Lifetime Data Analysis manuscript No.
(will be inserted by the editor)

A global kernel estimator for partially linear varying coefficient additive hazards models

Abstract We study kernel-based estimation methods for partially linear varying coefficient additive hazards models, where the effects of one type of covariates can be modified by another. Existing kernel estimation methods for varying coefficient models often use a "local" approach, where only a small local neighborhood of subjects are used for estimating the varying coefficient functions. Such a local approach, however, is generally inefficient as information about some non-varying nuisance parameter from subjects outside the neighborhood is discarded. In this paper, we develop a "global" kernel estimator that simultaneously estimates the varying coefficients over the entire domains of the functions, leveraging the non-varying nature of the nuisance parameter. We establish the consistency and asymptotic normality of the proposed estimators. The theoretical developments are substantially more challenging than those of the local methods, as the dimension of the global estimator increases with the sample size. We conduct extensive simulation studies to demonstrate the feasibility and superior performance of the proposed methods compared with existing local methods and provide an application to a motivating cancer genomic study.

Keywords Censored data · Kernel smoothing · Semiparametric model · Survival analysis

1 Introduction

In biomedical studies, we are often interested in the association between covariates and the time to some disease event, such as disease recurrence or death. A popular approach is to model the effect of the covariates on the event time through the hazard function, which represents the instantaneous risk of the event at a time point given that the event has not yet occurred. Such models capture the dynamics of the event occurrence process. Among such models, the most popular choice is the proportional hazards model (Cox, 1972), which assumes that the covariates act multiplicatively on the hazard function. The proportional hazards model is popular due to its intuitive interpretation of regression parameters and the availability of the partial likelihood, which enables a simple procedure for estimation and inference of the regression parameters without parametric assumptions on the baseline hazard function. Another model that directly captures the effects of covariates on the hazard is the additive hazards model (Aalen, 1989; Lin and Ying, 1994), which assumes that the covariates act additively to the hazard function. In particular, the conditional hazard function of an event time given a possibly time-varying covariate $\boldsymbol{X}(\cdot)$ takes the form

$$\lambda(t \mid \boldsymbol{X}) = \lambda(t) + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}(t),$$

where $\lambda(\cdot)$ is an unspecified positive function and $\boldsymbol{\beta}$ represents the covariate effect. Since the proportional hazards and additive hazards models assume different structures of association between the covariates and event time, both models would be of practical interest, especially in the data exploration stage. While the choice between the models should be based on the scientific question of interest and empirical data fit, the additive hazards model has the advantage of having closed-form estimators for the regression parameters.

In some applications, the outcome are influenced by the interplay among various covariates. For example, disease outcomes are often influenced by complex biological processes that involve various genomic features. Also, Hunter (2005) suggested that the effect of a genotype on the risk of disease development may differ based on environmental factors. In addition, Landi et al. (2008) demonstrated that the effects of some gene expressions on the risk of lung cancer mortality vary with tobacco consumption. To account for such interaction effects among covariates, we can formulate $\boldsymbol{\beta}$ as a function of another covariate $\boldsymbol{W}$. This varying coefficient structure has been widely applied to various models, including (generalized) linear models (Hastie and Tibshirani, 1993; Fan and Zhang, 1999; Fan et al., 2006), proportional hazards models (Tian et al., 2005), and additive hazards models (Aalen, 1989; McKeague and Sasieni, 1994; Yin et al., 2008; Li et al., 2007; Qu et al., 2018).

Varying coefficient additive hazards models generally take the form

$$\lambda(t \mid \boldsymbol{W}, \boldsymbol{X}) = \lambda(t) + \boldsymbol{\beta}(\boldsymbol{W})^{\mathrm{T}}\boldsymbol{X}(t), \tag{1}$$

where $\boldsymbol{\beta}(\boldsymbol{w})$ is a vector of coefficient functions that represents the effect of $\boldsymbol{X}(t)$ at $\boldsymbol{W} = \boldsymbol{w}$. In this model, the effect of $\boldsymbol{X}(t)$ is modified by $\boldsymbol{W}$, representing a potentially nonlinear interaction between $\boldsymbol{W}$ and each component of $\boldsymbol{X}(t)$. Model (1) includes various models as special cases. For example, when W is a univariate variable that represents time, that is, $W = t$, the model reduces to the additive hazards model with time-varying coefficients (Aalen, 1989). This model allows the covariate effects to vary over time, capturing time-dependent relationships between the covariates and the hazard rate. When the coefficient $\boldsymbol{\beta}$ is constant in model (1), the model simplifies to the additive hazards model with constant coefficients (Lin and Ying, 1994). Partially linear models can also be considered within the framework of model (1) by setting some components of $\boldsymbol{\beta}$ to be constant (McKeague and Sasieni, 1994; Yin et al., 2008).

In this paper, we focus on kernel smoothing techniques for the estimation of varying coefficient additive hazard models. The basic idea underlying kernel methods is to estimate a function at any given value based on subjects associated with a local neighborhood of the value. Kernel-based methods have been extensively applied to the estimation of varying coefficient survival models, such as the proportional hazards models (Tian et al., 2005; Fan et al., 2006) and the additive hazards models (Li et al., 2007; Yin et al., 2008; Qu et al., 2018). In these methods, the proposed estimators are obtained by taking existing estimators (for non-varying coefficient models) and including the kernel function as a weight to each observation. Because these estimators for the varying coefficient function at any point are based on a small local neighborhood of subjects, we generally refer to them as "local" kernel estimation methods. While local methods are easy to implement and can be straight-forwardly applied to various semiparametric models, they may suffer efficiency loss because information contained in subjects outside the neighborhood is discarded. To reduce the efficiency loss, Chen, Lin, and Zhou (2012) proposed global kernel estimation methods for the varying coefficient proportional hazards models that includes all subjects in the estimation of the varying coefficients. While their estimators improve efficiency, their methods could be computationally demanding due to the iterative nature of the algorithm.

We propose a global kernel estimator for the (partially linear) varying coefficient additive hazards model. Instead of separately estimating the varying coefficient function at any given covariate value using a local neighborhood of subjects, the proposed method estimates the varying coefficient function over a range of covariate values simultaneously and thus involves all subjects. The local method suffers efficiency loss because, effectively, only a small neighborhood of subjects are used to simultaneously estimate the varying coefficient function at any given value and the baseline hazard function. However, the baseline hazard function is shared among all subjects and could be more efficiently estimated by using all subjects. The proposed method takes into account this sharing of a common baseline hazard function to improve efficiency. The proposed estimator is the solution to a system of linear equations and has a closed-form solution, and thus it is computationally more efficient than existing global kernel estimators.

A major challenge in deriving the asymptotic properties of the proposed estimator is that it solves a system of linear equations, and the estimator at a specific covariate value depends on the estimators at other covariate values. Also, the number of covariate values considered, and thus the number equations to be solved, increases with the sample size. In contrast, the local estimator for the varying coefficient function at any point solves a single linear equation, allowing direct applications of standard counting process and martingale theories to establish asymptotic normality.

The remainder of this paper is structured as follows. In Section 2, we propose kernel estimation methods for the varying coefficient additive hazards models. In Section 3, we establish the asymptotic properties of the proposed estimator under some regularity conditions. The empirical performance of the proposed method is examined and compared to alternative methods through extensive simulation studies in Section 4. In Section 5, we demonstrate the application of the proposed method to cancer genomic data. We conclude with a discussion of our findings and topics for further research in Section 6. Some technical details and additional simulation results are provided in the Online Resource.

2 Model and estimation

2.1 Additive hazards model with varying covariate effects

We first consider an additive hazards model with only varying covariate effects. Let $\boldsymbol{W} \equiv (W_1, \dots, W_q)^{\mathrm{T}}$ and $\boldsymbol{X} \equiv (X_1, \dots, X_p)^{\mathrm{T}}$ be two sets of covariates and $\widetilde{T}$ be the event time. We assume the following hazard function for $\widetilde{T}$:

$$\lambda(t \mid \boldsymbol{W}, \boldsymbol{X}) = \lambda(t) + \boldsymbol{\beta}(\boldsymbol{W})^{\mathrm{T}} \boldsymbol{X}. \tag{2}$$

If $X_j = 1$, then we set $\beta_j(\boldsymbol{w}_0) = 0$ for some $\boldsymbol{w}_0$ within the support of $\boldsymbol{W}$ for model identifiability. Note that in contrast with model (1), we focus on time-independent covariates. We allow $\widetilde{T}$ to be right-censored and let C be the censoring time. Assume that C and $\widetilde{T}$ are conditionally independent given $(\boldsymbol{W}, \boldsymbol{X})$. Let $T \equiv \min(\widetilde{T}, C)$ be the observed time and $\Delta \equiv I(\widetilde{T} \leq C)$ be the event indicator. For a sample of size n, we observe $(T_i, \Delta_i, \boldsymbol{W}_i, \boldsymbol{X}_i)$ for $i = 1, \dots, n$. Define $N_i(t) \equiv \Delta_i I(T_i \leq t)$ and $Y_i(t) \equiv I(T_i \geq t)$ as the observed event process and at-risk process for the ith subject at time t, respectively. In model (2), if $\boldsymbol{\beta}$ is constant, then we can estimate it using the estimator of Lin and Ying (1994):

$$\left[\sum_{i=1}^{n} \int_0^{\tau} Y_i(t) \left\{ \boldsymbol{X}_i - \overline{\boldsymbol{X}}(t) \right\}^{\otimes 2} \mathrm{d}t \right]^{-1} \sum_{i=1}^{n} \int_0^{\tau} \left\{ \boldsymbol{X}_i - \overline{\boldsymbol{X}}(t) \right\} \mathrm{d}N_i(t),$$

where τ is the maximum follow-up time, $\overline{\boldsymbol{X}}(t) \equiv \sum_{i=1}^{n} \boldsymbol{X}_i Y_i(t) / \sum_{i=1}^{n} Y_i(t)$ is the average covariate vector among subjects at risk at time t with the convention that $0/0 = 0$, and for any vector $\boldsymbol{a}$, $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}$.

To motivate the proposed estimation method for model (2), we first consider the special case that $\boldsymbol{W}$ takes discrete values $\boldsymbol{w}_1, \dots, \boldsymbol{w}_m$. In this case, we can define a vector of covariates $\boldsymbol{B} \equiv (I(\boldsymbol{W} = \boldsymbol{w}_1)\boldsymbol{X}^{\mathrm{T}}, \dots, I(\boldsymbol{W} = \boldsymbol{w}_m)\boldsymbol{X}^{\mathrm{T}})^{\mathrm{T}}$. Fitting the standard additive hazards model with covariate $\boldsymbol{B}$ is equivalent to fitting model (2), and the Lin and Ying estimator can be applied to estimate $\boldsymbol{\beta} \equiv (\boldsymbol{\beta}(\boldsymbol{w}_1)^{\mathrm{T}}, \dots, \boldsymbol{\beta}(\boldsymbol{w}_m)^{\mathrm{T}})^{\mathrm{T}}$. The explicit formulation of the resulting estimator is provided in Appendix A1. For the case of a continuous $\boldsymbol{W}$, we mimic the estimator in the discrete case and apply the kernel smoothing technique. We define a grid $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m)$ over the support of $\boldsymbol{W}$ and estimate the values of $\boldsymbol{\beta}$ over this grid simultaneously. In particular, we estimate $(\boldsymbol{\beta}(\boldsymbol{w}_1)^{\mathrm{T}}, \ldots, \boldsymbol{\beta}(\boldsymbol{w}_m)^{\mathrm{T}})^{\mathrm{T}}$ by $\widehat{\boldsymbol{\beta}} \equiv \boldsymbol{V}^{-1}\boldsymbol{b}$, where $\boldsymbol{V}$ is a block matrix with $\boldsymbol{V}(\boldsymbol{w}_k, \boldsymbol{w}_\ell)$ as its (k, ℓ)th block, $\boldsymbol{b} = (\boldsymbol{b}(\boldsymbol{w}_1)^{\mathrm{T}}, \ldots, \boldsymbol{b}(\boldsymbol{w}_m)^{\mathrm{T}})^{\mathrm{T}}$, and

$$\boldsymbol{b}(\boldsymbol{w}_k) = \frac{1}{nm}\sum_{i=1}^n \int_0^\tau \left\{ K_H(\boldsymbol{W}_i - \boldsymbol{w}_k)\boldsymbol{X}_i - \sum_{j=1}^m K_H(\boldsymbol{W}_i - \boldsymbol{w}_j)\overline{\boldsymbol{X}}(t, \boldsymbol{w}_k) \right\} \mathrm{d}N_i(t),$$

$$\begin{aligned}\boldsymbol{V}(\boldsymbol{w}_k, \boldsymbol{w}_\ell) = {} & \frac{I(\boldsymbol{w}_k = \boldsymbol{w}_\ell)}{nm}\sum_{i=1}^n \int_0^\tau K_H(\boldsymbol{W}_i - \boldsymbol{w}_k)Y_i(t)\boldsymbol{X}_i^{\otimes 2}\,\mathrm{d}t \\ & - \frac{1}{nm}\sum_{i=1}^n \int_0^\tau \sum_{j=1}^m K_H(\boldsymbol{W}_i - \boldsymbol{w}_j)Y_i(t)\overline{\boldsymbol{X}}(t, \boldsymbol{w}_k)\overline{\boldsymbol{X}}(t, \boldsymbol{w}_\ell)^{\mathrm{T}}\,\mathrm{d}t,\end{aligned}$$

$$\overline{\boldsymbol{X}}(t, \boldsymbol{w}_k) = \frac{\sum_{i=1}^n K_H(\boldsymbol{W}_i - \boldsymbol{w}_k)Y_i(t)\boldsymbol{X}_i}{\sum_{i=1}^n \sum_{j=1}^m K_H(\boldsymbol{W}_i - \boldsymbol{w}_j)Y_i(t)}.$$

In the above formulation, $K_H(\boldsymbol{x}) \equiv |\boldsymbol{H}|^{-1/2}K(\boldsymbol{H}^{-1/2}\boldsymbol{x})$ is a kernel function with bandwidth matrix $\boldsymbol{H}$. Essentially, we take the estimator for the discrete case (given in Appendix A1) and replace the summations over the subjects with $\boldsymbol{W}_i = \boldsymbol{w}_k$ by a weighted sum over all subjects, where the weight is defined based on the distance between individual $\boldsymbol{W}_i$'s and $\boldsymbol{w}_k$. For summations over all subjects in the estimator of the discrete case, one possible approach is to keep them unchanged in the continuous case. Nevertheless, we propose to multiply $m^{-1}\sum_{j=1}^m K_H(\boldsymbol{W}_i - \boldsymbol{w}_j)$ to each term in the summation. Although the two resulting estimators are asymptotically equivalent, in our experience, the estimator with $m^{-1}\sum_{j=1}^m K_H(\boldsymbol{W}_i - \boldsymbol{w}_j)$ is more robust to the choice of the grid.

From the proof of Theorem 1 presented in Appendix A3, we can see that a consistent estimator of $\boldsymbol{\beta}(\cdot)$ can be obtained by solving an equation $\mathcal{V}\boldsymbol{\beta} = \mathcal{C}$, where $\mathcal{V}$ is a linear operator and $\mathcal{C}$ is a function that depends on the observed data. The proposed method effectively takes this equation and replaces the left-hand side by the product between a matrix and $\boldsymbol{\beta}$ evaluated at the grid points and replaces the right-hand side by a vector that consists of $\mathcal{C}$ evaluated at the grid points. Essentially, we approximate the integral equations involved in $\mathcal{V}\boldsymbol{\beta}$ by summations over the grid points and solve for $\widehat{\boldsymbol{\beta}}$ at these grid points only. Increasing the number of grid points generally reduces the approximation error but increases the computational cost. To preserve computational efficiency, we choose a set of sparse grid points that adequately represents the distribution of $\boldsymbol{W}$. We suggest using a quantile-based grid, in which the grid points are positioned at the empirical quantiles of $\boldsymbol{W}$. Alternatively, an evenly spaced grid that uniformly covers the support of $\boldsymbol{W}$ can be adopted. For values of $\boldsymbol{\beta}$ outside the grid points, we can estimate them using linear interpolation based on the values of $\widehat{\boldsymbol{\beta}}(\boldsymbol{w})$ at the nearest grid points.

When $m = 1$, the proposed estimator reduces to a local estimator, which is similar in spirit to the estimator of Yin et al. (2008). In particular, $\boldsymbol{\beta}(\boldsymbol{w})$ is estimated by $\boldsymbol{V}(\boldsymbol{w},\boldsymbol{w})^{-1}\boldsymbol{b}(\boldsymbol{w})$, with

$$\boldsymbol{b}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} K_H(\boldsymbol{W}_i-\boldsymbol{w})\left\{\boldsymbol{X}_i - \frac{\sum_{j=1}^{n} K_H(\boldsymbol{W}_j-\boldsymbol{w})Y_j(t)\boldsymbol{X}_j}{\sum_{j=1}^{n} K_H(\boldsymbol{W}_j-\boldsymbol{w})Y_j(t)}\right\}\mathrm{d}N_i(t),$$

$$\boldsymbol{V}(\boldsymbol{w},\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} K_H(\boldsymbol{W}_i-\boldsymbol{w})Y_i(t)\left\{\boldsymbol{X}_i - \frac{\sum_{j=1}^{n} K_H(\boldsymbol{W}_j-\boldsymbol{w})Y_j(t)\boldsymbol{X}_j}{\sum_{j=1}^{n} K_H(\boldsymbol{W}_j-\boldsymbol{w})Y_j(t)}\right\}^{\otimes 2}\mathrm{d}t.$$

This approach takes the Lin and Ying estimator and attaches a weight $K_H(\boldsymbol{W}_i-\boldsymbol{w})$ to the ith subject ($i=1,\ldots,n$); if $K_H(\boldsymbol{W}_i-\boldsymbol{w})$ is a constant for all $\boldsymbol{w}$, then the estimator reduces to the Lin and Ying estimator. The local kernel estimation method "profiles out" the baseline hazard function separately for each $\boldsymbol{w}$, thus implicitly assuming that the baseline hazard function varies with $\boldsymbol{W}$. The local method suffers loss of efficiency as it does not make use of the fact that the baseline hazard function is indeed shared. In the proposed method, $\boldsymbol{\beta}(\boldsymbol{w}_k)$'s are not estimated separately but are obtained simultaneously from a set of linear equations, because all subjects are jointly used to profile out the common baseline hazard function. Therefore, the global estimator tends to be more efficient than the local estimator.

2.2 Additive hazards model with mixed covariate effects

In practice, certain covariate effects may be known to be constant. We can generally include a covariate vector $\boldsymbol{Z}$ that has a constant effect on the hazard rate and consider the following partially linear varying coefficient additive hazards model:

$$\lambda(t \mid \boldsymbol{W},\boldsymbol{X},\boldsymbol{Z}) = \lambda(t) + \boldsymbol{\beta}(\boldsymbol{W})^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Z},$$

where $\boldsymbol{\alpha}$ is a vector of regression coefficients. We assume that the covariates $\boldsymbol{X}$ and $\boldsymbol{Z}$ do not overlap. In analogy with the kernel-weighted estimator motivated by the block-wise formulation for estimating the varying covariate effect, we propose to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ simultaneously by

$$\begin{pmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{\alpha}}\end{pmatrix} = \begin{pmatrix}\boldsymbol{V} & \boldsymbol{V}_{\beta\alpha}\\ \boldsymbol{V}_{\beta\alpha}^{\mathrm{T}} & \boldsymbol{V}_{\alpha\alpha}\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{b}\\ \boldsymbol{b}_{\alpha}\end{pmatrix}, \tag{3}$$

where $\boldsymbol{V}_{\beta\alpha} = (\boldsymbol{V}_{\beta\alpha}(\boldsymbol{w}_1)^{\mathrm{T}},\ldots,\boldsymbol{V}_{\beta\alpha}(\boldsymbol{w}_m)^{\mathrm{T}})^{\mathrm{T}}$ with

$$\boldsymbol{b}_{\alpha} = \frac{1}{nm}\sum_{i=1}^{n}\int_0^{\tau}\sum_{j=1}^{m} K_H(\boldsymbol{W}_i-\boldsymbol{w}_j)\big\{\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}(t)\big\}\,\mathrm{d}N_i(t),$$

$$\boldsymbol{V}_{\alpha\alpha} = \frac{1}{nm}\sum_{i=1}^{n}\int_0^{\tau}\sum_{j=1}^{m} K_H(\boldsymbol{W}_i-\boldsymbol{w}_j)Y_i(t)\big\{\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}(t)\big\}^{\otimes 2}\,\mathrm{d}t,$$

$$\boldsymbol{V}_{\beta\alpha}(\boldsymbol{w}_k) = \frac{1}{nm}\sum_{i=1}^{n}\int_0^{\tau} K_H(\boldsymbol{W}_i-\boldsymbol{w}_k)Y_i(t)\boldsymbol{X}_i\big\{\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}(t)\big\}^{\mathrm{T}}\,\mathrm{d}t,$$

$$\overline{\boldsymbol{Z}}(t)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}K_H(\boldsymbol{W}_i-\boldsymbol{w}_j)Y_i(t)\boldsymbol{Z}_i}{\sum_{i=1}^{n}\sum_{j=1}^{m}K_H(\boldsymbol{W}_i-\boldsymbol{w}_j)Y_i(t)}.$$

Although $\boldsymbol{\alpha}$ does not vary with $\boldsymbol{W}$, the estimator $\widehat{\boldsymbol{\alpha}}$ depends on the chosen set of grid points for $\boldsymbol{W}$. To obtain a more robust estimator, we could update the estimator of $\boldsymbol{\alpha}$ after obtaining an estimator for $\boldsymbol{\beta}(\boldsymbol{W}_i)$ for each $i=1,\ldots,n$. The updated estimator is essentially the Lin and Ying estimator with $\boldsymbol{\beta}(\boldsymbol{W}_i)$ fixed at $\widehat{\boldsymbol{\beta}}(\boldsymbol{W}_i)$. In particular, we define the estimator as

$$\widetilde{\boldsymbol{\alpha}}\equiv\left[\sum_{i=1}^{n}\int_0^\tau Y_i(t)\big\{\boldsymbol{Z}_i-\widetilde{\boldsymbol{Z}}(t)\big\}^{\otimes 2}\,\mathrm{d}t\right]^{-1}\sum_{i=1}^{n}\int_0^\tau\big\{\boldsymbol{Z}_i-\widetilde{\boldsymbol{Z}}(t)\big\}\big\{\mathrm{d}N_i(t)-Y_i(t)\widehat{\boldsymbol{\beta}}(\boldsymbol{W}_i)^{\mathrm{T}}\boldsymbol{X}_i\,\mathrm{d}t\big\},$$

where $\widetilde{\boldsymbol{Z}}(t)=\sum_{i=1}^{n}Y_i(t)\boldsymbol{Z}_i/\sum_{i=1}^{n}Y_i(t)$. This updated estimator $\widetilde{\boldsymbol{\alpha}}$ is more robust to the choice of grid. The cumulative baseline hazard function $\Lambda(t)\equiv\int_0^t\lambda(s)\,\mathrm{d}s$ can then be estimated by

$$\widehat{\Lambda}(t)=\int_0^t\frac{\sum_{i=1}^{n}\big\{\mathrm{d}N_i(s)-Y_i(s)\widehat{\boldsymbol{\beta}}(\boldsymbol{W}_i)^{\mathrm{T}}\boldsymbol{X}_i\,\mathrm{d}s-Y_i(s)\widetilde{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{Z}_i\,\mathrm{d}s\big\}}{\sum_{i=1}^{n}Y_i(s)}.$$

In the numerical studies, we used linear interpolation to approximate $\widehat{\boldsymbol{\beta}}(\boldsymbol{W}_i)$ for $i=1,\ldots,n$. It is also possible to construct a grid specific to each $\boldsymbol{W}_i$ and estimate $\boldsymbol{\beta}(\boldsymbol{W}_i)$ based on that particular grid. In our experience, the resulting estimates based on a single grid (combined with interpolation) and multiple grids (one for each $\boldsymbol{W}_i$) are very similar, given that the single grid is sufficiently dense. For computational efficiency, we suggest using a single grid to estimate $\boldsymbol{\beta}(\boldsymbol{W}_i)$'s.

2.3 Construction of confidence band for $\beta(\cdot)$

To construct a confidence interval for $\boldsymbol{\beta}$ at a specific $\boldsymbol{w}$, we can use the asymptotic distribution of $\widehat{\boldsymbol{\beta}}(\boldsymbol{w})$ given in Theorem 1 in Section 3. For the entire function $\boldsymbol{\beta}(\cdot)$, we can construct a simultaneous confidence band using the perturbation method of Tian et al. (2005). Suppose that $q=1$, and we are interested in constructing a confidence band for $\boldsymbol{\beta}(w)$ over an interval $[a,b]$. First, we obtain a large-sample approximation to the distribution of

$$\mathcal{S}_k=\sup_{w\in[a,b]}\widehat{v}_k(w)\left|\widehat{\beta}_k(w)-\beta_{0k}(w)\right|\quad\text{for } k=1,\ldots,p,$$

where $\widehat{v}_k(w)$ is a positive weight function that converges uniformly to a deterministic function. In particular, we approximate the distribution of the standardized $\mathcal{S}_k$ by considering the stochastic perturbation of the asymptotic linear form of $\widehat{\boldsymbol{\beta}}(w)$ defined by

$$\widetilde{\boldsymbol{M}}_n(w)=\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\Bigg(K_H(W_i-w)\boldsymbol{X}_i-\sum_{\ell=1}^{m}K_H(W_i-w_\ell)\overline{\boldsymbol{X}}(t,w)$$

$$
\begin{aligned}
&-\left[\frac{1}{n}\sum_{j=1}^{n}\int_0^{\tau} K_H(W_j-w)Y_j(t)\boldsymbol{X}_j\big\{\boldsymbol{Z}_j-\overline{\boldsymbol{Z}}(t)\big\}^{\mathrm{T}}\,\mathrm{d}t\right]\\
&\times\left[\frac{1}{n}\sum_{j=1}^{n}\int_0^{\tau}\sum_{\ell=1}^{m} K_H(W_j-w_\ell)Y_j(t)\big\{\boldsymbol{Z}_j-\overline{\boldsymbol{Z}}(t)\big\}^{\otimes 2}\,\mathrm{d}t\right]^{-1}\\
&\times\sum_{\ell=1}^{m} K_H(W_i-w_\ell)\big\{\boldsymbol{Z}_i-\overline{\boldsymbol{Z}}(t)\big\}\Bigg)\,\mathrm{d}N_i(t)\psi_i,
\end{aligned}
$$

where $\{\psi_i, i=1,\dots,n\}$ is an iid sample from the standard normal distribution. Note that $\widetilde{\boldsymbol{M}}_n(w)$ is obtained by replacing $M_i(t)$ in a version of $\boldsymbol{M}_n(w)$ (defined in the proof of Theorem 1) by $N_i(t)\psi_i$ for $i=1,\dots,n$. We then define

$$
\widetilde{\mathcal{S}}_k = \sup_{w\in[a,b]}\left|\widehat{v}_k(w)\widetilde{M}_{nk}(w)\right|,
$$

where $\widetilde{M}_{nk}(w)$ is the kth component of the vector $\boldsymbol{D}_n(w)^{-1}\widetilde{\boldsymbol{M}}_n(w)$ and $\boldsymbol{D}_n(w) = n^{-1}\sum_{i=1}^{n}\int_0^{\tau} K_H(W_i-w)Y_i(t)\boldsymbol{X}_i^{\otimes 2}\,\mathrm{d}t$. By repeatedly generating samples of $\{\psi_i, i=1,\dots,n\}$, we obtain an empirical distribution of $\widetilde{\mathcal{S}}_k$, which can serve as an approximation of the distribution of $\mathcal{S}_k$.

Let $c_{\alpha k}$ be the $(1-\alpha)$th empirical quantile of $\widetilde{\mathcal{S}}_k$, where $0<\alpha<1$. A $(1-\alpha)$ confidence band for $\{\widehat{\beta}_{0k}(w), w\in[a,b]\}$ can be constructed as

$$
\{\widehat{\beta}_k(w)\pm c_{\alpha k}\widehat{v}_k(w)^{-1}, a\le w\le b\}.
$$

For $\widehat{v}_k(w)$, we set $\{\widehat{v}_k(w)\}^{-1}$ to be the estimated standard error of $\widehat{\beta}_k(w)$, which is the square root of the kth diagonal element of $(nh)^{-1}\widehat{\boldsymbol{\Sigma}}(\boldsymbol{w})$, where $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{w})$ is defined in Appendix A4.

3 Asymptotic theories

We establish asymptotic normality of the estimators of the regression parameters and the cumulative baseline hazard function. In the theoretical development, we assume that $m\to\infty$ as $n\to\infty$, so the grid expands as the sample size increases. The major challenge in the derivations is that $\boldsymbol{b}$ and $\boldsymbol{V}$ increase in dimension as $m\to\infty$, unlike in the local methods where the estimators of $\boldsymbol{\beta}(\boldsymbol{w}_k)$'s can be considered separately. Let $\boldsymbol{\beta}_0\equiv(\boldsymbol{\beta}_0(\boldsymbol{w}_1)^{\mathrm{T}},\dots,\boldsymbol{\beta}_0(\boldsymbol{w}_m)^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\alpha}_0$, and Λ_0 denote the true values of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and Λ, respectively. For simplicity, we assume $\boldsymbol{W}\in[0,1]^q$ with probability one, $\boldsymbol{H}=h\boldsymbol{I}$, and we set $K_h(\boldsymbol{W}-\boldsymbol{w})=\prod_{j=1}^{q}h^{-1}K(h^{-1}(W_j-w_j))$, where h is a positive tuning parameter.

We impose the following conditions:

(C1) The marginal density function of $\boldsymbol{W}$, denoted by f, is twice continuously differentiable and satisfies $\inf_{\boldsymbol{w}\in[0,1]^q} f(\boldsymbol{w})>0$.

(C2) The conditional density of $(\boldsymbol{X}, \boldsymbol{Z})$ given $\boldsymbol{W} = \boldsymbol{w}$ is twice continuously differentiable with respect to $\boldsymbol{w}$. Also, there exists an M such that $P(\|\boldsymbol{X}\| + \|\boldsymbol{Z}\| < M) = 1$.

(C3) The function $\boldsymbol{\beta}_0(\cdot)$ has continuous second derivative on $[0,1]^q$, and $\int_0^\tau \lambda_0(t)\,\mathrm{d}t < \infty$.

(C4) The kernel function K is a density function that is symmetric around 0 with bounded derivative, $\int_{\mathbb{R}} x^2 K(x)\,\mathrm{d}x < \infty$, and $\int_y^\infty K(x)\,\mathrm{d}x \le M \exp(-ay^2)$ for some $M, a > 0$, and large enough y.

(C5) As $n \to \infty$, we have $h \to 0$, $m \to \infty$, $h^q m \to \infty$, $\log h/(\sqrt{n}h^q) \to 0$, and $\sqrt{nh^q}/(h^q m) \to 0$.

(C6) The q-dimension grid $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m)$ is the Cartesian product of q grids over the interval $[0,1]$. Each grid consists of a set of points $(a_1, a_2, \ldots, a_{m_k})$ such that $(\max_i |a_{i+1} - a_i|)/(\min_i |a_{i+1} - a_i|) < M$ for all n and some $M < \infty$.

(C7) If $\boldsymbol{g}(\boldsymbol{W})^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{a}$ is a constant almost surely for some $\boldsymbol{g} \in L^\infty([0,1]^q)$ and real vector $\boldsymbol{a}$, then $\boldsymbol{g}(\cdot) = \boldsymbol{0}$ and $\boldsymbol{a} = \boldsymbol{0}$.

(C8) The term nh^4 is bounded.

Conditions (C1)–(C4) are standard conditions in theoretical developments pertaining to kernel methods. Condition (C1) preserves certain smoothness of f so as to apply second order Taylor's expansion for f. Without loss of generality, we consider $\boldsymbol{w} \in [0,1]^q$ in Condition (C1). The symmetry of K in Condition (C4) implies that $\int_{\mathbb{R}} xK(x)\,\mathrm{d}x = 0$. Condition (C7) is for the purpose of model identifiability.

We first study the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ at a specific $\boldsymbol{w}_0$ that is in the interior of the support of $\boldsymbol{W}$. Assume that the sets of grid points are nested as $m \to \infty$ and that each set of grid points includes $\boldsymbol{w}_0$. For $\boldsymbol{w} \notin \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m\}$, we define $\widehat{\boldsymbol{\beta}}(\boldsymbol{w}) = \widehat{\boldsymbol{\beta}}(\boldsymbol{w}_k)$ for $w_{k,j} \le w_j < w_{k+1,j}$, where w_j and $w_{k,j}$ represent the jth component of $\boldsymbol{w}$ and $\boldsymbol{w}_k$, respectively, for $j = 1, \ldots, q$.

Theorem 1 *Under Conditions (C1)–(C7),*

$$\sqrt{nh^q}\big\{\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_0) - \boldsymbol{\beta}_0(\boldsymbol{w}_0)\big\} \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}(\boldsymbol{w}_0)),$$

where $\xrightarrow{d}$ *denotes convergence in distribution and the definition of* $\boldsymbol{\Sigma}(\cdot)$ *is provided in Appendix A3.*

Theorem 2 *Under Conditions (C1)–(C8), for any bounded function* $\boldsymbol{\phi}$,

$$\sqrt{n}\int_{\mathcal{W}} \boldsymbol{\phi}(\boldsymbol{w})^{\mathrm{T}}\big\{\widehat{\boldsymbol{\beta}}(\boldsymbol{w}) - \boldsymbol{\beta}_0(\boldsymbol{w})\big\}\,\mathrm{d}\boldsymbol{w} \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\phi),$$

where the definition of $\boldsymbol{\Sigma}_\phi$ *is provided in the Online Resource.*

Theorem 3 *Under Conditions (C1)–(C8),*

$$\sqrt{n}(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\widetilde{\alpha}}),$$

where the definition of $\boldsymbol{\Sigma}_{\widetilde{\alpha}}$ *is provided in the Online Resource.*

Estimators of $\boldsymbol{\Sigma}(\boldsymbol{w}_0)$, $\boldsymbol{\Sigma}_\phi$, and $\boldsymbol{\Sigma}_{\widetilde{\alpha}}$ are given in Appendix A4.

Theorem 4 *Under Conditions (C1)–(C8),* $\sqrt{n}\{\widehat{\Lambda}(t) - \Lambda_0(t)\}$ *converges in distribution to a Gaussian process in* $L_\infty([0,\tau])$.

The proof of Theorem 1 is provided in Appendix A3, and the proofs of Theorems 2–4 are provided in the Online Resource. Note that $\int_{\mathcal{W}} \boldsymbol{\phi}(\boldsymbol{w})^{\mathrm{T}}\widehat{\boldsymbol{\beta}}(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w}$ converges at the rate of $\sqrt{n}$, whereas $\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_0)$ only converges at the rate of $\sqrt{nh^q}$; the latter rate is the square root of the size of the local neighborhood at any $\boldsymbol{w}_0$. As discussed in Section 2.1, a larger grid size m generally improves the approximation of integrals in the estimating equation $\mathcal{V}\boldsymbol{\beta} = \mathcal{C}$. In the proof of Theorem 1, we can see that this approximation results in an error term of rate $\max(\sqrt{nh^{-q}m^{-2}}, \sqrt{nh^{q+4}})$ in the expansion of $\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_0)$, so we require $m \gg \sqrt{nh^{-q}}$, as specified in Condition (C5). If $m \gg \sqrt{h^{-3q-4}}$, then the error term in the expansion of $\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_0)$ is no longer dominated by this approximation error, so further enlarging m does not yield any asymptotic improvement. For the asymptotic normality of $\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_0)$, any positive integer value of q is valid as long as Condition (C5) is satisfied. By contrast, if the asymptotic normality of $\widetilde{\boldsymbol{\alpha}}$ is required, then Condition (C8) is needed, which requires $q = 1$.

4 Simulation studies

We assess the performance of the proposed methods through simulation studies. We considered a three-dimensional $\boldsymbol{X}$, whose effect on the outcome is modified by a univariate W ($q = 1$) or a bivariate $\boldsymbol{W}$ ($q = 2$). We considered a bivariate $\boldsymbol{Z}$ that has a constant effect on the hazard rate. The components of $\boldsymbol{W}$, $\boldsymbol{X}$, and $\boldsymbol{Z}$ were generated independently from the Uniform$(0,1)$ distribution. We generated the outcome variable from the following hazard function:

$$
\begin{aligned}
&\lambda(t \mid \boldsymbol{W} = \boldsymbol{w}, \boldsymbol{X}, \boldsymbol{Z}) \\
&= \lambda(t) + \beta_1(\boldsymbol{w})X_1 + \beta_2(\boldsymbol{w})X_2 + \beta_3(\boldsymbol{w})X_3 + \alpha_1 Z_1 + \alpha_2 Z_2 \\
&= t + \frac{1}{1 + \exp\{-20(\overline{w} - 0.5)\}}X_1 + \{1 - \sin(\pi\overline{w})\}X_2 + 0.2X_3 + 0.2Z_1 + 0.2Z_2,
\end{aligned}
$$

where $\overline{w} = q^{-1}\sum_{j=1}^{q} w_j$. The functions $\beta_1(\boldsymbol{w})$, $\beta_2(\boldsymbol{w})$, and $\beta_3(\boldsymbol{w})$ are plotted in Figure 1. We set the censoring time to follow an exponential distribution with the mean chosen to yield a censoring rate of about 30%. We considered sample sizes of $n = 200$, 500, and 1000.

In addition to the proposed method, we considered three alternative methods: the additive hazards model on linear predictors $\boldsymbol{X}$ and $\boldsymbol{Z}$, the varying coefficient additive hazards model estimated using the local kernel method, and the varying coefficient additive hazards model, with the varying coefficients approximated using B-spline functions. For the local method, we first consider $\boldsymbol{\alpha}$ as varying and estimate $(\boldsymbol{\alpha}, \boldsymbol{\beta}(\boldsymbol{W}_i))$ for each $\boldsymbol{W}_i$ ($i = 1, \ldots, n$). Following Yin et al. (2008), we set the final estimator of $\boldsymbol{\alpha}$ as a weighted average across the estimators $\widehat{\boldsymbol{\alpha}}(\boldsymbol{W}_i)$ at each $\boldsymbol{W}_i$. For the proposed method, under

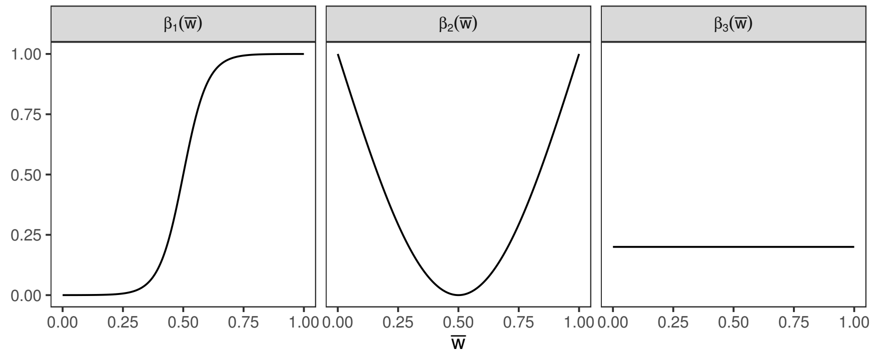

Fig. 1 True varying coefficient functions in the simulation studies

a univariate W, we obtained $\widehat{\boldsymbol{\beta}}$ at an evenly spaced grid over $[0, 1]$ with size 5, 9, or 13. Under a bivariate $\boldsymbol{W}$, we considered an evenly spaced grid over $[0, 1]$ with size 5 in each dimension. We used linear interpolation to estimate $\boldsymbol{\beta}$ at points off the grid. For the spline method, when $q = 1$, we approximate $\beta_j(w)$ by $\sum_{k=1}^{4} \gamma_{jk} B_k(w)$, where $(B_1, \ldots, B_4)$ are cubic B-spline basis functions with (boundary) knots at 0 and 1, and $(\gamma_{j1}, \ldots, \gamma_{j4})$ are the corresponding regression parameters. When $q = 2$, we approximate $\beta_j(w_1, w_2)$ by $\sum_{k=1}^{4} \sum_{\ell=1}^{4} \gamma_{jk\ell} B_k(w_1) B_\ell(w_2)$, where $\gamma_{jk\ell}$'s are regression parameters.

We consider a Gaussian kernel of $K(\boldsymbol{x}) \propto \exp(-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}/2)$. A full bandwidth matrix $\boldsymbol{H}$ gives more flexibility but also quadratically increases the number of bandwidth parameters to be chosen, precisely $q(q + 1)/2$. This complicates the bandwidth selection as q grows. A common simplification is to consider a diagonal bandwidth matrix. Here, we adopt a diagonal bandwidth matrix of $\boldsymbol{H} = \mathrm{diag}(h_1^2, \ldots, h_q^2)$ and follow Silverman (1986, p.45) to set $h_j = \widehat{\sigma}_j [4/\{n(q+2)\}]^{1/(q+4)}$, where $\widehat{\sigma}_j$ is the empirical standard deviation of $\boldsymbol{W}_{\cdot j}$ for $j = 1, \ldots, q$. If computationally feasible, cross-validation can be applied to choose the bandwidth.

The performance of each method is assessed using the following measures. We report the mean-squared error (MSE), defined as $\mathrm{E}[\{\widehat{\boldsymbol{\beta}}(\boldsymbol{W}) - \boldsymbol{\beta}_0(\boldsymbol{W})\}^{\mathrm{T}}\boldsymbol{X} + (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)^{\mathrm{T}}\boldsymbol{Z}]^2$, where $\widehat{\boldsymbol{\beta}}(\boldsymbol{W})$ and $\boldsymbol{\beta}_0(\boldsymbol{W})$ denote the estimated and true values of $\boldsymbol{\beta}$ evaluated at $\boldsymbol{W}$, respectively, and $\widehat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}_0$ denote the estimated and true values of $\boldsymbol{\alpha}$, respectively. Also, we compute the concordance index (C-index) (Uno et al., 2011) between $\widehat{\boldsymbol{\beta}}(\boldsymbol{W})^{\mathrm{T}}\boldsymbol{X} + \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{Z}$ and the event time. For both the MSE and C-index, we generated a large independent sample of $(T, \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Z})$ and approximated these two measures empirically on this sample. We present the sample standard deviations of the estimates and the average estimated standard errors, computed based on asymptotic approximation. We provide the coverage rates of the 95% confidence intervals. We considered 500 simulation replicates. The results for $q = 1$ are summarized in Tables 1–2 and Tables S1–S2 in the Online Resource. Figure 2 shows the average estimated

Table 1 MSE and C-index for $q = 1$

	Method	$n = 200$	$n = 500$	$n = 1000$
MSE	Constant	0.317	0.192	0.157
	Local	0.579	0.228	0.133
	Proposed ($m = 5$)	0.303	0.121	0.066
	Proposed ($m = 9$)	0.309	0.123	0.064
	Proposed ($m = 13$)	0.312	0.125	0.066
	Spline	0.440	0.148	0.072
C-index	Constant	0.532	0.541	0.545
	Local	0.548	0.565	0.575
	Proposed ($m = 5$)	0.568	0.582	0.590
	Proposed ($m = 9$)	0.568	0.582	0.590
	Proposed ($m = 13$)	0.568	0.582	0.591
	Spline	0.566	0.581	0.590

Note: Constant, regression model on linear predictors $\boldsymbol{X}$ and $\boldsymbol{Z}$; Local, the varying coefficient model where the local kernel estimation method is adopted; Proposed, the varying coefficient model estimated using the proposed method; Spline, the varying coefficient model estimated using the spline approximation.

Table 2 Estimation of regression coefficients for $q = 1$

		α_1				α_2			
n	m	Bias	SD	SE	CR	Bias	SD	SE	CR
200	5	0.016	0.380	0.350	0.962	0.028	0.361	0.349	0.970
	9	0.017	0.381	0.350	0.962	0.030	0.363	0.349	0.972
	13	0.017	0.382	0.350	0.962	0.030	0.363	0.349	0.970
500	5	0.015	0.216	0.219	0.978	−0.012	0.216	0.218	0.976
	9	0.016	0.216	0.219	0.976	−0.012	0.216	0.218	0.976
	13	0.016	0.216	0.219	0.976	−0.011	0.217	0.218	0.976
1000	5	−0.011	0.154	0.154	0.976	0.000	0.158	0.154	0.970
	9	−0.011	0.155	0.154	0.976	0.001	0.159	0.154	0.974
	13	−0.011	0.155	0.154	0.978	0.001	0.159	0.154	0.974

Note: SD, sample standard deviations of the estimates; SE, estimated standard errors based on asymptotic approximation; CR, coverage rates of the 95% confidence intervals.

values of β_1, β_2, and β_3 under $n = 1000$. The simulation results under other settings for $q = 1$ are plotted in Figures S1–S3 in the Online Resource. The simulation results for $q = 2$ are presented in Tables S3–S6 and Figure S4 in the Online Resource.

Overall, the proposed method yields higher prediction accuracy than the alternatives. For $q = 1$, the proposed method yields superior prediction performance in all simulation settings, with the lowest MSE and highest C-index values, as shown in Table 1. The proposed method outperforms the local method by effectively incorporating the shared baseline hazard function for all subjects. Also, the proposed method yields much higher predictive power than the local method in small sample scenarios, highlighting the importance of utilizing the information of the baseline hazard function from all subjects,

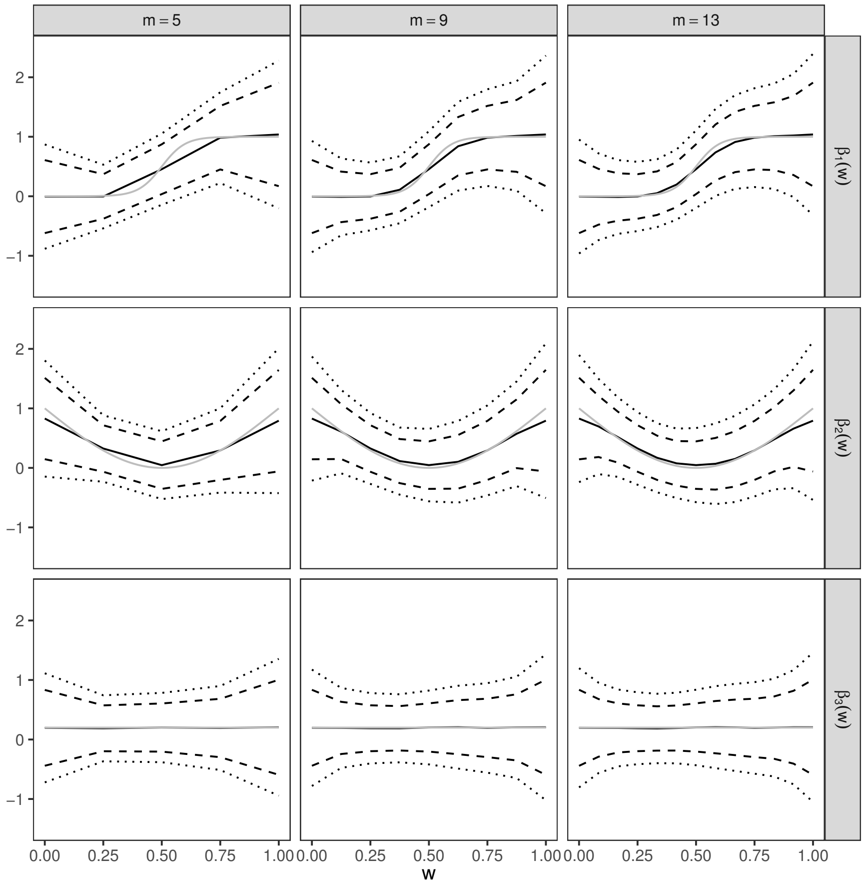

Fig. 2 The estimated coefficient functions (solid in black), true coefficient functions (solid in gray), 95% simultaneous confidence band (dotted), and 95% pointwise confidence intervals (dashed) for $q = 1$ under $n = 1000$

especially when the sample size is limited. As shown in Tables S1 and S5 of the Online Resource, the standard deviations of the estimates under the proposed method are smaller than those under the local method. The average estimated coefficient functions from the local method are virtually indistinguishable from those of the global method and thus are not presented in Figures 2, S1, and S2.

For $q = 2$, the kernel methods yield larger MSE values than the constant method when $n = 200$ and 500, as shown in Table S3. The biases for the grid points near the boundaries are relatively large due to limited samples. In the cases with small samples, this may lead to larger MSE values for the kernel methods. The biases reduce when the sample size increases. It is important to note that a sufficiently large sample is required when a multivariate index variable is involved. Nevertheless, the proposed method outperforms the local method in all simulation settings by demonstrating smaller MSE values and reduced standard deviations of the estimated varying coefficients.

For $m = 5$, 9, and 13, the mean-squared error and concordance index values are almost the same. As shown in Figure 2, when the number of grid points increases, the biases reduce as more interpolation grids can generally capture the curvilinear structure of the true function. Nevertheless, the improvement in the estimation accuracy is not substantial for β_1, which is a monotone function. If the true function is steadily increasing or decreasing, then a small number of grid points may be sufficient to capture the variation in the varying coefficient, and the prediction performance is not sensitive to the number of grids. For β_2, there is a considerable improvement in bias when we increase the number of grids from 5 to 13. If the true function has (multiple) turns, then more grid points are required to capture the variation of the varying coefficient over different values of w.

As shown in Tables S2 and S6 of the Online Resource, the confidence bands yield coverage rates over 95%, suggesting that the proposed method is effective in capturing the true underlying coefficient functions. Similar to the confidence intervals for regression parameters, the confidence bands tend to be conservative, due to some overestimation of the standard errors of the estimators. As shown in Table 2 and Table S4 of the Online Resource, the proximity of SD and SE suggests the validity of the inferential procedures on the constant covariate effects. Also, the coverage rates of confidence intervals exceed 95%, demonstrating the effectiveness of the proposed method in estimating these effects.

5 Real data analysis

We investigated the role of the mechanistic target of rapamycin (mTOR) signaling pathway in cancer progression. The mTOR signaling pathway is often activated in tumors to maintain the growth, survival, and proliferation of tumor cells (Zou et al., 2020). The activation subsequently leads to tumor cell metastasis, drug resistance, and ultimately affects the prognosis of cancer patients. Several studies have demonstrated that the overexpression or activation of two key downstream effectors of mTOR, namely 4E binding protein 1 (4EBP1) and eukaryotic initiation factor 4E (eIF4E) proteins, is associated with poor prognosis in many major types of cancers (Coleman et al., 2009; No et al., 2011; Inamdar et al., 2009). Understanding protein expressions in addition to gene expressions is of interest since the mTOR signaling pathway is tightly regulated by proteins. In a recent study, Nelson et al. (2024) suggested that high expression level of the Eukaryotic Translation Initiation Factor 4E Binding Protein 1 (EIF4EBP1), a protein-coding gene, could serve as a poor prognostic marker by activating mTOR signaling pathways and promoting increased tumorigenesis in bulk breast tumors through upregulation of cell cycling. While extensive literature has demonstrated the prognostic value of the mTOR signaling pathway in cancer progression, the specific impact of

Table 3 Preliminary analysis results

Variable(s)	C-index
$\boldsymbol{X}$	0.607
W	0.636
$\boldsymbol{Z}$	0.629
$\boldsymbol{X} + \boldsymbol{Z}$	0.651
$\boldsymbol{X} + W + \boldsymbol{Z}$	0.669
$\boldsymbol{X}I(W > \text{median}(W)) + \boldsymbol{X}I(W < \text{median}(W)) + \boldsymbol{Z}$	0.658
$\boldsymbol{X}I(W > \text{median}(W)) + \boldsymbol{X}I(W < \text{median}(W)) + W + \boldsymbol{Z}$	0.674

EIF4EBP1 gene expression in tumors through mTOR signaling pathways remains unclear. Therefore, in this study, we investigate the interplay between gene expression and the mTOR signaling pathway on cancer progression.

We examined the effects of important downstream effects of the mTOR signaling pathway and EIF4EBP1 gene expression on the time to death since initial diagnosis. We used a dataset of kidney renal clear cell carcinoma (KIRC) patients from The Cancer Genome Atlas (TCGA) (The Cancer Genome Atlas Network, 2013), available at `https://gdac.broadinstitute.org`. We set $\boldsymbol{X}$ to be the expressions of the proteins 4E-BP1_pS65, 4E-BP1_pT37_T46, and 4E-BP1_pT70, set W to be the expression of the gene EIF4EBP1, and set $\boldsymbol{Z}$ to be the clinical variables consisting of age and gender (0 for female and 1 male). After removing subjects with missing data, the sample size is 475. The median time to censoring or death is 3.22 years, and the censoring rate is 65.26%.

We first conducted a preliminary analysis to examine the prognostic value of different combinations of data types on time to death since initial diagnosis. We repeatedly split the data set into pairs of training and testing sets 500 times, with a ratio of sample sizes of 7:3. In each split, we fit an additive hazards model (with constant coefficients) on the training set and computed the C-index of the estimated hazard on the testing set. The average C-index values under different combination of data types across splits are presented in Table 3.

The results indicate that integrating proteomic, genomic, and clinical features tends to improve prognosis compared to using a single type of features. Also, models allowing the effects of protein expressions to vary with gene expression yield higher C-index values than models assuming constant effects for protein expressions, suggesting the presence of varying covariate effects for protein expressions.

Next, we applied the proposed method to investigate the effects of protein expressions, gene expressions, and clinical variables on time to death since initial diagnosis, considering the modification of protein expression effects by gene expression. We used an evenly-spaced grid of points over the lower and upper quantile of EIF4EBP1 expression and employed a kernel function with a chosen bandwidth, similar to the simulations. For comparison, we considered two alternative methods: the additive hazards model on linear predictors

Table 4 Estimated constant covariate effect and standard error (in parentheses) for the KIRC analysis

Variable	Constant	Local	Proposed
EIF4EBP1	0.033 (0.008)	-	-
4E-BP1_pS65	−0.070 (0.022)	-	-
4E-BP1_pT37_T46	0.026 (0.015)	-	-
4E-BP1_pT70	−0.018 (0.020)	-	-
Age	0.003 (0.001)	0.000 (0.000)	0.003 (0.001)
Gender	−0.004 (0.016)	−0.015 (0.004)	0.006 (0.016)

Note: See Note to Table 1.

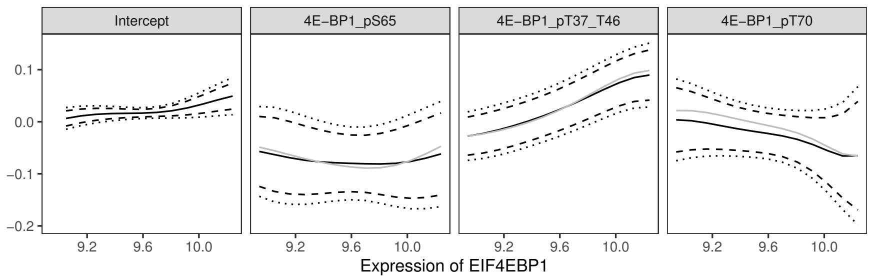

Fig. 3 The estimated coefficient functions using the proposed method (solid in black), 95% simultaneous confidence band (dotted), 95% pointwise confidence intervals (dashed) for the KIRC analysis, and The estimated coefficient functions (solid in black), the estimated coefficient functions using the local method (solid in gray)

W, $\boldsymbol{X}$, and $\boldsymbol{Z}$ estimated using the Lin and Ying estimator and the varying coefficient additive hazards model estimated using the local kernel estimation method. The estimated constant coefficients under different methods are shown in Table 4 and the estimated varying coefficients are shown in Figure 3. The effects of 4E-BG1_pS65 and 4E-BP1_pT70 tend to remain constant. The effect of 4E-BP1_pT37_T46 appears to increase as the expression of EIF4EBP1 increases, which suggests a potential varying covariate effect.

Overall, our real data analysis confirms the findings from the preliminary analysis, indicating that protein expressions, particularly 4E-BP1_pT37_T46, have distinct effects on survival depending on the level of gene expression. Unlike the preliminary analysis, which dichotomized the gene expression, the proposed method incorporates a continuous effect modifier and provides a more comprehensive understanding of the interplay between gene expression and the mTOR signaling pathway on survival.

6 Discussion

In this paper, we study kernel-based estimation methods for the varying coefficient additive hazards models. Many existing kernel-based methods for varying coefficient survival models are direct extensions of methods for corresponding models without varying coefficients that incorporate a kernel weight to each subject in existing estimators. Such local methods suffer efficiency loss, because they do not account for the fact that there are some nuisance parameters (in the cases of the proportional hazards and additive hazards models, the baseline hazard function) shared among all subjects. The proposed global kernel estimation method, by contrast, uses all subjects to profile out the nuisance parameters and thus has higher efficiency.

In the literature, estimation methods for varying coefficient models are mainly based on the kernel smoothing or spline approximation. Although the two approaches may yield similar empirical results for univariate varying coefficient functions, the kernel approach is more flexible under a multivariate $\boldsymbol{W}$. In particular, for kernel methods, we simply need to define the distance between two values of $\boldsymbol{W}$, which is not necessarily the Euclidean distance. This distance can then be used to determine the subject weights in estimation, and $\boldsymbol{\beta}(\boldsymbol{W})$ at two $\boldsymbol{W}$ values close to each other (according to the distance) would be encouraged to be similar. By contrast, it would be more challenging to define (multivariate) spline functions that encourage $\boldsymbol{\beta}(\boldsymbol{w}_1)$ and $\boldsymbol{\beta}(\boldsymbol{w}_2)$ to be similar for $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ close to each other under some non-Euclidean distance.

The proposed estimator is motivated by first considering a standard estimator of the proposed model under a discrete $\boldsymbol{W}$ and then "kernelizing" it, that is, suitably incorporating kernel weights to the estimator. This general approach can be applied to other varying coefficient models that contain parameters not varying with $\boldsymbol{W}$, such as the (varying coefficient) Cox proportional hazards model or transformation model (Cheng et al., 1995), to yield global estimators. In fact, such an approach for the varying coefficient Cox model would yield an estimator that is asymptotically equivalent to that of Chen et al. (2012), although they proposed an alternative computation method that iteratively updates (the counterparts of) $\boldsymbol{\beta}(\boldsymbol{W}_i)$ and $\boldsymbol{\alpha}$ for $i = 1, \ldots, n$.

The current work can be extended in several directions. First, the proposed method can be easily extended to estimate model (1), which contains time-varying covariates. In fact, we only need to replace the covariates in formulation of the estimator by their time-varying counterparts. The theoretical development is essentially the same. Second, we can allow the varying covariate effect for each component of $\boldsymbol{X}$ to depend on different components of $\boldsymbol{W}$; in practical applications, we may have prior knowledge that the effect of a covariate depends on specific variables. This can be achieved by considering different kernel matrices for each component of $\boldsymbol{X}$. Third, we can consider high-dimensional $\boldsymbol{X}$ and $\boldsymbol{Z}$, which is relevant to omics studies where omics features are often high-dimensional. In practice, it is often the case that only a small subset of features are relevant to a particular outcome. The proposed approaches can be extended to include penalization on the regression parameters/functions, so variable selection can be performed.

Appendix A1 Varying coefficient estimator for discrete $\boldsymbol{W}$

Let $N_i^{(k)}(t)$ and $Y_i^{(k)}(t)$ denote the event process and at-risk process for the ith subject in the kth group at time t, respectively, where the kth group refers to the set of subjects with $\boldsymbol{W} = \boldsymbol{w}_k$. Let $\boldsymbol{X}_i^{(k)}$ be the observed covariate vector for the ith subject in the kth group ($i = 1, \ldots, n_k$) and $\boldsymbol{B}_i = (I(\boldsymbol{W}_i = \boldsymbol{w}_1)\boldsymbol{X}_i^{\mathrm{T}}, \ldots, I(\boldsymbol{W}_i = \boldsymbol{w}_m)\boldsymbol{X}_i^{\mathrm{T}})^{\mathrm{T}}$. The average covariate vector at time t is

$$\overline{\boldsymbol{B}}(t) = \frac{\sum_{i=1}^n Y_i(t)\boldsymbol{B}_i}{\sum_{i=1}^n Y_i(t)} = \left(\frac{\sum_{i=1}^{n_1} Y_i^{(1)}(t)\boldsymbol{X}_i^{(1)\mathrm{T}}}{\sum_{i=1}^n Y_i(t)}, \ldots, \frac{\sum_{i=1}^{n_m} Y_i^{(m)}(t)\boldsymbol{X}_i^{(m)\mathrm{T}}}{\sum_{i=1}^n Y_i(t)} \right)^{\mathrm{T}}$$
$$\equiv \left(\overline{\boldsymbol{X}}^{(1)}(t)^{\mathrm{T}}, \ldots, \overline{\boldsymbol{X}}^{(m)}(t)^{\mathrm{T}} \right)^{\mathrm{T}}.$$

The Lin and Ying estimator for the regression parameters of the covariate $\boldsymbol{B}$ is $\boldsymbol{V}_{\mathrm{B}}^{-1}\boldsymbol{b}_{\mathrm{B}}$, where $\boldsymbol{b}_{\mathrm{B}} \equiv n^{-1}\sum_{i=1}^n \int_0^\tau \{\boldsymbol{B}_i - \overline{\boldsymbol{B}}(t)\}\,\mathrm{d}N_i(t)$ is a vector with the kth subvector given by

$$\frac{1}{n}\sum_{i=1}^{n_k}\int_0^\tau \boldsymbol{X}_i^{(k)}\mathrm{d}N_i^{(k)}(t) - \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \overline{\boldsymbol{X}}^{(k)}(t)\,\mathrm{d}N_i(t),$$

and $\boldsymbol{V}_{\mathrm{B}} \equiv n^{-1}\sum_{i=1}^n \int_0^\tau Y_i(t)\{\boldsymbol{B}_i - \overline{\boldsymbol{B}}(t)\}^{\otimes 2}\,\mathrm{d}t$ is a block matrix with the (k,ℓ)th block given by

$$\frac{I(k=\ell)}{n}\int_0^\tau \sum_{i=1}^{n_k} Y_i^{(k)}(t)\boldsymbol{X}_i^{(k)}\boldsymbol{X}_i^{(\ell)\mathrm{T}}\,\mathrm{d}t - \frac{1}{n}\int_0^\tau \sum_{i=1}^{n} Y_i(t)\overline{\boldsymbol{X}}^{(k)}(t)\overline{\boldsymbol{X}}^{(\ell)}(t)^{\mathrm{T}}\,\mathrm{d}t.$$

Appendix A2 Technical results

We introduce some notations to be used in the following theoretical development. Let $\boldsymbol{V} = (1, \boldsymbol{X}^{\mathrm{T}}, \boldsymbol{Z}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{V}_i$ be $\boldsymbol{V}$ for the ith subject. For $t \in [0,\tau]$ and $\boldsymbol{w} \in [0,1]^q$, we define

$$\boldsymbol{S}_n(t,\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^n K_h(\boldsymbol{W}_i - \boldsymbol{w})Y_i(t)\boldsymbol{V}_i^{\otimes 2} \equiv \begin{pmatrix} S_{n0}(t,\boldsymbol{w}) & \boldsymbol{S}_{nX}(t,\boldsymbol{w})^{\mathrm{T}} & \boldsymbol{S}_{nZ}(t,\boldsymbol{w})^{\mathrm{T}} \\ \boldsymbol{S}_{nX}(t,\boldsymbol{w}) & \boldsymbol{S}_{nX^2}(t,\boldsymbol{w}) & \boldsymbol{S}_{nZX}(t,\boldsymbol{w})^{\mathrm{T}} \\ \boldsymbol{S}_{nZ}(t,\boldsymbol{w}) & \boldsymbol{S}_{nZX}(t,\boldsymbol{w}) & \boldsymbol{S}_{nZ^2}(t,\boldsymbol{w}) \end{pmatrix},$$
$$\boldsymbol{S}_{nX^2}^*(t,\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^n K_h(\boldsymbol{W}_i - \boldsymbol{w})Y_i(t)\lambda(t \mid \boldsymbol{W}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)\boldsymbol{X}_i^{\otimes 2},$$

and their population counterparts

$$\boldsymbol{s}(t,\boldsymbol{w}) = f(\boldsymbol{w})\mathrm{E}\big(P(T \ge t \mid \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Z})\boldsymbol{V}^{\otimes 2} \mid \boldsymbol{W} = \boldsymbol{w}\big)$$
$$\equiv \begin{pmatrix} s_0(t,\boldsymbol{w}) & \boldsymbol{s}_X(t,\boldsymbol{w})^{\mathrm{T}} & \boldsymbol{s}_Z(t,\boldsymbol{w})^{\mathrm{T}} \\ \boldsymbol{s}_X(t,\boldsymbol{w}) & \boldsymbol{s}_{X^2}(t,\boldsymbol{w}) & \boldsymbol{s}_{ZX}(t,\boldsymbol{w})^{\mathrm{T}} \\ \boldsymbol{s}_Z(t,\boldsymbol{w}) & \boldsymbol{s}_{ZX}(t,\boldsymbol{w}) & \boldsymbol{s}_{Z^2}(t,\boldsymbol{w}) \end{pmatrix},$$
$$\boldsymbol{s}_{X^2}^*(t,\boldsymbol{w}) = f(\boldsymbol{w})\mathrm{E}\big(P(T \ge t \mid \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Z})\lambda(t \mid \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Z})\boldsymbol{X}^{\otimes 2} \mid \boldsymbol{W} = \boldsymbol{w}\big).$$

Similarly, for $t \in [0, \tau]$, we define

$$S_{n0}(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t), \quad s_0(t) = \mathrm{E}\big(P(T \geq t \mid \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Z})\big),$$

$$\boldsymbol{S}_{nZ}(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\boldsymbol{Z}_i, \quad \boldsymbol{s}_Z(t) = \mathrm{E}\big(P(T \geq t \mid \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Z})\boldsymbol{Z}\big),$$

$$\boldsymbol{S}_{nZ^2}(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\boldsymbol{Z}_i^{\otimes 2}, \ \boldsymbol{s}_{Z^2}(t) = \mathrm{E}\big(P(T \geq t \mid \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Z})\boldsymbol{Z}^{\otimes 2}\big).$$

Denote

$$\nu_0 = \int_{\mathbb{R}^q} K(\boldsymbol{z})^2 \,\mathrm{d}\boldsymbol{z}, \quad \boldsymbol{D}_n(\boldsymbol{w}) = \int_0^\tau \boldsymbol{S}_{nX^2}(t, \boldsymbol{w})\,\mathrm{d}t, \quad \boldsymbol{D}(\boldsymbol{w}) = \int_0^\tau \boldsymbol{s}_{X^2}(t, \boldsymbol{w})\,\mathrm{d}t,$$

$$\begin{aligned}\boldsymbol{\Psi}_n(\boldsymbol{w}; \boldsymbol{u}) = &\int_0^\tau \frac{\boldsymbol{S}_{nX}(t, \boldsymbol{w})\boldsymbol{S}_{nX}(t, \boldsymbol{u})^{\mathrm{T}}}{S_{n0}(t)}\,\mathrm{d}t + \int_0^\tau \left(\boldsymbol{S}_{nZX}(t, \boldsymbol{w})^{\mathrm{T}} - \frac{\boldsymbol{S}_{nX}(t, \boldsymbol{w})\boldsymbol{S}_{nZ}(t)^{\mathrm{T}}}{S_{n0}(t)}\right)\mathrm{d}t \\ &\times\left\{\int_0^\tau \left(\boldsymbol{S}_{nZ^2}(t) - \frac{\boldsymbol{S}_{nZ}(t)^{\otimes 2}}{S_{n0}(t)}\right)\mathrm{d}t\right\}^{-1}\left\{\int_0^\tau \left(\boldsymbol{S}_{nZX}(t, \boldsymbol{u})^{\mathrm{T}} - \frac{\boldsymbol{S}_{nX}(t, \boldsymbol{u})\boldsymbol{S}_{nZ}(t)^{\mathrm{T}}}{S_{n0}(t)}\right)\mathrm{d}t\right\}^{\mathrm{T}}\end{aligned}$$

$$\begin{aligned}\boldsymbol{\Psi}(\boldsymbol{w}; \boldsymbol{u}) = &\int_0^\tau \frac{\boldsymbol{s}_X(t, \boldsymbol{w})\boldsymbol{s}_X(t, \boldsymbol{u})^{\mathrm{T}}}{s_0(t)}\,\mathrm{d}t + \int_0^\tau \left(\boldsymbol{s}_{ZX}(t, \boldsymbol{w})^{\mathrm{T}} - \frac{\boldsymbol{s}_X(t, \boldsymbol{w})\boldsymbol{s}_Z(t)^{\mathrm{T}}}{s_0(t)}\right)\mathrm{d}t \\ &\times\left\{\int_0^\tau \left(\boldsymbol{s}_{Z^2}(t) - \frac{\boldsymbol{s}_Z(t)^{\otimes 2}}{s_0(t)}\right)\mathrm{d}t\right\}^{-1}\left\{\int_0^\tau \left(\boldsymbol{s}_{ZX}(t, \boldsymbol{u})^{\mathrm{T}} - \frac{\boldsymbol{s}_X(t, \boldsymbol{u})\boldsymbol{s}_Z(t)^{\mathrm{T}}}{s_0(t)}\right)\mathrm{d}t\right\}^{\mathrm{T}}.\end{aligned}$$

For any $(a \times b)$ matrices $\boldsymbol{A}_n$ and $\boldsymbol{A}$, we use $\boldsymbol{A}_n = \boldsymbol{A} + O(r_n)$ to denote that $A_{n,ij} = A_{ij} + O(r_n)$ for $i = 1, \ldots, a$ and $j = 1, \ldots, b$. We present three lemmas that will be used in the proofs of the theorems. The proofs of the lemmas are provided in the Online Resource.

Lemma 1 *Under Conditions (C1)–(C5),*

$$\sup_{t\in[0,\tau], \boldsymbol{w}\in[0,1]^q} |\boldsymbol{S}_n(t, \boldsymbol{w}) - \boldsymbol{s}(t, \boldsymbol{w})| = O_p\left(\frac{\log h}{\sqrt{n}h^q}\right) + O(h^2),$$

$$\sup_{t\in[0,\tau], \boldsymbol{w}\in[0,1]^q} |\boldsymbol{S}^*_{nX^2}(t, \boldsymbol{w}) - \boldsymbol{s}^*_{X^2}(t, \boldsymbol{w})| = O_p\left(\frac{\log h}{\sqrt{n}h^q}\right) + O(h^2).$$

Lemma 2 *The following results hold:*

$$\sup_{t\in[0,\tau]} |\boldsymbol{S}_{nZ^2}(t) - \boldsymbol{s}_{Z^2}(t)| = O_p(n^{-1/2}),$$

$$\sup_{t\in[0,\tau]} |\boldsymbol{S}_{nZ}(t) - \boldsymbol{s}_Z(t)| = O_p(n^{-1/2}),$$

$$\sup_{t\in[0,\tau]} |S_{n0}(t) - s_0(t)| = O_p(n^{-1/2}).$$

Lemma 3 *Let $\mathcal{I}$ be the identity operator. For $\boldsymbol{\phi} \in L^2([0,1]^q)$, define the operator $\mathcal{A}$ by*

$$\mathcal{A}(\boldsymbol{\phi})(\boldsymbol{w}) = \boldsymbol{D}(\boldsymbol{w})^{-1} \int_{\mathcal{W}} \boldsymbol{\Psi}(\boldsymbol{w};\boldsymbol{u})\boldsymbol{\phi}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}.$$

Under Condition (C7), the inverse operator $(\mathcal{I}-\mathcal{A})^{-1}$ exists.

The operator $\mathcal{A}(\boldsymbol{\phi})(\boldsymbol{w})$ can be estimated by $\widehat{\mathcal{A}}(\boldsymbol{\phi})(\boldsymbol{w}) = \boldsymbol{D}_n(\boldsymbol{w})^{-1}m^{-1}\sum_{k=1}^{m}\boldsymbol{\Psi}_n(\boldsymbol{w},\boldsymbol{w}_k)\boldsymbol{\phi}(\boldsymbol{w}_k)$.

Appendix A3 Proof of Theorem 1

A key step of the proof is to show that for $k = 1,\ldots,m$, $\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_k) - \boldsymbol{\beta}_0(\boldsymbol{w}_k)$ satisfies the Fredholm integral equation

$$\begin{aligned}\sqrt{nh^q}\big\{\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_k) - \boldsymbol{\beta}_0(\boldsymbol{w}_k)\big\} &= \boldsymbol{D}(\boldsymbol{w}_k)^{-1}\int_{\mathcal{W}}\boldsymbol{\Psi}(\boldsymbol{w}_k;\boldsymbol{u})\sqrt{nh^q}\big\{\widehat{\boldsymbol{\beta}}(\boldsymbol{u}) - \boldsymbol{\beta}_0(\boldsymbol{u})\big\}\,\mathrm{d}\boldsymbol{u}\\ &\quad + \boldsymbol{D}(\boldsymbol{w}_k)^{-1}\sqrt{nh^q}\boldsymbol{M}_n(\boldsymbol{w}_k)\\ &\quad + O_p(\sqrt{nh^{-q}m^{-2}}) + O(\sqrt{nh^{q+4}}) + O_p\left(\frac{\log h}{\sqrt{nh^q}}\right) + O(h^2),\end{aligned}$$

where $\boldsymbol{M}_n(\cdot)$ is defined by (9) on page 22.

Define

$$\begin{aligned}\widetilde{\boldsymbol{b}}(\boldsymbol{w}_k) &= \frac{1}{nm}\sum_{i=1}^{n}\int_0^{\tau}\big\{K_h(\boldsymbol{W}_i - \boldsymbol{w}_k)\boldsymbol{X}_i - m\widetilde{\boldsymbol{X}}(t,\boldsymbol{w}_k)\big\}\,\mathrm{d}N_i(t),\\ \widetilde{\boldsymbol{b}}_{\alpha} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\big\{\boldsymbol{Z}_i - \widetilde{\boldsymbol{Z}}(t)\big\}\,\mathrm{d}N_i(t),\\ \widetilde{\boldsymbol{V}}(\boldsymbol{w}_k,\boldsymbol{w}_\ell) &= \frac{I(\boldsymbol{w}_k = \boldsymbol{w}_\ell)}{nm}\sum_{i=1}^{n}\int_0^{\tau}K_h(\boldsymbol{W}_i - \boldsymbol{w}_k)Y_i(t)\boldsymbol{X}_i^{\otimes 2}\,\mathrm{d}t\\ &\quad - \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}Y_i(t)\widetilde{\boldsymbol{X}}(t,\boldsymbol{w}_k)\widetilde{\boldsymbol{X}}(t,\boldsymbol{w}_\ell)^{\mathrm{T}}\,\mathrm{d}t,\\ \widetilde{\boldsymbol{V}}_{\beta\alpha}(\boldsymbol{w}_k) &= \frac{1}{nm}\sum_{i=1}^{n}\int_0^{\tau}K_h(\boldsymbol{W}_i - \boldsymbol{w}_k)Y_i(t)\boldsymbol{X}_i\big\{\boldsymbol{Z}_i - \widetilde{\boldsymbol{Z}}(t)\big\}^{\mathrm{T}}\,\mathrm{d}t,\\ \widetilde{\boldsymbol{V}}_{\alpha\alpha} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}Y_i(t)\big\{\boldsymbol{Z}_i - \widetilde{\boldsymbol{Z}}(t)\big\}^{\otimes 2}\,\mathrm{d}t,\\ \widetilde{\boldsymbol{X}}(t,\boldsymbol{w}_k) &= \frac{\sum_{i=1}^{n}K_h(\boldsymbol{W}_i - \boldsymbol{w}_k)Y_i(t)\boldsymbol{X}_i}{m\sum_{i=1}^{n}Y_i(t)}.\end{aligned}$$

The above terms are obtained by replacing $m^{-1}\sum_{j=1}^{m}K_h(\boldsymbol{W}_i - \boldsymbol{w}_j)$ in (2.3) with 1. Note that

$$m^{-1}\sum_{j=1}^{m}K_h(\boldsymbol{W}_i - \boldsymbol{w}_j) = \int_{\mathcal{W}}K_h(\boldsymbol{W}_i - \boldsymbol{w})\,\mathrm{d}\boldsymbol{w} + O_p(h^{-q}m^{-1}) = 1 + O_p(\exp(-ah^{-2})) + O_p(h^{-q}m^{-1}),$$

for $i = 1, \ldots, n$. For simplicity, we assume that on the right-hand side above, the second term is dominated by the third term. Thus, it can be verified that $\boldsymbol{b}(\boldsymbol{w}_k) - \widetilde{\boldsymbol{b}}(\boldsymbol{w}_k) = O_p(h^{-q}m^{-2})$, $\boldsymbol{V}(\boldsymbol{w}_k, \boldsymbol{w}_\ell) - \widetilde{\boldsymbol{V}}(\boldsymbol{w}_k, \boldsymbol{w}_\ell) = O_p(h^{-q}m^{-3})$, $\boldsymbol{V}_{\beta\alpha}(\boldsymbol{w}_k) - \widetilde{\boldsymbol{V}}_{\beta\alpha}(\boldsymbol{w}_k) = O_p(h^{-q}m^{-2})$, $\boldsymbol{V}_{\alpha\alpha} - \widetilde{\boldsymbol{V}}_{\alpha\alpha} = O_p(h^{-q}m^{-1})$, and $\boldsymbol{b}_\alpha - \widetilde{\boldsymbol{b}}_\alpha = O_p(h^{-q}m^{-1})$.

The proposed estimator is obtained by solving the following system of equations:

$$\begin{pmatrix} \sum_{j=1}^m \boldsymbol{V}(\boldsymbol{w}_1, \boldsymbol{w}_j)\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_j) + \boldsymbol{V}_{\beta\alpha}(\boldsymbol{w}_1)\widehat{\boldsymbol{\alpha}} \\ \vdots \\ \sum_{j=1}^m \boldsymbol{V}(\boldsymbol{w}_m, \boldsymbol{w}_j)\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_j) + \boldsymbol{V}_{\beta\alpha}(\boldsymbol{w}_m)\widehat{\boldsymbol{\alpha}} \\ \sum_{j=1}^m \boldsymbol{V}_{\beta\alpha}(\boldsymbol{w}_j)^{\mathrm{T}}\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_j) + \boldsymbol{V}_{\alpha\alpha}\widehat{\boldsymbol{\alpha}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{b}(\boldsymbol{w}_1) \\ \vdots \\ \boldsymbol{b}(\boldsymbol{w}_m) \\ \boldsymbol{b}_\alpha \end{pmatrix}. \tag{4}$$

For $k = 1, \ldots, m$, the kth subvector of the left-hand side of (4) can be expressed as

$$\begin{aligned} &\sum_{j=1}^m \boldsymbol{V}(\boldsymbol{w}_k, \boldsymbol{w}_j)\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_j) + \boldsymbol{V}_{\beta\alpha}(\boldsymbol{w}_k)\widehat{\boldsymbol{\alpha}} \\ &= \sum_{j=1}^m \widetilde{\boldsymbol{V}}(\boldsymbol{w}_k, \boldsymbol{w}_j)\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_j) + \widetilde{\boldsymbol{V}}_{\beta\alpha}(\boldsymbol{w}_k)\widehat{\boldsymbol{\alpha}} + O_p(h^{-q}m^{-2}) \\ &= \frac{1}{m}\boldsymbol{D}_n(\boldsymbol{w}_k)\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_k) - \frac{1}{m^2}\sum_{j=1}^m \int_0^\tau \frac{\boldsymbol{S}_{nX}(t, \boldsymbol{w}_k)\boldsymbol{S}_{nX}(t, \boldsymbol{w}_j)^{\mathrm{T}}}{S_{n0}(t)}\,\mathrm{d}t\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_j) \\ &\quad + \frac{1}{m}\int_0^\tau \left\{\boldsymbol{S}_{nZX}(t, \boldsymbol{w}_k)^{\mathrm{T}} - \frac{\boldsymbol{S}_{nX}(t, \boldsymbol{w}_k)\boldsymbol{S}_{nZ}(t)^{\mathrm{T}}}{S_{n0}(t)}\right\}\mathrm{d}t\widehat{\boldsymbol{\alpha}} \\ &\quad + O_p(h^{-q}m^{-2}). \end{aligned} \tag{5}$$

By the Doob–Meyer decomposition (Fleming and Harrington, 2011, Corollary 1.4.1), the observed event process $N_i(t)$ can be uniquely decomposed so that for every i, $N_i(t) = M_i(t) + \int_0^t Y_i(s)\{\lambda_0(s) + \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0(\boldsymbol{W}_i) + \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\alpha}_0\}\,\mathrm{d}s$, where $M_i(t)$ is a martingale. Thus, the kth subvector of the right-hand side of (4) can be expressed as

$$\begin{aligned} \boldsymbol{b}(\boldsymbol{w}_k) &= \widetilde{\boldsymbol{b}}(\boldsymbol{w}_k) + O_p(h^{-q}m^{-2}) \\ &= \frac{1}{nm}\sum_{i=1}^n \int_0^\tau \left\{K_h(\boldsymbol{W}_i - \boldsymbol{w}_k)\boldsymbol{X}_i - m\widetilde{\boldsymbol{X}}(t, \boldsymbol{w}_k)\right\}\mathrm{d}N_i(t) + O_p(h^{-q}m^{-2}) \\ &= \frac{1}{nm}\sum_{i=1}^n \int_0^\tau \left\{K_h(\boldsymbol{W}_i - \boldsymbol{w}_k)\boldsymbol{X}_i - m\widetilde{\boldsymbol{X}}(t, \boldsymbol{w}_k)\right\}\mathrm{d}M_i(t) \\ &\quad + \frac{1}{nm}\sum_{i=1}^n \int_0^\tau K_h(\boldsymbol{W}_i - \boldsymbol{w}_k)Y_i(t)\boldsymbol{X}_i^{\otimes 2}\,\mathrm{d}t\boldsymbol{\beta}_0(\boldsymbol{W}_i) \\ &\quad - \frac{1}{m^2}\sum_{j=1}^m \int_0^\tau \frac{\boldsymbol{S}_{nX}(t, \boldsymbol{w}_k)\boldsymbol{S}_{nX}(t, \boldsymbol{w}_j)^{\mathrm{T}}}{S_{n0}(t)}\,\mathrm{d}t\boldsymbol{\beta}_0(\boldsymbol{w}_j) \end{aligned}$$

$$+\frac{1}{m}\int_0^\tau \left\{\boldsymbol{S}_{nZX}(t,\boldsymbol{w}_k)^{\mathrm{T}} - \frac{\boldsymbol{S}_{nX}(t,\boldsymbol{w}_k)\boldsymbol{S}_{nZ}(t)^{\mathrm{T}}}{S_{n0}(t)}\right\}\mathrm{d}t\boldsymbol{\alpha}_0$$
$$+O_p(h^{-q}m^{-2}) + O(h^2m^{-1}), \tag{6}$$

because $\boldsymbol{\beta}_0(\boldsymbol{W}_i) = m^{-1}\sum_{j=1}^m K_h(\boldsymbol{W}_i - \boldsymbol{w}_j)\boldsymbol{\beta}_0(\boldsymbol{w}_j) + O_p(h^{-q}m^{-1}) + O(h^2)$ for $i = 1,\ldots,n$. The last equation in (4) yields

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}} &= \boldsymbol{V}_{\alpha\alpha}^{-1}\Big\{\boldsymbol{b}_\alpha - \sum_{j=1}^m \boldsymbol{V}_{\beta\alpha}(\boldsymbol{w}_j)^{\mathrm{T}}\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_j)\Big\} \\
&= \widetilde{\boldsymbol{V}}_{\alpha\alpha}^{-1}\Big\{\widetilde{\boldsymbol{b}}_\alpha - \sum_{j=1}^m \widetilde{\boldsymbol{V}}_{\beta\alpha}(\boldsymbol{w}_j)^{\mathrm{T}}\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_j)\Big\} + O_p(h^{-q}m^{-1}) \\
&= \left[\int_0^\tau \left\{\boldsymbol{S}_{nZ^2}(t) - \frac{\boldsymbol{S}_{nZ}(t)^{\otimes 2}}{S_{n0}(t)}\right\}\mathrm{d}t\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^n\int_0^\tau \{\boldsymbol{Z}_i - \widetilde{\boldsymbol{Z}}(t)\}\,\mathrm{d}N_i(t)\right. \\
&\quad \left. -\frac{1}{m}\sum_{j=1}^m\int_0^\tau \left\{\boldsymbol{S}_{nZX}(t,\boldsymbol{w}_j)^{\mathrm{T}} - \frac{\boldsymbol{S}_{nX}(t,\boldsymbol{w}_j)\boldsymbol{S}_{nZ}(t)^{\mathrm{T}}}{S_{n0}(t)}\right\}^{\mathrm{T}}\mathrm{d}t\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_j)\right] \\
&\quad +O_p(h^{-q}m^{-1}).
\end{aligned} \tag{7}$$

Substituting (7) into (5) and setting it to equal (6), we have

$$\begin{aligned}
&\boldsymbol{D}_n(\boldsymbol{w}_k)\sqrt{nh^q}\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_k) - \frac{1}{n}\sum_{i=1}^n\int_0^\tau K_h(\boldsymbol{W}_i - \boldsymbol{w}_k)Y_i(t)\boldsymbol{X}_i^{\otimes 2}\,\mathrm{d}t\sqrt{nh^q}\boldsymbol{\beta}_0(\boldsymbol{W}_i) \\
&= \sqrt{nh^q}\boldsymbol{M}_n(\boldsymbol{w}_k) + \frac{1}{m}\sum_{j=1}^m\left[\int_0^\tau \frac{\boldsymbol{S}_{nX}(t,\boldsymbol{w}_k)\boldsymbol{S}_{nX}(t,\boldsymbol{w}_j)^{\mathrm{T}}}{S_{n0}(t)}\,\mathrm{d}t\right. \\
&\quad +\int_0^\tau\left\{\boldsymbol{S}_{nZX}(t,\boldsymbol{w}_k)^{\mathrm{T}} - \boldsymbol{S}_{nX}(t,\boldsymbol{w}_k)\frac{\boldsymbol{S}_{nZ}(t)^{\mathrm{T}}}{S_{n0}(t)}\right\}\mathrm{d}t\left[\int_0^\tau\left\{\boldsymbol{S}_{nZ^2}(t) - \frac{\boldsymbol{S}_{nZ}(t)^{\otimes 2}}{S_{n0}(t)}\right\}\mathrm{d}t\right]^{-1} \\
&\quad \left.\times\int_0^\tau\left\{\boldsymbol{S}_{nZX}(t,\boldsymbol{w}_j)^{\mathrm{T}} - \boldsymbol{S}_{nX}(t,\boldsymbol{w}_j)\frac{\boldsymbol{S}_{nZ}(t)^{\mathrm{T}}}{S_{n0}(t)}\right\}^{\mathrm{T}}\mathrm{d}t\right]\sqrt{nh^q}\{\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_j) - \boldsymbol{\beta}_0(\boldsymbol{w}_j)\} \\
&\quad +O(\sqrt{nh^{q+4}}) + O_p(\sqrt{nh^{-q}m^{-2}}),
\end{aligned} \tag{8}$$

where

$$\begin{aligned}
\boldsymbol{M}_n(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^n\int_0^\tau\Bigg(&K_h(\boldsymbol{W}_i - \boldsymbol{w})\boldsymbol{X}_i - m\widetilde{\boldsymbol{X}}(t,\boldsymbol{w}) \\
&-\left[\frac{1}{n}\sum_{j=1}^n\int_0^\tau K_h(\boldsymbol{W}_j - \boldsymbol{w})Y_j(t)\boldsymbol{X}_j\big\{\boldsymbol{Z}_j - \widetilde{\boldsymbol{Z}}(t)\big\}^{\mathrm{T}}\,\mathrm{d}t\right] \\
&\times\left[\frac{1}{n}\sum_{j=1}^n\int_0^\tau Y_j(t)\big\{\boldsymbol{Z}_j - \widetilde{\boldsymbol{Z}}(t)\big\}^{\otimes 2}\,\mathrm{d}t\right]^{-1}\big\{\boldsymbol{Z}_i - \widetilde{\boldsymbol{Z}}(t)\big\}\Bigg)\,\mathrm{d}M_i(t).
\end{aligned} \tag{9}$$

The terms in the outer square bracket of (8) are the sample counterpart of $\boldsymbol{\Psi}(\boldsymbol{w}_k, \boldsymbol{w}_j)$. By Lemma 1,

$$\begin{aligned}
&\int_0^\tau \frac{\boldsymbol{S}_{nX}(t,\boldsymbol{w}_k)\boldsymbol{S}_{nX}(t,\boldsymbol{w}_j)^{\mathrm{T}}}{S_{n0}(t)}\,\mathrm{d}t \\
&+\int_0^\tau \left\{\boldsymbol{S}_{nZX}(t,\boldsymbol{w}_k)^{\mathrm{T}} - \boldsymbol{S}_{nX}(t,\boldsymbol{w}_k)\frac{\boldsymbol{S}_{nZ}(t)^{\mathrm{T}}}{S_{n0}(t)}\right\}\mathrm{d}t\left[\int_0^\tau \left\{\boldsymbol{S}_{nZ^2}(t) - \frac{\boldsymbol{S}_{nZ}(t)^{\otimes 2}}{S_{n0}(t)}\right\}\mathrm{d}t\right]^{-1} \\
&\times\int_0^\tau \left\{\boldsymbol{S}_{nZX}(t,\boldsymbol{w}_j)^{\mathrm{T}} - \boldsymbol{S}_{nX}(t,\boldsymbol{w}_j)\frac{\boldsymbol{S}_{nZ}(t)^{\mathrm{T}}}{S_{n0}(t)}\right\}^{\mathrm{T}}\mathrm{d}t \\
=&\int_0^\tau \frac{\boldsymbol{s}_X(t,\boldsymbol{w}_k)\boldsymbol{s}_X(t,\boldsymbol{w}_j)^{\mathrm{T}}}{s_0(t)}\,\mathrm{d}t \\
&+\left\{\int_0^\tau \left(\boldsymbol{s}_{ZX}(t,\boldsymbol{w}_k)^{\mathrm{T}} - \frac{\boldsymbol{s}_X(t,\boldsymbol{w}_k)\boldsymbol{s}_Z(t)^{\mathrm{T}}}{s_0(t)}\right)\mathrm{d}t\right\}\left\{\int_0^\tau \left(\boldsymbol{s}_{Z^2}(t) - \frac{\boldsymbol{s}_Z(t)^{\otimes 2}}{s_0(t)}\right)\mathrm{d}t\right\}^{-1} \\
&\times\left\{\int_0^\tau \left(\boldsymbol{s}_{ZX}(t,\boldsymbol{w}_j)^{\mathrm{T}} - \frac{\boldsymbol{s}_X(t,\boldsymbol{w}_j)\boldsymbol{s}_Z(t)^{\mathrm{T}}}{s_0(t)}\right)\mathrm{d}t\right\}^{\mathrm{T}} + O_p\left(\frac{\log h}{\sqrt{nh^q}}\right) + O(h^2) \\
=&\,\boldsymbol{\Psi}(\boldsymbol{w}_k,\boldsymbol{w}_j) + O_p\left(\frac{\log h}{\sqrt{nh^q}}\right) + O(h^2).
\end{aligned}$$

Also, following similar arguments as in the proof of Lemma 1, it can be verified that

$$\frac{1}{n}\sum_{i=1}^n K_h(\boldsymbol{W}_i - \boldsymbol{w}_k)Y_i(t)\boldsymbol{X}_i^{\otimes 2}\boldsymbol{\beta}_0(\boldsymbol{W}_i) = \boldsymbol{s}_{X^2}(t,\boldsymbol{w}_k)\boldsymbol{\beta}_0(\boldsymbol{w}_k) + O_p\left(\frac{\log h}{\sqrt{nh^q}}\right) + O(h^2).$$

Thus, (8) can be expressed as

$$\begin{aligned}
&\boldsymbol{D}(\boldsymbol{w}_k)\sqrt{nh^q}\big\{\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_k) - \boldsymbol{\beta}_0(\boldsymbol{w}_k)\big\} \\
=&\sqrt{nh^q}\boldsymbol{M}_n(\boldsymbol{w}_k) + \frac{1}{m}\sum_{j=1}^m \boldsymbol{\Psi}(\boldsymbol{w}_k,\boldsymbol{w}_j)\sqrt{nh^q}\big\{\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_j) - \boldsymbol{\beta}_0(\boldsymbol{w}_j)\big\}\,\mathrm{d}\boldsymbol{w} \\
&+O_p(\sqrt{nh^{-q}m^{-2}}) + O(\sqrt{nh^{q+4}}) + O_p\left(\frac{\log h}{\sqrt{nh^q}}\right) + O(h^2).
\end{aligned}$$

For any function $\boldsymbol{\phi}$ with bounded derivative,

$$\frac{1}{m}\sum_{j=1}^m \boldsymbol{\phi}(\boldsymbol{w}_j)^{\mathrm{T}}\sqrt{nh^q}\big\{\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_j)-\boldsymbol{\beta}_0(\boldsymbol{w}_j)\big\} = \int_{\mathcal{W}} \boldsymbol{\phi}(\boldsymbol{w})^{\mathrm{T}}\sqrt{nh^q}\big\{\widehat{\boldsymbol{\beta}}(\boldsymbol{w})-\boldsymbol{\beta}_0(\boldsymbol{w})\big\}\,\mathrm{d}\boldsymbol{w}+O(m^{-1}).$$

We can write

$$\begin{aligned}
&\sqrt{nh^q}\big\{\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_k) - \boldsymbol{\beta}_0(\boldsymbol{w}_k)\big\} \\
=&\ \boldsymbol{D}(\boldsymbol{w}_k)^{-1}\sqrt{nh^q}\boldsymbol{M}_n(\boldsymbol{w}_k) + \boldsymbol{D}(\boldsymbol{w}_k)^{-1}\int_{\mathcal{W}} \boldsymbol{\Psi}(\boldsymbol{w}_k,\boldsymbol{w})\sqrt{nh^q}\big\{\widehat{\boldsymbol{\beta}}(\boldsymbol{w}) - \boldsymbol{\beta}_0(\boldsymbol{w})\big\}\,\mathrm{d}\boldsymbol{w} \\
&+O_p(\sqrt{nh^{-q}m^{-2}}) + O(\sqrt{nh^{q+4}}) + O_p\left(\frac{\log h}{\sqrt{nh^q}}\right) + O(h^2),
\end{aligned}$$

and hence

$$\begin{aligned}&\sqrt{nh^q}(\mathcal{I}-\mathcal{A})\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\big)(\boldsymbol{w}_k)\\&=\boldsymbol{D}(\boldsymbol{w}_k)^{-1}\sqrt{nh^q}\boldsymbol{M}_n(\boldsymbol{w}_k)+O_p(\sqrt{nh^{-q}m^{-2}})+O(\sqrt{nh^{q+4}})+O_p\left(\frac{\log h}{\sqrt{nh^q}}\right)+O(h^2).\end{aligned}$$

By Lemma 3, $(\mathcal{I}-\mathcal{A})$ is invertible, so

$$\begin{aligned}\sqrt{nh^q}\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\big)(\boldsymbol{w}_k)&=\sqrt{nh^q}(\mathcal{I}-\mathcal{A})^{-1}\boldsymbol{D}(\boldsymbol{w}_k)^{-1}\boldsymbol{M}_n(\boldsymbol{w}_k)\\&\quad+O_p(\sqrt{nh^{-q}m^{-2}})+O(\sqrt{nh^{q+4}})+O_p\left(\frac{\log h}{\sqrt{nh^q}}\right)+O(h^2).\end{aligned}$$

For brevity, we use $(\mathcal{I}-\mathcal{A})^{-1}\boldsymbol{D}(\boldsymbol{w}_k)^{-1}\boldsymbol{M}_n(\boldsymbol{w}_k)$ to represent the operation involving the inverse operator $(\mathcal{I}-\mathcal{A})^{-1}$ applied to the product of functions $\boldsymbol{D}^{-1}$ and $\boldsymbol{M}_n$, followed by evaluation at $\boldsymbol{w}_k$. Since $(\mathcal{I}-\mathcal{A})^{-1}=\mathcal{I}+(\mathcal{I}-\mathcal{A})^{-1}\mathcal{A}$, we have

$$\begin{aligned}\sqrt{nh^q}\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\big)(\boldsymbol{w}_k)=&\ \boldsymbol{D}(\boldsymbol{w}_k)^{-1}\sqrt{nh^q}\boldsymbol{M}_n(\boldsymbol{w}_k)+\sqrt{nh^q}(\mathcal{I}-\mathcal{A})^{-1}\mathcal{A}\boldsymbol{D}(\boldsymbol{w}_k)^{-1}\boldsymbol{M}_n(\boldsymbol{w}_k)\\&+O_p(\sqrt{nh^{-q}m^{-2}})+O(\sqrt{nh^{q+4}})+O_p\left(\frac{\log h}{\sqrt{nh^q}}\right)+O(h^2).\end{aligned}\tag{10}$$

By the martingale central limit theorem (Fleming and Harrington, 2011, Theorem 5.1.1), for $k=1,\ldots,m$, $\sqrt{nh^q}\boldsymbol{M}_n(\boldsymbol{w}_k)$ is asymptotically normal with mean zero and variance $\nu_0\int_0^\tau \boldsymbol{s}^*_{X^2}(t,\boldsymbol{w}_k)\,\mathrm{d}t$. By similar arguments, we can show that for $\boldsymbol{w}\in[0,1]^q$, $\sqrt{n}\mathcal{A}(\boldsymbol{D}(\boldsymbol{w})^{-1}\boldsymbol{M}_n(\boldsymbol{w}))$ is asymptotically normally distributed. For $\boldsymbol{w},\boldsymbol{w}'\in[0,1]^q$,

$$\begin{aligned}&\mathcal{A}(\boldsymbol{D}(\boldsymbol{w})^{-1}\boldsymbol{M}_n(\boldsymbol{w}))-\mathcal{A}(\boldsymbol{D}(\boldsymbol{w}')^{-1}\boldsymbol{M}_n(\boldsymbol{w}'))\\&=\int_{\mathcal{W}}\{\boldsymbol{D}(\boldsymbol{w})^{-1}-\boldsymbol{D}(\boldsymbol{w}')^{-1}\}\boldsymbol{\Psi}(\boldsymbol{w};\boldsymbol{u})\boldsymbol{D}(\boldsymbol{u})^{-1}\boldsymbol{M}_n(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}\\&\quad+\int_{\mathcal{W}}\boldsymbol{D}(\boldsymbol{w}')^{-1}\{\boldsymbol{\Psi}(\boldsymbol{w};\boldsymbol{u})-\boldsymbol{\Psi}(\boldsymbol{w}';\boldsymbol{u})\}\boldsymbol{D}(\boldsymbol{u})^{-1}\boldsymbol{M}_n(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}.\end{aligned}$$

Since $\boldsymbol{D}(\boldsymbol{w})$ and $\boldsymbol{\Psi}(\boldsymbol{w};\boldsymbol{u})$ are uniformly continuous in $w\in[0,1]^q$, so the process
$\{\mathcal{A}(\boldsymbol{D}(\boldsymbol{w})^{-1}\boldsymbol{M}_n(\boldsymbol{w})):\boldsymbol{w}\in[0,1]^q\}$ is uniformly equicontinuous. By Theorem 1.5.4 of van der Vaart and Wellner (1996), we conclude that the process $\sqrt{n}\mathcal{A}(\boldsymbol{D}(\boldsymbol{w})^{-1}\boldsymbol{M}_n(\boldsymbol{w}))$ converges to a Gaussian process, and thus $\sqrt{nh^q}(\mathcal{I}-\mathcal{A})^{-1}\mathcal{A}\boldsymbol{D}(\boldsymbol{w}_k)^{-1}\boldsymbol{M}_n(\boldsymbol{w}_k)=O_p(h^{q/2})$. Therefore, all but the first term on the right-hand side of (10) are $o_p(1)$, and

$$\sqrt{nh^q}\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\big)(\boldsymbol{w}_k)\xrightarrow{d}\mathrm{N}(\boldsymbol{0},\boldsymbol{\Sigma}(\boldsymbol{w}_k)),$$

where $\boldsymbol{\Sigma}(\boldsymbol{w}_k)=\boldsymbol{D}(\boldsymbol{w}_k)^{-1}\nu_0\int_0^\tau \boldsymbol{s}^*_{X^2}(t,\boldsymbol{w}_k)\,\mathrm{d}t\boldsymbol{D}(\boldsymbol{w}_k)^{-1}$.

Appendix A4 Estimators of asymptotic variances of regression parameter estimators

Based on the linear approximations of $\widehat{\boldsymbol{\beta}}(\boldsymbol{w}_0)$, $\int_{\mathcal{W}} \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{w})\widehat{\boldsymbol{\beta}}(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w}$, and $\widetilde{\boldsymbol{\alpha}}$ given in the proofs of Theorems 1–3, we can estimate the asymptotic variances $\boldsymbol{\Sigma}(\boldsymbol{w}_0)$, $\boldsymbol{\Sigma}_{\phi}$, and $\boldsymbol{\Sigma}_{\widetilde{\alpha}}$ by

$$
\begin{aligned}
\widehat{\boldsymbol{\Sigma}}(\boldsymbol{w}_0) = \; & \boldsymbol{D}_n(\boldsymbol{w}_0)^{-1}\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\Bigg(K_H(\boldsymbol{W}_i-\boldsymbol{w}_0)\boldsymbol{X}_i-\sum_{\ell=1}^{m}K_H(\boldsymbol{W}_i-\boldsymbol{w}_\ell)\overline{\boldsymbol{X}}(t,\boldsymbol{w}_0)\\
& -\frac{1}{n}\sum_{j=1}^{n}\int_0^{\tau}K_H(\boldsymbol{W}_j-\boldsymbol{w}_0)Y_j(t)\boldsymbol{X}_j\big\{\boldsymbol{Z}_j-\overline{\boldsymbol{Z}}(t)\big\}^{\mathrm{T}}\,\mathrm{d}t\\
& \times\left[\frac{1}{n}\sum_{j=1}^{n}\int_0^{\tau}\sum_{\ell=1}^{m}K_H(\boldsymbol{W}_j-\boldsymbol{w}_\ell)Y_j(t)\big\{\boldsymbol{Z}_j-\overline{\boldsymbol{Z}}(t)\big\}^{\otimes 2}\,\mathrm{d}t\right]^{-1}\\
& \times\sum_{\ell=1}^{m}K_H(\boldsymbol{W}_i-\boldsymbol{w}_\ell)\big\{\boldsymbol{Z}_i-\overline{\boldsymbol{Z}}(t)\big\}\Bigg)^{\otimes 2}\mathrm{d}N_i(t)\boldsymbol{D}_n(\boldsymbol{w}_0)^{-1},
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_{\phi} = & \int_0^{\tau}\frac{1}{n}\sum_{i=1}^{n}\Bigg\{Y_i(t)\Bigg(\boldsymbol{\phi}(\boldsymbol{W}_i)^{\mathrm{T}}(\mathcal{I}-\widehat{\mathcal{A}})^{-1}\boldsymbol{D}_n(\boldsymbol{W}_i)^{-1}\boldsymbol{X}_i\\
& -\frac{1}{n}\sum_{j=1}^{n}Y_j(t)\boldsymbol{\phi}(\boldsymbol{W}_j)^{\mathrm{T}}(\mathcal{I}-\widehat{\mathcal{A}})^{-1}\boldsymbol{D}_n(\boldsymbol{W}_j)^{-1}\boldsymbol{X}_jS_{n0}(t)^{-1}\\
& -\int_0^{\tau}\frac{1}{n}\sum_{j=1}^{n}Y_j(t)\boldsymbol{\phi}(\boldsymbol{W}_j)^{\mathrm{T}}(\mathcal{I}-\widehat{\mathcal{A}})^{-1}\boldsymbol{D}_n(\boldsymbol{W}_j)^{-1}\boldsymbol{X}_j\left\{\boldsymbol{Z}_j-\frac{\boldsymbol{S}_{nZ}(t)}{S_{n0}(t)}\right\}^{\mathrm{T}}\mathrm{d}t\\
& \times\left[\int_0^{\tau}\left\{\boldsymbol{S}_{nZ^2}(t)-\frac{\boldsymbol{S}_{nZ}(t)^{\otimes 2}}{S_{n0}(t)}\right\}\mathrm{d}t\right]^{-1}\left\{\boldsymbol{Z}_i-\frac{\boldsymbol{S}_{nZ}(t)}{S_{n0}(t)}\right\}\Bigg)^{\otimes 2}\Bigg\}\,\mathrm{d}N_i(t),
\end{aligned}
$$

and

$$
\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_{\widetilde{\alpha}} = & \left[\int_0^{\tau}\left\{\boldsymbol{S}_{nZ^2}(t)-\frac{\boldsymbol{S}_{nZ}(t)^{\otimes 2}}{S_{n0}(t)}\right\}\mathrm{d}t)\right]^{-1}\\
& \times\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\Bigg\{Y_i(t)\Bigg[\boldsymbol{Z}_i-\frac{\boldsymbol{S}_{nZ}(t)}{S_{n0}(t)}-\widehat{\boldsymbol{\phi}}(\boldsymbol{W}_i)^{\mathrm{T}}(\mathcal{I}-\widehat{\mathcal{A}})^{-1}\boldsymbol{D}_n(\boldsymbol{W}_i)^{-1}\boldsymbol{X}_i\\
& +\frac{1}{n}\sum_{j=1}^{n}\big\{Y_j(t)\widehat{\boldsymbol{\phi}}(\boldsymbol{W}_j)^{\mathrm{T}}(\mathcal{I}-\widehat{\mathcal{A}})^{-1}\boldsymbol{D}_n(\boldsymbol{W}_j)^{-1}\boldsymbol{X}_j\big\}S_{n0}(t)^{-1}\\
& +\int_0^{\tau}\frac{1}{n}\sum_{j=1}^{n}Y_j(t)\widehat{\boldsymbol{\phi}}(\boldsymbol{W}_j)^{\mathrm{T}}(\mathcal{I}-\widehat{\mathcal{A}})^{-1}\boldsymbol{D}_n(\boldsymbol{W}_j)^{-1}\boldsymbol{X}_j\left\{\boldsymbol{Z}_j-\frac{\boldsymbol{S}_{nZ}(t)}{S_{n0}(t)}\right\}^{\mathrm{T}}\mathrm{d}t\\
& \times\left[\int_0^{\tau}\left\{\boldsymbol{S}_{nZ^2}(t)-\frac{\boldsymbol{S}_{nZ}(t)^{\otimes 2}}{S_{n0}(t)}\right\}\mathrm{d}t\right]^{-1}\left\{\boldsymbol{Z}_i-\frac{\boldsymbol{S}_{nZ}(t)}{S_{n0}(t)}\right\}\Bigg]^{\otimes 2}\Bigg\}\,\mathrm{d}N_i(t)
\end{aligned}
$$

$$\times \left[\int_0^\tau \left\{ \boldsymbol{S}_{nZ^2}(t) - \frac{\boldsymbol{S}_{nZ}(t)^{\otimes 2}}{S_{n0}(t)} \right\} \mathrm{d}t) \right]^{-1},$$

with $\widehat{\boldsymbol{\phi}}(\boldsymbol{w}) = \int_0^\tau \{\boldsymbol{S}_{nZX}(t, \boldsymbol{w})^{\mathrm{T}} - \boldsymbol{S}_{nX}(t, \boldsymbol{w})\boldsymbol{S}_{nZ}(t)^{\mathrm{T}}/S_{n0}(t)\} \, \mathrm{d}t$, respectively.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Aalen OO (1989) A linear regression model for the analysis of life times. Statistics in Medicine 8:907–925

Chen K, Lin H, Zhou Y (2012) Efficient estimation for the Cox model with varying coefficients. Biometrika 99:379–392

Cheng S, Wei LJ, Ying Z (1995) Analysis of transformation models with censored data. Biometrika 82:835–845

Coleman L, Peter M, Teall T, Brannan R, Hanby A, Honarpisheh H, et al. (2009) Combined analysis of eIF4E and 4E-binding protein expression predicts breast cancer survival and estimates eIF4E activity. British Journal of Cancer 100:1393–1399

Cox DR (1972) Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) 34:187–202

Fan J, Zhang W (1999) Statistical estimation in varying coefficient models. The Annals of Statistics 27:1491–1518

Fan J, Lin H, Zhou Y (2006) Local partial-likelihood estimation for lifetime data. The Annals of Statistics 34:290–325

Fleming TR, Harrington DP (2011) Counting Processes and Survival Analysis. John Wiley & Sons

Hastie T, Tibshirani R (1993) Varying–coefficient models. Journal of the Royal Statistical Society: Series B (Methodological) 55:757–779

Hunter DJ (2005) Gene–environment interactions in human diseases. Nature Reviews Genetics 6:287–298

Inamdar KV, Romaguera JE, Drakos E, Knoblock RJ, Garcia M, Leventaki V, Medeiros J, Rassidakis G (2009) Expression of eukaryotic initiation factor 4E predicts clinical outcome in patients with mantle cell lymphoma treated with hyper-CVAD and rituximab, alternating with rituximab, high-dose methotrexate, and cytarabine. Cancer 115:4727–4736

Landi MT, Dracheva T, Rotunno M, Figueroa JD, Liu H, Dasgupta A, Mann FE, Fukuoka J, Hames M, Bergen AW, Murphy SE, Yang P, Pesatori AC, Consonni D, Bertazzi PA, Wacholder S, Shih JH, Caporaso NE, Jin J (2008) Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. PLoS One 3:e1651

Li H, Yin G, Zhou Y (2007) Local likelihood with time–varying additive hazards model. Canadian Journal of Statistics 35:321–337

Lin DY, Ying Z (1994) Semiparametric analysis of the additive risk model. Biometrika 81:61–71

McKeague IW, Sasieni PD (1994) A partly parametric additive risk model. Biometrika 81:501–514

Nelson ED, Benesch MG, Wu R, Ishikawa T, Takabe K (2024) High EIF4EBP1 expression reflects mTOR pathway activity and cancer cell proliferation and is a biomarker for poor breast cancer prognosis. American Journal of Cancer Research 14:227–242

No JH, Jeon YT, Park IA, Kim YB, Kim JW, Park NH, Kang SB, Han JY, Lim JM, Song YS (2011) Activation of mTOR signaling pathway associated with adverse prognostic factors of epithelial ovarian cancer. Gynecologic Oncology 121:8–12

Qu L, Song X, Sun L (2018) Identification of local sparsity and variable selection for varying coefficient additive hazards models. Computational Statistics & Data Analysis 125:119–135

Silverman BW (1986) Density Estimation for Statistics and Data Analysis. CRC press

The Cancer Genome Atlas Network (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499:43–49

Tian L, Zucker D, Wei L (2005) On the Cox model with time-varying regression coefficients. Journal of the American Statistical Association 100:172–183

Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ (2011) On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in Medicine 30:1105–1117

van der Vaart AW, Wellner JA (1996) Weak Convergence and Empirical Processes. Springer

Yin G, Li H, Zeng D (2008) Partially linear additive hazards regression with varying coefficients. Journal of the American Statistical Association 103:1200–1213

Zou Z, Tao T, Li H, Zhu X (2020) mTOR signaling pathway and mTOR inhibitors in cancer: progress and challenges. Cell & Bioscience 10:31