Statistical Generalized Derivative Applied to the Profile Likelihood Estimation in a Mixture of Semiparametric Models

Yuichi Hirose * and Ivy Liu

School of Mathematics and Statistics, Victoria University of Wellington, P.O. Box 600, 6140 Wellington, New Zealand; ivy.liu@vuw.ac.nz
* Correspondence: Yuichi.Hirose@vuw.ac.nz; Tel: +64-4463-6421

Received: 9 January 2020; Accepted: 25 February 2020; Published: 28 February 2020

Abstract: There is a difficulty in finding an estimate of the standard error (SE) of the profile likelihood estimator in the joint model of longitudinal and survival data. The difficulty is on the differentiation of an implicit function that appear in the profile likelihood estimation. We solve the difficulty by introducing the “statistical generalized derivative”. The derivative is used to show the asymptotic normality of the estimator with the SE expressed in terms of the profile likelihood score function.

Keywords: efficiency; efficient information bound; efficient score; implicitly defined function; profile likelihood; semi-parametric model; joint model; EM algorithm

1. Introduction

This paper proposes a method to show asymptotic normality of the maximum profile likelihood estimator (profile likelihood MLE) in a mixture of semiparametric models with the EM-algorithm. We derive an expression for the standard error (SE) of the estimator using the profile likelihood score function. As an example, we consider a joint model of ordinal responses and the proportional hazards model with finite mixture. Through this example, we demonstrate solving the theoretical challenge in a joint model of survival and longitudinal data stated by [1]: “No distributional or asymptotic theory is available to date, and even the standard errors (SEs), defined as the standard deviations of the parametric estimators, are difficult to obtain.” The difficulty of the problem is to deal with an implicit function which is difficult to differentiate. In the profile likelihood approach, we profile out the baseline hazard function by plugging in an estimate of the hazard function to the likelihood function. This estimator of the hazard function is an implicit function in our problem (see also [2], p67).

Here, we review some of the related works. For a more complete review of the joint models please see [2] and [3]. The paper [4] proposed a profile likelihood approach to a joint model with unobserved random effects. The EM-algorithm was applied for an estimation of parameters in the model. This approach became one of the standard models and has been adopted by many (for example, [5–7]). In a series of studies, ([8–11]), also adopted a profile likelihood approach in [4] and showed the asymptotic normality of the profile likelihood MLE. As a result, they obtained the SE of the estimator. Another well known approach was proposed by [12]. They used "an approximate least favorable submodel" to show the asymptotic normality of the profile likelihood MLE. The SE of the profile likelihood estimator was obtained from the asymptotic normality.

All of the existing work mentioned here avoided dealing with the implicit function in the profile likelihood. For example, the works of [8–11] used the equality between the maximum likelihood estimator (MLE) and the maximum profile likelihood estimator (profile likelihood MLE). They have shown the asymptotic normality of the MLE in the joint models and the result is used to show the asymptotic normality of the profile likelihood MLE. This indirect proof avoided the differentiation.
of the implicit function in the profile likelihood. In the case of [12], “an approximate least favorable submodel” was used to approximate the profile likelihood function. Then, this approximation was used to prove the result without differentiating the profile likelihood function directly.

In summary, the existing methods showed the asymptotic normality of the profile likelihood MLE using methods to avoid differentiating a profile likelihood function. They have shown that the variance of the profile likelihood MLE is the inverse of an efficient information matrix. However, they have not shown the efficient information matrix in terms of the score function in the profile likelihood.

In this paper, we introduce “the statistical generalized derivative” to deal with the differentiation of implicit function in the profile likelihood under consideration. Our approach enabled us to expand the profile likelihood function to show the asymptotic normality of the profile likelihood MLE with the efficient information matrix expressed as a variance of the profile likelihood score function.

The results of this paper give us an analytical understanding of profile likelihood estimation and a method of computing the SE of the profile likelihood MLE in terms of the profile likelihood score function.

2. Mixture of Semiparametric Models and Generalized Statistical Derivative

2.1. Introduction of Mixture Model and Notations

We consider a mixture of semiparametric models whose density is of the form

\[ p(x; \theta, \eta, \pi) = \sum_{r=1}^{R} \pi_r p_r(x; \theta_r, \eta_r), \]

where for each \( r = 1, \ldots, R \), \( p_r(x; \theta_r, \eta_r) \) is a semiparametric model with a finite dimensional parameter \( \theta_r \in \Theta_r \subset \mathbb{R}^{m_r} \) and an infinite dimensional parameter \( \eta_r \in H_r \) where \( H_r \) is a subset of Banach space \( B_r \), and \( \pi_1, \ldots, \pi_R \) are mixture probabilities. We assume that \( \pi_r > 0 \) for each \( r \) and \( \sum_{r=1}^{R} \pi_r = 1 \). We denote \( \theta = (\theta_1, \ldots, \theta_R), \eta = (\eta_1, \ldots, \eta_R) \) and \( \pi = (\pi_1, \ldots, \pi_R) \). Once we observe iid data \( X_1, \ldots, X_n \) from the mixture model, the joint probability function of the data \( X = (X_1, \ldots, X_n) \) is given by

\[ p(X; \theta, \eta, \pi) = \prod_{i=1}^{n} \prod_{r=1}^{R} \pi_r p_r(X_i; \theta_r, \eta_r). \]

We consider \( \theta \) is the parameters of interest, and \( \eta \) and \( \pi \) are nuisance parameters.

To discuss the EM-algorithm, we further introduce notations (we use notations from [13]). Let \( Z_i = (Z_{i1}, \ldots, Z_{iR}) \) be a group indicator variable for the subject \( i \): for each \( r \), \( Z_{ir} = 0 \) or \( 1 \) with \( P(Z_{ir} = 1) = \pi_r \), and \( \sum_{r=1}^{R} Z_{ir} = 1 \). Let \( Z = (Z_1, \ldots, Z_n) \). The joint probability function of the complete data \( (X, Z) \) is

\[ p(X, Z; \theta, \eta, \pi) = \prod_{i=1}^{n} \prod_{r=1}^{R} [\pi_r p_r(X_i; \theta_r, \eta_r)]^{Z_{ir}}. \]

Since \( Z_{ir} \) are unobserved, it is common to replace with its expected value. The expected complete data log-likelihood under \( p(Z; X, \theta, \eta, \pi) \) is

\[ \sum_{i=1}^{n} \sum_{r=1}^{R} E(Z_{ir})[\log \pi_r + \log p_r(X_i; \theta_r, \eta_r)] = \sum_{i=1}^{n} \sum_{r=1}^{R} \gamma(Z_{ir})[\log \pi_r + \log p_r(X_i; \theta_r, \eta_r)]. \]

where

\[ \gamma(Z_{ir}) = \frac{\pi_r p_r(X_i; \theta_r, \eta_r)}{\sum_{j=1}^{R} \pi_j p_j(X_i; \theta_j, \eta_j)}, \quad r = 1, \ldots, R. \]
2.2. Introducing Profile Likelihood, the Efficient Score Function and the Efficient Information Matrix

The efficient score function and information matrix in the mixture model: The score function for θ and score operator for η in the mixture model given in (1) are, respectively,

\[ \hat{\ell}(x; \theta, \eta) = \frac{\partial}{\partial \theta} \log \left( \sum_{r=1}^{R} \pi_r p_r(x; \theta_r, \eta_r) \right) = \sum_{r=1}^{R} \gamma_r(z_r) \frac{\partial}{\partial \theta} \log p_r(x; \theta_r, \eta_r), \]

(6)

and

\[ B(x; \theta, \eta) = d_\eta \log \left( \sum_{r=1}^{R} \pi_r p_r(x; \theta_r, \eta_r) \right) = \sum_{r=1}^{R} \gamma_r(x) d_\eta \log p_r(x; \theta_r, \eta_r) \]

(7)

where γ_r(z_r) is given in (5) with Z_{ir} replaced with z_r. The notation d_\eta is the Hadamard derivative operator with respect to the parameter η.

Let \( \theta_0, \eta_0 \) be the true values of θ, η and denote \( \hat{\ell}_0(x) = \hat{\ell}(x; \theta_0, \eta_0) \) and \( B_0(x) = B(x; \theta_0, \eta_0) \). Then, it follows from the standard theory ([14], page 374) that the efficient score function \( \hat{\ell}_0 \) and the efficient information matrix \( \hat{I}_0 \) in the semiparametric mixture model are given by

\[ \hat{\ell}_0(x) = (I - B_0(B_0^*)^{-1}B_0^*) \hat{\ell}_0(x), \]

(8)

and

\[ \hat{I}_0 = E[\hat{\ell}_0 \hat{\ell}_0^\top]. \]

(9)

Note: Equations (6) and (7) show that the score functions in the semiparametric mixture model (1) coincide with the ones for the expected complete data likelihood (4).

Introduction of the profile likelihood and its score functions: In the estimation of (θ, η), we use the profile likelihood approach. Let F be a cdf function which belongs to a set containing the empirical cdf \( F_\hat{\theta} \), and the true cdf \( F_0 \). In the profile likelihood approach, we obtain a function \( (\theta, F) \rightarrow \hat{\eta}_{\theta,F} = (\hat{\eta}_1, \ldots, \hat{\eta}_r, \ldots, \hat{\eta}_R) \) whose values are in the space of the parameter \( \eta = (\eta_1, \ldots, \eta_R) \). Then the profile likelihood function is defined by

\[ p(X; \theta, F, \pi) = \prod_{i=1}^{n} \sum_{r=1}^{R} \pi_r p_r(X_i; \theta_r, \hat{\eta}_{r,\theta,F}). \]

(10)

We also define the score functions for the profile likelihood in the model

\[ \phi(x; \theta, F) = \frac{\partial}{\partial \theta} \log \left( \sum_{r=1}^{R} \pi_r p_r(x; \theta_r, \eta_{r,\theta,F}) \right) = \sum_{r=1}^{R} \gamma_r(z_r) \frac{\partial}{\partial \theta} \log p_r(x; \theta_r, \hat{\eta}_{r,\theta,F}) \]

(11)

and

\[ \phi(x; \theta, F) = d_F \log \left( \sum_{r=1}^{R} \pi_r p_r(x; \theta_r, \eta_{r,\theta,F}) \right) = \sum_{r=1}^{R} \gamma_r(z_r) d_F \log p_r(x; \theta_r, \hat{\eta}_{r,\theta,F}), \]

(12)

where γ_r(z_r) is given in (5) with Z_{ir} replaced with z_r and η_r with \( \hat{\eta}_{r,\theta,F} \).

2.3. The EM-Algorithm to Obtain the Profile Likelihood MLE

Here, we describe the EM-algorithm applied to the profile likelihood function to obtain the profile likelihood MLE \( \hat{\theta} \) of θ:
The E-step: In the E-step, we use the current parameter estimates \( \hat{\theta} \) of \( \theta \) to find the expected values of \( Z_r \):

\[
\gamma(Z_r) = \frac{\pi_r p_r(X_i; \hat{\theta}, \hat{\eta}, \hat{\phi}, \hat{F}_n)}{\sum_{j=1}^R \pi_j p_j(X_i; \hat{\theta}, \hat{\eta}, \hat{\phi}, \hat{F}_n)}, \quad r = 1, \ldots, R.
\] (13)

The M-step: In the M-step,

1. Calculate the estimates of \( \pi_r \)

\[
\hat{\pi}_r = \frac{\sum_{i=1}^n \gamma(Z_i)}{n}.
\]

2. Keeping \( \gamma(Z_r) \) as a constant, we maximize the expected complete data log-likelihood

\[
\sum_{i=1}^n \sum_{r=1}^R \gamma(Z_r) \log p_r(X_i; \theta, \eta, \phi, F_n).
\] (14)

with respect to \( \theta \) to obtain new estimates \( \hat{\theta} \).

The estimated parameters from the M-step are returned into the E-step until the value of \( \hat{\theta} \) converges. The resulting estimator is the profile likelihood MLE.

2.4. Generalized Statistical Derivative and Asymptotic Normality of the Profile Likelihood MLE

In this section, we show the main result. We show the asymptotic normality of the profile likelihood MLE in Theorem 2. From the asymptotic normality, we can get an estimate of the SE of the estimator. Toward this goal, we introduce the statistical generalized derivative in Theorem 1.

Assumptions: We list assumptions used for Theorem 1 and Theorem 2 given below.

Let us denote \( \theta_0, \eta_0 \) and \( F_0 \) are the true values of the parameters \( \theta, \eta \) and cdf \( F \).

On the set of cdf functions \( \mathcal{F} \), we use the sup-norm, i.e., for \( F, F_0 \in \mathcal{F} \),

\[
\|F - F_0\| = \sup_x |F(x) - F_0(x)|.
\]

For \( \rho > 0 \), let

\[
\mathcal{C}_\rho = \{F \in \mathcal{F} : \|F - F_0\| < \rho\}.
\]

We denote the density function for the profile likelihood in the mixture model by

\[
p(x; \theta, F) = \sum_{r=1}^R \pi_r p_r(x; \theta, \eta, \phi, F).
\] (15)

We assume that:

(R1) The density function \( p(x; \theta, F) \) is bounded away from 0, i.e., there is a constant \( c > 0 \) such that for each \( x \) and \( (\theta, F) \in \Theta \times \mathcal{F} \), \( p(x; \theta, F) > c > 0 \). More over, the density function \( p(x; \theta, F) \) is continuously differentiable with respect to \( \theta \) and Hadamard differentiable with respect to \( F \) for all \( x \). We denote derivatives by \( \phi(x; \theta, F) = \frac{\partial}{\partial \theta} \log p(x; \theta, F) \) and \( \psi(x; \theta, F) = \frac{\partial}{\partial F} \log p(x; \theta, F) \) and they are given in (11) and (12).

(R2) We assume \( \hat{\eta}_{\theta, F} \) satisfies \( \hat{\eta}_{\theta_0, F_0} = \eta_0 \) and the function

\[
\tilde{\ell}_0(x) := \phi(x; \theta_0, F_0)
\]

is the efficient score function. Further, we assume \( n^{1/2}(F_n - F_0) = O_P(1) \) and \( (\hat{\theta}_{\theta_0, F_n} - \eta_0) = o_P(1) \) if \( \hat{\theta}_{\eta} \) is a consistent estimator of \( \theta_0 \).
The efficient information matrix \( \tilde{J}_0 = E[\tilde{I}_0 \tilde{I}_0^T] = E[\phi \phi^T (X; \theta_0, F_0)] \) is inevitable.

The score function \( \phi(x; \theta, F) \) defined in (11) takes the form

\[
\phi(x; \theta, F) = \phi(x; \theta, F, \eta_{0,F}),
\]

where, by assumption (R2), the efficient score function is given by

\[
\tilde{I}_0(x) = \phi(x; \theta_0, F_0) = \phi(x; \theta_0, F_0, \eta_0).
\]

We assume that there exists a \( \rho > 0 \) and neighborhoods \( \Theta \) and \( H \) of \( \theta_0 \) and \( \eta_0 \), respectively, such that \( C_\rho \) and \( H \) are Donsker and the class of functions \( \{ \phi(x; \theta, F, \eta) : (\theta, F, \eta) \in \Theta \times C_\rho \times H \} \) has a square integrable envelope function and it is Lipschitz in the parameters \( (\theta, F, \eta) \):

\[
\| \phi(x; \theta', F', \eta') - \phi(x; \theta, F, \eta) \| \leq M'(x) (\| \theta' - \theta \| + \| F' - F \| + \| \eta' - \eta \|)
\]

where \( M'(x) \) is a \( P_0 \)-square integrable function.

Note: Since the assumption (R4) is unusual condition, it requires an explanation. In our profile likelihood problem the estimate \( \eta_{0,F} \) of the parameter \( \eta \) is an implicit function. Therefore, the map \( (\theta, F) \rightarrow \eta_{0,F} \) is hard to differentiate. It follows that the profile likelihood score function \( \phi(x; \theta, F) \) defined in (11) is hard to differentiate. However it is often the case that the function \( \phi(x; \theta, F) \) takes the form in (16). In the examples, the map \( (\theta, \eta, F) \rightarrow \phi(x; \theta, \eta, F) \) is in a closed form and therefore it is easy to differentiate. This differentiability became handy when we need to expand the function \( \phi(x; \theta, F) \) through the differentiability of \( \phi(x; \theta, F, \eta) \). We use the assumption (R4) in the proof of Equation (20) in Theorem 1. In the example in Section 3, the function in (44) corresponds to \( \eta_{0,F} \) and the function in (51) corresponds to \( \phi(x; \theta, F, \eta) \).

Note: To calculate the second derivative of the score function \( \phi(x; \theta, F) \) given in (11), we use the idea similar to the derivative of generalized functions ([15]). Let \( \phi \rightarrow (f, \varphi) = \int_{-\infty}^\infty f(x) \varphi(x) \, dx \) be a generalized function, where \( \varphi \) vanishes outside of some interval. Then if \( f \) and \( \varphi \) are differentiable with derivative \( f' \) and \( \varphi' \), then by integration by parts,

\[
(f', \varphi) = \int_{-\infty}^\infty f'(x) \varphi(x) \, dx = -\int_{-\infty}^\infty f(x) \varphi'(x) \, dx = -(f, \varphi').
\]

We define the derivative \( (f', \varphi) \) of the generalized function \( \varphi \rightarrow (f, \varphi) \) by \(- (f, \varphi') \). This definition is valid even if \( f \) is not differentiable, provided \( \varphi \) is differentiable.

A similar idea can be applied in our problem. Suppose the density for the profile likelihood \( p(x; \theta, F) \) given in (10) is twice differentiable with respect to \( \theta \), then by differentiating

\[
\int \left\{ \frac{\partial}{\partial \theta} \log p(x; \theta, F) \right\} p(x; \theta, F) \, dx = 0,
\]

with respect to \( \theta \) at \( (\theta, F) = (\theta_0, F_0) \), we get equivalent expressions for the efficient information matrix in terms of the score function \( \phi(x; \theta_0, F_0) \):

\[
\tilde{I}_0 = E[\phi \phi^T (X; \theta_0, F_0)] = -E \left[ \frac{\partial}{\partial \theta} \phi(X; \theta_0, F_0) \right].
\]

From this equation we are motivated to define the expected derivative of the score function

\[
- E \left[ \frac{\partial}{\partial \theta} \phi(X; \theta_0, F_0) \right] \text{ by } E[\phi \phi^T (X; \theta_0, F_0)].
\]

In the following theorem, we show that the definition is valid even when the derivative of the score function \( \frac{\partial}{\partial \theta} \phi(x; \theta, F) \) does not exist.

**Theorem 1.** (Statistical generalized derivative) Let \( p(x; \theta, F) = \sum_{r=1}^R \pi_r p_r(x; \theta_r, \eta_r, F_r), \phi(x; \theta, F) = \frac{\partial}{\partial \theta} \log p(x; \theta, F), \) and \( \psi(x; \theta, F) = d_F \log p(x; \theta, F) \) as defined in (15), (11) and (12), respectively.
Suppose (R1) and (R4), then, for \( \theta_t \to \theta_0 \) and \( F_t \to F_0 \) as \( t \to 0 \), we have that
\[
E \left[ t^{-1} \{ \phi(X; \theta_t, F_t) - \phi(X; \theta_0, F_0) \} \right] = -E \left[ \phi(X; \theta_0, F_0) \phi'(X; \theta_0, F_0) \right] \{ t^{-1}(\theta_t - \theta_0) \} + o(1), \tag{19}
\]
and
\[
E \left[ t^{-1} \{ \phi(X; \theta_t, F_t) - \phi(X; \theta_t, F_0) \} \right] = -E[\phi(X; \theta_0, F_0)\psi(X; \theta_0, F_0)] \{ t^{-1}(F_t - F_0) \} + O\left\{ t^{-1}(\| \theta_t - \theta_0 \| + \| F_t - F_0 \|) (\| F_t - F_0 \| + \| \eta_{\theta,t,F_t} - \eta_{\theta_0,F_0} \|) \right\} + o(1 + \| \theta_t - \theta_0 \| + \| F_t - F_0 \| + \| \eta_{\theta,t,F_t} - \eta_0 \|). \tag{20}
\]

Note: Note that even when the derivative \( \frac{\partial}{\partial \theta} \phi(x; \theta, F) \) does not exist the Equation (19) shows that the derivative of the map \( \theta \to E[\phi(x; \theta, F)] \) exists. We may call the derivative the statistical generalized derivative. A similar comment for (20) holds.

**Proof.** In (R1) we assumed the density function \( p(x; \theta, F) \) is bounded away from 0 and it is differentiable with respect to \( \theta \) and \( F \). It follows that
\[
\frac{t^{-1}\{p(x; \theta_t, F_0) - p(x; \theta_0, F_0)\}}{p(x; \theta_0, F_0)} = \phi(x; \theta_0, F_0) \{ t^{-1}(\theta_t - \theta_0) \} + o(1), \tag{21}
\]
and
\[
\frac{t^{-1}\{p(x; \theta_t, F_t) - p(x; \theta_t, F_0)\}}{p(x; \theta_t, F_0)} = \psi(x; \theta_t, F_0) \{ t^{-1}(F_t - F_0) \} + o(1), \tag{22}
\]
and
\[
\frac{t^{-1}\{p(x; \theta_t, F_t) - p(x; \theta_0, F_0)\}}{p(x; \theta_0, F_0)} = \phi(x; \theta_0, F_0) t^{-1}(\theta_t - \theta_0) + \psi(x; \theta_0, F_0) t^{-1}(F_t - F_0) + o(1). \tag{23}
\]

We prove the first Equation (19). For each \( t \), the equality
\[
0 = t^{-1} \left\{ \int \phi(x; \theta_t, F_0) p(x; \theta_t, F_0) dx - \int \phi(x; \theta_0, F_0) p(x; \theta_0, F_0) dx \right\} = \int t^{-1} \{ \phi(x; \theta_t, F_0) - \phi(x; \theta_0, F_0) \} p(x; \theta_0, F_0) dx + \int \phi(x; \theta_t, F_0) t^{-1}\{p(x; \theta_t, F_0) - p(x; \theta_0, F_0)\} dx
\]
holds. It follows that, for each \( t \), we have that
\[
\int t^{-1}\{\phi(x; \theta_t, F_0) - \phi(x; \theta_0, F_0)\} p(x; \theta_0, F_0) dx = -\int \phi(x; \theta_t, F_0) t^{-1}\{p(x; \theta_t, F_0) - p(x; \theta_0, F_0)\} dx. \tag{24}
\]

By the dominated convergence theorem with (21), the right hand side of (24) is, as \( t \to 0 \),
\[
-\int \phi(x; \theta_t, F_0) t^{-1}\{p(x; \theta_t, F_0) - p(x; \theta_0, F_0)\} dx
= -\int \phi(x; \theta_t, F_0) t^{-1}\{p(x; \theta_t, F_0) - p(x; \theta_0, F_0)\} \frac{p(x; \theta_0, F_0)}{p(x; \theta_0, F_0)} dx
= -\int \phi(x; \theta_0, F_0) \phi'(x; \theta_0, F_0) p(x; \theta_0, F_0) dx \left\{ t^{-1}(\theta_t - \theta_0) \right\} + o(1).
\]
It follows that, we have (19):

\[
\int t^{-1}\{\phi(x; \theta_1, F_0) - \phi(x; \theta_0, F_0)\} p(x; \theta_0, F_0)dx
= - \int \phi(x; \theta_0, F_0)\phi^T(x; \theta_0, F_0)p(x; \theta_0, F_0)dx \{t^{-1}(\theta_1 - \theta_0)\} + o(1).
\]

Now we prove the second Equation (20). Similar to the beginning of the proof of (19), for each \(t\), the following equation holds:

\[
\int t^{-1}\{\phi(x; \theta_1, F_1) - \phi(x; \theta_1, F_0)\} p(x; \theta_1, F_1)dx
= - \int \phi(x; \theta_1, F_0)t^{-1}\{p(x; \theta_1, F_1) - p(x; \theta_1, F_0)\}dx.
\]  

(25)

Using (23) with the dominated convergence theorem, the left hand side of (25) is, as \(t \to 0\),

\[
\left\| \int t^{-1}\{\phi(x; \theta_1, F_1) - \phi(x; \theta_1, F_0)\} p(x; \theta_1, F_1)dx - \int t^{-1}\{\phi(x; \theta_1, F_1) - \phi(x; \theta_1, F_0)\} p(x; \theta_1, F_0)dx \right\|
= \left\| \int \{\phi(x; \theta_1, F_1) - \phi(x; \theta_1, F_0)\} t^{-1}\{p(x; \theta_1, F_1) - p(x; \theta_1, F_0)\} p(x; \theta_1, F_0)dx \right\|
\leq \left\{ \left\| M'(x)\{\phi(x; \theta_0, F_0)t^{-1}(\theta_1 - \theta_0) + \psi(x; \theta_0, F_0)t^{-1}(F_1 - F_0)\} p(x; \theta_0, F_0)dx \right\| + o(1) \right\}
\times (\left\| F_1 - F_0 \right\| + \left\| \hat{\eta}_{\theta_1, F_1} - \hat{\eta}_{\theta_1, F_0} \right\|)
= O\{t^{-1}(\left\| \theta_1 - \theta_0 \right\| + \left\| F_1 - F_0 \right\|)(\left\| F_1 - F_0 \right\| + \left\| \hat{\eta}_{\theta_1, F_1} - \hat{\eta}_{\theta_1, F_0} \right\|)\} + o(\left\| F_1 - F_0 \right\| + \left\| \hat{\eta}_{\theta_1, F_1} - \hat{\eta}_{\theta_1, F_0} \right\|),
\]  

(26)

where we used (17): with a \(P_0\)-square integrable function \(M'(x)\),

\[
\left\| \phi(x; \theta_1, F_1) - \phi(x; \theta_1, F_0) \right\| = \left\| \phi(x; \theta_1, F_1, \hat{\eta}_{\theta_1, F_1}) - \phi(x; \theta_1, F_0, \hat{\eta}_{\theta_1, F_0}) \right\|
\leq M'(x)(\left\| F_1 - F_0 \right\| + \left\| \hat{\eta}_{\theta_1, F_1} - \hat{\eta}_{\theta_1, F_0} \right\|).
\]

By the dominated convergence theorem with (22), it follows that the integral in the right hand side of the Equation (25) is

\[
\int \phi(x; \theta_1, F_0)t^{-1}\{p(x; \theta_1, F_1) - p(x; \theta_1, F_0)\}dx
= \int \phi(x; \theta_0, F_0)\psi(x; \theta_0, F_0)t^{-1}(F_1 - F_0)p(x; \theta_0, F_0)dx
\]

\[
+ O\{t^{-1}\left\| F_1 - F_0 \right\|((\left\| \theta_1 - \theta_0 \right\| + \left\| \hat{\eta}_{\theta_1, F_0} - \eta_0 \right\|) + o(1(\left\| \theta_1 - \theta_0 \right\| + \left\| \hat{\eta}_{\theta_1, F_0} - \eta_0 \right\|)),
\]

(27)

here again we used (17): there is a \(P_0\)-square integrable function \(M'(x)\) such that

\[
\left\| \phi(x; \theta_1, F_0) - \phi(x; \theta_0, F_0) \right\| = \left\| \phi(x; \theta_1, F_0, \hat{\eta}_{\theta_1, F_0}) - \phi(x; \theta_0, F_0, \eta_0) \right\|
\leq M'(x)(\left\| \theta_1 - \theta_0 \right\| + \left\| \hat{\eta}_{\theta_1, F_0} - \eta_0 \right\|).
\]
Since \( O(\hat{\eta}_{\theta_0,F} - \hat{\eta}_{\theta_0,F_0}) = O(\hat{\eta}_{\theta_0,F_0} - \hat{\eta}_{\theta_0,F_0}) = O(\hat{\eta}_{\theta_0,F} - \eta_0) \), by combining (26) and (27), the equality (25) is equivalent to
\[
\int t^{-1}\{\phi(x; \theta, F) - \phi(x; \theta_0, F_0)\} p(x; \theta_0, F_0) \, dx
\]
\[= - \int \phi(x; \theta_0, F_0) \psi(x; \theta, F) p(x; \theta_0, F_0) \, dx \left\{ t^{-1}(F_t - F_0) \right\}
+ O\left\{ t^{-1}(\|\theta_t - \theta_0\| + \|F_t - F_0\|) (\|F_t - F_0\| + \|\hat{\eta}_{\theta_t,F_t} - \hat{\eta}_{\theta_0,F_0}\|) \right\}
+ o(1 + \|\theta_t - \theta_0\| + \|F_t - F_0\| + \|\hat{\eta}_{\theta_t,F_t} - \eta_0\|).
\]

The (20) follows from this. \( \square \)

Using the result in Theorem 1, we show the asymptotic normality of the profile likelihood MLE:

**Theorem 2.** (Asymptotic normality of the profile likelihood MLE) Suppose the set of assumptions (R1) – (R4) holds. If the profile likelihood MLE (described in Section 2.3) is consistent, then it is an asymptotically linear estimator for \( \theta_0 \):
\[
\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_0^{-1} \bar{t}_0(X_i) + o_p(1).
\]

Hence we have that
\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I_0^{-1}) \text{ as } n \to \infty.
\]

**Proof.** The profile likelihood MLE described in Section 2.3 is solution to the estimating equation
\[
\sum_{i=1}^{n} \phi(X_i; \hat{\theta}_n, F_n) = 0
\]
where \( \phi(x; \theta, F) \) is the profile likelihood score function for \( \theta \) defined in (11).

In (R4) we assumed \( C_p \) and \( H \) are Donsker and the function \( \bar{\phi}(x; \theta, F, \eta) \) is Lipschitz in the parameters \( (\theta, F, \eta) \) with a \( P_0 \)-square integrable function \( M'(x) \) given in (17). By Corollary 2.10.13 in [16], the class \( \{\bar{\phi}(x; \theta, F, \eta) : (\theta, F, \eta) \in \Theta \times C_p \times H\} \) is Donsker. Moreover, we assumed \( \hat{\theta}_n - \theta_0 = o_p(1) \). In (R2), we assumed \( n^{1/2}(F_n - F_0) = O_p(1) \) and \( \hat{\eta}_{\theta_0,F_n} - \eta_0 = o_p(1) \). By the dominated convergence theorem with (17), we have \( E(\bar{\phi}(X; \hat{\theta}_n, F_n, \hat{\eta}_{\theta_0,F_n}) - \bar{\phi}(X; \theta_0, F_0, \eta_0))^2 = o_p(1) \).

By Lemma 19.24 in [14], it follows that
\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\bar{\phi}(X_i; \hat{\theta}_n, F_n, \hat{\eta}_{\theta_0,F_n}) - \bar{\phi}(X_i; \theta_0, F_0, \eta_0)\} = \sqrt{n} E\{\bar{\phi}(X; \hat{\theta}_n, F_n, \hat{\eta}_{\theta_0,F_n}) - \bar{\phi}(X; \theta_0, F_0, \eta_0)\} + o_p(1).
\]

Using (16), this is equivalent to
\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\phi(X_i; \hat{\theta}_n, F_n) - \phi(X_i; \theta_0, F_0)\} = \sqrt{n} E\{\phi(X; \hat{\theta}_n, F_n) - \phi(X; \theta_0, F_0)\} + o_p(1).
\]

From (19) it follows that
\[
\sqrt{n} E\{\phi(X; \hat{\theta}_n, F_n) - \phi(X; \theta_0, F_0)\} = -I_0 \sqrt{n}(\hat{\theta}_n - \theta_0) + o_p(1),
\]
where \( I_0 = E[\bar{t}_0^T] = E\{\phi(X; \theta_0, F_0)\phi^T(X; \theta_0, F_0)\} \).
Using (20),
\[ \sqrt{n}E\{ \phi(X; \hat{\theta}_n, F_n) - \phi(X; \hat{\theta}_n, F_0) \} \]
\[ = -E[\phi(X; \theta_0, F_0)\psi(X; \theta_0, F_0)]\{ \sqrt{n}(F_n - F_0) \}
+ O\{ \sqrt{n}(\|\hat{\theta}_n - \theta_0\| + \|F_n - F_0\|)(\|F_n - F_0\| +\|\hat{\eta}_{\theta_n,F_n} - \eta_0\|) \}
+ o(1 + \|\hat{\theta}_n - \theta_0\| + \|F_n - F_0\| +\|\hat{\eta}_{\theta_n,F_n} - \eta_0\|) \]
\[ = o_P(1 + \sqrt{n}(\hat{\theta}_n - \theta_0)), \] (31)

where we used:

1. Since \( \psi(x; \theta_0, F_0) \) is in the nuisance tangent space and \( \phi(x; \theta_0, F_0) \) is the efficient score function, we have
\[ E[\phi(x; \theta_0, F_0)\psi(x; \theta_0, F_0)] = 0. \] (32)

2. Using consistency of \( \hat{\theta}_n \) with assumptions \( n^{1/2}(F_n - F_0) = O_P(1) \) and \( \hat{\eta}_{\theta_n,F_n} - \eta_0 = o_P(1) \) in (R2), it follows that
\[ O\{ \sqrt{n}(\|\hat{\theta}_n - \theta_0\| + \|F_n - F_0\|)(\|F_n - F_0\| +\|\hat{\eta}_{\theta_n,F_n} - \eta_0\|) \} = o_P(1 + \sqrt{n}(\hat{\theta}_n - \theta_0)) \]
and
\[ o(1 + \|\hat{\theta}_n - \theta_0\| + \|F_n - F_0\| +\|\hat{\eta}_{\theta_n,F_n} - \eta_0\|) = o_P(1). \]

Using (30) and (31), the right hand side of (29) is
\[ \sqrt{n}E\{ \phi(X; \hat{\theta}_n, F_n) - \phi(X; \theta_0, F_0) \} \]
\[ = \sqrt{n}E\{ \phi(X; \hat{\theta}_n, F_0) - \phi(X; \theta_0, F_0) \} + \sqrt{n}E\{ \phi(X; \hat{\theta}_n, F_n) - \phi(X; \hat{\theta}_n, F_0) \} \]
\[ = -I_0\sqrt{n}(\hat{\theta}_n - \theta_0) + o_P\{1 + \sqrt{n}(\hat{\theta}_n - \theta_0)\}. \] (33)

Finally, (29) together with (33) and \( \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(X_i; \hat{\theta}_n, F_n) = 0 \) imply that
\[ \sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}I_0^{-1}\phi(X_i; \theta_0, F_0) + o_P(1). \]

\[ \square \]

3. Joint Mixture Model of Survival and Longitudinal Ordered Data

In the paper [17], we treated “the joint model of ordinal responses and the proportional hazards with the finite mixture” with real data and simulated data examples. In that paper we had the method of calculating the SE of the profile likelihood MLE applied to the examples. However it was not justified with proofs. This is the motivation to write this paper to give the proofs. In this section, we use the theorem 1 and theorem 2 to show the asymptotic normality of the profile likelihood MLE to the model and prove the method of calculating SE in the paper [17]. Please see [17] for real data and simulated data examples.

**Ordinal Response Models:** Let \( Y_{ijm} \) be the ordered categorical response from 1 (poor) to \( L \) (excellent) on item (or question) \( j \) for subject \( i \) at the \( m^{th} \) protocol-specified time point, where \( i = 1, 2, \ldots, n, \ j = 1, 2, \ldots, J \) and \( m = 1, 2, \ldots, M \). In total, there are \( J \) items in the questionnaire related to patients quality of life, collected at times \( t_1, t_2, \ldots, t_M \). Given that subject \( i \) belongs to group \( r \), an ordered stereotype model can be written as
\[
\log \left[ \frac{P(Y_{ijm} = \ell \mid \theta_r)}{P(Y_{ijm} = 1 \mid \theta_r)} \right] = a_\ell + \phi_\ell(b_j + \theta_r), \quad r = 1, \ldots, R,
\]
where \( a_\ell \) is a response level intercept parameter with \( \ell = 2, \ldots, L \), \( b_j \) is an item effect, and \( \theta_r \) is associated with the discrete latent variable, with \( a_1 = 0, b_1 = 0, \phi_1 = 0 \) and \( \theta_1 = 0 \). The parameter \( \theta_r \) can be referred to as a group effect of the quality of life for patients in group \( r \). However, the group memberships are unknown. The \( \{ \phi_r \} \) parameters can be regarded as unknown scores for the outcome categories. Because \( \phi_r(b_j + \theta_r) = (\Delta \phi_r((b_j + \theta_r)/A)) \) for any constant \( A \neq 0 \), for identifiability, we need to impose monotone scores on \( \{ \phi_i \} \) to treat \( Y_{ijm} \) as ordinal. Therefore, the model has the constraint \( 0 = \phi_1 \leq \phi_2 \leq \ldots \leq \phi_L = 1 \). The ordinal response part of likelihood function for the \( i \)th subject is

\[
P(Y_i \mid \theta_r, \alpha) = \prod_{m=1}^{M_i} \prod_{j=1}^J \prod_{\ell=1}^L \left( \frac{\exp(a_\ell + \phi_\ell(b_j + \theta_r))}{1 + \sum_{k=2}^L \exp(a_k + \phi_k(b_j + \theta_r))} \right)^{Y_{ijm}}
\]

(34)

where \( \alpha = (a, b, \phi) \). Each follow-up time point may have a different number of observations because some patient responses are missing.

**The Cox Proportional Hazards Model:** We consider the Cox proportional hazards model for the survival part in the joint model. Let \( X \) be a time-independent covariate. The hazard function for the failure time \( T_i \) of the \( i \)th subject is of the form

\[
\lambda(t \mid X_i, \theta_r, \delta) = \lambda_0(t) \exp(\theta_r \delta_0 + X_i \delta_1)
\]

(35)

where \( \lambda_0(t) \) is the baseline hazard function. The latent variable \( \theta_r \) is linked with the ordinal response model and \( \delta = (\delta_0, \delta_1) \) are coefficients.

For the estimation of the baseline hazard function \( \lambda_0(t) \), we use the method of nonparametric maximum likelihood described in [18, section 4.3]. Let \( \lambda_i \) be the hazard at time \( t_i \), where \( t_1 < t_2 < \ldots < t_n \) are the ordered observed times. Assume that the hazard is zero between adjacent times so that the survival time is discrete. The corresponding cumulative hazard function \( \Lambda_0(t_i) = \sum_{p \leq i} \lambda_p \) is a step function with jumps at the failure time \( t_i \). Then the survival part likelihood function of subject \( i \) is

\[
P(T_i, d_i \mid \lambda_i, \theta_r, \delta) = \left( \lambda_i \exp(\theta_r \delta_0 + X_i \delta_1) \right)^{d_i} \times \exp \left( - \sum_{p \leq i} \lambda_p \exp(\theta_r \delta_0 + X_i \delta_1) \right),
\]

(36)

where the \( d_i \) is an indicator of censorship for individual \( i \): if we observe failure time, then \( d_i = 1 \), otherwise \( d_i = 0 \).

**The Full Likelihood Function:** The joint likelihood function is obtained by combining the probability function from ordinal response model (34), and the proportional hazards model (36), by assuming the two models are independent given latent discrete random variables.

Let \( \tau_r \) be the unknown probability \( (r = 1, \ldots, R) \) that a subject lies in group \( r \), and \( (\Theta, \lambda) = ((\theta, \alpha, \delta), \lambda) \) be all the unknown parameters of the joint model. The mixture model likelihood function is

\[
L(\Theta, \lambda \mid Y, T, D) = \prod_{i=1}^n \left( \sum_{r=1}^R P(Y_i \mid \theta_r, \alpha) P(T_i, d_i \mid \lambda, \theta_r, \delta) \pi_r \right).
\]

(37)

Let \( Z_{ij} \) be the group indicator, where \( Z_{ij} = 1 \) if the \( i \)th individual was from the \( r \)th group and 0 otherwise. The complete data likelihood can be written as

\[
L(\Theta, \lambda \mid Y, T, d, Z) = \prod_{i=1}^n \prod_{r=1}^R \left( P(Y_i \mid \theta_r, \alpha) P(T_i, d_i \mid \lambda, \theta_r, \delta) \pi_r \right)^{Z_{ij}}.
\]

(38)
The expected complete data log likelihood under \( q(Z) = P(Z|Y, T, d) \) is
\[
\sum_Z q(Z) \log L(\Theta, \lambda|Y, T, d, Z) = \sum_{i=1}^{n} \sum_{r=1}^{R} \gamma(Z_{ir}) \left\{ \log \pi_r + \log P(Y_i|\theta_r, \alpha) + \log P(T_i, d_i|\lambda, \theta_r, \delta) \right\} \tag{39}
\]
where \( \gamma(Z_{ir}) \), \( P(Y_i|\theta_r, \alpha) \) and \( P(T_i, d_i|\lambda, \theta_r, \delta) \) are defined in equations (42), (34) and (36) respectively.

To estimate all parameters and the baseline hazards simultaneously, we combine the EM algorithm and the method of nonparametric maximum likelihood.

3.1. Estimation Procedure: Profile Likelihood with EM Algorithm

**Baseline Hazard Estimation:** Before starting the EM-step, we profile out the baseline hazard function \( \lambda_0(t) \). The survival part of Equation (39) can be separately maximized with respect to \( \lambda \):
\[
\sum_{i=1}^{n} \sum_{r=1}^{R} \gamma(Z_{ir}) \log P(T_i, d_i|\lambda, \theta_r, \delta) = \sum_{i=1}^{n} \sum_{r=1}^{R} \gamma(Z_{ir}) \left\{ d_i (\log \lambda_i + \theta_r \delta_0 + X_i \delta) - \sum_{p \leq i} \lambda_p \exp(\theta_r \delta_0 + X_i \delta) \right\}. \tag{40}
\]

By solving \( \frac{\partial}{\partial \lambda} \sum_{i=1}^{n} \sum_{r=1}^{R} \gamma(Z_{ir}) \log P(T_i, d_i|\lambda, \theta_r, \delta) = 0, l = 1, \ldots, n \), we find the maximizer \( \hat{\lambda}_l \) of (40) by holding (\( \theta, \delta \)) fixed, and it is given by
\[
\hat{\lambda}_l(\theta, \delta) = \frac{d_i}{\sum_{p \geq i} \sum_{r=1}^{R} \gamma(Z_{ir}) \exp(\theta_r \delta_0 + X_i \delta)}. \tag{41}
\]

Denote \( \hat{\lambda}(\theta, \delta) = (\hat{\lambda}_1(\theta, \delta), \ldots, \hat{\lambda}_n(\theta, \delta)) \).

**The E-step:** In the E-step, we use the current parameter estimates \( \Theta = (\theta, \alpha, \delta) \) to find the expected values of \( Z_{ir} \):
\[
\gamma(Z_{ir}) = E(Z_{ir}|Y_i, T_i, d_i) = \pi_r \frac{P(Y_i|\theta_r, \alpha) P(T_i, d_i|\hat{\lambda}(\theta, \delta), \theta_r, \delta)}{\sum_{g=1}^{K} \pi_g P(Y_i|\theta_g, \alpha) P(T_i, d_i|\hat{\lambda}(\theta, \delta), \theta_g, \delta)}. \tag{42}
\]

**The M-step:** In the M-step, we maximize Equation (39) with respect to \( \pi_r \) and \( \Theta = (\theta, \alpha, \delta) \). Due to the fact that there is no relationship between \( \pi_r \) and \( \Theta \), they can be estimated separately.

1. Calculate the estimates of \( \pi_r \)
\[
\bar{\pi}_r = \frac{\sum_{i=1}^{n} \gamma(Z_{ir})}{n}.
\]

2. We maximize the second and third parts of Equation (39) (with \( \hat{\lambda}(\theta, \delta) \) in the place of \( \lambda \))
\[
\sum_{i=1}^{n} \sum_{r=1}^{R} \gamma(Z_{ir}) \left\{ \log P(Y_i|\theta_r, \alpha) + \log P(T_i, d_i|\hat{\lambda}(\theta, \delta), \theta_r, \delta) \right\} \tag{43}
\]
with respect to \( \Theta = (\theta, \alpha, \delta) \) to obtain \( \hat{\Theta} \).

The estimated parameters from the M-step are returned into the E-step until the value of \( \hat{\Theta} \) converges.
3.2. Asymptotic Normality of the Profile Likelihood MLE and Its Asymptotic Variance

From (41), an estimator of the cumulative hazard function in the counting process notation is

\[
\tilde{\Lambda}(t) = \int_0^t \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n Y_i(u) \sum_{r=1}^R \gamma(Z_{ir}) \exp(\theta_0 + X_i \delta_1)}
\]

where \( N_i(u) = 1\{T_i \leq u, d_i = 1\} \) and \( Y_i(u) = 1\{T_i \geq u\} \).

Let us denote \( E_{\theta} f = \int f dF_n \). Then the above \( \tilde{\Lambda}(t) \) can be written as

\[
\tilde{\Lambda}(t; \Theta, F_n) = \int_0^t \frac{E_{\theta} dN(u)}{E_{\Theta} Y(u) \sum_{r=1}^R \gamma(Z_r) \exp(\theta_0 + X \delta_1)}
\]

(44)

where \( N(u) = 1\{T \leq u, d = 1\} \), \( Y(u) = 1\{T \geq u\} \).

Note: From now on we will deal with the cumulative hazard function \( \Lambda(t) = \int_0^t \lambda(s) ds \) instead of the hazard function \( \lambda(t) \). The function \( \tilde{\lambda}(\Theta, F_n) \) will be replaced with \( \tilde{\Lambda}(\Theta, F_n) \).

Note: Note that the function \( \tilde{\Lambda} \) in (44) depends on \( \gamma(Z_r) \). On the other hand, the function \( \gamma(Z_r) \) in (42) depends on \( \tilde{\Lambda} \). This circular relationship shows the function \( \tilde{\Lambda} \) in (44) is an implicit function.

Equation (43) gives the profile likelihood function for \( \Theta = (\theta, \alpha, \delta) \). The log-profile likelihood function for one observation is

\[
\log P(Y, T, d | \Theta, F_n) = \sum_{r=1}^R \gamma(Z_{ir}) \left\{ \log P(Y_i | \theta_0, \alpha) + \log P(T_i, d_i | \tilde{\Lambda}(\Theta, F_n), \theta_0, \delta) \right\}
\]

(45)

where

\[
\sum_{r=1}^R \gamma(Z_{ir}) \log P(Y_i | \theta_0, \alpha)
\]

(46)

and

\[
\sum_{r=1}^R \gamma(Z_{ir}) \log P(T_i, d_i | \tilde{\Lambda}(\Theta, F_n), \theta_0, \delta)
\]

(47)

In the above log-likelihood we set \( a_1 = b_1 = \phi_1 = \theta_1 = 0 \).

**Score functions:** The score functions for the profile likelihood are

\[
\phi(Y, T, d | \Theta, F_n) = \phi_O(Y_i | \Theta) + \phi_S(T_i, d_i | \Theta, F_n)
\]

\[
= \sum_{r=1}^R \gamma(Z_{ir}) \frac{\partial}{\partial \Theta} \log P(Y_i | \theta_0, \alpha) + \sum_{r=1}^R \gamma(Z_{ir}) \frac{\partial}{\partial \Theta} \log P(T_i, d_i | \tilde{\Lambda}(\Theta, F_n), \theta_0, \delta),
\]

\[
\psi(Y, T, d | \Theta, F_n) = \sum_{r=1}^R \gamma(Z_{ir}) d_F \log P(T_i, d_i | \tilde{\Lambda}(\Theta, F_n), \theta_0, \delta).
\]

(48)

Here all derivatives are calculated treating \( \gamma(Z_r) \) as constant. We call \( \phi_O \) is the score function for the ordinal response model and \( \phi_S \) is the one for the survival model.
Theorem 3. (The efficient score function) We drop subscript \( i \) in Equation (48). We have the followings: at the true value of \((\Theta, F)\),

1. \( \tilde{\Lambda}(t; \Theta, F) = \Lambda(t) \), the true cumulative hazard function, and;
2. the score function \( \phi(Y, T, d|\Theta, F) \) defined in (48) is the efficient score function in the model.

The proof is given in Appendix A.

Theorem 4. (Asymptotic normality of the profile likelihood MLE) We assume the followings:

(A1) The maximal right-censoring time \( \tau > 0 \) is finite and satisfies \( S(\tau) = P(T > \tau) > 0 \).
(A2) The covariate \( X \) is bounded and the parameter \( \Theta = (\theta, \alpha, \delta) \) is in a compact set. This implies that, for some \( 0 < M < \infty \), we have \( \|X\| \leq M \) and \( \|\Theta\| \leq M \).
(A3) The empirical cdf \( F_n \) is \( \sqrt{n} \)-consistent: \( \sqrt{n}(F_n - F_0) = O_p(1) \).
(A4) The efficient information matrix \( \tilde{I} \) is invertible.

Then the profile likelihood MLE \( \hat{\Theta}_n \) obtained by the procedure described in Section 3.1 is asymptotically normal:

\[
\sqrt{n}(\hat{\Theta}_n - \Theta_0) \sim N(0, \tilde{I}^{-1}),
\]

where \( \tilde{I} = E(\phi\phi^T) \) is the efficient information with \( \phi \) is defined in (48).

Proof. We check conditions (R1)-(R4) in Section 2.4 so that Theorem 1 and Theorem 2 can be used to get result stated in the theorem.

Since the ordinal response data part is a parametric model, we mainly discuss for the survival part of the model. The survival part of the profile log-likelihood function for a one observation is given in (47).

To express the survival part of the score function \( \phi_S(T, d|\Theta, F) \) in the form given in condition (R4), we introduce a few notations.

Let

\[
\gamma(Z_r|\Theta, \Lambda) = \frac{\pi_r P(Y|\theta_r, \alpha) P(T, d|\Lambda, \theta_r, \delta)}{\sum_{g=1}^R \pi_g P(Y|\theta_g, \alpha) P(T, d|\Lambda, \theta_g, \delta)}.
\]

The function \( \gamma(Z_r|\Theta, \Lambda) \) is differentiable with respect to \( \Theta \) and \( \Lambda \). Then the function \( \gamma(Z_r) \) in (42) can be expressed as

\[
\gamma(Z_r) = \gamma(Z_r|\Theta, \tilde{\Lambda}(\Theta, F)).
\]

Let

\[
M_0(t|\Theta, F, \Lambda) = E_Y(u) \sum_{r=1}^R \gamma(Z_r|\Theta, \Lambda) \exp(\theta_0 + X\delta_1)
\]

\[
M_1(t|\Theta, F, \Lambda) = E_Y(u) \sum_{r=1}^R \gamma(Z_r|\Theta, \Lambda) \begin{pmatrix} \delta_0 \\ \theta_r \\ X \end{pmatrix} \exp(\theta_0 + X\delta_1).
\]
Then the score function for the survival part \( \phi_S(T, d| \Theta, F) \) is
\[
\tilde{\phi}_S(T, d| \Theta, F, \hat{\Lambda}(\Theta, F)) = \frac{R}{r=1} \gamma(Z_r| \Theta, \hat{\Lambda}(\Theta, F)) \frac{\partial}{\partial \Theta} \log P(T, d| \hat{\Lambda}(\Theta, F), \theta_r, \delta) \\
= \frac{R}{r=1} \gamma(Z_r| \Theta, \hat{\Lambda}(\Theta, F)) \left\{ d \left[ \begin{array}{c} \delta_0 \\ \theta_r \\ X \end{array} \right] - \frac{M_1(T| \Theta, F, \hat{\Lambda}(\Theta, F))}{M_0(T| \Theta, F, \hat{\Lambda}(\Theta, F))} \right\} \\
+ \exp(\theta_r \delta_0 + X \delta_1) \int_0^T \frac{E_F dN(u)}{M_0(u| \Theta, F, \Lambda(\Theta, F))} \left[ \begin{array}{c} \delta_0 \\ \theta_r \\ X \end{array} \right] - \frac{M_1(u| \Theta, F, \hat{\Lambda}(\Theta, F))}{M_0(u| \Theta, F, \Lambda(\Theta, F))} \right\}.
\]

We will check condition (R4) using the function defined by
\[
\tilde{\phi}_S(T, d| \Theta, F, \Lambda) = \frac{R}{r=1} \gamma(Z_r| \Theta, \Lambda) \left\{ d \left[ \begin{array}{c} \delta_0 \\ \theta_r \\ X \end{array} \right] - \frac{M_1(T| \Theta, F, \Lambda)}{M_0(T| \Theta, F, \Lambda)} \right\} \\
+ \exp(\theta_r \delta_0 + X \delta_1) \int_0^T \frac{E_F dN(u)}{M_0(u| \Theta, F, \Lambda)} \left[ \begin{array}{c} \delta_0 \\ \theta_r \\ X \end{array} \right] - \frac{M_1(u| \Theta, F, \Lambda)}{M_0(u| \Theta, F, \Lambda)} \right\}.
\]

\[\Box\]

**Condition (R1):** With assumptions (A1) and (A2), a straight forward calculation can show that the probability functions \( P(T, d| \lambda, \theta_r, \delta) \) in (36) and \( P(Y| \theta_r, \alpha) \) in (34) are bounded from below.

We calculated the survival part score function \( \tilde{\phi}_S(T, d| \Theta, F, \hat{\Lambda}(\Theta, F)) \) in (50). The ordinal response data part is a parametric model, it is differentiable with respect to the parameter \( \Theta \) (we omit the calculation).

We calculate the score function \( \psi(T, d| \Theta, F) = \sum_{r=1}^R \gamma(Z_r) d_f \log P(T, d| \hat{\Lambda}(\Theta, F), \theta_r, \delta) \). For an integrable function \( h \) with the same domain as the cdfs \( F_i \),
\[
\psi(Y, T, d| \Theta, F)h = \sum_{r=1}^R \gamma(Z_r) d_f \log P(T, d| \hat{\Lambda}(\Theta, F), \theta_r, \delta)h \\
= \sum_{r=1}^R \gamma(Z_r) \left\{ d \left( \frac{E_h dN(T)}{E_F dN(T)} - \frac{E_h Y(T) \sum_{r=1}^R \gamma(Z_r) \exp(\theta_r \delta_0 + X \delta_1)}{E_F Y(T) \sum_{r=1}^R \gamma(Z_r) \exp(\theta_r \delta_0 + X \delta_1)} \right) \\
- \exp(\theta_r \delta_0 + X \delta_1) \int_0^T \frac{E_F dN(u)}{E_F Y(u) \sum_{r=1}^R \gamma(Z_r) \exp(\theta_r \delta_0 + X \delta_1)} \right\}.
\]

**Condition (R2):** In (A3), we assumed \( \sqrt{n}(F_n - F_0) = O_p(1) \) (\( \sqrt{n} \)-consistency of \( F_n \)). In Appendix C, we showed consistencies of \( \hat{\Theta}_n \) and \( \hat{\Lambda}(\hat{\Theta}_n, F_n) \): \( \hat{\Theta}_n - \Theta_0 = o_p(1) \) and \( \hat{\Lambda}(\hat{\Theta}_n, F_n) - \Lambda_0 = o_p(1) \). In Theorem 3 we verified the rest of conditions in (R2).

**Condition (R3):** In (A4), we assumed the efficient information matrix \( I \) is invertible.

**Condition (R4):** The score function \( \tilde{\phi}_S(T, d| \Theta, F, \Lambda) \) given in (51) is differentiable with respect to the parameters \( (\Theta, F, \Lambda) \) (with assumptions (A1) and (A2), the derivative is a bounded function). It follows Condition (R4).
4. Conclusions

We proposed a “statistical generalized derivative” to deal with an implicit function that appears in the profile likelihood estimation in semiparametric model examples. The example includes the joint model of ordinal responses and the proportional hazards model with a finite mixture which we treated in this paper. Using the “statistical generalized derivative” we expanded the profile likelihood function and showed asymptotic normality of the profile likelihood MLE. Moreover, this approach enabled us to express the SE of the estimator in terms of the profile likelihood score function. This contributes to the profile likelihood estimation methods where, otherwise, no direct expansions of the profile likelihood were possible. Our approach can be applied to not only the example in this paper, but also to many models with implicit function in the profile likelihood estimation. As a next step, we are working on applying our method to those examples.

Author Contributions: I.L. proposed the project and constructed the model which is in the example section of the paper. Y.H. was motivated to have an estimation of the SE of the profile likelihood estimator in the example. The proof part of the paper is done by the Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: Ivy Liu’s work was supported by the Marsden Fund (Award Number E2987-3648) administrated by the Royal Society of New Zealand.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proof of Theorem 3 (The Efficient Score Function)

Proof. For ease notation we denote the true values of the parameters by \( (\Theta, F, \Lambda) \). From (44), replacing \( F_n \) by \( F \), we have

\[
\hat{\Lambda}(t; \Theta, F) = \int_0^t \frac{EdN(u)}{EY(u) \sum_{r=1}^R \gamma(Z_r) \exp(\theta_r \delta_0 + X \delta_1)}
\]

where \( E \) is the expectation with respect to the true distribution \( F \). Since, at the true value of the parameters \( (\Theta, F, \Lambda) \),

\[
EdN(u) = E[Y(u) \sum_{r=1}^R \gamma(Z_r) \exp(\theta_r \delta_0 + X \delta_1)]d\Lambda(u), \tag{A1}
\]

we have that \( \hat{\Lambda}(t; \Theta, F) = \Lambda(t) \).

The score function \( \phi(Y, T, d|\Theta, F) = \phi_O(Y|\Theta) + \phi_S(T, d|\Theta, F) \) in (48) has two parts: the score function for the ordinal response model \( \phi_O(Y|\Theta) \) and the score function for the survival model \( \phi_S(T, d|\Theta, F) \). Since the score function for the ordinal response model does not involve the parameter \( \Lambda \), we will only work on the survival part of score function.

We treat the part \( \gamma(Z_r) \) as constant in terms of the parameters.

Let

\[
M_1(t) = E \sum_{r=1}^R \gamma(Z_r) \begin{pmatrix} \delta_0 \\ \theta_r \\ X \end{pmatrix} \exp(\theta_r \delta_0 + X \delta_1) I(t \leq T)
\]

\[
M_0(t) = E \sum_{r=1}^R \gamma(Z_r) \exp(\theta_r \delta_0 + X \delta_1) I(t \leq T) \tag{A2}
\]
Then the score function in the survival part of the model at the true value of parameters $\Theta$ and $F$ is
\[
\phi_2(T, d | \Theta, F) = \sum_{r=1}^{R} \gamma(Z_r) \frac{\partial}{\partial \Theta} \log P(T, d | \lambda(\Theta, F), \theta_r, \delta) \\
= \sum_{r=1}^{R} \gamma(Z_r) \frac{\partial}{\partial \Theta} \left\{ d \left( \log \frac{EdN(T)}{EY(T) \sum_{r=1}^{R} \gamma(Z_r) \exp(\theta_r \delta_0 + X \delta_1)} + \theta_r \delta_0 + X \delta_1 \right) \\
- \exp(\theta_r \delta_0 + X \delta_1) \int_{0}^{T} \frac{EdN(u)}{EY(u) \sum_{r=1}^{R} \gamma(Z_r) \exp(\theta_r \delta_0 + X \delta_1)} \right\} \\
= \sum_{r=1}^{R} \gamma(Z_r) \left[ \frac{\delta_0}{\theta_r} - \frac{M_1(T)}{M_0(T)} \right] \\
- \sum_{r=1}^{R} \gamma(Z_r) \left[ \exp(\theta_r \delta_0 + X \delta_1) \int_{0}^{T} \left( \frac{\delta_0}{\theta_r} - \frac{M_1(u)}{M_0(u)} \right) d\Lambda(u) \right] \\
(A3)
\]
where we used Equation (A1). The last expression is the efficient score function in the survival part of the model derived in Equation (A4), Appendix B.

Appendix B. Derivation of Efficient Score Function in the Joint Model

In this appendix, we derive the efficient score function in the joint model using (8). We denote $P_r(\Theta, \Lambda, T, d) = P(T, d | \Lambda, \theta_r, \delta)$. Again the true values of the parameters are denoted by ($\Theta, F, \Lambda$).

The survival part of log-likelihood function for a one observation is
\[
\sum_{r=1}^{R} \gamma(Z_r) \log P_r(\Theta, \Lambda, T, d) = \sum_{r=1}^{R} \gamma(Z_r) \{ d \left( \log \lambda(T) + \theta_r \delta_0 + X \delta_1 \right) - \Lambda(T) \exp(\theta_r \delta_0 + X \delta_1) \} .
\]
The score function for $\Theta$ is
\[
\hat{\ell}_{\Theta, \Lambda} = \sum_{r=1}^{R} \gamma(Z_r) \frac{\partial}{\partial \Theta} \log P_r(\Theta, \Lambda, T, d) = \sum_{r=1}^{R} \gamma(Z_r) \left( \frac{\delta_0}{\theta_r} - \frac{M_1(T)}{M_0(T)} \right) \{ d - \Lambda(T) \exp(\theta_r \delta_0 + X \delta_1) \} .
\]
Let $h : [0, \tau] \rightarrow R$ be a function on $[0, \tau]$. The path defined by
\[
d\Lambda_s = (1 + sh) d\Lambda
\]
is a submodel passing through $\Lambda$ at $s = 0$. The corresponding path for the $\lambda$ is
\[
\lambda_s(t) = \frac{d\Lambda_s(t)}{dt} = (1 + sh) \lambda(t).
\]
The derivative of the log-likelihood function
\[
\sum_{r=1}^{R} \gamma(Z_r) \log P_r(\Theta, \Lambda_s, T, d) = \sum_{r=1}^{R} \gamma(Z_r) \{ d \left( \log \lambda_s(T) + \theta_r \delta_0 + X \delta_1 \right) - \Lambda_s(T) \exp(\theta_r \delta_0 + X \delta_1) \} .
\]
with respect to $s$ at $s = 0$ is the score operator for $\Lambda$:
\[
B_{\Theta, \Lambda}h = \frac{d}{ds} \bigg|_{s=0} \sum_{r=1}^{R} \gamma(Z_r) \log P_r(\Theta, \Lambda_s, T, d) = \sum_{r=1}^{R} \gamma(Z_r) \left( dh(T) - \exp(\theta_r \delta_0 + X \delta_1) \int_{0}^{T} h(u) d\Lambda(u) \right) .
\]
Information operator $B_{\Theta,\Lambda}^* B_{\Theta,\Lambda}$

For functions $g, h : [0, \tau] \to R$, define a path $d\Lambda_s = (1 + sg + th + stgh) d\Lambda$. Then

$$\frac{\partial}{\partial s} \bigg|_{(s,t)=(0,0)} \sum_{r=1}^{R} \gamma(Z_r) \log P_{r,\Theta,\Lambda_s}(T,d) = B_{\Theta,\Lambda} g,$$

$$\frac{\partial}{\partial t} \bigg|_{(s,t)=(0,0)} \sum_{r=1}^{R} \gamma(Z_r) \log P_{r,\Theta,\Lambda_s}(T,d) = B_{\Theta,\Lambda} h.$$ 

Using these we have

$$\langle B_{\Theta,\Lambda} g, B_{\Theta,\Lambda} h \rangle_{L_2(P)} = E \{ (B_{\Theta,\Lambda} g)(B_{\Theta,\Lambda} h) \}$$

$$= -E \left\{ \frac{\partial^2}{\partial t \partial s} \bigg|_{(s,t)=(0,0)} \sum_{r=1}^{R} \gamma(Z_r) \log P_{r,\Theta,\Lambda_s}(T,d) \right\}$$

$$= -E \left\{ \frac{\partial}{\partial t} \bigg|_{t=0} B_{\Theta,\Lambda_0} g \right\}$$

$$= E \left\{ \sum_{r=1}^{R} \gamma(Z_r) \exp(\theta_r \delta_0 + X \delta_1) \int_{0}^{T} I(u \leq T) g(u) h(u) d\Lambda(u) \right\}$$

$$= \int_{0}^{T} h(u) E \left\{ \sum_{r=1}^{R} \gamma(Z_r) \exp(\theta_r \delta_0 + X \delta_1) I(u \leq T) g(u) d\Lambda(u) \right\}$$

$$= \left\langle E \sum_{r=1}^{R} \gamma(Z_r) \exp(\theta_r \delta_0 + X \delta_1) I(u \leq T) g(u), h(u) \right\rangle_{L_2(\Lambda)}$$

Since

$$\langle B_{\Theta,\Lambda} g, B_{\Theta,\Lambda} h \rangle_{L_2(P)} = \langle B_{\Theta,\Lambda}^* B_{\Theta,\Lambda} g, h \rangle_{L_2(\Lambda)},$$

we have the information operator

$$B_{\Theta,\Lambda}^* B_{\Theta,\Lambda} g = E \sum_{r=1}^{R} \gamma(Z_r) \exp(\theta_r \delta_0 + X \delta_1) I(t \leq T) g(t).$$

Since the operator multiplies a number, the inverse is

$$(B_{\Theta,\Lambda}^* B_{\Theta,\Lambda})^{-1} g = \left[ E \sum_{r=1}^{R} \gamma(Z_r) \exp(\theta_r \delta_0 + X \delta_1) I(t \leq T) \right]^{-1} g(t).$$

Calculation of $B_{\Theta,\Lambda}^* t_{\Theta,\Lambda}$

Consider a paths $(s, t) \to (\Theta + sa, \Lambda_t)$ with $d\Lambda_s = (1 + th) d\Lambda$. Then

$$\frac{\partial}{\partial s} \bigg|_{(s,t)=(0,0)} \sum_{r=1}^{R} \gamma(Z_r) \log P_{r,\Theta+s,\Lambda_t}(T,d) = a^T t_{\Theta,\Lambda}$$

$$\frac{\partial}{\partial t} \bigg|_{(s,t)=(0,0)} \sum_{r=1}^{R} \gamma(Z_r) \log P_{r,\Theta+s,\Lambda_t}(T,d) = B_{\Theta,\Lambda} h.$$
Using these we compute that

\[
(a^T  \ell_{\Theta, \Lambda} h)_{L_2(P)} = E \{ (a^T  \ell_{\Theta, \Lambda}) (B_{\Theta, \Lambda} h) \}
\]

\[
= -E \left\{ \frac{\partial^2}{\partial s \partial t} \left|_{(s, t) = (0, 0)} \sum_{r = 1}^{R} \gamma(Z_r) \log P_{r, \Theta + s a_\Lambda}(T, d) \right. \right\}
\]

\[
= -E \left\{ \frac{\partial}{\partial t} \big|_{t = 0} a^T  \ell_{\Theta, \Lambda} \right\}
\]

\[
= a^T E \sum_{r = 1}^{R} \gamma(Z_r) \left( \begin{array}{c} \delta_0 \\ \theta_r \\ X \end{array} \right) \{ \exp(\theta_r \delta_0 + X \delta_1) \int_{0}^{T} I(u \leq T) h(u) d\Lambda(u) \}
\]

\[
= a^T \int_{0}^{T} E \sum_{r = 1}^{R} \gamma(Z_r) \left( \begin{array}{c} \delta_0 \\ \theta_r \\ X \end{array} \right) \exp(\theta_r \delta_0 + X \delta_1) I(u \leq T) h(u) d\Lambda(u)
\]

\[
= \left\langle a^T E \sum_{r = 1}^{R} \gamma(Z_r) \left( \begin{array}{c} \delta_0 \\ \theta_r \\ X \end{array} \right) \exp(\theta_r \delta_0 + X \delta_1) I(u \leq T), h \right\rangle_{L_2(\Lambda)}
\]

Since

\[
(a^T  \ell_{\Theta, \Lambda} h)_{L_2(P)} = (a^T B_{\Theta, \Lambda}^*  \ell_{\Theta, \Lambda} h)_{L_2(\Lambda)},
\]

we have that

\[
B_{\Theta, \Lambda}^*  \ell_{\Theta, \Lambda} = E \sum_{r = 1}^{R} \gamma(Z_r) \left( \begin{array}{c} \delta_0 \\ \theta_r \\ X \end{array} \right) \exp(\theta_r \delta_0 + X \delta_1) I(u \leq T).
\]

**Efficient score function**

Then the efficient score function for the survival part of the model is given by

\[
\tilde{\ell}_{\Theta, \Lambda} = \ell_{\Theta, \Lambda} - B_{\Theta, \Lambda} (B_{\Theta, \Lambda}^* B_{\Theta, \Lambda})^{-1} B_{\Theta, \Lambda}^* \ell_{\Theta, \Lambda}
\]

\[
= \sum_{r = 1}^{R} \gamma(Z_r) \left( \begin{array}{c} \delta_0 \\ \theta_r \\ X \end{array} \right) \{ d - \Lambda(T) \exp(\theta_r \delta_0 + X \delta_1) \} - B_{\Theta, \Lambda} \frac{M_1(t)}{M_0(t)}
\]

\[
= \sum_{r = 1}^{R} \gamma(Z_r) \left( \begin{array}{c} \delta_0 \\ \theta_r \\ X \end{array} \right) \{ d - \Lambda(T) \exp(\theta_r \delta_0 + X \delta_1) \}
\]

\[
- \sum_{r = 1}^{R} \gamma(Z_r) \left( \begin{array}{c} \delta_0 \\ \theta_r \\ X \end{array} \right) \left( \frac{d M_1(T)}{M_0(T)} - \exp(\theta_r \delta_0 + X \delta_1) \int_{0}^{T} \frac{M_1(u)}{M_0(u)} d\Lambda(u) \right)
\]

\[
= \sum_{r = 1}^{R} \gamma(Z_r) \left[ \left( \begin{array}{c} \delta_0 \\ \theta_r \\ X \end{array} \right) - \frac{M_1(T)}{M_0(T)} \right]
\]

\[
- \sum_{r = 1}^{R} \gamma(Z_r) \left( \exp(\theta_r \delta_0 + X \delta_1) \int_{0}^{T} \left[ \left( \begin{array}{c} \delta_0 \\ \theta_r \\ X \end{array} \right) - \frac{M_1(u)}{M_0(u)} \right] d\Lambda(u) \right)
\]

(A4)

where $M_1(t)$ and $M_0(t)$ are defined in (A2).
Appendix C. Proof of Consistency

We outline proof of consistencies of \( \hat{\Theta}_n \) and \( \hat{\Lambda} (\hat{\Theta}_n, F_n) \). These consistencies were required in the verification of Condition (R2) in Theorem 4.

We use the proof of Theorem 5.7 in the book [14] (page 45) and adopt the notations from the book.

Proof. Let us denote the log-likelihood function \( M_n (\Theta, \Lambda) = P_n \log P_{\Theta, \Lambda} \) and the corresponding limit function \( M (\Theta, \Lambda) = P_{\Theta_0, \Lambda_0} \log P_{\Theta, \Lambda} \). We denote the MLE and the true value of \((\Theta, \Lambda)\) by \((\Theta_n, \Lambda_n)\) and \((\Theta_0, \Lambda_0)\), respectively.

Then by the construction of profile likelihood function

\[
M_n (\hat{\Theta}_n, \hat{\Lambda}_n) = M_n (\hat{\Theta}_n, \hat{\Lambda} (\hat{\Theta}_n, F_n))
\]

and \( \hat{\Lambda}_n = \hat{\Lambda} (\hat{\Theta}_n, F_n) \). It follows that the consistency of the MLE \((\hat{\Theta}_n, \hat{\Lambda}_n)\) implies the consistency of \((\hat{\Theta}_n, \hat{\Lambda} (\hat{\Theta}_n, F_n))\).

By (A2), \( \Theta \) is bounded and \( \Lambda (t) \) is in the set of monotone increasing function. The parameter \((\Theta, \Lambda)\) is in a set of Donsker class. Since the function \( \log P_{\Theta, \Lambda} \) is differentiable in the parameter \((\Theta, \Lambda)\) with square integrable envelope function, the set \\{log \( P_{\Theta, \Lambda} \) : \((\Theta, \Lambda)\)in the Donsker class\} is Donsker as well (by Corollary 2.10.13 in [16]). It follows that

\[
M_n (\hat{\Theta}_n, \hat{\Lambda}_n) - M (\hat{\Theta}_n, \hat{\Lambda}_n) = o_p (1).
\]

By the property of Kullback-Leibler divergence, we have

\[
M (\Theta, \Lambda) < M (\Theta_0, \Lambda_0)
\]

for all \((\Theta, \Lambda) \neq (\Theta_0, \Lambda_0)\).

By definition of the MLE also we have

\[
M_n (\hat{\Theta}_n, \hat{\Lambda}_n) \geq M_n (\Theta_0, \Lambda_0).
\]

By the proof of Theorem 5.7 in [14] (page 45), the consistency of the MLE \((\hat{\Theta}_n, \hat{\Lambda}_n) - (\Theta_0, \Lambda_0) = o_p (1) \) follows. \( \square \)

References

1. Hsieh, F.; Tseng, Y.K.; Wang, J.L. Joint modeling of survival and longitudinal data: Likelihood approach revisited. *Biometrics* **2006**, 62, 1037–1043. [CrossRef]
2. Rizopoulos, D. *Joint Models for Longitudinal and Time-to-Event Data With Applications in R*; CRC Press: Boca Raton, FL, USA, 2012.
3. Tsiatis, A.A.; Davidian, M. Joint modeling of longitudinal and time to event data: An overview. *Stat. Sin.* **2004**, 14, 809–834.
4. Wulfsohn, M.S.; Tsiatis, A.A. A joint model for survival and longitudinal data measured with error. *Biometrics* **1997**, 53, 330–339. [CrossRef]
5. Henderson, R.; Duggle, P.; Dobson, A. Identification and efficiency of longitudinal markers for survival. *Biostatistics* **2002**, 3, 33–50. [CrossRef] [PubMed]
6. Ratcliffe, S.J.; Guo, W.; Ten Have, T.R. Joint modeling of longitudinal and survival data via a common frailty. *Biometrics* **2004**, 60, 892–899. [CrossRef]
7. Song, X.; Davidian, M.; Tsiatis, A.A. A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrika* **2002**, 58, 742–753. [CrossRef]
8. Zeng, D.; Cai, J. Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *Ann. Stat.* **2005**, 33, 2132–2163. [CrossRef]
9. Zeng, D.; Lin, D.Y. Maximum likelihood estimation in semiparametric regression models with censored data. *J. R. Stat. Soc.: Series B* **2007**, 69, 507–564. [CrossRef]
10. Zeng, D.; Lin, D.Y. Semiparametric Transformation Models With Random Effects for Recurrent Events. *J. Am. Stat. Assoc.* **2007**, *102*, 167–180. [CrossRef]
11. Zeng, D.; Lin, D.Y. A general asymptotic theory for maximum likelihood estimation in semiparametric regression models with censored data. *Stat. Sin.* **2010**, *20*, 871–910.
12. Murphy, S.A.; van der Vaart, A.W. On profile likelihood (with discussion). *J. Amer. Statist. Assoc.* **2000**, *95*, 449–485. [CrossRef]
13. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006.
14. van der Vaart, A.W. *Asymptotic Statistics*; Cambridge University Press: Cambridge, UK, 1998.
15. Kolmogorov, A.N.; Fomin, S.V. *Introductory Real Analysis*; Dover Publication: New York, NY, USA, 1975.
16. van der Vaart, A.W.; Wellner, J.A. *Weak Convergence and Empirical Processes*; Springer: Berlin/Heidelberg, Germany, 1996.
17. Preedalikit, K.; Liu, I.; Hirose, Y.; Sibanda, N.; Fernández, D. Joint modeling of survival and longitudinal ordered data using a semiparametric approach. *Aust. New Zealand J. Stat.* **2016**, *58*, 153–172. [CrossRef]
18. Kalbfleisch, J.D.; Prentice, R.L. *The Statistical Analysis of Failure Time Data*, 2nd ed.; John Wiley & Sons: New York, NY, USA, 2002.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).