convergence of $G_K$, and the following integration by parts

$$\int_0^t \frac{1}{W_K(\Lambda, \beta, \mu, q; u)} \,\mathrm{d}(G_K(u) - G(u))$$

$$= \frac{G_K(u) - G(u)}{W_K(\Lambda, \beta, \mu, q; u)}\bigg|_0^t - \int_0^t (G_K(u) - G(u))\mathrm{d}\left(\frac{1}{W_K(\Lambda, \beta, \mu, q; u)}\right),$$

we immediately obtain the desired result.

## A.4. Three results due to van der Vaart and Wellner

We quote three important results used in this paper.

**Theorem A.1 (van der Vaart and Wellner 1996, Theorem 2.10.6).** *Let $\mathcal{F}_1, \ldots, \mathcal{F}_k$ be Donsker classes with $\|P\|_{\mathcal{F}_i} = \sup_{f_i\in\mathcal{F}_i}|Pf_i| < \infty$ for each i. Let $\phi : \mathfrak{R}^k \mapsto \mathfrak{R}$ satisfy*

$$|\phi \circ f(x) - \phi \circ g(x)|^2 \leqslant \sum_{l=1}^{k} (f_l(x) - g_l(x))^2,$$

*for every $f, g \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_k$ and x. Then the class $\phi \circ (\mathcal{F}_1, \ldots, \mathcal{F}_k)$ is Donsker, provided $\phi \circ (f_1, \ldots, f_k)$ is square integrable for at least one $(f_1, \ldots, f_k)$.*

The following theorem and lemma are applied in this paper with the same notation, apart from changing $n$ into $K$, and $P$ into E.

**Theorem A.2 (van der Vaart and Wellner 1996, Theorem 3.3.1).** *Let $\Psi_n$ and $\Psi$ be random maps and a fixed map, respectively, from $\Theta$ into a Banach space such that*

$$\sqrt{n}(\Psi_n - \Psi)(\hat{\theta}_n) - \sqrt{n}(\Psi_n - \Psi)(\theta_0) = o_{p^*}(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|),$$

*and such that the sequence $\sqrt{n}(\Psi_n - \Psi)(\theta_0)$ converges in distribution to a tight random element Z. Let $\theta \mapsto \Psi(\theta)$ be Fréchet differentiable at $\theta_0$ with a continuously invertible derivative $\dot{\Psi}_0$. If $\Psi(\theta_0) = 0$, $\hat{\theta}_n$ satisfies $\Psi_n(\hat{\theta}_n) = o_{p^*}(n^{-1/2})$, and converges in outer probability to $\theta_0$, then*

$$\sqrt{n}\,\dot{\Psi}_{\theta_0}(\hat{\theta}_n - \theta_0) = -\sqrt{n}(\Psi_n - \Psi)(\theta_0) + o_{p^*}(1).$$

*Consequently,* $\sqrt{n}\,(\hat{\theta}_n - \theta_0) \to -\dot{\Psi}_{\theta_0}^{-1} Z$.

In the case of independent and identically distributed observations, the theorem may be applied with $\Psi_n(\theta)\mathbf{h} = \mathbb{P}_n\,\phi_{\theta,\mathbf{h}}$ and $\Psi(\theta)\mathbf{h} = P\,\phi_{\theta,\mathbf{h}}$ for given measurable functions $\phi_{\theta,\mathbf{h}}$, indexed by $\Theta$ and arbitrary index set $\mathcal{H}$. In this case, $\sqrt{n}(\Psi_n - \Psi)(\theta) = \{G_n\,\phi_{\theta,\mathbf{h}} : \mathbf{h} \in \mathcal{H}\}$ is the empirical process indexed by the classes of functions $\{\phi_{\theta,\mathbf{h}} : \mathbf{h} \in \mathcal{H}\}$.

**Lemma A.1 (van der Vaart and Wellner 1996, Lemma 3.3.5).** *Suppose that the class of functions*

$$\{\phi_{\theta,\mathbf{h}} - \phi_{\theta_0,\mathbf{h}} : \|\theta - \theta_0\| < \delta,\ \mathbf{h} \in \mathcal{H}\}$$

*is P-Donsker for some* $\delta > 0$ *and*

$$\sup_{\mathbf{h}\in\mathcal{H}} P(\phi_{\theta,\mathbf{h}} - \phi_{\theta_0,\mathbf{h}})^2 \to 0, \text{ as } \theta \to \theta_0.$$

*If* $\hat{\theta}_n$ *converges in outer probability to* $\theta_0$*, then*

$$\|G_n(\phi_{\hat{\theta}_n,\mathbf{h}} - \phi_{\theta_0,\mathbf{h}})\|_{\mathcal{H}} = o_{p^*}(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|).$$

## Acknowledgements

This work is partially supported by NSC Grant 90-2118-M-400-001. We thank Yuh-Jenn Wu for helpful conversation during the preparation of this paper. We are grateful to an anonymous referee, whose valuable comments and suggestions led to significant improvement of this paper. We also thank both Professor van Zwet and Professor McCullagh for their suggestions regarding the presentation of the paper.

## References

Andersen, P.K., Borgan, Ø., Gill, R.D. and Keiding, N. (1993) *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

Bickel, P.J. and Ritov, Y. (2000) Discussion of S. Murphy and A. van der Vaart, 'On profile likelihood'. *J. Amer. Statist. Assoc.*, **95**, 466–468.

Bickel, P.J., Klaassen, C., Ritov, Y. and Wellner, J. (1993) *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore, MD: Johns Hopkins University Press.

Chang, I.-S., Hsiung, C.A., Wang, M.-C. and Wen, C.-C. (2004a) Nonparametric maximum likelihood estimator in the Cox-gene model. Technical Report DBB-2004-1, Division of Biostatistics and Bioinformatics, National Health Research Institutes, Taipei, Taiwan.

Chang, I.-S., Hsiung, C.A., Wen, C.-C., Wu, Y.-J. and Yang, C.-C. (2004b) Simulation studies of the nonparametric maximum likelihood estimate in the Cox-gene model with a fast algorithm. Preprint.

Elbers, C. and Ridder, G. (1982) True and spurious duration dependence: the identifiability of the proportional hazard model. *Rev. Econom. Stud.*, **49**, 403–409.

Guo, S.W. and Thompson, E.A. (1992) A Monte Carlo method for combined segregation and linkage analysis. *Amer. J. Hum. Genetics*, **51**, 1111–1126.

Heckman, J.J. and Singer, B. (1984) The identifiability of the proportional hazard model. *Rev. Econom. Stud.*, **51**, 231–241.

Heckman, J.J. and Taber, C.R. (1994) Econometric mixture models and more general models for unobservables in duration analysis. *Statist. Methods Medical Res.*, **3**, 279–299.

Johansen, S. (1983) An extension of Cox's regression model. *Internat. Statist. Rev.*, **51**, 165–174.

Kiefer, J. and Wolfowitz, J. (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.*, **27**, 887–906.

Kortram, R.A., van Rooij, A.C.M., Lenstra, A.J. and Ridder, G. (1995) Constructive identification of the mixed proportional hazards model. *Statist. Neerlandica*, **49**, 269–281.

Kosorok, M., Lee, B. and Fine, J. (2004) Robust inference for proportional hazards univariate frailty regression models. *Ann. Statist.*, **32**, 1448–1491.

Li, H. and Thompson, E.A. (1997) Semiparametric estimation of major gene and family-specific random effects for age of onset. *Biometrics*, **53**, 282–293.

Li, H., Thompson, E.A. and Wijsman, E.M. (1998) Semiparametric estimation of major gene effects for age of onset. *Genetic Epidemiology*, **15**, 279–298.

Mack, W., Langholz B. and Thomas, D.C. (1990) Survival models for familial aggregation of cancer. *Environ. Health Perspect.*, **87**, 27–35.

Murphy, S.A. (1994) Consistency in a proportional hazards model incorporating a random effect. *Ann. Statist.*, **22**, 712–731.

Murphy, S.A. (1995) Asymptotic theory for the frailty model. *Ann. Statist.*, **23**, 182–198.

Murphy, S.A. and van der Vaart, A.W. (1999) Observed information in semi-parametric models. *Bernoulli*, **5**, 381–412.

Murphy, S.A. and van der Vaart, A.W. (2000) On profile likelihood. *J. Amer. Statist. Assoc.*, **95**, 449–465.

Murphy, S.A., Rossini, A.J. and van der Vaart, A.W. (1997) Maximum likelihood estimation in the proportional odds model. *J. Amer. Statist. Assoc.*, **92**, 968–976.

Parner, E. (1998) Asymptotic theory for the correlated gamma-frailty model. *Ann. Statist.*, **26**, 183–214.

Rudin, W. (1973) *Functional Analysis*. New York: McGraw-Hill.

Rudin, W. (1976) *Principles of Mathematical Analysis*, 3rd edn, New York: McGraw-Hill.

Siegmund, K. and McKnight, B. (1998) Modelling hazard functions in families. *Genetic Epidemiology*, **15**, 147–171.

van de Geer, S. (2000) *Applications of Empirical Process Theory*. Cambridge: Cambridge University Press.

van der Vaart, A.W. (1994) Maximum likelihood estimation with partially censored observations. *Ann. Statist.*, **22**, 1896–1916.

van der Vaart, A.W. (1995) Efficiency of infinitely dimensional M-estimators. *Statist. Neerlandica*, **49**, 9–30.

van der Vaart, A.W. (1996) On a model of Hasminskii and Ibragimov. In I.A. Ibragimov and A. Yu. Zaitsev (eds), *Probability Theory and Mathematical Statistics*, pp. 297–308. Amsterdam: Gordon and Breach.

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

van der Vaart, A.W. and Wellner, J.A. (1996) *Weak Convergence and Empirical Processes*. New York: Springer-Verlag.

Wheeden, R.L. and Zygmund, A. (1997) *Measure and Integral: An Introduction to Real Analysis*. New York: Marcel Dekker.

Received February 2001 and revised January 2005