# Supplementary Material for "Minorize-maximize algorithm for the generalized odds rate model for clustered current status data"

Tong WANG[1,2], Kejun HE[3,†], Wei MA[3], Dipankar BANDYOPADHYAY[4] and Samiran SINHA[2,*]

[1]School of Statistics and Data Science, Nankai University, Tianjin, China
[2]Department of Statistics, Texas A&M University, College Station, TX, U.S.A.
[3]Center for Applied Statistics, Institute of Statistics and Big Data, Renmin University of China, Beijing, China
[4]Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, U.S.A.

## S.1 Results of Section 3

### S.1.1 Proof of Theorem 1

#### S.1.1.1 Proof of part i)

*Proof.* Define $f(u) = \log[\{1-(1+ru)^{-1/r}\}/\{1-(1+ru_0)^{-1/r}\}]$, $r > 0, u > 0$. Then $f(u_0) = 0$. Define

$$A_1(u) = \frac{\partial f(u)}{\partial u} = \frac{(1+ru)^{-(1+1/r)}}{1-(1+ru)^{-1/r}}$$

and

$$A_2(u) = -0.5\frac{\partial^2 f(u)}{\partial u^2} = 0.5\left[\frac{(1+r)(1+ru)^{-(2+1/r)}}{1-(1+ru)^{-1/r}} + \frac{(1+ru)^{-(2+2/r)}}{\{1-(1+ru)^{-1/r}\}^2}\right].$$

Consider the Taylor series expansion of $f(u)$ about $u = u_0$,

$$f(u) = (u-u_0)A_1(u_0) - (u-u_0)^2 A_2(u^*),$$

for some $|u^* - u| < |u_0 - u|$. Clearly, $A_2(u)$ is decreasing in $u$ for any $r > 0$. Therefore, for $u > u^* > u_0$ and any $r > 0$, $A_2(u) < A_2(u^*) < A_2(u_0)$, and then

$$\begin{aligned} f(u) &\geq (u-u_0)A_1(u_0) - (u-u_0)^2 A_2(u_0) \\ &= (u-u_0)A_1(u_0) - (u-u_0)^2 A_2(u_0) + \kappa\left\{\log\left(\frac{u_0}{u}\right) + \log\left(\frac{u}{u_0}\right)\right\} \end{aligned}$$

†Both the first and second authors contributed equally.
*Corresponding author: sinha@stat.tamu.edu

$$\geq \quad (u-u_0)A_1(u_0) - (u-u_0)^2 A_2(u_0) + \kappa \left\{ \log\left(\frac{u_0}{u}\right) + \left(1 - \frac{u_0}{u}\right) \right\}. \qquad \text{(S.1)}$$

In fact the result (S.1) holds for either choices of $\kappa$ mentioned in Theorem 1.

To prove the result when $u < u_0$, let us define $g(u) = f(u) - (u-u_0)A_1(u_0) + (u-u_0)^2 A_2(u_0) - \kappa\{\log(u_0/u) + 1 - u_0/u\}$. Then

$$\begin{aligned} g^{'}(u) &= A_1(u) - A_1(u_0) + 2(u-u_0)A_2(u_0) + \kappa \frac{u-u_0}{u^2} \\ &= (u-u_0)\left\{ -2A_2(u_\dagger) + 2A_2(u_0) + \frac{\kappa}{u^2} \right\}, \end{aligned}$$

where the last equality is obtained by applying the Taylor series expansion on $A_1(u)$ about $u = u_0$, $A_1(u) = A_1(u_0) - 2(u-u_0)A_2(u_\dagger)$, for some $u_\dagger \in [u, u_0)$.

Next we consider the case of $0 < r \leq 1$ with $\kappa = 1/r$. Define

$$\begin{aligned} B_1 \equiv \frac{1}{ru^2} - 2A_2(u_\dagger) &= \frac{1}{ru^2} - \frac{(1+ru_\dagger)^{-1/r-2}[1 + r\{1 - (1+ru_\dagger)^{-1/r}\}]}{\{1 - (1+ru_\dagger)^{-1/r}\}^2} \\ &= \frac{\{1 - (1+ru_\dagger)^{-1/r}\}^2 - ru^2(1+ru_\dagger)^{-1/r-2}[1 + r\{1 - (1+ru_\dagger)^{-1/r}\}]}{ru^2\{1 - (1+ru_\dagger)^{-1/r}\}^2}, \end{aligned}$$

and $B_2 = B_1 ru^2\{1 - (1+ru_\dagger)^{-1/r}\}^2$. Then

$$\begin{aligned} B_2 &= \{1 - (1+ru_\dagger)^{-1/r}\}\left[1 - (1+ru_\dagger)^{-1/r} - r^2u^2(1+ru_\dagger)^{-1/r-2}\right] - ru^2(1+ru_\dagger)^{-1/r-2} \\ &= \left\{ \frac{1 - (1+ru_\dagger)^{-1/r}}{(1+ru_\dagger)^{1/r+2}} \right\} \left\{ (1+ru_\dagger)^{1/r+2} - (1+ru_\dagger)^2 - r^2u^2 \right\} - ru^2(1+ru_\dagger)^{-1/r-2} \\ &= \frac{1}{(1+ru_\dagger)^{1/r+2}}\left[ \{1 - (1+ru_\dagger)^{-1/r}\}\left\{(1+ru_\dagger)^{1/r+2} - (1+ru_\dagger)^2 - r^2u^2\right\} - ru^2 \right]. \end{aligned}$$

Using the Bernoulli inequality we have for $0 < r \leq 1$,

$$\begin{aligned} (1+ru_\dagger)^{1/r+2} = (1+ru_\dagger)^{1/r}(1+ru_\dagger)^2 &\geq (1+u_\dagger)(1+ru_\dagger)^2 \\ &= (1+ru_\dagger)^2 + u_\dagger(1+ru_\dagger)^2. \end{aligned}$$

Now, using this inequality in the numerator of $B_2$ we obtain

$$\begin{aligned} B_2 &\geq \frac{1}{(1+ru_\dagger)^{1/r+2}}\left[ \{1 - (1+ru_\dagger)^{-1/r}\}\left\{(1+ru_\dagger)^2 + u_\dagger(1+ru_\dagger)^2 - (1+ru_\dagger)^2 - r^2u^2\right\} - ru^2 \right] \\ &= \frac{1}{(1+ru_\dagger)^{1/r+2}}\left[ \left\{1 - \frac{1}{(1+ru_\dagger)^{1/r}}\right\}\left\{u_\dagger(1+ru_\dagger)^2 - r^2u^2\right\} - ru^2 \right] \\ &\geq \frac{1}{(1+ru_\dagger)^{1/r+2}}\left[ \left(1 - \frac{1}{1+u_\dagger}\right)\left\{u_\dagger(1+ru_\dagger)^2 - r^2u^2\right\} - ru^2 \right] \end{aligned}$$

$$\begin{aligned} &= \frac{1}{(1+ru_\dagger)^{1/r+2}}\left[\left(\frac{u_\dagger}{1+u_\dagger}\right)\{u_\dagger(1+ru_\dagger)^2 - r^2u^2\} - ru^2\right] \\ &= \frac{B_3}{(1+ru_\dagger)^{1/r+2}(1+u_\dagger)}, \end{aligned}$$

where $B_3 = u_\dagger^2(1+ru_\dagger)^2 - r^2u^2u_\dagger - ru^2(1+u_\dagger)$. This last inequality pertaining to $B_2$ holds due to the application of the Bernoulli inequality for $0 < r \leq 1$ and $u_\dagger(1+ru_\dagger)^2 - r^2u^2 > 0$ for $u < u_\dagger$. Since $0 < r \leq 1$, $u_\dagger(1+ru_\dagger)^2 - r^2u^2 \geq u_\dagger(1+ru_\dagger)^2 - ru_\dagger^2 = u_\dagger + r^2u_\dagger^3 + 2r(u_\dagger^2 - u^2) > 0$. Now, we have

$$\begin{aligned} B_3 &= u_\dagger^2(1+ru_\dagger)^2 - r^2u^2u_\dagger - ru^2(1+u_\dagger) \\ &= u_\dagger^2 + r^2u_\dagger^4 + 2ru_\dagger^3 - r^2u^2u_\dagger - ru^2 - ru^2u_\dagger \\ &= u_\dagger^2\left\{1 - r\left(\frac{u}{u_\dagger}\right)^2\right\} + r^2u_\dagger^4 + ru_\dagger^3\left\{2 - (1+r)\left(\frac{u}{u_\dagger}\right)^2\right\}. \end{aligned}$$

Since $u \leq u_\dagger$, $u/u_\dagger \leq 1$ and consequently $\{1 - r(u/u_\dagger)^2\} \geq 0$ and $2 - (1+r)(u/u_\dagger)^2 \geq 0$ for $r \in (0,1]$. Hence, $B_3 > 0$ and $B_1, B_2 > 0$ as well. Since $A_2(u_0) > 0$ we have $B_1 + 2A_2(u_0) > 0$ and

$$g'(u) = (u - u_0)\left\{\frac{1}{ru^2} - 2A_2(u_\dagger) + 2A_2(u_0)\right\} < 0$$

for $0 < r \leq 1$ and $u \leq u_0$. This proves that $g(u)$ is decreasing for $u \leq u_0$. Note that $g(u_0) = 0$, so $g(u) \geq 0$ for $u \leq u_0$, and together with (S.1) we have $f(u) \geq (u-u_0)A_1(u_0) - (u-u_0)^2A_2(u_0) + (1/r)\{\log(u_0/u) + 1 - u_0/u\}$ for $0 < r \leq 1$.

Next consider the case of $r > 1$ with $\kappa = 1$. Here $g'(u) = (u - u_0)\{-2A_2(u_\dagger) + 2A_2(u_0) + 1/u^2\}$. Our goal is to show $g'(u) < 0$. To prove this it is sufficient to show

$$B_5 = \frac{1}{u_\dagger^2} - 2A_2(u_\dagger) = \frac{1}{u_\dagger^2} - \frac{(1+ru_\dagger)^{-1/r-2}[1 + r\{1 - (1+ru_\dagger)^{-1/r}\}]}{\{1 - (1+ru_\dagger)^{-1/r}\}^2} > 0, \qquad \text{(S.2)}$$

for $u \leq u_0$ because $g'(u) = (u - u_0)\{B_5 + 2A_2(u_0)\} + (u - u_0)(1/u^2 - 1/u_\dagger^2)$, $(u - u_0) < 0$, $(1/u^2 - 1/u_\dagger^2) > 0$ for $u \leq u_0$ and $u_\dagger \in [u, u_0)$, and $A_2(u_0) > 0$. Now, consider the following transformation of variables, $t = (1+ru_\dagger)^{1/r}$, so $u_\dagger = (t^r - 1)/r$. Then, showing inequality (S.2) is equivalent to show the following inequality,

$$\frac{r^2}{(t^r-1)^2} - \frac{(1+r)t - r}{t^{2r}(t-1)^2} > 0, \quad \forall r > 1, t > 1$$

$$
\begin{aligned}
&\iff r^2t^{2r}(t-1)^2 - \{(1+r)t-r\}(t^r-1)^2 > 0 \\
&\iff r^2(t-1)^2t^{2r} > (t^r-1)^2(t+rt-r) \\
&\iff 2\log(r) + 2\log(t-1) + 2r\log(t) - 2\log(t^r-1) > \log(t+rt-r). \qquad \text{(S.3)}
\end{aligned}
$$

Obviously $\lim_{t\to 1^+}\log(t+rt-r) = 0$, and

$$
\lim_{t\to 1^+}\{2\log(t-1) - 2\log(t^r-1)\} = 2\lim_{t\to 1^+}\log\left(\frac{t-1}{t^r-1}\right) = 2\log\left(\frac{1}{r}\right).
$$

Therefore, $\lim_{t\to 1^+}\{2\log(r) + 2\log(t-1) + 2r\log(t) - 2\log(t^r-1)\} = 0$. We thus have

$$
\begin{aligned}
& 2\log(r) + 2\log(t-1) + 2r\log(t) - 2\log(t^r-1) \\
= & \int_1^t \frac{\partial\{2\log(r) + 2\log(s-1) + 2r\log(s) - 2\log(s^r-1)\}}{\partial s},
\end{aligned}
$$

and

$$
\log(t+rt-r) = \int_1^t \frac{\partial\{\log(s+rs-r)\}}{\partial s}.
$$

Then, to prove (S.3), it suffices to show

$$
\begin{aligned}
& \frac{\partial\{2\log(r) + 2\log(t-1) + 2r\log(t) - 2\log(t^r-1)\}}{\partial t} > \frac{\partial\log(t+rt-r)}{\partial t}, \quad \forall r>1, t>1 \\
\iff & \frac{2}{t-1} - \frac{2r}{t(t^r-1)} > \frac{r+1}{t+rt-r} \\
\iff & \frac{1}{t-1} - \frac{2r}{t(t^r-1)} > (r+1)\left(\frac{1}{t+rt-r} - \frac{1}{(t-1)(r+1)}\right) \\
\iff & \frac{1}{t-1} - \frac{2r}{t(t^r-1)} + \frac{1}{(t+rt-r)(t-1)} > 0 \\
\iff & \frac{t+rt-r+1}{(t-1)(t+rt-r)} > \frac{2r}{t(t^r-1)} \\
\iff & \frac{(t^r-1)t}{(t-1)r} > \frac{2(t+rt-r)}{t+rt-r+1} \\
\iff & \frac{(t^r-1)t}{(t-1)r} > 1 + \frac{(t-1)(r+1)}{t+rt-r+1} \\
\iff & \frac{t^{r+1}-t-tr+r}{(t-1)r} > \frac{(t-1)(r+1)}{t+rt-r+1} \\
\iff & \frac{t^{r+1}-t-tr+r}{(t-1)^2r(r+1)} > \frac{1}{t+rt-r+1} \\
\iff & \frac{(t^{r+1}-1)/(t-1)-(r+1)}{(t-1)r(r+1)} > \frac{1}{t+rt-r+1}. \qquad \text{(S.4)}
\end{aligned}
$$

We now provide two useful statements, the first is

$$(t^{r+1}-1)/(t-1) = (r+1)\xi_1^r \geq (r+1)\left(\frac{t+1}{2}\right)^r \tag{S.5}$$

where the equality is obtained by the mean value theorem with $\xi_1 \in (1,t)$ and the inequality is obtained by

$$\xi_1^r = \frac{1}{t-1}\int_1^t s^r ds \geq \left(\frac{1}{t-1}\int_1^t sds\right)^r = \left(\frac{t+1}{2}\right)^r.$$

The last inequality is obtained by applying Jensen's inequality and noting that $x^r$ is a convex function for $r > 1$ and any generic $x > 0$. The second is

$$\frac{\{(t+1)/2\}^r - \{(t+1)/2\}^0}{r-0} = \left(\frac{t+1}{2}\right)^{\xi_2}\log\left(\frac{t+1}{2}\right), \tag{S.6}$$

where the equality is obtained by the mean value theorem for a $\xi_2 \in (0,r)$. Applying inequalities (S.5) and (S.6) to the left hand side of inequality (S.4), we have

$$\frac{(t^{r+1}-1)/(t-1)-(r+1)}{(t-1)r(r+1)} \geq \frac{\{(t+1)/2\}^r - 1}{(t-1)r} = \frac{\left(\frac{t+1}{2}\right)^{\xi_2}\log\left(\frac{t+1}{2}\right)}{t-1}.$$

Then, to prove (S.4), it is sufficient to show

$$\frac{\left(\frac{t+1}{2}\right)^{\xi_2}\log\left(\frac{t+1}{2}\right)}{t-1} > \frac{1}{t+rt-r+1} \iff \frac{t+1}{t-1}+r > \frac{1}{\left(\frac{t+1}{2}\right)^{\xi_2}\log\left(\frac{t+1}{2}\right)}. \tag{S.7}$$

Since $\log(x) \geq 1-1/x$ for any generic $x > 0$, we get $\log\{(t+1)/2\} \geq (t-1)/(t+1)$ and using this result to the right hand side of (S.7) we get

$$\frac{1}{\{(1+t)/2\}^{\xi_2}\log\{(t+1)/2\}} \leq \left(\frac{2}{1+t}\right)^{\xi_2} \times \left(\frac{t+1}{t-1}\right) < \left(\frac{t+1}{t-1}\right) < r + \left(\frac{t+1}{t-1}\right)$$

where the second last inequality follows as $t > 1$. The last inequality follows as $r > 1$. Hence (S.7) follows. Then the inequality (S.2) holds and $1/u^2 - 2A_2(u_\dagger) + 2A_2(u_0) > 0$ for $r > 1$. Consequently $g^{'}(u) < 0$ for $u \leq u_0$, and then $g(u) \geq g(u_0) = 0$ for $u \leq u_0$ and the desired result is obtained. □

#### S.1.1.2 Proof of part ii)

*Proof.* To prove the part ii) of Theorem 1, we first define $f(u) = \log[\{1-\exp(-u)\}/\{1-\exp(-u_0)\}]$. Observe that $f(u_0) = 0$. Let us consider the Taylor series expansion of $f(u)$ about

$u = u_0$

$$f(u) = (u - u_0)A_1(u_0) - (u - u_0)^2 A_2(u^*),$$

for some $u^* \in (u_0, u)$, where

$$A_1(u) = \frac{\partial f(u)}{\partial u} = \frac{\exp(-u)}{1 - \exp(-u)},$$

and

$$A_2(u) = -\frac{1}{2}\frac{\partial^2 f(u)}{\partial u^2} = \frac{\exp(-u)}{2\{1 - \exp(-u)\}} + \frac{\exp(-2u)}{2\{1 - \exp(-u)\}^2} = \frac{\exp(-u)}{2\{1 - \exp(-u)\}^2}.$$

Note that $A_2(u)$ is a decreasing function. So, for $u \geq u_0$, $A_2(u_0) = \max_{u \geq u_0} A_2(u)$. Hence, for $u \geq u_0$,

$$\begin{aligned} f(u) &\geq (u - u_0)A_1(u_0) - (u - u_0)^2 A_2(u_0) \\ &= (u - u_0)A_1(u_0) - (u - u_0)^2 A_2(u_0) + \log\left(\frac{u_0}{u}\right) + \log\left(\frac{u}{u_0}\right) \\ &\geq (u - u_0)A_1(u_0) - (u - u_0)^2 A_2(u_0) + \log\left(\frac{u_0}{u}\right) + \left(1 - \frac{u_0}{u}\right). \end{aligned} \quad \text{(S.8)}$$

To prove the result when $u \leq u_0$, let us define $g(u) = f(u) - (u - u_0)A_1(u_0) + (u - u_0)^2 A_2(u_0) - \log(u_0/u) - 1 + u_0/u$

$$\begin{aligned} g'(u) &= A_1(u) - A_1(u_0) + 2(u - u_0)A_2(u_0) + \frac{u - u_0}{u^2} \\ &= (u - u_0)\left\{-2A_2(u_\dagger) + 2A_2(u_0) + \frac{1}{u^2}\right\}, \end{aligned}$$

where the last equality is obtained by applying the Taylor series expansion on $A_1(u)$ about $u = u_0$, $A_1(u) = A_1(u_0) - 2(u - u_0)A_2(u_\dagger)$, for some $u_\dagger \in (u, u_0)$. As $A_2(u_0) > 0$, then for $u < u_\dagger$

$$\frac{1}{u^2} - 2A_2(u_\dagger) + 2A_2(u_0) > \frac{1}{u_\dagger^2} - 2A_2(u_\dagger) = \frac{1}{u_\dagger^2} - \frac{1}{\{1 - \exp(-u_\dagger)\}^2 \exp(u_\dagger)} \equiv h(u_\dagger).$$

Let us define

$$k(u_\dagger) \equiv \{1 - \exp(-u_\dagger)\}^2 \exp(u_\dagger) - u_\dagger^2 = \exp(u_\dagger) + \exp(-u_\dagger) - u_\dagger^2 - 2,$$

and investigate its properties. Note that

$$k(u_\dagger) = \left\{1 + u_\dagger + \frac{u_\dagger^2}{2} + \frac{u_\dagger^3}{3!} + \frac{u_\dagger^4}{4!} - \frac{u_\dagger^5}{5!} + \cdots\right\} + \left\{1 - u_\dagger + \frac{u_\dagger^2}{2} - \frac{u_\dagger^3}{3!} + \frac{u_\dagger^4}{4!} - \frac{u_\dagger^5}{5!} + \cdots\right\} - u_\dagger^2 - 2$$

$$= \quad 2\left\{\frac{u_\dagger^4}{4!}+\frac{u_\dagger^6}{6!}+\frac{u_\dagger^8}{8!}\cdots\right\}>0.$$

This result proves $g'(u) < 0$ ($g$ is a decreasing function) for any $u < u_0$. Thus, $g(u) \geq \min_{u \leq u_0} g(u) = \lim_{u \to u_0} g(u) = 0$. Hence, $g(u) = f(u) - (u-u_0)A_1(u_0) + (u-u_0)^2 A_2(u_0) - \log(u_0/u) - 1 + u_0/u \geq 0$ for $u \leq u_0$ and combined with (S.8) we now have part ii) of the theorem. □

### S.1.2 Proof of inequality (11)

This is the derivation of the minorization function for the $r > 0$ case with $\theta > 0$. Applying part (i) of Theorem 1 to the multiplier of $\Delta_{i,j}$ and result (i) of Lemma 1 to the multiplier of $(1-\Delta_{i,j})$, we have

$$\begin{aligned}
\ell_i(\boldsymbol{\xi}) \quad \geq \quad & \ell_i(\boldsymbol{\xi}_0) + \sum_k \omega_i^*(\boldsymbol{\xi}_0, a_k) \sum_{j=1}^{m_i}\Bigg(\Delta_{i,j}\{A_1(u_{i,j,k}(\boldsymbol{\xi}_0)) + 2A_2(u_{i,j,k}(\boldsymbol{\xi}_0))u_{i,j,k}(\boldsymbol{\xi}_0)\} \\
& \times u_{i,j,k}(\boldsymbol{\xi}_0)\exp\left[(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^\top \boldsymbol{W}_{i,j,k} + \log\left\{\frac{H_\psi(C_{i,j})}{H_{\psi_0}(C_{i,j})}\right\}\right] \\
& -\Delta_{i,j}A_2(u_{i,j,k}(\boldsymbol{\xi}_0))u_{i,j,k}^2(\boldsymbol{\xi}_0)\exp\left[2(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^\top \boldsymbol{W}_{i,j,k} + 2\log\left\{\frac{H_\psi(C_{i,j})}{H_{\psi_0}(C_{i,j})}\right\}\right] \\
& -(1-\Delta_{i,j})\frac{u_{i,j,k}(\boldsymbol{\xi}_0)}{1+ru_{i,j,k}(\boldsymbol{\xi}_0)}\exp\left[(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^\top \boldsymbol{W}_{i,j,k} + \log\left\{\frac{H_\psi(C_{i,j})}{H_{\psi_0}(C_{i,j})}\right\}\right] \\
& -\Delta_{i,j}\kappa\exp\left[(\boldsymbol{\alpha}_0-\boldsymbol{\alpha})^\top \boldsymbol{W}_{i,j,k} + \log\left\{\frac{H_{\psi_0}(C_{i,j})}{H_\psi(C_{i,j})}\right\}\right] \\
& -\Delta_{i,j}\kappa\left[\log\{H_\psi(C_{i,j})\} + \boldsymbol{\alpha}^\top \boldsymbol{W}_{i,j,k}\right] \\
& -\Delta_{i,j}A_1(u_{i,j,k}(\boldsymbol{\xi}_0))u_{i,j,k}(\boldsymbol{\xi}_0) - \Delta_{i,j}A_2(u_{i,j,k}(\boldsymbol{\xi}_0))u_{i,j,k}^2(\boldsymbol{\xi}_0) + (1-\Delta_{i,j})\frac{u_{i,j,k}(\boldsymbol{\xi}_0)}{1+ru_{i,j,k}(\boldsymbol{\xi}_0)} \\
& +\Delta_{i,j}\kappa + \Delta_{i,j}\kappa\log\{u_{i,j,k}(\boldsymbol{\xi}_0)\}\Bigg) \\
\geq \quad & \ell_i(\boldsymbol{\xi}_0) + \sum_k \omega_i^*(\boldsymbol{\xi}_0, a_k) \sum_{j=1}^{m_i}\Bigg(\Delta_{i,j}\{A_1(u_{i,j,k}(\boldsymbol{\xi}_0)) + 2A_2(u_{i,j,k}(\boldsymbol{\xi}_0))u_{i,j,k}(\boldsymbol{\xi}_0)\} \\
& \times u_{i,j,k}(\boldsymbol{\xi}_0)\left[1 + (\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^\top \boldsymbol{W}_{i,j,k} + \log\left\{\frac{H_\psi(C_{i,j})}{H_{\psi_0}(C_{i,j})}\right\}\right] \\
& -\Delta_{i,j}A_2(u_{i,j,k}(\boldsymbol{\xi}_0))u_{i,j,k}^2(\boldsymbol{\xi}_0)\left(\frac{1}{2}\right)\left[\exp\{4(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^\top \boldsymbol{W}_{i,j,k}\} + \left\{\frac{H_\psi(C_{i,j})}{H_{\psi_0}(C_{i,j})}\right\}^4\right] \\
& -(1-\Delta_{i,j})\frac{u_{i,j,k}(\boldsymbol{\xi}_0)}{1+ru_{i,j,k}(\boldsymbol{\xi}_0)}\left(\frac{1}{2}\right)\left[\exp\{2(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^\top \boldsymbol{W}_{i,j,k}\} + \left\{\frac{H_\psi(C_{i,j})}{H_{\psi_0}(C_{i,j})}\right\}^2\right]
\end{aligned}$$

$$
\begin{aligned}
& -\Delta_{i,j}\left(\frac{1}{2}\right)\kappa\left[\exp\{2(\boldsymbol{\alpha}_0-\boldsymbol{\alpha})^\top \boldsymbol{W}_{i,j,k}\}+\left\{\frac{H_{\psi_0}(C_{i,j})}{H_\psi(C_{i,j})}\right\}^2\right] \\
& -\Delta_{i,j}\kappa\left[\log\{H_\psi(C_{i,j})\}+\boldsymbol{\alpha}^\top \boldsymbol{W}_{i,j,k}\right] \\
& -\Delta_{i,j}A_1(u_{i,j,k}(\boldsymbol{\xi}_0))u_{i,j,k}(\boldsymbol{\xi}_0)-\Delta_{i,j}A_2(u_{i,j,k}(\boldsymbol{\xi}_0))u_{i,j,k}^2(\boldsymbol{\xi}_0)+(1-\Delta_{i,j})\frac{u_{i,j,k}(\boldsymbol{\xi}_0)}{1+ru_{i,j,k}(\boldsymbol{\xi}_0)} \\
& \left.+\Delta_{i,j}\kappa+\Delta_{i,j}\kappa\log\{u_{i,j,k}(\boldsymbol{\xi}_0)\}\right) \\
= \;& \ell_{\dagger,i}(\boldsymbol{\xi}|\boldsymbol{\xi}_0).
\end{aligned}
$$

The last inequality is obtained by using result (ii) and (iii) of Lemma 1. Thus $\ell_{\dagger,i}$ can be further written as $\ell_{\dagger,i}(\boldsymbol{\xi}|\boldsymbol{\xi}_0)=\ell_{\dagger,1,i}(\boldsymbol{\alpha}|\boldsymbol{\xi}_0)+\ell_{\dagger,2,i}(\boldsymbol{\psi}|\boldsymbol{\xi}_0)+\ell_{\dagger,3,i}(\boldsymbol{\xi}_0)$.

### S.1.3 Detailed derivation of Section 3.3

This is the details of the non-dependence case ($\theta=0$). Note that here

$$
\ell(\boldsymbol{\xi}) = \sum_{i=1}^n \ell_i(\boldsymbol{\xi}) = \ell(\boldsymbol{\xi}_0)+\sum_{i=1}^n \log\left[\frac{\{1-G_i(\boldsymbol{\xi})\}^{\Delta_i}\{G_i(\boldsymbol{\xi})\}^{1-\Delta_i}}{\{1-G_i(\boldsymbol{\xi}_0)\}^{\Delta_i}\{G_i(\boldsymbol{\xi}_0)\}^{1-\Delta_i}}\right].
$$

<u>Case of $r>0$</u>

For $r>0$, using the actual expressions of $G_i(\boldsymbol{\xi})$ and $G_i(\boldsymbol{\xi}_0)$, we have

$$
\ell(\boldsymbol{\xi}) = \ell(\boldsymbol{\xi}_0)+\sum_{i=1}^n\left(\Delta_i\log\left[\frac{1-\{1+ru_i(\boldsymbol{\xi})\}^{-1/r}}{1-\{1+ru_i(\boldsymbol{\xi}_0)\}^{-1/r}}\right]+(1-\Delta_i)\kappa\log\left\{\frac{1+ru_i(\boldsymbol{\xi})}{1+ru_i(\boldsymbol{\xi}_0)}\right\}\right).
$$

Using the same inequalities and techniques in Section S.1.2, we first obtain the minorization function $\ell_\dagger(\boldsymbol{\xi}|\boldsymbol{\xi}_0)$, such that $\ell(\boldsymbol{\xi})\geq \ell_\dagger(\boldsymbol{\xi}|\boldsymbol{\xi}_0)\equiv \ell_{\dagger,1}(\boldsymbol{\alpha}|\boldsymbol{\xi}_0)+\ell_{\dagger,2}(\boldsymbol{\psi}|\boldsymbol{\xi}_0)+\ell_{\dagger,3}(\boldsymbol{\xi}_0)$, where

$$
\begin{aligned}
\ell_{\dagger,1}(\boldsymbol{\alpha}|\boldsymbol{\xi}_0) = & \sum_{i=1}^n\Bigg[\Delta_i\Big\{A_1(u_i(\boldsymbol{\xi}_0))+2A_2(u_i(\boldsymbol{\xi}_0))u_i(\boldsymbol{\xi}_0)\Big\}u_i(\boldsymbol{\xi}_0)(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^\top\boldsymbol{X}_i \\
& -\left(\frac{\Delta_i}{2}\right)A_2(u_i(\boldsymbol{\xi}_0))u_i^2(\boldsymbol{\xi}_0)\exp\{4(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^\top\boldsymbol{X}_i\} \\
& -\left(\frac{1-\Delta_i}{2}\right)\frac{u_i(\boldsymbol{\xi}_0)}{1+ru_i(\boldsymbol{\xi}_0)}\exp\{2(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^\top\boldsymbol{X}_i\} \\
& -\left(\frac{\Delta_i\kappa}{2}\right)\exp\{2(\boldsymbol{\alpha}_0-\boldsymbol{\alpha})^\top\boldsymbol{X}_i\}-\Delta_i\kappa\boldsymbol{\alpha}^\top\boldsymbol{X}_i\Bigg], \\
\ell_{\dagger,2}(\boldsymbol{\psi}|\boldsymbol{\xi}_0) = & \sum_{i=1}^n\Bigg[\Delta_i\Big\{A_1(u_i(\boldsymbol{\xi}_0))+2A_2(u_i(\boldsymbol{\xi}_0))u_i(\boldsymbol{\xi}_0)\Big\} \\
& \times u_i(\boldsymbol{\xi}_0)\log\left\{\frac{H_\psi(C_i)}{H_{\psi_0}(C_i)}\right\}-\left(\frac{\Delta_i}{2}\right)A_2(u_i(\boldsymbol{\xi}_0))u_i^2(\boldsymbol{\xi}_0)\left\{\frac{H_\psi(C_i)}{H_{\psi_0}(C_i)}\right\}^4
\end{aligned}
$$

$$
\begin{aligned}
& -\left(\frac{1-\Delta_i}{2}\right)\frac{u_i(\boldsymbol{\xi}_0)}{1+ru_i(\boldsymbol{\xi}_0)}\left\{\frac{H_\psi(C_i)}{H_{\psi_0}(C_i)}\right\}^2 \\
& -\left(\frac{\Delta_i\kappa}{2}\right)\left\{\frac{H_{\psi_0}(C_i)}{H_\psi(C_i)}\right\}^2 - \Delta_i\kappa\log\{H_\psi(C_i)\}\Bigg], \\
\ell_{\dagger,3}(\boldsymbol{\xi}_0) \;=\; & \ell(\boldsymbol{\xi}_0) + \sum_{i=1}^n \Bigg(\Delta_i A_2(u_i(\boldsymbol{\xi}_0))u_i^2(\boldsymbol{\xi}_0) + (1-\Delta_i)\frac{u_i(\boldsymbol{\xi}_0)}{1+ru_i(\boldsymbol{\xi}_0)} \\
& + \Delta_i\kappa\left[1+\log\{u_i(\boldsymbol{\xi}_0)\}\right]\Bigg).
\end{aligned}
$$

Then, $\boldsymbol{\alpha}$ and $\boldsymbol{\psi}$ are estimated by the generic Newton-Raphson algorithm given in (12), where the needed quantities are

$$
\begin{aligned}
S(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) \;=\; & \sum_{i=1}^n \left\{\Delta_i A_1(u_i(\boldsymbol{\xi}^{(m-1)})) - \frac{(1-\Delta_i)}{1+ru_i(\boldsymbol{\xi}^{(m-1)})}\right\} u_i(\boldsymbol{\xi}^{(m-1)})\boldsymbol{X}_i, \\
S_\alpha(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) \;=\; & -\sum_{i=1}^n \Bigg[8\Delta_i A_2(u_i(\boldsymbol{\xi}^{(m-1)}))u_i^2(\boldsymbol{\xi}^{(m-1)}) \\
& + 2(1-\Delta_i)\frac{u_i(\boldsymbol{\xi}^{(m-1)})}{1+ru_i(\boldsymbol{\xi}^{(m-1)})} + 2\Delta_i\kappa\Bigg]\boldsymbol{X}_i^{\otimes 2}, \\
S(\boldsymbol{\psi}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) \;=\; & \sum_{i=1}^n \Bigg\{\Delta_i A_1(u_i(\boldsymbol{\xi}^{(m-1)})) \\
& - \frac{(1-\Delta_i)}{1+ru_i(\boldsymbol{\xi}^{(m-1)})}\Bigg\} u_i(\boldsymbol{\xi}^{(m-1)})\left[\frac{\partial\log\{H_\psi(C_i)\}}{\partial\boldsymbol{\psi}}\right]_{\psi=\psi^{(m-1)}}, \\
S_\psi(\boldsymbol{\psi}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) \;=\; & \sum_{i=1}^n \left\{\Delta_i A_1(u_i(\boldsymbol{\xi}^{(m-1)}))u_i(\boldsymbol{\xi}^{(m-1)}) - (1-\Delta_i)\frac{u_i(\boldsymbol{\xi}^{(m-1)})}{1+ru_i(\boldsymbol{\xi}^{(m-1)})}\right\} \\
& \times\left[\frac{\partial^2\log\{H_\psi(C_i)\}}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\mathsf{T}}\right]_{\psi=\psi^{(m-1)}} \\
& - \sum_{i=1}^n \left\{8\Delta_i A_2(u_i(\boldsymbol{\xi}^{(m-1)}))u_i^2(\boldsymbol{\xi}^{(m-1)}) + 2(1-\Delta_i)\frac{u_i(\boldsymbol{\xi}^{(m-1)})}{1+ru_i(\boldsymbol{\xi}^{(m-1)})} + 2\Delta_i\kappa\right\} \\
& \times\left(\left[\frac{\partial\log\{H_\psi(C_i)\}}{\partial\boldsymbol{\psi}}\right]_{\psi=\psi^{(m-1)}}\right)^{\otimes 2}.
\end{aligned}
$$

<u>Case of $r \to 0$</u>

For the $r \to 0$ case (actually the limiting case $r \to 0$) the minorization function is the total of the following terms

$$
\begin{aligned}
\ell_{\dagger,1}(\boldsymbol{\alpha}|\boldsymbol{\xi}_0) \;=\; & \sum_{i=1}^n \Bigg[\Delta_i\Big\{A_3(u_i(\boldsymbol{\xi}_0)) + 2A_4(u_i(\boldsymbol{\xi}_0))u_i(\boldsymbol{\xi}_0)\Big\} \times u_i(\boldsymbol{\xi}_0)(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^\mathsf{T}\boldsymbol{X}_i \\
& -\left(\frac{\Delta_i}{2}\right)A_4(u_i(\boldsymbol{\xi}_0))u_i^2(\boldsymbol{\xi}_0)\exp\{4(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^\mathsf{T}\boldsymbol{X}_i\}
\end{aligned}
$$

$$
\begin{aligned}
& \quad - \left(\frac{1-\Delta_i}{2}\right) u_i(\boldsymbol{\xi}_0) \exp\{2(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^\mathsf{T} \boldsymbol{X}_i\} \\
& \quad - \left(\frac{\Delta_i}{2}\right) \exp\{2(\boldsymbol{\alpha}_0-\boldsymbol{\alpha})^\mathsf{T} \boldsymbol{X}_i\} - \Delta_i \boldsymbol{\alpha}^\mathsf{T} \boldsymbol{X}_i \Bigg], \\
\ell_{\dagger,2}(\boldsymbol{\psi}|\boldsymbol{\xi}_0) &= \sum_{i=1}^n \Bigg[ \Delta_i \Big\{ A_3(u_i(\boldsymbol{\xi}_0)) + 2A_4(u_i(\boldsymbol{\xi}_0))u_i(\boldsymbol{\xi}_0) \Big\} \times u_i(\boldsymbol{\xi}_0) \log \left\{ \frac{H_\psi(C_i)}{H_{\psi_0}(C_i)} \right\} \\
& \quad - \left(\frac{\Delta_i}{2}\right) A_4(u_i(\boldsymbol{\xi}_0)) u_i^2(\boldsymbol{\xi}_0) \left\{ \frac{H_\psi(C_i)}{H_{\psi_0}(C_i)} \right\}^4 - \left(\frac{1-\Delta_i}{2}\right) u_i(\boldsymbol{\xi}_0) \left\{ \frac{H_\psi(C_i)}{H_{\psi_0}(C_i)} \right\}^2 \\
& \quad - \left(\frac{\Delta_i}{2}\right) \left\{ \frac{H_{\psi_0}(C_i)}{H_\psi(C_i)} \right\}^2 - \Delta_i \log\{H_\psi(C_i)\} \Bigg], \\
\ell_{\dagger,3}(\boldsymbol{\xi}_0) &= \ell(\boldsymbol{\xi}_0) + \sum_{i=1}^n \Big( \Delta_i A_4(u_i(\boldsymbol{\xi}_0)) u_i^2(\boldsymbol{\xi}_0) + (1-\Delta_i) u_i(\boldsymbol{\xi}_0) + \Delta_i \left[1 + \log\{u_i(\boldsymbol{\xi}_0)\}\right] \Big),
\end{aligned}
$$

and the terms needed in the Newton-Raphson algorithm (12) are

$$
\begin{aligned}
S(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) &= \sum_{i=1}^n \Big\{ \Delta_i A_3(u_i(\boldsymbol{\xi}^{(m-1)})) - (1-\Delta_i) \Big\} u_i(\boldsymbol{\xi}^{(m-1)}) \boldsymbol{X}_i, \\
S_\alpha(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) &= -\sum_{i=1}^n \Big[ 8\Delta_i A_4(u_i(\boldsymbol{\xi}^{(m-1)})) u_i^2(\boldsymbol{\xi}^{(m-1)}) + 2(1-\Delta_i) u_i(\boldsymbol{\xi}^{(m-1)}) + 2\Delta_i \Big] \boldsymbol{X}_i^{\otimes 2}, \\
S(\boldsymbol{\psi}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) &= \sum_{i=1}^n \Big\{ \Delta_i A_3(u_i(\boldsymbol{\xi}^{(m-1)})) - (1-\Delta_i) \Big\} u_{i,j}(\boldsymbol{\xi}^{(m-1)}) \left[ \frac{\partial \log\{H_\psi(C_i)\}}{\partial \boldsymbol{\psi}} \right]_{\psi=\psi^{(m-1)}}, \\
S_\psi(\boldsymbol{\psi}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) &= \sum_{i=1}^n \left\{ \Delta_i A_3(u_i(\boldsymbol{\xi}^{(m-1)})) - (1-\Delta_i) \right\} u_i(\boldsymbol{\xi}^{(m-1)}) \left[ \frac{\partial^2 \log\{H_\psi(C_i)\}}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^\mathsf{T}} \right]_{\psi=\psi^{(m-1)}} \\
& \quad - \sum_{i=1}^n \Big\{ 8\Delta_i A_4(u_i(\boldsymbol{\xi}^{(m-1)})) u_i^2(\boldsymbol{\xi}^{(m-1)}) + 2(1-\Delta_i) u_i(\boldsymbol{\xi}^{(m-1)}) + 2\Delta_i \Big\} \\
& \quad \times \left( \left[ \frac{\partial \log\{H_\psi(C_i)\}}{\partial \boldsymbol{\psi}} \right]_{\psi=\psi^{(m-1)}} \right)^{\otimes 2}.
\end{aligned}
$$

Since $H_\psi(C_i) = \sum_l M_l(C_i)\exp(\psi_l), \partial H_\psi(C_i)/\partial\psi_l = M_l(C_i)\exp(\psi_l)$, let us write $\partial H_\psi/\partial\boldsymbol{\psi} = \boldsymbol{D}_i \exp(\boldsymbol{\psi})$, where $\boldsymbol{D}_i = \mathrm{Diag}(M_1(C_i), \ldots, M_K(C_i))$ and $\exp(\boldsymbol{\psi}) = (\exp(\psi_1), \ldots, \exp(\psi_K))^\mathsf{T}$. Then we have

$$
\left[ \frac{\partial^2 \log\{H_\psi(C_i)\}}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^\mathsf{T}} \right]_{\psi=\psi^{(m-1)}} = \left[ \frac{\boldsymbol{D}_i \mathrm{Diag}(\exp(\boldsymbol{\psi}^{(m-1)}))}{H_{\psi^{(m-1)}}} - \frac{\boldsymbol{D}_i \exp(\boldsymbol{\psi}^{(m-1)}) \{\exp(\boldsymbol{\psi}^{(m-1)})\}^\mathsf{T} \boldsymbol{D}_i}{H^2_{\psi^{(m-1)}}} \right],
$$

and

$$
\left( \left[ \frac{\partial \log\{H_\psi(C_i)\}}{\partial \boldsymbol{\psi}} \right]_{\psi=\psi^{(m-1)}} \right)^{\otimes 2} = \frac{\boldsymbol{D}_i \exp(\boldsymbol{\psi}^{(m-1)}) \{\exp(\boldsymbol{\psi}^{(m-1)})\}^\mathsf{T} \boldsymbol{D}_i}{H^2_{\psi^{(m-1)}}(C_i)}.
$$

## S.2 Results of Section 4

### S.2.1 Background

**Notations:** To prove the main theorems more clearly, we first assume the subject-specific random effect $b$ is observed and investigate the asymptotic properties of the penalized complete ML estimator. The rate of convergence (Theorem 2) and semiparametric efficiency (Theorem 3) of the penalized observed ML estimator (4) can be proved with the similar arguments and presented at Subsection S.2.6.

Define $\boldsymbol{O}_* = (C_{*,1}, \ldots, C_{*,m_*}, \Delta_{*,1}, \ldots, \Delta_{*,m_*} \boldsymbol{X}_{*,1}^\top, \ldots, \boldsymbol{X}_{*,m_*}^\top, \boldsymbol{Z}_*)^\top$ as the observed data from a random cluster $*$, where $m_*$ is the cluster size. We also let $\mathbf{P}_{\boldsymbol{\iota}}$ be the distribution of the complete data $\boldsymbol{g} = (\boldsymbol{O}_*, b_*)^\top$ from a random cluster $*$ under the parameter vector $\boldsymbol{\iota}$, and $p_{\boldsymbol{\iota}}$ be the corresponding density with the dominating measure $\mu$. For simplicity, we define $\mathbf{P}_0 \equiv \mathbf{P}_{\boldsymbol{\iota}_0}$ and $p_0 \equiv p_{\boldsymbol{\iota}_0}$. Specifically, let $\mathcal{L}_c(\boldsymbol{\iota}; \boldsymbol{g})$ and $\ell_c(\boldsymbol{\iota}; \boldsymbol{g})$ be the likelihood and log-likelihood for one single complete observation, respectively. In other words,

$$
\begin{aligned}
\mathcal{L}_c(\boldsymbol{\iota}; \boldsymbol{g}) &= \prod_{j=1}^{m_*} \Big\{ 1 - S(C_{*,j} | \boldsymbol{X}_{*,j}, \boldsymbol{Z}_*, b_*) \Big\}^{\Delta_{*,j}} \Big\{ S(C_{*,j} | \boldsymbol{X}_{*,j}, \boldsymbol{Z}_*, b_*) \Big\}^{1-\Delta_{*,j}} \phi(b_*) \\
&= \phi(b_*) \prod_{j=1}^{m_*} \left( 1 - \Big[ 1 + rH(C_{*,j}) \exp\{\boldsymbol{\beta}^\top \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^\top \boldsymbol{Z}_* + \theta b_*\} \Big]^{-1/r} \right)^{\Delta_{*,j}} \\
&\quad \times \left( \Big[ 1 + rH(C_{*,j}) \exp\{\boldsymbol{\beta}^\top \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^\top \boldsymbol{Z}_* + \theta b_*\} \Big]^{-1/r} \right)^{1-\Delta_{*,j}}.
\end{aligned} \tag{S.9}
$$

Here we present the asymptotic properties of the penalized estimator when $r > 0$, the result for $r \to 0$ can be similarly obtained with the change of the expression $S(C_{*,j} | \boldsymbol{X}_{*,j}, \boldsymbol{Z}_*, b_*)$ in (S.9). Analogous to (4), we also define the penalized complete ML estimator as

$$
\begin{aligned}
\widehat{\boldsymbol{\iota}}_{c,n} &= (\widehat{\boldsymbol{\alpha}}_{c,n}{}^\top, \widehat{H}_{c,n})^\top \\
&= \underset{(\boldsymbol{\alpha}^\top, \sum_{k=1}^K M_k(t) \exp(\psi_k))^\top}{\arg\min} \left( \frac{1}{n} \sum_{i=1}^n \ell_c \Big\{ \boldsymbol{\alpha}, \sum_{k=1}^K M_k(t) \exp(\psi_k); \boldsymbol{g}_i \Big\} \right. \\
&\qquad \left. - \lambda \int_0^{T_0} \left[ \Big\{ \sum_{k=1}^K M_k(t) \exp(\psi_k) \Big\}^{(q)} \right]^2 \mathrm{d}t \right).
\end{aligned} \tag{S.10}
$$

To study the space spanned by $\{M_k(t)\}$, we let $\mathcal{S}_n(\boldsymbol{\tau}_n, L_n, d-1)$ denote the space of polynomial splines spanned by a degree $d-1$ B-spline basis with knots $\boldsymbol{\tau}_n = \{\tau_1, \tau_2, \ldots, \tau_L\}$ where $0 = \tau_0 <$

$\tau_1 < \tau_2 < \cdots < \tau_L < \tau_{L+1} = T_0$, $L \equiv L_n = O(n^{1/(2q+1)})$, with $d \geq q$. Furthermore, it is desirable to restrict the knots such that $\max_{0 \leq l \leq L} |\tau_{l+1} - \tau_l| = O(n^{-1/(2q+1)})$ as in Stone (1985). We also let $\mathcal{H}_n(\boldsymbol{\tau}_n, L_n, d)$ denote the space of polynomial splines spanned by $d$-degree I-spline basis, such that each basis function in $\mathcal{H}_n(\boldsymbol{\tau}_n, L_n, d)$ is the integration of the corresponding basis function in $\mathcal{S}_n(\boldsymbol{\tau}_n, L_n, d-1)$ over the domain $[0, T_0]$, and that all the coefficients are positive. In other words,

$$\mathcal{H}_n(\boldsymbol{\tau}_n, L_n, d) = \Bigg\{ \sum_{k=1}^{K} M_k(t) \exp(\psi_k) : \ M_k(t) = \int_0^t \mathcal{B}_k(s) \mathrm{d}s, \\ \mathcal{B}_k(s) \text{ is a basis function of } \mathcal{S}_n(\boldsymbol{\tau}_n, L_n, d-1),\ k = 1, \ldots, K \Bigg\},$$

where $K = L + d$. It is shown in de Boor (1978) that $\mathcal{H}_n(\boldsymbol{\tau}_n, L_n, d) \subset \mathcal{S}_n(\boldsymbol{\tau}_n, L_n, d)$. To simplify the notations, we also denote $\boldsymbol{\varphi} = \exp(\boldsymbol{\psi})$ with positive values, i.e., $\varphi_k = \exp(\psi_k)$, $k = 1, \ldots, K$. We first note that for a fixed $n$, letting the tuning parameter $\lambda \to 0$ implies an unpenalized estimate in the space spanned by the given polynomial space. On the other hand, letting $\lambda \to \infty$ forces convergence of the $q$th derivative of the spline function to zero. For example, when $q = 3$, the limiting transformation function will be quadratic with respect to $t$.

We introduce some further notation to be used in proving results. Given a random sample $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n$ with the probability measure $\mathbf{P}$, for a measurable function $f$, define $\mathbf{P}f = \int f \mathrm{d}\mathbf{P}$ as the expectation of $f$ under $\mathbf{P}$ and $\mathbb{P}_n f = (1/n) \sum_{i=1}^{n} f(\boldsymbol{g}_i)$ as the expectation of $f$ under the empirical measure $\mathbb{P}_n$. We write $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - \mathbf{P}) f$ for the empirical process $\mathbb{G}_n$ evaluated at $f$. Denote $\|\mathbb{G}_n\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\mathbb{G}_n f|$. Let $\|\cdot\|$ and $\|\cdot\|_\infty$ be the Euclidean norm of $\mathbb{R}^p$ and the supremum norm, respectively. We will use $v$ to denote a generic constant that may change values from context to context. For two sequences $\{a_{1,n}\}$ and $\{a_{2,n}\}$, we let $a_{1,n} \asymp a_{2,n}$ denote $a_{1,n} = O(a_{2,n})$ and $a_{2,n} = O(a_{1,n})$, simultaneously.

**Regularity conditions:** Here we present the regularity conditions required to study the asymptotic properties of the regularized semiparametric ML estimator.

(C1) The cluster size $m_*$ of a random cluster is completely random, and uniformly bounded above. In addition $\mathbf{P}(m_* \geq 1) > 0$.

(C2) The covariates $(\boldsymbol{X}_{*,1}^\top, \ldots, \boldsymbol{X}_{*,m_*}^\top, \boldsymbol{Z}_*^\top)^\top$ are uniformly bounded, that is, there exists a scalar $v$ such that $\mathbf{P}\{\|(\boldsymbol{X}_{*,1}^\top, \ldots, \boldsymbol{X}_{*,m_*}^\top, \boldsymbol{Z}_*^\top)\| \leq v\} = 1$, where $\|\cdot\|$ denotes the Euclidean norm. Moreover, all the eigenvalues of $E\Big[\big\{(\boldsymbol{X}_{*,1}^\top, \ldots, \boldsymbol{X}_{*,m_*}^\top, \boldsymbol{Z}_*^\top, b_*)^\top\big\}^{\otimes 2}\Big]$ are bounded away from zero and infinity, where $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^\top$ denotes the Gram matrix for any generic vector $\boldsymbol{a}$.

(C3) The conditional joint density of $(\boldsymbol{O}_*|b_*)$ has uniform positive lower and upper bound in the support of the joint random variables $\boldsymbol{O}_*$.

(C4) The $L_\infty$ norm of the true transformation function $H_0(t)$ is bounded away from 0 and $\infty$. Moreover, $H_0(\cdot)$ belongs to $\mathcal{H}$, a class of non-negative and monotonic functions, with zero values at $t = 0$ which are also continuously differentiable up to order $q$, $d \geq q \geq 2$, on $[0, T_0]$.

(C5) $\Theta$ is a compact subset of $\mathbb{R}^p$, where $p$ is the dimensionality of $\boldsymbol{\alpha}$. Furthermore, $\boldsymbol{\alpha}_0$ is an interior point of $\Theta$.

(C6) For any cluster size $m_*$, there exits some $\kappa \in (0, 1)$, such that

$$\begin{aligned}
&\boldsymbol{a}^\top \mathrm{var}\Big\{(\boldsymbol{X}_{*,1}^\top, \ldots, \boldsymbol{X}_{*,m_*}^\top, \boldsymbol{Z}_*^\top, b_*)^\top \big| C_{*,j}, 1 \leq j \leq m_*\Big\}\boldsymbol{a} \\
&\qquad \geq \kappa \boldsymbol{a}^\top E\Big[\big\{(\boldsymbol{X}_{*,1}^\top, \ldots, \boldsymbol{X}_{*,m_*}^\top, \boldsymbol{Z}_*^\top, b_*)^\top\big\}^{\otimes 2} \big| C_{*,j}, 1 \leq j \leq m_*\Big]\boldsymbol{a}
\end{aligned}$$

uniformly for all $\boldsymbol{a}$ with a suitable length.

Condition (C1), in the case of a completely random cluster size, can be found in Zeng et al. (2005). (C2)–(C6) are widely used in semiparametric modeling of survival analysis (see, for example, Huang and Wellner, 1997; Zhang et al., 2010) and usually satisfied in practice. Conditions (C1)–(C4) ensure the proposed model is identifiable. In particular, (C2) implies that for all $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \theta)$ and $v \in \mathbb{R}$,

$$\mathbf{P}(\boldsymbol{\beta}^\top \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^\top \boldsymbol{Z}_* + \theta b_* \neq v) > 0, \quad \forall j.$$

Condition (C3) suffices to prevent the joint distribution of the covariates and the inspection time from degeneration. For example, under (C3), we can show that

$$\begin{cases}\mathbf{P}(\Delta_{*,j}=1|C_{*,j}\neq 0)>0,\\ \mathbf{P}(\Delta_{*,j}=0|C_{*,j}\neq 0)>0.\end{cases}$$

Furthermore, it guarantees that the density function of $(C_{*,j}|b_*)$ is also bounded away from zero and infinity in its support. Condition (C4) regularizes the nonparametric function to be estimated. (C5) and (C6) are technical assumptions used to prove the rate of convergence and asymptotic normality. Although some of these conditions can be relaxed to a weaker version, it will make the proofs unnecessarily more complex.

The following theorem establishes the consistency of the penalized complete ML estimator.

**Theorem S.1.** *Suppose the regularity conditions (C1)–(C6) hold, $L = O(n^{1/(2q+1)})$, and the tuning parameter $\lambda$ satisfies $\lambda \asymp n^{-2q/(2q+1)}$. Then*

$$\mathrm{dist}(\widehat{\boldsymbol{\iota}}_{c,n}, \boldsymbol{\iota}_0) = O_p\big(n^{-q/(2q+1)}\big). \tag{S.11}$$

**Semiparametric efficiency bound:** For notational convenience, for a vector $\boldsymbol{\alpha}$ with suitable length, let $\dot{\ell}_{c,1}(\boldsymbol{\iota};\boldsymbol{g})$ denote the vector of partial derivatives of $\ell_c(\boldsymbol{\iota};\boldsymbol{g})$ with respect to $\boldsymbol{\alpha}$. For the nonparametric part, consider a parametric smooth submodel with parameter $(\boldsymbol{\alpha}^{\mathsf{T}}, H_{(s,w)})^{\mathsf{T}}$, such that $H_{(s,w)} = H + sw \in \mathcal{H}$ for $s$ in a small interval containing 0, with $H_{(0,w)} = H$ and $\{\partial H_{(s,w)}/\partial s\}|_{s=0} = w$. Let $\mathcal{W}$ be the class of functions $w$ defined by this equation. The score operator for $H$ begins with defining the Gâteaux (directional) derivative at $H$ along $w$: $\dot{\ell}_{c,2}(\boldsymbol{\iota};\boldsymbol{g})[w] = \{\partial l(\boldsymbol{\alpha}, H_{(s,w)};\boldsymbol{g})/\partial s\}|_{s=0}$. In addition, for $\boldsymbol{w} = (w_1,\ldots,w_p)^{\mathsf{T}}$ with $w_k \in \mathcal{W}$, $k = 1,\ldots,p$, let $\dot{\ell}_{c,2}(\boldsymbol{\iota};\boldsymbol{g})[\boldsymbol{w}]$ be the $p$-dimensional vector with its $k$th element $\dot{\ell}_{c,2}(\boldsymbol{\iota};\boldsymbol{g})[w_k]$. If $\boldsymbol{w}_c^* \in \mathcal{W}^p$ and satisfies

$$\boldsymbol{w}_c^* = \underset{\boldsymbol{w}\in\mathcal{W}^p}{\arg\min}\, E\big\|\dot{\ell}_{c,1}(\boldsymbol{\iota};\boldsymbol{g}) - \dot{\ell}_{c,2}(\boldsymbol{\iota};\boldsymbol{g})[\boldsymbol{w}]\big\|^2, \tag{S.12}$$

then $\boldsymbol{w}_c^*$ is called the *least favorable direction*, and by Theorem 1 in Bickel et al. (1993, pp. 70), the efficient score for $\boldsymbol{\alpha}$ is $\dot{\ell}_{c,1}(\boldsymbol{\iota};\boldsymbol{g}) - \dot{\ell}_{c,2}(\boldsymbol{\iota};\boldsymbol{g})[\boldsymbol{w}_c^*]$. According to the result in Bickel et al. (1993), the efficient information matrix of parameter $\boldsymbol{\alpha}$ for the complete likelihood is given by

$$I_c(\boldsymbol{\alpha}) = E\big\{\dot{\ell}_{c,1}(\boldsymbol{\iota};\boldsymbol{g}) - \dot{\ell}_{c,2}(\boldsymbol{\iota};\boldsymbol{g})[\boldsymbol{w}_c^*]\big\}^{\otimes 2}. \tag{S.13}$$

Analogously, the efficient information matrix of parameter $\boldsymbol{\alpha}$ for the observed likelihood is given by

$$I(\boldsymbol{\alpha}) = E\big\{\dot{\ell}_1(\boldsymbol{\iota}; \boldsymbol{U}_*) - \dot{\ell}_2(\boldsymbol{\iota}; \boldsymbol{O}_*)[\boldsymbol{w}^*]\big\}^{\otimes 2}, \tag{S.14}$$

where $\dot{\ell}_1$, $\dot{\ell}_2$, and $\boldsymbol{w}^*$ are the partial derivative of $\ell$ with respect to the parametric component $\boldsymbol{\alpha}$, Gâteaux (directional) derivative of $\ell$ with respect to the nonparametric component $H$, and the corresponding least favorable direction, respectively.

The next lemma shows the existence of the least favorable directions $\boldsymbol{w}_c^*$ and $\boldsymbol{w}^*$. Furthermore, the expressions of efficient information matrices $I_c(\boldsymbol{\alpha})$ in (S.13) and $I(\boldsymbol{\alpha})$ in (S.14) can be obtained.

**Lemma S.1.** *Under conditions (C1)–(C4), the least favorable directions $\boldsymbol{w}_c^*$ and $\boldsymbol{w}^*$ exist.*

For studying the asymptotic normality and efficiency, the least favorable direction must be estimable in the sense that its roughness penalty is bounded away from infinity, which leads to our last regularity condition.

(C7) The least favorable direction $\boldsymbol{w}_c^*$ for the complete likelihood satisfies $J(\boldsymbol{w}_c^*) < \infty$.

(C7') The least favorable direction $\boldsymbol{w}^*$ for the observed likelihood satisfies $J(\boldsymbol{w}^*) < \infty$.

**Theorem S.2.** *Suppose that all the assumptions given in Theorem 2 hold and the regularity condition (C7) is satisfied. Then, $n^{1/2}\big(\widehat{\boldsymbol{\alpha}}_{c,n} - \boldsymbol{\alpha}_0\big)$ converges to $\mathcal{N}\big(\mathbf{0}, I_c^{-1}(\boldsymbol{\alpha}_0)\big)$ in distribution, where $I_c(\boldsymbol{\alpha}_0)$ is the efficient information of $\boldsymbol{\alpha}$ with expected value at $\boldsymbol{\iota}_0$ for the complete likelihood, and is assumed to be non-singular.*

### S.2.2 Proof of Lemma 2

*Proof of Lemma 2.* Suppose that $(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}}, \widetilde{\theta}, \widetilde{H})$ gives the same observed likelihood function as of $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \theta_0, H_0)$. This supposition and and Condition (C1) imply that

$$\begin{aligned}
&\left(1 - \left[1 + r\widetilde{H}(C_{*,j}) \exp\{\widetilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_{*,j} + \widetilde{\boldsymbol{\gamma}}^\top \boldsymbol{Z}_* + \widetilde{\theta} b_*\}\right]^{-1/r}\right)^{\Delta_{*,j}} \\
&\qquad \times \left(\left[1 + r\widetilde{H}(C_{*,j}) \exp\{\widetilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_{*,j} + \widetilde{\boldsymbol{\gamma}}^\top \boldsymbol{Z}_* + \widetilde{\theta} b_*\}\right]^{-1/r}\right)^{1-\Delta_{*,j}} \\
&= \left(1 - \left[1 + rH_0(C_{*,j}) \exp\{{\boldsymbol{\beta}_0}^\top \boldsymbol{X}_{*,j} + {\boldsymbol{\gamma}_0}^\top \boldsymbol{Z}_* + \theta_0 b_*\}\right]^{-1/r}\right)^{\Delta_{*,j}} \\
&\qquad \times \left(\left[1 + rH_0(C_{*,j}) \exp\{{\boldsymbol{\beta}_0}^\top \boldsymbol{X}_{*,j} + {\boldsymbol{\gamma}_0}^\top \boldsymbol{Z}_* + \theta_0 b_*\}\right]^{-1/r}\right)^{1-\Delta_{*,j}}.
\end{aligned} \tag{S.15}$$

After using (C3) and choosing $\Delta_{*,j} = 0$ in (S.15), we then obtain

$$\begin{aligned}
&\left[1 + r\widetilde{H}(C_{*,j}) \exp\{\widetilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_{*,j} + \widetilde{\boldsymbol{\gamma}}^\top \boldsymbol{Z}_* + \widetilde{\theta} b_*\}\right]^{1/r} \\
&\qquad = \left[1 + rH_0(C_{*,j}) \exp\{{\boldsymbol{\beta}_0}^\top \boldsymbol{X}_{*,j} + {\boldsymbol{\gamma}_0}^\top \boldsymbol{Z}_* + \theta_0 b_*\}\right]^{1/r}.
\end{aligned}$$

From the monotonicity of $(1 + rx)^{1/r}$ $(r > 0)$ w.r.t. $x$, the aforementioned equation implies that

$$\widetilde{H}(C_{*,j}) \exp\{\widetilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_{*,j} + \widetilde{\boldsymbol{\gamma}}^\top \boldsymbol{Z}_* + \widetilde{\theta} b_*\} = H_0(C_{*,j}) \exp\{{\boldsymbol{\beta}_0}^\top \boldsymbol{X}_{*,j} + {\boldsymbol{\gamma}_0}^\top \boldsymbol{Z}_* + \theta_0 b_*\}. \tag{S.16}$$

We use Conditions (C3) and (C4) to get that with positive probability, we can fix $C_{*,j} \neq 0$ such that both $\widetilde{H}(C_{*,j})$ and $H_0(C_{*,j})$ are not equal to zero. (S.16) together with (C3) then imply that

$$\widetilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_{*,j} + \widetilde{\boldsymbol{\gamma}}^\top \boldsymbol{Z}_* + \widetilde{\theta} b_* = {\boldsymbol{\beta}_0}^\top \boldsymbol{X}_{*,j} + {\boldsymbol{\gamma}_0}^\top \boldsymbol{Z}_* + \theta_0 b_* + v$$

for some $v$. Using (C2), it shows that $(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}}, \widetilde{\theta}) = (\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \theta_0)$. The conclusion of $\widetilde{H} = H_0$ follows after plugging this result into (S.16).

□

### S.2.3 Proof of Theorem S.1

To prove Theorem S.1, we first need the following technical lemmas.

**Lemma S.2.** *If Conditions (C1)–(C7) hold, then, for a sufficiently small $\delta > 0$, there exists a constant $v > 0$ depending on $\mathbf{P}_0$ such that $\|H\|_\infty \leq v\{J(H)+1\}$ whenever $H \in \mathcal{H}$ and $\|H - H_0\|_2 < \delta$.*

*Proof of Lemma S.2.* Because $\|H - H_0\|_2 < \delta$ for a sufficiently small $\delta > 0$, it implies that there exist disjoint intervals $[a_i, b_i] \subset [0, T_0]$ such that $H(a_i) < H(b_i)$ and $\int_{[a_i,b_i]}\{H(t) - H_0(t)\}^2 \mathrm{d}t < \delta^2$ for each $i = 1, \ldots, k$. Therefore, there exists $t_i \in [a_i, b_i]$ satisfying $\{H(t_i) - H_0(t_i)\}^2 \leq v\delta^2$. In view of the fact that $H_0$ is uniformly bounded on $[0, T_0]$, it follows that $H(t_i) \leq K_\delta$ for some constant $K_\delta$ depending on $\delta$. For any $H \in \mathcal{H}$ with $J(H) < \infty$, Condition (C4) and $\|H - H_0\|_2 < \delta$ with sufficiently small $\delta$ imply that $J(H)$ is also bounded away from 0. Thus there exists a polynomial spline $\widetilde{H} \in \mathcal{S}(\boldsymbol{\tau}, L, d)$ such that $\|H - \widetilde{H}\|_\infty \leq vq^{-d} \leq J(H)$ (see, for example, the proof of Lemma 7.2 of Murphy and van der Vaart, 1999) with $d$ large enough. It follows that $\widetilde{H}(t_i) \leq J(H) + H(t_i) \leq J(H) + K_\delta$. Using the approximation property of polynomial spline (de Boor, 1978), $\|\widetilde{H}\|_\infty \leq v\{J(H) + K_\delta\}$, and $\|H\|_\infty$ is bounded by $v\{J(H)+1\}$ accordingly. □

**Lemma S.3.** *If Conditions (C1)–(C7) hold, then there exists a constant $v > 0$ such that*

$$\mathbf{P}\{\ell_c(\boldsymbol{\iota}; \boldsymbol{g}) - \ell_c(\boldsymbol{\iota}_0; \boldsymbol{g})\}^2 \geq v\|\boldsymbol{\iota} - \boldsymbol{\iota}_0\|_\Xi^2$$

*for $\boldsymbol{\iota}$ in a neighborhood of $\boldsymbol{\iota}_0$.*

*Proof of Lemma S.3.* From the complete likelihood function (S.9), it is shown that

$$\begin{aligned}
&\mathbf{P}\{\ell_c(\boldsymbol{\iota}; \boldsymbol{g}) - \ell_c(\boldsymbol{\iota}_0; \boldsymbol{g})\}^2 \\
&\quad = \int \Bigg( \sum_{j=1}^{m_*} (1 - \Delta_{*,j}) \big[ \log\{S_{\boldsymbol{\iota}}(C_{*,j}|\boldsymbol{X}_{*,j}, \boldsymbol{Z}_*, b_*)\} - \log\{S_{\text{sucp},\boldsymbol{\iota}_0}(C_{*,j}|\boldsymbol{X}_{*,j}, \boldsymbol{Z}_*, b_*)\} \big] \\
&\qquad + \sum_{j=1}^{m} \Delta_{*,j} \big[ \log\{1 - S_{\boldsymbol{\iota}}(C_{*,j}|\boldsymbol{X}_{*,j}, \boldsymbol{Z}_*, b_*)\} - \log\{1 - S_{\boldsymbol{\iota}_0}(C_{*,j}|\boldsymbol{X}_{*,j}, \boldsymbol{Z}_*, b_*)\} \big] \\
&\qquad + \big\{ \log\, \phi(b_*) - \log\, \phi(b_*) \big\} \Bigg)^2 \mathrm{d}\mathbf{P}, \qquad \text{(S.17)}
\end{aligned}$$

where $S_{\boldsymbol{\iota}}(C_{*,j}|\boldsymbol{X}_{*,j}, \boldsymbol{Z}_*, b_*)$ and $\phi(b)$ respectively denote the survival function of the time-to-event in the susceptible population given in (1) with parameter $\boldsymbol{\iota}$ and probability density function of

$b$ which is $\mathcal{N}(0,1)$. Using Conditions (C3) and (C5), to show (S.17) greater than or equal to $\|\boldsymbol{\iota}-\boldsymbol{\iota}_0\|_{\Xi}^2$, up to a constant, it suffices to show that

$$\int \Bigg( \sum_{j=1}^{m} \big[ \log\{1 - S_{\text{sucp},\boldsymbol{\iota}}(C_{*,j}|\boldsymbol{X}_{*,j}, \boldsymbol{Z}_*, b_*)\} - \log\{1 - S_{\text{sucp},\boldsymbol{\iota}_0}(C_{*,j}|\boldsymbol{X}_{*,j}, \boldsymbol{Z}_*, b_*)\}\big] \Bigg)^2 \mathrm{d}\mathbf{P} \\ \geq v\big\{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2 + \|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|^2 + (\theta-\theta_0)^2 + \|H - H_0\|_2^2\big\}, \tag{S.18}$$

for some constant $v > 0$.

Next, we first show the following simplified version of (S.18)

$$\int \big[ \log\{1 - S_{\text{sucp},\boldsymbol{\iota}}(C_{*,j}|\boldsymbol{X}_{*,j}, \boldsymbol{Z}_*, b_*)\} - \log\{1 - S_{\text{sucp},\boldsymbol{\iota}_0}(C_{*,j}|\boldsymbol{X}_{*,j}, \boldsymbol{Z}_*, b_*)\}\big]^2 \mathrm{d}\mathbf{P} \\ \geq v\big\{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2 + \|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|^2 + (\theta-\theta_0)^2 + \|H - H_0\|_2^2\big\}. \tag{S.19}$$

Let $g_1(s)$ denote

$$\log\Big[1 - \big\{1 + rH_s(C_{*,j})\exp(\boldsymbol{\beta}_s{}^{\top}\boldsymbol{X}_{*,j} + \boldsymbol{\gamma}_s{}^{\top}\boldsymbol{Z}_* + \theta_s b_*)\big\}^{-1/r}\Big],$$

where $H_s(C_{*,j}) = sH(C_{*,j}) + (1-s)H_0(C_{*,j})$, $\boldsymbol{\beta}_s = s\boldsymbol{\beta} + (1-s)\boldsymbol{\beta}_0$, $\boldsymbol{\gamma}_s = s\boldsymbol{\gamma} + (1-s)\boldsymbol{\gamma}_0$, and $\theta_s = s\theta + (1-s)\theta_0$, respectively. The term inside the integral of the left hand side of (S.19) is then equal to $\{g_1(1) - g_1(0)\}^2$. Application of the mean value theorem leads to $g_1(1) - g_1(0) = g_1'(\epsilon)$ for some $0 \leq \epsilon \leq 1$. It is shown that

$$\begin{aligned}
g_1'(\epsilon) &= \left( \frac{\big\{1 + rH_\epsilon(C_{*,j})\exp(\boldsymbol{\beta}_\epsilon{}^{\top}\boldsymbol{X}_{*,j} + \boldsymbol{\gamma}_\epsilon{}^{\top}\boldsymbol{Z}_* + \theta_\epsilon b_*)\big\}^{-1/r-1}\exp(\boldsymbol{\beta}_\epsilon{}^{\top}\boldsymbol{X}_{*,j} + \boldsymbol{\gamma}_\epsilon{}^{\top}\boldsymbol{Z}_* + \theta_\epsilon b_*)}{1 - \big[1 + r\big\{H_0 + \epsilon(H - H_0)\big\}(C_{*,j})\exp(\boldsymbol{\beta}_\epsilon{}^{\top}\boldsymbol{X}_{*,j} + \boldsymbol{\gamma}_\epsilon{}^{\top}\boldsymbol{Z}_* + \theta_\epsilon b_*)\big]^{-1/r}} \right) \\
&\quad \times \Big[(H - H_0)(C_{*,j}) + \big\{H_0 + \epsilon(H - H_0)\big\}(C_{*,j}) \\
&\qquad \times \big\{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^{\top}\boldsymbol{X}_{*,j} + (\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)^{\top}\boldsymbol{Z}_* + (\theta-\theta_0)b_*\big\}\Big] \\
&= \left( \frac{\big\{1 + rH_\epsilon(C_{*,j})\exp(\boldsymbol{\beta}_\epsilon{}^{\top}\boldsymbol{X}_{*,j} + \boldsymbol{\gamma}_\epsilon{}^{\top}\boldsymbol{Z}_* + \theta_\epsilon b_*)\big\}^{-1/r-1}\exp(\boldsymbol{\beta}_\epsilon{}^{\top}\boldsymbol{X}_{*,j} + \boldsymbol{\gamma}_\epsilon{}^{\top}\boldsymbol{Z}_* + \theta_\epsilon b_*)}{1 - \big[1 + r\big\{H_0 + \epsilon(H - H_0)\big\}(C_{*,j})\exp(\boldsymbol{\beta}_\epsilon{}^{\top}\boldsymbol{X}_{*,j} + \boldsymbol{\gamma}_\epsilon{}^{\top}\boldsymbol{Z}_* + \theta_\epsilon b_*)\big]^{-1/r}} \right) \\
&\quad \times \Big[(H - H_0)(C_{*,j})\left\{1 + \epsilon(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^{\top}\boldsymbol{X}_{*,j} + \epsilon(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)^{\top}\boldsymbol{Z}_* + \epsilon(\theta-\theta_0)b_*\right\} \\
&\qquad + H_0(C_{*,j})\big\{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^{\top}\boldsymbol{X}_{*,j} + (\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)^{\top}\boldsymbol{Z}_* + (\theta-\theta_0)b\big\}\Big] \\
&:= g_{1,\epsilon}(C_{*,j}, \boldsymbol{X}_{*,j}, \boldsymbol{Z}_*, b_*) \cdot \Big[(H - H_0)(C_{*,j})\left\{1 + \epsilon(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^{\top}\boldsymbol{X}_{*,j} + \epsilon(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)^{\top}\boldsymbol{Z}_* + \epsilon(\theta-\theta_0)b_*\right\} \\
&\qquad + H_0(C_{*,j})\big\{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^{\top}\boldsymbol{X}_{*,j} + (\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)^{\top}\boldsymbol{Z}_* + (\theta-\theta_0)b_*\big\}\Big],
\end{aligned}$$

where $g_{1,\epsilon}$ is a function of random variables $(C_{*,j}, \boldsymbol{X}_{*,j}, \boldsymbol{Z}_*, b_*)$. From the application of the mean value theorem and Conditions (C2)–(C5), we have

$$\begin{aligned}
&\int \big[\log\{1 - S_{\text{sucp},\boldsymbol{\iota}}(C_{*,j}|\boldsymbol{X}_{*,j}, \boldsymbol{Z}_*, b_*)\} - \log\{1 - S_{\text{sucp},\boldsymbol{\iota}_0}(C_{*,j}|\boldsymbol{X}_{*,j}, \boldsymbol{Z}_*, b_*)\}\big]^2 \mathrm{d}\mathbf{P} \\
&\quad \geq \int \Big[(H - H_0)(C_{*,j})\left\{1 + \epsilon(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \boldsymbol{X}_{*,j} + \epsilon(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^\top \boldsymbol{Z}_* + \epsilon(\theta - \theta_0) b_*\right\} \\
&\qquad + H_0(C_{*,j})\big\{(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \boldsymbol{X}_{*,j} + (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^\top \boldsymbol{Z}_* + (\theta - \theta_0) b_*\big\}\Big]^2 \mathrm{d}\mathbf{P}
\end{aligned} \tag{S.20}$$

up to a constant. To simplify the notations, we let $g_2(C_{*,j}, \boldsymbol{X}_{*,j}, \boldsymbol{Z}_*) = \big\{(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \boldsymbol{X}_{*,j} + (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^\top \boldsymbol{Z}_* + (\theta - \theta_0) b_*\big\} H_0(C_{*,j})$, $g_3(C_{*,j}) = (H - H_0)(C_{*,j})$, and $\vartheta(C_{*,j}) = 1 + \epsilon(H - H_0)(C_{*,j})/H_0(C_{*,j})$, respectively. To show (S.19), it thus suffices to verify

$$\mathbf{P}(g_2\vartheta + g_3)^2 \geq \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 + \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|^2 + (\theta - \theta_0)_2^2 + \|H - H_0\|_2^2 \tag{S.21}$$

up to a constant. To apply Lemma 25.86 of van der Vaart (1998), we need to bound $\{\mathbf{P}(g_2 g_3)\}^2$ by a constant less than $\mathbf{P}(g_2^2)\mathbf{P}(g_3^2)$. By then computing conditionally on $C_{*,j}$, we have

$$\begin{aligned}
\{\mathbf{P}(g_2 g_3)\}^2 &= \big[\mathbf{P}\{\mathbf{P}(g_2 g_3 | C_{*,j})\}\big]^2 \\
&\leq \mathbf{P}(g_3^2)\mathbf{P}\big[\{\mathbf{P}(g_2^2 | C_{*,j})\}^2\big] \\
&= \mathbf{P}(g_3^2)\mathbf{P}\Big[H_0^2(C_{*,j})\big\{\big((\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top, (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^\top, \theta - \theta_0\big) \\
&\qquad \times \big[\{\mathbf{P}(\boldsymbol{X}_{*,j}^\top, \boldsymbol{Z}_*^\top, b_*)^\top | C_{*,j}\}^{\otimes 2}\big]\big((\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top, (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^\top, \theta - \theta_0\big)^\top\big\}\Big] \\
&\leq (1 - \kappa)\mathbf{P}(g_3^2)\mathbf{P}\Big\{H_0^2(C_{*,j})\big((\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top, (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^\top, \theta - \theta_0\big) \\
&\qquad \times \mathbf{P}\big[\{(\boldsymbol{X}_{*,j}^\top, \boldsymbol{Z}_*^\top, b_*)^\top\}^{\otimes 2} | C_{*,j}\big]\big((\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top, (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^\top, \theta - \theta_0\big)^\top\Big\} \\
&= (1 - \kappa)\mathbf{P}(g_3^2)\mathbf{P}(g_2^2),
\end{aligned}$$

where the first and second inequalities follow from the Cauchy-Schwarz inequality and Condition (C6), respectively. Thus by Lemma 25.86 of van der Vaart (1998) and Conditions (C2)–(C4),

$$\mathbf{P}(g_2\vartheta + g_3)^2 \gtrsim \mathbf{P}(g_2^2) + \mathbf{P}(g_3^2) \gtrsim \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 + \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|^2 + (\theta - \theta_0)^2 + \|H - H_0\|_2^2,$$

where $\gtrsim$ denotes $\geq$ up to a constant.

The last step is to show (S.18) from its simplified version (S.19). Indeed, it can be completed by using Condition (C1) and similar arguments as shown in the proof of (S.21). □

*Proof of Theorem S.1.* To prove the claimed rate of convergence, we first show the consistency of the penalized estimator. Define

$$m_{\boldsymbol{\iota},\lambda} = \log\left(\frac{p_{\boldsymbol{\iota}} + p_0}{2p_0}\right) - \frac{\lambda}{2}\{J^2(H) - J^2(H_0)\}.$$

Under the order assumption of $\lambda$, we may assume that $\lambda \in \boldsymbol{\lambda}_n = [\widetilde{\lambda}_n, \infty)$ for

$$\widetilde{\lambda}_n = n^{-2q/(1+2q)}. \tag{S.22}$$

By the concavity of the logarithmic function, the relationship between $p_{\boldsymbol{\iota}}$ and $\ell_c(\boldsymbol{\iota}; \boldsymbol{g})$, and the definition of $\widehat{\boldsymbol{\iota}}_{c,n}$,

$$\mathbb{P}_n m_{\widehat{\boldsymbol{\iota}}_{c,n},\lambda} \geq \frac{1}{2}\mathbb{P}_n \log\left(\frac{p_{\widehat{\boldsymbol{\iota}}_{c,n}}}{p_0}\right) - \frac{\lambda}{2}\{J^2(\widehat{H}_{c,n}) - J^2(H_0)\} \geq 0 = \mathbb{P}_n m_{\boldsymbol{\iota}_0,\lambda}.$$

It can also be shown that

$$\mathbf{P}_0(m_{\boldsymbol{\iota},\lambda} - m_{\boldsymbol{\iota}_0,\lambda}) = \int \log \frac{p_{\boldsymbol{\iota}} + p_0}{2p_0} p_0 \mathrm{d}\mu - \frac{\lambda}{2}\{J^2(H) - J^2(H_0)\}.$$

Since $\log(x) \leq 2(x^{1/2} - 1)$ for $x > 0$, it follows that

$$\frac{1}{2}\int \log\left(\frac{p_{\boldsymbol{\iota}} + p_0}{2p_0}\right) p_0 \mathrm{d}\mu \leq \int \left(\frac{p_{\boldsymbol{\iota}} + p_0}{2p_0}\right)^{1/2} p_0 \mathrm{d}\mu - 1 = -\frac{1}{4}h^2(p_{\boldsymbol{\iota}} + p_0, 2p_0),$$

where $h(p_{\boldsymbol{\iota}}, p_0)$ is the Hellinger distance defined as $h^2(p_{\boldsymbol{\iota}}, p_0) = \int (p_{\boldsymbol{\iota}}^{1/2} - p_0^{1/2})^2 \mathrm{d}\mu$. Hence,

$$\mathbf{P}_0(m_{\boldsymbol{\iota},\lambda} - m_{\boldsymbol{\iota}_0,\lambda}) \leq -\frac{1}{2}h^2(p_{\boldsymbol{\iota}} + p_0, 2p_0) - \frac{\lambda}{2}\{J^2(H) - J^2(H_0)\}.$$

Using page 328 of van der Vaart and Wellner (1996), we have that

$$h(p_{\boldsymbol{\iota}} + p_0, 2p_0) \leq h(p_{\boldsymbol{\iota}}, p_0) \leq 2h(p_{\boldsymbol{\iota}} + p_0, 2p_0).$$

Thus the squared Hellinger distance $h^2(p_{\boldsymbol{\iota}} + p_0, 2p_0)$ is equivalent $h^2(p_{\boldsymbol{\iota}}, p_0)$, up to a constant. Theorem 3.4.4 of van der Vaart and Wellner (1996) and Condition (C3) imply that

$$\mathbf{P}_0\{\log(p_{\boldsymbol{\iota}}) - \log(p_0)\}^2 \leq v h^2(p_{\boldsymbol{\iota}}, p_0),$$

for some constant $v$. Hence, in view of Lemma S.3 and Condition (C3), it follows that

$$\mathbf{P}_0(m_{\boldsymbol{\iota},\lambda} - m_{\boldsymbol{\iota}_0,\lambda}) \lesssim -\|\boldsymbol{\iota} - \boldsymbol{\iota}_0\|_{\Xi}^2 - \lambda J^2(H) + \lambda,$$

where $\lesssim$ denotes $\leq$ up to a constant. This suggests the choice of

$$d_\lambda(\boldsymbol{\iota} - \boldsymbol{\iota}_0) = \left\{\|\boldsymbol{\iota} - \boldsymbol{\iota}_0\|_\Xi^2 + \lambda J^2(H)|\right\}^{1/2} \tag{S.23}$$

in Theorem 25.81 of van der Vaart (1998). Next, using the same arguments as those in Lemma 7.2 of Murphy and van der Vaart (1999), it can be shown that

$$\sup_Q \log N_{[]}\big(\epsilon, \{m_{\boldsymbol{\iota},0}, \boldsymbol{\alpha} \in \Theta, J(H) \leq M\}, L_2(Q)\big) \leq v\left(\frac{1+M}{\epsilon}\right)^{1/q}, \tag{S.24}$$

where $N_{[]}$ denotes the bracketing number of the metric space (the minimum number of $\epsilon$-brackets in $L_2(Q)$ needed to ensure that every function in $\{m_{\boldsymbol{\iota},0}, \boldsymbol{\alpha} \in \Theta, J(H) \leq M\}$ is contained in at least one bracket). Under the choice of (S.23), $d_\lambda(\boldsymbol{\iota} - \boldsymbol{\iota}_0) < \delta$ implies that $J(H) \leq \delta/\widetilde{\lambda}_n^{1/2}$. Using this fact, Lemma 2.1 of van de Geer (2000), Theorem 2.14.1 of van der Vaart and Wellner (1996), and (S.24) together imply that

$$\mathbf{P}_0 \sup_{d_\lambda(\boldsymbol{\iota}-\boldsymbol{\iota}_0)<\delta, \lambda \in \boldsymbol{\lambda}_n} \left|\mathbb{G}_n(m_{\boldsymbol{\iota},\lambda} - m_{\boldsymbol{\iota}_0,\lambda})\right| \leq v\left(1 + \frac{\delta}{\widetilde{\lambda}_n^{1/2}}\right)^{1/(2q)}.$$

Theorem 25.81 of van der Vaart (1998) yields $d_\lambda(\widehat{\boldsymbol{\iota}}_{c,n} - \boldsymbol{\iota}_0) = O_p(\delta_n + n^{-q/(1+2q)})$ for any $\delta_n \downarrow 0$ and $\delta_n \geq (n^{2q}\widetilde{\lambda}_n)^{-1/(8q-2)}$, which concludes the consistency of $\widehat{\boldsymbol{\iota}}_{c,n}$ by (S.22).

To show the rate of convergence, using Lemma S.2, it is reasonable to restrict $H$ to the set $\mathcal{H}_n = \{H : \|H\|_\infty \leq v(J(H) + 1)\}$ for a large constant $v$. If $d_\lambda(\boldsymbol{\iota} - \boldsymbol{\iota}_0) < \delta$ and $\lambda \in \boldsymbol{\lambda}_n$, then $\|\boldsymbol{\iota} - \boldsymbol{\iota}_0\|_\Xi < \delta$, $J(H) < \delta/\widetilde{\lambda}_n^{1/2}$, and hence, $\|H\|_\infty \leq v(\delta/\widetilde{\lambda}_n^{1/2} + 1)$. Using Taylor expansion along with condition (C1), (C2), and (C5), it can be shown that the parametric part of $m_{\boldsymbol{\iota},0}$ is essentially Lipschitz with respect to $\boldsymbol{\alpha}$. The above two facts and Example 19.10 of van der Vaart (1998) imply that

$$\log N_{[]}\big(\epsilon, \{m_{\boldsymbol{\iota},0} : \lambda \in \boldsymbol{\lambda}_n, H \in \mathcal{H}_n, d_\lambda(\boldsymbol{\iota} - \boldsymbol{\iota}_0) < \delta\}, L_2(\mathbf{P}_0)\big) \leq v\left(\frac{1 + \delta/\widetilde{\lambda}_n^{1/2}}{\epsilon}\right)^{1/q}.$$

Thus, Lemma 19.36 of van der Vaart (1998) shows that

$$\mathbf{P}_0 \sup_{d_\lambda(\boldsymbol{\iota}-\boldsymbol{\iota}_0)<\delta, \lambda \in \boldsymbol{\lambda}_n} \left|\mathbb{G}_n(m_{\boldsymbol{\iota},\lambda} - m_{\boldsymbol{\iota}_0,\lambda})\right| \leq vJ_n(\delta)\left\{1 + \frac{J_n(\delta)}{\delta^2 n^{1/2}}\right\},$$

where

$$J_n(\delta) = \int_0^\delta \left(\frac{1 + \delta/\widetilde{\lambda}_n^{1/2}}{\epsilon}\right)^{1/(2q)} d\epsilon = v\left(1 + \frac{\delta}{\widetilde{\lambda}_n^{1/2}}\right)^{1/(2q)} \delta^{1-1/(2q)} = v\left\{\delta^{1-1/(2q)} + \delta n^{1/2(2q+1)}\right\},$$

for some constant $v$. Therefore, Theorem 25.81 of van der Vaart (1998) implies

$$\|\widehat{\boldsymbol{\iota}}_{c,n} - \boldsymbol{\iota}_0\|_\Xi = O_p(\delta_n + \widetilde{\lambda}_n) = O_p(\delta_n + n^{-q/(1+2q)}), \tag{S.25}$$

with $\delta_n$ satisfying

$$J_n(\delta_n)\left\{1 + \frac{J_n(\delta_n)}{\delta^2 n^{1/2}}\right\} \leq n^{1/2}\delta_n^2.$$

Brief calculation shows that the optimal rate of $\delta_n$ in the aforementioned equation is $n^{-q/(1+2q)}$. This result together with (S.25) completes the proof of Theorem S.1. □

### S.2.4 Proof of Lemma S.1

Using Condition (C1), we directly calculate that

$$\begin{aligned}
&\dot{\ell}_{c,1}(\boldsymbol{\iota}; \boldsymbol{g}) \\
&= \sum_{j=1}^{m_*}(1-\Delta_{*,j})\left\{\frac{-H(C_{*,j})\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^\top \boldsymbol{Z}_* + \theta b_*)({\boldsymbol{X}_{*,j}}^\top, {\boldsymbol{Z}_*}^\top, b_*)^\top}{1 + rH(C_{*,j})\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^\top \boldsymbol{Z}_* + \theta b_*)}\right\} \\
&\quad + \sum_{j=1}^{m_*}\Delta_{*,j}\left(\frac{\left\{1 + rH(C_{*,j})\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^\top \boldsymbol{Z}_* + \theta b_*)\right\}^{-1/r-1}}{1 - \left\{1 + rH(C_{*,j})\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^\top \boldsymbol{Z}_* + \theta b_*)\right\}^{-1/r}}\right) \\
&\qquad \times H(C_{*,j})\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^\top \boldsymbol{Z}_* + \theta b_*)({\boldsymbol{X}_{*,j}}^\top, {\boldsymbol{Z}_*}^\top, b_*)^\top \\
&= \sum_{j=1}^{m_*} H(C_{*,j})\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^\top \boldsymbol{Z}_* + \theta b_*)({\boldsymbol{X}_{*,j}}^\top, {\boldsymbol{Z}_*}^\top, b_*)^\top \\
&\quad \times \left[\frac{\Delta_{*,j}\left\{1 + rH(C_{*,j})\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^\top \boldsymbol{Z}_* + \theta b_*)\right\}^{1/r}}{1 - \left\{1 + rH(C_{*,j})\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^\top \boldsymbol{Z}_* + \theta b_*)\right\}^{-1/r}}\right. \\
&\qquad \left. - \frac{1}{1 + rH(C_{*,j})\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^\top \boldsymbol{Z}_* + \theta b_*)}\right].
\end{aligned} \tag{S.26}$$

After denoting

$$\begin{aligned}
&Q_{c,j}(C_{*,j}, X_{*,j}, Z_*, b_*; \boldsymbol{\iota}) \\
&= \exp(\boldsymbol{\beta}^\top \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^\top \boldsymbol{Z}_* + \theta b_*)\left[\frac{\Delta_{*,j}\left\{1 + rH(C_{*,j})\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^\top \boldsymbol{Z}_* + \theta b_*)\right\}^{1/r}}{1 - \left\{1 + rH(C_{*,j})\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^\top \boldsymbol{Z}_* + \theta b_*)\right\}^{-1/r}}\right. \\
&\qquad \left. - \frac{1}{1 + rH(C_{*,j})\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^\top \boldsymbol{Z}_* + \theta b_*)}\right],
\end{aligned}$$

(S.26) can be written as

$$\dot{\ell}_{c,1}(\boldsymbol{\iota}; \boldsymbol{g}) = \sum_{j=1}^{m_*} H(C_{*,j})Q_{c,j}(C_{*,j}, X_{*,j}, Z_*, b_*; \boldsymbol{\iota})({\boldsymbol{X}_{*,j}}^\top, {\boldsymbol{Z}_*}^\top, b_*)^\top. \tag{S.27}$$

Similarly, differentiating the complete log-likelihood function $\ell_c(\boldsymbol{\iota};\boldsymbol{g})$ at $H$ along $w$ yields

$$\dot{\ell}_{c,2}(\boldsymbol{\iota};\boldsymbol{g})[w]=\sum_{j=1}^{m_*}w(C_{*,j})Q_{c,j}(C_{*,j},X_{*,j},Z_*,b_*;\boldsymbol{\iota}),\tag{S.28}$$

where $w\in\mathcal{W}$ be the class of functions such that $H+sw\in\mathcal{H}$ for $s$ in a small interval containing 0. Moreover, for $\boldsymbol{w}=(w_1,\ldots,w_p)^\top$ with $w_k\in\mathcal{W}$, $k=1,\cdots,p$, let $\dot{\ell}_{c,2}(\boldsymbol{\iota};\boldsymbol{g})[\boldsymbol{w}]=(\dot{\ell}_{c,2}(\boldsymbol{\iota};\boldsymbol{g})[w_1],\ldots,\dot{\ell}_{c,2}(\boldsymbol{\iota};\boldsymbol{g})[w_p])^\top$. To see that $\boldsymbol{w}_c^*$ exists in (S.12), we only need to show that the normal equation

$$E\dot{\ell}^*_{c,2}(\boldsymbol{\iota};\boldsymbol{g})\dot{\ell}_{c,1}(\boldsymbol{\iota};\boldsymbol{g})-E\dot{\ell}^*_{c,2}(\boldsymbol{\iota};\boldsymbol{g})\dot{\ell}_{c,2}(\boldsymbol{\iota};\boldsymbol{g})[\boldsymbol{w}_c^*]=0$$

has a solution, where $\dot{\ell}^*_{c,2}(\boldsymbol{\iota};\boldsymbol{g})$ is the adjoint operator of $\dot{\ell}_{c,2}(\boldsymbol{\iota};\boldsymbol{g})$ (van der Vaart, 2002). Expression (S.28) implies that $\dot{\ell}_{c,2}(\boldsymbol{\iota};\boldsymbol{g})$ is self-adjoint, and thus, writing that $\boldsymbol{C}_*=(C_{*,1},\ldots,C_{*,m_*})^\top$,

$$\boldsymbol{w}_c^*=\frac{E\big[\dot{\ell}_{c,1}(\boldsymbol{\iota};\boldsymbol{g})\{\sum_{j=1}^{m_*}Q_{c,j}(C_{*,j},X_{*,j},Z_*,b_*;\boldsymbol{\iota})\}|\boldsymbol{C}_*\big]}{E\big[\{\sum_{j=1}^{m_*}Q_{c,j}(C_{*,j},X_{*,j},Z_*,b_*;\boldsymbol{\iota})\}^2|\boldsymbol{C}_*\big]}\tag{S.29}$$

exists, provided (C1)–(C4) hold, where $\dot{\ell}_{c,1}(\boldsymbol{\iota};\boldsymbol{g})$ is given in (S.27).

The efficient score of $\boldsymbol{\alpha}$ for the complete likelihood is $\dot{\ell}_{c,1}(\boldsymbol{\iota};\boldsymbol{g})-\dot{\ell}_{c,2}(\boldsymbol{\iota};\boldsymbol{g})[\boldsymbol{w}_c^*]$. The efficient information of $\boldsymbol{\alpha}$ for the complete likelihood thus takes the form of

$$\begin{aligned}I_c(\boldsymbol{\alpha})&=E\big\{\dot{\ell}_{c,1}(\boldsymbol{\iota};\boldsymbol{g})-\dot{\ell}_{c,2}(\boldsymbol{\iota};\boldsymbol{g})[\boldsymbol{w}_c^*]\big\}^{\otimes 2}\\&=E\left[\sum_{j=1}^{m_*}Q_{c,j}(C_{*,j},X_{*,j},Z_*,b_*;\boldsymbol{\iota})\big\{H(C_{*,j})(\boldsymbol{X}_{*,j}{}^\top,\boldsymbol{Z}_*{}^\top,b_*)^\top-\boldsymbol{w}_c^*\big\}\right]^{\otimes 2},\end{aligned}$$

with $\boldsymbol{w}_c^*$ given in (S.29).

Using similar arguments above and a different expression of $Q_{c,j}$ in (S.29) based on the observed likelihood function, we can prove the existence of $\boldsymbol{w}^*$ and obtain a form of $I(\boldsymbol{\alpha})$. The proof is thus omitted.

### S.2.5 Proof of Theorem S.2

*Proof of Theorem S.2.* We first notice that $\widehat{\boldsymbol{\iota}}_{c,n}$ maximizes the penalized (complete) likelihood (S.10) rather than an ordinary likelihood, thus $\widehat{\boldsymbol{\iota}}_{c,n}$ does not satisfy the efficient score equation

$$\mathbb{P}_n\big\{\dot{\ell}_{c,1}(\boldsymbol{\iota};\boldsymbol{g})-\dot{\ell}_{c,2}(\boldsymbol{\iota};\boldsymbol{g})[\boldsymbol{w}_c^*]\big\}=0.$$

However, if we can show that the distance between $\widehat{\boldsymbol{\alpha}}_{c,n}$ and the efficient estimator is bounded above by $o_p(n^{-1/2})$, then the result follows.

To show this, we first show that

$$\mathbb{P}_n\big\{\dot{\ell}_{c,1}(\widehat{\boldsymbol{\iota}}_{c,n};\boldsymbol{g}) - \dot{\ell}_{c,2}(\widehat{\boldsymbol{\iota}}_{c,n};\boldsymbol{g})[\boldsymbol{w}_c^*]\big\} = o_p(n^{-1/2}) \tag{S.30}$$

which can begin with studying the upper bound of the penalization term. Indeed, if we plug $\big((\widehat{\boldsymbol{\alpha}}_{c,n}+s\boldsymbol{a})^\mathsf{T}, \widehat{H}_{c,n}-sw\big)^\mathsf{T}$ with $w \in \mathcal{W}\cap\mathcal{H}_n$ satisfying $J(w) < \infty$, into the penalized log-likelihood function (S.10), where $\boldsymbol{a}$ is a $p$-dimensional vector. Differentiating at $s = 0$, it is shown that

$$\mathbb{P}_n\big\{\dot{\ell}_{c,1}(\widehat{\boldsymbol{\iota}}_{c,n};\boldsymbol{g})^\mathsf{T}\boldsymbol{a} - \dot{\ell}_{c,2}(\widehat{\boldsymbol{\iota}}_{c,n};\boldsymbol{g})[w]\big\} + \lambda\int(\widehat{H}_{c,n})^{(q)}(t)w^{(q)}(t)\mathrm{d}t = 0. \tag{S.31}$$

Using the Cauchy-Schwarz inequality, the $\lambda\int(\widehat{H}_{c,n})^{(q)}(t)w^{(q)}(t)\mathrm{d}t$ is bounded by $\lambda J(\widehat{H}_{c,n})J(w)$. In Theorem S.1, it has been shown that

$$J(\widehat{H}_{c,n}) = O_p(1).$$

Readers are also referred to Lemma 7.1 of Murphy and van der Vaart (1999) for additional auxiliary results. Moreover, it is assumed that $\lambda = o_p(n^{-1/2})$, thus it follows that

$$\lambda J(\widehat{H}_{c,n})J(w) = o_p(n^{-1/2}). \tag{S.32}$$

As a result, the penalized estimator $\widehat{\boldsymbol{\iota}}_{c,n}$ satisfies the efficient score equation, up to a negligible $o_p(n^{-1/2})$ term. It is obvious to show that (S.31) is free of $\boldsymbol{a}$ and thus

$$\mathbb{P}_n\{\dot{\ell}_{c,1}(\widehat{\boldsymbol{\iota}}_{c,n};\boldsymbol{g})\} = \mathbf{0}. \tag{S.33}$$

(S.31) and (S.32) together imply that for any $w \in \mathcal{W}\cap\mathcal{H}_n$,

$$\mathbb{P}_n\big\{\dot{\ell}_{c,2}(\widehat{\boldsymbol{\iota}}_{c,n};\boldsymbol{g})[w]\big\} = o_p(n^{-1/2}). \tag{S.34}$$

We next only need to verify $\mathbb{P}_n\{\dot{\ell}_{c,2}(\widehat{\boldsymbol{\iota}}_{c,n};\boldsymbol{g})[\boldsymbol{w}_c^*]\} = o_p(n^{-1/2})$ for least favorable direction $\boldsymbol{w}_c^*$. Because each component of $\boldsymbol{w}_c^*$ has a bounded derivative, it is also a function with bounded variation. Using the arguments in Billingsley (1995, pp. 415–416) for functions with bounded variation and Jackson's Theorem in de Boor (1978, pp. 149), it can be shown that there exits a $\boldsymbol{w}_n \in (\mathcal{W} \cap \mathcal{H}_n)^p$ such that $\|\boldsymbol{w}_n - \boldsymbol{w}_c^*\|_2 = O(n^{-1/(2q+1)})$. Furthermore, we have

$$\mathbf{P}\big\{\ell_c\big(\boldsymbol{\alpha}_0, H_0 + s\boldsymbol{a}^\mathsf{T}(\boldsymbol{w}_c^* - \boldsymbol{w}_n); \boldsymbol{g}\big)\big\} \le \mathbf{P}\big\{\ell_c(\boldsymbol{\alpha}_0, H_0; \boldsymbol{g})\big\}$$

for $s$ with small absolute value and $\boldsymbol{a} \in \mathbb{R}^p$, then $\mathbf{P}\{\dot{\ell}_{c,2}(\boldsymbol{\iota}_0; \boldsymbol{g})[\boldsymbol{w}_c^* - \boldsymbol{w}_n]\} = \mathbf{0}$. Therefore we can write

$$\mathbb{P}_n\big\{\dot{\ell}_{c,2}(\widehat{\boldsymbol{\iota}}_{c,n}; \boldsymbol{g})[\boldsymbol{w}_c^*]\big\} = I_{1,n} + I_{2,n},$$

where

$$I_{1,n} = (\mathbb{P}_n - \mathbf{P})\big\{\dot{\ell}_{c,2}(\widehat{\boldsymbol{\iota}}_{c,n}; \boldsymbol{g})[\boldsymbol{w}_c^* - \boldsymbol{w}_n]\big\}$$

and

$$I_{2,n} = \mathbf{P}\big\{\dot{\ell}_{c,2}(\widehat{\boldsymbol{\iota}}_{c,n}; \boldsymbol{g})[\boldsymbol{w}_c^* - \boldsymbol{w}_n] - \dot{\ell}_{c,2}(\boldsymbol{\iota}_0; \boldsymbol{g})[\boldsymbol{w}_c^* - \boldsymbol{w}_n]\big\}.$$

Let $I_{1,n,k}$ be $k$-th component of $I_{1,n}$ and denote

$$\boldsymbol{A}_{1,k} = \big\{\dot{\ell}_{c,2}(\boldsymbol{\iota}; \boldsymbol{g})[w_{c,k}^* - w_{n,k}] : \boldsymbol{\iota} \in \Theta \times \mathcal{H}_n,\ w_{n,k} \in \mathcal{W} \cap \mathcal{H}_n \text{ and } \|w_{c,k}^* - w_{n,k}\|_2 \le vn^{-1/(2q+1)}\big\},$$

$k = 1, \ldots, p$. It can be argued that the $\epsilon$-bracketing numbers associated with $L_2(\mathbf{P})$-norm for $\Theta$, $\mathcal{H}_n$, and $\{w_{n,k} \in \mathcal{W} \cap \mathcal{H}_n : \|w_{c,k}^* - w_{n,k}\|_2 \le vn^{-1/(2q+1)}\}$ are $v(1/\epsilon)^p$, $v(1/\epsilon)^{vn^{1/(2q+1)}}$, and $v(1/\epsilon)^{vn^{1/(2q+1)}}$, respectively. Therefore, the $\epsilon$-bracketing number for $\boldsymbol{A}_{1,k}$ is bounded by $v(1/\epsilon)^p(1/\epsilon)^{vn^{1/(2q+1)}}(1/\epsilon)^{vn^{1/(2q+1)}}$, which results in a $\mathbf{P}$-Donsker class for $\boldsymbol{A}_{1,k}$ by Theorem 19.5 in van der Vaart (1998), $k = 1, \ldots, p$. Since

$$\dot{\ell}_{c,2}(\widehat{\boldsymbol{\iota}}_{c,n}; \boldsymbol{g})[w_{c,k}^* - w_{n,k}] \in \boldsymbol{A}_{1,k}$$

and as $n \to \infty$,

$$\mathbf{P}\big\{\dot{\ell}_{c,2}(\widehat{\boldsymbol{\iota}}_{c,n}; \boldsymbol{g})[w_{c,k}^* - w_{n,k}]\big\}^2 \le v\|w_{c,k}^* - w_{n,k}\|_\infty^2 \to 0,$$

then by Corollary 2.3.12 of van der Vaart and Wellner (1996) we have

$$I_{1,n,k} = o_p(n^{-1/2}) \quad k = 1, \ldots, p. \tag{S.35}$$

By the Cauchy-Schwarz inequality and Conditions (C2)–(C5), it can be shown that each component of $I_{2,n}$,

$$\begin{aligned} I_{2,n,k} &= \mathbf{P}\big\{\dot{\ell}_{c,2}(\widehat{\boldsymbol{\iota}}_{c,n}; \boldsymbol{g})[w_{c,k}^* - w_{n,k}] - \dot{\ell}_{c,2}(\boldsymbol{\iota}_0; \boldsymbol{g})[w_{c,k}^* - w_{n,k}]\big\} \\ &\le v \cdot \operatorname{dist}(\widehat{\boldsymbol{\iota}}_{c,n}, \boldsymbol{\iota}_0)\|w_{c,k}^* - w_{n,k}\|_\infty = o_p(n^{-1/2}), \end{aligned} \tag{S.36}$$

$k = 1, \ldots, p$. (S.35) and (S.36) imply that

$$\mathbb{P}_n\{\dot{\ell}_{c,2}(\widehat{\boldsymbol{\iota}}_{c,n}; \boldsymbol{g})[w^*_{c,k}]\} = I_{1,n,k} + I_{2,n,k} = o_p(n^{-1/2}), \quad k = 1, \ldots, p. \tag{S.37}$$

Thus, (S.31), (S.33), (S.34), and (S.37) together show that (S.30) holds.

We then show the asymptotic normality and efficiency of the estimator $\widehat{\boldsymbol{\alpha}}_{c,n}$ using Theorem 25.54 in van der Vaart (1998). For notational convenience, in the following, let $\widetilde{\ell}_{c,\boldsymbol{\alpha},H}(\boldsymbol{g})$ denote the semiparametric efficient score function under general $\boldsymbol{\alpha}$ and $H$ for the complete data likelihood. We also write $\mathbf{P}_{\boldsymbol{\iota}}\widetilde{\ell}_{c,\widehat{\boldsymbol{\alpha}},\widehat{H}}$ as an abbreviation for $\int \widetilde{\ell}_{c,\widehat{\boldsymbol{\alpha}},\widehat{H}}(\boldsymbol{g})\mathrm{d}\mathbf{P}_{\boldsymbol{\iota}}$, which is an integration taken with respect to $\boldsymbol{g}$ only and not with respect to $\widehat{\boldsymbol{\alpha}}$ nor $\widehat{H}$. Under the result of (S.30), we only need to verify conditions

$$\mathbf{P}_{\widehat{\boldsymbol{\alpha}}_{c,n},H_0}\widetilde{\ell}_{c,\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}} = o_p(n^{-1/2} + \|\widehat{\boldsymbol{\alpha}}_{c,n} - \boldsymbol{\alpha}_0\|), \tag{S.38}$$

and

$$\mathbf{P}_0\left\|\widetilde{\ell}_{c,\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}} - \widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0}\right\|^2 \xrightarrow{\mathbf{P}} 0, \qquad \mathbf{P}_{\widehat{\boldsymbol{\alpha}}_{c,n},H_0}\|\widetilde{\ell}_{\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}}\|^2 = O_p(1). \tag{S.39}$$

For (S.38), in view of the fact that $\mathbf{P}_{\boldsymbol{\alpha},H}\widetilde{\ell}_{\boldsymbol{\alpha},H} = 0$ for all $(\boldsymbol{\alpha}, H)$, write

$$\begin{aligned}
\mathbf{P}_{\widehat{\boldsymbol{\alpha}}_{c,n},H_0}\widetilde{\ell}_{c,\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}} &= (\mathbf{P}_0 - \mathbf{P}_{\boldsymbol{\alpha}_0,\widehat{H}_{c,n}})\widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0} + (\mathbf{P}_{\widehat{\boldsymbol{\alpha}}_{c,n},H_0} - \mathbf{P}_{\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}})(\widetilde{\ell}_{c,\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}} - \widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0}) \\
&\quad + (\mathbf{P}_{\boldsymbol{\alpha}_0,\widehat{H}_{c,n}} - \mathbf{P}_0 - \mathbf{P}_{\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}} + \mathbf{P}_{\widehat{\boldsymbol{\alpha}}_{c,n},H_0})\widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0} \\
&= I_{3,n} + I_{4,n} + I_{5,n}.
\end{aligned} \tag{S.40}$$

The definition of efficient score in van der Vaart (1998, pp. 369) shows that $\widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0}$ is orthogonal to all functions in the span of $\dot{\ell}_{c,2}(\boldsymbol{\iota}_0)$. It then yields

$$(\mathbf{P}_0 - \mathbf{P}_{\boldsymbol{\alpha}_0,\widehat{H}_{c,n}})\widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0} = \mathbf{P}_0\widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0}\left\{\frac{p_0 - p_{\boldsymbol{\alpha}_0,\widehat{H}_{c,n}}}{p_0} - \dot{\ell}_{c,2}(\boldsymbol{\alpha}_0, H_0)(H_0 - \widehat{H}_{c,n})\right\}.$$

Using the Taylor expansion, it is possible to show that

$$\left|(\mathbf{P}_0 - \mathbf{P}_{\boldsymbol{\alpha}_0,\widehat{H}_{c,n}})\widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0}\right| \leq \int \left|\widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0}\right| \left|\frac{\mathrm{d}^2}{\mathrm{d}s^2}p_{\boldsymbol{\alpha}_0,H_0+s(\widehat{H}_{c,n}-H_0)}\right| \mathrm{d}\mu$$

for $0 < s < 1$. Straightforward differentiation and Condition (C3) imply that

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2}p_{\boldsymbol{\alpha}_0,H_0+s(\widehat{H}_{c,n}-H_0)}$$

can be upper bounded by $v(\widehat{H}_{c,n} - H_0)^2$ for a positive constant $v$ independent with $\boldsymbol{g}$ and all $s$. It follows that $I_{3,n} = O_p(1)\|\widehat{H}_{c,n} - H_0\|_2^2$. By the Taylor expansion, $I_{4,n}$ can be written as

$$\int (\widetilde{\ell}_{c,\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}} - \widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0})\dot{\ell}_2(\widehat{\boldsymbol{\alpha}}_{c,n}, H_0)(H_0 - \widehat{H}_{c,n})p_0 \mathrm{d}\mu$$
$$- \frac{1}{2}\int (\widetilde{\ell}_{c,\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}} - \widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0})\frac{\mathrm{d}^2}{\mathrm{d}s^2}p_{\widehat{\boldsymbol{\alpha}}_{c,n},H_0+s(\widehat{H}_{c,n}-H_0)}\mathrm{d}\mu.$$

Since $\widehat{\boldsymbol{\alpha}}_{c,n}$ converges to $\boldsymbol{\alpha}_0$ as shown in Theorem S.1, $|\dot{\ell}_{c,2}(\widehat{\boldsymbol{\alpha}}_{c,n}, H_0)(H_0 - \widehat{H}_{c,n})|$ is upper bounded by $|\widehat{H}_{c,n} - H_0|$, up to a constant not depending on $\boldsymbol{g}$, with probability approaching 1. This, along with Conditions (C2) and (C5), implies that

$$|\widetilde{\ell}_{c,\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}} - \widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0}| \le v \cdot \mathrm{dist}(\widehat{\boldsymbol{\iota}}_{c,n}, \boldsymbol{\iota}_0)^2$$

on an event with probability approaching 1. Moreover, $(\mathrm{d}^2/\mathrm{d}s^2)p_{\widehat{\boldsymbol{\alpha}}_{c,n},H_0+s(\boldsymbol{w}_{c,n}hH_{c,n}-H_0)}$ is bounded above by $(\widehat{H}_{c,n} - H_0)^2$, up to a constant, with probability approaching 1. It thus follows that $I_{4,n} = O_p\big(\|\widehat{H}_{c,n} - H_0\|_2^2 + \|\widehat{\boldsymbol{\alpha}}_{c,n} - \boldsymbol{\alpha}_0\|\|\widehat{H}_{c,n} - H_0\|_2\big)$. We further use the Taylor expansion and the Cauchy-Schwarz inequality to obtain that $I_{5,n} = O_p\big(\|\widehat{\boldsymbol{\iota}}_{c,n} - \boldsymbol{\iota}_0\|_2^2 + \|\widehat{\boldsymbol{\alpha}}_{c,n} - \boldsymbol{\alpha}_0\|\|\widehat{H}_{c,n} - H_0\|_2\big)$. Therefore, (S.38) follows from the rate of convergence of $\widehat{\boldsymbol{\alpha}}_{c,n}$ and $\widehat{H}_{c,n}$ as shown in Theorem S.1.

For (S.39), we first use the dominated convergence theorem and the consistency of $\widehat{\boldsymbol{\iota}}_{c,n}$ to obtain that $\mathbf{P}_0 \big\|\widetilde{\ell}_{\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}} - \widetilde{\ell}_{\boldsymbol{\alpha}_0,H_0}\big\|^2 \to 0$ in probability. Furthermore, by the consistency of $\widehat{\boldsymbol{\alpha}}_{c,n}$, it can be shown that $\mathbf{P}_{\widehat{\boldsymbol{\alpha}}_{c,n},H_0}\|\widetilde{\ell}_{\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}}\|^2 = O_p(1)$ with similar arguments as those shown in (S.35). As a result, (S.39) holds. To sum up, it is possible to use the results in Theorem 25.54 of van der Vaart (1998), and thus $\widehat{\boldsymbol{\alpha}}_{c,n}$ is efficient. □

### S.2.6 Proof of Theorems 2 and 3

We only need to show the similar result as in Lemma S.3 such that

$$\mathbf{P}\{\ell(\boldsymbol{\iota}; \boldsymbol{g}) - \ell(\boldsymbol{\iota}_0; \boldsymbol{g})\}^2 \ge v\|\boldsymbol{\iota} - \boldsymbol{\iota}_0\|_{\Xi}^2\,, \tag{S.41}$$

whenever $\mathrm{dist}(\boldsymbol{\iota}, \boldsymbol{\iota}_0) < \varepsilon$ for some constant $\varepsilon > 0$. Indeed, the left hand side of (S.41) can be written as

$$\mathbf{P}\bigg[\log\bigg\{\int \mathcal{L}_c(\boldsymbol{\iota}; \boldsymbol{g})\mathrm{d}b_*\bigg\} - \log\bigg\{\int \mathcal{L}_c(\boldsymbol{\iota}_0; \boldsymbol{g})\mathrm{d}b_*\bigg\}\bigg]^2 \ge v\|\boldsymbol{\iota} - \boldsymbol{\iota}_0\|_{\Xi}^2\,. \tag{S.42}$$

Next consider $\mathcal{L}_c\{s\boldsymbol{\iota} + (1-s)\boldsymbol{\iota}_0; \boldsymbol{g}\}$, and then following the proof of Lemma S.3, it can be shown that the left hand side of (S.42) is bounded below by

$$\mathbf{P}\left(\frac{(\partial/\partial s)\left[\int \mathcal{L}_c\{s\boldsymbol{\iota} + (1-s)\boldsymbol{\iota}_0; \boldsymbol{g}\}\mathrm{d}b_* - \int \mathcal{L}_c(\boldsymbol{\iota}_0; \boldsymbol{g})\mathrm{d}b_*\right]\big|_{s=\epsilon}}{\int \mathcal{L}_c\{\epsilon\boldsymbol{\iota} + (1-\epsilon)\boldsymbol{\iota}_0; \boldsymbol{g}\}\mathrm{d}b_*}\right)^2,$$

for some $\epsilon \in [0, 1]$. By Conditions (C3)–(C5), it thus suffices to show

$$\mathbf{P}\left(\int \frac{\partial}{\partial s}\left[\mathcal{L}_c\{s\boldsymbol{\iota} + (1-s)\boldsymbol{\iota}_0; \boldsymbol{g}\} - \mathcal{L}_c(\boldsymbol{\iota}_0; \boldsymbol{g})\right]\Big|_{s=\epsilon}\mathrm{d}b_*\right)^2 \geq v\|\boldsymbol{\iota} - \boldsymbol{\iota}_0\|_{\Xi}^2.$$

Using the mean value theorem and the proof in van der Vaart (2002, pp. 431), the aforementioned equation is satisfied, which completes the proof of (S.41) as a consequence. The rest of the proof follows the same arguments as in Theorems S.1 and S.2, and are thus omitted.

# References


Bickel, P. J., Klaassen, C. A. and Wellner, J. A. (1993) *Efficient and adaptive estimation for semiparametric models*. Baltimore: Johns Hopkins University Press.

Billingsley, P. (1995) *Probability and Measure*. New York: Wiley.

de Boor, C. (1978) *A Practical Guide to Splines*, vol. 27. Springer-Verlag.

van de Geer, S. A. (2000) *Empirical Processes in M-estimation*. Cambridge, UK; New York: Cambridge University Press.

Huang, J. and Wellner, J. A. (1997) Interval censored survival data: A review of recent progress. In *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis* (ed. T. R. Lin, D. Y. and Fleming), 123–169. New York: Springer-Verlag.

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

Stone, C. J. (1985) Additive regression and other nonparametric models. *The Annals of Statistics*, **13**, 689–705.

van der Vaart, A. (2002) Semiparameric statistics. *Lectures on Probability Theory and Statistics*, 331–457.

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

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

Zeng, D., Lin, D. Y. and Yin, G. (2005) Maximum likelihood estimation for the proportional odds model with random effects. *Journal of the American Statistical Association*, **100**, 470–483.

Zhang, Y., Hua, L. and Huang, J. (2010) A spline-based semiparametric maximum likelihood estimation method for the cox model with interval-censored data. *Scandinavian Journal of Statistics*, **37**, 338–354.

## S.3 Additional simulation results

Table 1: Results of the simulation study for $\beta = -1$ and $\gamma = -1$ with 5 interior knots. Here RB, $\widetilde{\text{RB}}$, SD, SE, and CP denote the relative mean bias, the relative median bias, the standard deviation, the median of estimated standard error, and the 95% coverage probability, respectively. PAR: Parameter.

| | PAR | $\theta = 2$ | | | | | $\theta = 1$ | | | | | $\theta = 0.5$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | RB | $\widetilde{\text{RB}}$ | SD | SE | CP | RB | $\widetilde{\text{RB}}$ | SD | SE | CP | RB | $\widetilde{\text{RB}}$ | SD | SE | CP |
| $n = 300$ | | | | | | | | | | | | | | | | |
| $r \to 0$ | | | | | | | | | | | | | | | | |
| | $\beta$ | −0.01 | −0.02 | 0.14 | 0.13 | 0.92 | −0.01 | −0.01 | 0.11 | 0.10 | 0.94 | 0.00 | 0.00 | 0.09 | 0.09 | 0.95 |
| | $\gamma$ | −0.02 | −0.02 | 0.28 | 0.25 | 0.94 | 0.00 | 0.00 | 0.15 | 0.14 | 0.94 | 0.00 | 0.00 | 0.10 | 0.10 | 0.95 |
| | $\theta$ | −0.01 | −0.03 | 0.27 | 0.19 | 0.90 | −0.01 | −0.02 | 0.12 | 0.10 | 0.93 | −0.02 | −0.02 | 0.08 | 0.08 | 0.96 |
| $r = 1$ | | | | | | | | | | | | | | | | |
| | $\beta$ | −0.01 | −0.01 | 0.16 | 0.15 | 0.95 | 0.01 | 0.00 | 0.14 | 0.13 | 0.94 | 0.00 | 0.00 | 0.12 | 0.12 | 0.95 |
| | $\gamma$ | −0.02 | −0.01 | 0.27 | 0.26 | 0.94 | 0.02 | 0.01 | 0.19 | 0.16 | 0.93 | 0.01 | 0.00 | 0.14 | 0.13 | 0.92 |
| | $\theta$ | −0.01 | −0.02 | 0.26 | 0.18 | 0.93 | 0.00 | −0.03 | 0.21 | 0.12 | 0.93 | −0.03 | −0.04 | 0.17 | 0.15 | 0.94 |
| $r = 2$ | | | | | | | | | | | | | | | | |
| | $\beta$ | 0.00 | 0.00 | 0.23 | 0.20 | 0.92 | −0.01 | −0.02 | 0.19 | 0.18 | 0.93 | 0.01 | 0.00 | 0.17 | 0.17 | 0.96 |
| | $\gamma$ | 0.01 | −0.01 | 0.30 | 0.29 | 0.95 | 0.03 | 0.02 | 0.21 | 0.20 | 0.93 | 0.00 | −0.01 | 0.19 | 0.18 | 0.94 |
| | $\theta$ | 0.01 | −0.02 | 0.33 | 0.20 | 0.91 | −0.01 | −0.03 | 0.24 | 0.18 | 0.93 | −0.03 | −0.04 | 0.26 | 0.24 | 0.93 |
| $n = 1000$ | | | | | | | | | | | | | | | | |
| $r \to 0$ | | | | | | | | | | | | | | | | |
| | $\beta$ | −0.01 | −0.02 | 0.08 | 0.07 | 0.93 | 0.00 | 0.00 | 0.05 | 0.05 | 0.95 | 0.00 | 0.00 | 0.05 | 0.05 | 0.94 |
| | $\gamma$ | 0.01 | 0.01 | 0.15 | 0.13 | 0.93 | −0.01 | 0.00 | 0.08 | 0.08 | 0.95 | 0.00 | 0.00 | 0.06 | 0.06 | 0.94 |
| | $\theta$ | −0.01 | −0.02 | 0.17 | 0.10 | 0.92 | 0.00 | 0.00 | 0.06 | 0.06 | 0.95 | 0.00 | 0.00 | 0.05 | 0.04 | 0.95 |
| $r = 1$ | | | | | | | | | | | | | | | | |
| | $\beta$ | −0.01 | −0.02 | 0.09 | 0.08 | 0.94 | 0.00 | −0.01 | 0.08 | 0.07 | 0.94 | 0.00 | 0.00 | 0.07 | 0.07 | 0.94 |
| | $\gamma$ | −0.01 | −0.01 | 0.14 | 0.14 | 0.95 | 0.00 | −0.01 | 0.09 | 0.09 | 0.95 | 0.00 | 0.00 | 0.07 | 0.07 | 0.94 |
| | $\theta$ | 0.00 | 0.00 | 0.15 | 0.10 | 0.90 | 0.00 | −0.01 | 0.09 | 0.07 | 0.94 | −0.02 | −0.02 | 0.08 | 0.08 | 0.97 |
| $r = 2$ | | | | | | | | | | | | | | | | |
| | $\beta$ | −0.01 | 0.00 | 0.11 | 0.11 | 0.95 | 0.00 | −0.01 | 0.11 | 0.10 | 0.94 | 0.00 | 0.00 | 0.10 | 0.09 | 0.94 |
| | $\gamma$ | 0.00 | −0.01 | 0.16 | 0.15 | 0.95 | −0.01 | −0.02 | 0.13 | 0.11 | 0.93 | 0.00 | −0.01 | 0.11 | 0.10 | 0.95 |
| | $\theta$ | −0.01 | −0.02 | 0.15 | 0.11 | 0.91 | 0.01 | −0.01 | 0.18 | 0.10 | 0.92 | 0.02 | 0.02 | 0.16 | 0.13 | 0.91 |

Table 2: Results of the simulation study for $\theta = 3.5$, $\beta = 2$, $\gamma = -2$, and $r = 2$ with 5 interior knots. Here RB, $\widetilde{\text{RB}}$, SD, SE, and CP denote the relative mean bias, the relative median bias, the standard deviation, the median of estimated standard error, and the 95% coverage probability, respectively. PAR: Parameter.

| PAR | $n = 300$ | | | | | $n = 1000$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | RB | $\widetilde{\text{RB}}$ | SD | SE | CP | RB | $\widetilde{\text{RB}}$ | SD | SE | CP |
| $\beta$ | $-0.01$ | $-0.02$ | 0.22 | 0.23 | 0.96 | $-0.01$ | $-0.02$ | 0.12 | 0.13 | 0.96 |
| $\gamma$ | $-0.05$ | $-0.05$ | 0.43 | 0.43 | 0.94 | $-0.03$ | $-0.03$ | 0.24 | 0.24 | 0.93 |
| $\theta$ | $-0.04$ | $-0.04$ | 0.24 | 0.32 | 0.96 | $-0.03$ | $-0.03$ | 0.13 | 0.18 | 0.96 |

Table 3: Results of the simulation study for $\beta = -1$, $\gamma = -1$, and $r \to 0$. The proposed MM algorithm and regular Newton-Raphson (NR) method are used to estimate the parameters. Here RB, $\widetilde{\text{RB}}$, SD, SE, and CP denote the relative mean bias, the relative median bias, the standard deviation, the median of estimated standard error, and the 95% coverage probability, respectively. PAR: Parameter. The results are obtained based on 100 replications.

| P | $n = 300$ | | | | | | | | | | $n = 1000$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | MM | | | | | NR | | | | | MM | | | | | NR | | | | |
| R | RB | $\widetilde{\text{RB}}$ | SD | SE | CP | RB | $\widetilde{\text{RB}}$ | SD | SE | CP | RB | $\widetilde{\text{RB}}$ | SD | SE | CP | RB | $\widetilde{\text{RB}}$ | SD | SE | CP |
| 2 knots | | | | | | | | | | | | | | | | | | | | |
| $\theta = 0.5$ | | | | | | | | | | | | | | | | | | | | |
| $\beta$ | −0.01 | −0.01 | 0.08 | 0.09 | 0.97 | −0.01 | 0.00 | 0.09 | 0.08 | 0.92 | 0.00 | 0.00 | 0.05 | 0.05 | 0.95 | 0.01 | 0.02 | 0.06 | 0.05 | 0.79 |
| $\gamma$ | −0.01 | −0.02 | 0.11 | 0.10 | 0.95 | 0.00 | 0.01 | 0.11 | 0.10 | 0.92 | −0.01 | −0.01 | 0.05 | 0.06 | 0.97 | 0.00 | 0.01 | 0.06 | 0.05 | 0.86 |
| $\theta$ | −0.01 | −0.01 | 0.08 | 0.08 | 0.97 | 0.01 | 0.00 | 0.09 | 0.08 | 0.92 | −0.01 | −0.01 | 0.04 | 0.04 | 0.99 | 0.01 | 0.02 | 0.05 | 0.04 | 0.93 |
| $\theta = 1$ | | | | | | | | | | | | | | | | | | | | |
| $\beta$ | 0.00 | −0.01 | 0.11 | 0.10 | 0.90 | −0.01 | −0.01 | 0.12 | 0.10 | 0.90 | 0.00 | 0.00 | 0.05 | 0.05 | 0.96 | 0.02 | 0.01 | 0.05 | 0.05 | 0.93 |
| $\gamma$ | 0.00 | 0.01 | 0.14 | 0.14 | 0.92 | 0.01 | 0.03 | 0.14 | 0.14 | 0.94 | −0.01 | 0.00 | 0.08 | 0.08 | 0.97 | 0.01 | 0.01 | 0.07 | 0.08 | 0.98 |
| $\theta$ | −0.01 | 0.00 | 0.13 | 0.10 | 0.92 | 0.00 | 0.00 | 0.13 | 0.10 | 0.91 | −0.01 | −0.01 | 0.06 | 0.05 | 0.91 | 0.01 | 0.00 | 0.07 | 0.06 | 0.84 |
| $\theta = 2$ | | | | | | | | | | | | | | | | | | | | |
| $\beta$ | −0.03 | −0.02 | 0.13 | 0.12 | 0.92 | −0.03 | −0.03 | 0.13 | 0.12 | 0.92 | 0.01 | −0.01 | 0.09 | 0.07 | 0.92 | −0.01 | −0.01 | 0.08 | 0.07 | 0.94 |
| $\gamma$ | −0.01 | −0.02 | 0.24 | 0.25 | 1.00 | 0.00 | −0.01 | 0.25 | 0.25 | 0.99 | 0.01 | −0.01 | 0.16 | 0.14 | 0.93 | −0.03 | −0.03 | 0.14 | 0.13 | 0.93 |
| $\theta$ | 0.00 | −0.01 | 0.24 | 0.19 | 0.91 | −0.01 | −0.01 | 0.20 | 0.19 | 0.94 | −0.01 | −0.01 | 0.16 | 0.11 | 0.90 | −0.02 | −0.03 | 0.20 | 0.10 | 0.89 |
| 5 knots | | | | | | | | | | | | | | | | | | | | |
| $\theta = 0.5$ | | | | | | | | | | | | | | | | | | | | |
| $\beta$ | −0.01 | −0.01 | 0.08 | 0.09 | 0.98 | −0.01 | −0.01 | 0.07 | 0.08 | 0.98 | 0.00 | 0.00 | 0.05 | 0.05 | 0.98 | 0.00 | −0.01 | 0.06 | 0.05 | 0.86 |
| $\gamma$ | −0.01 | −0.01 | 0.10 | 0.10 | 0.95 | 0.00 | −0.01 | 0.12 | 0.10 | 0.93 | 0.00 | −0.01 | 0.05 | 0.06 | 0.98 | 0.00 | 0.00 | 0.07 | 0.05 | 0.88 |
| $\theta$ | −0.02 | −0.01 | 0.08 | 0.08 | 0.97 | 0.03 | 0.03 | 0.08 | 0.08 | 0.96 | 0.00 | 0.01 | 0.04 | 0.04 | 0.99 | 0.01 | 0.01 | 0.04 | 0.04 | 0.95 |
| $\theta = 1$ | | | | | | | | | | | | | | | | | | | | |
| $\beta$ | −0.02 | −0.02 | 0.10 | 0.10 | 0.93 | −0.02 | −0.03 | 0.12 | 0.10 | 0.90 | 0.00 | 0.00 | 0.07 | 0.05 | 0.96 | −0.01 | 0.00 | 0.05 | 0.05 | 0.94 |
| $\gamma$ | 0.00 | 0.01 | 0.14 | 0.14 | 0.94 | 0.01 | 0.02 | 0.15 | 0.14 | 0.90 | −0.01 | −0.02 | 0.07 | 0.08 | 0.97 | 0.00 | 0.00 | 0.08 | 0.08 | 0.94 |
| $\theta$ | 0.00 | −0.02 | 0.11 | 0.10 | 0.95 | −0.01 | −0.02 | 0.12 | 0.10 | 0.92 | 0.01 | −0.01 | 0.06 | 0.05 | 0.92 | −0.02 | −0.03 | 0.06 | 0.05 | 0.88 |
| $\theta = 2$ | | | | | | | | | | | | | | | | | | | | |
| $\beta$ | −0.03 | −0.02 | 0.14 | 0.12 | 0.91 | −0.03 | −0.05 | 0.15 | 0.13 | 0.90 | −0.01 | −0.01 | 0.07 | 0.07 | 0.94 | −0.03 | −0.03 | 0.09 | 0.07 | 0.86 |
| $\gamma$ | 0.00 | 0.00 | 0.26 | 0.25 | 0.98 | 0.04 | 0.01 | 0.27 | 0.25 | 0.97 | −0.01 | −0.01 | 0.12 | 0.14 | 0.95 | 0.00 | −0.01 | 0.16 | 0.13 | 0.91 |
| $\theta$ | −0.01 | −0.02 | 0.25 | 0.19 | 0.91 | −0.02 | −0.05 | 0.34 | 0.19 | 0.91 | 0.00 | −0.01 | 0.17 | 0.11 | 0.90 | −0.05 | −0.05 | 0.22 | 0.10 | 0.77 |

Table 4: Comparison of the average computation time in seconds for the proposed MM algorithm and direct maximization using the regular Newton-Raphson (NR) method. This comparison is for $r \to 0$ (the PH model), $\beta = -1$, $\gamma = -1$, and $\theta = 0.5$.

| | $k = 2$ | | $k = 5$ | |
|---|---|---|---|---|
| $n$ | MM | NR | MM | NR |
| 300 | 19.11 | 191.15 | 29.22 | 846.97 |
| 1000 | 65.57 | 1030.70 | 114.45 | 5262.27 |