Biometrics 000, 1–10 DOI: 000
October 2020

Supplement to 'Penalized Deep Partially Linear Cox Models with Application to CT Scans of Lung Cancer Patients'

Yuming Sun[1], Jian Kang[2,*], Chinmay Haridas[3], Nicholas Mayne[4], Alexandra Potter[3], Chi-Fu Yang[3], David C. Christiani[5] and Yi Li [2,]**

[1]Department of Mathematics, William & Mary, Williamsburg, VA 23185

[2]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109

[3]Division of Thoracic Surgery, Department of Surgery, Massachusetts General Hospital, Boston, MA 02114

[4]Department of Medicine, Duke University, Durham, NC 27710

[5]Department of Environmental Health and Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115

**email:* jiankang@umich.edu

***email:* yili@umich.edu

This paper has been submitted for consideration for publication in *Biometrics*

Composite Hölder Class of Smooth Functions

With constants $a, M > 0$ and a positive integer d, we define a Hölder class of smooth functions as

$$\mathcal{H}_d^a(\mathbb{D}, M) = \{f : \mathbb{D} \subset \mathbb{R}^d \to \mathbb{R} : \sum_{\upsilon:|\upsilon|<a} \|\partial^\upsilon f\|_\infty + \sum_{\upsilon:|\upsilon|=\lfloor a \rfloor} \sup_{x,y\in\mathbb{D},x\neq y} \frac{|\partial^\upsilon f(x) - \partial^\upsilon f(y)|}{\|x-y\|_\infty^{a-\lfloor a \rfloor}} \leqslant M\},$$

where $\mathbb{D}$ is a bounded subset of $\mathbb{R}^d$, $\lfloor a \rfloor$ is the largest integer smaller than a, $\partial^\upsilon := \partial^{\upsilon_1} \ldots \partial^{\upsilon_r}$ with $\upsilon = (\upsilon_1, \ldots, \upsilon_d) \in \mathbb{N}^d$, and $|\upsilon| := \sum_{j=1}^d \upsilon_j$.

For a positive integer q, let $\alpha = (\alpha_1, \ldots, \alpha_q) \in \mathbb{R}_+^q$, and $\mathbf{d} = (d_1, \ldots, d_{q+1}) \in \mathbb{N}_+^{q+1}$, $\tilde{\mathbf{d}} = (\tilde{d}_1, \ldots, \tilde{d}_q) \in \mathbb{N}_+^q$ with $\tilde{d}_j \leqslant d_j$. We then define a composite Hölder smooth function class as

$$\mathcal{H}(q, \alpha, \mathbf{d}, \tilde{\mathbf{d}}, M) = \{f = f_q \circ \cdots \circ f_1 : f_i = (f_{i1}, \ldots, f_{id_{i+1}})^\top, f_{ij} \in \mathcal{H}_{\tilde{d}_i}^{\alpha_i}([a_i, b_i]^{\tilde{d}_i}, M), |a_i|, |b_i| \leqslant M\}, \tag{A.1}$$

where $[a_i, b_i]$ is the bounded domain for each Hölder smooth function.

More Notation

Denote $a_n \lesssim b_n$ as $a_b \leqslant cb_n$ for some $c > 0$ when n is sufficiently large; $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Let $\eta(\cdot, \cdot) = (\boldsymbol{\beta}^\top \cdot, g(\cdot)) : \mathbb{R}^p \times \mathbb{R}^r \to \mathbb{R}^2$ denote the collection of a linear operator and a nonlinear operator. In this section, denote by $\mathbf{v} = (\mathbf{x}^\top, \mathbf{z}^\top)^\top$ the random vector underlying the observed IID data of $\mathbf{v}_i = (\mathbf{x}_i^\top, \mathbf{z}_i^\top)^\top$, and (T, Δ) the random vector underlying the observed IID data of $(T_i, \Delta_i), i = 1, \ldots, n$. Let $N(t) = I(T \leqslant t, \Delta = 1)$ and $N_i(t) = I(T_i \leqslant t, \Delta_i = 1)$. To simplify notation, we denote by $\eta(\mathbf{v}) = \boldsymbol{\beta}^\top \mathbf{x} + g(\mathbf{z})$. Denote the truth of $\eta(\cdot, \cdot)$ by $\eta_0(\cdot, \cdot) = (\boldsymbol{\beta}_\mathbf{0}^\top \cdot, g_0(\cdot))$. For two operators, say, $\eta_1(\cdot, \cdot) = (\boldsymbol{\beta}_1^\top \cdot, g_1(\cdot))$ and $\eta_2(\cdot, \cdot) = (\boldsymbol{\beta}_2^\top \cdot, g_2(\cdot))$, define their distance as

$$d^2(\eta_1, \eta_2) := \mathbb{E}[\{\eta_1(\mathbf{v}) - \eta_2(\mathbf{v})\}^2] = \int \{\eta_1(\mathbf{t}) - \eta_2(\mathbf{t})\}^2 f_\mathbf{v}(\mathbf{t}) d\mathbf{t},$$

and the corresponding norm

$$\|\eta\|^2 := \mathbb{E}[\eta^2(\mathbf{v})] = \int \eta^2(\mathbf{t}) f_{\mathbf{v}}(\mathbf{t}) d\mathbf{t}.$$

For the notational ease, we write $\eta = (\boldsymbol{\beta}, g)$ in the following.

With $Y(t) = 1(T \geqslant t)$ and $Y_i(t) = 1(T_i \geqslant t)$, define

$$S_{0n}(t,\eta) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\exp\{\eta(\mathbf{v}_i)\}, \qquad S_0(t,\eta) = \mathbb{E}[Y(t)\exp\{\eta(\mathbf{v})\}],$$

and for any vector function $\mathbf{h}$ of $\mathbf{v}$ define

$$S_{1n}(t,\eta,\mathbf{h}) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\mathbf{h}(\mathbf{v}_i)\exp\{\eta(\mathbf{v}_i)\}, \qquad S_1(t,\eta,\mathbf{h}) = \mathbb{E}[Y(t)\mathbf{h}(\mathbf{v})\exp\{\eta(\mathbf{v})\}],$$

where the expectation is taken with respect to the joint distribution of T and $\mathbf{v}$.

Let

$$l_n(t,\mathbf{v},\eta) = \eta(\mathbf{v}) - \log S_{0n}(t,\eta), \qquad l_0(t,\mathbf{v},\eta) = \eta(\mathbf{v}) - \log S_0(t,\eta).$$

Then the partial likelihood in (2)

can be written as

$$\ell(\eta) = \frac{1}{n}\sum_{i=1}^{n}\{\Delta_i l_n(T_i,\mathbf{v}_i,\eta) - \Delta_i \log n\}.$$

Since $\sum_{i=1}^{n} \Delta_i \log n$ does not involve unknown parameters and can be dropped in optimization, we replace below $\ell(\eta)$ by $\frac{1}{n}\sum_{i=1}^{n}\{\Delta_i l_n(T_i,\mathbf{v}_i,\eta)\}$.

Finally, for any function h of $(\mathbf{v},\Delta,T)$, where (Δ,T) is the random vector underlying (Δ_i,T_i), define

$$\mathbb{P}_n\{h(\mathbf{v},\Delta,T)\} = \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{v}_i,\Delta_i,T_i), \qquad \mathbb{P}\{h(\mathbf{v},\Delta,T)\} = \mathbb{E}\{h(\mathbf{v},\Delta,T)\},$$

and in particular, we define $L_n(\eta) = \mathbb{P}_n\{\Delta l_n(T,\mathbf{v},\eta)\}$ and $L_0(\eta) = \mathbb{P}\{\Delta l_0(T,\mathbf{v},\eta)\}$. Here, the expectation is taken with respect to the joint distribution of T, Δ and $\mathbf{v}$.

Proof of Theorem 1

Define $\alpha_n = \gamma_n \log^2 n + a_n = \tau_n + a_n$. For some $D > 0$, let $\mathbb{R}^p_D \coloneqq \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_\infty < D\}$ and $\mathcal{G}_D \coloneqq \mathcal{G}(L, \mathbf{p}, s, D)$, and define

$$\hat{\eta}_D = \underset{\eta \in \mathbb{R}^p_D \times \mathcal{G}_D}{\operatorname{argmax}}\ PL(\eta).$$

Further, denote by $\hat{\eta} = (\hat{\boldsymbol{\beta}}, \hat{g})$ a local maximizer of $PL(\eta)$ over $\mathbb{R}^p \times \mathcal{G}$, that is, by setting $D = \infty$ in $\mathbb{R}^p_D$ and $\mathcal{G}_D$. As in Zhong et al. (2022), it can be shown that if $\max(||\beta||, ||g||_\infty) \to \infty$, $PL(\eta) \to -\infty$; hence, when D is sufficiently large, $\hat{\eta} = \hat{\eta}_D$ almost surely. Therefore, in the following, we show that $d(\hat{\eta}_D, \eta_0) = O_p(\alpha_n)$, when D is sufficiently large.

To do so, it suffices to show that for any $\epsilon > 0$, there exists a C such that

$$\mathrm{P}\left\{\sup_{\eta \in \mathcal{N}_c} PL(\eta) < PL(\eta_0)\right\} \geqslant 1 - \epsilon, \tag{A.2}$$

where $\mathcal{N}_c = \{\eta \in \mathbb{R}^p_D \times \mathcal{G}_D : d(\eta, \eta_0) = C\alpha_n\}$. If it holds, it implies with probability at least $1-\epsilon$ that there exists a $C > 0$ such that a local maximum exists and is inside the ball $\mathcal{N}_c$. Hence, there exists a local maximizer such that $d(\hat{\eta}, \eta_0) = O_p(\alpha_n)$.

Without loss of generality, we assume that η satisfies $\mathbb{E}\{\eta(\mathbf{v})\} = \mathbb{E}\{\eta_0(\mathbf{v})\}$, implying $\mathbb{E}\{g(\mathbf{z})\} = 0$; if not, we can always centralize it. To see this, consider any $\eta = (\beta, g)$ in the ball $B_C = \{\eta \in \mathbb{R}^p_D \times \mathcal{G}_D : d(\eta, \eta_0) \leqslant C\alpha_n\}$, its centralization $\eta' = (\beta, g - \mathbb{E}\{\eta(\mathbf{v}) - \eta_0(\mathbf{v})\})$ is also in the ball B_C, satisfying $\mathbb{E}\{\eta'(\mathbf{v})\} = \mathbb{E}\{\eta_0(\mathbf{v})\}$ and $PL(\eta') = PL(\eta)$.

Because of the sparsity of the β-coefficients, we arrange the indices of the covariates $(x_1, \ldots, x_p)$ so that $\beta_{j0} = 0$ when $j > s_\beta$. We consider

$$\begin{aligned} & PL(\eta) - PL(\eta_0) \\ = \ & \{L_n(\eta) - L_n(\eta_0)\} - \sum_{j=1}^{p} \{p_\lambda(|\beta_j|) - p_\lambda(|\beta_{j0}|)\} \\ \leqslant \ & \{L_n(\eta) - L_n(\eta_0)\} - \sum_{j=1}^{s_\beta} \{p_\lambda(|\beta_j|) - p_\lambda(|\beta_{j0}|)\}, \end{aligned} \tag{A.3}$$

where the inequality holds because $p_\lambda(|\beta_j|) - p_\lambda(0) > 0$ when $j > s_\beta$.

We first deal with

$$\begin{aligned} L_n(\eta) - L_n(\eta_0) =& \{L_0(\eta) - L_0(\eta_0)\} \\ &+ \{L_n(\eta) - L_0(\eta)\} - \{L_n(\eta_0) - L_0(\eta_0)\}. \end{aligned} \tag{A.4}$$

According to Lemma 2 in Zhong et al. (2022), we know that

$$L_0(\eta) - L_0(\eta_0) \asymp -d^2(\eta, \eta_0).$$

Since $d(\eta, \eta_0) = C\alpha_n$, the first term in the right hand side of A.4 is of the order $C^2\alpha_n^2$.

After some calculation,

$$\begin{aligned} (L_n - L_0)(\eta) - (L_n - L_0)(\eta_0) =& (\mathbb{P}_n - \mathbb{P})\{\Delta l_0(T, \mathbf{v}, \eta) - \Delta l_0(T, \mathbf{v}, \eta_0)\} \\ &+ \mathbb{P}_n \Big\{ \Delta \log \frac{S_0(T, \eta)}{S_0(T, \eta_0)} - \Delta \log \frac{S_{0n}(T, \eta)}{S_{0n}(T, \eta_0)} \Big\} \\ =& I + II. \end{aligned} \tag{A.5}$$

According to the proof of Theorem 3.1 in Zhong et al. (2022), with $\mathcal{A}_\delta = \{(\beta, g) \in \mathbb{R}^p_D \times \mathcal{G}_D : \delta/2 \leqslant d(\eta, \eta_0) \leqslant \delta\}$, it follows that

$$\sup_{\eta \in \mathcal{A}_\delta} |I| = O(n^{-1/2}\phi_n(\delta)),$$

$$\sup_{\eta \in \mathcal{A}_\delta} |II| \leqslant O(n^{-1/2}\phi_n(\delta)),$$

where $\phi_n(\delta) = \delta\sqrt{s \log \frac{\mathcal{U}}{\delta}} + \frac{s}{\sqrt{n}} \log \frac{\mathcal{U}}{\delta}$ and $\mathcal{U} = L \prod_{l=1}^{L}(p_l + 1) \sum_{l=1}^{L} p_l p_{l+1}$.Then by Assumption 1, when $\delta = C(\tau_n + a_n)$, we can show that $n^{-1/2}\phi_n\{C(\tau_n + a_n)\} \leqslant C(\tau_n + a_n)^2 = C\alpha_n^2$.

By the Taylor expansion and the Cauchy-Schwarz inequality, the second term on the right-hand side of (A.3) is bounded by

$$\sqrt{s_\beta} a_n \|\boldsymbol{\beta} - \boldsymbol{\beta_0}\| + \frac{1}{2} b_n \|\boldsymbol{\beta} - \boldsymbol{\beta_0}\|^2.$$

Since $d(\eta, \eta_0) = C\alpha_n$, and therefore $\|\boldsymbol{\beta} - \boldsymbol{\beta_0}\|$ is of the order $C\alpha_n$. Hence, this upper bound is dominated by the first term in (A.4) as $b_n \to 0$ by the assumption.

Therefore, for any $\epsilon > 0$, there exist sufficiently large $C, D > 0$ so that (A.2) holds, and hence $d(\hat{\eta}_D, \eta_0) = O_p(\alpha_n)$, which gives $d(\hat{\eta}, \eta_0) = O_p(\alpha_n)$, where we recall $\hat{\eta}$ is the local maximizer of $PL(\eta)$ over $\mathbb{R}^p \times \mathcal{G}$. We note that

$$\begin{aligned} d^2(\hat{\eta}, \eta_0) &= \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \{\mathbf{x} - \mathbb{E}(\mathbf{x}|\mathbf{z})\} + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \mathbb{E}(\mathbf{x}|\mathbf{z}) + \{\hat{g}(\mathbf{z}) - g_0(\mathbf{z})\}]^2 \\ &= \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \{\mathbf{x} - \mathbb{E}(\mathbf{x}|\mathbf{z})\}]^2 + \mathbb{E}[\{\hat{g}(\mathbf{z}) - g_0(\mathbf{z})\} + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \mathbb{E}(\mathbf{x}|\mathbf{z})]^2, \end{aligned}$$

where the second equality holds because, by the definition of $d(\cdot,\cdot)$, $\mathbb{E}$ is taken with respect to the joint density of $\mathbf{v} = (\mathbf{x}^\top, \mathbf{z}^\top)^\top$, which is independent of the observed data, and hence, $\hat{\boldsymbol{\beta}}$ and $\hat{g}$. By Assumptions 2-4, it follows $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(\alpha_n)$ and $\|\hat{g} - g_0\|_{L^2} = O_p(\alpha_n)$.

Proof of Theorem 2

For the claims made in Theorem 2, it suffices to show that, with probability tending to 1, for any given $\eta = (\boldsymbol{\beta}, g)$ satisfying that $||\eta - \eta_0|| = O(\gamma_n \log^2 n)$, where $\eta_0 = (\boldsymbol{\beta}_0, g_0)$, and some constant $C > 0$,

$$PL\{(\boldsymbol{\beta}_1^\top, \mathbf{0}^\top)^\top, g\} = \max_{\|\boldsymbol{\beta}_2\| \leqslant C\gamma_n \log^2 n} PL\{(\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top, g\},$$

where $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_{s_\beta})^\top$ and $\boldsymbol{\beta}_2 = (\beta_{s_\beta+1}, \ldots, \beta_p)^\top$. We only need to show that, for any $j = s_\beta + 1, \ldots, p$,

$$\begin{aligned} &\partial PL(\boldsymbol{\beta}, g)/\partial\beta_j < 0, \quad \text{for } 0 < \beta_j < C\gamma_n \log^2 n; \\ &\partial PL(\boldsymbol{\beta}, g)/\partial\beta_j > 0, \quad \text{for } -C\gamma_n \log^2 n < \beta_j < 0. \end{aligned}$$

To proceed, we note that $\partial PL(\boldsymbol{\beta}, g)/\partial\beta_j = \partial\ell(\eta)/\partial\beta_j - sign(\beta_j)p'_\lambda(|\beta_j|)$ for $j = s_\beta + 1, \ldots, p$. Denote by $F_j(\eta)$ the partial derivative of $\ell(\eta)$ w.r.t. β_j, i.e.

$$F_j(\eta) = \frac{\partial\ell(\eta)}{\partial\beta_j} = \frac{1}{n}\sum_{1=1}^{n}\int_0^\tau \left\{x_{i,j} - \frac{\sum_{k=1}^n Y_k(s) x_{k,j} \exp(\boldsymbol{\beta}^\top \mathbf{x}_k + g(\mathbf{z}_k))}{\sum_{k=1}^n Y_k(s) \exp(\boldsymbol{\beta}^\top \mathbf{x}_k + g(\mathbf{z}_k))}\right\} dN_i(s),$$

where $x_{k,j}$ (or $x_{i,j}$) is the jth element of $\mathbf{x}_k$ (or $\mathbf{x}_i$). As part of η is a functional, we consider a functional expansion of $F_j(\eta)$ around its truth, η_0. Specifically, for a real number $0 \leqslant e \leqslant 1$, we define $\mathcal{F}_j(e) = F_j\{\eta_0 + e(\eta - \eta_0)\}$, a function of the scalar e only. Obviously, $\mathcal{F}_j(1) = F_j(\eta)$ and $\mathcal{F}_j(0) = F_j(\eta_0)$.

Taking the Taylor expansion of $\mathcal{F}_j(1)$ around 0 gives

$$\mathcal{F}_j(1) = \mathcal{F}_j(0) + \mathcal{F}_j'(0) + \mathcal{F}_j''(e^*), \tag{A.6}$$

where e^* is between 0 and 1. By some calculation,

$$\mathcal{F}_j'(e) = -\frac{1}{n}\sum_{i=1}^n \int_0^\tau \Big[\frac{\sum_k Y_k(s)\xi_e(\mathbf{v}_k)x_{k,j}(\eta-\eta_0)(\mathbf{v}_k)}{\sum_k Y_k(s)\xi_e(\mathbf{v}_k)} - \frac{\{\sum_k Y_k(s)\xi_e(\mathbf{v}_k)x_{k,j}\}\{\sum_k Y_k(s)\xi_e(\mathbf{v}_k)(\eta-\eta_0)(\mathbf{v}_k)\}}{\{\sum_k Y_k(s)\xi_e(\mathbf{v}_k)\}^2}\Big] dN_i(s),$$

where $\mathbf{v}_k = (\mathbf{x}_k^\top, \mathbf{z}_k^\top)^\top$, $\xi_e(\mathbf{v}_k) = \exp(\{\eta_0 + e(\eta - \eta_0)\}(\mathbf{v}_k))$ and $(\eta - \eta_0)(\mathbf{v}_k) = (\beta - \beta_0)^\top \mathbf{x}_k + (g - g_0)(\mathbf{z}_k)$, and

$$\begin{aligned}\mathcal{F}_j''(e) = -\frac{1}{n}\sum_{i=1}^n \int_0^\tau \Big[&\frac{\sum_k Y_k(s)\xi_e(\mathbf{v}_k)x_{k,j}(\eta-\eta_0)^2(\mathbf{v}_k)}{\sum_k Y_k(s)\xi_e(\mathbf{v}_k)} \\ &- \frac{2\{\sum_k Y_k(s)\xi_e(\mathbf{v}_k)x_{k,j}(\eta-\eta_0)(\mathbf{v}_k)\}\{\sum_k Y_k(s)\xi_e(\mathbf{v}_k)(\eta-\eta_0)(\mathbf{v}_k)\}}{\{\sum_k Y_k(s)\xi_e(\mathbf{v}_k)\}^2} \\ &- \frac{\{\sum_k Y_k(s)\xi_e(\mathbf{v}_k)x_{k,j}\}\{\sum_k Y_k(s)\xi_e(\mathbf{v}_k)(\eta-\eta_0)^2(\mathbf{v}_k)\}}{\{\sum_k Y_k(s)\xi_e(\mathbf{v}_k)\}^2} \\ &+ \frac{2\{\sum_k Y_k(s)\xi_e(\mathbf{v}_k)x_{k,j}\}\{\sum_k Y_k(s)\xi_e(\mathbf{v}_k)(\eta-\eta_0)(\mathbf{v}_k)\}^2}{\{\sum_k Y_k(s)\xi_e(\mathbf{v}_k)\}^3}\Big] dN_i(s).\end{aligned}$$

It follows that $\mathcal{F}_j(0)$ in (A.6) is equal to

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \int_0^\tau \Big\{x_{i,j} - \frac{\sum_{k=1}^n Y_k(s)x_{k,j}\exp(\boldsymbol{\beta}_0^\top \mathbf{x}_k + g_0(\mathbf{z}_k))}{\sum_{k=1}^n Y_k(s)\exp(\boldsymbol{\beta}_0^\top \mathbf{x}_k + g_0(\mathbf{z}_k))}\Big\} dN_i(s) \\ = &\frac{1}{n}\sum_{1=1}^n \int_0^\tau \Big\{x_{i,j} - \frac{\sum_{k=1}^n Y_k(s)x_{k,j}\exp(\boldsymbol{\beta}_0^\top \mathbf{x}_k + g_0(\mathbf{z}_k))}{\sum_{k=1}^n Y_k(s)\exp(\boldsymbol{\beta}_0^\top \mathbf{x}_k + g_0(\mathbf{z}_k))}\Big\} dM_i(s),\end{aligned}$$

where $dM_i(s) = dN_i(s) - \lambda_0(s)Y_i(s)\exp(\boldsymbol{\beta}_0^\top \mathbf{x}_i + g_0(\mathbf{z}_i))ds$ is the martingale with respect to the history up to time s. Hence, $n^{1/2}\mathcal{F}_j(0)$ converges in distribution to a normal distribution by the martingale central limit theorem (Fleming and Harrington, 2013), and therefore, $\mathcal{F}_j(0) = O_p(n^{-1/2})$.

We then consider

$$
\begin{aligned}
\mathcal{F}_j'(0) &= -\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\Big[\frac{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)x_{k,j}(\eta-\eta_0)(\mathbf{v}_k)}{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)} \\
&- \frac{\{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)x_{k,j}\}\{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)(\eta-\eta_0)(\mathbf{v}_k)\}}{\{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)\}^2}\Big]dN_i(s) \\
&= -\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\Big[\frac{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)x_{k,j}(\eta-\eta_0)(\mathbf{v}_k)}{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)} \\
&- \frac{\{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)x_{k,j}\}\{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)(\eta-\eta_0)(\mathbf{v}_k)\}}{\{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)\}^2}\Big]dM_i(s) \\
&- \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\Big[\frac{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)x_{k,j}(\eta-\eta_0)(\mathbf{v}_k)}{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)} \\
&- \frac{\{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)x_{k,j}\}\{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)(\eta-\eta_0)(\mathbf{v}_k)\}}{\{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)\}^2}\Big]Y_i(s)\xi_0(\mathbf{v}_i)\lambda_0(s)ds \\
&= I_1 + I_2,
\end{aligned}
$$

where $\xi_0(\mathbf{v}_k) = \exp(\eta_0(\mathbf{v}_k)) = \exp(\boldsymbol{\beta}_0^\top \mathbf{x}_i + g_0(\mathbf{z}_i))$. It follows that each summed item in I_1, i.e.,

$$
\begin{aligned}
&\int_0^{\tau}\Big[\frac{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)x_{k,j}(\eta-\eta_0)(\mathbf{v}_k)}{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)} \\
&- \frac{\{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)x_{k,j}\}\{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)(\eta-\eta_0)(\mathbf{v}_k)\}}{\{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)\}^2}\Big]dM_i(s),
\end{aligned}
$$

is a square integrable martingale (Fleming and Harrington, 2013). Hence, by the law of large numbers for martingales (Hall and Heyde, 2014), $I_1 \to 0$ in probability.

Also, I_2 can be shown to be equal to

$$
\begin{aligned}
&-\int_0^{\tau}\frac{1}{n}\Big[\sum_k Y_k(s)\xi_0(\mathbf{v}_k)x_{k,j}(\eta-\eta_0)(\mathbf{v}_k) \\
&- \frac{\{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)x_{k,j}\}\{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)(\eta-\eta_0)(\mathbf{v}_k)\}}{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)}\Big]\lambda_0(s)ds.
\end{aligned}
$$

Define

$$
S_{x_j,1}(s,\eta-\eta_0) = \mathbb{E}[Y_k(s)\xi_0(\mathbf{v}_k)x_{k,j}(\eta-\eta_0)(\mathbf{v}_k)], \quad S_{x_j}(s,\eta_0) = \mathbb{E}[Y_k(s)\xi_0(\mathbf{v}_k)x_{k,j}],
$$

$$
S_1(s,\eta-\eta_0) = \mathbb{E}[Y_k(s)\xi_0(\mathbf{v}_k)(\eta-\eta_0)(\mathbf{v}_k)], \qquad S_0(s,\eta_0) = \mathbb{E}[Y_k(s)\xi_0(\mathbf{v}_k)].
$$

Applying the empirical process arguments (Pollard, 1990; Wellner, 2005) yields that

$$\begin{aligned} & \sup_{s\in[0,\tau]} \Big|\frac{1}{n}\Big[\sum_k Y_k(s)\xi_0(\mathbf{v}_k)x_{k,j}(\eta-\eta_0)(\mathbf{v}_k) \\ - & \frac{\{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)x_{k,j}\}\{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)(\eta-\eta_0)(\mathbf{v}_k)\}}{\sum_k Y_k(s)\xi_0(\mathbf{v}_k)}\Big] \\ - & S_{x_j,1}(s,\eta-\eta_0) + \frac{S_{x_j}(s,\eta_0)S_1(s,\eta-\eta_0)}{S_0(s,\eta_0)}\Big| \to 0 \end{aligned}$$

in probability, which implies that

$$I_2 \to -\int_0^\tau \left\{S_{x_j,1}(s,\eta-\eta_0) - \frac{S_{x_j}(s,\eta_0)S_1(s,\eta-\eta_0)}{S_0(s,\eta_0)}\right\}\lambda_0(s)ds$$

in probability. Collecting all these terms, we thus have that

$$\mathcal{F}_j'(0) = -\int_0^\tau \left\{S_{x_j,1}(s,\eta-\eta_0) - \frac{S_{x_j}(s,\eta_0)S_1(s,\eta-\eta_0)}{S_0(s,\eta_0)}\right\}\lambda_0(s)ds + o_p(1).$$

We now bound $\mathcal{F}_j'(0)$. First note that, for any $s \in [0,\tau]$,

$$\begin{aligned} S_{x_j,1}(s,\eta-\eta_0) & \leqslant \mathbb{E}[\xi_0(\mathbf{v}_k)|x_{k,j}||(\eta-\eta_0)(\mathbf{v}_k)|] \\ & \leqslant \max_{\mathbf{v}_k\in\mathbb{D}}\{\xi_0(\mathbf{v}_k)|x_{k,j}|\}\int_{\mathbb{D}}|(\eta-\eta_0)(\mathbf{v})|f_{\mathbf{v}_k}(\mathbf{v})d\mathbf{v} \\ & \leqslant C_1\left\{\int_{\mathbb{D}}(\eta-\eta_0)^2(\mathbf{v})f_{\mathbf{v}_k}(\mathbf{v})d\mathbf{v}\right\}^{1/2} \\ & = C_1||\eta-\eta_0||, \end{aligned}$$

where $C_1 > 0$ is a constant, $f_{\mathbf{v}_k}(\cdot)$ is the density function of the random vector $\mathbf{v}_k$, and the last inequality stems from the Cauchy-Schwartz inequality, in conjunction with the boundedness assumptions on the covariates (i.e., $\mathbb{D}$ is bounded) and η_0 (Conditions 2 and 3 in the main text). Similarly, we can show that, for any $s \in [0,\tau]$,

$$|S_1(s,\eta-\eta_0)| \leqslant C_2||\eta-\eta_0||,\ \ |S_{x_j}(s,\eta_0)| \leqslant C_3,\ \ S_0(s,\eta_0) \geqslant C_4,$$

where $C_2, C_3, C_4 > 0$ are constants. The last inequality holds because at τ, there is at least probability of $\delta > 0$ of observing subjects at risk (Condition 4 in the main text), implying that $\min_{s\in[0,\tau]}\mathbb{E}(Y_k(s)|\mathbf{v}_k) \geqslant \delta > 0$ a.s.

As

$$
\begin{aligned}
& |\int_0^\tau \left\{ S_{x_j,1}(s,\eta-\eta_0) - \frac{S_{x_j}(s,\eta_0)S_1(s,\eta-\eta_0)}{S_0(s,\eta_0)} \right\} \lambda_0(s)ds| \\
\leqslant\ & \int_0^\tau |\{S_{x_j,1}(s,\eta-\eta_0)|\lambda_0(s)ds + \int_0^\tau \frac{|S_{x_j}(s,\eta_0)||S_1(s,\eta-\eta_0)|}{S_0(s,\eta_0)}\lambda_0(s)ds \\
\leqslant\ & (C_1 + C_2C_3C_4^{-1})\Lambda_0(\tau)||\eta-\eta_0||,
\end{aligned}
$$

where $\Lambda_0(\tau) = \int_0^\tau \lambda_0(s)ds < \infty$. Therefore, $\mathcal{F}_j'(0) = O_p(||\eta - \eta_0||)$.

Similarly, using the explicit form of $\mathcal{F}_j''(e)$, some calculation can show that $\mathcal{F}_j''(e^*) = o_p(||\eta - \eta_0||)$. Then we conclude that

$$
\begin{aligned}
\partial PL(\boldsymbol{\beta}, g)/\partial\beta_j &= \partial\ell(\eta)/\partial\beta_j - sign(\beta_j)p_\lambda'(|\beta_j|) \\
&= \lambda[\lambda^{-1}(\mathcal{F}_j(0) + \mathcal{F}_j'(0) + \mathcal{F}_j''(e^*)) - sign(\beta_j)\lambda^{-1}p_\lambda'(|\beta_j|)].
\end{aligned}
$$

With the assumptions of $\lambda^{-1}\gamma_n \log^2(n) \to 0$ and $\lambda^{-1}n^{-1/2} \to 0$, it follows that

$$
\lambda^{-1}(\mathcal{F}_j(0) + \mathcal{F}_j'(0) + \mathcal{F}_j''(e^*)) = o_p(1).
$$

On the other hand, using the condition of $\liminf_{n\to\infty} \liminf_{\theta\to 0+} \lambda^{-1}p_\lambda'(\theta) > 0$, it follows that the sign of $\partial PL(\boldsymbol{\beta}, g)/\partial\beta_j$ is the opposite sign of β_j with probability going to 1. Hence, the claims follow.

[Figure S.1 about here.]

[Figure S.2 about here.]

[Figure S.3 about here.]

References

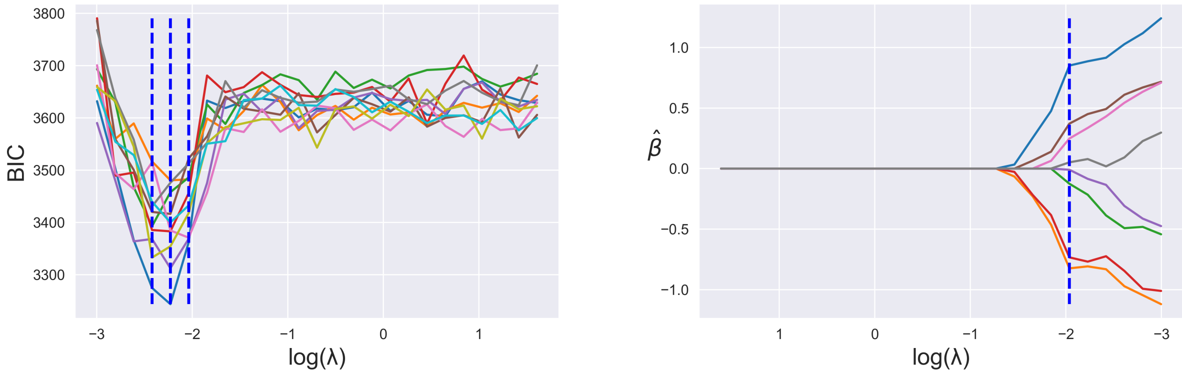
Fleming, T. R. and Harrington, D. P. (2013). *Counting processes and survival analysis*, volume 625. John Wiley & Sons.

Hall, P. and Heyde, C. C. (2014). *Martingale limit theory and its application.* Academic press.

Pollard, D. (1990). Empirical processes: theory and applications. Ims.

Wellner, J. A. (2005). Empirical processes: Theory and applications. *Notes for a course given at Delft University of Technology* page 17.

Zhong, Q., Mueller, J., and Wang, J.-L. (2022). Deep learning for the partially linear Cox model. *The Annals of Statistics* **50,** 1348–1375.



(a) Selection of λ for 10 simulated datasets

(b) Selection path for the non-zero coefficients

Figure S.1: Selection of λ in Penalized DPLC using BIC.

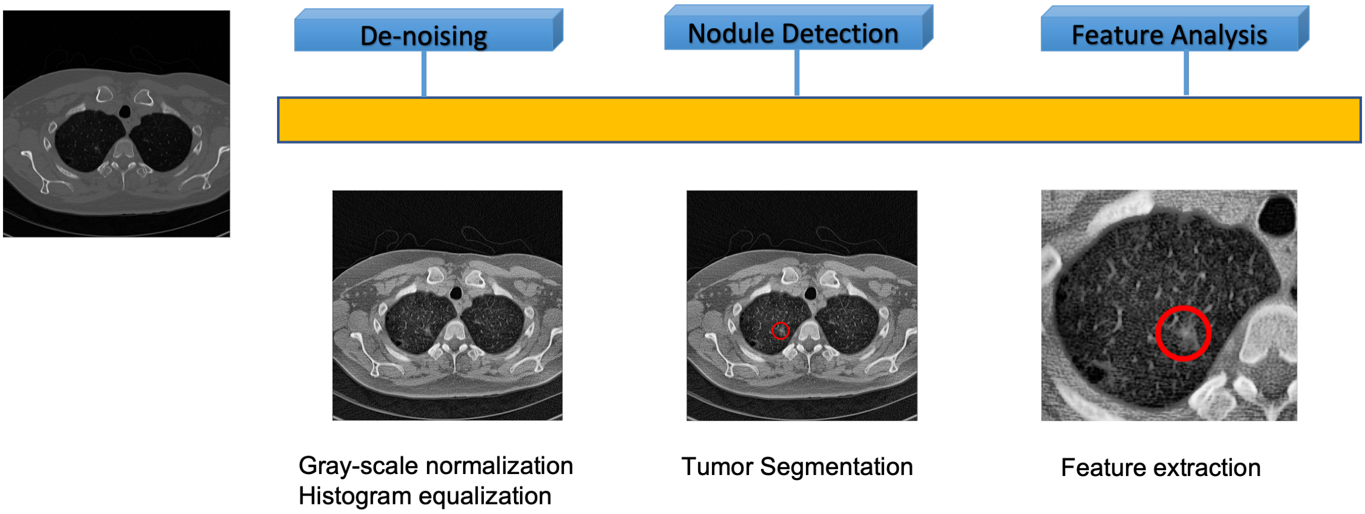


Figure S.2: **Image Preprocessing Pipeline**

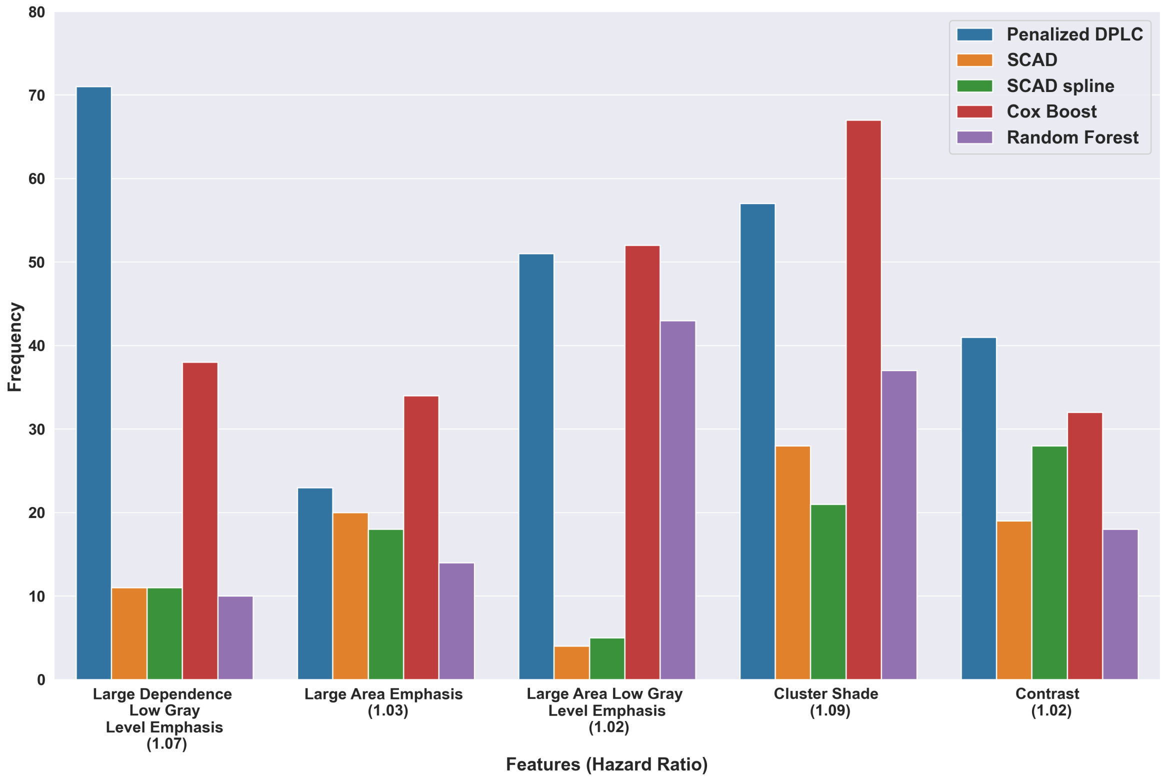


Figure S.3: **Selection Frequency and Hazard Ratio of Selected Features:** The selection frequency of the most frequently selected five texture features is reported. The hazard ratio is the average of 100 experiments