Supplementary Materials for Competing Risk Modeling with Bivariate Varying Coefficients to Understand the Dynamic Impact of COVID-19

Wenbo Wu, John D. Kalbfleisch, Jeremy M. G. Taylor, Jian Kang, Kevin He

Appendix A Gradient and Hessian of $\ell_{jg}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j)$

For $g = 1, \ldots, G$, $i = 1, \ldots, n_g$, and $j = 1, \ldots, m$, we define

$$S^{(u)}_{jgi}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j, X_{gi}) \coloneqq \sum_{r \in R_g(X_{gi})} \exp\{\mathbf{L}^\top_{gr}(X_{gi})\boldsymbol{\gamma}_j + \mathbf{W}^\top_{gr}\boldsymbol{\theta}_j\} \begin{bmatrix} \mathbf{L}_{gr}(X_{gi}) \\ \mathbf{W}_{gr} \end{bmatrix}^{\odot u}, \quad u = 0, 1, 2,$$

where $\mathbf{L}_{gr}(X_{gi}) \coloneqq \mathbf{Z}_{gr} \otimes \breve{\mathbf{B}}(\breve{X}_{gr}) \otimes \mathbf{B}(X_{gi})$, and for a vector $\mathbf{v} \in \mathbb{R}^p$, $\mathbf{v}^{\odot 0} \coloneqq 1$, $\mathbf{v}^{\odot 1} \coloneqq \mathbf{v}$, and $\mathbf{v}^{\odot 2} \coloneqq \mathbf{v}\mathbf{v}^\top$. The gradient $\dot{\ell}_{jg}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j)$ and Hessian $\ddot{\ell}_{jg}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j)$ of $\ell_{jg}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j)$ are hence given by

$$\dot{\ell}_{jg}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j) = \sum_{i=1}^{n_g} \Delta_{jgi} \left\{ \begin{bmatrix} \mathbf{L}_{gi}(X_{gi}) \\ \mathbf{W}_{gi} \end{bmatrix} - \mathbf{U}^{(1)}_{jgi}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j, X_{gi}) \right\}, \tag{1}$$

$$\ddot{\ell}_{jg}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j) = -\sum_{i=1}^{n_g} \Delta_{jgi} \mathbf{V}_{jgi}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j, X_{gi}), \tag{2}$$

in which

$$\mathbf{U}^{(w)}_{jgi}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j, X_{gi}) \coloneqq \frac{S^{(w)}_{jgi}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j, X_{gi})}{S^{(0)}_{jgi}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j, X_{gi})}, \quad w = 1, 2,$$

$$\mathbf{V}_{jgi}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j, X_{gi}) \coloneqq \mathbf{U}^{(2)}_{jgi}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j, X_{gi}) - \{\mathbf{U}^{(1)}_{jgi}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j, X_{gi})\}^{\odot 2}.$$

Appendix B Tensor-Product Proximal Newton Algorithm

Let $X_{jg1} < \cdots < X_{jgn_{jg}}$ denote the n_{jg} distinct times of type j failures within stratum g. For failure time X_{jgb}, $b = 1, \ldots, n_{jg}$, let $\mathbf{Z}_{jgb}$, $\mathbf{W}_{jgb}$, and $\breve{X}_{jgb}$ denote $\mathbf{Z}_{gi}$, $\mathbf{W}_{gi}$, and $\breve{X}_{gi}$, respectively, such that $\Delta_{jgi} = 1$ and $X_{gi} = X_{jgb}$. The tensor-product proximal Newton algorithm is outlined as Algorithm 1. Readers are referred to Wu et al. (2022) for theoretical arguments justifying the convergence of the algorithm. In the algorithm, ξ is used to control the modified Hessian in Line 18; a large value of 10^8 is used as default since it leads to a slight modification of the Hessian. The parameter δ is used to control the expansion of the series of ξ across iterations. By default, 1 is used, indicating that ξ remains constant across iterations. One may consider any value greater than 1 to obtain an increasing sequence of ξ so that the modification to the Hessian is shrinking. The parameter ϵ indicates the tolerance level with respect to the squared Newton increment η^2, with 10^{-10} as the default.

Algorithm 1: Tensor-Product Proximal Newton

```
for $j \leftarrow 1$ to $m$ do                                                   // $m$ failure types
    initialize $s \leftarrow 0$, $\xi_s > 0$, $\boldsymbol{\gamma}_j^{(s)} = \mathbf{0}$, and $\boldsymbol{\theta}_j^{(s)} = \mathbf{0}$;
    set $\phi \in (0, 0.5)$, $\psi \in (0.5, 1)$, $\delta \geq 1$ and $\epsilon > 0$;
    do
        for $g \leftarrow 1$ to $G$ do                                           // $G$ distinct strata
            for $b \leftarrow 1$ to $n_{jg}$ do                                  // $n_{jg}$ distinct failure times
                for $u \leftarrow 0$ to 2 do
                    $S_{jgb}^{(u)}(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}, X_{jgb}) = \sum_{r \in R_g(X_{jgb})} \exp\{\mathbf{L}_{gr}^\top(X_{jgb})\boldsymbol{\gamma}_j^{(s)} + \mathbf{W}_{gr}^\top \boldsymbol{\theta}_j^{(s)}\} \begin{bmatrix} \mathbf{L}_{gr}(X_{jgb}) \\ \mathbf{W}_{gr} \end{bmatrix}^{\odot u}$;
                end
                for $w \leftarrow 1$ to 2 do
                    $\mathbf{U}_{jgb}^{(w)}(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}, X_{jgb}) = S_{jgb}^{(w)}(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}, X_{jgb}) / S_{jgb}^{(0)}(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}, X_{jgb})$;
                end
                $\mathbf{V}_{jgb}(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}, X_{jgb}) = \mathbf{U}_{jgb}^{(2)}(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}, X_{jgb}) - \left[\mathbf{U}_{jgb}^{(1)}(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}, X_{jgb})\right]^{\odot 2}$;
            end
        end
        $\dot{\ell}_j(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}) = \sum_{g=1}^{G} \sum_{q=1}^{n_j} \left\{ \begin{bmatrix} \mathbf{L}_{jgb}(X_{jgb}) \\ \mathbf{W}_{jgb} \end{bmatrix} - \mathbf{U}_{jgb}^{(1)}(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}, X_{jgb}) \right\}$;
        $\ddot{\ell}_j(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}) = -\sum_{g=1}^{G} \sum_{q=1}^{n_j} \mathbf{V}_{jgb}(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}, X_{jgb})$;
        $\begin{bmatrix} \Delta\boldsymbol{\gamma}_j^{(s)} \\ \Delta\boldsymbol{\theta}_j^{(s)} \end{bmatrix} = \left[\mathbf{Q}_j(\boldsymbol{\mu}_j, \breve{\boldsymbol{\mu}}_j) + n\mathbf{I}/\xi_s - \ddot{\ell}_j(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)})\right]^{-1} \left\{ \dot{\ell}_j(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}) - \mathbf{Q}_j(\boldsymbol{\mu}_j, \breve{\boldsymbol{\mu}}_j) \begin{bmatrix} \boldsymbol{\gamma}_j^{(s)} \\ \boldsymbol{\theta}_j^{(s)} \end{bmatrix} \right\}$;      // Newton step
        $\eta^2 = n^{-1} \left\{ \dot{\ell}_j(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}) - \mathbf{Q}_j(\boldsymbol{\mu}_j, \breve{\boldsymbol{\mu}}_j) \begin{bmatrix} \boldsymbol{\gamma}_j^{(s)} \\ \boldsymbol{\theta}_j^{(s)} \end{bmatrix} \right\}^\top \begin{bmatrix} \Delta\boldsymbol{\gamma}_j^{(s)} \\ \Delta\boldsymbol{\theta}_j^{(s)} \end{bmatrix}$;      // $\eta$:  Newton increment
        $\nu \leftarrow 1$;
        while $\ell_j^{(\mathrm{P})}(\boldsymbol{\gamma}_j^{(s)} + \nu\Delta\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)} + \nu\Delta\boldsymbol{\theta}_j^{(s)}; \boldsymbol{\mu}_j, \breve{\boldsymbol{\mu}}_j) < \ell_j^{(\mathrm{P})}(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}; \boldsymbol{\mu}_j, \breve{\boldsymbol{\mu}}_j) + n\phi\nu\eta^2$ do $\nu \leftarrow \psi\nu$;
        $\boldsymbol{\gamma}_j^{(s+1)} = \boldsymbol{\gamma}_j^{(s)} + \nu\Delta\boldsymbol{\gamma}_j^{(s)}$;
        $\boldsymbol{\theta}_j^{(s+1)} = \boldsymbol{\theta}_j^{(s)} + \nu\Delta\boldsymbol{\theta}_j^{(s)}$;
        $\xi_{s+1} = \delta\xi_s$;
        $s \leftarrow s + 1$;
    while $\eta^2 \geq 2\epsilon$;
end
```

Appendix C Proof of Proposition 1

Proposition 1. *Under* $H_0^{(t)} : \mathbf{C}^{(t)}\mathrm{vec}(\boldsymbol{\gamma}_{jl}^\top) = \mathbf{0}$, *the test statistic*

$$\{\mathrm{vec}(\tilde{\boldsymbol{\gamma}}_{jl}^\top) - \tilde{\mathbf{b}}_{jl}\}^\top \{\mathbf{C}^{(t)}\}^\top \left[\mathbf{C}^{(t)}\boldsymbol{\Omega}_{jl}\{\mathbf{C}^{(t)}\}^\top\right]^{-1} \mathbf{C}^{(t)}\{\mathrm{vec}(\tilde{\boldsymbol{\gamma}}_{jl}^\top) - \tilde{\mathbf{b}}_{jl}\}$$

asymptotically follows a distribution characterized by

$$\sum_{u=1}^{K\breve{K} \times K\breve{K}} \mu_u G_u^2,$$

where G_u*'s are independent standard normal random variables, and* μ_u*'s are the possibly identical eigenvalues of the matrix product of* $[\mathbf{C}^{(t)}\boldsymbol{\Omega}_{jl}\{\mathbf{C}^{(t)}\}^\top]^{-1}$ *and the variance of* $\mathbf{C}^{(t)}\{\mathrm{vec}(\tilde{\boldsymbol{\gamma}}_{jl}^\top) - \tilde{\mathbf{b}}_{jl}\}$*.*

Proof. Let $\mathbf{M} := [\mathbf{C}^{(t)}\boldsymbol{\Omega}_{jl}\{\mathbf{C}^{(t)}\}^\top]^{-1}$, let $\boldsymbol{\Sigma}$ denote the variance of $\mathbf{x} := \mathbf{C}^{(t)}\{\mathrm{vec}(\tilde{\boldsymbol{\gamma}}_{jl}^\top) - \tilde{\mathbf{b}}_{jl}\}$, and let $Q := \mathbf{x}^\top\mathbf{M}\mathbf{x}$ denote the Wald test statistic. Since $\boldsymbol{\Sigma}$ is orthogonally diagonalizable, there exists an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}\boldsymbol{\Sigma}\mathbf{P}^\top = \boldsymbol{\Psi}$, with $\boldsymbol{\Psi}$ being a diagonal matrix of positive eigenvalues of $\boldsymbol{\Sigma}$. Let $\mathbf{R} := \boldsymbol{\Psi}^{-1/2}\mathbf{P}$, a nonsingular matrix. Then $\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}^\top = \mathbf{I}$. Since $(\mathbf{R}^\top)^{-1}\mathbf{M}\mathbf{R}^{-1}$ is symmetric and orthogonally diagonalizable, there exists another orthogonal matrix $\mathbf{T}$ such that $\mathbf{T}(\mathbf{R}^\top)^{-1}\mathbf{M}\mathbf{R}^{-1}\mathbf{T}^\top = \boldsymbol{\Phi}$ is a diagonal matrix sharing the same eigenvalues $\mu_1, \ldots, \mu_{K\breve{K}\times K\breve{K}}$ as those of $(\mathbf{R}^\top)^{-1}\mathbf{M}\mathbf{R}^{-1}$. Let $\mathbf{z} := \mathbf{T}\mathbf{R}\mathbf{x}$. Then under the null $H_0^{(t)}$, we have $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Since $\mathbf{T}\mathbf{R}$ is nonsingular, $\mathbf{x} = \mathbf{R}^{-1}\mathbf{T}^\top\mathbf{z}$. It follows that $Q = \mathbf{z}^\top\boldsymbol{\Phi}\mathbf{z} = \sum_{u=1}^{K\breve{K}\times K\breve{K}} \mu_u G_u^2$, where G_u's independently follow the standard normal distribution. Observe that

$$(\mathbf{R}^\top\mathbf{T}^\top)^{-1}\mathbf{M}\boldsymbol{\Sigma}\mathbf{R}^\top\mathbf{T}^\top = \mathbf{T}(\mathbf{R}^\top)^{-1}\mathbf{M}\boldsymbol{\Sigma}\mathbf{R}^\top\mathbf{T}^\top = \mathbf{T}(\mathbf{R}^\top)^{-1}\mathbf{M}\mathbf{R}^{-1}\mathbf{T}^\top = \boldsymbol{\Phi}.$$

This implies that $\mathbf{M}\boldsymbol{\Sigma}$ and $\boldsymbol{\Phi}$ have the same set of eigenvalues (since the mapping $\mathbf{A} \mapsto \mathbf{B}^{-1}\mathbf{A}\mathbf{B}$ preserves eigenvalues). □

Appendix D Test of Significance

In addition to tests of coefficient variation, a Wald statistic for the test of significance can be derived. Following the notation in Section 2.2, one may consider the null hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta}_j(t, \breve{x}) = \mathbf{0}$, in which $\mathbf{C}$ is a given $c \times p$ contrast matrix. Note that H_0 can be rewritten as $[\mathbf{C} \otimes \breve{\mathbf{B}}^\top(\breve{x}) \otimes \mathbf{B}^\top(t)]\boldsymbol{\gamma}_j = \mathbf{0}$, where $\boldsymbol{\gamma}_j = \mathrm{vec}(\boldsymbol{\Gamma}_j^\top)$ with $\boldsymbol{\Gamma}_j = [\mathrm{vec}(\boldsymbol{\gamma}_{j1}^\top), \ldots, \mathrm{vec}(\boldsymbol{\gamma}_{jp}^\top)]^\top$. The test statistic is thus given by

$$\{\tilde{\boldsymbol{\gamma}}_j - \hat{\mathbf{b}}_j\}^\top[\mathbf{C}^\top \otimes \breve{\mathbf{B}}(\breve{x}) \otimes \mathbf{B}(t)]\{[\mathbf{C} \otimes \breve{\mathbf{B}}^\top(\breve{x}) \otimes \mathbf{B}^\top(t)]\boldsymbol{\Omega}_j[\mathbf{C}^\top \otimes \breve{\mathbf{B}}(\breve{x}) \otimes \mathbf{B}(t)]\}^{-1}[\mathbf{C} \otimes \breve{\mathbf{B}}^\top(\breve{x}) \otimes \mathbf{B}^\top(t)]\{\tilde{\boldsymbol{\gamma}}_j - \hat{\mathbf{b}}_j\},$$

where $\boldsymbol{\Omega}_j$ denotes an arbitrary $pK\breve{K} \times pK\breve{K}$ symmetric and positive-definite matrix, e.g., the top-left $pK\breve{K} \times pK\breve{K}$ block of $\widetilde{\mathbf{V}}_j^{\mathrm{S}}$ or $\widetilde{\mathbf{V}}_j^{\mathrm{M}}$, and $\hat{\mathbf{b}}_j$ is a subvector of $\tilde{\mathbf{b}}_j$ consisting of the first $pK\breve{K}$ rows of $\tilde{\mathbf{b}}_j$. The asymptotic distribution of the test statistic can be determined following Proposition 1. That is, the test statistic asymptotically follows a distribution characterized by

$$\sum_{u=1}^{c} \mu_u G_u^2,$$

where G_u's are independent standard normal random variables, and μ_u's are the possibly identical eigenvalues of the matrix product of $\{[\mathbf{C} \otimes \breve{\mathbf{B}}^\top(\breve{x}) \otimes \mathbf{B}^\top(t)]\boldsymbol{\Omega}_j[\mathbf{C}^\top \otimes \breve{\mathbf{B}}(\breve{x}) \otimes \mathbf{B}(t)]\}^{-1}$ and the variance of $[\mathbf{C} \otimes \breve{\mathbf{B}}^\top(\breve{x}) \otimes \mathbf{B}^\top(t)]\{\tilde{\boldsymbol{\gamma}}_j - \hat{\mathbf{b}}_j\}$.

Based on 1000 simulated data replicates, Web Figure 7 displays a scatter plot of the probability of rejecting the null hypothesis that $\beta_1(t, \breve{x}) = 0$ versus the true value of $\beta_1(t, \breve{x}) = \sin(3\pi t/4)\exp(-0.5\breve{x})$. As expected, the probability of rejection increases with $|\beta_1(t, \breve{x})|$. For $\beta_1(t, \breve{x}) = 0$, the type I error rate is mostly less than 0.05, while for $|\beta_1(t, \breve{x})| > 0.2$, the statistical power is generally above 0.8.

Appendix E Proof of Proposition 2

Proposition 2. *Let $\hat{\lambda}_{0jg}(\cdot)$ be the estimated baseline hazard function derived from the unpenalized bivariate varying coefficient model. Let*

$$\tilde{M}_{jgi} := \Delta_{jgi} - \exp(\mathbf{W}_{gi}^\top \tilde{\boldsymbol{\theta}}_j^{-f}) \int_0^{X_{gi}} \exp\left\{\mathbf{Z}_{gi}^\top \tilde{\boldsymbol{\beta}}_j^{-f}(t, \breve{X}_{gi})\right\} \hat{\lambda}_{0jg}(t)\,\mathrm{d}t$$

be the martingale residual for subject i in the gth stratum, where $\tilde{\boldsymbol{\beta}}_j^{-f}(\cdot,\cdot)$ and $\tilde{\boldsymbol{\theta}}_j^{-f}$ are the penalized estimates from the corresponding fold f to which subject i in the gth stratum belongs. Then the deviance residual for subject i in the gth stratum with respect to the jth failure type is written as

$$d_{jgi} := \operatorname{sign}(\tilde{M}_{jgi}) \sqrt{-2\left[\Delta_{jgi}\left\{\mathbf{Z}_{gi}^\top \tilde{\boldsymbol{\beta}}_j^{-f}(X_{gi}, \breve{X}_{gi}) + \mathbf{W}_{gi}^\top \tilde{\boldsymbol{\theta}}_j^{-f} + \log \int_0^{X_{gi}} \hat{\lambda}_{0jg}(t)\,\mathrm{d}t\right\} + \tilde{M}_{jgi}\right]}.$$

Proof. Given estimates $\hat{\boldsymbol{\theta}}_j$, $\hat{\boldsymbol{\beta}}_j(\cdot,\cdot)$ for the bivariate varying coefficient model (1), the martingale residuals can be defined as

$$\hat{M}_{jgi} := \hat{M}_{jgi}(\infty, \breve{X}_{gi}) = \Delta_{jgi} - \exp(\mathbf{W}_{gi}^\top \hat{\boldsymbol{\theta}}_j) \int_0^{X_{gi}} \exp\left\{\mathbf{Z}_{gi}^\top \hat{\boldsymbol{\beta}}_j(t, \breve{X}_{gi})\right\} \hat{\lambda}_{0jg}(t)\,\mathrm{d}t,$$

where the baseline hazard estimates $\hat{\lambda}_{0jg}(\cdot)$ are determined via the Breslow estimator. Further, the log-likelihood with respect to the jth failure type can be written as

$$\begin{aligned}
&\sum_{g=1}^{G}\sum_{i=1}^{n_g}\{\Delta_{jgi}\log\lambda_{jgi}(X_{gi} \mid \mathbf{Z}_{gi}, \mathbf{W}_{gi}, \breve{X}_{gi}) + \log S_{jgi}(X_{gi} \mid \mathbf{Z}_{gi}, \mathbf{W}_{gi}, \breve{X}_{gi})\} \\
&= \sum_{g=1}^{G}\sum_{i=1}^{n_g}\Big[\Delta_{jgi}\{\mathbf{Z}_{gi}^\top \boldsymbol{\beta}_j(X_{gi}, \breve{X}_{gi}) + \mathbf{W}_{gi}^\top \boldsymbol{\theta}_j + \log\lambda_{0jg}(X_{gi}) \\
&\qquad - \int_0^{X_{gi}} \exp\{\mathbf{Z}_{gi}^\top \boldsymbol{\beta}_j(t, \breve{X}_{gi}) + \mathbf{W}_{gi}^\top \boldsymbol{\theta}_j\}\lambda_{0jg}(t)\,\mathrm{d}t\Big],
\end{aligned}$$

where $S_{jgi}(t \mid \mathbf{Z}_{gi}, \mathbf{W}_{gi}, \breve{X}_{gi})$ is the corresponding survivor function. Assuming that the baseline hazard $\lambda_{0jg}(\cdot)$ is known, we have the deviance D written as

$$
\begin{aligned}
D = 2 \sup_{\boldsymbol{\beta}_{jgi}, \boldsymbol{\theta}_{jgi}} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \Bigg\{ & \Delta_{jgi} [\mathbf{Z}_{gi}^{\top} \{ \boldsymbol{\beta}_{jgi} - \hat{\boldsymbol{\beta}}_j(X_{gi}, \breve{X}_{gi}) \} + \mathbf{W}_{gi}^{\top} (\boldsymbol{\theta}_{jgi} - \hat{\boldsymbol{\theta}}_j)] \\
& - \int_0^{X_{gi}} \Big[\exp(\mathbf{Z}_{gi}^{\top} \boldsymbol{\beta}_{jgi} + \mathbf{W}_{gi}^{\top} \boldsymbol{\theta}_{jgi}) - \exp\{ \mathbf{Z}_{gi}^{\top} \hat{\boldsymbol{\beta}}_j(t, \breve{X}_{gi}) + \mathbf{W}_{gi}^{\top} \hat{\boldsymbol{\theta}}_j \} \Big] \lambda_{0jg}(t) \, \mathrm{d}t \Bigg\},
\end{aligned}
$$

where $\boldsymbol{\beta}_{jgi}$ and $\boldsymbol{\theta}_{jgi}$ are subject-cause-specific estimates allowed in a saturated model. Now, we have the first order condition

$$
\Delta_{jgi} = \exp(\mathbf{Z}_{gi}^{\top} \boldsymbol{\beta}_{jgi} + \mathbf{W}_{gi}^{\top} \boldsymbol{\theta}_{jgi}) \int_0^{X_{gi}} \lambda_{0jg}(t) \, \mathrm{d}t, \quad g = 1, \ldots, G, \; i = 1, \ldots, n_g.
$$

With this condition, the deviance D reduces to

$$
\begin{aligned}
D &= -2 \sum_{g=1}^{G} \sum_{i=1}^{n_g} \left\{ \Delta_{jgi} \log \frac{\exp\{ \mathbf{Z}_{gi}^{\top} \hat{\boldsymbol{\beta}}_j(X_{gi}, \breve{X}_{gi}) + \mathbf{W}_{gi}^{\top} \hat{\boldsymbol{\theta}}_j \} \int_0^{X_{gi}} \lambda_{0jg}(t) \, \mathrm{d}t}{\Delta_{jgi}} + \tilde{M}_{jgi} \right\} \\
&= -2 \sum_{g=1}^{G} \sum_{i=1}^{n_g} \left[\Delta_{jgi} \left\{ \mathbf{Z}_{gi}^{\top} \hat{\boldsymbol{\beta}}_j(X_{gi}, \breve{X}_{gi}) + \mathbf{W}_{gi}^{\top} \hat{\boldsymbol{\theta}}_j + \log \int_0^{X_{gi}} \lambda_{0jg}(t) \, \mathrm{d}t \right\} + \tilde{M}_{jgi} \right],
\end{aligned}
$$

where

$$
\tilde{M}_{jgi} := \tilde{M}_{jgi}(\infty, \breve{X}_{gi}) = \Delta_{jgi} - \exp(\mathbf{W}_{gi}^{\top} \hat{\boldsymbol{\theta}}_j) \int_0^{X_{gi}} \exp\left\{ \mathbf{Z}_{gi}^{\top} \hat{\boldsymbol{\beta}}_j(t, \breve{X}_{gi}) \right\} \lambda_{0jg}(t) \, \mathrm{d}t
$$

is the martingale residual with known baseline hazard $\lambda_{0jg}(\cdot)$. Then the deviance residual d_{jgi} for subject i in the gth stratum with respect to the jth failure type can be written as

$$
d_{jgi} = \mathrm{sign}(\hat{M}_{jgi}) \sqrt{-2 \left[\Delta_{jgi} \left\{ \mathbf{Z}_{gi}^{\top} \hat{\boldsymbol{\beta}}_j(X_{gi}, \breve{X}_{gi}) + \mathbf{W}_{gi}^{\top} \hat{\boldsymbol{\theta}}_j + \log \int_0^{X_{gi}} \hat{\lambda}_{0jg}(t) \, \mathrm{d}t \right\} + \hat{M}_{jgi} \right]},
$$

where $\hat{M}_{jgi}$ is the martingale residual $\tilde{M}_{jgi}$ with $\lambda_{0jg}(\cdot)$ replaced by $\hat{\lambda}_{0jg}(\cdot)$. □

Appendix F Alternative Cross-Validation Methods

F.1 Fold-constrained (FC) cross-validated partial likelihood

In this approach, the cross-validation error (CVE) is proportional to the sum of fold-specific log-partial likelihood functions in which risk sets are constrained by the corresponding folds, i.e.,

$$
\mathrm{CVE}_j := -2 \sum_{f=1}^{F} \ell_j^f (\tilde{\boldsymbol{\eta}}_j^{-f}).
$$

F.2 Complementary fold-constrained (CFC) cross-validated partial likelihood

As the name suggests, the CVE is proportional to the sum of complementary fold-constrained log-partial likelihood functions, i.e.,

$$\mathrm{CVE}_j \coloneqq -2 \sum_{f=1}^{F} \{\ell_j(\tilde{\boldsymbol{\eta}}_j^{-f}) - \ell_j^{-f}(\tilde{\boldsymbol{\eta}}_j^{-f})\}.$$

This approach was applied in Verweij and Van Houwelingen (1993) and Simon et al. (2011).

F.3 Unconstrained (UC) cross-validated partial likelihood

First introduced by Breheny and Huang (2011) as cross-validated linear predictors, this approach features risk set construction unconstrained by folds in that fold-specific estimates $\tilde{\boldsymbol{\eta}}_j^{-f}$'s are assigned to all units of the sample according to their fold identities. With a slight abuse of notation, the CVE is written as

$$\mathrm{CVE}_j \coloneqq -2\ell_j(\tilde{\boldsymbol{\eta}}_j^{-1}, \ldots, \tilde{\boldsymbol{\eta}}_j^{-F}),$$

where $\tilde{\boldsymbol{\eta}}_j^{-f}$ is assigned to observations of fold f.

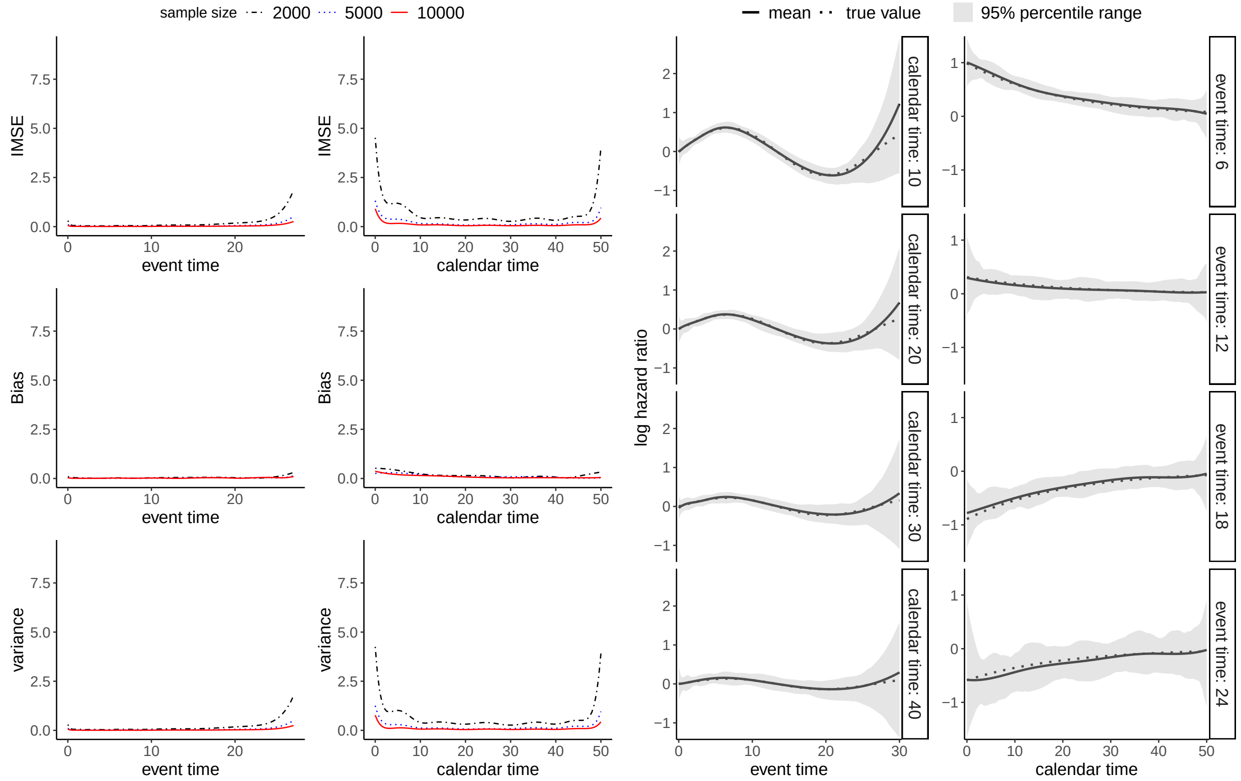
F.4 Generalized cross-validation (GCV)

Extending the approach of Yan and Huang (2012) to this setting with bivariate varying coefficients, we can write the CVE for the jth failure type as

$$\mathrm{CVE}_j = -\frac{\ell_j(\boldsymbol{\eta}_j)}{n(1 - f_j(\boldsymbol{\mu}_j, \breve{\boldsymbol{\mu}}_j)/n)^2},$$

where $f_j(\boldsymbol{\mu}_j, \breve{\boldsymbol{\mu}}_j) \coloneqq \operatorname{trace}\left(\{\ddot{\ell}_j^{(\mathrm{P})}(\boldsymbol{\eta}_j; \boldsymbol{\mu}_j, \breve{\boldsymbol{\mu}}_j)\}^{-1}\ddot{\ell}_j(\boldsymbol{\eta}_j)\right)$, i.e., the number of effective parameters (Yan and Huang, 2012), or the "degrees of freedom" of the model (Gray, 1992).

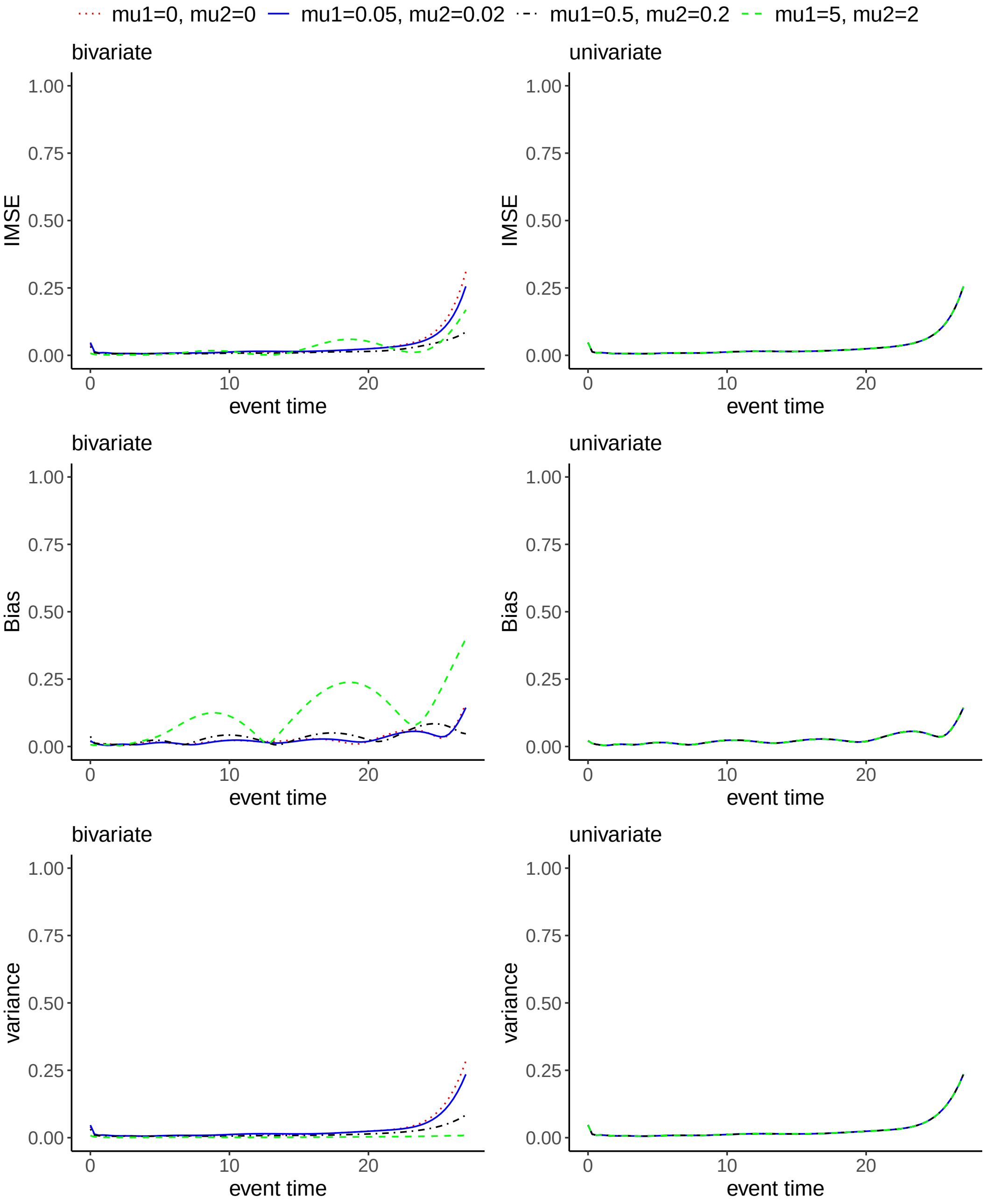
Appendix G Supplementary Figures



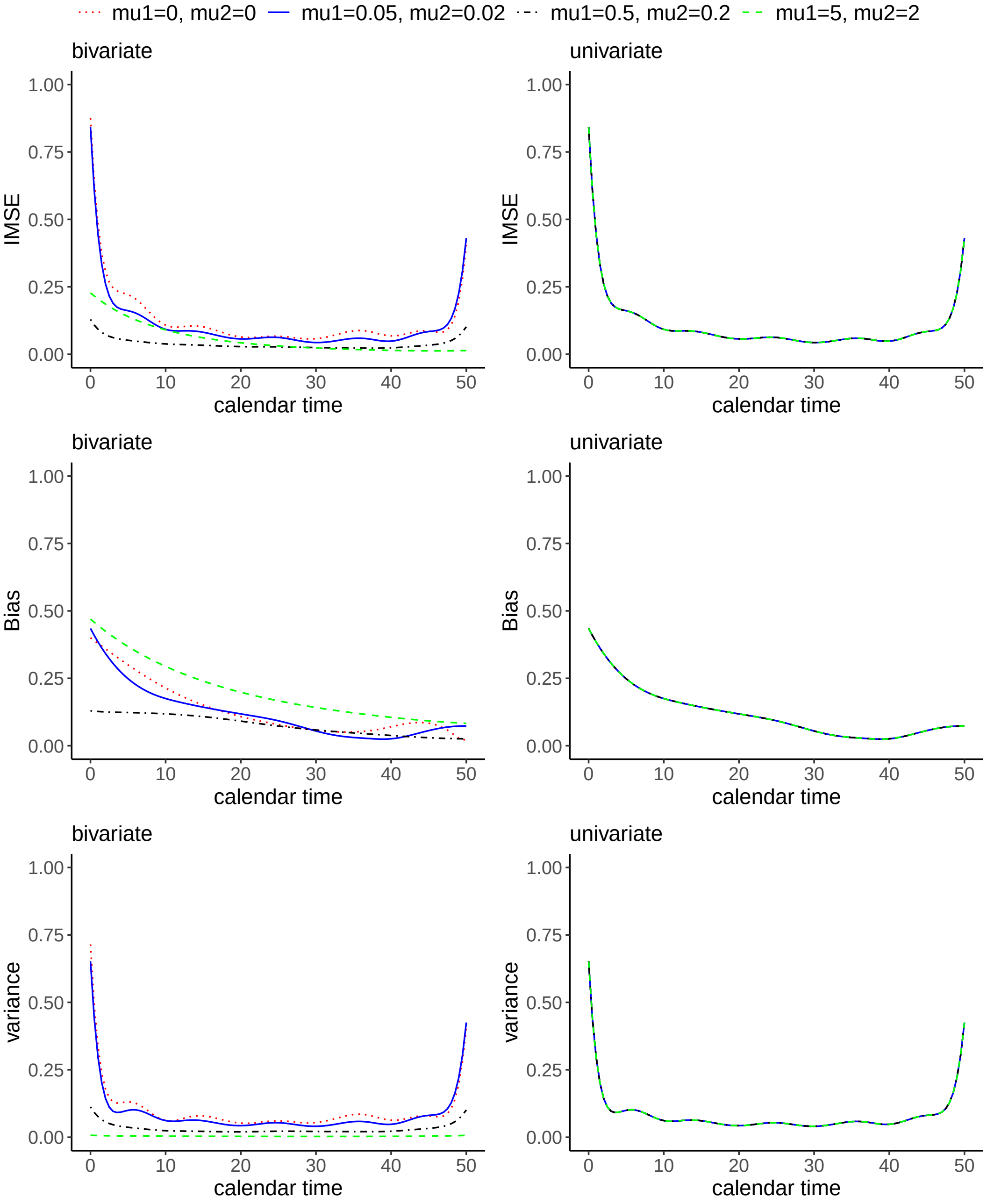
(a) IMSE, bias, and variance

(b) Mean versus true value

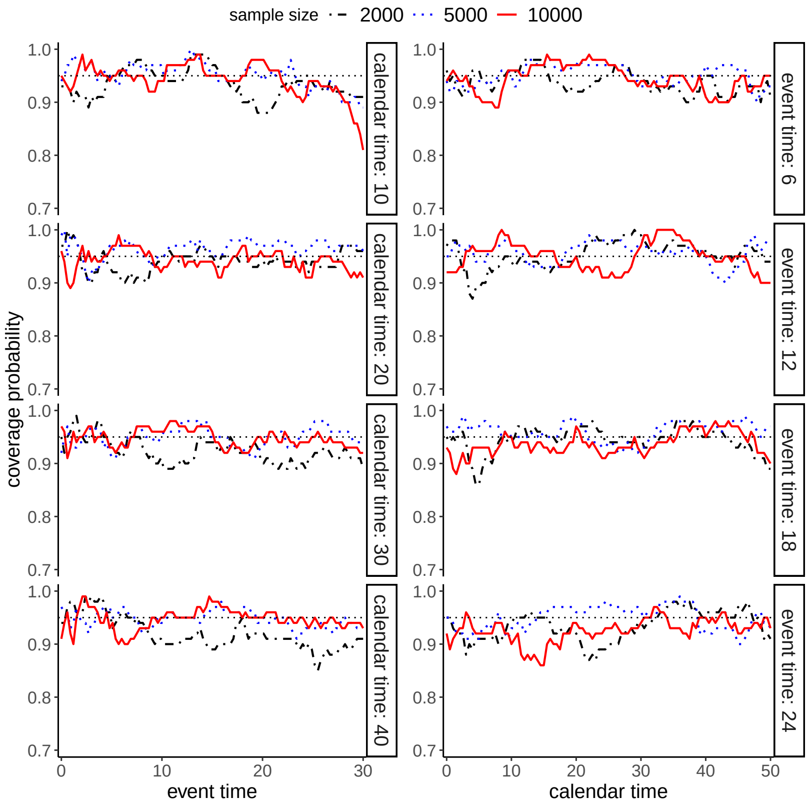
Supplementary Figure 1: (a) Integrated mean squared error (IMSE), average bias, and average variance of the estimated surface $\hat{\beta}_1(t, \breve{x})$ with varied sample sizes on event and calendar timescales. In each scenario, 100 data replicates were generated. On both timescales, $K = \breve{K} = 7$ cubic $(d = \breve{d} = 3)$ B-spline functions form a basis. True values are $\beta_1(t, \breve{x}) = \sin(3\pi t/4)\exp(-0.5\breve{x})$ and $\beta_2 = 1$. (b) Mean and 95% percentile range (2.5th and 97.5th percentiles as lower and upper limits) of pointwise estimates of $\beta_1(t, \breve{x})$ at selected event times and calendar times. In each scenario, 100 data replicates were generated with sample size equal to 10,000. On both timescales, $K = \breve{K} = 7$ cubic $(d = \breve{d} = 3)$ B-spline functions form a basis. True values are $\beta_1(t, \breve{x}) = \sin(3\pi t/4)\exp(-0.5\breve{x})$ and $\beta_2 = 1$. An unpenalized approach was used in (a) and (b).



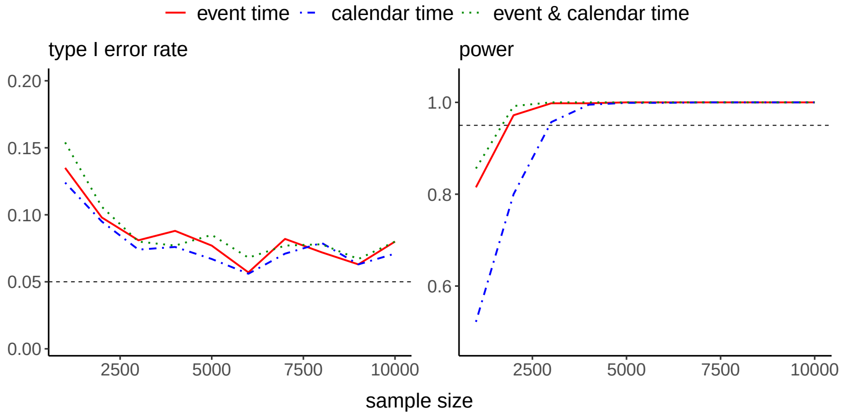
Supplementary Figure 2: Integrated mean squared error (IMSE), average bias, and average variance of the estimated surface $\hat{\beta}_1(t, \breve{x})$ across event time with sample size fixed at 10,000. In each scenario, 100 data replicates were generated. On both timescales, $K = \breve{K} = 7$ cubic ($d = \breve{d} = 3$) B-spline functions form a basis. True values are $\beta_1(t, \breve{x}) = \sin(3\pi t/4) \exp(-0.5\breve{x})$ and $\beta_2 = 1$. Various levels of penalization were introduced to $\beta_1(\cdot, \cdot)$, where `mu1` and `mu2` denote tuning parameters for calendar and event time, respectively, as in (4) of the manuscript. Both the bivariate varying coefficient model and the univariate time-varying coefficient model (Wu et al., 2022) were considered.



Supplementary Figure 3: Integrated mean squared error (IMSE), average bias, and average variance of the estimated surface $\hat{\beta}_1(t, \breve{x})$ across calendar time with sample size fixed at 10,000. In each scenario, 100 data replicates were generated. On both timescales, $K = \breve{K} = 7$ cubic ($d = \breve{d} = 3$) B-spline functions form a basis. True values are $\beta_1(t, \breve{x}) = \sin(3\pi t/4)\exp(-0.5\breve{x})$ and $\beta_2 = 1$. Various levels of penalization were introduced to $\beta_1(\cdot, \cdot)$, where `mu1` and `mu2` denote tuning parameters for calendar and event time, respectively, as in (4) of the manuscript. Both the bivariate varying coefficient model and the univariate time-varying coefficient model (Wu et al., 2022) were considered. In the latter model, the estimated coefficient was constant across calendar time.

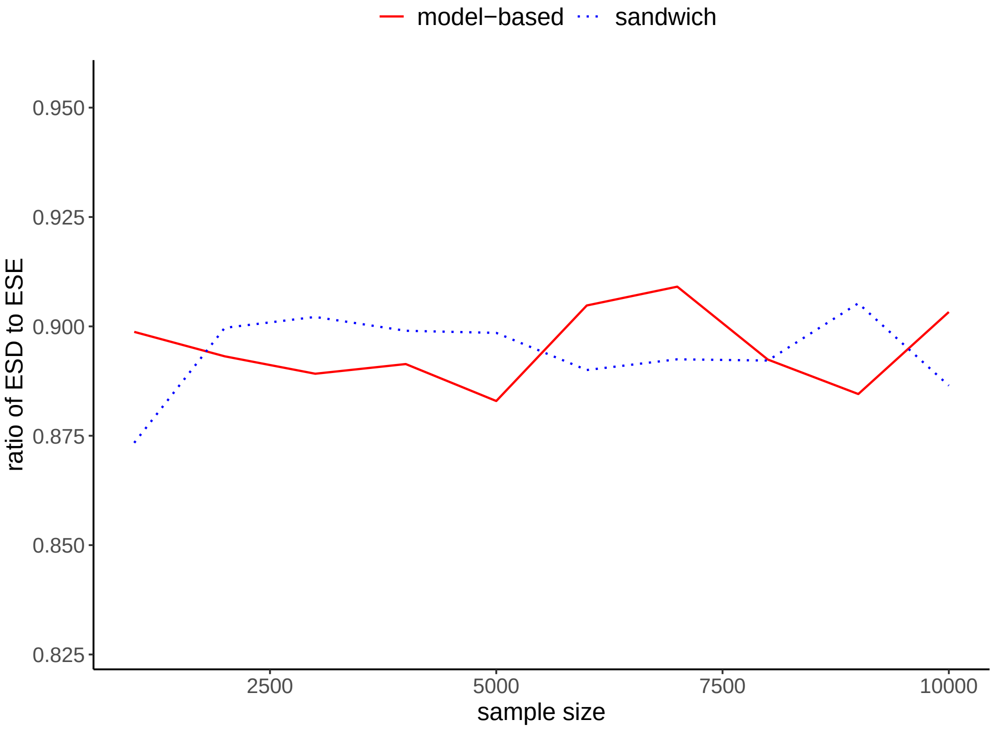


(a) Coverage probability

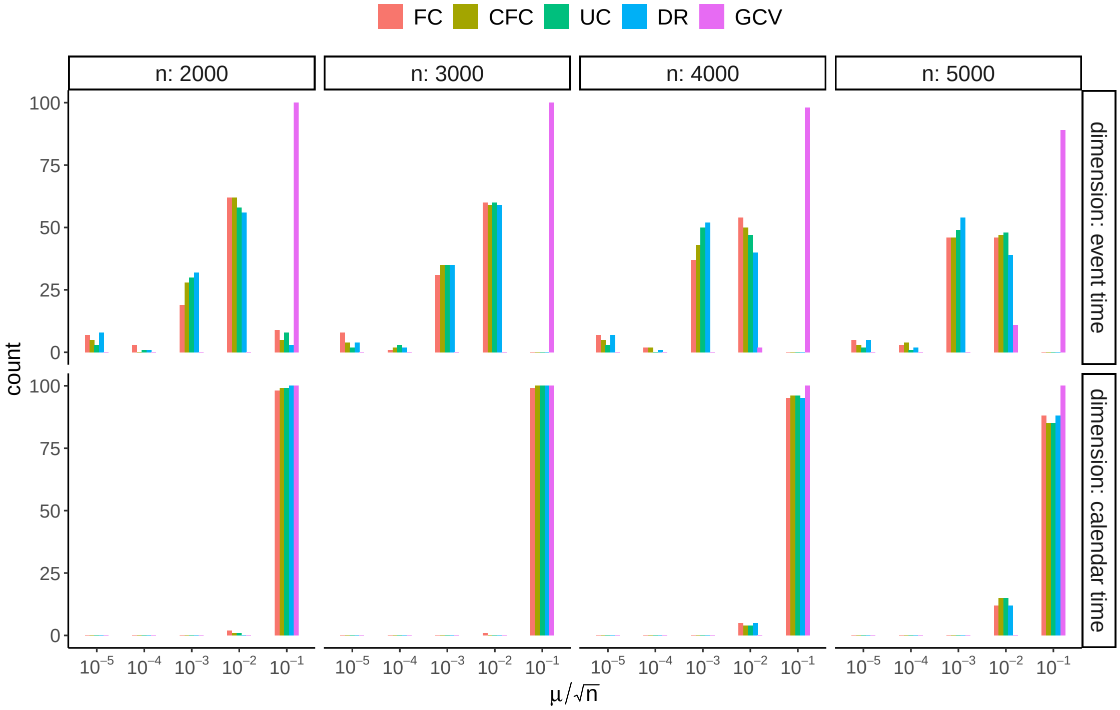


(b) Type I error rate and power

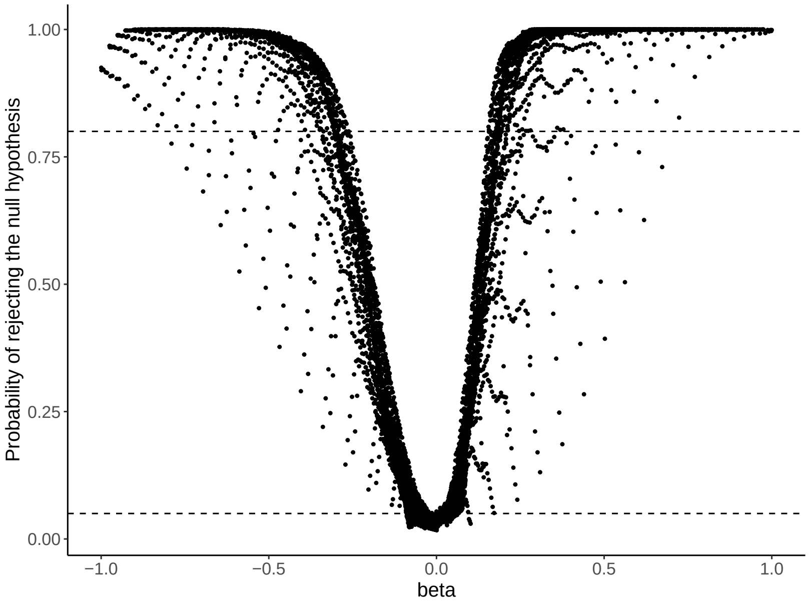
Supplementary Figure 4: (a) Coverage probability curves of $\beta_1(t, \breve{x})$ via pointwise 95% confidence intervals on event and calendar time scales, with varied sample sizes. In each scenario, 100 data replicates were generated with sample size equal to 10,000. On both timescales, $K = \breve{K} = 7$ cubic ($d = \breve{d} = 3$) B-spline functions form a basis. True values are $\beta_1(t, \breve{x}) = \sin(3\pi t/4)\exp(-0.5\breve{x})$ and $\beta_2 = 1$. (b) Type I error rate and power curves for tests of univariate and bivariate variation with varied sample sizes. In each scenario, 1,000 data replicates were generated. On both timescales, $K = \breve{K} = 7$ cubic ($d = \breve{d} = 3$) B-spline functions form a basis. True values are $\beta_1(t, \breve{x}) = 1$ and $\beta_2 = 1$ in the left panel, and $\beta_1(t, \breve{x}) = \sin(3\pi t/4)\exp(-0.5\breve{x})$ and $\beta_2 = 1$ in the right panel. An unpenalized approach was used in (a) and (b).



Supplementary Figure 5: The ratio of the empirical standard deviation to the estimated standard error for $\beta_1(t, \breve{x})$ averaged across grid points. In each scenario, 1,000 data replicates were generated. True values are $\beta_1(t, \breve{x}) = 0.5$ and $\beta_2 = 1$. A sandwich and a model-based variance estimator were used. Throughout all experiments, 7 cubic B-splines form a basis on both timescales, and tuning parameters vary with sample size, i.e., $\mu = n^{1/8}/500$ and $\breve{\mu} = n^{1/8}/200$.



Supplementary Figure 6: A comparison of the distribution of selected tuning parameters for five cross-validation methods: fold-constrained (FC), complementary fold-constrained (CFC), and fold-unconstrained (UC) cross-validated partial likelihood, cross-validated deviance residuals (DR), and generalized cross-validation (GCV). In each scenario, 100 training and validation data replicates were generated independently. A 5-by-5 grid of tuning parameters was formed such that $\mu/\sqrt{n}$ (with n denoting sample size) and $\breve{\mu}/\sqrt{n}$ varied from 10^{-5} to 10^{-1}. Each cross-validation method was applied to a training data replicate to determine the optimal tuning parameters. True values were $\beta_1(t,\breve{x}) = \sin(3\pi t/4)\exp(-0.5\breve{x})$ and $\beta_2(t,\breve{x}) = 1$.



Supplementary Figure 7: A scatter plot of the probability of rejecting the null hypothesis that $\beta_1(t, \breve{x}) = 0$ versus the true value of $\beta_1(t, \breve{x}) = \sin(3\pi t/4)\exp(-0.5\breve{x})$. In the experiment, 1000 data replicates were generated with sample size $n = 10,000$ and $\beta_2 = 1$, and 7 cubic B-splines were used to form a basis on both timescales. A sandwich variance estimator was used with test statistics approximately following a chi-squared distribution. Tuning parameters were set as `mu1`=0.5 and `mu1`=0.2. Two dashed horizontal lines correspond to 0.05 and 0.8, respectively.

Appendix H Supplementary Table

References

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, 5(1):232–253.

Davies, R. B. (1980). Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(3):323–333.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87(420):942–951.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13.

Verweij, P. J. and Van Houwelingen, H. C. (1993). Cross-validation in survival analysis. *Statistics in Medicine*, 12(24):2305–2314.

Table 1: Type I error rate and power for tests of univariate and bivariate variation with different test statistics and varied sample sizes. In each scenario, 1,000 data replicates were generated. True values are $\beta_1(t, \breve{x}) = \sin(3\pi t/4)$ and $\beta_2 = 1$ in Panel A, and $\beta_1(t, \breve{x}) = \exp(-0.5\breve{x})$ and $\beta_2 = 1$ in Panel B. In the first and second columns of each sub-panel, a sandwich and a model-based variance estimator were used with test statistics approximately following a chi-squared distribution. In the third column of each sub-panel, the test statistic in Gray (1992) was compared with a distribution of a linear combination of chi-squared random variables (Davies, 1980). We used 7 cubic B-splines to form a basis on both timescales, and tuning parameters vary with sample size, i.e., $\mu = n^{1/8}/500$ and $\breve{\mu} = n^{1/8}/200$. SC, sandwich estimator with chi-squared distribution; MC, model-based estimator with chi-squared distribution; MD, model-based estimator with a distribution of a linear combination of chi-squared random variables.

Panel A: $\beta_1(t, \breve{x}) = \sin(3\pi t/4)$

sample size	power (event time)			type I error rate (calendar time)			power (event & calendar time)		
	SC	MC	MD	SC	MC	MD	SC	MC	MD
1000	1	1	1	0.006	0.006	0.010	1	1	0.999
2000	1	1	1	0.006	0.008	0.004	1	1	1
3000	1	1	1	0.003	0.002	0.009	1	1	1
4000	1	1	1	0.004	0.002	0.004	1	1	1
5000	1	1	1	0.007	0.002	0.003	1	1	1
6000	1	1	1	0.003	0.003	0.002	1	1	1
7000	1	1	1	0.002	0.001	0.003	1	1	1
8000	1	1	1	0.006	0.001	0.002	1	1	1
9000	1	1	1	0.001	0.001	0.003	1	1	1
10000	1	1	1	0.002	0.003	0.006	1	1	1

Panel B: $\beta_1(t, \breve{x}) = \exp(-0.5\breve{x})$

sample size	type I error rate (event time)			power (calendar time)			power (event & calendar time)		
	SC	MC	MD	SC	MC	MD	SC	MC	MD
1000	0.008	0.005	0.003	0.602	0.58	0.59	0.596	0.607	0.596
2000	0.005	0.009	0.005	0.974	0.97	0.973	0.970	0.973	0.973
3000	0.001	0.003	0.003	1	0.999	1	1	0.997	1
4000	0.004	0.004	0.003	1	1	1	1	1	1
5000	0.004	0.002	0.002	1	1	1	1	1	1
6000	0.004	0.003	0.001	1	1	1	1	1	1
7000	0.006	0.004	0.002	1	1	1	1	1	1
8000	0.001	0.003	0.004	1	1	1	1	1	1
9000	0.002	0.004	0.002	1	1	1	1	1	1
10000	0.002	0.002	0.005	1	1	1	1	1	1

Wu, W., Taylor, J. M., Brouwer, A. F., Luo, L., Kang, J., Jiang, H., and He, K. (2022). Scalable proximal methods for cause-specific hazard modeling with time-varying coefficients. *Lifetime Data Analysis*, 28(2):194–218.

Yan, J. and Huang, J. (2012). Model selection for Cox models with time-varying coefficients. *Biometrics*, 68(2):419–428.