Supplementary Materials to "Identification and Efficient Estimation in Regression Analysis with Response Missing Not At Random"

Qinglong Tian[1], Donglin Zeng[2], Jiwei Zhao[3]

[1] *University of Waterloo,* [2] *University of Michigan*

[3] *University of Wisconsin-Madison*

Abstract

Section S1 provides proofs for all the lemmas and theorems in the main paper. Section S2 provides more details on the Gauss-Hermite Quadrature we used in the main paper.

S1 Proofs of Lemmas and Theorems

Lemma 1. Firstly, the conditional density function of (R, Y) given $\mathbf{X}$, from one observation, is

$$\left\{\pi(y,\mathbf{u})f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta})\right\}^r\left\{1-\int\pi(t,\mathbf{u})f_{Y|\mathbf{X}}(t,\mathbf{x};\boldsymbol{\beta})dt\right\}^{1-r}.$$

Thus, we need to show, if

$$\left\{\pi_1(y,\mathbf{u})f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta}_1)\right\}^r\left\{1-\int\pi_1(t,\mathbf{u})f_{Y|\mathbf{X}}(t,\mathbf{x};\boldsymbol{\beta}_1)dt\right\}^{1-r} \quad \text{(S1.1)}$$

$$= \left\{\pi_2(y,\mathbf{u})f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta}_2)\right\}^r\left\{1-\int\pi_2(t,\mathbf{u})f_{Y|\mathbf{X}}(t,\mathbf{x};\boldsymbol{\beta}_2)dt\right\}^{1-r} \quad \text{(S1.2)}$$

for all r, y and $\mathbf{x}$, then $\pi_1(y,\mathbf{u}) = \pi_2(y,\mathbf{u})$ and $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$. When (S1.1)=(S1.2), let $r = 1$, it implies $\pi_1(y,\mathbf{u})f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta}_1) = \pi_2(y,\mathbf{u})f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta}_2)$; let $r = 0$, it implies

$$\int\pi_1(t,\mathbf{u})f_{Y|\mathbf{X}}(t,\mathbf{x};\boldsymbol{\beta}_1)dt = \int\pi_2(t,\mathbf{u})f_{Y|\mathbf{X}}(t,\mathbf{x};\boldsymbol{\beta}_2)dt,$$

which is implied by $\pi_1(y,\mathbf{u})f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta}_1) = \pi_2(y,\mathbf{u})f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta}_2)$. Thus we have

$$\log f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta}_1) - \log f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta}_2) = \log\pi_2(y,\mathbf{u}) - \log\pi_1(y,\mathbf{u}).$$

Note that the above expression is a function not depending on z any more. We denote it as $m(y,\mathbf{u})$. Since random variable Z is one-dimensional, we simply take the bounded set $\mathcal{I}$ in condition (A2) as the closed interval

$[z_l, z_u]$. Similar arguments naturally follow for other forms of $\mathcal{I}$. As $t \to 0$, we define a sequence of continuously differentiable (with respect to z) functions $q_t(y, \mathbf{u}, z)$ as

$$\begin{aligned} q_t(y, \mathbf{u}, z) &= I(z_l - t < z \le z_l) K_l\left(\frac{z - z_l}{t}\right) m(y, \mathbf{u}) z_l \\ &\quad + I(z_l < z \le z_u) m(y, \mathbf{u}) z \\ &\quad + I(z_u < z \le z_u + t) K_r\left(\frac{z - z_u}{t}\right) m(y, \mathbf{u}) z_u, \end{aligned}$$

where

$$K_l(x) = \begin{cases} 0, & x \le -1, \\ 2(1+x)^2, & -1 < x \le -\frac{1}{2}, \\ 1 - 2x^2, & -\frac{1}{2} < x \le 0, \\ 1, & x > 0, \end{cases} \quad \text{and } K_r(x) = \begin{cases} 1, & x \le 0, \\ 1 - 2x^2, & 0 < x \le \frac{1}{2}, \\ 2(1-x)^2, & \frac{1}{2} < x \le 1, \\ 0, & x > 1. \end{cases}$$

One can verify that, when $z \notin \mathcal{I}$, for every fixed t, q_t is defined such that $q_t = 0$ uniformly in $(y, \mathbf{u})$ when $z \to \infty$ or $z \to -\infty$. Also when $t \to 0$, q_t is defined such that $\int_{z \notin \mathcal{I}} \partial q_t / \partial z dz \to 0$. Then we have, for every t,

$$\begin{aligned} &\int \frac{\partial q_t}{\partial z}(y, \mathbf{u}, z) \{\log f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta}_1) - \log f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta}_2)\} dz \\ = \; &\{\log \pi_2(y, \mathbf{u}) - \log \pi_1(y, \mathbf{u})\} q_t(y, \mathbf{u}, z) \mid_{z=-\infty}^{z=\infty} = 0. \end{aligned}$$

Additionally, by the definition of q_t, one can easily verify

$$\begin{aligned} &\int_{z \notin \mathcal{I}} \frac{\partial q_t}{\partial z}(y, \mathbf{u}, z) \{\log f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta}_1) - \log f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta}_2)\} dz \\ = \; &\{\log \pi_2(y, \mathbf{u}) - \log \pi_1(y, \mathbf{u})\} \int_{z \notin \mathcal{I}} \frac{\partial q_t}{\partial z}(y, \mathbf{u}, z) dz \to 0, t \to 0. \end{aligned}$$

Therefore, we also have

$$\begin{aligned} & \int_{z\in\mathcal{I}} \frac{\partial q_t}{\partial z}(y,\mathbf{u},z)\{\log f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta}_1)-\log f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta}_2)\}dz \\ = & \int_{z\in\mathcal{I}} \{\log f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta}_1)-\log f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta}_2)\}^2 dz \\ = & \{\log f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta}_1)-\log f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta}_2)\}^2(z_u-z_l)\to 0, t\to 0. \end{aligned}$$

Hence, we must have $\log f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta}_1)-\log f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta}_2)=0$ when $z\in\mathcal{I}$. Following the condition (A3), $\boldsymbol{\beta}_1=\boldsymbol{\beta}_2$. Then, we must also have $\pi_1(y,\mathbf{u})=\pi_2(y,\mathbf{u})$. This completes the model identifiability. □

Theorem 1. This proof consists of three major steps.

In the first step, we construct some functions in $\mathcal{S}(m,K_n,M_n)$ to approximate the true parameters. To do that, we need the following general result. From the properties of B-spline functions (Schumaker, 2007), we can define a linear operator $\mathcal{Q}$ mapping $W^{k,\infty}(\mathcal{D})$ to the sieve space; that is, for any $g\in W^{k,\infty}(\mathcal{D})$,

$$\mathcal{Q}[g]=\sum_{l_1,\ldots,l_{p+1}=-m+1}^{K_n}\Gamma_{l_1,\ldots,l_{p+1}}[g]B_{l_1}^m(y)\cdots B_{l_{p+1}}^m(u_p),$$

where $\Gamma_{l_1,\ldots,l_{p+1}}$ are the linear functionals in $L_\infty(\mathcal{D})$, where $L_\infty(\cdot)$ is the set of all measurable functions that are bounded almost everywhere. Moreover,

$$\sum_{l_1,\ldots,l_{p+1}=-m+1}^{K_n}|\Gamma_{l_1,\ldots,l_{p+1}}[g]|\le(2m+1)^{p+1}9^{(p+1)(m-1)}\|g\|_{L_\infty(\mathcal{D})},$$

and according to Schumaker (2007),

$$\|\mathcal{Q}[g] - g\|_{L_\infty(\mathcal{D})} \leq O(m)K_n^{-k}.$$

Thus we define $\eta_n(y, \mathbf{u}) = \mathcal{Q}[\text{logit}\pi_0(y, \mathbf{u})]$. Corresponding, we obtain

$$\pi_n(y, \mathbf{u}) = \text{expit}(\eta_n(y, \mathbf{u})).$$

As a result, we have the bound

$$\|\pi_n(y, \mathbf{u}) - \pi_0(y, \mathbf{u})\|_{L_\infty(\mathcal{D})} \leq O(K_n^{-k}).$$

For the second step, we first define $\mathbf{P}_n$ as the empirical measure based on the n i.i.d. observations, $\mathbf{P}$ as the corresponding expectation, and $\mathbf{G}_n$ as the empirical process $\sqrt{n}(\mathbf{P}_n - \mathbf{P})$. Recall that, in Section 2.2 we denoted the likelihood function from one generic observation as

$$p(r, y, \mathbf{x}; \boldsymbol{\beta}, \pi) \equiv \left\{\pi(y, \mathbf{u}) f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta})\right\}^r \left\{1 - \int \pi(t, \mathbf{u}) f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta}) dt\right\}^{1-r}.$$

Then we have

$$\mathbf{P}_n\{\text{log}p(R, Y, \mathbf{X}; \widehat{\boldsymbol{\beta}}_n, \widehat{\pi}_n)\} \geq \mathbf{P}_n\{\text{log}p(R, Y, \mathbf{X}; \boldsymbol{\beta}_0, \pi_n)\}.$$

Equivalently, we have

$$n^{-1/2}\mathbf{G}_n \left\{\text{log}\frac{p(R, Y, \mathbf{X}; \widehat{\boldsymbol{\beta}}_n, \widehat{\pi}_n)}{p(R, Y, \mathbf{X}; \boldsymbol{\beta}_0, \pi_n)}\right\} \geq \mathbf{P}\left\{\text{log}\frac{p(R, Y, \mathbf{X}; \boldsymbol{\beta}_0, \pi_n)}{p(R, Y, \mathbf{X}; \boldsymbol{\beta}_0, \pi_0)}\right\} + \mathbf{P}\left\{\text{log}\frac{p(R, Y, \mathbf{X}; \boldsymbol{\beta}_0, \pi_0)}{p(R, Y, \mathbf{X}; \widehat{\boldsymbol{\beta}}_n, \widehat{\pi}_n)}\right\} \quad \text{(S1.3)}$$

In the next we will first bound the left hand side of (S1.3) using empirical process theory. For this purpose, we consider a collection of functions $\mathcal{L}_n$

defined as

$$\mathcal{L}_n = \left\{ \log \frac{p(R, Y, \mathbf{X}; \widetilde{\boldsymbol{\beta}}_n, \widetilde{\pi}_n)}{p(R, Y, \mathbf{X}; \boldsymbol{\beta}_0, \pi_n)} : (\widetilde{\boldsymbol{\beta}}_n, \widetilde{\pi}_n) \in \mathcal{S}(m, K_n, M_n) \right\}.$$

Since $\|B_i^m(\cdot)\|_{L_\infty([0,1])} = 1$, any function of $\widetilde{\pi}_n$ given in $\mathcal{L}_n$ is bounded by 1. By the assumption, $p(R, Y, \mathbf{X}; \boldsymbol{\beta}_0, \pi_n)$ is bounded below by zero. Hence the class $\mathcal{L}_n$ has an upper bound $O_p(M_n)$. Also it can be verified that all the functions in $\mathcal{L}_n$ are Lipschitz continuous and the Lipschitz constant is bounded by $O_p(e^{c_0 M_n})$, and they lie in a hypercube of a real space $\mathcal{R}^{N_n}$ with $N_n = (K_n + m)^{p+1} + d$. Therefore, for any $\epsilon > 0$, if we partition this hypercube into subcubes with scale length ϵ, the total number of subcubes is at most $O((M_n/\epsilon)^{N_n})$. According to the Lipschitz property of the functions in $\mathcal{L}_n$, the L_∞-distance between any two functions of $\mathcal{L}_n$ with respective indexes in the same subcube is no more than $O_p(e^{c_0 M_n}) N_n \epsilon$. Consequently, we obtain that the bracketing number for $\mathcal{L}_n$ satisfies

$$N_{[\cdot]}(O_p(e^{c_0 M_n}) N_n \epsilon, \mathcal{L}_n, L_\infty) \leq O((M_n/\epsilon)^{N_n}).$$

According to van der Vaart (1998), we have, in probability,

$$\begin{aligned} \sqrt{n} E_P^* \|\mathbf{P}_n - \mathbf{P}\|_{\mathcal{L}_n} &\leq O_p(1) \int_0^{O(M_n)} \sqrt{\log \left(\frac{M_n e^{c_0 M_n} N_n}{\epsilon} \right)^{N_n}} d\epsilon \\ &\leq O_p(1) M_n^2 K_n^{\frac{p+1}{2}} \log K_n. \end{aligned}$$

Thus the left hand side of (S1.3) is bounded above by $O_p(M_n^2 K_n^{\frac{p+1}{2}} \log K_n / \sqrt{n})$.

We now bound below the right hand side of (S1.3). Firstly, since the function $p(\cdot)$ is Lipschitz continuous with each argument, we have the first component

$$\mathbf{P}\left\{\log\frac{p(R,Y,\mathbf{X};\boldsymbol{\beta}_0,\pi_n)}{p(R,Y,\mathbf{X};\boldsymbol{\beta}_0,\pi_0)}\right\} \geq -O_p(1)\|\pi_n-\pi_0\|_{L_\infty} \geq -O_p(1)K_n^{-k}.$$

The second component is the Kullback-Leibler divergence. We linearize the last term. The first order term in the expansion vanishes while the second order term in the expansion is bounded below by

$$O(e^{-c_1M_n})\|p(R,Y,\mathbf{X};\boldsymbol{\beta}_0,\pi_0)-p(R,Y,\mathbf{X};\widehat{\boldsymbol{\beta}}_n,\widehat{\pi}_n)\|_{L_2}^2.$$

Combining all the above results, we have

$$O_p(1)\left(\frac{e^{c_1M_n}}{K_n^k}+\frac{e^{c_1M_n}M_n^2K_n^{\frac{p+1}{2}}\log K_n}{\sqrt{n}}\right) \geq \|p(R,Y,\mathbf{X};\boldsymbol{\beta}_0,\pi_0)-p(R,Y,\mathbf{X};\widehat{\boldsymbol{\beta}}_n,\widehat{\pi}_n)\|_{L_2}^2 \quad \text{(S1.4)}$$

Using some Taylor expansion arguments, although the constant c_1 changes slightly, we can still obtain the same bound for $\|\log p(R,Y,\mathbf{X};\boldsymbol{\beta}_0,\pi_0)-\log p(R,Y,\mathbf{X};\widehat{\boldsymbol{\beta}}_n,\widehat{\pi}_n)\|_{L_2}^2$, i.e.,

$$O_p(1)\left(\frac{e^{c_1M_n}}{K_n^k}+\frac{e^{c_1M_n}M_n^2K_n^{\frac{p+1}{2}}\log K_n}{\sqrt{n}}\right) \geq \int\{\log p(r,y,\mathbf{x};\boldsymbol{\beta}_0,\pi_0)-\log p(r,y,\mathbf{x};\widehat{\boldsymbol{\beta}}_n,\widehat{\pi}_n)\}^2drdyd\mathbf{x} \quad \text{(S1.5)}$$

In the third step, we aim to obtain the L_2-convergence of the estimators, hence the consistency. In (S1.5), if we let $r=1$, then

$$\begin{aligned}&\int\{\log\widehat{\pi}_n(y,\mathbf{u})+\log f_{Y|\mathbf{X}}(y,\mathbf{x};\widehat{\boldsymbol{\beta}}_n)-\log\pi_0(y,\mathbf{u})-\log f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta}_0)\}^2dyd\mathbf{x}\\ \leq\ & O_p(1)\left(\frac{e^{c_1M_n}}{K_n^k}+\frac{e^{c_1M_n}M_n^2K_n^{\frac{p+1}{2}}\log K_n}{\sqrt{n}}\right).\end{aligned}$$

To proceed, we temporarily denote the right hand side bound in the above expression as A_n. Similar to the proof of Lemma 1, we take the compact set $\mathcal{I}$ in condition (A3) as the closed interval $[z_l, z_u]$. Note that, for each fixed n, $\log f_{Y|\mathbf{X}}(y, \mathbf{x}; \widehat{\boldsymbol{\beta}}_n) - \log f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta}_0)$ is a function of $y, \mathbf{u}, z$ and we temporarily denote it as $l(y, \mathbf{u}, z)$. We denote one of its original functions (with respect to z) as $L(y, \mathbf{u}, z)$ such that $\partial L(y, \mathbf{u}, z)/\partial z = l(y, \mathbf{u}, z)$. As $t \to 0$, we define a sequence of continuously differentiable (with respect to z) functions $w_t(y, \mathbf{u}, z)$ as

$$\begin{aligned} w_t(y, \mathbf{u}, z) &= I(z_l - t < z \le z_l) K_l\left(\frac{z - z_l}{t}\right) l(y, \mathbf{u}, z_l) \\ &\quad + I(z_l < z \le z_u) L(y, \mathbf{u}, z) \\ &\quad + I(z_u < z \le z_u + t) K_r\left(\frac{z - z_u}{t}\right) l(y, \mathbf{u}, z_u), \end{aligned}$$

where K_l and K_r were defined in the proof of Lemma 1.

One can easily verify, when $z \in \mathcal{I}$, $\partial w_t/\partial z = l(y, \mathbf{u}, z) = \log f_{Y|\mathbf{X}}(y, \mathbf{x}; \widehat{\boldsymbol{\beta}}_n) - \log f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta}_0)$ for any t. When $z \notin \mathcal{I}$, for every fixed t, w_t is defined such that $w_t = 0$ uniformly in $(y, \mathbf{u})$ when $z \to \infty$ or $z \to -\infty$; also when $t \to 0$, w_t is defined such that $\int_{z \notin \mathcal{I}} \partial w_t/\partial z \; l(y, \mathbf{u}, z) dz \to 0$ for any bounded function $l(y, \mathbf{u}, z)$.

Firstly, we have

$$\int \frac{\partial w_t}{\partial z} \{\log \widehat{\pi}_n(y, \mathbf{u}) + \log f_{Y|\mathbf{X}}(y, \mathbf{x}; \widehat{\boldsymbol{\beta}}_n) - \log \pi_0(y, \mathbf{u}) - \log f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta}_0)\} dy d\mathbf{u} dz \le A_n^{1/2} \left\| \frac{\partial w_t}{\partial z} \right\|_{L_2}.$$

Because

$$\int \frac{\partial w_t}{\partial z}\{\log\widehat{\pi}_n(y,\mathbf{u}) - \log\pi_0(y,\mathbf{u})\}dyd\mathbf{u}dz = \int \{\log\widehat{\pi}_n(y,\mathbf{u}) - \log\pi_0(y,\mathbf{u})\}w_t \mid_{z=-\infty}^{z=\infty} dyd\mathbf{u} = o_p(1),$$

we have

$$\int \frac{\partial w_t}{\partial z}\{\log f_{Y|\mathbf{X}}(y,\mathbf{x};\widehat{\boldsymbol{\beta}}_n) - \log f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta}_0)\}dyd\mathbf{u}dz \leq A_n^{1/2}\left\|\frac{\partial w_t}{\partial z}\right\|_{L_2}.$$

By the definition of w_t, we then have

$$\int_{\mathcal{I}} \frac{\partial w_t}{\partial z}\{\log f_{Y|\mathbf{X}}(y,\mathbf{x};\widehat{\boldsymbol{\beta}}_n) - \log f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta}_0)\}dyd\mathbf{u}dz \leq A_n^{1/2}\left\|\frac{\partial w_t}{\partial z}\right\|_{L_2}.$$

Let $t \to 0$, this is equivalent to

$$\int_{\mathcal{I}} \{\log f_{Y|\mathbf{X}}(y,\mathbf{x};\widehat{\boldsymbol{\beta}}_n) - \log f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta}_0)\}^2 dyd\mathbf{u}dz \leq A_n = O_p(1)\left(\frac{e^{c_1 M_n}}{K_n^k} + \frac{e^{c_1 M_n} M_n^2 K_n^{\frac{p+1}{2}}\log K_n}{\sqrt{n}}\right).$$

Therefore, if we choose a subsequence $f_{Y|\mathbf{X}}(y,\mathbf{x};\widehat{\boldsymbol{\beta}}_n)$ and we suppose $\widehat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}^*$, from the above inequality as well as the condition (A3), we must have $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$. Hence, we obtain

$$\begin{aligned} &\int_{\mathcal{I}} \{\log f_{Y|\mathbf{X}}(y,\mathbf{x};\widehat{\boldsymbol{\beta}}_n) - \log f_{Y|\mathbf{X}}(y,\mathbf{x};\boldsymbol{\beta}_0)\}^2 dyd\mathbf{u}dz \\ \leq\ & O_p(1)\left(\frac{e^{c_1 M_n}}{K_n^k} + \frac{e^{c_1 M_n} M_n^2 K_n^{\frac{p+1}{2}}\log K_n}{\sqrt{n}}\right), \end{aligned}$$

as well as

$$\begin{aligned} &\int_{\mathcal{I}} \{\log\widehat{\pi}_n(y,\mathbf{u}) - \log\pi_0(y,\mathbf{u})\}^2 dyd\mathbf{u}dz \\ \leq\ & O_p(1)\left(\frac{e^{c_1 M_n}}{K_n^k} + \frac{e^{c_1 M_n} M_n^2 K_n^{\frac{p+1}{2}}\log K_n}{\sqrt{n}}\right). \end{aligned}$$

Then, using the Taylor's expansion and provided that condition (A3) is satisfied, we can achieve the same bound for $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2$ and $\|\widehat{\pi}_n - \pi_0\|_{L_2}^2$.

On the other hand, from Schumaker (2007) and the condition (A4), we have

$$\|\nabla_y^{k_y} \nabla_{\mathbf{u}}^{k_u} \widehat{\eta}_n(y, \mathbf{u})\|_{L_\infty} \leq C K_n^k \sum_{l_1, \ldots, l_{p+1}} |\tau_{l_1, \ldots, l_{p+1}}| \leq O(M_n K_n^k),$$

thus

$$\|\nabla_y^{k_y} \nabla_{\mathbf{u}}^{k_u} \widehat{\pi}_n(y, \mathbf{u})\|_{L_\infty} \leq C e^{(k+1)M_n} M_n K_n^k \leq C e^{(k+2)M_n} K_n^k,$$

where $k_y + k_u = k$. According to the Sobolev interpolation inequality (Adams and Fournier, 2003), we have

$$\|\nabla(\widehat{\pi}_n - \pi_0)\|_{L_\infty} \leq C e^{(k+2)M_n \lambda} K_n^{k\lambda} \left(\frac{e^{c_1 M_n}}{K_n^k} + \frac{e^{c_1 M_n} M_n^2 K_n^{\frac{p+1}{2}} \log K_n}{\sqrt{n}} \right)^{(1-\lambda)/2},$$

where $\lambda = (p+3)/(2k) < 1/3$. The choice of K_n and M_n will guarantee this terms goes to zero. Hence we proved that

$$\|\widehat{\pi}_n(y, \mathbf{u}) - \pi_0(y, \mathbf{u})\|_{W^{1,\infty}} \xrightarrow{p} 0.$$

This completes the proof of the consistency theorem. □

Theorem 2. We will show that

$$\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2 + \|\widehat{\pi}_n - \pi_0\|_{L_2}^2 \leq O_p(K_n^{-2k}) + o_p(n^{-1/2}).$$

The proof is similar to that of Theorem 1. We define $\mathcal{L}_n$ as before, but the functions in $\mathcal{L}_n$ are indexed by $(\boldsymbol{\beta}, \pi)$, which belongs to a bounded set in

$R \times W^{1,\infty}$. Thus, $\mathcal{L}_n$ has a bounded covering function and the integration of the entropy for the class $\mathcal{L}_n$ is finite. Moreover, the function in the left hand side of (S1.3) converges to zero uniformly. Thus we apply Theorem 2.11.23 of van der Vaart and Wellner (1996), to obtain that the left hand side of (S1.3) is bounded by $o_p(n^{-1/2})$.

For the right hand side of (S1.3), we still perform Taylor expansion at the true parameters. Since each parameter is in a small neighborhood of the true parameters, the right hand side of (S1.3) is bounded below by

$$-O_p(1)\|\widehat{\pi}_n - \pi_0\|_{L_2}^2 + O_p(1)\|p(R, Y, \mathbf{X}; \boldsymbol{\beta}_0, \pi_0) - p(R, Y, \mathbf{X}; \widehat{\boldsymbol{\beta}}_n, \widehat{\pi}_n)\|_{L_2}^2.$$

Therefore we have

$$o_p(1)n^{-1/2} + O_p(1)K_n^{-2k} \geq \|p(R, Y, \mathbf{X}; \boldsymbol{\beta}_0, \pi_0) - p(R, Y, \mathbf{X}; \widehat{\boldsymbol{\beta}}_n, \widehat{\pi}_n)\|_{L_2}^2. \quad \text{(S1.6)}$$

Note that (S1.6) is the same as (S1.4) in the proof of Theorem 1, except for the bound on the left hand side. In what follows we use the same argument as in the proof of Theorem 1 and then the proof is complete. □

Theorem 3. We will show that $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ can be written as a linear functional of the empirical process $\mathbf{G}_n$, and that $\widehat{\boldsymbol{\beta}}_n$ is semiparametrically efficient.

We have the score function for parameter $\boldsymbol{\beta}$ is

$$l_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \pi) = r\mathbf{S}_{\boldsymbol{\beta}}(y, \mathbf{x}; \boldsymbol{\beta}) - (1-r)\frac{\int \mathbf{S}_{\boldsymbol{\beta}}(t, \mathbf{x}; \boldsymbol{\beta}) f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta})\pi(t, \mathbf{u})dt}{1 - \int f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta})\pi(t, \mathbf{u})dt},$$

where $\mathbf{S}_{\boldsymbol{\beta}}(y, \mathbf{x}; \boldsymbol{\beta}) = \partial \log f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta})/\partial \boldsymbol{\beta}$ is the score vector, and the score function for nuisance $\pi(y, \mathbf{u})$ along the submodel $\pi_t(y, \mathbf{u}) = \{1+th(y, \mathbf{u})\}\pi_0(y, \mathbf{u})$ is

$$l_\pi(\boldsymbol{\beta}, \pi) \quad = \quad rh(y, \mathbf{u}) - (1-r)\frac{\int h(t, \mathbf{u}) f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta}_0)\pi_0(t, \mathbf{u})dt}{1 - \int f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta}_0)\pi_0(t, \mathbf{u})dt}.$$

After some algebra, we have

$$l_\pi^* l_{\boldsymbol{\beta}} = \mathbf{S}_{\boldsymbol{\beta}}(y, \mathbf{x}; \boldsymbol{\beta}) f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta})\pi(y, \mathbf{u}) + \frac{\int \mathbf{S}_{\boldsymbol{\beta}}(t, \mathbf{x}; \boldsymbol{\beta}) f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta})\pi(t, \mathbf{u})dt}{1 - \int f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta})\pi(t, \mathbf{u})dt} f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta})\pi(y, \mathbf{u}),$$

and

$$l_\pi^* l_\pi a(y, \mathbf{u}) = a(y, \mathbf{u}) f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta})\pi(y, \mathbf{u}) + \frac{\int a(t, \mathbf{u}) f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta})\pi(t, \mathbf{u})dt}{1 - \int f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta})\pi(t, \mathbf{u})dt} f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta})\pi(y, \mathbf{u}),$$

where l_π^* is the adjoint operator of l_π. The information operator $l_\pi^* l_\pi$ is the sum of an invertible operator and a compact operator from the space $\mathbb{BV}$ to itself, where $\mathbb{BV}$ contains all the functions of y (for each fixed $\mathbf{x}$) with bounded variation. By the Fredholm theory (Rudin, 1973), the information operator is invertible if it is one to one, or equivalently, the Fisher information along any nontrivial submodel is zero.

Suppose that the Fisher information is zero along some submodel $(\boldsymbol{\beta}_0 + t\mathbf{w}, \{1 + th(y, \mathbf{u})\}\pi_0(y, \mathbf{u}))$. Then the score function along this submodel, $l_{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{w} + l_\pi(h(y, \mathbf{u}))$, is zero. Set $R = 1$, we have $\mathbf{S}_{\boldsymbol{\beta}}(y, \mathbf{x}; \boldsymbol{\beta})^{\mathrm{T}}\mathbf{w} + h(y, \mathbf{u}) = 0$, for any $(y, \mathbf{u}, z) \in \mathcal{P}$. Therefore, for any $(y, \mathbf{u}, z_i)$, $i = 1, 2$, we have

$$[\partial \log\{f_{Y|\mathbf{X}}(y, \mathbf{u}, \mathbf{z}_1; \boldsymbol{\beta}_0)/f_{Y|\mathbf{X}}(y, \mathbf{u}, \mathbf{z}_2; \boldsymbol{\beta}_0)\}/\partial \boldsymbol{\beta}]^{\mathrm{T}}\mathbf{w} = 0.$$

By the local identifiability condition (A5), $\mathbf{w} = 0$, hence $h = 0$ with probability one. Thus, the information operator is invertible. Consequently, there exists a function h such that

$$l_\pi^* l_\pi(h) = l_\pi^* l_{\boldsymbol{\beta}}.$$

This means that the least favorable direction for $\boldsymbol{\beta}_0$ exists. In addition, by using the arguments in the proof of Theorem 3.4 of Zeng (2005) and condition (A3), we can show that h belongs to the $W^{k,\infty}$ space.

We now construct the projection of $\pi(y, \mathbf{u})$ on the tangent space of the sieve space. First, by simple computation, the tangent vector $h_n(y, \mathbf{u})$ for the nuisance parameter at $\widehat{\pi}_n$ has the form $\widehat{\pi}_n \eta(y, \mathbf{u})$, where $\eta(y, \mathbf{u})$ has the same form as $\eta(y, \mathbf{u})$ in the sieve space. Then, one good approximation to the pseudo least favorable direction is to choose $h_n(y, \mathbf{u})$ such that its corresponding $\eta(y, \mathbf{u})$ satisfies $\eta(y, \mathbf{u}) = \mathcal{Q}[h_n/\pi_0]$. Thus the previous results imply that

$$\|h_n(y, \mathbf{u}) - h(y, \mathbf{u})\|_{L_2}^2 \leq O_p(K_n^{-2k}) + o_p(n^{-1/2}).$$

Since $(\widehat{\boldsymbol{\beta}}_n, \widehat{\pi}_n)$ maximizes the objective function in the sieve space, the score with respect to t along the submodel $(\widehat{\boldsymbol{\beta}}_n + t\mathbf{w}, (1 + th_n)\widehat{\pi}_n)$ must be zero, where h_n is the projection of h onto the tangent space of the sieve

space. Then it holds that

$$\mathbf{G}_n\{l_{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}_n, \widehat{\pi}_n) - l_\pi(\widehat{\boldsymbol{\beta}}_n, \widehat{\pi}_n)[h_n]\} = -\sqrt{n}\mathbf{P}\{l_{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}_n, \widehat{\pi}_n) - l_\pi(\widehat{\boldsymbol{\beta}}_n, \widehat{\pi}_n)[h_n]\}.$$

For the left hand side of the above equation, we apply Theorem 2.11.23 of van der Vaart and Wellner (1996). Note that the function in the left hand side, indexed by both $(\widehat{\pi}_n, h_n) \in W^{1,\infty}$ and $\widehat{\boldsymbol{\beta}}_n \in [-M, M]^d$, belongs to a P-Donsker class.

Moreover, we linearize the right hand side at the true parameters and approximate h_n by h. Since

$$\mathbf{P}\{l_{\pi\pi}(\boldsymbol{\beta}_0, \pi_0)[\widehat{\pi}_n - \pi_0, h] + l_{\boldsymbol{\beta}\pi}(\boldsymbol{\beta}_0, \pi_0)[\widehat{\pi}_n - \pi_0]\} = 0,$$

we obtain that

$$\begin{aligned}
& -\mathbf{P}\{l_{\pi\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \pi_0)[h] + l_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \pi_0)\}\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \\
= \; & \mathbf{G}_n\{l_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \pi_0) - l_\pi(\boldsymbol{\beta}_0, \pi_0)[h]\} + \sqrt{n}O_p(1)\{\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2 + \|\widehat{\pi}_n - \pi_0\|_{L_2}^2 + \|h_n - h\|_{L_2}^2\} \\
= \; & \mathbf{G}_n\{l_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \pi_0) - l_\pi(\boldsymbol{\beta}_0, \pi_0)[h]\} + o_p(1),
\end{aligned}$$

where $l_{\boldsymbol{\beta}\boldsymbol{\beta}}$ is the derivative of $l_{\boldsymbol{\beta}}$ with respect to $\boldsymbol{\beta}$, $l_{\boldsymbol{\beta}\pi}[h]$ is the derivative of $l_{\boldsymbol{\beta}}$ with respect to π along the direction h, $l_{\pi\boldsymbol{\beta}}[h]$ is the derivative of $l_\pi[h]$ with respect to $\boldsymbol{\beta}$, and $l_{\pi\pi}[h_1, h_2]$ is the derivative of $l_\pi[h_1]$ with respect to π along the direction h_2. Next we can show that

$$-\mathbf{P}\{l_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \pi_0) - l_{\pi\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \pi_0)[h]\} = \mathbf{P}\left[\{l_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \pi_0) - l_\pi(\boldsymbol{\beta}_0, \pi_0)[h]\}^{\otimes 2}\right]$$

is invertible. Therefore, the influence function of $\widehat{\boldsymbol{\beta}}_n$ is

$$\boldsymbol{\phi}(\boldsymbol{\beta}_0, \pi_0) \equiv -\left[\mathbf{P}\{l_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \pi_0) - l_{\pi\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \pi_0)[h]\}\right]^{-1}\{l_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \pi_0) - l_{\pi}(\boldsymbol{\beta}_0, \pi_0)[h]\},$$

which belongs to the space spanned by the scores. Thus, the asymptotic variance of $\widehat{\boldsymbol{\beta}}_n$ is $\boldsymbol{\Sigma} \equiv \mathbf{P}\left\{\boldsymbol{\phi}(\boldsymbol{\beta}_0, \pi_0)\boldsymbol{\phi}(\boldsymbol{\beta}_0, \pi_0)^{\mathrm{T}}\right\}$, and it is semiparametrically efficient. This completes the proof. □

S2 Gauss-Hermite Quadrature

The Gauss-Hermite quadrature (e.g., Liu and Pierce 1994) can be used to calculate generic integrals in the form of

$$\int_{-\infty}^{\infty} h(y)\phi(y; \mu, \sigma^2)dy,$$

where $\phi(\cdot; \mu, \sigma^2)$ is a normal pdf with mean μ and variance σ^2. The integral is approximated by

$$\int_{-\infty}^{\infty} h(y)\phi(y; \mu, \sigma^2)dy = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} \exp(-x^2)h(\sqrt{2}\sigma x+\mu)dx \approx \frac{1}{\sqrt{\pi}}\sum_{i=1}^{G} \omega_i h(\sqrt{2}\sigma x_i+\mu),$$

where $\{x_i, i = 1, \ldots, G\}$ and the associated weights $\{\omega_i, i = 1, \ldots, G\}$ are known constants given G. And a larger G will lead to a more precise approximation (but computationally heavier). In Equation (4.10), the integral is an expectation with respect to $f_{Y|\boldsymbol{X}}(y, \boldsymbol{X}_i; \widehat{\boldsymbol{\beta}}^{(t)})$, which is a normal pdf

with mean (suppose $\boldsymbol{X}$ and $\boldsymbol{\beta}$ contain the intercept)

$$\widehat{\mu}^{(t)} = \boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}^{(t)}$$

and variance $(\widehat{\sigma}^{(t)})^2$. The expectant is

$$\log[f_{Y|\boldsymbol{X}}(y, \boldsymbol{X}_i; \boldsymbol{\beta})] h_i(y).$$

Thus, we can see that (4.11) is the result of directly applying the aforementioned Gauss-Hermite quadrature.

Bibliography

Adams, R. A. and Fournier, J. J. (2003). *Sobolev Spaces*, volume 140. Elsevier.

Liu, Q. and Pierce, D. A. (1994). A note on gauss-hermite quadrature. *Biometrika*, 81(3):624–629.

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

Schumaker, L. (2007). *Spline Functions: Basic Theory*. Cambridge University Press.

van der Vaart, A. W. (1998). *Asymptotic Statistics*, volume 3. Cambridge University Press.

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

Zeng, D. (2005). Likelihood approach for marginal proportional hazards regression in the presence of dependent censoring. *The Annals of Statistics*, 33(2):501–521.