INDEPENDENT SCREENING FOR NONPARAMETRIC ADDITIVE COX MODEL

by

Sha Yu

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of philosophy in
Applied Mathematics

Charlotte

2020

Approved by:

Dr. Jiancheng Jiang

Dr. Zhiyi Zhang

Dr. Qingning Zhou

Dr. James Amburgey

Abstract

SHA YU. INDEPENDENT SCREENING FOR NONPARAMETRIC ADDITIVE COX MODEL. (Under the direction of DR. JIANCHENG JIANG)

Survival data with ultrahigh dimensional covariates are increasingly common recently due to the rapid development in technologies. It is challenging to model them using survival models in order to understand the association between covariate information and clinical information. In this paper, we focus on the nonparametric additive Cox's proportional model and propose an independent screening method for ultrahigh dimensional data. The proposed screening method is based on the favored bandwidth of the local partial likelihood estimator. Moreover, we develop a two-step procedure to recover all important covariates. This procedure first captures important variables with nonlinear impacts, and then identifies important variables with linear impacts. We further prove that the nonlinear step screening achieves the model selection consistency. Monte Carlo simulations are carried out to evaluate the performance of the proposed screening procedure, which provides evidence supporting the theory. Furthermore, we demonstrate the proposed methodology via a real data example.

ACKNOWLEDGEMENTS

First of all, I would like to express my deepest gratitude to my advisor Dr. Jiancheng Jiang for his consistent support throughout this long journey. I am so grateful for having him as my advisor and he has been a tremendous mentor for me.

I am deeply grateful for his guidance and patience for advising me, for his encouragement and inspiration. He navigated me to conquer the major tasks in my research. He inspired me to think and to learn. He watched me grow from a girl to a wife and to a mother. Beyond the academic support, he cares about my life, my family, and my daughter. Words can not express my gratitude. What I have learned from him will benefit me for the rest of my life.

Moreover, I would like to thank my committee members, Dr. Zhiyi Zhang, Dr. Qingning Zhou and Dr. James Amburgey, for their time and insightful comments.

I gratefully acknowledge the financial support received towards my Ph.D. from the Graduate School and Mathematics department. I would like to extend my thanks to my loving friends and all the people who help me throughout my Ph.D. journey.

Last but not the least, I would like to express my thanks to my family for their endless support.

Contents

List of Tables

List of Figures

CHAPTER 1: INTRODUCTION

Due to the development of information technology, a massive amount of covariate information has been collected in survival analysis. However, it is likely that only a fraction of the covariates are associated with clinical time. In practice, a parsimonious model may be preferred for improving the predictability and interpretability of the model. Thus, how to select relevant covariates is crucial in modeling, which has paved the way for variable selection in survival analysis.

Survival analysis is used to analyze the time until an event of interest occurs. The time can be measured in years, months, weeks, etc, and it records the time duration starting from a predefined time point until an event occurs. This time variable is called survival time. The event of interest can refer to death, relapse, credit default, or any designated experience of interest that might happen. The main goal of survival analysis is to study the association between survival time T and a vector of covariate variables $\mathbf{X}$. This goal is often achieved by studying the conditional hazard function of T given $\mathbf{X} = \mathbf{x}$, denoted by $\lambda(t|\mathbf{x})$. The definition of $\lambda(t|\mathbf{x})$ is

$$\lambda(t|\mathbf{x}) = \lim_{\Delta t \downarrow 0} \frac{P\{t \leq T < t + \Delta t | T \geq t, \mathbf{X} = \mathbf{x}\}}{\Delta t}.$$

It gives the instantaneous hazard rate at time t conditionally on that the individual has survived up to time t and a given value of covariate.

The Cox model introduced by Cox (1972)[1] is the most widely used model in survival analysis. The model has a simple form and the convenience in dealing with censoring contributes to its popularity. The model is given by

$$\lambda(t|\mathbf{x}) = \lambda_0(t)\Psi(\mathbf{x}).$$

The first factor $\lambda_0(t)$ represents the baseline hazard function and it is the conditional function of T given $\mathbf{X} = 0$ when $\Psi(0) = 1$. The second factor $\Psi(\mathbf{x})$ is the covariate effect. The Cox model assumes that the ratio of the hazards for any two individuals is constant over time.

Taking the preparametrization $\Psi(\mathbf{x}) = exp(\psi(\mathbf{x}))$, Cox proposed the proportional hazards regression model

$$\lambda(t|\mathbf{x}) = \lambda_0(t)exp(\psi(\mathbf{x})), \tag{1.1}$$

where $\psi(\mathbf{x})$ is referred as the risk function. See Fleming and Harrington (1991)[2], Anderson et al. (1993)[3] and references therein for more detailed literature review concerning this model. In this paper, we focus on the Cox proportional model.

Various methods have been developed for variable selection in survival analysis. For instance, Faraggi and Simon (1997)[4] worked on the Bayesian analysis of the Cox's proportional model to make inference about the parameters and proposed a variable selection method based on the Bayesian approach. Lee et al. (2011)[5] proposed a Bayesian variable selection scheme for a Bayesian semiparametric survival model. For the Bayesian variable selection procedure, it is computational intense to calculate the posterior probabilities in a high-dimensional setting. Many effective variable selection criterion are based on the penalization framework. Some classical variable selection techniques in linear regression models have been extended to survival analysis, such as AIC (Akaika, 1974[6]) and BIC (Schwartz[7]). In these methods, subset selection such as step wise selection and best subset selection are required. However, the subset selection suffers from the lack of stability (Breiman,1996[8]) and lack of interpretability in its theoretic properties. Tibshirani (1997)[9] extended the LASSO variable selection procedure to Cox model. Fan and Li (2001)[10] proposed a variable selection method with the smoothly clipped absolute deviation (SCAD) penalty based on a non-concave penalization likelihood and further extended it to Cox proportional model (Fan and LI, 2002[11]). In addition, they proved that the penalized likeli-

hood estimator possesses the oracle property. Cai, Fan and Li (2005)[12] worked on variable selection for multivariate survival data. Zhang and Lu (2007)[13] further developed the adaptive LASSO for Cox's proportional model. These penalization-based methods are showed to perform well in variable selection.

Many variable selection methods for Cox's proportional model relies on the linear assumption, which may be unrealistic in many practical situations. This motivates the research about Cox's proportional model with semi-parametric and nonparametric relative risk. Plenty of smoothing techniques, such as spline and local polynomial smoothing, have been applied to this model. Particularly, Fan, Gijbels and King (1997)[14] adopted the local polynomial smoothing methods and developed a local partial likelihood approach. Then risk function is obtained by maximizing the local partial likelihood. This approach is adopted in our work.

For Cox's proportional model with nonparametric risk function, it is rather challenging to deal with the variable selection. For nonparametric Cox model, Hastie and Tibshirani (1990)[15] modified the step-wise selection by considering several nonlinear model selection. Gray (1992, 1994 [16][17]) applied the spline smoothing technique to estimate the covariate effect and then performed model selection with hypothesis testing procedures. However, these approaches assume the Cox model with fixed dimensionality and thus cannot deal with ultrahigh dimensional data.

Due to the development of technology, high or ultrahigh dimensional data are increasingly common in many fields recently, and the demand for variable selection is more urgent. The definition of ultrahigh dimension introduced by Fan and Lv (2008)[18] is that the dimensionality grows exponentially with the sample size, i.e. $logp = O(n^{\alpha})$ for some $\alpha \in (0, 1/2)$. To deal with the ultrahigh dimensional data, Bradic, Fan and Jiang (2011)[19] generalized the penalized partial likelihood method to Cox's proportional model. Huang, et al. (2013)[20] worked on the absolute penalized maximum partial likelihood estimator in sparse, high-dimensional Cox pro-

portional hazards regression models. However, these variable selection methods may have problem in stability, accuracy and computation efficiency when dealing with ultrahigh dimensional data. This motivates the development of screening techniques, which can effectively reduce the number of covariates under consideration. Fan and Lv (2008)[18] proposed sure independent screening (SIS) for linear regression. They performed marginal regression on each predictor and ranked the variables based on the marginal correlation between each predictor and the response. They further proved that the SIS keeps all important predictor variables with the probability going to one. This idea has been extended to plenty of more general settings including generalized linear model (Fan and Song, 2010[21]), additive model (Fan et al., 2011[22]), varying-coefficient model (Liu et al., 2014[23]), quantile regression (He, Wang, and Hong, 2013[24]; Wu and Yin 2015[25]) and so forth. Fan, Feng and Wu (2010)[21] also extended the SIS to Cox's proportional model with the linear risk effect model assumption. Still for Cox model, Zhao and Li (2012)[26] proposed a principled sure independent screening method, in which the importance of predictors are quantified by the t-value of the estimated coefficient obtained by maximizing the marginal partial likelihood. However, all of these procedures are proposed under linear or parametric assumptions. To the best of our knowledge, there is no work on screening for nonparametric additive Cox's model with high or ultrahigh dimensional data. This motivates us to develop a marginal screening procedure for nonparametric additive Cox's model.

In this dissertation, we introduce a penalization form statistic to quantify the nonlinearity impact of each covariate, and we call it information criteria (IC). This IC is inspired by the likelihood cross validation (LCV) [27]. In this method, we first perform the marginal nonparametric smoothing on each covariate and get the maximum local partial likelihood estimator. Then we get the estimate of global partial likelihood by evaluating the maximum local partial likelihood estimator at each observation and

the negative global partial likelihood serves as the first term in IC. The penalty term is especially desired to achieve the goal of distinguishing the covariates with nonlinear impact from those with linear impact. For each covariate, the favored bandwidth can be attained by minimizing this IC. Then we rank the covariate based on its favored bandwidth from the smallest to the largest and keep the top ones to recover the covariates with nonlinear impacts. For the covariate not selected in this step, we further fit a parametric cox model and get the estimate of the coefficient. Then we divide the estimated coefficient by its standard deviation and get the corresponding z value. Similarly, we rank the covariate by the estimated z value in descending order and keep the top-ranked covariates. Theses covariates are regarded as important variables with linear impact. Combining these two steps, we accomplish the goal of identifying all the important covariates.

The contribution of this paper arises in two aspects. First, we propose the information criteria (IC) to quantify the nonlinear impact of covariate for nonparametric additive Cox model. It is also proved that the screening step based on this information criteria (IC) achieves the model selection consistency. Second, we develop a two-step screening method to identify all the important covariates for nonparametric additive Cox model in ultrahigh dimensional case.

The remainder of the thesis is organized as follows. In section 2, we first introduce the local partial likelihood and the maximum local partial likelihood estimator of the Cox's proportional model. Then we derive an approximation form of the likelihood cross validation. After this, we propose the information criteria (IC) and non-linearity measure and then develop a two-step screening procedure. At last, we derive the influence function for nonparametric Cox's proportional model, which dramatically improve the computing efficiency of the proposed screening procedure. Five theorems are presented in Section 3 in order to establish the model selection consistency. In Section 4, we conduct extensive simulations to assess the performance of proposed

procedure. We conclude this dissertation in Section 6. Proofs of the theorems and derivation of the influence function are relegated to the Appendix.

CHAPTER 2: METHODOLOGY

We consider the multivariate data set $\{(\mathbf{X}_i, T_i) : i = 1, \ldots, n\}$, which forms an i.i.d. sample from the population $(\mathbf{X}, T)$, where $\mathbf{X} = (X_1, \cdots, X_d)^T$ and $\mathbf{X}_i = (X_{i1}, \cdots, X_{id})^T$ is a column vector of covariates for individual i. In practice, the survival times $T_1, \cdots, T_n$ are not fully observed for a variety of reasons, for instance, the termination of study or the lost of follow-up during the study period. We refer to this kind of incomplete observations as right censored. We consider the independent censoring scheme here. That is, we assume the censoring times $C_1, \ldots, C_n$ are independent of the survival times $T_1, \ldots, T_n$ given the covariates X. We denote the observed event time for individual i as $Z_i = min(T_i, C_i)$ with a censoring indicator $\delta_i = I\{T_i \leq C_i\}$. $\delta_i = 1$ means the survival time is observed and $\delta_i = 0$ means the survival time is censored. Now we form the observed data as

$$\{(\mathbf{X}_i, Z_i, \delta_i) : \mathbf{X}_i \in \mathbb{R}^d, \delta_i \in \{0, 1\}, i = 1, \ldots, n\},$$

which are i.i.d samples from the population

$$(\mathbf{X}, min(T, C), \mathcal{I}\{T \leq C\}).$$

In this paper, we assume that the random variables T and C are continuous. Without loss of generality, we assume that each continuous covariate is in the range of $[0, 1]$.

For the ultrahigh dimensional case, that is, d grows that an exponential rate of the sample size, we work on the univariate Cox's proportional model on covariate $\mathbf{X}_k$ for

$k = 1, \ldots, d$,

$$\lambda_k(t|\mathbf{X}_k) = \lambda_{0,k}(t) \exp\{\psi_k(\mathbf{X}_k)\}, \tag{2.1}$$

where $\lambda_{0,k}(t)$ is the baseline hazard function. We do not assume any parametric form about the risk effect $\psi_k(\cdot)$.

Then by Fan, Gijbels and King (1997)[14], under the proportional model 2.1,

$$\Psi_k(x) = \frac{E\{\delta|\mathbf{X}_k = \mathbf{x}_k\}}{E\{\Lambda_{0,k}(Z)|\mathbf{X}_k = \mathbf{x}_k\}}, \tag{2.2}$$

where $\Lambda_{0,k}(t) = \int_0^t \lambda_{0,k}(u)du$ is the cumulative baseline hazard function. This indicates that the function $\Psi_k(x)$ can be estimated using regression techniques if the baseline hazard function is known.

Consider for now the case $\psi_k(\mathbf{x}_k) = \psi_k(\mathbf{x}_k; \boldsymbol{\beta}_k)$ and $\lambda_{0,k}(t) = \lambda_{0,k}(t; \theta)$, then the log likelihood for the proportional hazards model 2.1 is

$$logL_k = \sum_{i=1}^{n} \left[\delta_i\{\lambda_{0,k}(Z_i; \theta) + \psi_k(X_{ik}; \boldsymbol{\beta}_k)\} - \Lambda_{0,k}(Z_i; \theta) exp\{\psi_k(X_{ik}; \boldsymbol{\beta}_k)\}\right]. \tag{2.3}$$

However, for robustness of inference we do not consider this situation with known baseline. Instead we do not pre-assume knowing its parametric form. This leads us to consider the local partial likelihood estimation (Fan, Gijbels and King, 1997[14]).

2.1 Local Partial Likelihood

For Cox's model, a standard approach to estimate the risk function $\psi(x)$ is the partial likelihood method (Cox, 1975[28]). Let $t_1^0 < \cdots < t_N^0$ denoted the ordered observed failure times and let (j) denote the label of the item failing at time t_j^0. Let $\mathcal{R}_j$ be the risk set right before time t_j^0: $\mathcal{R}_j = \{i : Z_i \geq t_j^0\}$. The 'least informative' nonparametric modeling of $\Delta_0(t)$ assumes that $\Delta_0(t)$ has a jump of size θ_j at time t_j^0:

$\Lambda_0(t;\theta) = \sum_{i=1}^{N} \theta_j \mathcal{I}\{t_j^0 \leq t\}$. Then

$$\Lambda_0(t;\theta) = \sum_{i=1}^{N} \theta_j \mathcal{I}\{i \in R_j\}. \tag{2.4}$$

Using the Breslow estimator of the baseline hazard function (Breslow 1972[29])

$$\hat{\theta}_j = \big[\sum_{i \in R_j} exp\{\psi_k(X_{ik};\boldsymbol{\beta}_k)\}\big]^{-1}. \tag{2.5}$$

Substituting 2.4 and 2.5 into 2.3 leads to the log partial likelihood function (Cox 1975)

$$\sum_{j=1}^{N} \Big[\psi_k\{X_{(j)k};\boldsymbol{\beta}_k\} - log\big\{ \sum_{i \in R_j} exp(\psi_k\{X_{ik};\boldsymbol{\beta}_k\})\big\}\Big]. \tag{2.6}$$

Let $Y_j(t) = \mathcal{I}(Z_j \geq t)$, then 2.6 is equivalent to

$$\sum_{i=1}^{n} \delta_i \Big[\psi_k\{X_{ik};\boldsymbol{\beta}_k\} - log\big\{ \sum_{j=1}^{n} Y_j(Z_i) exp(\psi_k\{X_{jk};\boldsymbol{\beta}_k\})\big\}\Big] \tag{2.7}$$

Note the partial likelihood in 2.6 and 2.7 is a profile likelihood and it can be derived from the full likelihood with the least informative nonparametric modelling of the baseline hazard function in 2.4.

Now suppose the form of $\psi_k(x)$ is unknown and the pth order derivative of $\psi_k(x)$ at the point x exists. Then for X in the neighborhood of x, by pth order Taylor's expansion,

$$\psi_k(X) \approx \psi_k(x) + \psi_k'(x)(X-x) + \cdots + \frac{\psi_k^{(p)}(x)}{p!}(X-x)^p.$$

We further define

$$\widetilde{\mathbf{X}}_k = \{1, X_k - x, \ldots, (X_k - x)^p\}^T \quad \text{and} \quad \widetilde{\mathbf{X}}_{ik} = \{1, X_{ik} - x, \ldots, (X_{ik} - x)^p\}^T,$$

where T denotes the transpose of a vector. Let h be the bandwidth parameter that controls the size of the local neighborhood and $K(\cdot)$ be a kernel function with compact support. Then locally around x, as $h \to 0$,

$$\psi_k(X_k) \approx \widetilde{\mathbf{X}}_k^T \boldsymbol{\beta}_k, \tag{2.8}$$

where

$$\boldsymbol{\beta}_k = (\beta_{0k}, \ldots, \beta_{pk})^T = \{\psi_k(x), \ldots, \frac{\psi_k^{(p)}(x)}{p!}\}^T.$$

With the local approximation 2.8, for covariate X_k, Fan, Gijbels and King (1997)[14] introduced the local partial likelihood

$$\mathcal{L}_x(\boldsymbol{\beta}_k) = \frac{1}{n}\sum_{i=1}^{n} \delta_i K_h(X_{ik} - x)\Big[\widetilde{\mathbf{X}}_{ik}^T \boldsymbol{\beta}_k - log\big\{\sum_{j=1}^{n} Y_j(Z_i) exp(\widetilde{\mathbf{X}}_{jk}^T \boldsymbol{\beta}_k) K_h(X_{jk} - x)\big\}\Big], \tag{2.9}$$

which is a localized version of the partial likelihood 2.7.

Let $\hat{\boldsymbol{\beta}}_k$ maximize 2.9 with respect to $\boldsymbol{\beta}_k = (\beta_{0k}, \ldots, \beta_{pk})^T$. Since 2.9 does not contain the intercept β_{0k}, then $\psi_k(x)$ cannot be directly estimated. Note that the derivative of $\psi_k(x)$ is estimated by $\hat{\beta}_{1k}$, then we get the estimate of $\psi_k(x)$

$$\hat{\psi}_k(x) = \int_0^x \hat{\beta}_{1k}(t)dt.$$

In practice, this integral can be approximated by Trapezoidal rule as suggested by Tibshirani and Hastie (1987)[30].

2.2 Maximum Local Partial Likelihood Estimator

Recall the local partial likelihood 2.9 does not involve the intercept $\beta_{0k} = \psi_k(x)$, we now define

$$\boldsymbol{\beta}_k^* = (\beta_{1k}, \ldots, \beta_{pk})^T, \quad \hat{\boldsymbol{\beta}}_k^* = (\hat{\beta}_{1k}, \ldots, \hat{\beta}_{pk})^T \quad \text{and}$$

$$\widetilde{\mathbf{X}}_{ik}^* = \{X_{ik} - x, \ldots, (X_{ik} - x)^p\}^T.$$

Then, let $\hat{\boldsymbol{\beta}}_k^*$ be the maximizer of the local partial likelihood

$$\mathcal{L}_x(\boldsymbol{\beta}_k^*) = \frac{1}{n}\sum_{i=1}^{n} \delta_i K_h(X_{ik} - x)\Big[\widetilde{\mathbf{X}}_{ik}^{*T}\boldsymbol{\beta}_k^* - log\big\{\sum_{j=1}^{n} Y_j(Z_i) exp(\widetilde{\mathbf{X}}_{jk}^{*T}\boldsymbol{\beta}_k^*) K_h(X_{jk} - x)\big\}\Big], \tag{2.10}$$

where $Y_j(t) = \mathcal{I}(Z_j \geq t)$.

Tibshirani[9] applied the iterative reweighted least squares (IRLS) strategy to obtain the LASSO estimator of parametric Cox's model. We extend the idea to non-parametric Cox model and solve for the estimator $\hat{\beta}_k^*$. To this end we define

$$\eta_{ik} = \widetilde{\mathbf{X}}_{ik}^{*T}\boldsymbol{\beta}_k^*, \quad \mathbf{W}_k = (\widetilde{X}_{1k}^*, \ldots, \widetilde{X}_{nk}^*)^T$$

,and

$$\boldsymbol{\eta}_k = (\eta_{1k}, \ldots, \eta_{nk})^T = \mathbf{W}_k\boldsymbol{\beta}_k^*.$$

Then

$$\mathcal{L}_x'(\boldsymbol{\beta}_k^*) = \frac{\partial \mathcal{L}_x(\boldsymbol{\beta}_k^*)}{\partial \boldsymbol{\beta}_k^*} = \mathbf{W}_k^T \frac{\partial \mathcal{L}_x(\boldsymbol{\eta}_k)}{\partial \boldsymbol{\eta}_k} \equiv \mathbf{W}_k^T \mathbf{B}_k,$$

$$\mathcal{L}_x{}''(\boldsymbol{\beta}_k^*) = \mathbf{W}_k^T \frac{\partial^2 \mathcal{L}_x(\boldsymbol{\eta}_k)}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\eta}_k^T} \mathbf{W}_k \equiv \mathbf{W}_k{}^T \mathbf{A}_k \mathbf{W}_k,$$

where $\mathbf{B}_k = \frac{\partial \mathcal{L}_x(\boldsymbol{\eta}_k)}{\partial \boldsymbol{\eta}_k}$ and $\mathbf{A}_k = -\frac{\partial \mathcal{L}_x^2(\boldsymbol{\eta}_k)}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\eta}_k^T}$.

Obviously, the maximum local likelihood estimator $\hat{\boldsymbol{\beta}}_k^*$ is attained by solving the

equation $\mathcal{L}_n'(\hat{\boldsymbol{\beta}}_k^*) = 0$. Applying the first order Taylor expansion of $\mathcal{L}_x'(\boldsymbol{\beta}_k^*)$ for a given initial value $\widetilde{\boldsymbol{\beta}}_k^0$, we obtain that

$$\begin{aligned} 0 = \mathcal{L}_x'(\boldsymbol{\beta}_k^*) &\approx \mathcal{L}_x'(\widetilde{\boldsymbol{\beta}}_k^0) + \mathcal{L}_x{}''(\widetilde{\boldsymbol{\beta}}_k^0)(\hat{\boldsymbol{\beta}}_k^* - \widetilde{\boldsymbol{\beta}}_k^0) \\ &= \mathbf{W}_k^T\mathbf{B}_k - \mathbf{W}_k{}^T\mathbf{A}_k\mathbf{W}_k(\hat{\boldsymbol{\beta}}_k^* - \widetilde{\boldsymbol{\beta}}_k^0), \end{aligned}$$

which is equivalent to

$$\hat{\boldsymbol{\beta}}_k^* = \widetilde{\boldsymbol{\beta}}_k^0 + (\mathbf{W}_k^T\mathbf{A}_k\mathbf{W}_k)^{-1}\mathbf{W}_k^T\mathbf{B}_k \equiv (\mathbf{W}_k^T\mathbf{A}_k\mathbf{W}_k)^{-1}\mathbf{W}_k^T\mathbf{A}_k\mathbf{C}_k, \tag{2.11}$$

where $\mathbf{C}_k = \boldsymbol{\eta}_k + \mathbf{A}_k{}^{-1}\mathbf{B}_k$. Therefore, $\hat{\boldsymbol{\beta}}_k^*$ minimizes $(\mathbf{C}_k - \boldsymbol{\eta}_k)^T\mathbf{A}_k(\mathbf{C}_k - \boldsymbol{\eta}_k)$.

The procedures for solving $\hat{\boldsymbol{\beta}}_k^*$ are described as follows:

1. Fix h and initialize $\widetilde{\boldsymbol{\beta}}_k^0$.
2. Compute $\boldsymbol{\eta}_k$, $\mathbf{A}_k$, $\mathbf{B}_k$ and $\mathbf{C}_k$ based on $\widetilde{\boldsymbol{\beta}}_k^0$.
3. Minimize $(\mathbf{C}_k - \boldsymbol{\eta}_k)^T\mathbf{A}_k(\mathbf{C}_k - \boldsymbol{\eta}_k)$ and update $\hat{\boldsymbol{\beta}}_k{}^*$.
4. Repeat step 2 and step 3 until $\hat{\boldsymbol{\beta}}_k{}^*$ does not change. At convergence, $\mathbf{A}_k$, $\mathbf{B}_k$ and $\mathbf{C}_k$ are calculated at $\hat{\boldsymbol{\beta}}_k{}^*$.

It could be time consuming to compute the n by n full matrix A_k and its inverse in above procedures. As suggested by Tibshirani (1997)[9], we replace the matrix A_k with a diagonal matrix, which has the same diagonal elements of A_k to ease the computation burden. Hastie and Tibishirani (1990)[15] argued that this modification has little impact on the performance.

Since $\hat{\psi}_k'(x) = \hat{\beta}_{1k} = e_1^T\hat{\boldsymbol{\beta}}_k{}^*$, where $e_1 = (1, 0, \ldots, 0)^T \in \mathbb{R}^p$. Assume $\psi_k(0) = 0$ and $X_{1k}, \ldots, X_{nk}$ in $[0, 1]$, then we obtain

$$\hat{\psi_k}(X_{ik}) = \int_0^{X_{ik}} \hat{\boldsymbol{\beta}}_k(t)dt = \int_0^{X_{ik}} e_1^T\hat{\boldsymbol{\beta}}_k{}^*(t)dt$$

for $j = 1, \ldots, n$ for predictor X_k.

2.3 Likelihood Cross Validation and Its Approximation

One major problem in local likelihood estimation is the choice of the bandwidth. One possible approach is to maximize the likelihood cross validation. We introduce the likelihood cross validation in this section and derive its approximation. The likelihood cross validation is the extension of cross validation and it is natural to be considered for the methods based on likelihood.

Deviance is a measure of goodness of fit and can be constructed by the likelihood. Now define the parameter vector $\hat{\boldsymbol{\psi}}_k = (\hat{\psi}_k(X_{1k}), \ldots, \hat{\psi}_k(X_{nk}))^T$, and let $\mathcal{L}(\hat{\boldsymbol{\psi}}_k)$ denoted the averaged log partial likelihood for predictor X_k as defined in 2.7. Then, we have

$$\mathcal{L}(\hat{\boldsymbol{\psi}}_k) = \frac{1}{n}\sum_{i=1}^{n} l(\delta_i, z_i, \hat{\psi}_k(X_{ik})) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \Big[\hat{\psi}_k(X_{ik}) - log\big\{\sum_{j=1}^{n} Y_j(Z_i) exp(\hat{\psi}_k(X_{jk}))\big\}\Big]. \tag{2.12}$$

The deviance is introduced to measure the difference of the log-likelihood between the fitted model and the saturated model. For a singe observation (X_{ik}, δ_i, z_i), the deviance is defined as [31]

$$D(\delta_i, z_i, \hat{\psi}_k(X_{ik})) = 2\Big(sup_{\psi_k} l(\delta_i, z_i, \psi_k(X_{ik})) - l(\delta_i, z_i, \hat{\psi}_k(X_{ik}))\Big).$$

Since the partial likelihood of fitted model $l(\delta_i, z_i, \hat{\psi}_k(X_{ik}))$ is always smaller than the partial likelihood of saturated model. As a consequence, the deviance is always non-negative. The total deviance is defined by

$$\sum_{i=1}^{n} D(\delta_i, z_i, \hat{\psi}_k(X_{ik})). \tag{2.13}$$

It is a generalization of the residual sum of squares for regression models [32].

Now we introduce the definition of likelihood cross validation (LCV) proposed by Habbema et al. (1974) [27]. The likelihood cross validation criterion is defined by substituting the leave-x_i-out estimate $\hat{\psi}_{k,-i}(X_{ik})$ in the total deviance 2.13,

$$\begin{aligned} LCV(\hat{\psi}_k) &= \sum_{i=1}^{n} D(\delta_i, z_i, \hat{\psi}_{k,-i}(X_{ik})) \\ &= C - 2\sum_{i=1}^{n} l(\delta_i, z_i, \hat{\psi}_{k,-i}(X_{ik})), \end{aligned} \tag{2.14}$$

where C depends on the observations (δ_i, z_i), but not the estimate $\hat{\psi}_k(X_{ik})$ and hence not the bandwidth.

It could be time consuming to compute the n leave-x_i-out estimates and thus approximations must be developed. One possible approach is to build connection between $l(\delta_i, z_i, \hat{\psi}_{k,-i}(X_{ik}))$ and $l(\delta_i, z_i, \hat{\psi}_k(X_{ik}))$. This is referred to as the method of infinitesimal perturbations and was first studied by Cook (1997)[33] for linear models. Next we first review the existing method of infinitesimal perturbations for local likelihood model and then propose our approach for nonparametric Cox's model.

2.3.1 Existing Method of Infinitesimal Perturbations for Local Likelihood Model

Loader[32] considered the likelihood regressing model, in which each response variable is assumed to have a density

$$Y_i \sim f(y, \theta(x_i)),$$

where $\theta(x_i)$ is an unknown function of the covariate x_i.

For local likelihood model, $\theta(x)$ is not assumed to have a parametric form and a polynomial is fitted locally instead. Let $l(y, \theta) = log(f(y, \theta)$, then the local polynomial

log-likelihood is defined as[32]:

$$\mathcal{L}_x(a) = \sum_{i=1}^{n} w_i(x) l(Y_i, < a, A(x_i - x) >),$$

where a is the coefficient vector,$A(\cdot)$ denotes a vector of fitting functions and $w_i(x) = W(\frac{x_i - x}{h})$ with a weight function $W(\cdot)$ and a bandwidth h. Then the estimate $\hat{a}$ is attained by maximizing the above equation.

Now define

$$\mathbf{X} = (A(x_1 - x), \ldots, A(x_n - x))^T,$$

$$\mathbf{W} = diag(w_1(x), \ldots, w_n(x)) \quad \text{and} \quad \mathbf{Y} = (Y_1, \ldots, Y_n)^T.$$

Loader [32] worked on the modified local likelihood equation

$$\mathbf{X}^T \mathbf{W} \dot{l}(Y, \mathbf{X}a) - \lambda W(0) e_1 \dot{l}(Y_i, < a, A(0) >) = 0, \tag{2.15}$$

where $\dot{l}(y, \theta)$ denotes the first partial derivative of $l(y, \theta)$ with respect to θ. According to above equation, $\hat{a}$ is a function of λ, denoted as $\hat{a}(\lambda)$. Specifically, $\hat{a}(0)$ is local likelihood estimator with all observations and $\hat{a}(1)$ is estimated by eliminating x_i.

Equation 2.15 is essentially the partial derivative of the full local likelihood subtracted by the ith observation's contribution to the local likelihood. This approach cannot be directly applied to Cox's model since the contribution of the ith observation is related to others and thus cannot be isolated. Next, we propose a new method to approximate the leave-x_i-out estimate with the full estimate for nonparametric Cox's model.

2.3.2 Method of Infinitesimal Perturbations for Nonparametric Cox's Model

Next, we work on method of infinitesimal perturbations for nonparametric Cox's model, that is, relate the estimate $\hat{\psi}_{k,-i}(X_{ik})$ with the estimate $\hat{\psi}_k(X_{ik})$.

Following the same notations as in Fan, Gijbels and King (1997)[14], we define

$$\boldsymbol{\beta}_k^0 = \{\psi_k'(x), \ldots, \psi_k^{(p)}(x)\}^T, \quad \hat{\boldsymbol{\alpha}}_k = H^*(\hat{\boldsymbol{\beta}}_k^* - \boldsymbol{\beta}_k^0) \quad \text{and} \quad \mathbf{U}_{ik} = (H^*)^{-1}\widetilde{\mathbf{X}}_{ik}^{*T},$$

where $H^* = diag(h, \ldots, h^P)$. Then, by 2.10, $\hat{\boldsymbol{\alpha}}_k$ maximizes

$$\begin{aligned}\mathcal{L}_x(\boldsymbol{\alpha}_k) = \frac{1}{n}\sum_{i=1}^{n} \delta_i K_h(X_{ik} - x)\Big[\widetilde{\mathbf{X}}_{ik}^{*T}\boldsymbol{\beta}_k^0 + \mathbf{U}_{ik}^T\boldsymbol{\alpha}_k \\ - log\big\{\sum_{j=1}^{n} Y_j(Z_i) exp(\widetilde{\mathbf{X}}_{ik}^{*T}\boldsymbol{\beta}_k^0 + \mathbf{U}_{ik}^T\boldsymbol{\alpha}_k)K_h(X_{jk} - x)\big\}\Big]\end{aligned} \tag{2.16}$$

with respect to α_k. We now propose a more general form. Let α_k maximizes

$$\begin{aligned}\mathcal{L}_x(\boldsymbol{\alpha}_k, \tau) = \frac{1}{n}\sum_{i=1}^{n} \delta_i K_h(X_{ik} - x)\mathcal{I}\{Z_i \le \tau\}\Big[\widetilde{\mathbf{X}}_{ik}^{*T}\boldsymbol{\beta}_k^0 + U_{ik}^T\boldsymbol{\alpha}_k - \\ log\big\{\sum_{j=1}^{n} Y_j(Z_i) exp(\widetilde{\mathbf{X}}_{ik}^{*T} + \mathbf{U}_{ik}^T\boldsymbol{\alpha}_k)K_h(X_{jk} - x)\big\}\Big],\end{aligned}$$

where τ denotes the observation ending time. Then in our case, $\tau = \infty$.

To simply the notation, we define

$$N_i(u) = \mathcal{I}\{Z_i \le u, \delta_i = 1\}, \quad Y_i(u) = \mathcal{I}\{Z_i \ge u\},$$

$$S_{nk}^{(0)}(\boldsymbol{\alpha}_k, u) = \frac{1}{n}\sum_{i=1}^{n} Y_i(u) exp(\widetilde{\mathbf{X}}_{ik}^{*T}\boldsymbol{\beta}_k^0 + \mathbf{U}_{ik}^T\boldsymbol{\alpha}_k)K_h(X_{ik} - x),$$

$$S_{nk}^{(1)}(\boldsymbol{\alpha}_k, u) = \frac{1}{n}\sum_{i=1}^{n} Y_i(u) exp(\widetilde{\mathbf{X}}_{ik}^{*T}\boldsymbol{\beta}_k^0 + \mathbf{U}_{ik}^T\boldsymbol{\alpha}_k)K_h(X_{ik} - x)\mathbf{U}_{ik},$$

and

$$S_{nk}^{(2)}(\boldsymbol{\alpha}_k, u) = \frac{1}{n}\sum_{i=1}^{n} Y_i(u) exp(\widetilde{\mathbf{X}}_{ik}^{*T}\boldsymbol{\beta}_k^0 + \mathbf{U}_{ik}^T\boldsymbol{\alpha}_k) K_h(X_{ik} - x)\mathbf{U}_{ik}\mathbf{U}_{ik}^T.$$

Using the notation of counting process, we get the local partial likelihood

$$\begin{aligned}\mathcal{L}_x(\boldsymbol{\alpha}_k, \tau) = \frac{1}{n}\int_o^\tau \sum_{i=1}^{n} K_h(X_{ik} - x)\Big[\widetilde{\mathbf{X}}_{ik}^{*T}\boldsymbol{\beta}_k^0 + \\ \mathbf{U}_{ik}^T\boldsymbol{\alpha}_k - log\big\{nS_{nk}^{(0)}(\boldsymbol{\alpha}_k, u)\big\}\Big] dN_i(u).\end{aligned} \tag{2.17}$$

To evaluate influence of perturbing the ith observation to the local likelihood, we consider the modified local likelihood equation,

$$(1-\lambda)\frac{\partial}{\partial\boldsymbol{\alpha}_k}\mathcal{L}(\boldsymbol{\alpha}_k, \tau) + \lambda\frac{\partial}{\partial\boldsymbol{\alpha}_k}\mathcal{L}_{x,-i}(\boldsymbol{\alpha}_k, \tau) = 0, \tag{2.18}$$

where $\mathcal{L}_{x,-i}(\boldsymbol{\alpha}_k, \tau)$ denotes the local partial likelihood with X_{ik} left out and λ is a parameter and the solution is $\hat{\boldsymbol{\alpha}}_k(\lambda)$. Note that $\hat{\boldsymbol{\alpha}}_k(0)$ is the full local partial log-likelihood estimate, while $\hat{\boldsymbol{\alpha}}_k(1)$ is the leave-x_i-out estimate. Taking derivative over λ on both sides of 2.18, we get

$$\begin{aligned}-\frac{\partial}{\partial\hat{\boldsymbol{\alpha}}_k}\mathcal{L}(\hat{\boldsymbol{\alpha}}_k, \tau) + (1-\lambda)\frac{\partial^2}{\partial\hat{\boldsymbol{\alpha}}_k\partial\hat{\boldsymbol{\alpha}}_k^T}\mathcal{L}_x(\hat{\boldsymbol{\alpha}}_k, \tau)\frac{\partial}{\partial\lambda}\hat{\boldsymbol{\alpha}}_k(\lambda) + \frac{\partial}{\partial\hat{\boldsymbol{\alpha}}_k}\mathcal{L}_{x,-i}(\hat{\boldsymbol{\alpha}}_k, \tau) \\ \lambda\frac{\partial^2}{\partial\hat{\boldsymbol{\alpha}}_k\partial\hat{\boldsymbol{\alpha}}_k^T}\mathcal{L}_{x,-i}(\hat{\boldsymbol{\alpha}}_k, \tau)\frac{\partial}{\partial\lambda}\hat{\boldsymbol{\alpha}}_k(\lambda) = 0.\end{aligned}$$

Letting $\lambda = 0$, we get

$$\frac{\partial}{\partial\hat{\boldsymbol{\alpha}}_k}[\mathcal{L}_{x,-i}(\hat{\boldsymbol{\alpha}}_k, \tau) - \mathcal{L}_x(\hat{\boldsymbol{\alpha}}_k, \tau)] + \frac{\partial^2}{\partial\hat{\boldsymbol{\alpha}}_k\partial\hat{\boldsymbol{\alpha}}_k^T}\mathcal{L}_x(\hat{\boldsymbol{\alpha}}_k, \tau)\frac{\partial}{\partial\lambda}\hat{\boldsymbol{\alpha}}_k(\lambda)|_{\lambda=0} = 0. \tag{2.19}$$

Note that

$$\frac{\partial^2}{\partial \boldsymbol{\alpha}_k \partial \boldsymbol{\alpha}_k^T} \mathcal{L}_x(\boldsymbol{\alpha}_k, \tau) \equiv \mathcal{L}_x''(\boldsymbol{\alpha}_k, \tau) = -\int_0^\tau \frac{S_{nk}^{(2)}(\boldsymbol{\alpha}_k, \tau) S_{nk}^{(0)}(\boldsymbol{\alpha}_k, \tau) - S_{nk}^{(1)}(\boldsymbol{\alpha}_k, \tau)^{\otimes 2}}{S_{nk}^{(0)}(\boldsymbol{\alpha}_k, \tau)^{\otimes 2}} \frac{1}{n} \sum_{i=1}^n K_h(X_{ik} - x) dN_i(u),$$

where $\otimes$ denotes the kronecker product.

We now define

$$\begin{aligned}
Q_{ik}(x, \tau) &\equiv \frac{\partial}{\partial \boldsymbol{\alpha}_k} \big[\mathcal{L}_x(\boldsymbol{\alpha}_k, \tau) - \mathcal{L}_{x,-i}(\boldsymbol{\alpha}_k, \tau)\big]|_{\boldsymbol{\alpha}_k = \hat{\boldsymbol{\alpha}}_k} \\
&= \frac{1}{n} \int_0^\tau \sum_{l=1}^n K_h(X_{lk} - x) \big[\mathbf{U}_{lk} - \frac{S_{nk}^{(1)}(\boldsymbol{\alpha}_k, \tau)}{S_{nk}^{(0)}(\boldsymbol{\alpha}_k, \tau)}\big] dN_l(u) - \\
&\quad - \frac{1}{n} \int_0^\tau \sum_{l=1, l \neq i}^n K_h(X_{lk} - x) \big[\mathbf{U}_{lk} - \frac{S_{nk,-i}^{(1)}(\boldsymbol{\alpha}_k, \tau)}{S_{nk,-i}^{(0)}(\boldsymbol{\alpha}_k, \tau)}\big] dN_l(u) \\
&= \frac{1}{n} \delta_i K_h(X_{ik} - x) \mathbf{U}_{ik} - \frac{1}{n} \int_0^\tau \sum_{l=1, l \neq i}^n K_h(X_{lk} - x) \big[\frac{S_{nk}^{(1)}(\boldsymbol{\alpha}_k, \tau)}{S_{nk}^{(0)}(\boldsymbol{\alpha}_k, \tau)}\big] dN_l(u) \\
&\quad + \frac{1}{n} \int_0^\tau \sum_{l=1, l \neq i}^n K_h(X_{lk} - x) \big[\frac{S_{nk,-i}^{(1)}(\boldsymbol{\alpha}_k, \tau)}{S_{nk,-i}^{(0)}(\boldsymbol{\alpha}_k, \tau)}\big] dN_l(u),
\end{aligned}$$

where

$$S_{nk,-i}^{(0)}(\boldsymbol{\alpha}_k, u) = \frac{1}{n} \sum_{l=1, l \neq i}^n Y_l(u) exp(\widetilde{\mathbf{X}}_{lk}^{*T} \boldsymbol{\beta}_k^0 + \mathbf{U}_{lk}^T \boldsymbol{\alpha}_k) K_h(X_{lk} - x) \quad \text{and}$$

$$S_{nk,-i}^{(1)}(\boldsymbol{\alpha}_k, u) = \frac{1}{n} \sum_{l=1, l \neq i}^n Y_l(u) exp(\widetilde{\mathbf{X}}_{lk}^{*T} \boldsymbol{\beta}_k^0 + \mathbf{U}_{lk}^T \boldsymbol{\alpha}_k) K_h(X_{lk} - x) \mathbf{U}_{lk}.$$

We also define $H_k(x, \tau) \equiv -\mathcal{L}_x''(\boldsymbol{\alpha}_k, \tau)|_{\boldsymbol{\alpha}_k = \hat{\boldsymbol{\alpha}}_k}$. Then from equation 2.19, we have

$$\frac{\partial}{\partial \lambda} \hat{\boldsymbol{\alpha}}_k(\lambda)|_{\lambda=0} = -H_k(x, \tau)^{-1} Q_{ik}(x, \tau).$$

Using the first order Taylor expansion, we get

$$\hat{\boldsymbol{\alpha}}_k(1) \approx \hat{\boldsymbol{\alpha}}_k(0) + \frac{\partial}{\partial\lambda}\hat{\boldsymbol{\alpha}}_k(\lambda)|_{\lambda=0}.$$

That is,

$$\hat{\boldsymbol{\alpha}}_{k,-i}(x) \approx \hat{\boldsymbol{\alpha}}_k(x) - H_k(x,\tau)^{-1}Q_{ik}(x,\tau). \tag{2.20}$$

For the case $p = 1$, we have

$$\hat{\alpha}_k(x) = h(\hat{\beta}_k^*(x) - \beta_k^0(x)) = h(\hat{\psi}_k'(x) - \psi_k'(x)). \tag{2.21}$$

Define

$$r_{ik}(x,\tau) = \int_0^x H_k(t,\tau)^{-1}Q_{ik}(t,\tau)dt. \tag{2.22}$$

In our case, $\tau = \infty$. To simply the notation, denote $r_{ik}(x) \equiv r_{ik}(x,\infty)$. Then combining 2.20, 2.21 and 2.22 gives

$$\begin{aligned}\hat{\psi}_{k,-i}(X_{ik}) &= \hat{\psi}_k(X_{ik}) - \frac{1}{h}\int_0^{X_{ik}} H_k(t,\tau)^{-1}Q_{ik}(t,\tau)dt \\ &\equiv \hat{\psi}_k(X_{ik}) - \frac{1}{h}r_{ik}(X_{ik}).\end{aligned} \tag{2.23}$$

Note that equation 2.23 helps achieve the goal of approximating the estimate $\hat{\psi}_{k,-i}(X_{ik})$ with the estimate $\hat{\psi}_k(X_{ik})$. And this significantly improves the computation efficiency. Since at each observation X_{ik}, the risk function is estimated by $\hat{\psi}_k(X_{ik}) = \int_0^{X_{ik}} \hat{\beta}_k(t)dt$. We apply trapezoidal rule to approximate this integral using the observation $\{X_{jk} : j = 1,\ldots,n \quad \text{and} \quad X_{jk} \le X_{ik}\}$ as the grid points. Before the approximation, we need to go through the iterative procedure as described in Section 2.1 multiple times to obtain the estimate $\hat{\beta}_{k,-i}(X_{jk})$ for all $j = 1,\ldots,n$ such that $X_{jk} \le X_{ik}$. But using the approximation 2.23, theses estimates can be approximated by $\hat{\beta}_k(X_{jk})$ directly without going through the iterative procedure.

Substituting 2.23 into $l(\delta_i, z_i, \hat{\psi}_{k,-i}(X_{ik}))$ gives

$$\begin{aligned} l(\delta_i, z_i, \hat{\psi}_{k,-i}(X_{ik})) &= \delta_i \Big[\hat{\psi}_{k,-i}(X_{ik}) - log\big\{ \sum_{j=1}^{n} Y_j(Z_i) exp(\hat{\psi}_{k,-i}(X_{jk})) \big\} \Big] \\ &= \delta_i \Big[\hat{\psi}_k(X_{ik}) - \frac{1}{h} r_{ik}(X_{ik}) - log\big\{ \sum_{j=1}^{n} Y_j(Z_i) e^{\hat{\psi}_k(X_{jk})} e^{-\frac{1}{h} r_{ik}(X_{jk})} \big\} \Big]. \end{aligned}$$

Since $\hat{\psi}_{k,-i}(X_{ik}) \to \hat{\psi}_k(X_{ik})$ as $n \to \infty$. Then by Taylor expansion of $r_{ik}(x)$ in the neighborhood of 0, we have

$$e^{-r_{ik}(x)} \approx 1 - r_{ik}(x).$$

Thus when n is large enough,

$$\begin{aligned} l(\delta_i, z_i, \hat{\psi}_{k,-i}(X_{ik})) \approx l(\delta_i, z_i, \hat{\psi}_k(X_{ik})) - \delta_i \Big[& \frac{1}{h} r_{ik}(X_{ik}) + \\ & log\big\{ 1 - \frac{\sum_{j=1}^{n} Y_j(Z_i) exp\big(\hat{\psi}_k(X_{jk}) r_{ik}(X_{jk})\big)}{h \sum_{j=1}^{n} Y_j(Z_i) exp(\hat{\psi}_k(X_{jk}))} \big\} \Big]. \end{aligned} \tag{2.24}$$

As a result, we could approximate LCV_k by

$$\begin{aligned} LCV_k(h) \approx C - 2n\mathcal{L}(\hat{\psi}_k) + 2 \sum_{i=1}^{n} \delta_i \Big[& \frac{1}{h} r_{ik}(X_{ik}) + \\ & log\big\{ 1 - \frac{\sum_{j=1}^{n} Y_j(Z_i) exp\big(\hat{\psi}_k(X_{jk}) r_{ik}(X_{jk})\big)}{h \sum_{j=1}^{n} Y_j(Z_i) exp(\hat{\psi}_k(X_{jk}))} \big\} \Big]. \end{aligned}$$

2.4 Information Criteria and Non-linearity Measure

Inspired by the likelihood cross validation, we now introduce the following statistic to extract the information from data concerning the optimal bandwidth. We name

the proposed statistic as information formation criteria (IC) and it is defined as

$$IC_k(h) = -\mathcal{L}(\hat{\psi}_k) + \frac{1}{n}\sum_{i=1}^{n}\delta_i\Big[\frac{1}{h}r_{ik}(X_{ik}) + log\big\{1 - \frac{\sum_{j=1}^{n} Y_j(Z_i)exp\big(\hat{\psi}_k(X_{jk})r_{ik}(X_{jk})\big)}{h\sum_{j=1}^{n} Y_j(Z_i)exp(\hat{\psi}_k(X_{jk}))}\big\}\Big]\tau(\frac{nlogd}{h})^{\frac{1}{2}}, \tag{2.25}$$

where τ is imposed to control the penalty level. And this specific penalty is specially designed to help group the covariates with small favored bandwidth and those with infinite favored bandwidth. The order in the penalty term is specially designed to represent the uniform property across d predictors. We will establish the theoretical property of IC in Section 3.

For each predictor X_k, we obtain the favored bandwidth according to $IC_k(h)$ as follows

$$\hat{h}_k = \underset{h}{\operatorname{argmin}}\, IC_k(h).$$

When $h = \infty$, to get the estimator $\hat{\psi}_k$, we need to maximize the local partial likelihood 2.10. Consider the case $p = 1$, then the local partial likelihood becomes

$$\frac{1}{nh\sqrt{2\pi}}\sum_{i=1}^{n}\delta_i\Big[X_{ik}\beta_k^* - log\big\{\sum_{j=1}^{n} Y_j(Z_i)\mathcal{R}_i exp(X_{jk}\beta_k^*)\big\}\Big] + \frac{log(h\sqrt{2\pi})}{nh\sqrt{2\pi}}\sum_{i=1}^{n}\delta_i.$$

Maximizing the above equation is equivalent to maximize the global partial likelihood of Cox model with parametric risk effect since the local term cancels out. The estimated coefficient is a constant and we denote it by $\widetilde{\beta}_k$. As a result, the corresponding risk function, denoted by $\widetilde{\psi}_k$, is a linear function of the predictor. That is, $\widetilde{\psi}_k(x) = \widetilde{\beta}_k x$. Now we define $\widetilde{\boldsymbol{\psi}}_k$ as

$$\begin{aligned}\widetilde{\boldsymbol{\psi}}_k &= (\psi_k(X_{1k}), \ldots, \psi_k(X_{nk}))^T \\ &= (\widetilde{\beta}_k X_{1k}, \ldots, \widetilde{\beta}_k X_{nk})^T.\end{aligned}$$

Then we take a look at the penalty term in the definition of IC_k and it equals 0 when $h = \infty$. Thus, we get

$$IC_k(\infty) = -\mathcal{L}(\widetilde{\boldsymbol{\psi}}_k) = -\frac{1}{n}\sum_{i=1}^{n}\delta_i\Big[\widetilde{\psi}_k(X_{ik}) - log\big\{\sum_{j=1}^{n}Y_j(Z_i)exp(\widetilde{\psi}_k(X_{jk}))\big\}\Big]. \quad (2.26)$$

Since the first term in $IC_k(h)$ is the likelihood and it is a measurement of the goodness of fit for the local partial likelihood estimator $\hat{\varphi}_k(\cdot)$. If the variable X_k has nonlinear impact, then the likelihood will decrease as h increases since a larger smoothing bias is yielded with a bigger smoothing bandwidth. This in fact implies that variables with nonlinear impact favor small bandwidth. On the other hand, the likelihood is not expected to present a big change as h varies if the variable has a linear impact. And an infinite bandwidth is preferred for these variables. Intuited by this, we can rank the variables by the favored bandwidth from smallest to the largest and keep those with small favored bandwidth. This proposed information criteria is inspired by Feng et al.[34]. They worked on the general nonparametric model and proposed a penalized log residual sum of squares with the Nadaraya-Waston (NW) estimator such that the optimal bandwidth obtained by minimizing this statistic can be treated as a measure of the variable importance. However, their approach cannot be applied to the Cox's proportional model. In our research, the information criteria is based on the likelihood cross validation with the local partial likelihood estimator. This IC involves the first and second derivative of the local partial likelihood of the Cox's proportional model, which have a much more complex form and thus brings challenge in establishing the theoretic results.

It could be time consuming to search over all possible values of h to get the optimal one for all predictors.Therefore, as suggest in Feng et al. (2018)[34], we evaluate the $IC_k(h)$ at two candidate values $h = h^* = (\frac{logp}{n})^{1/5}$ and $h = \infty$ for all predictors. Then the estimated index set is defined as $\hat{S} = \{k|IC_k(h^*) < IC_k(\infty)\}$. The theoretical

property will be studied in Section 3.

Note that the super parameter τ in $IC_k(h)$ needs to be set properly and it is challenging in practice. Motivated by this, we propose the non-linearity measure N_L for each predictor X_k,

$$N_L(k) = \frac{\mathcal{L}(\hat{\psi}_{k,h^*}(X_{ik})) - \mathcal{L}(\widetilde{\psi}_k(X_{ik}))}{\frac{1}{n}\sum_{i=1}^{n}\delta_i\Big[\frac{1}{h^*}r_{ik}(X_{ik}) + log\big\{1 - \frac{\sum_{j=1}^{n}Y_j(Z_i)e^{\hat{\psi}_k(X_{jk})}r_{ik}(X_{jk})}{h^*\sum_{j=1}^{n}Y_j(Z_i)e^{\hat{\psi}_k(X_{jk})}}\big\}\Big](\frac{nlogp}{h})^{\frac{1}{2}}}, \quad (2.27)$$

where $\hat{\psi}_{k,h^*}(\cdot)$ is the corresponding local partial likelihood estimator estimated at the bandwidth $h = c(\frac{logd}{n})^{\frac{1}{5}}$, where c is a constant. The numerator measures the difference of global partial likelihood between two choices of bandwidth. The denominator is designed to adjust the numerator by taking into account of the degrees of freedom. This non-linearity measure can be used to rank the non-linearity impact of the predictor. The larger the $N_L(k)$, the more non-linearity impact the predictor has.

2.5 Two-step Screening

With the definition of non-linearity measure, we now propose a two-step screening method to perform screening for Cox's proportional model with nonlinear risk effect. The two-step screening procedure is implemented is described as follows.

1. For each predictor, compute non-linearity measure $N_L(k)$. Rank the predictors by the value of $N_L(k)$ from the largest to the smallest. Keep the top ranked predictors using the threshold $\lfloor n/log(n) \rfloor$ as suggested by Fan and Lv (2008). These predictors are regarded as the important variables with nonlinear impact;

2. For the variables not selected in step 1, fit marginal parametric Cox model for each predictor to get parameter estimate $\widetilde{\beta}$ and variance estimate $I(\widetilde{\beta})^{-1}$. Then we compute the absolute z value for each predictor as $I(\widetilde{\beta})^{1/2}|\widetilde{\beta}|$ and rank the covariates by the absolute z values from the largest to the smallest. Keep the $\lfloor n/log(n) \rfloor$ top ranked covariates. These predictors are regarded as important

variables with linear impact.

2.6 Influence Function of Nonparametric Cox's Proportional Model

Influence function is a useful tool in identifying the influence of the observations. It can be used to quantify the effect of removing an observation of a statistic without recalculating it. N. Reid and H. Crepeau (1985)[35] presented the influence function for Cox's proportional model with linear from of risk effect. In this section, we present the influence function for nonparametric Cox's proportional model. It is useful for calibrating the influence of each observation to estimating the nonparametric Cox model. Further, it can be used to approximate the r_{ik} defined in 2.22 and thus tremendously improve the computation efficiency of computing the proposed non-linearity measure N_L.

The flow of the derivation is first to rewrite the local partial likelihood as a function of the empirical cumulative distribution function and then get the gateaux derivative of the function. The detailed works are presented in Appendix C.

The empirical influence function of nonparametric Cox model at the observation (t_i, X_{ik}, δ_i) at the point x_0 is obtained by solving the following equation,

$$A^*(\hat{\beta}_k^*)I\hat{F}_{ik} = \delta_i K_h(X_{ik}-x_0)\big[\widetilde{X}_{ik}^* - \frac{\sum_{j=1}^n Y_j(Z_i)K_h(X_{jk}-x_0)exp(\widetilde{X}_{jk}^{*T}\hat{\beta}_k^*)\widetilde{X}_{jk}^*}{\sum_{j=1}^n Y_j(Z_i)K_h(X_{jk}-x_0)exp(\widetilde{X}_{jk}^{*T}\hat{\beta}_k^*)}\big] + C_i^*(\hat{\beta}_k^*), \tag{2.28}$$

where

$$\begin{aligned} A^*(\hat{\beta}_k^*) = \frac{1}{n}\sum_{i=1}^n \delta_i K_h(X_{ik}-x_0)\big[&\frac{\sum_{j=1}^n Y_j(Z_i)K_h(X_{jk}-x_0)exp(\widetilde{X}_{jk}^{*T}\hat{\beta}_k^*)\widetilde{X}_{jk}^*\widetilde{X}_{jk}^{*T}}{\sum_{j=1}^n Y_j(Z_i)K_h(X_{jk}-x_0)exp(\widetilde{X}_{jk}^{*T}\hat{\beta}_k^*)} \\ &- \big(\frac{\sum_{j=1}^n Y_j(Z_i)K_h(X_{jk}-x_0)exp(\widetilde{X}_{jk}^{*T}\hat{\beta}_k^*)\widetilde{X}_{jk}^*}{\sum_{j=1}^n Y_j(Z_i)K_h(X_{jk}-x_0)exp(\widetilde{X}_{jk}^{*T}\hat{\beta}_k^*)}\big)^{\otimes 2}\big], \end{aligned}$$

and

$$C_i^*(\hat{\beta}_k^*) = K_h(X_{ik} - x_0)exp(\widetilde{X}_{ik}^{*T}\hat{\beta}_k^*)\big[- \widetilde{X}_{ik}^* \sum_{z_j \le z_i} \frac{\delta_j K_h(X_{jk} - x_0)}{\sum_{l=1}^n Y_l(Z_j)K_h(X_{lk} - x_0)exp(\widetilde{X}_{lk}^{*T}\hat{\beta}_k^*)})$$
$$+ \sum_{z_j \le z_i} \frac{\delta_j K_h(X_{jk} - x_0)\sum_{l=1}^n Y_l(Z_j)K_h(X_{lk} - x_0)exp(\widetilde{X}_{lk}^{*T}\hat{\beta}_k^*)\widetilde{X}_{lk}^*}{\{\sum_{l=1}^n Y_l(Z_j)K_h(X_{lk} - x_0)exp(\widetilde{X}_{lk}^{*T}\hat{\beta}_k^*)\}^2}\big].$$

For parametric Cox model, N. Reid and H. Crepeau (1985)[35] stated the connection among influence function, full parameter $\hat{\beta}$ and leave-x_i-out estimate $\hat{\beta}_{-i}$ as $I\hat{F}_{-i} \approx (n-1)(\hat{\beta} - \hat{\beta}_{-i})$. But the detailed verification were not provided in their work. We show that the similar conclusion holds for local likelihood estimator of nonparametric Cox's model, that is,

$$I\hat{F}_{ik} \approx (n-1)(\hat{\beta}_k^* - \hat{\beta}_{k,-i}^*), \tag{2.29}$$

where $\hat{\beta}_{k,-i}^*$ is the estimate of β_k^* obtained when eliminating the ith observation. The details are included in Appendix C. And the computational cost of computing influence function is much less than that of $\hat{\beta}_{k,-i}^*$.

Based on 2.29, when $p = 1$, we get

$$\hat{\psi}_k(x) - \hat{\psi}_{k,-i} \approx \frac{1}{n-1}\int_0^x I\hat{F}_{ik}(t)dt.$$

Together with 2.23, we have

$$r_{ik}(x) \approx \frac{h}{n-1}\int_0^x I\hat{F}_{ik}(t)dt. \tag{2.30}$$

For predictor X_k, in order to compute $r_{ik}(X_{jk})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, n$, we first form the order statistic $(X_{(1)k}, \ldots, X_{(n)k})$. Then we apply the following equation

to speed up the computation

$$r_{ik}(X_{(j)k}) = r_{ik}(X_{(j-1)k}) + \int_{X_{(j-1)k}}^{X_{(j)k}} I\hat{F}_{ik}(t)dt. \tag{2.31}$$

Now we study the computation efficiency between **??** and 2.22. Using 2.30, for each predictor X_k, we need to obtain the matrix

$$I\hat{F}_k = \begin{pmatrix} I\hat{F}_1(X_{1k}) & \dots & I\hat{F}_1(X_{nk}) \\ \vdots & & \vdots \\ I\hat{F}_n(X_{1k}) & \dots & I\hat{F}_n(X_{nk}). \end{pmatrix}$$

This matrix can be computed column by column since for each column, the local likelihood estimator $\hat{\beta}_k$ is the same and a whole column of C_i can be attained with one computation. However, with 2.22, r_{ik} need to be computed element-wise. As a result, the computational complexity is reduced from n^2 to n. Next we compare the computational cost between 2.31 and 2.30. Since the integral is approximated by the Trapezoidal rule, for each given X_{ik}, the observations $\{X_{jk} : j = 1, \dots, n \quad \text{and} \quad X_{jk} \leq X_{ik}\}$ are taken as the grid points. With 2.30, we need to solve the local likelihood estimator $\hat{\beta}$ totally $n(n+1)/2$ times for each predictor. However, we only need n such computations with 2.31 and this further improves the computation efficiency. Together, computing $r_{ik}(x)$ with 2.31 is much faster than using 2.22 and thus can significantly improve the computing efficiency of the non-linearity measure $N_L(k)$

CHAPTER 3: ASYMPTOTIC PROPERTIES

In Chapter 2, we introduce the non-linearity measure $N_L(k)$ to quantify the nonlinear impact of predictor X_k and propose the two-step screening method. In this chapter, we illustrate the sure independent screening property for the proposed method. Chapter 3 are organized as follows. In Section 1, we define related notations. Section 2 presents all theorems that we establish concerning the asymptotic properties of our estimator.

3.1 Notations

- Define $N_i(u) = \mathcal{I}\{Z_i \leq u, \delta_i = 1\}$ and $Y_i(u) = \mathcal{I}\{Z_i \geq u\}$.
- Put $P(u|x) = P\{Z \geq u|X = x\}$, $\Lambda(t,x) = \int_0^t P(u|x)\lambda_0(u)du$, and $\Lambda_k(\tau,x) = \int_0^\tau P(Z \geq z|X = x_k)\lambda_0(u)du$.
- Let $\mu_1 = \int uK(u)du$ and $v_1 = \int u^2K(u)du - {\mu_1}^2$.
- Denote

$$\Sigma_k(\tau,x) = f_k(x)\Psi_k(x)\Lambda_k(\tau,x)\int K^2(u)du$$

and

$$\widetilde{\Sigma}_k(\tau,x) = v_1 f_k(x)\Psi_k(x)\Lambda_k(x).$$

- Let the filtration $\mathcal{F}_{ik}$ be the statistical information accruing during the time $[0,\tau]$, that is,

$$\mathcal{F}_{ik} = \sigma\{X_{ik}, N_i(u), Y_i(u), i = 1,\ldots,n, 0 \leq u \leq \tau\}.$$

- Under the independent censoring scheme,

$$M_{ik}(u) = N_i(u) - \int_0^\tau Y_i(u) exp\{\psi_k(X_{ik})\}\lambda_0(u)du$$

is an orthogonal local square integrable martingale with respect to $\mathcal{F}_{ik}$.

- Let

$$S_{nk}^{(0)}(\boldsymbol{\alpha}_k, u) = \frac{1}{n}\sum_{i=1}^{n} Y_i(u) exp(\widetilde{\mathbf{X}}_{ik}^{*T}\boldsymbol{\beta}_k^0 + \mathbf{U}_{ik}^T\boldsymbol{\alpha}_k)K_h(X_{ik} - x),$$

$$S_{nk}^{(1)}(\boldsymbol{\alpha}_k, u) = \frac{1}{n}\sum_{i=1}^{n} Y_i(u) exp(\widetilde{\mathbf{X}}_{ik}^{*T}\boldsymbol{\beta}_k^0 + \mathbf{U}_{ik}^T\boldsymbol{\alpha}_k)K_h(X_{ik} - x)\mathbf{U}_{ik},$$

$$S_{nk}^{(2)}(\boldsymbol{\alpha}_k, u) = \frac{1}{n}\sum_{i=1}^{n} Y_i(u) exp(\widetilde{\mathbf{X}}_{ik}^{*T}\boldsymbol{\beta}_k^0 + \mathbf{U}_{ik}^T\boldsymbol{\alpha}_k)K_h(X_{ik} - x)\mathbf{U}_{ik}\mathbf{U}_{ik}^T.$$

- Define

$$E_{nk}(\boldsymbol{\alpha}_k, u) = \frac{S_{nk}^{(1)}(\boldsymbol{\alpha}_k, u)}{S_{nk}^{(0)}(\boldsymbol{\alpha}_k, u)},$$

$$V_{nk}(\boldsymbol{\alpha}_k, u) = \frac{S_{nk}^{(2)}(\boldsymbol{\alpha}_k, u)}{S_{nk}^{(0)}(\boldsymbol{\alpha}_k, u)} - E_{nk}(\boldsymbol{\alpha}_k, u)^{\otimes 2}.$$

- Let $\xi_k(x) = \sum_{i=1}^n \int_o^\tau K_h(X_{ik} - x)\{\mathbf{U}_{ik} - E_{nk}(0, u)\}dM_{ik}(u)$.

- Put

$$g_{nk}(Z_i, X_{ik}) = \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} Y_j(Z_i) exp(\psi_k(X_{jk}) \int_0^{(X_{ik}-X_{jk})/h} (\mu_1 - u)K(u)du,$$

and

$$d_{nk}(Z_i) = \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} Y_j(Z_i) exp(\psi_k(X_{jk}).$$

- Let $D_{ik} = E\{d_{nk}(Z_i)|Z_i\}$ and $G_{ik} = E\{g_{nk}(Z_i, X_{ik})|Z_i\}$. Define

$$\Omega_k = E\{\frac{1}{n}\sum_{i=1}^{n} \delta_i \widetilde{\Sigma}_k(\tau, X_{ik})^{-1} G_{ik}/D_{ik}\}.$$

3.2 Asymptotic Properties

In this section, we establish the asymptotic properties for the proposed screening procedure. We prove that the screening step to identify the nonlinear impact predictors possesses the model selection consistency. We impose some technical conditions for asymptotic properties and the conditions are presented in Appendix A. The proofs for lemmas, propositions and theorems are presented in the Appendix B.

Since the nonlinear impact predictors are captured based on the value of IC. The main challenges arise in how to develop the uniform deviation results of IC over all k predictors for ultrahigh dimensional data. We first present a Bahadur representation of the local likelihood estimator for nonparametric Cox model in Theorem 1. And then introduce a uniform exponential inequality for the martingale in the representation in Theorem 2. After this, we investigate the uniform deviation results of the local partial likelihood estimator $\hat{\psi}_k(\cdot)$ as in Theorem 3 based on the results of Theorem 1 and Theorem 2. Then we derive the uniform property of the second term in IC as in Theorem 4. Finally we show the model selection consistency property in Theorem 5.

For $p = 1$, we have

$$\hat{\psi}_k(x) - \psi_k(x) = \int_0^x \frac{1}{h}\hat{\alpha}_k(t)dt.$$

To work on the uniform result of $\hat{\psi}_k$ towards its limit, we first express the $\hat{\alpha}_k(x)$ as the sum of a martingales summation and a bias term based on the asymptotic results of $\hat{\alpha}_k(x)$ proved by Fan and Gijbels (1997)[14]. And the result is summarized in the following Theorem 1.

Theorem 1. Assume the Conditions 1-5 hold. Then for $p = 1$, we have

$$\hat{\alpha}_k = \frac{1}{n}\widetilde{\Sigma}_k(\tau, x)^{-1}\xi_k(x) + B_{2k}(\tau, x),$$

where

$$\xi_k(x) = \sum_{i=1}^{n} \int_0^{\tau} K_h(X_{ik} - x)\{U_{ik} - E_{nk}(0, u)\} dM_{ik}(u),$$

$$B_{2k}(\tau, x) = f_k(x)\Psi_k(x)\frac{\psi_k^{(2)}(x)}{2}\Lambda_k(\tau, x) \int K^2(u)(u - \mu_1)u^2 du h^2 + o_p(h^2),$$

and $\widetilde{\Sigma}_k(\tau, x) = v_1 f_k(x)\Psi_k(x)\Lambda_k(x)$.

In Theorem 1, the first term in $\hat{\alpha}_k$ is the variance term and the second term represents the bias. Note that $\widetilde{\Sigma}_k(\tau, x)$ is a constant. In the following, we work on deriving the uniform deviation of ξ_k over k predictors. The result is presented in Theorem 2. For multivariate data case, Bradic, Fan and Jiang (2011)[19] proved the uniform deviation of the score vector of the penalized log partial likelihood function. Inspired by their approach, we prove the following theorem.

Theorem 2. Assume conditions 1 to 7 in Appendix A hold. For any given positive sequence u_n bounded away from 0, we have

$$P(|\xi_k(\tau, x)| > n^{\frac{1}{2}} h^{-\frac{1}{2}} u_n) \leq c_0 exp\{-c_1 u_n\}$$

uniformly over k for given x, where c_0 and c_1 are positive constants. Further more,

$$\sup_{k=1,\ldots,d} |\xi_k| = O_p(a_n),$$

where $a_n = (\frac{nlogp}{n})^{\frac{1}{2}}$.

We have established a uniform exponential inequality for martingales in Theorem 2. This result plays a crucial role in establishing the model selection consistency for ultrahigh dimensional data. Next we form the uniform deviation result of the local

log partial likelihood estimator $\hat{\psi}_k$ over all predictors and the domain of x.

Theorem 3. Under the conditions of Theorem 2, we have

$$\sup_{x\in[0,1]} \sup_{k=1,\ldots,d} |\hat{\psi}_k(x) - \psi_k(x)| = O_p(a_n^*),$$

where $a_n^* = (\frac{logp}{nh^3})^{\frac{1}{2}} + h$.

Theorem 4 characterizes the uniform order of the P_k over k predictors, which contributes in building the uniform deviation property of the penalty term in IC.

Theorem 4. Assume Conditions 1 to 8 in Appendix A are satisfied. To simply the notations, we define

$$P_k = \frac{1}{n}\sum_{i=1}^{n} \delta_i \Big[\frac{1}{h} r_{ik}(X_{ik}) + log\big\{1 - \frac{\sum_{j=1}^{n} Y_j(Z_i) e^{\hat{\psi}_k(X_{jk})} r_{ik}(X_{jk})}{h\sum_{j=1}^{n} Y_j e^{\hat{\psi}_k(X_{jk})}}\big\}\Big]$$

then $IC_k(h)$ in (2.24) can be written as

$$IC_k(h) = -\mathcal{L}(\hat{\psi}_k) + P_k\tau(\frac{nlogp}{h})^{\frac{1}{2}}.$$

Then we have the following result,

$$\sup_{k=1,\ldots,d} |P_k - \frac{\Omega_k}{nh}| \xrightarrow{\text{P}} 0,$$

where $\Omega_k = E\{\frac{1}{n}\sum_{i=1}^{n} \Sigma_k(\tau, X_{ik})^{-1} G_{ik}/D_{ik}\}$ with G_{ik} and D_{ik} as defined in Section 3.1.

Now we present the selection consistency result in Theorem 5.

Theorem 5. Assume Conditions 1 to 9 in Appendix A are satisfied. Let $\hat{S} = \{k : IC_k(h^*) < IC_k(\infty)\}$. Let $S \subset \{1, \ldots, d\}$ be the index set of predictors with nonlinear impact. Then,

$$P(\hat{S} = S) \to 1 \quad \text{as} \quad n \to \infty.$$

From Theorem 5, we conclude that the non-linearity screening step possesses the sure screening property. More specifically, by sure screening we mean a property that all the important variables survive after variable screening with probability tending to one. It is desired for a screening method. The proofs are attached in the Appendix.

CHAPTER 4: SIMULATIONS

In this section, we conduct numerical simulations to evaluate the performance of proposed independent screening method for nonparametric addititive Cox's proportional model. The event time T is generated from the following transformed regression model (Fan, Gijbels and King, 1997[14]):

$$\log \Lambda_0(T) = -\psi(X) + \epsilon$$

where ϵ has standard exponential distribution. It is easy to generate data for Cox model using above model. We use the Weibull baseline hazard function of the form $\lambda_0(t) = 3\lambda t^2$ with $\lambda = \frac{1}{3}$.

In the simulation, X is taken to be marginally uniform distributed over $[0, 1]$. The correlation among the d covariates are designed following the autoregressive structure. More specifically, we first generate multivariate normally distributed variables $(\widetilde{Z}_1, \ldots, \widetilde{Z}_d)^T$ with mean $(0, \ldots, 0)^T$ and correlation structure as $\Sigma_{ij} = \rho^{|i-j|}$ .Then the cumulative distribution function of the variables $\widetilde{Z}_k$ with $k = 1, \ldots, d$ follows uniform distribution $[0, 1]$.

We denote

$$\psi_1(x) = 5sin(2\pi x - \frac{\pi}{2}) + 5,$$

$$\psi_2(x) = \frac{5sin(4\pi x)}{1.1 - sin(4\pi x)} + 8(x - 0.5)^2 - 2 \quad \text{and} \quad \psi_3(x) = \frac{1}{2}x.$$

The censoring time C is simulated from an exponential distribution with mean $U * exp(\psi(x))$, where U is randomly generated from uniform distribution on $[1, c]$ and the constant c is chosen such that the total censoring rate is about 20% - 30%. The

Gaussian kernel is used for all simulations. We consider two correlation cases with $\rho = 0.25$ and $\rho = 0.5$. We fix $d = 100$ and run 100 simulations with sample size $n = 100$ for each example. Example 1-4 are designed to evaluate the performance of screening variables with nonlinear-impact with the proposed non-linearity measure. Example 5 and 6 are designed to evaluate the total performance of the proposed two-step screening procedure.

Example 1. Generate data from the following model:

$$\psi(x) = \psi_1(X_1)$$

with $\rho = 0.25$.

Example 2. The setting is the same as Example 1 but with $\rho = 0.5$.

Example 3. Generate data from the following model:

$$\psi(x) = \psi_2(X_1)$$

with $\rho = 0.25$.

Example 4. The setting is the same as Example 3 but with $\rho = 0.5$.

Example 5. Generate data from the following model:

$$\psi(x) = 2\psi_1(X_1) + 1.2\psi_2(X_2) + 3\psi_3(X_3)$$

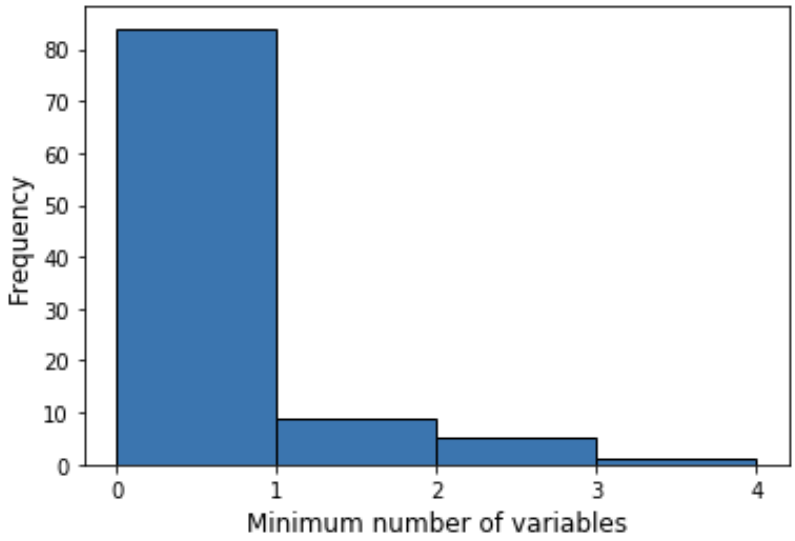
with $\rho = 0.25$.

Example 6. The setting is the same as Example 5 but with $\rho = 0.5$.

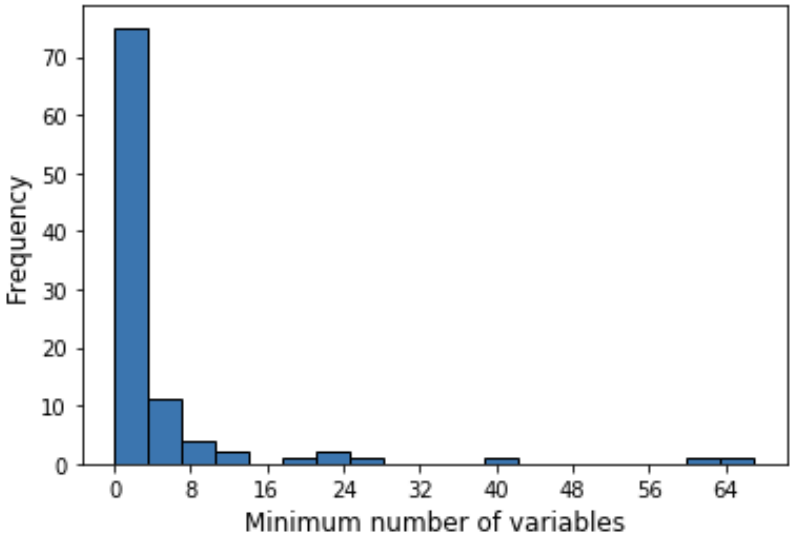
Table 4.1: Simulation results of Examples 1-4: Accuracy of proposed two-step screening in including the true model $\{X_1\}$.

Example	n	d	ρ	Probability
Ex 1	100	100	0.25	1.00
Ex 2	100	100	0.5	0.93
Ex 3	100	100	0.25	1.00
Ex 4	100	100	0.5	1.00

For Example 1-4, the only effect variable in the model has a nonlinear impact. Thus we only perform step 1 in the proposed two-step screening procedure. We compute the nonlinearity measure N_L for each covariate in each model. Then we rank the values of N_L from smallest to the largest and select the top $\lfloor n/log(n) \rfloor = 21$ covariate. We evaluate the performance of the method by computing the accuracy in including the true model $\{X_1\}$ and the results are summarized in Table 4.1. It shows that our method performs very well in capturing the important variables. More pertinently, the proposed non-linearity measure N_L identifies the nonlinear impact variable with high probability. Apart from the table, we also present the distribution of the smallest model size required to include the true model $\{X_1\}$ in Figure 4.1 and Figure 4.2. It is clear that reducing the dimensionality to $\lfloor n/log(n) \rfloor$ covariates can still retain the information in the model. For instance, for example 1, we can reduce the the model with 100 covariates to only 4 covariates to include the true model with probability 1. And for example 2, reducing the model size to $\lfloor n/log(n) \rfloor$ contains the true model with high probability.

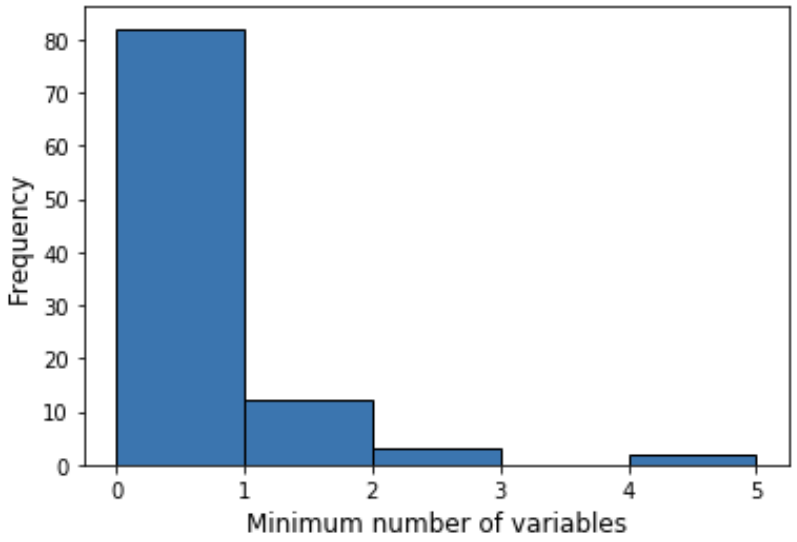


(a) Frequency vs. Smallest model size to cover the true model: Example 1

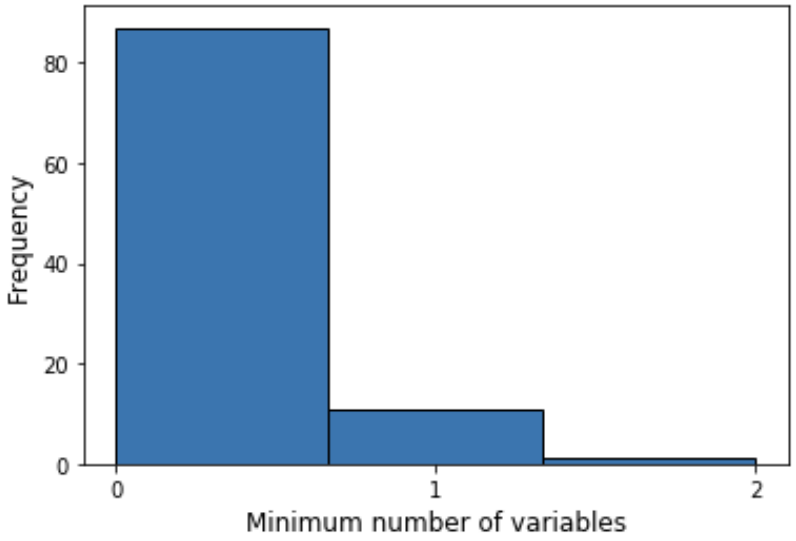


(b) Frequency vs. Smallest model size required to cover the true model: Example 2

Figure 4.1: Distribution of the smallest model size required to cover the true model $\{X_1\}$: Example 1 & 2.



(a) Frequency vs. Smallest model size to cover the true model: Example 3



(b) Frequency vs. Smallest model size to cover the true model: Example 4

Figure 4.2: Distribution of the Smallest model size required to cover the true model $\{X_1\}$: Example 3 & 4.

Next, we discuss the simulation results of example 5 and 6. The true model in these examples consists of two nonlinear impact covariates $\{X_1, X_2\}$ and one linear-impact covariate X_3. The performance of the screening procedure is evaluated by the following two terms:

1. P_I: the proportion that the individual important covariate is selected among the 100 simulations.

2. P_A: the proportion that all the important covariates are selected among the 100 simulations.

Table 4.2 indicates that the proposed screening procedure captures the important covariates with high probability. The correlation among the correlation impact the accuracy of screening. Specially, we observe a decrease in the screening accuracy as the correlation increases.

Table 4.2: Simulation results of Example 5 and 6.

Example	n	d	ρ	P_I			P_A
				X_1	X_2	X_3	all
Ex 5	100	100	0.25	0.94	0.96	0.90	0.88
Ex 6	100	100	0.5	0.89	0.90	0.88	0.85

CHAPTER 5: REAL EXAMPLE

In this section, we use a real data to demonstrate the performance of the proposed method. We adopt the Neuroblastoma data set in Oberthuer et al (2006)[36]. Neuroblastoma counts for up to 6% of all children cancer in the United States and it is the third most common type of cancer in children. Each year, there are about 800 new diagnostics of neuroblastoma in the United States. The average age of children when they are diagnosed is between 1 and 2 years. The data set was obtained from the MicroArray Quality Control phase-II (MAQC-II) project.

This data set contains 130 patients and the gene expression at 10,167 probe sites. The minimum age of all the patients is 3 days and the maximum age is 8983 days. The median of the age among all patients is 487 days. The clinical information including event free survival and overall survival is also available. In our study, we focus on the overall survival. Among the 130 patients, 87 of them are censored, which makes the censoring rate as high as about 67%.

The scale each predictor to make the value be between 0 and 1. Then we apply the proposed two-step screening method to the scaled data. To identify the covariate with nonlinear impact, we perform step 1 and keep the $\lfloor n/log(n) \rfloor = 26$ top variables. The selected genes with probe site names are summarized in Table 5.1. Then we perform the second step to screen the covariates with linear impact and the selected genes with probe site names are displayed in Table 5.2.

Table 5.1: Probe site names of genes selected by step 1 screening.

AF275813	RNASEP1	NDUFA1	AL133022
FLJ20516	I_960852	AB075859	LOC93081
MRPL3	LOC134492	SPINK2	WDR12
I_959809	AF060511	Hs94090.1	EEF1E1
Hs406351	PRDX4	MGC5528	MMP7
ARL2	GMNN	UBL1	QRSL1
GAJ	E2F3		

Table 5.2: Probe site names of genes selected by step 2 screening.

NNG1_exon5	NUDT5	SLC25A5	AHCY
STK6	MCTS1	C14orf166	NM_017669
AF117235	AL133641	ENST00000317847	SSRP1
Nup37	HSPC163	NOLA1	PAICS
BC006406	SNRPE	SNRPG	Hs155462.1
Hs108854.8	AHCY.1	SSR4	PX19
PPAT.1	AK057899		

We apply LASSO method for Cox's proportional model (Tibshirani, 1997) with the covariates selected in the two-step screening. Table 5.3 reports the 8 genes with probe site names that are kept in the final model.

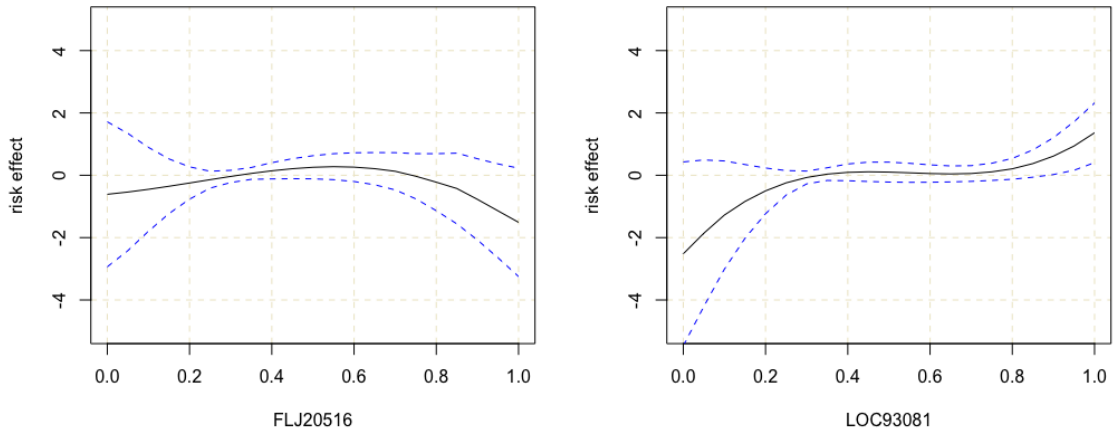
Table 5.3: Probe site names of genes kept in the final model.

SLC25A5	NUDT5	NM_017669	AF117235
FLJ20516	LOC93081	MRPL3	Hs94090.1

To obtain the confidence intervals of parameters, we take 1000 bootstrap samples from the empirical cumulative distribution of the sample (Burr, 1994)[37] The consistency of this approach was proved by Jiang (2011)[38]. That is, each time we randomly select n indexes with replacement from $\{1, \ldots, n\}$, denoted by B^*, and form the corresponding bootstrap sample $\{T_l, \delta_l, X_{lk}\}_{l \in B^*}$ based on the index. For the covariates with linear impacts, we calculate the standard deviations of the estimators of the coefficients and construct the confidence intervals as shown in Table 5.4. For the covariates with nonlinear impacts, we construct the 95% CIs for the estimation of the risk function $\psi(\cdot)$. The estimated risk effect along with the 95% point-wise confidence interval for each selected gene is plotted in Figure 5.1. We can tell that the hazard does not present a linear combination of the covariates. These plots indicate that the nonparametric estimate and the standard error estimates achieve satisfactory performance.

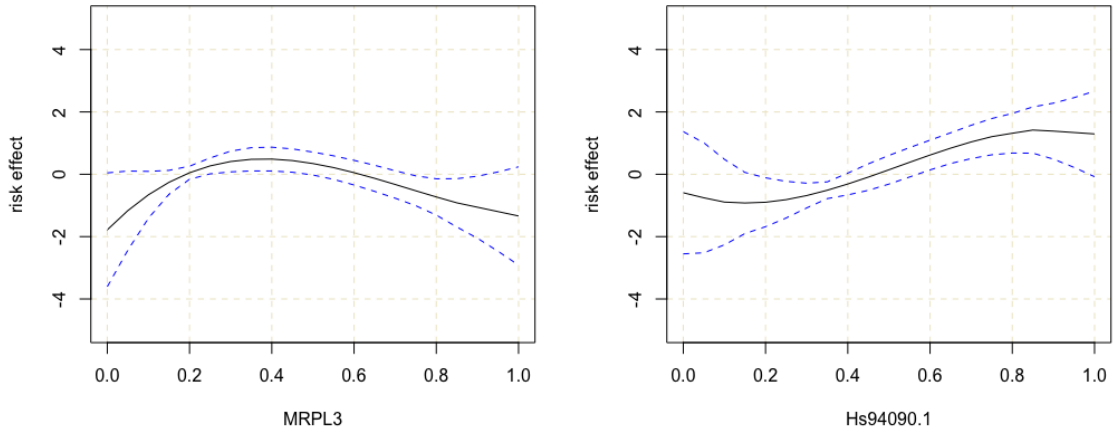
Table 5.4: Estimated parameters of selected covariates with linear impact. LCI and UCI are the lower and upper bounds of the 95% confidence interval, respectively.

Probe ID	Estimated coefficient	Standard error	LCI	UCI
SLC25A5	4.1958	0.3782	3.4546	4.9371
NUDT5	2.4012	0.3452	1.7246	3.0778
NM_017669	2.9499	0.2764	2.4082	3.4916
AF117235	-0.1726	0.2926	-0.7461	0.4009



(a) Gene site name: FLJ20516

(b) Gene site name: LOC93081



(c) Gene site name: MRPL3

(d) Gene site name: Hs94090.1

Figure 5.1: Estimated risk effect of selected nonlinear impact covariates. Black solid lines are the estimates and blue dashed lines are 95% confidence intervals.

Next, we try to provide more insight into the importance of the eight selected genes. The log partial likelihood of Cox's proportional model fitted with the eight genes is -130.4526. We remove one gene at each time, refit the Cox's proportional model, and compute the log partial likelihood of the refit model. The results together with the corresponding likelihood ratio test are summarized in Table 5.5. All the likelihood

ratio test values are significantly greater than the critical value 3.841, the χ^2 square value with degree of freedom 1 and 5% significance level and we conclude that the selected 8 genes are very important.

Table 5.5: Partial likelihood and likelihood ratio test of removing selected gene.

Probe ID of removed gene	Log partial likelihood	Likelihood ratio test
SLC25A5	-135.6423	10.379298
NUDT5	-137.2219	13.538451
NM_017669	-136.0623	11.219315
AF117235	-133.9759	7.046618
FLJ20516	-135.9180	10.930835
LOC93081	-138.4352	15.965157
MRPL3	-134.7404	8.575468
Hs94090.1	-134.7404	7.046010

CHAPTER 6: DISCUSSIONS

In this research, we propose a non-linearity measure to quantify the nonlinear impact of the covariates for Cox's proportional model with nonparametric additive risk effect. Then we introduce a two-step screening procedure to quickly reduce the dimensionality for ultra high dimensional data. We further establish the theoretical property that the nonlinear step screening possesses the sure independent screening property. Moreover, we derive the influence function for nonparametric Cox's proportional model and it can speed up the screening process dramatically. Simulation studies are carried out to assess the performance of the proposed screening procedure. We also use the Neuroblastoma data to shed more light on the application of the proposed screening method on real data.

Our future research work includes but not limited to the following two aspects. First, we can develop an IC for the 2nd step screening and compare it with the current one. Further, we can establish the theoretical result of the whole two-step screening. Second, it is interesting to investigate the post-selection statistical inference of the selected model. Some key works about this topic are Fithian et al. (2015)[39] and Tibshirani et al. (2016)[40].

Bibliography

[1] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[2] T. Fleming and D. Harrington, "Counting processes and survival analysis john wiley & sons," *Inc. New York*, 1991.

[3] P. Anderson, Ø. Borgan, R. Gill, and N. Keiding, "Statistical models based on counting processes," *Biometrics*, vol. 24, pp. 100–101, 1993.

[4] D. Faraggi and R. Simon, "Large sample bayesian inference on the parameters of the proportional hazard models," *Statistics in medicine*, vol. 16, no. 22, pp. 2573–2585, 1997.

[5] K. H. Lee, S. Chakraborty, J. Sun, *et al.*, "Bayesian variable selection in semiparametric proportional hazards model for high dimensional survival data," *The International Journal of Biostatistics*, vol. 7, no. 1, pp. 1–32, 2011.

[6] H. Akaike, "A new look at the statistical model identification," *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 716–723, 1974.

[7] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[8] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[9] R. Tibshirani, "The lasso method for variable selection in the cox model," *Statistics in medicine*, vol. 16, no. 4, pp. 385–395, 1997.

[10] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[11] J. Fan and R. Li, "Variable selection for cox's proportional hazards model and frailty model," *Annals of Statistics*, pp. 74–99, 2002.

[12] J. Cai, J. Fan, R. Li, and H. Zhou, "Variable selection for multivariate failure time data," *Biometrika*, vol. 92, no. 2, pp. 303–316, 2005.

[13] H. H. Zhang and W. Lu, "Adaptive lasso for cox's proportional hazards model," *Biometrika*, vol. 94, no. 3, pp. 691–703, 2007.

[14] J. Fan, I. Gijbels, M. King, *et al.*, "Local likelihood and local partial likelihood in hazard regression," *The Annals of Statistics*, vol. 25, no. 4, pp. 1661–1690, 1997.

[15] T. Hastie and R. Tibshirani, "Exploring the nature of covariate effects in the proportional hazards model," *Biometrics*, pp. 1005–1016, 1990.

[16] R. J. Gray, "Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis," *Journal of the American Statistical Association*, vol. 87, no. 420, pp. 942–951, 1992.

[17] R. J. Gray, "Spline-based tests in survival analysis," *Biometrics*, pp. 640–652, 1994.

[18] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 5, pp. 849–911, 2008.

[19] J. Bradic, J. Fan, and J. Jiang, "Regularization for cox's proportional hazards model with np-dimensionality," *Annals of statistics*, vol. 39, no. 6, p. 3092, 2011.

[20] J. Huang, T. Sun, Z. Ying, Y. Yu, and C.-H. Zhang, "Oracle inequalities for the lasso in the cox model," *Annals of statistics*, vol. 41, no. 3, p. 1142, 2013.

[21] J. Fan, R. Song, *et al.*, "Sure independence screening in generalized linear models with np-dimensionality," *The Annals of Statistics*, vol. 38, no. 6, pp. 3567–3604, 2010.

[22] J. Fan, Y. Feng, and R. Song, "Nonparametric independence screening in sparse ultra-high-dimensional additive models," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 544–557, 2011.

[23] J. Liu, R. Li, and R. Wu, "Feature selection for varying coefficient models with ultrahigh-dimensional covariates," *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 266–274, 2014.

[24] X. He, L. Wang, H. G. Hong, *et al.*, "Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data," *The Annals of Statistics*, vol. 41, no. 1, pp. 342–369, 2013.

[25] Y. Wu and G. Yin, "Conditional quantile screening in ultrahigh-dimensional heterogeneous data," *Biometrika*, vol. 102, no. 1, pp. 65–76, 2015.

[26] S. D. Zhao and Y. Li, "Principled sure independence screening for cox models with ultra-high-dimensional covariates," *Journal of multivariate analysis*, vol. 105, no. 1, pp. 397–411, 2012.

[27] J. Habbema, H. JDF, K. Van den Broek, *et al.*, "A stepwise discriminant analysis program using density estimation.," 1974.

[28] D. R. Cox, "Partial likelihood," *Biometrika*, vol. 62, no. 2, pp. 269–276, 1975.

[29] N. Breslow, "Covariance analysis of censored survival data," *Biometrics*, pp. 89–99, 1974.

[30] R. Tibshirani and T. Hastie, "Local likelihood estimation," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 559–567, 1987.

[31] J. A. Nelder and R. W. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society: Series A (General)*, vol. 135, no. 3, pp. 370–384, 1972.

[32] C. Loader, *Local regression and likelihood*. Springer Science & Business Media, 2006.

[33] R. D. Cook, "Detection of influential observation in linear regression," *Technometrics*, vol. 19, no. 1, pp. 15–18, 1977.

[34] Y. Feng, Y. Wu, and L. A. Stefanski, "Nonparametric independence screening via favored smoothing bandwidth," *Journal of Statistical Planning and Inference*, vol. 197, pp. 1–14, 2018.

[35] N. Reid and H. Crépeau, "Influence functions for proportional hazards regression," *Biometrika*, vol. 72, no. 1, pp. 1–9, 1985.

[36] A. Oberthuer, F. Berthold, P. Warnat, B. Hero, Y. Kahlert, R. Spitz, K. Ernestus, R. Konig, S. Haas, R. Eils, *et al.*, "Customized oligonucleotide microarray gene expression–based classification of neuroblastoma patients outperforms current clinical risk stratification," *Journal of clinical oncology*, vol. 24, no. 31, pp. 5070–5078, 2006.

[37] D. Burr, "A comparison of certain bootstrap confidence intervals in the cox model," *Journal of the American Statistical Association*, vol. 89, no. 428, pp. 1290–1302, 1994.

[38] J. Jiang and X. Jiang, "Inference for partly linear additive cox models," *Statistica Sinica*, pp. 901–921, 2011.

[39] W. Fithian, D. Sun, and J. Taylor, "Optimal inference after model selection," *arXiv preprint arXiv:1410.2597*, 2014.

[40] R. J. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani, "Exact post-selection inference for sequential regression procedures," *Journal of the American Statistical Association*, vol. 111, no. 514, pp. 600–620, 2016.

[41] A. W. Van Der Vaart and J. A. Wellner, "Weak convergence," in *Weak convergence and empirical processes*, pp. 16–28, Springer, 1996.

[42] S. van de Geer, "Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes," *The Annals of Statistics*, pp. 1779–1801, 1995.

[43] W. S. Krasker and R. E. Welsch, "Efficient bounded-influence regression estimation," *Journal of the American statistical Association*, vol. 77, no. 379, pp. 595–604, 1982.

[44] F. R. Hampel, "The influence curve and its role in robust estimation," *Journal of the american statistical association*, vol. 69, no. 346, pp. 383–393, 1974.

APPENDIX A: CONDITIONS

The following regularity conditions are needed for our asymptotic results. Note that conditions 1 to 5 are similar to those in Fan and Gijbels (1997)[14]. Conditions 6 and 7 are imposed to facilitate the development of the uniform deviation results of ξ_k, which is summarized in Theorem 2. Specifically, condition 7 is the consequence of martingale representation of the score function for the Cox model. Condition 8 is needed to establish the uniform deviation result of the penalty term in IC. Condition 9 assures that the signals of the covariates with nonlinear impact and those of the covariates with linear impact are separable.

Condition 1. The kernel function $K(\cdot) \geq 0$ is a bounded density function with compact support. Let $\mu_1 = \int uK(u)du$ and $v_1 = \int u^2 k(u)du - {\mu_1}^2$.

Condition 2. The function $\psi(\cdot)$ has a continuous second-order derivative around the point x.

Condition 3. The density function $f(\cdot)$ of X is continuous as at the point x and $f(x) > 0$.

Condition 4. The conditional probability $P(u|\cdot)$ is equi-continuous at point x.

Condition 5. Bandwidth h satisfies $nh \to \infty$ and nh^5 is bounded.

Condition 6. There exists a compact neighborhood $\mathcal{B}$ of 0 that satisfies each of the conditions.

1) On $\mathcal{B} \times [0, \tau] \times [0, 1]$, there exists scalar, vector and matrix function $s_k^{(l)}$ such that,

$$\sup_{0 \leq u \leq \tau} \sup_{\alpha \in \mathcal{B}_1} \left\| S_{nk}^{(l)}(\boldsymbol{\alpha}_k, u) - s_k^{(l)}(\boldsymbol{\alpha}_k, u) \right\| \to 0$$

as $n \to \infty$ for $\boldsymbol{\alpha}_k \in \mathcal{B}_1$, $\mathcal{B}_1 \in \mathcal{B}$, $l = 0, 1, 2$.

2) Suppose that $C_{nk} = \sup_{0 \leq u \leq \tau} |E_{nk}(0, u) - \mu_1|$, then there exits constant c such that $\sup_k C_{nk} < c$ almost surely.

Condition 7. Define $\epsilon_{ik} = \int_0^\tau K_h(X_{ik} - x)(U_{ik} - \mu_1)dM_{ik}(u)$. Suppose ϵ_{ik} satisfies the Cramer condition, that is,

$$E|\epsilon_{ik}|^m \leq m!M^{m-2}\sigma_k^2/2$$

for all k, where $m \geq 2$, M is a positive constant and $\sigma_k^2 = var(\epsilon_{ik}) < \infty$.

Condition 8. Let $P_{ik}^* = \delta_i \widetilde{\Sigma}_k(\tau, X_{ik})^{-1} G_{ik}/D_{ik} - \Omega_k$. Assume $E|P_{ik}^*|^m < m!M^{m-2}v_i/2$ for every $m \geq 2$ and all i and some constants M and v_i.

Condition 9. There exists sequences $\widetilde{C}_n$ and $\widetilde{D}_n$ such that $\widetilde{C}_n \gg (\frac{logp}{n})^{\frac{1}{5}} \geq \widetilde{D}_n$. For all $k \in S$, $E\{\delta[\psi_k(x_k) - log(s_{0k}(z))]\} - E\{\delta[x_k\widetilde{\beta}_k - log(s_{0k}^*(z))]\} > \widetilde{C}_n$ and for all $k \notin S$, $E\{\delta[\psi_k(x_k) - log(s_{0k}(z))]\} - E\{\delta[x_k\widetilde{\beta}_k - log(s_{0k}^*(z))]\} < \widetilde{D}_n$.

APPENDIX B: PROOFS OF THEOREMS

To facilitate our arguments for proofs, we first introduce some lemmas and propositions.

Lemma 1. (Besrstein's inequality, Lemma 2.2.11, Van Der Vaar and Wellner (1996)[41]). Let $Y_1, \ldots, Y_n$ be independent random variables with zero mean such that $E|Y_i|^m \leq m! M^{m-2} v_i/2$, for every $m \geq 2$ (and all i) and some constants M and v_i. Then

$$P(|Y_1 + \ldots Y_n| > x) \leq 2exp\{-x^2/(2(v + Mx))\},$$

for $v \geq v_1 + \ldots + v_n$.

Lemma 2. (Lemma 1, Fan and Gijbels (1997)[14]). Suppose that function K is bounded and compactly supported. If $g(\cdot)$ is continuous at the point x and $P(t|\cdot)$ is equi-contious at the point x and $h \to 0$ in such a way that $nh/log(n) \to \infty$, then

$$\sup_{0 \leq t \leq \tau} |c_n(t) - c(t)| \xrightarrow{\mathrm{P}} 0,$$

where $c_n(t) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t) g(X_i) K_h(X_i - x)$, and $c(t) = f(x)g(x)P(t|x) \int K(u)du$ with $Y_i(t) = \mathcal{I}(Z_i \geq t)$.

Lemma 3. (Martingale Inequality, Van de Geer (1995)[42]). Let $\{M_t\}_{t \geq 0}$ be a locally square integrable martingale with respect to the filtration $\{\mathcal{F}_t\}_{t \geq 0}$. Denote the jump of $\{M_t\}$ by $\Delta M_t = M_t - M_{t^-}$ and the predictable variation by $V_t = < M_t >$. Suppose that $|\Delta M_t| \leq m$ for all $t > 0$ and some $0 \leq m < \infty$. Then for each $a > 0$, $b > 0$,

$$P(M_t \geq a \text{ and } V_t^2 \leq b^2 \text{ for some t}) \leq exp\Big\{\frac{-a^2}{2(am + b^2)}\Big\}.$$

Next we study the uniform property of $\mathcal{L}(\hat{\psi}_k)$ defined in (2.12), where $\mathcal{L}(\cdot)$ denotes the global partial likelihood.

Proposition 1. Assume Conditions 1-7 holds. We define $\hat{\psi}_k(x)$ as the estimator of local partial likelihood of Cox's proportional model estimated at point x. And for $k = 1, \cdots, d$, we define

$$\mathcal{L}(\hat{\psi}_k) = \frac{1}{n}\sum_{i=1}^{n} \delta_i[\hat{\psi}_k(X_{ik}) - log\{\sum_{j=1}^{n} Y_j(Z_i)exp(\hat{\psi}_k(X_{jk}))\}],$$

$$\mathcal{L}(\psi_k) = \frac{1}{n}\sum_{i=1}^{n} \delta_i[\psi_k(X_{ik}) - log\{\sum_{j=1}^{n} Y_j(Z_i)exp(\psi_k(X_{jk}))\}],$$

where $\boldsymbol{\psi}_k = (\psi_k(X_{1k}), \cdots, \psi_k(X_{nk}))^T$ is the true risk function. There exists a set $\mathcal{A}_1$ with $P(\mathcal{A}_1) \to 1$ and a constant $A_1 > 0$ such that on set $\mathcal{A}_1$, for $k = 1, \ldots, d$,

$$|\mathcal{L}(\hat{\psi}_k) - \mathcal{L}(\psi_k)| \leq A_1 a_n^*,$$

where $a_n^* = \sqrt{\frac{logp}{nh^3}} + h$.

Proof. It follows from Theorem 3 that there exists a set $\mathcal{A}_1$ with $P(\mathcal{A}_1) \to 1$ such that on set $\mathcal{A}_1$, for $k = 1, \ldots, d$ and any $i = 1, \ldots, n$,

$$|\hat{\psi}_k(X_{ik}) - \psi_k(X_{ik})| \leq a_n^*.$$

Define

$$\hat{L}_{ik} = \delta_i\big[\hat{\psi}_k(X_{ik}) - log\{\sum_{j=1}^{n} Y_j(Z_i)exp(\hat{\psi}_k(X_{jk}))\}\big]$$

and

$$L_{ik} = \delta_i\big[\psi_k(X_{ik}) - log\{\sum_{j=1}^{n} Y_j(Z_i)exp(\psi_k(X_{jk}))\}\big].$$

Then

$$|\mathcal{L}(\hat{\psi}_k) - \mathcal{L}(\psi_k)| \leq \frac{1}{n}\sum_{i=1}^{n} \big|\hat{L}_{ik} - L_{ik}\big|.$$

We observe that

$$\begin{aligned}
\big|\hat{L}_{ik} - L_{ik}\big| &\leq \big|\delta_i[\hat{\psi}_k(X_{ik}) - \psi_k(X_{ik}) - log\{\sum_{j=1}^{n} Y_j(Z_i)exp(\hat{\psi}_k(X_{jk}))\} \\
&\quad + log\{\sum_{j=1}^{n} Y_j(Z_i)exp(\psi_k(X_{jk}))\}]\big| \\
&\leq \big|[\hat{\psi}_k(X_{ik}) - \psi_k(X_{ik})]\big| + \big|log\{\frac{\sum_{j=1}^{n} Y_j(Z_i)exp(\psi_k(X_{jk}))}{\sum_{j=1}^{n} Y_j(Z_i)exp(\hat{\psi}_k(X_{jk}))}\}\big|.
\end{aligned}$$

Since

$$\begin{aligned}
log\big\{\frac{\sum_{j=1}^{n} Y_j(Z_i)exp(\psi_k(X_{jk}))}{\sum_{j=1}^{n} Y_j(Z_i)exp(\hat{\psi}_k(X_{jk}))}\big\} &= log\big\{\sum_{j=1}^{n} Y_j(Z_i)exp[\psi_k(X_{jk}) - \hat{\psi}_k(X_{jk})]exp(\hat{\psi}_k(X_{jk}))\big\} \\
&\quad - log\big\{\sum_{j=1}^{n} Y_j(Z_i)exp(\hat{\psi}_k(X_{jk}))\big\} \\
&\leq log\big\{exp(a_n^*)\frac{\sum_{j=1}^{n} Y_j(Z_i)exp(\hat{\psi}_k(X_{jk}))}{\sum_{j=1}^{n} Y_j(Z_i)exp(\hat{\psi}_k(X_{jk}))}\big\} \\
&= a_n^*.
\end{aligned}$$

Then, we get

$$\begin{aligned}
\big|\hat{L}_{ik} - L_{ik}\big| &\leq \big|[\hat{\psi}_k(X_{ik}) - \psi_k(X_{ik})]\big| + a_n^* \\
&\leq 2a_n^*.
\end{aligned}$$

As a result, on set $\mathcal{A}_1$, there exist a constant $A_1 > 0$ such that $|\hat{\psi}_k(X_{ik}) - \psi_k(X_{ik})| \leq A_1 a_n^*$ for $k = 1, \ldots, d$. □

Proposition 2. We now define

$$S_{0k}(Z_i) = \frac{1}{n}\sum_{j=1}^{n} Y_j(Z_i)exp(\psi_k(X_{jk})) \quad \text{and} \quad s_{0k}(z) = E\{Y(z)exp(\psi_k(x_k))\}.$$

Then

$$\mathcal{L}(\boldsymbol{\psi}_k) = -log(n)E\delta(1+o_p(1)) + E\{\delta[\psi_k(x_k) - log(s_{0k}(z))]\},$$

where $\mathcal{L}(\cdot)$ denotes the global partial likelihood.

Proof. Based on the definition of $\mathcal{L}(\cdot)$ in 2.12, we have

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\psi}_k) &= \frac{1}{n}\sum_{i=1}^{n}\delta_i\big[\psi_k(X_{ik}) - log\{\sum_{j=1}^{n}Y_j(Z_i)exp(\psi_k(X_{jk}))\}\big] \\
&= \frac{1}{n}\sum_{i=1}^{n}\delta_i\big[\psi_k(X_{ik}) - log\{\frac{1}{n}\sum_{j=1}^{n}Y_j(Z_i)exp(\psi_k(X_{jk}))\} - log(n)\big] \\
&= -log(n)E\delta(1+o_p(1)) + \frac{1}{n}\sum_{i=1}^{n}\delta_i\big[\psi_k(X_{ik}) - log(s_{0k}(z)) + log\frac{s_{0k}(z)}{S_{0k}(Z_i)}\big].
\end{aligned}$$

Since $s_{0k}(z)/S_{0k}(Z_i) \xrightarrow{\text{P}} 1$, then we prove that

$$\mathcal{L}(\boldsymbol{\psi}_k) = -log(n)E\delta(1+o_p(1)) + E\{\delta[\psi_k(x_k) - log(s_{0k}(z))]\}.$$

□

Next we study the property of the global partial likelihood estimated at $h = \infty$. Recall that maximizing the local partial likelihood when $p = 1$ is equivalent to maximize the global partial likelihood of Cox model with parametric risk effect, thus the estimate of coefficient is a constant. We denote the estimated coefficient by $\widetilde{\beta}_k$ and the corresponding estimated risk function by $\widetilde{\psi}_k$, then we have $\widetilde{\psi}_k(x) = \widetilde{\beta}_k x$.

Proposition 3. Let $\widetilde{\boldsymbol{\psi}}_k$ denote the estimated risk function for the case $h = \infty$, that is $\widetilde{\boldsymbol{\psi}}_k = (\widetilde{\psi}_k(X_{1k}), \ldots, \widetilde{\psi}_k(X_{nk}))^T$ and $\widetilde{\psi}_k(X_{ik}) = X_{ik}\widetilde{\beta}_k$ for $i = 1, \ldots, n$. We further define

$$S^*_{0k}(Z_i) = \frac{1}{n}\sum_{j=1}^{n}Y_j(Z_i)exp(X_{jk}\widetilde{\beta}_k) \quad \text{and} \quad s^*_{0k}(z) = E[Y(z)exp(x_k\widetilde{\beta}_k)].$$

Then we get

$$\mathcal{L}(\widetilde{\boldsymbol{\psi}}_k) = -log(n)E\delta(1 + o_p(1)) + E\{\delta[x_k\widetilde{\beta}_k - log(s^*_{0k}(z))]\}$$

Proof. Proposition 3 can be proved following the similar arguments as in Proposition 2. □

Remark. Based on the results of Proposition 2 and Proposition 3, we can reach the following conclusions:

$$\mathcal{L}(\boldsymbol{\psi}_k) - \mathcal{L}(\widetilde{\boldsymbol{\psi}}_k) = E\{\delta[\psi_k(x_k) - log(s_{0k}(z))]\} - E\{\delta[x_k\widetilde{\beta}_k - log(s^*_{0k}(z))]\}.$$

1. If the true risk function $\psi_k(\cdot)$ is linear, then $\mathcal{L}(\boldsymbol{\psi}_k) - \mathcal{L}(\widetilde{\boldsymbol{\psi}}_k) = o_p(1)$.

2. If the true risk function $\psi_k(\cdot)$ is not linear, then $\mathcal{L}(\boldsymbol{\psi}_k) - \mathcal{L}(\widetilde{\boldsymbol{\psi}}_k) \neq 0$.

Theorem 1

Proof. **of Theorem 1.** Recall that $\hat{\boldsymbol{\alpha}}_k$ is the maximizer of

$$\mathcal{L}_x(\boldsymbol{\alpha}_k, \tau) = \frac{1}{n}\int_o^\tau \sum_{i=1}^n K_h(X_{ik} - x)\Big[\widetilde{X}^*_{ik}\boldsymbol{\beta}^0_k + \mathbf{U}^T_{ik}\boldsymbol{\alpha}_k - log\big\{\sum_{j=1}^n Y_j(Z_i)exp(\widetilde{\mathbf{X}}^{*T}_{ik} + \mathbf{U}^T_{ik}\boldsymbol{\alpha}_k)K_h(X_{jk} - x)\big\}\Big]dN_i(u).$$

with respect to $\boldsymbol{\alpha}_k$ with $\tau = \infty$, where τ denotes the observation ending time. We consider the local linear estimator, that is, $p = 1$, then $\hat{\alpha}_k$ is a scalar in this case.

Fan (1997) provides the consistency for $\hat{\alpha}_k$ as

$$\hat{\alpha}_k \xrightarrow{\text{P}} 0.$$

By Taylor Expansion round 0,

$$0 = \mathcal{L}'_x(\hat{\alpha}_k, \tau) = \mathcal{L}'_x(0, \tau) + \mathcal{L}''_x(\hat{\alpha}^*_k, \tau)(\hat{\alpha}_k - 0),$$

where $\hat{\alpha}^*_k$ is between 0 and $\hat{\alpha}_k$. Thus,

$$\hat{\alpha}_k = -\mathcal{L}''_x(\hat{\alpha}^*_k, \tau)^{-1}\mathcal{L}'_x(0, \tau). \tag{B.1}$$

It is easy to see that $\hat{\alpha}^*_k \xrightarrow{\text{P}} 0$. Under the continuity assumption, we have $\mathcal{L}''_x(\hat{\alpha}^*_k, \tau) = \mathcal{L}''_x(0, \tau) + o_p(1)$. Recall that,

$$\mathcal{L}''_x(\alpha_k, \tau) = -\int_0^\tau \frac{S^{(2)}_{nk}(\alpha_k, \tau)S^{(0)}_{nk}(\alpha_k, \tau) - S^{(1)}_{nk}(\alpha_k, \tau)^2}{S^{(0)}_{nk}(\alpha_k, \tau)^2}\frac{1}{n}\sum_{i=1}^n K_h(X_{ik} - x)dN_i(u).$$

By Lemma 2, we have

$$\sup_{0\le u\le\tau}|\frac{S_{nk}^{(2)}(\alpha_k,\tau)S_{nk}^{(0)}(\alpha_k,\tau)-S_{nk}^{(1)}(\alpha_k,\tau)^2}{S_{nk}^{(0)}(\alpha_k,\tau)^2}-v_1|=o_p(1),$$

where $v_1=\int u^2k(u)du-{\mu_1}^2$. Then according to the result in Fan and Gijbels (1997)[14], we have

$$\mathcal{L}_x''(0,\tau)=-v_1f_k(x)\Psi_k(x)\Lambda_k(\tau,x)+o_p(1)\equiv-\widetilde{\Sigma}_k(\tau,x)+o_p(1). \tag{B.2}$$

Now we work on $\mathcal{L}_x'(0,\tau)$. We see that

$$\mathcal{L}_x'(0,\tau)=\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau K_h(X_{ik}-x)\{U_{ik}-E_{nk}(0,u)\}dN_i(u). \tag{B.3}$$

Since $M_{ik}(u)=N_i(u)-\int_0^\tau Y_i(u)exp\{\psi_k(X_{ik})\}\lambda_0(u)du$ is a martingale with respect to $\mathcal{F}_{ik}$. Then we can rewrite (B.3) as

$$\begin{aligned}\mathcal{L}_x'(0,\tau)=&\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau K_h(X_{ik}-x)\{U_{ik}-E_{nk}(0,u)\}dM_{ik}(u)\\&+\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau K_h(X_{ik}-x)\{U_{ik}-E_{nk}(0,u)\}Y_i(u)exp\{\psi_k(X_{ik})\}\lambda_0(u)du.\end{aligned}$$

Denote the first term by $B_{1k}(\tau,x)$ and the second term by $B_{2k}(\tau,x)$. By Lemma 2,

$$\sup_{0\le u\le\tau}|E_{nk}(0,u)-\mu_1|\xrightarrow{\mathrm{P}}0,$$

where $\mu_1=\int uK(u)du$. Based on the results in Fan and Gijbels (1997)[14], we get

$$B_{2k}(\tau,x)=f_k(x)\Psi_k(x)\frac{\psi_k^{(2)}(x)}{2}\Lambda_k(\tau,x)\int K^2(u)(u-\mu_1)u^2duh^2+o_p(h^2)=O_p(h^2). \tag{B.4}$$

Then (B.3) can be written as

$$\mathcal{L}'_x(0,\tau) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau K_h(X_{ik}-x)\{U_{ik}-E_{nk}(0,u)\}dM_{ik}(u) + O_p(h^2). \tag{B.5}$$

Combine (B.1), (B.2) and (B.5), we get

$$\hat{\alpha}_k = \widetilde{\Sigma}_k(\tau,x)^{-1}\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau K_h(X_{ik}-x)\{U_{ik}-E_{nk}(0,u)\}dM_{ik}(u) + O_p(h^2).$$

Recall $\xi_k(x) \equiv \sum_{i=1}^n \int_o^\tau K_h(X_{ik}-x)\{U_{ik}-E_{nk}(0,u)\}dM_{ik}(u)$. As a result,

$$\hat{\alpha}_k = \frac{1}{n}\widetilde{\Sigma}_k(\tau,x)^{-1}\xi_k(x) + B_{2k}(\tau,x) = \frac{1}{n}\widetilde{\Sigma}_k(\tau,x)^{-1}\xi_k(x) + O_p(h^2). \tag{B.6}$$

□

Theorem 2

Proof. **of Theorem 2.** Recall $\xi_k(x) = \sum_{i=1}^n \int_o^\tau K_h(X_{ik}-x)\{U_{ik}-E_{nk}(0,u)\}dM_{ik}(u)$. Then $\xi_k(x)$ can be written as,

$$\begin{aligned}\xi_k(x) &= \sum_{i=1}^n \int_0^\tau K_h(X_{ik}-x)\{U_{ik}-\mu_1\}dM_{ik}(u) \\ &\quad - \sum_{i=1}^n \int_0^\tau K_h(X_{ik}-x)\{E_{nk}(0,u)-\mu_1\}dM_{ik}(u) \\ &\equiv \xi_{k1}(\tau,x) - \xi_{k2}(\tau,x).\end{aligned}$$

To establish the exponential inequality for $\xi_k(x)$, we work on establishing the exponential inequalities for $\xi_{k1}(\tau,x)$ and $\xi_{k2}(\tau,x)$ in the following.

Note that $\xi_{k1}(\tau,x) = \sum_{i=1}^n \epsilon_{ik}$, where $\{\epsilon_{ik}\}_{i=1}^n$ is a sequence of i.i.d. random variables with mean 0. Then by Condition 7, we have the Berstein's exponential inequality,

$$P(|\xi_{k1}(x)| > a) \leq 2exp\{-a^2/2(n\sigma_k^2 + Ma)\}. \tag{B.7}$$

Now consider the counting process $N_i(t)$. For covariate X_k, the intensity process is defined as $\lambda_{ik}(t) = Y_i(t)exp\{\psi_k(X_{ik})\}\lambda_0(t)$. The continuous compensator for martingale $M_i(t)$ is defined as $\Lambda_{ik}(t) = \int_o^t \lambda_{ik}(u)du$. We denote $\bar{\Lambda}_k(t) = \sum_{i=1}^n \Lambda_{ik}(t)$ and by the continuity of compensator, we have $|\Delta\bar{\Lambda}_k(t)| = 0$.

Let $\bar{N}(t) = \sum_{i=1}^n N_i(t)$, then $\Delta\bar{N}(t) = \sum_{i=1}^n \Delta N_i(t)$, where $\Delta N_i(t) = N_i(t) - N_i(t^-)$ indicates the number of events that occurs at time t. Since no two counting process jump at the same time, we have $|\Delta\bar{N}(t)| \leq 1$. Note that $\bar{M}(t) = \bar{N}(t) - \bar{\Lambda}(t)$, then, we get

$$\begin{aligned}|\Delta(h^{\frac{1}{2}}n^{-\frac{1}{2}}\xi_{k2}(u,x))| &\leq h^{\frac{1}{2}}n^{-\frac{1}{2}} \max_{i=1,\ldots,n} |K_h(X_{ik}-x)| \sup_{0\leq u\leq\tau} |E_{nk}(0,u)-\mu_1| \\ &\equiv h^{\frac{1}{2}}n^{-\frac{1}{2}} \max_{i=1,\ldots,n} |K_h(X_{ik}-x)|C_{nk}.\end{aligned}$$

By Condition 6 (2), this is bounded almost surely.

We denote the quadratic variation of martingale of $M_i(t)$ by $< M_i(t) >$. Then the predictable quadratic variation of $h^{\frac{1}{2}}n^{-\frac{1}{2}}\xi_{k2}(t,x)$ is

$$\begin{aligned} < h^{\frac{1}{2}}n^{-\frac{1}{2}}\xi_{k2}(u,x) > &= hn^{-1}\sum_{i=1}^{n}\int_0^u K_h^2(X_{ik}-x)\{E_{nk}(0,t)-\mu_1\}^2 d < M_i(t) > \\ &= hn^{-1}\sum_{i=1}^{n}\int_0^u K_h^2(X_{ik}-x)\{E_{nk}(0,t)-\mu_1\}^2 Y_i(t) \\ &\quad exp\{\psi_k(X_{ik})\}\lambda_0(t)dt \\ &\equiv b_{nk}^2(u). \end{aligned}$$

It is clear that

$$b_{nk}^2(u) \le b_{nk}^2(\tau) \le C_{nk}^2 hn^{-1}\sum_{i=1}^{n}\int_0^\tau K_h^2(X_{ik}-x)Y_i(u)exp\{\psi_k(X_{ik})\}\lambda_0(u)du.$$

By Lemma 2,

$$\begin{aligned} &hn^{-1}\sum_{i=1}^{n}\int_0^\tau K_h^2(X_{ik}-x)Y_i(u)exp\{\psi_k(X_{ik})\}\lambda_0(u)du \\ &\qquad = f_k(x)\Psi_k(x)\Lambda_k(\tau,x)\int K^2(u)du + o_p(1) \equiv \Sigma_k(\tau,x) + o_p(1), \end{aligned}$$

where $\Lambda_k(\tau,x) = \int_0^\tau P(Z \ge z|X = x_k)\lambda_0(u)du$.

Assume $f_k(x)$ and $\Psi_k(x)$ are bounded for any k. Therefore, there exits positive contants $0 \le b < \infty$ and $0 < d < \infty$ such that

$$|\Delta(h^{\frac{1}{2}}n^{-\frac{1}{2}}\xi_{k2}(u,x))| < b \quad \text{and} \quad < h^{\frac{1}{2}}n^{-\frac{1}{2}}\xi_{k2}(u,x) > \le d^2.$$

As a result, we apply the exponential inequality for martingales as in Lemma 3,

for $u_n > 0$,

$$P(|\xi_{k2}(\tau,x)| > n^{\frac{1}{2}}h^{-\frac{1}{2}}u_n) = P(h^{-\frac{1}{2}}n^{\frac{1}{2}}\xi_{k2}(\tau,x)) \leq 2exp\{-\frac{u_n^2}{2(bu_n+d^2)}\}.$$

Therefore, by Condition 7, then there exists constant $c > 0$ such that

$$P(|\xi_{k2}(\tau,x)| > n^{\frac{1}{2}}h^{-\frac{1}{2}}u_n) \leq 2exp\{-cu_n\} \tag{B.8}$$

uniformly over k. Note that

$$\begin{aligned} P(|\xi_k(\tau,x)| > n^{\frac{1}{2}}h^{-\frac{1}{2}}u_n) \leq\ & P(|\xi_{k1}(\tau,x)| > 0.5n^{\frac{1}{2}}h^{-\frac{1}{2}}u_n) \\ & + P(|\xi_{k2}(\tau,x)| > 0.5n^{\frac{1}{2}}h^{-\frac{1}{2}}u_n). \end{aligned}$$

Then by (B.7) and (B.8), we have

$$\begin{aligned} P(|\xi_k(\tau,x)| > n^{\frac{1}{2}}h^{-\frac{1}{2}}u_n) \leq\ & 2exp\{-\frac{u_n}{4h(2\sigma_k^2 {u_n}^{-1} + Mn^{-\frac{1}{2}}h^{-\frac{1}{2}})}\} \\ & + 2exp\{-cu_n\}. \end{aligned} \tag{B.9}$$

Then there exists positive constants c_0 and c_1 such that

$$P(|\xi_k(\tau,x)| > n^{\frac{1}{2}}h^{-\frac{1}{2}}u_n) \leq c_0 exp\{-c_1u_n\} \tag{B.10}$$

uniformly over k for given x.

Taking appropriate u_n, we get

$$\sup_{k=1,\ldots,d} |\xi_k(\tau,x)| = O_p(n^{\frac{1}{2}}h^{-\frac{1}{2}}u_n).$$

Let $u_n = c\sqrt{logp}$. Consider the ultrahigh-dimensional framework, that is, the dimen-

sionality grows exponentially with the sample size. Then, we get

$$P(|\xi_k(\tau, x)| > c\sqrt{nlogp/h}) \to 0$$

uniformly in k for given x. That is $P(|\xi_k(\tau, x)\sqrt{h/nlogp}| > c) \to 0$. Then we have,

$$\sup_{k=1,\ldots,d} |\xi_k(\tau, x)\sqrt{h/nlogp}| = O_p(1).$$

Thus,

$$\sup_{k=1,\ldots,d} |\xi_k(\tau, x)| = O_p(\sqrt{nlogp/h}).$$

Let $a_n = \sqrt{nlogp/h}$, then we get

$$\sup_{k=1,\ldots,d} |\xi_k(\tau, x)| = O_p(a_n).$$

□

Theorem 3

Theorem 3 shows the uniform consistency results for local partial likelihood estimate over k predictors and the domain of x.

Proof. **of Theorem 3.** Recall that when $p = 1$,

$$\hat{\psi}'_k(x) - \psi'_k(x) = \hat{\beta}^*_k(x) - \beta^0_k(x) = \frac{1}{h}\hat{\alpha}_k(x)$$

and

$$\hat{\alpha}_k = \frac{1}{n}\widetilde{\Sigma}_k(\tau, x)^{-1}\xi_k(x) + O_p(h^2).$$

Then

$$\hat{\psi}'_k(x) - \psi'_k(x) = \frac{1}{nh}\widetilde{\Sigma}_k(\tau, x)^{-1}\xi_k(x) + O_p(h).$$

Suppose there exists positive constant c' such that $\widetilde{\Sigma}_k(\tau, x) \leq c' < \infty$ uniformly in k. Then by Theorem 2,

$$\sup_{k=1,\ldots,d} |\hat{\psi}'_k(x) - \psi'_k(x)| = O_p(\sqrt{\frac{logp}{nh^3}}) + O_p(h). \tag{B.11}$$

As a result, for $a^*_n = \sqrt{\frac{logp}{nh^3}} + h$, we have

$$\sup_{k=1,\ldots,d} |\hat{\psi}'_k(x) - \psi'_k(x)| = O_p(a^*_n)$$

for given $x \in [0, 1]$.

Now partition $[0, 1]$ into M intervals by selecting points $\{x_0, \ldots, x_M\}$ with $x_0 = 0$ and $x_M = 1$ such that $|x_s - x_{s-1}| < \epsilon$ for some positive ϵ for s in $1, \ldots, M$. Then,

$$\max_{1\leq s\leq M} \sup_{k=1,\ldots,d} |\hat{\psi}'_k(x) - \psi'_k(x)| = O_p(a^*_n). \tag{B.12}$$

Note that

$$\begin{aligned}\sup_{x\in[0,1]}\sup_{k=1,\ldots,d}|\hat{\psi}_k'(x)-\psi_k'(x)| &\le \max_{1\le s\le M}\sup_{k=1,\ldots,d}|\hat{\psi}_k'(x_s)-\psi_k'(x_s)| \\ &+ \max_{1\le s\le M}\sup_{x\in[0,1]}\sup_{k=1,\ldots,d}\big|\hat{\psi}_k'(x)-\psi_k'(x)-(\hat{\psi}_k'(x_s)-\psi_k'(x_s))\big| \\ &= O_p(a_n^*)+\max_{1\le s\le M}\sup_{x\in[0,1]}\sup_{k=1,\ldots,d}\big|\frac{1}{nh}\widetilde{\Sigma}_k(\tau,x)^{-1}\xi_k(x)-\frac{1}{nh}\widetilde{\Sigma}_k(\tau,x_s)^{-1}\xi_k(x_s)\big|.\end{aligned} \tag{B.13}$$

Now we work on the second term on the right-hand side of (B.13). We have

$$\begin{aligned}&\max_{1\le s\le M}\sup_{x\in[0,1]}\sup_{k=1,\ldots,d}\big|\frac{1}{nh}\widetilde{\Sigma}_k(\tau,x)^{-1}\xi_k(x)-\frac{1}{nh}\widetilde{\Sigma}_k(\tau,x_s)^{-1}\xi_k(x_s)\big| \\ &\quad=\max_{1\le s\le M}\sup_{x\in[0,1]}\sup_{k=1,\ldots,d}\big|\frac{1}{nh}(\widetilde{\Sigma}_k(\tau,x)^{-1}-\widetilde{\Sigma}_k(\tau,x_s)^{-1})\xi_k(x_s) \\ &\qquad+\frac{1}{nh}\widetilde{\Sigma}_k(\tau,x)^{-1}(\xi_k(x)-\xi_k(x_s))\big| \\ &\quad\le\max_{1\le s\le M}\sup_{x\in[0,1]}\sup_{k=1,\ldots,d}\big|\frac{1}{nh}(\widetilde{\Sigma}_k(\tau,x)^{-1}-\widetilde{\Sigma}_k(\tau,x_s)^{-1})\xi_k(x_s)\big| \\ &\qquad+\max_{1\le s\le M}\sup_{x\in[0,1]}\sup_{k=1,\ldots,d}\big|\frac{1}{nh}(\widetilde{\Sigma}_k(\tau,x)^{-1}-\widetilde{\Sigma}_k(\tau,x_s)^{-1})\xi_k(x_s)\big|,\end{aligned}$$

which tends to zero as $\epsilon\to 0$ by the continuity of $\widetilde{\Sigma}_k(\tau,x)^{-1}$ and $\xi_k(x)$.

Thus we have

$$\sup_{x\in[0,1]}\sup_{k=1,\ldots,d}|\hat{\psi}_k'(x)-\psi_k'(x)|=O_p(a_n^*). \tag{B.14}$$

Since

$$|\hat{\psi}_k(x)-\psi_k(x)|\le\int_0^x|\hat{\psi}_k'(t)-\psi_k'(t)|dt\le\sup_{x\in[0,1]}|\hat{\psi}_k'(x)-\psi_k'(x)|$$

for any k. As a result, we prove the result

$$|\hat{\psi}_k(x)-\psi_k(x)|=O_p(a_n^*).$$

This result shows the uniform consistency of local linear estimator of local partial

likelihood for Cox's proportional model over dimension k and domain of x. $\square$

Theorem 4

Proof. **of Theorem 4.** Recall that

$$P_k = \frac{1}{n}\sum_{i=1}^{n}\delta_i\Big[\frac{1}{h}r_{ik}(X_{ik}) + log\big\{1 - \frac{\sum_{j=1}^{n}Y_j(Z_i)e^{\hat{\psi}_k(X_{jk})}r_{ik}(X_{jk})}{h\sum_{j=1}^{n}Y_j(Z_i)e^{\hat{\psi}_k(X_{jk})}}\big\}\Big],$$

where

$$r_{ik}(x) = \int_0^x H_k(t,\tau)^{-1}Q_{ik}(t,\tau)dt,$$

$$H_k(x,\tau) = -\mathcal{L}_x''(\boldsymbol{\alpha}_k,\tau)|_{\boldsymbol{\alpha}_k=\hat{\boldsymbol{\alpha}}_k}$$

and

$$\begin{aligned}
Q_{ik}(x,\tau) &= \frac{\partial}{\partial\boldsymbol{\alpha}_k}\big[\mathcal{L}_x(\boldsymbol{\alpha}_k,\tau) - \mathcal{L}_{x,-i}(\boldsymbol{\alpha}_k,\tau)\big]|_{\boldsymbol{\alpha}_k=\hat{\boldsymbol{\alpha}}_k} \\
&= \frac{1}{n}\delta_i K_h(X_{ik}-x)\mathbf{U}_{ik} - \frac{1}{n}\int_0^\tau\sum_{l=1}^{n}K_h(X_{lk}-x)\big[\frac{S_{nk}^{(1)}(\boldsymbol{\alpha}_k,\tau)}{S_{nk}^{(0)}(\boldsymbol{\alpha}_k,\tau)}\big]dN_l(u) \\
&\quad + \frac{1}{n}\int_0^\tau\sum_{l=1,l\neq i}^{n}K_h(X_{lk}-x)\big[\frac{S_{nk,-i}^{(1)}(\boldsymbol{\alpha}_k,\tau)}{S_{nk,-i}^{(0)}(\boldsymbol{\alpha}_k,\tau)}\big]dN_l(u).
\end{aligned}$$

In our case $\tau=\infty$. Consider the case $p=1$, since $\hat{\alpha}_k \xrightarrow{\text{P}} 0$, we now define

$$H_k^*(x) = -\mathcal{L}_x''(0,\infty),$$

$$\begin{aligned}
Q_{ik}^*(x) &= \frac{\partial}{\partial\alpha_k}\big[\mathcal{L}_x(\alpha_k,\infty) - \mathcal{L}_{x,-i}(\alpha_k,\infty)\big]|_{\alpha_k=0} \\
&= \frac{1}{n}\delta_i K_h(X_{ik}-x)U_{ik} - \frac{1}{n}\int_0^\infty\sum_{l=1}^{n}K_h(X_{lk}-x)\big[\frac{S_{nk}^{(1)}(0,\infty)}{S_{nk}^{(0)}(0,\infty)}\big]dN_l(u) \\
&\quad + \frac{1}{n}\int_0^\infty\sum_{l=1,l\neq i}^{n}K_h(X_{lk}-x)\big[\frac{S_{nk,-i}^{(1)}(0,\infty)}{S_{nk,-i}^{(0)}(0,\infty)}\big]dN_l(u).
\end{aligned}$$

and

$$r_{ik}^*(x) \equiv r_{ik}^*(x,\infty) = \int_0^x H_k^*(t)^{-1}Q_{ik}^*(t)dt.$$

Then P_k can be written as

$$P_k = \frac{1}{n}\sum_{i=1}^{n}\delta_i\Big[\frac{1}{h}r_{ik}^*(X_{ik}) + log\big\{1 - \frac{\sum_{j=1}^{n}Y_j(Z_i)e^{\hat{\psi}_k(X_{jk})}r_{ik}^*(X_{jk})}{h\sum_{j=1}^{n}Y_j(Z_i)e^{\hat{\psi}_k(X_{jk})}}\big\}\Big](1+o_p(1))$$
$$\equiv P_k^*(1+o_p(1)).$$

In the following, we work on P_k^*. By Lemma 2, we get

$$\sup_{0\le u\le\tau}\Big|\frac{S_{nk}^{(1)}(0,u)}{S_{nk}^{(0)}(0,u)} - \mu_1\Big| \xrightarrow{\mathrm{P}} 0,$$

where $\mu_1 = \int uK(u)du$. Then

$$\begin{aligned}Q_{ik}^*(x) &= \frac{1}{n}\delta_i K_h(X_{ik}-x)U_{ik} - \frac{1}{n}\int_0^\tau\sum_{l=1}^{n}K_h(X_{lk}-x)\mu_1 dN_l(u)\\ &\quad + \frac{1}{n}\int_0^\tau\sum_{l=1,l\neq i}^{n}K_h(X_{lk}-x)\mu_1 dN_l(u) + o_p(1)\\ &= \frac{1}{nh}\delta_i K_h(X_{ik}-x)(X_{ik}-x) - \frac{1}{n}\delta_i K_h(X_{ik}-x)\mu_1 + o_p(1).\end{aligned}$$

Define $K_h^{(1)}(x) = K_h(x/h)(x/h)$. Then $Q_{ik}^*(x)$ can be represented as

$$Q_{ik}^*(x) = \frac{1}{n}\delta_i K_h^{(1)}(X_{ik}-x) - \frac{1}{n}\delta_i K_h(X_{ik}-x)\mu_1. \tag{B.15}$$

Recall that

$$\mathcal{L}_x''(0,\tau) = -\widetilde{\Sigma}_k(\tau,x) + o_p(1),$$

where $\widetilde{\Sigma}_k(\tau, x) = v_1 f_k(x)\Psi_k(x)\Lambda_k(\tau, x)$. Thus,

$$\begin{aligned} r^*_{ik}(x) &= \int_0^x H^*_k(t)^{-1} Q^*_{ik}(t)dt \\ &= \int_0^x (\widetilde{\Sigma}_k(\tau,t)^{-1} + o_p(1))[\frac{1}{n}\delta_i K^{(1)}_h(X_{ik}-t) - \frac{1}{n}\delta_i K_h(X_{ik}-t)\mu_1]dt \\ &= \int_0^x \widetilde{\Sigma}_k(\tau,t)^{-1}[\frac{1}{n}\delta_i K^{(1)}_h(X_{ik}-t) - \frac{1}{n}\delta_i K_h(X_{ik}-t)\mu_1]dt \\ &\quad + o_p(1)\frac{1}{n}\delta_i \int_0^x [K^{(1)}_h(X_{ik}-t) - K_h(X_{ik}-t)\mu_1]dt. \end{aligned}$$

Let $u \equiv (X_{ik}-t)/h$, then $t = X_{ik} - uh$ and $(X_{ik}-x)/h \le u \le X_{ik}/h$. Thus

$$\begin{aligned} r^*_{ik} &= \int_{(X_{ik}-x)/h}^{X_{ik}/h} -\widetilde{\Sigma}_k(\tau, X_{ik}-uh)^{-1}\big[\frac{1}{n}\delta_i K(u)u - \frac{1}{n}\delta_i\mu_1 K(u)\big]du + o_p(\frac{1}{n}) \\ &= \frac{1}{n}\delta_i \int_{(X_{ik}-x)/h}^{X_{ik}/h} \widetilde{\Sigma}_k(\tau, X_{ik}-uh)^{-1} K(u)(\mu_1 - u)du + o_p(\frac{1}{n}). \end{aligned}$$

Since $h \to 0$, by Taylor expansion, we have

$$\widetilde{\Sigma}_k(\tau, X_{ik}-uh)^{-1} = \widetilde{\Sigma}_k(\tau, X_{ik})^{-1} + O_p(uh).$$

As a result, we obtain

$$r^*_{ik} = \frac{1}{n}\delta_i \widetilde{\Sigma}_k(\tau, X_{ik})^{-1} \int_{(X_{ik}-x)/h}^{X_{ik}/h} (\mu_1 - u)K(u)du + o_p(\frac{1}{n}). \tag{B.16}$$

Plug the result (B.15) and (B.16) back into the definition of P^*_k and by Taylor

expansion of $log(1-x)$,

$$
\begin{aligned}
P_k^* &= \frac{1}{n}\sum_{i=1}^{n}\delta_i\big[\frac{1}{nh}\delta_i\widetilde{\Sigma}_k(\tau,X_{ik})^{-1}\int_0^{X_{ik}/h}(\mu_1-u)K(u)du \\
&\quad - \frac{\frac{1}{n}\sum_{j=1}^{n}Y_j(Z_i)exp(\psi_k(X_{jk}))\delta_i\widetilde{\Sigma}_k(\tau,X_{ik})^{-1}\int_{(X_{ik}-X_{jk})/h}^{X_{ik}/h}(\mu_1-u)K(u)du}{h\sum_{j=1}^{n}Y_j(Z_i)exp(\psi_k(X_{jk}))}\big]+o_p(\frac{1}{nh}) \\
&= \frac{1}{n^2h}\sum_{i=1}^{n}\delta_i\widetilde{\Sigma}_k(\tau,X_{ik})^{-1}\big[\int_0^{X_{ik}/h}(\mu_1-u)K(u)du \\
&\quad - \frac{\sum_{j=1}^{n}Y_j(Z_i)exp(\psi_k(X_{jk}))\int_{(X_{ik}-X_{jk})/h}^{X_{ik}/h}(\mu_1-u)K(u)du}{\sum_{j=1}^{n}Y_j(Z_i)exp(\psi_k(X_{jk}))}\big]+o_p(\frac{1}{nh}).
\end{aligned}
$$

Note that

$$
\begin{aligned}
\int_{(X_{ik}-X_{jk})/h}^{X_{ik}/h}(\mu_1-u)K(u)du &= \int_0^{X_{ik}/h}(\mu_1-u)K(u)du \\
&\quad - \int_0^{(X_{ik}-Xjk)/h}(\mu_1-u)K(u)du.
\end{aligned}
$$

Then

$$
\begin{aligned}
&\int_0^{X_{ik}/h}(\mu_1-u)K(u)du - \frac{\sum_{j=1}^{n}Y_j(Z_i)exp(\psi_k(X_{jk}))\int_{(X_{ik}-X_{jk})/h}^{X_{ik}/h}(\mu_1-u)K(u)du}{\sum_{j\in R_i}exp(\psi_k(X_{jk}))} \\
&\qquad = \frac{\sum_{j=1}^{n}Y_j(Z_i)exp(\psi_k(X_{jk}))\int_0^{(X_{ik}-X_{jk})/h}(\mu_1-u)K(u)du}{\sum_{j=1}^{n}Y_j(Z_i)exp(\psi_k(X_{jk}))}.
\end{aligned}
$$

As a result,

$$
\begin{aligned}
P_k^* = \frac{1}{n}\sum_{i=1}^{n}\delta_i\big[&\frac{1}{nh}\delta_i\widetilde{\Sigma}_k(\tau,X_{ik})^{-1} \\
&\frac{\sum_{j=1}^{n}Y_j(Z_i)exp(\psi_k(X_{jk}))\int_0^{(X_{ik}-X_{jk})/h}(\mu_1-u)K(u)du}{\sum_{j=1}^{n}Y_j(Z_i)exp(\psi_k(X_{jk}))}\big]+o_p(\frac{1}{nh}).
\end{aligned}
\tag{B.17}
$$

Now we define

$$
g_{nk}(Z_i,X_{ik}) = \frac{1}{n-1}\sum_{j=1,j\neq i}^{n}Y_j(Z_i)exp(\psi_k(X_{jk}))\int_0^{(X_{ik}-X_{jk})/h}(\mu_1-u)K(u)du
$$

and

$$d_{nk}(Z_i) = \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} Y_j(Z_i) exp(\psi_k(X_{jk})).$$

Then

$$\begin{aligned} P_k^* &= \frac{1}{nh^2} \sum_{i=1}^{n} \delta_i \widetilde{\Sigma}_k(\tau, X_{ik})^{-1} \Bigg[\frac{E\{g_{nk}(Z_i, X_{ik})|Z_i\}}{E\{d_{nk}(Z_i)|Z_i\}} + \Big(\frac{g_{nk}(Z_i, X_{ik})}{d_{nk}(Z_i)} \\ &\quad - \frac{E\{g_{nk}(Z_i, X_{ik})|Z_i\}}{E\{d_{nk}(Z_i)|Z_i\}} \Big) \Bigg] + o_p(\frac{1}{nh}) \\ &= \frac{1}{nh^2} \delta_i \widetilde{\Sigma}_k(\tau, X_{ik})^{-1} \frac{E\{g_{nk}(Z_i, X_{ik})|Z_i\}}{E\{d_{nk}(Z_i,)|Z_i\}} + o_p(\frac{1}{nh}). \end{aligned}$$

Note that

$$\begin{aligned} E\{d_{nk}(Z_i)|Z_i\} &= E\{\mathcal{I}(Z_j \geq Z_i) exp(\psi_k(X_{jk}))|Z_i\} \\ &= E\{E\{\mathcal{I}(Z_j \geq Z_i|X_{jk}\} exp(\psi_k(X_{jk}))\} \\ &= E\{P(Z_j \geq Z_i|X_{jk}) exp(\psi_k(X_{jk})\} \\ &= \int \int_{Z_i}^{\infty} \widetilde{f}_k(z|x) exp(\psi_k(x)) f_k(x) dz dx \\ &\equiv D_{ik}, \end{aligned}$$

where $\widetilde{f}_k(z|x)$ is the conditional density of z given x for predictor X_k.

Similarly,

$$\begin{aligned} E\{g_{nk}(Z_i, X_{ik})|Z_i\} &= E\{P(Z_j \geq Z_i|X_{jk}) exp(\psi_k(X_{jk}) \int_0^{\frac{(X_{ik}-X_{jk})}{h}} (\mu_1 - u) K(u) du\} \\ &= \int \int_{Z_i}^{\infty} \int_0^{\frac{(X_{ik}-x)}{h}} \widetilde{f}_k(z|x) exp(\psi_k(x)) (\mu_1 - u) K(u) f_k(x) du dz dx \\ &\equiv G_{ik}. \end{aligned}$$

As a result,

$$P_k^* = \frac{1}{nh^2}\sum_{i=1}^{n}\delta_i\widetilde{\Sigma}_k(\tau, X_{ik})^{-1}G_{ik}/D_{ik} + o_p(\frac{1}{nh})$$
$$= \frac{1}{nh}(\frac{1}{n}\sum_{i=1}^{n}\delta_i\widetilde{\Sigma}_k(\tau, X_{ik})^{-1}G_{ik}/D_{ik} + o_p(1)).$$

Put

$$\Omega_k = E\{\frac{1}{n}\sum_{i=1}^{n}\delta_i\widetilde{\Sigma}_k(\tau, X_{ik})^{-1}G_{ik}/D_{ik}\}$$

and

$$P_{ik}^* = \delta_i\widetilde{\Sigma}_k(\tau, X_{ik})^{-1}G_{ik}/D_{ik} - \Omega_k,$$

then

$$P_{ik}^* - \frac{1}{nh}\Omega_k = \frac{1}{nh}(\frac{1}{n}\sum_{i=1}^{n}P_{ik}^* + o_p(1)).$$

It is clear that

$$E\{P_{ik}^*\} = E\{E(P_{ik}^*|X)\} = 0.$$

Then assume Condition 8 holdes and we apply the Berstien exponential inequality , we have for $k = 1, \ldots, d$

$$P(|\sum_{i=1}^{n}P_{ik}^*| > x) \leq 2exp(-\frac{x^2}{2[E(P_{ik}^{*2} + Mx]}).$$

Choose $x = n^2hc$, then

$$P(\frac{1}{nh^2}|\sum_{i=1}^{n}P_{ik}^*| > c) \leq 2exp(-\frac{n^3h^2c^2}{2[E(P_{ik}^{*2} + nhMc]}),$$

That is, for $k = 1, \ldots, d$

$$P(|P_k^* - \frac{1}{nh}\Omega_k| > c) \leq 2exp(-\frac{n^3h^2c^2}{2[E(P_{ik}^{*2} + nhMc]}).$$

Then we can prove that

$$\sup_{k=1,\ldots,d} |P_k - \frac{\Omega_k}{nh}| \xrightarrow{\mathrm{P}} 0.$$

□

Theorem 5

Proof. **of Theorem 5.** We decompose $\hat{S} = S$ into d terms,

$$\begin{aligned} P(\hat{S} = S) &= P(\cap_{k\in S}[IC_k(h^*) < IC_k(\infty)] \cap_{k\notin S} [IC_k(h^*) > IC_k(\infty)] \\ &\geq 1 - \sum_{k\in S} P(IC_k(h^*) > IC_k(\infty)) - \sum_{k\notin S} P(IC_k(h^*) < IC_k(\infty)) \\ &\geq 1 - \sum_{k=1}^{d} P(F_k), \end{aligned}$$

where $F_k = \{IC_k(h^*) > IC_k(\infty)\}$ if $k \in S$ and $F_k = \{IC_k(h^*) < IC_k(\infty)\}$ if $k \notin S$. Then it suffices to prove for $k \in S$,$IC_k(h^*) < IC_k(\infty)$ with high probability and for $k \notin S$, $IC_k(h^*) > IC_k(\infty)$ with high probability.

Recall the definition of IC as in (2.25). When $h = \infty$,

$$IC_k(\infty) = -\mathcal{L}(\widetilde{\psi}_k) = -\frac{1}{n}\sum_{i=1}^{n} \delta_i \Big[\widetilde{\psi}_k(X_{ik}) - log\big\{\sum_{j=1}^{n} Y_j(Z_i) exp(\widetilde{\psi}_k(X_{jk}))\big\}\Big].$$

When $h = h^* = (\frac{logp}{n})^{\frac{1}{5}}$, we have

$$\begin{aligned} IC_k(h) &= -\mathcal{L}(\hat{\psi}_k) + \frac{1}{n}\sum_{i=1}^{n} \delta_i \Big[\frac{1}{h} r_{ik}(X_{ik}) + \\ &\qquad log\big\{1 - \frac{\sum_{j=1}^{n} Y_j(Z_i) e^{\hat{\psi}_k(X_{jk})} r_{ik}(X_{jk})}{h\sum_{j=1}^{n} Y_j(Z_i) e^{\hat{\psi}_k(X_{jk})}}\big\}\Big] \tau (\frac{nlogp}{h})^{\frac{1}{2}} \\ &= -\mathcal{L}(\hat{\psi}_k) + P_k \tau (\frac{nlogp}{h})^{\frac{1}{2}}. \end{aligned}$$

We have the uniform deviation of $\mathcal{L}(\hat{\psi}_k)$ and $\mathcal{L}(\widetilde{\psi}_k)$ from Proposition 1 that there exists a set $\mathcal{B}_1$ with $P(\mathcal{B}_1) \to 1$ and a universal constant $B_1 > 0$ such that on set $\mathcal{B}_1$, we have

$$|\mathcal{L}(\hat{\boldsymbol{\psi}}_k) - \mathcal{L}(\psi_k)| \leq B_1 a_n^*, \tag{B.18}$$

for all $k = 1, \ldots, d$.

From Lemma 4, for any given $B_2 > 0$, there exists a set $\mathcal{B}_2$ with $P(\mathcal{B}_2) \to 1$ as $n \to \infty$ such that on the set $\mathcal{B}_2$, we have

$$|P_k - \frac{\Omega_k}{nh}| \leq B_2, \tag{B.19}$$

for all $k = 1, \ldots, d$.

Then for all $k = 1, \ldots, d$ on set $\mathcal{B} = \mathcal{B}_1 \cap \mathcal{B}_2$,

$$-\mathcal{L}(\psi_k) + \frac{\tau\Omega_k a_n}{nh} - B_1 a_n^* < IC_k(h) < -\mathcal{L}(\psi_k) + \frac{\tau\Omega_k a_n}{nh} + B_1 a_n^*, \tag{B.20}$$

where $a_n = (\frac{nlogp}{h})^{\frac{1}{2}}$.

Now for each variable, we compare the IC for the two different bandwidth $h = h^*$ and $h = \infty$. Assume $h \to 0$ and $nh \to \infty$, then for $k \in S$, we have

$$IC_k(h) - IC_k(\infty) < \mathcal{L}(\widetilde{\psi}_k) - \mathcal{L}(\psi_k) + \frac{\tau\Omega_k a_n}{nh} + B_1 a_n^*.$$

With $h = (\frac{logp}{n})^{\frac{1}{5}}$ and for large n, we have

$$IC_k(h) - IC_k(\infty) < \mathcal{L}(\widetilde{\psi}_k) - \mathcal{L}(\psi_k) + \tau\Omega_k(\frac{logp}{n})^{\frac{1}{5}} + B_1(\frac{logp}{n})^{\frac{1}{5}}. \tag{B.21}$$

From Proposition 2 and Proposition 3, we get

$$\mathcal{L}(\widetilde{\psi}_k) - \mathcal{L}(\psi_k) = E\{\delta[x_k^T\widetilde{\beta}_k - logS_{0k}^*(z)]\} - E\{\delta[\psi_k(x_k) - logS_{0k}(z)]\}.$$

Using Condition 9 on the signal level, $IC_k(h) - IC_k(\infty) < 0$ for all $k \in S$ with high probability. This proves that for the variables with nonlinear impact, the favored bandwidth is $h = h^*$. Next, we consider $k \notin S$. Similarly, for $k \notin S$,

$$IC_k(h) - IC_k(\infty) > \mathcal{L}(\widetilde{\psi}_k) - \mathcal{L}(\psi_k) + \frac{\tau\Omega_k a_n}{nh} - B_1 a_n^*.$$

With $h = (\frac{logp}{n})^{\frac{1}{5}}$ and for large n, we have

$$IC_k(h) - IC_k(\infty) > \mathcal{L}(\widetilde{\psi}_k) - \mathcal{L}(\psi_k) + (\tau\Omega_k - B_1)(\frac{logp}{n})^{\frac{1}{5}}. \tag{B.22}$$

For $k \notin S$, on set $\mathcal{B}$ if $\tau\Omega_k > B_1$, using Condition 9 for $k \notin S$, we get $IC_k(h) - IC_k(\infty) > 0$ with high probability. This shows that for variables with linear impact, the favored bandwidth is $h = \infty$. Together, we prove that the proposed method achieves the selection consistency in the scenario that that the dimensionality grows at an exponential rate of sample size, i.e., $p = o(exp(n^\alpha))$ with $0 < \alpha < 1$. □

APPENDIX C: DERIVATION OF THE INFLUENCE FUNCTION

We extend the idea in N. Reid and H. Crepeau (1985)[35] to derive the influence function for Cox's proportional model with nonparametric risk effect. We apply the method of random covariates used in regression models (Krasker and Welsch, 1982[43]) to nonparametric proportional hazard model.

To simplify the notations, we consider the case for a general X. Denote $H(z, x, \delta)$ as the joint cumulative distribution function for (Z, X, δ), where Z is the survival time, X is the covariate and δ is the censoring indicator. Denote $H_n(z, x, \delta)$ as the corresponding empirical distribution function. Similarly, Denote $H(z, x)$ as the joint marginal distribution function of (Z, X) and $H_n(z, x)$ as the corresponding distribution function.

Then at a local point x_0, the local partial likelihood estimator β^* satisfies

$$\sum_{i=1}^{n} \delta_i K_h(X_i - x_0)\big[\widetilde{X}_i^* - \frac{\sum_{j=1}^{n} Y_j(Z_i) K_h(X_j - x_0) exp(\widetilde{X}_j^{*T}\beta^*)\widetilde{X}_j^*}{\sum_{j=1}^{n} Y_j(Z_i) K_h(X_j - x_0) exp(\widetilde{X}_j^{*T}\beta^*)}\big] = 0. \qquad \text{(C.1)}$$

To simply the notation, we rewrite

$$\widetilde{X}_i^* = \{X_i - x_0, \ldots, (X_i - x_0)^p\}^T \equiv X_i(x_0).$$

Since

$$H_n(z, x, \delta) = \frac{1}{n}\sum_{i=1}^{n} \delta_{(Z_i, X_i, \delta_i)}(z, x, \delta) \quad \text{and} \quad H_n(\widetilde{z}, \widetilde{x}) = \frac{1}{n}\sum_{i=1}^{n} \delta_{(Z_i, X_i)}(z, x),$$

where

$$\delta_{(Z_i, X_i, \delta_i)}(z, x, \delta) = \begin{cases} 1 & \text{if} \quad z = Z_i, x = X_i, \delta = \delta_i \\ 0 & \text{otherwise} \end{cases}$$

and

$$\delta_{(Z_i,X_i)}(z,x) = \begin{cases} 1 & \text{if} \quad z = Z_i, x = X_i \\ 0 & \text{otherwise.} \end{cases}$$

Then (C.1) becomes

$$\int \delta K_h(x-x_0)\big[x(x_0) - \frac{\int K_h(\widetilde{x}-x_0)exp(\widetilde{x}(x_0)^T\beta^*)\widetilde{x}(x_0)\mathcal{I}(\widetilde{z}\geq z)dH_n(\widetilde{z},\widetilde{x})}{\int K_h(\widetilde{x}-x_0)exp(\widetilde{x}(x_0)^T\beta^*)\mathcal{I}(\widetilde{z}\geq z)dH_n(\widetilde{z},\widetilde{x})}\big] \qquad (C.2)$$
$$dH_n(z,x,\delta) = 0.$$

Replace the empirical distribution function in (C.2) by the population distribution function, we get the infinite sample version as:

$$\int \delta K_h(x-x_0)\big[x(x_0) - \frac{\int_{\widetilde{z}\geq z} K_h(\widetilde{x}-x_0)exp(\widetilde{x}(x_0)^T\beta^*(H))\widetilde{x}(x_0)dH(\widetilde{z},\widetilde{x})}{\int_{\widetilde{z}\geq z} K_h(\widetilde{x}-x_0)exp(\widetilde{x}(x_0)^T\beta^*(H))dH(\widetilde{z},\widetilde{x})}\big] \qquad (C.3)$$
$$dH(z,x,\delta) = 0.$$

The above equation defines β^* as a functional $\beta^*(H)$.

The definition of influence function of T at F is (Hample, 1974[44]):

$$\hat{IF}(X_i) = \lim_{\epsilon\to 0}\{T[(1-\epsilon)F + \epsilon\delta_x] - T(F)\}/\epsilon$$

at the point x, where δ_x is a point mass at x.

Replace H in (C.3) by $(1-\epsilon)H + \epsilon\delta_{(Z_i,X_i,\delta_i)}$, then $\beta^*[(1-\epsilon)H + \epsilon\delta_{(Z_i,X_i,\delta_i)}]$ satisfies

$$\int \delta K_h(x-x_0)\Bigg[x(x_0) - \int_{\widetilde{z}\geq z} K_h(\widetilde{x}-x_0)exp(\widetilde{x}(x_0)^T\beta^*[(1-\epsilon)H + \epsilon\delta_{(Z_i,X_i,\delta_i)}])\widetilde{x}(x_0)[(1-\epsilon)$$
$$dH(\widetilde{x},\widetilde{z}) + \epsilon d\delta_{Z_i,X_i,\delta_i}(\widetilde{x},\widetilde{z})]\Big/ \int_{\widetilde{z}\geq z} K_h(\widetilde{x}-x_0)exp(\widetilde{x}(x_0)^T\beta^*[(1-\epsilon)H + \epsilon\delta_{(Z_i,X_i,\delta_i)}])$$
$$[(1-\epsilon)dH(\widetilde{x},\widetilde{z}) + \epsilon d\delta_{(Z_i,X_i,\delta_i)}(\widetilde{x},\widetilde{z})]\Bigg]\big[(1-\epsilon)dH(z,x,\delta) + \epsilon d\delta_{(Z_i,X_i,\delta_i)}(z,x,\delta)\big] = 0$$

To simply the notations, we now write

$$\beta^*(\epsilon) \equiv \beta^*[(1-\epsilon)H + \epsilon\delta_{(Z_i,X_i,\delta_i)}],$$

$$\alpha_0(z,\beta^*(\epsilon),x_0) \equiv \int_{\widetilde{z}\geq z} K_h(\widetilde{x}-x_0)exp(\widetilde{x}(x_0)^T\beta^*(\epsilon)dH(\widetilde{z},\widetilde{x}),$$

$$\alpha_1(z,\beta^*(\epsilon),x_0) \equiv \frac{\partial}{\partial\beta^*}\alpha_0(z,\beta^*(\epsilon),x_0) \quad \text{and} \quad \alpha_2(z,\beta^*(\epsilon),x_0) \equiv \frac{\partial}{\partial\beta^*}\alpha_1(z,\beta^*(\epsilon),x_0).$$

Then we have

$$\int \delta K_h(x-x_0)\Big[x(x_0) - \big\{(1-\epsilon)\alpha_1(z,\beta^*(\epsilon),x_0) + \epsilon K_h(X_i-x_0)exp(\widetilde{X}_i^{*T}\beta^*)\mathcal{I}(z_i\geq z)\widetilde{X}_i^*\big\}/ \big\{(1-\epsilon)\alpha_0(z,\beta^*(\epsilon),x_0) + \epsilon K_h(X_i-x_0)exp(\widetilde{X}_i^{*T}\beta^*)\mathcal{I}(z_i\geq z)\big\}\Big] \big[dH(z,x,\delta)(1-\epsilon) + \epsilon d\delta_{(Z_i,X_i,\delta_i)}\big] = 0.$$

Further, we get

$$\begin{aligned}(1-\epsilon)\int \delta K_h(x-x_0)\Big[&x(x_0) - \big\{(1-\epsilon)\alpha_1(z,\beta^*(\epsilon),x_0) + \epsilon K_h(X_i-x_0)exp(\widetilde{X}_i^{*T}\beta^*) \\ &\mathcal{I}(z_i\geq z)\widetilde{X}_i^*\big\}/\big\{(1-\epsilon)\alpha_0(z,\beta^*(\epsilon),x_0) + \epsilon K_h(X_i-x_0)exp(\widetilde{X}_i^{*T}\beta^*)\mathcal{I}(z_i\geq z)\big\}\Big] \\ &dH(z,x,\delta) + \epsilon\delta_i K_h(X_i-x_0)\Big[\widetilde{X}_i^* - \big\{(1-\epsilon)\alpha_1(z,\beta^*(\epsilon),x_0) + \epsilon K_h(X_i-x_0) \\ &exp(\widetilde{X}_i^{*T}\beta^*)\widetilde{X}_i^*\big\}/\big\{(1-\epsilon)\alpha_0(z,\beta^*(\epsilon),x_0) + \epsilon K_h(X_i-x_0)exp(\widetilde{X}_i^{*T}\beta^*)\big\}\Big] = 0.\end{aligned} \tag{C.4}$$

Denote the first term in (C.4) as $K_1(\beta^*,\epsilon,x_0)$ and the second term as $K_2(\beta^*,\epsilon,x_0)$. Then take derivative with respect to ϵ on both sides of (C.4) and evaluate at $\epsilon = 0$, we have

$$\Big[\Big(\frac{\partial K_1(\beta^*,\epsilon,x_0)}{\partial\beta^{*T}} + \frac{\partial K_2(\beta^*,\epsilon,x_0)}{\partial\beta^{*T}}\Big)\frac{\partial\beta^*}{\partial\epsilon} + \frac{\partial K_1(\beta^*,\epsilon,x_0)}{\partial\epsilon} + \frac{\partial K_1(\beta^*,\epsilon,x_0)}{\partial\epsilon}\Big]\Big|_{\epsilon=0} = 0. \tag{C.5}$$

Since

$$\begin{aligned}\frac{\partial K_1(\beta^*,\epsilon,x_0)}{\partial\epsilon}\Big|_{\epsilon=0} &= -\int \delta K_h(X_i-x_0)\big[x(x_0)-\frac{\alpha_1(z,\beta^*(H),x_0)}{\alpha_0(z,\beta^*(H),x_0)}\big]dH(z,x,\delta)\\ &\quad -\int \delta K_h(X_i-x_0)\Big\{\big[\alpha_1(z,\beta^*(H),x_0)+K_h(X_i-x_0)exp(\widetilde{X}_i^{*T}\beta^*)\\ &\quad \mathcal{I}(z_i\geq z)\widetilde{X}_i^*\big]\alpha_0(z,\beta^*(H),x_0)-\alpha_1(z,\beta^*(H),x_0)\big[-\alpha_0(z,\beta^*(H),x_0)\\ &\quad +K_h(X_i-x_0)exp(\widetilde{X}_i^{*T}\beta^*)\mathcal{I}(z_i\geq z)\Big\}/\alpha_0^2(z,\beta^*(H),x_0)dH(z,x,\delta)\\ &= -\int \frac{\delta K_h(x-x_0)}{\alpha_0(z,\beta^*(H),x_0)}\Big[K_h(X_i-x_0)exp(\widetilde{X}_i^{*T}\beta^*)\mathcal{I}(z_i\geq z)\widetilde{X}_i^*-\\ &\quad \frac{\alpha_1(z,\beta^*(H),x_0)}{\alpha_0(z,\beta^*(H),x_0)}K_h(X_i-x_0)exp(\widetilde{X}_i^{*T}\beta^*)\mathcal{I}(z_i\geq z)\Big]dH(z,x,\delta),\end{aligned}$$

$$\frac{\partial K_2(\beta^*,\epsilon,x_0)}{\partial\epsilon}\Big|_{\epsilon=0} = \delta_i K_h(X_i-x_0)\big[\widetilde{X}_i^*-\frac{\alpha_1(z,\beta^*(H),x_0)}{\alpha_0(z,\beta^*(H),x_0)}\big],$$

and

$$\begin{aligned}\frac{\partial K_1(\beta^*,\epsilon,x_0)}{\partial\beta^{*T}}\Big|_{\epsilon=0} &= -\int \delta K_h(X_i-x_0)\Big[\frac{\alpha_2(z,\beta^*(H),x_0)}{\alpha_0(z,\beta^*(H),x_0)}-\frac{\alpha_1(z,\beta^*(H),x_0)}{\alpha_0(z,\beta^*(H),x_0)}\\ &\quad (\frac{\alpha_1(z,\beta^*(H),x_0)}{\alpha_0(z,\beta^*(H),x_0)})^T\Big]dH(z,x,\delta)\\ &\equiv A^*.\end{aligned}$$

Plug these into (C.5) produces

$$\begin{aligned}A^*\frac{\partial\beta^*[(1-\epsilon)H+\epsilon\delta_{(Z_i,X_i,\delta_i)}]}{\partial\epsilon}\Big|_{\epsilon=0} &= \delta_i K_h(X_i-x_0)\big[\widetilde{X}_i^*-\frac{\alpha_1(z,\beta^*(H),x_0)}{\alpha_0(z,\beta^*(H),x_0)}\big]-\\ \int \big[K_h(X_i-x_0)exp(\widetilde{X}_i^{*T}\beta^*)\mathcal{I}(z_i\geq z)\widetilde{X}_i^*-\frac{\alpha_1(z,\beta^*(H),x_0)}{\alpha_0(z,\beta^*(H),x_0)}K_h(X_i-x_0)\\ exp(\widetilde{X}_i^{*T}\beta^*)\mathcal{I}(z_i\geq z)\big]\frac{\delta K_h(X_i-x_0)}{\alpha_0(z,\beta^*(H),x_0)}dH(z,x,\delta) &= 0.\end{aligned}\tag{C.6}$$

Replace $H(z,x,\delta)$ by $H_n(z,x,\delta)$ in (C.6), then we get the finite sample version of $I\hat{F}$

as

$$A^*(\hat{\beta}^*)I\hat{F}_i = \delta_i K_h(X_i - x_0)\big[\widetilde{X}_i^* - \frac{\sum_{j=1}^n Y_j(Z_i)K_h(X_j - x_0)exp(\widetilde{X}_j^{*T}\hat{\beta}^*)\widetilde{X}_j^*}{\sum_{j=1}^n Y_j(Z_i)K_h(X_j - x_0)exp(\widetilde{X}_j^{*T}\hat{\beta}^*)}\big] + C_i^*(\hat{\beta}^*), \tag{C.7}$$

where

$$A^*(\hat{\beta}^*) = \frac{1}{n}\sum_{i=1}^n \delta_i K_h(X_i - x_0)\big[\frac{\sum_{j=1}^n Y_j(Z_i)K_h(X_j - x_0)exp(\widetilde{X}_j^{*T}\hat{\beta}^*)\widetilde{X}_j^*\widetilde{X}_j^{*T}}{\sum_{j=1}^n Y_j(Z_i)K_h(X_j - x_0)exp(\widetilde{X}_j^{*T}\hat{\beta}^*)} - (\frac{\sum_{j=1}^n Y_j(Z_i)K_h(X_j - x_0)exp(\widetilde{X}_j^{*T}\hat{\beta}^*)\widetilde{X}_j^*}{\sum_{j=1}^n Y_j(Z_i)K_h(X_j - x_0)exp(\widetilde{X}_j^{*T}\hat{\beta}^*)})^{\otimes 2}\big],$$

and

$$C_i^*(\hat{\beta}^*) = K_h(X_i - x_0)exp(\widetilde{X}_i^{*T}\hat{\beta}^*)\big[-\widetilde{X}_i^*\sum_{z_j \le z_i}\frac{\delta_j K_h(X_j - x_0)}{\sum_{l=1}^n Y_l(Z_j)K_h(X_l - x_0)exp(\widetilde{X}_l^{*T}\hat{\beta}^*)}) + \sum_{z_j \le z_i}\frac{\delta_j K_h(X_j - x_0)\sum_{l=1}^n Y_l(Z_j)K_h(X_l - x_0)exp(\widetilde{X}_j^{*T}\hat{\beta}^*)\widetilde{X}_l^*}{\{\sum_{l=1}^n Y_l(Z_j)K_h(X_l - x_0)exp(\widetilde{X}_l^{*T}\hat{\beta}^*)\}^2}\big].$$

Remark. Since $\hat{\beta}^* = \beta^*(H_n)$, then by the definition of influence function,

$$I\hat{F}_i = \lim_{\epsilon \to 0}\frac{\beta^*[(1-\epsilon)H_n + \epsilon\delta_{(Z_i,X_i,\delta_i)}] - \beta^*(H_n)}{\epsilon}.$$

When n is large enough, $\frac{1}{n-1} \to 0$. As a result, letting $\epsilon = -\frac{1}{n-1}$, we have

$$\begin{aligned}(1-\epsilon)H_n + \epsilon\delta_{Z_i,X_i,\delta_i} &= (1 + \frac{1}{n-1})\frac{\sum_{h=1}^n \delta_(Z_j, X_j, \delta_j)}{n} - \frac{1}{n-1}\delta_{(Z_i,X_i,\delta_i)} \\ &= \frac{1}{n-1}\sum_{j=1, j\neq i}^n \delta_{(Z_j,X_j,\delta_j)} \equiv H_{n,-i}.\end{aligned}$$

Therefore, when n is large enough, we have

$$\begin{aligned}I\hat{F}_i &\approx -(n-1)[\beta^*(H_{n,-i}) - \beta^*(H_n)] \\ &= (n-1)(\hat{\beta}^* - \hat{\beta}_{-i}).\end{aligned}$$