Latent Supervised Learning for Survival Data

Susan Wei and Michael R. Kosorok

August 19, 2013

Abstract

Latent supervised learning is a machine learning technique for performing binary classification using a surrogate variable for the unobserved training label. We extend latent supervised learning to the case when the surrogate variable is a right-censored survival time. A motivating application for the proposed methodology is to stratify patients into two risk groups given a set of biomarkers. Sieve maximum likelihood estimation is employed for model estimation with special care taken to account for censoring. Consistency of the proposed estimator is established. Simulations show that the proposed estimator is accurate under a range of settings. Applications to real data examples demonstrate its advantages over a competing method; the proposed method produces more significant separation in survival on both training sets and held-out independent test sets.

Keywords: Censoring; Classification and Clustering; Cox Model; Inverse Probability of Censoring Weighted; Proportional Hazards; Random Forest; Statistical Learning; Sieve Maximum Likelihood Estimation; Sliced Inverse Regression; Survival Analysis.

[1]Susan Wei is a doctoral student, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 (Email: susanwe@live.unc.edu). Michael R. Kosorok is Professor and Chair, Department of Biostatistics, and Professor, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 (Email: kosorok@unc.edu). The first author was funded through the NSF Graduate Fellowship. The second author was funded in part by NIH grant CA142538.

1 Introduction

Latent supervised learning is a machine learning technique for performing binary classification using a surrogate variable when labeled training data is unavailable. Wei and Kosorok (2013) first introduced this idea and applied it to a model wherein the surrogate variable is Gaussian distributed. Here we extend the methodology to the surrogate variable that is a right-censored survival time.

The proposed methodology is particularly motivated by the problem of stratifying a population into two risk groups based on a set of biomarkers. One possibility for risk stratification is to cluster the patients based solely on biomarkers. This approach ensures patients with similar biomarker values are assigned to the same risk group. There is no guarantee, however, that the resulting clusters will exhibit different survival experiences. Another approach is to stratify patients based solely on survival time. This, however, is likely to have the undesirable effect of producing subgroups of patients in the same risk group with dissimilar biomarker patterns.

The proposed methodology can be applied to discover subgroups that are both biologically and clinically meaningful. An index is constructed which divides the population into two groups using the the binary rule "index $<$ cutpoint." The index is a linear combination of the biomarkers to be estimated from the data. Our model takes the classic Cox model as a starting point. Let T denote the true (unobservable) lifetime and C the censoring time. The observed data consists of $Y = \min(T, C)$, $\delta = 1\{T \leq C\}$ and a real p-dimensional covariate vector X. The proposed model assumes the linear hyperplane in the covariate space defined by $\omega_0^T x - \gamma_0 = 0$, where $\omega_0 \in \mathbb{R}^p$ and $\gamma_0 \in \mathbb{R}$, "separates" the survival times into two distributions with proportional hazards. Accordingly, the conditional hazard function is given by

$$h(t|x) = \exp(\beta_0 1\{\omega_0^T x - \gamma_0 \geq 0\}) h_0(t), \tag{1}$$

where $h_0(t)$ is the baseline hazard function and β_0 is the log hazards ratio, assumed to be nonzero for identifiability. It is further assumed that censoring C is independent of survival T, conditional on X.

In the estimation of Model (1), the primary parameters of interest are ω_0 and γ_0. The estimation procedure used in the Gaussian model in Wei and Kosorok (2013) will be adapted for Model (1) to account for censoring. The estimation is obtained by maximizing the Cox proportional hazards partial likelihood over a data-driven sieve, an approximating space constructed to grow dense as sample size increases.

1.1 Related work

Tian and Tibshirani (2011) proposed the adaptive index model which constructs a score that is the sum of several binary rules such as "age > 60" or "blood pressure > 120 mm Hg". The conditional hazard function in an adaptive index model is given by

$$h(t|x) = \exp(\beta_0 + \beta \sum_{k=1}^{K} 1\{x_k^* \leq c_k\}) h_0(t), \tag{2}$$

where x_k^* are from the set $\{\pm x_1, \ldots, \pm x_p\}$ and K, the number of binary rules, is no bigger than p, the dimension of the covariate vector X.

In both Model (1) and (2), the covariate space is divided into regions with different survival experiences. As linear boundaries offer an especially rich and flexible array of binary partitions in high dimensions, Model (1) may in this regard have certain advantages over the adaptive index model in which boundaries always remain parallel to the coordinate axes. However, as is typical in statistics, there is no single "correct" model. Scores such as the International Prognostic Index (IPI), based on age, disease stage, etc., used for risk classification in Non-Hodgkins lymphoma is one example of a setting where clinicians may prefer the adaptive index model. On the other hand if the individual covariates are less interpretable, the rules $x_k^* \leq c_k$ become less meaningful and Model (1) is preferable. There is also no need to pre-specify in the proposed model the number of covariates to include in the final stratification rule (K in Model (2)). Variable selection is built into the proposed methodology since all covariates are assigned a weight which give an indication of the importance of a variable.

An alternative to the single binary partition considered in Model (1) is to use tree-based methods and perform recursive binary partitions. The Classification and Regression Tree (CART) methodology of Breiman et al. (1984) is a seminal work in this area. CART and other tree-based methods have many advantages over traditional linear methods in classification and regression. For instance, tree-based methods perform well even if the assumptions deviate from the true model, i.e they are robust. Several authors have extended tree-based methods in the setting of censored survival data (Leblanc and Crowley, 1993; Banerjee and Noone, 2007). Another generalization is multivariate trees which allow for a linear combination of variables at decision and leaf nodes rather than a single variable. In Gama (2004) a method for constructing multivariate survival trees is given.

Trees combine binary rules constructed using univariate variables. In contrast, the proposed methodology constructs a single binary split based on a linear combination of all variables. While tree-based methods allow for finer risk stratification, the model studied

here is more parsimonious. The richness of linear decision boundaries in high dimensions makes a compelling case for studying a parsimonious model as in Model (1). Also, there are fewer tuning parameters involved in the proposed methodology whereas in many tree-based methods, careful decisions have to be made regarding growing the tree and then pruning it back.

There are also several related work in the literature that deal with risk stratification using non-tree approaches. One such methodology was put forth in Bair and Tibshirani (2004). There, a continuous predictor of survival based on gene expression is constructed and a threshold on the predictor is used to identify two subgroups. The procedure has two steps: 1) a subset of genes with the highest Cox scores is selected, and 2) principal components analysis is performed on this subset of the gene expression data, and a proportional hazards model based on the first few principal components is used to obtain a continuous predictor of survival. A disadvantage of this two-step procedure is that genes which do not play a strong individual role but play an important role when considered in an ensemble of genes will be completely removed in the first step. Another procedure that suffers from the same drawback is the method proposed in Wu et al. (2008), which also involves two steps. In the first step, a subset of genes is selected using correlation and liquid association. Liquid association is used in studying coexpression patterns between three genes. Although this is likely an improvement on a univariate approach to selecting genes, important genes may still be omitted from the analysis.

In the simulations and data examples found later in this paper, we will directly compare the proposed methodology to a method proposed in Li et al. (1999), which will be referred to as Li's double-slicing method. This method is based on the dimension reduction technique Sliced Inverse Regression (SIR) (Li, 1991). The SIR model assumes the response variable $T \in \mathbb{R}$ is related to the covariate vector $X \in \mathbb{R}^p$ in the following manner:

$$T = g(\omega_{0,1}^T X, \ldots, \omega_{0,K}^T X, \epsilon). \tag{3}$$

where $\omega_{0,k}$ are unit vectors in $\mathbb{R}^p$, $k = 1, \ldots, K$ for some integer K. One feature of Li's model is that the function g is completely unspecified as is the distribution of the error term ϵ.

Li et al. (1999) proposed the double-slicing method for censored survival data. This is a two-step procedure that first involves reducing the dimension of the covariate space by slicing simultaneously on survival and censoring. A kernel-based approach is then used to estimate the inverse regression curve adjusting for the presence of censoring. The double-

slicing procedure critically assumes censoring follows the SIR assumption, i.e.

$$C = h(\omega_{0,1}^T X, \ldots, \omega_{0,K}^T X, \delta). \tag{4}$$

for some function h and error δ. It will be seen in simulations that Li's double-slicing method is rather sensitive to departures from the assumption in (4).

1.2 Outline

The methodology is described in Section 2 and consistency of the estimator is established in Section 3. Section 4 contains simulation findings comparing two variations of the proposed method and Li's double-slicing method. Applications to real datasets are presented in Section 5. The paper concludes with a discussion in Section 6.

2 Methodology

There are three parameters in Model (1) to estimate – the parameters involved in the hyperplane, ω_0 and γ_0, and the log ratio hazard β_0. An obvious estimation procedure is simply to maximize the Cox proportional hazards partial likelihood over the parameter space of these three variables. However direct maximization is computationally challenging when the dimension of the covariate is high. The proposed method performs maximization over a data-driven approximating space that grows dense as the sample size increases. Such a sequence of approximating spaces is referred to as a sieve in the literature, following the terminology in Grenander (1981). A sieve maximum likelihood approach was also used for the Gaussian Latent Supervised Learning model studied in Wei and Kosorok (2013). We follow their procedure of constructing a preliminary sieve based on information in the covariate space and then updating the sieve by incorporating the surrogate variable. Since the surrogate variable – survival time – can be censored, there are many challenges here not faced in the completely observed Gaussian model studied previously.

2.1 The Estimator

The log proportional hazards partial likelihood is given by

$$L_n(\omega, \gamma, \beta) = n^{-1} \sum_{i=1}^{n} \delta_i \{\beta 1\{\omega^T x_i - \gamma \geq 0\} - \log n^{-1} \sum_{j: y_j \geq y_i} \exp(\beta 1\{\omega^T x_j - \gamma \geq 0\})\}.$$

The factor n^{-1} was added to be consistent with the empirical processes notation in Section 3. Given ω and γ, the estimated log hazard ratio $\hat{\beta}_n(\omega, \gamma) = \arg\max_\beta L_n(\omega, \gamma, \beta)$ can be

found via the standard Newton-Raphson approach. The profile likelihood is given by

$$M_n(\omega, \gamma) = L_n(\omega, \gamma, \hat{\beta}_n(\omega, \gamma)).$$

The profile likelihood $M_n(\omega, \gamma)$ is maximized over a data-driven sieve $\hat{\Omega}_n \subset \mathbb{S}^p$ where $\mathbb{S}^p = \{\omega \in \mathbb{R}^p : ||\omega|| = 1\}$ is the unit sphere in $\mathbb{R}^p$. The proposed estimator, called the sieve estimator, is given by

$$(\hat{\omega}_n^s, \hat{\gamma}_n^s) := \underset{\omega \in \hat{\Omega}_n, \gamma \in \mathbb{R}}{\arg\max}\ M_n(\omega, \gamma). \tag{5}$$

The next two sections describe the construction of the sieve $\hat{\Omega}_n$.

2.2 The Simple Sieve

This section details the construction of the preliminary sieve, also referred to as the simple sieve. The simple sieve is based on the Mean Difference (MD) discrimination rule applied to the covariates x. The MD, also known as the nearest centroid method (see Chapter 1 of Scholkopf and Smola (2001)), is a forerunner to the shrunken nearest centroid method of Tibshirani et al. (2002). It is based on the class sample mean vectors. A new data vector is assigned to the the positive (negative) class if it is closer to the mean vector of the positive (negative) class. Thus the MD discrimination method results in a separating hyperplane with normal vector

$$(\text{ positive class mean vector } - \text{ negative class mean vector }).$$

The simple sieve consists of MD directions formed in the following manner:

1. Partition the covariate space X into K regions. Let $S_k \subset \{1, \ldots, n\}$ be the index set for region k.

2. Let $\mathcal{P}_k$ denote the collection of partitions of the set S_k into two parts. For $P \in \mathcal{P}_k$, let P_1 and P_2 be the parts of the partition, i.e. $P_1 \cup P_2 = S_k$ and $P_1 \cap P_2 = \emptyset$.

3. For each $P \in \bigcup_k \mathcal{P}_k$, calculate the Mean Difference direction $\omega^{MD}(P)$ — the vector connecting the centroids of the two classes $\{X_i : i \in P_1\}$ and $\{X_i : i \in P_2\}$,

$$\omega^{MD}(P) = \frac{\bar{X}_{P_1} - \bar{X}_{P_2}}{||\bar{X}_{P_1} - \bar{X}_{P_2}||},$$

where $\bar{X}_{P_1}$ and $\bar{X}_{P_2}$ are the sample means of X's in P_1 and P_2 respectively.

To obtain a partition of the covariate space, K-means clustering can be used. If K-means returns clusters that are very large, sample a sub-population of the cluster. The parameter K should be chosen to ensure the cardinality of the sieve is not too big. Setting K to be roughly $n/10$ works well in practice. This choice results in the sieve having approximately $\sum_{k=1}^{K} 2^{|S_k|} = n2^{10}/10$ elements, which grows linearly in n and is quite manageable computationally.

2.3 Incorporating the survival data

This section describes the process by which the simple sieve is updated. Let $0 = t_1 < t_2 < \cdots < t_H < \infty = t_{H+1}$ be a partition of the observed survival times and $I_h = [t_h, t_{h+1})$ for $h = 1, \ldots, H$. The covariate X is standardized to have mean zero and unit covariance, denoted by $Z = \Sigma_{xx}^{-1/2}(X - EX)$. The sample version is $z_i = \hat{\Sigma}_{xx}^{-1/2}(x_i - \bar{x})$, for $i = 1, \ldots, n$, where $\bar{x}$ and $\hat{\Sigma}_{xx}$ are the sample mean and sample covariance matrix, respectively.

Under certain conditions, the largest eigenvector of the covariance matrix of the inverse regression curve $E(Z|T\})$ lies in the direction of $\omega_0^T \Sigma_{xx}^{1/2}$. This follows from the theoretical properties of Sliced Inverse Regression established in Li (1991). To see this, we first need Condition 3.1 in Li (1991):

For any $b \in \mathbb{R}^p$, the conditional expectation $E(b^T X|\omega_0^T X) = c\omega_0^T X$ for some constant c.

Under this assumption, Theorem 3.1 in Li (1991) guarantees

$$E(Z|T) \text{ falls into the space generated by } \omega_0^T \Sigma_{xx}^{1/2}. \tag{6}$$

To estimate the inverse regression curve $E(Z|T)$ and its covariance, we can slice on the variable T as is done in SIR. However, Wei and Kosorok (2013) demonstrated that there are certain advantages of slicing on both the variable T and the variable $1\{\omega^T X - \gamma \geq 0\}$ over slicing on T alone for the type of model we study here. Let $V(\omega, \gamma)$ be the weighted covariance matrix of $E(Z|T \in I_h, 1\{\omega^T X - \gamma \geq 0\})$ given by

$$V(\omega, \gamma) = \sum_{h=1}^{H} p_{h,1}(\omega, \gamma) m_{h,1}(\omega, \gamma) m_{h,1}(\omega, \gamma)^T + \sum_{h=1}^{H} p_{h,2}(\omega, \gamma) m_{h,2}(\omega, \gamma) m_{h,2}(\omega, \gamma)^T, \tag{7}$$

where

$$m_{h,1}(\omega, \gamma) = E(Z|T \in I_h, \omega^T X - \gamma \geq 0)$$

and

$$p_{h,1}(\omega, \gamma) = P(T \in I_h, \omega^T X - \gamma \geq 0),$$

and similarly for $m_{h,2}(\omega,\gamma)$ and $p_{h,2}(\omega,\gamma)$ where the inequality is switched. Note that

$$m_{h,1}(\omega,\gamma) = E(Z|T \in I_h, 1\{\omega^T X - \gamma \geq 0\}) = E(E(Z|T,\omega^T X)|T \in I_h, 1\{\omega^T X - \gamma \geq 0\})$$

for any $\omega \in \mathbb{S}^d$ and $\gamma \in \mathbb{R}$ such that $P(\omega^T X - \gamma \geq 0) > 0$. Thus it follows from the properties of SIR that the largest eigenvector of $V(\omega,\gamma)$ is in the direction of $\omega_0^T \Sigma_{xx}^{1/2}$ under Condition (6).

The estimation of the various components in $V(\omega,\gamma)$ in Equation (7) is complicated by the fact that T may not have been observed. Censoring must be handled carefully to construct an unbiased estimate of the weighted covariance matrix. In what follows, a weight function w is used to adjust for the presence of censoring. A method called RIST is used to estimate w (Zhu and Kosorok, 2012). Further detail on RIST is given in Web Appendix A. The detailed derivation of the following estimate for $V(\omega,\gamma)$ is given in Web Appendix A. Here we give its final expression:

$$\hat{V}_n(\omega,\gamma) = \sum_{h=1}^{H} \hat{p}_{h,1}\hat{m}_{h,1}(\omega,\gamma)\hat{m}_{h,1}(\omega,\gamma)' + \sum_{h=1}^{H} \hat{p}_{h,2}\hat{m}_{h,2}(\omega,\gamma)\hat{m}_{h,2}(\omega,\gamma)' \tag{8}$$

where

$$\begin{aligned}\hat{m}_{h,1}(\omega,\gamma) = \frac{1}{n\hat{p}_{h,1}(\omega,\gamma)} \sum z_i \{ & 1\{t_h \leq y_i \leq t_{h+1}, \omega^T x_i - \gamma_0 \geq 0\} \\ & + \hat{w}(y_i, t_h, x_i) 1\{y_i < t_h, \delta_i = 0, \omega^T x_i - \gamma_0 \geq 0\} \\ & - \hat{w}(y_i, t_{h+1}, x_i) 1\{y_i < t_{h+1}, \delta_i = 0, \omega^T x_i - \gamma_0 \geq 0\}\}\end{aligned}$$

and

$$\begin{aligned}\hat{p}_{h,1}(\omega,\gamma) = n^{-1} \sum & 1\{t_h \leq y_i \leq t_{h+1}, \omega^T x_i - \gamma_0 \geq 0\} \\ & + \hat{w}(y_i, t_h, x_i) 1\{y_i < t_h, \delta_i = 0, \omega^T x_i - \gamma_0 \geq 0\} \\ & - \hat{w}(y_i, t_{h+1}, x_i) 1\{y_i < t_{h+1}, \delta_i = 0, \omega^T x_i - \gamma_0 \geq 0\}.\end{aligned}$$

The expressions for $\hat{m}_{h,2}$ and $\hat{p}_{h,2}$ can be found by switching the inequality in the indicator functions.

Let $\hat{\nu}_n(\omega,\gamma)$ be the largest eigenvector of $\hat{V}_n(\omega,\gamma)$. The individual components in $\hat{V}_n(\omega,\gamma)$ are uniformly consistent and thus $\hat{V}_n(\omega,\gamma)$ itself is uniformly consistent for $V(\omega,\gamma)$. It then follows that $\hat{\nu}_n(\omega,\gamma)$ is consistent for $\nu(\omega,\gamma)$, the largest eigenvector of $V(\omega,\gamma)$.

The updated sieve $\hat{\Omega}_n$ is formed by applying $\hat{\nu}_n$ to the simple sieve of Mean Difference directions. It is given by

$$\hat{\Omega}_n := \left\{ \hat{\nu}_n(\omega^{MD}(P), \gamma^{MD}(P))\hat{\Sigma}_{xx}^{-1/2} : P \in \bigcup_{k=1}^{K} \mathcal{P}_k \right\}. \tag{9}$$

where $\gamma^{MD}(P)$ is the intercept that maximizes the profile likelihood $M_n(\omega, \gamma)$ given $\omega^{MD}(P)$. The term $\hat{\Sigma}_{xx}^{-1/2}$ is necessary to transform the estimate back to the original scale.

The sieve estimate based on the RIST adjustment will be called the *sieve RIST* estimator. In Web Appendix B, we outline an alternative way to estimate $V(\omega, \gamma)$ using the Inverse Probability of Censoring Weighting (IPCW). Our motivation for this is the successful application of IPCW in Nadkarni et al. (2011) to estimate SIR-like directions in a regression setting. Following their procedure, the IPCW is estimated using a kernel conditional Kaplan Meier estimate. The sieve estimate based on the IPCW will be called the *sieve IPCW* estimator.

Finally we note that experience indicates the sieve estimation is not sensitive to the choice of H, the number of slices; setting $H = n/10$ works well in most applications. The computational complexity of the sieve estimate almost completely reduces to the choice of K, the number of regions the covariate space is partitioned into in the simple sieve.

3 Consistency

In this section, the sieve estimator is shown to be consistent. This is done by applying Theorem 14.1 (Argmax Theorem) in Chapter 14 of Kosorok (2008). The following list of assumptions is needed:

A1 The intercept γ_0 is known to lie in a bounded interval $[a, b]$.

A2 The log hazards ratio β_0 is non-zero.

A3 For any $b \in \mathbb{R}^p$ the conditional expectation $E(bX|\omega_0^T X)$ is linear in $\omega_0^T X$.

A4 The variable $\omega_0^T X$ has a strictly bounded and positive density f over $[a, b]$ with $P(\omega_0^T X < a) > 0$ and $P(\omega_0^T X > b) > 0$

A5 The covariate X has a continuous distribution.

A6 $P(C = 0) = 0, P(C \geq \tau|X) = P(C = \tau|X) > 0$, almost surely for some $0 < \tau < \infty$, and censoring is independent of T given X.

The interval $[a, b]$ in A1 may be estimated from the data by first calculating the direction of maximal variation of the sample covariates X, and next considering the range of the resulting projections. Assumption A2 ensures the model is identifiable. Li showed in the original SIR paper (Li, 1991) that A3 is necessary to ensure the consistency of SIR. Any elliptical

distributions, the Gaussian in particular, satisfy the condition. Condition A4 proved useful in establishing consistency of the sieve estimator for the Gaussian Latent Supervised Learning model studied in Wei and Kosorok (2013). Condition A5 may be relaxed at the cost of more complicated proofs. Condition A6 is standard in survival analysis.

Theorem 1 (Consistency)**.** *Let* $(x_1, y_1, \delta_1), \ldots, (x_n, y_n, \delta_n)$ *be iid from Model* (1)*. Under A1 – A6, the sieve estimator* $(\hat{\omega}_n^s, \hat{\gamma}_n^s)$ *is consistent for* (ω_0, γ_0)*.*

Proof. Let $\mathbb{P}$ denote the probability measure of $Z = (X, Y, \delta)$ under Model (1). Define the empirical measure to be $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{Z_i}$ where δ_z is the measure that assigns mass 1 at z and zero elsewhere. For a measurable function f, we denote $\mathbb{P}_n f = n^{-1} \sum_{i=1}^n f(Z_i)$ and $\mathbb{P} f = \int f \, dP$. Using the empirical processes notation described above, the profile log proportional hazards likelihood $M_n(\omega, \gamma)$ in Section 2.1 can be rewritten as

$$M_n(\omega, \gamma) = \mathbb{P}_n \delta \{ \hat{\beta}_n(\omega, \gamma) 1\{\omega^T X - \gamma \geq 0\} - \log F_n(Y, \omega, \gamma, \hat{\beta}_n(\omega, \gamma)) \}, \tag{10}$$

where

$$F_n(t, \omega, \gamma, \beta) = \mathbb{P}_n Y(t) \exp(\beta 1\{\omega^T X - \gamma \geq 0\})$$

and

$$Y(t) = 1\{Y \geq t\}.$$

The related population quantities are now defined. First, let $\tilde{Z} = (\tilde{X}, \tilde{Y}, \tilde{\delta})$ be an independent realization of Z and let $\tilde{\mathbb{P}}$ be a copy of the probability measure $\mathbb{P}$. The theoretical log proportional hazards likelihood $L(\omega, \gamma, \beta)$ is given by

$$L(\omega, \gamma, \beta) = \tilde{\mathbb{P}} \tilde{\delta} \{ \beta 1\{\omega^T \tilde{X} - \gamma \geq 0\} - \log \mathbb{P} Y(\tilde{Y}) \exp(\beta 1\{\omega^T X - \gamma \geq 0\}) \}. \tag{11}$$

Given ω and γ, the population log hazard ratio is defined to be $\beta(\omega, \gamma) = \arg\max_\beta L(\omega, \gamma, \beta)$. The theoretical profile likelihood is given by

$$M(\omega, \gamma) = L(\omega, \gamma, \beta(\omega, \gamma))$$

which can be written in empirical processes notation to mirror expression (10):

$$M(\omega, \gamma) = \mathbb{P} \delta \{ \beta(\omega, \gamma) 1\{\omega^T X - \gamma \geq 0\} - \log F_0(Y, \omega, \gamma, \beta(\omega, \gamma)) \} \tag{12}$$

where

$$F_0(t, \omega, \gamma, \beta) = \mathbb{P} Y(t) \exp(\beta 1\{\omega^T X - \gamma \geq 0\}).$$

Following Theorem 14.1 (Argmax Theorem) in Kosorok (2008), the following conditions must be satisfied to obtain consistency: 1) The sequence $(\hat{\omega}_n^s, \hat{\gamma}_n^s)$ is uniformly tight; 2) The

map $(\omega, \gamma) \mapsto M(\omega, \gamma)$ is upper semi-continuous with a unique maximum at (ω_0, γ_0); 3) M_n converges to M uniformly over compact subsets K of $\mathbb{S}^p \times [a, b]$; and 4) The sieve estimator nearly maximizes the objective function, i.e.,

$$M_n(\hat{\omega}_n^s, \hat{\gamma}_n^s) \geq M_n(\omega_0, \gamma_0) - o_P(1).$$

Each of these conditions is checked in turn:

1) The first condition is easily seen to hold since $||\hat{\omega}_n^s|| = 1$ and $\hat{\gamma}_n^s$ is constrained to lie in the interval $[a, b]$.

2a) The profile likelihood $M(\omega, \gamma)$ will actually be shown to be continuous. Let (ω_n, γ_n) be a sequence converging to (ω, γ). Since X is continuous by A5, we have

$$\begin{aligned}
&|\mathbb{P}\delta 1\{\omega_n^T X - \gamma_n \leq 0\} - \delta 1\{\omega^T X - \gamma \geq 0\}| \\
&\leq \mathbb{P}\delta |1\{\omega_n^T X - \gamma_n \leq 0\} - \delta 1\{\omega^T X - \gamma \geq 0\}| \\
&= \mathbb{P}\delta |1\{\omega_n^T X - \gamma_n \leq 0\} - \delta 1\{\omega^T X - \gamma \geq 0\}| 1\{|\omega_n^T X - \gamma_n - \omega^T X - \gamma_0| \leq \epsilon\} \\
&+ \delta |1\{\omega_n^T X - \gamma_n \leq 0\} - \delta 1\{\omega^T X - \gamma \geq 0\}| 1\{|\omega_n^T X - \gamma_n - \omega^T X - \gamma_0| > \epsilon\} \\
&\to 0
\end{aligned}$$

If $\beta(\omega_n, \gamma_n) \to \beta(\omega, \gamma)$ then $F_0(t, \omega_n, \gamma_n, \beta(\omega_n, \gamma_n)) \to F_0(t, \omega, \gamma, \beta(\omega, \gamma))$ almost surely. Note that $F_0(t, \omega_n, \gamma_n, \beta(\omega_n, \gamma_n)) \leq \max(\exp(beta), 1)\mathbb{P}Y(t)$ and is thus bounded by an integrable function under A6. This gives $\mathbb{P}\delta \log F_0(Y, \omega_n, \gamma_n, \beta(\omega_n, \gamma_n) \to \mathbb{P}\delta \log F_0(Y, \omega, \gamma, \beta(\omega, \gamma))$. Thus to show $M(\omega, \gamma)$ is continuous, the continuity of $\beta(\omega, \gamma)$ must be established.

We next show $\beta(\omega, \gamma)$ is continuous. First it is easy to see $L(\omega, \gamma, \beta)$ is continuous with respect to ω, γ, and β using the arguments above. Next we establish $L(\omega, \gamma, \beta)$ has a unique maximum in the β argument for every pair (ω, γ). Consider the partial derivative of L with respect to β

$$\frac{dL}{d\beta} = \tilde{\mathbb{P}}\tilde{\delta}\left\{1\{\omega^T \tilde{X} - \gamma \geq 0\} - \frac{\mathbb{P}Y(\tilde{Y})\exp(\beta 1\{\omega^T X - \gamma \geq 0\})1\{\omega^T X - \gamma \geq 0\}}{\mathbb{P}Y(\tilde{Y})\exp(\beta 1\{\omega^T X - \gamma \geq 0\})}\right\}$$

A straightforward calculation shows the second partial derivative with respect to β is strictly less than 0. Thus combined with the continuity of $L(\omega, \gamma, \beta)$, we get $\beta(\omega_n, \gamma_n) \to \beta(\omega, \gamma)$.

2b) In the full likelihood, replace $\lambda(t)$ with $\lambda_s(t) = (1 + sh(t))\lambda(t)$, h is for now an unspecified bounded function, and take the derivative of the full likelihood with respect to s (the Gateaux derivative). Let $N(t) = 1\{Y \leq t, \delta = 0\}$ be the counting process.

Using the fact that $PdN(t) = P\left[Y(t)\exp\left(\beta_0 1\{\omega_0' X - \gamma_0 \leq 0\}\right)\right]$, we obtain that the resulting derivative is:

$$\int_0^\tau h(t)P[Y(t)b(X,\theta_0)]d\Lambda_0(t) - \int_0^\tau h(t)P[Y(t)b(X,\theta)]d\Lambda(s) \tag{13}$$

where $b(X,\theta) = \exp\left(\beta 1\{\omega' X - \gamma \leq 0\}\right)$, $\theta = (\beta, \omega, \gamma)$ and $\theta_0 = (\beta_0, \omega_0, \gamma_0)$. Now if we replace Λ in 13 with $\Lambda_s(t) = \int_0^t (1 + sg(u))d\Lambda(u)$, for some other function g, and differentiate with respect to s again, we obtain the second Gateaux derivative is

$$-\int_0^\tau h(t)g(t)P[Y(t)b(X,\theta)]d\Lambda(t)$$

which is strictly negative, implying that for fixed θ, any Λ which satisfies 13 for a rich enough collection of h is a maximizer over all Λ for fixed θ. Choose $h(t) = 1\{t \leq u\}$, plug into 13, and allow u to range over $[0, \tau]$, and we obtain that the profile maximizer of the full expected log-likelihood over Λ satisfies

$$\int_0^u P[Y(t)b(X,\theta_0)]d\Lambda_0(t) - \int_0^u P[Y(t)b(X,\theta)]d\Lambda(t) = 0, \text{ for all } u \in [0,\tau].$$

Hence

$$\frac{d\Lambda(t)}{d\Lambda_0(t)} = \frac{P[Y(t)b(X,\theta_0)]}{P[Y(t)b(X,\theta)]}.$$

Plugging this back into the full likelihood, and removing additive terms which are constants with respect to θ, we obtain that the profile log-likelihood is

$$P\left[\int_0^\tau \left(b(X,\theta) - \log(P[Y(t)b(X,\theta)])\right) dN(t)\right], \tag{14}$$

which is equal to the expected log-likelihood in Equation (11).

Now suppose θ_1 maximizes 14. Then, by the fact that 14 is the profile log-likelihood, there exists a Λ_1 such that the joint parameter (θ_1, Λ_1) maximizes the full likelihood. By the property of the Kullback-Leibler discrepancy and model identifiability, this implies that $\theta_1 = \theta_0$.

Hence 14 has a unique maximizer at θ_0.

3) First let $m_{\omega,\gamma,\beta}(x,y,\delta) = \delta(\beta 1\{\omega^T X - \gamma \leq 0\} - \log F_n(y,\omega,\gamma,\beta))$ and consider the class of functions $\{m_{\omega,\gamma,\beta}(x,y,\delta) : (\omega,\gamma,\beta) \in K\}$ where K is a compact subset of the

parameter space $\mathbb{S}^p \times [a,b] \times \mathbb{R}$. The argument that this class is Donsker and therefore also Glivenko-Cantelli is as follows. The at-risk process Y is Donsker by Lemma 4.1 in Kosorok (2008). Trivially, the class $\{\beta\}$ is Donsker. The class $\{1\{\omega^T X - \gamma \geq 0\}\}$ is also Donsker by way of the example in Section 4.1.1 in Chapter 4 of Kosorok (2008). Therefore the product $\{\beta 1\{\omega^T X - \gamma \geq 0\}\}$ is Donsker since products of bounded Donsker classes are Donsker. Now the class $\exp \beta 1\{\omega^T X - \gamma \geq 0\}$ is Donsker since exponentiation is Lipschitz continuous on compacts. Repeating these arguments shows $\log F_n(y, \omega, \gamma, \beta)$ is Donsker and hence the class $\{m_{\omega,\gamma,\beta}(x,y,\delta)$ is Donsker.

Now, let $m_{\omega,\gamma}(x,y,\delta) = \delta\{\hat{\beta}_n(\omega,\gamma)1\{\omega^T x - \gamma \geq 0\} - \log F_n(y,\omega,\gamma,\hat{\beta}_n(\omega,\gamma))\}$. The estimated log ratio hazard $\hat{\beta}_n(\omega,\gamma)$ lives in a compact set in $\mathbb{R}$ for all ω, γ in compact K, and thus the class $\{m_{\omega,\gamma}(x,y,\delta)\}$ is contained in a Donsker class which implies it is also a Glivenko-Cantelli class. Writing $M_n(\omega,\gamma) = \mathbb{P}_n m_{\omega,\gamma}(x,y,\delta)$, we have

$$\sup_{\omega \in \mathbb{S}^p, \gamma \in \mathbb{R}} |M_n(\omega,\gamma) - \mathbb{P}m_{\omega,\gamma}(x,y,\delta)| \to 0$$

in probability. The uniform convergence of $\hat{\beta}_n(\omega,\gamma)$ to $\beta(\omega,\gamma)$ and $\hat{F}_n(t,\omega,\gamma)$ to $F_0(t,\omega,\gamma)$ follows from standard arguments. Thus we have $\mathbb{P}m_{\omega,\gamma}(x,y,\delta)$ converges uniformly to $M(\omega,\gamma) = \mathbb{P}\delta\{\beta(\omega,\gamma)1\{\omega^T X - \gamma \geq 0\} - \log F_0(Y)\}$ over $\mathbb{S}^p \times \mathbb{R}$. Finally, this gives $M_n(\omega,\gamma)$ converging uniformly to $M(\omega,\gamma)$ over $\mathbb{S}^p \times \mathbb{R}$.

4) The main text in Section 2.3 already established the sieve $\hat{\Omega}_n$ is dense. In other words, there exists a sequence $(\omega_n, \gamma_n) \in \hat{\Omega}_n \times [a,b]$ that converges to ω_0, γ_0. By definition we have

$$M_n(\hat{\omega}_n^s, \hat{\gamma}_n^s) \geq M_n(\omega_n, \gamma_n).$$

By the continuity of $M(\omega,\gamma)$, we have $M_n(\omega_0,\gamma_0) - M_n(\omega_n,\gamma_n) = o_P(1)$ and thus

$$M_n(\hat{\omega}_n^s, \hat{\gamma}_n^s) \geq M_n(\omega_0,\gamma_0) - o_P(1).$$

Note the conditions of the Argmax theorem are met, and the desired consistency follows.

□

4 Simulations

In this section we examine the performance of the sieve RIST estimator, the sieve IPCW estimator, and Li's double-slicing method. Since Li's double-slicing method produces a direction estimate only, the profile likelihood $M_n(\omega,\gamma)$ is used to estimate an intercept in

order to make a direct performance comparison with the sieve estimators. The specific tuning parameters used for each of the methods, necessary to reproduce the simulation results below, are given in Web Appendix C.

Let T_1 be the distribution of T when $\omega_0^T X - \gamma_0 \geq 0$ and T_2 be the distribution of T when $\omega_0^T X - \gamma_0 < 0$. The following distributions are considered for T_1 and T_2: 1) exponential distributions satisfying proportional hazards (Exponential PH), 2) Weibull distributions satisfying proportional hazards (Weibull PH), and 3) Weibull distributions satisfying the accelerated failure time model (Weibull AFT). Let $exp(\lambda)$ denote the exponential distribution with mean $1/\lambda$. Let $Weibull(\lambda, \nu)$ denote the Weibull distribution with scale parameter $\lambda > 0$ and shape parameter $\nu > 0$ where the density function is given by $f(t) = \lambda\nu t^{\nu-1}\exp(-\lambda t^{\nu})$. The specific survival time distributions considered are

1. Exponential PH: $T_1 \sim exp(\lambda\exp(\beta))$ and $T_2 \sim exp(\lambda)$.
2. Weibull PH: $T_1 \sim weibull(\lambda\exp(\beta), \nu)$ and $T_2 \sim weibull(\lambda, \nu)$.
3. Weibull AFT: $T_1 \sim weibull(\lambda\exp(\nu\beta), \nu)$ and $T_2 \sim weibull(\lambda, \nu)$.

The survival parameters are set to $\lambda = 1/10, \beta = \log 10$, and $\nu = 2$.

Various censoring mechanisms are also considered: 1) independent – censoring completely independent of X, 2) linear – censoring dependent on X only through $\omega_0^T X - \gamma_0$, and 3) nonlinear – censoring dependent on X in a non-linear manner. Specifically, the three censoring distributions considered are

1. independent: $C \sim unif(0, \tau)$
2. linear: $C \sim \min(unif(0, \tau_1), a)1\{\omega_0^T X - \gamma_0 \geq 0\} + \min(unif(0, \tau_2), b)1\{\omega_0^T X - \gamma_0 < 0\}$
3. nonlinear: $C \sim exp(ae^{X_1 + X_2^2 + \log|X_3|})$

The parameters $\tau, \tau_1, \tau_2, a, b$ are set to different values for each survival model considered. Their values, along with the overall censoring percentage for each censoring and survival setting, are given in Web Appendix C.

The covariate vector X is generated from the standard p-variate Gaussian distribution. The first $p/2$ components of ω_0 are set to $-p^{-1/2}$ and the rest to $p^{-1/2}$. The intercept γ_0 is set to $1/4$ which results roughly in a 60/40 split of the data. The sample size is fixed at 100. Four dimensions are considered $p = 5, 10, 25, 50$. The average classification error over 100 Monte Carlo simulations is reported for each of the three methods. The error rate is obtained by generating a large independent test set and calculating the number of

misclassifications resulting from the estimated hyperplane. The average angle between the estimate and the true direction ω_0 and the average intercept estimate error can be found in Web Appendix D.

Web Appendix D also reports other performance measurements include the the average angle between the estimated direction and the true direction ω_0 and the distance between the estimated intercept and the true intercept.

Exponential proportional hazards Figure 1 shows the performance of the three methods for the Exponential PH survival setting. The average classification error rate is given as a function of the dimension of X. We see that Li's method performs worse than the RIST sieve estimator in the independent censoring setting. This is expected as there is no benefit to slicing on the censoring variable. We also see that the sieve IPCW performs the worst in high dimensions.

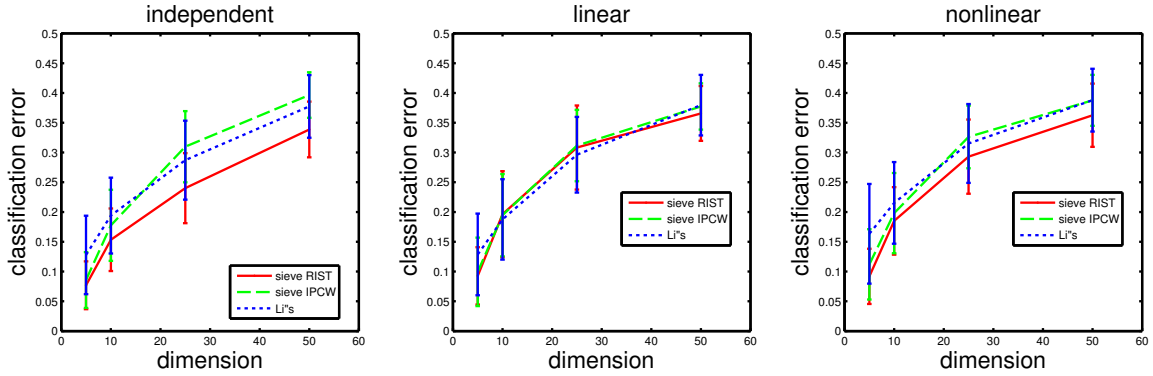


Figure 1: Exponential proportional hazards – classification error as a function of dimension for 100 Monte Carlo simulations with error bars for each of the three censoring mechanisms considered – independent, linear, and nonlinear.

In the linear censoring setting, all methods perform similarly. Since the linear censoring setting satisfies Li's double-slicing censoring assumption, we do not expect the sieve RIST method to provide a substantial improvement. In the nonlinear censoring setting, the censoring time cannot be written as a function of a linear combination of the covariates. Hence, it does not satisfy Li's double-slicing assumption. We see that Li's estimate is quite sensitive to departure from this assumption. The sieve RIST estimate outperforms Li's estimate and the sieve IPCW estimator. Note that under the nonlinear censoring mechanism, the estimation of the individual slice means must adjust for the presence of censoring unlike in the independent censoring setting. This is because, for instance, when

x_1, x_2 and x_3 are large, the censoring time tends to be small. Thus if the slice means were estimated using the sample mean of observed survival times, the estimate would be biased in favor of small values of x_1, x_2 and x_3.

Weibull proportional hazards Figure 2 shows the performance of the three methods for the Weibull PH survival setting. In the independent censoring setting, the sieve RIST estimate is more accurate than the other two methods. In the linear censoring setting, all three methods perform similarly. In the nonlinear censoring setting, the sieve RIST estimate performs the best as expected. The performance patterns of the three methods for the Weibull PH survival setting is quite similar to those in the preceding Exponential PH setting. One interesting difference is the sieve IPCW actually outperforms Li's estimate in the nonlinear censoring setting here.

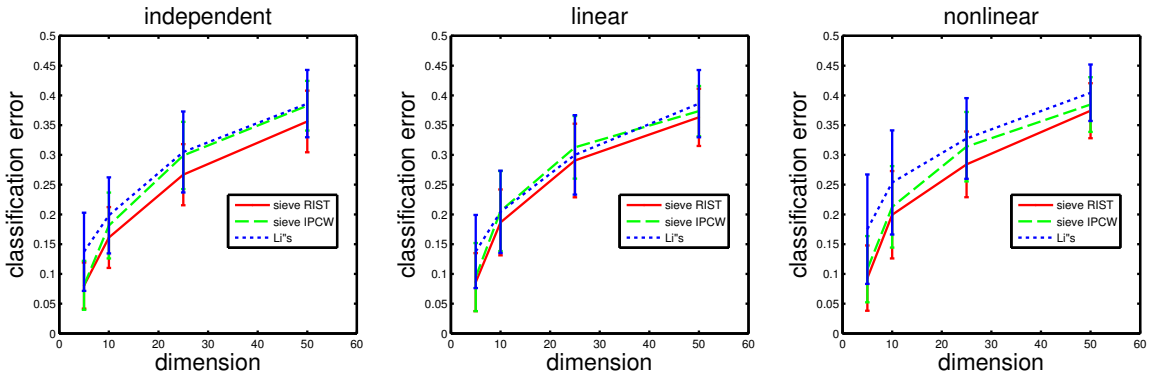


Figure 2: Weibull proportional hazards – classification error as a function of dimension for 100 Monte Carlo simulations with error bars for each of the three censoring mechanisms considered – independent, linear, and nonlinear.

Weibull accelerated failure time Figure 3 shows the performance of the three methods for the Weibull AFT survival setting. The Weibull AFT model violates the assumption of proportional hazards in Model (1) and thus provides an opportunity to assess the robustness of the sieve estimators under distributional departures. In all three censoring settings, the sieve IPCW performs the worst across all dimensions while the sieve RIST estimate and Li's double slicing estimate perform very similarly across all three censoring settings. Li's double-slicing method makes no distributional assumptions unlike our sieve methods. Fortunately, the sieve RIST estimate seems to be quite robust to departure from the proportional hazards assumption.

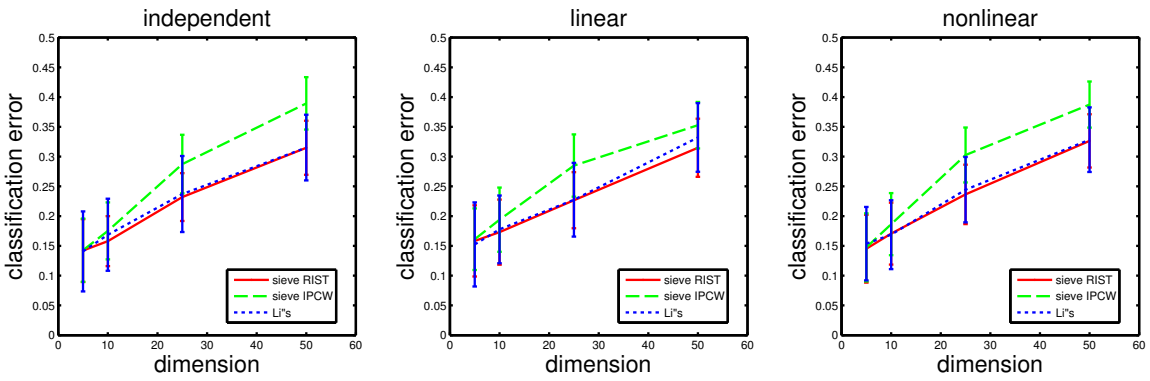


Figure 3: Weibull accelerated failure time – classification error as a function of dimension for 100 Monte Carlo simulations with error bars.

The simulations in this section suggest that the sieve RIST estimator performs similarly to the Li's double-slice method when Li's censoring assumption is satisfied (as in the linear censoring setting) but outperforms the double-slicing method when the assumption is not satisfied (as in the nonlinear censoring setting) or satisfied trivially (as in the independent censoring setting). The simulations also suggest that the sieve RIST method is more accurate than the sieve IPCW method, especially for higher dimensions. In addition, the sieve RIST estimate which assumes the proportional hazards model is seen to be robust when the survival distribution actually follows the accelerated failure time model. Finally, the simulations reveal that the performance of all three methods deteriorate in high dimensions. Regularization may be promising for extending the proposed method to high dimensions.

5 Examples

The sieve RIST method and Li's double-slice method is demonstrated on two datasets. The sieve IPCW method is omitted in the following analysis because the simulation findings revealed the sieve RIST estimator consistently outperformed it. The tuning parameters used in the sieve RIST estimate and Li's double-slice estimate are the same as those used in the simulations section, they are given in Web Appendix C.

5.1 Diffuse large B-cell lymphoma

Diffuse large B-cell lymphoma (DLBCL) is a cancer of white blood cells and the most common type of non-Hodgkin lymphoma among adults. In Rosenwald et al. (2002), hierarchical clustering based on gene expression was used to identify three DLBCL subtypes: 1)

Germinal-center B-cell-like (GBC), 2) Activated B-cell-like (ABC), and Type 3. Although these subgroups are biologically meaningful, there is no guarantee that their survival experiences differ. Indeed, Figure 4a, which displays the estimated survival function in each subgroup and the log-rank test p-values, reveals there is little difference between the survival experiences of the ABC subgroup and the GBC subgroup.

The DLBCL data was obtained from

`http://llmpp.nih.gov/DLBCL/DLBCL_patient_data_NEW.txt`

and consists of 240 patients with 138 patient deaths at follow-up, resulting in 42% censoring. There are five covariates in the data that are used in the analysis here. The covariates include the average values of four gene expression signatures. The signatures are 1)Germinal center B cell, 2)Lymph node, 2)Proliferation, and 4) MHC class II. The fifth covariate is the gene expression value of the BMP6 gene. Removing cases with missing values leaves 222 cases, of which 73 are randomly set aside for a test set. The remaining 149 cases comprise the training set. The training set has 47% censoring, and the test set 34% censoring.

The sieve RIST estimator and Li's double-slicing estimator are now applied to the DLBCL dataset. Figures 4b and 4c show the estimated survival functions of the subgroups discovered by each method and the estimated log hazards ratio β in the DLBCL training set. The p-value for the log-rank test $\beta = 0$ is also displayed. Both methods identify subgroups that are significantly different with respect to survival with the sieve RIST estimate producing a larger (in magnitude) log hazard ratio and more significant p-value. As a model diagnostic, we apply the sieve RIST and Li's double-slicing estimates to the held out test set. Figure 5 shows the estimated survival functions of the subgroups identified by each method on the DLBCL test set. The p-value is very significant for the subgroups produced by the sieve RIST estimate. On the other hand, the subgroups identified by Li's estimate are not significant at the 0.01 level. Thus, the sieve RIST estimator was capable of finding a more significant separation in survival in both the training and test sets than Li's estimate. Furthermore, both methods identify subgroups of patients that are both biologically and clinically relevant as compared to Rosenwald's discovered subgroups.

5.2 Primary Biliary Cirrhosis

Primary Biliary Cirrhosis (PBC) is an autoimmune disease of the liver. The data studied here is from the Mayo Clinic trial in PBC of the liver conducted between 1974 and 1984. This dataset has been extensively studied in survival analysis. The PBC data was obtained from

http://mayoresearch.mayo.edu/mayo/research/biostat/upload/therneau_upload/pbc.html

The first 312 cases in the data set participated in a randomized trial and contain largely complete data. This will form our training set. The additional 112 did not participate in the clinical trial but have basic measurements recorded and were followed for survival. This will form our test set. After removing cases with missing data, we have 308 patients in the training set with 60% censoring and 91 patients in the test set with 71% censoring. We conduct our analysis with the following eight covariates: 1) age, 2) sex, 3) edema, 4) bili, 5) albumin, 6) platelet, 7) protime, and 8) stage.

The sieve RIST method and Li's double-slicing method are now applied to the PBC dataset. For the PBC training set, Figure 6 shows the estimated survival functions of the subgroups discovered by each method and the p-value for the log rank test $\beta = 0$. Both methods identify subgroups that are significantly different with respect to survival, with the sieve RIST estimate producing subgroups with larger (in magnitude) log hazards ratio and a more significant p-value. As a model diagnostic, we examine the performance of these estimates on the held out test set. Figure 7 shows the estimated survival functions of the subgroups identified by each method on the PBC test set. The p-value for β is very significant for the subgroups produced by the sieve estimate. The subgroups identified by Li's estimate are less significant but still very significant. Overall, the sieve RIST estimator seems to be fulfilling the purpose for which it was designed, i.e. to find two subgroups with maximally different proportional hazards.

6 Discussion

Latent supervised learning is a machine learning technique for performing binary classification using a surrogate variable for the unobserved training label. The concept of latent supervised learning bridges the gap between unsupervised and supervised learning. It was applied here to tackle the problem of identifying two subgroups in the covariate space whose survival distributions have maximally different proportional hazards. The model considered is parsimonious and easy to interpret. The applicability of this method is immediate to areas such as drug discovery or personalized medicine where it is desirable to identify subgroups that are both clinically and biologically meaningful.

The simulations conducted here demonstrate the proposed method performs similarly to Li's double-slicing method when Li's censoring assumption is satisfied and outperforms Li's estimate when this assumption is violated. In the DLBCL dataset, the proposed method

produced subgroups that were significantly different with respect to survival on an independent test set while Li's estimate produced a less significant result. In the PBC data, both methods produced subgroups in the test set with significantly different survival, but the proposed method produced subgroups whose survival distributions had a larger log hazards ratio than Li's estimate.

Consistency for the estimation procedure was established. Performing inference for the parameters ω and γ, however, is still an open problem. As a model diagnostic in the data examples, we applied the estimated separating hyperplane to an independently held out test set and assessed the significance of the survival separation. This in some sense wastes part of the data as we have to set aside a test set to perform model diagnostics. A theoretical basis for performing inference would be preferable. Establishing the weak convergence of the sieve estimator is thus an important endeavor.

Model mis-specification is another important issue. For instance in the true underlying model there may be more than two subgroups. We are currently working on extending the methodology to an arbitrary number of subgroups using ideas from multi-class classification methods in machine learning. An interesting question is how to discover the correct number of groups. For this endeavor, we can borrow from the literature on how to determine the appropriate number of clusters in clustering analysis.

Another important extension of the work is to high dimensional low sample size data, i.e. when $p > n$. Currently covariance and inverse covariance estimation limit the application of the methodology to this setting. Extension of the methodology to ultra-high dimensions, i.e. when p is much bigger than n, would also be of practical interest. Some type of regularization of the coefficients in the linear combination may be promising.

References

Bair, E. and Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology*, 2(4):e108.

Banerjee, M. and Noone, A.-M. (2007). *Tree-Based Methods for Survival Data*, pages 265–285. John Wiley & Sons, Inc.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. (1984). *Classification and Regression Trees*. Wadsworth, New York.

Gama, J. (2004). Functional trees. *Machine Learning*, 55:219–250.

Grenander, U. (1981). *Abstract Inference*. Wiley, New York.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York, 1 edition.

Leblanc, M. and Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association*, 88(422):457–467.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316– 327.

Li, K.-C., Wang, J.-L., and Chen, C.-H. (1999). Dimension reduction for censored regression data. *The Annals of Statistics*, 27(1):pp. 1–23.

Nadkarni, N. V., Zhao, Y., and Kosorok, M. R. (2011). Inverse regression estimation for censored data. *Journal of the American Statistical Association*, 106(493):178–190.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., Hurt, E. M., Zhao, H., Averett, L., Yang, L., Wilson, W. H., Jaffe, E. S., Simon, R., Klausner, R. D., Powell, J., Duffey, P. L., Longo, D. L., Greiner, T. C., and It Et Al (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine*, 346:1937–1947.

Scholkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA.

Tian, L. and Tibshirani, R. (2011). Adaptive index models for marker-based risk stratification. *Biostatistics*, 12(1):68–86.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572.

Wei, S. and Kosorok, M. R. (2013). Latent supervised learning. *Journal of The American Statistical Association*, In press.

Wu, T., Sun, W., Yuan, S., Chen, C.-H., and Li, K.-C. (2008). A method for analyzing censored survival phenotype with gene expression data. *BMC Bioinformatics*, 9(1):417.

Zhu, R. and Kosorok, M. R. (2012). Recursively imputed survival trees. *Journal of The American Statistical Association*, 107(497):331–340.

A Estimation of $V(\omega, \gamma)$ in the sieve RIST estimator

We derive the estimation of the components of $V(\omega, \gamma)$ in Equation (7). The derivation is given for the individual slice mean $m_{h,1}$ from which it should be clear how to estimate the probabilities $p_{h,1}$. The individual slice mean can be written as

$$m_{h,1}(\omega, \gamma) = \frac{E(Z1\{T \geq t_j\}1\{\omega^T X - \gamma \geq 0\}) - E(Z1\{T \geq t_{j+1}\}1\{\omega^T X - \gamma \geq 0\})}{E(1\{T \geq t_j\}1\{\omega^T X - \gamma \geq 0\}) - E(1\{T \geq t_{j+1}\}1\{\omega^T X - \gamma \geq 0\})}$$

and similarly for $m_{h,2}(\omega, \gamma)$. Consider the following decomposition

$$\begin{aligned} &E(Z1\{T \geq t\}1\{\omega^T X - \gamma \geq 0\}) \\ &= E(Z1\{Y \geq t_j\}1\{\omega^T X - \gamma \geq 0\}) + E(Z1\{T \geq t, C < t\}1\{\omega^T X - \gamma \geq 0\}). \end{aligned}$$

The second term can be further expressed as follows

$$\begin{aligned} &E(Z1\{T \geq t, C < t\}1\{\omega^T X - \gamma \geq 0\}) \\ &= E(Z1\{Y \geq t, \delta = 0\}1\{T \geq t\}1\{\omega^T X - \gamma \geq 0\}) \\ &= E(Z1\{Y \geq t, \delta = 0\}1\{\omega^T X - \gamma \geq 0\}E(1\{T \geq t\}|Y, \delta = 0, Z)) \\ &= E(Z1\{Y \geq t, \delta = 0\}1\{\omega^T X - \gamma \geq 0\}E(1\{T \geq t\}|C, T > C, Z)) \\ &= E(Z1\{Y \geq t, \delta = 0\}1\{\omega^T X - \gamma \geq 0\}E(1\{T \geq t\}|C, T > Y, Z)) \\ &= E(Z1\{Y < t, \delta = 0\}1\{\omega^T X - \gamma \geq 0\}w(Y, t, Z)) \end{aligned}$$

where

$$w(Y, t, Z) = P(T \geq t|Z)/P(T \geq Y|Z)$$

is the weight adjustment for the presence of censoring. Let $\hat{w}$ be an estimate of w to be discussed below. Putting these pieces together leads to the expression in Equation (8).

To estimate the weight adjustment, it is necessary to estimate the conditional survival function of T given X. For this task, we choose a method called Recursively Imputed Survival Trees (RIST) (Zhu and Kosorok, 2012). RIST can be viewed as a type of Monte Carlo EM algorithm which generates extra diversity in the fitting process. Let $\hat{w}(y_i, t_h, z_i)$ be the estimated RIST weight for $i = 1, \ldots, n$ and $h = 1, \ldots, H + 1$.

Since RIST is known to be an unbiased estimator for the conditional survival function, the proposed estimate $\hat{V}_n(\omega, \gamma)$ is consistent for $V(\omega, \gamma)$.

B Estimating the IPCW

We consider an alternative way to estimate $V(\omega, \gamma)$ in Equation (7) using Inverse Probability of Censoring Weight (IPCW) as in Nadkarni et al. (2011). Let $\hat{m}_{h,1}^{IPCW}(\omega, \gamma)$ be the IPCW-

adjusted weighted average of the Z's associated with observed survival times in the h-th slice that are above the hyperplane $\omega^T x - \gamma \geq 0$:

$$\hat{m}_{h,1}^{IPCW}(\omega, \gamma) = \frac{\sum_{i=1}^{n} \delta_i \frac{z_i}{\hat{P}(C > y_i | z_i, \omega^T x_i - \gamma \geq 0)} 1\{y_i \in I_h\} 1\{\omega^T x_i - \gamma \geq 0\}}{\hat{p}_{h,1}(\omega, \gamma)}, \tag{15}$$

where

$$\hat{p}_{h,1}^{IPCW}(\omega, \gamma) = \sum_{i=1}^{n} \delta_i \frac{1}{\hat{P}(C > y_i | z_i, \omega^T x_i - \gamma \geq 0)} 1\{y_i \in I_h\} 1\{\omega^T x_i - \gamma \geq 0\}.$$

The expressions for $\hat{m}_{h,2}^{IPCW}(\omega, \gamma)$ and $\hat{p}_{h,2}^{IPCW}(\omega, \gamma)$ are similar.

The inverse probability censoring weight $\hat{P}(C > t|z)$ can be estimated as follows. Let $N_i^C(t)$ and $Y_i^C(t)$ denote the counting process and at-risk process respectively for the i-th observation: $N_i^C(t) = 1\{y_i \leq t, \delta_i = 0\}$ and $Y_i^C(t) = 1\{y_i > t\}$. The conditional censoring distribution $P(C > t|z)$ is estimated using a kernel conditional Kaplan Meier estimate:

$$\hat{P}(C > t|z) = \phi \left(- \int_0^t \frac{\sum_{i=1}^{n} K(||z - z_i||/h) \, dN_i^C(t)}{\sum_{j=1}^{n} K(||z - z_j||/h) Y_j^C(t)} \right). \tag{16}$$

Here ϕ is the product integral functional and K is a kernel function and h is the bandwidth. The integral on the right hand side of Equation (16) can be simplified as follows:

$$\begin{aligned}
\int_0^t \frac{\sum_{i=1}^{n} K(||z - z_i||/h) \, dN_i^C(t)}{\sum_{i=1}^{n} K(||z - z_i||/h) Y_i^C(t)} &= \sum_{i=1}^{n} \int_0^t \frac{K(||z - z_i||/h) \, dN_i^C(t)}{\sum_{j=1}^{n} K(||z - z_j||/h) Y_j^C(t)} \\
&= \sum_{i:\delta_i = 0} \int_0^t \frac{K(||z - z_i||/h) \, dN_i^C(t)}{\sum_{j=1}^{n} K(||z - z_j||/h) Y_j^C(t)} \\
&= \sum_{i:\delta_i = 0} \frac{K(||z - z_i||/h) 1\{Y_i \leq t\}}{\sum_{j=1}^{n} K(||z - z_j||/h) Y_j^C(Y_i)} \\
&= \sum_{i:\delta_i = 0, Y_i \leq t} \frac{K(||z - z_i||/h)}{\sum_{j:Y_j > Y_i} K(||z - z_j||/h)}
\end{aligned}$$

which gives

$$\hat{P}(C > t|z) = \prod_{i:\delta_i = 0, Y_i \leq t} 1 - \frac{K(||z - z_i||/h)}{\sum_{j:Y_j > Y_i} K(||z - z_j||/h)}. \tag{17}$$

It should now be clear how to derive the expressions for $\hat{P}(C > t|z, \omega^T x - \gamma \geq 0)$ and $\hat{P}(C > t|z, \omega^T x - \gamma < 0)$.

C Simulations

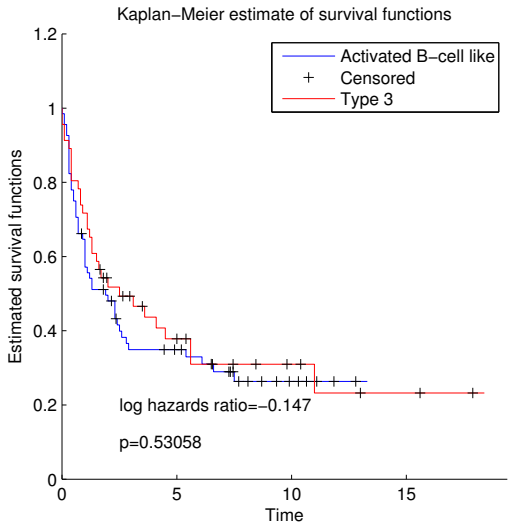
For the sieve RIST estimator, K-means clustering was used in the preliminary sieve where K is set to $n/10$ and the number of slices in the updated sieve H is also set to $n/10$. The

RIST procedure contains several tuning parameters including 1) the number of covariates considered per split, 2) the minimum number of observed data in each node, 3) the number of trees in each fold, and 4) the number of folds. We set these parameters to $5, 6, 50$ and 1, respectively. For the sieve IPCW estimator, K and H are also each set to $n/10$, and a standard Gaussian kernel is used in the IPCW estimation. For Li's double-slicing method, we used Wei Sun's default implementation available at

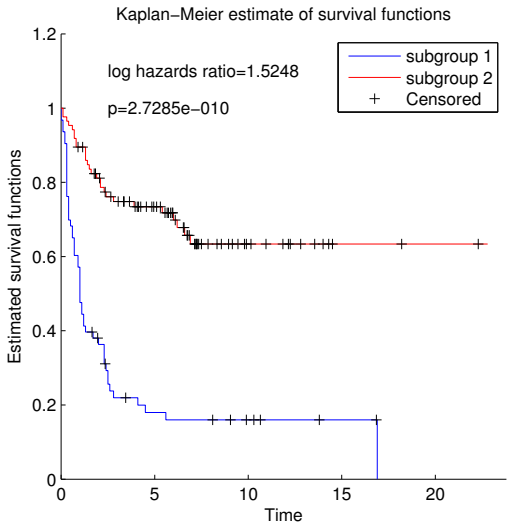
`http://www.bios.unc.edu/~weisun/software.htm`

For each censoring setting, the parameters (and censoring percentage) for the Exponential PH, the Weibull PH, and the Weibull AFT, are respectively

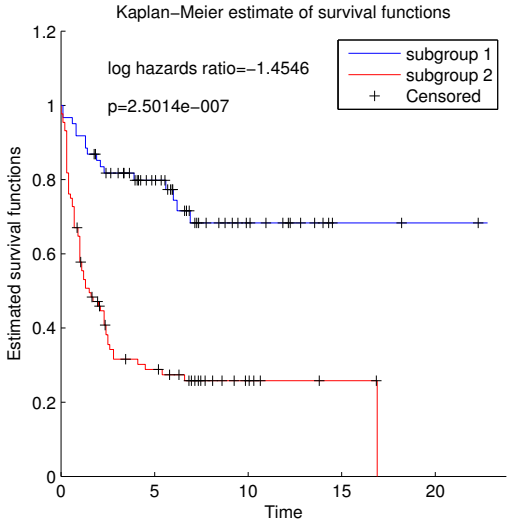
1. independent: $\tau = 10$ (42%); $\tau = 20$ (32%); $\tau = 10$ (58%)
2. linear: $\tau_1 = 31.97, a = 20, \tau_2 = 3.2, b = 2$ (32%); $\tau_1 = 30, a = 15, \tau_2 = 9, b = 5$ (34%); $\tau_1 = 15, a = 5, \tau_2 = 4, b = 2$ (40%)
3. nonlinear: $a = 1/10$ (40%); $a = 1/20$ (39%); $a = 1/20$ (55%)



(a) Rosenwald subgroups

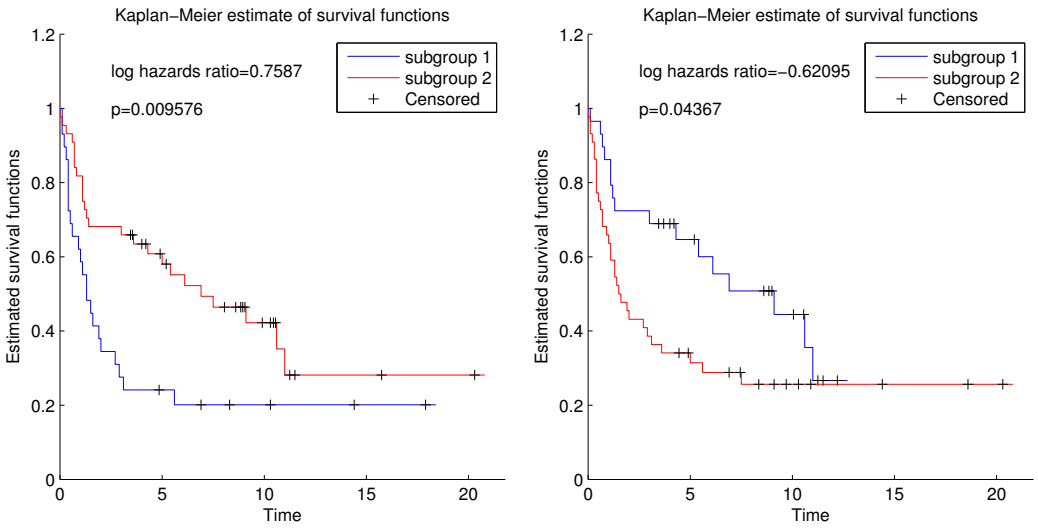


(b) Sieve RIST estimate



(c) Li's double-slicing SIR

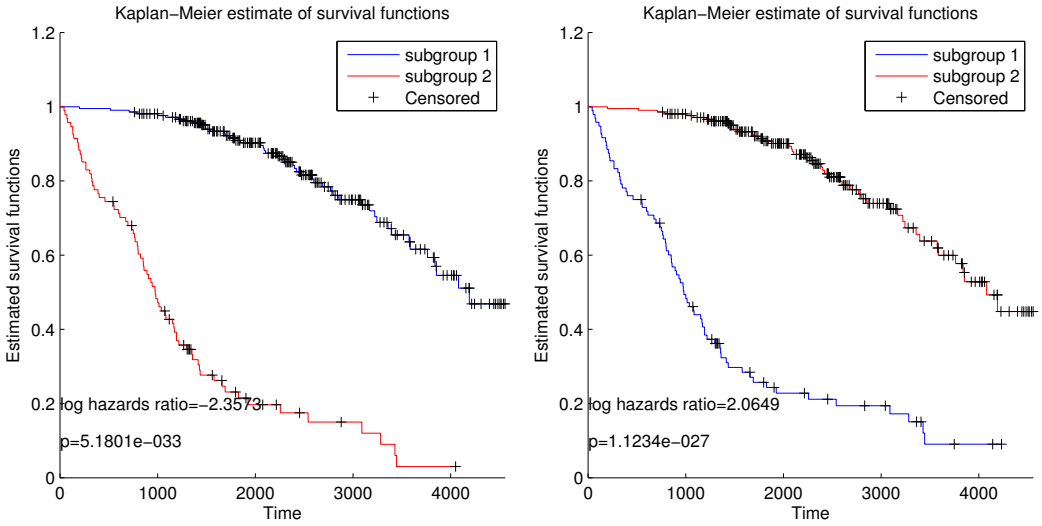
Figure 4: The first panel shows the estimated survival functions of the ABC and Type 3 subgroups. The subgroups are not significantly different with respect to survival. The middle and right panels regard the DLBCL training set and display the estimated survival functions of the subgroups identified in the training set, by the sieve estimate and double-slicing SIR estimate, respectively. Both estimates produce subgroups in the training set which are significantly different with respect to survival.



(a) Sieve RIST estimator

(b) Li's double-slicing

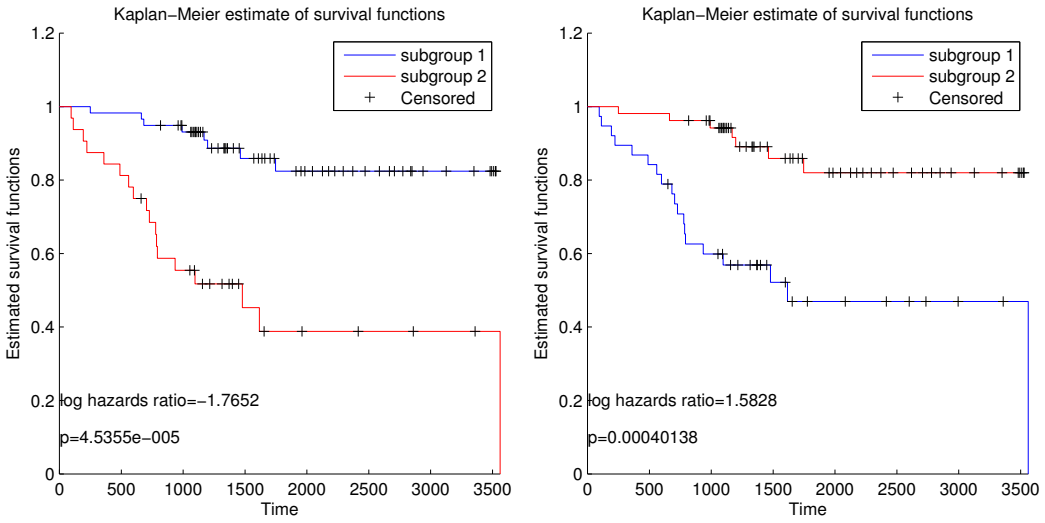
Figure 5: DLBCL test set. The left and right panels display the estimated survival functions of the subgroups identified in the test set, by the sieve RIST estimate and Li's double-slice estimate, respectively. The sieve RIST estimate produces subgroups in the test set which are more significantly different with respect to survival.



(a) Sieve RIST estimate

(b) Li's double-slicing SIR

Figure 6: PBC training set. The left and right panels display the estimated survival functions of the subgroups identified in the training set, by the sieve RIST estimate and Li's double-slicing estimate, respectively. Both pairs of subgroups are significantly different with respect to survival. The sieve RIST estimate, however, results in subgroups with a higher log hazards ratio (in terms of magnitude).



(a) Sieve RIST estimator

(b) Li's double-slicing

Figure 7: PBC test set. The left and right panels display the estimated survival functions of the subgroups identified in the test set, by the sieve RIST estimate and Li's double-slice estimate, respectively. Both pairs of subgroups are significantly different. The sieve RIST estimate, however, results in subgroups with a higher log hazards ratio (in terms of magnitude).