Maximum Likelihood Estimation for the Proportional Odds Model with Random Effects

DONGLIN ZENG, D. Y. LIN, and GUOSHENG YIN

In this article, we study the semiparametric proportional odds model with random effects for correlated, right-censored failure time data. We establish that the maximum likelihood estimators for the parameters of this model are consistent and asymptotically Gaussian. Furthermore, the limiting variances achieve the semiparametric efficiency bounds and can be consistently estimated. Simulation studies show that the asymptotic approximations are accurate for practical sample sizes and that the efficiency gains of the proposed estimators over those of Cai, Cheng and Wei (2002, *JASA*) can be substantial. A real example is provided to illustrate the proposed methods.

KEY WORDS: Correlated failure time data; Frailty model; Linear transformation model; Proportional hazards; Semiparametric efficiency; Survival data.

Donglin Zeng is Assistant Professor and D. Y. Lin is Dennis Gillings Distinguished Professor, Department of Biostatistics, CB# 7420, University of North Carolina, Chapel Hill, NC 27599-7420. Guosheng Yin is Assistant Professor, Department of Biostatistics, M. D. Anderson Cancer Center, Houston, TX 77030. This research was supported by the National Institutes of Health.

1. INTRODUCTION

In many scientific studies, there exists natural or artificial clustering of study subjects such that the survival times or failure times of the subjects within the same cluster are correlated. A common approach to accommodating the intra-class dependence is to incorporate an unobserved random effect, the so-called frailty, into the Cox (1972) proportional hazards model. Specifically, the hazard function for the jth subject of the ith cluster associated with a d_1-vector of covariates $\mathbf{X}_{ij}$ is postulated to take the form

$$\lambda(t|\mathbf{X}_{ij}, \xi_i) = \xi_i \lambda_0(t) e^{\mathbf{X}_{ij}^T \boldsymbol{\alpha}}, \quad i = 1, \ldots, n; j = 1, \ldots, n_i, \tag{1}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function, $\boldsymbol{\alpha}$ is a vector of unknown regression parameters, and ξ_i is the unobserved frailty for the ith cluster. Although various parametric distributions for the frailty have been suggested, the existing literature has been focused on the simple case of gamma frailty. The consistency and asymptotic distribution of the maximum likelihood estimator for the gamma frailty model have been rigorously studied by Murphy (1994; 1995) for the case of no covariates and by Parner (1998) for the case with covariates.

Model (1) imposes a common gamma frailty on all members of the same cluster. Several authors have extended this shared gamma frailty model to accommodate more flexible dependence among cluster members. In particular, Pertersen (1998) allowed different additive frailties for different members of the same cluster. Parner (1998) assumed that the frailty for each cluster consists of two independent components: a common cluster-level effect and a subject-specific effect, and showed that the maximum likelihood estimator is efficient.

Under model (1), the conditional hazard functions given frailties are required to be proportionate over time among different sets of covariate values. This assumption of proportional hazards may not be satisfied in certain applications. For independent failure time data, an attractive alternative to the proportional hazards model is the proportional odds model (Pettitt 1982; Bennett 1983). The proportional odds model constraints the ratio of the odds of survival associated with two sets of covariate values to be constant over time, and consequently the ratio of the hazards to converge to unity as time increases. By contrast, the proportional hazards model constraints the hazard ratio to be constant while the odds ratio tends to 0 or infinity. Physical and biological rationale behind the proportional odds model was provided by Bennett (1983) and others. Statistical inference is much more challenging under the proportional odds model than under the proportional hazards model. Important contributions have been made

by Bennett (1983), Pettitt (1984), Cuzick (1988), Dabrowska and Doksum (1988), Cheng, Wei and Ying (1995), Wu (1995), Murphy, Rossini and van der Vaart (1997), Shen (1998), Lam and Leung (2001), and Chen, Jin and Ying (2002) among others.

In this article, we consider the proportional odds model with random effects for correlated failure time data. Specifically,

$$-\text{logit}\{S(t|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i)\} = G(t) + \mathbf{X}_{ij}^T\boldsymbol{\beta} + \mathbf{Z}_{ij}^T\mathbf{b}_i, \quad i = 1, \ldots, n; j = 1, \ldots, n_i, \tag{2}$$

where $\mathbf{X}_{ij}$ is a d_1-vector of covariates, as defined earlier, $\mathbf{Z}_{ij}$ is a d_2-vector of covariates, which usually contains 1 and part of $\mathbf{X}_{ij}$, $G(\cdot)$ is an unspecified strictly increasing function, $\boldsymbol{\beta}$ is a set of unknown regression parameters, $\mathbf{b}_i$ is a set of unobserved random effects, and $S(\cdot|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i)$ is the survival function conditional on $\mathbf{X}_{ij}$ $\mathbf{Z}_{ij}$, and $\mathbf{b}_i$. We assume that $\mathbf{b}_i$ follows a normal distribution with mean zero and unknown covariance matrix $\boldsymbol{\Sigma}$. Note that model (2) allows covariate-specific or subject-specific random effects whereas model (1) only allows a cluster-specific frailty.

Two recent articles are concerned with special versions of model (2). Specifically, Cai, Cheng and Wei (2002) studied model (2) with a scalar random effect (i.e., $\mathbf{Z}_{ij} \equiv 1$). The parameter estimators are obtained by minimizing the empirical sum of squares of the differences between certain observed quantities and their expected values. The estimators are not asymptotically efficient, and the variance estimation is computationally demanding. The censoring mechanism is required to be purely random and independent of covariates. Lam, Lee and Leung (2002) considered the proportional odds model with scalar random effects μ_{ij}, $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. Within the ith cluster, μ_{ij}, $j = 1, \ldots, n_i$, are multivariate normal with a specific covariance structure. Lam et al. (2002) obtained the estimators for the regression parameters by maximizing a marginal likelihood based on the ranks of the failure times. They did not provide formal asymptotic results or consider the problem of survival function estimation.

In this article, we study the maximum likelihood estimation of model (2). The estimators are shown to be consistent and asymptotically efficient. The asymptotic distributions of the estimators and consistent variance estimators are also obtained. Numerical studies reveal that the proposed estimators perform well for practical sample sizes and the efficiency gains over the estimators of Cai et al. (2002) can be substantial.

We describe in greater detail the data structure and model assumptions in the next section, and develop the estimation theory in Section 3. We then present the results of our numerical studies in Section 4 and provide an application to a real medical study in Section 5. Some

concluding remarks are given in Section 6. Most of the technical details are relegated to the appendix.

2. DATA STRUCTURE AND MODEL ASSUMPTIONS

Suppose that there is a random sample of n clusters with potentially different sizes. For $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$, let T_{ij} and C_{ij}^* be the latent failure time and censoring time for the jth member of the ith cluster, and let $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ be the corresponding d_1- and d_2-vectors of covariates. The regression relationship between T_{ij} and $(\mathbf{X}_{ij}, \mathbf{Z}_{ij})$ is given by model (2). The data consist of $(Y_{ij}, \Delta_{ij}, \mathbf{X}_{ij}, \mathbf{Z}_{ij})$ $(i = 1, \ldots, n; j = 1, \ldots, n_i)$, where $Y_{ij} = T_{ij} \wedge C_{ij}$, $\Delta_{ij} = I(T_{ij} \leq C_{ij})$, and $C_{ij} = C_{ij}^* \wedge \tau$. Here and in the sequel, $a \wedge b = \min(a, b)$, $a \vee b = \max(a, b)$, $I(\cdot)$ is the indicator function, and τ is a fixed constant denoting the end of the study.

We impose the following regularity conditions.

C.1. Conditional on covariates $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$, the censoring time C_{ij}^* is independent of the failure time T_{ij} and random effects $\mathbf{b}_i$.

C.2. There exists some positive constant δ_0 such that $\Pr(C_{ij}^* \geq \tau | \mathbf{X}_{ij}, \mathbf{Z}_{ij}) \geq \delta_0$ almost surely.

C.3. All the $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ are bounded. In addition, if there exist a constant vector $\mathbf{c}$ and a symmetric matrix $\mathbf{\Sigma}$ such that

$$[1, \mathbf{X}_{ij}^T]\mathbf{c} + \mathbf{Z}_{ij}^T \mathbf{\Sigma} \mathbf{Z}_{ij} = 0, \quad j = 1, \ldots, n_i,$$

and

$$\mathbf{Z}_{ij}^T \mathbf{\Sigma} \mathbf{Z}_{ij'} = 0, \quad j \neq j'; j, j' = 1, \ldots, n_i$$

almost surely, then $\mathbf{c} = \mathbf{0}$ and $\mathbf{\Sigma} = \mathbf{0}$.

C.4. The true value $G_0(t)$ of $G(t)$ is a strictly increasing function in $[0, \tau]$ and is continuously differentiable. In addition, $G_0(0) = -\infty, de^{G_0(t)}/dt|_{t=0+} > 0$, and $G_0(\tau) < \infty$.

C.5. The true values of $\boldsymbol{\beta}$ and $\mathbf{\Sigma}$, $\boldsymbol{\beta}_0$ and $\mathbf{\Sigma}_0$, belong to the interior of a known compact set

$$\begin{aligned} \Theta \ = \{ \ & (\boldsymbol{\beta}, \mathbf{\Sigma}) : |\boldsymbol{\beta}| \leq \mathcal{B} \text{ for some constant } \mathcal{B}, \ \mathbf{\Sigma} \text{ is positive definite} \\ & \text{and its eigenvalues are bounded away from 0 and } \infty \}. \end{aligned}$$

C.6. The cluster size is completely random. In addition, there exists a positive integer n_0 such that $1 \leq n_i \leq n_0$ and $\Pr(n_i \geq 2) > 0$.

Remark 1. Conditions C.3, C.4 and C.6 ensure the identifiability of the parameters in model (2). If $\mathbf{Z}_{ij} = \mathbf{Z}_{ij'}$ in condition C.3 for continuous covariates, then the two displays in this condition are equivalent to the linear independence of $[1, \mathbf{X}_{ij}^T]$ and the linear independence of $\mathbf{Z}_{ij}$. In condition C.4, the equality $G_0(0) = -\infty$ follows from the fact that $S(0|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i) = 1$, and the inequality $G_0(\tau) < \infty$ implies that $\Pr(T_{ij} > \tau|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i) > 0$. The bound $G_0(\tau)$ is unknown in practice. Condition C.6 implies that the cluster size is bounded and some clusters have at least two subjects.

3. MAXIMUM LIKELIHOOD ESTIMATION

Define $H(t) = e^{G(t)}$ and $H_0(t) = e^{G_0(t)}$. Note that $H_0(0) = 0$. Under model (2) and condition C.1, the likelihood function for the parameters $(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, H_0)$ is proportional to

$$\prod_{i=1}^{n} \left[\int_{\mathbf{b}} \prod_{j=1}^{n_i} \left\{ \frac{e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}+\mathbf{Z}_{ij}^T\mathbf{b})}}{H(Y_{ij}) + e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}+\mathbf{Z}_{ij}^T\mathbf{b})}} \right\}^{1-\Delta_{ij}} \right.$$
$$\left. \times \left\{ \frac{e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}+\mathbf{Z}_{ij}^T\mathbf{b})} H'(Y_{ij})}{(H(Y_{ij}) + e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}+\mathbf{Z}_{ij}^T\mathbf{b})})^2} \right\}^{\Delta_{ij}} |\boldsymbol{\Sigma}|^{-1/2} e^{-\mathbf{b}^T\boldsymbol{\Sigma}^{-1}\mathbf{b}/2} d\mathbf{b} \right],$$

where $H'(t)$ is the derivative of $H(t)$. It would seem natural to calculate the maximum likelihood estimators of $(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, H_0)$ by maximizing the above likelihood function. The maximum of this function, however, is infinity since we can always choose some function $H(\cdot)$ with fixed values at the Y_{ij} while letting $H'(Y_{ij})$ go to infinity for some Y_{ij} with $\Delta_{ij} = 1$. Thus, we relax $H(\cdot)$ to be right-continuous and allow $H(\cdot)$ to have jumps at the Y_{ij}. We then maximize the following function

$$L_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}, H) \equiv \prod_{i=1}^{n} \left[\int_{\mathbf{b}} \prod_{j=1}^{n_i} \left\{ \frac{e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}+\mathbf{Z}_{ij}^T\mathbf{b})}}{H(Y_{ij}) + e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}+\mathbf{Z}_{ij}^T\mathbf{b})}} \right\}^{1-\Delta_{ij}} \right.$$
$$\left. \times \left\{ \frac{e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}+\mathbf{Z}_{ij}^T\mathbf{b})} H\{Y_{ij}\}}{(H(Y_{ij}) + e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}+\mathbf{Z}_{ij}^T\mathbf{b})})^2} \right\}^{\Delta_{ij}} |\boldsymbol{\Sigma}|^{-1/2} e^{-\mathbf{b}^T\boldsymbol{\Sigma}^{-1}\mathbf{b}/2} d\mathbf{b} \right], \quad (3)$$

where $H\{t\}$ denotes the jump size of $H(t)$ at t. To be specific, we maximize $L_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}, H)$ over the parameter space

$$\{(\boldsymbol{\beta}, \boldsymbol{\Sigma}, H) : (\boldsymbol{\beta}, \boldsymbol{\Sigma}) \in \Theta,\ H(t) \text{ is an increasing right-continuous function in } [0, \tau] \text{ with } H(0) = 0\}.$$

The resulting estimators, denoted by $\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\Sigma}}_n$ and $\widehat{H}_n$, are referred to as the nonparametric maximum likelihood estimators (NPMLEs) (Parner, 1998) or the sieve maximum likelihood estimators (Huang and Rossini, 1997; Murphy, Rossini and van der Vaart, 1997).

The existence of the maximizers follows from the following arguments. First, for any $(\boldsymbol{\beta}, \boldsymbol{\Sigma}, H)$ in the parameter space, the ith term on the right-hand side of (3) is bounded by

$$\max_{\mathbf{x}_{ij}, \mathbf{z}_{ij}, (\boldsymbol{\beta}, \boldsymbol{\Sigma}) \in \Theta} \int_{\mathbf{b}} \prod_{j=1}^{n_i} e^{\mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \mathbf{b}} |\boldsymbol{\Sigma}|^{-1/2} e^{-\mathbf{b}^T \boldsymbol{\Sigma}^{-1} \mathbf{b}/2} d\mathbf{b} < \infty,$$

where the inequality follows from the boundedness of $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$, and the compactness of Θ. Second, for any H, we can always construct a new increasing function H^* which is a step function with jumps only at the Y_{ij} for which $\Delta_{ij} = 1$ such that $H^*(Y_{ij}) = H(Y_{ij})$. Clearly, $H^*\{Y_{ij}\} \geq H\{Y_{ij}\}$ for $\Delta_{ij} = 1$ so that $L_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}, H^*) \geq L_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}, H)$. This implies that the function H which maximizes $L_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}, H)$ should be a step function with positive jumps only at the Y_{ij} for which $\Delta_{ij} = 1$. Third, if $H\{Y_{ij}\} = \infty$ for some Y_{ij}, then it is easy to see that $L_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}, H) = 0$. Therefore, we conclude that the maximizers exist.

The above arguments imply that $\widehat{H}_n(t)$ is a step function with jumps only at the Y_{ij} for which $\Delta_{ij} = 1$. Thus, the NPMLEs for $(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, H_0)$ can be obtained by maximizing $L_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}, H)$ over the parameter space $(\boldsymbol{\beta}, \boldsymbol{\Sigma}) \in \Theta$ and the jump sizes of H at the Y_{ij}. This maximization can be realized via many optimization algorithms such as the large-scale unconstrained optimization function *fminunc* in MATLAB, which is described in the next section.

The asymptotic properties of the proposed estimators are stated in the following theorems.

Theorem 1. Under conditions C.1∼C.6, $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| \to 0$, $\|\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_0\| \to 0$ and $\sup_{t \in [0,\tau]} |\widehat{H}_n(t) - H_0(t)| \to 0$ almost surely, where $\|\cdot\|$ is the Euclidean norm.

Theorem 2. Under conditions C.1∼C.6, the random element $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^T - \boldsymbol{\beta}_0^T, \widehat{\boldsymbol{\Sigma}}_n^T - \boldsymbol{\Sigma}_0^T, \widehat{H}_n(\cdot) - H_0(\cdot))^T$ converges weakly to a zero-mean Gaussian process in the metric space $\mathcal{R}^{d_1} \times \mathcal{R}^{d_2(d_2+1)/2} \times l^\infty[0,\tau]$, where $\widehat{\boldsymbol{\Sigma}}_n$ and $\boldsymbol{\Sigma}_0$ are treated as extended column vectors consisting of the upper triangle elements, and $l^\infty[0,\tau]$ is a normed space consisting of all the bounded functions and the norm is defined as the supremum norm on $[0,\tau]$. Furthermore, $\widehat{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{\Sigma}}_n$ are asymptotically efficient.

Remark 2. Theorem 1 presents the consistency of the maximum likelihood estimators. In conditions C.1∼C.6, $H(\cdot)$ is not assumed to be a bounded function, which means that the weak-compactness of the parameter $H(\cdot)$ is not assumed. Thus, obtaining a bound for the maximum likelihood estimator $\widehat{H}_n(\cdot)$ is a key to the proof of Theorem 1. The proof of Theorem 1 adopts some ideas from Murphy's (1994) proof of the consistency for the gamma frailty model, but the technical details are quite different. Once the consistency is established, the asymptotic

distributions of the maximum likelihood estimators stated in Theorem 2 can be derived along the lines of Murphy (1995) and Parner (1998), although the verification of the continuous invertibility of the information operator is substantially different from theirs. In the statement of Theorem 2, asymptotically efficient estimators mean that the asymptotic variances attain the semiparametric efficiency bounds as defined in Bickel et al. (1993, Ch. 3). The proofs of Theorems 1 and 2 are given in the appendix.

It is essential to estimate the asymptotic covariance matrices of $\widehat{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{\Sigma}}_n$. Intuitively, the variation in estimating the parameter $H(\cdot)$ arises from the variation in estimating the jump sizes of $H(\cdot)$ at the Y_{ij} for which $\Delta_{ij} = 1$. Thus, we can regard the observed likelihood function as a likelihood function indexed by the parameters $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, and the parameters which represent the jump sizes of $H(\cdot)$ at the Y_{ij} for which $\Delta_{ij} = 1$. From the Fisher information theory in the parametric setting, the asymptotic covariance matrix in Theorem 2 can be estimated by the inverse of the observed information matrix for all the parameters. Specifically, for any constant vector $(\mathbf{h}_1, \mathbf{h}_2) \in \mathcal{R}^{d_1} \times \mathcal{R}^{d_2(d_2+1)/2}$ and any bounded function h_3, the asymptotic variance of $\mathbf{h}_1^T\widehat{\boldsymbol{\beta}}_n + \mathbf{h}_2^T\widehat{\boldsymbol{\Sigma}}_n + \int_0^\tau h_3(t)d\widehat{H}_n(t)$ is equal to the asymptotic variance of $\mathbf{h}_1^T\widehat{\boldsymbol{\beta}}_n + \mathbf{h}_2^T\widehat{\boldsymbol{\Sigma}}_n + \sum_{\Delta_{ij}=1} h_3(Y_{ij})\widehat{H}_n\{Y_{ij}\}$ so that it can be estimated by $\mathbf{h}_n^T\mathbf{J}_n^{-1}\mathbf{h}_n$, where $\mathbf{h}_n$ is the vector comprising of $\mathbf{h}_1$, $\mathbf{h}_2$ and the $h_3(Y_{ij})$ for which $\Delta_{ij} = 1$, and $\mathbf{J}_n$ is the negative Hessian matrix of $\log L_n(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}, \widehat{H})$ with respect to $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ and the jump sizes of H at the Y_{ij} for which $\Delta_{ij} = 1$. The next theorem formalizes this approximation.

Theorem 3. Let $V(\mathbf{h}_1, \mathbf{h}_2, h_3)$ be the asymptotic variance of the random variable $n^{1/2}\{\mathbf{h}_1^T(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + \mathbf{h}_2^T(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_0) + \int_0^\tau h_3(t)d(\widehat{H}_n(t) - H_0(t))\}$. Under conditions C.1∼C.6, the estimator $n\mathbf{h}_n^T\mathbf{J}_n^{-1}\mathbf{h}_n \to V(\mathbf{h}_1, \mathbf{h}_2, h_3)$ uniformly in $(\mathbf{h}_1, \mathbf{h}_2, h_3)$ in probability.

Theorem 3 implies that, when the number of uncensored observations is not too large, one can simply invert the observed information matrix for all the parameters, including $\boldsymbol{\beta}, \boldsymbol{\Sigma}$, and the $H\{Y_{ij}\}$ for which $\Delta_{ij} = 1$ to calculate the variances and covariances. Our numerical studies revealed that this approximation is satisfactory for practical sample sizes.

4. NUMERICAL STUDIES

Simulation studies were conducted to evaluate the finite-sample properties of the proposed methods. In the first set of studies, the failure times were generated from the following special case of model (2):

$$-\text{logit}\{S(t|X_{1ij}, X_{2ij}, b_i)\} = \log t - X_{1ij} + X_{2ij} + b_i, \quad i = 1, \ldots, n; j = 1, 2,$$

where $X_{1i1} = 0$, $X_{1i2} = 1$, $X_{2i1} \equiv X_{2i2}$ is a uniform$(0,1)$ random variable, and b_i is zero-mean normal with variance σ^2. The censoring times were generated from the uniform$(0,15)$ distribution, corresponding to approximately 33% censoring rate.

We used the optimization algorithm *fminunc* in the optimization toolbox of MATLAB to obtain the maximum likelihood estimates of β_1, β_2, σ and H. When the gradients and the Hessian derivatives of the likelihood function are provided, the search algorithm is a subspace trust region method and is based on the interior-reflective Newton method described in Coleman and Li (1994; 1996). In each iteration of the search, a large linear system is approximately solved by using the method of preconditioned conjugate gradients. The algorithm converges when the search step size and the norm of the search gradients are smaller than certain thresholds. To avoid negative estimates of the jump sizes for H or negative estimates of σ, we used the logarithms of the jumps sizes and $\log \sigma$ as the parameters during the search. The starting values for $(\beta_1, \beta_2, \sigma)$ were set to be $(0, 0, 1)$. The starting value for the jump size $H\{Y_{ij}\}$ at the failure time Y_{ij} was given by expression (A.1) in Appendix A.1, on the right-hand side of which the values for $(\beta_1, \beta_2, \sigma)$ were set to be the initial values and $H(t)$ was set to be t. In general, the search algorithm converged within 10 iterations. After the algorithm converged, the variance estimates were calculated by inverting the observed information matrix.

Table 1 displays the results of these simulation studies with $n = 200$. The maximum likelihood estimators for all the parameters show little bias. The proposed standard error estimators agree well with the empirical standard errors, and the confidence intervals provide reasonable coverages.

For comparisons, we also computed the estimates based on the method of Cai et al. (2002), which minimizes the following criterion function

$$n\rho \sum_{i=1}^{n} \sum_{l \neq k=1}^{n_i} \left[\frac{\Delta_{il} I(Y_{il} \leq Y_{ik} \wedge t_0)}{\widehat{G}_w(Y_{il})} - \int_b \int_{-\infty}^{\alpha} \frac{e^{-(t+\mathbf{X}_{ik}^T \boldsymbol{\beta}+b)}}{1+e^{-(t+\mathbf{X}_{ik}^T \boldsymbol{\beta}+b)}} \frac{e^{-(t+\mathbf{X}_{il}^T \boldsymbol{\beta}+b)}}{\{1+e^{-(t+\mathbf{X}_{il}^T \boldsymbol{\beta}+b)}\}^2} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{b^2}{2\sigma^2}} dt db \right]^2$$

$$+ \sum_{j \neq i=1}^{n} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} \left[\frac{\Delta_{jl} I(Y_{jl} \leq Y_{ik} \wedge t_0)}{\widehat{G}_c^2(Y_{jl})} \right.$$

$$\left. - \int_{b,\widetilde{b}} \int_{-\infty}^{\alpha} \frac{e^{-(t+\mathbf{X}_{ik}^T \boldsymbol{\beta}+b)}}{1+e^{-(t+\mathbf{X}_{ik}^T \boldsymbol{\beta}+b)}} \frac{e^{-(t+\mathbf{X}_{jl}^T \boldsymbol{\beta}+\widetilde{b})}}{\{1+e^{-(t+\mathbf{X}_{jl}^T \boldsymbol{\beta}+\widetilde{b})}\}^2} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{b^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\widetilde{b}^2}{2\sigma^2}} dt db d\widetilde{b} \right]^2, \qquad (4)$$

where t_0 is the minimum of the 95th percentile of the observed Y_{ij} and the 98th percentile of the observed Y_{ij} for which $\Delta_{ij} = 1$. In (4), ρ is chosen to minimize the asymptotic variance of the estimator for $\boldsymbol{\beta}_0$, $\widehat{G}_c(t) = e^{-\widehat{\Lambda}_c(t)}$ and $\widehat{G}_w(t) = e^{-\widehat{\Lambda}_w(t)}$, where $\widehat{\Lambda}_c$ is the Nelson-Aalen estimator

based on $(Y_{ij}, 1-\Delta_{ij})$, $i=1,\ldots,n; j=1,\ldots,n_i$, and $\widehat{\Lambda}_w$ is the Nelson-Aalen estimator based on $\{Y_{ik}\wedge Y_{il}, 1-(I(Y_{ik}\le Y_{il})\Delta_{ik}+I(Y_{ik}>Y_{il})\Delta_{il})\}$, $i=1,\ldots,n, 1\le k<l\le n_i$. The mean squared errors for Cai et al.'s estimators of β_1, β_2 and σ turned out to be 0.096, 0.749 and 0.935 under $\sigma=3$, and 0.055, 0.192 and 0.311 under $\sigma=1$. Thus, these estimators can be considerably less efficient than the maximum likelihood estimators, especially when there is strong intra-class dependence. It would be interesting to make comparisons with Lam et al.'s method. Their estimators, however, are not easy to program.

Table 1. Summary Statistics for the Simulation Studies With One Random Effect

	Paramter	Value	Mean	SE	SEE	95% CP	MSE
$\sigma=3$	β_1	1.00	0.981	0.205	0.207	0.956	0.042
	β_2	-1.00	-1.020	0.797	0.809	0.950	0.634
	σ	3.00	2.918	0.317	0.297	0.924	0.107
	$H(\tau/4)$	3.75	4.176	2.519	2.101	0.946	6.517
	$H(\tau/2)$	7.50	8.306	4.842	4.334	0.944	24.044
	$H(3\tau/4)$	11.25	12.667	7.550	6.919	0.956	58.894
	$H(\tau)$	15.00	16.229	10.583	9.763	0.948	113.299
$\sigma=1$	β_1	1.00	0.989	0.185	0.191	0.958	0.034
	β_2	-1.00	-1.014	0.390	0.400	0.952	0.152
	σ	1.00	0.949	0.211	0.207	0.980	0.047
	$H(\tau/4)$	3.75	3.856	1.037	1.063	0.962	1.085
	$H(\tau/2)$	7.50	7.724	2.347	2.365	0.962	5.545
	$H(3\tau/4)$	11.25	11.519	4.029	3.964	0.952	16.274
	$H(\tau)$	15.00	14.837	6.877	6.306	0.934	47.220

NOTE: Mean and SE stand for the mean and standard error of the estimator. SEE is the mean of the standard error estimator, and 95% CP is the coverage probability of the 95% confidence interval. MSE is the mean squared error. Each entry is based on 500 simulated data sets.

In our second set of studies, we considered bivariate normal random effects. The failure times were generated from the following model:

$$-\text{logit}\{S(t|X_i,b_{1i},b_{2i})\}=\log\left\{(1+t/2)^2-1\right\}+0.5X_{1ij}-0.5X_{2ij}+b_{1i}+X_{2ij}b_{2i},$$

$$i=1,\ldots,n; j=1,2,$$

where $X_{1i1}=0$, $X_{1i2}=1$, $X_{2i1}\equiv X_{2i2}$ is a uniform$(0,1)$ random variable, and (b_{1i},b_{2i}) has a bivariate normal distribution with zero means, unit variances and covariance -0.4. The censoring times were set to be $\min(3,C^*)$, where C^* is uniform$(3/8,11/8)$, so that approximately

34% of the failure times were censored. The optimization algorithm *fminunc* was again used to find the maximum likelihood estimates. We used the logarithms of the jump sizes of H and the elements in the square-root of the covariate matrix of the random effects as the parameters during the search. To ensure that the covariance matrix estimate is positive definite, we let the objective function be a large negative value (i.e., -10^5) if any condition for a positive definite matrix was violated. This penalization essentially restricts the search within the meaningful regions of the parameters. As before, both the gradients and the Hessian derivatives of the objective function were supplied in the search algorithm. The starting values for the regression parameters and the covariance matrix were zeros and the identity matrix respectively, while the starting values for the jump sizes of H were determined by (A.1), in which the parametric components were set to be the initial values and $H(t)$ to be t.

Table 2. Summary Statistics for the Simulation Studies With Two Random Effects

	Parameter	Value	Mean	SE	SEE	95% CP	MSE
$n = 200$	β_1	0.50	0.498	0.186	0.188	0.958	0.035
	β_2	-0.50	-0.520	0.391	0.412	0.956	0.153
	σ_{11}	0.979	0.861	0.341	0.440	0.978	0.130
	σ_{12}	-0.204	-0.324	0.454	0.620	0.968	0.221
	σ_{22}	0.979	1.069	0.837	1.222	0.946	0.708
	$H(\tau/4)$	0.891	0.920	0.231	0.237	0.962	0.054
	$H(\tau/2)$	2.063	2.119	0.540	0.568	0.964	0.295
	$H(3\tau/4)$	3.517	3.630	1.069	1.052	0.956	1.153
	$H(\tau)$	5.250	5.412	1.821	1.796	0.952	3.336
$n = 400$	β_1	0.50	0.504	0.135	0.133	0.950	0.018
	β_2	-0.50	-0.499	0.299	0.291	0.948	0.089
	σ_{11}	0.979	0.891	0.263	0.295	0.982	0.077
	σ_{12}	-0.204	-0.258	0.381	0.475	0.938	0.148
	σ_{22}	0.979	1.019	0.731	0.984	0.938	0.535
	$H(\tau/4)$	0.891	0.903	0.169	0.163	0.952	0.029
	$H(\tau/2)$	2.063	2.088	0.419	0.393	0.946	0.176
	$H(3\tau/4)$	3.517	3.527	0.761	0.716	0.944	0.578
	$H(\tau)$	5.250	5.250	1.355	1.222	0.932	1.834

NOTE: See Note to Table 1.

Table 2 displays the results for $n = 200$ and $n = 400$. For $n = 200$, the search usually converged after about 10 iterations, and it took less than 2 hours to complete 500 repetitions on 20 1.4 GHz Athlon machines. For $n = 400$, it took about 10 hours to complete. In the table,

σ_{11}, σ_{22} and σ_{12} are the elements in the square-root of the covariance matrix for b_1 and b_2 so that $\sigma_{11} = \sigma_{22} = 0.979$ and $\sigma_{12} = -0.204$. For the regression parameters and the function H, the maximum likelihood estimators show little bias and the proposed standard error estimators agree well with the empirical standard errors. The parameters for the covariance matrix of the random effects are estimated less well, although there is a trend for improvement as n increases.

5. AN EXAMPLE

We now illustrate the proposed methods with the well-known Diabetic Retinopathy Study (Huster et al. 1989), which was conducted to assess the effectiveness of laser photocoagulation in delaying visual loss among patients with diabetic retinopathy. One eye of each patient was randomly selected to receive the laser treatment while the other eye was used as a control. The failure time of interest is the time to visual loss as measured by visual acuity less than 5/200. Following previous authors, we confine our attention to a subset of 197 high-risk patients, and consider three covariates: X_{1ij} indicates, by the values 1 versus 0, whether or not the jth eye ($j = 1$ for the left eye and $j = 2$ for the right eye) of the ith patient was treated with laser photocoagulation, $X_{2i1} \equiv X_{2i2}$ indicates, by the values 1 versus 0, whether the ith patient had adult-onset or juvenile-onset diabetics, and $X_{3ij} = X_{1ij} * X_{2ij}$. We fit model (2) with these three covariates, along with random effects b_i to account for the correlation between the two eyes of the same patient. We used the *fminunc* function with the starting values described in the previous section. The results of the analysis are shown in Table 3. There is a high degree of dependence between the failure times of the two eyes from the same patient. Both the treatment indicator and the interaction term are significant, whereas the diabetic type is not. Cai et al. (2002) reported estimates of β_1, β_2 and β_3 of -0.46, 0.74 and -1.41 with estimated standard errors of 0.30, 0.38 and 0.54. The conclusions based on the two sets of results would be somewhat different.

Table 3. Maximum Likelihood Estimates of the Random-Effect Proportional Odds Model for the Diabetic Retinopathy Study

Parameter	Estimate	SE	Est/SE	p-value
β_1	-0.659	0.295	-2.233	0.025
β_2	0.496	0.345	1.438	0.150
β_3	-1.234	0.466	-2.650	0.008
σ	1.296	0.251	5.168	< 0.001

NOTE: SE is the estimated standard error, and p-value pertains to the two-sided test of zero parameter value.

To compare the fit of competing models, Cai et al. (2002, p. 516) proposed a distance measure which summarizes the differences between the observed and fitted values of the failure times. This distance measure turns out to be 2.233×10^4 for the proportional odds model with normal random effect as opposed to 2.241×10^4 for the proportional hazards model with gamma frailty. Thus, the former model appears to fit the data slightly better than the latter.

One important application of random-effects models is to predict the future survival experience of one member given the survival history of the other members of the same cluster. In the Diabetic Retinopathy Study, one may be interested in estimating, for example, the conditional survival probabilities of the treated eye given that it has not failed before 30 months while the untreated eye failed between 24 and 30 months, i.e., $\Pr(T_2 > t | T_2 > 30, 24 < T_1 < 30, X_{11} = 0, X_{12} = 1, X_2)$ for $t > 30$, where T_2 is the failure time for the treated eye and T_1 is the failure time for the untreated eye, X_{1k} is the treatment status for the kth eye, and X_2 is the diabetic type for this patient. It is straightforward to show that

$$\begin{aligned} &\Pr(T_2 > t | T_2 > 30, 24 < T_1 < 30, X_{11} = 0, X_{12} = 1, X_2) \\ &= \frac{\int_u g(u, t, 1, X_2; \boldsymbol{\beta}, \sigma, H)\{g(u, 24, 0, X_2; \boldsymbol{\beta}, \sigma, H) - g(u, 30, 0, X_2; \boldsymbol{\beta}, \sigma, H)\}\phi(u)du}{\int_u g(u, 30, 1, X_2; \boldsymbol{\beta}, \sigma, H)\{g(u, 24, 0, X_2; \boldsymbol{\beta}, \sigma, H) - g(u, 30, 0, X_2; \boldsymbol{\beta}, \sigma, H)\}\phi(u)du}, \end{aligned} \tag{5}$$

where $\phi(\cdot)$ is the standard normal density function, and

$$g(u, t, X_1, X_2; \boldsymbol{\beta}, \sigma, H) = \frac{e^{-\beta_1 X_1 - \beta_2 X_2 - \beta_3 X_1 X_2 - \sigma u}}{H(t) + e^{-\beta_1 X_1 - \beta_2 X_2 - \beta_3 X_1 X_2 - \sigma u}}.$$

We can easily estimate this probability function by replacing β_1, β_2, σ and H in (5) by their respective maximum likelihood estimates and then evaluating the integration via the Gaussian-quadrature formula. The variance function is given by $\mathbf{D}^T \mathbf{J}_n^{-1} \mathbf{D}$, where $\mathbf{D}$ is the derivative of (5) with respect to $(\beta_1, \beta_2, \sigma)$ and the jump sizes of H at the Y_{ij} for which $\Delta_{ij} = 1$. Figure 1 displays the estimated survival curves along with the 95% confidence intervals for the two diabetic types.

6. DISCUSSION

We have developed consistent and efficient estimators for the proportional odds model with random effects, which is a useful alternative to the popular proportional hazards model with

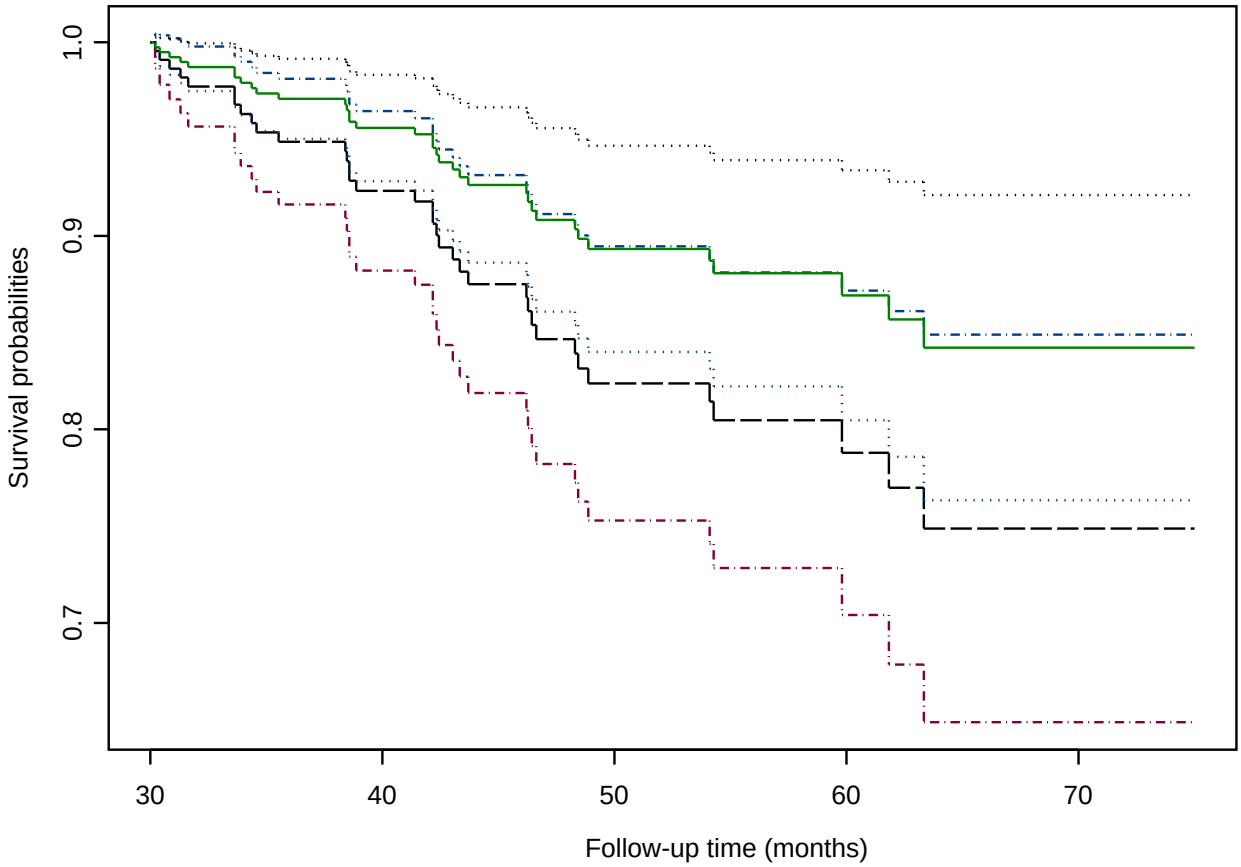

Figure 1: Estimated conditional survival probabilities for the diabetic retinopathy patients: the "——" curve represents the point estimate of the survival function for the adult-onset diabetics, and the "· · · · · ·" curves the corresponding 95% confidence limits; the "– – –" curve represents the the point estimate of the survival function for the juvenile-onset diabetics, and the "– · – · –" curves the corresponding 95% confidence limits.

gamma frailty. The proposed estimators are more efficient than those of Cai et al. (2002). It is computationally less demanding to evaluate the variances of the proposed estimators than those of Cai et al.'s estimators as the latter requires a multi-layer summation.

The proposed numerical algorithm does not guarantee a global maximum. This is a common problem for all maximum likelihood estimators in complex settings. Our experience, however, indicates that the proposed algorithm works well in practice. One approach to increase one's confidence in the estimates is to employ different starting values. We have tried different starting values in our simulated and real data and obtained very similar answers. Note that the existing ad hoc estimating equations may have multiple solutions as well.

For the variance estimation, we invert the observed information matrix on the basis of Theorem 3. When the number of uncensored observations is large, the matrix inversion may potentially be unstable. An alternative approach is to use the numerical differentiation of the profile log-likelihood function, as implemented by Huang and Rossini (1997) and Murphy, Rossini and van der Vaart (1997). In the latter approach, the choice of the neighborhood is arbitrary and no variance estimates are available for the survival function estimators.

It would be worthwhile to study the maximum likelihood estimation for a general class of linear transformation models with random effects

$$\psi\{S(t|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i)\} = G(t) + \mathbf{X}_{ij}^T\boldsymbol{\beta} + \mathbf{Z}_{ij}^T\mathbf{b}_i, \tag{6}$$

where ψ is a given link function. A versatile family of link functions is $\psi(s) = \log\{\lambda^{-1}(s^{-\lambda}-1)\}$, $\lambda \geq 0$, which contains both the proportional hazards model ($\lambda = 0$) and the proportional odds model ($\lambda = 1$). General linear transformation models have been studied by Bickel (1986), Dabrowka and Doksum (1988), Cheng et al. (1995) and Chen et al. (2003) among others for independent failure time data and by Cai et al. (2002) for clustered failure time data (with a scalar random effect), although asymptotically efficient estimators have yet to be developed. It is expected that the asymptotic normality and the efficiency of the maximum likelihood estimators for this class of models depend on the smoothness property of ψ. We are currently investigating the conditions for ψ and developing the requisite asymptotic theory.

The proposed methods are based on the normality of the random effects. The normality assumption may not be satisfied in some applications. It would be desirable to relax this assumption and to require only that the random effects have zero means. One possible approach is to approximate the density of random effects with a truncated series expansion (Davidian and Giltinana 1995, Ch. 7). We will pursue this generalization in our future work.

APPENDIX. PROOFS OF THEOREMS

A.1 Proof of Theorem 1

The proof of Theorem 1 mimics Murphy's (1994) proof of consistency for the proportional hazards model with gamma frailty. Substantial technical complications arise from the fact that, unlike the gamma frailty model, the random effects in our setting cannot be integrated out explicitly. The proof consists of two major steps: in the first step, we show that $\widehat{H}_n(\cdot)$ has an upper bound in $[0, \tau]$ with probability one; in the second step, we show that any convergent subsequence of $(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\Sigma}}_n, \widehat{H}_n)$ must converge to $(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, H_0)$.

Step 1. We prove that $\widehat{H}_n(\cdot)$ has an upper bound in $[0, \tau]$ with probability one. Our approach is to show that since $\widehat{H}_n$ maximizes L_n it cannot diverge. Let $l_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}, H) = \log L_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}, H)$. By definition, $l_n(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\Sigma}}_n, \widehat{H}_n) - l_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}, H) \geq 0$ for any $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$ and H. We wish to show that if $\widehat{H}_n$ diverges, then the difference in the log-likelihood must be negative, which will be a contradiction. If H is continuous, $l_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}, H)$ will be infinite for finite n. Thus, the choice of $H = H_0$ is excluded. The key is to construct a suitable function $\widetilde{H}_n$ that uniformly converges to H_0. Suppose that $\widehat{H}_n(\tau) \to \infty$ in some sample space with positive probability. We will show that $n^{-1}\{l_n(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\Sigma}}_n, \widehat{H}_n) - l_n(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, \widetilde{H}_n)\}$ diverges to $-\infty$ if $\widehat{H}_n(\tau) \to \infty$.

We will construct the function $\widetilde{H}_n$ by imitating $\widehat{H}_n$. By differentiating $l_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}, H)$ with respect to $H\{Y_{ij}\}$ and setting the derivative to zero, we see that $\widehat{H}_n\{Y_{ij}\}$ satisfies the equation

$$\frac{\Delta_{ij}}{H\{Y_{ij}\}} = \sum_{k=1}^{n} \left\{ \frac{\int_{\mathbf{b}} R_{1k}(\widehat{\boldsymbol{\beta}}_n, H, \mathbf{b}) R_{2k}(Y_{ij}, \widehat{\boldsymbol{\beta}}_n, H, \mathbf{b}) e^{-\mathbf{b}^T \widehat{\boldsymbol{\Sigma}}_n^{-1} \mathbf{b}/2} |\widehat{\boldsymbol{\Sigma}}_n|^{-1/2} d\mathbf{b}}{\int_{\mathbf{b}} R_{1k}(\widehat{\boldsymbol{\beta}}_n, H, \mathbf{b}) e^{-\mathbf{b}^T \widehat{\boldsymbol{\Sigma}}_n^{-1} \mathbf{b}/2} |\widehat{\boldsymbol{\Sigma}}_n|^{-1/2} d\mathbf{b}} \right\}, \tag{A.1}$$

where

$$R_{1k}(\boldsymbol{\beta}, H, \mathbf{b}) = \prod_{l=1}^{n_k} \frac{e^{-(\mathbf{X}_{kl}^T \boldsymbol{\beta} + \mathbf{Z}_{kl}^T \mathbf{b})}}{\left\{ H(Y_{kl}) + e^{-(\mathbf{X}_{kl}^T \boldsymbol{\beta} + \mathbf{Z}_{kl}^T \mathbf{b})} \right\}^{1+\Delta_{kl}}},$$

$$R_{2k}(t, \boldsymbol{\beta}, H, \mathbf{b}) = \sum_{l=1}^{n_k} \frac{(1+\Delta_{kl}) I(Y_{kl} \geq t)}{H(Y_{kl}) + e^{-(\mathbf{X}_{kl}^T \boldsymbol{\beta} + \mathbf{Z}_{kl}^T \mathbf{b})}}.$$

Thus, we define $\widetilde{H}_n(t)$ as a step function with jumps only at the Y_{ij} for which $\Delta_{ij} = 1$ and the jump size $\widetilde{H}_n\{Y_{ij}\}$ satisfies the equation

$$\frac{\Delta_{ij}}{\widetilde{H}_n\{Y_{ij}\}} = \sum_{k=1}^{n} \left\{ \frac{\int_{\mathbf{b}} R_{1k}(\boldsymbol{\beta}_0, H_0, \mathbf{b}) R_{2k}(Y_{ij}, \boldsymbol{\beta}_0, H_0, \mathbf{b}) e^{-\mathbf{b}^T \boldsymbol{\Sigma}_0^{-1} \mathbf{b}/2} |\boldsymbol{\Sigma}_0|^{-1/2} d\mathbf{b}}{\int_{\mathbf{b}} R_{1k}(\boldsymbol{\beta}_0, H_0, \mathbf{b}) e^{-\mathbf{b}^T \boldsymbol{\Sigma}_0^{-1} \mathbf{b}/2} |\boldsymbol{\Sigma}_0|^{-1/2} d\mathbf{b}} \right\}. \tag{A.2}$$

Specifically, $\widetilde{H}_n(t) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} I(Y_{ij} \leq t) \widetilde{H}_n\{Y_{ij}\}$.

We will show that $\widetilde{H}_n(t)$ converges to $H_0(t)$ uniformly in $t \in [0, \tau]$ with probability one. By the Glivenko-Cantelli theorem (van der Vaart and Wellner 1996, p. 122), $\widetilde{H}_n(t)$ converges almost surely to $E\left\{\sum_{j=1}^{n_i} I(Y_{ij} \le t)\Delta_{ij}/\mu(Y_{ij})\right\}$, where

$$\mu(y) = E\left\{\frac{\int_{\mathbf{b}} R_{1k}(\boldsymbol{\beta}_0, H_0, \mathbf{b}) R_{2k}(y, \boldsymbol{\beta}_0, H_0, \mathbf{b}) e^{-\mathbf{b}^T \boldsymbol{\Sigma}_0^{-1}\mathbf{b}/2} |\boldsymbol{\Sigma}_0|^{-1/2} d\mathbf{b}}{\int_{\mathbf{b}} R_{1k}(\boldsymbol{\beta}_0, H_0, \mathbf{b}) e^{-\mathbf{b}^T \boldsymbol{\Sigma}_0^{-1}\mathbf{b}/2} |\boldsymbol{\Sigma}_0|^{-1/2} d\mathbf{b}}\right\}$$
$$= E\left[\sum_{l=1}^{n_k} E\left\{\frac{(1+\Delta_{kl}) I(Y_{kl} \ge y)}{H_0(Y_{kl}) + e^{-(\mathbf{X}_{kl}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{kl}^T \mathbf{b})}}\middle| n_k\right\}\right].$$

If we denote by $S_c(\cdot|\mathbf{X}_{kl}, \mathbf{Z}_{kl})$ the survival function of C_{kl} given $(\mathbf{X}_{kl}, \mathbf{Z}_{kl})$, then

$$\begin{aligned}
& E\left\{\frac{(1+\Delta_{kl}) I(Y_{kl} \ge y)}{H_0(Y_{kl}) + e^{-(\mathbf{X}_{kl}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{kl}^T \mathbf{b})}}\middle| n_k\right\} \\
= \; & E\left[2\int_y^\infty \frac{1}{H_0(t) + e^{-(\mathbf{X}_{kl}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{kl}^T \mathbf{b})}} \frac{H_0'(t)}{\left\{H_0(t) + e^{-(\mathbf{X}_{kl}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{kl}^T \mathbf{b})}\right\}^2} S_c(t|\mathbf{X}_{kl}, \mathbf{Z}_{kl}) dt\right] \\
& - E\left\{\int_y^\infty \frac{1}{H_0(t) + e^{-(\mathbf{X}_{kl}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{kl}^T \mathbf{b})}} \frac{1}{H_0(t) + e^{-(\mathbf{X}_{kl}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{kl}^T \mathbf{b})}} dS_c(t|\mathbf{X}_{kl}, \mathbf{Z}_{kl})\middle| n_k\right\} \\
= \; & E\left[\frac{S_c(y|\mathbf{X}_{kl}, \mathbf{Z}_{kl})}{\{H_0(y) + e^{-(\mathbf{X}_{kl}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{kl}^T \mathbf{b})}\}^2}\middle| n_k\right],
\end{aligned}$$

where the second equality follows from integration by part. Thus,

$$E\left\{\sum_{j=1}^{n_i} \frac{I(Y_{ij} \le t)\Delta_{ij}}{\mu(Y_{ij})}\right\} = E\left(\sum_{j=1}^{n_i} E\left[\int_0^t \frac{S_c(y|\mathbf{X}_{ij}, \mathbf{Z}_{ij}) H_0'(y)}{\mu(y)\{H_0(y) + e^{-(\mathbf{X}_{ij}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{ij}^T \mathbf{b})}\}^2} dy\middle| n_i\right]\right)$$
$$= \int_0^t H_0'(y) dy = H_0(t).$$

Consequently, $\widetilde{H}_n(t)$ uniformly converges to $H_0(t)$ in $[0, \tau]$.

By plugging equation (A.1) into $l_n(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\Sigma}}_n, \widehat{H}_n)$, we obtain

$$l_n(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\Sigma}}_n, \widehat{H}_n) = \sum_{i=1}^n \log\left\{\int_{\mathbf{b}} R_{1i}(\widehat{\boldsymbol{\beta}}_n, \widehat{H}_n, \mathbf{b}) |\widehat{\boldsymbol{\Sigma}}_n|^{-1/2} e^{-\mathbf{b}^T \widehat{\boldsymbol{\Sigma}}_n^{-1}\mathbf{b}/2} d\mathbf{b}\right\}$$
$$- \sum_{i=1}^n \sum_{j=1}^{n_i} \Delta_{ij} \log\left\{\sum_{k=1}^n \frac{\int_{\mathbf{b}} R_{1k}(\widehat{\boldsymbol{\beta}}_n, \widehat{H}_n, \mathbf{b}) R_{2k}(Y_{ij}, \widehat{\boldsymbol{\beta}}_n, \widehat{H}_n, \mathbf{b}) e^{-\mathbf{b}^T \widehat{\boldsymbol{\Sigma}}_n^{-1}\mathbf{b}/2} |\widehat{\boldsymbol{\Sigma}}_n|^{-1/2} d\mathbf{b}}{\int_{\mathbf{b}} R_{1k}(\widehat{\boldsymbol{\beta}}_n, \widehat{H}_n, \mathbf{b}) e^{-\mathbf{b}^T \widehat{\boldsymbol{\Sigma}}_n^{-1}\mathbf{b}/2} |\widehat{\boldsymbol{\Sigma}}_n|^{-1/2} d\mathbf{b}}\right\}.$$

Likewise, by plugging equation (A.2) into $l_n(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, \widetilde{H}_n)$ and applying the Glivenko-Cantelli theorem, we see that $n^{-1} l_n(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, \widetilde{H}_n) = O(1) - n^{-1} \sum_{i=1}^n \sum_{j=1}^{n_i} \Delta_{ij} \log(n)$, where $O(1)$ denotes a random variable bounded away from infinity almost surely. Thus,

$$n^{-1}\{l_n(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\Sigma}}_n, \widehat{H}_n) - l_n(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, \widetilde{H}_n)\} = O(1) + n^{-1} \sum_{i=1}^n \log\left\{\int_{\mathbf{b}} R_{1i}(\widehat{\boldsymbol{\beta}}_n, \widehat{H}_n, \mathbf{b}) |\widehat{\boldsymbol{\Sigma}}_n|^{-1/2} e^{-\mathbf{b}^T \widehat{\boldsymbol{\Sigma}}_n^{-1}\mathbf{b}/2} d\mathbf{b}\right\}$$

$$-n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\Delta_{ij}\log\left\{n^{-1}\sum_{k=1}^{n}\frac{\int_{\mathbf{b}}R_{1k}(\widehat{\boldsymbol{\beta}}_n,\widehat{H}_n,\mathbf{b})R_{2k}(Y_{ij},\widehat{\boldsymbol{\beta}}_n,\widehat{H}_n,\mathbf{b})e^{-\mathbf{b}^T\widehat{\boldsymbol{\Sigma}}_n^{-1}\mathbf{b}/2}|\widehat{\boldsymbol{\Sigma}}_n|^{-1/2}d\mathbf{b}}{\int_{\mathbf{b}}R_{1k}(\widehat{\boldsymbol{\beta}}_n,\widehat{H}_n,\mathbf{b})e^{-\mathbf{b}^T\widehat{\boldsymbol{\Sigma}}_n^{-1}\mathbf{b}/2}|\widehat{\boldsymbol{\Sigma}}_n|^{-1/2}d\mathbf{b}}\right\}. \tag{A.3}$$

We will show that if $\widehat{H}_n(\tau)\to\infty$, then the right-hand side of (A.3) will diverge to $-\infty$. To this end, we will bound each term in (A.3). Let m and M be constants such that $0<m\le e^{-\mathbf{X}_{kl}^T\widehat{\boldsymbol{\beta}}_n}\le M<\infty$ almost surely for all $k=1,\ldots,n; l=1,\ldots,n_k$. Since

$$\widehat{H}_n(y)+e^{-(\mathbf{X}_{kl}^T\widehat{\boldsymbol{\beta}}_n+\mathbf{Z}_{kl}^T\mathbf{b})}\ge\begin{cases}\widehat{H}_n(y)+e^{-\mathbf{X}_{kl}^T\widehat{\boldsymbol{\beta}}_n}, & \text{if } \mathbf{Z}_{kl}^T\mathbf{b}\le 0,\\ e^{-\mathbf{Z}_{kl}^T\mathbf{b}}\left\{\widehat{H}_n(y)+e^{-\mathbf{X}_{kl}^T\widehat{\boldsymbol{\beta}}_n}\right\}, & \text{if } \mathbf{Z}_{kl}^T\mathbf{b}>0,\end{cases}$$

we have $\widehat{H}_n(y)+e^{-(\mathbf{X}_{kl}^T\widehat{\boldsymbol{\beta}}_n+\mathbf{Z}_{kl}^T\mathbf{b})}\ge e^{-|\mathbf{Z}_{kl}^T\mathbf{b}|}\{\widehat{H}_n(y)+m\}$. Similarly, $\widehat{H}_n(y)+e^{-(\mathbf{X}_{kl}^T\widehat{\boldsymbol{\beta}}_n+\mathbf{Z}_{kl}^T\mathbf{b})}\le e^{|\mathbf{Z}_{kl}^T\mathbf{b}|}\{\widehat{H}_n(y)+M\}$. Thus, there exist constants C_1 and C_2 such that the following results hold:

$$\begin{aligned}\int_{\mathbf{b}}R_{1i}(\widehat{\boldsymbol{\beta}}_n,\widehat{H}_n,\mathbf{b})|\widehat{\boldsymbol{\Sigma}}_n|^{-1/2}e^{-\mathbf{b}^T\widehat{\boldsymbol{\Sigma}}_n^{-1}\mathbf{b}/2}d\mathbf{b}&\le\int_{\mathbf{b}}\prod_{j=1}^{n_i}\frac{e^{-(\mathbf{X}_{ij}^T\widehat{\boldsymbol{\beta}}_n+\mathbf{Z}_{ij}^T\mathbf{b})+(1+\Delta_{ij})|\mathbf{Z}_{ij}^T\mathbf{b}|}}{(\widehat{H}_n(Y_{ij})+m)^{1+\Delta_{ij}}}|\widehat{\boldsymbol{\Sigma}}_n|^{-1/2}e^{-\mathbf{b}^T\widehat{\boldsymbol{\Sigma}}_n^{-1}\mathbf{b}/2}d\mathbf{b}\\&\le C_1\prod_{j=1}^{n_i}\frac{1}{(\widehat{H}_n(Y_{ij})+m)^{1+\Delta_{ij}}},\end{aligned} \tag{A.4}$$

$$\begin{aligned}&\int_{\mathbf{b}}R_{1k}(\widehat{\boldsymbol{\beta}}_n,\widehat{H}_n,\mathbf{b})R_{2k}(Y_{ij},\widehat{\boldsymbol{\beta}}_n,\widehat{H}_n,\mathbf{b})|\widehat{\boldsymbol{\Sigma}}_n|^{-1/2}e^{-\mathbf{b}^T\widehat{\boldsymbol{\Sigma}}_n^{-1}\mathbf{b}/2}d\mathbf{b}\\&\ge\int_{\mathbf{b}}\prod_{l=1}^{n_k}\frac{e^{-(\mathbf{X}_{kl}^T\widehat{\boldsymbol{\beta}}_n+\mathbf{Z}_{kl}^T\mathbf{b})-(1+\Delta_{kl})|\mathbf{Z}_{kl}^T\mathbf{b}|}}{(\widehat{H}_n(Y_{kl})+M)^{1+\Delta_{kl}}}\left\{\sum_{l'=1}^{n_k}\frac{(1+\Delta_{kl'})I(Y_{kl'}\ge Y_{ij})e^{-|\mathbf{Z}_{kl'}^T\mathbf{b}|}}{\widehat{H}_n(Y_{kl'})+M}\right\}|\widehat{\boldsymbol{\Sigma}}_n|^{-1/2}e^{-\mathbf{b}^T\widehat{\boldsymbol{\Sigma}}_n^{-1}\mathbf{b}/2}d\mathbf{b}\\&\ge C_2\sum_{l'=1}^{n_k}\frac{I(Y_{kl'}\ge Y_{ij})}{\widehat{H}_n(Y_{kl'})+M}\prod_{l=1}^{n_k}\frac{1}{(\widehat{H}_n(Y_{kl})+M)^{1+\Delta_{kl}}}.\end{aligned} \tag{A.5}$$

After plugging (A.4) and (A.5) into (A.3), we obtain

$$\begin{aligned}n^{-1}\{l_n(\widehat{\boldsymbol{\beta}}_n,\widehat{\boldsymbol{\Sigma}}_n,\widehat{H}_n)-l_n(\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_0,\widetilde{H}_n)\}&=O(1)-n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(1+\Delta_{ij})\log(\widehat{H}_n(Y_{ij})+m)\\&\quad-n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\log\left[n^{-1}\sum_{k=1}^{n}\sum_{l'=1}^{n_k}\left\{\frac{I(Y_{kl'}\ge Y_{ij})}{\widehat{H}_n(Y_{kl'})+M}\right\}\frac{\prod_{l=1}^{n_k}C_2(\widehat{H}_n(Y_{kl})+m)^{1+\Delta_{kl}}}{\prod_{l=1}^{n_k}C_1(\widehat{H}_n(Y_{kl})+M)^{1+\Delta_{kl}}}\right].\end{aligned}$$

Because there exists a constant C_3 such that

$$\frac{\prod_{l=1}^{n_k}C_2(\widehat{H}_n(Y_{kl})+m)^{1+\Delta_{kl}}}{\prod_{l=1}^{n_k}C_1(\widehat{H}_n(Y_{kl})+M)^{1+\Delta_{kl}}}\ge C_3>0,$$

we conclude that

$$n^{-1}\{l_n(\widehat{\boldsymbol{\beta}}_n,\widehat{\boldsymbol{\Sigma}}_n,\widehat{H}_n)-l_n(\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_0,\widetilde{H}_n)\}=O(1)-n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(1+\Delta_{ij})\log(\widehat{H}_n(Y_{ij})+m)$$

$$-n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\Delta_{ij}\log\left\{n^{-1}\sum_{k=1}^{n}\sum_{l=1}^{n_k}\frac{I(Y_{kl}\geq Y_{ij})}{\widehat{H}_n(Y_{kl})+M}\right\}. \qquad (A.6)$$

It remains to show that, if $\widehat{H}_n(\tau)\to\infty$, the right-hand side of (A.6) diverges to $-\infty$. To this end, we choose a partition of $[0,\tau]$ as follows: with $s_0=\tau$, choose $s_1<s_0$ such that

$$\frac{1}{2}E\left\{\sum_{j=1}^{n_i}(1+\Delta_{ij})I(Y_{ij}=s_0)\right\}>E\left\{\sum_{j=1}^{n_i}\Delta_{ij}I(Y_{ij}\in[s_1,s_0))\right\}.$$

By conditions C.2 and C.4, such an s_1 exists. Define a constant $\epsilon\in(0,1)$ such that

$$\frac{\epsilon}{1-\epsilon}<\frac{E\left\{\sum_{j=1}^{n_i}I(Y_{ij}\in[s_1,s_0))\right\}}{E\left\{\sum_{j=1}^{n_i}\Delta_{ij}I(Y_{ij}\in[0,\tau))\right\}}.$$

If $s_1>0$, we can choose $s_2\equiv\max(0,s)$ such that s is the minimum value less than s_1 satisfying that

$$(1-\epsilon)E\left\{\sum_{j=1}^{n_i}(1+\Delta_{ij})I(Y_{ij}\in[s_1,s_0))\right\}\geq E\left\{\sum_{j=1}^{n_i}\Delta_{ij}I(Y_{ij}\in[s,s_1))\right\}.$$

Clearly, s_2 exists under condition C.4, and $s_2<s_1$. This process is continued so that we obtain a sequence: $\tau\equiv s_0>s_1>s_2>\ldots\geq 0$ such that

$$\begin{aligned}\frac{1}{2}E\left\{\sum_{j=1}^{n_i}(1+\Delta_{ij})I(Y_{ij}=s_0)\right\} &\geq E\left\{\sum_{j=1}^{n_i}\Delta_{ij}I(Y_{ij}\in[s_1,s_0))\right\},\\ (1-\epsilon)E\left\{\sum_{j=1}^{n_i}(1+\Delta_{ij})I(Y_{ij}\in[s_p,s_{p-1}))\right\} &\geq E\left\{\sum_{j=1}^{n_i}\Delta_{ij}I(Y_{ij}\in[s_{p+1},s_p))\right\},\quad p\geq 1.\end{aligned}$$

We claim that such a sequence cannot be infinite, i.e., there exists a finite N such that $s_{N+1}=0$; otherwise, $s_p\to s^*$ for some $s^*\in[0,\tau)$. By the definition of s_p,

$$(1-\epsilon)E\left\{\sum_{j=1}^{n_i}(1+\Delta_{ij})I(Y_{ij}\in[s_p,s_{p-1}))\right\}=E\left\{\sum_{j=1}^{n_i}\Delta_{ij}I(Y_{ij}\in[s_{p+1},s_p))\right\},\quad p\geq 1.$$

We sum the above equations over $p=1,2,\ldots$, and by the continuity of true densities, we obtain

$$(1-\epsilon)E\left\{\sum_{j=1}^{n_i}(1+\Delta_{ij})I(Y_{ij}\in[s^*,\tau))\right\}=E\left\{\sum_{j=1}^{n_i}\Delta_{ij}I(Y_{ij}\in[s^*,s_1))\right\}.$$

Thus,

$$(1-\epsilon)E\left\{\sum_{j=1}^{n_i}I(Y_{ij}\in[s^*,\tau))\right\}\leq\epsilon E\left\{\sum_{j=1}^{n_i}\Delta_{ij}I(Y_{ij}\in[s^*,s_1))\right\},$$

which contradicts the choice of ϵ. Therefore, the sequence is finite: $\tau = s_0 > \ldots > s_{N+1} = 0$. Now the right-hand side of (A.6) can be bounded by

$$
\begin{aligned}
& -n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i} I(Y_{ij}=\tau)(1+\Delta_{ij})\log(\widehat{H}_n(\tau)+m) \\
& -\sum_{p=0}^{N} n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(1+\Delta_{ij})I(Y_{ij}\in[s_{p+1},s_p))\log(\widehat{H}_n(s_{p+1})+m) \\
& -\sum_{p=0}^{N} n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\Delta_{ij}I(Y_{ij}\in[s_{p+1},s_p))\log\left\{n^{-1}\sum_{k=1}^{n}\sum_{l=1}^{n_k}\frac{I(Y_{kl}\geq Y_{ij},Y_{kl}\in[s_{p+1},s_p))}{\widehat{H}_n(s_p)+M}\right\}+O(1) \\
\leq\ & -\frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(1+\Delta_{ij})I(Y_{ij}=\tau)\log(\widehat{H}_n(\tau)+m) \\
& -\left\{\frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(1+\Delta_{ij})I(Y_{ij}=\tau)\log(\widehat{H}_n(\tau)+m)\right. \\
& \qquad \left. -n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\Delta_{ij}I(Y_{ij}\in[s_1,s_0))\log(\widehat{H}_n(\tau)+M)\right\} \\
& -\sum_{p=1}^{N}\left\{n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(1+\Delta_{ij})I(Y_{ij}\in[s_p,s_{p-1}))\log(\widehat{H}_n(s_p)+m)\right. \\
& \qquad \left. -n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\Delta_{ij}I(Y_{ij}\in[s_{p+1},s_p))\log(\widehat{H}_n(s_p)+M)\right\} \\
& -\sum_{p=0}^{N} n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\Delta_{ij}I(Y_{ij}\in[s_{p+1},s_p))\log\left\{n^{-1}\sum_{k=1}^{n}\sum_{l=1}^{n_k}I(Y_{kl}\geq Y_{ij},Y_{kl}\in[s_{p+1},s_p))\right\}+O(1).
\end{aligned}
\tag{A.7}
$$

The first term on the right-hand side of (A.7) diverges to $-\infty$ as $\widehat{H}_n(\tau)\to\infty$. The second term is negative as n is large due to the choice of s_1. By the selection of $s_p, p=1,\ldots,N$, the third term cannot diverge to $+\infty$. Finally, the fourth term is bounded because of the Glivenko-Cantelli theorem. Hence, the right-hand side of (A.7) diverges to $-\infty$. This contradicts the fact that the left-hand side (A.6) is non-negative.

In conclusion, we have shown that $\widehat{H}_n(\tau)$ has an upper bound with probability 1. Thus, it follows from Helly's selection theorem that there exists a convergent subsequence, still denoted by $\widehat{H}_n(\cdot)$, which converges point-wise to a monotone function $H^*(\cdot)$ in $[0,\tau]$. Since $\widehat{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{\Sigma}}_n$ belong to a compact set, by choosing a further subsequence, we can assume that $\widehat{\boldsymbol{\beta}}_n\to\boldsymbol{\beta}^*$ and $\widehat{\boldsymbol{\Sigma}}_n\to\boldsymbol{\Sigma}^*$ for some random vectors $\boldsymbol{\beta}^*$ and $\boldsymbol{\Sigma}^*$.

Step 2. We will show that $\boldsymbol{\beta}^*=\boldsymbol{\beta}_0, \boldsymbol{\Sigma}^*=\boldsymbol{\Sigma}_0$ and $H^*(t)=H_0(t)$. Define

$$
R_{3k}(t,\boldsymbol{\beta},\boldsymbol{\Sigma},H)=\frac{\int_{\mathbf{b}} R_{1k}(\boldsymbol{\beta},H,\mathbf{b})R_{2k}(t,\boldsymbol{\beta},H,\mathbf{b})e^{-\mathbf{b}^T\boldsymbol{\Sigma}^{-1}\mathbf{b}/2}d\mathbf{b}}{\int_{\mathbf{b}} R_{1k}(\boldsymbol{\beta},H,\mathbf{b})e^{-\mathbf{b}^T\boldsymbol{\Sigma}^{-1}\mathbf{b}/2}d\mathbf{b}}.
$$

In view of equations (A.1) and (A.2), we see that $\widehat{H}_n(t)$ is absolutely continuous with respect to $\widetilde{H}_n(t)$, and

$$\widehat{H}_n(t) = \int_0^t \frac{\sum_{k=1}^n R_{3k}(u, \boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, H_0)}{\sum_{k=1}^n R_{3k}(u, \widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\Sigma}}_n, \widehat{H}_n)} d\widetilde{H}_n(u).$$

By taking the limits on both sides of the above display, we conclude that $H^*(t)$ is absolutely continuous with respect to $H_0(t)$ so that $H^*(t)$ is differentiable with respect to t. In addition, $d\widehat{H}_n(t)/d\widetilde{H}_n(t)$ converges to $dH^*(t)/dH_0(t)$ uniformly in t. On the other hand,

$$\begin{aligned} 0 &\leq n^{-1}\{l_n(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\Sigma}}_n, \widehat{H}_n) - l_n(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, \widetilde{H}_n)\} \\ &= n^{-1}\sum_{i=1}^n \log\left\{\int_b R_{1i}(\widehat{\boldsymbol{\beta}}_n, \widehat{H}_n, \mathbf{b})|\widehat{\boldsymbol{\Sigma}}_n|^{-1/2} e^{-\mathbf{b}^T \widehat{\boldsymbol{\Sigma}}_n^{-1}\mathbf{b}/2} d\mathbf{b}\right\} \\ &\quad -n^{-1}\sum_{i=1}^n \log\left\{\int_b R_{1i}(\boldsymbol{\beta}_0, \widetilde{H}_n, \mathbf{b})|\boldsymbol{\Sigma}_0|^{-1/2} e^{-\mathbf{b}^T \boldsymbol{\Sigma}_0^{-1}\mathbf{b}/2} d\mathbf{b}\right\} \\ &\quad +n^{-1}\sum_{i=1}^n\sum_{j=1}^{n_i} \Delta_{ij} \log\left(\widehat{H}_n\{Y_{ij}\}/\widetilde{H}_n\{Y_{ij}\}\right). \end{aligned} \tag{A.8}$$

By letting $n \to \infty$ in (A.8), we have

$$0 \leq E\left\{\log \frac{\int_{\mathbf{b}} R_{1i}(\boldsymbol{\beta}^*, H^*, \mathbf{b})|\boldsymbol{\Sigma}^*|^{-1/2} e^{-\mathbf{b}^T \boldsymbol{\Sigma}^{*-1}\mathbf{b}/2} d\mathbf{b} \prod_{j=1}^{n_i} H^{*\prime}(Y_{ij})^{\Delta_{ij}}}{\int_{\mathbf{b}} R_{1i}(\boldsymbol{\beta}_0, H_0, \mathbf{b})|\boldsymbol{\Sigma}_0|^{-1/2} e^{-\mathbf{b}^T \boldsymbol{\Sigma}_0^{-1}\mathbf{b}/2} d\mathbf{b} \prod_{j=1}^{n_i} H_0{}'(Y_{ij})^{\Delta_{ij}}}\right\}.$$

Because the right-hand side is the negative Kullback-Leibler information, we have

$$\begin{aligned} &\prod_{j=1}^{n_i} H^{*\prime}(Y_{ij})^{\Delta_{ij}} \int_{\mathbf{b}} R_{1i}(\boldsymbol{\beta}^*, H^*, \mathbf{b})|\boldsymbol{\Sigma}^*|^{-1/2} e^{-\mathbf{b}^T \boldsymbol{\Sigma}^{*-1}\mathbf{b}/2} d\mathbf{b} \\ = &\prod_{j=1}^{n_i} H_0{}'(Y_{ij})^{\Delta_{ij}} \int_{\mathbf{b}} R_{1i}(\boldsymbol{\beta}_0, H_0, \mathbf{b})|\boldsymbol{\Sigma}_0|^{-1/2} e^{-\mathbf{b}^T \boldsymbol{\Sigma}_0{}^{-1}\mathbf{b}/2} d\mathbf{b} \end{aligned}$$

almost surely. In other words,

$$\begin{aligned} &\int_{\mathbf{b}} \prod_{j=1}^{n_i} \frac{e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}^* + \mathbf{Z}_{ij}^T\mathbf{b})} H^{*\prime}(Y_{ij})^{\Delta_{ij}}}{\left\{H^*(Y_{ij}) + e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}^* + \mathbf{Z}_{ij}^T\mathbf{b})}\right\}^{1+\Delta_{ij}}} |\boldsymbol{\Sigma}^*|^{-1/2} e^{-\mathbf{b}^T \boldsymbol{\Sigma}^{*-1}\mathbf{b}/2} d\mathbf{b} \\ = &\int_{\mathbf{b}} \prod_{j=1}^{n_i} \frac{e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}_0 + \mathbf{Z}_{ij}^T\mathbf{b})} H_0{}'(Y_{ij})^{\Delta_{ij}}}{\left\{H_0(Y_{ij}) + e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}_0 + \mathbf{Z}_{ij}^T\mathbf{b})}\right\}^{1+\Delta_{ij}}} |\boldsymbol{\Sigma}_0|^{-1/2} e^{-\mathbf{b}^T \boldsymbol{\Sigma}_0{}^{-1}\mathbf{b}/2} d\mathbf{b}. \end{aligned} \tag{A.9}$$

We will show that equation (A.9) entails that $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0, \boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_0$ and $H^* = H_0$. Fix an integer k such that $1 \leq k \leq n_i$. We let $\Delta_{ij} = 1, Y_{ij} = 0$ in (A.9) for $j = 1, \ldots, k$; for those

j such that $j > k$, we perform the following action on the jth term on both sides of (A.9): if $\Delta_{ij} = 0$, we replace Y_{ij} with τ; if $\Delta_{ij} = 1$, we integrate Y_{ij} from 0 to τ. Thus, we obtain

$$\int_{\mathbf{b}} \prod_{j=1}^{k} \left\{ H^{*\prime}(0) e^{\mathbf{X}_{ij}^T \boldsymbol{\beta}^* + \mathbf{Z}_{ij}^T \mathbf{b}} \right\} \prod_{j=k+1}^{n_i} \left\{ \frac{H^*(\tau) e^{\mathbf{X}_{ij}^T \boldsymbol{\beta}^* + \mathbf{Z}_{ij}^T \mathbf{b}}}{H^*(\tau) e^{\mathbf{X}_{ij}^T \boldsymbol{\beta}^* + \mathbf{Z}_{ij}^T \mathbf{b}} + 1} \right\}^{\Delta_{ij}} \left\{ \frac{1}{H^*(\tau) e^{\mathbf{X}_{ij}^T \boldsymbol{\beta}^* + \mathbf{Z}_{ij}^T \mathbf{b}} + 1} \right\}^{1-\Delta_{ij}}$$

$$\times |\boldsymbol{\Sigma}^*|^{-1/2} e^{-\mathbf{b}^T {\boldsymbol{\Sigma}^*}^{-1} \mathbf{b}/2} d\mathbf{b}$$

$$= \int_{\mathbf{b}} \prod_{j=1}^{k} \left\{ {H_0}'(0) e^{\mathbf{X}_{ij}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{ij}^T \mathbf{b}} \right\} \prod_{j=k+1}^{n_i} \left\{ \frac{H_0(\tau) e^{\mathbf{X}_{ij}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{ij}^T \mathbf{b}}}{H_0(\tau) e^{\mathbf{X}_{ij}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{ij}^T \mathbf{b}} + 1} \right\}^{\Delta_{ij}} \left\{ \frac{1}{H_0(\tau) e^{\mathbf{X}_{ij}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{ij}^T \mathbf{b}} + 1} \right\}^{1-\Delta_{ij}}$$

$$\times |\boldsymbol{\Sigma}_0|^{-1/2} e^{-\mathbf{b}^T {\boldsymbol{\Sigma}_0}^{-1} \mathbf{b}/2} d\mathbf{b}. \qquad (A.10)$$

Since $\{\Delta_{ij} : j = k+1, \ldots, n_i\}$ are arbitrary, we sum the two sides of (A.10) over all possible $\Delta_{ij}, j = k+1, \ldots, n_i$ to yield

$$\int_{\mathbf{b}} \prod_{j=1}^{k} \left\{ H^{*\prime}(0) e^{\mathbf{X}_{ij}^T \boldsymbol{\beta}^* + \mathbf{Z}_{ij}^T \mathbf{b}} \right\} |\boldsymbol{\Sigma}^*|^{-1/2} e^{-\mathbf{b}^T {\boldsymbol{\Sigma}^*}^{-1} \mathbf{b}/2} d\mathbf{b}$$

$$= \int_{\mathbf{b}} \prod_{j=1}^{k} \left\{ {H_0}'(0) e^{\mathbf{X}_{ij}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{ij}^T \mathbf{b}} \right\} |\boldsymbol{\Sigma}_0|^{-1/2} e^{-\mathbf{b}^T \boldsymbol{\Sigma}_0^{-1} \mathbf{b}/2} d\mathbf{b}.$$

Thus,

$$\exp\left\{ \sum_{j=1}^{k} \mathbf{X}_{ij}^T \boldsymbol{\beta}^* + \frac{(\sum_{j=1}^{k} \mathbf{Z}_{ij})^T \boldsymbol{\Sigma}^* (\sum_{j=1}^{k} \mathbf{Z}_{ij})}{2} \right\} H^{*\prime}(0)^k$$

$$= \exp\left\{ \sum_{j=1}^{k} \mathbf{X}_{ij}^T \boldsymbol{\beta}_0 + \frac{(\sum_{j=1}^{k} \mathbf{Z}_{ij})^T \boldsymbol{\Sigma}_0 (\sum_{j=1}^{k} \mathbf{Z}_{ij})}{2} \right\} H_0'(0)^k. \qquad (A.11)$$

Condition C.4 implies that $H^{*\prime}(0) > 0$. Note that the index set $\{1, \ldots, k\}$ in equation (A.11) can be replaced by any subset of $\{1, \ldots, n_i\}$. Thus, it is easy to derive from (A.11) that

$$\mathbf{Z}_{ij}^T \boldsymbol{\Sigma}^* \mathbf{Z}_{ij'} = \mathbf{Z}_{ij}^T \boldsymbol{\Sigma}_0 \mathbf{Z}_{ij'}, \quad j \neq j';\ j, j' = 1, \ldots, n_i,$$

and

$$\mathbf{X}_{ij}^T \boldsymbol{\beta}^* + \frac{\mathbf{Z}_{ij}^T \boldsymbol{\Sigma}^* \mathbf{Z}_{ij}}{2} + \log H^{*\prime}(0) = \mathbf{X}_{ij}^T \boldsymbol{\beta}_0 + \frac{\mathbf{Z}_{ij}^T \boldsymbol{\Sigma}_0 \mathbf{Z}_{ij}}{2} + \log H_0'(0), \quad j = 1, \ldots, n_i.$$

According to condition C.3, $\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_0$, $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$ and $H^{*\prime}(0) = H_0'(0)$.

To show that $H^* = H_0$, we let $\Delta_{i1} = 1$ in (A.10) and integrate Y_{i1} from 0 to y; we also perform the following action on the jth term on both sides of (A.10) for $j = 2, \ldots n_i$: if $\Delta_{ij} = 0$,

we replace Y_{ij} with τ; if $\Delta_{ij} = 1$, we integrate Y_{ij} from 0 to τ. Then we sum the resulting equalities over all possible $\{\Delta_{ij} : j = 2, \ldots, n_i\}$ to yield

$$\int_{\mathbf{b}} \left\{ \frac{H^*(y)e^{\mathbf{X}_{i1}^T\boldsymbol{\beta}^* + \mathbf{Z}_{i1}^T\mathbf{b}}}{H^*(y)e^{\mathbf{X}_{i1}^T\boldsymbol{\beta}^* + \mathbf{Z}_{i1}^T\mathbf{b}} + 1} \right\} |\boldsymbol{\Sigma}^*|^{-1/2} e^{-\mathbf{b}^T {\boldsymbol{\Sigma}^*}^{-1}\mathbf{b}/2} d\mathbf{b}$$

$$= \int_{\mathbf{b}} \left\{ \frac{H_0(y)e^{\mathbf{X}_{i1}^T\boldsymbol{\beta}_0 + \mathbf{Z}_{i1}^T\mathbf{b}}}{H_0(y)e^{\mathbf{X}_{i1}^T\boldsymbol{\beta}_0 + \mathbf{Z}_{i1}^T\mathbf{b}} + 1} \right\} |\boldsymbol{\Sigma}_0|^{-1/2} e^{-\mathbf{b}^T {\boldsymbol{\Sigma}_0}^{-1}\mathbf{b}/2} d\mathbf{b}.$$

Because the two sides of the above equation are strictly monotone in $H^*(y)$ and $H_0(y)$, respectively, we have $H^*(y) = H_0(y)$.

Combining the results from *Step 1* and *Step 2*, we conclude that, almost surely, $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| \to 0$, $\|\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_0\| \to 0$, and $|\widehat{H}_n(y) - H_0(y)| \to 0$. The uniform convergence of $\widehat{H}_n$ to H_0 follows from the fact that H_0 is a continuous function.

A.2. Proof of Theorem 2

Consider the set

$$\begin{aligned} \mathcal{H} \;=\; & \{(\mathbf{h}_1, \mathbf{h}_2, h_3) : \mathbf{h}_1 \in \mathcal{R}^{d_1}, \mathbf{h}_2 \in \mathcal{R}^{d_2(d_2+1)/2}, h_3(\cdot) \text{ is a function on } [0,\tau]; \\ & |\mathbf{h}_1| \le 1, |\mathbf{h}_2| \le 1, \|h_3\|_V \le 1\}, \end{aligned}$$

where $\|h_3\|_V$ denotes the total variation of $h_3(\cdot)$ in $[0,\tau]$. We define a sequence of maps S_n mapping a neighborhood of $(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, H_0)$, denoted by $\mathcal{U}$, in the parameter space for $(\boldsymbol{\beta}, \boldsymbol{\Sigma}, H)$ into $l^\infty(\mathcal{H})$ (i.e., the space consisting of bounded functionals on $\mathcal{H}$) as follows:

$$\begin{aligned} & S_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}, H)[\mathbf{h}_1, \mathbf{h}_2, h_3] \\ \equiv\; & n^{-1}\frac{d}{d\epsilon} l_n(\boldsymbol{\beta} + \epsilon\mathbf{h}_1, \boldsymbol{\Sigma} + \epsilon\mathbf{h}_2, H(t) + \epsilon\int_0^t h_3(s)dH(s))\Big|_{\epsilon=0} \\ \equiv\; & A_{n1}[\mathbf{h}_1] + A_{n2}[\mathbf{h}_2] + A_{n3}[h_3], \end{aligned}$$

where A_{np}, $p = 1, 2, 3$, are linear functionals on $\mathcal{R}^{d_1}, \mathcal{R}^{d_2(d_2+1)/2}$, and $BV[0,\tau]$, respectively, and $BV[0,\tau]$ is the space of functions with finite total variation in $[0,\tau]$. In fact, if we let $l_{\boldsymbol{\beta}}, l_{\boldsymbol{\Sigma}}, l_H[h_3]$ be the score function for $\boldsymbol{\beta}$, the score function for $\boldsymbol{\Sigma}$, and the score for H along the path $H(t) + \epsilon\int_0^t h_3(s)dH(s)$ for a single cluster, then

$$A_{n1}[\mathbf{h}_1] = \mathcal{P}_n[\mathbf{h}_1^T l_{\boldsymbol{\beta}}],\; A_{n2}[\mathbf{h}_2] = \mathcal{P}_n[\mathbf{h}_2^T l_{\boldsymbol{\Sigma}}],\; A_{n3}[h_3] = \mathcal{P}_n[l_H[h_3]],$$

where $\mathcal{P}_n$ denotes the empirical measure based on n independent clusters.

We can explicitly write the functionals A_{np},$p = 1, 2, 3$, as follows. Define the operation " $\cdot$ " between two matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ of the same size as the trace of $(\mathbf{M}_1\mathbf{M}_2^T)$, and for each $\mathbf{h}_2 \in \mathcal{R}^{d_2(d_2+1)/2}$, let $\mathcal{D}(\mathbf{h}_2)$ be the symmetric matrix such that the extended vector taken from $\mathcal{D}(\mathbf{h}_2)$ is the same as $\mathbf{h}_2$. Then,

$$
\begin{aligned}
A_{n1}[\mathbf{h}_1] &= n^{-1}\sum_{i=1}^{n}\frac{\int_{\mathbf{b}} R_{1i}(\boldsymbol{\beta}, H, \mathbf{b})\sum_{j=1}^{n_i}\mathbf{X}_{ij}^T\mathbf{h}_1[(1+\Delta_{ij})/\{1+H(Y_{ij})e^{\mathbf{X}_{ij}^T\boldsymbol{\beta}+\mathbf{Z}_{ij}^T\mathbf{b}}\}-1]e^{-\mathbf{b}^T\boldsymbol{\Sigma}^{-1}\mathbf{b}/2}d\mathbf{b}}{\int_{\mathbf{b}} R_{1i}(\boldsymbol{\beta}, H, \mathbf{b})e^{-\mathbf{b}^T\boldsymbol{\Sigma}^{-1}\mathbf{b}/2}d\mathbf{b}},\\
A_{n2}[\mathbf{h}_2] &= n^{-1}\sum_{i=1}^{n}\frac{\int_{\mathbf{b}} R_{1i}(\boldsymbol{\beta}, H, \mathbf{b})e^{-\mathbf{b}^T\boldsymbol{\Sigma}^{-1}\mathbf{b}/2}\{\mathbf{b}^T\boldsymbol{\Sigma}^{-1}\mathcal{D}(\mathbf{h}_2)\boldsymbol{\Sigma}^{-1}\mathbf{b}/2-\boldsymbol{\Sigma}^{-1}\cdot\mathcal{D}(\mathbf{h}_2)/2\}d\mathbf{b}}{\int_{\mathbf{b}} R_{1i}(\boldsymbol{\beta}, H, \mathbf{b})e^{-\mathbf{b}^T\boldsymbol{\Sigma}^{-1}\mathbf{b}/2}d\mathbf{b}},\\
A_{n3}[h_3] &= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\Bigg\{\Delta_{ij}h_3(Y_{ij})\\
&\quad -(1+\Delta_{ij})\int_0^{Y_{ij}} h_3(y)dH(y)\frac{\int_{\mathbf{b}} R_{1i}(\boldsymbol{\beta}, H, \mathbf{b})/(H(Y_{ij})+e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}+\mathbf{Z}_{ij}^T\mathbf{b})})e^{-\mathbf{b}^T\boldsymbol{\Sigma}^{-1}\mathbf{b}/2}d\mathbf{b}}{\int_{\mathbf{b}} R_{1i}(\boldsymbol{\beta}, H, \mathbf{b})e^{-\mathbf{b}^T\boldsymbol{\Sigma}^{-1}\mathbf{b}/2}d\mathbf{b}}\Bigg\}.
\end{aligned}
$$

Correspondingly, we define the limit map $S : (\boldsymbol{\beta}, \boldsymbol{\Sigma}, H) \to l^\infty(\mathcal{H})$ as

$$S(\boldsymbol{\beta}, \boldsymbol{\Sigma}, H)[\mathbf{h}_1, \mathbf{h}_2, h_3] = A_1[\mathbf{h}_1] + A_2[\mathbf{h}_2] + A_3[h_3],$$

where the linear functionals A_p, $p = 1, 2, 3$, are obtained by replacing the the empirical sum in the A_{np} by the expectation. Clearly, $S_n(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\Sigma}}_n, \widehat{H}_n) = 0$, and $S(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, H_0) = 0$.

The desired asymptotic normality will follow if we can verify the four conditions stated in Theorem 2 of Murphy (1995). The first condition that $\sqrt{n}(S_n(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, H_0) - S(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, H_0))$ weakly converges to a tight Gaussian process on $l^\infty(\mathcal{H})$ holds since $\mathcal{H}$ is a Donsker class and the functionals A_{np} are bounded Lipschitz functionals with respect to $\mathcal{H}$. By the smoothness of $S(\boldsymbol{\beta}, \boldsymbol{\Sigma}, H)$, the Fréchet differentiability holds and the derivative of $S(\boldsymbol{\beta}, \boldsymbol{\Sigma}, H)$ at $(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, H_0)$, denoted by $\dot{S}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, H_0)$, is a map from the space

$$\{(\boldsymbol{\beta}-\boldsymbol{\beta}_0, \boldsymbol{\Sigma}-\boldsymbol{\Sigma}_0, H-H_0) : (\boldsymbol{\beta}, \boldsymbol{\Sigma}, H) \text{ is in the neighborhood } \mathcal{U} \text{ of } (\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, H_0)\}$$

to $l^\infty(\mathcal{H})$. The approximation condition that

$$
\begin{aligned}
&\sup_{(\mathbf{h}_1,\mathbf{h}_2,h_3)\in\mathcal{H}}\left|(S_n-S)(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\Sigma}}_n, \widehat{H}_n)[\mathbf{h}_1, \mathbf{h}_2, h_3] - (S_n-S)(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, H_0)[\mathbf{h}_1, \mathbf{h}_2, h_3]\right|\\
&= o_p\left(n^{-1/2} \vee \left\{\|\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\| + \|\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{\Sigma}_0\| + \sup_{y\in[0,\tau]}|\widehat{H}_n(y)-H_0(y)|\right\}\right)
\end{aligned}
$$

can be verified along the lines of Murphy (1995, appendix).

It remains to show that the linear map $\dot{S}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, H_0)$, denoted by Λ, is continuously invertible on its range. Note that Λ maps $(\boldsymbol{\beta} - \boldsymbol{\beta}_0, \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0, H - H_0)$ to a bounded functional on $\mathcal{H}$. Algebraic manipulations yield

$$\begin{aligned} & \Lambda(\boldsymbol{\beta} - \boldsymbol{\beta}_0, \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0, H - H_0)[\mathbf{h}_1, \mathbf{h}_2, h_3] \\ = \; & (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \mathcal{Q}_1(\mathbf{h}_1, \mathbf{h}_2, h_3) + (\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0)^T \mathcal{Q}_2(\mathbf{h}_1, \mathbf{h}_2, h_3) + \int_0^\tau \mathcal{Q}_3(\mathbf{h}_1, \mathbf{h}_2, h_3) d(H - H_0), \end{aligned}$$

where

$$\begin{aligned} \mathcal{Q}_1(\mathbf{h}_1, \mathbf{h}_2, h_3) &= \mathbf{B}_1 \begin{pmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{pmatrix} + \int_0^\tau h_3(y) D_1(y) dy, \\ \mathcal{Q}_2(\mathbf{h}_1, \mathbf{h}_2, h_3) &= \mathbf{B}_2 \begin{pmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{pmatrix} + \int_0^\tau h_3(y) D_2(y) dy, \\ \mathcal{Q}_3(\mathbf{h}_1, \mathbf{h}_2, h_3) &= \mathbf{B}_3 \begin{pmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{pmatrix} + b_4 h_3(y) + \int_0^\tau h_3(t) D_3(t, y) dt, \end{aligned}$$

$\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3$ are constant matrices, $D_1(y), D_2(y), D_3(t, y)$ are continuously differentiable functions depending on the true densities, and $b_4 > 0$. Therefore, the operator $\mathcal{Q}(\mathbf{h}_1, \mathbf{h}_2, h_3) \equiv (\mathcal{Q}_1(\mathbf{h}_1, \mathbf{h}_2, h_3), \mathcal{Q}_2(\mathbf{h}_1, \mathbf{h}_2, h_3), \mathcal{Q}_3(\mathbf{h}_1, \mathbf{h}_2, h_3))^T$ can be considered as a sum of a continuously invertible linear operator and a compact operator from the linear span of $\mathcal{H}$ to itself.

The invertibility of $\Lambda \equiv \dot{S}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, H_0)$ is equivalent to the invertibility of the linear operator $\mathcal{Q}(\mathbf{h}_1, \mathbf{h}_2, h_3)$. It suffices to prove that $\mathcal{Q}(\mathbf{h}_1, \mathbf{h}_2, h_3)$ is a one-to-one map (Rudin 1973, pp. 99-103). If $\mathcal{Q}(\mathbf{h}_1, \mathbf{h}_2, h_3) = \mathbf{0}$, then $\Lambda(\boldsymbol{\beta} - \boldsymbol{\beta}_0, \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0, H - H_0)[\mathbf{h}_1, \mathbf{h}_2, h_3] = \mathbf{0}$ for any $(\boldsymbol{\beta}, \boldsymbol{\Sigma}, H)$ in the neighborhood $\mathcal{U}$. In particular, we choose

$$\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \epsilon \mathbf{h}_1, \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 + \epsilon \mathbf{h}_2, H(y) = H_0(y) + \epsilon \int_0^y h_3(t) dH_0(t)$$

for a small constant ϵ. By the definition of Λ,

$$0 = \Lambda(\boldsymbol{\beta} - \boldsymbol{\beta}_0, \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0, H - H_0)[\mathbf{h}_1, \mathbf{h}_2, h_3] = \epsilon E\{(l_{\boldsymbol{\beta}}[\mathbf{h}_1] + l_{\boldsymbol{\Sigma}}[\mathbf{h}_2] + l_H[h_3])^2\}.$$

Thus, $l_{\boldsymbol{\beta}}[\mathbf{h}_1] + l_{\boldsymbol{\Sigma}}[\mathbf{h}_2] + l_H[h_3] = 0$ almost surely. After writing out the expression of this equation, we obtain

$$\begin{aligned} & \sum_{j=1}^{n_i} \int_{\mathbf{b}} R_{4i}(\boldsymbol{\beta}_0, H_0, \mathbf{b}) \left[\mathbf{X}_{ij}^T \mathbf{h}_1 \left\{ -1 + \frac{(1 + \Delta_{ij})}{(H_0(Y_{ij}) e^{(\mathbf{X}_{ij}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{ij}^T \mathbf{b})} + 1)} \right\} \right] d_{\mathbf{b}} N(0, \Sigma_0) \\ & + \int_{\mathbf{b}} \left\{ -\frac{1}{2} {\boldsymbol{\Sigma}_0}^{-1} \cdot \mathcal{D}(\mathbf{h}_2) + \frac{1}{2} \mathbf{b}^T {\boldsymbol{\Sigma}_0}^{-1} \mathcal{D}(\mathbf{h}_2) {\boldsymbol{\Sigma}_0}^{-1} \mathbf{b} \right\} R_{4i}(\boldsymbol{\beta}_0, H_0, \mathbf{b}) d_{\mathbf{b}} N(0, \Sigma_0) \\ & + \sum_{j=1}^{n_i} \int_{\mathbf{b}} R_{4i}(\boldsymbol{\beta}_0, H_0, \mathbf{b}) \left\{ \Delta_{ij} h_3(Y_{ij}) - \frac{(1 + \Delta_{ij}) \int_0^{Y_{ij}} h_3(y) dH_0(y)}{(H_0(Y_{ij}) + e^{-(\mathbf{X}_{ij}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{ij}^T \mathbf{b})})} \right\} d_{\mathbf{b}} N(0, \Sigma_0) = 0, \end{aligned} \quad (A.12)$$

where $R_{4i}(\boldsymbol{\beta}, H, \mathbf{b}) = R_{1i}(\boldsymbol{\beta}, H, \mathbf{b}) \prod_{j=1}^{n_i} \{H_0'(Y_{ij})\}^{\Delta_{ij}}$.

We will show that (A.12) entails $\mathbf{h}_1 = \mathbf{0}, \mathbf{h}_2 = \mathbf{0}$ and $h_3 = 0$ by adopting the ideas used in the proof of the identifiability for Theorem 1. First, we let $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ be fixed. Then for a fixed k such that $1 \le k \le n_i$, we define measures $\mu_1, \ldots \mu_{n_i}$ on the set $\{0,1\} \times [0,\tau]$ as follows: for any Borel set $A \subset [0,\tau]$,

$$\mu_m(\{0\} \times A) = 0, \quad \mu_m(\{1\} \times A) = I(0 \in A), \quad m \le k,$$

$$\mu_m(\{0\} \times A) = I(\tau \in A), \quad \mu_m(\{1\} \times A) = \int I_A dx, \quad m > k.$$

We integrate both sides of (A.12) over $\{(\Delta_{i,1}, Y_1), \ldots, (\Delta_{i,n_i}, Y_{i,n_i})\}$ with respect to the product measure $d\mu_1 \cdots d\mu_{n_i}$. In other words, we let $\Delta_{im} = 1$ and $Y_{im} = 0$ for $m \le k$; we sum all the equalities of (A.12) for all possible combinations of $\{\Delta_{i,k+1}, \ldots, \Delta_{i,n_i}\} \in \{0,1\}^{n_i - k}$, in which we choose $Y_{im} = \tau$ if $\Delta_{im} = 0$ and integrate Y_{im} from 0 to τ if $\Delta_{im} = 1$. The resulting integration is 0.

We study the integral of each term on the left-hand side of (A.12) with respect to the measure $\prod_{m=1}^{n_i} \mu_m$. For the first term on the left-hand side of (A.12), from the expression of $R_{4i}(\boldsymbol{\beta}_0, H_0, \mathbf{b})$, we have that for any $\mathbf{b}$, if $j \le k$,

$$\int \left[R_{4i}(\boldsymbol{\beta}_0, H_0, \mathbf{b}) \mathbf{X}_{ij}^T \mathbf{h}_1 \left\{ -1 + \frac{(1+\Delta_{ij})}{(H_0(Y_{ij}) e^{(\mathbf{X}_{ij}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{ij}^T \mathbf{b})} + 1)} \right\} \right] d(\prod_{m=1}^{n_i} \mu_m)$$

$$= \mathbf{X}_{ij}^T \mathbf{h}_1 \prod_{m \le k} \left\{ H_0'(0) e^{\mathbf{X}_{im}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{im}^T \mathbf{b}} \right\}$$

$$\times \left[\sum_{\delta_{im} \in \{0,1\}, m>k} \prod_{m>k} \left\{ \frac{1}{(H_0(\tau) e^{\mathbf{X}_{im}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{im}^T \mathbf{b}} + 1)^{1-\delta_{im}}} \left(1 - \frac{1}{H_0(\tau) e^{\mathbf{X}_{im}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{im}^T \mathbf{b}} + 1} \right)^{\delta_{im}} \right\} \right]$$

$$= \mathbf{X}_{ij}^T \mathbf{h}_1 \prod_{m \le k} \left\{ H_0'(0) e^{\mathbf{X}_{im}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{im}^T \mathbf{b}} \right\};$$

if $j > k$, it holds that

$$\int \left[R_{4i}(\boldsymbol{\beta}_0, H_0, \mathbf{b}) \mathbf{X}_{ij}^T \mathbf{h}_1 \left\{ -1 + \frac{(1+\Delta_{ij})}{(H_0(Y_{ij}) e^{(\mathbf{X}_{ij}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{ij}^T \mathbf{b})} + 1)} \right\} \right] d(\prod_{m=1}^{n_i} \mu_m)$$

$$= \mathbf{X}_{ij}^T \mathbf{h}_1 \prod_{m \le k} \left\{ H_0'(0) e^{\mathbf{X}_{im}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{im}^T \mathbf{b}} \right\}$$

$$\times \left[\sum_{\delta_{im} \in \{0,1\}, m>k, m \ne j} \prod_{m>k, m \ne j} \left\{ \frac{1}{(H_0(\tau) e^{\mathbf{X}_{im}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{im}^T \mathbf{b}} + 1)^{1-\delta_{im}}} \left(1 - \frac{1}{H_0(\tau) e^{\mathbf{X}_{im}^T \boldsymbol{\beta}_0 + \mathbf{Z}_{im}^T \mathbf{b}} + 1} \right)^{\delta_{im}} \right\} \right]$$

$$\times \sum_{\delta_{ij}\in\{0,1\}} \left\{ (1-\delta_{ij}) \frac{1}{(H_0(\tau)e^{\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b}}+1)} \left(-1 + \frac{1}{H_0(\tau)e^{\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b}}+1} \right) \right.$$

$$\left. + \delta_{ij} \int_0^\tau \frac{H_0'(t)e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b})}}{(H_0(t)+e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b})})^2} \left(-1 + \frac{2}{H_0(t)e^{\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b}}+1} \right) dt \right\}$$

$$= \mathbf{X}_{ij}^T\mathbf{h}_1 \prod_{m\le k} \left\{ H_0'(0)e^{\mathbf{X}_{im}^T\boldsymbol{\beta}_0+\mathbf{Z}_{im}^T\mathbf{b}} \right\}$$

$$\times \sum_{\delta_{ij}\in\{0,1\}} \left\{ (1-\delta_{ij}) \frac{1}{(H_0(\tau)e^{\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b}}+1)} \left(-1 + \frac{1}{H_0(\tau)e^{\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b}}+1} \right) \right.$$

$$\left. + \delta_{ij} \int_0^\tau \frac{H_0'(t)e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b})}}{(H_0(t)+e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b})})^2} \left(-1 + \frac{2}{H_0(t)e^{\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b}}+1} \right) dt \right\}$$

$$= 0.$$

Therefore,

$$\int \sum_{j=1}^{n_i} \int_{\mathbf{b}} R_{4i}(\boldsymbol{\beta}_0, H_0, \mathbf{b}) \left[\mathbf{X}_{ij}^T\mathbf{h}_1 \left\{ -1 + \frac{(1+\Delta_{ij})}{(H_0(Y_{ij})e^{(\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b})}+1)} \right\} \right] d_{\mathbf{b}}N(0,\Sigma_0) d(\prod_{m=1}^{n_i} \mu_m)$$

$$= \sum_{j\le k} \mathbf{X}_{ij}^T\mathbf{h}_1 \int_{\mathbf{b}} \prod_{m\le k} \left\{ H_0'(0)e^{\mathbf{X}_{im}^T\boldsymbol{\beta}_0+\mathbf{Z}_{im}^T\mathbf{b}} \right\} d_{\mathbf{b}}N(0,\Sigma_0). \tag{A.13}$$

Likewise,

$$\int\int_{\mathbf{b}} \left\{ -\frac{1}{2}{\boldsymbol{\Sigma}_0}^{-1} \cdot \mathcal{D}(\mathbf{h}_2) + \frac{1}{2}\mathbf{b}^T{\boldsymbol{\Sigma}_0}^{-1}\mathcal{D}(\mathbf{h}_2){\boldsymbol{\Sigma}_0}^{-1}\mathbf{b} \right\} R_{4i}(\boldsymbol{\beta}_0, H_0, \mathbf{b}) d_{\mathbf{b}}N(0,\Sigma_0) d(\prod_{m=1}^{n_i} \mu_m)$$

$$= \int_{\mathbf{b}} \left\{ -\frac{1}{2}{\boldsymbol{\Sigma}_0}^{-1} \cdot \mathcal{D}(\mathbf{h}_2) + \frac{1}{2}\mathbf{b}^T{\boldsymbol{\Sigma}_0}^{-1}\mathcal{D}(\mathbf{h}_2){\boldsymbol{\Sigma}_0}^{-1}\mathbf{b} \right\} \prod_{m\le k} \left\{ H_0'(0)e^{\mathbf{X}_{im}^T\boldsymbol{\beta}_0+\mathbf{Z}_{im}^T\mathbf{b}} \right\} d_{\mathbf{b}}N(0,\Sigma_0). \tag{A.14}$$

Furthermore, if $j \le k$,

$$\int R_{4i}(\boldsymbol{\beta}_0, H_0, \mathbf{b}) \left\{ \Delta_{ij}h_3(Y_{ij}) - \frac{(1+\Delta_{ij})\int_0^{Y_{ij}} h_3(y)dH_0(y)}{H_0(Y_{ij})+e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b})}} \right\} d(\prod_{m=1}^{n_i} \mu_m)$$

$$= h_3(0) \prod_{m\le k} \left\{ H_0'(0)e^{\mathbf{X}_{im}^T\boldsymbol{\beta}_0+\mathbf{Z}_{im}^T\mathbf{b}} \right\};$$

if $j > k$,

$$\int R_{4i}(\boldsymbol{\beta}_0, H_0, \mathbf{b}) \left\{ \Delta_{ij}h_3(Y_{ij}) - \frac{(1+\Delta_{ij})\int_0^{Y_{ij}} h_3(y)dH_0(y)}{H_0(Y_{ij})+e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b})}} \right\} d(\prod_{m=1}^{n_i} \mu_m)$$

$$= \prod_{m\le k} \left\{ H_0'(0)e^{\mathbf{X}_{im}^T\boldsymbol{\beta}_0+\mathbf{Z}_{im}^T\mathbf{b}} \right\} \left[\sum_{\delta_{ij}\in\{0,1\}} \left\{ -(1-\delta_{ij}) \frac{1}{H_0(\tau)e^{\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b}}+1} \frac{\int_0^\tau h_3(y)dH_0(y)}{H_0(\tau)+e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b})}} \right. \right.$$

$$+\delta_{ij}\int_0^\tau \frac{H_0'(t)e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b})}}{(H_0(t)+e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b})})^2}\left(h_3(t)-\frac{2\int_0^t h_3(y)dH_0(y)}{H_0(t)+e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b})}}\right)dt\Bigg\}\Bigg]=0.$$

Thus,

$$\int\sum_{j=1}^{n_i}\int_{\mathbf{b}} R_{4i}(\boldsymbol{\beta}_0,H_0,\mathbf{b})\left\{\Delta_{ij}h_3(Y_{ij})-\frac{(1+\Delta_{ij})\int_0^{Y_{ij}}h_3(y)dH_0(y)}{(H_0(Y_{ij})+e^{-(\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b})})}\right\}d_{\mathbf{b}}N(0,\Sigma_0)d(\prod_{m=1}^{n_i}\mu_m)$$

$$=\sum_{j\le k}h_3(0)\int_{\mathbf{b}}\prod_{m\le k}\left\{H_0'(0)e^{\mathbf{X}_{im}^T\boldsymbol{\beta}_0+\mathbf{Z}_{im}^T\mathbf{b}}\right\}d_{\mathbf{b}}N(0,\Sigma_0). \qquad (A.15)$$

Combining (A.13), (A.14) and (A.15) and integrating over $\mathbf{b}$, we obtain

$$\sum_{j=1}^{k}\mathbf{X}_{ij}^T\mathbf{h}_1+\frac{1}{2}\Big(\sum_{j=1}^{k}\mathbf{Z}_{ij}\Big)^T\mathcal{D}(\mathbf{h}_2)\Big(\sum_{j=1}^{k}\mathbf{Z}_{ij}\Big)+kh_3(0)=0.$$

Since the order for the subscripts $j=1,\ldots,k$ is arbitrary, it holds that

$$\sum_{j=k_1+1}^{k_2}\mathbf{X}_{ij}^T\mathbf{h}_1+\frac{1}{2}\Big(\sum_{j=k_1+1}^{k_2}\mathbf{Z}_{ij}\Big)^T\mathcal{D}(\mathbf{h}_2)\Big(\sum_{j=k_1+1}^{k_2}\mathbf{Z}_{ij}\Big)+(k_2-k_1)h_3(0)=0$$

for any $1\le k_1<k_2\le n_i$. Thus, $\mathbf{Z}_{ij}^T\mathcal{D}(\mathbf{h}_2)\mathbf{Z}_{ij'}=0$ for $j\ne j'$ and $\mathbf{X}_{ij}^T\mathbf{h}_1+\mathbf{Z}_{ij}^T\mathcal{D}(\mathbf{h}_2)\mathbf{Z}_{ij}/2+h_3(0)=0$. By condition C.3, $\mathcal{D}(\mathbf{h}_2)=0$. As a result, $\mathbf{h}_2=\mathbf{0}$ and $\mathbf{h}_1=\mathbf{0}$. In (A.13), we set $Y_{ij}=0$, $j=2,\ldots,n_i$, and $\Delta_{ij}=1$, $j=1,\ldots,n_i$, so as to obtain

$$h_3(Y_{i1})=$$
$$2\int_0^{Y_{i1}}h_3(y)dH_0(y)\frac{\int_{\mathbf{b}}e^{-(\mathbf{X}_{i1}^T\boldsymbol{\beta}_0+\mathbf{Z}_{i1}^T\mathbf{b})+\sum_{j=2}^{n_i}(\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b})}e^{-\mathbf{b}^T\boldsymbol{\Sigma}_0^{-1}\mathbf{b}/2}/(H_0(Y_{i1})+e^{-(\mathbf{X}_{i1}^T\boldsymbol{\beta}_0+\mathbf{Z}_{i1}^T\mathbf{b})})^3d\mathbf{b}}{\int_{\mathbf{b}}e^{-(\mathbf{X}_{i1}^T\boldsymbol{\beta}_0+\mathbf{Z}_{i1}^T\mathbf{b})+\sum_{j=2}^{n_i}(\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b})}e^{-\mathbf{b}^T\boldsymbol{\Sigma}_0^{-1}\mathbf{b}/2}/(H_0(Y_{i1})+e^{-(\mathbf{X}_{i1}^T\boldsymbol{\beta}_0+\mathbf{Z}_{i1}^T\mathbf{b})})^2d\mathbf{b}}.$$

That is, $g(y)\equiv\int_0^y h_3(t)dH_0(t)$ satisfies the homogeneous equation

$$\frac{g'(y)}{H_0'(y)}-g(y)\frac{\int_{\mathbf{b}}e^{-(\mathbf{X}_{i1}^T\boldsymbol{\beta}_0+\mathbf{Z}_{i1}^T\mathbf{b})+\sum_{j=2}^{n_i}(\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b})}e^{-\mathbf{b}^T\boldsymbol{\Sigma}_0^{-1}\mathbf{b}/2}/(H_0(Y_{i1})+e^{-(\mathbf{X}_{i1}^T\boldsymbol{\beta}_0+\mathbf{Z}_{i1}^T\mathbf{b})})^3d\mathbf{b}}{\int_{\mathbf{b}}e^{-(\mathbf{X}_{i1}^T\boldsymbol{\beta}_0+\mathbf{Z}_{i1}^T\mathbf{b})+\sum_{j=2}^{n_i}(\mathbf{X}_{ij}^T\boldsymbol{\beta}_0+\mathbf{Z}_{ij}^T\mathbf{b})}e^{-\mathbf{b}^T\boldsymbol{\Sigma}_0^{-1}\mathbf{b}/2}/(H_0(Y_{i1})+e^{-(\mathbf{X}_{i1}^T\boldsymbol{\beta}_0+\mathbf{Z}_{i1}^T\mathbf{b})})^2d\mathbf{b}}=0$$

with boundary condition $g(0)=0$. Thus, it is clear that $g(y)=0$, i.e., $h_3(y)=0$. Hence, we have verified that $\mathcal{Q}$ is one-to-one map and have thus shown the invertibility of $\dot{S}(\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_0,H_0)$.

The asymptotic distribution stated in Theorem 2 now follows from Theorem 2 of Murphy (1995). Furthermore,

$$\begin{aligned}&\sqrt{n}\dot{S}(\boldsymbol{\beta}_0,\boldsymbol{\Sigma}_0,H_0)(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0,\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{\Sigma}_0,\widehat{H}_n-H_0)[\mathbf{h}_1,\mathbf{h}_2,h_3]\\ =\;&\sqrt{n}(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0)^T\mathcal{Q}_1(h)+\sqrt{n}(\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{\Sigma}_0)^T\mathcal{Q}_2(h)+\sqrt{n}\int_0^\tau\mathcal{Q}_3(h)d(\widehat{H}_n-H_0)\\ =\;&\sqrt{n}(\mathcal{P}_n-\mathcal{P})[\mathbf{h}_1^Tl_{\boldsymbol{\beta}}+\mathbf{h}_2^Tl_{\boldsymbol{\Sigma}}+l_H[h_3]]+o_p(1) \qquad (A.16)\end{aligned}$$

uniformly in $\mathbf{h}_1$, $\mathbf{h}_2$ and h_3, where $h = (\mathbf{h}_1, \mathbf{h}_2, h_3)$. Since it has already been shown that $\mathcal{Q} \equiv (\mathcal{Q}_1, \mathcal{Q}_2, \mathcal{Q}_3)^T$ is invertible, we can find $N \equiv d_1 + d_2(d_2+1)/2$ unique directions $\omega_1 \equiv (\omega_{11}, \omega_{12}, \omega_{13}), \ldots, \omega_N \equiv (\omega_{N1}, \omega_{N2}, \omega_{N3}) \in \mathcal{H}$ such that

$$\begin{aligned} & (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^T(\mathcal{Q}_1(\omega_1), \ldots, \mathcal{Q}_1(\omega_N)) + (\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_0)^T(\mathcal{Q}_2(\omega_1), \ldots, \mathcal{Q}_2(\omega_N)) \\ & + \int_0^\tau (\mathcal{Q}_3(\omega_1), \ldots, \mathcal{Q}_3(\omega_N)) d(\widehat{H}_n - H_0) \\ = \; & ((\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^T, (\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_0)^T). \end{aligned}$$

For such ω's,

$$\begin{aligned} & \sqrt{n}((\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^T, (\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_0)^T) \\ = \; & \sqrt{n}(\mathcal{P}_n - \mathcal{P})(\omega_{11}^T l_{\boldsymbol{\beta}} + \omega_{12}^T l_{\boldsymbol{\Sigma}} + l_H[\omega_{13}], \ldots, \omega_{N1}^T l_{\boldsymbol{\beta}} + \omega_{N2}^T l_{\boldsymbol{\Sigma}} + l_H[\omega_{N3}]) + o_p(1). \end{aligned}$$

Thus, $\widehat{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{\Sigma}}_n$ are asymptotically linear estimators for $\boldsymbol{\beta}_0$ and $\boldsymbol{\Sigma}_0$, respectively, and their influence functions belong to the space spanned by the score functions. It follows that $(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\Sigma}}_n)$ are semiparametrically efficient (Bickel et al. 1993, Ch. 3).

A.3. Proof of Theorem 3

The proof of Theorem 3 parallels the proof of Theorem 3 in Parner (1998) and will be kept brief. The left-hand side of equation (A.16) is equal to $\sqrt{n}$ times the expectation of the second derivative of the log-likelihood function along the directions of $(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0, \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_0, \widehat{H}_n - H_0)$ and the direction $(\mathbf{h}_1, \mathbf{h}_2, \int h_3 dH_0)$. This second derivative can be approximated uniformly in $(\mathbf{h}_1, \mathbf{h}_2, h_3) \in \mathcal{H}$ by

$$(\,\mathbf{h}_1^T, \mathbf{h}_2^T, \vec{h}_3^T\,)\,(\mathbf{J}_n/n) \begin{pmatrix} \widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 \\ \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_0 \\ \left\{\widehat{H}_n\{Y_{ij}\} - \delta H_0(Y_{ij}) : \Delta_{ij} = 1\right\} \end{pmatrix},$$

where $\vec{h}_3$ denotes the vector of $\{h(Y_{ij}) : \Delta_{ij} = 1\}$, and $\delta H_0(Y_{ij}) = H_0(Y_{ij}) - \max_{Y_{kl} < Y_{ij}, \Delta_{kl}=1} H_0(Y_{kl})$. On the other hand, for large n, the distribution of the right-hand side of (A.14) approximates $(\,\mathbf{h}_1^T, \mathbf{h}_2^T, \vec{h}_3^T\,)\,(\mathbf{J}_n/n)^{1/2}\mathbf{G}$, where $\mathbf{G}$ is standard multivariate normal. It follows that

$$\sqrt{n}\,(\,\mathbf{h}_1^T, \mathbf{h}_2^T, \vec{h}_3^T\,)\,(\mathbf{J}_n/n) \begin{pmatrix} \widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 \\ \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_0 \\ \left\{\widehat{H}_n\{Y_{ij}\} - \delta H_0(Y_{ij}) : \Delta_{ij} = 1\right\} \end{pmatrix} \overset{d}{\approx} (\,\mathbf{h}_1^T, \mathbf{h}_2^T, \vec{h}_3^T\,)\,(\mathbf{J}_n/n)^{1/2}\mathbf{G},$$

where "$X \overset{d}{\approx} Y$" means that X and Y have the same asymptotic distribution. The replacement of $(\,\mathbf{h}_1^T, \mathbf{h}_2^T, \vec{h}_3^T\,)$ by $(\,\mathbf{h}_1^T, \mathbf{h}_2^T, \vec{h}_3^T\,)\,(\mathbf{J}_n/n)^{-1}$ yields

$$\sqrt{n}\,(\,\mathbf{h}_1^T, \mathbf{h}_2^T, \vec{h}_3^T\,) \begin{pmatrix} \widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 \\ \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_0 \\ \left\{\widehat{H}_n\{Y_{ij}\} - \delta H_0(Y_{ij}) : \Delta_{ij} = 1\right\} \end{pmatrix} \overset{d}{\approx} (\,\mathbf{h}_1^T, \mathbf{h}_2^T, \vec{h}_3^T\,)\,(\mathbf{J}_n/n)^{-1/2}\mathbf{G}.$$

Thus, $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^T \mathbf{h}_1 + \sqrt{n}(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_0)^T \mathbf{h}_2 + \int_0^\tau h_3(t) d(\widehat{H}_n(t) - H_0(t))$ converges to a zero-mean normal distribution whose variance is the limit of $n\,(\,\mathbf{h}_1^T, \mathbf{h}_2^T, \vec{h}_3^T\,)\,\mathbf{J}_n^{-1}\,(\,\mathbf{h}_1^T, \mathbf{h}_2^T, \vec{h}_3^T\,)$. Hence, the conclusion of Theorem 3 holds.

REFERENCES

Bennett, S. (1983), "Analysis of Survival Data by the Proportional Odds Model," *Statistics in Medicine*, 2, 273-277.

Bickel, P. J. (1986), "Efficient Testing in a Class of Transformation Models," in *Papers on Semiparametric Models at the ISI Centenary Session*, Amsterdam, pp. 63-81.

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

Cai, T, Cheng, S. C., and Wei, L. J. (2002), "Semiparametric Mixed-effects Models for Clustered Failure Time Data," *Journal of the American Statistical Association*, 97, 514-522.

Chen, K., Jin, Z., and Ying, Z. (2002), "Semiparametric Analysis of Transformation Models With Censored Data," *Biometrika*, 89, 659-668.

Cheng, S. C., Wei, L. J., and Ying, Z. (1995), "Analysis of Transformation Models With Censored Data," *Biometrika*, 82, 835-845.

Coleman, T. F and Li, Y. (1996), "An Interior, Trust Region Approach for Nonlinear Minimization Subject to Bounds," *SIAM Journal on Optimization*, 6, 418-445.

Coleman, T. F. and Li, Y. (1994), "On the Convergence of Reflective Newton Methods for Large-Scale Nonlinear Minimization Subject to Bounds," *Mathematical Programming*, 67, 189-224.

Cox, D. R. (1972), "Regression Models and Life Tables" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 34, 187-220.

Cuzick, J. (1988), "Rank Regression," *The Annals of Statistics*, 16, 1369-1389.

Dabrowska, D. M. and Doksum, K. A. (1988), "Estimation and Testing in the Two-Sample Generalized Odds-Rate Model," *Journal of the American Statistical Association*, 83, 744-

749.

Davidian, M. and Giltinana, D. M. (1995), *Nonlinear Models for Repeated Measurement*, London: Chapman and Hall.

Huang, J. and Rossini, A. J. (1997), "Sieve Estimation for the Proportional Odds Failure-time Regression Model with Interval Censoring," *Journal of the American Statistical Association*, 92, 960-967.

Huster, W. J, Brookmeyer, R., and Self, S. G. (1989), "Modelling Paired Survival Data With Covariates," *Biometrics*, 45, 145-156.

Lam, K. F., Lee, Y. W., and Leung, T. L. (2002), "Modeling Multivariate Survival Data by a Semiparametric Random Effects Proportional Odds Models," *Biometrics*, 58, 316-323.

Lam, K. F. and Leung, T. L. (2001), "Marginal Likelihood Estimation for Proportional Odds Models with Right Censored Data," *Lifetime Data Analysis*, 7, 39-54.

Murphy, S. A. (1994), "Consistency in a Proportional Hazards Model Incorporating a Random Effect," *The Annals of Statistics*, 22, 712-731.

Murphy, S. A. (1995), "Asymptotic Theory for the Frailty Model," *The Annals of Statistics*, 23, 182-198.

Murphy, S. A. and van der Vaart, A. W. (2000), "On the Profile Likelihood ," *Journal of the American Statistical Association*, 449-465.

Murphy, S. A., Rossini, A. J., and van der Vaart, A. W. (1997), "Maximal Likelihood Estimate in The Proportional Odds Model," *Journal of the American Statistical Association*, 92, 968-976.

Parner, E. (1998), "Asymptotic Theory for the Correlated Gamma-Frailty Model," *The Annals of Statistics*, 26, 183-214.

Petersen, J. H. (1998), "An Additive Frailty Model for Correlated Life Times," *Biometrics*, 54, 646-661.

Pettitt, A. N. (1984), "Proportional Odds Models for Survival Data and Estimates Using Ranks," *Applied Statistics*, 33, 169-175.

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

Shen, X. (1998), " Proportional Odds Regression and Sieve Maximum Likelihood Estimation," *Biometrika*, 85, 165-177.

van der Vaart, A. W. and Wellner, J. A. (1995), *Weak Convergence and Empirical Processes*, New York: Springer-Verlag.

Wu, C. O. (1995), "Estimating the Real Parameter in a Two-Sample Proportional Odds Model," *The Annals of Statistics*, 23, 376-395.