Journal of the Korean Statistical Society (2024) 53:509–536
https://doi.org/10.1007/s42952-024-00256-5

Online ISSN 2005-2863
Print ISSN 1226-3192

RESEARCH ARTICLE

Gradient-based kernel variable selection for support vector hazards machine

Sanghun Jeong[1] · **Kyungjun Kang**[1] · **Hojin Yang**[1]

Received: 10 July 2023 / Accepted: 2 January 2024 / Published online: 15 February 2024

Abstract
This study aims to improve the predictive performance for the event time through the machine learning model and find informative variables in the time-to-event data, simultaneously. To address this issue, after regarding the time-to-event data as the dichotomized counting processes data for predicting survival time, we consider the time-dependent support vector machine (SVM) framework for the dichotomized counting process data, where the decision function in this framework consists of the time-independent risk score and time-dependent intercept. Also, we consider the empirical partial derivative of the risk score function with respect to each marginal predictor as the indicator for the important predictor. Through this approach, it is possible to predict survival time and find variables that affect on the survival time at the same time. Simulation studies were conducted to confirm the performance of the model, and real data analysis was conducted by predicting the survival time of the lung cancer after the diagnosis and selecting genes associate with lung cancer through human gene data.

Keywords Counting process · Prediction · Reproducing kernel · Risk process · Variable selection · Weight

1 Introduction

Several parametric and nonparametric approaches have been developed in the survival analysis area to account for the occurrence or survival time of an event. Most of these methods focus on estimating conditional and unconditional survival

✉ Hojin Yang
hjyang@pusan.ac.kr

Sanghun Jeong
shjeong95@pusan.ac.kr

Kyungjun Kang
kangjun205@gmail.com

[1] Department of Statistics, Pusan National University, Busan, South Korea

probability up to a certain point in time and predicting the conditional survival time using covariates (Kalbfleisch & Prentice, 2011; Fleming & Harrington, 2011; Ibrahim et al., 2001; Lawless, 2002). The outcome variable is the observed time, which can be either of the survival time or the censoring time away from the observation, wherein the observed time's actual meaning is denoted by the censoring indicator. Such a characteristic of the outcome variable requires a distinct approach—one that differs from the conventional statistical models, such as the Cox proportional hazards model (Cox, 1972), which is the representative method widely used for the time-to-event data, particularly in terms of the hazard function associated with the survival outcome as the multiplicative structure between the baseline hazard function and the exponential function of the linear predictor. Estimators of the regression coefficients are obtained by numerically maximizing the partial likelihood describing the observed data; these estimated coefficients can be used not only for the inference on the effect of the covariate's effect, but also for predicting survival time in the context of the covariate's specific level.

Although estimating the regression coefficients of the Cox proportional hazards model helps to understand the characteristics of the hazard function, the proportional hazards assumption that the hazard ratio of any level of the predictor is constant over time, must be premised. An accelerated failure time model (Wei, 1992) accounting for the relationship between the log-transformed survival time and predictors in the presence of the error term followed from a specific probability distribution can be an alternative to the Cox proportional hazards model in the survival analysis. However, it also requires the premised belief called the accelerated failure time assumption that any level of predictor additively effects on the log survival time. Likewise these two typical models, because most of survival models tend to accompany with the model-based assumptions, the issue for predicting survival time depends on the model-based assumption. The performance of the prediction for the survival time may be inaccurate if the assumed model is incorrect in reality. Following the work of Wang et al. (2016), a support vector hazard machine (SVHM) which estimates the hyperplane maximizing the margin to classify the censored or uncensored observations at each survival time can be used for the prediction on the survival time without any model-based assumptions.

When the dimension of the covariates is large, many statistical models may also suffer from difficulties, such as computational stability, noise accumulation, and variable selection problems (Clarke et al., 2009). The regularized solution minimizing the objective function comprising the empirical risk and penalty terms has great advantages in that it leads to numerical stability and avoids the problem of overfitting (Tibshirani, 1997). However, this is an insufficient approach when the dimensions of the covariates are extremely large. Various feature screening methods with specific modeling assumptions have been suggested to filter the massive number of non-informative covariates, which have complemented regularized methods. Fan and Lv (2010) recommended a two-stage approach that screens out the non-informative covariates in the first stage of variable selection and uses the regularization method in the second stage. However, these existing variable selection methods may not be optimal when prediction and variable selection for time-to-event data must be achieved simultaneously. Specifically, if the negative partial log-likelihood

loss function is replaced by another loss function: Owing to the prediction purpose, the approach that estimates the optimal decision function under the modified loss function may conflict with the important covariates selected under the partial likelihood framework. However, if the aforementioned approach can estimate the optimal decision function under the modified loss function and find clues for the important covariates from this decision function, it would be more desirable and accessible to simultaneously achieve the two objectives of prediction and variable selection.

This study aims to predict the survival time and develop a variable selection method for the time-to-event dataset. We use the SVHM framework with the two different weights to estimate the time-dependent intercept and time-independent risk score after considering the observed time-to-event data as dichotomized counting processes. Then, we predict the survival time of the event based on the pair including the ranks of the estimated risk scores estimated and observed survival time, Finally, we measure the contribution of each marginal covariate effect to the optimal decision function, through the gradient information (He et al., 2021; Park & Park, 2021; Xia, 2007; Xia et al., 2002; Fukumizu & Leng, 2014). The selection method based on the gradient information basically assumes that the corresponding partial derivative of the optimal decision function must be zero if a particular covariate exhibits a negligible effect on the survival outcome. Following this belief, we computed the optimal decision function's partial derivatives for each marginal covariate producing the time-independent risk score to conduct the variable selection.

We consider that there are several contributions in terms of statistical point of views. While the existing SVHM can not conduct the variable selection, our proposed method contributes to adding the role of variable selection to the SVHM method, which yields the unified framework in the survival analysis. Specifically, the proposed method can differ from the conventional prediction and variable selection methods in that it does not requires any premised belief such as the proportional hazards or the accelerated failure time assumptions. Also, the predicted survival time and selected variables from our proposed method can be more systematically consistent results because these results are established within the identical loss function framework, within the same kernel choice, and within one stage estimation compared with the existing two stages models. For instance, the survival time predicted by the hinge loss function in the SVHM and the important variables selected by the negative partial likelihood loss function in the Cox model seems to be somewhat inconsistent. As another aspect, we propose the inverse probability weight to deal with the unbalanced time-to-event data, which provides the possibility being able to incorporate the current suggested approach with the various weighted models.

The remainder of this paper is organized as follows. In Sect. 2, we introduce relevant notations and fundamental results of the SVHM, and variable selection approaches. In Sect. 3, we perform quantitative studies using the proposed approach and provide illustrative example based on real data in Sect. 4. We present concluding remarks in Sect. 5.

2 Methodology

2.1 Support vector hazards machine

We start with the mathematical notation and fundamental concept of the SVHM (Wang et al., 2016) in this subsection. Let a random vector $X = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$ be a set of predictors, and random variables $T^* \in \mathbb{R}^+$ and $C \in \mathbb{R}^+$ be the survival time and censoring time, respectively. The observed time and censoring indicator are defined as $T = \min(T^*, C)$ and the $\Delta = I(T^* \leq C)$, respectively, where $I(\cdot)$ is an indicator function and Δ is the event indicator. Assumedly, the survival time is independent of the censoring time given by predictors X. We observed a sample of n subjects given by $\{(X_i, T_i, \Delta_i) : i = 1, \ldots, n\}$, where X_i, T_i and Δ_i denote a p-dimensional vector of covariates, observed time, and censoring indicator of the ith individual, respectively. Let $N_i(t) = \Delta_i I(T_i \leq t)$ be the counting process and $Y_i(t) = I(T_i \geq t)$ be the at-risk process for the ith individual for any $t \in \mathcal{T} = [0, \tau]$, where τ denotes the stopping time. $dN_i(t)$ is the jump size of the counting process in a small time interval $[t, t + dt]$. Thus, $dN_i(t) = 1$ if $T_i \in [t, t + dt]$, and $dN_i(t) = 0$ otherwise. Let $t_{(1)} \leq t_{(2)} \leq \cdots, \leq t_{(q)}$ be an order statistics for q distinct event times obtained from the observed dataset, where it is assumed that there are no ties in the event times.

We defined the dichotomized variable for each subject and event time as $\delta N_i(t_{(j)}) = 2(N_i(t_{(j)}) - N_i(t_{(j)}-)) - 1$, which takes the value of 1 if the survival time of the ith subject is observed, and -1 otherwise.

We considered the time-dependent risk score $f_0(t, X)$, where $f_0 : \mathcal{T} \times \mathbb{R}^p \to \mathbb{R}$ is a nonparametric smoothed function of the covariates at a specific time. Suppose that such a general risk score comprises the intercept term, $\mu : \mathcal{T} \to \mathbb{R}$, as a function of time, and the nonparametric risk score term, $f : \mathbb{R}^p \to \mathbb{R}$ as a function of the covariates, that is, $f_0(t, X) = \mu(t) + f(X)$. We used the time-dependent risk score to predict whether the corresponding subject experienced a failure event at the next immediate time. Specifically, when the ith subject is still contained in the risk set at time t, we predicted the ith subject to have an event if $f_0(t, X) \geq 0$, or not to have the event if $f_0(t, X) < 0$. When a symmetric positive definite kernel $k(\cdot, \cdot) : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is available, we can construct a feature map with the kernel, that is, $\phi(X) = k(X, \cdot)$ such that $\phi(\cdot) : \mathbb{R}^p \to \mathcal{H}$. When endowing the inner product between the feature maps as the value of the kernel, $k(X, X') = \langle \phi(X), \phi(X') \rangle_{\mathcal{H}}$, we can establish a unique RKHS, denoted by $(\mathcal{H}, \langle \cdot, \cdot \rangle)$, such that the reproducing property $f(X) = \langle f, k(\cdot, X) \rangle$ holds for all $f \in \mathcal{H}$, and $X \in \mathbb{R}^p$.

The optimal separating hyperplane between subjects with and without the event at each time is the hyperplane that can create the largest margin between the two classes. Following Wang et al. (2016), we can express such an optimization problem as follows:

$$\begin{aligned} \min_{\mu, f} \ & \frac{1}{2}\|f\|_{\mathcal{H}}^2 + \lambda^{-1} \sum_{i=1}^{n} \sum_{j=1}^{q} w_i(t_j) Y_i(t_j) \xi_i(t_j), \\ \text{subject to } \ & Y_i(t_j) \delta N_i(t_j) \big\{ \mu(t_j) + f(X_i) \big\} \geq Y_i(t_j) \big\{ 1 - \xi_i(t_j) \big\}, \\ & Y_i(t_j) \xi_i(t_j) \geq 0, \\ \text{for } \ & i = 1, \ldots, n, \quad j = 1, \ldots, q, \end{aligned} \tag{1}$$

where the slack variable $\xi_i(t_j)$ allows the prediction of the ith subject on the wrong side of its margin at the jth event time, the weight $w_i(t_j)$ adjusts the imbalance for the dichotomized response of the ith subject at the jth event time, the cost variable C controls the total sum of the weighted slack variables, and $\|f\|_{\mathcal{H}}$ denotes the norm in the RKHS. Problem (1) is a convex optimization problem with inequality constraints, and using Lagrange multipliers leads to the primal function

$$L_p = \frac{1}{2}\|f\|_{\mathcal{H}}^2 + \lambda^{-1}\sum_{i=1}^{n}\sum_{j=1}^{q} w_i(t_j)Y_i(t_j)\xi_i(t_j) - \sum_{i=1}^{n}\sum_{j=1}^{q}\gamma_{ij}Y_i(t_j)\xi_i(t_j)$$
$$- \sum_{i=1}^{n}\sum_{j=1}^{q}\alpha_{ij}\left[Y_i(t_j)\delta N_i(t_j)\left\{\mu(t_j)+f(X_i)\right\} - Y_i(t_j)\left\{1-\xi_i(t_j)\right\}\right],$$

where it must be minimized with respect to μ, f, and $\xi_i(t_j)$, where $\alpha_{ij} \geq 0$ and $\gamma_{ij} \geq 0$ are the corresponding Lagrange multipliers, and $\|f\|_{\mathcal{H}}$ denotes the norm in the RKHS. By the reproducing property (Aronszajn, 1950), we can represent $f \in \mathcal{H}$ as the evaluation functional, $f(X) = \langle f, \phi(X)\rangle$. Substituting this representation into the primal function and considering the partial derivatives as zeros, we obtain

$$f = \sum_{i=1}^{n}\sum_{j=1}^{q}\alpha_{ij}Y_i(t_j)\delta N_i(t_j)\phi(X_i),$$
$$0 = \sum_{i=1}^{n}\alpha_{ij}Y_i(t_j)\delta N_i(t_j),$$
$$\alpha_{ij}Y_i(t_j) = \lambda^{-1}w_i(t_j)Y_i(t_j) - \mu_{ij}Y_i(t_j), \quad \text{for } i = 1,\ldots,n,\ j = 1,\ldots,q.$$

By substituting the above results into the primal function, we obtain the dual function

$$L_D = \sum_{i=1}^{n}\sum_{j=1}^{q}\alpha_{ij}Y_i(t_j) - \frac{1}{2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{j=1}^{q}\sum_{j'=1}^{q}\alpha_{ij}\alpha_{i'j'}Y_i(t_j)Y_{i'}(t_{j'})\delta N_i(t_j)\delta N_{i'}(t_{j'})k(X_i,X_{i'}),$$

for which it must be maximized with respect to the multipliers α_{ij} subject to $0 \leq \alpha_{ij} \leq \lambda^{-1}w_i(t_j)$ and $\sum_{i=1}^{n}\alpha_{ij}Y_i(t_j)\delta N_i(t_j) = 0$ for $i = 1,\ldots,n$ and $j = 1,\ldots,q$. The Karush–Kuhn–Tucker includes the constraints

$$\left[Y_i(t_j)\delta N_i(t_j)\left\{\mu(t_j)+f(X_i)\right\} - Y_i(t_j)\left\{1-\xi_i(t_j)\right\}\right] \geq 0$$
$$\alpha_{ij}\left[Y_i(t_j)\delta N_i(t_j)\left\{\mu(t_j)+f(X_i)\right\} - Y_i(t_j)\left\{1-\xi_i(t_j)\right\}\right] = 0,$$
$$\gamma_{ij}\xi_i(t_j) = 0 \quad \text{for } i = 1,\ldots,n\ j = 1,\ldots,q$$

characterizing the optimal solution for the above primal and dual objective functions. The solution for f has the form $\widehat{f} = \sum_{i=1}^{n}\sum_{j=1}^{q}\widehat{\alpha}_{ij}Y_i(t_j)\delta N_i(t_j)\phi(X_i)$, which yields the estimator for the smoothed risk score, given by

$$\widehat{f}(\mathbf{x}) = \left\langle \sum_{i=1}^{n}\sum_{j=1}^{q} \widehat{\alpha}_{ij} Y_i(t_j)\delta N_i(t_j)\phi(X_i), \phi_k(\mathbf{x}) \right\rangle = \sum_{i=1}^{n}\sum_{j=1}^{q} \widehat{\alpha}_{ij} Y_i(t_j)\delta N_i(t_j) k(X_i, \mathbf{x}), \tag{2}$$

where $\mathbf{x} \in \mathbb{R}^p$ denotes a test point for any covariate. Complementary slackness is summarized as the following equations

$$\begin{aligned} &\alpha_{ij}\left[Y_i(t_j)\delta N_i(t_j)\left\{\mu(t_j) + f(X_i)\right\} - Y_i(t_j)\left\{1 - \xi_i(t_j)\right\}\right] = 0, \\ &(\lambda^{-1} w_i(t_j) - \alpha_{ij})\xi_i(t_j) Y_i(t_j) = 0 \quad \text{for } i = 1, \ldots, n \ \ j = 1, \ldots, q, \end{aligned}$$

These conditions allow for the ith subject to be well separated at the jth time if $\alpha_{ij} = 0$, $\xi_i(t_j) = 0$, and $Y_i(t_j) = 1$ are satisfied, whereas they allow for the ith subject at the jth time to be contained in the support vectors lying on the edge of the margin when $0 < \alpha_{ij} \leq \lambda^{-1} w_i(t_j)$, $\xi_i(t_j) = 0$, and $Y_i(t_j) = 1$ are satisfied, and lying on the wrong side of the margin when $\alpha_{ij} = \lambda^{-1} w_i(t_j)$, and $Y_i(t_j) = 1$ are satisfied. According to the complementary slackness, the estimator for f in (2) is characterized by the support vectors across all subjects and event times. The optimal time-varying intercept is also estimated by using complementary slackness. At the jth time point t_j, all subjects lying on the edge of the margin satisfy the condition $Y_i(t_j)\delta N_i(t_j)\left\{\mu(t_j) + f(X_i)\right\} - Y_i(t_j) = 0$, which is equivalent to $\widehat{\mu}(t_j) = 1/\delta N_i(t_j) - f(X_i)$. In practice, we used the average of all subjects at the time.

2.2 Gradient-based variable selection

The sparsity assumption that only a few covariates have been associated with the survival outcome is more practical than relating all covariates with the outcome in the real world. Such a sparsity assumption has been widely employed within various models for the last two decades in the variable selection areas. In this subsection, we intend to develop a variable selection approach for the SVHM within the sparsity assumption. Suppose that function f is continuously differentiable. If there exists a covariate strongly related to the risk score f associated with the survival time, guessing that a small change in the value of the corresponding covariate will cause a large change in the value, the risk score is not unreasonable. Following the work of (Park & Park, 2021), the partial derivative of f with respect to the jth predictor x_k given by

$$\frac{\partial f(X)}{\partial x_j} = \frac{\partial f(x_1, \ldots, x_p)}{\partial x_j} \equiv \partial_j f$$

can serve as the aforementioned criterion by capturing the relative importance of the jth predictor at a fixed point. Suppose that the inner product in the reproducing kernel Hilbert space (RKHS) is computed with the partial differential operator, and the partial differential operator is a self-adjoint operator. It follows from the reproducing property that:

$$\partial_j f = \frac{\partial}{\partial x_j}\langle f, k(\cdot, X)\rangle_{\mathcal{H}} = \langle f, \frac{\partial}{\partial x_j}k(\cdot, X)\rangle_{\mathcal{H}} = \langle \frac{\partial}{\partial x_j}f, k(\cdot, X)\rangle_{\mathcal{H}}$$

holds for all $f \in \mathcal{H}$ and $X \in \mathbb{R}^p$. From the Cauchy–Schwarz inequality, we have an upper bound:

$$\langle f, \frac{\partial}{\partial x_j}k(\cdot, X)\rangle_{\mathcal{H}} \leq \|f\|_{\mathcal{H}}\|\frac{\partial}{\partial x_j}k(\cdot, X)\|_{\mathcal{H}}.$$

Thereafter, we observe that the partial derivative of f with respect to any predictor belongs to the RKHS when the partial derivative of the reproducing kernel is bounded, which reveals that the reproducing kernel remains an essential tool for obtaining the partial derivative of f. The above partial derivative depends on the fixed point x_j, which might make it difficult to understand the importance of covariate X_j across the entire sample space. If the marginal probability density function for X, denoted by $P_X(x)$, is available, the L_2 norm with respect to $P_X(x)$ is given by

$$\|\partial_j f\|^2_{L_2(P_X)} = \int_{\mathcal{X}} (\partial_j f)^2 dP_X(x). \tag{3}$$

provides a better understanding of the importance of covariate X_j across possible realizations. If the jth predictor does not have a relationship with the risk score associated with the survival outcome, then the L_2 norm in (3) tends to be close to zero for any fixed point x_j. Let

$$\mathcal{M} = \{1 \leq j \leq p \,:\, \|\partial_j f\|^2_{L_2(P_X)} \neq 0\}$$

be a true set of indices containing the important predictors, and $s = |\mathcal{M}|$ be the number of elements in the true set, for which it is assumed that s is a smaller integer compared with the sample size n, and dimension p. We use the estimator $\widehat{f}$ mentioned in Sect. 2.1 for the risk score f to compute the empirical partial derivatives for each predictor and the empirical probability measure as the counter part of the marginal probability measure of the predictor to obtain the index set estimating the true sparse set. Taking the partial derivative of $\widehat{f}$ mentioned in (2) with respect to the jth predictor,

$$\partial_j \widehat{f} = \sum_{i=1}^{n}\sum_{j=1}^{q} \widehat{\alpha}_{ij} Y_i(t_j)\delta N_i(t_j)\frac{\partial k(X_i, \mathbf{x})}{\partial x_j}$$

and applying the empirical probability to the L_2 norm in (3), we compute the estimator for the magnitude of the risk score with respect to the jth expressed as:

$$\begin{aligned}\|\partial_j \widehat{f}\|_n^2 &= \frac{1}{n}\sum_{l=1}^{n}\left(\frac{\partial \widehat{f}(\mathbf{x}_l)}{\partial x_j}\right)^2 = \frac{1}{n}\sum_{l=1}^{n}\left(\langle \frac{\partial}{\partial x_j}\widehat{f}, k(\cdot,\mathbf{x}_l)\rangle_{\mathcal{H}}\right)^2 \\ &= \frac{1}{n}\sum_{l=1}^{n}\left(\left\langle \sum_{i=1}^{n}\sum_{j=1}^{q}\widehat{\alpha}_{ij}Y_i(t_j)\delta N_i(t_j)\frac{\partial k(\cdot,X_i,)}{\partial x_j}, k(\cdot,\mathbf{x}_l)\right\rangle_{\mathcal{H}}\right)^2 \\ &= \frac{1}{n}\sum_{l=1}^{n}\left(\sum_{i=1}^{n}\sum_{j=1}^{q}\widehat{\alpha}_{ij}Y_i(t_j)\delta N_i(t_j)\frac{\partial k(X_i,\mathbf{x}_l)}{\partial x_j}\right)^2.\end{aligned} \tag{4}$$

We utilized the empirical norm of the partial derivative for the risk score presented in (4) to select informative variables,

$$\widehat{\mathcal{M}}(\gamma_n) = \{1 \le j \le p : \|\partial_j \widehat{f}\|_n^2 \text{ is amongst the first } \gamma_n \text{ largest of all values}\}, \tag{5}$$

where γ_n is the predefined threshold value. Numerous literature on the variable selection issues (Fan & Lv, 2010; Jeong et al., 2023) have popularly used the value of γ_n as $n-1$ or $[n/\log(n)]$, and we use this value as our predefined threshold value, where $[a]$ denotes the greatest integer value less than a.

2.3 Properties of gradient based kernel selection

Let $\mathcal{X} \subset \mathbb{R}^p$ be a boundedly connected set including all possible value of random vector X and $\mathcal{H}_k$ be a RKHS induced by the kernel k. Let $\mathcal{M} = \{1 \le j \le p \mid \|\partial_j f^*\|_{P_X}^2 > 0\}$ and $\widehat{\mathcal{M}}(\rho) = \{1 \le j \le p \mid \|\partial_j \widehat{f}\|_n^2 > \rho\}$ be the true set of indices including the important predictors and the estimated set of indices based on the the empirical norm of the partial derivative function for each predictor, where ρ is pre-defined threshold value. Since there is one-to-one relationship between γ_n in (5) and ρ in $\widehat{\mathcal{M}}(\rho)$, we mainly focus on γ_n in this subsection. Define the empirical risk

$$\mathcal{R}_n(f,\mu) = \min_{\mu,f}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{q}Y_i(t_j)\left[1-(\mu(t_j)+f(X_i))\delta N_i(t_j)\right]_+ \tag{6}$$

where $[a]_+ = \max\{0,a\}$. Then, define the empirical solution as

$$(\widehat{\mu},\widehat{f}) = \arg\min_{\mu,f}\mathcal{R}_n(f,\mu) + \frac{1}{2}\lambda\|f\|^2$$

and the true decision function as $(\mu, f^*) = \arg\min_{\mu,f}\mathcal{R}(f,\mu)$, respectively, where

$$\begin{aligned}\mathcal{R}(f,\mu) = E\left[\int Y(t)[1-\mu(t_j)-f(X))]_+ dN(t)\right] \\ + \int \frac{E(Y(t))[1+\mu(t_j)+f(X))]_+}{E(Y(t))}E(dN(t)).\end{aligned}$$

To conveniently develop theoretical justification, instead of using the risk functions aforementioned $\mathcal{R}_n(f,\mu)$ and $\mathcal{R}(f,\mu)$, we use the profile risk function excluding the effect of the time-dependent function μ. Following Wang et al. (2016), the profile risk can be derived as

$$\mathcal{PR}(f) = \mathcal{R}(f,\mu^*) = E\left[\Delta \frac{\tilde{P}I(\tilde{Y} \geq Y)[2 - f(\tilde{X}) - f(X)]_+}{\tilde{P}I(\tilde{Y} \geq Y)}\right],$$

where $\mu^* = \min_\mu \mathcal{R}(f,\mu)$ and $\tilde{P}$ is the probability measure with respect to $(\tilde{Y},\tilde{X},\tilde{\Delta})$.

The following conditions are required to establish the properties for our method.

(A1) For each $f \in \mathcal{H}_k$, f is continuous and continuously differentiable.

(A2) There exists a constant τ such that for all $j \in \{1,2,\ldots,p\}$

$$\sup_{X\in\mathcal{X}} \|k(X,\cdot)\| \leq \tau \quad \text{and} \quad \sup_{X\in\mathcal{X}} \|\partial_j k(X,\cdot)\| \leq \tau.$$

(A3) There exists a constant c_1 such that for $\widehat{f}, f^* \in \mathcal{H}_k$

$$\|\widehat{f} - f^*\| \leq c_1 |\mathcal{PR}(\widehat{f}) - \mathcal{PR}(f^*)|$$

and exists a constant c_2 such that for some positive δ

$$P(|\mathcal{PR}(\widehat{f}) - \mathcal{PR}(f^*)| \geq c_2 n^{-(\frac{q}{q+1})} \leq \frac{\delta}{2}.$$

(A4) There exists a positive constant $\eta < \frac{q}{q+1}$ such that for some positive constant κ_2

$$\min_{j\in\mathcal{M}^*} \|\partial_j f^*\|^2 > \kappa_2 (\log p) n^{-\eta}.$$

Condition (A1) enables us to exclude the discontinuously differentiable functions strongly associated with the survival time. For instance, there may exit a piecewise constant function defined as the survival time on some region of the jth predictor and zeros on the other regions though $\|\partial_j f\| = 0$. By the condition (A1), we can be more confident that the jth predictor is not important when observing $\|\partial_j f\| = 0$. Condition (A2) provides the boundedness of the reproducing kernel and the partial derivative function with the respect to all marginal predictors. Condition (A2) is always satisfied for the typical kernels such as the Guassian, polynomial and Sobolev kernels defined on a bounded domain. The first assertion of the condition (A3) implies that the discrepancy between the empirical estimator and true solution has an upper bound consisting of the discrepancy between the corresponding profiled risk functions. The second assertion of the condition (A3) assumes the specific probability of the difference between the profiled risk functions evaluated at the estimator and true function, respectively. The rate of the convergence for $\mathcal{PR}(\widehat{f}) - \mathcal{PR}(f^*)$ to the zero is known to be $O_p(n^{-\frac{q}{q+1}})$ when taking $\lambda = n^{-q/(q+1)}$ with some other conditions as introduced in Wang et al. (2016) or Remark 1, where $q = 1/(4/\xi + 1)$ and $\xi \in (0,2)$.

We adopt this result as the basic condition to develop the theoretical justification for the proposed method. Consequently, condition (A3) allows us to approximate the probability of the tails of $\|\widehat{f} - f^*\|$ although it seems to be somewhat strong assumption. Condition (A4) assumes that the true gradient function for the important predictors contains sufficient information as the value being able to discriminate from the uninformative predictors.

Remark 1 For the convergence rate of $\mathcal{PR}(\widehat{f}) - \mathcal{PR}(f^*)$, the literature (Wang et al., 2016) assumes that λ and σ go to zero, that $n\lambda\sigma^{p(2/\xi-1/2))}$ goes to infinity, and $E[Y(t^*)|X]$ is bounded away from zero, where t^* is the stopping time, and that σ is a scale factor in the Gaussian kernel defined as $k(z_1, z_2) = \exp\{-\|z_1 - z_2\|^2/\sigma\}$.

We establish two results as the justification of the proposed selection method.

Proposition 1 *Suppose that assumptions* 1–3 *are satisfied. Then with the probability at least* $1 - \delta$, *there holds*

$$\max_{1\le j\le p}\left|\|\partial_j\widehat{f}\|_n^2 - \|\partial_j f^*\|_{P_X}^2\right| \le \kappa(\log p)n^{-\frac{q}{q+1}} \tag{7}$$

where $q = 1/(4/\xi + 1)$, $\xi \in (0, 2)$, *and* κ *is some positive constant.*

Proposition 1 lays out the rate of the convergence for the maximum discrepancy between the empirical norm for the partial derivative of the estimated SVHM and the L_2 norm for the partial derivative of the true function for all predictors. This result is important because it implies that $\|\partial_j\widehat{f}\|_n^2$ converges to $\|\partial_j f^*\|_{P_X}^2$ and contributes to establish the asymptotic selection consistency as the followings.

Proposition 2 *Suppose that assumptions* 1–4 *are satisfied and Proposition* 1 *holds. Let* $\rho = \frac{\kappa}{2}(\log p)n^{-\eta}$. *Then we have*

$$P(\widehat{\mathcal{M}}(\rho) = \mathcal{M}^*) \longrightarrow 1 \tag{8}$$

as n goes to infinity.

Proposition 2 demonstrate the asymptotic selection consistency of our proposed gradient based selection method. With the condition (A4), the specific rate of the selection consistency is determined by the result in Proposition 1. Thereby, the proposed method keeps the informative predictors and filter out the uninformative predictors in the selection procedure with the overwhelming probability.

2.4 Prediction, weight and tuning parameters

Following the work of Wang et al. (2016), we predicted the survival time for the risk score evaluated for the specific predictors. Specifically, we first sorted the observed survival time $t_{(q)} \ge t_{(q-1)} \ge \cdots, \ge t_{(1)}$ in decreasing order. Thereafter, we computed

the risk score $\widehat{f}(\mathbf{x}_{i_k})$ for the corresponding subjects with the observed event time, $k = 1, 2, \ldots, q$, and sorted the risk scores $\widehat{f}(\mathbf{x}_{(i_1)}) \leq \widehat{f}(\mathbf{x}_{(i_2)}) \leq \cdots, \leq \widehat{f}(\mathbf{x}_{(i_q)})$ in increasing order. We redefined the reference pair for the prediction of survival time given by $\{(\widehat{f}(\mathbf{x}_{(i_1)}), t_{(q)}), (\widehat{f}(\mathbf{x}_{(i_2)}), t_{(q-1)}), \ldots, (\widehat{f}(\mathbf{x}_{(i_q)}), t_{(1)})\}$. as $\{(\widehat{f}(\mathbf{x}_{(i_k)}), \widetilde{t}_{(k)}) : k = 1, \ldots, q\}$ We predicted the survival time for the new observation $\mathbf{x}$, given by $\widehat{T} = \sum_{k \in N(h)} \widetilde{t}_{(k)} / |N(h)|$, where $N(h)$ is the set of pairs $(\widehat{f}(\mathbf{x}), \widehat{f}(\mathbf{x}_{(i_k)}))$ approximately distance h depart for $k = 1, \ldots, q$, and $N(h)$ denotes the number of pairs in $N(h)$. For computational simplicity, we selected the first three closest observed survival times to replace the value of h.

In classification problems, an imbalance of the response variables reduces the performance of the classifier. Although the introduced nonparametric risk score has played a role in classifying whether or not an event occurs at a specific time, the situation is extremely similar to the situation of unbalanced data in the classification problem, when considering the general situation that there is usually a single event at a specific time, while the rest of the observations remain at risk, which means that the event is free at a specific time. Therefore, we need subject- and time-specific weight to balance the occurrence and nonoccurrence of the event. We considered two types of weights expressed as

$$w_i(t_j) = I(\delta N_i(t_j) = 1) \frac{\sum_{i=1} Y_i(t_j) - 1}{\sum_{i=1} Y_i(t_j)} + I(\delta N_i(t_j) = -1) \frac{1}{\sum_{i=1} Y_i(t_j)}$$

proposed by Wang et al. (2016), and

$$w_i(t_j) = I(\delta N_i(t_j) = 1) \frac{1}{\exp\{-G(t_j)\}} + I(\delta N_i(t_j) = -1)$$

proposed by Yang et al. (2021), where

$$G(t_j) = \int_0^{t_j} \frac{\sum_{i=1}^n d\widetilde{N}_i(s)}{\sum_{i=1}^n Y_i(s)},$$

and $\widetilde{N}_i(t) = (1 - \Delta_i) I(T_i \leq t)$ presents the counting process for any $t \in \mathcal{T} = [0, \tau]$. Notably, the first type of weight increases the occurrence of the event up to the size of the risk set at a specific time, and reduces the number of nonoccurrence events to one at the time. The second weight is an inverse probability-of-censoring weight that enables adjustment of the imbalance in the number of event occurrences by increasing it to the expected number of trials wherein the survival time is observed at a specific time.

For the choice of the tuning parameters including the cost variable λ^{-1}, and scale factor σ associated with the kernel, the grid search method can be used to find the optimal tuning parameter estimates $\widehat{\lambda}_{cv}^{-1}$ and $\widehat{\sigma}_{cv}$ that minimizes the k-fold cross-validation errors defined as the empirical root mean squared error as mentioned in (10), after splitting the dataset into training and test datasets. However, because the grid search method needs the expensive cost in terms of the computational time, it is appropriate

for the data sets with the low dimensional predictors or the high censoring rate. For the data sets with the high dimensional predictors and the low censoring rate, a value of the decreasing sequence less than one can be used as $\widehat{\lambda}$ because $\lambda \to 0$ is assumed in Remark 1, where the results depending on these values did not yield dramatic change. For instance, we observed that $\widehat{\lambda}^{-1} = 1000$ and $\widehat{\lambda}^{-1} = 100$ leads to the identical support vectors in the simulation study.

3 Simulation

We conducted four sets of numerical simulations to examine the finite-sample performance of the prediction and variable selection for our proposed method at different censoring rates, sample sizes, and the number of covariates. In the first scenario, we generated p-dimensional random predictors $X_i = (x_{i1}, \ldots, x_{ip})^T$ using a uniform distribution. Specifically, we independently generated the first five predictors $x_{i1}, x_{i2}, \ldots, x_{i5} \sim \text{Uniform}(0, 5)$, and the other predictors $x_{i6}, x_{i7}, \ldots, x_{ip} \sim \text{Uniform}(0, 1)$. We set the true proportional hazards regression coefficients as $\beta = (1, 1, 1, 1, 1, 0, 0, \cdots, 0)^T$ for the random predictors, that is, we set the true set of indices indicating the specific predictors associated with the outcome variable as $\mathcal{M} = \{1, 2, 3, 4, 5\}$ for the variable selection. After we generated the true survival probability $U_i \sim \text{Uniform}(0, 1)$ for each ith observation, we generated the true survival time of the event with probability, that is,

$$T_i^* = \frac{1}{\lambda_0^{1/a}} \left[\frac{-\log(U_i)}{\exp\{\beta^T X_i\}} \right]^{1/a}, \tag{9}$$

where λ_0 denotes the baseline hazard function, and $\lambda_0 = 0.25$, and $a = 1$ were used.

The censoring time of the ith observation was independently generated from the exponential distribution, $C_i \sim \text{Exponential}(\lambda_i)$, where $\lambda_i = \frac{cr}{1-cr}\lambda_0 / \exp\{\beta^T X_i\}$ for censoring rate, denoted by cr. We generated the observed survival time $T_i = \min(T_i^*, C_i)$ and the observed censoring indicator $\Delta_i = I(T_i^* \le C_i)$ for each ith observation. Then, we obtained sample of n observations expressed as $\{(X_i, T_i, \Delta_i) : i = 1, \ldots, n\}$ as the dataset for the simulation. In the second scenario, we generated the important predictors $x_{i1}, x_{i2}, \ldots, x_{i5} \sim \text{Uniform}(0, 5)$, and other unimportant predictors $x_{i6}, x_{i7}, \ldots, x_{ip} \sim \text{Uniform}(0, 1)$, respectively. We defined the logarithm of the hazard ratio as a polynomial function given by

$$g(X_i) = x_{i1} \cdot x_{i2} \cdot x_{i3} + x_{i4}^2 + 5x_{i5}$$

and generated the true survival time, whereas the polynomial function $g(X_i)$ was substituted for the linear function term $\beta^T X_i$ in (9), where an identical value was used as the baseline hazard function. We used the correlated predictors in the third scenario. Specifically, the predictors were generated from $N_p(\mu, \Sigma)$, where $\mu = (3 \cdot 1_5^T, 0_{p-5}^T)^T$ and $\Sigma = \Sigma_5 \oplus \Sigma_{p-5}$, with $\Sigma_5 = 0.5 \cdot I_5 + 0.5 \cdot 1_5 1_5^T$ and $\Sigma_{p-5} = 0.5 \cdot I_{p-5} + 0.5 \cdot 1_{p-5} 1_{p-5}^T$ for which $\oplus$ denotes the direct sum, and 0_k is a

$k \times 1$ column vector of zeros. The corresponding survival times were generated from the model (9) in the setting with $a = 2$ and $\lambda = 0.1$. In the fourth scenario, we generated the first five predictors $x_{i1}, x_{i2}, \ldots, x_{i5} \sim$ Uniform$(0, 5)$, and the other predictors $x_{i6}, x_{i7}, \ldots, x_{ip} \sim$ Uniform$(0, 1)$, and used the following nonlinear function

$$g(X_i) = x_{i1} \cdot x_{i2} + 5cos(x_{i3}) + x_{i4}^2 + 3x_{i5}$$

as the logarithm of the hazard ratio.

Following a similar procedure, we generated censoring time, observed survival time, and censoring indicator variables. For each simulation scenario, we considered sample sizes of 50 and 100 (denoted by n), censoring rates of 20%, 40%, and 60% (denoted by cr), while we changed the number of predictors with various settings of 50, 200, and 1000 (denoted by p). Additionally, for each scenario, we simulated 100 datasets, for which each dataset consisted of training data with a sample size of n and test data with a sample size of 500 (denoted by n^{test}). Notably, the first five covariates were set as the important covariates associated with variable selection for all scenarios.

For each new observation, denoted by $(X_i^{\text{test}}, T_i^{\text{test}}, \Delta_i^{\text{test}})$ in the test dataset, we predicted survival time of the event, $\widehat{T}_i^{\text{test}}$ after computing the predicted risk score $\widehat{f}(X_i^{\text{test}})$ and estimating the smoothed risk score $\widehat{f}(\cdot)$ based on the training dataset for each iteration. For the prediction accuracy of the proposed method, we considered two performance measures: the empirical root mean squared error (RMSE) term

$$RMSE = \sqrt{\sum_{i=1}^{n^{\text{test}}} \Delta_i^{\text{test}} (T_i^{\text{test}} - \widehat{T}(\widehat{f}_i^{\text{test}}))^2 \Big/ \sum_{i=1}^{n^{\text{test}}} \Delta_i^{\text{test}}} \tag{10}$$

and the empirical concordance index term (CCI)

$$CCI = \frac{\sum_{i=1}^{n^{\text{test}}} \Delta_i^{\text{test}} \left\{ \sum_{T_j^{\text{test}} > T_i^{\text{test}}} I(\widehat{f}(X_j^{\text{test}}) > \widehat{f}(X_i^{\text{test}})) \right\}}{\sum_{i=1}^{n^{\text{test}}} \Delta_i^{\text{test}} \left\{ \sum_{T_j^{\text{test}} > T_i^{\text{test}}} 1 \right\}},$$

respectively. This implies that the proportion of pairs wherein the ranking of predicted values is accurately arranged among all comparable pairs of time-to-event data. Furthermore, we computed the empirical partial derivative of the smoothed risk score $\widehat{f}(X)$ with respect to all marginal covariates, and, thereby, obtained the estimated set of indices $\widehat{\mathcal{M}}$ for the true important subset $\mathcal{M}$, as mentioned in (5). For the accuracy associated with the proposed method's variable selection, we considered three performance measures: the empirical true positive rate, defined as $TPR = |\mathcal{M} \cap \widehat{\mathcal{M}}|/|\mathcal{M}|$, the empirical false positive rate, defined as $FPR = |\mathcal{M}^c \cap \widehat{\mathcal{M}}|/|\mathcal{M}^c|$, and the number of elements for the smallest set of indices $\widehat{\mathcal{M}}$ being able to include all the true important covariates (denoted by $\widehat{d}$), where $\mathcal{A}^c$

denotes the complement set of $\mathcal{A}$ and $|\mathcal{A}|$ denotes the number of elements contained in a set $\mathcal{A}$.

For our proposed methods, we used the SVHM with the weight based on the risk process and weight based on the censoring distribution's inverse survival probability, as described in Sect. 2.3, denoted by Models (1) and (2), respectively. For comparison, we employed support vector regression (SVR) method introduced in (Khan & Zubek, 2008) and support vector machine based on the ranking (SVMR) introduced in Van Belle et al. (2011) for the time-to-event dataset, denoted by Models (3) and (4), respectively. Notably, the approach in Khan and Zubek (2008) directly predicted the survival time by applying the variant epsilon-insensitive hinge loss function depending on the censoring type, whereas Van Belle et al. (2011) predicted the survival time by voting in the nearest observed survival times, which were computed based on the risk score and its ranking as mentioned in Sect. 2.3.

We focused on using the Gaussian radial basis function (RBF) kernel, $k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2/\sigma)$ for all methods. As the gradient of the smoothed risk score is represented by the linear combination of the selected kernel's partial derivative, we had to choose a kernel that can have a varied values of partial derivatives, and the Gaussian RBF kernel sufficiently satisfies this condition. For the choice of the scale parameter, we used $\sigma = 1/p$ as the fixed value to control noise accumulation due to the high dimensionality of the predictors in the L_2 distance of the exponent. To compute the inverse probability weight, we used the Kaplan–Meier estimates for the survival function of the censoring distribution for each observation.

Tables 1, 2, 3 and 4 present the empirical average values of the total iterations for the prediction performance including the RMSE and CCI for sample size of 100 in all simulation scenarios, where the numbers in the parenthesis indicate the empirical standard deviation. Clearly, the SVHMs with the weight based on the risk process (Model (1)) outperformed the other methods across all cases of $p = 50$, 200, and 1, 000 in terms of RMSE for all simulation scenarios. The weight based on the inverse probability (Model (2)) showed similar or slightly inferior performance compared with Model (1) across the different dimension settings in the linear, polynomial, and nonlinear scenarios with the independent predictors from Tables 1, 2 and 4. Although the performances of the SVR (Model (3)) for RMSE were not better than Model (1) and (2) in Tables 1, 2 and 3 for the first three scenarios, these results were not extremely poor, as much as the results of Model (3) shown in Table 4 for the nonlinear scenario. Interestingly, we found that Model (3) showed good result when the censoring rate and dimension were set to be 60% and $p = 1,000$ in Table 3 for the correlated predictors scenario. The support vector machine based on the censoring type (Model (4)) exhibited unsatisfactory results. We observed that Models (1) and (2) outperformed the other methods across all dimension cases and four different simulation scenarios in terms of CCI. We also see that Model (1) provided more stable results for the different censoring rates and dimension settings. Overall, for all methods, when the sample size increases, the CCI tends to improve, and this measure tends to be worse despite their slight difference when the censoring rate increases. The corresponding results for the sample size of 50 were summarized in the supplementary material.

Table 1 Prediction measures in the first scenario ($n = 100$)

p	Censoring rate	Model (1)		Model (2)		Model (3)		Model (4)	
		RMSE	CCI	RMSE	CCI	RMSE	CCI	RMSE	CCI
50	20%	2.51	0.874	2.77	0.826	2.92	0.375	336.49	0.359
		(22.23)	(0.01)	(22.18)	(0.024)	(6.43)	(0.047)	(235.01)	(0.074)
	40%	2.55	0.867	2.6	0.824	2.12	0.382	406.61	0.352
		(22.47)	(0.014)	(22.23)	(0.028)	(4.96)	(0.047)	(274.31)	(0.083)
	60%	2.57	0.851	2.73	0.821	2.39	0.38	456.36	0.343
		(22.48)	(0.021)	(22.53)	(0.032)	(11.54)	(0.046)	(360.86)	(0.088)
200	20%	0.88	0.867	0.88	0.828	2.81	0.328	145.08	0.396
		(3.38)	(0.013)	(3.04)	(0.023)	(4.09)	(0.043)	(104.55)	(0.059)
	40%	0.75	0.857	0.75	0.821	3.8	0.327	191.44	0.411
		(3.10)	(0.017)	(2.97)	(0.032)	(11.00)	(0.040)	(144.92)	(0.064)
	60%	0.55	0.843	0.83	0.821	2.72	0.319	229.04	0.387
		(2.80)	(0.023)	(3.13)	(0.037)	(6.29)	(0.042)	(182.41)	(0.064)
1000	20%	0.83	0.858	1.57	0.810	1.55	0.271	33.41	0.429
		(0.00)	(0.008)	(0.00)	(0.020)	(0.00)	(0.033)	(0.03)	(0.053)
	40%	0.04	0.842	0.53	0.772	0.71	0.255	56.92	0.403
		(0.00)	(0.017)	(0.00)	(0.038)	(0.00)	(0.031)	(0.04)	(0.052)
	60%	0.60	0.827	0.64	0.795	1.55	0.274	77.89	0.422
		(0.00)	(0.020)	(0.00)	(0.041)	(0.00)	(0.054)	(0.04)	(0.053)

Model (1) is SVHM, Model (2) is KM-inverse weight SVHM, Model (3) is SVR, and Model (4) is SVMR. The values in the table comprise average (standard deviation) (Units of RMSE: 1E−3)

Tables 5, 6, 7 and 8 report the empirical average and the corresponding standard deviation values for the true positive rate, false positive rate, and number of elements for the smallest set of indices containing all the true important predictors for each dimension of the predictors in the four simulation scenarios. We found that the selection accuracy of Model (1) outperformed the others across all different settings in the four simulation scenarios, and that it showed the best performance especially in Tables 5 and 7 for the two linear scenarios including the independent and dependent predictors. While model (2) using the inverse probability weight exhibited desirable results, for which it tended to be slightly lower than that of Model (1) in most cases (Tables 5, 6 and 8), it provided the inferior results compared with Model (3) in Table 7 when the predictors tended to be correlated and high dimensional settings. Model (4) revealed better performance than Model (3) in Table 5, 6, and 8 for the most situations. However, we observed the opposite results associated with the selection performance of Models (3) and (4) in Table 7. From four simulation studies, we confirmed that the proposed Models (1) and (2) can work better in both aspects of prediction and variable selection than the other methods when the time-to-event dataset is generated from a more complex design. The selection results for the sample size of 50 were contained in the supplementary material.

Four panels contained in Fig. 1 depict the trajectories of the time varying intercept, $\widehat{\mu}(t)$ in the SVHM for four different cases, denoted by A ($n = 50, p = 100, cr = 40\%$),

Table 2 Prediction measures in the second scenario ($n = 100$)

p	Censoring rate	Model (1)		Model (2)		Model (3)		Model (4)	
		RMSE	CCI	RMSE	CCI	RMSE	CCI	RMSE	CCI
50	20%	9.63	0.874	9.65	0.906	20.91	0.419	343.93	0.370
		(65.90)	(0.013)	(67.76)	(0.012)	(71.61)	(0.04)	(226.39)	(0.066)
	40%	9.72	0.863	9.55	0.894	18.42	0.419	419.65	0.353
		(68.22)	(0.018)	(67.40)	(0.018)	(68.79)	(0.04)	(320.75)	(0.089)
	60%	9.77	0.844	9.69	0.875	15.25	0.420	440.49	0.352
		(68.48)	(0.025)	(68.43)	(0.027)	(68.11)	(0.042)	(318.18)	(0.08)
200	20%	0.98	0.850	1.08	0.828	9.46	0.385	133.25	0.386
		(5.13)	(0.015)	(5.09)	(0.026)	(15.73)	(0.037)	(102.52)	(0.065)
	40%	0.88	0.841	1.05	0.826	7.85	0.383	193.01	0.400
		(4.97)	(0.02)	(4.94)	(0.028)	(14.46)	(0.036)	(162.88)	(0.055)
	60%	0.83	0.828	1.48	0.825	5.64	0.382	238.81	0.403
		(4.40)	(0.03)	(7.73)	(0.036)	(11.89)	(0.037)	(165.69)	(0.072)
1000	20%	0.01	0.841	0.01	0.786	11.25	0.330	11.81	0.374
		(0.00)	(0.011)	(0.00)	(0.038)	(0.02)	(0.064)	(0.01)	(0.033)
	40%	0.13	0.826	0.13	0.761	0.64	0.330	29.64	0.419
		(0.00)	(0.021)	(0.00)	(0.045)	(0.00)	(0.038)	(0.03)	(0.032)
	60%	1.83	0.803	2.40	0.802	4.23	0.332	82.94	0.412
		(0.01)	(0.021)	(0.01)	(0.025)	(0.01)	(0.041)	(0.07)	(0.076)

Model (1) is SVHM, Model (2) is KM-inverse weight SVHM, Model (3) is SVR, and Model (4) is SVMR. The values in the table comprise average (standard deviation). (Units of RMSE: 1E−3)

B $(n = 50, p = 100, cr = 60\%)$, C $(n = 50, p = 200, cr = 40\%)$, and D $(n = 50, p = 200, cr = 60\%)$, where the horizontal axis denotes the event time and the vertical axis denotes the value of the time varying intercept. We confirmed that these intercepts did not decrease as time increased, which can be interpreted as an increasing tendency of the hazard rates for all cases. The four panels included in Fig. 2 present the scatter plots between the estimated risk score and observed survival time for identical simulation cases, where the horizontal axis denotes event time, and the vertical axis denotes the value of the risk score in each panel. We observed that the risk score tended to decrease when the survival time tended to be longer in all simulation settings.

4 Analysis of real data

To illustrate its usefulness, a real data application was performed using four different prediction models. The data used for the analysis were human gene data collected through an oligonucleotide array in the work (Beer et al., 2002). We were biologically interested in predicting the time of occurrence of lung cancer and identifying the important genes that were mainly associated with the disease. The

Table 3 Prediction measures in the third scenario ($n = 100$)

p	Censoring rate	Model (1)		Model (2)		Model (3)		Model (4)	
		RMSE	CCI	RMSE	CCI	RMSE	CCI	RMSE	CCI
50	20%	6.77	0.860	10.15	0.809	14.33	0.242	321.47	0.323
		(0.02)	(0.011)	(0.02)	(0.025)	(0.03)	(0.029)	(0.26)	(0.047)
	40%	1.95	0.837	3.94	0.789	7.45	0.251	361.09	0.304
		(0.00)	(0.026)	(0.00)	(0.030)	(0.01)	(0.054)	(0.22)	(0.063)
	60%	7.60	0.826	8.19	0.787	5.69	0.251	231.78	0.318
		(0.02)	(0.031)	(0.01)	(0.043)	(0.01)	(0.052)	(0.23)	(0.061)
200	20%	8.19	0.766	9.88	0.724	11.11	0.255	273.84	0.388
		(0.02)	(0.062)	(0.02)	(0.045)	(0.01)	(0.022)	(0.19)	(0.046)
	40%	10.96	0.701	19.57	0.707	13.39	0.252	199.00	0.388
		(0.03)	(0.060)	(0.04)	(0.060)	(0.03)	(0.021)	(0.16)	(0.050)
	60%	8.05	0.670	7.19	0.695	11.53	0.262	186.53	0.374
		(0.01)	(0.051)	(0.01)	(0.059)	(0.02)	(0.039)	(0.15)	(0.042)
1000	20%	5.07	0.592	39.88	0.591	9.52	0.358	241.34	0.463
		(0.01)	(0.046)	(0.07)	(0.032)	(0.01)	(0.038)	(0.16)	(0.027)
	40%	10.64	0.578	13.57	0.591	13.00	0.355	222.58	0.444
		(0.01)	(0.044)	(0.02)	(0.040)	(0.01)	(0.036)	(0.19)	(0.036)
	60%	7.84	0.577	5.52	0.594	5.08	0.340	236.65	0.458
		(0.01)	(0.034)	(0.01)	(0.040)	(0.00)	(0.028)	(0.20)	(0.015)

Model (1) is SVHM, Model (2) is KM-inverse weight SVHM, Model (3) is SVR, and Model (4) is SVMR. The values in the table comprise average (standard deviation) (Units of RMSE: 1E−3)

data comprised 86 subjects ($n = 86$), the predictors were composed of 7129 genes ($p = 7129$), and the outcome was composed of the lung cancer diagnosis time and corresponding censoring indicator, for which 62 of the 86 subjects were observed with the survival time, whereas 24 subjects were observed with the censoring time; consequently, a censoring rate of 28% was calculated.

After the entire dataset was divided into five folds data sets, each of which comprised the training and test datasets, we attempted to fit the four different prediction models based on the margin maximization methods to the training datasets to estimate the functional structure of the risk score. Thereafter, we predicted the risk score for each subject contained in the test datasets, computed the CCI values for each test dataset as a measure of the prediction performance, and compared the prediction performance of our proposed methods with those of the others. Finally, we selected the important genes through the gradient-based variable selection method, where the threshold number for the variable selection was set as $[\log(n)/n] = [18.2] = 18$, the Gaussian RBF kernel was employed, and the bandwidth size was used as mentioned in the work (Wang, 2012). When dividing the data into five folds, the training datasets were generated with an almost identical censoring rate to that of the original dataset.

Table 4 Prediction measures in the fourth scenario ($n = 100$)

p	censoring rate	Model (1)		Model (2)		Model (3)		Model (4)	
		RMSE	CCI	RMSE	CCI	RMSE	CCI	RMSE	CCI
50	20%	1.19	0.889	2.78	0.909	83.24	0.388	215.72	0.351
		(0.00)	(0.010)	(0.01)	(0.012)	(0.11)	(0.033)	(0.20)	(0.078)
	40%	36.51	0.879	36.45	0.896	120.26	0.366	377.83	0.329
		(0.16)	(0.013)	(0.15)	(0.017)	(0.16)	(0.047)	(0.22)	(0.077)
	60%	3.40	0.874	93.50	0.890	71.13	0.388	325.65	0.361
		(0.01)	(0.022)	(0.40)	(0.024)	(0.11)	(0.045)	(0.24)	(0.092)
200	20%	0.96	0.873	3.75	0.852	416.44	0.349	160.03	0.418
		(0.00)	(0.013)	(0.01)	(0.011)	(1.05)	(0.045)	(0.12)	(0.064)
	40%	3.40	0.860	4.60	0.837	145.76	0.355	210.02	0.423
		(0.01)	(0.023)	(0.01)	(0.031)	(0.19)	(0.041)	(0.11)	(0.057)
	60%	0.08	0.846	3.00	0.827	82.79	0.362	189.79	0.403
		(0.00)	(0.029)	(0.01)	(0.037)	(0.16)	(0.042)	(0.15)	(0.076)
1000	20%	8.06	0.837	8.02	0.817	97.49	0.284	19.83	0.407
		(0.03)	(0.010)	(0.03)	(0.016)	(0.11)	(0.032)	(0.02)	(0.052)
	40%	0.67	0.844	30.26	0.788	60.67	0.295	156.42	0.404
		(0.00)	(0.019)	(0.13)	(0.037)	(0.10)	(0.028)	(0.46)	(0.063)
	60%	18.22	0.837	24.97	0.792	148.36	0.298	66.86	0.400
		(0.05)	(0.026)	(0.06)	(0.037)	(0.17)	(0.034)	(0.04)	(0.053)

Model (1) is SVHM, Model (2) is KM-inverse weight SVHM, Model (3) is SVR, and Model (4) is SVMR. The values in the table comprise average (standard deviation) (Units of RMSE: 1E−3)

For each prediction model, we conducted variable selection with the gradient value estimate for each marginal predictor and refitted the prediction methods for the selected predictors.

Table 9 presents the average CCI values and RMSE performance for all the methods prior and posterior to the variable section. We observed that the three CCI values for Models (1), (2), and (3) slightly decreased after applying the variable selection, while the CCI for Model (4) increased after applying the variable selection. In the contrast to the results of the CCI performance, we found that the prediction performance for Models (1) and (2) presented a meaningful result in that their corresponding RMSE values tended to be improved after conducting our proposed kernel-based variable selection approach. The prediction performance of the rank-based SVM, Model (4) was also improved compared with the result before the variable selection while Model (3) was not improved even after applying the variable selection method.

Figure 3 contains four box plots, each of which shows the distribution of the RMSE value for each prediction method, with the weight based on the risk set. The weight based on the inverse probability, support vector regression,and support vector machine based on the ranking are denoted by Models (1), (2), (3), and (4), respectively. As the empirical average of the prediction performance, the RMSE values 5.305 and 4.884 were obtained for our proposed Models (1) and

Table 5 Variable selection measures in the first scenario ($n = 100$)

p	cr	Model (1)			Model (2)			Model (3)			Model (4)		
		TPR	FPR	$\hat{d}$	TPR	FPR	$\hat{d}$	TPR	FPR	$\hat{d}$	TPR	FPR	$\hat{d}$
50	20%	1.000	0.000	5.000	0.592	0.045	10.768	0.131	0.097	27.396	0.708	0.032	13.271
		(0.000)	(0.000)	(0.000)	(0.269)	(0.030)	(5.474)	(0.150)	(0.017)	(8.165)	(0.171)	(0.019)	(6.442)
	40%	1.000	0.000	5.000	0.608	0.044	11.558	0.127	0.097	26.708	0.694	0.034	12.938
		(0.000)	(0.000)	(0.000)	(0.266)	(0.030)	(6.575)	(0.167)	(0.019)	(8.567)	(0.181)	(0.020)	(6.114)
	60%	0.996	0.000	5.031	0.648	0.039	11.844	0.108	0.099	26.990	0.698	0.034	13.812
		(0.029)	(0.003)	(0.227)	(0.269)	(0.030)	(7.405)	(0.130)	(0.014)	(7.512)	(0.188)	(0.021)	(7.043)
200	20%	1.000	0.000	5.000	0.781	0.006	12.186	0.010	0.025	118.698	0.600	0.010	77.062
		(0.000)	(0.000)	(0.000)	(0.256)	(0.007)	(13.417)	(0.045)	(0.001)	(32.054)	(0.199)	(0.005)	(48.698)
	40%	1.000	0.000	5.000	0.746	0.007	20.602	0.010	0.025	121.448	0.596	0.010	84.635
		(0.000)	(0.000)	(0.000)	(0.278)	(0.007)	(28.584)	(0.045)	(0.001)	(34.159)	(0.195)	(0.005)	(50.356)
	60%	0.991	0.000	7.756	0.759	0.006	25.663	0.012	0.025	108.552	0.577	0.011	87.219
		(0.041)	(0.001)	(13.783)	(0.256)	(0.007)	(37.034)	(0.049)	(0.001)	(33.072)	(0.221)	(0.006)	(51.579)
1000	20%	1.000	0.000	5.000	0.940	0.000	6.800	0.680	0.002	210.900	0.600	0.002	419.000
		(0.000)	(0.000)	(0.000)	(0.097)	(0.000)	(3.736)	(0.140)	(0.001)	(206.824)	(0.298)	(0.001)	(338.988)
	40%	1.000	0.000	5.000	0.860	0.001	82.600	0.680	0.002	141.400	0.420	0.003	543.700
		(0.000)	(0.000)	(0.000)	(0.190)	(0.001)	(136.579)	(0.193)	(0.001)	(223.913)	(0.305)	(0.002)	(292.212)
	60%	0.940	0.000	5.500	0.940	0.000	23.500	0.760	0.001	150.800	0.560	0.002	414.000
		(0.097)	(0.000)	(0.972)	(0.097)	(0.000)	(53.736)	(0.227)	(0.001)	(225.140)	(0.246)	(0.001)	(313.878)

Model (1) is SVHM, Model (2) is KM-inverse weight SVHM, Model (3) is SVR, and Model (4) is SVMR. The values in the table consist of average (standard deviation)

Table 6 Variable selection measures in the second scenario ($n = 100$)

p	cr	Model (1)			Model (2)			Model (3)			Model (4)		
		TPR	FPR	$\hat{d}$	TPR	FPR	$\hat{d}$	TPR	FPR	$\hat{d}$	TPR	FPR	$\hat{d}$
50	20%	1.000	0.000	5.000	0.987	0.001	5.074	0.065	0.104	35.229	0.688	0.035	13.792
		(0.000)	(0.000)	(0.000)	(0.049)	(0.005)	(0.300)	(0.107)	(0.012)	(6.081)	(0.181)	(0.020)	(7.141)
	40%	0.996	0.000	5.022	0.933	0.007	5.583	0.071	0.103	34.552	0.733	0.030	12.458
		(0.029)	(0.003)	(0.146)	(0.129)	(0.014)	(1.427)	(0.116)	(0.013)	(6.132)	(0.180)	(0.020)	(6.568)
	60%	0.981	0.002	5.189	0.857	0.016	6.747	0.060	0.104	34.333	0.698	0.034	13.562
		(0.059)	(0.007)	(0.689)	(0.179)	(0.020)	(2.779)	(0.109)	(0.012)	(6.210)	(0.186)	(0.021)	(7.294)
200	20%	1.000	0.000	5.000	0.784	0.006	14.489	0.004	0.026	142.969	0.612	0.010	83.875
		(0.000)	(0.000)	(0.000)	(0.215)	(0.006)	(16.993)	(0.029)	(0.001)	(23.265)	(0.192)	(0.005)	(51.992)
	40%	0.995	0.000	6.670	0.773	0.006	17.611	0.002	0.026	142.312	0.629	0.010	77.354
		(0.030)	(0.001)	(11.023)	(0.242)	(0.006)	(24.112)	(0.020)	(0.001)	(24.671)	(0.217)	(0.006)	(50.795)
	60%	0.966	0.001	9.355	0.779	0.006	20.277	0.002	0.026	142.271	0.623	0.010	74.833
		(0.081)	(0.002)	(16.867)	(0.246)	(0.006)	(31.619)	(0.020)	(0.001)	(24.158)	(0.207)	(0.005)	(54.127)
1000	20%	1.000	0.000	5.000	0.900	0.001	155.125	0.450	0.003	464.125	0.500	0.003	422.375
		(0.000)	(0.000)	(0.000)	(0.151)	(0.001)	(291.031)	(0.141)	(0.001)	(318.374)	(0.239)	(0.001)	(319.776)
	40%	0.950	0.000	65.375	0.850	0.001	251.625	0.475	0.003	469.500	0.550	0.002	603.500
		(0.093)	(0.000)	(169.557)	(0.177)	(0.001)	(385.606)	(0.183)	(0.001)	(236.353)	(0.207)	(0.001)	(162.208)
	60%	0.940	0.000	70.700	0.900	0.001	41.200	0.500	0.003	493.100	0.500	0.003	391.600
		(0.097)	(0.000)	(138.209)	(0.141)	(0.001)	(68.264)	(0.170)	(0.001)	(281.070)	(0.216)	(0.001)	(326.075)

Model (1) is SVHM, Model (2) is KM-inverse weight SVHM, Model (3) is SVR, and Model (4) is SVMR. The values in the table consist of average (standard deviation)

Table 7 Variable selection measures in the third scenario ($n = 100$)

p	cr	Model (1)			Model (2)			Model (3)			Model (4)		
		TPR	FPR	$\hat{d}$	TPR	FPR	$\hat{d}$	TPR	FPR	$\hat{d}$	TPR	FPR	$\hat{d}$
50	20%	1.000	0.000	5.000	0.716	0.032	13.368	0.705	0.033	13.789	0.116	0.098	35.842
		(0.000)	(0.000)	(0.000)	(0.154)	(0.017)	(10.930)	(0.168)	(0.019)	(6.909)	(0.121)	(0.013)	(8.585)
	40%	1.000	0.000	5.000	0.660	0.038	13.350	0.650	0.039	14.050	0.160	0.093	38.000
		(0.000)	(0.000)	(0.000)	(0.206)	(0.023)	(9.826)	(0.182)	(0.020)	(6.533)	(0.167)	(0.019)	(11.107)
	60%	1.000	0.000	5.000	0.770	0.026	11.950	0.700	0.033	11.900	0.150	0.094	34.850
		(0.000)	(0.000)	(0.000)	(0.187)	(0.021)	(7.345)	(0.178)	(0.020)	(5.767)	(0.143)	(0.016)	(9.063)
200	20%	1.000	0.000	5.000	0.750	0.006	24.562	0.700	0.008	15.438	0.075	0.024	125.312
		(0.000)	(0.000)	(0.000)	(0.225)	(0.006)	(26.069)	(0.231)	(0.006)	(16.219)	(0.100)	(0.003)	(41.514)
	40%	1.000	0.000	5.000	0.750	0.006	24.200	0.890	0.003	7.850	0.060	0.024	133.950
		(0.000)	(0.000)	(0.000)	(0.193)	(0.005)	(40.069)	(0.152)	(0.004)	(5.613)	(0.114)	(0.003)	(33.404)
	60%	1.000	0.000	5.000	0.747	0.006	19.053	0.842	0.004	13.579	0.105	0.023	126.053
		(0.000)	(0.000)	(0.000)	(0.248)	(0.006)	(20.118)	(0.107)	(0.003)	(20.865)	(0.193)	(0.005)	(47.512)
1000	20%	1.000	0.000	5.000	0.580	0.002	106.600	0.920	0.000	7.100	0.080	0.005	636.600
		(0.000)	(0.000)	(0.000)	(0.319)	(0.002)	(222.874)	(0.140)	(0.001)	(3.957)	(0.140)	(0.001)	(292.546)
	40%	1.000	0.000	5.000	0.700	0.002	62.200	0.940	0.000	5.900	0.020	0.005	713.700
		(0.000)	(0.000)	(0.000)	(0.254)	(0.001)	(82.192)	(0.097)	(0.000)	(1.912)	(0.063)	(0.000)	(235.135)
	60%	1.000	0.000	5.000	0.680	0.002	29.400	0.900	0.001	6.400	0.040	0.005	608.600
		(0.000)	(0.000)	(0.000)	(0.286)	(0.001)	(39.093)	(0.141)	(0.001)	(2.119)	(0.126)	(0.001)	(225.811)

Model (1) is SVHM, Model (2) is KM-inverse weight SVHM, Model (3) is SVR, and Model (4) is SVMR. The values in the table consist of average (standard deviation)

Table 8 Variable selection measures in the fourth scenario ($n = 100$)

p	cr	Model (1)			Model (2)			Model (3)			Model (4)		
		TPR	FPR	$\hat{d}$	TPR	FPR	$\hat{d}$	TPR	FPR	$\hat{d}$	TPR	FPR	$\hat{d}$
50	20%	1.000	0.000	5.000	0.890	0.012	5.850	0.060	0.104	35.200	0.720	0.031	11.950
		(0.000)	(0.000)	(0.000)	(0.137)	(0.015)	(1.226)	(0.131)	(0.015)	(5.415)	(0.136)	(0.015)	(5.790)
	40%	0.979	0.002	5.263	0.811	0.021	6.895	0.084	0.102	30.895	0.684	0.035	14.368
		(0.063)	(0.007)	(0.933)	(0.156)	(0.017)	(1.997)	(0.101)	(0.011)	(7.838)	(0.168)	(0.019)	(6.593)
	60%	0.979	0.002	5.368	0.716	0.032	8.947	0.095	0.101	33.158	0.747	0.028	12.842
		(0.063)	(0.007)	(1.383)	(0.224)	(0.025)	(4.916)	(0.122)	(0.014)	(6.212)	(0.147)	(0.016)	(7.313)
200	20%	0.965	0.001	7.294	0.729	0.007	17.294	0.012	0.025	142.471	0.600	0.010	75.941
		(0.079)	(0.002)	(8.950)	(0.121)	(0.003)	(15.814)	(0.049)	(0.001)	(24.600)	(0.212)	(0.005)	(58.062)
	40%	0.979	0.001	7.842	0.653	0.009	34.737	0.011	0.025	141.789	0.611	0.010	90.421
		(0.063)	(0.002)	(8.745)	(0.198)	(0.005)	(39.546)	(0.046)	(0.001)	(21.865)	(0.156)	(0.004)	(60.379)
	60%	0.895	0.003	29.368	0.632	0.009	54.737	0.011	0.025	136.789	0.642	0.009	66.053
		(0.139)	(0.004)	(40.247)	(0.233)	(0.006)	(45.423)	(0.046)	(0.001)	(29.421)	(0.217)	(0.006)	(53.642)
1000	20%	0.971	0.000	5.429	0.957	0.000	30.429	0.529	0.002	308.071	0.543	0.002	560.286
		(0.073)	(0.000)	(1.158)	(0.085)	(0.000)	(83.435)	(0.127)	(0.001)	(275.865)	(0.287)	(0.001)	(333.682)
	40%	0.916	0.000	38.632	0.832	0.001	233.053	0.495	0.003	527.211	0.589	0.002	277.211
		(0.154)	(0.001)	(79.096)	(0.153)	(0.001)	(274.386)	(0.122)	(0.001)	(305.133)	(0.156)	(0.001)	(281.078)
	60%	0.956	0.000	51.778	0.800	0.001	267.556	0.500	0.003	424.944	0.511	0.002	464.278
		(0.086)	(0.000)	(190.284)	(0.137)	(0.001)	(284.047)	(0.124)	(0.001)	(168.946)	(0.230)	(0.001)	(310.886)

Model (1) is SVHM, Model (2) is KM-inverse weight SVHM, Model (3) is SVR, and Model (4) is SVMR. The values in the table consist of average (standard deviation)

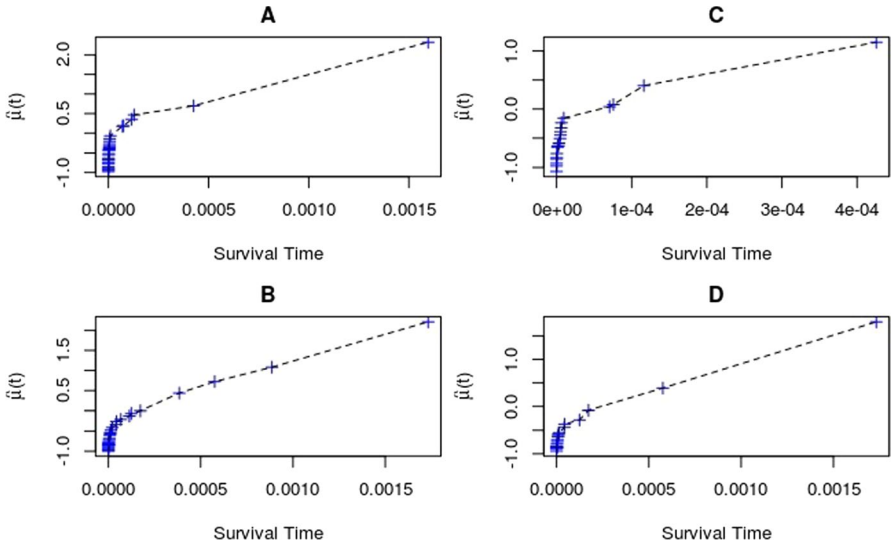

Fig. 1 The blue plus points (+) denotes the survival time, $\hat{\mu}(t)$, and the black dashed line is the line that connects the blue points. The results of SVHM were plotted, where **A** is $n = 50, p = 100$, censoring rate $= 0.4$, **B** is $n = 50, p = 100$, censoring rate $= 0.6$, **C** is $n = 50, p = 200$, censoring rate $= 0.4$, and **D** is $n = 50, p = 200$, censoring rate $= 0.6$

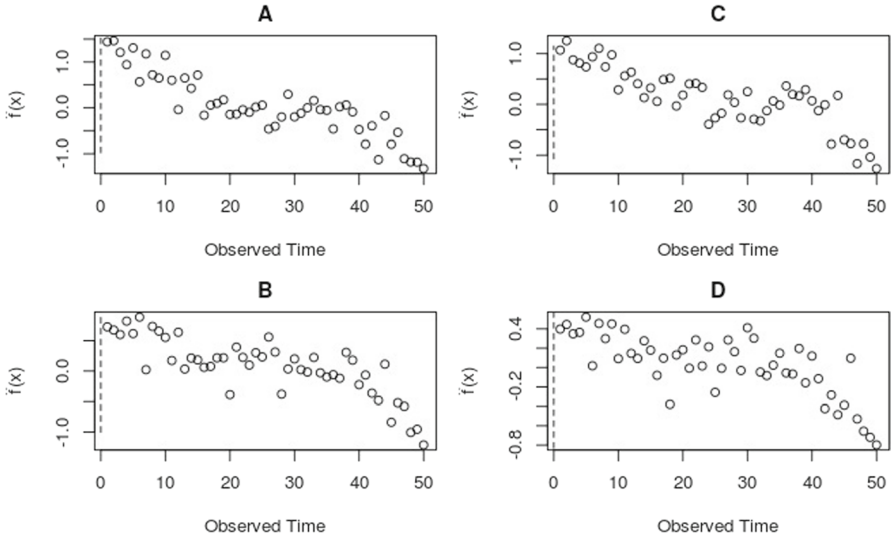

Fig. 2 Scatter plots of rank of observed time and risk score. The results of SVHM were plotted, where **A** is $n = 50, p = 100$, censoring rate $= 0.4$, **B** is $n = 50, p = 100$, censoring rate $= 0.6$, **C** is $n = 50, p = 200$, censoring rate $= 0.4$, and **D** is $n = 50, p = 200$, censoring rate $= 0.6$

(2), respectively, which indicates better performance compared with the others in terms of the prediction. Additionally, we observed that the spread of the SVHM

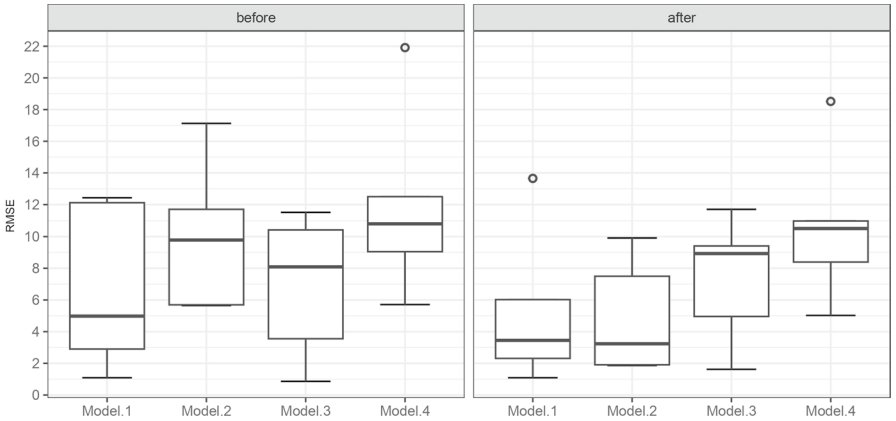

Fig. 3 Box plots for the results for RMSE of each model. Model (1) is SVHM, Model (2) is KM-inverse weight SVHM, Model (3) is SVR, and Model (4) is SVMR. The results before the variable selection are displayed on the left and the results after that are displayed on the right

Table 9 Prediction measures in the real data analysis

Model	Concordance Index		RMSE	
	Before	After	Before	After
Model (1)	0.693 (0.084)	0.652 (0.087)	6.704 (5.271)	5.305 (5.011)
Model (2)	0.723 (0.109)	0.671 (0.198)	9.992 (4.775)	4.884 (3.634)
Model (3)	0.331 (0.078)	0.326 (0.117)	6.886 (4.548)	7.325 (4.006)
Model (4)	0.260 (0.096)	0.513 (0.181)	11.996 (6.087)	10.686 (4.969)

Model (1) is SVHM, Model (2) is KM-inverse weight SVHM, Model (3) is SVR, and Model (4) is SVMR. The values in the table consist of average (standard deviation)

with the inverse probability based on the KM estimators in Model (2) tends to be smaller than that of the SVHM with the weight based on the risk set in Model (1).

Two panels of Fig. 4 depict the time-varying intercept $\widehat{\mu}(t)$ and time-independent risk score $\widehat{f}(\cdot)$ estimated for the lung cancer data set. With the increase in the survival time, the intercept tended to increase, which reveals an appropriately estimated appearance because the risk needs to be increased over the time. Although the time-independent risk score showed a slightly scattered appearance, we observed a functional tendency to decrease to some degree.

Table 10 shows the results of the variable selection through the gradient information for each method. We observed that some genetic predictors were commonly selected by the prediction models employed, and there were also a few differences in the selected genetic predictors between the prediction models. The indices of the

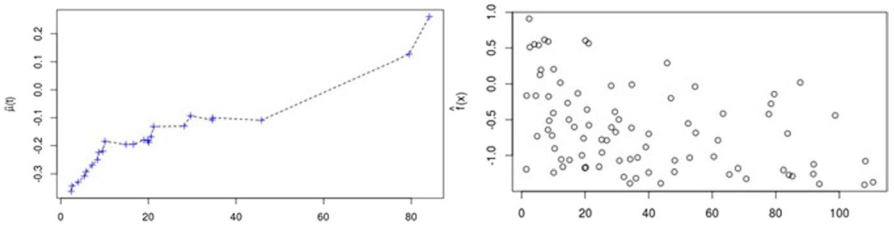

Fig. 4 (Left) Scatter plot of estimated intercept and survival time. The black dashed line is the line that connects the blue points. (Right) Scatter plot of rank of observed time and risk score

Table 10 Results of variable selection for each model

Model	Variable index
Model (1)	144 146 148 564 1107 1188 1430 2429 2433 2778 3001 3255 3481 4548 5474 5491 5909 6061 6678
Model (2)	133 146 148 564 892 1037 1107 2429 2778 2992 4551 5463 5474 5491 5521 5622 5909 6061 6150
Model (3)	127 133 142 148 564 1107 1188 1221 2429 2433 2753 3001 3059 5474 5475 5782 5909 6061 6678
Model (4)	131 144 146 148 564 686 1226 2326 2429 2753 2778 2873 2999 3001 3481 3596 5474 5903 5909

Model (1) is SVHM, Model (2) is KM-inverse weight SVHM, Model (3) is SVR, and Model (4) is SVMR

genetic predictors selected by three or more prediction models among the genes selected through the proposed SVHM contained (146, 148, 564, 1107, 2429, 2778, 3001, 5474, 5909, and 6061), whereas the indices of the genetic predictors selected only by SVHM contained (3255, and 4548). To ascertain whether the selected genetic predictors can be a biologically meaningful result, we looked for a study with the association between the selected genes and lung cancer. Peng et al. (2022) found that the non-small cell lung cancer could be worse through the GAPDH (gene index 146). Carleo et al. (2020) found an association between the gene IGKC (gene index 3001) and the lung carcinogenesis in idiopathic pulmonary fibrosis patients. Ma et al. (2017) investigated the expression and epigenetic regulation of CSTB (gene index 4548) in lung cancer, where this gene has been included only in our proposed method. Although it is not directly related with the lung cancer but may not be irrelevant to the lung cancer, there was a study Gustafsson et al. (2008) that showed an association the gene IGHG (gene index 3255) and asthma severity, where it has been detected only by our method. Through these cases, we confirmed that our proposed method could be biologically meaningful.

5 Discussion

In this study, we developed a method to improve the prediction performance of the survival time and to select the important predictors in the time-to-event dataset. Specifically, we considered the counting process for each subject in the time-to-event data as a time-varying dichotomized outcome and, thereafter, adopted the SVHM and gradient-based variable selection methods to achieve two purposes namely, prediction and variable selection. Through simulation studies, we found that not only the existing margin-based methods, such as SVR and SVMR, but also the SVHM with the two different weights could present desirable prediction performance in terms of MSE, whereas the SVHM with the weights outperformed the others in terms of CCI. Moreover, we observed that the finite performance of the prediction measure for the SVHM approach with weights tended to be better for the complicated scenario in the simulation study. We believe that the proposed framework can be practically used to solve the problem of predicting the time of occurrence of an event and choosing variables in time-to-event data.

For the real data application, we used the gene expression values of the microarray data of the patients with lung cancer as high-dimensional predictors and survival time to death as the outcome. We demonstrated that both our SVHM approaches with two weights provided better prediction performance, and that such a prediction performance did not decrease significantly and was maintained even after using the gradient-based variable selection method. Using the results obtained by applying the gradient-based variable selection method to each prediction model, it was possible to identify genes that could be found in common and genes that can be uniquely discovered using each method. We confirmed that the genes identified by our proposed method were biologically meaningful and demonstrated that the proposed method is scientifically valid.

We highlighted that the amount of computational time and time-independent covariates are limitations of our proposed approach. Moreover, we could not use time-dependent covariate because we were considering time-dependent risk scores and covariate-dependent risk scores for the time-independent covariate. Additionally, we must optimize the regularized empirical risk at each survival time, which requires heavy computations. We believe that developing a scalable SVHM method is necessary and will leave this work for future research.

Acknowledgements We thank the Editor, Associate Editor and two reviewers, whose questions and insightful comments have led to a much improved paper.

Funding This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF2021R1C1C1007023).

Data availability We used Beer's microarray data, which is available with *LungCancer3* function in the R package "*GSCA*".

References

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society, 68*, 337–404.

Beer, D. G., Kardia, S. L., Huang, C.-C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., et al. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine, 8*, 816–824.

Carleo, A., Landi, C., Prasse, A., Bergantini, L., d'Alessandro, M., Cameli, P., Janciauskiene, S., Rottoli, P., Bini, L., & Bargagli, E. (2020). Proteomic characterization of idiopathic pulmonary fibrosis patients: Stable versus acute exacerbation. *Monaldi Archives for Chest Disease, 90*, 180–190.

Clarke, B. S., Fokoué, E., & Zhang, H. H. (2009). *Principles and Theory for Data Mining and Machine Learning*. Springer.

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

Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica, 20*(1), 101–148.

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

Fukumizu, K., & Leng, C. (2014). Gradient-based kernel dimension reduction for regression. *Journal of the American Statistical Association, 109*, 359–370.

Gustafsson, P. M., Oxelius, V.-A., Nilsson, S., & Kjellman, B. (2008). Association between gm allotypes and asthma severity from childhood to young middle age. *Respiratory Medicine, 102*, 266–272.

He, X., Wang, J., & Lv, S. (2021). Efficient kernel-based variable selection with sparsistency. *Statistica Sinica, 31*, 2123–2151.

Ibrahim, J. G., Chen, M.-H., & Sinha, D. (2001). *Bayesian Survival Analysis*. Springer.

Jeong, S., Kim, C., & Yang, H. (2023). Wasserstein filter for variable screening in binary classification in the reproducing kernel Hilbert space. *Journal of Nonparametric Statistics*, 1–20 **(in press)**

Kalbfleisch, J. D., & Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data* (2nd ed.). Wiley.

Khan, F. M., & Zubek, V. B. (2008). Support Vector Regression for Censored Data (SVRc): A Novel Tool for Survival Analysis (pp. 863–868). IEEE, IEEE International Conference on Data Mining.

Lawless, J. F. (2002). *Statistical Models and Methods for Lifetime Data*. Wiley.

Ma, Y., Chen, Y., & Petersen, I. (2017). Expression and epigenetic regulation of cystatin b in lung cancer and colorectal cancer. *Pathology-Research and Practice, 213*, 1568–1574.

Park, B., & Park, C. (2021). Kernel variable selection for multicategory support vector machines. *Journal of Multivariate Analysis, 186*, 104800.

Peng, J., Li, W., Tan, N., Lai, X., Jiang, W., & Chen, G. (2022). Usp47 stabilizes bach1 to promote the Warburg effect and non-small cell lung cancer development via stimulating hk2 and gapdh transcription. *American Journal of Cancer Research, 12*, 91–107.

Tibshirani, R., et al. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine, 16*, 385–395.

Van Belle, V., Pelckmans, K., Van Huffel, S., & Suykens, J. A. (2011). Support vector methods for survival analysis: A comparison between ranking and regression approaches. *Artificial Intelligence in Medicine, 53*, 107–118.

Wang, Q. (2012). Kernel principal component analysis and its applications in face recognition and active shape models. arXiv:1207.3538

Wang, Y., Chen, T., & Zeng, D. (2016). Support vector hazards machine: A counting process framework for learning risk scores for censored outcomes. *The Journal of Machine Learning Research, 17*, 5825–5861.

Wei, L.-J. (1992). The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in Medicine, 11*, 1871–1879.

Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *The Annals of Statistics, 35*, 2654–2690.

Xia, Y., Tong, H., Li, W., & Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*, 363–410.

Yang, H., Zhu, H., Ahn, M., & Ibrahim, J. G. (2021). Weighted functional linear cox regression model. *Statistical Methods in Medical Research, 30*, 1917–1931.