Statistica Sinica **30** (2020), 1605-1632
doi:https://doi.org/10.5705/ss.202017.0375

# TIME-VARYING HAZARDS MODEL FOR INCORPORATING IRREGULARLY MEASURED HIGH-DIMENSIONAL BIOMARKERS

Xiang Li[1], Quefeng Li[2], Donglin Zeng[2],
Karen Marder[1], Jane Paulsen[3] and Yuanjia Wang[1]

[1] *Columbia University,* [2] *University of North Carolina, Chapel Hill*
*and* [3] *University of Iowa*

*Abstract:* Clinical studies with time-to-event outcomes often collect measurements of a large number of time-varying covariates over time (e.g., clinical assessments or neuroimaging biomarkers) in order to build a time-sensitive prognostic model. However, resource-intensive or invasive (e.g., lumbar puncture) data-collection processes mean that biomarkers may be measured infrequently and, thus, not be available at every observed event time point. Therefore, leveraging all available time-varying biomarkers is important to improving our models event occurrence. We propose a kernel smoothing-based approach that borrows information across subjects to remedy the problem of infrequent and unbalanced biomarker measurements under a time-varying hazards model. A penalized pseudo-likelihood function is proposed for estimation, and an efficient augmented penalization minimization algorithm related to the alternating direction method of multipliers is adopted for computation. Given several regularity conditions, used to control the approximation bias and stochastic variability, we show that even in the presence of ultrahigh dimensionality, the proposed method selects important biomarkers with high probability. We use simulation studies to show that our method outperforms existing methods in terms of estimation and selection performance. Finally, we apply the proposed method to real data to model time-to-disease conversion using longitudinal, whole-brain structural magnetic resonance imaging biomarkers. The results show substantial improvement in performance over that of current standards, including using baseline measures only.

*Key words and phrases:* Biomarker studies, high-dimensional covariates, irregular measurements, kernel-weighted estimation, neurological disorders, time-varying hazards model.

## 1. Introduction

Time-varying biomarkers are often collected in studies on disease mechanisms and when building time-varying prognostic models for time-to-event outcomes, such as disease onset. Technological advancements have made repeated measurements of high-dimensional time-varying biomarkers possible at the individual level. However, while these are potentially useful in terms of improving the power of predictions, several statistical and computational challenges have emerged. First, collecting certain biomarkers may be resource-intensive or invasive (e.g., neuroimaging measures involving radiation exposure). As a result, such measurements tend to be infrequent and irregularly spaced over time for each subject. Thus, the biomarkers as covariates may not be available at every observed event time point. Second, an extensive body of literature (Desikan et al. (2006); Chen et al. (2014); Paulsen et al. (2014a); Ryan et al. (2015)) suggests that biomarker effects on neurological disorders vary with age, time, and an individual's disease progression. For example, a large natural history study of a neurological disorder (Paulsen et al. (2014a)), Huntington's disease (HD), showed that regional brain atrophy measures, considered important HD biomarkers, manifest differential rates of decline at distinct disease stages. Then, a work on neurobiological processes revealed an age-dependent effect pattern for brain activation, as measured by neuroimaging biomarkers (Ryan et al. (2015)). Third, biomarkers often exhibit a biological network structure (e.g., structural covariation network, He, Chen and Evans (2008); structural and functional brain networks, Bullmore and Bassett (2011); gene co-expression network, Stuart et al. (2003)), where linked biomarkers may indicate a similar likelihood of disease diagnosis or prognosis, owing to their sharing of disease pathways. Given these challenges, a valid statistical method that includes the biomarkers for prediction should take into account all available non-frequent biomarker measurements and time-varying effects as well as the high dimensionality and informative network structure among covariates.

The methods commonly used to associate time-dependent biomarkers with hazards of time-to-event outcomes require that biomarkers be completely observed at each event time (Honda and Härdle (2014); Honda and Yabe (2017)). As a result, they cannot be applied when the biomarkers are measured infrequently and irregularly. An easy solution is to treat this as a missing covariates problem, which means we can impute missing biomarkers at an event time using the last value carried forward (LVCF; Andersen and Liestol (2003)). However, while straightforward to implement, this approach may induce bias and lead to incorrect inferences, especially when the biomarkers show substantial change (Prentice (1982); Tsiatis and Davidian (2001)). Several sophisticated approaches (Tsiatis and Davidian (2004); Gould et al. (2014); Taylor et al. (2013)) have been proposed as alternatives, that link group-average (rather than individual) biomarker trajectories with time-to-event outcome models. However, the predictive performance of the models obtained from these approaches may not reflect the true predictivity of the biomarker measures collected on each individual. Other methods link unobserved random effects to time-to-event models using a measurement error model and joint modeling (e.g., Rizopoulos (2011)). However, measurement error models are not practically relevant when biomarker variation is due to true biological variability rather than random errors. Thus, we need to be able to directly associate observed biomarker values on each individual (rather than group-average or random effects) with the event times. Recent works along this line include those of Cao, Zeng and Fine (2014); Cao et al. (2015), although they do not deal with time-varying effects or with high-dimensional biomarkers.

In addition, the estimation of high-dimensional, time-dependent effect profile functions for biomarkers on event outcomes using limited data is an ambitious goal. Thus, incorporating biological information is crucial to reducing model space complexity and stabilizing the estimation. The following strong biological evidence has been observed for neurodegenerative disorders: (1) signals are clustered in networks (He, Chen and Evans (2008); Eidelberg and Surmeier (2011); Parikshak, Gandal and Geschwind (2015)); (2) biomarker signals evolve with disease progression and/or age (Desikan et al. (2006); Chen et al. (2014); Paulsen et al. (2014a)); and (3) signals are expected to be sparse (e.g., Liu et al. (2014)). Existing methods used to select functions for hazards models (e.g., Yan and Huang (2012); Liu et al. (2013)) or transformation models (Liu and Zeng (2013)) do not simultaneously handle irregularly measured time-dependent covariates and incorporate biological information. In our motivating study and in other applications, subjects' biomarker assessments were scheduled less frequently scheduled than clinical visits and, thus, did not necessarily coincide, rendering the aforementioned works nonapplicable.

In this article, we propose a unified method for estimating the time-varying effects in a hazards model using high-dimensional time-varying biomarkers that are measured irregularly. Our first contribution is that we resolve the complication of unavailable biomarkers at some event times without excluding subjects with missing biomarkers. To this end, we adopt local kernel smoothing to pool observations across event times and subjects. Second, to facilitate the selection of the entire profile function, after approximating each function using B-splines, we incorporate a group sparsity penalty on the spline coefficients, inspired by the work of Huang, Horowitz and Wei (2010). Furthermore, to incorporate the biological network structure among the biomarkers, we include an additional regularization to encourage strongly linked biomarkers to yield similar prognosis effects. It is challenging to control the selection and estimation accuracy when dealing with high-dimensional functions, because estimation noise at any given time point may cause a biomarker to enter the model. Thus, our third contribution is to propose an efficient computational algorithm that achieves $\ell_0$-penalty-like sparsity of the functions by modifying popular augmented penalization methods, including the alternating direction method of multipliers (ADMM; Boyd et al. (2011)) algorithm. Fourth, our examination of the theoretical properties includes establishing a high-dimensional oracle selection for functions (instead of scalar parameters) in the presence of the kernel approximation. This requires techniques to appropriately control for approximation bias and stochastic variability, which are not available in existing theories.

The remainder of the paper is organized as follows. In Section 2, we describe a time-varying hazards model with time-varying biomarkers, an approximated likelihood function used to borrow information, and an efficient algorithm for implementation. In Section 3, we provide theories showing that, under a general class of penalty functions, our method admits the "oracle property" in terms of selecting and estimating the true time-dependent effects. In Section 4, we extend the proposed algorithm to incorporate the biological network structure in the model regularization. In Section 5, we present simulation studies that examine the finite-sample performance of the proposed method, demonstrating that it outperforms alternative approaches. In Section 6, we apply our method to data from a study (Paulsen et al. (2014b)) that uses whole-brain structural magnetic resonance imaging (MRI) data to estimate a network-regularized biomarker signature in order to predict the time-to-onset of HD. We show that the predictive performance of the proposed model is significantly better than that of the LVCF, baseline-only analyses, and current standards developed independently in the recent literature (Long et al. (2016)). Finally, we conclude the paper in Section 7.

## 2. Methodologies

### 2.1. Model and estimation

Assume that data are collected from $n$ independent and identically distributed (i.i.d.) subjects. For subject $i$, let $\boldsymbol{X}_i(t)$ denote a $p_n$-dimensional vector of covariates including time-dependent biomarkers, and let $T_i$ denote the time-to-event of interest (e.g., age-at-onset of a disease). To model the time-varying effects of covariates on the event, we propose a time-varying hazards model in which we assume that the conditional hazard rate function of $T_i = t$, given the covariate history by time $t$, is

$$\lambda(t|\boldsymbol{X}_i(s), s \leq t) = \lambda_0(t) \exp\left\{\boldsymbol{\beta}^T(t)\boldsymbol{X}_i(t)\right\}, \tag{2.1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, and $\boldsymbol{\beta}(t)$ is a vector of covariate effects at time $t$. Note that (2.1) can also include additional components of the covariate history, for instance, lagging effects, by expanding $\boldsymbol{X}_i(t)$ to include the covariate history.

Assume that $\boldsymbol{X}_i(t)$ is only measured at $n_i$ discrete time points: $t_{i1}, \ldots, t_{in_i}$. The observed data consist of $\Big\{\widetilde{T}_i = \min(T_i, C_i), \Delta_i = I(T_i \leq C_i), \boldsymbol{X}_i(t_{i1}), \ldots, \boldsymbol{X}_i(t_{in_i})\Big\}$, for $i = 1, \ldots, n$, where $C_i$ denotes the censoring time, assumed to be conditionally independent of $T_i$ given $\boldsymbol{X}_i(t)$, $\widetilde{T}_i$ is the observed event time or censoring time, and $\Delta_i$ the censoring indicator. Furthermore, let $Y_i(t) = I(\widetilde{T}_i \geq t)$ and $N_i(t) = I(\widetilde{T}_i \leq t, \Delta_i = 1)$ denote the at-risk process and the observed counting process, respectively.

If a complete history of the covariate processes $\boldsymbol{X}_i(t)$ is available for all $t < \widetilde{T}_i$ and all $i$, then the classic log-partial likelihood (Fleming and Harrington (2011)), defined as

$$n^{-1}\sum_{i=1}^{n}\int_0^{\tau}\left(\boldsymbol{\beta}^T(t)\boldsymbol{X}_i(t) - \log\left[n^{-1}\sum_{j=1}^{n}Y_j(t)\exp\{\boldsymbol{\beta}^T(t)\boldsymbol{X}_j(t)\}\right]\right)\mathrm{d}N_i(t), \tag{2.2}$$

can be maximized to estimate the coefficients, where $\tau$ is the duration of a study. Because $\boldsymbol{X}_i(t)$ is only observable at some distinct time points, we need to approximate each term in the above log-partial likelihood function using the observed data. Note that the objective function (2.2) relies on some empirical average of the functionals of $\boldsymbol{X}_i(t)$. Thus, the approximation does not need to be accurate for each subject's $\boldsymbol{X}_i(t)$; instead, an accurate approximation of this empirical average $\boldsymbol{X}_i(t)$ is sufficient. This motivates us to adopt kernel smoothing by pooling observations from the same trajectory on subject $i$, but also from the other subjects when approximating $\boldsymbol{X}_i(t)$. Specifically, consider the kernel smoothing proposed in Andersen and Liestol (2003), and weight the subjects in the pooled data, where the weights are based on the distance between the observed measurement times and $t$; the resulting approximated objective function is given as

$$
\begin{aligned}
l_n^s(\boldsymbol{\beta}) =& n^{-1}\sum_{i=1}^{n}\int_0^{\tau}\sum_{v=1}^{n_i} K_{h_n}(t-t_{iv})\Bigg(\boldsymbol{\beta}^T(t)\boldsymbol{X}_i(t_{iv}) \\
&-\log\left[n^{-1}\sum_{j=1}^{n}\sum_{k=1}^{n_j} K_{h_n}(t-t_{jk})Y_j(t)\exp\{\boldsymbol{\beta}^T(t)\boldsymbol{X}_j(t_{jk})\}\right]\Bigg)\mathrm{d}N_i(t),
\end{aligned}
$$

where $K_{h_n}(t) = h_n^{-1}K(t/h_n)$ for some symmetric kernel function $K(\cdot)$ and bandwidth $h_n$.

Because $\boldsymbol{\beta}(t)$ is fully nonparametric, maximization is not feasible. Thus, we use a B-spline approximation for each $\beta_j(t)$, $j = 1,\ldots,p_n$. Specifically, let $\phi_{1,m}(t),\ldots,\phi_{q_n,m}(t)$ be the B-spline basis functions of order $m$ associated with $(q_n - m)$ equally spaced interior knots $0 = t_0 < t_1 < \cdots < t_{q_n-m} < t_{q_n-m+1} = \tau$, where $\tau$ is the study duration. The basis functions can be generated using the Cox–de Boor recursion formula $\phi_{\ell,0}(t) := 1$ if $t_\ell \le t < t_{\ell+1}$, and $\phi_{\ell,0}(t) := 0$ otherwise. Furthermore, $\phi_{\ell,k}(t) := (t-t_\ell)/(t_{\ell+k-1}-t_\ell)\phi_{\ell,k-1}(t)+(t_{\ell+k}-t)/(t_{\ell+k}-t_{\ell+1})\phi_{\ell+1,k-1}(t)$, for $1 \le k \le m$.

To simplify the notation, we write $\phi_{\ell,m}(t)$ as $\phi_\ell(t)$. Define $\boldsymbol{\phi}(t) = (\phi_1(t),\ldots,\phi_{q_n}(t))^T$. Then, for the $j$th component of $\boldsymbol{\beta}(t)$, we approximate $\beta_j(t)$ by

$$
\beta_j(t) \approx \boldsymbol{\gamma}_j^T\boldsymbol{\phi}(t), \quad j = 1,\ldots,p_n, \tag{2.3}
$$

where $\boldsymbol{\gamma}_j = (\gamma_{j1},\ldots,\gamma_{jq_n})^T$ is a coefficient vector for the B-spline approximation. Consequently, define $\boldsymbol{Z}_i(t,u) = \boldsymbol{X}_i(t)\otimes\boldsymbol{\phi}(u)$, where $\otimes$ is the Kronecker product, and $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}_1^T,\ldots,\boldsymbol{\gamma}_{p_n}^T\right)^T$. Then, we propose maximizing

$$
\begin{aligned}
l_n(\boldsymbol{\gamma}) =& n^{-1}\sum_{i=1}^{n}\int\sum_{v=1}^{n_i} K_{h_n}(t-t_{iv})\Bigg(\boldsymbol{\gamma}^T\boldsymbol{Z}_i(t_{iv},t) \\
&-\log\left[n^{-1}\sum_{j=1}^{n}\sum_{k=1}^{n_j} I(\widetilde{T}_j \ge t)K_{h_n}(t-t_{jk})\exp\left\{\boldsymbol{\gamma}^T\boldsymbol{Z}_j(t_{jk},t)\right\}\right]\Bigg)dN_i(t) \\
=& n^{-1}\sum_{i=1}^{n}\sum_{v=1}^{n_i}\Delta_i K_{h_n}(\widetilde{T}_i-t_{iv})\Bigg(\boldsymbol{\gamma}^T\boldsymbol{Z}_i(t_{iv},\widetilde{T}_i) \\
&-\log\left[n^{-1}\sum_{j=1}^{n}\sum_{k=1}^{n_j} I(\widetilde{T}_j \ge \widetilde{T}_i)K_{h_n}(\widetilde{T}_i-t_{jk})\exp\left\{\boldsymbol{\gamma}^T\boldsymbol{Z}_j(t_{jk},\widetilde{T}_i)\right\}\right]\Bigg)
\end{aligned}
\tag{2.4}
$$

to estimate $\boldsymbol{\gamma}$, and thus $\boldsymbol{\beta}(t)$.

## 2.2. Sparsity regularization

With a large number of biomarkers, directly maximizing $l_n(\boldsymbol{\gamma})$ may lead to high variability in the time-varying effects, and may even be infeasible. Furthermore, because most of the biomarkers are expected to be noninformative in terms of disease prognosis, it is important that we identify which ones contribute to the underlying biological mechanism. Thus, we impose a regularization to stabilize the computation and for variable selection. Specifically, we propose minimizing the following penalized function:

$$\widehat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \left\{ -l_n(\boldsymbol{\gamma}) + p(\boldsymbol{\gamma}; \nu_n) \right\}, \tag{2.5}$$

where $p(\boldsymbol{\gamma}; \nu_n)$ is a prespecified penalty function with tuning parameter $\nu_n$.

Because we aim to select the important $\beta_j(t)$ as functions in $[0, \tau]$, or equivalently, $\boldsymbol{\gamma}_j$ as a vector, the penalty is imposed on the Euclidean norm of $\boldsymbol{\gamma}_j$, rather than on each component of $\boldsymbol{\gamma}_j$. Furthermore, to encourage oracle selection, following Zhang and Zhang (2012), one may wish to choose a concave penalty function. Therefore, we choose the penalty term $p(\boldsymbol{\gamma}; \nu_n) = \nu_n \sum_{j=1}^{p_n} \sqrt{q_n} \rho(\|\boldsymbol{\gamma}_j\|_2)$, where $\rho(\cdot)$ is a general penalty function imposed on $\|\boldsymbol{\gamma}_j\|_2$, the Euclidean norm of $\gamma_j$. For example, $\rho(t) = t$ gives the LASSO penalty,

$$\rho(t) = \int_0^t I(u \le \nu_n) + \frac{(a\nu_n - u)_+}{(a-1)\nu_n} I(u \ge \nu_n) du$$

gives the SCAD penalty (Fan and Li (2001)), and

$$\rho(t) = \int_0^t \frac{(a\nu_n - u)_+}{a\nu_n} du$$

yields the MCP penalty (Zhang (2010)). As an extreme case, we can choose $p(\boldsymbol{\gamma}; \nu_n)$ to be the $\ell_0$-penalty, which is defined as $\nu_n \sum_{j=1}^{p_n} \sqrt{q_n} \|\boldsymbol{\gamma}_j\|_{G0}$, with $\|\boldsymbol{\gamma}_j\|_{G0} = I(\|\boldsymbol{\gamma}_j\|_2 \ne 0)$. A concave penalty may lead to a nonconvex minimization; in the next section, we describe a unified algorithm to facilitate the computation.

## 2.3. Computational algorithm

We propose a unified computational algorithm for the optimization (2.5). The algorithm is motivated by a class of proximal methods that perform augmentation and splitting, including the ADMM algorithm (Boyd et al. (2011)).

Specifically, additional slack variables $\boldsymbol{\theta}$ of the same dimension as the target variables $\boldsymbol{\gamma}$ are introduced to facilitate efficient computation and scaling.

Following Li et al. (2017), we first approximate (2.5) by the following constrained optimization problem:

$$\underset{\boldsymbol{\gamma},\boldsymbol{\theta}}{\arg\min} \; -l_n(\boldsymbol{\gamma}) + p(\boldsymbol{\theta};\nu_n) \quad \text{subject to} \quad \sum_{j=1}^{p_n} \|\boldsymbol{\gamma}_j - \boldsymbol{\theta}_j\|_2 \le c_n, \tag{2.6}$$

where $c_n$ is some constant that controls the difference between $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$. Note that the ADMM algorithm is a special case when $c_n = 0$ in (2.6). We further propose solving the equivalent Lagrangian problem

$$\underset{\boldsymbol{\gamma},\boldsymbol{\theta}}{\arg\min} \; -l_n(\boldsymbol{\gamma}) + p(\boldsymbol{\theta};\nu_n) + \phi_n \sum_{j=1}^{p_n} \sqrt{q_n}\|\boldsymbol{\gamma}_j - \boldsymbol{\theta}_j\|_2, \tag{2.7}$$

where $\phi_n$ is the Lagrangian multiplier. We delegate the proof of the equivalence between (2.6) and (2.7) to the Supplementary Material. Thus, we can minimize (2.7) for given $\nu_n$ and $\phi_n$ by iteratively updating all parameters using the following algorithm: at the $k$th iteration,

$$\boldsymbol{\gamma}^{k+1} = \underset{\boldsymbol{\gamma}}{\arg\min} -l_n(\boldsymbol{\gamma}) + \phi_n \sum_{j=1}^{p_n} \sqrt{q_n}\|\boldsymbol{\gamma}_j - \boldsymbol{\theta}_j^k\|_2, \tag{2.8}$$

$$\boldsymbol{\theta}^{k+1} = \underset{\boldsymbol{\theta}}{\arg\min}\, p(\boldsymbol{\theta};\nu_n) + \phi_n \sum_{j=1}^{p_n} \sqrt{q_n}\|\boldsymbol{\gamma}_j^{k+1} - \boldsymbol{\theta}_j\|_2. \tag{2.9}$$

Note that the above update (2.8) is similar to updating a regularized regression using a group LASSO, where the objective function is convex. For (2.9), when $p(\boldsymbol{\theta};\nu_n) = \nu_n \sum_{j=1}^{p_n} \sqrt{q_n}\rho(\|\boldsymbol{\gamma}_j\|_2)$, we can perform a groupwise minimization, which often results in an explicit solution. Therefore, each iteration of the proposed algorithm involves one step of convex minimization and one step of simple calculation. The tuning parameters $\nu_n$ and $\phi_n$ can be chosen using likelihood-based cross-validation.

For example, when $p(\boldsymbol{\gamma};\nu_n)$ is chosen as the $\ell_0$-penalty, the second step in each iteration of the above algorithm becomes

$$\boldsymbol{\theta}^{k+1} = \underset{\boldsymbol{\theta}}{\arg\min}\, \nu_n \sum_{j=1}^{p_n} \sqrt{q_n}\|\boldsymbol{\theta}_j\|_{G0} + \phi_n \sum_{j=1}^{p_n} \sqrt{q_n}\|\boldsymbol{\gamma}_j^{k+1} - \boldsymbol{\theta}_j\|_2.$$

In Section 4, we show the groupwise $\ell_0$-penalty acts as a hard threshold for the estimates obtained in the first step. Simple algebra gives, for $j = 1, \ldots, p_n$,

$$\boldsymbol{\theta}_j^{k+1} = \boldsymbol{\gamma}_j^{k+1} I \left( \|\boldsymbol{\gamma}_j^{k+1}\|_2 > \frac{\nu_n}{\phi_n} \right). \tag{2.10}$$

Note that, the challenge in selecting informative biomarkers from a large candidate pool arises because each component of the effect profiles $\boldsymbol{\beta}(t)$ is a function. Thus, the estimation noise on the entire range of $t$ needs to be controlled, and a penalty needs to be imposed on its norm.

Furthermore, when the dimension of $\boldsymbol{X}_i(t)$ is high, the computation in (2.8) can still be intensive, even if it is a convex minimization. In Section 4, we suggest a coordinate-descent approach, in which we first approximate $l_n(\boldsymbol{\gamma})$ using a summation of the quadratic functions for each $\boldsymbol{\gamma}_j$. Thus, the computation can easily scale up to high-dimensional scenarios.

## 3. Theoretical Properties

The main challenge in proving theoretical properties is to appropriately control for the approximation bias resulting from the local kernel smoothing in the approximated likelihood (2.3) and stochastic variability. Let $\boldsymbol{\beta}^*(t)$ denote the true value of $\boldsymbol{\beta}(t)$, and let $\beta_j^*(t)$ denote its $j$th element. We provide a nonasymptotic result showing that, with large probability, the nonzero $\beta_j^*(t)$ can be correctly selected and consistently estimated by $\hat{\beta}_j(t)$, where $\hat{\beta}_j(t) := \hat{\boldsymbol{\gamma}}_j^T \boldsymbol{\phi}(t)$ and $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\gamma}}_1^T, \ldots, \hat{\boldsymbol{\gamma}}_{p_n}^T)^T$ is the estimator given by (2.5). In particular, we allow $p_n$ and $q_n$ to diverge to infinity and $\nu_n$ to converge to zero with $n$.

### 3.1. Technical notation

Before presenting regularity conditions, we need the following notation. Let

$$S_n^{(l)}(\boldsymbol{\gamma}, t) = n^{-1} \sum_{i=1}^{n} \sum_{v=1}^{n_i} K_{h_n}(t - t_{iv}) Y_i(t) \{\boldsymbol{Z}_i(t_{iv}, t)\}^{\otimes l} \exp\{\boldsymbol{\gamma}^T \boldsymbol{Z}_i(t_{iv}, t)\}, \quad l = 0, 1, 2. \tag{3.1}$$

Then, the approximated log-partial likelihood for optimization can be rewritten as

$$l_n(\boldsymbol{\gamma}) = \frac{1}{n} \sum_{i=1}^{n} \int_0^{\tau} \sum_{v=1}^{n_i} K_{h_n}(t - t_{iv}) \left[ \boldsymbol{\gamma}^T \boldsymbol{Z}_i(t_{iv}, t) - \log\{S_n^{(0)}(\boldsymbol{\gamma}, t)\} \right] dN_i(t).$$

Denote $\boldsymbol{E}_n(\boldsymbol{\gamma}, t) = S_n^{(1)}(\boldsymbol{\gamma}, t)/S_n^{(0)}(\boldsymbol{\gamma}, t)$. Then, the gradient vector, denoted by $\boldsymbol{U}_n(\boldsymbol{\gamma})$, and the negative Hessian matrix, denoted by $\boldsymbol{I}_n(\boldsymbol{\gamma})$, of $l_n(\boldsymbol{\gamma})$ are given by

$$\boldsymbol{U}_n(\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\sum_{v=1}^{n_i} K_{h_n}(t - t_{iv})\{\boldsymbol{Z}_i(t_{iv}, t) - \boldsymbol{E}_n(\boldsymbol{\gamma}, t)\}dN_i(t),$$
$$\boldsymbol{I}_n(\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\sum_{v=1}^{n_i} K_{h_n}(t - t_{iv})\left\{\frac{S_n^{(2)}(\boldsymbol{\gamma}, t)}{S_n^{(0)}(\boldsymbol{\gamma}, t)} - \boldsymbol{E}_n(\boldsymbol{\gamma}, t)^{\otimes 2}\right\}dN_i(t).$$

In addition, we define

$$\bar{s}^{(l)}(\boldsymbol{\gamma}, t) = \mathrm{E}[Y(t)\{\boldsymbol{Z}(t, t)\}^{\otimes l}\exp\{\boldsymbol{\gamma}^T\boldsymbol{Z}(t, t)\}], \text{ for } l = 0, 1, 2;$$

$\boldsymbol{e}(\boldsymbol{\gamma}, t) = \bar{s}^{(1)}(\boldsymbol{\gamma}, t)/\bar{s}^{(0)}(\boldsymbol{\gamma}, t)$, and

$$\boldsymbol{\Sigma}(\boldsymbol{\gamma}, t) = \int_0^{\tau}\left\{\frac{\bar{s}^{(2)}(\boldsymbol{\gamma}, t)}{\bar{s}^{(0)}(\boldsymbol{\gamma}, t)} - \boldsymbol{e}(\boldsymbol{\gamma}, t)^{\otimes 2}\right\}\bar{s}^{(0)}(\boldsymbol{\gamma}, t)\lambda_v(t)d\Lambda_0(t),$$

where we assume $\{t_{i1}, \ldots, t_{in_i}\}$ follow an independent counting process with intensity function $\lambda_v(t)$.

In the penalized partial likelihood (2.5), we assume $\rho(t)$ belongs to a general class of folded-concave functions, as discussed in Fan and Lv (2011). Such a class of functions will be characterized by Condition 11 in Section 3.2. In particular, denote $\kappa(\rho, \boldsymbol{u})$ as the "local concavity" of $\rho(\cdot)$ at a general vector $\boldsymbol{u} = (u_1, \ldots, u_s)^T \in \mathcal{R}^s$, yielding

$$\kappa(\rho, \boldsymbol{u}) = \lim_{\varepsilon\to 0_+}\max_{1\le j\le s}\sup_{t_1<t_2\in(|u_j|-\varepsilon,|u_j|+\varepsilon)} -\frac{\rho'(t_2) - \rho'(t_1)}{t_2 - t_1}.$$

For example, for the LASSO penalty, $\kappa(\rho, \boldsymbol{u}) = 0$, whereas for the SCAD penalty,

$$\kappa(\rho, \boldsymbol{u}) = \begin{cases}(a-1)^{-1}\nu_n^{-1}, & \text{if there exists a } u_j \text{ such that } \nu_n \le |u_j| \le a\nu_n;\\ 0, & \text{otherwise.}\end{cases}$$

Our subsequent regularity condition for the penalty function is expressed in terms of $\kappa(\rho, \boldsymbol{u})$.

Lastly, for a vector $\boldsymbol{a}$, let $\|\boldsymbol{a}\|_\infty = \max_j |a_j|$ denote its sup-norm. For a matrix $\boldsymbol{A}$, let $\|\boldsymbol{A}\|_\infty = \max_i \sum_j |a_{ij}|$ denote its matrix sup-norm, where $a_{ij}$ is the $(i, j)$th element of $\boldsymbol{A}$. Let $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ be the minimal and maximal eigenvalues of $\boldsymbol{A}$, respectively.

### 3.2. Regularity conditions

We define the unique projection of $\beta_j^*(t)$ on the sieve space consisting of $\boldsymbol{\phi}(t)$ as $\boldsymbol{\gamma}_j^{*T}\boldsymbol{\phi}(t)$, where $\boldsymbol{\gamma}_j^*$ is a $q_n$-dimensional vector. Denote $\mathcal{M} = \{j : \beta_j^*(t) \neq 0, t \in [0,\tau]\}$ as the set of active $\beta_j^*(t)$. We assume that when $q_n$ is sufficiently large, $\mathcal{M}$ is equivalent to the support of $\boldsymbol{\gamma}^*$; that is, $\mathcal{M} = \left\{j : \boldsymbol{\gamma}_j^* \neq \mathbf{0}\right\}$. In other words, the important covariates with $\beta_j^*(t) \neq 0$ are fully characterized by the nonzero $\boldsymbol{\gamma}^*$ vectors when we choose a sufficient number of spline bases. Let $r_n = |\mathcal{M}|$ denote its cardinality. Denote $\mathcal{A} = \{j_l : j \in \mathcal{M} \text{ and } 1 \leq l \leq q_n\}$. Note that $|\mathcal{A}| = r_n q_n$. Denote $d_n = \min_{j\in\mathcal{M}}\|\boldsymbol{\gamma}_j^*\|_2$ as the minimal signal strength. For a set $S$, denote $\boldsymbol{a}_S$ as the subvector of $\boldsymbol{a}$ with indices in $S$, and $\boldsymbol{A}_{SS}$ as the submatrix with row and column indices in $S$. Let $\mathcal{B}_0 := \{\boldsymbol{\gamma} \in \mathcal{R}^{p_n q_n} : \|\boldsymbol{\gamma}_{\mathcal{A}} - \boldsymbol{\gamma}_{\mathcal{A}}^*\|_\infty \leq d_n \text{ and } \boldsymbol{\gamma}_{\mathcal{A}^c} = \mathbf{0}\}$. We assume that $M$ is a universal positive constant and that the following conditions hold.

Condition **1.** $\Lambda_0(\tau) = \int_0^\tau \lambda_0(t)dt < \infty$ and $P\{C \geq \tau\} > 0$.

Condition **2.** $\sup_{t\in[0,\tau]} |X_j(t)| \leq M$, for all $1 \leq j \leq p$; $\sup_{t\in[0,\tau]} |(\boldsymbol{\beta}^*(t))^T \boldsymbol{X}(t)| \leq M$.

Condition **3.** $\lambda_v(t)$ is bounded and twice continuously differentiable with a bounded second derivative.

Condition **4.** The kernel function $K(x)$ is symmetric and has a finite second moment.

Condition **5.** $|\mathrm{E}[X_j(t)]| \leq M$ and $\mathrm{E}[X_j(t)]$ is twice continuously differentiable and
$|(\mathrm{E}[X_j(t)])''| \leq M$. In addition, $\mathrm{E}[X_j(t) - X_j(s)]^2 \leq M(t-s)^2$ holds almost everywhere for $t, s \in [0,\tau]$.

Condition **6.** There exists some positive constant $\alpha$ such that $\sup_{t\in[0,\tau]} |\beta_j^*(t) - (\boldsymbol{\gamma}_j^*)^T\boldsymbol{\phi}(t)| \leq c_\alpha q_n^{-\alpha}$ for all $1 \leq j \leq p_n$, where $c_\alpha$ is some positive constant.

Condition **7.** $(nh_n)^{-1/2} = O(1)$.

Condition **8.** $r_n q_n c_n d_n^{1/2} = o(n)$, where $c_n := r_n q_n^2 h_n^{-1} \vee h_n^{-2}$.

Condition **9.** $\sup_{\boldsymbol{\gamma}\in\mathcal{B}_0} \|\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}(\boldsymbol{\gamma})^{-1}\|_\infty \leq M$.

Condition **10.** $\sup_{\boldsymbol{\gamma}\in\mathcal{B}_0} \|\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}(\boldsymbol{\gamma})\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}(\boldsymbol{\gamma})^{-1}\|_\infty \leq (1-\zeta)\rho'(0+)/\rho'(d_n/2)$, for some $\zeta \in (0,1)$.

Condition **11.** $\rho(t)$ is increasing and concave in $t \in [0,\infty)$, and has a continuous derivative $\rho'(t)$, with $0 < \rho'(0+) < M$. $\sup_{\boldsymbol{\gamma}\in\mathcal{B}_0} |\nu_n \kappa(\rho, \boldsymbol{\gamma})| \leq M$.

Condition 1 is a common condition when using the Cox model. Condition

2 is used to establish the exponential type of concentration inequalities for the gradient vector. Condition 3 is a smoothness condition on $\lambda_v(t)$. Condition 4 is imposed on the kernel function and is satisfied for common kernel functions, such as the Gaussian kernel and Epanechnikov kernel. Condition 5 is a smoothing assumption on $\mathrm{E}[X_j(t)]$. Because this condition is imposed on the population average, it allows the individual realization of $X_{ij}(t)$ to be nonsmooth or even discontinuous. For example, the trajectory of $X_{ij}(t)$ may have a discontinuity point that is from a continuous distribution in $[0, \tau]$. Condition 6 requires that $\beta_j^*(t)$ be sufficiently smooth that it can be well approximated by splines. If $\beta_j^*(t)$ has bounded $k$th derivatives, then $\alpha = k - 1$ (see Schumaker (2007)). Condition 9 and Condition 10 are imposed on the population information matrix $\boldsymbol{\Sigma}(\boldsymbol{\gamma})$. Similar conditions also appear in derivations of the oracle properties for parametric models (Fan and Lv (2011)). Condition 10 is an irrepresentable-type condition. If some concave penalties (e.g., SCAD or MCP) are used, the upper bound in Condition 10 is allowed to diverge to infinity at a polynomial rate of $n$ (Fan and Lv (2011)). If the $\ell_1$-penalty is used, the upper bound reduces to $1 - \zeta$, which is exactly the irrepresentable condition needed for the LASSO (Zhao and Yu (2006)). Under Conditions 9 and 10, similar results for the sample information matrix $\boldsymbol{I}_n(\boldsymbol{\gamma})$ can be shown to hold with high probability (See Lemma S5 in the Supplementary Material). Condition 11 is imposed on the penalty function, and is satisfied by commonly used penalty functions, such as the LASSO, SCAD, and MCP.

### 3.3. Main results

We first give a concentration inequality for the gradient vector $\boldsymbol{U}_n(\boldsymbol{\gamma}^*)$, which will play a key role in establishing the main result.

**Lemma 1.** *Under conditions* 1 *to* 8*, there exist positive constants* $C_1$*,* $C_2$*,* $C_3$*,* $C_4$*, and* $D$ *such that, for any* $x > 0$ *and* $\epsilon > 0$*, it holds with probability no less than* $1 - \epsilon - C_1 \exp(-C_2 n h_n^6 x^2) - C_3 \exp(-C_4 x)$ *that*

$$|U_{n,j}(\boldsymbol{\gamma}^*)| \leq D\{(nh_n^2)^{-1/2} x + \pi_n\},$$

*where* $U_{n,j}(\boldsymbol{\gamma}^*)$ *is the* $j$*th element of* $\boldsymbol{U}_n(\boldsymbol{\gamma}^*)$ *and* $\pi_n = (r_n q_n c_n d_n^{1/2}/n)^{1/2} + h_n^2 + r_n q_n^{-\alpha}$.

**Remark 1.** In contrast to existing studies on the ordinary Cox model in high-dimensional settings (Bradic, Fan and Jiang (2011)), the expectation of the gradient vector $\boldsymbol{U}_n(\boldsymbol{\gamma}^*)$ is no longer equal to zero, owing to the spline approximation of $\boldsymbol{\beta}^*(t)$ and the local smoothing. The term $\pi_n$ quantifies such a bias. In the expression of $\pi_n$, $h_n^2$ is the result of local smoothing, $r_n q_n^{-\alpha}$ is the result of the approximation of $(\boldsymbol{\gamma}_j^*)^T \boldsymbol{\phi}(t)$ by $\beta_j^*(t)$, and $(r_n q_n c_n d_n^{1/2}/n)^{1/2}$ is introduced by $S_n^{(l)}(\boldsymbol{\gamma}, t)$. However, owing to the conditions of $p_n, q_n$, and $d_n$, the bias $\pi_n$ vanishes as the sample size $n$ grows, and so it does not affect the variable selection given in Theorem 1 below.

**Remark 2.** Under condition 3, $n_i = O_P(1)$; that is, for any $\epsilon > 0$, there is a constant $M_\epsilon$ such that $P(n_i > M_\epsilon) \leq \epsilon$. Such an $\epsilon$ appears in Lemma 1. All other probabilities are calculated conditioning on the event $\{n_i \leq M_\epsilon\}$. We also condition on such an event when calculating the exception probabilities in Theorem 1 and the additional lemmas in the Supplementary Material.

Our main theoretical result considers the variable selection property of our method. Specifically, we show that the estimator $\hat{\boldsymbol{\beta}}(t) = (\hat{\beta}_1(t), \ldots, \hat{\beta}_p(t))^T$ possesses the "weak oracle property", as discussed in Fan and Lv (2011) for the generalized linear model; that is, it is variable selection consistent and it consistently estimates the nonzero components of $\boldsymbol{\beta}^*(t)$, the component functions that are not identically equal to zero.

**Theorem 1.** *Suppose conditions* 1 *to* 11 *hold, and*

$$\frac{n^2 h_n^8 (\nu_n \sqrt{q_n} - \pi_n)^2}{\log(p_n q_n)} \to \infty, \quad \frac{(n h_n^2)^{1/2} (\nu_n \sqrt{q_n} - \pi_n)}{\log(p_n q_n)} \to \infty, \tag{3.2}$$

$$\frac{n h_n^2 (r_n q_n)^{-1}}{\log(p_n r_n q_n^2)} \to \infty, \quad d_n > 2M \nu_n q_n. \tag{3.3}$$

*There exist universal positive constants* $C_1$*,* $C_2$*,* $C_3$*,* $C_4$*,* $C_5$*, and* $C_6$ *such that, for any* $\epsilon > 0$*, with probability at least*

$$\begin{aligned} &1 - \epsilon - C_1 p_n q_n \exp\{-C_2 n^2 h_n^8 (\nu_n \sqrt{q_n} - \pi_n)^2\} \\ &\quad - C_3 p_n q_n \exp\{-C_4 (n h_n^2)^{1/2} (\nu_n \sqrt{q_n} - \pi_n)\} - C_5 p_n r_n q_n^2 \exp\{-C_6 n h_n^2 (r_n q_n)^{-1}\}, \end{aligned} \tag{3.4}$$

*there is a solution to* (2.5) *that yields* $\hat{\boldsymbol{\beta}}(t)$*, satisfying*
$(a)$*(variable selection consistency):* $\{j : \hat{\beta}_j(t) \not\equiv 0\} = \{j : \beta_j^*(t) \not\equiv 0\}$*;*
$(b)$*(*$L_\infty$ *error):* $\max_{j \in \mathcal{M}} \sup_{0 \leq t \leq \tau} |\hat{\beta}_j(t) - \beta_j^*(t)| \leq M(\nu_n q_n^{3/2} + q_n^{-\alpha})$*, where* $M$ *is a positive constant.*

The result in (a) guarantees the variable selection consistency with a high probability. The term $M \nu_n q_n^{3/2}$ in statement (b) gives the upper bound of

$\max_{j_l \in \mathcal{A}} |\hat{\gamma}_{j_l} - \gamma^*_{j_l}|$, and the term $Mq_n^{-\alpha}$ corresponds to the approximation error of $\boldsymbol{\gamma}_j^{*T}\boldsymbol{\phi}(t)$ to $\beta_j^*(t)$. The first assumption in (3.3) provides a constraint on the divergence rates of $p_n$, $r_n$, and $q_n$. They are allowed to diverge, as long as $r_n q_n \log(p_n r_n q_n^2) = o(nh_n^2)$. The minimal signal strength $d_n$ is allowed to converge to zero, given that the second term in (3.3) holds.

The conditions in Theorem 1 also restrict the choice of the tuning parameter $\nu_n$ in the penalty function and the bandwidth $h_n$ in the kernel smoothing. Specifically, by (3.2) and (3.3), the tuning parameter $\nu_n$ needs to satisfy

$$\pi_n q_n^{-1/2} \vee \frac{\log(p_n q_n)}{n^{1/2} h_n q_n^{1/2}} \vee \frac{\sqrt{\log(p_n q_n)}}{n h_n^4 q_n} \ll \nu_n < \frac{d_n}{2Mq_n}.$$

Moreover, the lower bound for $h_n$ is given by

$$h_n \gg \frac{\log(p_n r_n q_n^2)}{n(r_n q_n)^{-1}} \vee \frac{\{\log(p_n q_n)\}^2}{n q_n \nu_n^2} \vee \frac{\{\log(p_n q_n)\}^{1/4}}{\sqrt{n q_n \nu_n}}.$$

For example, suppose that $r_n$ and $d_n$ are fixed and independent of $n$ and $\alpha \geq 2$. Then, for $p_n = \exp\{O(n^{1/8})\}$, once we choose $q_n \asymp n^{1/8}$, $h_n \asymp n^{-1/8}$, and $n^{-5/16} \ll \nu_n \ll n^{-3/16}$, all requirements in Theorem 1 are met such that the probability in (3.4) becomes arbitrarily close to one and the upper bound in Theorem 1(b) converges to zero.

## 4. Incorporating the Network Structure

In many applications, biomarkers such as structural/functional brain measures (He, Chen and Evans (2008); Alexander-Bloch, Giedd and Bullmore (2013)) and co-expression of genes (Stuart et al. (2003)) exhibit network structures and, hence, can be naturally described by a graph $G = (V, E, \mathcal{W})$, where $V$ is the set of vertices corresponding to the biomarkers, $E = \{j \sim k\}$ is the set of edges that indicate connected vertices, and $\mathcal{W} = \{w_{jk} > 0 : (j,k) \in E\}$ is the set of edge weights. Note that $\mathcal{W}$ is a $p_n \times p_n$ matrix with zero diagonal entries; that is, $w_{jj} = 0$, for $j = 1, \ldots, p_n$.

When two biomarkers are highly linked, that is $w_{jk}$ is large, they are likely to be involved in similar disease pathways, which means the corresponding $\boldsymbol{\beta}_j$ and $\boldsymbol{\beta}_k$ are similar. Our proposed method can be extended easily to incorporate such network information in order to encourage this pattern in the analysis. Specifically, we add to (2.5) a Laplacian quadratic penalty on the effect size using the $\ell_2$-norm vector ${}_{|}\boldsymbol{\gamma}_{|_G} = \left(\|\boldsymbol{\gamma}_1\|_2, \ldots, \|\boldsymbol{\gamma}_{p_n}\|_2\right)^T$ to encourage smoothness of the functions over the network graph. Such a penalty takes the form of ${}_{|}\boldsymbol{\gamma}_{|_G}^T \mathcal{L}_n^* {}_{|}\boldsymbol{\gamma}_{|_G}$, where $\mathcal{L}_n^*$ is a positive semidefinite matrix associated with the graph $G$. We can choose $\mathcal{L}_n^* = \boldsymbol{D} - \mathcal{W}$, where $\boldsymbol{D} = \operatorname{diag}(d_1, \ldots, d_{p_n})$ with $d_j = \sum_{l:(j,l)\in E} w_{jl}$, or its normalized version given by $\mathcal{L}_n^* = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\mathcal{W}\boldsymbol{D}^{-1/2}$. The former choice of $\mathcal{L}_n^*$ yields the penalty term

$$\sum_{(j,l)\in E} w_{jl}\big(\|\boldsymbol{\gamma}_j\|_2 - \|\boldsymbol{\gamma}_l\|_2\big)^2, \tag{4.1}$$

whereas the latter gives

$$\sum_{(j,l)\in E} w_{jl}\left(\frac{\|\boldsymbol{\gamma}_j\|_2}{\sqrt{d_j}} - \frac{\|\boldsymbol{\gamma}_l\|_2}{\sqrt{d_l}}\right)^2.$$

With this additional penalty, the previous algorithm in Section 2.2 is still applicable. In particular, the first step of each iteration solves

$$\begin{aligned} \widehat{\boldsymbol{\gamma}} = \arg\min\nolimits_{\boldsymbol{\gamma}} \Big\{ &- l_n(\boldsymbol{\gamma}) + \lambda_1 \textstyle\sum_{j=1}^{p_n} \sqrt{q_n}\|\boldsymbol{\gamma}_j - \boldsymbol{\theta}_j\|_2 \\ &+(\lambda_2/2)\textstyle\sum_{(j,l)\in E} w_{jl}\left(\frac{\|\boldsymbol{\gamma}_j\|_2}{\sqrt{d_j}} - \frac{\|\boldsymbol{\gamma}_l\|_2}{\sqrt{d_l}}\right)^2 \Big\}. \end{aligned} \tag{4.2}$$

When the dimension $p_n$ is large, it is not straightforward to directly find the minimum of the objective function (4.2), so we propose a majorization-minimization (MM) approach (Lange (2013)), and employ a groupwise descent algorithm to solve (4.2). The minimization is achieved by a cyclic descent over each group, where we choose a target group $\boldsymbol{\gamma}_j$ to minimize and consider other group coefficients $\boldsymbol{\gamma}_k = \widehat{\boldsymbol{\gamma}}_k$, for $k \neq j$, as fixed from the previous iteration. The details are given as follows.

Denote by $\nabla_j l_n(\boldsymbol{\gamma}_j)$ the gradient taken over $\boldsymbol{\gamma}_j$. Using a second-order Taylor expansion on $\boldsymbol{\gamma}_j$ centered at a point $\widetilde{\boldsymbol{\gamma}}_j$, and replacing the Hessian matrix with a suitable matrix $\boldsymbol{H}$, we first majorize $-l_n(\boldsymbol{\gamma}_j)$ using a surrogate function $M(\boldsymbol{\gamma}_j|\widetilde{\boldsymbol{\gamma}}_j)$, as follows:

$$M(\boldsymbol{\gamma}_j|\widetilde{\boldsymbol{\gamma}}_j) = -\left\{ l_n(\widetilde{\boldsymbol{\gamma}}_j) + (\boldsymbol{\gamma}_j - \widetilde{\boldsymbol{\gamma}}_j)^T \nabla_j l_n(\widetilde{\boldsymbol{\gamma}}_j) + \frac{1}{2}(\boldsymbol{\gamma}_j - \widetilde{\boldsymbol{\gamma}}_j)^T \boldsymbol{H}(\boldsymbol{\gamma}_j - \widetilde{\boldsymbol{\gamma}}_j) \right\},$$

with $-l_n(\boldsymbol{\gamma}_j) \leq M(\boldsymbol{\gamma}_j|\widetilde{\boldsymbol{\gamma}}_j)$ and $-l_n(\widetilde{\boldsymbol{\gamma}}_j) = M(\widetilde{\boldsymbol{\gamma}}_j|\widetilde{\boldsymbol{\gamma}}_j)$. We further assume that the matrix $\boldsymbol{H}$ has the form $-\boldsymbol{H} = t^{-1}\boldsymbol{I}_{q_n}$, where $\boldsymbol{I}_{q_n}$ is an identity matrix with dimension $q_n$, and $t$ is sufficiently small such that the quadratic term $(2t)^{-1}\|\boldsymbol{\gamma}_j - \widetilde{\boldsymbol{\gamma}}_j\|_2^2$ dominates the negative Hessian matrix of $l_n(\boldsymbol{\gamma}_j)$. Thus, in each gradient step to update $\widehat{\boldsymbol{\gamma}}_j$, we solve

$$\begin{aligned}\underset{\boldsymbol{\gamma}_j}{\arg\min}\Bigg\{&\frac{1}{2t}\big\|\boldsymbol{\gamma}_j-(\widetilde{\boldsymbol{\gamma}}_j+t\nabla_j l_n(\widetilde{\boldsymbol{\gamma}}_j))\big\|_2^2+\lambda_1\sqrt{q_n}\|\boldsymbol{\gamma}_j-\boldsymbol{\theta}_j\|_2\\&+\frac{\lambda_2}{2}\sum_{l:(j,l)\in E}w_{jl}\left(\frac{\|\boldsymbol{\gamma}_j\|_2}{\sqrt{d_j}}-\frac{\|\widehat{\boldsymbol{\gamma}}_l\|_2}{\sqrt{d_l}}\right)^2\Bigg\},\end{aligned}\tag{4.3}$$

which can be carried out easily using the Newton–Raphson method. Note that a similar approach is used in Meier, Van De Geer and Bühlmann (2008) and Simon et al. (2013).

Finally, we propose using a $K$-fold cross-validation to choose all tuning parameters simultaneously. For each fixed bandwidth, the cross-validation criterion is the summation of the change in the log-partial likelihood function after omitting one fold. To choose the bandwidth, we let $h_n = Cn^{-1/8}$, start from a small positive constant $C$, and increase by a step size until the increment of the cross-validation is below a prespecified threshold. All algorithms have been implemented in an R package, which is available upon request.

## 5. Simulation Studies

In this section, we report on the extensive simulations we used to evaluate the performance of the proposed method. We set the study duration to $\tau = 1$. First, we considered $p_n$ time-dependent covariates, each with piecewise constant trajectories given by

$$X_{ij}(t)=\sum_{l=1}^{20}I\left\{\frac{(l-1)}{20}\le t<\frac{l}{20}\right\}Z_{ijl},$$

where $\{Z_{ijl} : l = 1,\ldots,20\}$ are from a multivariate normal distribution with mean zero and covariance $\mathrm{Cov}(Z_{ijl}, Z_{ijl'}) = e^{-|l-l'|/20}$, for $l, l' = 1,\ldots,20$. We also imposed a network structure on these covariates by assuming that there are links only within each block of four consecutive covariates, $\{X_{i1}, X_{i2}, X_{i3}, X_{i4}\}$, $\{X_{i5}, X_{i6}, X_{i7}, X_{i8}\}$, and so on. Furthermore, within each block of four covariates, the edge weight for each linked pair is set to be 0.5. Next, conditional on all covariates, the survival time $T_i$ was generated from model (1) with $\lambda_0(t) = 2$, and $\boldsymbol{\beta}(t)$ is given for either of the following two scenarios:

(a) $\beta_1(t) = \cdots = \beta_4(t) = 2\exp\{-10(t-0.1)^2\}$, $\beta_5(t) = \cdots = \beta_8(t) = -1$ and $\beta_9(t) = \cdots = \beta_{p_n}(t) = 0$;

(b) $\beta_i(t) = (-1)^{i+1} 2 \exp\{-10(t-i/10)^2\}$ for $i = 1, \ldots, 4$, $\beta_i(t) = (-1)^{i+1}(i-4)/2$ for $i = 5, \ldots, 8$ and $\beta_9(t) = \cdots = \beta_{p_n}(t) = 0$.

Therefore, for either scenario, only the first eight covariates are informative. Additionally, in scenario (a), the linked important covariates have the same time-varying effects, but this is not the case in scenario (b).

To simulate irregular measurements of the covariates, for each subject, we generated measurement times $t_{i1}, \ldots, t_{in_i}$ as ordered uniform distributed times in $[0, \tau]$, where $n_i$ is from a Poisson distribution with mean eight. Thus, the average number of measurements per subject is eight. Furthermore, to generate right-censored observations, we generated $C_i$ from a uniform distribution in $[0, c]$, where $c$ was chosen to yield about a 30% censoring rate.

In the simulation studies, we set $p_n = 20, 50, 1,000$ and $n = 100, 200$. When applying the proposed method to the simulated data, we used the Epanechnikov kernel function and quadratic B-splines, with two interior knots fixed at sample quantiles of the observed failure times; therefore, $q_n = 5$. When using the penalty form in (4.2), we re-parameterized $\lambda_1 = \lambda_n \alpha$ and $\lambda_2 = \lambda_n(1 - \alpha)$. We set $\alpha = 0.2, 0.5, 0.8, 1.0$, and for each $\alpha$, we selected a path for $\lambda_n$, as in Friedman, Hastie and Tibshirani (2010). Specifically, $\lambda_n$ decreases from $\lambda_{\max}$, which ensures all parameters are zero, to a portion of $\lambda_{\max}$ (i.e., $0.01 \times \lambda_{\max}$), with a length of 10 values. For the bandwidth, we chose from $h_n = 0.05, 0.1, 0.15, 0.2$. To select the tuning parameters and bandwidth, five-fold cross-validation was used. Simulations were repeated 100 times.

To evaluate the estimation performance, we computed the sum of squared errors (SSE) for the estimated $\boldsymbol{\beta}$. The numbers of true positive covariates (TP) and false positive (FP) covariates are used as measures of the variable selection performance. Moreover, we compared the performance of the proposed method (DB-hazard) with that of the LVCF. We also compared different penalty functions, including the group LASSO penalty (gLasso), group LASSO with network penalty (gNet), and $\ell_0$-regularization (2.10) with network penalty ($\ell_0$Net). We also compared the performance with that of the LVCF (imputing missing covariates using the last observed values) under various penalty functions.

Tables 1 and 2 summarize the simulation results for both settings of $\boldsymbol{\beta}(t)$. In all cases, DB-hazard, using all available longitudinal measurements, significantly improves the estimation relative to that of the LVCF in terms of yielding a much smaller SSE. This implies that the kernel smoothing method that uses all available measurements of the covariates exhibits less finite sample bias and is more efficient than using the last observed value to impute the covariate values

Table 1. Setting (a): Comparison of estimation and selection performance of the proposed DB-hazard method with that of the LVCF under various penalty functions.

| | DB-hazard | | | LVCF | | |
|---|---|---|---|---|---|---|
| | gLasso† | gNet‡ | $\ell_0$Net* | gLasso | gNet | $\ell_0$Net |
| $n = 100,\ p_n = 20$ | | | | | | |
| SSE[1] | 4.38 | 3.76 | 3.06 | 5.43 | 5.19 | 4.30 |
| TP[2] | 8.0 | 8.0 | 8.0 | 8.0 | 8.0 | 7.9 |
| FP[3] | 6.3 | 9.2 | 1.1 | 5.9 | 8.3 | 0.8 |
| $n = 100,\ p_n = 50$ | | | | | | |
| SSE | 5.19 | 4.30 | 3.12 | 6.58 | 5.74 | 4.28 |
| TP | 8.0 | 8.0 | 8.0 | 8.0 | 8.0 | 7.9 |
| FP | 14.2 | 23.9 | 1.5 | 11.4 | 21.9 | 1.4 |
| $n = 100,\ p_n = 1,000$ | | | | | | |
| SSE | 8.34 | 6.25 | 4.57 | 9.23 | 7.53 | 5.25 |
| TP | 7.7 | 8.0 | 8.0 | 7.7 | 8.0 | 7.8 |
| FP | 33.2 | 127.2 | 1.6 | 29.5 | 137.6 | 1.2 |
| $n = 200,\ p_n = 20$ | | | | | | |
| SSE | 2.69 | 2.55 | 1.92 | 4.26 | 4.17 | 3.40 |
| TP | 8.0 | 8.0 | 8.0 | 8.0 | 8.0 | 8.0 |
| FP | 7.7 | 9.4 | 0.8 | 6.1 | 8.1 | 0.8 |
| $n = 200,\ p_n = 50$ | | | | | | |
| SSE | 3.51 | 3.10 | 2.16 | 4.92 | 4.70 | 3.39 |
| TP | 8.0 | 8.0 | 8.0 | 8.0 | 8.0 | 8.0 |
| FP | 16.3 | 24.8 | 1.1 | 14.2 | 22.2 | 0.9 |
| $n = 200,\ p_n = 1,000$ | | | | | | |
| SSE | 5.04 | 4.17 | 2.83 | 6.52 | 5.73 | 4.06 |
| TP | 8.0 | 8.0 | 8.0 | 8.0 | 8.0 | 8.0 |
| FP | 57.1 | 149.0 | 1.7 | 41.8 | 137.6 | 1.4 |

†: group Lasso; ‡: group Lasso with a Laplacian penalty; *: $\ell_0$-regularization penalty (2.10); [1]: sum of squared errors; [2]: number of true positives; [3]: number of false positives.

at the observed event times. Using the $\ell_0$-penalty in our method is superior to using either the group LASSO or network regularization, indicated by a smaller SSE, much better FP, and comparable TP in all cases. The results indicate the benefits of iteratively performing hard thresholding and considering the network structure among the variables. A comparison between gLasso and gNet without hard thresholding shows that gNet has a slightly better SSE and TP, but much worse FP, which could be explained by the grouping effect of using the Laplacian penalty: selecting a noninformative covariate makes it more likely that, other highly linked covariates will as well, resulting in many more covariates being identified and poor performance in terms of variable selection.

Table 3 summarizes the performance of the bandwidth selection. We specified a range of candidate bandwidths and performed five-fold cross-validation

Table 2. Setting (b): Comparison of estimation and selection performance of the proposed DB-hazard method with that of the LVCF under various penalty functions.

| | DB-hazard | | | LVCF | | |
|---|---|---|---|---|---|---|
| | gLasso$^\dagger$ | gNet$^\ddagger$ | $\ell_0$Net$^*$ | gLasso | gNet | $\ell_0$Net |
| | | | $n = 100,\ p_n = 20$ | | | |
| SSE[1] | 6.12 | 5.96 | 4.96 | 8.93 | 8.65 | 7.49 |
| TP[2] | 7.6 | 7.9 | 7.4 | 7.3 | 7.8 | 7.0 |
| FP[3] | 7.5 | 9.2 | 0.8 | 6.0 | 8.0 | 0.8 |
| | | | $n = 100,\ p_n = 50$ | | | |
| SSE | 8.67 | 8.00 | 6.08 | 10.76 | 10.33 | 8.26 |
| TP | 6.8 | 7.7 | 7.2 | 6.4 | 7.3 | 6.6 |
| FP | 14.4 | 24.2 | 1.2 | 11.3 | 18.9 | 1.1 |
| | | | $n = 100,\ p_n = 1,000$ | | | |
| SSE | 14.14 | 13.91 | 12.59 | 14.51 | 14.27 | 13.05 |
| TP | 2.1 | 3.3 | 3.5 | 1.9 | 3.4 | 3.4 |
| FP | 14.8 | 38.9 | 5.2 | 8.0 | 31.0 | 3.7 |
| | | | $n = 200,\ p_n = 20$ | | | |
| SSE | 4.04 | 3.94 | 3.22 | 6.61 | 6.67 | 5.61 |
| TP | 7.9 | 8.0 | 7.8 | 7.9 | 7.9 | 7.6 |
| FP | 8.5 | 9.5 | 0.8 | 7.9 | 8.7 | 0.6 |
| | | | $n = 200,\ p_n = 50$ | | | |
| SSE | 5.62 | 5.53 | 3.78 | 8.30 | 8.25 | 6.21 |
| TP | 7.8 | 7.9 | 7.7 | 7.7 | 7.8 | 7.4 |
| FP | 18.8 | 24.9 | 0.4 | 16.4 | 20.7 | 0.6 |
| | | | $n = 200,\ p_n = 1,000$ | | | |
| SSE | 10.43 | 10.02 | 8.06 | 12.33 | 11.77 | 9.21 |
| TP | 5.9 | 7.1 | 7.4 | 5.6 | 7.3 | 7.1 |
| FP | 48.2 | 133.6 | 1.0 | 26.8 | 75.3 | 0.7 |

$^\dagger$: group Lasso; $^\ddagger$: group Lasso with a Laplacian penalty; $^*$: $\ell_0$-regularization penalty (2.10); [1]: sum of squared errors; [2]: number of true positives; [3]: number of false positives.

to select the optimal bandwidth. We compared our selection approach to the method with the smallest SSE among all candidates, denoted as "Best" in Table 3. The results show that our selected bandwidths are very close to the "Best" bandwidth, indicating satisfactory performance of our data-driven procedure.

To ease the computational burden, we implemented several techniques to speed up our algorithms. We use warm starts to estimate $\boldsymbol{\beta}(t)$ along a regularization path, and use a sparse data structure to save memory and to reduce the time taken to search for the nonzero coefficients in a sparse $\boldsymbol{\beta}$. Thus, the computing time for our method is highly manageable. Figure S1 in the Supplementary Materials shows the running time of the proposed method with the $\ell_0$-regularization penalty, based on $\lambda$ with length ten and fixed $\alpha$ and $h$. Overall, the computation time increased linearly with the number of covariates. When

Table 3. Performance of the bandwidth selection procedure for DB-hazard.

| | Setting (a) | | Setting (b) | | Setting (a) | | Setting (b) | |
|---|---|---|---|---|---|---|---|---|
| | Selected | Best[1] | Selected | Best | Selected | Best | Selected | Best |
| | $n = 100,\ p_n = 20$ | | | | $n = 200,\ p_n = 20$ | | | |
| Bandwidth | 0.080 | 0.096 | 0.093 | 0.100 | 0.090 | 0.081 | 0.093 | 0.071 |
| SSE[2] | 3.06 | 2.67 | 4.96 | 4.37 | 1.92 | 1.66 | 3.22 | 2.83 |
| | $n = 100,\ p_n = 50$ | | | | $n = 200,\ p_n = 50$ | | | |
| Bandwidth | 0.080 | 0.089 | 0.081 | 0.095 | 0.089 | 0.080 | 0.088 | 0.084 |
| SSE | 3.12 | 2.65 | 6.08 | 5.24 | 2.16 | 1.73 | 3.78 | 3.25 |
| | $n = 100,\ p_n = 1,000$ | | | | $n = 200,\ p_n = 1,000$ | | | |
| Bandwidth | 0.056 | 0.085 | 0.055 | 0.113 | 0.059 | 0.086 | 0.061 | 0.104 |
| SSE | 4.57 | 3.89 | 12.59 | 11.31 | 2.83 | 2.19 | 8.06 | 6.90 |

[1]: defined as the bandwidth leading to the smallest SSE; [2]: sum of squared errors.

$p_n = 1,000$ and $n = 200$, the running time is 634 seconds, with a total of $p_n q_n = 5,000$ parameters.

We also evaluated the performance of the proposed method based on a different kernel function, namely, a Gaussian kernel; similar results were obtained. In addition, the impact of various numbers of basis functions was considered by using quadratic B-splines with 5, 7, and 10 interior knots, corresponding to $q_n = 8,\ 10,\ 13$, respectively. We observed increases in the SSE and the number of identified variables as the number of basis functions increased. Note that $\beta_j(t)$ is a linear combination of basis functions. To obtain $\beta_j(t) = 0$, all elements in the coefficient vector $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jq_n})^T$ have to be zero. Thus, it is more likely that we will obtain nonzero estimates if we have a greater number of basis functions. After increasing $n$ to 200, the performance improved, which may suggest we need greater sample sizes when describing a more complicated function $\beta_j(t)$ with more basis functions. Details of the above numerical studies are given in the Supplementary Material.

## 6. Application

Recent research has suggested that brain imaging biomarkers play an important role in predicting the onset of neurodegenerative disorders, and particularly HD (e.g., Feigin et al. (2007); Paulsen (2010)). A diagnosis of HD is made based on a neurological examination that indicates with 99% confidence that the extrapyramidal movement disorder is consistent with HD. By the time clinical symptoms are apparent, subjects may already be in an advanced disease stage. Therefore, identifying biomarkers that may be informative for the early prediction of disease onset preceding clinical diagnosis has important implications for recruiting pre-symptomatic subjects for clinical trials related to early intervention (Paulsen (2010)). Current research indicates that neuroimaging biomarkers are among the most promising for predicting the time to HD onset (Paulsen et al. (2014a,b)). In this work, we analyze data collected from a newly completed, large natural history study on the disease progression, PREDICT-HD (Paulsen et al. (2014b)), in individuals who carry expanded CAG repeats and are destined to develop HD. CAG repeat length is inversely related to age at onset, but the exact onset age varies. We aim to predict the time-to-onset of HD using structural MRI region of interest (ROI) volumetric measures for subjects without a diagnosis at the baseline, but with expanded CAG repeats. The regional summary volumetric measures were created using a fully automated procedure and were preprocessed using Freesurfer 5.2 (`http://surfer.nmr.mgh.harvard.edu`). Details on the imaging biomarker preprocessing can be found in Paulsen et al. (2014a).

Our analysis data consist of 866 subjects who had expanded CAG repeats at the huntingtin gene (MacDonald et al. (1993)) without a clinical diagnosis at the baseline. These subjects will develop HD during their lifetime owing to the expanded repeat length at the HD gene, but the exact age of onset is unknown. The median follow-up time was 4.0 years, with an average of 1.9 follow-ups per subject. Imaging biomarkers were measured approximately bi-annually, with some random variation and, thus, were obtained less frequently than the clinical measures of the time-to-diagnosis outcome (assessed annually). Figure S2 in the Supplementary Materials displays the number of subjects with available clinical measures (time-to-diagnosis outcome) and longitudinal imaging measurements at follow ups, indicating the shows sparse measurements of imaging biomarkers at times (e.g., 18 months after the baseline). The biomarkers and clinical assessments included in the analyses are as follows: baseline CAP score (scaled product of CAG repeats length at the HD gene and baseline age; Zhang et al. (2011)) eight demographic or clinical measures (gender, baseline total motor score, TMS, from the United Huntington's Disease Rating Scale; and cognitive and functioning measures); and whole-brain MRI volumetric biomarkers, including 58 subcortical region of interest (ROI) measures and 68 cortical ROIs.

Figure S3 in the Supplementary Materials shows the heatmaps of the 136 features measured at the baseline and at the last visit for 142 subjects diagnosed with HD during the study (converters) and 390 subjects who remained free of HD (nonconverters). The goal is to simultaneously select informative biomarkers, estimate their time-dependent effect profiles, and indentifying the combination that tracks with HD conversion. In Figure S3, no single feature can definitively distinguish converters from nonconverters, suggesting that a multi-dimensional approach that considers all features will outperform univariate analyses. However, covariation among features is also prevalent; thus, a multi-dimensional approach needs to account for high dimensionality and covariation patterns through regularization. Most biomarkers and subjects show a smooth trend between the first and last visits. Appropriately smoothing over time and borrowing information from nearby measurements is essential to predicting the conversion events, especially for time points where imaging measurements are sparse. Lastly, several features show subtle differences in discriminating converters from nonconverters at the first and last visits, suggesting that their prognostic power may vary over time (i.e., greater discriminant power when using recent imaging biomarker measurements).

In the analysis, we applied DB-hazard with an $\ell_0$-penalty and Laplacian network regularization to the data. We constructed the imaging biomarkers' covariation network (He, Chen and Evans (2008)) based on an independent control group (no expanded repeat length at the huntingtin gene) in PREDICT-HD. The obtained covariation pattern was introduced in the Laplacian regularization. For DB-hazard, the estimation follows the same procedure as that described in Section 2, using all longitudinal imaging measurements over time. We used five-fold cross-validation to select the bandwidth and tuning parameters. We compared DB-hazard with using baseline data alone ("Baseline"), and with the last value carried forward ("LVCF"). All features were standardized before fitting the model. The running time of the proposed DB-hazard with the $\ell_0$-penalty for this analysis is 927 seconds.

Table S4 in Supplementary Material summarizes the area under the ROC curve (AUC), time-dependent sensitivity (SEN), specificity (SPE), positive predictive value (PPV), and negative predictive value (NPV) at a given time, where the threshold is obtained by optimizing Youden's index. The results show that, overall, DB-hazard outperforms the two alternatives. For example, the AUC of DB-hazard is the highest of the three methods at all time points. DB-hazard outperforms LVCF significantly, which may be due to the bias of LVCF. When compared with using only baseline measurements, DB-hazard performs better, especially at later years (e.g., year 6), demonstrating the advantages of using current values of longitudinal biomarkers to update longer-term predictions. In a recent independent study, Long et al. (2016) compared Harrell's C-index (i.e., AUC) for ten models, including data from four studies on the progression of HD: PREDICT-HD, TRACK (Tabrizi et al. (2013)), COHORT (Dorsey et al.

(2012)), and REGISTRY (Handley et al. (2012)). The best model had a median AUC of 0.87, which is similar to our baseline-only analyses, but lower than using DB-hazard to incorporate all available longitudinal measures. This comparison further supports the information from incorporating all available time-dependent structural MRI biomarkers and clinical assessments. Our analysis is the first to use longitudinal imaging biomarkers to track HD conversion.

With regard to other measures, the specificity for DB-hazard by year 6 is 0.873, whereas it is only 0.651 and 0.810 for Baseline and LVCF, respectively. Similarly, the PPV estimated by DB-hazard by year 6 (0.540) is higher than those of the other two methods (Baseline: 0.278; LVCF: 0.414). The high time-dependent sensitivity and specificity of the DB-hazard combined HD biomarker signature suggests that it is a valuable tool for tracking clinically defined disease onset. The higher PPV at years 4 and 6 demonstrate the valuable information gain from using longitudinal imaging measures to improve predictions. However, the moderate magnitude of the PPV implies that additional biomarkers (e.g., genomic or proteomic biomarkers; Langfelder et al. (2016)) may need to be identified to further improve the prediction performance.

From 136 features, DB-hazard identified six nonimaging covariates (i.e., CAP, total functional capacity (TFC), baseline total motor score (baseline TMS), symbol digit modality test (SDMT), Stroop color naming total, and Stroop word reading total), and six imaging biomarkers (i.e., left Caudate, left and right Putamen, left Pallidum, left Accumbens, left lateral occipital volume) as informative for predicting time-to-onset of HD, including five subcortical measures and one cortical measure (left lateral occipital volume). Figure S4 in the Supplementary Material shows the heatmaps of the selected features, where they are seen to better distinguish converters from nonconverts than the nonselected noise features in Figure S3 do. In addition, through the use of network regularization, redundancy among the features was removed, with 12 features achieves a high AUC. The discriminant power of some features changes between the first and last visits changes at the first and last visit (e.g., TFC). It is interesting that a greater number of subcortical ROIs were selected and only one cortical ROI was selected when both were included in DB-hazard as candidates. This result is consistent with clinical research suggesting that regional atrophy of subcortical grey matter is an important biological feature of HD progression (Ross et al. (2014)). All subcortical ROIs identified have been ranked as top candidate biomarkers in existing clinical research (Paulsen et al. (2014a)); however, the cortical measure has not been reported previously. In Figure S5, we present the effects of top-ranked measures, estimated by DB-hazard, and the corresponding 95% confidence intervals, obtained by bootstrapping 100 times. The baseline TMS has the strongest effect and a similar shape to that of the baseline CAP. The effects of TMS, TFC, and SDMT increase in the first two years, with the largest effect recorded between years 2 and 3. Two imaging biomarkers, left Caudate and left Putamen, show a similar effect size to that of the baseline CAP, TFC, and SDMT. The measure with the largest effect is the baseline TMS.

## 7. Discussion

We have proposed methods for fitting a time-varying hazards model using sparsely measured time-dependent covariates. In contrast to existing methods (e.g., Paulsen et al. (2014b); Long et al. (2016)), we use all longitudinal measures (both imaging biomarkers and clinical measures at follow-ups) to perform analyses, which was not possible previously owing to imbalanced assessments of imaging measures. Our simulation studies show that smoothing over longitudinal measurements across subjects improves performance over that of the commonly used LVCF and baseline-only analysis. In addition, the proposed DB-hazard with $\ell_0$-penalty, solved using a two-step procedure, substantially outperforms methods without the hard-thresholding or when using the group LASSO alone, in terms of both estimation and selection accuracy. We prove the theoretical oracle property under local kernel smoothing, which has not been investigated in the literature previously. We also demonstrate substantial improvement compared with current standards by applying our method to data from a real-world study (PREDICT-HD).

Here, we assume a constant network structure in equation (4.1). It would be interesting and challenging to explore a time-varying network $\mathcal{L}_n^*(t)$. One method of doing so would be to incorporate a time-varying Gaussian graphical model (Zhou, Lafferty and Wasserman (2010); Razavian et al. (2010)), where the time-varying network $\widehat{\mathcal{L}_n}^*(t)$ can be obtained from $\widehat{\mathcal{L}_n}^*(t)^{-1} = \arg\max_\Theta \log|\Theta(t)| - \mathrm{tr}(\boldsymbol{R}(t)\Theta(t)) - \rho\|\Theta(t)\|_1$, where $\boldsymbol{R}(t)$ is the weighted correlation matrix and can be calculated from the weighted covariance matrix

$$\boldsymbol{R}'(t) = \frac{\sum_{i=1}^n \sum_{v=1}^{n_i} K_{h_n}(t_{iv} - t)\boldsymbol{X}_i(t_{iv})\boldsymbol{X}_i(t_{iv})^T}{\sum_{i=1}^n \sum_{v=1}^{n_i} K_{h_n}(t_{iv} - t)}.$$

Another extension would be to study the effect of a time-varying network on the disease outcome, and to distinguish between the effects of the longitudinal measurements and those of their time-varying network.

Lastly, we focus on time-to-event data. However, the proposed approach can be extended easily to other types of outcomes. For example, we cold would replace the log-partial likelihood function with the least squares loss function for a continuous outcome, or with an appropriate likelihood for generalized outcomes.

## Supplementary Material

The online Supplementary Material includes proofs of Lemma 1 and Theorem 1, and additional information related to the simulation and real-data analysis.

## Acknowledgments

This research is supported by the U.S. NIH grants NS073671, GM124104, and MH117458.

## References

Alexander-Bloch, A., Giedd, J. N. and Bullmore, E. (2013). Imaging structural co-variance between human brain regions. *Nature Reviews Neuroscience* **14**, 322–336.

Andersen, P. K. and Liestol, K. (2003). Attenuation caused by infrequently updated covariates in survival analysis. *Biostatistics* **4**, 633–649.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* **3**, 1–122.

Bradic, J., Fan, J. and Jiang, J. (2011). Regularization for coxs proportional hazards model with np-dimensionality. *The Annals of Statistics* **39**, 3092–3120.

Bullmore, E. T. and Bassett, D. S. (2011). Brain graphs: graphical models of the human brain connectome. *Annual Review of Clinical Psychology,* **7**, 113–140.

Cao, H., Churpek, M. M., Zeng, D. and Fine, J. P. (2015). Analysis of the proportional hazards model with sparse longitudinal covariates. *Journal of the American Statistical Association* **110**, 1187–1196.

Cao, H., Zeng, D. and Fine, J. P. (2014). Regression analysis of sparse asynchronous longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**, 775–776.

Chen, T., Wang, Y., Chen, H., Marder, K. and Zeng, D. (2014). Targeted local support vector machine for age-dependent classification. *Journal of the American Statistical Association* **109**, 1174–1187.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Alberti,M. S. and Killiany,R. S. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31**, 968–980.

Dorsey, E. R., Investigators, H. S. G. C. et al. (2012). Characterization of a large group of individuals with huntington disease and their relatives enrolled in the cohort study. *PLoS*

*One* **7**, e29522.

Eidelberg, D. and Surmeier, D. J. (2011). Brain networks in Huntington disease. *Journal of Clinical Investigation* **121**, 484–492.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *Information Theory, IEEE Transactions* **57**, 5467–5484.

Feigin, A., Tang, C., Ma, Y., Mattis, P., Zgaljardic, D., Guttman, M., Paulsen, J., Dhawan, V. and Eidelberg, D. (2007). Thalamic metabolism and symptom onset in preclinical huntington's disease. *Brain* **130**, 2858–2867.

Fleming, T. R. and Harrington, D. P. (2011). *Counting Processes and Survival Analysis*. John Wiley & Sons.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical Software* **33**, 1–22.

Gould, L. A., Boye, M. E., Crowther, M. J., Ibrahim, J. G., Quartey, G., Micallef, S. and Bois, F. Y. (2014). Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. *Statistics in Medicine* **34**, 2181–2195.

Handley, O., Landwehrmeyer, B., Committee, R. S., Investigators, E. R. et al. (2012). European huntington's disease network registry: current status. *Journal of Neurology, Neurosurgery & Psychiatry* **83**, A47–A47.

He, Y., Chen, Z. and Evans, A. (2008). Structural insights into aberrant topological patterns of large-scale cortical networks in alzheimer's disease. *The Journal of Neuroscience* **28**, 4756–4766.

Honda, T. and Härdle, W. K. (2014). Variable selection in cox regression models with varying coefficients. *Journal of Statistical Planning and Inference* **148**, 67–81.

Honda, T. and Yabe, R. (2017). Variable selection and structure identification for varying coefficient cox models. *Journal of Multivariate Analysis* **161**, 103–122.

Huang, J., Horowitz, J. L. and Wei, F. (2010). Variable selection in nonparametric additive models. *The Annals of Statistics* **38**, 2282–2313.

Lange, K. (2013). The MM algorithm. *Optimization* 185–219. Springer.

Langfelder, P., Cantle, J. P., Chatzopoulou, D., Wang, N., Gao, F., Al-Ramahi, I., Lu, X.-H., Ramos, E. M., El-Zein, K., Zhao, Y., Deverasetty, S., Tebbe, A., Schaab, C., Lavery D. J., Howland, D., Kwak, S., Botas, J., Aaronson J. S., Rosinski, J., Coppola, G ., Horvath, S., Yang, X. W.(2016). Integrated genomics and proteomics define huntingtin cag length-dependent networks in mice. *Nature Neuroscience* **19**, 623–633.

Li, X., Xie, S., Zeng, D. and Wang, Y. (2017). Efficient $\ell_0$-norm feature selection based on augmented and penalized minimization. *Statistics in Medicine*, in press.

Liu, J., Huang, J., Ma, S. and Wang, K. (2013). Incorporating group correlations in genome-wide association studies using smoothed group lasso. *Biostatistics* **14**, 205–219.

Liu, M., Zhang, D., Shen, D., Initiative, A. D. N. et al. (2014). Identifying informative imaging biomarkers via tree structured sparse learning for ad diagnosis. *Neuroinformatics* **12**, 381–394.

Liu, X. and Zeng, D. (2013). Variable selection in semiparametric transformation models for right-censored data. *Biometrika* **100**, 859–876.

Long, J. D., Langbehn, D. R., Tabrizi, S. J., Landwehrmeyer, B. G., Paulsen, J. S., Warner, J. and Sampaio, C. (2016). Validation of a prognostic index for huntington's disease. *Movement Disorders* **32**, 256–263.

MacDonald, M. E., Ambrose, C. M., Duyao, M. P., Myers, R. H., Lin, C., Srinidhi, L., Barnes, G., Taylor, S. A., James, M., Groot, N. et al. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on huntington's disease chromosomes. *Cell* **72**, 971–983.

Meier, L., Van De Geer, S. and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 53–71.

Parikshak, N. N., Gandal, M. J. and Geschwind, D. H. (2015). Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nature Reviews Genetics* **16**, 441–458.

Paulsen, J. S. (2010). Early detection of huntington's disease. *Future Neurology* **5**, 85–104.

Paulsen, J. S., Long, J. D., Johnson, H. J., Aylward, E. H., Ross, C. A., Williams, J. K., Nance, M. A., Erwin, C. J., Westervelt, H. J., Harrington, D. L. et al. (2014a). Clinical and biomarker changes in premanifest huntington disease show trial feasibility: a decade of the predict-hd study. *Frontiers in Aging Neuroscience* **6**, 78.

Paulsen, J. S., Long, J. D., Ross, C. A., Harrington, D. L., Erwin, C. J., Williams, J. K., Westervelt, H. J., Johnson, H. J., Aylward, E. H., Zhang, Y. et al. (2014b). Prediction of manifest huntington's disease with clinical and imaging measures: a prospective observational study. *The Lancet Neurology* **13**, 1193–1201.

Prentice, R. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69**, 331–342.

Razavian, N. S., Moitra, S., Kamisetty, H., Ramanathan, A. and Langmead, C. J. (2010). Time-varying gaussian graphical models of molecular dynamics data. *Technical Report.*

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* **67**, 819–829.

Ross, C. A., Aylward, E. H., Wild, E. J., Langbehn, D. R., Long, J. D., Warner, J. H., Scahill, R. I., Leavitt, B. R., Stout, J. C., Paulsen, J. S. et al. (2014). Huntington disease: natural history, biomarkers and prospects for therapeutics. *Nature Reviews Neurology* **10**, 204–216.

Ryan, N. P., Catroppa, C., Cooper, J. M., Beare, R., Ditchfield, M., Coleman, L., Silk, T., Crossley, L., Beauchamp, M. H. and Anderson, V. A. (2015). The emergence of age-dependent social cognitive deficits after generalized insult to the developing brain: A longitudinal prospective analysis using susceptibility-weighted imaging. *Human Brain Mapping* **36**, 1677–1691.

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

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics* **22**, 231–245.

Stuart, J. M., Segal, E., Koller, D. and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255.

Tabrizi, S. J., Scahill, R. I., Owen, G., Durr, A., Leavitt, B. R., Roos, R. A., Borowsky, B., Landwehrmeyer, B., Frost, C., Johnson, H. et al. (2013). Predictors of phenotypic progression and disease onset in premanifest and early-stage huntington's disease in the track-hd study: analysis of 36-month observational data. *The Lancet Neurology* **12**, 637–649.

Taylor, J. M., Park, Y., Ankerst, D. P., Proust-Lima, C., Williams, S., Kestin, L., Bae, K., Pickles, T. and Sandler, H. (2013). Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics* **69**, 206–213.

Tsiatis, A. A. and Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika* **88**, 447–458.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* **14**, 809–834.

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

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**, 894–942.

Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science* **4**, 576–593.

Zhang, Y., Long, J. D., Mills, J. A., Warner, J. H., Lu, W. and Paulsen, J. S. (2011). Indexing disease progression at study entry with individuals at-risk for huntington disease. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **156**, 751–763.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541–2563.

Zhou, S., Lafferty, J. and Wasserman, L. (2010). Time varying undirected graphs. *Machine Learning* **80**, 295–319.

Department of Biostatistics, Mailman School of Public Health, Columbia University, NY 10032, USA.

E-mail: XL2473@cumc.columbia.edu

Department of Biostatistics, University of North Carolina at Chapel Hill, NC 27599, USA.

E-mail: quefeng@email.unc.edu

Department of Biostatistics, University of North Carolina at Chapel Hill, NC 27599, USA

E-mail: dzeng@email.unc.edu

Department of Neurology, College of Physicians and Surgeons, Columbia University, NY 10032, USA.

E-mail: ksm1@columbia.edu

Department of Neurology, Psychiatry and Psychology, University of Iowa, IA 52242, USA

E-mail: jane-paulsen@uiowa.edu

Department of Biostatistics, Mailman School of Public Health, Columbia University, NY 10032.

E-mail: yw2016@columbia.edu

(Received August 2017; accepted September 2018)