VARIABLE SELECTION AND STATISTICAL LEARNING FOR CENSORED DATA

Xiaoxi Liu

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2014

Approved by:

Donglin Zeng

Gerardo Heiss

Danyu Lin

Yuanjia Wang

Michael Wu

ABSTRACT

XIAOXI LIU: Variable Selection and Statistical Learning for Censored Data
(Under the direction of Donglin Zeng)

This dissertation focuses on (1) developing an efficient variable selection method for a class of general transformation models; (2) developing a support vector based method for predicting failure times allowing the coarsening at random assumption for the censoring distribution; (3) developing a statistical learning method for predicting recurrent events.

In the first topic, we propose a computationally simple method for variable selection in a general class of transformation models with right-censored survival data. The proposed algorithm reduces to maximizing a weighted partial likelihood function within an adaptive lasso framework. We establish the asymptotic properties for the proposed method, including selection consistency and semiparametric efficiency of parameter estimators. We conduct simulation studies to investigate the small-sample performance. We apply the method to data sets from a primary biliary cirrhosis study and the Atherosclerosis Risk in Communities (ARIC) Study, and demonstrate its superior prediction performance as compared to existing risk scores.

In the second topic, we develop a novel support vector hazard regression approach for predicting survival outcomes. Our method adapts support vector machines to predict dichotomous outcomes of the counting processes among subjects at risk, and allows censoring times to depend on covariates without modeling the censoring distribution. The formulation can be solved conveniently using any convex quadratic programming

package. Theoretically, we show that the decision rule is equivalent to maximizing the discrimination power based on hazard functions, and establish the consistency and learning rate of the predicted risk. Numerical experiments demonstrate a superior performance of the proposed method to existing learning methods. Real data examples from a study of Huntington's disease and the ARIC Study are used to illustrate the proposed method.

In the third topic, we adapt support vector machines in the context of the counting process to handle time-varying covariates and predict recurrent events. We conduct extensive simulation studies to compare performances of the proposed method to the Andersen and Gill proportional intensity model for the prediction of multiple recurrences. The extension of theoretical properties is described. We illustrate the proposed method by analyzing the data set from a bladder cancer study.

ACKNOWLEDGMENTS

This dissertation would not have been possible without the help of so many people in so many ways. First and foremost, I would like to thank my advisor Dr. Donglin Zeng, for his generous support, guidance, and encouragement throughout my five-year graduate study. Not only did he provide me the initial research assistantship, but he also motivated my interest and passion to be involved in both methodology and collaborative research. Dr. Zeng always was available to discuss my work and I owe him a great debt of gratitude. In addition, I am appreciative of my committee members, Drs. Gerardo Heiss, Danyu Lin, Yuanjia Wang, and Michael Wu, for providing valuable and insightful comments on my dissertation.

I would also like to thank my supervisors Drs. Woody Chambless and Lisa Wruck at Collaborative Studies Coordinating Center (CSCC). I truly enjoyed working with them on the ARIC study, and I appreciate their understanding and patience beyond the mere financial support. I will also give special thanks to my manager Dr. Bob Rodriguez and my supervisors Drs. Guixian Lin and Warren Kuhfeld at SAS Institute for exposing me to interesting statistical methods and research in analytic software development.

Finally, I want to extend my gratitude to all my friends in the Department for their friendship and emotional support, and all the faculty and staff for their help in various ways during the whole process of my dissertation work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER1: INTRODUCTION

Statistical model building is a challenging task for censored data when there are a large number of concomitant covariates. Existing methods tend to make strong assumptions on the covariate effects or the censoring mechanism, making them unsuitable for the task of predicting future outcomes accurately. For example, the Cox proportional hazards model assumes that the hazard functions between two subjects are proportional over time. Although the model allows for time-varying covariates, it is apparent that the model excludes many complex covariate patterns. In this dissertation, we will develop statistical methods that are less dependent on the restrictive assumptions than the existing methods. Specifically, we generalize the efficient variable selection method to a class of transformation models; we adapt the popular support vector machines technique for statistical learning to the censored data that are represented by counting processes; also, we generalize this approach to recurrent event data.

1.1 Variable Selection in Semiparametric Transformation Models

In the first topic, we study variable selection in general transformation models for right-censored data. The models studied can incorporate external time-varying covariates, and they include the proportional hazards model and the proportional odds model as special cases. We propose an estimation method that involves minimizing a weighted negative partial loglikelihood function plus an adaptive lasso penalty, with the initial values obtained from nonparametric maximum likelihood estimation. The objective function is parametric and convex, so the minimization is easy to implement

and guaranteed to converge numerically. Under the regularity conditions in Zeng and Lin (2006), we show that our selection has oracle properties and that the estimator is semiparametrically efficient. We demonstrate the small-sample performance of the proposed method via simulations, and we use the method to analyze data from the Atherosclerosis Risk in Communities Study and Primary Biliary Cirrhosis Study.

1.2 Support Vector Hazard Regression for Predicting Survival Outcomes

In the second topic, we develop a novel support vector hazards regression (SVHR) approach to predict time-to-event outcomes using right-censored data. Our method is based on predicting the counting process via a series of support vector machines that maximally separate the event and non-event subjects among all subjects at risk. Introducing counting processes to represent the time-to-event data leads to an intuitive connection of the proposed method with both support vector machines in standard supervised learning and hazard regression models in standard survival analysis. The resulting optimization is a convex quadratic programming problem that can easily incorporate non-linearity using kernel machines. We demonstrate an interesting connection of the profiled empirical risk function with the Cox partial likelihood which sheds lights on the optimality of SVHR. We formally show that the SVHR is optimal in discriminating the covariate-specific hazard function from the population average hazard function, and establish the consistency and learning rate of the predicted risk. Simulation studies demonstrate much improved prediction accuracy of the event times using SVHR compared to existing machine learning methods. Finally, we apply our method to analyze data from the Huntington's Disease Study and the Atherosclerosis Risk in Communities Study to demonstrate superiority of SVHR in practical settings.

1.3 Support Vector Machines for Predicting Recurrent Events

In the third topic, we describe a generalization of support vector machines to predict recurrent event times. The prediction of recurrence using censored data has not been discussed in other statistical learning works, as all of them adapt the standard learning techniques to censored data based on survival times and are not able to handle multiple records for a subject. Similar to the support vector hazard regression, we integrate the support vector machines in the framework of counting process. As a result, there is a straightforward application to handle both recurrent events and time-varying covariates. The resulting formulation is a convex optimization problem that has a unique global solution. We present extensive simulation results comparing the performance of our method with the Anderson and Gill (1982) proportional intensity model under different scenarios, including adding dependence among recurrences and adding baseline noise variables. The data from a bladder cancer study is used to illustrate the proposed method.

CHAPTER2: LITERATURE REVIEW

In this chapter, we review literature on statistical methods for semiparametric survival models in Section 2.1, for traditional and penalized variable selection in Section 2.2, and for statistical supervised learning and outcome prediction in Section 2.3.

2.1 Semiparametric Models for Censored Data

In many medical trials, outcome of interest is survival time and is subject to censoring, where the exact survival time may be longer than the duration of the trial period and is therefore unknown. Typical examples include time to death from the start of a diagnosis, response time to a particular medical treatment, and time to recurrence of cancer tumor. It is often of interest to study whether certain clinical characteristics are related to occurrence of certain events and then examine the predictive values of survival in terms of these covariates. Since the distributional assumption on the survival times is not valid in many situations, semiparametric methods are widely used.

The most popular semiparametric model for data fitting is the Cox (1972) proportional hazards model. Given the vector of covariates Z, this model is specified by a hazard function

$$\lambda(t|Z) = \lambda_0(t)\exp(\beta^T Z) \tag{2.1}$$

where β is a vector of unknown regression coefficients and $\lambda_0(t)$ is an unknown baseline hazard function. The covariate effects act multiplicatively on the hazard function, and the exponential of the coefficient β gives the constant hazard rate ratio for an increase of

one unit for the covariate in question. To efficiently estimate the regression coefficients, Cox (1972, 1975) introduced the partial likelihood principle to eliminate the infinite-dimensional baseline hazard function, and the resulting estimator was a function of the survival times only through their ranks. In the discussion of Cox's paper (1972), Breslow (1972) proposed a nonparametric maximum likelihood estimator (NPMLE) for the arbitrary baseline hazard in (2.1) using the joint full likelihood and this estimator reduces to the Kaplan-Meier product limit estimator when there is no covariate effect. In a seminal paper, Andersen and Gill (1982) extended the Cox proportional hazard model to general counting processes to allow for recurrent event and established the asymptotic properties of the maximum partial likelihood estimator and the associated Breslow (1972) estimator of the cumulative baseline hazard function via the elegant counting-process martingale theory.

The commonly used graphical and numerical ways to check the proportional hazard assumption include the plot of logarithm of cumulative hazard functions (Andersen, 1982), the plot of Schoenfeld residuals (Schoenfeld, 1982), and the introduction of interaction between time and covariates (Lee and Go, 1997). When the proportional hazards assumption is violated, one remedy is to stratify the data into subgroups and apply the model for each stratum (Lee and Go, 1997). A drawback of this approach is that the effect of the stratifying variable cannot be estimated. Another strategy is to consider the time-varying covariates. In this case, the covariates in model (2.1) are indexed by time and the setup of the partial likelihood functions is still applicable. However, Fisher and Lin (1999) argued that time-varying covariates must be used with caution since they involve constructing a function of time that is usually not self-evident and may be suggested by biological hypotheses. They gave several examples to illustrate the complexity of choosing the functional form and the misleading results when the function form is misspecified.

A concise alternative to capture the non-proportionality is the proportional odds model (Bennett, 1983a; 1983b). Under this model, the hazard ratio between two sets of covariate values converges to unity rather than staying constant as time increases. The survival function S_Z, given the vector of covariates Z, is parameterized by

$$-\log\left\{\frac{S_Z(t)}{1-S_Z(t)}\right\} = G(t) + \beta^T Z \tag{2.2}$$

where G is an arbitrary baseline log-odds and β is a vector of regression coefficients. The nonparametric maximum likelihood estimation (NPMLE) for this model was proposed by Bennett (1983b). Bennett's estimator of β is the maximum profile likelihood estimator of β, with the baseline log-odds function being profiled out. Murphy et al. (1997) showed that this maximum profile likelihood estimator was consistent, asymptotically normal, and semiparametrically efficient. Further Murphy et al. (1997) demonstrated that the profile likelihood could be treated as a parametric likelihood and provided the asymptotic distribution for the profile likelihood ratio statistic. Another method to estimate β is maximizing the marginal likelihood of the ranks (Pettitt, 1983; 1984). Since this marginal likelihood cannot be calculated explicitly for all β, Pettitt (1983, 1984) used an approximation based on a Taylor expansion on the logarithm of the marginal likelihood at $\beta = 0$, however, the resulting estimator is biased and inconsistent. Lam and Leung (2001) employed the technique of importance sampling to express the marginal likelihood as an expectation with respect to some distribution. Their method is computationally intensive since the importance sampling is a Markov chain Monte Carlo (MCMC) algorithm. In addition, the theoretical properties of the maximum marginal likelihood estimator were not investigated.

Both the proportional hazards and proportional odds models belong to the class of linear transformation models, which relates an unknown transformation of the survival time linearly to covariates. Let T be the survival time and Z a corresponding vector of

covariates. This model can be written as

$$H(T) = -\beta^T Z + \epsilon \tag{2.3}$$

where H is an unknown monotone transformation function, ϵ is a random variable with a known distribution and is independent of Z, and β is a vector of unknown regression coefficients. If ϵ follows the extreme value distribution, model (2.3) becomes to the proportional hazards model, while if ϵ follows the standard logistic distribution, model (2.3) becomes to the proportional odds model. Several papers proposed general estimators for the regression coefficients. Dabrowska and Doksum (1988) provided estimators in the two sample problem based on the marginal likelihood of ranks and the MCMC method. Their estimators suffer from severe bias under heavy censoring and the bias cannot be reduced by increasing the size of Monte Carlo simulation (Lam and Leung, 1997). Cheng et al. (1995) derived inference procedures from a class of generalized estimating equations based on dichotomous variables of pairwise ranks. They adjusted censoring by the inverse weight of the Kaplan-Meier estimator for the survival function of the censoring variable, so the validity of their procedures relies on the assumption that the censoring variable is independent of covariates. Chen et al. (2002) mentioned that such an assumption was often too restrictive, even for randomized clinical trials. Chen et al. (2002) proposed an estimator using martingale-based estimating equations and the estimating equations precisely become to the Cox partial likelihood score equation for the proportional hazards model. Although all these methods established the consistency and asymptotic normality for their estimators, none of them is semiparametrically efficient. In addition, the class of linear transformation models is confined to traditional survival (i.e., single-event) data and time-invariant covariates.

To accommodate recurrent events and time-varying covariates, Zeng and Lin (2006, 2007a) proposed a class of general transformation models using the counting-process

notation. Let $N^*(t)$ be the counting process recording the number of events that have occurred by time t and $Z(t)$ be a vector of possibly time-varying covariates. The cumulative intensity function for $N^*(t)$ conditional on $Z(t)$ takes the form

$$\Lambda_Z(t) = G\left\{\int_0^t Y^*(t)e^{\beta^T Z(s)}d\Lambda(s)\right\} \tag{2.4}$$

where G is a thrice continuously differentiable and strictly increasing function with $G(0) = 0$, $G'(0) > 0$ and $G(\infty) = \infty$, $Y^*(.)$ is a predictable process and $\Lambda(.)$ is an unspecified increasing function, and β is a vector of unknown regression coefficients. For survival data, $Y^*(t) = I(T \geq t)$, where T is the survival time and $I(.)$ is the indicator function; for recurrent event, $Y^*(.) = 1$. When $N^*(.)$ has a single jump at survival time T and covariates Z is time-invariant, model (2.4) reduces to the linear transformation model (2.3). Specifying the function G while leaving the function Λ unspecified is equivalent to specifying the distribution of ϵ while leaving the function H unspecified. Zeng and Lin (2006) developed nonparametric maximum likelihood estimators for the regression coefficients and cumulative intensity functions of these models. The estimators were shown to be consistent and asymptotically normal. The limiting variances for the estimators of the regression coefficients achieved the semiparametric efficient bounds. Later Zeng and Lin (2007a) proposed a technique to implement the inference procedures by the simple and stable expectation-maximization (EM) algorithm. The trick is to use the Laplace transformation to convert the general transformation model into the proportional hazards model with a random effect, in which random effects pertain to missing data. The EM algorithm asserts the increase in the likelihood of successive iterations and the convergence can be guaranteed (Dempster et al., 1977). On convergence the Louis (1982) formula is used to calculate the observed information matrix.

Another important alternative to the Cox proportional hazard model is the accelerated failure time model. This model provides a natural formulation of the effects of covariates on potentially censored response variable. Let T be the survival time and Z a corresponding vector of covariates. The model can be written as

$$\log T = -\beta^T Z + \epsilon \tag{2.5}$$

where ϵ is a measurement error independent of Z and β is a vector of unknown regression coefficients. Note that model (2.5) does not belong to the linear transformation model, which has unknown H and known distribution of ϵ. Buckley and James (1979) proposed the least square estimator, where they used the least square normal equation and replaced a censored observation by its conditional mean based on the residuals and product limit estimator. Jin and Lin (2003) studied a broad class of rank-based monotone estimating functions, including the Gehan-type weight function and weighted log-rank estimating equation. Later Zeng and Lin (2007b) proposed an extension of model (2.5) to incorporate time-varying covariates, and the extended model no longer took the log-linear form. They used kernel smoothing to construct a smooth approximation to the profile likelihood for the regression coefficients, and established that the resulting estimators were consistent, asymptotically normal, and semiparametrically efficient. They also provided an explicit estimator for the error distribution.

Independent sample is assumed for all the models reviewed above; however, this assumption may be violated in medical research. For example, siblings or parents and offspring are likely to be correlated, and the times between the recurrent tumor occurrences thus not independent. Zeng and Lin (2007a) further extended the general transformation model (2.5) to characterize the dependence of recurrent events, multiple types of events and clusters through random effects or frailty. They also studied

joint models of repeated measures and survival time in longitudinal studies. The nonparametric maximum likelihood estimators of all the proposed models were shown to be consistent, asymptotically normal and semiparametrically efficient via the modern empirical process theory. In their paper, Zeng and Lin (2007a) emphasized the flexible modeling capability and accurate prediction of semiparametric transformation models, and suggested using them in the practice of survival analysis.

2.2 Variable Selection for Censored Data

With the improvement of modern technologies in epidemiologic and genetic studies, researchers are able to collecting many variables, such as patients' characteristics, biomarkers and genotypes, to predict clinical outcomes. When a large number of variables is included in prediction models it often causes over-fitting and results in low prediction power. On the other hand, it is commonly believed that only a few important variables exhibit strong effects. Hence, it is desirable to identify those few important variables in the model building process. Prior knowledge from the scientific literature is formally seen as the most important rationale for including or excluding variables from a statistical analysis, which is not always available for all research questions, and too often involves only the iterative imposition of exact (typically exclusion) restrictions on individual variables (Smith and Campbell, 1980; Walter and Tiemeier, 2009). Comparatively, data-driven variable selection is more flexible and as a result a commonly used method in practice.

2.2.1 Variable Selection Methods

Among the data-driven variable selection approaches, stepwise selection remains a commonly used technique in epidemiologic research (Walter and Tiemeier, 2009), which is carried out either by trying out one independent variable at a time and including it

if statistically significant, or by eliminating those that are not statistically significant, simultaneously adjusting for the other variables in the regression model. Despite the ease of implementation, disadvantages of stepwise selection are known from separate studies (Derksen and Keselman, 1992; Harrell et al., 1996; Steyerberg et al., 1999), including that arbitrary definitions of thresholds may lead to bias and overfitting; binary decisions about the inclusion of variables cause information to be lost; true predictors may be excluded in small data sets because of a lack of power; noise variables may be selected because of multiple comparisons problems; and that the solution may be only locally optimal. An alternative approach leading to the global optimal solution is the best-subset selection, which chooses a small subset of the predictor variables that yields the most accurate model when the regression is restricted to that subset. However, the best-subset selection is computationally infeasible when the number of predictors is large and extremely unstable since it is a discrete process where variables are either retained or dropped from the model (Breiman, 1996). Moreover, many variable selection approaches in use, such as stepwise selection and best-subset selection, make inferences as if a model is known to be true when it has, in fact, been selected from the same data to be used for estimation purposes. Ignoring the model uncertainty causes non-trivial biases in coefficient estimates and underestimation of the variability of estimated coefficients in the resulting model (Chatfield, 1995; Harrell et al., 1996; Steyerberg et al., 1999).

In recent years, regularized/penalized variable selection methods have been extensively studied. These methods select variables and estimate coefficients simultaneously, which enable us to construct confidence intervals for the estimated coefficients (Fan and Li, 2001). For linear regression with continuous outcome, the corresponding coefficients

$\beta = (\beta_0, \ldots, \beta_d)^T$ minimize a penalized residual sum of squares,

$$\sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{d} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{d} p(|\beta_j|), \tag{2.6}$$

where $(x_{i1}, \ldots, x_{id})$, $i = 1, \ldots, N$ are predictor variables, y_i, $i = 1, \ldots, N$ are responses, $p(|.|)$ is the penalty function, and $\lambda \geq 0$ is regularization/tuning parameters. Larger value of λ leads to greater amount of shrinkage. The intercept has been left out of the penalty functions, since penalization of the intercept would make the procedure dependent on the origin chosen for response (Hastie et al., 2009, page 63-64). Many penalty functions have been proposed. Ridge regression was introduced as a method for stabilizing regression estimates in the presence of extreme collinearity in predictors. The ridge penalty takes the form $p(|\beta_j|) = \beta_j^2$, $j = 1, \ldots, d$ and stochastically shrink the estimates towards zero. However, it does not give an easily interpretable model and is not scale invariant (Smith and Campbell, 1980; Frank and Friedman, 1980). In addition, it cannot give accurate predictions when there is a mix of large and small coefficients (Breiman, 1996). A scale invariant alternative to ridge regression is the nonnegative garrote, which has $p(|\beta_j|) = |\beta_j|/|\beta_j^0|$, $j = 1, \ldots, d$, with additional sign constraints $\beta_j\beta_j^0 \geq 0 \ \forall j$, where β_j^0 is the ordinary least square (OLS) estimate (Breiman, 1995; Zou, 2006). The garrote eliminates some variables, shrinks others, and is relatively stable. A drawback of the garrote is that its solution depends on both the sign and the magnitude of the OLS estimates (Tibshirani, 1996). To avoid the explicit use of the OLS estimates, Tibshirani (1996) proposed the popular technique of least absolute shrinkage (lasso) for simultaneous estimation and variable selection. The lasso penalty is $p(|\beta_j|) = |\beta_j|$, $j = 1, \ldots, d$. Like the ridge regression, lasso is not scale-invariant and requires initial standardization of the regressors. To solve for the lasso estimator, Tibshirani (1996) used a combined quadratic programming algorithm and Fu (1998) developed the simple shooting algorithm based on theoretical results of the

structure of the bridge estimators. In these algorithms, the tuning parameter λ needs to be searched over a grid of values using some criteria, such as cross-validation, generalized cross-validation (Craven and Wahba, 1979), Akaike information criterion (AIC) (Akaike, 1974), and Bayesian information criterion (BIC) (Schwarz, 1978). Later, Efron et al. (2004) proposed an extremely efficient algorithm Least Angle Regression (LARS) for computing the entire lasso path, which was proved to be a one-dimensional path of prediction vectors growing piecewise linearly from the origin to the full least-squares solution.

Frank and Friedman (1993), and Fu (1998) considered a more general class of regression estimators that minimized function (2.6) with bridge penalty $p(|\beta_j|) = |\beta_j|^q$ for $0 < q \leq \infty$, $j = 1, \ldots, d$. The value $q \to 0$ corresponds to the best-subset selection, as the penalty simply counts the number of nonzero coefficients and expresses no preference for particular variables; $q = 1$ corresponds to the lasso, while $q = 2$ to the ridge regression. As illustrated in Figure 2.1, the estimators for $q \leq 1$ have the potentially attractive feature of being exactly 0 thus combining coefficient estimation and model selection, while the bridge penalty for $q \geq 1$ has a convex structure that will make the computation simple (Tibshirani, 1996). When the regression matrix of predictor variables is assumed to be orthonormal, the minimization problem is equivalent to minimizing componentwise, with explicit forms of special cases given in Table 2.1. It is indicated that ridge regression gives a proportional shrinkage, and lasso translate each coefficient by a constant factor λ, truncating at zero. Figure 2.2(a)-(e) depict and compare the values of the bridge estimator with the OLS estimator, whose value is plotted on the diagonal. It is shown that the bridge regression of large value of q ($q \geq 2$) tends to retain small coefficients while the small value of q ($q < 2$) tends to shrink small coefficients to 0. Therefore, it can be implied that if the true model includes many small but nonzero regression coefficients, the lasso will perform poorly but the bridge of large q value will

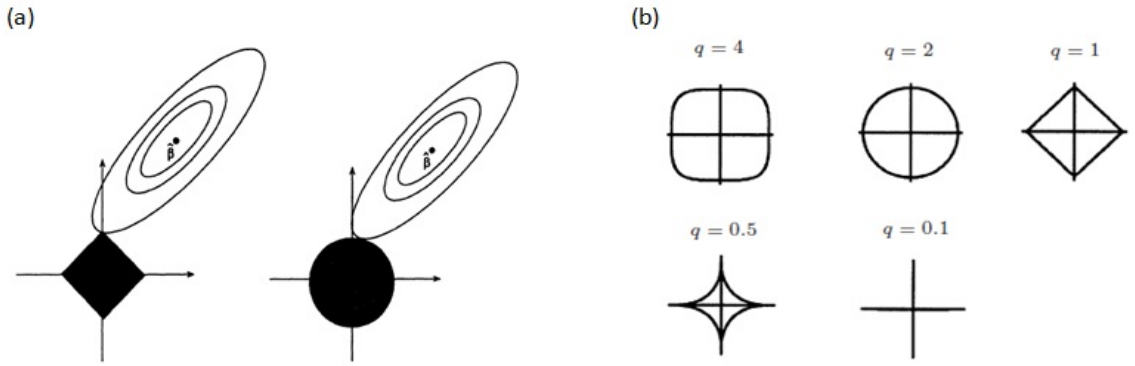

Figure 2.1: (a) Estimation picture in two dimensions for the lasso (left) and ridge regression (right). Shown are contours of the objective and constraint functions. The solid areas are the constraint regions, while the ellipses are the contours of the least square objective function. (b) Contours of constant value of the constraint regions $\sum_{j=1}^{2} |\beta_j|^q$ for given values of q.

perform well.

Table 2.1: Estimator of β_j in the case of orthornormal regression matrix. M and λ are constants chosen by the corresponding technique; sign denotes the sign of its arguments ($\pm$) and x_+ denotes "positive part" of x.

Estimator	Formula
Best-subset (size M)	$\hat{\beta}_j^0 I(\lvert\hat{\beta}_j^0\rvert \geq \lvert\hat{\beta}_{(M)}^0\rvert)$
Ridge	$\hat{\beta}_j^0/(1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j^0)(\lvert\hat{\beta}_j^0\rvert - \lambda)_+$

From the theoretical perspective, Fan and Li (2001, 2002) argued that a good penalty function should result in an estimator with the following three properties: unbiasedness for a large true coefficient to avoid excessive estimation bias, sparsity (estimating a small coefficient as zero) to reduce model complexity, and continuity to avoid unnecessary variation in model prediction. They demonstrated that bridge penalties did not satisfy all three properties. In particular, when $q > 1$, it does not produce sparse solution; when $q < 1$, the solution is not continuous; the only sparse and continuous

solution in this family is lasso ($q = 1$), but this comes at the price of shifting the resulting estimator by a constant λ. Furthermore, Fan and Li (2001, 2002) introduced the concept of oracle properties, that is, with appropriate choice of the regularization parameter, the true regression coefficients that are zero are automatically estimated as zero, and the remaining coefficients are estimated as well as if the correct submodel were known in advance. Knight and Fu (2000) showed that the bridge estimator with $0 < q < 1$ possessed the oracle properties. Zou (2006) proved that the lasso ($q = 1$) could not be an oracle procedure, except in some simple settings such as orthonormal regression matrix or only two predictor variables, otherwise, a nontrivial condition was required for the underlying model to make the lasso selection consistent.

To overcome the limitations of bridge penalties, Fan and Li (2001) proposed the smoothly clipped absolute deviation penalty (SCAD), defined by its first derivative,

$$p'_\lambda(\beta_j) = \lambda \left\{ I(\beta_j \le \lambda) + \frac{(a\lambda - \beta_j)_+}{(a-1)\lambda} I(\beta_j > \lambda) \right\}$$

for some $a > 2$ and $\beta_j > 0$, with $p_\lambda(0) = 0$, where $p_\lambda(.)$ is the $\lambda p(|.|)$ in (1), $j = 1, \ldots, d$. From the Bayesian statistical point of view, they suggested using a=3.7. The SCAD improves the lasso via penalizing large coefficients equally (e.g., see Figure 2.2(f)), and as a result, it has all the precedingly discussed theoretical properties. However, the nonconvex form of SCAD penalty makes its optimization challenging in practice, and the solutions may suffer from numerical instability (Zhang and Lu, 2007). Later, Zou (2006) proposed a new version of the lasso, called the adaptive lasso, where adaptive weights are used for penalizing different coefficients in the lasso penalty, that is, $p(|\beta_j|) = w_j|\beta_j|$, $j = 1, \ldots, d$. The weights are data-dependent and in the form of $w_j = 1/|\hat{\beta}_j|^\gamma$ with $\gamma > 0$, where any consistent estimator of β_j can be used, $j = 1, \ldots, d$. As the sample size grows, the weights for zero coefficient predictors get inflated (to infinity), whereas the weights for nonzero coefficient predictors converge to finite constant, which in some

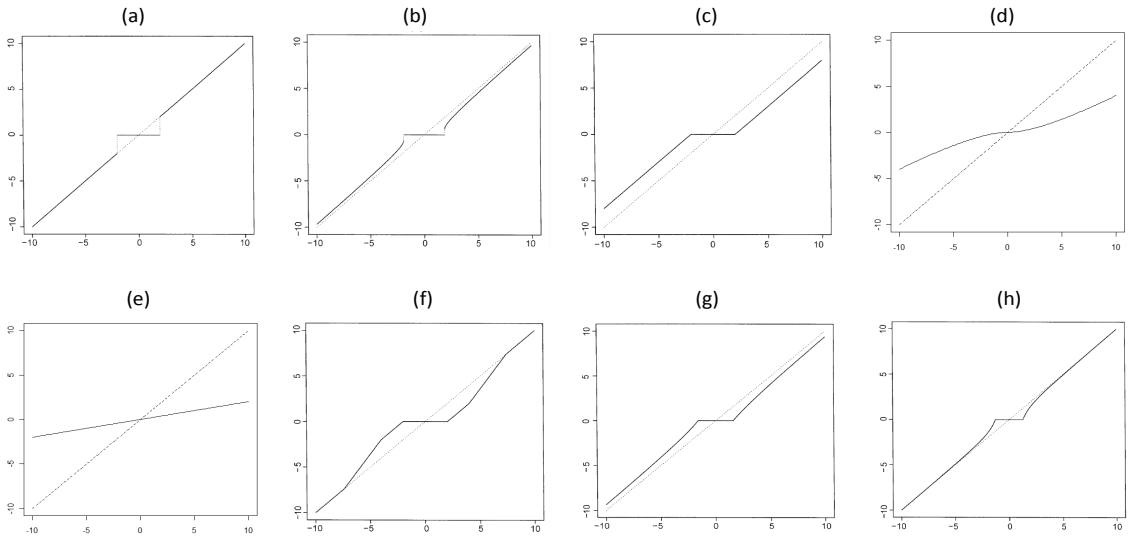

Figure 2.2: Plot of shrinkage functions with $\lambda = 2$ for (a) the best-subset; (b) the bridge, $q = 0.5$; (c) the lasso; (d) the bridge, $q = 1.5$; (e) the ridge; (f) the SCAD, a=3.7; (g) the adaptive lasso, $\gamma = 0.5$; (h) the adaptive lasso, $\gamma = 2$. The shrinkage functions are estimated under orthonormal regression matrix by minimizing $\frac{1}{2}(\beta_j^0 - \beta_j)^2 + p_\lambda(|\beta_j|)$, where β_j^0 is the OLS estimate plotted on the diagonal.

sense is the same rationale behind the SCAD (e.g., see Figure 2.2(g)-(h)), and as a result, the adaptive lasso is also an oracle procedure with continuity. Computationally, the adaptive lasso is a convex penalty, so the optimization problem does not suffer from the multiple local minima issue. Moreover, it is essentially a lasso penalization method so that all the current efficient algorithms for solving the lasso can be used to compute the adaptive lasso estimates.

Because of the attractive properties, the lasso penalty has been generalized to improve its performance in some special problems. Zou and Hastie (2005) proposed the elastic net, where the penalty included both lasso-type thresholding and ridge-type shrinkage. The elastic net enjoys a sparsity of representation, and also encourages a grouping effect, where strongly correlated predictors tend to be or out of the model together. Meanwhile, Tibshirani et. al (2005) proposed the fused lasso, designed for problems with features that could be ordered in some meaningful way. The fused lasso

penalizes the L_1 norm of both the coefficients and their successive differences, and the sparsity property applies to the number of sequences of identical nonzero coefficients. Later, Meinshausen (2006) proposed the relaxed lasso, which used the lasso to select the set of nonzero predictors, and then applied the lasso again, but using only the selected predictors from the first step. For data where there are a very large number of noise variables, the relaxed lasso has sparser estimates and much more accurate predictions than lasso. Recently, Radchenko and James (2011) proposed a method that could adaptively adjust the level of shrinkage, not just on the final model coefficients, as used in the relaxed lasso, but also during the selection of potential candidate models. They call this method forward-lasso adaptive shrinkage, which incorporates both forward selection and lasso as special cases, and can work well in situations where neither forward selection nor lasso succeeds.

2.2.2 Application of Variable Selection Methods to Censored Data

Extending penalized variable selection to survival analysis presents a number of challenges because of the complicated data structure, and therefore receives much attention in the recent literature. For Cox proportional hazards model, the objective function is parametric partial likelihood, denoted by $l_i(y_i; \delta_i; z_i^T\beta)$, where the collected data (y_i, δ_i, z_i) are independent samples, y_i is the minimum of the failure time and censoring time, and δ_i is the censoring indicator. A general form of penalized partial likelihood is

$$\sum_{i=1}^{n} l_i(y_i; \delta_i; z_i^T\beta) - n\sum_{j=1}^{d} p_\lambda(|\beta_j|).$$

Different penalties have been applied such as lasso (Tibshirani, 1997), SCAD (Fan and Li, 2002), and adaptive lasso (Zhang and Lu, 2007). With an appropriate choice of the regularization parameter λ, the corresponding SCAD and adaptive lasso estimators were shown to be root-n consistent and have the oracle properties.

For accelerated failure time model, Johnson (2008) explored the penalized weighted rank-based statistics and the penalized Buckley-James statistics. Both are challenging because of the discontinuity and non-monotone in the regression coefficients. Motivated by these, Johnson et al. (2008) established the general theory for a board class of penalized estimating functions. Suppose that $U(\beta) \equiv (U_1(\beta), \ldots, U_d(\beta))^T$ is an estimating function for β based on a random sample of size n. They mainly studied the situations where $U(\beta)$ was not a score function or the derivative of any objective function. A penalized estimating function is defined as

$$U^P(\beta) = U(\beta) - nq_\lambda(|\beta|)\text{sign}(\beta),$$

where $q_\lambda(|\beta|)$ are coefficient-dependent continuous functions, and the second term is the component-wise product of q_λ and $\text{sign}(\beta)$. With the commonly used SCAD or adaptive lasso penalty, the resulting estimators were shown to be root-n consistent and enjoy the oracle properties. Johnson (2009) further improved the approximate zero-crossing of penalized Buckley-James estimating function by using a one-step imputation and a principled initial value. The one-step estimator is an exact zero-crossing, and with lasso penalty, it reduces to Tibshirani's lasso as the proportion of censored observations approaches zero.

Comparatively, there are very few literatures on variable selection in transformation models, mainly due to the computational difficulties from non-concave likelihood functions and the presence of infinite-dimensional nuisance parameters. Particularly, Lu and Zhang (2007) studied the proportional odds model by maximizing the marginal likelihood of ranks subject to a shrinkage penalty. Based on the notation of model

(2.3), the marginal likelihood is represented as

$$L_{n,M}(\beta) = \int_{V_{(1)}<\ldots<V_{(K)}} \int \prod_{i=1}^{n} \{\lambda(V_{(k_i)} + \beta^T Z_i)\}^{\delta_i} e^{-\Lambda(V_{(k_i)}+\beta^T Z_i)} \prod_{k=1}^{K} dV_{(k)},$$

where $V_{(k)} = H(T_{(k)})$, $T_{(1)} < \ldots < T_{(K)}$ are ordered uncensored failure times in the sample, δ_i is the censoring indicator, $\Lambda(x)$ is the cumulative hazard function of ϵ, and $\lambda(x) = d\Lambda(x)/dx$. Since the marginal likelihood does not have a closed form, they approximated the high dimensional integrals and implemented the procedure by the computationally intensive MCMC algorithm, and did not give corresponding large sample properties.

Later, Zhang et al. (2010) proposed a penalized estimating equation estimator for linear transformation models (2.3). The estimator was constructed based on the martingale difference equation for the unknown transformation function and the martingale integral equation for regression coefficients as in Chen et al. (2002). To tackle the difficulties of infinite dimensional parameter H, Zhang et al. (2010) introduced the notion of the 'profiled' score, which was computed by plugging in the solution $\tilde{H}$ using the current estimate of β. Let $N_i(t)$ and $Y_i(t)$ respectively denote the counting and at-risk process, and the 'profile' score is

$$U_n(\beta) = \sum_{i=1}^{n} \int_0^{\tau} Z_i[dN_i(t) - Y_i(t)d\Lambda\{\beta^T Z_i + \tilde{H}(t;\beta)\}].$$

Then they used U_n and its variance estimate to construct a loss function as

$$D_n(\beta) = U_n'(\beta)\tilde{V}_n^{-1}U_n(\beta),$$

where the inverse variance $\tilde{V}_n^{-1}$ of the profiled score $U_n(\beta)$ is the weight matrix. To

achieve the sparse estimation, they finally proposed minimizing

$$Q_n(\beta) = D_n(\beta) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|).$$

By adopting the adaptive lasso penalty, they proved the root-n consistency and oracle properties for the resulting estimator. However, their method has the following limitations: (i) the implementation needs to solve the nuisance parameters by iteration; (ii) the resulting adaptive lasso estimator is not asymptotically efficient.

Recently, Li and Gu (2012) extended the approach of penalized marginal likelihood of ranks (Lu and Zhang, 2007) to a class of general transformation models with the form of

$$S_Z(t) = \Phi(S_0(t), Z, \beta),$$

where $S_Z(t)$ is the conditional survival function of failure time T given covariate vector Z; $S_0(t)$ is a completely unspecified baseline survival function; $\Phi(u, v, w)$ is a known monotonically increasing function with respect to u satisfying $\Phi(0, v, w) = 0$ and $\Phi(0, v, w) = 1$ for any v and w. Denote k_n as the total number of uncensored failure times, R_n^* as the partial ranking among the k_n uncensored failure times and the specified observations between each pair of uncensored observations, and L_{i_r} as the set of labels corresponding to those observations censored in interval $[T_{i_r}, T_{i_{r+1}})$. The rank-based marginal likelihood function is defined by

$$L_n(\beta) = \Pr(T_n \in C_n | R_n^*, Z) = \int_{t \in C_n} (-1)^n \prod_{i=1}^{n} \phi(S_0(t_i), Z_i, \beta) \prod_{i=1}^{n} dS_0(t_i),$$

where $C_n = \{(t_1, \ldots, t_n) : t_{i_1} < \ldots < t_{i_{k_n}}, t_j \geq t_{i_r}, \text{ for } j \in L_{i_r} \text{ and } 0 \leq r \leq k_n\}$ and $\phi(u, v, w) = \partial\Phi(u, v, w)/\partial u$. Under certain regularity conditions, Li and Gu (2012) resolved the theoretical limitation of Lu and Zhang's procedure by giving large sample properties. Specifically, using the adaptive lasso penalty, the corresponding estimator

was shown to be root-n consistent and satisfy oracle properties. However, their implementation is still based on the MCMC algorithm, which is computationally intensive. In addition, it is not clear whether time-varying covariates can be included. Therefore, the current methods for variable selection in transformation models still leave a lot to be desired.

2.3 Statistical Learning for Censored Data

The science of learning plays a key role in the fields of statistics, data mining and artificial intelligence. In a typical scenario, we have a training set of data in which we observe the outcome and feature measurement for a set of objects. The goal is to build a prediction model, or learner, which will enable us to predict the outcome for new unseen objects. This is called supervised learning because of the presence of the outcome variable to guide the learning process and a good learner is one that accurately predicts such an outcome. A review of some popular supervised learning methods (Hastie et al., 2009) and their applications to censored data is given below.

2.3.1 Supervised Learning Methods

In supervised learning we seek a function $f(X)$ for predicting Y given values of the input X. We also need a loss function $L(Y, f(X))$ for penalizing errors in prediction. For particular data sets, our goal is to find a useful approximation $\hat{f}(x)$ to the function $f(x)$ that underlies the predictive relationship between the inputs and outputs, however, minimizing the empirical loss functions may lead to infinitely many solutions. Hence, we must restrict the eligible solutions of $f(x)$ to a smaller set of functions. These restrictions are sometimes encoded via the parametric representation of f or may be built into the learning method itself, either implicitly or explicitly. In general the constraints imposed by most learning methods can be described as some

kind of regular behavior in small neighborhoods of the input space. The larger the size of the neighborhood, the stronger the constraint, and the more sensitive the solution is to the particular choice of constraint.

The linear model $f(x) = x^T\beta$ makes stringent assumptions about the structure and yields stable but possibly inaccurate predictions. It relies heavily on the assumption that a linear decision boundary is appropriate. Comparatively, the method of k-nearest neighbors is essentially model-free and assumes $f(x)$ is well approximated by a locally constant function. The resulting prediction is often accurate but can be unstable. These two simple procedures are the basis for a large subset of popular techniques, such as kernel smoothing, basis expansions, generalized additive model, projection pursuit regression (PPR) model, neural network and so forth. Neural network is a two-stage regression or classification model, typically represented by a network diagram. Interpretation of the fitted model is usually difficult, because each input enters into the model in a complex and multifaceted way. As a result, it is most useful for prediction, but not very useful for producing an understandable model for data.

Unlike the neural network, tree-based methods often yield classification and prediction rules that are relatively easy to interpret for a wide variety of applications and became popular due in great part to the development of the CART (tree-based regression and classification) paradigm (Bou-Hamad et al., 2011). The basic idea of a tree is to partition the covariate space recursively to form groups (nodes in the tree) of subjects which are similar according to the outcome of interest. The typical algorithm starts at the root node with all observations; perform an exhaustive search through all potential binary splits with the covariates; and selects the one by minimizing a measure of node impurity. In the CART approach, the process is repeated recursively on the children nodes until a stopping criterion is met (often until a minimum node size is attained). This tends to produce a large tree that usually overfits the data. Then this

large tree is pruned using cost complexity pruning.

One major problem with trees is their high variability, that is, often a small change in the data can result in a very different series of splits. Bagging is a technique for reducing the variance by fitting the same tree many times to bootstrap-sampled versions of the training data. Another popular ensemble method is random forest, which improves the variance reduction of bagging by reducing the correlation between the trees. This is achieved in the tree-growing process through random selection of the input variables as candidates for splitting. As in bagging, the bias of a random forest is the same as the bias of any of the individual sampled trees. Hence, the improvement in prediction obtained by bagging or random forests is solely a result of variance reduction.

Another popular learning machine is support vector machines (SVMs) that produce decision boundaries for classification. This method has been applied in many areas, such as financial time series forecasting, determination of the layered structure of the earth, identification of human genes, content based image retrieval, intrusion detection of computer networks and so forth (Tay and Cao, 2001; Hidalgo et al., 2003; Fernandez and Miranda-Saavedra, 2012; Rao et al., 2010; Ganapathy et al., 2012). SVMs differ radically from comparable approaches such as neural networks, since their training always finds a global minimum and their simple geometric interpretation provides fertile ground for further investigation (Burges, 1998). Suppose that the training data consist of N pairs (x_i, y_i), $i = 1, \ldots, N$, with $y_i \in \{-1, 1\}$, and define a hyperplane by $\{x : f(x) = x^T\beta + \beta_0 = 0\}$, then a classification rule induced by $f(x)$ is $\text{sign}[x^T\beta + \beta_0]$. The goal is to find the hyperplane that explicitly tries to separate the data into different classes 1 and -1 as well as possible, else finds the hyperplane that minimizes some measure of

overlap in the training data (Figure 2.3(a)). This concept is captured by

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\zeta_i$$
$$\text{subject to } \zeta_i \geq 0,\ y_i(x_i^T\beta + \beta_0) \geq 1 - \zeta_i, \forall i$$

where the value ζ_i is the proportional amount by which the prediction $f(x_i)$ is on the wrong side of its margin and misclassifications occur when $\zeta_i > 1$; the parameter C is 'cost' parameter and the separable case corresponds to $C = \infty$. This is a convex quadratic programming problem, since the objective function is itself convex, and those points which satisfy the constraints also form a convex set. This problem can be converted to its dual form by differentiating the corresponding Lagrangian function with respect to β, β_0 and ζ_i, solving the results, and substituting the expressions back, and the dual objective function is

$$L_D = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\alpha_i\alpha_{i'}y_iy_{i'}x_i^Tx_{i'},$$

where α_is are non-negative parameters. In this formulation, the training data will only appear in the form of inner products between vectors, so $x_i^Tx_i$ can be replaced by a kernel function $K(x_i, x_{i'}) = \langle h(x_i), h(x_{i'})\rangle$ to map data into a richer feature space including non-linear features and allows SVMs to form nonlinear boundaries. The transformation h needs not be specified at all and only knowledge of the kernel function is required. In the solution of this problem, those points for which $\alpha_i > 0$ are called support vectors, which lie closest to the decision hyperplance, are most difficult to classify, and would change the position of the decision hyperplane if removed.

On the other hand, with $f(x) = h(x)^T\beta + \beta_0$, the optimization problem can be

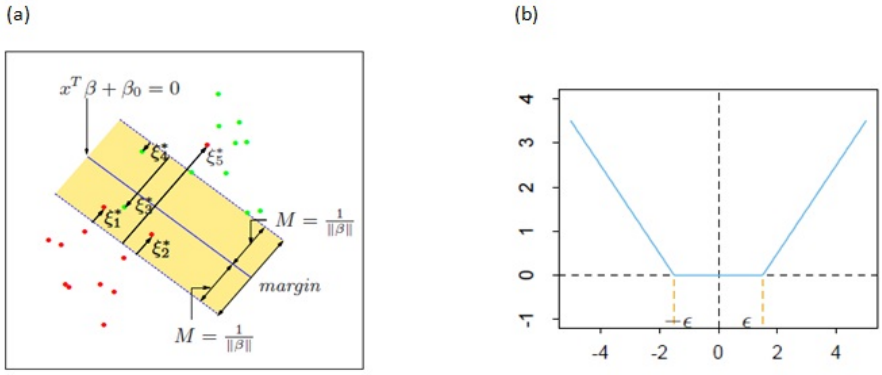

Figure 2.3: (a) Nonseparable support vector machine for classification. (b) ϵ-insensitive error function used by the support vector regression.

written as a penalization method,

$$\min_{\beta,\beta_0} \sum_{i=1}^{N} [1 - y_i f(x_i)]_+ \frac{\lambda}{2} \|\beta\|^2 ,$$

where the subscript '+' indicates the positive part of the function, and $\lambda = 1/C$. The loss function $L(y, f) = [1 - yf]_+$ is called 'hinge' loss and is reasonable for two-class classification when compared to other more traditional loss functions. The SVMs can also be adapted for regression with a quantitative response by using the ϵ-insensitive loss $L(y, f) = [|f - y| - \epsilon]_+$, $\epsilon > 0$ (Figure 2.3(b)). A smaller ϵ leads to more support vectors and an increased complexity. The ϵ-insensitive loss is zero as long as the absolute difference between the actual and predicted values is less than ϵ, and grows linearly when this absolute difference exceeds ϵ. Perhaps the biggest limitation of support vector approaches lies in the choice of the kernel (Burges, 1998; Scholkopf et al., 1998). This choice, and hence of the feature space to work in, is of both theoretical and practical interest. In addition, there is still missing an application where support vector methods significantly outperform any other available algorithm or solve a problem that has so far been impossible to tackle (Scholkopf et al., 1998).

2.3.2 Application of Supervised Learning Methods to Censored Data

Many problems of medical prediction involve the use of right censored survival data, and censoring in the data is the main reason why standard supervised machine learning techniques are hard to use for modeling survival. Ripley and Ripley (2001) and Ripley et al. (2004) discussed and described models for survival analysis which is based on neural network. These models allow non-linear predictors to be fitted implicitly and the effect of the covariates to vary over time.

- In a discrete survival time context, most neural network survival methods are based on dividing up the survival time into discrete intervals, and estimating the probability of an event in each interval. With two intervals, survival is considered binary and this is an extension of logistic regression, where each censored patient is included twice, once as an event and once as a non-event. With more than two intervals, one way is to divide the survival time into one of a set of non-overlapping intervals, and view the outputs of the network as the absolute probability of an event in a particular interval. Another alternative is to model the conditional probability of an event given no events in the previous interval. Biganzoli et al. (1998) considered the feed forward neural networks with one input node assigned to each explanatory variable and an additional input for the time interval. This approach used entropy error function and could be easily implemented using software packages based on back-propagation.

- In a continuous survival time context, the models are based on the observed likelihood function. One approach is to use the parametric survival distributions with logarithm of the hazard replaced by the output of a neural network. Alternatively, Faraggi and Simon (1995) suggested a non-linear proportional hazards

model based on the input-output relationship associated with a simple feed forward network. They replaced the linear function βx in the partial likelihood by the output of the network and obtained the maximum likelihood estimates of the parameters of the neural network using the Newton-Raphson method.

Survival trees are popular nonparametric alternatives to (semi) parametric models. They offer great flexibility and can automatically detect certain types of interactions without the need to specify them beforehand (Bou-Hamad et al. 2011). In recent years considerable research effort has been dedicated toward extending classical trees to the case of censored data. These researches focused on utilizing different splitting and pruning criteria to involve survival time and censoring information. Segal (1988) replaced the conventional splitting rules with rules based on the Tarone-Ware or Harrington-Fleming classes of two-sample statistics, which measured the between-node separation instead of the within-node homogeneity. Leblanc and Crowley (1993) further generalized Segal's method by introducing a new algorithm that automatically chose the size of a tree and gave optimally pruned subtrees. They defined a measure of tree performance analogous to the cost complexity of CART for recursive partitions based on two-sample statistics and called it split complexity. Alternatively, Leblanc and Crowley (1992) proposed a tree-structured method that adopted the proportional hazards model and gave the relative risk estimates for censored survival data. In particular, the splitting criterion was based on a node deviance measure between a saturated model log-likelihood and a maximized log-likelihood, which was a measure of within-node error. The advantage of this method is that it can be implemented easily in any recursive partitioning software for Poisson trees (Bou-Hamad et al. 2011).

Survival trees are ideal candidates for combination by means of an ensemble method and can thus be transformed into very powerful predictive tools (Bou-Hamad et al.

2011). Hothorn et al. (2004) improved predicted survival probability functions of censored event free survival by bagging survival trees. They computed a set a of survival trees based on bootstrap samples using the Leblanc and Crowley (1992) method, and then defined the aggregated Kaplan-Meier curve of a new observation by the Kaplan-Meier curve of all observations identified by the leaves containing the new observation. Later, Ishwaran et al. (2008) introduced the random survival forests for right censored data. Specifically, using independent bootstrap samples, each tree was grown by randomly selecting a subset of variables at each node and then splitting the node using a survival criterion involving the survival time and the censoring status information, and a tree was considered fully grown when each terminal node had no fewer than certain amount of unique deaths. Besides the several papers discussed here, extensive research on tree-based methods for the analysis of survival data with censoring was published over the last 25 years, reviewed by Bou-Hamad et al. (2011). The authors also covered more complex models, more specialized methods, and more specific problems such as multivariate data, the use of time-varying covariates, and discrete-scale survival data.

The appeal of support vector approaches derives from the fact that they are easy to compute and they enable estimation under weak or no assumptions on the distribution. Different methods have been suggested to adapt the support vector learning to censored data. Shivaswamy et al. (2007) proposed a support vector technique for regression on censored targets by generalizing the ϵ-insensitive loss function (Figure 2.4(a)). They considered the data set including censored targets that have covariates x_i and are within open-end intervals (l_i, u_i) with $l_i < u_i$, $i = 1, \ldots, n$, and penalized only if the predicted value $f(x_i)$ was more than u_i or if it is less than l_i. Thus, they gave the loss function for this case by

$$c(f(x_i), l_i, u_i) = \max(l_i - f(x_i), f(x_i) - u_i).$$

When $l_i = -\infty$ (left censored) or $u_i = +\infty$ (right censored), this loss function became one sided. Suppose that f is linear, $f(x_i) = w^T x_i + b$; the formulation proposed for the censored dataset is:

$$\begin{aligned}
&\min_{w,b,\zeta,\zeta^*} \frac{1}{2}\|w\|^2 + C\left(\sum_{i\in U}\zeta_i + \sum_{i\in L}\zeta_i^*\right) \\
&\text{subject to } w_i^T x_i + b - u_i \le \zeta_i, \ \forall i \in U \\
&\qquad l_i - w_i^T x_i - b \le \zeta_i^*, \ \forall i \in L \\
&\qquad \zeta_i \ge 0 \ \forall i \in U; \ \zeta_i^* \ge 0 \ \forall i \in L,
\end{aligned}$$

where L contains the indices of those samples whose targets have a finite lower bound while U contains the indices of those having a finite upper bound. This formulation was also shown to be equivalent to the support vector machine and the support vector regression by setting l_i and u_i appropriately. For non-censored targets, the support vector regression was used in their method. However, this method penalized incorrect predictions for left (right) censored data only if the prediction is higher (lower) than the observed censoring time, and penalized incorrect predictions the same whether the prediction was higher or lower than the observed event time (Van Belle et al. 2011b). Later, Khan and Zubek (2008) proposed an asymmetric modification to the ϵ-insensitive loss function which allowed censored data to be processed and accounted for the differences between censored and event instances (Figure 2.4(b)). Their method provided different losses for events and censored data and for predictions higher and lower than the observed time, and correspondingly in the formulation used different costs C and slack variables ζ for different situations. As a result, the major drawback was the large number of parameters to be estimated.

Alternatively, Van Belle et al. (2009) proposed the use of a least-squares support vector machine for right censored survival data. For event data, the same constraints

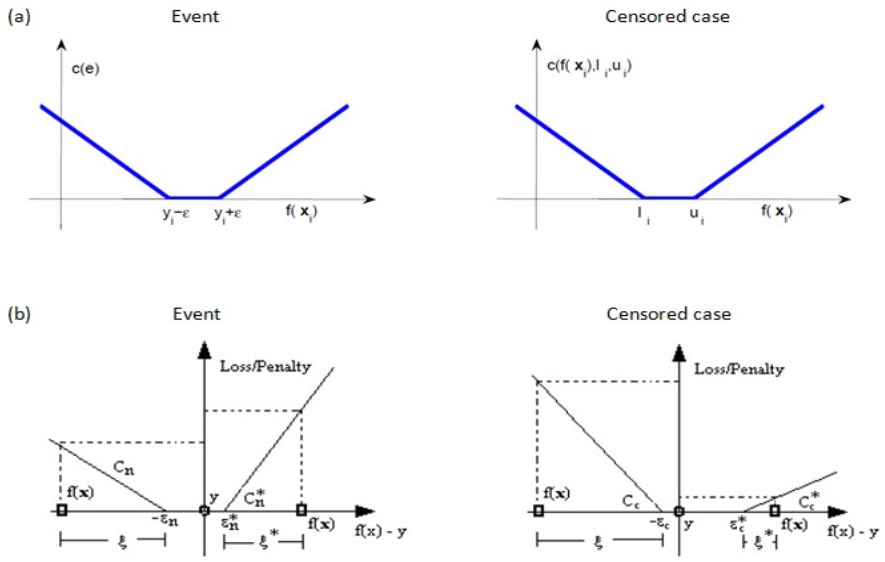

Figure 2.4: (a) Loss functions as defined by Shivaswamy et al. (2007). (b) Loss functions as defined by Khan and Zubek (2008).

accounted as in the standard support vector regression method. To handle censored data, they adopted the concept of concordance index and added the ranking constraints for all comparable data pairs. A data pair is said to be comparable whenever the order of their observed times is known, such as two events, an event and a right censored instances for which the censoring time of the latter is later than the event time of the former, and so forth. They considered the data points (x_i, y_i, δ_i), $i = 1\ldots, n$, where x_is are covariates, y_is are observed times and δ_is are censoring indicators, and assumed that the observed times were ordered ($y_i < y_j$ for $i < j$), then the ranking constraints for predicted values were defined by

$$f(x_j) - f(x_i) \geq 1 - \zeta_{ij}, \ \ \forall i < j,$$

where slack variables $\zeta_{ij} \geq 0$ were allowed for misranking and the sum of ζ_{ij} over all comparable pairs were minimized in the formulation. Later, Van Belle at al. (2011a) proposed a computationally simplified approach by modifying the ranking constraints

to be

$$f(x_i) - f(x_{\tilde{j}(i)}) \geq 1 - \zeta_i, \ \forall i,$$

where $\tilde{j}(i)$ was the data point comparable with data point i and with the largest y_j smaller than y_i. Also, their formulation included only the ranking constraints and was for the problems whose primary interests were in defining risk groups instead of prediction of survival times. To evaluate the performance of support vector approaches for survival data, Van Belle et al. (2011b) compared several models based on ranking constraints, based on regression constraints and based on both ranking and regression constraints, and their results indicated a significant better performance for models including regression constraints than models only based on ranking constraints. However, the prediction rules to obtain the event times in these methods are not clear, and none of the above intuitive methods has theoretical justification. For example, the rank-based methods may not fully use observed event information, and it is unclear whether Van Belle et al. (2011b) is valid if the censoring time depends on the subject's covariates.

From another perspective, Park and Jeong (2011) proposed a technique called recursive support vector censored regression to make a direct prediction of survival time. Their approach replaced the censored observations by the corresponding Buckley-James estimates and conducted the estimation through a recursive procedure. It is computationally intensive and the theoretical properties were not studied. Later, Goldberg and Kosorok (2012b) developed a unified support vector approach for right censored survival data, and the general methodology to estimation was applied for the truncated mean, median, quartiles, and for classification problems. The core idea was to use the inverse-probability-of-censoring weighting to correct the bias induced by censoring, that is,

$$L(z, Y(u), s) \times \frac{\delta}{\hat{G}_n(u|z)},$$

where $L(.)$ was the original loss function, δ was the censoring indictor and $\hat{G}_n$ was a generalized Kaplan-Meier estimator for the survival function G. As a result, in their method, a different loss function was defined for each data set and minimizing the empirical loss no longer consisted of minimizing a sum of independent and identically distributed observations. They also showed that the proposed method was well defined and measurable, and derived finite sample bounds on the deviation from the optimal risk. However, their method may suffer from severe bias when the censoring distribution is misspecified. Additionally, the weights used in inverse weighting can become large in some situations. As a result, the computation of this method becomes numerically unstable and even infeasible.

By applying a similar idea, Goldberg and Kosorok (2012a) proposed a Q-learning algorithm for right censored data. Q-learning is a reinforcement learning algorithm that assigns values to action-state pairs, and learns, based on state at each decision point, how best to choose an action to maximize the expected sum of incremental rewards. This algorithm has a so-called Q function which calculates the quality of a state-action combination. Goldberg and Kosorok (2012a) adjusted the Q function by the inverse-probability-of-censoring weighting to take into account the censored observations. For a theoretical justification, they provided finite sample bounds on the average difference in expected survival time between the optimal dynamic treatment regime and the dynamic treatment regime obtained by the proposed Q-learning algorithm.

CHAPTER3: VARIABLE SELECTION IN SEMIPARAMETRIC TRANSFORMATION MODELS

3.1 Methodology

3.1.1 Transformation Models

We let $Z(\cdot) = \{Z_1(\cdot), \ldots, Z_d(\cdot)\}^T$ denote a vector of d-dimensional possibly time-varying covariates used for predicting survival outcome T. A general transformation model assumes that the cumulative hazard function of T given $Z(\cdot)$ is

$$\Lambda\{t \mid Z(\cdot)\} = G\left\{\int_0^t e^{\beta^T Z(s)} d\Lambda(s)\right\}, \tag{3.1}$$

where $\Lambda(\cdot)$ is a completely unspecified cumulative hazard function, and $\beta = (\beta_1, \ldots, \beta_d)^T$ is an unknown vector of regression coefficients. If all covariates are time-invariant, the above model is equivalent to $\log \Lambda(T) = -\beta^T Z + \log G^{-1}(-\log \epsilon_0)$, where ϵ_0 has a uniform distribution.

The transformation G is assumed to have the form $G(x) = -\log \int_0^\infty e^{-x\zeta}\phi(\zeta)d\zeta$, where $\phi(\zeta)$ is a known density function on $[0, \infty)$. A commonly used choice of $\phi(\zeta)$ is the gamma density with unit mean and variance r. Then $G(x)$ arises from a class of logarithmic transformations (Chen et al., 2002):

$$G(x) = \begin{cases} \log(1 + rx)/r, & r > 0, \\ x, & r = 0. \end{cases}$$

If $r = 0$, the transformation model is exactly the Cox proportional hazards model; if $r =$

1, $G(x) = \log(1+x)$ results in the proportional odds model. This class of transformations is commonly used, although any transformation induced by some density $\phi(\cdot)$ with support in $[0, \infty)$ is applicable.

Suppose a random sample of n subjects is chosen. Let T_i denote the failure time and C_i denote the censoring time of the ith subject, respectively. Define the observed time $Y_i = \min(T_i, C_i)$ and the censoring indicator $\Delta_i = I(T_i \leq C_i)$. Let $Z_i(\cdot) = \{Z_{i1}(\cdot), \ldots, Z_{id}(\cdot)\}^T$ be the corresponding vector of time-varying covariates for the ith subject. Thus, the observed data consist of $\{Y_i, \Delta_i, Z_i(\cdot)\}$, for $i = 1, \ldots, n$. Here we consider only external time-varying covariates, that is, the whole trajectory of $Z_i(\cdot)$ is observable. Assume that T_i and C_i are conditionally independent given $Z_i(\cdot)$, and the censoring mechanism is noninformative. Under the transformation model (3.1), the likelihood function for the observed data is

$$\prod_{i=1}^{n} \left[\Lambda'(Y_i) e^{\beta^T Z_i(Y_i)} G' \left\{ \int_0^{Y_i} e^{\beta^T Z_i(s)} d\Lambda(s) \right\} \right]^{\Delta_i} \times \exp\left[-G\left\{ \int_0^{Y_i} e^{\beta^T Z_i(s)} d\Lambda(s) \right\} \right], \quad (3.2)$$

where $\Lambda'(Y_i)$ is the derivative of Λ at Y_i. Expression (3.2) involves both β and the infinite dimensional parameter Λ, and may not be concave in these parameters. Also, there is no partial likelihood function available due to the transformation G. Thus, directly applying the penalized methods in Fan and Li (2002) or Zhang and Lu (2007) for variable selection is no longer feasible.

To resolve this difficulty, we adopt the method proposed by Zeng and Lin (2007). The idea is to treat ζ as a latent variable, in which case model (3.1) is equivalent to the survival time T with cumulative hazard function

$$\Lambda\{t \mid Z(\cdot), \zeta\} = \zeta \int_0^t e^{\beta^T Z(s)} d\Lambda(s), \quad (3.3)$$

because

$$\begin{aligned}
\text{pr}\{T > t \mid Z(\cdot)\} &= E[\text{pr}\{T > t \mid Z(\cdot), \zeta\} \mid Z(\cdot)] \\
&= E\left[\exp\left\{-\zeta \int_0^t e^{\beta^T Z(s)} d\Lambda(s)\right\} \mid Z(\cdot)\right] \\
&= \int_0^\infty \exp\left\{-\zeta \int_0^t e^{\beta^T Z(s)} d\Lambda(s)\right\} \phi(\zeta) d\zeta \\
&= \exp\left[-G\left\{\int_0^t e^{\beta^T Z(s)} d\Lambda(s)\right\}\right].
\end{aligned}$$

That is, conditional on the covariates $Z(\cdot)$ and the latent variable ζ, the survival time T follows a Cox proportional hazards model with ζ missing. Thus, instead of working on the observed data for variable selection, we work on the complete data so that the method for variable selection in the Cox proportional hazards model may be used.

The expectation-maximization algorithm is used to fit model (3.3) based on complete data, $\{Y_i, \Delta_i, Z_i(\cdot), \zeta_i\}$ $(i = 1, \ldots, n)$. In this setting, the likelihood function (3.2) becomes

$$\prod_{i=1}^n \left\{\zeta_i \delta\Lambda(Y_i) e^{\beta^T Z_i(Y_i)}\right\}^{\Delta_i} \times \exp\left\{-\zeta_i \int_0^{Y_i} e^{\beta^T Z_i(s)} d\Lambda(s)\right\} \times \phi(\zeta_i),$$

where $\Lambda'(Y_i)$ is replaced by $\delta\Lambda(Y_i)$, the jump size of Λ at Y_i, in the nonparametric maximum likelihood estimation.

The expectation-maximization algorithm consists of an expectation step and a maximization step: see the appendix of Zeng and Lin (2007). The first step computes the expected log-likelihood based on the current estimates of all the parameters, conditional on the observed data. Specifically, it computes the posterior expectation of latent variables, such as $E\{\zeta_i \mid Y, \Delta, Z(\cdot), \tilde{\beta}_k, \delta\tilde{\Lambda}_k(Y)\}$ $(i = 1, \ldots, n)$, at the kth iteration, based on the posterior density of ζ. The second step computes the estimates maximizing the

expected log-likelihood obtained in the expectation step, which is

$$\sum_{i=1}^{n}\left[\Delta_i\left\{\log\delta\Lambda(Y_i)+\beta^T Z_i(Y_i)\right\}-E\{\zeta_i\mid Y,\Delta,Z,\tilde{\beta}_k,\delta\tilde{\Lambda}_k(Y)\}\sum_{Y_j\le Y_i}e^{\beta^T Z_i(Y_j)}\delta\Lambda(Y_j)\right]. \quad (3.4)$$

After convergence, we obtain the maximum likelihood estimates $\tilde{\beta}$ and $\delta\tilde{\Lambda}(Y_i)$ $(i = 1,\ldots,n)$. The algorithm is guaranteed to converge, since the objective function (3.4) in the maximization step is increased in each iteration, and is only unchanged at convergence.

3.1.2 Variable Selection

The objective function (3.4) takes a very similar form to the Cox log-likelihood function. Based on the maximum likelihood estimates, we compute the posterior expectation $E\{\zeta_i \mid Y,\Delta,Z(\cdot),\tilde{\beta},\delta\tilde{\Lambda}(Y)\}$, denoted as the weight $\tilde{c}_i$ $(i = 1,\ldots,n)$. These weights are data-dependent. As given in the appendix of Zeng and Lin (2007),

$$\tilde{c}_i=\begin{cases} G'\left\{\sum_{j=1}^{n} I(Y_j\le Y_i)e^{\beta^T Z_i(Y_j)}\delta\Lambda(Y_j)\right\}, & \Delta_i=0,\\ -\frac{G''\left\{\sum_{j=1}^{n} I(Y_j\le Y_i)e^{\beta^T Z_i(Y_j)}\delta\Lambda(Y_j)\right\}}{G'\left\{\sum_{j=1}^{n} I(Y_j\le Y_i)e^{\beta^T Z_i(Y_j)}\delta\Lambda(Y_j)\right\}}+G'\left\{\sum_{j=1}^{n} I(Y_j\le Y_i)e^{\beta^T Z_i(Y_j)}\delta\Lambda(Y_j)\right\}, & \Delta_i=1. \end{cases}$$

Given $\tilde{c}_i$, we differentiate function (3.4) with respect to $\delta\Lambda(Y_i)$ and set it to be zero, giving

$$\delta\Lambda(Y_i)=\frac{\Delta_i}{\sum_{j=1}^{n} I(Y_j\ge Y_i)\tilde{c}_j e^{\beta^T Z_j(Y_i)}}.$$

Substituting $\delta\Lambda(Y_i)$ back into function (3.4), we obtain a weighted version of the partial log-likelihood function,

$$l_n(\beta)=\sum_{i=1}^{n}\Delta_i\left[\beta^T Z_i(Y_i)-\log\left\{\sum_{j=1}^{n} I(Y_j\ge Y_i)\tilde{c}_j e^{\beta^T Z_j(Y_i)}\right\}\right]. \quad (3.5)$$

We use function (3.5) to accommodate penalties for variable selection. This function

is the objective function in the maximization step of the expectation-maximization algorithm, which results in the efficient maximum likelihood estimator $\tilde{\beta}$ if maximized without penalties (Zeng and Lin, 2006, 2007). An important advantage of function (3.5) is its strict concavity, as shown in the appendix. For the Cox proportional hazards model, (3.5) reduces to the partial likelihood function. These properties enable us to adopt similar procedures for the implementation to those for the Cox model, and to derive nice theoretical results for the estimator after variable selection.

Although many penalties can be applied with function (3.5), here we use the convex adaptive lasso penalty for computational simplicity. This penalty adapts each coefficient with a weight to reflect the importance of the corresponding covariate, which is equivalent to using different tuning parameters for different coefficients. The coefficients of unimportant covariates are assigned larger weights so that they can be shrunk to zero more easily, leading to the oracle property (Zou, 2006). The reciprocal of any consistent estimator β can be used as the adapting weights; here we take the maximum likelihood estimator $\tilde{\beta}$. Writing $\beta = (\beta_1, \ldots, \beta_d)$, the corresponding penalized objective function is

$$-l_n(\beta) + \lambda \sum_{j=1}^{d} |\beta_j| / |\tilde{\beta}_j|^{\gamma}, \tag{3.6}$$

where γ is a given positive constant.

To obtain the adaptive lasso estimates $\hat{\beta}$, we need to minimize function (3.6). Assume that the covariates $Z_{ij}(\cdot)$ are standardized so that $\sum_{i=1}^{n} Z_{ij}(\cdot)/n = 0$ and $\sum_{i=1}^{n} Z_{ij}^2(\cdot)/n = 1$. We modify the computational algorithm of Zhang and Lu (2007) for the proportional hazards model. The strategy is to approximate the weighted partial likelihood function as an iterative least squares step using a Newton–Raphson update. Define the gradient vector $\nabla l(\beta) = -\partial l_n(\beta)/\partial\beta$ and the Hessian matrix $\nabla^2 l(\beta) = -\partial^2 l_n(\beta)/\partial\beta\partial\beta^T$. Consider the Cholesky decomposition of $\nabla^2 l(\beta)$, i.e., $\nabla^2 l(\beta) = X^T X$, and set the pseudo response vector $W = (X^T)^{-1}\{\nabla^2 l(\beta)\beta - \nabla l(\beta)\}$. Then a second-order

Taylor expansion for $-l_n(\beta)$ has the form

$$\frac{1}{2}(W - X\beta)^T(W - X\beta). \tag{3.7}$$

Hence to minimize the original problem (3.6) for any fixed λ, we use the following procedure:

Step 1. Use the expectation-maximization algorithm to compute $\tilde{\beta}$ and $\delta\tilde{\Lambda}(Y_i)$, and then compute the weights $\tilde{c}_i$ $(i = 1, \ldots, n)$.

Step 2. Initialize by setting $\hat{\beta} = \tilde{\beta}$.

Step 3. Compute ∇l, $\nabla^2 l$, X and W based on the current values of $\hat{\beta}$.

Step 4. Use the modified shooting algorithm (Zhang and Lu, 2007) to minimize the function (3.7) plus the penalty $\lambda \sum_{j=1}^{d} |\beta_j|/|\tilde{\beta}_j|^{\gamma}$.

Step 5. Repeat Steps 3 and 4 until the convergence criterion is met.

The initialization in Step 2 reduces the number of iterations compared with setting $\hat{\beta} = 0$, since $\tilde{\beta}$ is already consistent. In addition, with $\hat{\beta} = \tilde{\beta}$ in Step 2, the estimates from the one-step iteration are fairly close to those from iteration until convergence. The minimization in Step 3 is based on a quadratic least squares function, so the path-based algorithms in the least squares setting for solving the adaptive lasso can be applied to compute the whole solution path for the one-step estimates.

In the proposed algorithm, there is a data-dependent tuning parameter λ. Like Zhang and Lu (2007), we use generalized cross validation (Craven and Wahba, 1979) to select λ. We consider λ for a set of grid points, and for each, we approximate the number of effective parameters in the adaptive lasso estimator by $p(\lambda) = \mathrm{tr}\{(\tilde{G} + \lambda A)^{-1}\tilde{G}\}$, so the generalized cross validation criterion is $-l_n(\hat{\beta})/[n\{1 - p(\lambda)/n\}^2]$, where $\tilde{G} = \nabla^2 l(\hat{\beta})$ and $A = \mathrm{diag}(|\hat{\beta}_1|^{-1}|\tilde{\beta}_1|^{-\gamma}, \ldots, |\hat{\beta}_d|^{-1}|\tilde{\beta}_d|^{-\gamma})$. The best choice of the tuning parameter λ is that yielding the smallest value of this criterion. A more stable method to select λ is V-fold cross validation, usually with $V = 5$ or 10. However, for this method, we must

partition the data into V subsets with equal sizes. For a given λ, we must compute the coefficients using $V-1$ subsets and function (3.7) plus the adaptive lasso penalty using the Vth subset V times, which is computationally much more complicated to implement than generalized cross validation. The simulation studies in Section 3.3 show that generalized cross validation works well for our models.

After variable selection, we suggest refitting model (3.3) to obtain the maximum likelihood estimates for the coefficients of selected covariates. As shown in the previous literature, although the adaptive lasso estimator is consistent, its finite sample bias can be non-negligible.

3.1.3 Standard Errors

Our method treats the transformation as missing latent variables, so the Louis formula (Louis, 1982) is used to obtain the standard errors for the maximum likelihood estimates. The Louis formula computes the observed information within the expectation-maximization framework. To apply it, we need to consider both the desired parameter β and the nuisance parameter Λ, where Λ is evaluated at each observed time point. Denote all the parameters as $\theta = \{\beta_1, \ldots, \beta_d, \delta\Lambda(Y_1), \ldots, \delta\Lambda(Y_n)\}^T$ and the log-likelihood using complete data as f_i $(i = 1, \ldots, n)$. Then the covariance matrix of θ is

$$\left(-\sum_{i=1}^{n} E\left\{ \frac{\partial^2 f_i(\theta)}{\partial \theta^2} \mid Y, \Delta, Z(\cdot) \right\} - \sum_{i=1}^{n} E\left[\left\{ \frac{\partial f_i(\theta)}{\partial \theta} \right\}^{\otimes 2} \mid Y, \Delta, Z(\cdot) \right] + \sum_{i=1}^{n} \left[E\left\{ \frac{\partial f_i(\theta)}{\partial \theta} \mid Y, \Delta, Z(\cdot) \right\} \right]^{\otimes 2} \right)^{-1}. \tag{3.8}$$

We also apply formula (3.8) to approximate the standard errors for the adaptive lasso estimates. Instead of plugging in $\tilde{\beta}$ and $\tilde{\Lambda}$, we plug in the adaptive lasso estimates $\hat{\beta}$ and the updated estimates of the nuisance parameter $\delta\hat{\Lambda}(Y)$ using $\hat{\beta}$. After obtaining

the standard errors, we set those for zero estimates to be zero, assuming that the corresponding covariates are unimportant. Compared with the sandwich formula used in Zhang and Lu (2007), formula (3.8) does not have the tuning parameter λ. Intuitively, the information of λ is carried by the adaptive lasso estimates, and λ is small in most cases. This method can work well since the adaptive lasso estimator is asymptotically efficient, and reaches the same efficiency bound as the maximum likelihood estimator. Correspondingly, this efficiency bound can be consistently estimated by the covariance matrix of the adaptive lasso estimates.

3.2 Theoretical Properties

In this section we provide asymptotic properties for our estimators. We consider the penalized objective function based on n samples: $Q_n(\beta) = l_n(\beta) - n\lambda_n \sum_{j=1}^{d} |\beta_j|/|\tilde{\beta}_j|^{\gamma}$. Denote the true values of β and Λ by β_0 and Λ_0. We write β_0 as $(\beta_{10}^T, \beta_{20}^T)^T$, where β_{10} consists of all q non-zero components and β_{20} consists of the remaining zero components. Correspondingly, we have the adaptive lasso estimator $\hat{\beta}_n = (\hat{\beta}_{1n}^T, \hat{\beta}_{2n}^T)^T$ and the maximum likelihood estimator after variable selection $\check{\beta}_n = (\check{\beta}_{1n}^T, 0)^T$. Also, we write the time-varying covariates $Z(\cdot)$ as $\{Z_1(\cdot)^T, Z_2(\cdot)^T\}^T$, where $Z_1(\cdot)$ denotes the important covariates and $Z_2(\cdot)$ denotes the unimportant covariates.

We require the following regularity conditions.

Condition 1. The function $\Lambda_0(t)$ is strictly increasing and continuously differentiable, and β_0 lies in the interior of a compact set.

Condition 2. With probability one, $Z(\cdot)$ has bounded total variation in $[0, \tau]$. In addition, if there exists a vector α and a deterministic function $\alpha_0(t)$ such that $\alpha_0(t) + \alpha^T Z(t) = 0$ with probability one, then $\alpha = 0$ and $\alpha_0(t) = 0$.

Condition 3. With probability one, there exists a positive constant a_0 such that $\mathrm{pr}(C \geq \tau \mid Z) > a_0$ and $\mathrm{pr}(T \geq \tau \mid Z) > a_0$.

Condition 4. $\limsup_{x\to\infty}\{G(m_0x)\}^{-1}\log\{x\sup_{y\le x}G'(y)\} = 0$ for any positive constant m_0.

The same conditions are used in Zeng and Lin (2006). No additional conditions are needed for the adaptive lasso estimator. Conditions 1 and 3 are standard in survival models. Condition 2 is equivalent to saying that the design matrix $\{1, Z(t)\}$ is full rank with some positive probability for all $t \in [0,\tau]$, and is used to show the strict concavity of the objective function $l_n(\beta)$. Condition 4 specifies the tail behavior of the transformation function $G(x)$. It is easy to check that the logarithmic transformation satisfies this condition.

Under Conditions 1–4, we claim the asymptotic results for our estimators.

Theorem 3.2.1. *If $n^{1/2}\lambda_n = O_p(1)$, then the adaptive lasso estimator satisfies $\|\hat\beta_n - \beta_0\| = O_p(n^{-1/2})$, where $\|\cdot\|$ denotes the Euclidean norm.*

Theorem 3.2.2. *If $n^{1/2}\lambda_n \to 0$ and $n^{(\gamma+1)/2}\lambda_n \to \infty$, then under Theorem 3.2.1, the adaptive lasso estimator $\hat\beta_n$ satisfies the following:*

(i) $\hat\beta_{2n} = 0$ with probability tending to 1;

(ii) $n^{1/2}(\hat\beta_{1n} - \beta_{10}) = n^{1/2}(P_n - P)S_{\beta_1}\{Y, \Delta, Z_1(\cdot), \beta_{10}, \Lambda_0\} + o_p(1)$, where P_n is the empirical measure, with P being the expectation, S_{β_1} is the efficient influence function for β_1 as given implicitly in Zeng and Lin (2006), and $o_p(1)$ denotes the random element converging to zero in probability in the metric space R^q. Consequently, $\hat\beta_{1n}$ is semiparametrically efficient.

Theorem 3.2.3. *The maximum likelihood estimator after variable selection $\check\beta_{1n}$ satisfies $n^{1/2}(\check\beta_{1n} - \beta_{10}) = n^{1/2}(P_n - P)S_{\beta_1}\{Y, \Delta, Z_1(\cdot), \beta_{10}, \Lambda_0\} + o_p(1)$.*

Theorem 3.2.1 indicates that the adaptive lasso estimator is consistent for the true value at the rate $n^{1/2}$. Theorem 3.2.2 indicates that the adaptive lasso estimator has the oracle property, so it behaves as if the unimportant variables were known. In addition, it is asymptotically normal and efficient for important variables. The efficiency is

due to the fact that the weighted partial likelihood function (3.5) is the objective function in the last maximization step of the expectation-maximization algorithm, so its maximizer without penalization is exactly the original maximum likelihood estimator. The additional penalization is not dominating, so it should not affect the asymptotic efficiency except by producing sparse estimation. Theorem 3.2.3 gives the theoretical properties for the maximum likelihood estimator of selected important variables after refitting the model without the adaptive lasso penalty after variable selection. Theorem 3.2.3 is from Zeng and Lin (2006). Proofs of Theorems 3.2.1 and 3.2.2 are given in the Appendix.

3.3 Simulation Studies

3.3.1 Simulation Setup

Consider the logarithmic transformation for G. We consider three transformation models with $r = 0$, $r = 0.5$ and $r = 1$, where $r = 0$ yields the proportional hazards model and $r = 1$ yields the proportional odds model. We take ten covariates in the regression model, with true $\beta = (0.3, 0.5, 0.7, 0, 0, 0, 0, 0, 0, 0)^T$, so only the first three covariates have non-zero effects. The associated ten covariates $Z = (Z_1, \ldots, Z_{10})$ are marginally standard normal with pairwise correlation $\text{corr}(Z_j, Z_k) = \rho^{|j-k|}$, where $\rho = 0.5$. The failure times T are generated from the transformation model (3.1). The Weibull distribution is assumed for the baseline cumulative hazard function, with $\Lambda(t) = at^b$ $(a, b > 0)$. To generate T, we first generate a random variable U from the uniform distribution (0, 1), and then let $T = [\{(1/U)^r - 1\} \exp(\beta^T Z)/(ar)]^{\{1/b\}}$. The censoring times are generated from a uniform distribution $(0, u_0)$, where u_0 is chosen to obtain the desired censoring ratios, and we consider censoring ratios 20% and 40%. For γ in the adaptive lasso penalty, we use $\gamma = 1$ for all the simulation studies.

For each simulated data set, we apply our method for estimation and variable selection. We first apply the expectation-maximization algorithm in Section 3.1.1 to obtain initial estimates then implement the adaptive lasso procedure in Section 3.1.2 for selecting the non-zero coefficients. To compare the performance, we also use the lasso penalty in the proposed procedure. We consider a grid $0.5, 1, 5, 10, 15, 20, 20 + (n-20)/10, \ldots, n$ for the tuning parameter λ, where n is the sample size, and report the results that generate the smallest value of the generalized cross validation criterion. After variable selection, the expectation-maximization algorithm is reapplied to the models with only selected covariates. We repeat the simulation 1000 times and consider sample sizes, $n = 100, 400$.

3.3.2 Simulation Results

Table 3.1 gives the average numbers of correct and incorrect zero coefficients, and the median of mean squared errors $(\hat{\beta} - \beta)^T \Sigma (\hat{\beta} - \beta)$, where Σ is the population covariance matrix. The adaptive lasso method performs well for all three models. It outperforms the method with lasso penalty with respect to variable selection, and gives more accurate prediction when the censoring ratio is 20%. Table 3.2 summarizes the proportions of variable selection for the adaptive lasso method, where the columns of signal-noise ratios are true β divided by the sample standard errors of the initial values. Larger ratios lead to the higher probabilities of selecting important covariates. Slightly better results are observed for 20% censoring than for 40% censoring. In particular, the important covariates almost stay in the models when the sample size is 400, and the capability of shrinking zero coefficients to zero is improved as the sample size increases from 100 to 400, which agree with the oracle property of Section 3.2.

Tables 3.3 and 3.4 test the accuracy of non-zero coefficient estimates and the proposed standard error formula for the 20% and 40% censoring cases. The results of the

two cases have similar trends. The adaptive lasso estimates are slightly biased and the bias can be considerably reduced by refitting the selected models with maximum likelihood estimation when the sample size is 400. The resulting maximum likelihood estimators also have smaller standard errors compared with those before variable selection. Inference based on the adaptive lasso estimator is not very accurate for small coefficients, and it becomes more reasonable as the coefficients get larger. When the sample size is 400, the maximum likelihood estimation has small biases and the estimated standard errors are close to the sample standard errors. The 95% confidence intervals for the maximum likelihood method based on the estimated coefficients and standard errors have accurate coverage for the true parameters. Interestingly, the maximum likelihood estimators after variable selection perform noticeably better numerically than the adaptive lasso estimators, even if both are theoretically efficient according to Theorems 3.2 and 3.3. One possible reason for this is that, in a small sample, estimation after variable selection estimates fewer parameters, so it gains more degrees of freedom in fitting data.

3.3.3 Simulation under Misspecified Transformation

Our method assumes that the transformation function is known, so we study its performance when the transformation is misspecified. We conduct simulation using the same parameter settings as in Section 3.3.1. The true data are generated from the proportional odds model, $r = 1$, but we fit the proportional hazards model, $r = 0$. The signal-noise ratios of important variables are similar to those with $r = 0$ in Table 3.2.

Table 3.5 summarizes the variable selection results and the median of mean squared errors using the adaptive lasso. Even if the proportional odds model is misspecified by the proportional hazards model, the proposed method is still able to select the correct set of important variables most of the time when the sample size is 400. However,

Table 3.1: Average numbers of correct and incorrect zero coefficients and median mean square errors from 1000 simulated data sets

			Adaptive lasso method			Lasso method		
Censoring	r	n	Corr. [a]	Incorr. [b]	MMSE [c]	Corr.	Incorr.	MMSE
20%	0	100	6.71	0.40	0.08	4.33	0.11	0.09
		400	6.95	0.01	0.03	6.09	0.00	0.11
	0.5	100	6.09	0.49	0.10	3.62	0.16	0.15
		400	6.57	0.02	0.03	6.26	0.01	0.11
	1	100	5.41	0.56	0.15	2.89	0.17	0.22
		400	6.05	0.03	0.04	5.66	0.01	0.09
40%	0	100	6.26	0.41	0.09	4.19	0.06	0.10
		400	6.84	0.01	0.03	4.01	0.00	0.02
	0.5	100	5.61	0.51	0.14	3.47	0.12	0.16
		400	6.39	0.04	0.04	3.21	0.00	0.04
	1	100	4.93	0.58	0.21	3.10	0.21	0.22
		400	5.82	0.08	0.05	2.81	0.01	0.06

[a]Corr., average number of correct zeros;

[b]Incorr., average number of incorrect zeros;

[c]MMSE, median of mean squared errors.

Table 3.2: Proportions of each covariate being selected and signal-noise ratios for important covariates based on 1000 simulated data sets for the adaptive lasso method

			Proportions of variable selection										Signal-noise ratios		
Censoring	r	n	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}	t_1	t_2	t_3
20%	0	100	68	92	100	5	4	4	4	4	4	4	2.00	2.50	3.33
		400	99	100	100	1	0	1	1	1	1	1	5.00	6.25	8.75
	0.5	100	65	89	97	14	13	11	12	14	15	12	1.58	2.00	2.69
		400	98	100	100	6	7	6	6	7	7	5	3.75	4.54	6.36
	1	100	65	84	96	23	23	23	23	23	23	23	1.43	1.67	2.41
		400	98	100	100	13	14	14	14	13	16	12	3.33	4.17	5.83
40%	0	100	70	90	99	11	10	11	10	11	11	11	1.58	2.08	2.80
		400	99	100	100	2	3	3	2	2	3	2	4.29	5.00	7.00
	0.5	100	68	84	96	19	20	21	18	21	21	19	1.43	1.72	2.41
		400	96	100	100	8	9	9	9	9	10	8	3.33	4.55	5.83
	1	100	68	81	93	28	30	32	30	29	28	30	1.15	1.52	2.06
		400	93	99	100	16	17	17	17	17	17	17	2.73	3.33	4.67

Table 3.3: Estimates of coefficients, their standard errors, and coverage probabilities for nominal 95% confidence intervals from 1000 simulated data sets for censoring ratio 20%

			MLE[a] without variable selection				Adaptive lasso estimator				MLE after variable selection			
r	n	β	Bias	SE[b]	SEE[c]	CP[d]	Bias	SE	SEE	CP	Bias	SE	SEE	CP
0	100	β_1	0.03	0.15	0.14	92	−0.08	0.13	0.13	93	0.08	0.11	0.12	93
		β_2	0.06	0.20	0.18	92	−0.05	0.20	0.17	88	0.09	0.18	0.16	89
		β_3	0.08	0.21	0.19	92	−0.10	0.20	0.18	87	0.04	0.18	0.16	92
	400	β_1	0.01	0.06	0.06	95	−0.08	0.07	0.06	69	0.01	0.06	0.06	96
		β_2	0.02	0.08	0.08	95	−0.03	0.10	0.08	86	0.01	0.08	0.08	95
		β_3	0.02	0.08	0.08	95	−0.05	0.08	0.08	91	0.01	0.07	0.07	96
0.5	100	β_1	0.01	0.19	0.17	94	−0.04	0.15	0.17	98	0.11	0.13	0.16	93
		β_2	0.06	0.25	0.23	93	−0.02	0.25	0.22	93	0.12	0.22	0.21	90
		β_3	0.05	0.26	0.23	93	−0.10	0.24	0.22	90	0.03	0.23	0.20	93
	400	β_1	0.01	0.08	0.08	95	−0.06	0.09	0.08	80	0.01	0.08	0.08	96
		β_2	0.01	0.11	0.11	95	−0.03	0.13	0.10	87	0.01	0.11	0.10	93
		β_3	0.01	0.11	0.11	94	−0.04	0.11	0.11	93	0.01	0.10	0.10	94
1	100	β_1	0.01	0.21	0.21	95	−0.01	0.18	0.20	98	0.12	0.16	0.19	92
		β_2	0.04	0.30	0.27	92	0.03	0.28	0.26	95	0.14	0.25	0.25	90
		β_3	0.04	0.29	0.27	94	−0.06	0.27	0.27	94	0.05	0.26	0.24	94
	400	β_1	0.01	0.09	0.10	96	−0.05	0.11	0.10	87	0.01	0.09	0.10	97
		β_2	0.01	0.12	0.13	95	−0.02	0.15	0.12	89	0.01	0.12	0.12	95
		β_3	0.01	0.12	0.13	95	−0.03	0.13	0.13	94	0.00	0.12	0.12	94

[a]MLE, maximum likelihood estimator;

[b]SE, standard error;

[c]SEE, mean of standard error estimator;

[d]CP, coverage probability for nominal 95% confidence interval.

Table 3.4: Estimates of coefficients, their standard errors, and coverage probabilities for nominal 95% confidence intervals from 1000 simulated data sets for censoring ratio 40%

			MLE[a] without variable selection				Adaptive lasso estimator				MLE after variable selection			
r	n	β	Bias	SE[b]	SEE[c]	CP[d]	Bias	SE	SEE	CP	Bias	SE	SEE	CP
0	100	β_1	0.05	0.19	0.16	91	−0.04	0.15	0.16	98	0.11	0.13	0.15	90
		β_2	0.07	0.24	0.21	92	−0.03	0.23	0.20	93	0.11	0.20	0.19	90
		β_3	0.10	0.25	0.22	91	−0.07	0.23	0.21	91	0.06	0.21	0.18	91
	400	β_1	0.01	0.07	0.07	95	−0.08	0.08	0.07	75	0.01	0.06	0.07	96
		β_2	0.03	0.10	0.09	93	−0.03	0.12	0.09	87	0.02	0.09	0.09	93
		β_3	0.03	0.10	0.10	94	−0.04	0.09	0.09	93	0.02	0.09	0.09	95
0.5	100	β_1	0.04	0.21	0.20	94	−0.01	0.17	0.19	99	0.13	0.16	0.18	92
		β_2	0.05	0.29	0.26	93	−0.01	0.26	0.25	97	0.13	0.23	0.24	92
		β_3	0.06	0.29	0.26	93	−0.07	0.27	0.26	93	0.05	0.26	0.23	92
	400	β_1	0.01	0.09	0.09	94	−0.06	0.10	0.09	84	0.01	0.08	0.09	96
		β_2	0.01	0.11	0.12	96	−0.03	0.14	0.12	89	0.01	0.12	0.12	95
		β_3	0.02	0.12	0.12	95	−0.04	0.12	0.12	93	0.01	0.11	0.11	94
1	100	β_1	0.03	0.26	0.23	92	0.04	0.21	0.23	97	0.15	0.20	0.22	90
		β_2	0.03	0.33	0.30	93	0.02	0.29	0.29	97	0.13	0.26	0.28	93
		β_3	0.06	0.34	0.31	92	−0.03	0.30	0.30	95	0.08	0.30	0.27	93
	400	β_1	0.01	0.11	0.11	94	−0.04	0.12	0.11	90	0.03	0.10	0.11	97
		β_2	0.01	0.15	0.14	94	−0.02	0.17	0.14	88	0.02	0.14	0.14	93
		β_3	0.02	0.15	0.14	95	−0.03	0.15	0.14	93	0.01	0.14	0.13	94

[a]MLE, maximum likelihood estimator;

[b]SE, standard error;

[c]SEE, mean of standard error estimator;

[d]CP, coverage probability for nominal 95% confidence interval.

Table 3.5: Variable selection proportions, average numbers of correct and incorrect zero coefficients, and median mean squared errors from 1000 simulated data sets for the misspecified models using the adaptive lasso method

Censoring	n	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}	Corr. [a]	Incorr. [b]	MMSE [c]
20%	100	41	65	84	9	9	7	7	6	6	6	6.50	1.10	0.21
	400	65	92	100	3	2	3	3	3	1	1	6.85	0.43	0.38
40%	100	49	66	85	13	14	13	13	13	11	11	6.12	1.01	0.18
	400	71	90	100	3	4	5	5	4	4	3	6.72	0.39	0.31

[a]Corr., average number of correct zeros;

[b]Incorr., average number of incorrect zeros;

[c]MMSE, median of mean squared errors.

the robust performance of the misspecified model in variable selection is at the cost of prediction accuracy, as measured by the median mean squared errors, which are much larger than those from the models with correct transformation.

3.4 Application

3.4.1 Atherosclerosis Risk in Communities Study Data

We consider data from the Atherosclerosis Risk in Communities Study, a prospective investigation of the etiology of atherosclerosis and its clinical sequelae and variation in cardiovascular risk factors, medical care and disease by race, gender, location, and date (The ARIC Investigators, 1989). The study includes five examinations. The baseline examination of the cohort was conducted from 1987 to 1989, and enrolled 15792 participants with ages 45–64 years from four U.S. communities. In this example we apply our method to part of the baseline data, where participants are African American males living in Jackson, Mississippi or Forsyth County, North Carolina. We study the traditional cardiovascular risk factors for incident heart failure until 2005.

Our analysis consists of 1332 participants after excluding those with missing covariates. Incident heart failure occurred in 196 men through 2005, with a median follow-up time of 16.5 years. The proportional hazards assumption is not satisfied for these data, so the Cox model is not appropriate. We analyze the data using transformation models. To determine the best transformation for fitting the data, we consider logarithmic transformations with $r = 0, 0.1, \ldots, 6$. We apply the expectation-maximization algorithm to estimate the parameters for each r and profile the log-likelihood values in Figure 3.1: r=3 yields the largest log-likelihood. Under this model, we apply our variable selection procedure. The tuning parameter λ is chosen to be 6 via generalized cross validation. After variable selection, we refit the transformation model with selected covariates. The results are given in Table 3.6. Incident heart failure is associated with age, diabetes, hypertension, systolic blood pressure, serum albumin, heart rate, left ventricular hypertrophy, bundle branch block, prevalent coronary heart disease, valvular heart disease, high-density lipoprotein, pack years of smoking and current smoking status.

We assess the prediction capability of the selected risk set via the area under the receiver operating characteristic curve. This statistic is often used for model comparison and extended to accommodate the time-dependence and censoring for the survival outcomes (Chambless and Diao, 2006). Under the logarithmic transformation with r=3, the 10-year area under the curve for selected covariates is 0.85, the same as for all 18 covariates, which indicates that the selected model performs as well as the full model. For further comparison, we consider an external risk score from the Health, Aging, and Body Composition Study (Butler et al., 2008), which was obtained in an elderly population using the Cox model and backward elimination. When directly applying it to these data, the 10-year area under the curve is 0.77, smaller than for our selected model.

To compare the performance of our method under different transformations, we

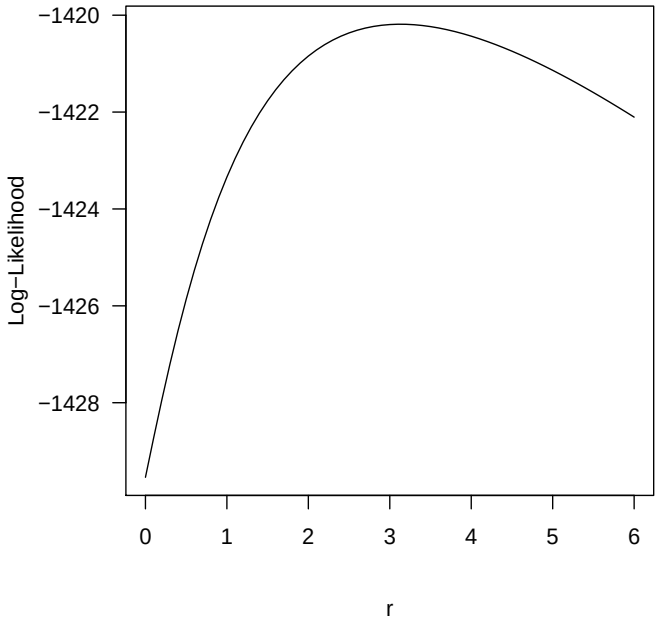

Figure 3.1: Fitted observed log-likelihood values for logarithmic transformation parameter r in the Atherosclerosis Risk in Communities data.

consider another criterion to select the best transformation. For each r, instead of using the log-likelihood with maximum likelihood estimates before variable selection, we conduct variable selection to obtain the adaptive lasso estimates, update the cumulative hazards with these estimates, and compute the log-likelihood minus the adaptive lasso penalty. The maximum of these penalized log-likelihood values corresponds to the transformation r=2.6. Under this model, we select the same set of covariates as in Table 3.6, with slightly different estimates and standard errors; as a result, the 10-year area under the curve for these covariates is also 0.85.

3.4.2 Primary Biliary Cirrhosis Data

As a second example, we apply our method to the primary biliary cirrhosis data, which were collected in the Mayo Clinic trial of primary biliary liver cirrhosis, conducted between 1974 and 1984. For each patient in the trial, clinical, biochemical, serological,

Table 3.6: Estimated coefficients and standard errors for Atherosclerosis Risk in Communities data

Covariate [a]	MLE[b] without variable selection	Adaptive lasso estimator	MLE after variable selection
Age (in years)	0.076 (0.019)	0.070 (0.015)	0.078 (0.019)
Diabetes	1.001 (0.345)	1.102 (0.335)	1.176 (0.244)
Hypertension	0.625 (0.261)	0.566 (0.250)	0.627 (0.257)
BMI (kg/m^2)	0.009 (0.021)	0 (–)	0 (–)
SBP (mm of Hg)	0.014 (0.006)	0.013 (0.006)	0.015 (0.006)
Fasting glucose (mg/dL)	0.002 (0.003)	0 (–)	0 (–)
Serum albumin (g/dL)	−1.563 (0.397)	−1.470 (0.266)	−1.528 (0.385)
Serum creatinine (mg/dl)	0.198 (0.497)	0 (–)	0 (–)
Heart rate (beats/minute)	0.037 (0.010)	0.034 (0.009)	0.037 (0.010)
Left ventricular hypertrophy	0.997 (0.389)	0.850 (0.383)	0.975 (0.385)
Bundle branch block	1.186 (0.406)	1.054 (0.399)	1.202 (0.399)
Prevalent CHD	2.171 (0.444)	2.103 (0.438)	2.185 (0.442)
Valvular heart disease	1.476 (0.585)	1.270 (0.580)	1.502 (0.582)
HDL (mg/dl)	−0.026 (0.008)	−0.023 (0.008)	−0.028 (0.008)
LDL (mg/dl)	0.002 (0.003)	0 (–)	0 (–)
Pack years of smoking	0.013 (0.005)	0.012 (0.005)	0.014 (0.005)
Current smoking status	0.646 (0.320)	0.385 (0.304)	0.492 (0.237)
Former smoking status	0.142 (0.301)	0 (–)	0 (–)

[a]BMI, body mass index; SBP, systolic blood pressure; CHD, coronary heart disease; HDL, high-density lipoprotein; LDL, low-density lipoprotein.

[b]MLE, maximum likelihood estimator;

and histological parameters are collected. A description of the clinical background is provided in Fleming and Harrington (2005, p. 2), and a more extended discussion can be found in Dickson et al. (1989). In this example, we consider 312 out of 424 patients who agreed to participate in the randomized trial. We have 276 patients for analysis after excluding the data with missing covariates, and 111 of them died before the end of trial. The median follow-up time is 4.9 years. We study the dependence of the survival time on all seventeen covariates: trt (0/1 for placebo/D-penicillamine), age (in years), sex (0/1 for male/female), ascites (presence of ascites), hepato (presence of hepatomegaly or enlarged liver), spiders (presence of blood vessel malformations), edema (0/0.5/1 for no edema/untreated or successfully treated/edema despite diuretic therapy), bili (serum bilirubin in mg/dl), chol (serum cholesterol in mg/dl), albumin (serum albumin in g/dl), copper (urine copper in ug/day), alkphos (alkaline phosphotase in U/liter), ast (aspartate aminotransferase in U/ml), trig (triglycerides in mg/dl), platelet (platelet count), protime (standardized blood clotting time), and stage (histologic stage of disease).

We analyze the data following the same procedure as in Section 3.4.1. First, we select the transformation using the observed log-likelihood function with maximum likelihood estimates before variable selection. As illustrated in Figure 3.2, the logarithmic transformation with r=1 maximizes the profile, and it corresponds to the proportional odds model. The results for variable selection and coefficient estimation are given in Table 3.7, with the tuning parameter determined as 7 via generalized cross validation. The covariates selected into the predictive model are: age, sex, ascites, spiders, edema, bili, albumin, copper, ast, protime, and stage. In Table 3.7, we also provide the adaptive lasso estimates using the same tuning parameter based on the penalized marginal likelihood of ranks for the proportional odds model (Lu and Zhang, 2007). Their method tends to give more shrinkage of the coefficients toward zero, and consequently,

the adaptive lasso estimates are smaller than those from our method and have more zeros, including some with large initial effects, such as sex and ascites. Table 3.8 summarizes the results under the transformation model selected from the maximal value of penalized log-likelihood with adaptive lasso estimates. The selected transformation r=0.6 is similar to the proportional odds model using the previous criterion, leading to similar results of variable selection and coefficient estimation. Using the Least Angle Regression algorithm, we give the whole solution paths for the one-step adaptive lasso estimates under transformation r=1 and r=0.6 in Figure 3.3.

In addition, we fit these data using the Cox proportional hazards model ($r = 0$) to compare our method with others. Under this model, the covariates sex, ascites, and spiders have initial estimates much closer to zero and are shrunk to zero by the penalty. We select the same set of covariates as the adaptive lasso using the partial likelihood function (Zhang and Lu, 2007) and the martingale estimating equations (Zhang et al., 2010).

3.5 Remarks

Although we focus on the adaptive lasso penalty, it is rather straightforward to extend our method to other commonly used penalties and show that the asymptotic properties still hold. In practice, the transformation function is unknown and needs to be selected. We used the log-likelihood and penalized log-likelihood for the real example here. Other criteria may also work, and as a result, several transformations may be appropriate for a certain dataset. Interesting future work would be to study the performance of the proposed method when the transformation is misspecified. In one of our simulation studies, we find that the variable selection is still robust, even though we misspecify the proportional odds model. However, in the primary biliary cirrhosis example given in the supplementary material, the variable selection under

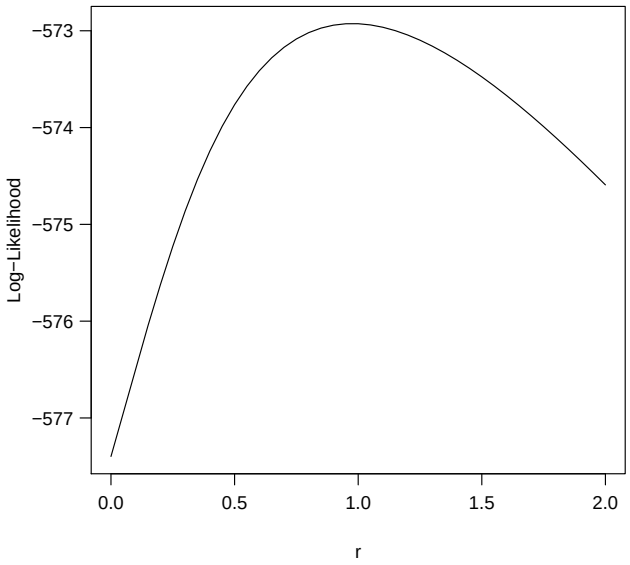

Figure 3.2: Fitted observed log-likelihood values for logarithmic transformation parameter r in the primary biliary cirrhosis data.

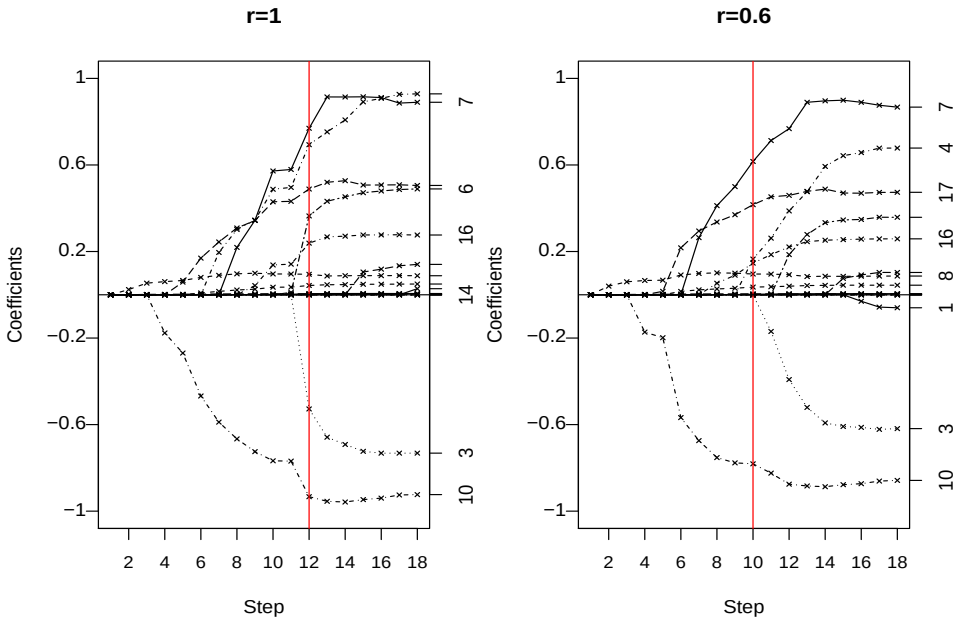

Figure 3.3: Solution path for primary biliary cirrhosis data under selected transformation models

Table 3.7: Estimated coefficients and standard errors for primary biliary cirrhosis data under the proportional odds model

Covariate	MLE[a] without variable selection	Adaptive lasso estimator	MLE after variable selection	Lu and Zhang's adaptive lasso estimator
trt	0.028 (0.295)	0 (–)	0 (–)	0 (–)
age	0.050 (0.016)	0.039 (0.014)	0.046 (0.015)	0.031 (0.013)
sex	−0.732 (0.419)	−0.265 (0.387)	−0.773 (0.408)	0 (–)
ascites	0.929 (0.684)	0.560 (0.666)	0.849 (0.659)	0 (–)
hepato	0.140 (0.338)	0 (–)	0 (–)	0 (–)
spiders	0.489 (0.342)	0.179 (0.328)	0.524 (0.330)	0 (–)
edema	0.890 (0.661)	0.726 (0.635)	0.831 (0.614)	0.724 (0.536)
bili	0.088 (0.037)	0.095 (0.036)	0.097 (0.031)	0.088 (0.029)
chol	0.001 (0.001)	0 (–)	0 (–)	0 (–)
albumin	−0.924 (0.404)	−0.860 (0.302)	−1.008 (0.392)	−0.580 (0.346)
copper	0.005 (0.002)	0.005(0.002)	0.004 (0.002)	0.004 (0.001)
alkphos	0.000 (0.000)	0 (–)	0 (–)	0 (–)
ast	0.006 (0.003)	0.004 (0.002)	0.006 (0.003)	0.002 (0.002)
trig	−0.001 (0.002)	0 (–)	0 (–)	0 (–)
platelet	−0.000 (0.002)	0 (–)	0 (–)	0 (–)
protime	0.276 (0.149)	0.208 (0.123)	0.279 (0.146)	0.185 (0.137)
stage	0.505 (0.222)	0.484 (0.204)	0.530 (0.195)	0.398 (0.172)

[a]MLE, maximum likelihood estimator.

Table 3.8: Estimated coefficients and standard errors for primary biliary cirrhosis data under the transformation model with $r = 0.6$

Covariate	MLE[a] without variable selection	Adaptive lasso estimator	MLE after variable selection
trt	−0.060 (0.266)	0 (–)	0 (–)
age	0.044 (0.014)	0.036 (0.012)	0.040 (0.014)
sex	−0.619 (0.378)	−0.087 (0.354)	−0.546 (0.357)
ascites	0.678 (0.580)	0.151 (0.568)	0.446 (0.542)
hepato	0.103 (0.308)	0 (–)	0 (–)
spiders	0.358 (0.308)	0 (–)	0 (–)
edema	0.867 (0.562)	0.744 (0.543)	0.975 (0.498)
bili	0.087 (0.033)	0.095 (0.031)	0.095 (0.027)
chol	0.001 (0.001)	0 (–)	0 (–)
albumin	−0.858 (0.359)	−0.813 (0.269)	−0.941 (0.349)
copper	0.004 (0.002)	0.004 (0.001)	0.004 (0.001)
alkphos	0.000 (0.000)	0 (–)	0 (–)
ast	0.005 (0.002)	0.003 (0.002)	0.005 (0.002)
trig	−0.001 (0.002)	0 (–)	0 (–)
platelet	−0.000 (0.001)	0 (–)	0 (–)
protime	0.259 (0.133)	0.190 (0.110)	0.280 (0.129)
stage	0.474 (0.204)	0.459 (0.187)	0.538 (0.174)

[a]MLE, maximum likelihood estimator.

these two models differs. This phenomenon and its formal justification need to be further investigated.

3.6 Appendix: Proof of Theorems

We define the counting process $N_i(s) = \Delta_i I(Y_i \leq s)$, where $s \in [0, \tau]$ and τ is the follow-up time. Then $l_n(\beta)$ can be written as

$$l_n(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \beta^T Z_i(s) dN_i(s) - \int_0^{\tau} \log \left\{ \sum_{j=1}^{n} I(Y_j \geq s) \tilde{c}_j e^{\beta^T Z_j(s)} \right\} d \left\{ \sum_{i=1}^{n} N_i(s) \right\},$$

where the weights $\tilde{c} = (\tilde{c}_1, \ldots, \tilde{c}_n)$ based on the maximum likelihood estimators $\tilde{\beta}$ and $\delta\tilde{\Lambda}(Y)$ are

$$\begin{aligned} \tilde{c}_j &= G'\left\{\int_0^\tau I(s \le Y_j) e^{\tilde{\beta}^T Z_j(s)} d\tilde{\Lambda}(s)\right\} - \int_0^\tau \frac{G''\left\{\int_0^s e^{\tilde{\beta}^T Z_j(t)} d\tilde{\Lambda}(t)\right\}}{G'\left\{\int_0^s e^{\tilde{\beta}^T Z_j(t)} d\tilde{\Lambda}(t)\right\}} dN_j(s) \\ &\equiv c_j\left\{Y_j, Z_j(\cdot), \Delta_j, \tilde{\beta}, \tilde{\Lambda}\right\}. \end{aligned}$$

To facilitate the proof of the theorems, we first claim the following lemmas under Conditions 1–4.

Lemma 3.6.1. *Denote the first-order derivative of $l_n(\beta)$ with respect to β as $U_n(\beta)$, then $n^{-1/2}U_n(\beta_0) = O_p(1)$, where $O_p(1)$ is bounded in probability.*

Lemma 3.6.2. *Denote the second-order derivative of $-l_n(\beta)$ with respect to β as $V_n(\beta)$, then $V_n(\beta)/n$ converges uniformly to a positive definite matrix $V(\beta)$, which does not depend on the data, and as a result, $l_n(\beta)$ is a strictly concave function when n is large.*

Proof of Lemma 3.6.1. Denote the true weight c_{0j} in the vector $c_0 = (c_{01}, \ldots, c_{0n})$ as $c_{0j} = c_j\{Y_j, Z_j(\cdot), \Delta_j, \beta_0, \Lambda_0\}$, where β_0 is the true value of β and Λ_0 is the true value of Λ. Let P_n be the empirical measure, with P being the expectation. Then the derivative

$U_n(\beta)$ at $\beta = \beta_0$ can be further written as

$$
\begin{aligned}
& n^{-1/2}U_n(\beta_0) \\
= \; & n^{-1/2}\sum_{i=1}^{n}\int_0^\tau \left\{ Z_i(s) - \frac{\sum_{j=1}^n I(Y_j \geq s)Z_j(s)\tilde{c}_j e^{\beta_0^T Z_j(s)}}{\sum_{j=1}^n I(Y_j \geq s)\tilde{c}_j e^{\beta_0^T Z_j(s)}} \right\} dN_i(s) \\
= \; & n^{1/2}(P_n - P)\int_0^\tau \left[Z(s) - \frac{E\left\{I(Y \geq s)Z(s)c_0 e^{\beta_0^T Z(s)}\right\}}{E\left\{I(Y \geq s)c_0 e^{\beta_0^T Z(s)}\right\}} \right] dN(s) \qquad (3.9) \\
+ \; & n^{1/2}E\left\{ \int_0^\tau \left[Z(s) - \frac{E\left\{I(Y \geq s)Z(s)c_0 e^{\beta_0^T Z(s)}\right\}}{E\left\{I(Y \geq s)c_0 e^{\beta_0^T Z(s)}\right\}} \right] dN(s) \right\} \qquad (3.10) \\
- \; & n^{1/2}\int_0^\tau \left[\frac{E\left\{I(Y \geq s)Z(s)\tilde{c}e^{\beta_0^T Z(s)}\right\}}{E\left\{I(Y \geq s)\tilde{c}e^{\beta_0^T Z(s)}\right\}} - \frac{E\left\{I(Y \geq s)Z(s)c_0 e^{\beta_0^T Z(s)}\right\}}{E\left\{I(Y \geq s)c_0 e^{\beta_0^T Z(s)}\right\}} \right] P_n dN(s) \qquad (3.11) \\
- \; & n^{1/2}\int_0^\tau \left[\frac{P_n\left\{I(Y \geq s)Z(s)\tilde{c}e^{\beta_0^T Z(s)}\right\}}{P_n\left\{I(Y \geq s)\tilde{c}e^{\beta_0^T Z(s)}\right\}} - \frac{E\left\{I(Y \geq s)Z(s)\tilde{c}e^{\beta_0^T Z(s)}\right\}}{E\left\{I(Y \geq s)\tilde{c}e^{\beta_0^T Z(s)}\right\}} \right] P_n dN(s) \qquad (3.12)
\end{aligned}
$$

For (3.10), since the intensity for $N_j(s)$ is $I(Y_j \geq s)e^{\beta_0^T Z_j(s)}\Lambda_0'(s)G'\left\{\int_0^s e^{\beta_0^T Z_j(t)}d\Lambda_0(t)\right\}$, (3.10) is equal to

$$
n^{1/2}E\left[\int_0^\tau \left\{ Z(s) - \frac{E\left[I(Y \geq s)Z(s)e^{\beta_0^T Z(s)}G'\left\{\int_0^s e^{\beta_0^T Z(t)}d\Lambda_0(t)\right\}\right]}{E\left[I(Y \geq s)e^{\beta_0^T Z(s)}G'\left\{\int_0^s e^{\beta_0^T Z(t)}d\Lambda_0(t)\right\}\right]} \right\} dN(s) \right] = 0.
$$

For (3.11), by the mean-value theorem, the integrand of (3.11) is equal to

$$
\begin{aligned}
& -n^{1/2}\left[\nabla_\beta \frac{E\left\{I(Y \geq s)Z(s)c_0 e^{\beta_0^T Z(s)}\right\}}{E\left\{I(Y \geq s)c_0 e^{\beta_0^T Z(s)}\right\}}(\tilde{\beta} - \beta_0) \right. \\
& \qquad \left. + \nabla_\Lambda \frac{E\left\{I(Y \geq s)Z(s)c_0 e^{\beta_0^T Z(s)}\right\}}{E\left\{I(Y \geq s)c_0 e^{\beta_0^T Z(s)}\right\}}(\tilde{\Lambda} - \Lambda_0) + o_p(n^{-1/2}) \right] \\
= \; & -n^{1/2}I(s)(\tilde{\beta} - \beta_0, \tilde{\Lambda} - \Lambda_0) + o_p(1),
\end{aligned}
$$

so (3.11) is equal to $-n^{1/2}(P_n - P)\int_0^\tau I(s)(S_\beta, S_\Lambda)PdN(s) + o_p(1)$, where ∇_β denotes the derivative with respect to β and ∇_Λ denotes the Hadamard derivative with respect

to Λ; I is the linear operator; $o_p(1)$ converges to zero in probability uniformly in s, $s \in [0, \tau]$; S_β and S_Λ are efficient influence functions for β_0 and Λ_0.

For (3.12), using the asymptotic results for $\tilde{\beta}$ and $\tilde{\Lambda}$ from Zeng and Lin (2006), by the mean-value theorem we have

$$\begin{aligned}\sup_{j=1,\ldots,n} |\tilde{c}_j - c_{j0}| \quad \leq \quad & \sup_{j=1,\ldots,n} |\nabla_\beta c_j \{Y_j, Z_j(\cdot), \Delta_j, \beta^*, \Lambda^*\}| \|\tilde{\beta} - \beta_0\| \\ & + \sup_{j=1,\ldots,n} |\nabla_\Lambda c_j \{Y_j, Z_j(\cdot), \Delta_j, \beta^*, \Lambda^*\}| \left\|\tilde{\Lambda} - \Lambda_0\right\|_{l^\infty[0,\tau]} \to 0\end{aligned}$$

almost surely, where $\|\cdot\|_{l^\infty[0,\tau]}$ denotes the supremum norm in $[0, \tau]$; β^* is between $\tilde{\beta}$ and β_0, and Λ^* is between $\tilde{\Lambda}$ and Λ_0 uniformly in t. Based on Theorem 2.10.3 and Theorem 2.10.6 of van der Vaart and Wellner (1996), the weight $c\{Y, Z(\cdot), \Delta, \beta, \Lambda\}$ is a bounded Donsker class. Then by the Glivenko–Cantelli theorem, the integrand of (3.12) is equal to

$$\begin{aligned}-n^{1/2} \Bigg(& \frac{(P_n - P)\left\{I(Y \geq s) Z(s) c_0 e^{\beta_0^T Z(s)}\right\}}{E\left\{I(Y \geq s) c_0 e^{\beta_0^T Z(s)}\right\}} \\ & - \frac{E\left\{I(Y \geq s) Z(s) c_0 e^{\beta_0^T Z(s)}\right\}}{\left[E\left\{I(Y \geq s) c_0 e^{\beta_0^T Z(s)}\right\}\right]^2} (P_n - P)\left\{I(Y \geq s) c_0 e^{\beta_0^T Z(s)}\right\} \Bigg) + o_p(1),\end{aligned}$$

so (3.12) is equal to $-n^{1/2}(P_n - P)\int_0^\tau S_1 P dN(s) + o_p(1)$, where S_1 is the influence function and $o_p(1)$ converges to zero in probability uniformly in s, $s \in [0, \tau]$.

Therefore, the normalized derivative $n^{-1/2} U_n(\beta_0)$ can be written as

$$\begin{aligned}n^{1/2}(P_n - P) \Bigg(& \int_0^\tau \left[Z(s) - \frac{E\left\{I(Y \geq s) Z(s) c_0 e^{\beta_0^T Z(s)}\right\}}{E\left\{I(Y \geq s) c_0 e^{\beta_0^T Z(s)}\right\}} dN(s) \right] \\ & - \int_0^\tau I(s)(S_\beta, S_\Lambda) P dN(s) - \int_0^\tau S_1 P dN(s) \Bigg) + o_p(1).\end{aligned}$$

By the Donsker theorem, $n^{-1/2} U_n(\beta_0) = O_p(1)$. □

Proof of Lemma 3.6.2. We have

$$n^{-1}V_n(\beta) = \int_0^\tau \left(\frac{P_n\left\{I(Y \geq s)\tilde{c}Z(s)^{\otimes 2}e^{\beta^T Z(s)}\right\}}{P_n\left\{I(Y \geq s)\tilde{c}e^{\beta^T Z(s)}\right\}} - \left[\frac{P_n\left\{I(Y \geq s)\tilde{c}Z(s)e^{\beta^T Z(s)}\right\}}{P_n\left\{I(Y \geq s)\tilde{c}e^{\beta^T Z(s)}\right\}}\right]^{\otimes 2}\right) P_n dN(s).$$

By the Glivenko–Cantelli theorem, the integrand converges uniformly to its asymptotic limit

$$\frac{E\left\{I(Y \geq s)c_0 Z(s)^{\otimes 2}e^{\beta^T Z(s)}\right\}}{E\left\{I(Y \geq s)c_0 e^{\beta^T Z(s)}\right\}} - \left[\frac{E\left\{I(Y \geq s)c_0 Z(s)e^{\beta^T Z(s)}\right\}}{E\left\{I(Y \geq s)c_0 e^{\beta^T Z(s)}\right\}}\right]^{\otimes 2}.$$

Define

$$V(\beta) = \int_0^\tau \left(\frac{E\left\{I(Y \geq s)c_0 Z(s)^{\otimes 2}e^{\beta^T Z(s)}\right\}}{E\left\{I(Y \geq s)c_0 e^{\beta^T Z(s)}\right\}} - \left[\frac{E\left\{I(Y \geq s)c_0 Z(s)e^{\beta^T Z(s)}\right\}}{E\left\{I(Y \geq s)c_0 e^{\beta^T Z(s)}\right\}}\right]^{\otimes 2}\right) P dN(s),$$

Then $\sup_\beta |n^{-1}V_n(\beta) - V(\beta)| \to 0$ almost surely.

To show that $V(\beta)$ is positive definite, we insert the intensity for $N(s)$, and then $V(\beta)$ can be written as

$$\int_0^\tau E\left[\left(Z(s) - \frac{E\left[I(Y \geq s)Z(s)e^{\beta^T Z(s)}G'\left\{\int_0^s e^{\beta_0^T Z(t)}d\Lambda_0(t)\right\}\right]}{E\left[I(Y \geq s)e^{\beta^T Z(s)}G'\left\{\int_0^s e^{\beta_0^T Z(t)}d\Lambda_0(t)\right\}\right]}\right)^{\otimes 2} \times I(Y \geq s)e^{\beta^T Z(s)}G'\left\{\int_0^s e^{\beta_0^T Z(t)}d\Lambda_0(t)\right\}\right] \times \frac{E\left[I(Y \geq s)e^{\beta_0^T Z(s)}G'\left\{\int_0^s e^{\beta_0^T Z(t)}d\Lambda_0(t)\right\}\right]}{E\left[I(Y \geq s)e^{\beta^T Z(s)}G'\left\{\int_0^s e^{\beta_0^T Z(t)}d\Lambda_0(t)\right\}\right]} d\Lambda_0(s).$$

Thus, $V(\beta)$ is semi-positive definite. If $V(\beta)$ is not positive definite, there will exist a

vector α, which satisfies $\alpha \neq 0$ and $\alpha^T V(\beta)\alpha = 0$. This indicates that for all $s \in [0, \tau]$,

$$0 = \alpha^T Z(s) - \alpha^T \frac{E\left[I(Y \geq s)Z(s)e^{\beta^T Z(s)}G'\left\{\int_0^s e^{\beta_0^T Z(t)}d\Lambda_0(t)\right\}\right]}{E\left[I(Y \geq s)e^{\beta^T Z(s)}G'\left\{\int_0^s e^{\beta_0^T Z(t)}d\Lambda_0(t)\right\}\right]} = \alpha_0(s) + \alpha^T Z(s),$$

which is a contradiction with Condition 2. Therefore, $V(\beta)$ is positive definite, so $l_n(\beta)$ is strictly concave when n is large. □

Proof of Theorem 3.2.1. Consider the penalized objective function

$$Q_n(\beta) = l_n(\beta) - n\lambda_n \sum_{j=1}^{d} |\beta_j|/|\tilde{\beta}_j|^\gamma.$$

Since the penalty term is strictly convex, it follows from Lemma 3.6.2 that $Q_n(\beta)$ is strictly concave when n is large. Thus, there exists a unique maximiser $\hat{\beta}_n$ of $Q_n(\beta)$ for large n. It is sufficient to show that for any given $\epsilon > 0$, there exists a large constant C so that

$$P\left\{\sup_{\|u\|=C} Q_n(\beta_0 + n^{-1/2}u) < Q_n(\beta_0)\right\} \geq 1 - \epsilon. \tag{3.13}$$

This implies that, with probability at least $1 - \epsilon$, there exists a local maximum in the ball $\{\beta_0 + n^{-1/2} : \|u\| \leq C\}$, $C > 0$.

Furthermore, we have

$$\begin{aligned} D(u) &\equiv n^{-1}\left\{Q_n(\beta_0 + n^{-1/2}u) - Q_n(\beta_0)\right\} \\ &\leq n^{-1}\left\{l_n(\beta_0 + n^{-1/2}u) - l_n(\beta_0)\right\} - \lambda_n \sum_{j=1}^{q}\left(\frac{|\beta_{j0} + n^{-1/2}u_j|}{|\tilde{\beta}_j|^\gamma} - \frac{|\beta_{j0}|}{|\tilde{\beta}_j|^\gamma}\right) \\ &\leq n^{-1}\left\{l_n(\beta_0 + n^{-1/2}u) - l_n(\beta_0)\right\} - n^{-1/2}\lambda_n \sum_{j=1}^{q}\frac{|u_j|}{|\tilde{\beta}_j|^\gamma}. \end{aligned} \tag{3.14}$$

For the first term in (3.14), since for any β^* between $\tilde{\beta}$ and β_0, $\|\beta^* - \beta_0\| \leq \|\tilde{\beta} - \beta_0\| \to 0$ almost surely, it follows from Lemma 3.6.2 that $|n^{-1}V_n(\beta^*) - V(\beta_0)| \leq$

$|n^{-1}V_n(\beta^*) - V(\beta^*)| + |V(\beta^*) - V(\beta_0)| \to 0$ almost surely. That is, $n^{-1}V_n(\beta^*) = V(\beta_0) + o_p(1)$. Then by the Taylor expansion and Lemma 3.6.1, the first term is equal to

$$\begin{aligned} & n^{-1}u^T\{n^{-1/2}U_n(\beta_0)\} - (2n)^{-1}u^T\{n^{-1}V_n(\beta^*)\}u \\ = \; & n^{-1}O_p(1)\sum_{j=1}^{d}|u_j| - (2n)^{-1}u^T\{V(\beta_0) + o_p(1)\}u, \end{aligned}$$

where β^* is between β_0 and $\beta_0 + n^{-1/2}u$.

For the second term in (3.14), since $n^{1/2}\|\tilde{\beta} - \beta_0\| = O_p(1)$ from Zeng and Lin (2006), by the Taylor expansion, the second term is equal to

$$\begin{aligned} & n^{-1/2}\lambda_n \sum_{j=1}^{q}|u_j|\left\{\frac{1}{|\beta_{j0}|^\gamma} - \frac{\gamma \text{sign}(\beta_{j0})}{|\beta_{j0}|^{\gamma+1}}(\tilde{\beta}_j - \beta_{j0}) + o_p(|\tilde{\beta}_j - \beta_{j0}|)\right\} \\ = \; & n^{-1/2}\lambda_n \sum_{j=1}^{q}\left\{\frac{1}{|\beta_{j0}|^\gamma} + \frac{O_p(1)}{\sqrt{n}}\right\}|u_j| \\ = \; & \frac{1}{n}(n^{1/2}\lambda_n)O_p(1)\sum_{j=1}^{q}|u_j|. \end{aligned}$$

Since $n^{1/2}\lambda_n = O_p(1)$, we have

$$D(u) \le -(2n)^{-1}u^T\{V(\beta_0) + o_p(1)\}u + n^{-1}O_p(1)\sum_{j=1}^{d}|u_j| - n^{-1}O_p(1)\sum_{j=1}^{q}|u_j|.$$

By choosing a sufficiently large C, the first term is of the order C^2/n, and the second and third terms are of the order C/n, so the second and third terms are dominated by the first term. Therefore, the inequality (3.13) holds, and it completes the proof. □

Proof of Theorem 3.2.2. (i) For β_j, $j = q+1, \ldots, d$, we have,

$$0 = \nabla_{\beta_j}Q_n(\beta)|_{\beta=\hat{\beta}} = n^{1/2}\left\{\frac{\nabla_{\beta_j}l_n(\beta)|_{\beta=\hat{\beta}}}{n^{1/2}} - n\lambda_n\frac{\text{sign}(\hat{\beta}_j)}{n^{1/2}|\tilde{\beta}_j|^\gamma}\right\} \tag{3.15}$$

By the Taylor expansion, Lemma 3.6.1, and Lemma 3.6.2, (3.15) becomes to

$$\begin{aligned} 0 &= n^{1/2}\left\{n^{-1/2}U_{jn}(\beta_0) + n^{-1}V_{jjn}(\beta^*)n^{1/2}(\hat{\beta}_j - \beta_{j0}) - n^{(\gamma+1)/2}\lambda_n \frac{\text{sign}(\hat{\beta}_j)}{(n^{1/2}|\tilde{\beta}_j|)^\gamma}\right\} \\ &= n^{1/2}\left\{O_p(1) + V_{jj}(\beta_0)n^{1/2}(\hat{\beta}_j - \beta_{j0}) - n^{(\gamma+1)/2}\lambda_n \frac{\text{sign}(\hat{\beta}_j)}{(n^{1/2}|\tilde{\beta}_j|)^\gamma}\right\} \end{aligned}$$

where $U_{jn}(\beta_0)$ is the jth element of $U_n(\beta_0)$, $V_{jjn}(\beta^*)$ is the (j,j)th element of $V_n(\beta^*)$, and $V_{jj}(\beta_0)$ is the (j,j)th element of $V(\beta_0)$. Since $n^{1/2}(\tilde{\beta}_j - 0) = O_p(1)$ and $n^{1/2}(\hat{\beta}_j - \beta_{j0}) = O_p(1)$, we have

$$n^{1/2}\left\{O_p(1) - n^{(\gamma+1)/2}\lambda_n \text{sign}(\hat{\beta}_j)\right\} = 0.$$

Then $n^{(\gamma+1)/2}\lambda_n \to \infty$ implies that $\hat{\beta}_j = 0$ with probability tending to 1.

(ii) Let β_1 denote the β index for β_{10}. According to (i), $\text{pr}(\hat{\beta}_{2n} = 0) \to 1$; thus, we only need to derive the asymptotic expansion of $\hat{\beta}_{1n}$ in the probability set $\{\hat{\beta}_{2n} = 0\}$. For any probability sample in the latter set, $\nabla_{\beta_1} Q_n(\beta)|_{\beta=\{\hat{\beta}_{1n}^T, 0^T\}^T} = 0$. Let $U_{1n}(\beta)$ be the first q elements of $U_n(\beta)$ and $V_{11n}(\beta)$ be the first $q \times q$ submatrix of $V_n(\beta)$. Then

$$\begin{aligned} 0 &= \nabla_{\beta_1} Q_n(\beta)|_{\beta=\{\hat{\beta}_{1n}^T, 0^T\}^T} = \nabla_{\beta_1} l_n(\beta)|_{\beta=\{\hat{\beta}_{1n}^T, 0^T\}^T} - n\lambda_n \left\{\frac{\text{sign}(\hat{\beta}_1)}{|\tilde{\beta}_1|^\gamma}, \ldots, \frac{\text{sign}(\hat{\beta}_q)}{|\tilde{\beta}_q|^\gamma}\right\}^T \\ &= U_{1n}(\beta_0) - V_{11n}(\beta^*)(\hat{\beta}_{1n} - \beta_{10}) - n\lambda_n \left\{\frac{\text{sign}(\hat{\beta}_1)}{|\tilde{\beta}_1|^\gamma}, \ldots, \frac{\text{sign}(\hat{\beta}_q)}{|\tilde{\beta}_q|^\gamma}\right\}^T, \end{aligned}$$

where β^* is between $\hat{\beta}_n$ and β_0, and the last equation is implied by the Taylor expansion. Following the proof of Lemma 3.2, we can show that $V_{11n}(\beta^*)/n \to V_{11}(\beta_0)$, where $V_{11}(\beta_0)$ is the first $q \times q$ submatrix of $V(\beta_0)$. Since $n^{1/2}\lambda_n \to 0$, $|\tilde{\beta} - \beta_0| \xrightarrow{as} 0$, and $\text{sign}(\hat{\beta}_j) =$

$\text{sign}(\beta_{j0})$ for large n, $j = 1, \ldots, q$, we have

$$\begin{aligned} & n^{1/2}(\hat{\beta}_{1n} - \beta_{10}) \\ = \; & \{n^{-1}V_{11n}(\beta^*)\}^{-1}\left[n^{-1/2}U_{1n}(\beta_0) - n^{1/2}\lambda_n\left\{\frac{\text{sign}(\hat{\beta}_1)}{|\tilde{\beta}_1|^\gamma}, \ldots, \frac{\text{sign}(\hat{\beta}_q)}{|\tilde{\beta}_q|^\gamma}\right\}^T\right] \\ = \; & \{V_{11}(\beta_0)\}^{-1}\{n^{-1/2}U_{1n}(\beta_0)\} + o_p(1). \end{aligned}$$

On the other hand, from Lemma 3.1, we know that the influence function of $n^{-1/2}U_{1n}(\beta_0)$ can be expressed as the $n^{1/2}(P_n - P)E\{\nabla_{\beta_1} l_c \mid Y, Z(\cdot), \Delta\}$ plus a linear functional of $(\tilde{\beta} - \beta_0)$ and $(\tilde{\Lambda} - \Lambda_0)$, where $\nabla_{\beta_1} l_c$ is the score for β_1 using the complete log-likelihood, where ζ is missing data, and $(\tilde{\beta}, \tilde{\Lambda})$ are the initial nonparametric maximum likelihood estimators. According to Zeng and Lin (2006), the influence functions of $(\tilde{\beta} - \beta_0)$ and $(\tilde{\Lambda} - \Lambda_0)$ lie on the tangent space spanned by the scores. Moreover, $E\{\nabla_{\beta_1} l_c(\beta) \mid Y, Z(\cdot), \Delta\}$ is clearly on the same tangent space. Therefore, the influence function of $\hat{\beta}_{1n}$ is also on this space, so it must be the efficient influence function which is unique. In other words, $n^{1/2}(\hat{\beta}_{1n} - \beta_{10}) = n^{1/2}(P_n - P)S_{\beta_1}\{Y, \Delta, Z_1(\cdot), \beta_{10}, \Lambda_0\} + o_p(1)$, where S_{β_1} is the efficient influence function for the maximum likelihood estimator $\hat{\beta}$ corresponding to β_1 as given in Zeng and Lin (2006). Particularly, $\text{var}(S_{\beta_1})$ attains the semiparametric efficiency bound. It completes the proof. □

CHAPTER4: SUPPORT VECTOR HAZARD REGRESSION FOR PREDICTING SURVIVAL OUTCOMES

4.1 Support Vector Hazard Regression

4.1.1 General Methodology

Let T denote the failure time and X denote a vector of baseline covariates of d-dimension. Due to patient's drop-out or termination of the study, T is subject to right-censoring. Therefore, from a random sample of n subjects, the observed data consist of $\{T_i \wedge C_i, \Delta_i = I(T_i \leq C_i), X_i\}$ for $i = 1, \ldots, n$. Furthermore, we define the observed counting process as $N_i(t) = I(T_i \wedge C_i \leq t)$ and the observed at-risk process as $Y_i(t) = I(T_i \wedge C_i \geq t)$.

Since predicting T is equivalent to predicting its associated counting process, which can be treated as a sequence of binary outcomes (failure vs. no failure, or event vs. no event) over time, this motivates us to reformulate predicting the failure time as predicting the jumps of the counting process over a sequence of time points among the subjects still at risk at those time points. In other words, we will develop 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; equivalently, we wish to learn the hazard rate functions for the counting process of T. Similar to classical hazard regression models in survival analysis, the main advantages of learning through hazard rate functions are: first, we can use all the available information from both failure cases and censored cases; second, we allow censoring time C to depend on X but do not require modeling the distribution of C given X.

To formalize idea, we consider a general decision function $f(t,x)$ at time t for a subject with $X = x$. In other words, if this subject is still at risk at time t, we predict the subject to fail at the next immediate time if $f(t,x) > 0$ or not fail otherwise. Empirically, suppose that there are m distinct ordered failure 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 $f(t,x)$ becomes a sequence of classification problems over t_j's. Ideally, the best decision function should minimize the total classification errors, the sum of $I(\delta N_i(t_j) f(t_j, X_i) < 0)$ over all subjects i and time t_j when subject i is still at risk at t_j, i.e., $Y_i(t_j) = 1$. However, in practice, we most likely observe that only one subject experiences failure at t_j while the rest of subjects who are still at risk do not. To account for this imbalance between the failures and non-failures at each time t_j, we need to give more weights to the failure cases while less for the non-failure cases. Specially, at each t_j and for subject i at risk at t_j, we apply the following weight related to the size of risk set

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

In other words, if subject i has a failure event at t_j, we assign a weight close to 1; otherwise, we assign a weight equal to the reciprocal of the risk set size. By doing this, an optimal decision function thus minimizes the following weighted total classification error:

$$R_{0n}(f) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} w_i(t_j) Y_i(t_j) I(\delta N_i(t_j) f(t_j, X_i) < 0), \tag{4.1}$$

where the use of $Y_i(t_j)$ terms reflects that only subjects still at risk contribute towards prediction.

Minimizing (4.1) is infeasible due to the non-smoothness of the 0-1 loss in the expression $I(\delta N_i(t_j)f(t_j, X_i) > 0)$. Furthermore, no restriction on the complexity of f leads to potential overfitting. To handle these issues, we adopt the same idea in support vector machines for supervised learning where we replace the 0-1 loss in (4.1) by the hinge loss and place regularization to estimate f. Specifically, we propose to minimize the following regularized hinge loss:

$$R_n(f) + \lambda_n \|f\|^2,$$

$$\text{where } R_n(f) \equiv n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} w_i(t_j)[1 - f(t_j, X_i)\delta N_i(t_j)]_+, \tag{4.2}$$

where $[1-x]_+ = \max(1-x, 0)$ is the hinge loss, $\|f\|$ 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 failure and non-failure 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 $f(t, x)$ is essentially learning the hazard rate function, we refer our proposed method as "support vector hazard regression" (SVHR).

4.1.2 Additive Learning Rules

The functional form of $f(t, x)$ in (4.2) is fully nonparametric to ensure flexibility. However, prediction rules based on this general time-varying rule may not be practically useful. Instead, a desirable prediction rule would be based on a single risk score from subject's baseline covariates, X, without appealing to a complex and time-varying function $f(t, x)$. Particularly, such a decision function, $f(t, x)$, can take the following additive form

$$f(t, x) = \alpha(t) + g(x), \tag{4.3}$$

where both $\alpha(\cdot)$ and $g(\cdot)$ are assumed to be unknown. One major advantage of this additive structure is that only a single score $g(x)$ is required to perform prediction for subjects with $X = x$. For example, if $g(x) = x^T\beta$, then a prediction score is simply a linear combination of baseline covariates, and the coefficients, β, can be used to rank the importance of each covariate. Thus, in the following development, we focus on the decision function with the additive structure as in (4.3).

Next, we describe the computational algorithm to solve the minimization in (4.2). We do not impose any restriction on $\alpha(t)$, and assume $g(x)$ lies in a reproducing kernel Hilbert space H_n with a kernel function $K(x, x')$. Commonly used kernels include linear kernel, where $K(x, x') = x^T x'$; radial basis kernel, where $K(x, x') = \exp(-\|x - x'\|^2/\sigma)$; and lth-degree polynomial kernel, where $K(x, x') = (1 + \langle x, x' \rangle)^l$. Furthermore, we let $\|f\| = \|g\|_{H_n}$, which is the norm in the reproducing kernel Hilbert space H_n. Thus, the minimization in (4.2),

$$\min n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} w_i(t_j) Y_i(t_j) [1 - (\alpha(t_j) + g(X_i)) \delta N_i(t_j)]_+ + \lambda_n \|g\|_{H_n}, \tag{4.4}$$

is equivalent to

$$\min_{\alpha, g} \|g\|^2 + C_n \sum_{i=1}^{N} \sum_{j=1}^{m} w_i(t_j) Y_i(t_j) \zeta_i(t_j)$$

$$\text{subject to } \; Y_i(t_j)\zeta_i(t_j) \geq 0, \; Y_i(t_j)\delta N_i(t_j)\{\alpha(t_j) + g(X_i)\} \geq Y_i(t_j)\{1 - \zeta_i(t_j)\},$$

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.

From the KKT conditions, we can easily derive the dual objective function for (4.4)

as

$$L_D = \sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}Y_i(t_j) - \frac{1}{2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{j=1}^{m}\sum_{j'=1}^{m}\gamma_{ij}\gamma_{i'j'}Y_i(t_j)Y_{i'}(t_{j'})\delta N_i(t_j)\delta N_{i'}(t_{j'})K(X_i, X_{i'}). \tag{4.5}$$

We maximize L_D subject to $0 \le \gamma_{ij} \le w_i(t_j)C_n$ 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 the quadratic programming packages available in many software. The tuning parameter C_n is chosen using the cross-validation searching over a grid of values. Comparing the proposed algorithm (4.5) with existing standard support vector machine algorithms, we see that the objective function sums across all at-risk subjects and across time points for which they are at risk. Constraints are placed on those subjects and the time points. Therefore, SVHR acts as a time-varying support vector machine.

After computing $\widehat{\gamma}_{ij}$, from (4.5), we obtain the predicted score for a feature subject with baseline covariate x as

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

To obtain the predicted event time, we use a two-step approach. We first adopt the nearest-neighbor prediction: for a future subject with $X = x$, we find the non-censored subject in the training data whose predictive score is closest to $\widehat{g}(x)$, denoted as $\widehat{g}(x_j)$. Next, to maintain the monotone relationship between the event times and predictive scores, we sort the derived scores of non-censored subjects in the training data in descending order and find the rank of $\widehat{g}(x_j)$. Then we sort the event times of these derived scores in the training data in ascending order and find the event time with the same rank as the rank of $\widehat{g}(x_j)$, denoted as $T_{j'}$. We predict the future subject's event time to be $T_{j'}$.

4.1.3 Profile Empirical Risk

The function $\alpha(t)$ in (4.3) is analogous to the baseline hazard rate function in the proportional hazards model, which is treated as a nuisance parameter, and thus often profiled out for inference. Therefore, it will be similarly interesting to profile out $\alpha(t)$ in the minimization problem (4.4).

To this end, for a fixed $g(x)$, from the derivation similar to Hastie et al. (2009, p.421) and Abe (2010, p.77), 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

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

Otherwise, $\widehat{\alpha}(t_j)$ can be any value satisfying

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

For the latter case, taking $\widehat{\alpha}(t_j) = 1 - g(X_i)$ where $\delta N_i(t_j) = 1$ satisfying these constraints. After substituting $\widehat{\alpha}(t_j)$ in this form into (4.4), we obtain the following profile empirical risk for $g(\cdot)$:

$$\begin{aligned} PR_n(g) &= \frac{1}{n}\sum_{i=1}^{n}\int \frac{\sum_{k=1}^{n} Y_k(t)[2 - g(X_i) + g(X_k)]_+}{\sum_{k=1}^{n} Y_k(t)} dN_i(t) - \frac{2}{n}\sum_{i=1}^{n}\int \frac{dN_i(t)}{\sum_{k=1}^{n} Y_k(t)} \\ &= \frac{1}{n}\sum_{i=1}^{n}\Delta_i \frac{\sum_{k=1}^{n} I(Y_k \geq Y_i)[2 - g(X_i) + g(X_k)]_+}{\sum_{k=1}^{n} I(Y_k \geq Y_i)} - \frac{2}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\sum_{k=1}^{n} I(Y_k \geq Y_i)} \\ &= P_n\left(\Delta \frac{\widetilde{P}_n\{I(\widetilde{Y} \geq Y)[2 + g(\widetilde{X}) - g(X)]_+\}}{\widetilde{P}_n[I(\widetilde{Y} \geq Y)]}\right) - \frac{2}{n}P_n\left\{\frac{\Delta}{\widetilde{P}_n[I(\widetilde{Y} \geq Y)]}\right\}, \end{aligned}$$

where P_n denotes the empirical measure from n observations and $\widetilde{P}_n$ is the empirical measure applied to $(\widetilde{Y}, \widetilde{X}, \widetilde{\Delta})$. Thus, $\widehat{g}(x)$ minimizes $PR_n(g) + \lambda_n \|g\|^2_{H_n}$. If we let

$\widehat{f}(x,t) = \widehat{\alpha}(t) + \widehat{g}(x)$ be the function minimizing (4.4) over $g \in H_n$, then $R_n(\widehat{f}) = PR_n(\widehat{g})$. It is worthy to point out one interesting observation: $PR_n(g)$ takes a similar form as the partial likelihood function in survival analysis under a different loss function. This connection sheds lights on the optimality of SVHR which we prove in the next section.

4.2 Theoretical Properties

4.2.1 Risk Function and Optimal Decision Rule

In this section, we will derive the population risk function for the proposed SVHR. We will then drive the optimal decision rule for this risk function and show that this decision rule also optimizes the 0-1 loss corresponding to (4.1).

To this end, we first examine the population version of $R_n(f)$. By the definition, we can rewrite $R_n(f)$ as

$$\begin{aligned} R_n(f) &= \frac{1}{n}\sum_{i=1}^{n}\int [1 - f(t,X_i)]_+ dN_i(t) + \frac{1}{n}\int \frac{\sum_{i=1}^{n} Y_i(t)[1+f(t,X_i)]_+}{\sum_{i=1}^{n} Y_i(t)} d\left\{\sum_{i=1}^{n} N_i(t)\right\} \\ &\quad -\frac{1}{n}\sum_{i=1}^{n}\int \frac{1}{\sum_{i=1}^{n} Y_i(t)} \left([1 - f(t,X_i)]_+ + [1 + f(t,X_i)]_+\right) dN_i(t). \end{aligned}$$

Therefore, as n goes to infinity, note that the last term in $R_n(f)$ vanishes, so we obtain the asymptotic limit of $R_n(f)$, denoted as $R(f)$, to be

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

Similarly, when n goes to infinity, the empirical risk based on the prediction error in (4.1) converges to

$$R_0(f) = E\left(\int I\left[f(t,X) \le 0\right] dN(t)\right) + \int \frac{E\left(Y(t) I\left[f(t,X) \ge 0\right]\right)}{E\{Y(t)\}} E\{dN(t)\}.$$

Let $f^*(t,x)$ be the optimal function minimizing $R(f)$, the limit of the risk function in the SVHR method. The following theorem gives the property of $f^*(t,x)$.

Theorem 4.2.1. *Let $h(t,x)$ denote the conditional hazard rate function of $T = t$ given $X = x$ and let $\bar{h}(t) = E[dN(t)/dt]/E[Y(t)] = E[h(t,X)|Y(t) = 1]$ be the average hazard rate at time t. Then $f^*(t,x) = sign(h(t,x) - \bar{h}(t))$ minimizes $R(f)$. Furthermore, $f^*(t,x)$ also minimizes $R_0(f)$ and*

$$R_0(f^*) = P(T \le C) - \frac{1}{2}E\left[\int E\{Y(t)|X = x\}|h(t,x) - \bar{h}(t)|dt\right].$$

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

$$R_0(f) - R_0(f^*) \le R(f) - R(f^*)$$

for some constant c.

The proof of Theorem 4.1 is in the appendix. From Theorem 4.1, we see the best rule is essentially to predict whether an at-risk subject will have an event still by comparing the subject-specific hazard rate to the population-average hazard rate obtained from all the at-risk subjects. Since the minimizer of $R(f)$ also minimizes $R_0(f)$, this justifies the use of the hinge-loss in the SVHR method in order to minimize the weighted prediction error in $R_0(f)$. The last inequality in Theorem 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.

4.2.2 Asymptotic Properties of the Additive Learning Rules

In this section, we will study the asymptotic properties of the SVHR when the decision function takes the additive form in (4.3). We denote 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 $R(f)$, we consider the profile risk for $R(f)$ defined as

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

Then since for $f(t,x) = \alpha(t) + g(x)$,

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

it is easy to see that

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

minimizes $R(f)$. Therefore,

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

Clearly, $PR(g)$ is the asymptotic limit of $PR_n(g)$. Then the following theorem holds for the risk $PR(\hat{g})$.

Theorem 4.2.2. *Assume that X's support is compact and $E[Y(\tau)|X]$ 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*

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

The proof of Theorem 4.2 (see appendix) 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. We state two useful observations as remarks below.

Remark 4.1. From Theorem 4.2, if we choose $\sigma_n = (n\lambda_n)^{-1/[2d(1/p+1/4)]}$, it gives

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

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

Remark 4.2. If we choose $\lambda_n = n^{-q/(q+1)}$, then the optimal rate from Theorem 4.2 becomes

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

4.3 Simulation Studies

4.3.1 Simulation Setup

In this section, we illustrate the finite sample performance of the proposed method in various settings. In all scenarios, we generated both failure times and censoring times to be dependent on the covariates. First we simulated five covariates $X = (X_1, \ldots, X_5)$ which are marginally normal $N(0, 0.5^2)$ with pairwise correlation $\text{corr}(X_j, X_k) = \rho^{|j-k|}$, and $\rho = 0.5$. The failure times were generated from the Cox model with true $beta = (2, -1.6, 1.2, -0.8, 0.4)^T$ and the exponential distribution $0.25t$ was assumed for the baseline cumulative hazard function $\Lambda(t)$. 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, i.e., $\ln N(X^T\beta_c + a, 0.5^2)$, with true $\beta_c = (1, 1, 1, 1, 1)^T$. In the second type, the distribution of the censoring times follows the Cox model with true $\beta_c = (1, 1, 1, -2, -2)^T$ 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 failure times or censored times greater than u_0 were truncated at u_0, where u_0 is the 90th percentile of the failure 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, we use a randomly generated testing data set of size 10000 in each scenario including only the failure times to evaluate prediction performance. We experiment two sample sizes, 100 and 200.

For all scenarios, we compared the proposed SVHR 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 (IPCW) (Goldberg and Kosorok, 2013). We used linear kernel $K(x, x') = x^T x'$ in all three 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 model comparison criterion, we adapted mean squared error for censored data, which only sums up the mean squared differences between the fitted event times and observed event times if uncensored, and between fitted times and censoring times if censored and the predicted values are smaller than the observed values. The mean squared differences are assumed to be zero if censored and the predicted values are greater than the observed values. We divided the total sum of squares by the total number of observations. We repeated the simulation 500 times.

4.3.2 Simulation Results

Table 4.1 and 4.2 give the average Pearson correlations and root mean square errors $\{\sum(\widehat{T} - T)^2\}^{1/2}$ based on the fitted failure times and observed failure times T in the testing data set. Larger correlation and smaller root mean square error indicate better performance. SVHR outperforms the other two methods for all the simulation cases and sample sizes. The advantages are not affected by including 5 noise variables,

and the improvements become more evident when the censoring ratio is 60% or the censoring distribution follows the accelerated failure time model. The columns of the average correlations show that the modified SVR has the similar capability to capture the rank information as SVHR. However, it gives less accurate prediction of the exact failure times measured by the higher RMSEs. The IPCW methods have the worst performances among all the methods, no matter using the Kaplan-Meier estimator or fitting a 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 our method has the largest improvement with respect to the ratios of the average root mean squared errors. We also explored training the data with a Gaussian kernel for the sample size 100 and the computation is more intensive. The resulting average correlations and root mean square errors are similar to those in Table 4.1 and 4.2, and therefore not shown.

4.4 Application

4.4.1 Huntington's Disease Study Data

We apply our method to the data collected from a neurological disease (Huntington's disease, HD) study (Paulsen et al., 2008). HD is a severe dominant genetic disorder for which at risk subjects can be identified through a genetic testing of C-A-G expansion status at the ITI5 gene (Huntington's Study Investigators 1993). The availability of genetic testing and virtually complete penetrance of gene provides opportunity for early intervention. In the data we analyze here, pre-manifest HD subjects in the absence of experimental treatment were recruited (Paulsen et al., 2008). The goal of the study is to identify and combine salient clinical markers and biological markers sensitive enough to detect early indicators of gradual changes of patient disease progression before evident clinical signs of HD emerge. In this example, we have 705 subjects for

Table 4.1: Comparison of three support vector learning methods for right censored data using a linear kernel, with censoring times following the accelerated failure time model.

	# of		n = 100			n = 200		
Censoring	Noises	Method	Corr. [a]	RMSE [b]	Ratio [c]	Corr.	RMSE	Ratio
40%	0	Modified SVR	0.59	5.59 (0.60)	1.19	0.62	5.58 (0.58)	1.24
		IPCW-KM [d]	0.40	5.60 (0.52)	1.20	0.45	5.45 (0.41)	1.21
		IPCW-Cox	0.43	5.80 (0.64)	1.24	0.50	5.62 (0.57)	1.25
		SVHR	0.61	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
		SVHR	0.58	4.90 (0.35)	1.00	0.63	4.62 (0.20)	1.00
	95[e]	Modified SVR	0.21	6.65 (0.89)	1.10	0.30	6.32 (0.52)	1.10
		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
		SVHR	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
		SVHR	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
		SVHR	0.51	5.40 (0.48)	1.00	0.58	5.02 (0.33)	1.00
	95	Modified SVR	0.17	6.90 (1.08)	1.11	0.25	7.12 (1.42)	1.20
		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
		SVHR	0.17	6.22 (0.24)	1.00	0.26	5.94 (0.25)	1.00

[a]Corr., average number of correlations.

[b]RMSE, average number of root mean square errors.

[c]Ratio, ratio of average root mean square errors between the method used and our method.

[d]IPCW-KM, IPCW using the Kaplan-Meier estimator for the censoring distribution; IPCW-Cox, IPCW using the Cox model for the censoring distribution.

[e]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 4.2: Comparison of three support vector learning methods for right censored data using a linear kernel, with censoring times following the Cox proportional hazards model

	# of		n = 100			n = 200		
Censoring	Noises	Method	Corr. [a]	RMSE [b]	Ratio [c]	Corr.	RMSE	Ratio
40%	0	Modified SVR	0.59	5.15 (0.59)	1.11	0.62	5.09 (0.54)	1.12
		IPCW-KM [d]	0.53	5.16 (0.42)	1.11	0.55	5.08 (0.31)	1.12
		IPCW-Cox	0.52	5.31 (0.57)	1.14	0.56	5.09 (0.46)	1.12
		SVHR	0.61	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
		SVHR	0.58	4.89 (0.29)	1.00	0.62	4.65 (0.18)	1.00
	95[e]	Modified SVR	0.21	6.43 (0.92)	1.04	0.33	6.06 (0.59)	1.05
		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
		SVHR	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
		SVHR	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
		SVHR	0.53	5.23 (0.37)	1.00	0.59	4.96 (0.27)	1.00
	95	Modified SVR	0.18	6.47 (0.87)	1.05	0.26	6.36 (0.90)	1.06
		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
		SVHR	0.20	6.14 (0.38)	1.00	0.28	6.00 (0.35)	1.00

[a]Corr., average number of correlations.

[b]RMSE, average number of root mean square errors.

[c]Ratio, ratio of average root mean square errors between the method used and our method.

[d]IPCW-KM, IPCW using the Kaplan-Meier estimator for the censoring distribution; IPCW-Cox, IPCW using the Cox model for the censoring distribution.

[e]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.

analysis after excluding the data with missing covariates, and 126 of them developed HD during the 10-year course of study. For each subject, a wide range of measures on motor, psychiatric and cognitive signs are collected. The covariates cover important clinical and functional domains of HD including CAP score (a combination of age and C-A-G repeats length, Zhang et al, (2011)), symbol digital modality test, STROOP color, word and interference tests, total functional capacity scores, UHDRS total motor scores, various SCL-90 psychiatric scores and demographic variables such as gender and education in years.

We study the prediction capability of the above fifteen baseline markers predicting the age-at-onset of HD during the study period. We also evaluate the usefulness of the combined score in performing risk stratification. We apply the proposed SVHR, modified SVR, and IPCW to analyze the data and compare their performances. The covariates are normalized to the same scale for numeric stability. The predicted values of 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 written as $K(x, x') = \exp(-\gamma\|x - x'\|^2)$, the parameter γ is fixed to be 0.005. To compare the prediction capability, we computed several quantities using the predicted values of onset ages and the original values of onset ages at the disease diagnosis or at the censoring. Specifically, we report the concordance index defined as the percentage of correctly ordered pairs among all feasible pairs (C-index). In addition, to evaluate the ability of the fitted scores on performing risk stratification, we separated the data into two groups using various percentiles of the combined predictive scores as cut points. We report the Chi-square statistics from the logrank test and the hazard ratios comparing the hazard of developing HD between two groups from fitting a univariate Cox model based on percentile splitting.

The results are given in Table 4.3. The proposed SVHR significantly improves the

other methods with respect to all the quantities for both linear kernel and Gaussian kernel, and the performances are similar using different kernels. The logrank Chi-square statistics and hazard ratio of SVHR is much larger than all competing methods in all quantiles. In addition, the logrank statistics and concordance index indicate that the predictions of IPCW cannot capture the trend of the original onset ages. Figure 4.1 complements the results in the table by plotting the hazard ratios comparing two groups separated using a series of percentiles of the predicted values as cut points, and SVHR consistently has the largest hazard ratio across all percentiles among all methods. The improvement of SVHR increases at the higher percentiles indicating it is particularly effective in discriminating high risk subjects. This result is consistent with our theoretical results which reveal that SVHR is optimal in separating the individual covariate-specific hazard function, $h(t, x)$ given x, from the population average hazard function, $\bar{h}_n(t)$.

We show the fitted coefficients from SVHR of the markers in Table 4.4 and compare with fits from a Cox proportional hazards model. The top ranking markers with largest standardized effects from both model include 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 (Paulsen et al., 2008). SVHR suggests that the baseline total motor score appears to be slightly more predictive than CAP score in terms of predicting future HD diagnosis during the trial. The neuropsychological markers (Stroop color, Stroop word, SDMT) are predictive but not Stroop interference. The coefficients from Cox model however, suggest that SDMT is not important, which may not be consistent with the clinical literature (Paulsen, 2011). Lastly, SVHR 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 less informative for HD diagnosis due to reasons such as subjects may seek treatment. In contrast, Cox model

Table 4.3: Comparison of prediction capability for different methods using Huntington's disease data

			25th percentile		50th percentile		75th percentile	
Kernel	Method	C-index	Logrank χ^2 [a]	HR [b]	Logrank χ^2	HR	Logrank χ^2	HR
Linear	Modified SVR	0.71	42.50	3.08	27.53	2.98	13.41	3.82
	IPCW-KM	0.44	0.06	1.06	0.23	1.09	1.39	1.26
	IPCW-Cox	0.54	5.47	1.57	5.66	1.53	0.90	1.22
	SVHR	0.75	81.79	4.68	35.12	4.74	16.53	7.76
Gaussian	Modified SVR	0.72	46.53	3.27	30.24	3.34	14.33	3.67
	IPCW-KM	0.44	0.32	1.16	0.86	1.19	1.50	1.27
	IPCW-Cox	0.53	5.42	1.57	3.89	1.42	1.97	1.34
	SVHR	0.75	78.66	4.63	36.78	4.68	17.46	8.11

[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.

yields high weights for these markers.

4.4.2 Atherosclerosis Risk in Communities Study Data

As a second example, we consider 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). The study includes four examinations. The baseline examination of the cohort was conducted from 1987 to 1989, and enrolled 15792 participants of ages 45–64 from four U.S. communities. In this example, we apply our method to part of the baseline data, where participants are African-American males with hypertension living in Jackson, Mississippi. 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,

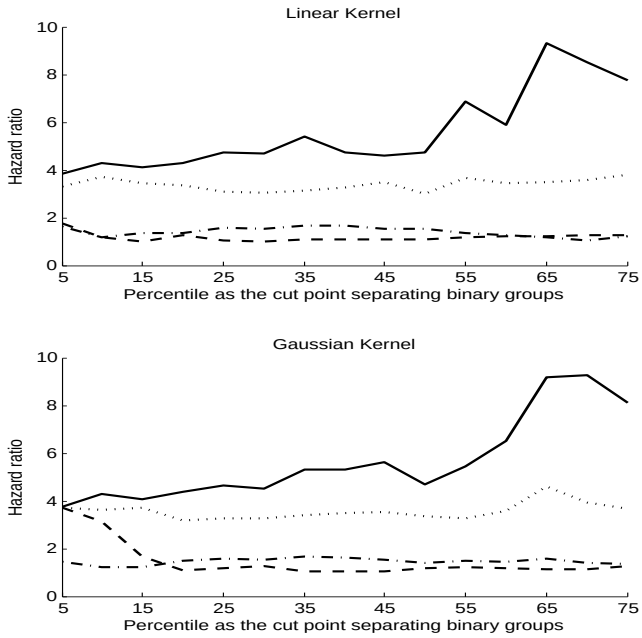

Figure 4.1: Hazard Ratios comparing two groups separated using percentiles of predicted values as cut points for Huntington's disease data. Dotted curve: Modified SVR; Dashed curve: IPCW-KM; Dashed-dotted curve: IPCW-Cox; Black solid curve: SVHR.

Table 4.4: Normalized coefficient estimates using linear kernel for Huntington's disease data

Marker	Normalized β	Cox model [a]
Total Motor Score	0.680	0.354 *
CAP	0.440	0.334 *
Stroop Color	-0.235	-0.247
Stroop Word	-0.208	-0.107
SDMT	-0.151	-0.076
Stroop Interference	0.034	0.271
FRSBE Total	0.246	0.242
UHDRS Psychiatric	0.197	0.270
SCL90 Depression	-0.062	-0.306
SCL90 GSI	-0.004	0.114
SCL90 PST	-0.081	-0.217
SCL90 PSDI	0.096	0.061
TFC	-0.054	-0.047
Education	-0.025	-0.092
Male Gender	-0.315	-0.392 *

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

heart 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 consists of 624 participants, after excluding those with missing risk factors. 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. We analyze the data following the same procedure as in Section 4.4.1. The results for prediction capability of different methods are given in Table 4.5. SVHR provides more accurate prediction than other methods using the linear 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 in most cases. In Table 4.6, we can see that all the risk factors have positive effects on the incident heart failure except HDL, serum albumin and

Table 4.5: Comparison of prediction capability for different methods using Atherosclerosis Risk in Communities data

			25th percentile		50th percentile		75th percentile	
Kernel	Method	C-index	Logrank χ^2 [a]	HR[b]	Logrank χ^2	HR	Logrank χ^2	HR
Linear	SURSVMR	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
	Our method	0.76	95.09	4.78	67.06	4.63	34.93	5.36
Gaussian	SURSVMR	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
	Our method	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.

former smoking status. We also present estimated coefficients from a Cox proportional hazards model as comparison in Table 4.6. Most coefficients are comparable in terms of size. However, note that higher fasting glucose level appears to be protective of heart failure using Cox model, which is the opposite of the expected direction. Contrary, fasting glucose has a positive sign using SVHR, which is consistent with the clinical literature.

4.5 Remarks

In this chapter, we propose a novel framework for predicting the event times using right-censored data by support vector hazards regression. 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

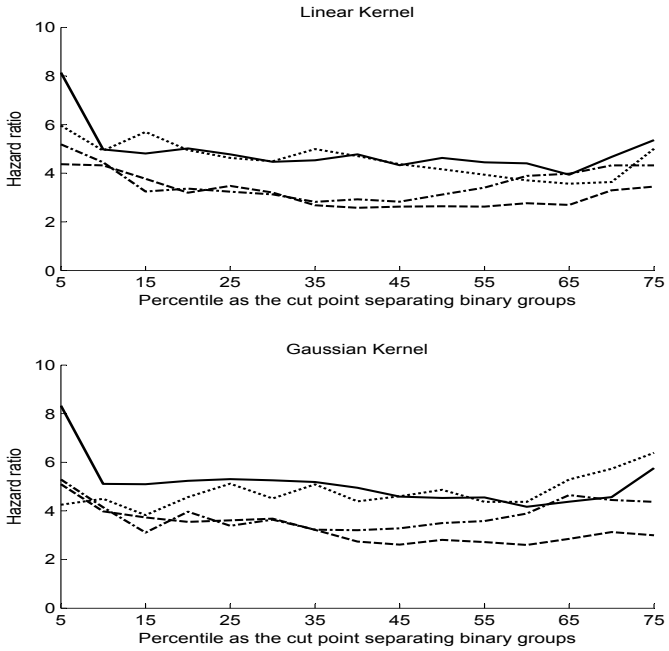

Figure 4.2: Hazard Ratios comparing two groups separated using percentiles of predicted values as cut points for Atherosclerosis Risk in Communities data. Dotted curve: Modified SVR; Dashed curve: IPCW-KM; Dashed-dotted curve: IPCW-Cox; Black solid curve: SVHR.

Table 4.6: Normalized coefficient estimates using linear kernel for Atherosclerosis Risk in Communities data

Covariate [a]	Normalized β	Cox model [b]
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]BMI, body mass index; SBP, systolic blood pressure; CHD, coronary heart disease; HDL, high-density lipoprotein; LDL, low-density lipoprotein.

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

function. The simulation studies and real data applications demonstrate satisfactory results in finite samples with much improved overall accuracy and stable prediction 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 success 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.

In practice, one potential challenge is the large number of parameters to be optimized and the fast growing dimensions of the quadratic programming optimization as the sample size increases. The latter part is a typical problem encountered by the standard support vector machines and the sequential minimal optimization algorithm (Platt, 1999) was developed to tackle the issue. However, this algorithm cannot be easily adapted to our method due to the time-specific intercepts $\alpha(t)$. To improve computational efficiency, one possible solution is to round the event times into some small number of distinct values. When predicting the event times, the only assumption we depend on is their monotone relationship with the fitted one-dimensional risk scores obtained from the learning algorithm. Although the nearest-neighbor method is adopted here and provides promising results, other methods based on this assumption such as linear regression or monotone kernel regression may also be reasonable choices. Interpolation may be needed when there are only few distinct survival times in the training data.

In the current framework, the time-specific prediction rules $f(t, X)$ being considered include only a class of additive rules. From the perspective of survival analysis, it may be generalized to be fully nonparametric. As a result, one would be able to predict the whole counting process instead of only the survival times, and the time-varying

covariates can be handled in an automatic way. 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.

4.6 Appendix: Proof of Theorems

Here we sketch the proofs of Theorem 4.1 and 4.2.

Proof of Theorem 4.1. Since $f^*(t,x)$ minimizes $R(f)$, conditional $X = x$, $f^*(t,x)$ also minimizes

$$E\left(\int [1 - f(t,X)]_+ dN(t) | X = x\right) + \int \frac{E\left(Y(t)[1 + f(t,X)]_+ | X = x\right)}{E\{Y(t)\}} E\{dN(t)\}. \quad (4.6)$$

Clearly, the value $f^*(t,x)$ should belong to the interval $[-1,1]$, because otherwise truncation of f at -1 or 1 gives a lower value. Assuming $-1 \le f(t,x) \le 1$, (4.6) becomes

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

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

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

Therefore, one optimal decision function minimizing $R_L(f)$ is

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

On other hand, we note

$$\begin{aligned} R_0(f) &= \int I[f(t,x) \le 0] E\left(Y(t)|X=x\right) h(t,x) dt \\ &+ \int I[f(t,x) \ge 0] E\left(Y(t)|X=x\right) \bar{h}(t) dt. \end{aligned}$$

Thus, any decision function has the same sign as $(h(t,x) - \bar{h}(t))$ minimizes $R_0(f)$ so $f^*(t,x)$ minimizes $R_0(f)$. Finally, under the optimal rule $f^*(t,x)$, the minimal value of the weighted 0-1 risk is given as

$$\begin{aligned} R_0(f^*) &= E\Big[\int E\{Y(t)|X=x\} \min\{h(t,x), \bar{h}(t)\} dt\Big] \\ &= \frac{1}{2} E\Big[\int E\{Y(t)|X=x\}\{h(t,x) + \bar{h}(t) - |h(t,x) - \bar{h}(t)|\} dt\Big] \\ &= P(T \le C) - \frac{1}{2} E\Big[\int E\{Y(t)|X=x\} |h(t,x) - \bar{h}(t)| dt\Big]. \end{aligned}$$

To show the last inequality in Theorem 4.1, we note hat for $-1 \le f(t,x) \le 1$,

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

and

$$\begin{aligned} R(f^*) &= 2P(T \le C) \\ &- E\Big[\int \text{sign}\{h(t,x) - \bar{h}_n(t)\} E\{Y(t)|X=x\}\{h(t,x) - \bar{h}(t)\} dt\Big]. \end{aligned}$$

Thus,

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

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

$$\begin{aligned} & R_0(f) - R_0(f^*) \\ = \; & E\left[\int E\{Y(t)|X = x\}\left(I[f(t,x) \le 0]h(t,x)\right)dt\right] \\ + \; & E\left[\int E\{Y(t)|X = x\}\left(I[f(t,x) \ge 0]\bar{h}(t) - \min\{h(t,x), \bar{h}(t)\}\right)dt\right] \\ = \; & E\left[\int E\{Y(t)|X = x\}\left|h(t,x) - \bar{h}(t)\right| \times I\left(\{h(t,x) - \bar{h}(t)\}\text{sign}\{f(t,x)\} < 0\right)dt\right]. \end{aligned}$$

Note that

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

We then obtain $R_0(f) - R_0(f^*) \le R(f) - R(f^*)$. □

Proof of Theorem 4.2. The proof Theorem 4.2 follows a similar procedure to the standard support vector machine theory. However, the main difference is that the proof handles $PR_n(f)$ instead of the simple empirical mean of the hinge-loss in the standard theory. Let g_{λ_n} be the function in H_n which minimizes $\lambda_n\|g\|^2_{H_n} + PR(g)$. The proof consists of the following steps.

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

$$\lambda_n\|g_{\lambda_n}\|^2_{H_n} + PR(g_{\lambda_n}) \le PR(0).$$

This gives $\|g_{\lambda_n}\|_{H_n} \leq \sqrt{c/\lambda}$ 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

$$\lambda_n \|\widehat{g}\|_{H_n}^2 + PR_n(\widehat{g}) \leq \lambda_n \|g_{\lambda_n}\|_{H_n}^2 + PR_n(g_{\lambda_n}),$$

we conclude $\|\widehat{g}\|_{H_n} \leq \sqrt{c/\lambda_n}$ 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 (4.2) to be in $\sqrt{c/\lambda_n} B_{H_n}$, where B_{H_n} be the unit ball in 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:

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

From Step 1, we conclude

$$\lambda_n \|\widehat{g}\|_H^2 + PR(\widehat{g}) - (\lambda_n \|g_{\lambda_n}\| + PR(g_{\lambda_n})) \leq 2 \sup_{\|g\|_{H_n} \leq \sqrt{c/\lambda_n}} |PR_n(g) - PR(g)|. \tag{4.7}$$

We derive a bound for the right-hand side of (4.7). First,

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

where

$$\begin{aligned} f_g(Y,X,\Delta) &= \Delta\frac{\widetilde{P}_n\{I(\widetilde{Y}\geq Y)[2+g(\widetilde{X})-g(X)]_+\}}{\widetilde{P}_n[I(\widetilde{Y}\geq Y)]} \\ &+ \widetilde{P}\left(\widetilde{\Delta}\frac{I(Y\geq\widetilde{Y})[2+g(X)-g(\widetilde{X})]_+}{\widetilde{P}_n[I(\widetilde{Y}\geq Y)]}\right) \\ &- \widetilde{P}\left(\widetilde{\Delta}\frac{I(Y\geq\widetilde{Y})P^*\{I(Y^*\geq\widetilde{Y})[2+g(X^*)-g(\widetilde{X})]_+\}}{P_n^*[I(Y^*\geq\widetilde{Y})]P^*[I(Y^*\geq\widetilde{Y})]}\right). \end{aligned}$$

Therefore,

$$\sup_{\|g\|_{H_n}\leq\sqrt{c/\lambda_n}}|PR_n(g)-PR(g)|\leq\sup_{\|g\|_{H_n}\leq\sqrt{c/\lambda_n}}|(P_n-P)f_g|+c/n.$$

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

$$\log N(\epsilon,\sqrt{c/\lambda_n}B_{H_n},l_\infty)\leq c_{p,d}\sigma_n^{(p/4-1)d}\left(\frac{\epsilon}{\sqrt{c/\lambda_n}}\right)^{-p}\leq c_{p,d}\sigma_n^{(p/4-1)d}\lambda_n^{-p/2}\epsilon^{-p},$$

where $N(\epsilon,F,l_\infty)$ is the ϵ-covering number of F under l_∞-norm, d is the dimension of 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

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

This implies

$$\log N(\epsilon,\{f_g/a_n:g\in\sqrt{c/\lambda_n}B_{H_n}\},l_\infty)\leq c_{p,d}\sigma_n^{(p/4-1)d}\epsilon^{-p},$$

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

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

for some constant c only depending on (p,d). Consequently, (4.7) gives

$$P(\lambda_n\|\widehat{g}\|_H^2+PR(\widehat{g})-(\lambda_n\|g_{\lambda_n}\|+PR(g_{\lambda_n}))>cn^{-1}+a_nn^{-1/2}x)\le e^{-cx^2}.$$

Hence, we have proved

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

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

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

Thus, if we define

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

then $\widetilde{g}\in H_n$ and

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

Therefore,

$$\inf_{g\in H_n}\{\lambda_n\|g\|_{H_n}+PR(g)\}\le\{\lambda_n\|\widetilde{g}\|_{H_n}+PR(\widetilde{g})\}\le PR(g^*)+c\sigma_n^{d/2}+c\lambda_n.$$

The result in Theorem 4.2 holds.

□

CHAPTER5: SUPPORT VECTOR MACHINES FOR PREDICTING RECURRENT EVENTS

5.1 Methodology

5.1.1 Generalization of Support Vector Machines

For a random sample of n subjects, let T_{ik} be the kth event time and C_i the censoring time for the ith subject. Let X_i denote the corresponding vector of baseline covariates and $Z_i(\cdot)$ the corresponding vector of time-varying covariates. Here we only consider $Z_i(\cdot)$ that depends on the prior recurrence history of the ith subject and changes at event times T_{ik}. Thus the observed data at kth recurrence consist of $\{T_{ik} \wedge C_i, I(T_{ik} \leq C_i), X_i, Z_i(T_{ik} \wedge C_i)\}$ for $i = 1, \ldots, n$. We first focus on using a linear score of X and $Z(\cdot)$ to predict the recurrent events. Define the observed counting process as $N_i(t) = \sum_k I(T_{ik} \wedge C_i \leq t)$ and define the observed at-risk process as $Y_i(t) = I(C_i \geq t)$. Assume there are d distinct ordered event times over all the observed recurrences, $t_1 < t_2 < \ldots, < t_d$ with $d = \sum_k \sum_{i=1}^{n} I(T_{ik} \leq C_i)$. At each time point t_j $(j = 1, \ldots, d)$ and for all the subjects still at risk, we identify a linear risk score

$$f(t_j, X_i, Z_i(\cdot)) = \alpha(t_j) + X_i^T \beta + Z_i^T(t_j)\gamma$$

to classify the time-varying binary outcome $\delta N_i(t_j) \equiv N_i(t_j) - N_i(t_j-)$ with maximal separation between the subjects who experience the event and those who do not. The time-varying intercept $\alpha(t_j)$ allows the classification boundary to vary with time, and is also used to identify multiple records of the same subjects at different event times.

Redefine $\delta N_i(t_j) \equiv 2[N_i(t_j) - N_i(t_j-)] - 1$ to be consistent with standard support vector machine. We maximize the margin M between subjects in the event and no-event classes subject to the constraints on the misclassification rate. This is, we solve the optimization problem

$$\begin{aligned}
&\max_{\alpha(t_j),\beta,\gamma,\|\beta,\gamma\|=1} M, \\
&\text{subject to } \; Y_i(t_j)\delta N_i(t_j)\{\alpha(t_j) + X_i^T\beta + Z_i^T(t_j)\gamma\} \geq Y_i(t_j)\{1-\zeta_i(t_j)\}, \\
&Y_i(t_j)\zeta_i(t_j) \geq 0, \quad \sum_{i=1}^{n}\sum_{j=1}^{d} w_i(t_j)Y_i(t_j)\zeta_i(t_j) \leq \tau_n, \; i = 1,\ldots,n, \; j = 1,\ldots,d,
\end{aligned}$$

where the value $\zeta_i(t_j)$ is the proportional amount by which the prediction $f(t_j, X_i, Z_i(\cdot))$ is on the wrong side of its margin, τ_n is a pre-specified constant, and

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

This is a nice convex optimization problem, and the prediction rules can be easily calculated by the quadratic programming algorithms. The weights $w_i(t_j)$s give large weights to events and small weights to non-events to adjust for the one vs. many problem at each event time.

To derive the dual form of the above maximization problem, note that it is equivalent to

$$\begin{aligned}
&\min_{\alpha(t_j),\beta,\gamma} \frac{1}{2}\|\beta\|^2 + \frac{1}{2}\|\gamma\|^2 + C_n\sum_{i=1}^{N}\sum_{j=1}^{d} w_i(t_j)Y_i(t_j)\zeta_i(t_j) \\
\text{subject to} \quad &Y_i(t_j)\zeta_i(t_j) \geq 0, \\
&Y_i(t_j)\delta N_i(t_j)\{\alpha(t_j) + X_i^T\beta + Z_i(t_j)^T\gamma\} \geq Y_i(t_j)\{1-\zeta_i(t_j)\},
\end{aligned}$$

where C_n is the cost parameter. We can further convert the above problem to its dual

form by using the corresponding Lagrangian function

$$\begin{aligned} L_p &= \frac{1}{2}\|\beta\|^2 + \frac{1}{2}\|\gamma\|^2 + C_n \sum_{i=1}^{n}\sum_{j=1}^{d} w_i(t_j)Y_i(t_j)\zeta_i(t_j) - \sum_{i=1}^{n}\sum_{j=1}^{d}\mu_{ij}Y_i(t_j)\zeta_i(t_j) \\ &- \sum_{i=1}^{n}\sum_{j=1}^{d}\eta_{ij}[Y_i(t_j)\delta N_i(t_j)\{\alpha(t_j) + X_i^T\beta + Z_i(t_j)^T\gamma\} - Y_i(t_j)\{1-\zeta_i(t_j)\}]. \end{aligned}$$

We minimize L_P with respect to β, γ, $\alpha(t_j)$, and $\zeta_i(t_j)$. Setting the respective derivatives to zero, we obtain

$$\beta = \sum_{i=1}^{n}\sum_{j=1}^{d}\eta_{ij}Y_i(t_j)\delta N_i(t_j)X_i^T, \tag{5.1}$$

$$\gamma = \sum_{i=1}^{n}\sum_{j=1}^{d}\eta_{ij}Y_i(t_j)\delta N_i(t_j)Z_i(t_j)^T, \tag{5.2}$$

$$\sum_{i=1}^{n}\eta_{ij}Y_i(t_j)\delta N_i(t_j) = 0, \tag{5.3}$$

$$C_n w_i(t_j)Y_i(t_j) - \mu_{ij}Y_i(t_j) = \eta_{ij}Y_i(t_j),\ i = 1,\ldots,n,\ j = 1,\ldots,d, \tag{5.4}$$

as well as the positivity constraints η_{ij}, μ_{ij}, $\zeta_i(t_j) \geq 0\ \forall i,j$. By substituting these back to L_p, the dual objective function is

$$\begin{aligned} L_D &= \sum_{i=1}^{n}\sum_{j=1}^{d}\eta_{ij}Y_i(t_j) \\ &- \frac{1}{2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{j=1}^{d}\sum_{j'=1}^{d}\eta_{ij}\eta_{i'j'}Y_i(t_j)Y_{i'}(t_{j'})\delta N_i(t_j)\delta N_{i'}(t_{j'})[X_i^T X_{i'} + Z_i(t_j)^T Z_{i'}(t_{j'})]. \end{aligned}$$

We maximize L_D subject to $0 \leq \eta_{ij} \leq w_i(t_j)C_n$ and $\sum_{i=1}^{n}\eta_{ij}Y_i(t_j)\delta N_i(t_j) = 0$ for $i = 1,\ldots,n$ and $j = 1,\ldots,d$. The tuning parameter C_n is chosen using the cross-validation searching over a grid of values. The Karush-Kuhn-Tucker (KKT) condition includes the constraints

$$\eta_{ij}[Y_i(t_j)\delta N_i(t_j)\{\alpha(t_j) + X_i^T\beta + Z_i(t_j)^T\gamma\} - Y_i(t_j)\{1-\zeta_i(t_j)\}] = 0, \tag{5.5}$$

$$\mu_{ij}\zeta_i(t_j) = 0, \tag{5.6}$$

$$Y_i(t_j)\delta N_i(t_j)\{\alpha(t_j) + X_i^T\beta + Z_i(t_j)^T\gamma\} - Y_i(t_j)\{1 - \zeta_i(t_j)\} \geq 0. \tag{5.7}$$

In the solution of the problem, those points for which $\hat{\eta}_{ij} > 0$ are support vectors, which determine $\hat{\beta}$ and $\hat{\gamma}$ using (5.1) and (5.2). At each t_j, $\hat{\alpha}(t_j)$ can be solved by using the constraints (5.3)-(5.6). Specifically, if there are some support vectors lying on the edge of the margin which are characterized by $0 < \hat{\eta}_{ij} < w_i(t_j)C_n$, $\hat{\alpha}(t_j) = 1/\delta N_i(t_j) - X_i^T\hat{\beta} - Z_i(t_j)^T\hat{\gamma}$ for these points, and we average of all the solutions for numerical stability. Otherwise, if all the support vectors at t_j are $\hat{\eta}_{ij} = C_n w_i(t_j)$, $\hat{\alpha}(t_j)$ is not unique and falls into a range

$$\min_{\substack{\hat{\eta}_{ij}=C_n w_i(t_j),\\ \delta N_i(t_j)=1}} \{1 - X_i^T\hat{\beta} - Z_i(t_j)^T\hat{\gamma}\} \geq \hat{\alpha}(t_j) \geq \max_{\substack{\hat{\eta}_{ij}=C_n w_i(t_j),\\ \delta N_i(t_j)=-1}} \{-1 - X_i^T\hat{\beta} - Z_i(t_j)^T\hat{\gamma}\}.$$

5.1.2 Prediction of Recurrent Events

In this section we use the learned information to predict the times of recurrent events for new subjects. We make use of the similarity between our proposed method and standard multicategory support vector machines from the prospective of supervised learning. In other words, the classification based on counting process results in d ordered categories with labels t_j $(j = 1, \ldots, d)$, i.e. $t_1 < \ldots < t_d$, and all the categories share the same linear risk score $X^T\beta + Z(\cdot)^T\gamma$. Thus, we are able to adapt the Max Wins algorithm for multicategory prediction in Friedman (1996).

Denote the kth event time of a new subject as $\tilde{T}_k$, and given $\tilde{T}_k$, we want to predict the time to the next recurrence $\tilde{T}_{k,k+1}$ using the subject's baseline covariates $\tilde{X}$ and time-varying covariates $\tilde{Z}(\tilde{T}_k)$. A two-step method is used. In the first step, we find out the set of $\hat{\alpha}(t)$s that are available to be used for the prediction conditional on $\tilde{T}_k$. In the training data set, we use only the subjects who have event times greater than

$\tilde{T}_k$, and use their smallest event times greater than $\tilde{T}_k$ and the corresponding $\hat{\alpha}(t)$s. For example, if we want to predict the first recurrence of a new subject, we use the first event times of all the subjects in the training data set; and if we want to predict the second recurrence of a new subject given his/her observed first recurrence $\tilde{T}_1$, we use only the event times greater than $\tilde{T}_1$ in the training data set, and we may have a smaller set of $\hat{\alpha}(t)$s to be used compared to the prediction of the first recurrence.

In the second step, we adapt the Max Wins algorithm based on the selected set of $\hat{\alpha}(t)$s. Suppose that we have d' ordered elements in the set, i.e., $\hat{\alpha}(t_1), \hat{\alpha}(t_2), \ldots, \hat{\alpha}(t_{d'})$ with $t_1 < t_2 < \ldots < t_{d'}$. We assign a score to each t using the signs of $\hat{f}(t, \tilde{X}, \tilde{Z}(\tilde{T}_k)) = \hat{\alpha}(t) + \tilde{X}\hat{\beta} + \tilde{Z}(\tilde{T}_k)\hat{\gamma}$. Specifically, if $\hat{f}(t_j, \tilde{X}, \tilde{Z}(\tilde{T}_k)) > 0$, the scores of event times in the selected set less than or equal to t_j add 1; otherwise, the scores of event times in the selected set larger than t_j add 1. At the end, we find the event time t_m in the selected set with the largest score, and predict the new subject's $(k+1)$th event time from kth recurrence $T_{k,k+1}$ to be $t_m - \tilde{T}_k$.

The method described so far adopts only the linear score $f(t, X, Z(\cdot))$. As an advantage of support vector based methods, we can make the procedure more flexible by considering a non-linear relationship $g_1(X)$ and $g_2(Z(\cdot))$ instead of $X^T\beta$ and $Z^T(\cdot)\gamma$. This is a straightforward extension because of the expression of the training data in the form of inner products in the dual objective function L_D. The inner products $X_i^T X_{i'}$ and $Z_i(t_j)^T Z_{i'}(t_{j'})$ can be replaced by kernel functions $K(X_i, X_{i'}) = \langle g_1(X_i)^T, g_1(X_{i'})\rangle$ and $K(Z_i(t_j), Z_{i'}(t_{j'})) = \langle g_2(Z_i(t_j))^T, g_2(Z_{i'}(t_{j'})\rangle$ to map data into a richer feature space. The transformation g_1 and g_2 do not need to be specified explicitly, and only the knowledge of the kernel function is required. Commonly used kernels are: linear kernel, $K(a, a') = a^T a'$; radial basis kernel, $K(a, a') = \exp(-\|a - a'\|^2/\sigma^2)$; and dth-degree polynomial kernel, $K(a, a') = (1 + \langle a, a'\rangle)^d$. In the non-linear situation, the

decision function $f(t, \tilde{X}, \tilde{Z}(\cdot))$ becomes to

$$\hat{\alpha}(t) + \sum_{i=1}^{n}\sum_{j=1}^{d} \hat{\eta}_{ij}\delta N_i(t_j)K(\tilde{X}, X_i) + \sum_{i=1}^{n}\sum_{j=1}^{d} \hat{\eta}_{ij}\delta N_i(t_j)K(\tilde{Z}(\cdot), Z_i(t_j)),$$

where $\tilde{X}$ and $\tilde{Z}(\cdot)$ are baseline and time-varying covariates of the new subject. Then we can use $f(t, \tilde{X}, \tilde{Z}(\cdot))$ and follow the same steps to predict the times of recurrence events.

5.2 Theoretical Properties

In this section we derive the optimal decision rule and Bayesian risk for the proposed method. By simple algebraic calculations, the optimization problem in Section 5.1 can be written as a regularization method,

$$\min \lambda_n(\|g_1\|^2_{H_{1n}} + \|g_2\|^2_{H_{2n}}) + \sum_{i=1}^{n}\sum_{j=1}^{d} Y_i(t_j)w_i(t_j)[1 - f(t_j, X_i, Z_i(\cdot))\delta N_i(t_j)]_+,$$

where the subsript '$+$' indicates the positive part of a function, and $\lambda_n = 1/2C_n$. In this formulation, the empirical risk is

$$\begin{aligned}
R_n(f) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} Y_i(t_j)w_i(t_j)[1 - f(t_j, X_i, Z_i(\cdot))\delta N_i(t_j)]_+ \\
&= \frac{1}{n}\sum_{i=1}^{n}\int [1 - f(t, X_i, Z_i(\cdot))]_+ dN_i(t) + \frac{1}{n}\int \frac{\sum_{i=1}^{n} Y_i(t)[1 + f(t, X_i, Z_i(\cdot))]_+}{\sum_{i=1}^{n} Y_i(t)} d\left\{\sum_{i=1}^{n} N_i(t)\right\} \\
&-\frac{1}{n}\sum_{i=1}^{n}\int \frac{1}{\sum_{i=1}^{n} Y_i(t)} \left([1 - f(t, X_i, Z_i(\cdot))]_+ + [1 + f(t, X_i, Z_i(\cdot))]_+\right) dN_i(t).
\end{aligned}$$

We refer the loss function of $R_n(f)$ as the integrated hinge loss. As n goes to infinity, the last term in the above equation vanishes, and we obtain the asymptotic limit of

$R_n(f)$, denoted as $R(f)$,

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

On the other hand, based on the similar rationale of standard support vector machines, we consider the integrated hinge loss as a convex surrogate loss function for the non-convex integrated 0-1 loss to make the optimization problem computationally feasible. As in Chapter 4, we define the empirical risk of the integrated 0-1 loss as

$$R_{n,0}(f) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} Y_i(t_j) w_i(t_j) I\left[f(t_j, X_i, Z_i(\cdot)) \delta N_i(t_j) \leq 0\right],$$

with the asymptotic limit

$$R_0(f) = E\left(\int I\left[f(t, X, Z(\cdot)) \leq 0\right] dN(t)\right) + \int \frac{E\left(Y(t) I\left[f(t, X, Z(\cdot)) \geq 0\right]\right)}{E\{Y(t)\}} E\{dN(t)\}.$$

To derive the optimal decision rule, we need to find $f^*(t, x, z(.))$ that minimizes the asymptotic limit $R(f)$. By plugging $f^*(t, x, z(.))$ into $R_0(f)$, we can obtain the Bayesian risk of the proposed method. The derivation takes the similar steps as in Chapter 4, and the following theorem gives the results.

Theorem 5.2.1. *Let $\lambda(t, x, z(\cdot))$ denote the conditional intensity function of $T = t$ given $X = x$ and $Z(\cdot) = z(\cdot)$. Let $\bar{\lambda}(t) = E[dN(t)/dt]/E[Y(t)] = E[\lambda(t, X, Z(\cdot))|Y(t) = 1]$ be the average intensity rate at time t. Then $f^*(t, x, z(\cdot)) = sign(\lambda(t, x, z(\cdot)) - \bar{\lambda}(t))$ minimizes $R(f)$. Furthermore, $f^*(t, x, z(\cdot))$ also minimizes $R_0(f)$ and*

$$R_0(f^*) = P(T \leq C) - \frac{1}{2} E\left[\int E\{Y(t)|X = x, Z(\cdot) = z(\cdot)\}|\lambda(t, x, z(\cdot)) - \bar{\lambda}(t)|dt\right].$$

Theorem 5.1 indicates that the prediction rule is optimal in comparing the subject-specific intensity to the average intensity rate for all the subjects still at risk when predicting the recurrence for a certain subject. In addition, it reveals the nature of our method from the perspective of survival analysis, which is relying on the intensity function instead of the cumulative intensity function as in traditional semiparametric survival models. As a result, this phenomenon has intuitively explained the reason why we use only one event time greater than $\tilde{T}_k$ per subject instead of all event times greater than $\tilde{T}_k$ in the training data set to predict $T_{k,k+1}$ in Section 5.1.2.

5.3 Simulation Studies

5.3.1 Simulation Setup

Simulations are conducted to illustrate the finite sample performance of the proposed method. For each subject we consider three recurrences, $k = 1, 2, 3$. We take five baseline covariates $X = (X_1, \ldots, X_5)$ which are marginally normal with a mean of 0, variance 0.25 and pairwise correlation $\text{corr}(Z_j, Z_k) = 0.5^{|j-k|}$. We use one time-varying covariate $Z(\cdot)$ indicating the time of prior recurrence, i.e., $Z(\cdot) = \log(T_{k-1})$ for the kth recurrence and $Z(\cdot) = 0$ for the first recurrence. The data are generated using a linear risk score $g(X, Z(.), v) = X^T\beta_0 + Z(\cdot)^T\gamma_0 + v$, where $\beta_0 = (2, -1.6, 1.2, -0.8, 0.4)^T$, $\gamma_0 = 1$, and v is a subject-specific frailty that is normally distributed with a mean of 0 and variance σ^2. We examine three different values of σ^2, 0, 1, and 2. We generate the gap times to three recurrences T_1, T_{12}, and T_{23} from three Cox models $\Lambda_{01}(t)\exp(g(X, 0, v))$, $\Lambda_{02}(t)\exp(g(X, \log(T_1), v))$, and $\Lambda_{03}(t)\exp(g(X, \log(T_1 + T_{12}), v))$, where the baseline cumulative hazards $\Lambda_{01}(t)$, $\Lambda_{02}(t)$, and $\Lambda_{03}(t)$ follow three Weibull distributions. Thus, the gap times to the second (third) recurrence depend on the times of the first (second) recurrence. Then the total times of the three recurrences are T_1, $T_2 = T_1 + T_{12}$, and $T_3 = T_2 + T_{23}$. The censoring times are also generated from a Cox model $\Lambda_{0c}(t)\exp(X^T\beta_c)$,

where $\beta_c = (1,1,1,1,1)^T$ and the baseline cumulative hazard $\Lambda_{0c}(t) = a_c t$ (a_c is a constant determining the percentages of events for each recurrence). For each subject there are three event times and one censoring time, and the observed times are the minimum of event times and the censoring time. We consider four cases:

(i) baseline cumulative hazards for three gap times are the same where $\Lambda_{01}(t) = \Lambda_{02}(t) = \Lambda_{03}(t) = 0.25t$, and censoring constant $a_c = 0.1$ which leads to about 60%, 46%, and 42% subjects with at least one, two and three recurrences;

(ii) baseline cumulative hazards for three gap times are the same where $\Lambda_{01}(t) = \Lambda_{02}(t) = \Lambda_{03}(t) = 0.25t$, and censoring constant $a_c = 0.5$ which leads to about 40%, 21%, and 18% subjects with at least one, two and three recurrences;

(iii) baseline cumulative hazards for three gap times are different, where $\Lambda_{01}(t) = 0.25t$, $\Lambda_{02}(t) = 0.5t$, and $\Lambda_{03}(t) = 0.75t^{0.75}$, and censoring constant $a_c = 0.1$ which leads to about 60%, 49%, and 47% subjects with at least one, two and three recurrences;

(iv) baseline cumulative hazards for three gap times are different, where $\Lambda_{01}(t) = 0.25t$, $\Lambda_{02}(t) = 0.5t$, and $\Lambda_{03}(t) = 0.75t^{0.75}$, and censoring constant $a_c = 0.5$ which leads to about 40%, 24%, and 22% subjects with at least one, two and three recurrences.

We truncate any observed time greater than 20 to be 20, which is above the 90th percentile of the first and second recurrence times, and above the 85th percentile of the third recurrence times. In addition, we explore and compare the performances of our method when adding some baseline noise variables. Besides the training data, we use a randomly generated testing data set of size 10000 without censoring to evaluate prediction performance.

We consider two sample sizes, 100 and 200. We use a linear kernel $K(x, x') = x^T x'$ in the simulation. For each simulated data set, we apply our method using the linear decision function $f(t, X, Z(\cdot)) = \alpha(t) + X^T\beta + Z^T(t)\gamma$, where $Z^T(t) = \log(T_{k-1})$. The tuning parameter C_n is chosen via 5-fold cross-validation among the set $\{2^{-16}, 2^{-15}, \ldots, 2^{15}, 2^{16}\}$

using the predicted and observed times of the first recurrence. As a model selection criterion, we use a mean squared error adapted for censoring, which sums up the mean squared difference between fitted times and observed event times if uncensored, and between fitted times and censoring times if censored and the predicted values are less than the observed values. We divide the total sum of squares by the total number of observations. We repeated the simulation 500 times.

5.3.2 Simulation Results

We compare the prediction of our method with the Andersen and Gill proportional intensity model (AG model) for recurrent events. This model assumes $\lambda(t) = \lambda_0(t)\exp(X^T\beta + Z(\cdot)^T\gamma)$, where the hazard function $\lambda_0(t)$ is the same for all the recurrences. Thus the AG model is not the correct model for the simulated data except when $\sigma^2 = 0$ in case 1, and we want to look at its performance for misspecification. For the AG model, after obtaining the estimates of β, γ, and $\lambda_0(t)$, we use the survival curve to obtain the predicted times of three recurrences. For a new subject with covariates $\tilde{X}$ and $\tilde{Z}(\cdot)$, given the kth event time $\tilde{T}_k$, the survival curve is $\hat{S}(t) = \exp[\hat{\Lambda}_0(t)\exp\{\tilde{X}^T\hat{\beta} + \log(\tilde{T}_k)\hat{\gamma}\}]$, where $\hat{\Lambda}_0(t) = \sum_{t'<t}\hat{\lambda}_0(t')$. Then we left truncate the survival curve at $\tilde{T}_k$, and predict the time to the $(k+1)$th event to be the first event time on the curve whose corresponding survival probability is less than $0.5\hat{S}(\tilde{T}_k)$. If there is no such event, we predict the time to the $(k+1)$th event to be last event time on the curve whose corresponding survival probability is greater than $0.5\hat{S}(\tilde{T}_k)$.

Table 5.1-5.4 summarize the results from simulation. These results are obtained using the gap times of three recurrences in the testing data, e.g. the time to the first recurrence, the time from the first recurrence to the second recurrence, and the time from the second recurrence to the third recurrence. We only use the subjects in the testing data whose all three recurrences can be predicted. For example, if a

subject's first recurrence time is larger than all the observed events in the training data, the time to the second recurrences cannot be predicted. The root mean square error (RMSE) is calculated as $\sqrt{\sum(T_{k-1,k} - \hat{T}_{k-1,k})^2/n}$, $k = 1, 2, 3$, and smaller RMSE indicates better predictive accuracy. We summarize the average of RMSEs over 500 replicates in the tables, with the corresponding sample standard deviation of RMSEs given in the parentheses.

In Table 5.1 and 5.2, Andersen and Gill proportional intensity model is the underlying true model when the variance of frailty is zero, and the corresponding predictions from AG model for all three recurrent event times tend to have smaller RMSE than our method, except when the sample size is 100 and there are 40 noise variables in Table 5.1 and 20 noise variables in Table 5.2. The sample standard deviations of RMSEs from our method are smaller than the ones from AG model in most of the cases, and one possible reason is that we use less number of distinct event times in the training data for the prediction of each recurrence. For example, when predicting the first recurrent time in the testing data, AG model uses all the distinct event times in the training data, while our method uses only all the subjects' first distinct event times. When the variance of frailty is not zero, AG model is no longer the underlying model, and the advantages of our method become more obvious as the variance of frailty increases. When the variance of frailty is two, our method gives more accurate prediction for the second and third recurrences for both sample sizes 100 and 200 and even without any noise variables. In addition, Table 5.1 and 5.2 show that our method is most appealing in the high-dimensional situations, i.e. the low signal-noise ratios and small sample sizes. Particularly, our method avoids the problems of non-convergence and indicates a significant improvement in the prediction accuracy compared with AG model when there are 20 and 40 noise variables for both sample 100 and 200.

Table 5.3 and 5.4 give the results for the cases that three recurrences have different

cumulative baseline hazards. In these cases, AG model is not the underlying model, and the trends of results are similar to Table 5.1 and 5.2. Our method always leads to smaller RMSEs for the prediction of the second and third recurrences when the frailty of variance is one or two. This phenomenon may be because the time-varying covariates are constantly zero for the first recurrence and the inclusion of time-varying covariates depending on event history affects more on the prediction of later recurrences. Another interesting point is the significant increase in the number of non-convergent replicates for AG model when the sample size is 100 and there are 40 noise variables in Table 5.3 compared with the corresponding part in Table 5.1. Hence, the noise variables and small sample size may cause more non-concavity of the partial likelihood when the underlying model is more complicated and more different from the AG model itself.

5.4 Application

We apply our method to analyze data from a bladder cancer study conducted by the Veterans Administration Cooperative Urological Research Group. In this study, all patients had bladder tumors when they entered the trial. These tumors were removed and patients may have multiple tumor recurrences during the study period. A description of the clinical background is provided in Byar (1980). In this example, we consider 85 out of 118 subjects who had nonzero follow-up and were assigned to either thiotepa treatment or placebo. The maximum number of recurrences is 4, and specifically there are 47 subjects with at least one recurrence, 29 subjects with at least two recurrences, 22 subjects with at least three recurrences, and 14 subjects with four recurrences. We study the prediction capability of treatment, the initial number of tumors, and the initial size of tumors for predicting the tumor recurrences. The logarithm of previous recurrence time is added as time-varying covariate, and as a result, we can only consider the subjects with first (second, third) recurrences for the prediction of the second

Table 5.1: Root mean square errors (RMSE) of comparing our method (GSVM) and Andersen and Gill proportional intensity model (AG) for the prediction of recurrent events (Case 1)

Variance			n = 100			n = 200		
of frailty	Noises	Method	1st Recur	2nd Recur	3rd Recur	1st Recur	2nd Recur	3rd Recur
0	0	AG	3.94(0.37)	3.02(0.27)	1.35(0.07)	3.93(0.22)	3.03(0.19)	1.29(0.02)
		GSVM	4.46(0.41)	3.36(0.29)	1.65(0.06)	4.58(0.28)	3.47(0.19)	1.65(0.03)
	10	AG	4.31(0.46)	3.14(0.34)	1.48(0.11)	4.12(0.24)	2.99(0.18)	1.38(0.03)
		GSVM	4.63(0.44)	3.37(0.29)	1.73(0.07)	4.59(0.26)	3.37(0.16)	1.72(0.03)
	20	AG	4.80(0.60)	3.37(0.42)	1.59(0.21)	4.26(0.31)	3.05(0.22)	1.37(0.04)
		GSVM	4.76(0.43)	3.46(0.32)	1.71(0.09)	4.64(0.27)	3.39(0.18)	1.69(0.03)
	40	AG[a]	5.91(0.87)	4.29(0.70)	2.31(0.56)	4.57(0.35)	3.31(0.25)	1.38(0.06)
		GSVM	4.89(0.48)	3.66(0.39)	1.71(0.16)	4.73(0.27)	3.54(0.19)	1.61(0.04)
1	0	AG	4.37(0.43)	4.31(0.63)	1.79(0.17)	4.42(0.25)	4.38(0.46)	1.68(0.08)
		GSVM	5.03(0.49)	4.06(0.54)	1.82(0.09)	5.19(0.31)	4.19(0.42)	1.80(0.04)
	10	AG	4.86(0.62)	4.84(0.75)	2.16(0.33)	4.50(0.26)	4.66(0.45)	1.82(0.11)
		GSVM	5.09(0.52)	4.33(0.60)	1.96(0.15)	5.15(0.33)	4.40(0.42)	1.88(0.06)
	20	AG	5.67(0.80)	5.13(0.76)	2.64(0.52)	4.92(0.42)	4.78(0.57)	2.02(0.22)
		GSVM	5.28(0.59)	4.39(0.74)	2.12(0.55)	5.27(0.36)	4.31(0.42)	1.93(0.08)
	40	AG [b]	7.35(1.01)	6.25(0.89)	4.04(0.85)	5.58(0.51)	5.19(0.50)	2.36(0.28)
		GSVM	5.45(0.60)	4.61(0.64)	2.23(0.41)	5.40(0.37)	4.52(0.43)	1.98(0.13)
2	0	AG	4.49(0.46)	5.28(0.89)	2.19(0.31)	4.56(0.23)	5.43(0.70)	2.03(0.18)
		GSVM	5.21(0.57)	4.75(0.77)	2.01(0.19)	5.42(0.36)	4.96(0.61)	1.97(0.12)
	10	AG	5.05(0.70)	5.81(0.90)	2.73(0.50)	4.63(0.34)	5.66(0.65)	2.25(0.25)
		GSVM	5.33(0.63)	5.03(0.81)	2.26(0.30)	5.39(0.41)	5.14(0.61)	2.10(0.16)
	20	AG	6.04(0.96)	6.18(0.95)	3.42(0.73)	5.11(0.45)	5.83(0.67)	2.54(0.34)
		GSVM	5.54(0.72)	5.04(0.81)	2.47(0.49)	5.55(0.41)	5.08(0.60)	2.18(0.19)
	40	AG[c]	7.91(1.21)	7.24(1.04)	5.00(1.02)	5.87(0.58)	6.22(0.58)	3.13(0.44)
		GSVM	5.61(0.76)	5.23(0.80)	2.60(0.56)	5.59(0.45)	5.27(0.56)	2.29(0.27)

[a] 5 out of 500 replicates do not converge for n = 100.

[b] 9 out of 500 replicates do not converge for n = 100.

[c] 7 out of 500 replicates do not converge for n = 100.

Table 5.2: Root mean square errors (RMSE) of comparing our method (GSVM) and Andersen and Gill proportional intensity model (AG) for the prediction of recurrent events (Case 2)

Variance			n = 100			n = 200		
of frailty	Noises	Method	1st Recur	2nd Recur	3rd Recur	1st Recur	2nd Recur	3rd Recur
0	0	AG	3.78(0.70)	3.00(0.57)	1.56(0.29)	3.85(0.46)	2.99(0.37)	1.39(0.15)
		GSVM	4.21(0.72)	3.17(0.52)	1.74(0.23)	4.43(0.48)	3.33(0.33)	1.68(0.11)
	10	AG	4.56(0.99)	3.44(0.77)	1.90(0.50)	4.12(0.55)	3.04(0.42)	1.51(0.13)
		GSVM	4.48(0.82)	3.35(0.62)	1.90(0.35)	4.51(0.52)	3.29(0.33)	1.76(0.09)
	20	AG [a]	5.53(1.33)	4.09(1.05)	2.42(0.79)	4.53(0.61)	3.25(0.45)	1.56(0.19)
		GSVM	4.60(0.82)	3.43(0.61)	1.91(0.31)	4.60(0.47)	3.36(0.31)	1.73(0.09)
1	0	AG	4.12(0.88)	4.07(1.01)	2.12(0.50)	4.26(0.50)	4.14(0.72)	1.85(0.23)
		GSVM	4.49(0.87)	3.62(0.84)	1.96(0.32)	4.85(0.54)	3.81(0.56)	1.83(0.15)
	10	AG	5.03(1.28)	4.80(1.28)	2.71(0.83)	4.58(0.63)	4.53(0.78)	2.10(0.36)
		GSVM	4.64(0.91)	3.96(0.94)	2.20(0.64)	4.85(0.56)	4.03(0.62)	1.94(0.20)
	20	AG	6.05(1.44)	5.27(1.26)	3.39(0.99)	5.14(0.78)	4.74(0.82)	2.39(0.45)
		GSVM	4.79(0.96)	3.99(0.88)	2.29(0.60)	4.98(0.60)	4.03(0.63)	2.00(0.20)
2	0	AG	4.13(1.00)	4.68(1.29)	2.55(0.72)	4.30(0.55)	4.82(0.94)	2.22(0.41)
		GSVM	4.52(0.96)	4.08(1.21)	2.21(0.91)	4.93(0.62)	4.26(0.77)	1.97(0.23)
	10	AG	5.16(1.39)	5.44(1.47)	3.28(1.01)	4.67(0.69)	5.23(0.97)	2.56(0.52)
		GSVM	4.68(0.99)	4.38(1.21)	2.49(0.91)	4.99(0.65)	4.51(0.83)	2.13(0.29)
	20	AG	6.30(1.51)	5.96(1.39)	4.05(1.13)	5.30(0.84)	5.55(0.96)	2.98(0.62)
		GSVM	4.89(1.07)	4.45(1.14)	2.64(0.79)	5.13(0.67)	4.58(0.82)	2.22(0.31)

[a] 3 out of 500 replicates do not converge for n = 100.

Table 5.3: Root mean square errors (RMSE) of comparing our method (GSVM) and Andersen and Gill proportional intensity model (AG) for the prediction of recurrent events (Case 3)

Variance			n = 100			n = 200		
of frailty	Noises	Method	1st Recur	2nd Recur	3rd Recur	1st Recur	2nd Recur	3rd Recur
0	0	AG	3.98(0.43)	3.07(0.51)	1.08(0.20)	4.04(0.27)	3.17(0.36)	0.94(0.06)
		GSVM	4.37(0.48)	2.80(0.47)	1.13(0.19)	4.56(0.34)	2.92(0.32)	1.04(0.06)
	10	AG	4.30(0.48)	3.20(0.53)	1.23(0.23)	4.19(0.31)	3.15(0.36)	1.04(0.07)
		GSVM	4.55(0.50)	2.89(0.49)	1.23(0.20)	4.54(0.31)	2.79(0.29)	1.11(0.06)
	20	AG [a]	4.76(0.65)	3.38(0.61)	1.41(0.29)	4.35(0.32)	3.12(0.39)	1.05(0.11)
		GSVM	4.71(0.50)	2.96(0.49)	1.26(0.19)	4.63(0.32)	2.79(0.32)	1.09(0.07)
	40	AG[b]	5.72(0.91)	4.24(0.82)	2.30(0.69)	4.62(0.34)	3.43(0.35)	1.11(0.12)
		GSVM	4.82(0.58)	3.26(0.57)	1.36(0.35)	4.71(0.31)	3.03(0.32)	1.05(0.10)
1	0	AG	4.40(0.52)	4.68(0.92)	1.74(0.30)	4.55(0.28)	4.95(0.69)	1.57(0.18)
		GSVM	4.98(0.57)	3.90(0.75)	1.51(0.24)	5.19(0.38)	4.05(0.60)	1.39(0.12)
	10	AG	4.82(0.61)	5.11(0.91)	2.15(0.42)	4.58(0.31)	5.22(0.60)	1.76(0.18)
		GSVM	5.06(0.58)	4.15(0.78)	1.73(0.31)	5.13(0.39)	4.20(0.57)	1.51(0.15)
	20	AG	5.52(0.83)	5.29(0.90)	2.69(0.61)	4.95(0.42)	5.24(0.66)	1.97(0.30)
		GSVM	5.25(0.70)	4.20(0.80)	1.92(0.46)	5.27(0.40)	4.14(0.57)	1.58(0.20)
	40	AG [c]	7.06(1.13)	6.22(1.05)	4.20(0.98)	5.47(0.53)	5.41(0.64)	2.37(0.33)
		GSVM	5.38(0.70)	4.40(0.80)	2.14(0.56)	5.39(0.43)	4.31(0.53)	1.74(0.27)
2	0	AG	4.46(0.54)	5.45(1.14)	2.25(0.46)	4.66(0.29)	5.83(0.94)	2.10(0.32)
		GSVM	5.17(0.66)	4.66(1.00)	1.93(0.43)	5.42(0.41)	4.85(0.75)	1.78(0.26)
	10	AG	4.91(0.69)	5.90(1.08)	2.83(0.61)	4.68(0.30)	6.07(0.81)	2.34(0.35)
		GSVM	5.31(0.69)	4.91(0.99)	2.28(0.53)	5.40(0.44)	4.98(0.76)	1.95(0.30)
	20	AG	5.79(0.98)	6.21(1.11)	3.56(0.79)	5.07(0.47)	6.14(0.78)	2.66(0.43)
		GSVM	5.50(0.82)	4.97(1.00)	2.51(0.64)	5.58(0.47)	5.00(0.76)	2.10(0.38)
	40	AG[d]	7.53(1.31)	7.12(1.21)	5.21(1.14)	5.66(0.56)	6.31(0.71)	3.20(0.47)
		GSVM	5.55(0.84)	5.11(1.01)	2.68(0.72)	5.60(0.50)	5.12(0.67)	2.26(0.45)

[a] 1 out of 500 replicates does not converge for n = 100.

[b] 52 out of 500 replicates do not converge for n = 100.

[c] 29 out of 500 replicates do not converge for n = 100.

[d] 19 out of 500 replicates do not converge for n = 100.

Table 5.4: Root mean square errors (RMSE) of comparing our method (GSVM) and Andersen and Gill proportional intensity model (AG) for the prediction of recurrent events (Case 4)

Variance			n = 100			n = 200		
of frailty	Noises	Method	1st Recur	2nd Recur	3rd Recur	1st Recur	2nd Recur	3rd Recur
0	0	AG	3.61(0.78)	2.90(0.83)	1.39(0.48)	3.80(0.50)	3.00(0.60)	1.13(0.27)
		GSVM	4.02(0.81)	2.65(0.70)	1.38(0.45)	4.30(0.56)	2.70(0.47)	1.18(0.25)
	10	AG	4.32(0.96)	3.34(0.93)	1.78(0.64)	4.02(0.56)	3.04(0.63)	1.24(0.28)
		GSVM	4.32(0.84)	2.89(0.77)	1.59(0.53)	4.37(0.58)	2.67(0.45)	1.26(0.26)
	20	AG [a]	5.12(1.30)	3.87(1.15)	2.34(0.90)	4.39(0.60)	3.17(0.60)	1.34(0.31)
		GSVM	4.45(0.90)	3.00(0.79)	1.68(0.58)	4.49(0.52)	2.74(0.46)	1.26(0.24)
1	0	AG	3.96(0.95)	4.00(1.21)	2.10(0.72)	4.20(0.57)	4.25(0.99)	1.74(0.35)
		GSVM	4.38(0.98)	3.34(1.05)	1.77(0.61)	4.76(0.62)	3.46(0.76)	1.49(0.28)
	10	AG	4.67(1.25)	4.58(1.41)	2.67(0.93)	4.43(0.66)	4.55(0.98)	2.00(0.45)
		GSVM	4.51(1.01)	3.63(1.05)	2.03(0.66)	4.76(0.65)	3.64(0.77)	1.63(0.34)
	20	AG [b]	5.56(1.44)	5.04(1.39)	3.41(1.15)	4.89(0.80)	4.71(0.98)	2.31(0.54)
		GSVM	4.66(1.04)	3.75(1.10)	2.24(0.87)	4.88(0.68)	3.71(0.82)	1.72(0.39)
2	0	AG	3.90(1.04)	4.25(1.40)	2.52(0.90)	4.20(0.64)	4.55(1.15)	2.17(0.56)
		GSVM	4.43(1.08)	3.77(1.27)	2.11(0.86)	4.84(0.73)	3.88(0.91)	1.79(0.41)
	10	AG	4.72(1.32)	5.05(1.57)	3.27(1.11)	4.46(0.71)	4.98(1.16)	2.53(0.61)
		GSVM	4.58(1.08)	4.09(1.31)	2.45(0.98)	4.90(0.78)	4.14(1.01)	2.00(0.53)
	20	AG	5.74(1.56)	5.59(1.57)	4.08(1.30)	4.99(0.86)	5.28(1.15)	2.95(0.70)
		GSVM	4.80(1.19)	4.24(1.34)	2.71(1.03)	5.08(0.76)	4.29(0.98)	2.13(0.55)

[a] 9 out of 500 replicates do not converge for n = 100.

[b] 2 out of 500 replicates do not converge for n = 100.

(third, forth) recurrences.

We compare the performances of our method and AG model. Due to the small sample size and discrete values of covariates, we use only linear kernel here to avoid the potential overfitting problem. The predicted times of tumor recurrences are obtained via three-fold cross validation, and the tuning parameter is chosen from the grid $2^{-16}, 2^{-15}, \ldots, 2^{16}$. To compare the prediction capability, we separate the data into two groups based on the 25th and 50th percentiles of the predicted times, and report a pseudo Chi-square statistics from the Logrank test and a pseudo hazard ratio from fitting a univariate Cox model. The results are given in Table 5.5. The pseudo Chi-square statistics indicate that the prediction of our method has better performance regarding to the risk stratification than the AG model for both the first and second recurrences, and the superiority of our method is more obvious for the first recurrence by the large values of both pseudo Chi-square statistics and pseudo hazard ratios. Table 5.6 gives the coefficient estimates that complement the results in Table 5.5. The estimates from both methods have the same signs, but the relative covariate effects differ. Particularly, our method leads to relatively large effect of the initial number of tumors and the prior recurrence time, while AG model gives the thiotepa treatment the largest estimate.

5.5 Remark

In this chapter, we propose a conceptually straightforward method to the prediction of recurrent event time data. This method adapts support vector machines to learn the counting process, and then use the learned information to make predictions. The newly developed prediction rule indicates the similarity between the proposed method and standard multicategory support vector machines. The time-specific intercepts $\alpha(t)$ are utilized in the prediction rule, and they may not be uniquely determined in certain cases. Different values of $\alpha(t)$ may lead to slightly different prediction results, however,

Table 5.5: Comparison of prediction capability for our method and Andersen and Gill proportional intensity model using bladder cancer data

Recurrence	Method	25th percentile		50th percentile	
		Pseudo Logrank χ^2 [a]	Pseudo HR [b]	Pseudo Logrank χ^2	Pseudo HR HR
1st	AG model	5.07	2.00	1.27	1.39
	Our method	9.68	2.49	6.29	2.11
2nd	AG model	1.54	1.83	0.46	1.30
	Our method	2.19	1.63	1.77	1.55

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

[b]Pseudo HR, pseudo hazard ratios comparing two groups separated using the 25th and 50th percentiles of predicted values.

Table 5.6: Coefficient estimates for bladder cancer data

Covariate	Normalized β of our method	β of AG model
Treatment	-0.215	-0.378
Initial tumor number	0.447	0.159
Initial tumor size	-0.100	-0.040
Prior recurrence	0.863	0.340

major differences of predictive accuracy are not expected. Another attraction of this method is the comparability with the intensity-based survival models. These models are essentially based on the cumulative intensity functions. Comparatively, the optimal rule of our method focuses on the intensity functions, which can be thought as a local view in parallel with the global view of the survival models. Simulation results reveal the superiority of our method to the Andersen and Gill proportional intensity model when this model is not the underlying model. In addition, our method is a convex quadratic programming algorithm, so it is particularly appealing to be applied in the high-dimensional situations for which the partial likelihood function often breaks down due to the occurrence of non-concavity.

The time-varying covariates are included and discussed in our method to make the prediction of the next event time based on not only baseline attributes but the event history. In practice, we may need to determine the type of time-varying covariates using the background information of the study, and we may want to explore multiple choices for comparison and conduct sensitivity analysis. The framework of counting process implicitly assumes that all time-varying covariates are predictable at the present point. For the case of general stochastic processes, the application of the proposed method cannot be fully justified unless they are treated as predictable time-varying covariates. As a result, we may be able to establish the asymptotic learning rate in a similar way to the one in Chapter 4 that includes only survival data and time-independent covariates. The formal derivation will be further investigated.

CHAPTER6: SUMMARY AND FUTURE RESEARCH

In this dissertation, we have studied the semiparametric and nonparametric statistical methods for variable selection and survival outcome prediction using censored data, which relax the assumptions on the underlying model and censoring mechanism of many existing approaches. Particularly, in Chapter 3, we proposed a penalized variable selection procedure in general transformation models. The Laplace transformation and expectation-maximization algorithm were used to obtain an objective function that removes the nonparametric estimation of baseline cumulative hazards and includes only the parameter of interest to incorporate penalties for variable selection. In Chapter 4, we developed a support vector hazards regression to predict the time to event. The failure times were presented in the notations of counting process so that the statuses of all the subjects still at risk at each event time become binary outcomes, and the support vector machines were adapted with restrictions on the covariate effects to learn the counting process. We found that the resulting optimal decision rule discriminates the covariate-specific hazard function from the population average hazard function. In Chapter 5, we generalized the support vector machines in the framework of the counting process to handle time-varying covariates and predict recurrent event times based on the event history. The proposed method allows the censoring mechanism to depend on covariates without specifying the censoring distribution.

Theoretically, we established the asymptotic selection consistency using the adaptive lasso penalty for the penalized variable selection procedure, and derived the asymptotic learning rate using the Gaussian kernel for the support vector hazards regression. The

proofs heavily rely on the modern empirical process theory. We also conducted extensive simulation studies to explore the small-sample performances of all the proposed methods, and demonstrated the comparability and superiority of our methods to existing approaches. Several real data examples were used to illustrate the proposed methods. Specifically, in Chapter 3, we used part of the baseline cohort data in the Atherosclerosis Risk in Communities study, including traditional cardiovascular risk factors for incident heart failure, and of the primary biliary cirrhosis data from Mayo Clinic trial of primary biliary liver cirrhosis. In Chapter 4, we used the Atherosclerosis Risk in Communities data again, and we also apply our method to the data collected from a neurological disease study. We analyzed data from a bladder cancer study conducted by the Veterans Administration Cooperative Urological Research Group in Chapter 5.

The proposed methods in this dissertation can be extended in several directions for future research. In Chapter 3, we used the adaptive lasso penalty due to computational advantages. In practice, other penalties may be more appropriate to be considered for handling some specific problems, for example, fused lasso (Tibshirani et. al, 2005) for problems with ordered features and elastic-net (Zou and Hastie, 2005) for problems with grouping effects. These penalties can be easily applied with the proposed objective function, however, the current computation algorithm may not work, and another algorithm may need to be developed before numerical application. In Chapters 4 and 5 we adapted support vector machines in the framework of the counting process to predict survival outcomes and recurrent events, and we considered the scenario of having right censoring. There are other complications with censored data, including left truncation and competing risk, that are commonly presented using counting process to make statistical inference. Our methods may also be extended to these data, for example, predicting a certain outcome in the setting of competing risk. In addition,

as described in the previous chapter, another formulation of our methods is as an empirical risk plus a regularization term, and we used the L_2 regularization by following the standard support vector machines. For specific future work, we may want to use or additionally include the L_1 regularization to conduct variable selection and make prediction simultaneously. However, this modification of our methods will no longer be a convex quadratic programming algorithm, and the issues of implementation need to be further studied.

REFERENCES

Akaike, H. (1974), "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

Andersen, P. K. (1982), "Testing goodness of fit of Cox's regression and life model," *Biometrics*, 38, 67–77.

Andersen, P. K. and D., G. R. (1982), "Cox's regression model for counting processes: a large sample study," *Annals of Statistics*, 10, 1100–1120.

Bennett, S. (1983a), "Analysis of survival data by the proportional odds model," *Statistics in Medicine*, 2, 273–2777.

— (1983b), "Log-logistic regression models for survival data," *Journal of the Royal Statistical Society: Series C*, 32, 165–171.

Biganzoli, E., Boracchi, P., Mariani, L., and Marubini, E. (1998), "Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach," *Statistics in Medicine*, 17, 1169–1186.

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

Breiman, L. (1995), "Better subset regression using the nonnegative garrote," *Technometrics*, 37, 373–384.

— (1996), "Heuristics of instability and stabilization in model selection," *Annals of Statistics*, 24, 2350–2383.

Breslow, N. (1972), "Contribution to the discussion of the paper by D. R. Cox." *Journal of the Royal Statistical Society: Series B*, 34, 187–220.

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

Burges, C. J. C. (1998), "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, 2, 121–167.

Butler, J., Kalogeropoulos, A., Georgiopoulou, V., Belue, R., Rodondi, N., Garcia, M., Bauer, D. C., Satterfield, S., Smith, A. L., Vaccarino, V., Newman, A. B., Harris, T. B., Wilson, P. W., and Kritchevsky, S. B. (2008), "Incident heart failure prediction

in the elderly: the health ABC heart failure score," *Circulation: Heart Failure*, 1, 125–133.

Chambless, L. and Diao, G. (2006), "Estimation of time-dependent area under the ROC curve for long-term risk prediction," *Statistics in Medicine*, 25, 3474–3486.

Chatfield, C. (1995), "Model uncertainty, data mining and statistical inference," *Journal of the Royal Statistical Society: Series A*, 158, 419–466.

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

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

Cox, D. R. (1972), "Regressioin Models and Life-Tables (with Discussion)," *Journal of Royal Statistical Society: Series B*, 34, 187–220.

— (1975), "Partial likelihood," *Biometrika*, 62, 269–276.

Craven, P. and Wahba, G. (1979), "Smoothing noisy data with spline functions," *Numerische Mathematik*, 31, 377–403.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm (with Discussion)," *Journal of the Royal Statistical Society: Series B*, 39, 1–38.

Derksen, S. and Keselman, H. J. (1992), "Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables," *British Journal of Mathematical and Statistical Psychology*, 45, 265–282.

Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D., and Langworthy, A. (1989), "Prognosis in primary biliary cirrhosis: model for decision making," *Hepatology*, 10, 1–7.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least angle regression," *Annals of Statistics*, 32, 407–499.

Fan, J. and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and

its oracle properties," *Journal of the American Statistical Association*, 96, 1348–1360.

— (2002), "Variable selection for Cox's proportional hazards model and frailty model," *Annals of Statistics*, 30, 74–99.

Faraggi, D. and Simon, R. (1995), "A neural network model for survival data," *Statistics in Medicine*, 14, 73–82.

Fernandez, M. and Miranda-Saavedra, D. (2012), "Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines," *Nucleic Acids Research*, 40, e77.

Fisher, L. D. and Lin, D. Y. (1999), "Time-dependent covariates in the Cox proportional-hazards regression model," *Annual Review of Public Health*, 20, 145–157.

Frank, I. E. and Friedman, J. H. (1993), "A statistical view of some chemometrics regression tools," *Technometrics*, 35, 109–135.

Friedman, J. (1996), "Another approach to polychotomous classification," *Technical report, Stanford University.*

Fu, W. (1998), "Penalized regressions: the bridge versus the lasso," *Journal of Computational and Graphical Statistics*, 7, 397–416.

Ganapathy, S., Yogesh, P., and Kannan, A. (2012), "Intelligent agent-based intrusion detection system using enhanced multiclass SVM," *Computational Intelligence and Neuroscience*, 2012, 1–10.

Goldberg, Y. and Kosorok, M. R. (2012), "Q-learning with censored data," *Annals of Statistics*, 40, 529–560.

— (2013), "Support vector regression for right censored data," *Unpublished manuscript.*

Harrell, F. E., Lee, K. L., and Mark, D. B. (1996), "Tutorial in Biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statistics in Medicine*, 15, 361–387.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The elements of statistical learning: data mining, inference, and prediction, second edition*, New York: Springer.

Hidalgo, H., Sosa Leon, S., and Gomez-Trevino, E. (2003), "Application of the kernel

method to the inverse geosounding problem," *Neural Networks*, 16, 349–353.

Hothorn, T., Lausen, B., Benner, A., and Radespiel-Troger, M. (2004), "Bagging survival trees," *Statistics in Medicine*, 23, 77–91.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008), "Random survival forests," *Annals of Applied Statistics*, 2, 841–860.

Jin, Z., Lin, D. Y., Wei, L. J., and Ying, Z. (2003), "Rank-based inference for the accelerated failure time model," *Biometrika*, 90, 341–353.

Johnson, B. A. (2008), "Variable selection in semiparametric linear regression with censored data," *Journal of the Royal Statistical Society: Series B*, 70, 351–370.

— (2009), "On lasso for censored data," *Electronic Journal of Statistics*, 3, 485–506.

Johnson, B. A., Lin, D. Y., and Zeng, D. (2008), "Penalized estimating functions and variable selection in semiparametric regression models," *Journal of the American Statistical Association*, 103, 672–680.

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

Knight, K. and Fu, W. (2000), "Asymptotics for lasso-type estimators," *Annals of Statistics*, 28, 1356–1378.

Lam, K. F. and Leung, T. L. (2001), "Marginal likelihood estimation for proportional odds models with right censored data," *Lifetime Data Analysis*, 7, 39–54.

Leblanc, M. and Crowley, J. (1992), "Relative risk trees for censored survival data," *Biometrics*, 48, 411–425.

— (1993), "Survival Trees by Goodness of Split," *Journal of the American Statistical Association*, 88, 457–467.

Lee, E. T. and Go, O. T. (1997), "Survival analysis in public health research," *Annual Review of Public Health*, 18, 105–134.

Li, J. and Gu, M. (2012), "Adaptive LASSO for general transformation models with right censored data," *Computational Statistics & Data Analysis*, 56, 2583–2597.

Louis, T. (1982), "Finding the observed information matrix when using the EM algorithm," *Journal of the Royal Statistical Society: Series B*, 44, 226–233.

Lu, W. and Zhang, H. H. (2007), "Variable selection for proportional odds model," *Statistics in Medicine*, 26, 3771–3781.

Meinshausen, N. (2007), "Relaxed lasso," *Computational Statistics & Data Analysis*, 52, 374–393.

Murphy, S. A., Rossini, A. J., and Van Der Vaart, A. W. (1997), "Maximum likelihood estimation in the proportional odds model," *Journal of the American Statistical Association*, 92, 968–76.

Park, J. I. and Jeong, M. K. (2011), "Recursive support vector censored regression for monitoring product quality based on degradation profiles," *Applied Intelligence*, 35, 63–74.

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

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

Pettitt, A. N. (1983), "Approximate methods using ranks for regression with censored data," *Biometrika*, 70, 121–132.

— (1984), "Proportional odds models for survival data and estimates using ranks," *Journal of the Royal Statistical Society: Series C*, 33, 169–175.

Radchenko, P. and James, G. M. (2011), "Improved variable selection with forward-LASSO adaptive shrinkage," *Annals of Applied Statistics*, 5, 427–448.

Rao, C. S., Kumar, S. S., and Mohan, B. C. (2010), "Content based image retrieval using exact legendre moments and support vector machine," *The International Journal of Multimedia and Its Application*, 2, 69–79.

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

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

Schoenfeld, D. (1982), "Partial residuals for the proportional hazards regression model," *Biometrika*, 69, 239–241.

Scholkopf, B., Dumais, S. T., Osuna, E., and Platt, J. (1998), "Support vector machines," *In IEEE Intelligent Systems Magazine, Trends and Controversies, Marti Hearst, editor*, 13, 18–28.

Schwarz, G. (1978), "Estimating the dimension of a model," *Annals of Statistics*, 6, 461–464.

Segal, M. R. (1988), "Regression trees for censored data," *Biometrics*, 44, 35–48.

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

Smith, G. and Campbell, F. (1980), "A critique of some ridge regression methods," *Journal of the American Statistical Association*, 75, 74–81.

Steinwart, I. and Christmann, A. (2008), *Support vector machines, first edition*, New York: Springer.

Steyerberg, E. W., Eijkemans, M. J., and Habbema, J. D. (1999), "Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis," *Journal of Clinical Epidemiology*, 52, 935–42.

Tay, F. E. H. and Cao, L. (2001), "Application of support vector machines in financial time series forecasting," *Omega: The International Journal of Management Science*, 29, 309–317.

The ARIC investigators (1989), "The atherosclerosis risk in communities study: Design and objectives," *American Journal of Epidemiology*, 129, 687–702.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B*, 58, 267–88.

— (1997), "The lasso method for variable selection in the Cox model," *Statistics in Medicine*, 16, 385–395.

Tibshirani, R., Saunders, M., Rossert, S., Zhu, J., and Knight, K. (2005), "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B*, 67, 91–108.

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

— (2011a), "Learning transformation models for ranking and survival analysis," *Journal of Machine Learning Research*, 12, 819–862.

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

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

Walter, S. and Tiemeier, H. (2009), "Variable selection: current practice in epidemiological studies," *European Journal of Epidemiology*, 24, 733–736.

Zeng, D. and Lin, D. Y. (2006a), "The adaptive Lasso and its oracle properties," *Journal of the American Statistical Association*, 101, 1418–1429.

— (2006b), "Efficient estimation of semiparametric transformation models for couting processes," *Biometrika*, 93, 627–640.

— (2007a), "Efficient estimation for the accelerated failure time model," *Journal of the American Statistical Association*, 102, 1387–1396.

— (2007b), "Maximum likelihood estimation in semiparametric regression models with censored data (with Discussion)," *Journal of the Royal Statistical Society: Series B*, 69, 507–564.

Zhang, H. H. and Lu, W. (2007), "Adaptive Lasso for Cox's proportional hazards model," *Biometrika*, 94, 691–703.

Zhang, H. H., Lu, W., and Wang, H. (2010), "On sparse estimation for semiparametric linear transformation models," *Journal of Multivariate Analysis*, 101, 1594–1606.

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

Genetics, 156B, 751–763.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B*, 67, 301–320.