Journal of Machine Learning Research 17 (2016) 1-37 Submitted 1/16; Revised 7/16; Published 8/16

Support Vector Hazards Machine: A Counting Process Framework for Learning Risk Scores for Censored Outcomes

Yuanjia Wang YW2016@COLUMBIA.EDU
Department of Biostatistics
Mailman School of Public Health
Columbia University
New York, NY 10032, USA

Tianle Chen TIANLE.CHEN@BIOGEN.COM
Biogen
300 Binney Street
Cambridge, MA 02142, USA

Donglin Zeng DZENG@EMAIL.UNC.EDU
Department of Biostatistics
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599, USA

Editor: Karsten Borgwardt

Abstract

Learning risk scores to predict dichotomous or continuous outcomes using machine learning approaches has been studied extensively. However, how to learn risk scores for time-to-event outcomes subject to right censoring has received little attention until recently. Existing approaches rely on inverse probability weighting or rank-based regression, which may be inefficient. In this paper, we develop a new support vector hazards machine (SVHM) approach to predict censored outcomes. Our method is based on predicting the counting process associated with the time-to-event outcomes among subjects at risk via a series of support vector machines. Introducing counting processes to represent time-to-event data leads to a connection between support vector machines in supervised learning and hazards regression in standard survival analysis. To account for different at risk populations at observed event times, a time-varying offset is used in estimating risk scores. The resulting optimization is a convex quadratic programming problem that can easily incorporate non-linearity using kernel trick. We demonstrate an interesting link from the profiled empirical risk function of SVHM to the Cox partial likelihood. We then formally show that SVHM is optimal in discriminating covariate-specific hazard function from population average hazard function, and establish the consistency and learning rate of the predicted risk using the estimated risk scores. Simulation studies show improved prediction accuracy of the event times using SVHM compared to existing machine learning methods and standard conventional approaches. Finally, we analyze two real world biomedical study data where we use clinical markers and neuroimaging biomarkers to predict age-at-onset of a disease, and demonstrate superiority of SVHM in distinguishing high risk versus low risk subjects.

Keywords: support vector machine, survival analysis, risk bound, risk prediction, neuroimaging biomarkers, early disease detection

1. Introduction

Time-to-event outcome is of interest in many scientific studies in which right censoring occurs when subjects' event times are longer than the duration of studies or subjects drop out of the study prematurely. One important goal in these studies is to use baseline covariates collected on a newly recruited subject to construct an effective risk score to predict likelihood an event of interest. For example, in one of our motivating studies analyzed in Section 4.2 (PREDICT-HD, Paulsen et al. 2008a), the research aim is to combine neuroimaging biomarkers with clinical markers measured at the baseline to provide risk stratification for time-to-onset of Huntington's disease (HD) to facilitate early diagnosis, where subjects who did not experience HD during the study had censored HD onset time. This critical goal of identifying prognostic markers predictive of disease onset is shared by research communities on other neurological disorders such as Alzheimer's disease and Parkinson's disease, and recognized as one of the primary aims in research initiatives such as Alzheimer's Disease Neuroimaging Initiative (Mueller et al., 2005) and Parkinson's Progression Markers Initiative (Marek et al., 2011).

Learning risk scores for binary or continuous outcomes are examined extensively in statistical learning literature (Hastie et al., 2009). However, learning risk scores for occurrence of an event subject to censoring is much less explored. Existing work on survival analysis focuses on estimating population-level quantities such as survival function or association parameters through hazard function. For example, the most popular model for the time-to-event analysis is the Cox proportional hazards model (Cox, 1972), which assumes the hazard ratio between two subjects with different covariate values stays as a constant as time progresses. A Cox partial likelihood function is maximized for estimation. When the proportional hazards assumption is violated, several alternative models have been proposed in statistics literature, including the proportional odds model (Bennett, 1983), the accelerated failure time model (Buckley and James, 1979), the linear transformation models (Dabrowska and Doksum, 1988; Cheng et al., 1995; Chen et al., 2002), and more recently general transformation models (Zeng and Lin, 2006, 2007). The above models are all likelihood-based which impose certain parametric or semiparametric relationship between the underlying hazard function and the covariates. In addition, they are designed to estimate the population-level parameters for the association between covariates and the time-to-event outcomes (and thus uses likelihood as the optimization function), but do not directly focus on individual risk scores for predicting an event time.

For non-censored outcomes, supervised learning plays an important role for risk prediction. In many applications, a large number of input variables with known output values are used to learn an unknown functional relationship between the inputs and outputs through a suitable algorithm, and the learned functional is used to predict the output value for future subjects from their input variables (Steinwart and Christmann, 2008). Many learning approaches have been developed for standard classification and regression problems, such as kernel smoothing, support vector machines (SVM), projection pursuit regression, neural network, and decision trees (Hastie et al., 2009). In particular, support vector machine is among one of the most popular and successful learning methods in practice (Mogueraza and Munoz, 2006; Orru et al., 2012). From the training data, support vector machine finds a hyperplane that separates the data into two classes as accurately as possible and has a simple

geometric interpretation. In addition, the algorithm can be written as a strictly convex optimization problem, which leads to a unique global optimum and incorporates non-linearity in an automatic manner using various kernel machines. By reformulating the algorithm into a minimization of regularized empirical risk, Steinwart (2002) established the universal consistency and learning rate on some functional space. Support vector machines have also been applied to continuous outcomes through regression (Smola and Schölkopf, 2004), multicategory discrete outcomes (Lee et al., 2004), and structured classification problems (Wang et al., 2011).

For time-to-event outcomes, right censoring makes developing supervised learning techniques challenging due to missing event times for censored subjects and a lack of standard prediction loss function. Ripley and Ripley (2001) and Ripley et al. (2004) discussed models for survival analysis based on neural network. Bou-Hamad et al. (2011) reviewed survival tree approaches in the recent work as non-parametric alternatives to semiparametric models. Compared to survival trees, effectively extending the support vector–based methods to censored data is still an on-going research. Shivaswamy et al. (2007) and Khan and Zubek (2008) proposed asymmetric modifications to the ϵ-insensitive loss function of support vector regression (SVR) to handle censoring. Specifically, they penalized the censored and non-censored subjects using different loss functions to extract incomplete information due to censoring. Van Belle et al. (2010) proposed a least-squares support vector machine, where they adopted the concept of concordance index and added rank constraints to handle censored data. In their method, the empirical risk of miss-ranking two data points with respect to their event times was minimized. Furthermore, Van Belle et al. (2011) conducted numerical experiments to compare some recent machine learning methods for censored data and proposed a modified procedure to adjust for censoring based on both rank and regression constraints. Their results indicate that including two types of constraints performs the best regarding the prediction accuracy. None of the above methods has theoretical justification and the relationship between their objective loss functions to be minimized and the goal of predicting survival time remains unclear. The rank-based methods only use feasible pairs of observations whose ranks are comparable so that it may result in potential selection bias when constructing prediction rules, especially when the censoring mechanism is not completely at random (e.g., censoring time depends/correlates with a subject's covariates). Recently, Goldberg and Kosorok (2013) used inverse-probability-of-censoring weighting to adapt standard support vector methods for complete data to censored data. However, inverse weighting is known to be inefficient (Robins et al., 1995) due to the fact that it discards useful information for some subjects known to survive longer than observed times, and in addition, this method may exhibit severe bias when the censoring distribution is misspecified. Additionally, the weights used in the inverse weighting can be large in some situations, and computation of Goldberg and Kosorok (2013) becomes numerically unstable and even infeasible.

In this work, we propose a new support vector hazards machine (SVHM) framework to learn risk scores for survival outcomes using the concept of counting process. We aim to maximally separate event and no-event subjects among all subjects at risk, and allow censoring times to depend on covariates without modeling the censoring distribution. One major challenge in predicting censored event times is the difficulty of defining a sensible loss function for prediction. Because of the equivalence of an event time to its counting process,

if a prediction rule can adequately predict the event time, the same rule should also predict the counting process at any given time that a subject is still at risk. We propose a flexible nonparametric decision function with an additive structure for the counting process, which gives the desirable risk scores but also includes a time-varying offset to account for different at-risk population as time progresses. Empirically, we transform the prediction of an event time to predicting a sequence of binary outcomes for which algorithm such as support vector machine (SVM) is standard and commonly used. This transformation allows for the successful statistical learning tools designed for classification and prediction of binary outcomes to be used for censored outcomes without modeling the censoring distribution. The developed algorithm formulation is similar to the standard support vector machines and can be solved conveniently using any convex quadratic programming packages. In addition, theoretical analysis shows that the optimal rule obtained from SVHM is equivalent to maximizing the difference between the instantaneous subject-specific hazards and population-average hazard, which intuitively links SVHM to the commonly used hazards regression models in traditional survival analysis. The profile loss shares similarity with Cox partial likelihood. Under some regularity conditions, we show the universal consistency of SVHM and derive corresponding finite sample bounds on the deviation from the optimal risk. Numeric simulations and applications to real world studies show superior performance in distinguishing high risk versus low risk subjects.

2. Learning Risk Scores with SVHM

In this section, we first introduce the population loss function that SVHM aims to optimize with infinite sample and its corresponding Bayes risk. Next, we lay out the algorithm to empirically learn the risk scores and assess the empirical risk.

2.1 Review of Survival Analysis and Introduction of Counting Process Framework for SVHM

We begin by briefly introducing basic concepts and notation of classical survival analysis (c.f. Fleming and Harrington, 1991). Survival analysis focuses on using covariates to predict time to event outcomes. The events of interest can be death, diagnosis of a disease, onset of cancer metastasis, or failure of a machine component. An event time of interest (i.e., age at onset of a disease) is usually denoted by T, and a vector of baseline covariates (e.g., genomic risk factors) is denoted by $\mathbf{X}$. The main goals of survival analysis are to understand association between $\mathbf{X}$ and T or predicting T from $\mathbf{X}$. A fundamental problem of survival analysis is to deal with incomplete observation of T due to that the event may not occur in some of the subjects due to study termination or subjects dropping out of the study. For example, in a study on predicting time to cancer metastasis, some subjects may not develop metastasis by the end of study period, and thus their T is not observed. These subjects are termed as being censored and their time to study termination is termed as censoring time, usually denoted by C. For each subject in the study, we observe either their event time T or censoring time C, whichever is smaller. This observation is usually denoted by $T_i \wedge C_i$, where the operator $(a \wedge b)$ denotes taking minimum of a and b. A usual assumption in survival analysis is that the censoring time C is independent of T given covariates $\mathbf{X}$. From a random sample of n subjects, the observed data consist of $\{T_i \wedge C_i, \Delta_i = I(T_i \leq C_i), \mathbf{X}_i\}$

for $i = 1, \ldots, n$, where $I(\cdot)$ is an indicator function and $I(T_i \leq C_i)$ is thus the event indicator. The central quantity of interest in a survival analysis is occurrence of an event over time. Such occurrences are equivalent to point processes described by counting the number of events as they occur by certain time point, termed as counting processes. That is, a counting process of the event on subject i counts the number of events that have occurred up to, and including t, and is denoted as $N_i(t) = I[(T_i \wedge C_i) \leq t]$. Corresponding to the counting process for the events, the at-risk process counts subjects who have not yet had an event by time t and thus who are still "at risk" of experiencing an event. Such process is denoted by $Y_i(t) = I[(T_i \wedge C_i) \geq t]$.

The fundamental idea to learn risk scores for T to distinguish high risk versus low risk subjects is to equivalently learn risk scores for the counting process associated with T at each time point. Since the latter can be treated as a sequence of binary outcomes (event vs. no event) over time, it motivates one to reformulate the problem as deriving the risk score for predicting the jumps of the counting process over a sequence of time points among subjects still at risk at those times. This amounts to developing a classification rule to predict whether a subject will experience an event in the next immediate time point given that the subject has not yet experienced an event. To account for different risk sets as time progresses (i.e., risk set at time t is the subset of subjects with $Y_i(t) = 1$) , it is necessary to include a time-varying offset for the nonparametric risk score. Thus, consider the following general form at time t for a subject with $\mathbf{X} = \mathbf{x}$,

$$f(t, \mathbf{x}) = \alpha(t) + g(\mathbf{x}), \tag{1}$$

where both $\alpha(\cdot)$ and $g(\cdot)$ are unknown nonparametric functions, $g(\mathbf{x})$ is the risk score, and $\alpha(t)$ is the time-varying offset. To understand (1), consider a risk score function at time t for a subject with $\mathbf{X} = \mathbf{x}$: if this subject is still at risk at time t, we predict the subject to experience the event at the next immediate time point if $f(t, \mathbf{x}) > 0$, and predict as event-free if $f(t, \mathbf{x}) \leq 0$. Thus, within a small time interval $[t, t + dt)$, where dt denotes a positive infinitesimal unit, a natural prediction loss counting rate of risk-misclassification is given by

$$Y(t)dN(t)I(f(t, \mathbf{X}) < 0) + Y(t)(1 - dN(t))I(f(t, \mathbf{X}) \geq 0),$$

where $Y(t)$ and $\mathbf{X}$ are the at risk process and covariates for a subject drawn from the population, respectively, and $dN(t)$ denotes the number of jumps of the counting process $N(t)$ in a small time interval $[t, t + dt)$. Equivalently, $dN(t) = 1$ if $T \in [t, t + dt)$ and $dN(t) = 0$ otherwise, so it is a binary variable taking value one if an event occurs in the interval $[t, t+dt)$ for subjects who are still at risk for experiencing an event. Thus, summing above loss function over subjects counts the number of at-risk subjects miss-classified by the prediction rule $f(t, \mathbf{X})$. The above prediction loss can be viewed as a natural extension of the 0-1 loss for binary case to capture the same information for an at-risk subject in a survival analysis: if the prediction function and the observed counting process at time t are inconsistent, a loss is incurred. However, at any time t, the probability of $dN(t) = 1$ is almost zero as compared to the probability of $dN(t) = 0$, which implies that the above prediction loss is completely dominated by non-event subjects in the risk set. In order to balance the contribution from subjects with and without events at any given time, borrowing from the weighted SVM for unbalanced classes, a sensible prediction loss is the following

weighted loss, where the ratio of weights for two unbalanced classes is proportional to $E[dN(t)]/E[Y(t)]$:

$$Y(t)dN(t)I(f(t,\mathbf{X}) \leq 0) + \frac{E[dN(t)]}{E[Y(t)]}Y(t)(1-dN(t))I(f(t,\mathbf{X}) \geq 0). \tag{2}$$

This weighting scheme can also be understood in the context of nested case-control design. That is, select one subject from the event class, $\{i : dN_i(t) = 1\}$, at this interval and another subject from the non-event class, $\{i : dN_i(t) = 0\}$, using $E[dN(t)]/E[Y(t)]$ as the sampling weights for the latter. Consequently, an overall weighted prediction loss for the proposed SVHM, which is the expectation of (2) and ignores infinitesimal terms, is

$$\mathcal{R}_0(f) = E\left(\int Y(t)I\left[f(t,\mathbf{X}) \leq 0\right] dN(t)\right) + \int \frac{E\left(Y(t)I\left[f(t,\mathbf{X}) \geq 0\right]\right)}{E(Y(t))}E(dN(t)),$$

where the expectation is with respect to random variables $Y(t)$ and $dN(t)$. Our goal of learning a prediction rule for T, or equivalently, $N(t)$, based on the censored data is to minimize the population loss $\mathcal{R}_0(f)$.

To define the empirical loss, suppose there are m distinct ordered event times, $t_1 < t_2 < \ldots < t_m$. We let

$$\delta N_i(t_j) \equiv 2(N_i(t_j) - N_i(t_j-)) - 1$$

so $\delta N_i(t_j)$ takes values 1 or -1 depending on whether the ith subject experiences an event at t_j or not. Learning $f(t,\mathbf{x})$ becomes a sequence of binary classification problems over t_j's. Furthermore, at each t_j and for subject i at risk at t_j, we use the following weight associated with the risk set size at t_j:

$$w_i(t_j) = I\left\{\delta N_i(t_j) = 1\right\}\left\{1 - \frac{1}{\sum_{i=1}^n Y_i(t_j)}\right\} + I\left\{\delta N_i(t_j) = -1\right\}\left\{\frac{1}{\sum_{i=1}^n Y_i(t_j)}\right\}.$$

Note that the weights $w_i(t_j)$ are the empirical version of the weights used in (2) with similar interpretation as the reciprocal of the empirical probability of remaining event free or experiencing an event at the observed event time. Such weights balance the differential size of event class and non-event class at time t_j. Then an optimal decision function that minimizes the empirical version of $\mathcal{R}_0(f)$ is to minimize the following weighted total misclassification error:

$$\mathcal{R}_{0n}(f) = n^{-1}\sum_{i=1}^n\sum_{j=1}^m w_i(t_j)Y_i(t_j)I(\delta N_i(t_j)f(t_j,\mathbf{X}_i) < 0), \tag{3}$$

where the term $Y_i(t_j)$ reflects that only subjects still at risk will contribute towards prediction.

Directly minimizing (3) is difficult due to non-smoothness of the 0-1 loss in the indicator function. Furthermore, no restriction on the complexity of f leads to potential overfitting. To handle these issues, we adopt the same idea as SVM for supervised learning to replace the 0-1 loss in (1) by the hinge loss, and impose regularization to estimate f. Specifically, we propose to minimize the following regularized SVHM loss:

$$\mathcal{R}_n(f) + \lambda_n\|f\|^2,$$

$$\text{with } \mathcal{R}_n(f) \equiv n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} w_i(t_j) Y_i(t_j) [1 - f(t_j, \mathbf{X}_i) \delta N_i(t_j)]_+, \tag{4}$$

where $[1-x]_+ = \max(1-x, 0)$ is the hinge loss, $\|\cdot\|$ is a suitable norm or semi-norm for f to be discussed in the following sections, and λ_n is the regularization parameter. This minimization is equivalent to maximizing the margin between subjects in the event and non-event classes subject to an upper bound on the misclassification rate. Since this learning method is a weighted version of the standard support vector machines and learning $f(t, \mathbf{x})$ is essentially learning the hazard rate function, we refer our proposed method as "support vector hazards machine".

2.2 Learning Algorithm

Next, we describe the computational algorithm to solve the optimization in (4). We do not impose any restriction on $\alpha(t)$, and assume $g(\mathbf{x})$ lies in a reproducing kernel Hilbert space $\mathcal{H}_n$ with a kernel function $K(\mathbf{x}, \mathbf{x}')$. Commonly used kernels include linear kernel, where $K(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle$; radial basis kernel, where $K(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2/\sigma)$; and lth-degree polynomial kernel, where $K(\mathbf{x}, \mathbf{x}') = (1 + \langle \mathbf{x}, \mathbf{x}' \rangle)^l$. Furthermore, let $\|f\| = \|g\|_{\mathcal{H}_n}$ which is the norm in the reproducing kernel Hilbert space $\mathcal{H}_n$. The minimization in (3),

$$\min_{\alpha, g} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} w_i(t_j) Y_i(t_j) [1 - (\alpha(t_j) + g(\mathbf{X}_i)) \delta N_i(t_j)]_+ + \lambda_n \|g\|^2_{\mathcal{H}_n}, \tag{5}$$

is equivalent to

$$\min_{\alpha, g} \frac{1}{2} \|g\|^2_{\mathcal{H}_n} + C_n \sum_{i=1}^{n} \sum_{j=1}^{m} w_i(t_j) Y_i(t_j) \zeta_i(t_j), \tag{6}$$

$$\begin{gathered} \text{subject to } \; Y_i(t_j) \zeta_i(t_j) \geq 0, i = 1, \cdots, n, j = 1, \cdots, m, \\ Y_i(t_j) \delta N_i(t_j) \{\alpha(t_j) + g(\mathbf{X}_i)\} \geq Y_i(t_j) \{1 - \zeta_i(t_j)\}, i = 1, \cdots, n, j = 1, \cdots, m, \end{gathered}$$

where the value $\zeta_i(t_j)$ is the proportional amount by which the prediction is on the wrong side of its margin at time t_j, and C_n is the cost parameter.

The constrained optimization in (6) is usually solved by turning it into its dual form (through including Lagrange multipliers of the constraints into the objective function). We convert the above problem to its dual form by using the corresponding Lagrangian function

$$\begin{aligned} L_p \;\; = \;\; & \frac{1}{2} \|g\|^2_{\mathcal{H}_n} + C_n \sum_{i=1}^{n} \sum_{j=1}^{m} w_i(t_j) Y_i(t_j) \zeta_i(t_j) - \sum_{i=1}^{n} \sum_{j=1}^{m} \mu_{ij} Y_i(t_j) \zeta_i(t_j) \\ - \;\; & \sum_{i=1}^{n} \sum_{j=1}^{m} \gamma_{ij} [Y_i(t_j) \delta N_i(t_j) \{\alpha(t_j) + g(\mathbf{X}_i)\} - Y_i(t_j) \{1 - \zeta_i(t_j)\}], \end{aligned}$$

where $\mu_{ij} \geq 0$ and $\gamma_{ij} \geq 0$ are the corresponding Lagrange multipliers. Let $\{\phi_1, \phi_2, ...\}$ be the orthornomal basis system in $\mathcal{H}_n$ and $g(\mathbf{X}) = \sum_{k=1}^{\infty} \beta_k \phi_k(\mathbf{X})$. Then after differentiating

the Lagrangian function with respect to β's, $\alpha(t_j)$'s and $\zeta_i(t_j)$'s, we obtain

$$\beta_k = \sum_{i=1}^{n}\sum_{j=1}^{m} \gamma_{ij} Y_i(t_j)\delta N_i(t_j)\phi_k(\mathbf{X}_i), \quad k = 1, 2, \cdots,$$

$$\sum_{i=1}^{n} \gamma_{ij} Y_i(t_j)\delta N_i(t_j) = 0,$$

$$C_n w_i(t_j) Y_i(t_j) - \mu_{ij} Y_i(t_j) = \gamma_{ij} Y_i(t_j),\ i = 1, \cdots, n,\ j = 1, \cdots, m,$$

as well as the positivity constraints γ_{ij}, μ_{ij}, $\zeta_i(t_j) \geq 0$ for all i and j. By substituting these back to L_p and noting that $\sum_{k=1}^{\infty} \phi_k(\mathbf{X}_i)\phi_k(\mathbf{X}) = K(\mathbf{X}_i, \mathbf{X})$ (Theorem 4.2, Steinwart (2002)), we obtain the dual objective function to be

$$L_D = \sum_{i=1}^{n}\sum_{j=1}^{m} \gamma_{ij} Y_i(t_j) - \frac{1}{2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{j=1}^{m}\sum_{j'=1}^{m} \gamma_{ij}\gamma_{i'j'} Y_i(t_j) Y_{i'}(t_{j'})\delta N_i(t_j)\delta N_{i'}(t_{j'}) K(\mathbf{X}_i, \mathbf{X}_{i'}), \tag{7}$$

and the optimization is carried out by maximizing L_D with respective to γ_{ij} subject to $0 \leq \gamma_{ij} \leq w_i(t_j)C_n$ and $\sum_{i=1}^{n} \gamma_{ij} Y_i(t_j)\delta N_i(t_j) = 0$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. This optimization can be solved using quadratic programming packages available in many softwares (for example, MOSEK toolbox in Matlab). The tuning parameter C_n is chosen by cross-validation searching over a grid of values. Denote $\widehat{\gamma}_{ij}$ as the solutions for γ_{ij} obtained from the optimization procedure in (7). Comparing (7) with existing standard support vector machine algorithms, we see that the objective function sums across all at-risk subjects and across all time points for which they are at risk. Constraints are placed on those subjects and time points.

Next, from the equalities between β_k's and γ_{ij}'s in the above duality derivation, the solutions for β_k (denoted as $\widehat{\beta}_k$) are given by

$$\widehat{\beta}_k = \sum_{i=1}^{n}\sum_{j=1}^{m} \widehat{\gamma}_{ij} Y_i(t_j)\delta N_i(t_j)\phi_k(\mathbf{X}_i), \quad k = 1, 2, \cdots.$$

Thus, the solution for g that minimzies (5), which is the risk score for a future subject with baseline covariates $\mathbf{x}$, is

$$\widehat{g}(\mathbf{x}) = \sum_{k=1}^{\infty} \widehat{\beta}_k \phi_k(\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{m} \widehat{\gamma}_{ij} Y_i(t_j)\delta N_i(t_j) \sum_{k=1}^{\infty} \phi_k(\mathbf{X}_i)\phi_k(\mathbf{x})$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} \widehat{\gamma}_{ij}\delta N_i(t_j) K(\mathbf{x}, \mathbf{X}_i).$$

It follows that those data points with $\widehat{\gamma}_{ij} > 0$ form support vectors and determine $g(\mathbf{X})$.

Furthermore, to determine the solution to $\alpha(t_j)$ at each t_j, denoted by $\widehat{\alpha}(t_j)$, we solve the Karush-Kuhn-Tucker (KKT) conditions

$$\gamma_{ij}[Y_i(t_j)\delta N_i(t_j)\{\alpha(t_j) + g(\mathbf{X}_i)\} - Y_i(t_j)\{1 - \zeta_i(t_j)\}] = 0,$$

$$Y_i(t_j)\zeta_i(t_j) \geq 0,$$
$$Y_i(t_j)\delta N_i(t_j)\{\alpha(t_j) + g(\mathbf{X}_i)\} - Y_i(t_j)\{1 - \zeta_i(t_j)\} \geq 0.$$

Specifically, if there are some support vectors lying on the edge of the margin which are characterized by $0 < \widehat{\gamma}_{ij} < w_i(t_j)C_n$, $\widehat{\alpha}(t_j) = 1/\delta N_i(t_j) - \widehat{g}(\mathbf{X}_i)$ for these points, and we take the average of all the solutions for numerical stability. If all the support vectors at t_j are $\widehat{\gamma}_{ij} = C_n w_i(t_j)$, $\hat{\alpha}(t_j)$ is not unique and falls into a range

$$\min_{\substack{Y_i(t_j)=1, \hat{\gamma}_{ij}=C_n w_i(t_j), \\ \delta N_i(t_j)=1}} \{1 - \widehat{g}(\mathbf{X}_i)\} \geq \hat{\alpha}(t_j) \geq \max_{\substack{Y_i(t_j)=1, \hat{\gamma}_{ij}=C_n w_i(t_j), \\ \delta N_i(t_j)=-1}} \{-1 - \widehat{g}(\mathbf{X}_i)\}.$$

In this case, we take $\widehat{\alpha}(t_j) = 1 - \widehat{g}(\mathbf{X}_i)$ where $\delta N_i(t_j) = 1$ for some i with $Y_i(t_j) = 1$.

Since a higher value of the prediction function $\widehat{\alpha}(t) + \widehat{g}(x)$ leads to a greater likelihood of having an event at an earlier time, the magnitude of $\widehat{g}(x)$ induces a natural ordering of the risks. Lastly, the learned risk scores can be used to predict the event time for any future subjects using their baseline covariates $\mathbf{x}$. To this end, consider the nearest-neighbor prediction: for a future subject with $\mathbf{X} = \mathbf{x}$, find k (k=1 or 3 in our applications) non-censored subjects in the training data whose predictive scores are closest to $\widehat{g}(\mathbf{x})$, denoted as $\widehat{g}(\mathbf{X}_j)$. To maintain the monotone relationship between the event times and predictive scores, sort these scores of non-censored subjects in the training data in descending order and identify the rank of $\widehat{g}(\mathbf{X}_j)$. Next, sort the event times of the derived scores in the training data in ascending order and find the event times with the same rank as the rank of $\widehat{g}(\mathbf{X}_j)$, denoted as $T_{j'}$. The event time for this subject is predicted as $T_{j'}$ (or the average of $T_{j'}$ for $k = 3$). We provide a detailed description of SVHM algorithm in Appendix A.

2.3 Connection with Existing Support Vector-Based Approaches

Support vector-based approaches in machine learning literature are motivated by the fact that they are easy to compute and enable estimation under weak or no assumptions on the distribution. Most of these approaches (Shivaswamy et al., 2007; Van Belle et al., 2010, 2011) adapt the ϵ-insensitive loss for SVR to account for incomplete observations in time-to-event data. To improve performance, modified SVR (Van Belle et al., 2011) further places ranking constraints under the ϵ-insensitive loss. The formulation of the problem is

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\epsilon},\boldsymbol{\xi},\boldsymbol{\xi}^*} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \lambda_1 \sum_i \epsilon_i + \lambda_2 \sum_i (\xi_i + \xi_i^*), \qquad (8) \\ \text{subject to } \mathbf{w}^T(\varphi(\mathbf{X}_i) - \varphi(\mathbf{X}_{j(i)}) \geq Y_i - Y_{j(i)} - \epsilon_i, i = 1, \cdots, n, \\ \mathbf{w}^T\varphi(\mathbf{X}_i) + b \geq Y_i - \xi_i, i = 1, \cdots, n, \\ \Delta_i(\mathbf{w}^T\varphi(\mathbf{X}_i) + b) \geq -\Delta_i Y_i - \xi_i^*, i = 1, \cdots, n, \\ \epsilon_i \geq 0, \xi_i \geq 0, \xi_i^* \geq 0, i = 1, \cdots, n, \end{aligned}$$

where $Y_i = T_i \wedge C_i$, $\varphi(\cdot)$ is the feature map that does not need to be specified explicitly in a kernel-based method, and $j(i)$ indicates the subject with the largest event time smaller than Z_i. The first set of constraints above aims at ensuring rank consistency to maximize C-index for predicting survival outcomes, and the second and third sets of constraints are the

same as the regression constraints in Shivaswamy et al. (2007) for the modified ϵ-insensitive loss for survival outcomes. One potential problem with the above optimization is that the observations contributing to these three sets of constraints may consist of a selected (non-censored) sample from the full data; thus, the derived prediction rule will likely favor those observations which contribute most.

Furthermore, comparing the modified SVR in (8) with SVHM in (6), we see that the loss function for the former is the ϵ-insensitivity loss plus the loss resulting from violating rank consistency, while for the latter it is sum of a sequence of hinge losses. The objective function and the slack variables (i.e., $\epsilon_i, \xi_i, \xi_i^*$) for the modified SVR, however, are time-invariant, while the slack variables for SVHM (i.e., $\zeta_i(t_j)$ in (8)) are time-sensitive. Thus, we expect better control of the prediction error by SVHM. Note that this advantage stems from the counting process formulation of SVHM transforming prediction of time-to-event outcomes (or survival outcomes) as a sequence of binary prediction problems over time.

2.4 Connection with the Cox Partial Likelihood

In classical survival analysis using Cox regression model (Cox, 1972), partial likelihood plays a central role since it only involves association parameter of interest (i.e., hazard ratios as regression coefficients) but not the nuisance parameter (i.e., baseline hazard function), and maximizing the partial likelihood directly estimates the hazard ratios. The partial likelihood is constructed by multiplying together the conditional probabilities of observing an event for individual i at time t, given the past and given that an event is observed at that time, over all observed event times. This conditional probability formulation shares some similarity with our hazard formulation for SVHM. Since maximizing Cox partial likelihood leads to regression estimators that enjoy optimal statistical property (i.e., being semiparametric efficient, Bickel et al. (1998)), it is worth to draw connection between SVHM and partial likelihood to shed lights on the theoretical properties of SVHM.

To this end, we further explore the optimization in (5) to compare the SVHM objective function and the Cox partial likelihood. First note that the function $\alpha(t)$ in (5) is analogous to the baseline hazard function in the Cox model (Cox, 1972), which is treated as a nuisance parameter and profiled out for inference. Thus, we also profile out $\alpha(t)$ to investigate the profile risk function for SVHM (e.g., substitute fitted $\alpha(t)$ in the original risk function). For a fixed $g(\mathbf{x})$, from the derivation similar to Hastie et al. (2009) (p421) and Abe (2010) (p77), we can show that at each t_j, if there are some support vectors lying on the edge of the margin which are characterized by $0 < \gamma_{ij} < w_i(t_j)C_n$, these margin points can be used to solve for $\alpha(t_j)$. This yields

$$\widehat{\alpha}(t_j) = 1 - g(\mathbf{X}_i), \quad \delta N_i(t_j) = 1.$$

Note that $\mathbf{X}_i$ is the covariate value for the subject who has an event at t_j. However, if γ_{ij} is not within $(0, w_i(t_j)C_n$, $\widehat{\alpha}(t_j)$ can be any value satisfying

$$\min_{\substack{\widehat{\gamma}_{ij}=C_n w_i(t_j), \\ \delta N_i(t_j)=1}} \{1 - g(\mathbf{X}_i)\} \geq \alpha(t_j) \geq \max_{\substack{\widehat{\gamma}_{ij}=C_n w_i(t_j), \\ \delta N_i(t_j)=-1}} \{-1 - g(\mathbf{X}_i)\}.$$

In this case, taking $\widehat{\alpha}(t_j) = 1 - g(\mathbf{X}_i)$ where $\delta N_i(t_j) = 1$ satisfies these constraints. Further note that optimizing (5) is equivalent to minimizing

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} Y_i(t_j)w_i(t_j)[1-(\alpha(t)+g(\mathbf{X}_i))\delta N_i(t_j)]_+ + \lambda_n\|g\|_{\mathcal{H}_n} \\
= \; & \frac{1}{n}\sum_{i=1}^{n}\int [1-(\alpha(t)+g(\mathbf{X}_i))]_+ dN_i(t) + \frac{1}{n}\int \frac{\sum_{i=1}^{n} Y_i(t)[1+(\alpha(t)+g(\mathbf{X}_i))]_+}{\sum_{i=1}^{n} Y_i(t)} d\left\{\sum_{i=1}^{n} N_i(t)\right\} \\
& -\frac{1}{n}\sum_{i=1}^{n}\int \frac{1}{\sum_{i=1}^{n} Y_i(t)}\left([1-(\alpha(t)+g(\mathbf{X}_i))]_+ + [1+(\alpha(t)+g(\mathbf{X}_i))]_+\right) dN_i(t) + \lambda_n\|g\|_{\mathcal{H}_n}.
\end{aligned}$$

After we plug the expression of $\widehat{\alpha}(t) = \sum_{i=1}^{n}(1-g(X_i))I(\delta N_i(t)=1)$ into the above expression, we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\int [1-(\widehat{\alpha}(t)+g(\mathbf{X}_i))]_+ dN_i(t) = \frac{1}{n}\sum_{i=1}^{n}\Delta_i[1-(1-g(\mathbf{X}_i)+g(\mathbf{X}_i))])+ = 0,$$

and similarly,

$$\frac{1}{n}\sum_{i=1}^{n}\int \frac{1}{\sum_{i=1}^{n} Y_i(t)}\left([1-(\widehat{\alpha}(t)+g(\mathbf{X}_i))]_+ + [1+(\widehat{\alpha}(t)+g(\mathbf{X}_i))]_+\right) dN_i(t) = \frac{2}{n}\sum_{i=1}^{n}\int \frac{dN_i(t)}{\sum_{i=1}^{n} Y_i(t)}.$$

Additionally,

$$\begin{aligned}
& \frac{1}{n}\int \frac{\sum_{i=1}^{n} Y_i(t)[1+(\widehat{\alpha}(t)+g(\mathbf{X}_i))]_+}{\sum_{i=1}^{n} Y_i(t)} d\left\{\sum_{i=1}^{n} N_i(t)\right\} \\
= & \frac{1}{n}\sum_{k=1}^{n}\Delta_k \frac{\sum_{i=1}^{n} I(Y_i \geq Y_k)[1+(\widehat{\alpha}(\mathbf{X}_k)+g(\mathbf{X}_j))]_+}{\sum_{i=1}^{n} I(Y_i \geq Y_k)} \\
= & \frac{1}{n}\sum_{k=1}^{n}\frac{\sum_{i=1}^{n} I(Y_i \geq Y_k)[2-g(X_k)+g(X_i)]_+}{\sum_{i=1}^{n} I(Y_i \geq Y_k)}\Delta_k.
\end{aligned}$$

The objective function (5) can be written as $\mathcal{PR}_n(g) + \lambda_n\|g\|^2_{\mathcal{H}_n}$, where

$$\begin{aligned}
\mathcal{PR}_n(g) &= \frac{1}{n}\sum_{i=1}^{n}\int \frac{\sum_{k=1}^{n} Y_k(t)[2-g(\mathbf{X}_i)+g(\mathbf{X}_k)]_+}{\sum_{k=1}^{n} Y_k(t)} dN_i(t) - \frac{2}{n}\sum_{i=1}^{n}\int \frac{dN_i(t)}{\sum_{k=1}^{n} Y_k(t)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\Delta_i \frac{\sum_{k=1}^{n} I(Y_k \geq Y_i)[2-g(\mathbf{X}_i)+g(\mathbf{X}_k)]_+}{\sum_{k=1}^{n} I(Y_k \geq Y_i)} - \frac{2}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\sum_{k=1}^{n} I(Y_k \geq Y_i)} \\
&= \mathbf{P}_n\left(\Delta\frac{\widetilde{\mathbf{P}}_n\{I(\widetilde{Y} \geq Y)[2+g(\widetilde{\mathbf{X}})-g(\mathbf{X})]_+\}}{\widetilde{\mathbf{P}}_n[I(\widetilde{Y} \geq Y)]}\right) - \frac{2}{n}\mathbf{P}_n\left\{\frac{\Delta}{\widetilde{\mathbf{P}}_n[I(\widetilde{Y} \geq Y)]}\right\}.
\end{aligned}$$

Here, $\mathbf{P}_n$ denotes the empirical measure from n observations and $\widetilde{\mathbf{P}}_n$ is the empirical measure applied to $(\widetilde{Y}, \widetilde{\mathbf{X}}, \widetilde{\Delta})$, an i.i.d copy of $(Y, \mathbf{X}, \Delta)$. Thus, $\widehat{g}(\mathbf{x})$ minimizes $\mathcal{PR}_n(g) + \lambda_n\|g\|^2_{\mathcal{H}_n}$.

If we let $\widehat{f}(x,t) = \widehat{\alpha}(t) + \widehat{g}(\mathbf{x})$ be the function minimizing (5) over $g \in \mathcal{H}_n$, then $\mathcal{R}_n(\widehat{f}) = \mathcal{PR}_n(\widehat{g})$.

In a Cox partial likelihood function, $g(\mathbf{X})$ is estimated by minimizing

$$\mathbf{P}_n \left(\Delta \log \frac{\widetilde{\mathbf{P}}_n\{I(\widetilde{Y} \geq Y) \exp\{g(\widetilde{\mathbf{X}}) - g(\mathbf{X})\}\}}{\widetilde{\mathbf{P}}_n[I(\widetilde{Y} \geq Y)]} \right).$$

Therefore, it is worthy to point out one interesting observation: both $\mathcal{PR}_n(g)$ and the Cox partial likelihood take a similar form which essentially evaluates a loss comparing the risk scores from the subjects at risk versus the one from the subject who has an event at the same time. SVHM uses a hinge loss while Cox partial likelihood uses an exponential loss and a logarithm transformation, which is similar to the contrast between SVM and logistic regression. The robustness of hinge loss compared to exponential loss suggests SVHM will be less sensitive to extreme observations. In addition, this connection sheds lights on the theoretical optimality of SVHM which we prove in the next section.

3. Theoretical Properties

In this section, we study the asymptotic properties of SVHM and the predicted risk. We first derive the population risk function for the proposed SVHM. Next, we derive the optimal fully nonparametric decision rule for this risk function and show that it also optimizes the 0-1 loss corresponding to (3). We highlight important differences in the theoretical proof that distinguish this work from the standard proofs in the statistical learning theories.

3.1 Risk Function and Optimal Risk Classification Rule

To derive the population risk function for SVHM, we first examine the population version (the expectation) of $\mathcal{R}_n(f)$. Recall the definition of $\mathcal{R}_n(f)$ is given in (4) as

$$\mathcal{R}_n(f) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{I\{\delta N_i(t_j) = 1\}(\sum_{l=1}^{n} Y_l(t_j) - 1)}{\sum_{l=1}^{n} Y_l(t_j)} [1 - f(t_j, \mathbf{X}_i)]_+$$

$$+ n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{I\{\delta N_i(t_j) = 1\}}{\sum_{l=1}^{n} Y_l(t_j)} [1 + f(t_j, \mathbf{X}_i)]_+.$$

After re-arranging the terms and adopting counting process notation, we can rewrite $\mathcal{R}_n(f)$ as

$$\begin{aligned} \mathcal{R}_n(f) \;=\; & \frac{1}{n} \sum_{i=1}^{n} \int Y_i(t)[1 - f(t, \mathbf{X}_i)]_+ dN_i(t) + \frac{1}{n} \int \frac{\sum_{i=1}^{n} Y_i(t)[1 + f(t, \mathbf{X}_i)]_+}{\sum_{i=1}^{n} Y_i(t)} d \left\{ \sum_{i=1}^{n} N_i(t) \right\} \\ & - \frac{1}{n} \sum_{i=1}^{n} \int \frac{1}{\sum_{i=1}^{n} Y_i(t)} \left([1 - f(t, \mathbf{X}_i)]_+ + [1 + f(t, \mathbf{X}_i)]_+ \right) dN_i(t). \end{aligned}$$

Note that the last term in $\mathcal{R}_n(f)$ is on the order of $O(1/n)$, so it vanishes as n goes to infinity. By the central limit theorem, we obtain the asymptotic limit of $\mathcal{R}_n(f)$, denoted as

$\mathcal{R}(f)$, to be

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

Likewise, similar arguements show that the empirical risk based on the prediction error in (1), i.e., $\mathcal{R}_{0n}(f)$, converges to $\mathcal{R}_0(f)$.

Let $f^*(t, \mathbf{x})$ denote the limit of the risk function estimated by SVHM (i.e., the optimal function minimizing $\mathcal{R}(f)$). Since the difference between $\mathcal{R}(f)$ and $\mathcal{R}_0(f)$ is the hinge loss versus the zero-one loss, one question is whether $f^*(t, \mathbf{x})$ also minimizes $\mathcal{R}_0(f)$. The following theorem gives such a result for $f^*(t, \mathbf{x})$.

Theorem 3.1 *Let $h(t, \mathbf{x})$ denote the conditional hazard rate function of $T = t$ given $\mathbf{X} = \mathbf{x}$ and let $\bar{h}(t) = E[dN(t)/dt]/E[Y(t)] = E[h(t, \mathbf{X})|Y(t) = 1]$ be the average hazard rate at time t. Then $f^*(t, \mathbf{x}) = sign(h(t, \mathbf{x}) - \bar{h}(t))$ minimizes $\mathcal{R}(f)$. Furthermore, $f^*(t, \mathbf{x})$ also minimizes $\mathcal{R}_0(f)$ and*

$$\mathcal{R}_0(f^*) = P(T \leq C) - \frac{1}{2} E\left[\int E(Y(t)|\mathbf{X} = \mathbf{x})|h(t, \mathbf{x}) - \bar{h}(t)|dt\right].$$

In addition, for any $f(t, \mathbf{x}) \in [-1, 1]$,

$$\mathcal{R}_0(f) - \mathcal{R}_0(f^*) \leq \mathcal{R}(f) - \mathcal{R}(f^*),$$

where $h(t, \mathbf{x})$ denotes the conditional hazard rate of $T = t$ given $\mathbf{X} = \mathbf{x}$ and $\bar{h}(t)$ is the population average hazard at time t,

$$\bar{h}(t) = \frac{E[dN(t)]/dt}{E[Y(t)]} = E[h(t, \mathbf{X})|Y(t) = 1].$$

The proof of Theorem 3.1 is provided in the Appendix B. Theorem 3.1 resembles the excess risk in most learning theories (Bartlett et al., 2006); however, the loss function in our case is some composite expectation, $\mathcal{R}_0(f)$, which is not covered by Bartlett et al. (2006). From Theorem 3.1, we see that the optimal rule is essentially to predict whether an at-risk subject will experience an event by comparing the subject-specific hazard rate depending on the covariate to the population-average hazard rate obtained from all at-risk subjects at a given time point. Since the minimizer of $\mathcal{R}(f)$ also minimizes $\mathcal{R}_0(f)$, this theory justifies the use of hinge-loss in SVHM to minimize the weighted prediction error in $\mathcal{R}_0(f)$. The last inequality in Theorem 3.1 proves that a decision function with a small excess hinge-loss–based risk will lead to a small excess 0-1 loss–based risk.

3.2 Asymptotic Properties

Here, we study the asymptotic properties of SVHM when the decision function takes the form in (1). Specifically, we examine a stochastic bound for the excess risk when using $\widehat{g}$, the estimator from n observations. This bound will be given in terms of the sample size n, the tuning parameter λ_n and the bandwidth of the kernel function σ_n. Denote $\mathcal{H}_n$ as

a reproducing kernel Hilbert space from a Gaussian kernel $k(x,x') = \exp\{-\|x-x'\|^2/\sigma_n\}$. Instead of considering the risk for $\mathcal{R}(f)$, we consider

$$\mathcal{PR}(g) = \min_{\alpha(t)} \mathcal{R}(\alpha(t) + g(\mathbf{x}))$$

and refer it as "profile risk", since $\alpha(t)$ is profiled out from the original risk function. In other words, $\mathcal{PR}(g)$ is the best expected risk for a given score $g(\mathbf{x})$ after accounting for $\alpha(t)$.

To obtain an explicit expression of $\mathcal{PR}(g)$, we first note that

$$\begin{aligned}
\mathcal{R}(\alpha(t)+g(\mathbf{x})) &= E\left(\int Y(t)[1-f(t,\mathbf{X})]_+ dN(t)\right) + \int \frac{E\left(Y(t)[1+f(t,\mathbf{X})]_+\right)}{E(Y(t))} E(dN(t)) \\
&= \int E[Y(t)h(t,\mathbf{X})]\left[\frac{E[Y(t)h(t,\mathbf{X})] - E[Y(t)g(\mathbf{X})h(t,\mathbf{X})]}{E[Y(t)h(t,\mathbf{X})]} - \alpha(t)\right]_+ dt \\
&\quad + \int \bar{h}(t) E[Y(t)]\left[\frac{E[Y(t)] + E[Y(t)g(\mathbf{X})]}{E[Y(t)]} + \alpha(t)\right]_+ dt.
\end{aligned}$$

Since $\alpha(t)$ is arbitrary and the integrand in the above expression is a piecewise linear function of $\alpha(t)$, simple algebra gives that

$$\alpha(t) = -\frac{E[Y(t)] + E[Y(t)g(\mathbf{X})]}{E[Y(t)]}$$

minimizes $\mathcal{R}(f)$. Therefore, after replacing $\alpha(t)$ by this minimizer in $\mathcal{R}(\alpha(t)+g(\mathbf{x}))$, we obtain

$$\mathcal{PR}(g) = E\left[\Delta \frac{\widetilde{\mathbf{P}} I(\tilde{Y} \geq Y)[2 - g(\widetilde{\mathbf{X}}) + g(\mathbf{X})]_+}{\widetilde{\mathbf{P}} I(\tilde{Y} \geq Y)}\right].$$

Clearly, $\mathcal{PR}(g)$ is the asymptotic limit of $\mathcal{PR}_n(g)$. The following theorem holds for the risk $\mathcal{PR}(\widehat{g})$.

Theorem 3.2 *Assume that* $\mathbf{X}$*'s support is compact and* $E[Y(\tau)|\mathbf{X}]$ *is bounded from zero where* τ *is the study duration. Furthermore, assume* λ_n *and* σ_n *satisfies* $\lambda_n, \sigma_n \to 0$*, and* $n\lambda_n \sigma_n^{(2/p-1/2)d} \to \infty$ *for some* $p \in (0,2)$*. Then it holds*

$$\lambda_n \|\widehat{g}\|^2_{\mathcal{H}_n} + \mathcal{PR}(\widehat{g}) \leq \inf_g \mathcal{PR}(g) + O_p\left\{\lambda_n + \sigma_n^{d/2} + \frac{\lambda_n^{-1/2}\sigma_n^{-(1/p-1/4)d}}{\sqrt{n}}\right\}.$$

The proof of Theorem 3.2 mostly follows the machinery for support vector machines. It mainly uses empirical process theories to control the stochastic error of the empirical risk functions and the approximation properties of the reproducing kernel Hilbert space based on the Gaussian kernel function. However, one major difference from the classical proof is that the empirical loss function we study here is some composite statistics instead of the summation of n i.i.d terms. This poses additional challenges to control stochastic variability. The constants in Theorem 3.2 imply that the bandwidth for the Gaussian

kernel and regularization parameter should converge to zero in certain rates depending on $\mathbf{X}$'s dimension, but not too fast to ensure stochastic variability is under control. Finally, we state two useful observations as remarks below.

Remark 1. From Theorem 3.2, if we choose $\sigma_n = (n\lambda_n)^{-1/[2d(1/p+1/4)]}$, it gives

$$\mathcal{PR}(\widehat{g}) - \mathcal{PR}(g^*) = O_p\left\{\lambda_n + (n\lambda_n)^{-q}\right\},$$

where $q = 1/(4/p+1)$ and g^* is the function minimizing $\mathcal{PR}(g)$.

Remark 2. Furthermore, if we choose $\lambda_n = n^{-q/(q+1)}$, then the optimal rate obtained from Theorem 3.2 becomes

$$\mathcal{PR}(\widehat{g}) - \mathcal{PR}(g^*) = O(n^{-q/(q+1)}).$$

4. Numeric Examples

In this section, we first present simulation results comparing SVHM to existing machine learning approaches and semiparametric approaches based on the Cox proportional hazards regression. Next, we provide applications to two real world empirical studies.

4.1 Simulation Studies

In all scenarios, we generated both event times and censoring times to be dependent on the covariates. First we simulated five covariates $\mathbf{Z} = (Z_1, \ldots, Z_5)$ which are marginally normal $N(0, 0.5^2)$ with pairwise correlation $\text{corr}(Z_j, Z_k) = \rho^{|j-k|}$, and $\rho = 0.5$. The event times were generated from the Cox proportional hazards model with true $\boldsymbol{\beta} = (2, -1.6, 1.2, -0.8, 0.4)^T$ and the exponential distribution with $\lambda = 0.25$ was assumed to compute the baseline cumulative hazard function $\Lambda(t) = \int_0^t \lambda_0(s)ds$, where $\lambda_0(s)$ is the baseline hazard function. We simulated two types of censoring distributions. In the first type, the censoring times were generated from an accelerated failure time model following the log-normal distribution, that is, $\log C \sim N(\mathbf{Z}^T\boldsymbol{\beta}_c + a, 0.5^2)$, with true $\boldsymbol{\beta}_c = (1, 1, 1, 1, 1)^T$. In the second type, the distribution of the censoring times follows the Cox proportional hazards model with true $\boldsymbol{\beta}_c = (1, 1, 1, -2, -2)^T$ and the baseline cumulative hazard function $\Lambda_c(t) = bt$ $(b > 0)$. The parameters a and b were chosen to obtain the desired censoring ratio. We considered the censoring ratios 40% and 60%. Any event times or censoring times greater than u_0 were truncated at u_0, where u_0 is the 90th percentile of the event times. Moreover, we explored some generalizations of the above scenarios to include more covariates in the regression models and include additional noise variables. Besides these training data sets (with a sample size of 100 or 200), we use a randomly generated testing data set with $10,000$ subjects in each scenario with no censoring to evaluate prediction performance of various methods.

For all scenarios, we compared SVHM with the modified support vector regression for right censored data based on the ranking constraints (modified SVR) (Van Belle et al., 2011) and the inverse-probability-of-censoring weighting with censoring distribution estimated using Kaplan-Meier (IPCW-KM) or estimated under a Cox model (IPCW-Cox) (Goldberg and Kosorok, 2013), whose objective function is defined as

$$n^{-1}\sum_{i=1}^{n}\frac{\Delta_i}{\widehat{S}(Y_i)}(\log Y_i - \mathbf{x}^T\phi(\mathbf{X}_i))^2$$

with $\widehat{S}(t)$ is the estimated survival probability for the censoring time. We used linear kernel $K(x, x') = \mathbf{x}^T\mathbf{x}'$ in all four methods, and used 5-fold cross-validation to choose the tuning parameters from the grid of $\{2^{-16}, 2^{-15}, \ldots, 2^{15}, 2^{16}\}$. As the model comparison criterion, we adapted mean squared error to censored data, which sums up the mean squared differences between the fitted event times and observed event times for uncensored subjects. For censored subjects, we sum up the squared differences between fitted times and censoring times if the former is smaller than the latter. Essentially, for these censored subjects, if their predicted event times were less than the observed censoring times, we imposed a penalty to measure how much under-estimation there is. The mean squared differences were assumed to be zero for censored subjects if their predicted values were greater than the observed censoring times. We divided the total sum of squares by the total number of observations. We repeated the simulation 500 times, since our results show that 500 repetitions are sufficient to obtain stable simulation results to draw conclusions on comparing performance of different methods while still achieving computational efficiency.

Table 1 and 2 give the average Pearson correlations and root mean square errors (RMSE) $\{\sum(\widehat{T} - T)^2\}^{1/2}$ based on the fitted event times and observed event times T on the testing data set. Larger correlation and smaller root mean squared error indicate better performance. The results show that SVHM outperforms the other methods for all the simulation cases and sample sizes. The advantages are not affected by including 5 or 15 noise variables, and the improvements become more evident when the censoring rate is 60% or the censoring distribution follows the accelerated failure time model. The columns of the average correlations show that the modified SVR has similar capability to capture the rank information as SVHM. However, it gives less accurate prediction of the exact event times as measured by the higher RMSEs. The IPCW methods have the worst performance, no matter using the Kaplan-Meier estimator or Cox model to estimate the censoring distribution, even when the censoring distribution follows the Cox model. The performances of all the methods are improved as the sample size increases from 100 to 200, and the proposed SVHM has the largest improvement with respect to the ratios of the average RMSEs. The RMSE of SVHM is significantly lower than the best competing method in all simulation settings in Table 1 and 2. Correlation between the risk scores and event times for SVHR is not significantly different from modified SVR, but in the first simulation setting it is significantly higher than two IPW-based methods except when there are 95 noise variables (Table 1). In the second simulation setting, difference between SVHR and IPW is smaller, with the former significantly greater for most cases with $n = 200$ (Table 2).

In conclusion, Table 1 shows that SVHM performs much better than Cox regression when the model assumption does not hold, and Table 2 shows that SVHM still maintains its advantage when the Cox proportional hazards assumption holds. This advantage may be due to that Cox model aims at maximizing the likelihood while SVHM directly aims at discriminating individual's risk and prediction.

We also explored SVHM with a Gaussian kernel for the sample size 100 and the computation is more intensive. The resulting average correlations and RMSEs are similar to those for linear kernel. For example, under the setting in Table 1 with 60% censoring rate, no noise variable and $n = 100$, using Gaussian kernel yields almost similar correlation of 0.48, 0.10, 0.15, and 0.53 for four competing methods (modified SVR, IPCW-KM, IPCW-Cox, SVHM), respectively. The corresponding RMSEs are 6.03 6.62, 6.75, and 5.26, respectively

Censoring	# of Noises	Method	$n = 100$			$n = 200$		
			CORR[a]	RMSE[b](SD[c])	Ratio[d]	CORR	RMSE (SD)	Ratio
40%	0	Modified SVR	0.59	5.59 (0.60)	1.19	0.62	5.58 (0.58)	1.24
		IPCW-KM[e]	0.40	5.60 (0.52)	1.20	0.45	5.45 (0.41)	1.21
		IPCW-Cox[f]	0.43	5.80 (0.64)	1.24	0.50	5.62 (0.57)	1.25
		SVHM	**0.61**[g]	**4.68 (0.27)**	1.00	**0.64**	**4.49 (0.17)**	1.00
	5	Modified SVR	0.55	5.64 (0.60)	1.15	0.61	5.63 (0.57)	1.22
		IPCW-KM	0.32	5.93 (0.47)	1.21	0.42	5.63 (0.44)	1.22
		IPCW-Cox	0.33	6.17 (0.54)	1.26	0.44	5.87 (0.57)	1.27
		SVHM	**0.58**	**4.90 (0.35)**	1.00	**0.63**	**4.62 (0.20)**	1.00
	15	Modified SVR	0.46	5.73 (0.47)	1.10	0.54	5.55 (0.50)	1.15
		IPCW-KM	0.21	6.12 (0.32)	1.18	0.31	5.86 (0.34)	1.22
		IPCW-Cox	0.20	6.47 (0.46)	1.24	0.32	6.09 (0.47)	1.26
		SVHM	**0.48**	**5.20 (0.36)**	1.00	**0.57**	**4.82 (0.23)**	1.00
	95[h]	Modified SVR	0.21	6.65 (0.89)	1.10	0.30	6.29 (0.47)	1.09
		IPCW-KM	0.06	6.33 (0.21)	1.05	0.10	6.28 (0.14)	1.09
		IPCW-Cox	0.08	6.59 (0.23)	1.09	0.11	6.61 (0.39)	1.15
		SVHM	**0.22**	**6.04 (0.32)**	1.00	**0.32**	**5.76 (0.25)**	1.00
60%	0	Modified SVR	0.55	6.00 (0.54)	1.16	0.60	6.07 (0.42)	1.24
		IPCW-KM	0.15	6.45 (0.41)	1.25	0.18	6.42 (0.37)	1.32
		IPCW-Cox	0.21	6.56 (0.47)	1.27	0.26	6.47 (0.48)	1.33
		SVHM	**0.57**	**5.18 (0.43)**	1.00	**0.61**	**4.88 (0.33)**	1.00
	5	Modified SVR	0.50	6.06 (0.53)	1.12	0.57	6.07 (0.50)	1.21
		IPCW-KM	0.11	6.61 (0.34)	1.22	0.15	6.56 (0.32)	1.31
		IPCW-Cox	0.15	6.77 (0.39)	1.25	0.21	6.66 (0.39)	1.33
		SVHM	**0.51**	**5.40 (0.48)**	1.00	**0.58**	**5.02 (0.33)**	1.00
	15	Modified SVR	0.39	6.14 (0.45)	1.10	0.49	5.97 (0.45)	1.15
		IPCW-KM	0.07	6.56 (0.30)	1.17	0.10	6.54 (0.24)	1.26
		IPCW-Cox	0.10	6.82 (0.30)	1.22	0.13	6.70 (0.27)	1.29
		SVHM	**0.40**	**5.60 (0.44)**	1.00	**0.51**	**5.20 (0.36)**	1.00
	95	Modified SVR	**0.17**	6.90 (1.08)	1.11	0.25	7.20 (1.52)	1.21
		IPCW-KM	0.01	6.53 (0.26)	1.05	0.03	6.54 (0.20)	1.10
		IPCW-Cox	0.02	6.87 (0.20)	1.10	0.04	6.86 (0.21)	1.15
		SVHM	**0.17**	**6.22 (0.24)**	1.00	**0.26**	**5.94 (0.25)**	1.00

[a] CORR, average value of correlations.
[b] RMSE, average value of root mean square errors.
[c] Empirical standard deviation of the RMSE across 500 repetitions
[d] Ratio, ratio of average root mean square errors between the method used and our method.
[e] IPCW-KM, IPCW using the Kaplan-Meier estimator for the censoring distribution.
[f] IPCW-Cox, IPCW using the Cox model for the censoring distribution.
[g] Entries in boldface highlight the best performance method.
[h] For the cases of 95 noises, the calculation of inverse weights in the IPCW-Cox method uses only five signal variables to fit the Cox model for the censoring times.

Table 1: Comparison of four support vector learning methods for right censored data using a linear kernel, with censoring times following the accelerated failure time (AFT) model

Censoring	# of Noises	Method	$n = 100$ CORR[a]	$n = 100$ RMSE[b](SD[c])	$n = 100$ Ratio[d]	$n = 200$ CORR	$n = 200$ RMSE (SD)	$n = 200$ Ratio
40%	0	Modified SVR	0.59	5.15 (0.59)	1.11	0.62	5.09 (0.54)	1.12
		IPCW-KM[e]	0.53	5.16 (0.42)	1.11	0.55	5.08 (0.31)	1.12
		IPCW-Cox[f]	0.52	5.31 (0.57)	1.14	0.56	5.09 (0.46)	1.12
		SVHM	**0.61**[g]	**4.66 (0.25)**	1.00	**0.63**	**4.53 (0.16)**	1.00
	5	Modified SVR	0.56	5.28 (0.51)	1.08	0.61	5.09 (0.50)	1.12
		IPCW-KM	0.46	5.58 (0.42)	1.14	0.52	5.27 (0.34)	1.13
		IPCW-Cox	0.44	5.73 (0.52)	1.17	0.51	5.41 (0.51)	1.16
		SVHM	**0.58**	**4.89 (0.29)**	1.00	**0.62**	**4.65 (0.18)**	1.00
	15	Modified SVR	0.47	5.43 (0.40)	1.04	0.55	5.14 (0.38)	1.06
		IPCW-KM	0.36	5.79 (0.34)	1.11	0.44	5.49 (0.30)	1.13
		IPCW-Cox	0.34	6.00 (0.40)	1.15	0.42	5.70 (0.43)	1.18
		SVHM	**0.49**	**5.21 (0.33)**	1.00	**0.57**	**4.84 (0.20)**	1.00
	95[h]	Modified SVR	0.21	6.43 (0.92)	1.04	0.33	6.03 (0.54)	1.04
		IPCW-KM	0.17	6.16 (0.21)	1.00	0.24	6.06 (0.18)	1.05
		IPCW-Cox	0.16	6.32 (0.23)	1.02	0.22	6.21 (0.22)	1.07
		SVHM	**0.23**	**6.18 (0.40)**	1.00	**0.34**	**5.78 (0.24)**	1.00
60%	0	Modified SVR	0.56	5.43 (0.56)	1.08	0.59	5.43 (0.47)	1.12
		IPCW-KM	0.44	5.68 (0.43)	1.13	0.46	5.62 (0.33)	1.16
		IPCW-Cox	0.42	5.83 (0.56)	1.16	0.47	5.67 (0.48)	1.17
		SVHM	**0.57**	**5.01 (0.37)**	1.00	**0.60**	**4.85 (0.25)**	1.00
	5	Modified SVR	0.50	5.61 (0.48)	1.07	0.57	5.40 (0.46)	1.09
		IPCW-KM	0.36	6.02 (0.38)	1.15	0.43	5.79 (0.35)	1.17
		IPCW-Cox	0.34	6.25 (0.44)	1.20	0.41	5.96 (0.47)	1.20
		SVHM	**0.53**	**5.23 (0.37)**	1.00	**0.59**	**4.96 (0.27)**	1.00
	15	Modified SVR	0.40	5.77 (0.42)	1.05	0.50	5.44 (0.38)	1.06
		IPCW-KM	0.27	6.07 (0.31)	1.10	0.35	5.94 (0.26)	1.16
		IPCW-Cox	0.25	6.39 (0.40)	1.16	0.32	6.16 (0.33)	1.20
		SVHM	**0.42**	**5.51 (0.40)**	1.00	**0.52**	**5.13 (0.29)**	1.00
	95	Modified SVR	0.18	6.47 (0.87)	1.05	0.27	6.31 (0.80)	1.05
		IPCW-KM	0.12	6.22 (0.29)	1.01	0.18	6.19 (0.21)	1.03
		IPCW-Cox	0.12	6.54 (0.26)	1.07	0.16	6.50 (0.23)	1.08
		SVHM	**0.20**	**6.14 (0.38)**	1.00	**0.28**	**6.00 (0.35)**	1.00

[a] CORR, average value of correlations.
[b] RMSE, average value of root mean square errors.
[c] Empirical standard deviation of the RMSE across 500 repetitions
[d] Ratio, ratio of average root mean square errors between the method used and our method.
[e] IPCW-KM, IPCW using the Kaplan-Meier estimator for the censoring distribution.
[f] IPCW-Cox, IPCW using the Cox model for the censoring distribution.
[g] Entries in boldface highlight the best performance method.
[h] For the cases of 95 noises, the calculation of inverse weights in the IPCW-Cox method uses only five signal variables to fit the Cox model for the censoring times.

Table 2: Comparison of four support vector learning methods for right censored data using a linear kernel, with censoring times following the Cox proportional hazards model

for each method. Under the setting in Table 2 with 60% censoring rate, no noise variable and $n = 100$, the correlations for the four methods are 0.52, 0.42, 0.40, and 0.55, respectively, and the RMSEs are 5.52, 5.76, 5.94, and 5.06.

In our next simulation experiment, we compare SVHM with Cox model based analysis and explore 1-nearest-neighbor (1-NN) prediction and the average of 3-nearest-neighbors (3-NN) prediction. In the first setting we generate five discrete covariates $\mathbf{Z} = (Z_1, \ldots, Z_5)$ with equal probability of taking each value: Z_1 takes values -5, -4, -2, -1 or 0; Z_2 takes values -1, 0 or 1; Z_3 takes integer values 1 to 10; Z_4 has a correlation of 0.5 with Z_1 and is also correlated with a random normal noise variable $N(0, 0.5)$, and Z_5 has a correlation of 0.3 with Z_1 and is also correlated with a random uniform noise variable $U(0, 0.5)$. Similar to the previous simulations, the event times were generated from Cox proportional hazards model with true $\boldsymbol{\beta} = (2, -1.6, 1.2, -0.8, 0.4)^T$ and the exponential distribution with $\lambda = 0.25$ was assumed for the baseline cumulative hazard function $\Lambda(t)$. The distribution of the censoring times followed Cox proportional hazards model with true $\boldsymbol{\beta}_c = (1, 1, 1, -2, -2)^T$ and the baseline hazard rate was a constant. In the second setting, we generated $Z_1, ..., Z_3$ independently from $U(0, 1)$ and Z_4 from a binary distribution with $P(Z_4 = 1) = P(Z_4 = -1) = 0.5$. Furthermore, both the event times and censoring times were generated from accelerated failure time models with both main effects and interactions:

$$\log T = -0.2 - 0.5Z_1 + 0.5Z_2 + 0.3Z_3 + 0.5Z_4 - 0.1Z_1Z_4 - 0.6Z_2Z_4 + 0.1Z_3Z_4 + N(0, 1),$$
$$\log C = 0.5 - 0.8Z_1 + 0.4Z_2 + 0.4Z_3 + 0.5_4 - 0.1Z_1Z_4 - 0.6Z_2Z_4 + 0.3Z_3Z_4 + N(0, 1).$$

The censoring ratio was around 30% in both settings. We experimented two sample sizes, 100 or 200, and two numbers of noise variables, 10 or 30.

The simulation results are summarized in Table 3. The same 1-NN or 3-NN method was applied to predict event times using the fitted scores derived from SVHM or Cox model. We can see that 1-NN performs slightly better than 3-NN in terms of a higher correlation and lower RMSE for both methods. In addition, when the event times were simulated from the Cox model, SVHM with 1-NN or 3-NN performs similarly to Cox model-based analysis. This is expected since proportional hazards assumption was satisfied for the Cox model based method. We also compared using 1-NN and 3-NN for prediction with using median survival times under a Cox model. We see 1-NN with SVHM or 1-NN with Cox model leads to superior performance than using median survival time. When the true model for the event times was accelerated failure time model (AFT), SVHM outperforms Cox model based analysis in terms of a higher correlation and lower RMSE. In the AFT model case, using the median survival time from the Cox model for prediction tends to be less accurate since the model assumption does not hold. Lastly, when the number of noise variables was 95, Cox model analysis did not converge in most simulations and thus the results were not included. In summary, results in Table 3 show that nearest neighbor based prediction rule performs better than using median survival time, and SVHM performs better than Cox model based methods when the model assumption does not hold.

4.2 PREDICT-HD Study

In the first real data analysis, we apply various methods to a study on Huntington's disease (HD). HD is a severe dominant genetic disorder for which at risk subjects can be identified

			Cox Model			SVHM	
Model	n	Index	1-NN	3-NN[a]	Median[b]	1-NN	3-NN
Cox1[c]	100	CORR	0.871	0.859	0.866	0.863	0.851
		RMSE	6.068	6.485	6.487	6.099	6.503
	200	CORR	0.896	0.890	0.871	0.885	0.879
		RMSE	5.755	6.168	6.226	5.781	6.186
Cox2[d]	100	CORR	0.841	0.831	0.839	0.854	0.844
		RMSE	6.146	6.548	6.546	6.139	6.546
	200	CORR	0.887	0.884	0.855	0.883	0.879
		RMSE	5.760	6.209	6.273	5.761	6.214
AFT1[e]	100	CORR	0.210	0.211	0.192	0.224	0.224
		RMSE	0.766	0.756	0.950	0.739	0.731
	200	CORR	0.275	0.275	0.262	0.277	0.277
		RMSE	0.720	0.717	0.879	0.709	0.706
AFT2[f]	100	CORR	0.129	0.129	0.110	0.174	0.175
		RMSE	0.859	0.841	1.050	0.753	0.745
	200	CORR	0.197	0.197	0.175	0.221	0.222
		RMSE	0.778	0.774	0.999	0.732	0.729

[a] Using mean of 3 nearest neighbors as predicted event time
[b] Using median survival time fitted from a Cox model as predicted event time.
[c] T and C simulated from Cox model with 10 noise variables.
[d] T and C simulated from Cox model with 30 noise variables.
[e] T and C simulated from AFT model with 10 noise variables.
[f] T and C simulated from AFT model with 10 noise variables.

Table 3: Comparison of SVHM with Cox model based methods

through a genetic testing of C-A-G expansion status at the IT15 gene (MacDonald et al., 1993). The availability of genetic testing and virtually complete penetrance of gene provides opportunity for early intervention. Currently a major research interest in HD is to combine salient clinical markers and biological markers sensitive enough to detect early indicators of patient disease diagnosis before evident clinical signs of HD emerge, and thus inform early interventions long before the clinical diagnosis. The hope of such early detection and intervention is to alter the disease course before substantial damage has occurred. The most promising markers thus far are brain imaging biomarkers and some cognitive markers which correlate with future clinical diagnosis (Paulsen, 2011; Paulsen et al., 2014).

We perform analysis using data collected in the PREDICT-HD study (Paulsen et al. 2008b; data available through dbGap: `http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000222.v3.p2`). PREDICT-HD is by far the largest and most comprehensive study of prodromal HD subjects that collects clinical, cognitive and structural MRI imaging biomarkers predictive of HD onset. Pre-manifest HD subjects in the absence of experimental treatment were recruited and followed to monitor HD symptom progression and assess HD onset. In our analyses, there were 647 subjects and 118 of them developed HD during the course of study. For each subject, a wide range of measures on motor, psychiatric, cognitive signs as well as MRI imaging markers were collected at the baseline visit. The covariates cover important clinical, cognitive, functional, psychiatric and imaging domains of HD including CAP score (a combination of age and C-A-G repeats length, Zhang et al. (2011)), symbol digital modality test (SDMT), STROOP color, word and interference tests, total functional capacity scores, UHDRS total motor scores, various SCL-90 psychiatric scores, demographic variables such as gender and education in years, and imaging measures based on regional brain volumes. The structural MRI T1-weighted imaging analysis of subcortical and cortical segmentations and cortical parcellations were based on a customized Freesurfer 5.2 pipeline developed at The University of Iowa. The details of imaging preprocessing and analysis are available in the online Supplementary Material of Paulsen et al. (2014). The subcortical volumetric measures of interest include nucleus accumbens, caudate, putamen, hippocampus, and thalamus (Paulsen et al., 2014).

We study the combined prediction capability of 31 baseline markers predicting the age-at-onset of HD diagnosis during the study period, and evaluate the usefulness of the fitted prediction score on performing risk stratification. The covariates are normalized to the same scale to achieve numeric stability and allow for comparing their relative importance. The predicted values of HD onset ages are obtained via three-fold cross validation, and the cost tuning parameter is chosen from the grid $2^{-16}, 2^{-15}, \ldots, 2^{16}$. We consider both linear kernel and Gaussian kernel. For the Gaussian kernel $K(x, x') = \exp(-\gamma\|x - x'\|^2)$, the parameter γ is fixed to be 0.005. To compare the prediction capability, we compute several quantities using the predicted values of onset ages and the observed onset ages or censoring ages. Specifically, we report the concordance index defined as the percentage of correctly ordered pairs among all feasible pairs (C-index) when including imaging markers. To evaluate the ability of the fitted scores on performing risk stratification, we separated subjects into two groups (high risk versus low risk group) based on whether their predicted scores are higher or lower than given percentiles computed from all subjects' fitted scores. We then calculate the Chi-square statistics from the logrank test and the hazard ratios comparing the hazard rate of developing HD between two groups.

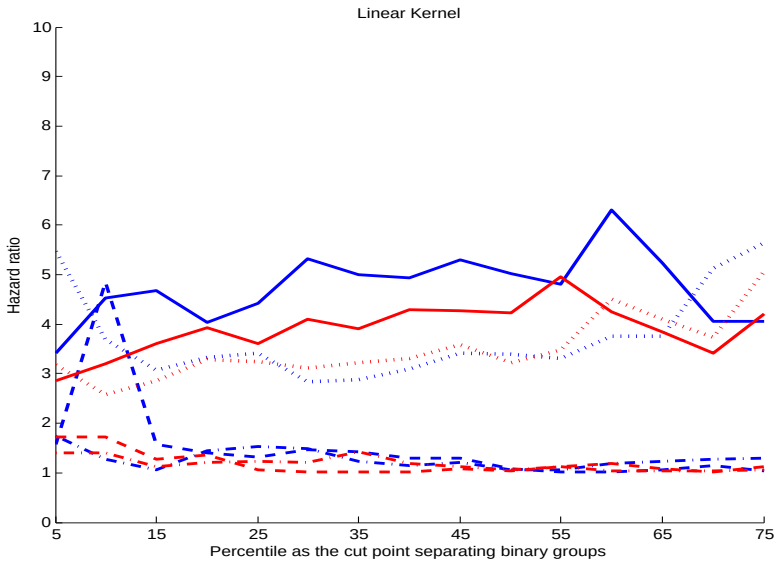

Figure 1: Hazard ratios comparing two groups separated using percentiles of predicted scores as cut points for PREDICT-HD data with linear kernel. Blue curves obtained from analyses with MRI imaging biomarkers and red curves without imaging biomarkers. Solid curve: SVHM; Dotted curve: Modified SVR; Dashed curve: IPCW-KM; Dashed-dotted curve: IPCW-Cox.

The analysis results are given in Table 4. SVHM improves over the other methods with respect to all the quantities for both linear kernel and Gaussian kernel, and the performances are similar using different kernels. For example, the logrank Chi-square statistics and hazard ratios of SVHM are much larger than the competing methods at most quantiles except at the right tail (e.g., over 65th percentile). A higher value of logrank Chi-square and a larger hazard ratio indicate greater difference between high risk and low risk subjects using a given percentile as a cut off value, and thus better discriminant ability of a risk score distinguishing high/low risk subjects. In addition, the predictions from IPCW cannot capture the trend of the original disease onset ages. Figure 1 complements the results in the table by plotting the hazard ratios comparing two groups separated using a series of percentiles of the predicted scores as cut points, and SVHM consistently has the largest hazard ratio across all percentiles among all methods. The improvement of SVHM increases at the higher percentiles, indicating that it is particularly effective in discriminating high risk subjects. This observation is consistent with our theoretical results which reveal that SVHM is optimal in separating the individual covariate-specific hazard function, $h(t, \mathbf{x})$ given $\mathbf{x}$, from the population average hazard function, $\bar{h}(t)$.

Additionally, we show the fitted coefficients from SVHM and other competing methods in Table 5 and compare with coefficients obtained from a Cox proportional hazards model.

Imaging[a]	Method	C-index	25th percentile		50th percentile		75th percentile	
			Logrank χ^2 [b]	HR[c]	Logrank χ^2	HR	Logrank χ^2	HR
No	Modified SVR	0.70	43.55	3.24	25.85	3.23	15.53	5.07
	IPCW-KM	0.47	0.04	1.05	0.04	1.04	0.31	1.12
	IPCW-Cox	0.48	1.05	1.23	0.05	1.05	0.08	1.06
	SVHM (Linear)[d]	0.73	53.41	3.61	32.72	4.22	9.02	4.21
Yes	Modified SVR	0.70	47.85	3.41	38.36	3.40	27.44	5.65
	IPCW-KM	0.47	0.76	1.31	0.14	1.08	0.04	1.04
	IPCW-Cox	0.47	1.47	1.57	0.07	1.05	0.90	1.49
	SVHM (Linear)	0.74	71.26	4.42	46.73	5.02	14.75	4.06
	SVHM (Gaussian)[e]	0.79	105.99	5.86	67.66	7.44	25.31	7.18

[a] Whether structural MRI imaging biomarkers were included in the analysis.
[b] Logrank χ^2: Chi-square statistics from Logrank tests for two groups separated using the 25th percentile, 50th percentile, and 75th percentile of predicted values.
[c] HR: Hazard Ratios comparing two groups separated using the 25th percentile, 50th percentile, and 75th percentile of predicted values.
[d] SVHM with linear kernel.
[e] SVHM with Gaussian kernel.

Table 4: Comparison of prediction capability for different methods using PREDICT-HD data with and without structural MRI imaging measures ($n = 647$)

Modified SVR yields coefficients in the same direction as SVHM, while two IPW methods give several coefficients in the opposite direction of other methods. SVHM suggests the top ranking markers with largest standardized effects to be the baseline total motor score and CAP score, which is consistent with the clinical literature on the importance of these markers on the diagnosis of HD (Zhang et al., 2011; Paulsen et al., 2014; Chen et al., 2014). The baseline total motor score as a measure of motor impairment appears to be more informative than CAP score in terms of predicting future HD diagnosis during the study. Several neuropsychological markers (Stroop color, Stroop word, SDMT) are predictive except for Stroop interference score. The coefficients from Cox model however, suggest that SDMT is not important, which is not consistent with the clinical literature (Paulsen, 2011; Paulsen et al., 2014). Note that SVHM gives psychiatric markers (SCL 90 depression, GSI, PST and PSDI) low weights which is consistent with clinical observations that the psychiatric markers are considered as noisy for predicting HD diagnosis due to reasons such as subjects may seek treatment for their psychiatric symptoms. In contrast, Cox model yields high weights for these markers which are deemed to be less informative.

In terms of neuroimaging markers, we see that pallidum, putamen, caudate, and thalamus show relatively strong predictive ability of HD onset, while accumbens and hippocampus show low predictive ability. Comparing SVHM and Cox model analysis, note that SVHM provides coefficients with similar magnitude for imaging measures on the life and right side of the same brain region, but Cox model sometimes produces substantially different results for left and right side. For example, left pallidum area is significant but not right pallidum area in Cox model. This observation suggests that SVHM may lead to more interpretable results especially for correlated variables. Another biomarker, cerebral spinal fluid, appears to be promising for predicting HD onset with a coefficient with moderate magnitude. To assess the added value of MRI imaging measures in terms of risk stratification, in Figure 1 we show the hazard ratio comparing high risk versus low risk group based on percentile split of the fitted scores obtained with and without imaging biomarkers. For SVHM with linear kernel, adding imaging measures leads to a larger hazard ratio and a greater difference between high and low risk groups at all percentiles, which demonstrates the ability of SVHM to extract information from imaging biomarkers and corroborates other findings suggesting their added values in predicting HD onset (Paulsen et al., 2014). When using Gaussian kernel for SVHM, we see further improvement of C-index and logrank chi-square statistics. Other methods such as modified SVR or IPCW do not show an advantage from including imaging measures, which may suggest their limitations in handling correlated biomarkers.

4.3 ARIC Study

As a second real world numeric example, we analyze data from the Atherosclerosis Risk in Communities Study, a prospective investigation of the aetiology of atherosclerosis and its clinical sequelae, as well as the variation in cardiovascular risk factors, medical care and disease by race, gender, location and date (The ARIC investigators, 1989; Lubin et al., 2016). We assess the prediction capability of some common cardiovascular risk factors for incident heart failure until 2005. Specifically, these risk factors include age, diabetes status, body mass index, systolic blood pressure, fasting glucose, serum albumin, serum creatinine, heart

Variable	Modified SVR ($\times 10^{-1}$)	IPCW-KM ($\times 10^{-2}$)	IPCW-Cox ($\times 10^{-3}$)	SVHM	Cox model[a]
CAP	0.051	0.936	0.202	0.255	0.058
TOTAL MOTOR SCORE	0.280	-1.083	0.529	0.519	0.398*
SDMT	-0.096	-0.411	0.076	-0.119	-0.190
STROOP COLOR	-0.042	0.412	-0.038	-0.153	-0.160
STROOP WORD	-0.227	0.488	0.188	-0.191	-0.217
STROOP INTERFERENCE	0.254	-0.432	-0.239	-0.000	0.328
TOTAL FUNCTIONAL CAPACITY	-0.062	0.175	0.142	-0.072	0.007
UHDRS PSYCH	0.168	0.137	-0.280	0.155	0.228
SCL90 DEPRESS	-0.285	-0.255	-0.132	-0.064	-0.618*
SCL90 GSI	0.316	-0.184	-0.182	0.007	0.618
SCL90 PST	-0.108	-0.265	-0.246	-0.057	-0.268
SCL90 PSDI	0.099	-0.379	-0.249	0.103	0.035
FRSBE TOTAL	-0.088	0.108	-0.136	0.112	0.115
Education Years	0.019	-1.057	0.349	-0.016	-0.053
Gender (Male)	0.178	-0.394	-0.202	0.376	0.344*
Right Putamen	-0.009	-0.395	-0.376	-0.134	-0.038
Left Putamen	-0.590	-0.165	-0.210	-0.116	-0.369
Right Pallidum	-0.015	-0.490	-0.151	-0.225	-0.049
Left Pallidum	-0.329	0.100	-0.189	-0.261	-0.626*
Right Caudate	-0.830	0.655	-0.160	-0.147	-0.943*
Left Caudate	0.397	0.738	-0.265	-0.079	0.306
Right Accumbens	0.282	-0.214	-0.470	0.051	0.220
Left Accumbens	-0.256	-0.568	-0.487	-0.057	-0.467*
Right Thalamus	0.099	-0.295	-0.710	0.172	0.260
Left Thalamus	0.258	-0.404	-0.636	0.219	0.138
Right Hippocampus	0.103	-1.152	-0.821	0.010	0.095
Left Hippocampus	-0.130	-1.087	-0.847	-0.082	-0.128
Third Ventricle	-0.101	1.071	0.841	-0.042	-0.046
Right Lateral Ventricle	0.140	2.794	1.409	-0.119	-0.016
Subcortical Gray Area	0.932	-0.868	-0.691	0.307	1.473*
Cerebral Spinal Fluid	-0.268	0.116	0.954	-0.113	-0.104

[a] The estimates from Cox model with significant p-values ($p < 0.05$) are marked with *.

Table 5: Normalized coefficients estimated from PREDICT-HD data (including imaging biomarkers) using Modified SVR, IPCW-KM, IPCW-Cox, SVHM with linear kernel and Cox model

rate, left ventricular hypertrophy, bundle branch block, prevalent coronary heart disease, valvular heart disease, high-density lipoprotein, pack-years of smoking, and current and former smoking status. The analysis sample consists of 624 participants who are African-American males living in Jackson, Mississippi. Incident heart failure occurred in 133 men through 2005, with a median follow-up time 16.2 years. Among those participants who did not develop heart failure, 324 were administratively censored on December 31st, 2005. The analysis follows the same procedure as in Section 4.2. The results for prediction capability of different methods are given in Table 6. SVHM provides more accurate prediction than other methods using the linear kernel or Gaussian kernel. It also has higher logrank test statistic and hazard ratio comparing high risk versus low risk group using various percentiles of the predictive scores as cut off points (Figure 2).

In Table 7, we can see that all the risk factors have positive effects on the incident heart failure except high-density lipoprotein, serum albumin and former smoking status. Risk factors for incident heart failure with largest standardized effects include HDL, age, prevalent CHD, and serum albumin level. We also present estimated coefficients from a Cox proportional hazards model as comparison in Table 7. Most coefficients are comparable in terms of size. However, note that higher fasting glucose level appears to be protective of heart rate failure using Cox model, which is the opposite of the expected direction.

5. Concluding Remarks

In this paper, we propose a new statistical framework to learn risk scores for event times using right-censored data by support vector hazards machine. We propose to view the prediction of time-to-event outcomes from a counting process point of view to avoid complications from specifying a censoring distribution. Asymptotically, we justify the associated universal consistency and learning rate through the structural risk minimization and show a natural link between the fitted decision function and the true hazard function: the fitted decision rule asymptotically minimizes the integrated difference between the covariate-specific hazard function and population average hazard function. Our theory shows that SVHM essentially compares events and non-events among the subjects at risk at each follow-up time; therefore, SVHM is sensitive to temporal difference between events and non-events which may not be reflected in either SVR or inverse weighted approaches. We also reveal a theoretical connection between SVHM and Cox partial likelihood function; the proposed method uses a hinge loss which should be robust to extreme observations in contrast to the exponential loss used in Cox partial likelihood. The simulation studies and real data applications demonstrate satisfactory results in finite samples with improved overall risk prediction accuracy in the presence of noise variables compared to other methods, especially when the censoring rate is high and the distribution of censoring times is unknown. The improved performance of our method is due to introducing counting processes to represent the time-to-event data, which leads to an intuitive connection of the method with both support vector machines in standard supervised learning and hazard regression models in standard survival analysis.

Since SVHM essentially learns hazard functions across subjects conditional on each risk set, the intercept term, $\alpha(t)$, is a non-informative nuisance parameter and allowed to be discontinuous over time. This feature is analogous to the estimation in Cox regression

Kernel	Method	C-index	25th percentile		50th percentile		75th percentile	
			Logrank χ^2 [a]	HR [b]	Logrank χ^2	HR	Logrank χ^2	HR
Linear	Modified SVR	0.74	90.52	4.63	59.11	4.16	31.85	5.01
	IPCW-KM	0.69	54.90	3.48	29.53	2.64	22.92	3.45
	IPCW-Cox	0.71	48.34	3.24	39.70	3.12	27.63	4.32
	SVHM	0.76	95.09	4.78	67.06	4.63	34.93	5.36
Gaussian	Modified SVR	0.76	105.10	5.12	70.41	4.87	37.66	6.39
	IPCW-KM	0.70	58.15	3.61	33.49	2.81	19.61	3.00
	IPCW-Cox	0.72	52.77	3.39	47.10	3.50	27.99	4.37
	SVHM	0.77	111.10	5.31	64.79	4.53	35.60	5.76

[a] Logrank χ^2, Chi-square statistics from Logrank tests for two groups separated using the 25th percentile, 50th percentile, and 75th percentile of predicted values.

[b] HR, Hazard Ratios comparing two groups separated using the 25th percentile, 50th percentile, and 75th percentile of predicted values.

Table 6: Comparison of prediction capability for different methods using Atherosclerosis Risk in Communities data

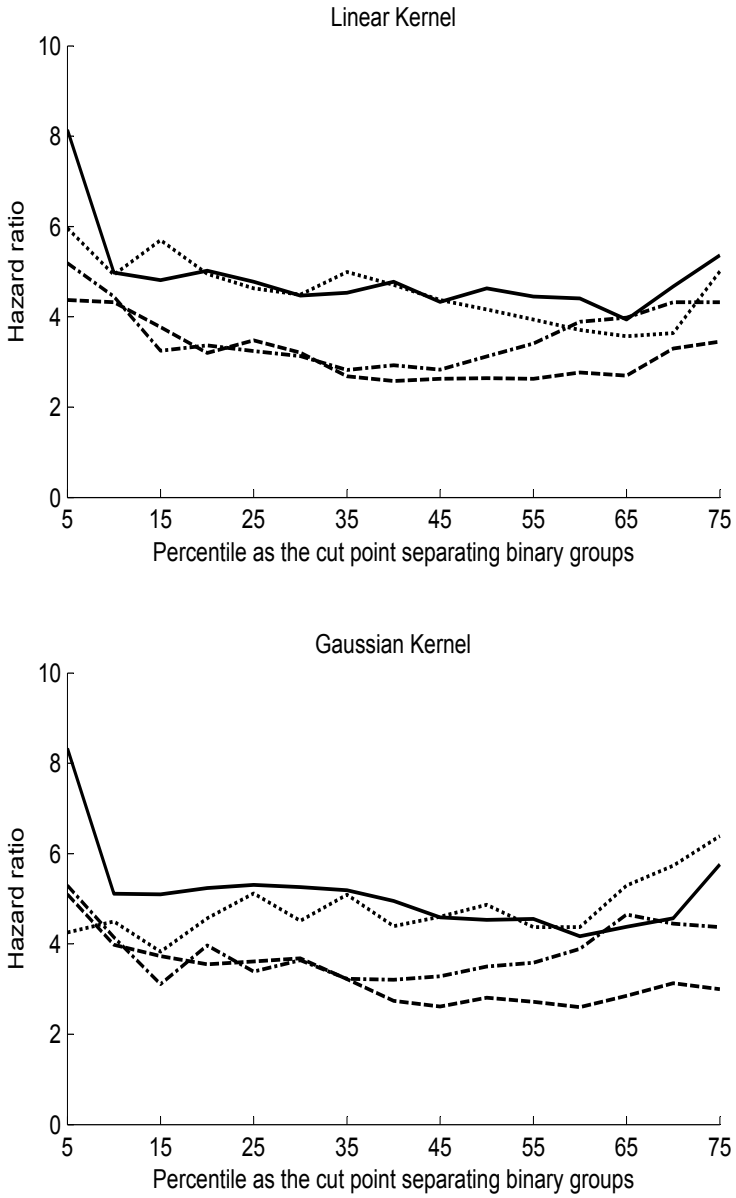

Figure 2: Hazard Ratios comparing two groups separated using percentiles of predicted values as cut points for Atherosclerosis Risk in Communities data. Solid curve: SVHM; Dotted curve: Modified SVR; Dashed curve: IPCW-KM; Dashed-dotted curve: IPCW-Cox.

Covariate	Normalized β	Cox model [a]
Age (in years)	0.363	0.328 *
Diabetes	0.288	0.221 *
BMI (kg/m^2)	0.150	0.136
SBP (mm of Hg)	0.172	0.178
Fasting glucose (mg/dL)	0.173	-0.093
Serum albumin (g/dL)	-0.363	-0.273 *
Serum creatinine (mg/dl)	0.007	0.029
Heart rate (beats/minute)	0.124	0.125
Left ventricular hypertrophy	0.250	0.158 *
Bundle branch block	0.341	0.242 *
Prevalent CHD	0.330	0.216 *
Valvular heart disease	0.200	0.169 *
HDL (mg/dl)	-0.287	-0.436 *
LDL (mg/dl)	0.016	0.051
Pack years of smoking	0.289	0.230 *
Current smoking status	0.210	0.022
Former smoking status	-0.133	-0.232 *

[a] The estimates from Cox model with significant p-value (p-value < 0.05) are marked with *.

Table 7: Normalized coefficient estimates using linear kernel for Atherosclerosis Risk in Communities data

through maximizing the Cox partial likelihood function, where the baseline hazard function is estimated to be non-continuous. Furthermore, due to the martingale property of the counting process, data from each time point can be viewed as independent in the learning method, despite that they may be from the same individual. Thus, we expect little efficiency loss even though some weighing scheme can be adopted to weight distinct risk sets differently over time.

In the current framework, the time-specific risk score $f(t, \mathbf{X})$ being considered includes a class of additive rules. They can be generalized to be fully nonparametric to learn dynamic risk profiles using a subject's time-varying covariates under the current set up. However, this generalization may lose the similarity of formulation to the standard support vector machines and cause numerical instability in the optimization algorithm. These challenging issues will be further investigated in future work. One limitation of the current nonparametric framework not specifying the event distribution is that no straightforward prediction formulae using distribution exist. We used nearest neighbors to perform prediction and simulation studies show that using less closer neighbors (3-NN instead of 1-NN) has little influence on the results. In our simulation studies, we found that a training sample size of $n = 100$ or $n = 200$ both yield stable estimation of correlation and RMSE (not sensitive to the choice of 1-NN or 3-NN). However, further work is needed to examine alternative prediction methods. Lastly, this work opens possibilities to use other powerful learning algorithms for binary and continuous outcomes to handle censored outcomes. For example, instead of using series of SVM to predict counting process as demonstrated here, other effective tools such as AdaBoosting and random forest can also be used. Gaussian process approaches (Barrett and Coolen, 2013) have been recently applied for survival data with competing risks so it will be interesting to compare SVHM with their approaches in terms of prediction performance and robustness.

Acknowledgments

We wish to acknowledge Dr. XiaoXi Liu's contribution to this work. We would like to acknowledge the National Institute of Health Grants NS073671 and NS082062, the NIH dbGap data repository (phs000222.v3.p2) and the PREDICT-HD investigators.

Appendix A. SVHM Algorithm

In this section, we provide a detailed description of the SVHM algorithm:

Algorithm: SVHM for Censored Outcomes

Input: Training data $(\mathbf{X}_i, T_i \wedge C_i, Y_i(t_j), \delta N_i(t_j))$ for $i = 1, \cdots, n, j = 1, \cdots, m$.

Step 1. Solve the quadratic programming problem:

$$\max_{\gamma_{ij}} \sum_{i=1}^{n}\sum_{j=1}^{m} \gamma_{ij} Y_i(t_j) - \frac{1}{2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{j=1}^{m}\sum_{j'=1}^{m} \gamma_{ij}\gamma_{i'j'} Y_i(t_j) Y_{i'}(t_{j'}) \delta N_i(t_j) \delta N_{i'}(t_{j'}) K(\mathbf{X}_i, \mathbf{X}_{i'})$$

$$\text{subject to: } 0 \le \gamma_{ij} \le w_i(t_j) C_n, \sum_{i=1}^{n} \gamma_{ij} Y_i(t_j) \delta N_i(t_j) = 0, i = 1, \ldots, n, j = 1, \ldots, m.$$

Denote the solutions as $\widehat{\gamma}_{ij}$.

Step 2. Compute the risk scores for non-censored subjects in the training data as

$$\widehat{g}(\mathbf{X}_s) = \sum_{i=1}^{n}\sum_{j=1}^{m} \widehat{\gamma}_{ij} \delta N_i(t_j) K(\mathbf{X}_s, \mathbf{X}_i).$$

Step 3. Predicting event time of a future subject with covariates $\mathbf{x}$ by k -nearest-neighbor:

(a) Compute the risk score for this subject as $\widehat{g}(\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{m} \widehat{\gamma}_{ij} \delta N_i(t_j) K(\mathbf{x}, \mathbf{X}_i)$.

(b) Find k non-censored subjects in the training data whose risk scores are closest to $\widehat{g}(\mathbf{x})$ and denote them as $\widehat{g}(\mathbf{X}_l)$ for $l = 1, \cdots, k$.

(c) Sort all $\widehat{g}(\mathbf{X}_s)$ in descending order and denote the rank of $\widehat{g}(\mathbf{X}_l)$ as r_l

(d) Sort event times T_s of all non-censored subjects in ascending order. Find the r_l-th event time and denote as T_l for $l = 1, \cdots, k$.

(e) The event time for this subject is predicted as $\widehat{T} = \frac{1}{k}\sum_l T_l$.

Output: For a subject with covariates $\mathbf{x}$, predict risk score as $\widehat{g}(\mathbf{x})$, and predict event time as $\widehat{T}$.

Appendix B. Proof of Theorems

In this section, we prove Theorem 3.1 and Theorem 3.2.

Proof (Theorem 3.1)

Since $f^*(t, \mathbf{x})$ minimizes $\mathcal{R}(f)$, conditional $\mathbf{X} = \mathbf{x}$, $f^*(t, \mathbf{x})$ also minimizes

$$E\left(\int Y(t)[1 - f(t, \mathbf{X})]_+ dN(t) | \mathbf{X} = \mathbf{x}\right) + \int \frac{E\left(Y(t)[1 + f(t, \mathbf{X})]_+ | \mathbf{X} = \mathbf{x}\right)}{E(Y(t))} E(dN(t)). \tag{A.1}$$

Clearly, the value $f^*(t, \mathbf{x})$ should belong to the interval $[-1, 1]$, because otherwise truncation of f at -1 or 1 gives a lower value. Assuming $-1 \le f(t, \mathbf{x}) \le 1$, (A.1) becomes

$$\int E(Y(t)|\mathbf{X} = \mathbf{x})\{h(t, \mathbf{x}) + \bar{h}(t)\}dt - \int f(t, \mathbf{x}) E(Y(t)|\mathbf{X} = \mathbf{x})\{h(t, \mathbf{x}) - \bar{h}(t)\}dt,$$

where recall that $h(t, \mathbf{x})$ is the conditional hazard rate of $T = t$ given $\mathbf{X} = \mathbf{x}$ and $\bar{h}(t)$ is the population average hazard at time t,

$$\bar{h}(t) = \frac{E[dN(t)]/dt}{E[Y(t)]} = E[h(t, \mathbf{X})|Y(t) = 1].$$

Therefore, one optimal decision function minimizing $\mathcal{R}_L(f)$ is

$$f^*(t, \mathbf{x}) = \text{sign}\{h(t, \mathbf{x}) - \bar{h}(t)\}.$$

On other hand, we note

$$\mathcal{R}_0(f) = \int I[f(t, \mathbf{x}) \le 0] E\left(Y(t)|\mathbf{X} = \mathbf{x}\right) h(t, \mathbf{x}) dt + \int I[f(t, \mathbf{x}) \ge 0] E\left(Y(t)|\mathbf{X} = \mathbf{x}\right) \bar{h}(t) dt.$$

Thus, any decision function has the same sign as $(h(t, \mathbf{x}) - \bar{h}(t))$ minimizes $\mathcal{R}_0(f)$ so $f^*(t, \mathbf{x})$ minimizes $\mathcal{R}_0(f)$. Finally, under the optimal rule $f^*(t, \mathbf{x})$, the minimal value of the weighted 0-1 risk is given as

$$\begin{aligned}
\mathcal{R}_0(f^*) &= E\left[\int E(Y(t)|\mathbf{X} = \mathbf{x}) \min\{h(t, \mathbf{x}), \bar{h}(t)\} dt\right] \\
&= \frac{1}{2} E\left[\int E(Y(t)|\mathbf{X} = \mathbf{x}) \{h(t, \mathbf{x}) + \bar{h}(t) - |h(t, \mathbf{x}) - \bar{h}(t)|\} dt\right] \\
&= P(T \le C) - \frac{1}{2} E\left[\int E(Y(t)|\mathbf{X} = \mathbf{x}) |h(t, \mathbf{x}) - \bar{h}(t)| dt\right].
\end{aligned}$$

To show the last inequality in Theorem 3.1, we note hat for $-1 \le f(t, \mathbf{x}) \le 1$,

$$\begin{aligned}
\mathcal{R}(f) &= E\left[\int E(Y(t)|\mathbf{X} = \mathbf{x}) \{h(t, \mathbf{x}) + \bar{h}(t)\} dt - \int f(t, \mathbf{x}) E(Y(t)|\mathbf{X} = \mathbf{x}) \{h(t, \mathbf{x}) - \bar{h}(t)\} dt\right] \\
&= 2P(T \le C) - E\left[\int f(t, \mathbf{x}) E(Y(t)|\mathbf{X} = \mathbf{x}) \{h(t, \mathbf{x}) - \bar{h}(t)\} dt\right],
\end{aligned}$$

and

$$\mathcal{R}(f^*) = 2P(T \le C) - E\left[\int \text{sign}\{h(t, \mathbf{x}) - \bar{\lambda_n}(t)\} E(Y(t)|\mathbf{X} = \mathbf{x}) \{h(t, \mathbf{x}) - \bar{h}(t)\} dt\right].$$

Thus,

$$\begin{aligned}
\mathcal{R}(f) - \mathcal{R}(f^*) &= E\left[\int E(Y(t)|\mathbf{X} = \mathbf{x}) \left\{\text{sign}\{h(t, \mathbf{x}) - \bar{h}(t)\} - f(t, \mathbf{x})\right\} \times \{h(t, \mathbf{x}) - \bar{h}(t)\} dt\right] \\
&= E\left[\int E(Y(t)|\mathbf{X} = \mathbf{x}) \left|f(t, \mathbf{x}) - \text{sign}\{h(t, \mathbf{x}) - \bar{h}(t)\}\right| \times |h(t, \mathbf{x}) - \bar{h}(t)| dt\right]
\end{aligned}$$

On the other hand, for the risk function based on the 0-1 loss, we have

$$\begin{aligned}
&\mathcal{R}_0(f) - \mathcal{R}_0(f^*) \\
= \; & E\left[\int E(Y(t)|\mathbf{X} = \mathbf{x}) \left(I[f(t, \mathbf{x}) \le 0] h(t, \mathbf{x}) + I[f(t, \mathbf{x}) \ge 0] \bar{h}(t) - \min\{h(t, \mathbf{x}), \bar{h}(t)\}\right) dt\right] \\
= \; & E\left[\int E(Y(t)|\mathbf{X} = \mathbf{x}) \left|h(t, \mathbf{x}) - \bar{h}(t)\right| \times I\left(\{h(t, \mathbf{x}) - \bar{h}(t)\} \text{sign}\{f(t, \mathbf{x})\} < 0\right) dt\right].
\end{aligned}$$

Note that

$$I\left(\{h(t,\mathbf{x})-\bar{h}(t)\}\text{sign}\{f(t,\mathbf{x})\}<0\right)\leq\left|f(t,\mathbf{x})-\text{sign}\{h(t,\mathbf{x})-\bar{h}(t)\}\right|.$$

We then obtain $\mathcal{R}_0(f)-\mathcal{R}_0(f^*)\leq\mathcal{R}(f)-\mathcal{R}(f^*)$. ■

Proof (Theorem 3.2)

The proof of Theorem 3.2 follows a similar procedure to the standard support vector machine theory. However, the main difference is that the proof handles $\mathcal{PR}_n(f)$ instead of the simple empirical mean of the hinge-loss in the standard theory. Let g_{λ_n} be the function in $\mathcal{H}_n$ which minimizes $\lambda_n\|g\|^2_{\mathcal{H}_n}+\mathcal{PR}(g)$. The proof consists of the following steps.

First, we derive a preliminary bound for some norms of $\widehat{g}$. Clearly,

$$\lambda_n\|g_{\lambda_n}\|^2_{\mathcal{H}_n}+\mathcal{PR}(g_{\lambda_n})\leq\mathcal{PR}(0).$$

This gives $\|g_{\lambda_n}\|_{\mathcal{H}_n}\leq\sqrt{c/\lambda}$ for some constant λ_n so by Lemma 4.23 (Steinwart and Christmann, 2008, p124), we obtain $\|g_{\lambda_n}\|_\infty\leq\sqrt{c/\lambda_n}$. Furthermore, using the fact

$$\lambda_n\|\widehat{g}\|^2_{\mathcal{H}_n}+\mathcal{PR}_n(\widehat{g})\leq\lambda_n\|g_{\lambda_n}\|^2_{\mathcal{H}_n}+\mathcal{PR}_n(g_{\lambda_n}),$$

we conclude $\|\widehat{g}\|_{\mathcal{H}_n}\leq\sqrt{c/\lambda_n}$ so $\|\widehat{g}\|_\infty\leq\sqrt{c/\lambda_n}$, where c may be another different constant (without confusion, we always use c to denote some constant). Therefore, we can restrict g in the minimization of (2) to be in $\sqrt{c/\lambda_n}\mathcal{B}_{\mathcal{H}_n}$, where $\mathcal{B}_{\mathcal{H}_n}$ be the unit ball in $\mathcal{H}_n$.

Second, we obtain a key inequality for comparing the risks of $\widehat{g}$ and g_{λ_n}. By the definition of $\widehat{g}$, the following fact holds:

$$\begin{aligned}
& \lambda_n\|\widehat{g}\|^2_H+\mathcal{PR}(\widehat{g})-(\lambda_n\|g_{\lambda_n}\|+\mathcal{PR}(g_{\lambda_n})) \\
\leq\ & \lambda_n\|\widehat{g}\|^2_H+\mathcal{PR}(\widehat{g})-(\lambda_n\|g_{\lambda_n}\|+\mathcal{PR}(g_{\lambda_n})) \\
& -\left[\lambda_n\|\widehat{g}\|^2_H+\mathcal{PR}_n(\widehat{g})-(\lambda_n\|g_{\lambda_n}\|+\mathcal{PR}_n(g_{\lambda_n}))\right] \\
=\ & \mathcal{PR}(\widehat{g})-\mathcal{PR}_n(\widehat{g})-\left\{\mathcal{PR}(g_{\lambda_n})-\mathcal{PR}_n(g_{\lambda_n})\right\}.
\end{aligned}$$

From Step 1, we conclude

$$\lambda_n\|\widehat{g}\|^2_H+\mathcal{PR}(\widehat{g})-(\lambda_n\|g_{\lambda_n}\|+\mathcal{PR}(g_{\lambda_n}))\leq 2\sup_{\|g\|_{\mathcal{H}_n}\leq\sqrt{c/\lambda_n}}|\mathcal{PR}_n(g)-\mathcal{PR}(g)|. \qquad (A.2)$$

We derive a bound for the right-hand side of (A.2). First,

$$\mathcal{PR}_n(g)-\mathcal{PR}(g)=(\mathbf{P}_n-\mathbf{P})f_g(Y,\mathbf{X},\Delta)-\frac{2}{n}\mathbf{P}_n\left\{\frac{\Delta}{\widetilde{\mathbf{P}}_n[I(\widetilde{Y}\geq Y)]}\right\},$$

where

$$f_g(Y,\mathbf{X},\Delta)=\Delta\frac{\widetilde{\mathbf{P}}_n\{I(\widetilde{Y}\geq Y)[2+g(\widetilde{\mathbf{X}})-g(\mathbf{X})]_+\}}{\widetilde{\mathbf{P}}_n[I(\widetilde{Y}\geq Y)]}+\widetilde{\mathbf{P}}\left(\widetilde{\Delta}\frac{I(Y\geq\widetilde{Y})[2+g(\mathbf{X})-g(\widetilde{\mathbf{X}})]_+}{\widetilde{\mathbf{P}}_n[I(\widetilde{Y}\geq Y)]}\right)$$

$$-\widetilde{\mathbf{P}}\left(\widetilde{\Delta}\frac{I(Y\geq\widetilde{Y})\mathbf{P}^*\{I(Y^*\geq\widetilde{Y})[2+g(\mathbf{X}^*)-g(\widetilde{\mathbf{X}})]_+\}}{\mathbf{P}^*_n[I(Y^*\geq\widetilde{Y})]\mathbf{P}^*[I(Y^*\geq\widetilde{Y})]}\right).$$

Therefore,

$$\sup_{\|g\|_{\mathcal{H}_n}\le\sqrt{c/\lambda_n}} |\mathcal{PR}_n(g) - \mathcal{PR}(g)| \le \sup_{\|g\|_{\mathcal{H}_n}\le\sqrt{c/\lambda_n}} |(\mathbf{P}_n - \mathbf{P})f_g| + c/n.$$

On the other hand, from Theorem 3.1 in Steinwart and Scovel (2007), we have

$$\log N(\epsilon, \sqrt{c/\lambda_n}\mathcal{B}_{\mathcal{H}_n}, l_\infty) \le c_{p,d}\sigma_n^{(p/4-1)d}\left(\frac{\epsilon}{\sqrt{c/\lambda_n}}\right)^{-p} \le c_{p,d}\sigma_n^{(p/4-1)d}\lambda_n^{-p/2}\epsilon^{-p},$$

where $N(\epsilon, \mathcal{F}, l_\infty)$ is the ϵ-covering number of $\mathcal{F}$ under l_∞-norm, d is the dimension of $\mathbf{X}$, p is any number in $(0, 2)$ and $c_{p,d}$ is a constant only depending on (p, d). Moreover, we note that by the property of the hinge-loss, f_g is the Lipschitz continuous in g and satisfies

$$|f_{g_1} - f_{g_2}| \le c|g_1 - g_2|.$$

This implies

$$\log N(\epsilon, \{f_g/a_n : g \in \sqrt{c/\lambda_n}\mathcal{B}_{\mathcal{H}_n}\}, l_\infty) \le c_{p,d}\sigma_n^{(p/4-1)d}\epsilon^{-p},$$

where $a_n = \sqrt{c/\lambda_n}\sigma_n^{-(1-p/4)d/p}$. Therefore, according to Theorem 2.14.10 in van der Vaart and Wellner (1996), we obtain

$$P\left(\sqrt{n}\sup_{\|g\|_{\mathcal{H}_n}\le\sqrt{c/\lambda_n}} |(\mathbf{P}_n - \mathbf{P})(f_g/a_n)| > x\right) \le e^{-cx^2}$$

for some constant c only depending on (p, d). Consequently, (A.2) gives

$$P(\lambda_n\|\widehat{g}\|_H^2 + \mathcal{PR}(\widehat{g}) - (\lambda_n\|g_{\lambda_n}\| + \mathcal{PR}(g_{\lambda_n})) > cn^{-1} + a_n n^{-1/2}x) \le e^{-cx^2}. \qquad (A.3)$$

Hence, we have proved

$$\lambda_n\|\widehat{g}\|_{\mathcal{H}_n}^2 + \mathcal{PR}(\widehat{g}) \le \inf_{g\in\mathcal{H}_n}\{\lambda_n\|g\|_{\mathcal{H}_n} + \mathcal{PR}(g)\} + O_p\left(\frac{\lambda_n^{-1/2}\sigma_n^{-(1/p-1/4)d}}{\sqrt{n}}\right). \qquad (A.3)$$

Let $g^* = \text{argmin}\mathcal{PR}(g)$. From the expression of $\mathcal{PR}(g)$, we note

$$|\mathcal{PR}(g) - \mathcal{PR}(g^*)| \le c\|g - g^*\|_{L_1(P)}.$$

Thus, if we define

$$\widetilde{g}(\mathbf{x}) = \frac{2\sigma_n^{-d/2}}{\pi^{d/4}}\int e^{-\|x-y\|^2/(2\sigma_n^2)}g^*(y)dy,$$

then $\widetilde{g} \in \mathcal{H}_n$ and

$$\|g - g^*\|_{\mathcal{H}_n} \le \|g - g^*\|_{L_2(P)} \le c\sigma_n^{d/2}.$$

Therefore,

$$\inf_{g\in\mathcal{H}_n}\{\lambda_n\|g\|_{\mathcal{H}_n} + \mathcal{PR}(g)\} \le \{\lambda_n\|\widetilde{g}\|_{\mathcal{H}_n} + \mathcal{PR}(\widetilde{g})\} \le \mathcal{PR}(g^*) + c\sigma_n^{d/2} + c\lambda_n,$$

and the result in Theorem 3.2 holds.

■

References

S. Abe. *Support Vector Machines for Pattern Classification, Second Edition*. Springer, London, 2010.

J. E. Barrett and A. C. C. Coolen. Gaussian process regression for survival data with competing risks. *arXiv preprint*, 1312.1591, 2013.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification and risk bounds. *The Journal of American Statistical Associations*, 101(473):138–156, 2006.

S. Bennett. Analysis of survival data by the proportional odds model. *Statistics in Medicine*, 2:273–277, 1983.

P. J. Bickel, C. A. J. Klaassen, Y. Ritov, J. A. Wellner, et al. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, 1998.

I. Bou-Hamad, D. Larocque, and H. Ben-Ameurm. A review of survival trees. *Statistics Surveys*, 5:44–71, 2011.

J. Buckley and I. James. Linear regression with censored data. *Biometrika*, 66:429–436, 1979.

K. Chen, Z. Jin, and Z. Ying. Semiparametric analysis of transformation models with censored data. *Biometrika*, 89:659–668, 2002.

T. Chen, Y. Wang, H. Chen, K. Marder, and D. Zeng. Targeted local support vector machine for age-dependent classification. *Journal of the American Statistical Association*, 109(507):1174–1187, 2014.

S. C. Cheng, L. J. Wei, and Z. Ying. Analysis of transformation models with censored data. *Biometrika*, 82:835–845, 1995.

D. R. Cox. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society: Series B*, 34:187–220, 1972.

D. M. Dabrowska and K. A. Doksum. Partial likelihood in transformation models with censored data. *Scandinavian Journal of Statistics*, 15:1–23, 1988.

Y. Goldberg and M. R. Kosorok. Support vector regression for right censored data. *Unpublished manuscript*, 2013.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer, New York, 2009.

F. M. Khan and V. B. Zubek. Support vector regression for censored data (SVRc): a novel tool for survival analysis. *In Eighth IEEE International Conference on Data Mining*, pages 863–868, 2008.

Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99:67–81, 2004.

J. H. Lubin, D Couper, P. L. Lutsey, M. Woodward, H. Yatsuya, and R. R. Huxley. Risk of cardiovascular disease from cumulative cigarette use and the impact of smoking intensity. *Epidemiology*, 27(3):395–404, 2016.

M. E. MacDonald, C. M. Ambrose, M. P. Duyao, R. H. Myers, C. Lin, L. Srinidhi, G. Barnes, S. A. Taylor, M. James, N. Groot, et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, 72(6):971–983, 1993.

K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury, et al. The parkinson progression marker initiative (PPMI). *Progress in neurobiology*, 95(4):629–635, 2011.

J. Mogueraza and A. Munoz. Support vector machines with applications. *Statistical Science*, 21(3):322–336, 2006.

S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's & Dementia*, 1(1):55–66, 2005.

G. Orru, W. Pettersson-Yeo, A.F. Marguand, and et al. Using support vector machine to identify imaging biomarkers of neurological and psychiatri disease: a critical review. *Neurosci Biobehav Rev*, 36(4):1140–1152, 2012.

J. S. Paulsen. Cognitive impairment in Huntington disease: diagnosis and treatment. *Current neurology and neuroscience reports*, 11(5):474–483, 2011.

J. S. Paulsen, D. R. Langbehn, J. C. Stout, E. Aylward, C. A. Ross, M. Nance, and et al. Detection of Huntington's disease decades before diagnosis: the Predict-HD study. *Journal of Neurology, Neurosurgery and Psychiatry*, 79:874–880, 2008a.

J. S. Paulsen, D. R. Langbehn, J. C. Stout, E. Aylward, C. A. Ross, M. Nance, M. Guttman, S. Johnson, M. MacDonald, L. J. Beglinger, K. Duff, E. Kayson, K. Biglan, I. Shoulson, D. Oakes, and M. Hayden. Detection of Huntington's disease decades before diagnosis: the Predict-HD study. *Journal of Neurology, Neurosurgery & Psychiatry*, 79(8):874–880, 2008b.

J. S. Paulsen, J. D. Long, H. J. Johnson, E. H. Aylward, C. A. Ross, J. K. Williams, M. A. Nance, C. J. Erwin, H. J. Westervelt, D. L. Harrington, et al. Clinical and biomarker changes in premanifest Huntington disease show trial feasibility: a decade of the PREDICT-HD study. *Frontiers in aging neuroscience*, 6:78:1–11, 2014.

B. D. Ripley and R. M. Ripley. Neural networks as statistical methods in survival analysis. *Clinical Application of Artificial Neural Network*, pages 237–255, 2001.

R. M. Ripley, A. L. Harris, and L. Tarassenko. Non-linear survival analysis using neural networks. *Statistics in Medicine*, 23(5):825–842, 2004.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

P. K. Shivaswamy, W. Chu, and M. Jansche. A support vector approach to censored targets. *In Seventh IEEE International Conference on Data Mining*, pages 655–660, 2007.

A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Journal of Statistics and Computing*, 14:199–222, 2004.

I. Steinwart. Support vector machines are universally consistent. *Journal of Complexity*, 18:768–791, 2002.

I. Steinwart and A. Christmann. *Support Vector Machines, First Edition*. Springer, New York, 2008.

I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *The Annals of Statistics*, 35(2):575–607, 2007.

The ARIC investigators. The Atherosclerosis Risk in Communities (ARIC) study: design and objectives. *American Journal of Epidemiology*, 129(4):687–702, 1989.

V. Van Belle, K. Pelckmans, J. A. K. Suykens, and S. Van Huffel. Additive survival least-squares support vector machines. *Statistics in Medicine*, 29(2):296–308, 2010.

V. Van Belle, K. Pelckmans, S. Van Huffel, and J. A. K. Suykens. Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artificial Intelligence in Medicine*, 53(2):107–118, 2011.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

H. Wang, X. Shen, and W. Pan. Large margin hierarchical classification with mutually exclusive class membership. *The Journal of Machine Learning Research*, 12:2721–2748, 2011.

D. Zeng and D. Y. Lin. Efficient estimation of semiparametric transformation models for counting processes. *Biometrika*, 93(3):627–640, 2006.

D. Zeng and D. Y. Lin. Maximum likelihood estimation in semiparametric models with censored data (with discussion). *Journal of the Royal Statistical Society, B*, 69(4):507–564, 2007.

Y. Zhang, J. D. Long, J. A. Mills, J. H. Warner, W. Lu, J. S. Paulsen, and et al. Indexing disease progression at study entry with individuals at-risk for Huntington disease. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 156B(7):751–763, 2011.